Guest editorial

Big data analytics and business process innovation

Big data analytics (BDA), which is defined as a “holistic process to manage, process and analyze 5 Vs (i.e., volume, variety, velocity, veracity and value) in order to create actionable insights for sustained competitive advantages” (Fosso Wamba et al., 2015), is now considered as a hot topic for both practitioners and scholars. BDA has a huge potential in transforming various organizations and industries seeking sustained competitive advantage (Davenport and Harris, 2007). While much has been written about the high operational and strategic potential of BDA, the literature dedicated to the empirical evidence of the real business value of BDA at the firm and supply chain levels on the one hand, and to the relationship between BDA and process innovation on the other hand, remains thin and even poor. The main objective of this special issue was to fill this knowledge gap in the relevant literature. Specifically, a good number of scholars and practitioners have been invited to explore the ways and means to identify and capture business value from BDA in terms of innovative business models, improved decision making, improved intra- and inter-organizational performance, and competitive advantage.

All articles included in this special issue went through a double-blind review process. The review of the introductory article was handled by the editor-in-chief of the journal.

Synopsis of articles

Here below, a synopsis of the articles included in this special issue is set out.

In the paper entitled “Big data integration with business processes: a literature review”, by Samuel Fosso Wamba and Deepa Mishra, the authors make a synthesis of studies published between 2006 and 2016 on the integration of business process management (BPM), business process re-engineering (BPR) and business process innovation (BPI) with big data. A total of 49 papers on big data, BPM, BPR and BPI from top journals in the retained period are profoundly reviewed. The most influential works were identified and retrieved based on the citations and PageRank methods. Through the network analysis, four major clusters providing potential opportunities for future investigation were identified. This paper ends up with the implications for both practitioners and scholars, and future research directions.

In the paper “Business intelligence serious game participatory development: lessons from ERPsim for big data”, by Elise Labonte-LeMoyne, Pierre-Majorique Leger, Jacques Robert, Gilbert Babin, Patrick Charland and Jean-François Michon, the authors present a serious business intelligence’s game participatory development by focusing on lessons from ERPsim for big data. They discuss educational methods to prepare business students to use business analytics tools when they reach the workplace. The authors introduce a new serious game and discuss how it was developed through collaborative efforts, together with a testing of an enactive approach to teach business intelligence.

This special issue was made possible thanks to the collaboration and active participation of several people. In the first place, the author would like to thank Professor Majed Al-Mashari, the Chief Editor of the journal, for having provided useful comments on the early version of the special issue proposal and for having accepted the final document. His strong support during the whole process was very much appreciated. Also, the author would like to acknowledge the active involvement of the whole editorial team, especially Emily Mitchelson and Claire Jackson. Finally, the author would like to thank the authors and reviewers without whom this special issue would have never been a reality.
concepts and provide hands-on experience to students in an authentic and dynamic business environment.

The paper titled “Examining the adoption of big data and analytics curriculum”, by Alexander J. McLeod, Michael Bliemel and Nancy Jones, examines the demand for big data and analytics curriculum. Such demand seems to be increasing as organizations are more and more eager for a highly skilled and data-savvy workforce. Connecting business processes and analytics is important to facilitate decision making, and yet, today, this means dealing with big data. Thus, it is crucial to examine and share how faculties come to adopt curriculums on evolving big data and analytics. To support the dissemination of such curricula, large software companies such as SAP, Oracle and Microsoft fund programs that provide “hands-on” curricula to be adopted by academic institutions. One such program, SAP University Alliances, supports the development of faculty curricula, subsidizes targeted curriculum development based on its products and provides community-based support for courses such as big data and analytics. This landscape continues to change as technological improvements make their way into businesses and academia. In this work, the authors explore the SAP University Alliances program to review the academic utilization of curricula on big data and analytics, and then consider academic usage trends and suggest a research agenda to support new curriculum development. The results show that there is a sizable shift toward business process curriculum in general and a larger demand for newer big data and analytics curriculum. Recommendations from business leaders in this area encourage the identification of business processes in the production of big data and analytics, so as to enable better organizational decision making. A faculty interested in creating or furthering their business process programs to include big data and analytics will find in this study a number of practical information, materials, suggestions, as well as a resourceful agenda on research and curriculum development.

The paper entitled “A big data framework for facilitating product innovation processes” is coauthored by Yuanzhu Zhan, Kim Hua Tan, Guojun Ji, Leanne Chung and Minglang Tseng; it seeks to explore how firms could use big data to facilitate product innovation processes, notably by shortening the time to market, improving customers’ product adoption and reducing costs. The research is based on a two-step approach. First, this research identifies four potential key success factors for organizations to integrate big data with a view to accelerating their product innovation processes. The proposed factors are further examined and developed by conducting interviews with different organization experts and academic researchers. Then, a framework is proposed based on the interview outputs. This framework sets out the key success factors involved in leveraging big data to reduce lead times and costs in product innovation processes. The findings show that the three determined key success factors are: accelerated process, customer connection, and a fast launch-and-improve ecosystem. We believe that the developed framework based on big data represents a paradigm shift. It can help firms to make new product development faster and less costly.

In the paper “A theoretical model of jump diffusion-mean reversion – constant proportion portfolio insurance strategy under the presence of transaction cost and stochastic floor”, by Anindya Chakrabarty, Zongwei Luo, Rameshwar Dubey and Shan Jiang, the authors develop a theoretical model for the jump diffusion-mean reversion – constant proportion portfolio insurance (JD-MR-CPPI) strategy with transaction cost and stochastic floor. From the existing literature on CPPI strategy, a common assumption has been that the investment in the risk-free assets grows at a constant rate in spite of frequent trading. Empirical evidence buttresses the fact that the interest rate follows a stochastic mean reverting behavior, and thus the frequent reshuffling of portfolio between risky and risk-free assets makes it impractical to assume that the investment in the money market account will grow at a constant rate along the entire investment horizon. Considering this gap, this paper modifies the CPPI algorithm by
redefining the floor of the algorithm in order to turn it into a stochastic mean reverting process which is guided by the movement of the short-term interest rate in the economy. This development is more relevant for two reasons. First, the short-term interest rate changes with time and, therefore, the constant yield during each rebalancing step is not practically feasible. Second, historical literatures have revealed that the short-term interest rate tends to move opposite that of the equity market. So, during the bear run, the floor will increase at a higher rate whereas the growth of the floor will stagnate at the bull phase which helps the model to capitalize on the upward potential at the growth phase and to cut down on the exposure at the crisis phase. In the theoretic model of the JD-MR-CPPI strategy in the presence of transaction cost and stochastic floor as opposed to the deterministic floor, the paper adopts the Merton’s jump diffusion (JD) model to simulate the price path followed by risky assets, but also the CIR MR model to simulate the path followed by the short-term interest rate. The floor of the CPPI strategy is linked to the stochastic process driving the value of a fixed income instrument whose yield follows the CIR mean reversion model. The developed model is benchmarked against CNXNIFTY 50 and is back tested using the Monte-Carlo simulation across the crisis and recovery phases of the 2008 recession regime, and this revealed that the portfolio performs better not only than the risky markets during the crisis by hedging effectively the downside risk, but also than the fixed income instruments during the growth phase by leveraging on the upside potential. This makes it a value than can enhance a proposal for the risk adverse investors.

In the paper “A bibliographic study on big data: concepts, trends and challenges”, the authors Deepa Mishra, Zongwei Luo, Shan Jiang, Thanos Papadopoulos and Rameshwar Dubey conduct a review of the literature on big data by using citation and co-citation analysis. Citation analysis enables researchers to understand when the major articles in a field were published and how their popularity has evolved over time, and to see whether an article is still useful for current research (Pilkington and Meredith, 2009). Co-citation analysis, on its part, can reveal the major research clusters within a particular field and show how they evolve and vary across different journals over time. The article has a two-pronged objective: providing a consolidated overview of the existing literature on “big data”; and describing the current trends and opening up various future directions that may be explored by researchers as further research inputs in this rapidly evolving field. This analysis involved an assessment of 57 articles published over a period of five years (2011-2015) in ten selected journals. The findings reveal that the number of articles devoted to the study of “big data” has increased rapidly in recent years. Moreover, the study identifies some of the most influential articles of this area. The findings of research can help researchers to understand the evolution of research trends in the field, and those articles that have been influential in shaping research in these years; and to reveal the intellectual structure of the field. It is important for companies to adopt BDA so as to better comprehend the trends in customer behaviors and provide them improved and customized services.

The paper entitled “Past, present and future of contact centers: a literature review”, by Morteza Saberi, Omar Khadeer Hussain and Elizabeth Chang, deals with intelligent contact centers in CRM. While contact centers play a pivotal role in organizations and form an important part of CRM operations with the changes in technology, their operation and structure have to evolve in order to meet the ongoing challenges. In this paper, the authors made a review of the state-of-the-art literature and described the challenges while identifying the gaps to be filled through intelligent contact centers. The current literature on contact centers is essentially made up of analytical and managerial studies. It has been found that current contact centers suffer from two main defects, namely the lack of interactive contact centers and the high amounts of unstructured data. Given the production of massive amounts of data in contact centers, especially unstructured data, the authors opted for discussing the benefit of using big data processing techniques in contact centers and presenting it as a new research agenda for
CCs. Based on the critical literature review that are being made on contact centers, these structures have the potential to receive more attention from data scientists in the next decade.

In the paper “Let’s stop trying to be ‘sexy’ – preparing managers for the (big) data-driven business era” by Kevin Daniel André Carillo, the author critically investigates the inadequacies of current business education in the tackling of the educational challenges inherent to big data and to the overall advent of a data-driven business era. Through a review of the literature and of secondary data, it analyzes the implications of digitization and BDA on organizations, with a special emphasis on decision-making processes and the function of managers. The paper argues that business schools and other educational institutions have well responded to the need to train future data scientists but have rather disregarded how the function of managers and decision makers shall evolve in data-driven organizations. In short, the development of analytics skills and mindset shall not pertain to data scientists only; it must rather become an organizational cultural component shared among all employees and, more specifically, among managers. In the data-driven business era, managers turn into manager-scientists who shall possess skills at the crossroads of data management, analytical/modeling techniques and tools, and business. The multidisciplinary nature of BDA and data science seems to collide with the dominant “functional silo design” that characterizes business schools which have been criticized for years for their lack of multidisciplinary integration and experiential components. The specificities of analytics and data science necessitate revisiting pedagogical models by developing approaches such as experiential learning or spiral-shaped pedagogy. The attention of scholars is needed as there is an array of unexplored research territories in which investigation will help bridge the gap between education and the industry. The results and recommendations of this paper will help practitioners and business education to develop effective trainings and programs that are suited to address the challenges faced by data-driven organizations.

The paper “Does big data analytics influence frontline employees in services marketing?” by Saradhi Motamarri, Shahriar Akter and Venkat Yanamandram, argues that frontline employees (FLEs) play a dual role as “voice of the firm” to the customer and “voice of the customer” to the firm. The FLEs need to adapt their service to suit the individual customer needs and thereby enhance the customer’s service experience. In high contact services, like financial, healthcare and airlines, FLEs need to deal with every other customer differently as the interactions are highly personal and variable in nature. Detailed information about customers and their path to service facilitate FLEs to adapt the service in an optimal fashion. BDA may provide insights about the customers’ preferences and market conditions, which facilitate service adaptation. For a financial consultant, a prior idea about a customer is invaluable in structuring an effective solution; as for a retailer, he can offer better discount to a loyal customer, while an airline may propose an optimal itinerary for a frequent traveler. Aptly, Ostrom et al. (2015) have identified big data as one of the 12 key research priorities for services marketing. The motivation for this research is to answer the question: “How does big data analytics enable frontline employees in effective service delivery?” In doing so, while synthesizing the extant literature, the paper explores the challenges of FLEs to: enhance service delivery, support informed customers, achieve mass-customization, and build deeper relationships with customers. From the perspective of FLEs, the role and necessity of BDA differ from one service type to another. For example, delivering a financial consultation is resource- and time-intensive in comparison to processing a withdrawal request at a teller counter. In recognition of this, the research develops a service typology to explore the research questions. Similar to variations in customers and service types, there exist variations among FLEs and it is vital to recognize their typology as well. In addition to these variations, firms also vary according to their maturity in deploying BDA across their business functions. One of the findings of this research is that all these typologies
intricately interact and influence BDA and impact the service delivery capabilities of frontlines. Our analysis also conveys that in knowledge-intensive, customizable and high contact service contexts with innovators, the type of FLEs in highly matured BDA firm settings (metamorphosis or Level 5) results in higher co-creation potential for both firms and customers. However, the review has identified significant knowledge gap in enabling the FLEs with BDA tools. Reconciling such shortcomings from the service-dominant logic perspective, it implies that managers ought to enhance the skill asymmetry between their frontlines and customers so that providers sustain their service portfolio. It also suggests that managers need to devise training programs to enable frontlines. Frontlines are to be oriented with customer linking and market-sensing capabilities and empower them to make adaptive decisions in real time. Ultimately, the better the frontlines deliver service, the better organizations sustain in the competitive markets. Last but not least, both firms and customers also need to be aware of the privacy and ethical concerns of big data.

In the paper “Big data in the Danish Industry: application and value creation” by Sune Dueholm Müller and Preben Jensen, the authors argue that companies are increasingly realizing the value creating potential in big data. This article focuses on how SMEs use big data to create value. The study compares the findings from an online survey among 457 Danish companies with the extant literature, in an effort to address the following research question: To what extent does the application of big data create value for small and medium-sized companies? The findings show that the application of big data correlates with value creation, but that it is highly dependent on the organizational context and managerial action. Companies should not focus on capturing, storing, and analyzing data through application of technology in isolation. Strong leadership is needed in terms of establishing and communicating clear business goals. IT capabilities, business processes, and analytical skills need to be aligned with and adapted to those goals. Thus, a holistic perspective on big data is advocated and implications are offered with an emphasis on how the application of big data creates value for SMEs.

In the paper “Employees’ reactions to IT-enabled process innovations in the age of data analytics in healthcare” by Hillol Bala and Viswanath Venkatesh, the authors claim that inter-organizational business process standards (IBPS) are IT-enabled process specifications that standardize, streamline and improve business processes related to inter-organizational relationships. There has been much interest in IBPS as organizations from different industries implement these process innovations that lead to successful organizational outcomes when they are integrated into and standardized with intra- and inter-organizational business processes. These process innovations enable BDA capabilities by facilitating new sources of inter-organizational process data. The purpose of this study is to unearth employees’ reactions to a new type of supply chain process innovations that involved an implementation of a new IBPS and supply chain management (SCM) system and associated analytical capabilities. The researchers gathered and analyzed qualitative data for a year from the employees of a healthcare supplier, a high-tech manufacturing organization, during the implementation of a SCM system and RosettaNet-based IBPS. In what they termed the initiation stage, there was quite a bit of confusion and unrest among employees regarding the relevance of the new process standards and associated analytical capabilities. With the passage of time, in the institutionalization stage, although the situation improved slightly, employees found workarounds that allowed them to appropriate just part of specific processes and the analytical capabilities. Finally, once routinized, employees felt comfortable in the situation but still did not appropriate the new supply chain processes faithfully. Overall, employees’ reactions toward the SCM system and associated analytical capabilities were different from their reactions toward the new business processes. Significant contribution to the literature is hereby made by the authors, as they offer novel insights on how employees react to and appropriate process innovations that change their work processes.
In the paper “Big data analytics: transforming data to action” by Daniel Bumblauskas, Herb Nold, Paul Bumblauskas and Amy Igou, the authors provide a conceptual model for the transformation of big data sets into actionable knowledge. The model introduces a framework for converting data to actionable knowledge and potential impediments for the organization. Case examples and the use of dashboards provide a practical application for analysis of big data. The authors developed the model and used industry experience and network resources to gain valuable insights into an effective BPM related to BDA in research and practice. Examples and cases have been provided to highlight the use of dashboards as a visual tool within the conceptual framework. They argue that the transitions required reaching the actionable knowledge state; virus identification process and dashboard visualization tools can all be deployed by practitioners in industry. Furthermore, information assurance, security, and the risk of large-scale data breaches are a contemporary problem in the society today. These topics have been addressed within the model framework.

In the paper “Digital competences of the workforce – a research topic?” by Matthias Murawski and Markus Bick, the authors deal with digital competences of the workforce. They discuss whether this field is a research topic or not. Their paper presents both pros and cons regarding this issue based on the experiences and findings gathered from literature analysis as well as practical projects. The authors argue that digital competences of the workforce are indeed a research topic. They conclude by providing a first proposal of a research agenda for the topic at hand. Although a comprehensive list of related literature is provided, this paper is a personal viewpoint and does not report on the outcomes of a structured research project. This paper is one of the very few contributions that are being made in the area of digital competences of the workforce; besides, it considers various perspectives like different IT generations, occupations, roles, and curricula design.

In the paper “Automated competitor analysis using big data analytics: evidence from the fitness mobile app business”, by Liang Guo, Ruchi Sharma, Lei Yin, Ruodan Lu and Ke Rong, the authors argue that competitor analysis is a key component in operations management. Most business decisions are rooted in the analysis of rival products inferred from the market structure. Relative to more traditional competitor-analysis methods, they provide operation managers with an innovative tool to monitor a firm’s market position and competitors in real time at higher resolution and lower cost, all of which more traditional competitor-analysis methods cannot enable. They combine the techniques of Web Crawler, Natural Language Processing and Machine Learning algorithms with data visualization to develop a big data competitor-analysis system that informs operations managers about competitors and meaningful relationships among them. The authors illustrate their approach by using the fitness mobile app business. Their study shows that the proposed system supports operational decision making both descriptively and prescriptively. In particular, their innovative probabilistic topic modeling algorithm combined with conventional multidimensional scaling, product feature comparison and market structure analyses reveal an app’s position in relation to its peers. The authors also develop a user segment overlapping index based on user’s social-media data. They combine this new index with the product functionality similarity index to map indirect and direct competitors with and without user lock-in. Their approach improves on previous approaches by fully automating information extraction from multiple online sources. The authors believe this is the first system of its kind. With limited human intervention, their proposed methodology can easily be adapted to different settings, giving quicker, more reliable real-time results. Their approach is also cost-effective for market analysis projects covering different data sources.

**Professor Samuel Fosso Wamba**

*Department of Information, Operations and Management Sciences, Toulouse Business School, Toulouse, France*
References


About the Guest Editor
Samuel Fosso Wamba, PhD, HDR, is a Full Professor at the Toulouse Business School, France. He earned an MSc Degree in Mathematics from the University of Sherbrooke in Canada, an MSc Degree in e-commerce from HEC Montreal, Canada, and a PhD Degree in Industrial Engineering for his work on RFID-enabled supply chain transformation from the Polytechnic School of Montreal, Canada. His current research focuses on business value of IT, business analytics, big data, inter-organizational system (e.g. RFID technology) adoption and use, e-government, IT-enabled social inclusion, IT and talent management, supply chain management, electronic commerce and mobile commerce. He has published papers in the proceedings of a number of international conferences (IEEE, AMCIS, HICSS, ICIS, and PACIS) and in renowned international journals, including the *European Journal of Information Systems, Production Planning and Control, the International Journal of Production Economics, Information Systems Frontiers, the International Journal of Production Research*, the *Business Process Management Journal*, etc. He has been organizing special issues on IT-related topics for top IS and OM journals. He is certified from CompTIA-RFID+. Professor Samuel Fosso Wamba can be contacted at: s.fosso-wamba@tbs-education.fr
Big data integration with business processes: a literature review

Samuel Fosso Wamba
Department of Information, Operations and Management Sciences, Toulouse Business School, Toulouse, France, and
Deepa Mishra
Indian Institute of Technology Kanpur, Kanpur, India

Abstract

Purpose – The purpose of this paper is to improve the understanding of the integration of business process management (BPM), business process re-engineering (BPR) and business process innovation (BPI) with big data. It focusses on synthesizing research published in the period 2006-2016 to establish both what the authors know and do not know about this topic, identifying areas for future research.

Design/methodology/approach – The research is based on a review of 49 published papers on big data, BPM, BPR and BPI in the top journals in the field 2006-2016.

Findings – In this paper, the authors have identified the most influential works based on citations and PageRank methods. Through network analysis the authors identify four major clusters that provide potential opportunities for future investigation.

Practical implications – It is important for practitioners to be aware of the benefits of big data, BPM, BPR and BPI integration. This paper provides valuable insights for practitioners.

Originality/value – This paper is based on a comprehensive literature review, which gives big data researchers the opportunity to understand business processes in depth. In addition, highlighting many gaps in the current literature and developing an agenda for future research, will save time and effort for readers looking to research topics within big data and business processes.

Keywords Research methodology, Network analysis, Bibliometric analysis, Big data analytics

Paper type Literature review

1. Introduction

The area of big data has received increasing attention from both the academic and the business communities over the past few decades. It helps to gain business insights, competitive advantage and transforms entire business processes (Wong, 2012; Oh et al., 2012; Mishra, Gunasekaran, Papadopoulos and Childe, 2016). It can “create significant value for the world economy”, which “enhances the productivity and competitiveness of companies and the public sector and creates a substantial economic surplus for consumers” (Manyika et al., 2011, p. 1). Several objectives can be fulfilled by analysing big data, so there is a need for analytical techniques (big data analytics or BDA) to deal with large and complex data sets. Some organizations view BDA as a tool that can help in making strategic decisions, while scholars use it as a basis for verifying existing models and theories (Muhtaroglu et al., 2013). Organizations can empower their customers and improve decisions if they harness the power of BDA effectively (Miller, 2013). By recognizing BDA’s potential, organizations improve the efficiency and quality of their business processes through effective business process management (BPM). BPM not only improves processes but also monitors the technological advances that can be integrated in the development of efficient processes through business process re-engineering (BPR) and business process innovation (BPI) (Anand et al., 2013). Thus, the successful integration of BDA and business processes may create a “new class of economic asset” and help the top-performing organizations redefine their business and outperform their competitors.

A number of literature reviews on big data have been completed in the past few years (Sagiroglu and Sinanc, 2013; Fosso Wamba et al., 2015; Gandomi and Haider, 2015;
Amiri Khorheh et al., 2015; Wang et al., 2016; Mishra, Gunasekaran, Papadopoulos and Childé, 2016). Anand et al. (2013) reviewed the literature on BPM, BPR and BPI. Nevertheless, there has been no effort to review the integration of big data and BPM, BPR and BPI, using rigorous bibliometric tools to make them useful for researchers and practitioners. There is therefore a need to synthesize the evidence about the usefulness of existing studies.

Motivated by this lack, our main objective with this study is to introduce the idea of using a bibliometric and network analysis technique to explore the world of big data and business process. We aim to: systematically and rigorously collect and analyse existing studies in this field to identify the top contributing authors, countries, affiliations and key research topics; and use network analysis to reveal future research gaps that can be pursued by the big data research community. We performed this analysis using the guidelines proposed by Fahimnia et al. (2015). A bibliometric and network analysis is a powerful tool for identifying established and emerging topical areas. This review collects and analyses 49 articles on big data and business processes published from January 2006 to October 2016. We believe that this review will be valuable for researchers who want to identify areas that have been thoroughly researched or where research is lacking, and for practitioners who want to stay up to date about the state of research and big data and business processes.

The paper is structured as follows. Section 2 outlines our research methodology, including protocol development, study selection, data extraction and analysis. Sections 3 and 4 report the results of the bibliometric and network analysis. Our findings, limitations and directions for future research are discussed in Section 5.

2. Methodology
This study is a bibliometric and network analysis review, that is, we document all the available studies relevant to a current area or a specific research question (Fahimnia et al., 2015). We determined on this methodology for a number of reasons: to identify the top contributing authors, organizations and countries related to the field; to compare citation and PageRank analysis; and to uncover current research gaps through data clustering.

To achieve our research objectives, we took a five-step approach, outlined in detail below: develop a review protocol; identify inclusion and exclusion criteria; explain the search strategy process; study the selection process; and use data extraction and analysis.

2.1 Review protocol
Our search began with the development of a comprehensive review protocol based on the guiding principles and procedures of the bibliometric and network analysis review. This protocol identifies the search strategy, research objectives, data extraction, criteria for article selection and data analysis.

2.2 Inclusion and exclusion criteria
To achieve our objectives, we set up inclusion and exclusion criteria so that the most relevant articles were extracted from the database. We considered research articles from peer-reviewed journals in the English language, published from January 2000 to October 2016 in the Web of Science (WoS) database. We eliminated conferences, workshops, editorials, meetings, notes and tutorial summaries and considered articles only related to big data and business processes.

2.3 Search strategy
We chose the WoS database as it is one of the largest bibliographic databases, providing access to articles published since 1970 and covering approximately 8,500 high-impact
research journals. We searched for specific keywords derived from our research objectives and the structure of this review to identify relevant articles. We searched in titles, abstracts and keywords of articles in the WoS database for “BPM”, “BPR”, “BPI”, “big data”, “business analytics” and “big data analytics”. The initial search resulted in 1,078 articles. The results were then saved in plain text format and contained basic information about the paper, such as title, authors’ names and affiliations, abstracts, keywords and references.

2.4 Study selection process
Of the 1,078 studies selected, 253 were duplicates and were removed using Endnote. To fulfill the objective of our study, we restricted our search to titles, abstracts, keywords and peer-reviewed journals (excluding grey literature – workshops, conference papers, notes, editorials, meetings, etc.). This elimination process resulted in 486 relevant documents, published during the 11-year period 2006-2016. The next step in the selection process was to consider articles published in the top ten journals (i.e. the journals with the maximum papers in the field according to WoS). We found 49 articles. The distribution of the primary studies throughout the period is presented in Figure 1.

We restricted the period studied to 2006-2016 to facilitate graphical representation. The number of papers (nine) for 2016 was estimated up to October 2016. The figure demonstrates the changing behaviour of publications in each year. It shows that the number of publications on big data and business processes increased slowly from 2006 to 2012, with a dramatic rise in publications after that date. It is clear that interest in integrating big data with business processes has increased rapidly in the past four years.

2.4.1 Distribution of articles per journal. Table I, which details the number of articles related to big data integration with business processes by journal, shows that the most popular is Decision Support Systems with 13 articles (26.53 per cent), followed some distance behind by Interfaces and Industrial Management Data Systems (5; 10.20 per cent) and the IBM Journal of Research and Development (4; 8.16 per cent).

2.4.2 Classification of articles: approaches used most. The distribution of articles by approach used is presented in Table II. The vast majority of papers (14; 28.57 per cent) are experiments/model-based papers, followed by review papers (10; 20.41 per cent) and survey studies (9; 18.37 per cent). The results also indicate that there is a shortage of simulations (3; 6.12 per cent) and case studies (6; 12.24 per cent).

2.4.3 Classification of articles: research areas. Table III shows the classification of articles based on research areas. The largest number of published articles is in computer science (35; 71.43 per cent), followed by operations research (27; 55.10 per cent), engineering
The high proportion of articles in computer science is not a surprise as this research area is at the heart of the big data revolution and has provided tools to analyse massive amounts of data.

2.5 Data extraction and synthesis

In the data extraction and synthesis stage, we read all 49 studies carefully and extracted relevant data using Endnote and Excel spreadsheets. Our main objective was to obtain full and precise content records of all the primary studies. The data related to authors, keywords, ISSN, study title, publication date, location and affiliation; cited references were extracted from the WoS core collection. Once the data from the primary studies had been extracted and recorded, we performed the analysis using BibExcel and Gephi.
3. Bibliometric analysis

In this section, we provide statistics about the 49 shortlisted articles. Specifically, we studied these articles in terms of their authors, keywords, affiliations and funding agencies. We conducted bibliometric analysis using BibExcel as it is highly flexible and capable of dealing with large volumes of data; it is also compatible with other applications such as Excel, Pajek and Gephi (Persson et al., 2009; Paloviita, 2009). An additional merit of BibExcel is that it generates data for future network analysis, which is not possible with other software like HistCite or Publish or Perish.

The data extracted from WoS in plain text format (containing all the necessary bibliographic information) was used as input into BibExcel. For the data analysis, the plain text format was reformatted to generate different file types, such as .net-file, .cit-file, .oux-file and .out-file.

3.1 Author influence

To analyse the influence of authors, we extracted the author field from the data file and recorded the frequency with which all authors appeared. Table IV presents a list of the top ten contributing authors and the number of publications they have authored or co-authored. As we can see, Chae, with three publications, dominates the list, followed by seven others with two publications.

3.2 Keywords

The keywords and words used in the titles of papers were extracted from WoS plain text format in BibExcel, and the frequency of their occurrence was recorded. The top 20 words used in titles and most popular keywords are presented in Tables V and VI. From these tables,

<table>
<thead>
<tr>
<th>Authors</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chae, B.</td>
<td>3</td>
</tr>
<tr>
<td>Zhao, J.L.</td>
<td>2</td>
</tr>
<tr>
<td>Trkman, P.</td>
<td>2</td>
</tr>
<tr>
<td>Sheu, C.</td>
<td>2</td>
</tr>
<tr>
<td>Olson, D.</td>
<td>2</td>
</tr>
<tr>
<td>Mccormack, K.</td>
<td>2</td>
</tr>
<tr>
<td>Liberatore, M.</td>
<td>2</td>
</tr>
<tr>
<td>De Oliveira, M.P.V</td>
<td>2</td>
</tr>
<tr>
<td>Zorrilla, M.</td>
<td>1</td>
</tr>
<tr>
<td>Zimbrao, G.</td>
<td>1</td>
</tr>
</tbody>
</table>

Table IV. Top ten contributing authors

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytics</td>
<td>15</td>
<td>Information</td>
<td>4</td>
</tr>
<tr>
<td>Data</td>
<td>10</td>
<td>Big</td>
<td>4</td>
</tr>
<tr>
<td>Business</td>
<td>8</td>
<td>Mining</td>
<td>3</td>
</tr>
<tr>
<td>Supply</td>
<td>6</td>
<td>System</td>
<td>3</td>
</tr>
<tr>
<td>Enterprise</td>
<td>6</td>
<td>Systems</td>
<td>3</td>
</tr>
<tr>
<td>Chain</td>
<td>5</td>
<td>Framework</td>
<td>3</td>
</tr>
<tr>
<td>Services</td>
<td>5</td>
<td>Process</td>
<td>3</td>
</tr>
<tr>
<td>Management</td>
<td>5</td>
<td>Modelling</td>
<td>3</td>
</tr>
<tr>
<td>Performance</td>
<td>4</td>
<td>Operational</td>
<td>3</td>
</tr>
<tr>
<td>Analysis</td>
<td>4</td>
<td>Impact</td>
<td>3</td>
</tr>
</tbody>
</table>

Table V. Top 20 words in titles
we see a uniformity in the words used in titles and in lists of keywords. For instance, both tables include big data, analytics, performance and information systems. This demonstrates clearly that the most popular keywords are actually the search words we chose for this study.

### 3.3 Affiliation statistics

To understand the impact of affiliation on the number of publications, authors’ affiliations were extracted from the WoS plain text file in BibExcel. The frequency with which these affiliations occurred was used to identify the top-performing organizations, shown in Table VII. The contribution of countries can also be identified in a similar way. Table VIII shows the top ten countries contributing to the field of big data. A comparison of Tables IV and VII reveal that top universities like Kansas State University and the University of Ljubljana are represented by

<table>
<thead>
<tr>
<th>Affiliation</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Nebraska</td>
<td>4</td>
</tr>
<tr>
<td>Kansas State University</td>
<td>4</td>
</tr>
<tr>
<td>University of Ljubljana</td>
<td>3</td>
</tr>
<tr>
<td>IBM Software Group</td>
<td>3</td>
</tr>
<tr>
<td>Villanova University</td>
<td>2</td>
</tr>
<tr>
<td>IBM Research Division</td>
<td>2</td>
</tr>
<tr>
<td>IBM Corporation</td>
<td>2</td>
</tr>
<tr>
<td>City University of Hong Kong</td>
<td>2</td>
</tr>
<tr>
<td>Vlerick Business School</td>
<td>1</td>
</tr>
<tr>
<td>Virginia Polytechnic Institute</td>
<td>1</td>
</tr>
</tbody>
</table>

Table VII. Top ten contributing organisations

<table>
<thead>
<tr>
<th>Countries</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>24</td>
</tr>
<tr>
<td>People's Republic of China</td>
<td>9</td>
</tr>
<tr>
<td>UK</td>
<td>5</td>
</tr>
<tr>
<td>Germany</td>
<td>4</td>
</tr>
<tr>
<td>Taiwan</td>
<td>3</td>
</tr>
<tr>
<td>Slovenia</td>
<td>3</td>
</tr>
<tr>
<td>Canada</td>
<td>3</td>
</tr>
<tr>
<td>Brazil</td>
<td>3</td>
</tr>
<tr>
<td>Austria</td>
<td>2</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1</td>
</tr>
</tbody>
</table>

Table VIII. Top ten contributing countries
the top contributing authors Chae, B. and Trkman, P. Thus, it may be concluded that the performance of one or two researchers is sufficient to improve the ranking of an institute. We also notice that the majority of work has been carried out in USA, followed by People’s Republic of China, while only a few studies have been done in Austria and Switzerland.

4. Network analysis
The literature records the use of various tools, such as Pajek, VOSviewer, HistCite Graph Maker, and Gephi, to perform network analysis. We chose Gephi for this study as it can handle various data formats and complex data sets and generate flexible, insightful visual aids. In Gephi, the published articles act as nodes and citations as arcs or edges. To generate graphs in Gephi, a .NET file is needed and can be created using BibExcel.

4.1 Citation analysis
Citation analysis examines the frequency with which an article is cited. The number of citations of a particular article reflects its importance in that area of research (Garfield, 1972). Thus, the importance of an article can be measured as high or low, depending on the number of citations it has received. This method helps the researchers to understand how the area of research has evolved over a period of time and which articles are the most popular (Pilkington and Meredith, 2009). Although citation analysis has been criticized, it is one of the most commonly used techniques for analyzing the literature (MacRoberts and MacRoberts, 2010).

Figure 2 shows the ten most influential works published between 2006 and 2016. The top score, 45 citations, is Trkman et al. (2010). These authors investigated the relationship between analytical capabilities and performance in the planning, sourcing, manufacturing and delivery areas of the supply chain using information system support and business process orientation as moderators. Another important contribution was made by Abrahams et al. (2012) who used a text mining technique to analyse popular online discussion forums used by motor vehicle enthusiasts. This work received 28 citations, which reflects the significance of the article in this field. The article by Jararweh et al. (2014), which has been cited 16 times, introduced a modelling and simulation environment for cloud computing known as CloudExp and integrated it with the MapReduce processing model to handle the processing of big data.

4.2 PageRank analysis
Although citation analysis is commonly used to measure the popularity of an article, Ding et al. (2009) claimed that it should not be the only criterion of an article’s significance. Prestige, which records the number of times an article has been cited by other highly cited articles, is another
important criterion. To account for both popularity and prestige, Brin and Page (1998) introduced PageRank, which is an excellent way to prioritize the results of web keyword searches (Mishra, Gunasekaran, Papadopoulos and Childe, 2016; Mishra, Gunasekaran, Childe, Papadopoulos, Dubey and Wamba, 2016). There may be situations where these two measures are positively correlated, but it is not essential for a highly cited article to be a prestigious article as well. If we compare Figure 2 and Table IX, Trkman et al. (2010) has shifted to fourth position in the list of top ten PageRank papers, which is dominated by LaValle et al. (2011), while none of the other articles in Figure 2 appears anywhere in Table IX.

4.3 Co-citation analysis

Through co-citation analysis, we can examine the relationship between groups of authors, topics, journals or keywords and explain how they are related to each other. Co-citation analysis can be based on authors or publications; author-based co-citation analysis helps depict social structure while publication-based co-citation analysis helps explain the intellectual structure of research field (Chen et al., 2010).

In this study, we used Gephi to perform co-citation analysis. When the .NET file for 49 articles is opened for the first time in Gephi, a random graph with no clear pattern is generated. To provide visibility in the graph, we used Gephi’s ForceAtlas layout, in which strongly connected nodes move to the centre of the network while less connected nodes move to its boundaries (Bastian et al., 2009). This means that co-cited articles remain connected together while articles that are rarely co-cited are distanced from them. The nodes or “outliers” isolated from the network are excluded for the purpose of data clustering, which is explained in the following section.

4.3.1 Data clustering

The data clustering method helps to group articles in different clusters (Radicchi et al., 2004; Mishra, Gunasekaran, Papadopoulos and Childe, 2016; Mishra, Gunasekaran, Childe, Papadopoulos, Dubey and Wamba, 2016) that has been used in literature for classifying given set of publications and also termed as modularity. The edges between the nodes in the same cluster are denser than those in different clusters. This density can be measured through modularity, an in-built tool in Gephi based on Louvain algorithm. The value of the modularity index lies in between $-1$ and $+1$ that measures the density of links inside communities vs the links between communities (Fahimnia et al., 2015). Using this algorithm, we created four major clusters and found the modularity index to be 0.49 (see Figure 3). This value indicates strong relationship between the nodes within each cluster and yet a relatively strong relationship between the nodes of different clusters.

According to Hjørland (2013), if two or more articles are cited together, they are more likely to share a similar area of interest. Therefore, we performed a detailed analysis of the papers within each cluster to identify their research area. In Table X, we record the top publications based on their PageRank co-citation.

<table>
<thead>
<tr>
<th>Article</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lavalle et al. (2011)</td>
<td>0.012076</td>
</tr>
<tr>
<td>Davenport and Harris (2007)</td>
<td>0.011495</td>
</tr>
<tr>
<td>Chen et al. (2012)</td>
<td>0.010823</td>
</tr>
<tr>
<td>Trkman et al. (2010)</td>
<td>0.009516</td>
</tr>
<tr>
<td>Manyika et al. (2011)</td>
<td>0.008468</td>
</tr>
<tr>
<td>Bharadwaj (2000)</td>
<td>0.008252</td>
</tr>
<tr>
<td>Pfeffer and Sutton (2006)</td>
<td>0.007304</td>
</tr>
<tr>
<td>Barney (1991)</td>
<td>0.006890</td>
</tr>
<tr>
<td>Adomavicius and Tuzhilin (2005)</td>
<td>0.006674</td>
</tr>
<tr>
<td>Feng et al. (2008)</td>
<td>0.006359</td>
</tr>
</tbody>
</table>

Table IX. PageRank’s top ten articles
The classification in Table X reveals that the articles in cluster 1 mainly focus on conceptual and theoretical studies of big data. They highlight the need for analytical tools to deal with the massive amount of data that is being generated through recent developments in technology. These works also inspired organizations to use analytical decisions for business problem-solving and competitive advantage. Motivated by the works in cluster 1, researchers in cluster 2 identify the role of business analytics in managing and solving supply chain-related problems. The majority of the articles in this cluster are empirical and focus on the techniques that help to improve supply chain performance. Cluster 3 mainly concentrates on developing methods and models that would be beneficial while dealing with

![Figure 3. Four major clusters](image-url)

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
</table>

| Table X. | Top ten papers in each cluster (PageRank co-citation measure) | | |

Big data integration with business processes
forecasting problems, while researchers in cluster 4 focus on recent advances and trends in the big data environment. The first two clusters are the more popular, while there is a scope for future work in clusters 3 and 4. This four-cluster classification provides a reliable guide for scholars looking for current research topics and future research opportunities.

5. Discussion and conclusions

In this study, we present an overview of the distribution of publications on big data and business processes by conducting a bibliometric and network analysis review of articles written during the period 2006-2016. To extract relevant studies, we searched for papers in the WoS database using predefined keywords. We screened papers by analysing their titles and abstracts and removed those that violated the inclusion criteria. To provide an overview of big data and business process status, we identified a primary set of 49 articles. The results of our study identify the key contributing authors, countries, affiliations and keywords across a broad spectrum of disciplines.

We can see from the bibliometric results that a large majority of the 49 primary studies were carried out in USA (approximately 50 per cent), while only 2-4 per cent were done in Switzerland and Austria. We therefore recommend that these countries should put more research effort into improving their business processes by recognizing the potential of big data. Our findings also note that relatively low number of publications have appeared in this field. From Figure 1 we can see a rapid increase in publication numbers in the field of big data and business processes since 2012. This clearly demonstrates a growing interest in this area, which is unsurprising for a relatively new concept.

Reviewing and summarising what we know in relation to big data, business processes and how organizations integrate them to their advantage, we believe that this study will be beneficial for a wide range of researchers and practitioners. Our findings may help researchers to identify new research questions, gain an overview of current research and position and align their own work. Our study also helps practitioners to understand the practical challenges when integrating big data with business processes. Young scholars may use these findings as a guide to where to locate and publish different types of related research and gain further insights into the emerging field of big data.

5.1 Limitations and directions for further research

The literature review conducted in this paper has several limitations. Even though we adopted an established methodology, it could have limited the results as it focussed only on articles that appeared in peer-reviewed academic journals published in English. This may have led to the exclusion of potentially relevant articles from the sample. We took care to include all past studies by consulting the WoS database but the selection process we used may have omitted some relevant research papers. Furthermore, we omitted a search for grey literature – this may provide material for further insights. Overall, this review provides a perspective on the state of big data research today.

References


Appendix


Corresponding author
Professor Samuel Fosso Wamba can be contacted at: s.fosso-wamba@tbs-education.fr

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Business intelligence serious game participatory development: lessons from ERPsim for big data

Elise Labonte-LeMoyne, Pierre-Majorique Leger, Jacques Robert and Gilbert Babin
Department of Information Technologies, HEC Montreal, Montreal, Canada
Patrick Charland
Department of Education, UQAM, Montreal, Canada, and
Jean-François Michon
ERPsim Lab, Montreal, Canada

Abstract
Purpose – A major trend in enterprise resource planning software (ERP) is to embed business analytics tools within user-centered roles in enterprise software. This integration allows business users to get better and faster insight to action. As a consequence, it is imperative for business students to learn how to use these new tools to adequately prepare them for new expectations in the industry. The paper aims to discuss these issues.
Design/methodology/approach – In this paper, the authors propose a new serious game, called ERPsim for big data, to enable the learner to acquire abilities at each level of the business analytics learning taxonomy. To maximize the pedagogical impact of the game, participatory design (PD) with professors as co-designers was used during game development.
Findings – This case study presents the PD approach and analyses the efficacy of the proposed new simulation.
Originality/value – The authors conclude by providing recommendations and lessons learned from this approach.
Keywords Participatory design, Learning, ERP, Business analytics, Problem-based learning
Paper type Research paper

1. Introduction
Building on new advances in business analytics, forward looking organizations are setting up technological infrastructure to leverage the potential of what is been referred to as big data (vom Brocke et al., 2014). However, there is a shortage of skilled workers who have the expertise necessary to optimize the use of those tools (Wixom et al., 2014). While there has been an important increase in students registering in big data classes at universities and colleges (Davis and Woratschek, 2015), new pedagogical approaches are needed to help develop the skills and attitudes toward business intelligence required of the future workforce (Davenport and Patil, 2012).

This paper presents a case study of the collaborative efforts deployed to develop an enactive approach to teach business intelligence concepts and to provide hands-on experience to students in an authentic and dynamic business environment. Using participatory design (PD), a well-known approach at leveraging the collaboration of designers and users to co-innovate in the development of an IT artifact (Halskov and Hansen, 2015), we developed a
business simulation game that allows players to experience the challenges of big data. This new serious game was developed upon previous work on a simulation called ERPsim (Léger, 2006; Léger et al., 2007). Through the business simulation, participants learn to use an enterprise resource planning (ERP) system, to collaborate as a team, to understand the business environment and to make and implement business decisions. Until now, ERPsim was mostly aimed at learning goals related to operational excellence (Léger et al., 2012), the focus on big data analytics was limited. This new version of ERPsim intends to bring big data learning goals at the forefront of the experience.

In this paper, we report our approach and discuss the appropriateness of using a PD development method in this context. This co-design approach has led to the development of ERPsim for big data, a new version of our simulation that, results suggest, provides analytical abilities that expand the competency of enterprise system learners. We also discuss the effectiveness of the proposed simulation with faculty users involved in the beta testing phase. We conclude by identifying the lessons learned and provide recommendations for PD to develop enactive pedagogical material to teach IT as a means to identify pedagogical needs and preferences of an academic community.

ERPsim is a serious game developed for business school students to learn business process management using an ERP system (Léger, 2006; Léger et al., 2007). In the ERPsim simulation, teams of students interact through a real SAP system. Teams compete in the same virtual market for customers and products. For example, each team chooses to order certain amounts of raw materials to create a product, to ship quantities of that product to different regions and to set the price in each region. According to these decisions, each team makes a profit and the team with the highest profits wins the simulation. Through the game, participants learn to use an ERP system, to collaborate as a team, to understand the business environment and to make and implement business decisions. It has been argued that ERPsim follows the problem-based approach and brings students to learn higher levels of cognitive abilities with respect to enterprise systems (Léger et al., 2012), enhanced supply chain collaboration skills (Caya et al., 2014) and emotional control in ERP decision making (Léger et al., 2014).

The ERPsim simulation for big data seeks to create a dynamic big data environment using SAP software that can simulate the velocity of big data in order to enable the learner to achieve the highest levels of the taxonomy of business analytics. To find its place in this era of big data, SAP put forth a new tool for database architecture, SAP HANA (SAP, USA). SAP HANA is an in-memory database platform that was released in 2010. It allows for real-time analytics of both structured and unstructured data. In-memory analytics refers to a new way through which computers are managing data and applications by keeping data in their main memory instead of regularly having to access the hard drive to retrieve them (vom Brocke et al., 2014). One of the benefits of the HANA technology is that it is no longer necessary to extract the transactional data into a data warehouse in order to perform analytical processes. It is now possible to perform big data analysis directly from the HANA infrastructure which supports both the transactional and analytical processes. New software technologies, such as SAP HANA design studio (creation of analytical views) and SAP LUMIRA (data visualization), enable the extraction and analysis of big transactional data. By combining these new tools with the ERPsim simulator, students can learn how to exploit big data. The ERPsim for big data game is also aligned with Bliemel’s definition of a good educational game (Bliemel and Ali-hassan, 2014) through the following elements: it presents enough of a challenge for most students, allows students control over actions, presents clear goals, shows feedback in the form of financial statements, is a cooperative experience and develops new skills.

2. The pedagogical challenges of teaching business intelligence
There is an extensive literature on end-user training and training methods in the IS literature, and among the various training approaches, enactive learning has been encouraged in recent
years (Gupta et al., 2010). Grounded in social cognitive theory (Bandura, 1986), enactive learning builds upon a problem-based approach, in which answers are not to be memorized, but rather in which learners are forced to determine what they need to know in order to find the answers by themselves (Hmelo-Silver et al., 2007). An enactive approach provides the user with the possibility, by trial-and-error, to receive constant feedback in response to their actions, thereby allowing learning by doing through detection and correction of errors (Léger et al., 2014). Optimally, enactive problem-based learning builds on a student-centered approach, where students are in a small group setting and have to collaborate on a very authentic problem (Léger et al., 2012). In the field of information systems, research has shown that enactive training is a very effective approach to acquire skills and competences in using a new business software (Gupta and Bostrom, 2013; Léger et al., 2014).

While the approach itself is gathering support, to our knowledge, most of the available enactive pedagogical material to assist in teaching business intelligence is incomplete. The current pedagogical material available for teaching business analytics does provide various opportunities to put into practice business analytical skills using data sets provided to the students. However, this data have often been pre-generated or adapted for pedagogical purposes and, in most cases, these data sets are static. They represent data available for a company at a specific point in time. Students are tasked to explore and discover opportunities in the data.

Undisputedly, these pedagogical activities can help achieve some pedagogical goals. However, they can fall short in inducing more complex forms of learning and could be improved. The hierarchical complexity of pedagogical objectives is illustrated in the Bloom’s taxonomy of cognitive processes (see Figure 1).

Like the original taxonomy, “the revision is a hierarchy in the sense that the six major categories of the cognitive process dimension are believed to differ in their complexity, with remember being less complex than understand, which is less complex than apply, and so on” (p. 215) (Krathwohl, 2002). Thus, educators or trainers should aim to have their students attain each ability level in order for them to succeed in using this knowledge in their future careers. Different types of teaching need to be put forward in order to guide the students through all phases of learning. It could be argued that class activities requiring the implementation of business intelligence methods on a static database can guide students to reach the first three levels of the taxonomy: “remember,” “understand” and “apply.” For example, they can remember the terminology, menus and procedures specific to the

![Figure 1. Revised Bloom learning taxonomy](source: Adapted from Forehand (2012))
software, understand the functionalities of the system and apply the statistical methods presented in class to generate results. However, in the field of business analytics, the higher level abilities, such as “analyze,” “evaluate” or “create,” are necessary for students to be appropriately trained in management, visualization and interpretation of big data. Well-trained students should be able to analyze the information to correctly identify a problem, to evaluate the problem and identify possible solutions, and to create and develop various ways of representing the data so others can understand it and collectively make better decisions in the long run (Oceans of Data Institute, 2014). These skills will be essential for their ability to handle the complexity of the tasks and real world problems they will encounter as business intelligence professionals (Charland et al., 2015). To obtain these skills while interacting with big data, enactive training approaches are likely to be required.

Big data is commonly defined as the combination of volume (a large quantity of data), variety (multiple types of data) and velocity (the speed at which data are created) (Laney, 2001). With static data sets, students can be provided with volume and variety of data, but not easily with velocity. Velocity is characterized not only by the speed at which data are generated, but also by the dynamic nature of data and how it evolves as a consequence of business decisions. By regularly feeding new data to the students, some degree of velocity may be achieved, but the data would typically be predefined and would not reflect actual decisions made by the students. In order to allow students to attain the highest levels of the taxonomy, the use of unstructured and complex dynamic problems with multiple solutions appears to be an interesting alternative to static data sets, as it links the data to analyze with the decisions they make while resolving the problems.

2.1 ERP technology as platform for enactive learning of BI concepts

In 2012, the Business Intelligence version of the ERPsim game was developed to allow quasi real-time data analytics. It relies on extract, transfer and load (ETL) to provide updated data at the end of each virtual day (each day is about one minute long) and stores the data in a Microsoft SQL database. Using Microsoft Excel, the students are able to extract views from the Microsoft SQL database and refresh these views every minute. This approach offers the ability to develop dashboards that connect to live data from the game, opening a new array of pedagogical possibilities. The Business Intelligence version of the ERPsim game can be used to teach dashboard design principles and business intelligence with a real enterprise system and a live data set.

In this paper, we present ERPsim for big data game as the next step. It enables real-time analytics on raw transactional data. Instead of using an ETL approach, where data used for analysis are available after the ETL process is completed, analysis can be performed directly on transactional data. The information cubes (analytical views in the HANA naming convention) are designed to provide meaningful data to students, data that they are free to explore and visualize as they desire. Visualization tools, such as SAP Lumira, allow this exploration and understanding of the data, and ultimately supports the decision-making process during the simulation.

3. ERPsim for big data: a PD approach

As of 2015, more than 1,000 instructors have been trained to use ERPsim for teaching in more than 220 universities and colleges worldwide. Several of these instructors have been using the original game in their classroom for many years and are often the source of incremental innovation in the simulation.

Early in the phases of development of ERPsim for big data, we took the decision to involve several of these experienced instructors as co-designers in order to ensure that the new simulation game would meet the pedagogical objectives presented in Section 2. These collaborators had extensive experience with ERPsim; they teach ERPsim on a regular
basis to a variety of levels and disciplines of study, and have identified key aspects of this knowledge that students need to experience in the game. It was determined that they would be useful in identifying the type of data needed in the cubes for this game as well as the optimal ways to consolidate a large variety of needs into an efficient set of cubes.

As the development of ERPsim for big data got underway, PD was identified as a useful method to involve these co-designers. PD originated in Scandinavia as a tradition to involve the end-user of a technological artifact in its design. It has been defined as a collaboration between the designers and the people who will be using the finished product through an iterative process that often involves brainstorming conversations and prototyping (Simonsen and Robertson, 2013). PD requires reciprocal learning for both the designers and users before stakeholders can engage in democratic and creative sessions (Carmel et al., 1993). All participants thus become co-designers as they have an equally vested interest in the outcome of the process.

For our purpose, two PD events were organized. A design workshop took place in the early phases of development to clearly identify and consolidate the needs of all the co-designers as well as jumpstart the development. Following that, a beta test was put into place along with an accompanying blog to facilitate the conversation between co-designers.

3.1 Participatory co-design workshop

3.1.1 Participants. The PD workshop was held between June 16 and June 19, 2014. In total, 31 participants from 24 different universities and six different countries (USA, Canada, Germany, Switzerland, Indonesia and Finland) participated in the workshop. The workshop was the occasion for long term and experienced users of the ERPsim simulations to learn about new technological improvements and pedagogical innovations in the ERPsim simulations games.

3.1.2 Procedure and methods. Step 1 – development of reciprocal learning by the co-developers. To kick off the workshop, researchers presented the revised bloom taxonomy and the overall objective to develop a serious game material to enable a business student to develop abilities at each level of the taxonomy. To create reciprocal knowledge with regard to the existing ERPsim game, we began the PD session by having the participants play the most recent version of the simulation game, the ERPsim Manufacturing game (Léger et al., 2011). In this game, the teams have to plan production and then purchase the required raw materials in order to get ready-to-eat muesli boxes available for sale. Teams compete for customers (i.e. grocery stores). During the workshop, participants used the Business Intelligence version of ERPsim. As explained earlier, this technology provides analytic visualization based on MS Excel and an SQL Server database. During the simulation game, the ERPsim simulator pushes data into an SQL Server database. This database thus contains the historical data pertaining to the ongoing simulations. A Microsoft Excel worksheet is then used to extract and manipulate views (or queries) already available on the SQL Server database. The views are displayed in the MS Excel file in the form of pivot tables. Using the pivot tables, the participants can create multiple graphs and tables, use conditional formatting to create alerts, etc. The use of MS Excel offers multiple benefits, as it is an easily accessible programming tool. However, the data download is slow and not designed with the intent to provide easy insight to act upon. Furthermore, data are only available after some delay (two to three minutes). The objective of initially running this game with the MS Excel technology in our workshop was to make sure that all participants were familiar with the latest version of the technology, shared a similar understanding of the concept of cubes, and of the benefits and limits of using existing analytical technology in ERPsim games.

Step 2 – ideation process. After this reciprocal learning activity, we moved on to the ideation stage of the process. The moderator, one of the creators of the game, first provided an overview of the ideation process in which the group would participate. The moderator then
asked each participant to participate in a brainstorm activity regarding the information architecture required for a new version. Specifically, the participants had to identify useful sets of measures and dimensions. The information cubes around which the business analytics are built refer to two main concepts: measures (quantity sold, sales revenues, etc.) and dimensions (time, geographic, type of products, etc.). So a typical combination of measure (bold) and dimensions (italic) is something like “**quantity sold** by **product** over **time**” “**quantity in stock** by **product** and **area**” or “**number of different products available** on the market at **any given time**.” All in all, about 40 distinct combinations were proposed by the participants.

Step 3 – consolidation of requirements. Building on requirements extraction approaches proposed by Hartson and Pyla (2012), the group then worked on consolidating the informational requirements. Specifically, the group of co-designer filtered, sorted and regrouped the proposed measure and dimension combinations. Each participant was asked to identify which combinations among the proposed list were the most important and useful. Combinations were also sorted by category. Some combinations could be grouped into one theme, cube or view. In the end, we collectively came up with seven different cubes, cubes that together contained a large number of the proposed measure and dimension combinations. The cubes were drawn on the board by the moderator (see Plate 1).

Step 4 – review existing information architecture. When developing information architecture, it is crucial to reuse existing structure developed by the software manufacturer. The main goal is to leverage existing design and minimize future maintenance of the architecture. Thus, the next step was to examine the existing cubes already available in SAP. Using Lumira, a data-visualization tool that connects to ERPsim for big data to retrieve real-time data from the SAP ERP system, users can create visualizations using a drag-and-drop interface. While the participants were seated in the conference room, one of the lead researchers on the ERPsim for big data project, acting as moderator, and the ERPsim lab operations manager (OM) were at the front of the room. The OM was using Lumira in order to visualize data from a live game, data that were provided by some of the existing cubes in SAP. The objective of the exercise was to come up with specific data visualizations within Lumira, and participants were invited to propose ideas and debate on how to reuse or extend existing informational architecture in SAP HANA. The role of the moderator was to lead the conversation and provide guidance. Once a consensus was reached, the OM implemented the suggested presentation live, leading to more comments from the audience and adjustments by the OM. There was no structured layout for the workshop so as to allow the conversation to flow, the moderator and OM improvised when to move on to a new topic and which topic to move on to.

Plate 1.
Example of a cube drawn at the workshop
Step 5 – implementation of co-design solution. The final step was to create seven new cubes. The conceptual definition of the seven cubes was drawn on the classroom board (see Plate 1). The proposed designs were used to create a fast prototype version in SAP HANA to ascertain the feasibility of the proposed cubes. The implementation was performed live by an analyst so that participant could learn and comment on the detailed configuration of the solution. Modeling in SAP HANA is essentially done by creating “attribute views,” “analytical views” and “calculation views,” which are referred to as cubes. Using the tables from the database as their source, the cubes are created using a business use case to model business logic and provide meaningful reports (SAP SE, 2014). When implementation was complete, participants were able to access the cubes and provide immediate feedback on the configuration. Immediate adjustments were made to increase the quality of the solution.

3.1.3 Results and discussion
The main benefits of the workshop format stemmed from the face-to-face group setting. This setting allowed not only a more pedagogically optimal version of the game to be developed, but also gave the developers ready answers if changes discussed in the workshop are ever requested by others. In a more traditional setting where software testers simply send comments to a developer, this type of reflection would not have taken place. For example, based on a suggestion during the workshop, an existing view that used to provide information on the market every five days was reconstructed to allow the report to be produced every day, making it possible to have a more granular analysis. However, an experienced user and collaborator commented that having that level of details will often cause students to simply follow the market without thinking and anticipating the market’s reactions. This, in turn, would lead to students not learning as much about the decision-making process. In the end, the consensus was to go back to the less granular, but more educational, five day report.

This type of group reflection leads to a large number of innovation iterations in a short period of time and makes the development much more efficient. Another advantage of a face-to-face workshop of this type, as opposed to online discussions where a very vocal minority can seem like they represent the majority, is that the moderator can encourage everyone to voice their opinion. Finally, participants are fully immersed and have no other distraction, as opposed to an at-home beta test where other obligations can limit availability for discussion. Unfortunately, this process might be difficult to reproduce for a new game that does not already have a significant number of users who are willing to participate in the design of a new iteration. It also requires participants to have a common ground in order to ensure they all “speak the same language.”

In total, eight data cubes were developed based on the results of the co-design workshop. The next stage was to have these cubes tested in live setting. The next section describes the second phase of the co-development process.

3.2 Beta test of co-design solution
In the fall of 2014, a beta test of ERPsim for big data was conducted with a small sample of experienced instructors. These instructors used the new game in their classrooms for the semester as they saw fit. The goal was for them to evaluate the newly developed solution and to collaborate with the design team to further optimize these informational cubes. Also, we aimed at assessing if ERPsim for big data would meet the pedagogical objectives targeted by this new simulation, i.e. the analytical abilities at the different stages of the taxonomy, and more specifically to what extent this new version better meets these objectives as compared to previous ERPsim games.

3.2.1 Participants. The first phase of the beta test included 11 instructors from different universities throughout Canada and the USA. All of the instructors were university
professors and have accepted to provide feedback from their experience by answering an online questionnaire and by participating in a co-design blog. The participating instructors were teaching to classes ranging from 15 to 100 students and spanning all levels of undergraduate and masters classes. Six classes were in IT/IS departments, one in management and four in other. The instructors had between 7 and 25 years of experience as teachers and all have been using SAP and ERPsim for over two years.

3.2.2 Procedure and method. To accompany the beta test, a collaborative blog was put in place. Each user of the beta test was granted access so they could create entries and post comments on other entries. No restrictions on content were imposed.

To document the students’ learning, a web questionnaire was set up on Unipark (Questback, Germany), so it could be sent by e-mail and answered online. The first portion documented the makeup of the class (number of students, level, duration of use in class). The second portion of the questionnaire was the Self-assessment scale of learning in the context of an ERP business simulation game reproduced from Cronan et al. (2012) which requires participants to rate, using a one to seven Likert scale, the abilities they believe students developed while using the traditional ERPsim game vs while using ERPsim for big data. That instrument has three sections: the enterprise systems management knowledge, the business process knowledge and the SAP transaction skills. A cumulative score was calculated for each section for each game. The third portion required participants to rate their agreement with statements on the ease of use of ERPsim for big data and the fourth portion asked for comments on that use. Finally, the fifth portion documented the professor’s experience teaching and using SAP and ERPsim. A paired Wilcoxon test was used to evaluate the comparison between the traditional ERPsim game and the new ERPsim for big data measures. We also used non-parametric correlations to evaluate the impact of the professor’s experience, the makeup of the class and the professor’s participation in the blog onto the measures.

3.2.3 Results and discussion. When comparing the learning that occurs with a traditional ERPsim game and with the new ERPsim for big data game, a significant difference can be seen for the enterprise systems knowledge \( (p = 0.02) \) with perceived student learning increased for the ERPsim for big data game. For business process knowledge \( (p = 0.22) \) and the SAP transaction skills \( (p = 0.15) \), we cannot conclude that there is a significant difference in learning. These results show that this new version of the game is an improvement on the traditional version and that the increase in enterprise system knowledge is not counterbalanced by a drop in another area of knowledge. Interestingly, the newly designed information architecture appears to help the learner get a broader understanding of the purpose of the enterprise system. The perception might be associated with the fact that students had a better sense of the insight to action when using the new real-time visualization software when making decision in the enterprise system.

For the teachers’ perceptions, four correlations were found to be significant. First, the number of years of experience with ERPsim was negatively correlated with the agreement with the statement “I found ERPsim for big data useful as a pedagogical tool” \( (r = -0.83, p = 0.003) \). Second, the number of blog posts by a professor was negatively correlated with most scores of the learning assessment for both games (BP_ERPsim: \( r = -0.65, p = 0.03 \); TS_ERPsim: \( r = -0.59, p = 0.05 \); ES_BigData: \( r = -0.61, p = 0.06 \); BP_BigData: \( r = -0.73, p = 0.02 \); TS_BigData: \( r = -0.77; p = 0.01 \)). Both of these could be explained by professors who are participating in the co-design process or who are more experienced with the game being more critical of their students’ learning as they have a more thorough understanding of the optimal learning that can be derived from these simulations. Third, the number of blog posts was also positively correlated with the agreement with the statement “I found the use of ERPsim for big data an improvement on my previous classes” \( (r = 0.76, p = 0.01) \),
this would suggest that involvement in the PD process influences their perception of the added value of the new game.

The following summarizes the comments made by the professors to the qualitative questions we asked. To the question “ERPsim for big data allows students to work with real-time analytical tools, as a user of the ERPsim game, what do you believe this new tool brings to your students’ skills?” most participants pointed out the crucial role of real-time data in decision making. Almost all respondent answered that ERPsim for big data helped students acquire the ability to leverage real-time data in order to create instant analysis and make better decisions, “this tool allowed students to better understand the simulation and recognize the value of analytics.” Some participants mentioned that this tool brought the development of “visualization skills” to their students using a “simple drag and drop technology.” Finally, one participant mentioned its students gained the “ability to share a common perception on the company” and the “ability to rapidly iterate in their collaborative problem solving.”

To the question: “What do you believe is the added value of real-time analytics to your classroom?,” a number of respondents referred to the experiential aspect of the class. “The real-time analytics is an essential part of the classroom. It highlights one of the key value of information system.” “Students are able to experience real time analytics in a dynamics context,” “[…] without this the students might think ERP has no ability to provide real-time analytics.” One participant added: “it gives my students familiarity with a new business tool so that they can speak confidently in a job interview.”

To the same question, some users mentioned the added value of the ease of use of the game. Finally, the central themes of immediacy (“the importance of utilizing high velocity data,” “Immediate feedback from business decisions”) and of visualization (“ability to visualize business in motion and see patterns”) are mentioned similarly to the previous question.

To the question “How did the use of ERPsim for big data change the enthusiasm/interest/motivation of your students? Do you think it affected the mood and attitudes of the class when compared to your previous use of ERPsim game?,” most participants observed an improvement in attitude. “They felt somewhat more under control” was the response of one participant. “Students were very excited to be able to see how their decisions were affecting the performance of their organization” was the answer of another. The ease of use and the flexibility of the reporting enabled the students to see the impact of their decision. “They are not trapped in the limitations of the SAP GUI and can adapt their own decision support system. Hence they are more engaged.” Two respondents, however, remained unsure about the impact of the change on their students’ perceptions.

To the question “How would you use it differently next time?,” the responses from participants vary considerably. One wishes to highlight the technology behind HANA. One plans no changes next time. One wants to explore predictive tools to run regression and statistical analysis. One “would like the students have the ability to develop cubes and then develop their own analytical approach.” Finally, one “intends to demonstrate to students that while SAP and HANA facilitates timelier and more accurate business decisions within the business enterprise, it will still be up to them, as managers and supervisors to communicate with each other.”

Finally, we sought comments on future iterations of the game design, here are some comments from the beta test participants that identify a commonly requested feature that will be discussed by the design team. We asked “What changes would you like to see in ERPsim for big data?” While some are uncertain about the changes they would like to see in ERPsim for big data. Others have very precise requests such as “put the news and raw material prices so they can be piped in through Lumira dashboards” or “create location data for customer so we can geomap sales.” Two participants proposed to use alternative visualization tools such as Tableau (Seattle, WA, USA): “Tableau would increase the appeal of
using ERPsim in a business intelligence class that already uses Tableau as the primary visualization tool.” Three respondents would like to see more material developed in order to make “ERPsim [for big data] as the cornerstone of an applied Business Analytics class.” In particular, the “ability to integrate predictive analysis features” and “ability for students to construct their own views and dig more into the development side of HANA” were proposed.

In short, while it is not possible to conclude that the new version completely addresses all the learning stages of the taxonomy, this new version of the game appears to provide students with more enterprise system knowledge and get a better appreciation for the role of real-time data in business decision making. Comments by participants identified a number of additional features that would be appreciated by users, in particular, the ability to create additional views and/or connect to alternative visualization tools. This shows that the beta test and the accompanying survey were beneficial for the design team, while the improvement in learning confirms that the game is reaching its goals and paves the way for the next steps in development. Following the beta test, the eight cubes were optimized in light of the collaborators comments.

4. Conclusions
By employing a PD workshop and a closely monitored beta test, we gained a better understanding of the needs of the end users as well as benefited from their knowledge and experience in complementary fields of expertise. This allowed the development of this new game to be much faster than otherwise envisioned. It also set the tone for a much more active adoption of the technology, with users quickly informing the designers of any bugs. Finally, by having all the co-designers in the same room at the workshop, their needs were more easily reconciled, leading to less redundant cubes and a more efficient development.

Following the 2014 summer workshop, the final comments from the discussion were integrated in the development and optimization of the cubes. Specifically, they oriented the selection of the business cubes or views that will be included in the general release. It also provided the development team with enough comments to confirm or invalidate the needs for information in some of the views.

In preparation for the general release, all of the cubes initially developed have been reengineered to match the requirements defined by the team and tested with the co-designers. Additionally, cubes that were ideated during the workshop have been provided as an initial release during the beta period, and are now being optimized. Improvements to the simulator are also needed to support the new requirements; this is being done in parallel of the cube development. September 2015 marked the release of ERPsim for big data. The final version is described in Michon et al. (2015).

We conclude with a number of recommendations for future serious game developers who wish to employ similar activities to optimize their development. In general, involving teachers and professors, rather than students, as co-designers allowed us to circumvent some of the communication difficulties identified for PD in serious game design, as “serious game designers must be fluent with both domain content and game design,” while end users “may lack one or both of these forms of knowledge” (Khaled and Vasalou, 2014). The downside, however, of such a process is that expectations were much higher and workshop participants hoped their desired applications would be included in the game.

4.1 Recommendations for a PD workshop activity
- Someone who is part of the design team and who has a very good understanding of what is currently available should do the moderation to lead the conversation in the right direction.
- Concerns about the implementation of the suggestions should not be discussed during the workshop to allow creative out-of-the-box thinking. As mentioned by the OM:
“say yes to everything, discuss feasibility later.” However, there should be an initial disclaimer that not all ideas will be kept to avoid disappointed expectations later on.

- According to the OM, it is important to “have an open mind and humility about your product, since the participants will challenge design decisions at some point. By providing them with an environment to freely express their opinions about the product, you allow yourself to get the most out of the activity, and discussions may explore possibilities that had not been considered initially.”

- To avoid concentrating on note taking, record the discussion.

- The more users are comfortable with the functional technology, the better they will be able to comment on the content. Participants should be given ample time to familiarize themselves with a prototype ahead of time.

- Having participants play a preliminary version of the game before the brainstorming session allowed them to really understand the possibilities of such a tool.

4.2 Recommendations for a closely monitored beta test

- A user-friendly way to report bugs in the system is very important; however, it might be preferable to have that separately from a comment section on possible design improvements.

- An objective empirical evaluation of the game should be planned out in advance and can be a great tool to demonstrate the pedagogical advantage such a game provides.

- Qualitative comments, however, are useful in identifying specific elements of the game that require improvement.

The development of this game was done with enactive learning in mind. The next step is to evaluate students’ learning and to perform a direct comparison with vicarious learning alone for the same topic to adequately identify which step of the Bloom Learning Taxonomy (Krathwohl, 2002) is reached by students using this game as well as whether or not students do learn more through enactive learning when applied to business intelligence and big data management. For example, with the suggestion of allowing students to create their own views that was brought up in the beta testing phase, this was implemented (Michon et al., 2015) and we believe by having them be creative they will attain the highest understanding of this topic.

This paper contributes to research and practice in the field of educational games development. PD with experienced instructors as co-designers has proved most helpful in the case of ERPsim for big data and the lessons learned from this development should be taken into consideration by others. It will help them design their games faster, help insure that the game is aligned with their pedagogical needs and help increase adoption once the game is released and using a structured PD method will help consolidate the needs of multiple co-designers without getting overwhelmed. It also contributes to research and practice on the teaching of big data, showing the importance of enactive learning and the importance of an authentic and dynamic business environment. With appropriate pedagogical tools that provide students with optimal skills and attitudes toward business intelligence, we can ensure a highly skilled future workforce.

References


Oceans of Data Institute (2014), Profile of a Big-Data-Enabled Specialist, Education Development Center, Waltham, MA.


For instructions on how to order reprints of this article, please visit our website: www.emeraldgroup-publishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com
Examining the adoption of big data and analytics curriculum

Alexander J. McLeod  
*Department of Health Information Management, Texas State University, San Marcos, Texas, USA*

Michael Bliemel  
*Rowe School of Business, Dalhousie University, Halifax, Canada, and*

Nancy Jones  
*School of Accountancy, San Diego State University, San Diego, California, USA*

**Abstract**

**Purpose** – The purpose of this paper is to explore the demand for big data and analytics curriculum, provide an overview of the curriculum available from the SAP University Alliances program, examine the evolving usage of such curriculum, and suggest an academic research agenda for this topic.

**Design/methodology/approach** – In this work, the authors reviewed recent academic utilization of big data and analytics curriculum in a large faculty-driven university program by examining school hosting request logs over a four-year period. The authors analyze curricula usage to determine how changes in big data and analytics are being introduced to academia.

**Findings** – Results indicate that there is a substantial shift toward curriculum focusing on big data and analytics.

**Research limitations/implications** – Because this research only considered data from one proprietary software vendor, the scope of this project is limited and may not generalize to other university software support programs.

**Practical implications** – Faculty interested in creating or furthering their business process programs to include big data and analytics will find practical information, materials, suggestions, as well as a research and curriculum development agenda.

**Keywords** Curriculum development, Big data, Research agenda, Analytics, Business process curriculum, University Alliances

**Paper type** Research paper

**Introduction**

Business analytics technologies such as web intelligence, web analytics, and mining of unstructured data continue to drive demand for greater insight into enterprise system data and have spawned a new era of business intelligence and analytics research (Chen *et al.*, 2012; Dubey and Gunasekaran, 2015; Wamba *et al.*, 2015). These relatively new technologies drive demand for analytically skilled knowledge workers. In the USA, McKinsey (2011) estimates a shortage of 140,000-190,000 knowledge workers with deep analytical skills by 2018. Roughly 50 percent of the respondents to a 2014 Accenture and GE survey, “Moving Toward the Future of the Industrial Internet” indicated that their organizations lacked sufficient talent to collect, analyze, and interpret big data and 87 percent of these responding corporate managers believe that big data and analytics are changing their competitive landscape in meaningful ways, compelling investment in big data and analytics technologies (Accenture and General Electric, 2015). These business leaders perceive that failure to analyze the vast volume of data and form appropriate real-time strategies is to fall behind in a digitally competitive environment. This sense of urgency has created the

The authors would like to thank the SAP University Alliances program and the world-wide peer hosting centers.
demand for individuals with big data and analytical skills. McKinsey Global Institute (Violino, 2014) predicts shortage of 140,000-190,000 data scientists and a lack of 1.5 million managers with analytical decision making ability by 2018. Organizations are increasingly turning to big data and analytics to provide unique insight and prescriptive understanding of data (Chen et al., 2012; Siemens and Long, 2011). Clearly, there is great demand for university graduates who can work with data in a meaningful way, thus the push by industry for big data and analytics-related curricula.

The accounting firm Price Waterhouse Cooper and others believe that to answer this demand, analytics should be integrated throughout a university student’s course of study. Business schools are reacting to these influencers by advancing various curricula (Colman and Paul, 2015). Some have suggested that academia is “behind the curve” in developing big data and analytics curriculum (Chiang et al., 2012; Chung, 2015). As big data and analytics curriculum develops to meet demand, it is important to analyze how curriculum changes and how what is taught improves student skills to match market needs (Gupta, 2000; Wagner and Ice, 2012). If academia is to provide better business analytics curriculum to support decision making, then new analytical materials and pedagogy development must occur (Wilder and Oزgur, 2015).

The SAP University Alliances program supports big data and analytics education by provisioning a variety of SAP products such as Business Warehousing (BW), Business Objects Explorer, Lumira, and Predictive Analytics to schools (SAP Predictive Analytics Product Tutorial, 2015). The supporting materials for these software products often include hands-on exercises, case studies, and/or data from organizations operating in the big data and analytics environment (Baxter, 2014). With the increase in the number of universities offering degrees related to big data and analytics, the curriculum provided by SAP University Alliances continues to grow and change to match the educational and employment markets. Observing these incremental changes, we questioned how curriculum adoption and usage has changed in recent years and sought to analyze how schools teaching big data and analytics in this program were adapting their coursework and provide guidance to those faculty seeking to include big data and analytics in their degree programs. The remainder of this paper is organized in the following manner. First, we discuss the big data and analytics curricula currently provided by SAP to member schools. Then, we analyze curriculum adoption and usage to see how school programs are changing. We follow with a big data and analytics research and curriculum development agenda and close with a conclusion.

Analytics curricula
Today’s SAP University Alliances curriculum is changing to include big data and analytics pedagogical objectives. This change is supported by professors in the program developing and sharing much of the global curriculum offered to the academic community. In addition to curriculum independently developed by faculty, there is also a good amount of SAP sponsored and developed curriculum. These big data and analytical courses help students understand how to analyze and model big data, create reports, visualize data, and predict or prescribe actions for organizations. Newer cloud solutions, in-memory database management systems, data visualization tools, and data analysis solutions bring in more current forms of content to supplement student learning. Flipped classrooms and massive open online courses (MOOCs) are also changing the way students study and learn. Large ERP systems require hosting whereas newer software such as data visualization (SAP Lumira) or advanced analytics (SAP Predictive Analytics) can be downloaded free of charge and installed on the user’s desktop. Software and courses in big data and analytics fall into five categories: data modeling and extraction, transformation and loading, (DM/ETL); analytical reporting tools; data visualization software; predictive analytics courseware; and big data curricula.
SAP began as an enterprise resource planning (ERP) system so its data modeling and warehousing tools are most applicable for the purpose of extracting transactional and aggregated business data. For this purpose, SAP Business Explorer (BEx) Query Designer and SAP NetWeaver BW (Business Warehouse or BW) are used as the first step in much of the analytics curriculum. The DM/ETL tools get the data stored in the system, prepared, and manipulated for subsequent purposes. These courses have a significant data staging process using data warehouse and query development software. Other data sources include Teradata, spreadsheet files, and various databases.

**Analytical reporting tools**

Reporting tools include software that allows the user to extract information from data and provide either a printed and/or a screen version of a report. Options for software products that provide reporting and dashboard capabilities when using large data sets include Business Objects Analysis for Office, Business Objects Design Studio, and Crystal Reports/Crystal Dashboards. These tools allow students to extract, aggregate, and display information in professionally acceptable formats. In some areas, students participate in dashboard development competitions to see who can design dashboards that explain and display information in the most pleasing and informative fashion. Thus, analytical reporting software provides an interesting introduction for students beginning studies in analytics.

**Data visualization software**

The third category of analytical software supplied by SAP is data visualization tools. Visualization tools typically are used to deliver some visualization of the data in an organized fashion allowing users to glean meaning from large data sets quickly and easily. The predominant data visualization tool in this group is SAP Lumira and an introduction to data visualization curriculum is available. There are several other classes dealing with big data and analytics visualization including those based on cloud versions of Lumira and BusinessObjects Explorer that have been shared by faculty.

**Predictive analytics courseware**

Analytical tools offer the greatest benefit but also require more work in terms of specification, query, modeling, and statistical programming. Organizations use predictive analysis to gain insight and expose opportunities by building models. Predictive analytics goes beyond reports and visualization by mining the data for in-depth understanding. Data preparation, visualization, automated analysis, application development, and model management are included in predictive analytics. SAP University Alliances offers faculty-developed courses in predictive analytics using a variety of data sources and hands-on data modeling tools. Software products are available on desktop, hosted, and in the cloud depending on faculty choice and curriculum adopted. Professors in the University Alliances have access to tools for predictive analytics, including SAP Predictive Analytics and SAP HANA. Because these are relatively new products, the curricula for these big data and analytics tools are still developing.

**Big data curricula**

The availability of inexpensive RAM supported the development of in-memory database management systems using columnar databases that perform real-time analytics and lightning fast queries with reduced data aggregation, caching, and hard drive storage. Big data software tools such as SAP HANA bring together an in-memory database management system, an application processing system, and the integration of analytical services on a single development platform.
Recently these tools have started integrating with open source software such as R and Hadoop, geospatial data providers such as ESRI, multiple data sources such as Teradata and others, and can include time series forecasting, outlier detection, trend analysis, classification analysis, segmentation analysis, and affinity analysis (SAP Predictive Analytics Product Tutorial, 2015). SAP University Alliances has supported SAP HANA with curriculum based around the development platform. For more information about faculty training opportunities and other resources, see Appendix.

In addition to tools for big data, there is also a variety of curriculum available for teaching big data concepts in a hands-on approach to cover the five Vs; volume, variety, velocity, veracity and value (Granville, 2014). For example, the volume concept of big data is taught with University of Arkansas hosted data sets that are made available to SAP University Alliances members. One of the most popular is the Sam’s Club data set, which contains point-of-sale data in six tables and over 55 million rows. Courseware on data mining this set allows students to look into shopping basket analysis for commonly purchased items, as well as explore retail returns fraud, and critically examine the veracity of the data. Here students can analyze data for outliers to detect if there are suspicious returns occurring or if the outliers are due to mistakes in the system or unusual situations, such as a new store being initially setup and thus having anomalous sales ratios.

Other curriculum uses the Tyson Foods data set from the University of Arkansas and requires students to examine profitability indicators, discard or explain data anomalies within the data set, and make recommendations to improve profitability. The Tyson Foods cube contains nearly 12 million rows of customer transactions for two years and more than 13,000 products shipped to 92 different sales districts in the USA.

In the past, the big data concept of variety has been covered in an ad hoc manner by individual instructors. Recently courseware has been introduced using spatial data and analysis tools of SAP HANA as well as map service providers such as Google, ESRI, and Galileo. Other variety of data courseware uses UI5 and the SAP HANA application development platform to blend map information and weather data feeds or analyze text information from social media feeds using sentiment analysis. Even more diverse data types and sources are used in the SAP DataGenius contest, where students, as well as users work with sports, scientific, geopolitical, and business data to tell a data-driven story using SAP Lumira or SAP Predictive Analytics. There is now a growing repository of data sources and examples available at https://ideas.sap.com/DataGenius.

Many of these preceding examples use Data-at-Rest, which makes teaching the big data concepts of velocity and value a challenge. These require not only Data-in-Motion, but also actionable data from which decisions can be made and consequences learned from outcomes. For this, the HEC ERP Simulation Game (ERPsim) curriculum is ideal (Léger, 2006; Léger et al., 2011). Here students work in teams in a business simulation where planning, procurement, production, logistics, sales, and financial processes are accomplished in an ERP system that includes a competitive market simulating the passing of one day every 60 seconds. Student teams make operational decisions in real time with data streaming at high speeds, providing an effective way to teach the velocity concept of big data. To cope with velocity, dashboards use calculated key metrics and targets – such as inventory turnover and profit margins. Well-designed dashboards provide a lot of value, a big data concept easily spoken to, but difficult to internalize without seeing the consequences of both good and bad data-driven decisions. Participants in the HEC ERPsim game, learn the value of data and all the analytics tools, both during the game at high velocity and between 12 simulated rounds when there is more time for detailed analysis to drive their actions in terms of what to make, how to market it, and what logistics have to be executed. As in the business world, consequences of these decisions are seen in the financial statements of these simulated businesses.
Comparing curricula adoption and usage

One goal of this paper is to compare curricula usage by member schools in recent years, to see how adoption of analytics-related curriculum is changing. In this section, we discuss our research methodology and present insight into trends in analytics-related curriculum adoption.

Methodology

The authors used course requests from the North American program hosting centers, which included course title, description, and the number of student logons requested by course. We coded data by academic year and term following the convention that an academic year such as 2011/2012, consisted of the Fall 2011, Spring 2012, and Summer 2012 semesters or terms. These tables were then merged to retain as much information as possible. We performed a data cleansing process to filter non-class related requests and eliminate outliers. The final data set used in our analysis consisted of 14,889 course requests which we felt confident represent actual curriculum usage for members over the four-year period.

Analysis of curriculum trends

Course data usage seen in Figure 1 compares curricula categories over four years. For a baseline, we used the number of students in the Introduction to ERP category – which grew 142 percent from 57,045 to 138,186 students. When comparing ERP to BW and Business Objects curricula, a similarly increasing trend is seen. Since it is common for students to register for multiple courses in the same semester, it is difficult to assess the exact number of unique students.

Another way to examine changes is to look at growth in the number of institutions using the curricula instead of the number of students. The typical Introduction to ERP course contains a large number of students per section when compared to big data and analytics classes. In Figure 2, one can see a side-by-side comparison of the adoption of analytics curriculum (such as BW and Business Objects) increasing over the four academic years from 2010/2011 to 2013/2014. Based on these data, by 2014 roughly 29 percent of all institutions have adopted BW analytics curricula, 23 percent had adopted Business Objects

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BW</td>
<td>5,267</td>
<td>6,455</td>
<td>8,300</td>
<td>8,821</td>
</tr>
<tr>
<td>Business Objects</td>
<td>1,875</td>
<td>4,042</td>
<td>6,251</td>
<td>6,431</td>
</tr>
<tr>
<td>HANA No.1</td>
<td></td>
<td></td>
<td></td>
<td>1,832</td>
</tr>
<tr>
<td>HEC ERPsim</td>
<td>22,821</td>
<td>30,259</td>
<td>32,271</td>
<td>48,311</td>
</tr>
<tr>
<td>Intro to ERP</td>
<td>57,045</td>
<td>76,481</td>
<td>105,789</td>
<td>138,186</td>
</tr>
</tbody>
</table>

Figure 1.
Number of students by curriculum category
curricula and 45 percent had adopted ERPsim – used by some to teach big data analytics in addition to business process integration.

It should be noted that these curricula are based on full membership hosted systems and do not represent desktop or cloud-based usage, which may be more easily adopted by individual faculty members since the cost is less and setup effort may be reduced. Increases in desktop and cloud-based usage would occur above and beyond hosted systems, thus making our data conservative.

A research agenda for big data and analytics curricula

To support curriculum development, we turned to business to inform our research agenda. Business leaders have identified business processes as areas where big data analysis can provide significant returns. Organizations are no longer looking to standardize and control repetitive work but to use the right analytical approaches to optimize outcomes and reduce risk (Cantara, 2014). Obviously then analytical approaches and contextual intelligence lend themselves to curriculum development.

According to an Accenture (2014) survey of 1,000 companies, the areas where big data analysis provide greatest results are: improvement in customer service and demand fulfillment, faster and more effective reaction time to supply chain issues, increase in supply chain efficiency, greater integration across the supply chain, optimization of inventory and asset productivity, and more effective operations process and decision making. Companies that embed big data analytics into their operations are more likely to generate business process benefits than those that do not (Accenture and General Electric, 2015). Research into operational efficiencies therefore should provide content for course creation.

Clearly, the gap between technological capabilities and the demand for individuals who can use these technologies provides academia a perfect opportunity to educate students to take full advantage of the new big data capabilities that will launch their organization into more informed decision making and competitive advantage in the areas related to business
Researchers in this area should look to non-ERP data sources and invoke more non-structured data in their work. Of particular note is the limited curriculum currently available for business processes other than sales. Most of the current big data and analytics curricula focus on sales-related data and do not address analytical questions related to other use cases such as operations, human resources, production, consumer sentiment, or customer churn, for example.

To develop a comprehensive educational program, faculty need to craft curriculum in key areas of business process management. In order to do this, the faculty will need access to data from non-traditional sources such as social media and web pages in conjunction with large sets of high-quality transactional data. Although some sources exist for non-transactional and unstructured data, shared and donated data are sometimes older and therefore less relevant. For example, UC Irvine maintains and shares data sets at its site http://archive.ics.uci.edu/ml/datasets. The Amazon data set at this site is dated from 2011 and while useful for teaching analytics techniques, data may not be as relevant for contemporary process analysis. The shortage of non-traditional data presents an opportunity for faculty in the more technical fields such as management information systems or computer information systems to develop curriculum that scrapes or otherwise acquires relevant data that can be handed off to faculty in various business process areas such as supply chain, accounting, production, and so on to create big data analysis curriculum. Acquisition of transactional data has also been problematic in the past as it is often difficult to convince real-world businesses to divulge their operating information. As a result, academics may need to resort to contrived data that may or may not represent real life scenarios. So the first step in development of the curriculum to meet industry needs is to locate realistic data sets in important business process areas. These data can readily be housed in an SAP ERP instance or an SAP BW at the University Alliances hosting sites to provide access to University Alliances member faculty and students. The University of Arkansas has already provided University Alliances members access to big data sets at its enterprise site. Sam’s Club and Tyson Foods were mentioned previously and contain sales-related data as do the Hallux and Dillard’s Department Store info cubes. Also available are over a million rows of detailed demographic data provided by Acxiom that could be used for supply chain and marketing big data curriculum. Potentially, data may also be collected from the more advanced ERPsim simulations that contain more robust operational scenarios.

Assuming availability of data sets, faculty will need to turn their attention to curriculum development for analysis of the data to examine efficiencies within internal operations and throughout the supply chain, asset management, plant management, customer service, and all other business processes. The sooner these curricula are developed, the quicker academia can help to fill the gap between existing industry big data and analysis skill sets and the demand for these skills now and in the near future.

Beyond the analytics aspects there are other competencies which were mapped in a recent study on data literacy at Dalhousie University (Ridsdale et al., 2015). Here, data literacy was defined as “the ability to collect, manage, evaluate, and apply data, in a critical manner” in a knowledge synthesis project. After a thematic review of competencies, knowledge, tasks, and skills related to data literacy, the researchers developed a framework illustrated in Figure 3, showing the knowledge areas and competencies. This was based on a much larger set of 64 tasks from 32 articles on the topic that were classified by the researchers to formulate a map for data literacy found in Appendix of their report.

This approach to mapping competencies is complementary to our topic of big data analytics and we should use it as a foundation to expand upon and include competencies around technologies, how they function, when to utilize what, and what management issues are associated around the big data issues. Applying the 5Vs of big data could be a starting point, to see how issues around volume, velocity, variety, veracity, and value have particular tasks, skills, and competency requirements. It will be worthwhile exploring this in future
research aimed at extending the data literacy to big data literacy in business. Another direction to expand could be to look at business knowledge, as the context is often critical to proper design and interpretation of analytics results. Knowledge of the business processes that generated the data as well as having an understanding of the operational features and limitations of technologies used to collect, store, and process the data is another dimension for big data literate people in business and therefore a topic for curriculum development.

Conclusion and implications for future research
The increased use of data analytic software in classes may be due in part to recommendations by industry partners and professional organizations to include analytics in curriculum. For example, the big four accounting firms have published white papers and promoted events to encourage faculty to incorporate more technology and particularly big data and data analytics into their courses. In addition, educational accrediting entities are pushing business schools to adopt analytics and other technologies, which might also explain the increase in introductory ERP curriculum adoption. As an example, the Association to Advance Collegiate Schools of Business mandates in its standard A7 that:

Consistent with mission, expected outcomes, and supporting strategies, accounting degree programs include learning experiences that develop skills and knowledge related to the integration of information technology in accounting and business. Included in these learning experiences is the development of skills and knowledge related to data creation, data sharing, data analytics, data mining, data reporting, and storage within and across organizations (AACSB, 2014).

In 2015, the American Accounting Association (AAA) held its first “Accounting IS Big Data” conference and has held numerous webinars to assist faculty in embedding analytics into their courses. With 7,000 members, the AAA has significant influence over course content at the university level. Other professional organizations such as the AIS educators have been promoting data analytics even longer.
The SAP University Alliances program is one of the largest faculty-driven communities offering competitive advantage to those schools ramping up big data and analytics degrees, certificates and programs. Obviously, the greatest advantage of the program is the faculty-developed training, curricula, and SAP supported conferences for faculty collaboration. This makes the SAP Community Network a viable place to look for course development support. There are many places to find big data and analytics curricula; however, few programs provide faculty and students the needed software, training, and course content, along with a highly supportive community of peers engaged in determining how to teach big data and analytics the best way possible.

The course offerings of the SAP University Alliances have increased over the last two to the three years with more modern courses being adopted including big data and analytics. Already by mid-2014, 29 percent of all Americas institutions used one or more of the big data or analytics curricula provided via membership, and 45 percent used the business simulation ERPsim where the velocity and value of data may be learned. Over the next several years, it is anticipated that the number of universities adopting these tools will increase dramatically given increasing demand for new hires knowledgeable in big data and analytics. More data scientist, business analyst, and other analytics supported employment opportunities will continue to lure schools into the big data and analytics domain.

While we have discussed many currently available big data and analytics curriculum offerings, we also identified opportunities where more curricula development and research can be done. We recognize the limitations and constraints encountered. This work provides a limited data perspective and a lack of knowledge or insight into non-hosted membership use. The period of analysis is relatively short, only having access to four years of data. In addition, there are a very small number of schools that are self-hosted in the Americas. While their numbers are counted in the hosted environment, many offer special classes that may not appear in the data.

There remain gaps to fill if universities are to meet the needs of industry talent in terms of the scope and scale of big data analytics in business knowledge. These must address the technologies for big data analytics, the generation and storage of data, as well as their application to business problems. Examining how the 5 Vs can be applied to the data literacy competencies (Ridsdale et al., 2015) is one such starting point. Examining how big data analytics can be applied in business process management curriculum is another example of what should be addressed in the future business school classrooms. Moving forward in this field will require collaboration between industry, academics, and the developers of technologies for big data analytics.

References


Granville, V. (2014), *Developing Analytic Talent: Becoming a Data Scientist*, John Wiley & Sons, Indianapolis, IN.


Further reading

Appendix

Faculty training opportunities in big data and analytics
While there are a number of big data and analytically related courses offered by SAP University Alliances, there are also faculty development opportunities three or four times a year at conferences and summer/winter workshops. Given the importance and present demand, faculty have developed materials and curriculum for the workshops and shared it via the SAP Community Network. These hands-on classes allow faculty to receive the same training and courseware that their students might receive in the classroom (see “Analytics functions covered in train the teacher workshops”). The focus of the most recent four-day long business analytics class included the following curriculum which was compiled and developed by faculty teaching these train-the-trainer workshops.

Analytics functions covered in train the teacher workshops.
Curriculum:

- dashboards and mobile apps in Design Studio;
- data mining using SAP BW and BEx to Discover Retail Fraud;
- Lumira Infographics, Visualization and Animation;
- Multidimensional Analysis of Teradata from the University of Arkansas;
- Multidimensional Analysis using SAP Business Objects for Office;
- Predictive Analytics: Automated Analytics (KXEN Infinite Insight);
- Predictive Analytics: Expert Analytics and data mining with R plugins;
- SAP Explorer data exploration for insights; and
- Web Scraping.

Each of these topics supports big data and analytics training in the classroom. These faculty-developed materials allow colleagues to the use analytics in the classroom with students.

Other resources for analytics
Aside from the global curriculum and the instructor adapted curriculum from the workshops, there are many other resources available through the SAP Community Network http://scn.sap.com. These include blogs, tutorials, and general information on a variety of big data and analytics applications, which are often moderated by Mentors who are SAP customers that have taken on a community leadership role to share their expertise in using different analytics tools in their workplaces. These also contain tips for data sets and links to free demos that can easily be assigned as class projects. For those looking for an overview of big data and analytics software trial versions in one place, http://go.sap.com/product/analytics.html is a good place to start.

Complementing the SAP Community Network blogs are several excellent product tutorials originally designed by SAP for customers to use for training purposes. These are available for faculty searching for the latest tutorials and are very handy when preparing a class or to assign students to get hands-on experience with different big data and analytical tools and methods. Table AI shows the available tutorials and associated links.
Additional instruction using big data and analytics can be obtained via MOOCs through OpenHPI and OpenSAP. These courses are available in either live mode where one participates in weekly assignments to get a grade and certificate or in recorded mode where one can simply view the lectures on demand. These are relatively new resources, but highly effective ways of learning what is possible and how to use the tools in a variety of business scenarios. Examples of recent analytics-related classes via OpenSAP (https://open.sap.com/courses) include “Text Analytics with SAP HANA Platform,” “Driving Business Results with Big Data,” “How the Internet of Things and Smart Services Will Change,” and “BI Clients and Applications on SAP HANA,” and from OpenHPI (https://open.hpi.de/courses) “In-Memory Data Management,” “Automated Visual Software Analytics,” and “Knowledge Engineering with Semantic Web Technologies.”

In addition to the MOOCs, there are self-paced video tutorials on analytics applications at www.youtube.com/user/saphanaacademy. Here, there are over 800 videos covering a variety of use cases and solutions using many different big data and analytics tools. What is particularly valuable about this resource is that it is comprehensive, covering a wide variety of topics. One can learn very advanced tools at the SAP HANA Academy. For example, one can learn about creating dashboards in SAP Lumira, creating real time social media analytics tools, or how to build analytical apps for mobile devices in SAP HANA Cloud Platform.

For more technical students in Computer Science or MIS, there are also developer resources and cloud system access through the student portal of the HANA Cloud Platform http://hcp.sap.com/students.html, which features additional self-study tutorials and learning objects that are very useful for student projects. Students can even publish their apps to the SAP App Center where businesses go to adopt their analytic solutions www.sapappcenter.com/. This can be a very interesting way to get noticed by potential employers.

This publication contains references to the products of SAP AG. SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, Clear Enterprise, SAP BusinessObjects Explorer, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP France in the USA and in other countries.

SAP AG is neither the author nor the publisher of this publication and is not responsible for its content. SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

### Tutorial links

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAP Predictive Analytics and SAP InfiniteInsight</td>
<td><a href="http://scn.sap.com/docs/DOC-32651">http://scn.sap.com/docs/DOC-32651</a></td>
</tr>
<tr>
<td>SAP Lumira</td>
<td><a href="http://scn.sap.com/docs/DOC-26507">http://scn.sap.com/docs/DOC-26507</a></td>
</tr>
<tr>
<td>SAP BusinessObjects Business Intelligence Suite</td>
<td><a href="http://scn.sap.com/docs/DOC-7725">http://scn.sap.com/docs/DOC-7725</a></td>
</tr>
<tr>
<td>SAP BusinessObjects Dashboards 4.x (Crystal Dashboard)</td>
<td><a href="http://scn.sap.com/docs/DOC-7946">http://scn.sap.com/docs/DOC-7946</a></td>
</tr>
<tr>
<td>SAP Crystal Reports</td>
<td><a href="http://scn.sap.com/docs/DOC-8514">http://scn.sap.com/docs/DOC-8514</a></td>
</tr>
<tr>
<td>SAP HANA</td>
<td><a href="http://scn.sap.com/docs/DOC-34403">http://scn.sap.com/docs/DOC-34403</a></td>
</tr>
</tbody>
</table>

### Table AI. Analytic product tutorials

**Corresponding author**
Alexander J. McLeod can be contacted at: am@txstate.edu

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
A big data framework for facilitating product innovation processes

Yuanzhu Zhan and Kim Hua Tan
Business School, University of Nottingham, Nottingham, UK
Guojun Ji
School of Management, Xiamen University, Xiamen, China
Leanne Chung
Business School, Cardiff University, Cardiff, UK, and
Minglang Tseng
Department of Business Administration, Lunghwa University of Science and Technology, Guishan, Taiwan

Abstract

Purpose – The purpose of this paper is to suggest how firms could use big data to facilitate product innovation processes, by shortening the time to market, improving customers’ product adoption and reducing costs.

Design/methodology/approach – The research is based on a two-step approach. First, this research identifies four potential key success factors for organisations to integrate big data in accelerating their product innovation processes. The proposed factors are further examined and developed by conducting interviews with different organisation experts and academic researchers. Then a framework is developed based on the interview outputs. The framework sets out the key success factors involved in leveraging big data to reduce lead times and costs in product innovation processes.

Findings – The three determined key success factors are: accelerated innovation process; customer connection; and an ecosystem of innovation. The authors believe that the developed framework based on big data represents a paradigm shift. It can help firms to make new product development dramatically faster and less costly.

Research limitations/implications – The proposed accelerated innovation processes demand a shift in traditional organisational culture and practices. It is, though, meaningful only for products and services with short life cycles. Moreover, the framework has not yet been widely tested.

Practical implications – This paper points to the vital role of big data in helping firms to accelerate product innovation processes. First of all, it allows organisations to launch new products to market as quickly as possible. Second, it helps organisations to determine the weaknesses of the product earlier in the development cycle. Third, it allows functionalities to be added to a product that customers are willing to pay a premium for, while eliminating features they do not want. Last, but not least, it identifies and then prioritises customer needs for specific markets.

Originality/value – The research shows that firms could harvest external knowledge and import ideas across organisational boundaries. An accelerated innovation process based on big data is characterised by a multidimensional process involving intelligence efforts, relentless data collection and flexible working relationships with team members.

Keywords Big data, Accelerated innovation, Key success factors, Product innovation processes, Rapid innovation

Paper type Research paper

1. Introduction

Big data is attracting considerable attention worldwide. According to Davenport (2013), the effective use of big data has underlying benefits in bringing about dramatic cost reductions and substantial improvements in the time required to perform a computing task, or in new
product innovation and service offerings. Taking advantage of valuable knowledge beyond big data will become the basis for competition for today’s enterprises (Barton and Court, 2012; Wamba et al., 2015). Several researchers have pointed out that big data can enhance firms’ innovation capabilities in many respects (Manyika et al., 2011; Gobble, 2014; Tan et al., 2015). For example, Manyika et al. (2011) report that predictive modelling using big data can cut three to five years off the approximately 13 years that pharmaceutical companies generally need to bring a new drug to market. Capgemini (2012) estimates that the process improvements enabled by big data may lead to an average 26 per cent performance improvement over a three-year period. What is more, the analysis of big data may have huge operational and strategic impact on business process innovation at the firm and supply chain levels, and therefore allow firms that adopt it to achieve competitive advantage.

Today, the global market is in a product “war” and management of innovation is the strategic weapon. According to Cooper (1990), product innovation – the development of new and improved products – is crucial to the survival and prosperity of the modern business. A new product is usually defined as one that has been on the market for three years or less and that, in the customer’s view, is visibly different from previous offerings, with new features, functionality or performance characteristics (Cooper and Kleinschmidt, 2011). A company’s development of new products can be much more quickly and efficiently helped with a bit of planning before development starts (Cooper, 1990). Facing increased competition from home and abroad, maturing markets and the heightened pace of technological change, corporations look to new products and business for sustained growth and competitive advantage. Reports and surveys identify that most companies count heavily on new product development for growth and profitability (LaValle et al., 2011). However, an empirical study has in fact shown that fully commercialised new products have a remarkable failure rate of 40-50 per cent, and this performance has not changed much over the past 20 years (Cierpicki et al., 2000). Therefore, the need for effective product innovation in organisations demands immediate attention. Ortt and Duin (2008) point out that current systematic innovation approaches are lacking in enough market focus and some are becoming too complex to manage efficiently and effectively. Some researchers also argue that current innovation approaches are too time-consuming; as well as having too many time wasters and too much cost ineffectiveness, some of them are bureaucratic and have no provision for focus (Cooper, 1994). Researchers believe that a good product innovation process should be adaptable, provide companies with a much more efficient roadmap, bring products to market faster and improve the use of scarce resources (Sheu and Lee, 2011; Wooder and Baker, 2012). Clearly, there is a lack of an effective way to support organisations to utilise big data and drive new product innovation from idea through to launch. With big data, firms can extract new ideas or understanding about their products, customers and markets, which are crucial to innovation. However, how could organisations use big data to better facilitate their product innovation? Product innovation is a process. And like other processes, innovation can be managed (Cooper, 1990; Ortt and Duin, 2008). The first and most important step is to first understand the key success factors – those factors that make the difference between winning and losing at product innovation.

This paper seeks to develop a big data framework which can facilitate organisations’ product innovation processes. To assist our understanding of harvesting big data to facilitate product innovation, this study is structured into three parts. First of all, this paper investigates how big data can be used to transform the development of new products, by shortening the time to market, improving customers’ product adoption and reducing costs. Then, drawing on a systematic literature review, this research identifies key success factors that help broaden the understanding of the most significant success factors in product innovation. Second, a series of interviews with leading academics and subject matter
experts from a number of industries and disciplines are conducted to improve and examine the key factors. Third, based on the literature and interview output, a big data innovation framework is developed and demonstrated. Finally, the implications for practitioners and academia are discussed and conclusions drawn.

2. Literature review

2.1 Big data transforming product innovation

Tan et al. (2015) define big data as a holistic approach to managing process; they analyse the 3Vs (volume, variety, velocity) in order to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantages. In 2000, only 800,000 petabytes (PB) of data were stored in the world (IBM, 2013). It is expected this number will reach 35 zettabytes (ZB) by 2020 (Wong, 2012). The explosion of data is a natural tendency and, if harvested properly, can provide companies with better product innovation. For example, Dell initiated the development of a database that includes 1.5 million records related to sales and advertisements (Davenport, 2006) and Tesco generates more than 1.5 billion new items of data every month to support their new product development (Manyika et al., 2011). Thus, product innovation can be facilitated by acquiring amounts of information from different sources to develop better innovation processes, and to quickly find out the market acceptance of new products, customers’ needs or even competitors’ market movements. It also provides organisations with big ideas which could lead to big concepts and big solutions – the growth engines of the future (Wamba et al., 2015). In short, it helps organisations to generate valuable insights, better decision making and finally achieve competitive advantage co-creation and realisation.

Moreover, there are many different types of data, such as texts, weblogs, GPS location information, sensor data, graphs, videos, audio data and more online data. Besides, data have become complex because the variety has shifted from traditional structured data to more semi-structured and unstructured data – from search indexes, e-mails, log files, social media forums, sensor data from systems, and so on (Zikopoulos and Eaton, 2011). In the digital economy, a firm’s success will rely on its ability to draw insights from the various kinds of data available to it, which includes both traditional and non-traditional. The ability to analyse all types of data will create more opportunity and more value for an enterprise (Dijcks, 2013; IBM, 2013). Big data analytics can integrate heterogeneous resources and tools from multi-disciplines to gain great advantages; these include increasing operational efficiency, informing strategic direction, developing better customer service, identifying and developing new products and services, identifying new customers and markets, etc. (Zhang et al., 2011; Chen et al., 2012; Lohr, 2012; Demirkan and Delen, 2013). For example, Tata Motors analyse four million text messages every month, spanning everything from product complaints to reminders about service appointments to announcements about new models, as well as connecting these with customer satisfaction polling (Agarwal and Weill, 2012); Procter and Gamble created a group consisting of more than 100 analysts from such functions as operations, supply chain, sales, consumer research and marketing to improve total business performance by analysing interrelationships among functional areas (Davenport, 2006). Therefore, big data is pushing traditional operations and product innovation to a higher generation, which can be more adaptable to complex situations, and also self-adjusted to changing conditions and unstable information to satisfy a wide range of customers (Zhong et al., 2016). Instead of collecting customer feedback via formal questionnaires, new product innovation relies more on mobile devices, social media platforms – including YouTube, Facebook and Twitter – and the internet in order to build better customer connections and get feedback quickly at a reduced cost.

In terms of velocity, huge amounts of data are generated every second and increasingly have a very short life (Xu et al., 2013; Zhong et al., 2016). In 2011, about four billion mobile
phone users were identified worldwide; about 12 per cent of them using smartphones and having the capability of turning themselves into data streams. Meanwhile, the video platform YouTube received 24 hours of video every 60 seconds (The Economist, 2011). On Facebook alone we send ten billion messages including photos and videos per day; we click the “share” button 4.5 billion times and upload 350 million new pictures each and every day (Thibeault and Wadsworth, 2014). In these circumstances, firms can easily track customers’ data, including clickstream data from the web, and can leverage details from their behavioural analysis to better support their new product innovations. For example, Amazon manages a constant flow of new products, suppliers, customers and promotions without compromising guaranteed delivery dates (Davenport, 2006). The velocity of big data can drive new product development dramatically faster and at less cost through responding to market feedback in a short time. Firms are now capable of gathering users’ feedback in near real time to track changes in customer behaviour and rapidly communicating this to the R&D team to ensure that a newly launched product is sufficiently flexible to incorporate new functionality quickly.

What are the most important success factors for harvesting big data in product innovation? Although big data can offer much more useful production innovation and can gain great competitive advantages, there are no success factors for managers to support organisations’ product innovation from idea through launch based on the values captured from big data. Therefore, instead of just generating vast amounts of information from big data, managers need success factors as guidelines to structure and utilise the information captured from big data to support their product innovation systematically, so that a better insight into the issue being analysed could be gained. Previous research shows that there are many existing success factors that organisations could implement to support product innovation (Cooper, 1990, 2014; Cooper and Kleinschmidt, 2011; Balbontin et al., 1999). However, these traditional success factors are limited and not necessarily optimised for big data harvesting tasks due to their general purposes. With big data, firms can gain a better understanding of their products, customers and markets – and this is crucial to innovation (Manyika et al., 2011; Wong, 2012). The main challenge for firms is how to use big data to dramatically hasten the development of new products and to make them less costly.

2.2 Key success factors in product innovation
According to Cooper and Kleinschmidt (1996), product innovation can potentially be more successful if a number of factors are improved and implemented. Failing to do so can have disastrous results for a firm. In order to determine the key success factors for product innovation – based on values harvested from big data – this research draws on different researchers who have systematically and comprehensively summarised such key success factors for successful product innovation. A series of interviews with leading academics and big data experts from a number of industries and disciplines were then conducted to further develop and examine the key factors.

As determined by previous research, the most important recognised factors in organisations to support product innovation are: pre-development research; an accelerated innovation process; customer connection; and an ecosystem of innovation (see Table I).

In terms of the pre-development research, Cooper (1994) points out that the seeds of success or failure are sown in the first few steps of the process (the pre-development stage). The pre-development activities are important because they qualify and define the project. Many projects are poorly defined when they enter the development phase. This is often the result of weak pre-development activities: the target user is not well understood, user needs and wants are vaguely defined and required product features and attributes are fuzzy. Those in R&D and design engineers are not mind readers. With a poorly defined project, they waste considerable time seeking definition, often recycling back several times to get the
Better project definition, the result of sound pre-development research, actually speeds up the development process. What is more, pre-development research up front encourages changes to occur earlier in the process rather than later, when they are more costly. The result is considerable savings in time and money at the back end of the project and a more efficient new product process.

The accelerated innovation process aims to use systematic methods to speed up the innovation process as much as possible. Speed yields competitive advantages: being the first on the market can result in a quicker realisation of profit, and there will be a lower risk that the competitive situation or market would have changed before the new product can be launched (Steinfeld and Beltoft, 2014). Systematic innovation is significant in the product innovation process; it can create value and secure competitive advantage for organisations by generating a series of innovations, rather than unplanned or haphazard activities (Mann and Jones, 2002; Sethi et al., 2001; Cooper and Kleinschmidt, 2011). Autonomy management is based on guaranteeing the freedom of the individual groups of employees to decide on basic issues; it can improve efficiency and increase the effects of employee job satisfaction and motivation to work. Cross-functional teams can improve integration and coordination, span organisational boundaries, and reduce the production cycle time in new product development (Cooper and Kleinschmidt, 2011). What is more, bringing people together from different disciplines can improve problem solving and lead to more thorough decision making, which makes it easier to achieve corporate goals and customer satisfaction at the same time (Sethi et al., 2001). Simultaneous development is a method of designing and developing products, in which the different stages run simultaneously, rather than consecutively. Applying such a method can result in great benefits for organisations, such as reduced overall programme costs, lower

<table>
<thead>
<tr>
<th>Key success factors</th>
<th>Essential elements</th>
<th>Literature support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-development research</td>
<td>Idea generation</td>
<td>1, 3, 5, 7, 8, 11, 12, 13, 15, 16, 17, 18, 19, 20, 22, 25</td>
</tr>
<tr>
<td></td>
<td>Initial screening</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Preliminary market assessment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Detailed market study</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Financial analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Well-defined product</td>
<td></td>
</tr>
<tr>
<td>Accelerated innovation process</td>
<td>Systematic innovation process</td>
<td>1, 2, 3, 4, 5, 8, 9, 10, 13, 14, 15, 18, 20, 21, 24, 26, 27</td>
</tr>
<tr>
<td></td>
<td>Autonomy management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-functional teams</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simultaneous development</td>
<td></td>
</tr>
<tr>
<td>Customer connection</td>
<td>Market orientation</td>
<td>1, 3, 4, 5, 6, 7, 9, 13, 16, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28, 29</td>
</tr>
<tr>
<td></td>
<td>Customer communication</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understanding of customers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Good relationship with customers</td>
<td></td>
</tr>
<tr>
<td>Ecosystem of innovation</td>
<td>Connection with customers and partners</td>
<td>1, 2, 3, 4, 7, 11, 14, 15, 16, 18, 19, 23, 24, 26, 27, 29</td>
</tr>
<tr>
<td></td>
<td>Proficiency of marketing test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fast development and launch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quick response to market</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Market and partner tests</td>
<td></td>
</tr>
</tbody>
</table>


Table I. Success factors for product innovation
manpower requirements, reduced potential risks, improved high-quality products and flexibility (Lovelace et al., 2001).

In terms of customer connection, a thorough connection with and understanding of customers is significant: the more the customer is understood, and the more that understanding is implemented in product design, the more positive the impact will be on market share, revenues and margins (Evanschitzky et al., 2012; Bohlmann et al., 2012). Therefore, a strong customer connection is critical to success in product innovation. With regard to customer connection, in terms of firms’ value creation in product innovation, this research identified four essential elements from literature (as Table I shows): market orientation; customer communication; understanding of customers; and keeping good relationship with customers. According to Chesbrough (2003), companies are increasingly rethinking the fundamental ways in which they generate ideas and bring them to market. Because R&D has long been a costly and inexact process (Anders and Ali, 2004), customer connection has been widely acclaimed in management rhetoric as a means to tighten the feedback loop between the cycles of consumption and production (Lundkvist and Yakhlef, 2004). Underlying most such views is the assumption that customers are sources of information and knowledge (Lacity and Willcocks, 2014) and that customer connection can enhance product concept effectiveness (Anders and Ali, 2004).

Adner (2006) defines an ecosystem of innovation as the collaborative arrangements through which organisations combine their individual offerings into a coherent, customer-facing solution. As one of the determined key success factors in product innovation, an ecosystem of innovation refers to building an innovative and market-testing environment that can support organisations to develop new products at dramatically fast speeds and with lower expenses. There are five essential elements identified in this factor, as Table I shows: connection with customers and partners; proficiency of marketing tests; fast development and launch; quick response to markets; and market and partner tests. Bogel et al. (2014) point out that the ecosystem allows organisations to generate a greater value that no single firm could have generated alone. It also bridges the gap between the need for new product definitions and the changeable market conditions as development proceeds (Gupta, 2013). For each project, instead of focussing on R&D internally, allocating resources externally from partners (e.g. with customers, universities or companies) can be far more effective because critical bottlenecks may reside outside the company (Lacity and Willcocks, 2014). Especially when enabled by information technologies that have drastically reduced the costs of coordination, innovation ecosystems have become a significant success factor in the growth strategies of organisations in a wide range of industries (Adner, 2006).

3. Methodology

So far, this research has presented a normative research approach to determine the key success factors for enhancing product innovation in a big data context. Starting from a systematic literature review, five key success factors were identified to support organisations in gaining competitive advantages by implementing product innovation successfully. In a second step, this research conducted qualitative research using semi-structured interviews (Hansen et al., 2009; Silverman, 2008) to validate the key success factors and develop a model based on the determined factors.

Using multiple interviews brought a richer portrait of each case (Yin, 1994) and also aided in mitigating bias in historical data interpretation. As previous research in the area of product innovation suggested addressing high-level corporate officers to gain the most accurate information (Hansen et al., 2009), this research conducted interviews with 15 high-level experts from leading organisations between May and September 2015. As Tables II and III show, these experts belong to two groups: corporate experts (eight executives) and research experts (seven senior researchers).
These organisations reflect a variety of different industry sectors (e.g. electronics, manufacturing, automotive, family apparatus). The diverse selection of interview partners guarantees a holistic perspective on the topic and uses very different experiences and opinions in the field of big data and product innovation. Each interview took approximately one-and-a-half hours and was recorded. The questionnaire itself was extensively pretested via personal interviews. Parts of the interview were translated into English, and then transcribed and analysed with qualitative data analysis.

4. Results
During the interviews we received a vast amount of feedback on the identified key success factors and their potential implementation in product innovation in a big data context. We recorded both broad agreements and controversial opinions towards the discussions. Table IV summarises the feedback obtained from the interviews of different organisation executives and research experts about the identified key success factors in using big data to support product innovation.

4.1 Accelerated innovation process
Overall, the managers and researchers felt that the key success factors are conceptually accurate in capturing the essence of product innovation in a big data environment. For the accelerated innovation process, we observed broad acceptance among organisations and research experts. By applying this concept, self-directed teams emerge as the way to drive innovation and deliver great products, and we observed that it can create high velocity in product innovation projects. Most interviewees commend autonomous teams as being more effective for addressing projects with high technological novelty or radical innovation, especially in a big data environment. Also, some interviewees found the cross-functional
teams and simultaneous management to be very “new to its [produce development’s] culture”, and can help communication more broadly, gain alignment more easily and build better products in a short time. It was highlighted that this factor can provide a better big data-supported product innovation approach:

By implementing the accelerated innovation process, people from different function departments are grouped together to work actively. It cuts across boundaries of different departments and there is no more marketing team or production team. Instead, every team member becomes involved in marketing, engineering, design, production or R&D. This movement can save us a lot of time and eliminated vast number of unnecessary double communication within various teams.

This observation adds to the fact that the elements of the accelerated innovation process are commonly known and integrated in numerous business and theoretical concepts.

4.2 Customer connection

Most of the organisation experts and researchers concurred with the customer connection factor. They referred to various examples of their current projects addressing the importance of customer connection in product innovation. One organisation’s marketing manager pointed out that the proposed customer connection factor provides more than merely an idea for product innovation. It supplies the firm with information on market needs or existing problems, product-related specifications or even a complete product design. It is generally believed by the interviewees that as one of the most important factors in product innovation, organisations need a process in place to pay close attention to their customers in every phase of the product innovation value chain, from idea generation to product development to marketing. According to firm F’s CEO:

Normally, companies just determine their main customers and potential customers, but big data will allow them to investigate more detailed aspects such as where they are, what problem they face, what they need, how they want to be contacted and when.

Therefore, the observation shows that customer connection is significant for every organisation for product innovation, and it can be further enhanced by harvesting values from big data.
4.3 Ecosystem of innovation

There was broad agreement on the innovation ecosystem concept among business executives and researchers, which underlines this concept’s high diffusion and acceptance. As a result, it costs the companies less time and money than would ordinarily be required, by concentrating on new product research and development rather than other non-value adding processes. According to the interviewees, by building a stable and diverse ecosystem of innovation, organisations can provide new products to meet their customers’ requirements in a much more efficient and effective way. One organisation executive outlined the ecosystem of innovation as a major business opportunity:

The company is already spending about a million dollars in cooperating with desirable partners among the entire supply chain to support its product development ecosystem and it helps the company gain an outsized competitive advantage such as offering the consumer a greater value than the competitors, as well as providing better products and services.

However, a few academic experts criticised the innovation ecosystem concept as being too simple. For example, it was argued that the concept is based on successful examples of agglomeration, whether in industrial, entrepreneurial, economic or geographic terms. As such, there is relatively little new about the innovation ecosystem compared with earlier concepts like development clusters or blocks. These diverse and sometimes controversial viewpoints underline the difficulties when trying to operationalise the concept of innovation ecosystem.

4.4 Pre-development research

Most of the interviewees were not satisfied by the pre-development research factor towards product innovation in a big data environment. Although many researchers argued that pre-development research is a key to success, most of the organisation experts in particular commented that it can be achieved in the early stage of customer connection via harvesting information from big data. As an R&D manager points out:

Companies should not spend too much time on pre-development research because people don’t know what they want until you show it to them.

Instead of spending a lot of time and resources on conducting pre-development research, big data can be used as the most important source in generating new ideas, capturing useful information, assessing target markets, introducing new product concepts and gathering feedback. Therefore, as pre-development research can be included as part of customer connection, it was removed from the identified key success factors.

4.5 Implementation in product innovation management

All interviewees outlined the advantage of the determined key success factors as an orientation framework for supporting product innovation in a big data context. Often, the key success factors were considered particularly helpful in organisations when employees needed to be introduced or sensitised to the concepts, such as autonomy and innovation ecosystem. Organisation experts initially criticised the lack of possibilities for quantitative methods/techniques to harvest big data, but later acknowledged the proposed key success factors’ role as a meta-method (Paterson et al., 2001). These observations underscore two facts. On the one hand, businesses seek accurate amounts of valuable information from big data as a basis to support their product innovation. On the other hand, large organisations may have already built up their big data toolbox of analytic techniques and methods to support product innovation, and hence they may not necessarily take advantage of the key success factors as a meta-method. Small- and medium-sized enterprises (SMEs), conversely, might not have enough resources to engage in the field of big data analytics. Here, the developed key success factors as a meta-method could significantly support organisations.
in providing guidelines for appropriate product innovation processes by harvesting values from big data. However, the particular value of the developed key success factors for SMEs is beyond the scope of this paper and is a suitable subject for further research.

To summarise, the developed key success factors are: accelerated innovation process, customer connection and innovation ecosystem, which gained significant support in all major elements of successful product innovation implementation in a big data context. These factors can be seen as an orientation framework and introduction to the field of big data, adding value to both big data and product innovation management.

5. Development of the framework
Based on the above research, a framework can be further developed to assist firms in product innovation through harvesting big data to shorten the lead time to market, improve customers’ product adoption and reduce costs (see Figure 1). It is termed the ACE framework because it is based on the principles of accelerated innovation process (A), customer connection (C), and ecosystem of innovation (E). We believe the proposed framework represents a paradigm shift. It can provide guidelines to firms in harvesting big data to better support their product innovation.

5.1 Accelerated process
One of the principles of the ACE framework is accelerated innovation process. Figure 2 shows the most important elements in achieving this.
According to Wynen et al. (2014), autonomy is the mother of motivation and creativity. The first principle in the proposed ACE framework is to give autonomy to the innovation teams. This means allowing R&D team members a high level of freedom to make decisions by themselves in their workplace. Autonomy here also implies that project teams work in parallel, rather than sequentially. At each stage of a project, many activities take place concurrently and involve different functions of the firm. Under autonomy management, a group leader is allocated to supervise the output of the project. The project approach begins with dividing the innovation process into many small elements. After that, the divided project activities are undertaken by cross-function teams (which mean a team of people from different functional areas) who work on different elements in parallel. By doing this, the so-called innovation “assembly line” can be accelerated and results can be delivered quickly (Davenport, 2013). Autonomy does not mean being separate: project teams need alignment with the core, using big data to share innovation portfolios as well as to cultivate a network of peers and relationships, to facilitate innovation (Chen et al., 2015).

In this situation, the innovation process is industrialised by assigning more people to the many small steps and project activities (Williamson and Yin, 2014). The total outlays for a given project can nonetheless be reduced, as these people are less highly trained than traditional R&D staff and are generally therefore paid less (McNeish and Hazra, 2014). For example, Lenovo overcomes the usual problems of implementation by: breaking down its product designs into separate modules linked by standardised interfaces; redesigning its software to be compatible across all activities associated with the new product; establishing short lines of communication where each team member can represent his or her respective functional department; and introducing open design processes where information is shared with the entire team as early as possible.

Big data plays a significant role. In terms of traditional innovation approaches, many companies have found it hard to implement the autonomy principle, because of barriers such as unwillingness by engineers to release information early and difficulties in coordinating multidisciplinary teams (Berglund and Sandström, 2013). Companies now can rely directly on big data to gather the latest information; team members are now working and living in a big data environment, which ensures their communication and knowledge sharing are both effective and efficient.

5.2 Customer connection

Barwise and Meehan (2012) believe that Apple built its success not as a pioneer, but as a good follower of its customers. The second cornerstone of the ACE framework is customer connection, i.e. a focus on building a close relationship with customers via big data (see Figure 3).

Innovation can be facilitated by evolving ideas while listening to the voice of customers; the product is better when potential customers can be identified and their needs satisfied (Prahalad and Ramaswamy, 2013; Steinfeld and Beltoft, 2014; Cooper, 2014). Many projects have poor customer connections, which results in a series of problems: customer

![Figure 3. Customer connection](image)
requirements and problems are vaguely defined; the product’s functions and features are fuzzy; and the target customers are not well understood (Dunn and Dahl, 2012). Engineers and R&D teams are not mind readers. With poor customer connections, they often have to back-track to make the product right. Thus, they waste considerable time in defining projects appropriately. This development process can be speeded up by building better customer connections. What is more, instead of making changes late in the project, customer connection encourages changes to occur earlier, when they are less expensive (Williamson and Yin, 2014). For example, the MIUI system developed by Xiaomi (now the world’s third largest manufacturer of smartphones) and the Talend big data platform by Lenovo (the largest personal computer vendor in China and the second largest in the world) are both good ways to build a close connection to customers. Additionally, Didi Dache (a young Chinese taxi service company, selected as one of China’s 13 most valuable start-ups) spent a lot of time building various platforms to connect to its users as well as the market (including using big data to clarify its product definition and to identify its main competitors, market size and customers’ problems and needs).

The involvement of customers is an emerging trend (Cooper, 2014; Dunn and Dahl, 2012; Steinfeld and Beltoft, 2014). The innovation process can be dramatically accelerated by using big data in the form of, for example, usage information, which is much more rapidly available than, say, the results of market surveys (Williamson and Yin, 2014). Big data in the form of feedback can be an important source of useful information and new ideas. Key questions need to be focussed at this stage, such as: who exactly is the target customer? What functionalities and features should be developed to give the product a differential advantage? What exactly should the product be to make it a winner? How should the product be positioned? By answering such questions, companies can gain a better understanding of their customers, markets and products.

5.3 Fast launch-and-improve ecosystem

The third cornerstone of the ACE framework is an innovation and market-testing environment to develop new products at dramatically fast speeds and lower costs. Adner (2006) argues that innovation ecosystems have become a core element in the growth strategies of organisations in a wide range of industries. The ecosystem element of the ACE framework indicates that the company network is used to acquire new requirements and components of product development processes externally or from intermediaries, in order to create a fast launch-and-improve environment that is able to launch a product quickly with less cost. A fast launch-and-improve ecosystem involves the concept of autonomy and customer connection. It helps the product team to move quickly to a market-winning product through a series of iterations: new product ideas, fast launch, feedback gathering, fast improvement and re-launch (see Figure 4).

The emergence of big data and the internet allow for the combination of organisations’ business strategies and those of outside suppliers within an ecosystem (Shih et al., 2014). For example, in order to further enhance its smartphone ecosystem and take ownership of
its future products roadmap, Lenovo Group in 2014 acquired world-renowned Motorola Mobility from Google Company, including the Motorola brand and Motorola Mobility’s portfolio of innovative smartphones. The acquisition of such an iconic brand immediately made Lenovo a powerful global competitor in smartphones, through scaling Motorola Mobility into a major player within its existing Android ecosystem and facilitating product innovation across the new Android ecosystem (Google Official Blog, 2014).

Enormous cost advantages can be acquired from the ecosystem and companies should identify the key competencies of their components and focus on them, and acquire the less important components from the ecosystem (Horn, 2005). For example, owing to its supportive ecosystem, Xiaomi is able to sell its products and a wide range of accessories at near-production cost to keep prices competitive and sell a large volume of goods. By integrating a range of key components and technologies, many high-class products can be invented and created. Producing such products in an open innovation ecosystem depends on contributions from across the network of suppliers and creates value for the eventual buyer (Boer et al., 2001).

The core competitive advantage of this ecosystem arises from the use of big data to attract and connect to a wide range of networks in each step of product development (McAfee and Brnjolffson, 2012). This might be through the presentation of mock-ups, images or videos of the new product to customers and thus the gathering of feedback early in the process of product development (Tuulennäki and Välikangas, 2011). For example, rather than spending time on internal R&D to make the product perfect, Didi Dache and Xiaomi tend to launch their new product ideas on the market quickly (and are able to do so through implementation of the autonomy principle) and then improve them through extremely fast and continuous rounds of commercial realisation and testing within their ecosystems. Hence, companies can earn a premium by staying abreast of competitors’ innovations and by having up-to-date products available in volume at affordable prices (Williamson and Yin, 2014). Moreover, nurturing interactions in the proposed ecosystem improves efficiency and creativity, and also makes innovation a cycle of continuous improvement and information transformation (Gupta, 2013). In the ecosystem, innovations are made from interrelated networks (Nieto and Santamaria, 2007) and these empower organisations to rapidly integrate useful feedback from customers and partners. Through repeated accelerated innovation cycles, project teams can iterate the product in sync with evolving market requirements and stay ahead of the competition. With innovation ecosystems, firms are better able to respond to today’s fluid, changeable information and evolving market conditions.

6. Discussion

The core competitive advantage of the proposed framework arises from the use of big data to attract and connect to a wide range of networks in each step of product development. This might be through the presentation of mock-ups, images or videos of the new product to customers and thus the gathering of feedback early in the process of product development (Tuulennäki and Välikangas, 2011). By implementing the framework, innovations are made from interrelated networks and these empower organisations to rapidly integrate useful feedback from customers and partners. Through repeated accelerated innovation cycles, project teams can iterate the product in sync with evolving market requirements and stay ahead of the competition. With the innovation ecosystems, firms are better able to respond to today’s fluid, changeable information and evolving market conditions (McAfee and Brnjolffson, 2012).

Traditionally, the new product development process involved inefficient sequential processing of information between functional specialties. The ACE framework allows firms to adapt and respond rapidly to changing market needs and to develop innovative products in
such an environment. Rather than spending years to exploit in-house capabilities, the ACE framework can be used to build a network to piece together production according to capabilities. Hence, it ensures the company remains at the cutting edge of product innovation. Proactive assessment of customer needs and behaviours is vital in today’s competitive environment (Brown and Bessant, 2003). The demand for intelligence on product defects, improvements and usage has never been greater, especially in high technology firms in the electronic and manufacturing industries (Salge et al., 2013). The accelerated approach is meaningful for products and services with short product life cycles, notably the consumer electronics industry, where demand is driven mainly by lifestyle trends. Moreover, the case also highlights that achieving innovation and flexibility requires considerable planning and coordination through the various phases of development. Thus, top-level management support through a product champion and tight interfacing with social media and the target market are essential components of accelerated innovation.

As Figure 5 shows, compared with traditional innovation approaches, the ACE framework places particular emphasis on efficiency and cost saving. There is no magic formula for innovation. However, firms could expand their existing innovation competence in many ways by tapping into the knowledge afforded by big data. The ACE framework provides a blueprint for using big data to make product innovation dramatically faster and less costly. By using the ACE framework, firms are leveraging big data analytics to embed customer sentiment in product development. This enables firms to move away from product-focussed innovation and to turn their attention to innovation around the customer experience. The proposed paradigm-shifting innovation processes enable firms to find ways to innovate – to make new product development dramatically faster and less costly.

7. Conclusion
The ACE framework proposed in this research is based on the key success factors elicited from the literature; these success factors were validated by conducting interviews with eight different leading organisations and seven university researchers. The framework presented in this paper can facilitate better planning and organisation of parallel work teams and groups that may be involved in rapid new product development. This paper extends the accelerated innovation boundaries pointed out by Williamson and Yin (2014), and provides further evidence of the vital role of the innovation ecosystem in new product development.

This paper also points to the vital role of big data in helping firms to accelerate innovation. The incorporation of big data into the fast launch-and-improve ecosystem can be significant. It allows organisations to launch new products to market as quickly as possible. What is more, it helps organisations to determine the weaknesses of the product earlier in the development cycle; and it allows functionalities to be added to a product that customers are willing to pay a premium for, while eliminating features they do not want. Furthermore, it identifies and then prioritises customer needs for specific markets.
However, this study has its limitations. First, since using big data to support accelerated innovation is new, there is little literature to build on. We have had to rely on investigation of some trends of increasing importance in product innovation evolution to identify the key success factors, and use semi-structured interviews with leading industry experts and academic researchers to verify them. Therefore, the issue of whether more industry experts and academic researchers would generate similar results, internationally, needs to be investigated. In addition, in order to better apply the ACE framework to business practice, further research could address how to integrate the framework into real product innovation processes. This implies defining an appropriate search focus, for which the ACE framework might serve as an ideal basis.

Second, developing a high-level framework for such a complicated phenomenon as accelerated innovation may highlight some obvious connections while failing to capture others. Future empirical studies might test the ACE framework across different industries. Also, this research could take a longitudinal approach to analysing ACE implementation in a firm over time. Learning and innovation processes can be properly studied only over a period of time. In particular, it is significant to address the problem of how to apply big data in each cornerstone of the ACE framework in detail. What hardware facilities are required to be constructed and what innovation capabilities are needed to be improved are both worth studying in future research. The findings could have useful implications for future research and policy design and implementation.

Lastly, this qualitative study interviewed industry experts from leading organisations as these organisations are highly engaged in big data harvesting and product innovation management. However, the ACE framework may also add value to SMEs by supporting their efforts in harvesting big data to support their new product development. The implementation of the ACE framework in SMEs may thus be an interesting field for further research.

References


Further reading


Corresponding author

Yuanzhu Zhan can be contacted at: svenchim@gmail.com

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com
A theoretical model of jump diffusion-mean reversion

Constant proportion portfolio insurance strategy under the presence of transaction cost and stochastic floor

Anindya Chakrabarty
Department of Quantitative Research, CRISIL Global Research & Analytics, Pune, India

Zongwei Luo
Department of Computer Science, Southern University of Science and Technology, Shenzhen, China

Rameshwar Dubey
Institute of Operations Management, Symbiosis International University, Nashik, India, and

Shan Jiang
Department of Computer Science, Southern University of Science and Technology, Shenzhen, China

Abstract

Purpose – The purpose of this paper is to develop a theoretical model of a jump diffusion-mean reversion constant proportion portfolio insurance strategy under the presence of transaction cost and stochastic floor as opposed to the deterministic floor used in the previous literatures.

Design/methodology/approach – The paper adopts Merton’s jump diffusion (JD) model to simulate the price path followed by risky assets and the CIR mean reversion model to simulate the path followed by the short-term interest rate. The floor of the CPPI strategy is linked to the stochastic process driving the value of a fixed income instrument whose yield follows the CIR mean reversion model. The developed model is benchmarked against CNX-NIFTY 50 and is back tested during the extreme regimes in the Indian market using the scenario-based Monte Carlo simulation technique.

Findings – Back testing the algorithm using Monte Carlo simulation across the crisis and recovery phases of the 2008 recession regime revealed that the portfolio performs better than the risky markets during the crisis by hedging the downside risk effectively and performs better than the fixed income instruments during the growth phase by leveraging on the upside potential. This makes it a value-enhancing proposition for the risk-averse investors.

Originality/value – The study modifies the CPPI algorithm by re-defining the floor of the algorithm to be a stochastic mean reverting process which is guided by the movement of the short-term interest rate in the economy. This development is more relevant for two reasons: first, the short-term interest rate changes with time, and hence the constant yield during each rebalancing steps is not practically feasible; second, the historical literatures have revealed that the short-term interest rate tends to move opposite to that of the equity market. Thereby, during the bear run the floor will increase at a higher rate, whereas the growth of the floor will stagnate during the bull phase which aids the model to capitalize on the upward potential during the growth phase and to cut down on the exposure during the crisis phase.

Keywords Stochastic process, CIR, CPPI, Jump diffusion model, Monte Carlo simulation, Portfolio insurance

Paper type Research paper

The authors are highly grateful to the Editor and two reviewers for their constructive inputs which have helped significantly to improve the authors’ revised paper. This work is partially supported by the GDNSF fund (2015A030313782) and SUSTech Startup fund (Y01236215).
1. Introduction
Higher integration of financial markets, surge of foreign institutional investments, and real-time information streaming, on the one hand, have facilitated the discovery of fair market price and the enhancement of market efficiencies, on the other hand, have resulted in higher market volatility. Contagion or the spill-over effect is more prominent in an integrated market than in a segmented one. These have led to an increasing concern both among the potential investors as well as the financial institutions over the assessment and management of risks. While the two-fund separation theorem asserts that a rational investor always allocates his/her endowments between risky asset and risk free asset in accordance to his/her degree of risk averseness, the same, however, changes with the change in market scenario. Empirically, it has been seen that investors in general tends to become more risk averse during the regimes of higher volatility and more risk loving during the counter regimes. Thus, the prime objective of utility maximization can be visualized as a combination of two sub-objectives: first being the protection of the investors’ wealth during the bad times and second, being the maximization of returns during the good times. Portfolio insurance strategies address both these needs. Traditionally, there are two categories of portfolio insurance strategies: static and dynamic. The former strategy chooses stock index options or futures to hedge the downside risk of the portfolio, while the latter one relies on continuous rebalancing of the portfolio between the risky and risk-free asset with an objective of insuring the investments from all possible erosions. While the option-based portfolio insurance (OBPI) strategy is the popular example of static portfolio insurance, constant-mix strategy, constant proportion portfolio insurance (CPPI), dynamic proportion portfolio insurance, time invariant portfolio protection, etc., are the popular examples of the dynamic portfolio insurance. Among them, however, the CPPI strategy is still the most popular and widely practiced (Pain and Rand, 2008). Investment schemes developed utilizing these strategies are generally coined as capital protection funds or capital guaranteed funds.

Among the existing literature on CPPI strategy a common assumption has been that the investment in the risk free asset grows at a constant rate in spite of frequent trading. Empirical evidences buttress the fact that interest rate follows a stochastic mean reverting behavior, and thus frequent reshuffling of portfolio between risky and risk-free asset makes it impractical to assume that the investment in the money market account will grow at a constant rate along the entire investment horizon. Considering this gap in the existing literatures, the paper proposes to construct a model of JD-MR-CPPI strategy under the presence of transaction cost and stochastic floor as opposed to the deterministic floor used in the previous literature and evaluates the effectiveness of the algorithm during the extreme regimes in the Indian market using the scenario-based Monte Carlo simulation technique.

The rest of the paper is organized as follows: Section 2 briefly probes into the existing literature on the CPPI strategy. Section 3 presents the theoretical framework of the JD-MR CPPI strategy, Section 4 empirically validates the model using the scenario-based simulation technique and Section 5 concludes the paper.

2. Literature review
The CPPI is introduced by Perold (1986) on fixed income assets. Black and Jones (1987) extended this method on equity-based underlying assets. Black and Perold (1992) further developed the algorithm by probing into how transaction costs and borrowing constraints impacts the insurance strategy. Their research revealed that in the absence of any transaction cost the CPPI is equivalent to investing in perpetual American call option and that the strategy is best for the HARA utility function under minimum consumption constraints. Their study further revealed that with the increase in the “multiple” value, the payoff under the CPPI approaches to that of the stop-loss strategy. CPPI strategies in
presence of jumps in stock prices were first considered by Prigent and Tahar (2005) in a diffusion model with finite intensity jumps. Following their work, Cont and Tankov (2009) quantified the gap risk of classical CPPI strategies with the assumption that the risky asset follows Merton’s jump diffusion (JD) model. Furthermore, their study probed into the analytical expression for the expected losses and the distribution of the losses given the gap event has occurred. Unlike the former research they incorporated infinite activity jumps and stochastic volatility in their algorithm. Estep and Kritzman (1988) came up with the time invariant portfolio insurance (TIPI) strategy where the floor is proposed to be dynamic proportional to the current wealth. Thereby, this strategy was tuned toward protection of the current wealth and not the floor value. Compared to the traditional CPPI, the TIPI strategy is more conservative in terms of restricted exposure during growth phases. In line with their work Chen and Liao (2006), proposes a goal-directed CPPI strategy to combine an investor’s goal-directed trading behavior with the traditional CPPI strategy. The objective is to maintain conservative exposure in the risky asset when the portfolio value approaches the pre-set goal and to take an aggressive exposure when the deviation from the goal is large. However, the approach suffers from one major drawback that it fails to utilize the upside potential to the fullest extent. Hainaut (2010) analyzes the influence of switches of assets regimes on the CPPI performance and risk exposure under the additional assumption that the dynamics of the risky asset is driven by a hidden Markov process. The paper shows how the value at risk and the tail VaR can be retrieved by inversion of the Fourier transform of the characteristic function of the return density. Another important line of research on the CPPI strategy concentrated on the determination of the “multiple” that guides the exposure to the risky asset, and hence the overall risk exposure of the portfolio. Authors like Bertrand and Prigent (2002) and Prigent and Tahar (2005) probed into the development of unconditional “multiple” estimates using the extreme value approach, while Hamidi et al. (2009) concentrated on conditional “multiple” determination where the multiplier is defined as a function of an extended expected value-at-risk with an objective to keep a constant exposure to risk. It is widely assumed in most of the literature on CPPI that the floor grows at a constant risk free rate. An alternative to this notion was introduced by Boulier and Kanniganti (1995) and later extended by Mkaouar and Prigent (2007), where they assumed that floor value at any given time is partially dependent on the portfolio value. The partial dependence can be explained by the fact that the floor value increases when the risky asset in the portfolio performs strongly but the same does not decreases during poor performance.

In contrast to the previous work, the current paper assumes the floor of the model to be a stochastic mean reverting process which is guided by the movement of the short-term interest rate in the economy. This development is more relevant for two reasons: first, the short-term interest rate changes with time, and hence the constant yield during each rebalancing steps is not practically feasible; second, the historical literature have revealed that the short-term interest rate tends to move opposite to that of the equity market. Thereby, during the bear run the floor will increase at a higher rate, whereas the growth of the floor will stagnate during the bull phase which helps the algorithm to capitalize on the upward potential during the growth phase and to cut down on the exposure during the crisis phase.

3. Theoretical framing
The JD-MR-CPI model: the CPPI strategy dynamically reallocates fund between a risky asset and a risk free money market account with an objective of protecting the investors’ initial capital along the investment horizon. The algorithm starts by setting a floor, which is normally kept equal to the present value of the initial investment, discounted at the risk free rate over the investment horizon. Capital allocated as floor today will grow at the risk free rate to the initial investment at maturity. The idea is that if through dynamic rebalancing, the fund manager can ensure the portfolio value to never fall below the floor
then irrespective of the price movement of the risky asset, the portfolio value will always
remain above or equal to the initial investment at maturity.

Suppose "\(v_0\)" is the initial investment, "\(r\)" is the risk free rate and "\(T\)" is the investment
horizon (in years). The initial floor (\(f_0\)) is set equal to an amount which when invested in the
risk free rate will grow to \(v_0\) at time \(T\).

The expression of the exposure is given by the following equation. The multiple (\(m\)) is an
important parameter in the algorithm, because it controls the exposure of the fund to the
risky asset. Higher is the multiple, higher is the exposure and higher will be the expected
return of the portfolio. But a high multiple also increases the probability of gap risk:

\[
E_0 = \min(v_0, m \times (v_0-f_0))
\]  

(2)

Once the exposure has been determined, the same amount is invested in the risky asset and
the remaining fund is parked into the money market account. Thus, the amount (\(B_0\))
invested in the money market account is, thus, given by:

\[
B_0 = v_0 - \min(v_0, m \times (v_0-f_0))
\]

(3)

Once the initial allocation has been done the next task is to decide upon the
rebalancing approach. There are two commonly used rebalancing approaches: the time-
based rebalancing, where rebalancing is done at a fixed time interval over the investment
horizon and the move-based rebalancing, where rebalancing is done once the percentage
change in exposure to the risky asset crosses a predetermined threshold value. Sometimes, a
combination of both the approaches is used. Move-based rebalancing is suitable in the world
with high transaction costs, as this method prevents unnecessary rebalancing during minor
fluctuations and thereby minimizes the transaction cost. However, in this approach the
threshold value has to be chosen carefully. A higher threshold value reduces the number of
rebalancing, and hence the total transaction cost, but at the same time increases the
probability of the portfolio value crashing down the floor. The optimal threshold limit
should be the one that minimizes both the transaction cost and the cost of gap risk. In case
of time-based rebalancing, the decision parameter is the rebalancing interval. Higher
rebalancing interval reduces the total transaction cost, but increases the cost of gap risk.
Hence, the optimal rebalancing interval is the one that minimizes both the total transaction
cost and the cost of gap risk.

During each rebalancing period (\(t\)) along the investment horizon the cushion is
recalculated by considering the difference between the portfolio value and the floor \(f_t\):

\[
c_t = \max(0, (v_t-f_t)) \quad \text{where} \quad 0 \leq t \leq T
\]

(4)

In the event, if the portfolio value goes below the floor, the cushion is set to 0 (Equation (4)).

The exposure is then calculated as:

\[
E_t = \min(v_t, m \times c_t)
\]  

(5a)

Replacing the value of \(c_t\) from Equation (4), we get:

\[
E_t = \min(v_t, m \times \max(0, (v_t-f_t)))
\]  

(5b)
The exposure can never be more than the portfolio value at any point of time. If the exposure at any rebalancing period exceeds the portfolio value, then the exposure is reset to the current portfolio value and the entire fund is invested in the risky asset. This explains the “min” function in Equations (5a) and (5b). After investing the exposure amount in the risky asset the difference ($V_t - E_t$) is invested in the money market account. The procedure is repeated at each of the rebalancing terminal till the maturity of the portfolio.

Geometric Brownian motion had been widely used for depicting the diffusion process of risky asset. But the empirical evidence of leptokurtic distribution and the presence of flat tail of the financial asset return distribution necessitated the search for alternatives. Merton (1976), for the first time, introduces the JD model where the diffusion process is assumed to be composed of two parts: a geometric Brownian motion with a constant drift and volatility and a compound Poisson process guiding the arrival of jumps. Merton further assumed that the jump size is log-normally distributed with a constant mean and variance. The jumps signify the arrival of news (both good and bad) that results in the sharp movement of the asset price within a short-time interval. Following the seminal work of Merton (1976), Kou (2002) delivered the double exponential jump diffusion model (DEJD), where the arrival of news is still guided by the Poisson distribution but the jump magnitude is depicted by the double exponential distribution. Ramezani and Zeng (1998) arrived at the Pareto-Beta jump diffusion (PBJD) model, where the jumps, caused by good news, are assumed to follow Pareto distribution and those caused by bad news are assumed to follow beta distribution. Though from conceptual point of view both the DEJD and PBJD models are alike, they, however, differ structurally. While Kou (2002) suggested using two exponential distributions with dissimilar parameters to define the jumps, Ramezani and Zeng (1998) assume that both good and bad news are produced by two autonomous Poisson processes with different intensities and that the corresponding jump magnitudes are drawn from Pareto and beta distributions, respectively. However, simplistic assumptions and ease of use makes the Merton’s JD model a popular modeling tool among the practitioners in comparison to the other complicated models. This model is used in the current study. Under the JD model, the incremental change in the price of the risky asset is given by:

$$dS_t = \left(\mu - \lambda K - \frac{\sigma^2}{2}\right)S_t dt + \sigma S_t dW_t + S_t (Y_t - 1) dN_t$$

(6)

where $W_t$ is a standard Weiner process with zero drift and variance equal to $dt$. The increments $dW_t$ are independent of one another in the interval $[0, T]$. $T$ is the maturity period of the portfolio, $\mu$ is the constant drift and $\sigma^2$ is the constant variance of the risky asset. The term $\sigma^2/2$ is used for convexity correction. $N_t$ is the compound Poisson process with intensity $\lambda$ signifying the number of jumps within the time interval $[0, t]$. $Y_t$ is log-normally distributed random process signifying the jump magnitude. For a small time interval $dt$, the asset price jumps from $S_t$ to $S_t Y_t$. Thus, the percentage change in the asset price caused by the jump is given by:

$$\frac{dS_t}{S_t} = \frac{Y_t S_t - S_t}{S_t} = Y_t - 1$$

(7)

The incremental change $dN_t$ gives jumps occurring within the incremental time $dt$ such that:

$$\mathbb{P}(dN_t = 1) = \lambda dt \quad \text{and} \quad \mathbb{P}(dN_t = 0) = 1 - \lambda dt$$

(8)

The log of the jump magnitude is $\sim$ i.i.d normal ($\mu_J$, $\sigma_J$). The relative jump magnitude $(Y_t - 1)$ is also log-normally distributed with the mean and variance given by:

$$\mathbb{E}(Y_t - 1) = e^{(\mu + \frac{1}{2}\sigma^2)} - 1 = K$$

(9)
Under the CPPI strategy, rebalancing is done frequently to avoid the loss because of gap risk and thereby the paper assumes that the amount invested in the money market account grows at the short-term interest rate. It is widely documented that short-term interest rates for any economy are neither constant nor do they follow random walk, but display a well-known phenomenon of mean reversion. It refers to the tendency that the interest rate drifts at a certain rate toward the long-term average. Empirically, this means that the change in interest rate should be significantly positively correlated with the deviation from the long-term mean. Several researchers like Vasicek (1977), Dothan (1978), Brennan and Schwartz (1979), Cox et al. (1985), Heath et al. (1992), and others contributed significantly toward the interest rate modeling. While Vasicek (1977) presented one of the earliest stochastic mean reverting model for interest rate where the author assumed that in their one-factor model the interest rate follows the Ornstein Uhlenbeck process, Cox et al. (1985) in their general equilibrium model (CIR model) improved upon the same to ensure the interest rate not to go below 0. Furthermore, unlike the Vasicek model, in the CIR model the short-term interest rate does not display a normal or lognormal distribution, but instead exhibit a non-central \( \chi^2 \) distribution. The paper adopts the CIR model to govern the interest rate process of the money market account. Under the CIR model the short-term interest rate diffusion process is given by:

\[
d r_t = \kappa(\theta - r_t)dt + \sigma \sqrt{r_t}d\omega_t
\]  

(11)

where \( \kappa \) is the speed of adjustment of the instantaneous interest rate toward the target \( \theta \). \( \sigma \) is the standard deviation of the interest rate and \( d\omega_t \) is standard Weiner process with zero drift and variance equal to \( dt \). The model also inculcates two set of restrictions, namely, \( \kappa, \theta, \sigma > 0 \) and \( 2\kappa \theta > \sigma^2 \), where the second restriction prevents the interest rate from going negative. Now, under the stochastic interest rate environment the initial floor of the CPPI strategy should be set equal to the value of a zero coupon bond that grows to the initial investment \( I \) at the stochastic interest rate \( r_t \) at maturity \( T \). According to the CIR model the price of a zero coupon bond at time \( t \) \((t \in \{0, T\})\) with a maturity value of \( I \) and maturity period of \( T \) is given by:

\[
B(t, T) = IA(t, T)e^{-r_tR(t, T)}
\]  

(12)

where \( I \) is the maturity value; \( r_t \) is the interest rate on the valuation date; \( T \) is the maturity period and the rest of the parameters are depicted below:

\[
A(t, T) = \left[ \frac{2h\theta^{h+\kappa}(T-t)/2}{2h + (h + \kappa)(\theta^{(T-t)}-1)} \right]^{2\theta/\sigma^2}
\]  

(13)

\[
R(t, T) = \left[ \frac{2(\theta^{h(T-t)}-1)}{2h + (h + \kappa)(\theta^{(T-t)}-1)} \right]
\]  

(14)

\[
h = \sqrt{\kappa^2 + 2\sigma^2}
\]  

(15)

Thus, the initial floor \( f_0 \) for the capital protection fund under stochastic interest rate is given by:

\[
f_0 = I\left[ \frac{2h\theta^{h+\kappa}(T)/2}{2h + (h + \kappa)(\theta^{(T)}-1)} \right]^{2\theta/\sigma^2} \times e^{-r_0} \left[ \frac{2(\theta^{h(T)}-1)}{2h + (h + \kappa)(\theta^{(T)}-1)} \right]
\]  

(16)

where \( r_0 \) is the spot interest rate at time \( t = 0 \) when the floor valuation is done.
The diffusion process of the zero coupon bond is given by:

$$dB_t = r_t B_t dt - \sigma \sqrt{r_t B_t} \frac{2(e^{hT-t} - 1)}{2h + (h + \kappa)(e^{hT-t} - 1)} d\omega_t$$

(17)

Coming back to the CPPI strategy at any rebalancing date \(t\) the exposure in the risky asset is computed as:

$$e_t = \min \left( v_t, m |v_t - f_t|^+ \right)$$

where \( |v_t - f_t|^+ = \max \{ 0, (v_t - f_t) \} \)

(18)

Thus, the amount invested in the risk free money market account is given by:

$$B_t = \left( v_t - \min \left( v_t, m |v_t - f_t|^+ \right) \right)$$

(19)

where \(B_t\) grows at a stochastic risk free rate \(r_t\) through time and its movement is depicted in Equation (17).

A vital assumption of the CPPI portfolio is self-financing. It means that for every small time increments, the incremental change of the risky asset holding and the incremental change of the risk free asset holding will only contribute to the incremental change of the portfolio value and no infusion of extra fund at any stage is made. The mathematical representation of the self-financing strategy is given by the following equation:

$$d v_t = \min \left( v_t, m |v_t - f_t|^+ \right) \left[ \frac{d S_t}{S_t} + \left( v_t - \min \left( v_t, m |v_t - f_t|^+ \right) \right) \frac{d B_t}{B_t} \right]$$

(20)

The terms \(d S_t/S_t\) and \(d B_t/B_t\) in Equation (20) can be replaced by the corresponding terms from Equations (6) and (17), respectively to obtain:

$$d v_t = \min \left( v_t, m |v_t - f_t|^+ \right) \left[ \left( \mu - \lambda K - \frac{\sigma^2}{2} \right) dt + \sigma d W_t + (Y_t - 1) d N_t \right]

+ \left[ v_t - \min \left( v_t, m |v_t - f_t|^+ \right) \right] \left[ r_t dt - \sigma \sqrt{r_t} \frac{2(e^{hT-t} - 1)}{2h + (h + \kappa)(e^{hT-t} - 1)} d \omega_t \right]$$

(21)

A slight rearrangement of the terms in Equation (21) results in the following equation:

$$d v_t = \left( \left( \mu - \lambda K - \frac{\sigma^2}{2} - r_t \right) \min \left( v_t, m |v_t - f_t|^+ \right) + r_t v_t \right) dt + \min \left( v_t, m |v_t - f_t|^+ \right) \sigma d W_t

- \left[ v_t - \min \left( v_t, m |v_t - f_t|^+ \right) \right] \sigma \sqrt{r_t} \frac{2(e^{hT-t} - 1)}{2h + (h + \kappa)(e^{hT-t} - 1)} d \omega_t

+ \min \left( v_t, m |v_t - f_t|^+ \right) (Y_t - 1) d N_t$$

(22)

Equation (22) represents the diffusion process of the JD-MR CPPI portfolio value. It consists of four components: a deterministic drift, a stochastic term representing the unpredictability of the risky asset investment, a stochastic term representing the Poisson distributed jump process, and finally, a stochastic term representing the randomness of the money market investment. The objective of rebalancing is to keep the portfolio value \(v_t\) above or equal to the floor \(f_t\). Once \(v_t\) touches the floor \((v_t = f_t)\) then \(v_t - f_t^+\) is set to 0 and so is the exposure \(\min (v_t, m |v_t - f_t|^+)\). As a result the stochastic component because of the jump and the
risky asset diffusion vanishes. The entire fund is now allocated to the risk free asset. The differential Equation (22) is now reduced to the diffusion process followed by the money market account:

\[
dv_t = r_t v_t dt - v_t \sigma \sqrt{h} \left( \frac{2(e^{h(T-t)}-1)}{2h + (h+\kappa)(e^{h(T-t)}-1)} \right) d\omega_t
\]  

(23)

However, if the rebalancing cannot be achieved when \( v_t \) touches \( f_t \) and \( v_t \) goes below \( f_t \) then the whole idea of protection will be compromised. Such a situation is more probable at times of steep fall in the underlying risky asset price and it gives rise to the so called “gap-risk.” Now, during the bull phase when the risky assets price surges and the exposure crosses the portfolio value, the entire fund is allocated to the risky asset. The boundary condition for its occurrence is:

\[
v_t = m(v_t - f_t)
\]

and which implies that:

\[
v_t = \frac{f_t}{1 - 1/m}
\]  

(24)

Under this condition, the exposure \( \min (v_t, m |v_t - f_t|) \) is set to \( v_t \) and the stochastic differential equation (Equation (13)) reduces to:

\[
dv_t = v_t \left( \mu - \lambda K - \frac{\sigma^2}{2} \right) dt + v_t \sigma dW_t + v_t (Y_t - 1) dN_t
\]  

(25)

Equation (25) reveals that the portfolio value follows the geometric Brownian motion with jump having the same drift and variance as that of the underlying risky asset. Higher is the expected return of the underlying risky asset higher will be the expected return of the portfolio during a bull run. Thus the JD-MR-CPPI strategy address the upside potential effectively and at the same time takes care of the downside risk by eliminating the stochastic component once \( v_t = f_t \) (see Equation (16)).

When the portfolio value lies within the interval given by the inequality (Equation (27)) the allocation is done both in risky and risk free asset. In that case the stochastic differential equation governing the portfolio value process is shown in Equation (27):

\[
\frac{f_t}{1 - 1/m} \leq v_t \leq \frac{f_t}{1 - 1/m}
\]

(26)

\[
dv_t = \left( \mu - \lambda K - \frac{\sigma^2}{2} - r_t \right) m (v_t - f_t) + r_t v_t dt + m(v_t - f_t) \sigma dW_t
\]

\[- [v_t - m(v_t - f_t)] \sigma \sqrt{h} \left( \frac{2(e^{h(T-t)}-1)}{2h + (h+\kappa)(e^{h(T-t)}-1)} \right) d\omega_t + m(v_t - f_t)(Y_t - 1) dN_t
\]  

(27)

Finally, the observations can be summarized as displayed in the Table I.

4. Empirical validation of the JD-MR-CPPI algorithm

4.1 Data analysis

For empirical analysis of the developed model, suitable proxies of the risky asset and the risk free assets are prerequisite. The National Stock Exchange of India, though relatively
new compared to the Bombay Stock Exchange, has witnessed a considerable growth during the last ten years both in terms of volume traded and number of companies listed. This qualifies CNX-NIFTY 50 to be a suitable proxy for the risky asset. Indian financial market witnessed a considerable swing during the financial crisis of 2008. CNX-NIFTY 50 plunged down from a record high of 6,288 on January 8, 2008 to 2,524 on October 27, 2008 (source: Yahoo Finance). Post-recession recovery of the Indian market was also quick compared to the developed economies. CNX-NIFTY 50 touched 6,312.45 again on November 5, 2010, which was the new high after such crisis. The paper proposes to capture these two phases of the market in order to compare the performance of the JD-MR-CPPI algorithm against the CNX-NIFTY 50 index (taken as a benchmark as well as the underlying asset) in both these phases. For this purpose the period from January 8, 2008 to October 27, 2008 is coined as the downswing phase and the period from October 27, 2008 to November 5, 2010 as the recovery phase.

We have stressed tested our model in these two historical extreme periods to check the boundary of our model. Further, using Monte Carlo simulation on these extreme regions we have derived a series of hypothetical stressed scenarios. We have subsequently stressed tested our model on all these extreme scenarios. Stress testing of any model on the historical and hypothetical stressed scenario is an efficient way of testing the model performance and robustness. This methodology has gained considerable importance after the 2008 subprime crisis and the regulators are increasingly considering stress-testing as a viable means of model validation and model risk management (see Dodd-Frank Act and Comprehensive Capital Adequacy and Review Guidelines of Federal Reserve System for details, Ref.: www.federalreserve.gov/bankinforeg/ccar.htm).

Daily price data of the CNX-NIFTY 50 is collected across both the phases. On the other hand, the low level of development of the Indian debt market coupled with illiquid instrument and lack of reliable data hinders the selection of a suitable proxy for the money market accounts. As a proxy for the short-term interest rate of the money market account, the call money rate is selected because of two reasons: first, the reliable quotations are available on a daily basis and second, previous experience with the Indian market have revealed that the interbank call rates influence significantly the interest rate of the economy. Daily call money rates across both the phase are collected from IFMR Data Centre (Source: www.ifmr.ac.in/). The summary statistics of the collected data are displayed in Tables II and III, respectively.

4.2 JD-MR CPPI model calibration

The diffusion models guiding the risky asset price and the money market account are calibrated against the CNX-NIFTY and the call money rate data during both the phase using the maximum likelihood estimation technique. For a given set of data and the assumed

<table>
<thead>
<tr>
<th>Range of the portfolio value</th>
<th>Allocation</th>
<th>Differential equation governing the portfolio value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_t = f_t$</td>
<td>Entirely in money market account</td>
<td>$dv_t = r v_t dt - v_t \sqrt{r_t^2 + \sigma^2} dW_t$</td>
</tr>
<tr>
<td>$v_t = \frac{f_t}{1 + \alpha}$</td>
<td>Entirely in risky asset</td>
<td>$dv_t = \left(\mu - \beta - \frac{\sigma^2}{2}\right)dt + v_t \sigma dW_t + \gamma(Y_t - 1)dN_t$</td>
</tr>
<tr>
<td>$f_t \leq v_t \leq \frac{f_t}{1 + \alpha}$</td>
<td>The exposure in risky asset and the balance in the money market account</td>
<td>$dv_t = \left(\mu - \beta - \frac{\sigma^2}{2}\right)dt + v_t \sigma dW_t + \gamma(Y_t - 1)dN_t$</td>
</tr>
</tbody>
</table>

Source: Authors' analysis
underlying model the MLE technique returns the optimal set of the model parameters that maximizes the probability or likelihood that the model output and the observed data will match. In MATLAB the same is achieved by maximizing the log likelihood function of the process against the set of parameters using the “fminsearch” non-linear optimization routine present in the optimization toolbox. The calibrated parameters for the risky asset and the money market account are displayed in Tables IV and V, respectively.

### Table II.
Summary statistics of the daily returns of CNX-NIFTY 50

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Recession phase</th>
<th>Recovery phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>−0.00465661</td>
<td>0.00191757</td>
</tr>
<tr>
<td>Median</td>
<td>−0.00308592</td>
<td>0.00173951</td>
</tr>
<tr>
<td>Minimum</td>
<td>−0.130142</td>
<td>−0.0793851</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.0672052</td>
<td>0.163343</td>
</tr>
<tr>
<td>SD</td>
<td>0.0272906</td>
<td>0.0195593</td>
</tr>
<tr>
<td>CV</td>
<td>5.86062</td>
<td>10.2000</td>
</tr>
<tr>
<td>Skewness</td>
<td>−0.460445</td>
<td>1.05220</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.24326</td>
<td>10.7251</td>
</tr>
</tbody>
</table>

**Source:** Authors’ analysis

### Table III.
Summary statistics of the annual returns of call money rates

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Recession phase</th>
<th>Recovery phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>7.8279</td>
<td>3.8533</td>
</tr>
<tr>
<td>Median</td>
<td>7.6700</td>
<td>3.2500</td>
</tr>
<tr>
<td>Minimum</td>
<td>4.1100</td>
<td>1.6800</td>
</tr>
<tr>
<td>Maximum</td>
<td>18.530</td>
<td>17.890</td>
</tr>
<tr>
<td>SD</td>
<td>1.9682</td>
<td>1.5860</td>
</tr>
<tr>
<td>CV</td>
<td>0.25143</td>
<td>0.41159</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.6620</td>
<td>4.4768</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>5.3297</td>
<td>31.110</td>
</tr>
</tbody>
</table>

**Source:** Authors’ analysis

### Table IV.
Parameters of the Merton’s jump diffusion model calibrated for the CNX-NIFTY 50 returns using maximum likelihood estimation

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Calibrated value during the recession phase</th>
<th>Calibrated value during the recovery phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>−0.9127</td>
<td>0.9166</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.3523</td>
<td>0.3954</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1.9466</td>
<td>1</td>
</tr>
<tr>
<td>$\mu_J$</td>
<td>−0.1047</td>
<td>0.1618</td>
</tr>
<tr>
<td>$\sigma_J$</td>
<td>1.1077e−006</td>
<td>4.6146e−007</td>
</tr>
</tbody>
</table>

**Source:** Authors’ analysis

### Table V.
Parameters of the CIR model calibrated for the call money rates using maximum likelihood estimation

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Calibrated value during the recession phase</th>
<th>Calibrated value during the recovery phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$</td>
<td>0.2886</td>
<td>0.2393</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.078782</td>
<td>0.037788</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.04859</td>
<td>0.03756</td>
</tr>
</tbody>
</table>

**Source:** Authors’ analysis
4.3 JD-MR CPPI model simulations

Once the model has been calibrated to the historical data the same is used to simulate 100,000 trajectories of the possible paths of the portfolio during each of the downswing and recovery phase. The expected return and risks are then calculated by taking the corresponding means across all the simulated paths. The initial investment in the portfolio is taken at Rs1,000 with a maturity period of one year. The initial floor is set to the price of the zero coupon bond that provides a maturity value of Rs1,000 after one year, with the interest rate following the CIR mean reversion process (given by Equation (11)). The multiple is set to 3 for the present study. Transaction cost is taken as 0.01 percent of the total volume transaction. It is also assumed that the same transaction cost prevails for buying and selling of the risky and risk-free asset in the financial market. Rebalancing frequency is kept at 200 times a year at equal interval. The initial NIFTY value is normalized to the initial investment for comparison purpose. The results are displayed in Table VI.

4.4 JD-MR CPPI model performance analysis

Table VI indicates that during the downswing phase when the aggregate market return was −75.11 percent, the JD-MR CPPI portfolio manages to maintain an average return of 1.2 percent. The 99 percent VaR of the portfolio is also significantly less than that of the benchmarked market index. During the downswing phase it can be deduced that for an initial investment of Rs1,000 the loss will not exceed Rs10.5251 for 99 percent of the cases if the investment is made in the JD-MR-CPPI portfolio; whereas, for an equal investment in the market during the downswing phase, the corresponding loss value increases to a whopping Rs907.6475. During the recovery phase the portfolio generates an aggregate return of 85.45 percent against a market return of 212.41 percent, but manages to control the VaR at Rs35.0508 as opposed to that of the market (Rs90.5709). Thus, the JD-MR-CPPI portfolio performs better that the risky market during the downswing and performs better than the fixed income market during the growth phase, which makes it a value enhancing proposition for the risk adverse investors. Figure 1 provides the path followed by the portfolio, the market and the floor for a particular simulation, the corresponding allocation in risky assets and the final histogram of the terminal value of the portfolio. The histograms for both the phases are right skewed indicating the hedging effectiveness under the extreme market environments. Figure 2 provides the terminal value of the portfolio (in green) and the terminal value of the floor (in red) for all the 100,000 simulations. The lower cut-off of

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Downswing phase</th>
<th>Recovery phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected annual returns over 100,000 simulations (%)</td>
<td>JD-MR CPPI portfolio 1.20</td>
<td>Nifty 50 85.45</td>
</tr>
<tr>
<td>Expected annual SD over 100,000 simulations (%)</td>
<td>6.4619</td>
<td>120.6026 288.8027</td>
</tr>
<tr>
<td>99% VAR</td>
<td>10.5251</td>
<td>907.6475 99.0508</td>
</tr>
<tr>
<td>99% VAR of the money market account 10.2710</td>
<td></td>
<td>99% VAR of the money account 10.2776</td>
</tr>
<tr>
<td>Mean terminal value over 100,000 simulations</td>
<td>JD-MR CPPI portfolio 1.012</td>
<td>Floor 1.8545</td>
</tr>
<tr>
<td>Mean net transaction costs</td>
<td>16.1162</td>
<td>119.9253 1,001.4</td>
</tr>
<tr>
<td>Source: Authors’ analysis</td>
<td></td>
<td>Table VI</td>
</tr>
</tbody>
</table>
the terminal value of the portfolio indicates the capital protective feature of the algorithm, while the volatility of the terminal value of the floor is expected because of its stochastic nature as defined in the algorithm. Figure 3 displays the cumulative average transaction cost curve across 100,000 simulations. The curve is concave and the value stabilizes near Rs16 during the market downswing phase. This is primarily because of low transactions and stable investment in the debt segment during the crisis period. During the recovery phase the curve displays convex characteristics primarily because of the heavy transactions and the average costs shoots up at increasing rate as the market rises.

5. Conclusion and relevance of the algorithm in the Indian market
The paper develops a theoretical model of a JD-MR-CPPI strategy under the presence of transaction cost and stochastic floor as opposed to the deterministic floor used in the previous literature. The model is validated via back testing during the extreme regimes in the Indian market using the scenario-based Monte Carlo simulation technique. Besides providing capital protection, the strategy is found to hedges the downside risk effectively during bad times and is also found to leverages the upside potentials during the good times. Coming to the Indian

![Figure 1. Path followed by the portfolio for a particular simulation, the corresponding allocation in risky assets and the final histogram of the terminal value of the portfolio](image1)

Source: Authors’ analysis using MATLAB

![Figure 2. Terminal value of the portfolio and floor for 100,000 simulated paths](image2)

Source: Authors’ analysis using MATLAB
context, till the year 2006 the Security and Exchange Board of India (SEBI) voted against the
classification of any capital protected schemes in the Indian market, although the same were widely
flourishing in the foreign counterparts. However, in 2006, following several rounds of
discussions and constant persuasions from the Association of Mutual Fund of India, SEBI
allowed the entry of the capital protected schemes by amendment of the SEBI (Mutual Fund)
Regulations, 1996 vide a circular dated August 14, 2006. As per the regulations, capital
protection schemes floated by the AMC should be close ended and should be mandatorily rated
by a registered credit rating agency to ascertain the degree of certainty of achieving the
objective of the fund. The regulation also clearly indicated that the asset management
companies can market the scheme as “Capital protection oriented” fund and not “Capital
Guaranteed fund” (ref: SEBI Circular No. SEBI/IMD/CIR No. 9/74364/06). The difference is also
evidently identified in the next line in the circular – “the orientation toward capital protection
initiates from the portfolio structure and not from any bank guarantee, insurance, cover, etc.”
(ref: SEBI Circular No. SEBI/IMD/CIR No. 9/74364/06). Thus embedded options, which invokes
guarantee by virtue of its design, are excluded and thereby the OBPI strategy was not
encouraged in the Indian market. As per the SEBI guideline, capital protection-oriented funds
are sought to be structured by suitable combination of risky and risk free assets and by
dynamic rebalancing the same through time with an objective of protecting the investor’s
initial fund. Thereby, only the CPPI strategy fits perfectly within the scope provided by SEBI.
Given this backdrop, the developed JD-MR-CPPI model will be best suited for engineering of
the structured products in the Indian market.

References
method”, Finance, Vol. 23 No. 2, pp. 69-86.
strategies”, 5th AFIR International Colloquium, pp. 1093-1124.
Brennan, M.J. and Schwartz, E.S. (1979), “A continuous time approach to the pricing of bonds”,


Appendix. MATLAB code for the JD-MR-CPPI model

% This program simulates the CPPI strategy under the assumption that the % risky asset follows the Merton's Jump Diffusion Model and the risk free % money market account follows the CIR mean reversion diffusion process. % Transaction cost is taken to be Rs 1 for every Rs 100 transaction. Time % is represented per year basis.

Initial_Investment = I;
Time_Horizon = T; % in years

Transaction_cost = 0.01; % assume the same transaction cost for buying and selling of the risky and risk-free asset in the financial market

Mean = 0.9166; % yearly expected return on the underlying
Sigma = 0.3954; % yearly expected percentage volatility on the stock index
r = 0.05; % Initial spot risk-free (money market) interest rate
Meanj = 0.1618;
Sigmaj = 4.6146e-007;
\[ \Lambda = 1; \]
\[ \Theta = 0.037788; \]
\[ \kappa = 0.2393; \]
\[ \Sigma_r = 0.03756; \]
\[ \Gamma = \sqrt{(\kappa)^2 + 2 \Sigma_r^2}; \]
\[ t = 0; \]
\[ \text{NumSimul} = 100,000; \]
\[ \text{NumStep} = 200; \]
\[ \text{Time}_{-Step} = T/\text{NumStep}; \]
\[ \text{Temp1} = (\Gamma + \kappa) \text{exp}(\Gamma(T-t)) - 1 + 2 \Gamma; \]
\[ \text{Temp2} = 2 \Gamma \text{exp}((\Gamma + \kappa)(T-t)/2); \]
\[ \text{Temp3} = 2 \text{exp}(\Gamma(T-t)) - 1; \]
\[ \text{Computation of A and B} \]
\[ A = (\text{Temp2}/\text{Temp1})^{2 \kappa \Theta / \Sigma_r^2}; \]
\[ B = \text{Temp3}/\text{Temp1}; \]
\[ \text{Rates' simulation} \]
\[ r = r + \kappa (\Theta - r) \text{Time}_{-Step} + \Sigma_r \sqrt{r} \text{sqrt(Ti} \text{me}_{-Step}) \text{randn}(1); \]
\[ \text{Interest}_{-Rate} = r; \]
\[ \text{floor today (will evolve at the risk-free rate), e.g.: 950} \]
\[ \text{Floor} = \text{I} \text{exp}((R/100)T) \text{A} \text{exp}(r \text{B}); \]
\[ t = \text{Time}_{-Step}; \]
\[ \text{leverage on the cushion between your money and the floor, e.g. 3} \]
\[ \text{Multiple}_{-CPPI} = M; \]
\[ \text{initialize values} \]
\[ \text{Underlying}_{-Index} = \text{Initial}_{-Investment}; \]
\[ \text{value of the underlying at starting time, normalized to equal investment} \]
\[ \text{Portfolio}_{-Value} = \text{Initial}_{-Investment}; \]
\[ \text{Cushion} = \max(0, \text{Portfolio}_{-Value} - \text{Floor}); \]
\[ \text{Underlyings}_{-in}_{-Portfolio} = \min(\text{Portfolio}_{-Value}, \max(0, \text{Multiple}_{-CPPI} \text{Cushion})); \]
\[ \text{Cash}_{-in}_{-Portfolio} = \text{Portfolio}_{-Value} - \text{Underlyings}_{-in}_{-Portfolio}; \]
\[ \text{Total}_{-transaction}_{-cost} = (\abs{\text{Underlyings}_{-in}_{-Portfolio}} + \abs{\text{Cash}_{-in}_{-Portfolio}}) \times \text{Transaction}_{-cost}; \]
\[ \text{Portfolio}_{-Value} = \text{Initial}_{-Investment} - \text{Total}_{-transaction}_{-cost}; \]
\[ \text{Underlying}_{-in}_{-Portfolio}_{-Percent} = \text{Underlyings}_{-in}_{-Portfolio} / \text{Portfolio}_{-Value}; \]
\[ a = \text{Underlyings}_{-in}_{-Portfolio}; \]
\[ b = \text{Cash}_{-in}_{-Portfolio}; \]
\[ c = \text{Floor}; \]
\[ e = \text{Underlying}_{-in}_{-Portfolio}_{-Percent}; \]
\[ f = \text{Cushion}; \]
\[ g = \text{Total}_{-transaction}_{-cost}; \]
\[ \text{initialize parameters for the plot (no theory in this)} \]
\[ \text{Cushion} = \text{zeros} \text{NumSimul}, \text{NumStep} + 1); \]
\[ \text{Cushion}(:, 1) = f \times \text{ones} \text{NumSimul}, 1); \]
\[ \text{Floor} = \text{zeros} \text{NumSimul}, \text{NumStep} + 1); \]
\[ \text{Floor}(:, 1) = c \times \text{ones} \text{NumSimul}, 1); \]
\[ \text{dW} = \sqrt{\text{Time}_{-Step}} \text{randn} \text{NumSimul}, \text{NumStep}); \]
\[ \text{dw} = \sqrt{\text{Time}_{-Step}} \text{randn} \text{NumSimul}, \text{NumStep}); \]
\[ \text{Index} = \text{zeros} \text{NumSimul}, \text{NumStep} + 1); \]
\[ \text{Poisson} = \text{poissrnd} \text{Lambda}/\text{NumStep} \text{NumSimul}, \text{NumStep}); \]
\[ \text{Poisson}_{-Jumps} = \text{zeros} \text{NumSimul}, \text{NumStep} + 1); \]
\[ k = \exp(-\text{Mean}) - 1; \]
Portfolio_Value = zeros(NumSimul,NumStep+1);
Portfolio_Value(:,1) = Underlying_Index*ones(NumSimul,1);
Total_Cost = zeros(NumSimul,NumStep+1);
Total_transaction_cost = zeros(NumSimul,NumStep+1);
Total_transaction_cost(:,1) = g.*ones(NumSimul,1);
Total_Cost(:,1) = Total_transaction_cost(:,1);
Underlying_in_Portfolio_Percent = zeros(NumSimul,NumStep+1);
Underlying_in_Portfolio_Percent(:,1) = e.*ones(NumSimul,1);
Change_in_cash = zeros(NumSimul,NumStep+1);
Change_in_Underlying = zeros(NumSimul,NumStep+1);
Underlying_in_Portfolio = zeros(NumSimul,NumStep+1);
Index(:,1) = Underlying_Index*ones(NumSimul,1);
Underlying_in_Portfolio(:,1) = a*ones(NumSimul,1);
Cash_in_Portfolio(:,1) = b*ones(NumSimul,1);
Old_Cash_in_Portfolio = zeros(NumSimul,NumStep+1);
Old_Underlying_in_Portfolio = zeros(NumSimul,NumStep+1);

for n = 1:NumStep

Poisson_Jumps(:,n) = sum(randn(Poisson(NumSimul,n),1),1);
Index(:,n+1) = Index(:,n).*(1+(Mean-Lambda*k-0.5*Sigma^2)*Time_Step+
Sigma*dW(:,n)+(Meanj-0.5*Sigmaj^2).*Poisson(:,n)+
Sigmaj.*Poisson_Jumps(:,n));
Underlying_in_Portfolio(:,n+1) = Underlying_in_Portfolio(:,n).*(1+(Mean-Lambda*k-0.5*Sigma^2)*Time_Step+
Sigma*dW(:,n)+(Meanj-0.5*Sigmaj^2).*Poisson(:,n)+
Sigmaj.*Poisson_Jumps(:,n));

Temp1 = (Gamma+Kappa)*(exp(Gamma*(T-t))-1)
+ 2*Gamma/Temp1;
Temp2 = 2*Gamma*exp((Gamma+Kappa)*(T-t)/2);
Temp3 = 2*(exp(Gamma*(T-t))-1);
A = (Temp2/Temp1)^(2*Kappa/Theta/Sigma_r^2);
B = Temp3/Temp1;

Interest_Rate = [Interest_Rate r];
Multip = (1+r*Time_Step-Sigma_r*B*sqrt(Time_Step)*randn(1);
Floor(:,n+1) = Floor(:,n).*Multiple(:,n);
Cash_in_Portfolio(:,n+1) = Cash_in_Portfolio(:,n).*Multiple(:,n);
Portfolio_Value(:,n+1) = Underlying_in_Portfolio(:,n)+Cash_in_Portfolio(:,n+1);
if abs(Index(1,n+1)-Index(1,n))/Index(1,n) > 0.001% #ok < BDSCA >
Cushion(:,n+1) = max(zeros(NumSimul,1),Portfolio_Value(:,n+1)-Floor(:,n+1));
Old_Cash_in_Portfolio(:,n+1) = Cash_in_Portfolio(:,n+1);
Old_Underlying_in_Portfolio(:,n+1) = Underlying_in_Portfolio(:,n+1);

Underlying_in_Portfolio(:,n+1) = min(Portfolio_Value(:,n),max(0, Multiple_CPPI*Cushion(:,n+1)));
Change_in_Underlying(:,n+1) = Underlying_in_Portfolio(:,n+1)-Old_Underlying_in_Portfolio(:,n+1);
Cash_in_Portfolio(:,n+1) = Portfolio_Value(:,n+1)-Underlying_in_Portfolio(:,n+1);
Change_in_cash(:,n+1) = Cash_in_Portfolio(:,n+1)-Old_Cash_in_Portfolio(:,n+1);
Total_transaction_cost(:,n+1) = (abs(Change_in_Underlying(:,n+1))+abs(Change_in_cash(:,n+1)))*Transaction_cost;
Portfolio_Value(:,n+1) = Underlying_in_Portfolio(:,n+1)+Cash_in_Portfolio(:,n+1);
Total_transaction_cost(:,n+1) = Total_Cost(:,n+1);

else

Underlying_in_Portfolio_Percent(:,n+1) = Underlying_in_Portfolio(:,n+1)/Portfolio_Value(:,n+1);
Total_Cost(:,n+1) = Total_Cost(:,n)+Total_transaction_cost(:,n+1);

end
end

% Output Section

Terminal_Value_P = mean(Portfolio_Value(:,NumStep));
Terminal_Floor = mean(Floor(:,NumStep));
Variance_Floor = var(Floor(:,NumStep));
Return_P = (Terminal_Value_P-Initial_Investment)/Initial_Investment;
Risk_P = mean(std(Portfolio_Value,0,1),2);
Net_Transaction_cost = mean(Total_Cost(:,NumStep));
Terminal_Value_Index = mean(Index(:,NumStep));
Return_I = (Terminal_Value_Index-Initial_Investment)/Initial_Investment;
Risk_I = mean(std(Index,0,1),2);

% Graphical Section

subplot(3,1,1);
plot(Portfolio_Value(1,:),"linewidth",2.5, "color", "b");
hold on;
plot(Floor(1,:),"linewidth",2, "color", "r");
hold on;
plot(Index(1,:),"linewidth",1, "color", "g");
ylabel("VALUE", "fontsize",12, "fontweight", "bold");
title("PORTFOLIO VALUE (blue), FLOOR (red), RISKY ASSET (green)", "fontsize",12, "fontweight", "bold")

subplot(3,1,2);
bar(Underlying_in_Portfolio_Percent(1,:), "stack", "r");
title("PERCENTAGE IN RISKY ASSET", "fontsize",12, "fontweight", "bold")

subplot(3,1,3);
hist(Portfolio_Value(:,NumStep),100);
title("HISTOGRAM OF THE TERMINAL VALUE OF THE PORTFOLIO", "fontsize",12, "fontweight", "bold")

figure

subplot(2,1,1);
plot(Portfolio_Value(:,NumStep), "linewidth",1, "color", "g");
title("TERMINAL VALUE OF THE PORTFOLIO FOR ALL THE SIMULATED RISKY ASSET PRICE", "fontsize",12, "fontweight", "bold")

subplot(2,1,2);
plot(Floor(:,NumStep), "linewidth",2, "color", "r");
figure
plot(mean(Total_Cost(),1), "linewidth",2, "color", "r");
ylabel("CUMULATIVE TRANSACTION COST (Rs)", "fontsize",12, "fontweight", "bold");
title("CUMULATIVE TRANSACTION COST (Rs)", "fontsize",12, "fontweight", "bold")

figure

subplot(2,1,1);
plot(Index(1,:), "linewidth",1, "color", "g");
subplot(2,1,2);
hist(Index(:,NumStep),100);
figure
subplot(2,1,1);
plot(Interest_Rate);
subplot(2,1,2);
hist(Interest_Rate,100);
end

Corresponding author
Zongwei Luo can be contacted at: luozw@sustc.edu.cn
A bibliographic study on big data: concepts, trends and challenges

Deepa Mishra
Indian Institute of Technology Kanpur, Kanpur, India
Zongwei Luo
Southern University of Science and Technology, Shenzhen, China
Shan Jiang
Department of Computer Science, Southern University of Science and Technology, Shenzhen, China
Thanos Papadopoulos
School of Business, Management and Economics, University of Sussex, Brighton, UK, and
Rameshwar Dubey
Institute of Operations Management, Symbiosis International University, Nashik, India

Abstract

Purpose – The purpose of this paper is twofold. First, it provides a consolidated overview of the existing literature on “big data” and second, it presents the current trends and opens up various future directions for researchers who wish to explore and contribute in this rapidly evolving field.

Design/methodology/approach – To achieve the objective of this study, the bibliographic and network techniques of citation and co-citation analysis was adopted. This analysis involved an assessment of 57 articles published over a period of five years (2011-2015) in ten selected journals.

Findings – The findings reveal that the number of articles devoted to the study of “big data” has increased rapidly in recent years. Moreover, the study identifies some of the most influential articles of this area. Finally, the paper highlights the new trends and discusses the challenges associated with big data.

Research limitations/implications – This study focusses only on big data concepts, trends, and challenges and excludes research on its analytics. Thus, researchers may explore and extend this area of research.

Originality/value – To the knowledge of the authors, this is the first study to review the literature on big data by using citation and co-citation analysis.

Keywords Big data, Citation and co-citation analysis

Paper type Research paper

1. Introduction

Due to the technological advancements in twenty-first century, the amount of data generated is doubling each year. Meng and Ci (2013) argued that technologies such as cloud computing, Internet of Things (IoT), and social networking mark the beginning of “big data” era. This data are no more measured in terms of gigabyte or terabyte (TB), but in petabyte (PB), exabyte (EB) and zettabyte (ZB) (where, 1PB = 2^{10} TB, 1EB = 2^{40} PB, and 1ZB = 2^{60} EB). In reference to this, Gantz and Reinsel (2012) forecasted that by the end of 2020, 40 ZB of data will be generated.

The evolution of big data has captured the interest of both academics and professionals. Gobble (2013) identified big data as the “next big thing in innovation” and was considered as “the fourth paradigm of science” (p. 34) by Strawn (2012). As per McKinsey & Co, big data is “the next frontier for innovation, competition and productivity”. In the words of McAfee and Brynjolfsson (2012b), big data is the “management revolution” and “brings a revolution in
science and technology” (Ann Keller et al., 2012). In this regard, Brown et al. (2011) claimed that the logic behind these facts lies in the capability of “big data” to change competition by “transforming processes, altering corporate ecosystems, and facilitating innovation”. In fact, big data not only influences competition and growth for individual companies, rather enhances productivity, innovation, and competitiveness for entire sectors and economies. Its significance in scientific research is reflected from the special issues published in Nature (2008) and Science (2011). In their study on big data, Perrey et al. (2013) noted that “retailers can achieve up to 15 to 20 per cent increase in ROI by putting big data into analytics”. Applications of big data can be seen in medical field, retail, finance, manufacturing, logistics, telecommunications, and other industries (Feng et al., 2013).

Scholars have attempted to review the existing literature (Wamba et al., 2015; Gandomi and Haider, 2015), but to the best of our knowledge no study has provided a systematic review using citation/co-citation analysis for understanding the wide variety of research on big data. To address this gap, in this paper we review the articles on big data published from 2011 to 2015. The search was restricted to this time period because the field of big data has witnessed rapid growth in the past five years. Furthermore the main attributes of big data, that is, 5V’s (volume, velocity, veracity, variety, and value) (Russom, 2011; Gartner, 2012; Gandomi and Haider, 2015) were manifested during this period. In reviewing the literature, we aim to comprehend the concept of big data by fulfilling the following objectives: understand the definitions of big data; systematically review the literature on big data using citation and co-citation analysis; synthesize the findings of the literature review; and identify future research directions.

For the purpose of this study, we have chosen the technique of bibliometrics as it provides a way to quantitatively analyse the literature by means of citations and co-citations (Pilkington and Meredith, 2009). In an attempt to analyse the present structure of research on big data, we conducted citation and co-citation analysis. Citation analysis is a quantitative technique that provides information on the degree of influence of a research article on a specific field; whereas, co-citation analysis traces the linkage and connection between the authors and their areas of research. Citation analysis enables researchers to understand when the major articles in a field were published and how their popularity has evolved over time, and hence if an article is still useful for the current research (Pilkington and Meredith, 2009). Co-citation analysis can reveal the major research clusters within a particular field and how they evolve and vary across different journals over time. Leydesdorff and Vaughan (2006 in Pilkington and Meredith, 2009) suggest that data received through co-citation “can be considered as such linkage data among texts, while cited references are variables attributed to texts […] one should realize that network data are different from attributes as data. From a network perspective, for example, one may wish to focus on how the network develops structurally over time”. In this paper, we follow the argument of Pilkington and Meredith (2009) and suggest that there is a need to look at the field “big data” more objectively and answer, “what articles are actually cited in research studies? And to reveal the structure of the interrelationships among articles, what works are commonly cited together (co-cited)” (p. 186).

The paper is structured as follows. Starting with presenting the methodology for reviewing the literature (Section 2), we conduct the literature review of big data in Section 3. It is followed by the presentation of results of citation and co-citation analysis (Section 4). Next, the discussion of current and future trends in big data on the basis of our results precedes the managerial implications of our review analysis (Section 5). Finally, in the last section, conclusions and limitations of the study are presented.

2. Methodology for reviewing the literature on “big data”
The analysis was performed in two stages:

Stage 1: citation analysis was performed to evaluate the citation frequency on a particular document. According to Garfield (1972), the total number of citations on a scientific journal
indicates its significance in that area of research. Moreover, scholars (Sharplin and Mabry, 1985; Culnan, 1986) have emphasized that the impact of heavily cited articles on scientific research is greater than that of less cited articles. Despite the critics of citation analysis, it is still regarded as one of the most commonly used techniques for analysing literature and identifying the most influential author, journal, or work in that particular area of research (MacRoberts and MacRoberts, 1989, 2010; Vokurka, 1996).

Stage 2: co-citation analysis was conducted to investigate the relationships between authors, topics, journals, or keywords, thus elucidating how these groups are related with each other (Small, 1973; Pilkington and Liston-Heyes, 1999). Chen et al. (2010) claimed that co-citation analysis can be conducted either on the basis of authors or publications, where, the former helps in manifesting the social structure and the latter reveals the intellectual structure of research field. For that reason, we considered both author and publications-based co-citation analysis. In publication-based analysis, the numbers of scientific articles which have cited any particular set of two documents are recorded and researchers decipher it as a measure for resemblance of content of the two documents (Figure 1).

Before moving on to citation and co-citation analysis, data were collected from different online databases such as, ISI Web of Science (WoS) and Scopus. Since the number of journals in the database of WoS is limited as compared to Scopus, we restricted ourselves to select relevant papers from Scopus database only. In fact, the process of citation and co-citation analysis has been considerably simplified due to the advancement in IT and online data storage. We searched for the publications that contained the term “big data” in their title, abstract, and keywords. The first document search resulted in 7,527 number of publications which belonged to different subject areas, journals, and languages. On restricting the subject area, we obtained 4,325 publications in the second search. Next, we excluded articles that and were in languages other than English. This search resulted in 4,021 articles. Following the objectives of our study, we restricted those articles to scientific publications (articles, reviews, articles in press, and conference papers) that appeared in peer reviewed journals as these can be considered as “certified knowledge” (Ramos-Rodriguez and Ruiz-Navarro, 2004). For data purification,

we excluded unpublished articles, working papers, and newspaper articles from the database. Finally, we obtained 57 relevant documents. Later on, references and citations were recorded in a database for future analysis. The distribution of articles by journal title is depicted in Table I.

The co-citation analysis was conducted as follows. We analysed the citations of scientific articles received from Step 1 to find out if any pair of reference has been cited together. This co-occurrence gives an indication that these scientific articles apparently share similar thoughts. In this regard, Pilkington and Meredith (2009) pointed that this collection of articles may be termed as “structural knowledge group”. As per Leydesdorff and Vaughan (2006), such groups delineate the intellectual structures of a field. The co-citation analysis was conducted using Bibexcel version 25-03-2014. It is a bibliometric toolbox developed by Olle Persson (Persson et al., 2009) through which connection with other softwares such as Pajek, Excel, and SPSS becomes easy and trouble-free network diagrams can be drawn. The first step in co-citation analysis is data preparation which has been discussed in the aforementioned paragraph. The downloaded data were saved as a text file which was further converted to Bibexcel format (.doc file). In the data pre-processing step, the duplicate documents were identified and removed from Bibexcel which resulted in .oux and .cit file. Finally, data analysis was done by selecting those articles which were highly co-cited and .net file was obtained. As in this study we are interested in performing co-citation analysis based on article, author and keywords, we repeated the previous steps and obtained three different .net files. These files were then exported to Pajek 2.05 where network diagrams were drawn. For better visualization, these diagrams were further refined by removing very thin lines and Kamada-Kawai diagrams were finally drawn (Kamada and Kawai, 1989).

3. Review of big data
In this section, we report on the literature review by discussing the definitions, concepts, and characteristics of big data.

3.1 Definitions and concepts
Although, the term “big data” is ubiquitous these days, its origin dates back to mid-1990s. In support of this argument, Diebold (2012) noted that the term “big data […] probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid-1990s, in which John Mashey figured prominently”. The popularity of big data can be attributed to the fact that this topic was Google searched 252,000 times in November 2011 (Flory, 2012) and then reached the impressive number of 80,10,00,000 hits in October 2015. Figure 2 depicts the Google search results for big data. The bold line represents the progress in the field till 2015 and dotted line reveals the future trend.

<table>
<thead>
<tr>
<th>Journals</th>
<th>Number of articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data Research (BDR)</td>
<td>13</td>
</tr>
<tr>
<td>International Journal of Production Economics (IJPE)</td>
<td>10</td>
</tr>
<tr>
<td>International Journal of Production Research (IJPR)</td>
<td>6</td>
</tr>
<tr>
<td>Harvard Business Review (HBR)</td>
<td>6</td>
</tr>
<tr>
<td>MIS Quarterly Executive (MIS-QE)</td>
<td>5</td>
</tr>
<tr>
<td>Supply Chain Management: An International Journal (SCMIJ)</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Business Research (JBR)</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Business Logistics (JBL)</td>
<td>4</td>
</tr>
<tr>
<td>MIT Sloan Management Review (MIT Sloan)</td>
<td>4</td>
</tr>
<tr>
<td>McKinsey Quarterly (McKQ)</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
</tr>
</tbody>
</table>

Table I. Distribution of the articles by journal title
In literature, several attempts have been made by the researchers to define big data. For instance, “Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse”. This definition by Manyika (2012) is not confined to data size, since data sets will increase in the future. It highlights the necessity of technology to cope up with the rapid growth in available data. Furthermore, there are other characteristics also that have been put forward to comprehend this concept in a better way. For example, the three dimensions, volume, variety, and velocity (or the three V’s) can be considered as both characteristics and challenges while dealing with big data. According to Gartner, Inc., “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making” (Gartner IT Glossary, n.d.). A similar definition was provided by TechAmerica Foundation’s Federal Big Data Commission (2012) where “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information”. Scholars (Chen et al., 2012; Kwon et al., 2014) claimed that these 3V’s have come up as a common framework for describing big data. Recently, Dubey et al. (2015) viewed big data as the “data whose sources are heterogeneous and autonomous; whose dimensions are diverse; whose size is beyond the capacity of conventional processes or tools to effectively and affordably capture, store, manage, analyze, and exploit; and whose relationships are complex, dynamic, and evolving”.

3.2 Big data characteristics

Volume reflects the magnitude of data that has increased drastically in the past few years. The size of big data may vary from multiple TB to PB. In 2011, Russom defined volume as the “the large amount of data that either consume huge storage or entail of large number of records data”. Schroek et al. (2012) noted that in the survey conducted during mid-2012 by IBM, just over half of the 1,144 respondents considered data sets over one TB to be big data. As the amount of data crossing the internet per second has increased tremendously, firms have got an opportunity to work with many petabytes of data in a single data set. This piece of information is evident from the example of Wal-Mart, where it is estimated that more than 2.5 PB of data every hour are collected from its customer transactions (McAfee and Brynjolfsson, 2012a, b).

Variety refers to the “structural heterogeneity in a dataset” (Gandomi and Haider, 2015). In the work of Russom (2011), variety in a big data is defined as the “data generated from greater variety of sources and formats contain multidimensional data fields”. Due to the advancement in technology, firms these days are using various types of data; structured, semi-structured, and unstructured. Structured data refer to the tabular data available in spreadsheets and amounts to only 5 per cent of all the existing data (Cukier, 2010), whereas,
unstructured data are comprised of text, images, audio, and video. A continuum between these two types of data is referred as semi-structured data which does not follow any particular standards. A classic example of semi-structured data is Extensible Markup Language which is used for exchanging data on internet.

Velocity refers to the “rate at which data are generated and the speed at which it should be analysed and acted upon” (Gandomi and Haider, 2015). Owing to the rapid growth in digitalization, data are getting generated at an exceptional rate which drives the need for real-time analytics and evidence-based planning. According to Cukier (2010), the retail company Wal-Mart deals with more than one million transactions per hour. Since the conventional data management systems are inefficient to handle large data sets, big data technologies act as a safeguard by helping the firms in creating real-time intelligence from high volumes of “perishable” data (Gandomi and Haider, 2015).

Besides these 3V’s, three other characteristics: veracity, variability, and value have been introduced. Veracity known as the fourth V reflects the “unreliability inherent in some sources of data” (Gandomi and Haider, 2015). IBM reported that data quality behaves as an important challenge for big data since the inherent unpredictability of data cannot be removed even by best data cleansing methods. As per White (2012), veracity highlights “the importance of data quality and the level of trust in a data source”.

Variability (and complexity) are the two dimensions of big data which were introduced by SAS. Usually, the velocity of big data is inconsistent and has regular ups and downs which reflects the variation in data flow rates. This variation is termed as variability of big data (Gandomi and Haider, 2015). Complexity in big data arises from innumerable sources. Thus, there exists a need to connect, match, cleanse, and transform data received from these sources (Gandomi and Haider, 2015).

Value reflects the “economic benefits from the data” (Forrester, 2012; Oracle, 2012). A typical big data contains substantial amount of information, hence it is necessary to acknowledge what is meaningful and extract that data for further analysis. To sum up, these characteristics of big data create a chance for firms to gain competitive advantage. This allows the firms to alter the way they interact and serve their customers.

Big data analytics is another important area which is getting recognition among research communities. For instance, McGahan (2013) argued that specialized analytics are required to deal with big data as it cannot be managed with the traditional software programmes such as Excel. In this direction, a consolidated study was done by Gandomi and Haider (2015) where they mainly focussed on analytical methods and presented various techniques and tools for improving decision-making abilities (i.e. text analytics, audio analytics, video analytics, social media analytics, and predictive analytics).

4. Big data: citation and co-citation analysis

In this section, we identify the influential scientific contributions in the field of big data. This section is divided in two sub-sections. In the first, we discuss the results of our citation analysis, whereas in the second section we present and comment on the results of the co-citation analysis.

4.1 Citation analysis (2011-present)

The most influential article during this period is the seminal work published by McAfee and Brynjolfsson (2012a, b) which has been cited 111 times. The authors highlighted the significance of big data by stating that it allows the managers to measure and thus, acquire thorough knowledge of the business which can be used to improve decision making and performance. They also claimed that big data enable firms to take decisions based on evidence rather depending upon instincts. In the same year, an important contribution was made by Davenport and Patil (2012) where the authors discussed about
the role of data scientist and considered them as the key to grasp the opportunities offered by big data. This work received 52 citations which reflects the significance of the article in this field. Furthermore, the first article of this period by Brown et al. (2011) which has been cited 36 times, emphasized on the strategic opportunities presented by big data and at the same time, they viewed it as a key factor in the prosperity of the nation. The peaks of Figure 3 demonstrate the influential works published between 2011 and 2015. The citation frequency of influential articles of the period (2011-2015) can be seen in Table AI.

Figure 4 demonstrates the changing pattern of publications in each year, starting from 2011 until the end of 2015. As can be clearly seen from the figure that major work on big data initiated in 2011 and since then it has been growing continuously. Interestingly, a dramatic rise in publications of this field can be observed after 2014, as the number of publications rose from 11 in 2014 to 27 in 2015. Therefore, we can sum up that this area has the ability to draw and retain the interest of scholars and practitioners.

Turning our focus now to identify the journal contributions within this particular area, we analyse the number of publications in the ten journals identified during our review from 2011 to 2015. It is evident from the bar graph (Figure 5) that the longest bar represents Big Data Research, which reflects that this journal has given the maximum contribution to this field.

Furthermore, the graph illustrates that McKinsey Quarterly, Harvard Business Review, International Journal of Production Economics, and International Journal of Production Research) are among the major contributing journals publishing articles on this area.

A bibliographic study on big data

Figure 3.
Frequency distribution of most cited articles (2011-2015)

Figure 4.
Frequency distribution of number of articles published during 2011-2015
4.2 Results of co-citation analysis

In this section, we first report the results of article co-citation analysis for the period 2011-2015 on big data (see Figure 6). In the figure, the different research works are presented as nodes and their relationships in arcs that have different width. This reflects the difference in the nature of relationship between these articles. The thick arcs extrapolate the strong relationship between the two co-cited articles. In contrast, the thin arcs indicate that the co-cited articles apparently do not share common ideas. For instance, the thick arc connecting McAfee and Brynjolfsson (2012a, b) and Manyika et al. (2011) reflects that these articles share common thoughts. A similar pattern can be noticed between McAfee and Brynjolfsson (2012a, b) and Barton and Court (2012). However, the thin arc between Manyika et al. (2011) and Davenport et al. (2012) reveal that they do not share common ideas. The maximum co-citation value of the publications can be seen in Table AII.

It is to be noted that the dark arcs in Figure 7 reflect that the authors belong to same area of research while thin arcs represent that they are less likely to have same area of interest. For instance, Wang and Li are connected by dark lines; whereas, Chen and Xu are connected...
by light lines. The figure also illustrates that McAfee and Brynjolfsson, Waller and Fawcett and Brown, Court and Willmott, belong to the same area of research when considered independently while differing when compared to each other.

The dense network in Figure 8 clearly indicates that “big” and “data” are the two most popular keywords in this area of research. Indeed it appears that significant associated work has been conducted in the field of supply chain and analytics. The figure emphasizes that researchers are also developing frameworks and exploring the challenges faced by big data.

5. Current and future trends in big data
Our results reveal that the concept of big data has evolved over the years. Due to the significant advances in technology, huge amount of data is being generated every year.
This proliferated data, termed as “big data”, have become a hot topic for discussion by capturing the attention of academics, professionals, and governments all around the world. Attempts were made to answer the question “What is big data?” To date, there is no clear or universally accepted definition of big data (Mayer-Schönberger and Cukier, 2013). As the name itself suggest, “size” was conceived as its main characteristic. But later on, Gartner, Inc. observed that size may not be the only criteria to adjudge a “data” as “big data”. They proposed 3Vs; volume, variety, and velocity as the new characteristics to describe big data. Another characteristic, veracity which is widely accepted in literature and is commonly known as the fourth V was added by IBM. In addition, value and variability (and complexity) came up as the fifth and sixth characteristics which were introduced by Oracle and SAS, respectively.

Taking the above points further, the next question that came up was “What are the challenges in handling big data?” The magnitude of data (volume) was not a big challenge as it can be resolved by using efficient computing systems. However, the actual challenge “was to deal with diversified data types (variety), timely response requirements (velocity), and uncertainties in the data (veracity)” (Mishra and Sharma, 2015, p. 28). Besides dealing with the traditional structured data, applications must be developed in order to handle both semi-structured and unstructured data including, for instance, text, images, video, and voice (Mishra and Sharma, 2015). Another challenge refers to the responses that are not received in a timely manner. This may be due to the lack of insufficient sources needed to gather, store, and analyse big data but within a particular time frame (Mishra and Sharma, 2015). Furthermore, it is a difficult task to identify the difference between reliable and unreliable data as the uncertainty inherent in data cannot be excluded even by implementing best data cleaning methods (Jin et al., 2015).

The next important issue related to big data is to identify its area of applications. It was in the year 2012, when McKinsey, the leading global strategic management consulting firm realized that big data has found its application in nearly every sector of global economy (Manyika et al., 2011). For instance, considering the significance and value of big data, Obama Administration invested around US$200 million to officially launch the Big Data Research and Development Initiative in March 2012. This initiative involved six federal government agencies; Department of Defense, Defense Advanced Research Projects Agency, Department of Energy, National Institutes of Health, National Science Foundation, and US Geological Survey (Kalil, 2012). In addition, scholars have attempted to identify its role in health care (Priya and Ranjith Kumar, 2015; Wu et al., 2015; Huang et al., 2015; Qin et al., 2015) and supply chain and logistics (Zhong et al., 2015, 2017; Schoenherr and Speier-Pero, 2015; Dubey et al., 2015). For a detailed overview of big data research, readers may refer to Table AI.

5.1 Limitations and future research directions
We believe that we have used bibliometric and network analysis tools to generate some useful insights on big data. However the bibliometric, citation, and co-citation analyses have the following limitations:

(1) The findings of the review are based on the last five-year (2011-2015) publications. This list is not exhaustive, but comprehensive, covering those scientific journals that contain highly cited and co-cited articles.

(2) Our method of conducting co-citation analysis is not the only method. There are different methods to conduct co-citation analysis (see Pilkington and Fitzgerald, 2006; Pilkington and Meredith, 2009; Colicchia and Strozzi, 2012). In this paper, we used the guidelines by Pilkington and Meredith (2009).
The findings were based on searches using the particular keyword “big data”. This technique has been used in the past by scholars (e.g. Eksoz et al., 2014; Gunasekaran et al., 2015).

We have limited our analysis to few journals and articles. Though the cited articles are considered as potential articles but we may not ignore the possibility of missing some potential articles which could have provided interesting insights into the emerging field.

The present study used the bibliographic technique of citation and co-citation analysis. This analysis involved the assessment of 57 articles published over a period of five years (2011-2015). This clearly suggests that big data in last five years has attracted significant attentions from diverse communities. The big data in future may be regarded as one of the research strands which revolve around the potential use of unstructured data, semi-structured data, and structured data to advance current management theories. The big data may pose some challenges due to heterogeneous nature; however we believe that this heterogeneity presents a unique opportunity to further investigate the applicability of existing theories on information management, or develop innovative theories that may explain the creation, storage, processing, and management of big data. The future studies are likely to be benefited from the possible integration of IoT, big data, and relational exchange theories to explain complex behaviour of supply chains, organizations, people, and technology. Finally, we submit that the big data has potential to redefine the existing research methods by bridging potential gaps between qualitative and quantitative research methods.

Below we outline potential research gaps with regards to big data research:

1. Big data can offer multiple insights into current debates surrounding sustainable operations and the need to search, assess, store, process, analyse, and manage data considering the social, economic, and environmental dimensions of sustainability.

2. Big data has immense applications in explaining coordination in humanitarian operations and supply chains, and disaster resilience.

3. The sustainable consumptions and production theories would need to be examined using large data sets. We believe that current field can be largely benefitted with the use of big data.

4. Big data can help providing insights in supply chain design by focussing on crucial properties of supply chain including agility, adaptability, alignment, and integration.

5. To be able to process and analyse big data, particular methods, based on both quantitative and qualitative research methods would need to be developed, using, for instance, methods from statistics and management science.

5.2 Managerial implications

We underline the importance for managers to attend to the big data era, and the different methods and analytics for big data, as well as to how to explore big data in order to achieve the sustainable competitive advantage. Processing of big data needs to abide to the objectives of each organization and supply chain. We also believe that it is not only processing and analysis of big data that needs to be considered; rather, particular performance measurements and metrics need to be assigned in order to understand whether the use of big data leads to competitive advantage, or in improved decision making at strategic, tactical, and operational levels. Furthermore, we underline the importance of appropriate and robust big data collection and analysis, as well as investments on infrastructure and human resources in order to be able to take advantage of the big data
analytics and put it into practice. Frequent auditing of such tools and their alignment with business objectives is also required to ensure any misalignments are adjusted. To achieve this, the participation of the relevant managers and senior executives is required.

6. Conclusions
Drawing broadly on bibliometric and citation and co-citation analysis, we presented an extensive review of literature on big data over the period of five years (2011-2015). We offered insights regarding the contributions of scientific journals towards advancing big data-related research and the contributions of researchers to the emerging field of big data. To our knowledge, this is the first study attempting to classify highly cited and co-cited works related to this field. However in other fields these tools have attracted growing attentions. The current research of ours has made significant contributions towards the theory by identifying fertile research areas. Although we have only cited potential applications from operations management field, the applications goes far beyond management fields and other interdisciplinary areas where we have not focussed. It is important for companies to adopt big data analytics in an attempt to comprehend the trend in customer behaviour and provide them improved and customized services. Despite the limitations, we believe that our study provides food for thought and encouragement for researchers to further investigate the field of big data.

References


Oracle (2012), Big Data for the Enterprise, Oracle, Redwood Shores, CA.


Further reading


(The Appendix follows overleaf.)
## Appendix

Table A1. Influential articles and their maximum citation count during the period (2011-2015)

<table>
<thead>
<tr>
<th>Article</th>
<th>Topic</th>
<th>Research strategy</th>
<th>Research method</th>
<th>Maximum citation count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown <em>et al.</em> (2011)</td>
<td>Discusses important ways in which big data could change competition: by transforming processes, altering corporate ecosystems, and facilitating innovation</td>
<td>Conceptual</td>
<td>n/a</td>
<td>36</td>
</tr>
<tr>
<td>Bughin <em>et al.</em> (2011)</td>
<td>Discusses four principles that will help CEOs and other corporate leaders in seizing the potential of big data</td>
<td>Conceptual</td>
<td>n/a</td>
<td>5</td>
</tr>
<tr>
<td>McAfee and Brynjolfsson (2012a, b)</td>
<td>Emphasizes that data-driven decisions are better decisions as by using big data, managers can decide on the basis of evidence rather than intuition</td>
<td>Conceptual</td>
<td>n/a</td>
<td>111</td>
</tr>
<tr>
<td>McAfee and Brynjolfsson (2012a, b)</td>
<td>Highlights the significance of big data by stating that it allows the managers to measure and thus, acquire thorough knowledge of the business which can be used to improve decision making and performance</td>
<td>Conceptual</td>
<td>n/a</td>
<td>66</td>
</tr>
<tr>
<td>Davenport and Patil (2012)</td>
<td>Highlights the role of data scientists and considers them as the key to realizing the opportunities presented by big data</td>
<td>Conceptual</td>
<td>n/a</td>
<td>52</td>
</tr>
<tr>
<td>Barton and Court (2012)</td>
<td>Offers a useful guide for leaders and managers who want to take a deliberative approach to big data</td>
<td>Conceptual</td>
<td>n/a</td>
<td>19</td>
</tr>
<tr>
<td>Davenport <em>et al.</em> (2012)</td>
<td>Focuses on how the insights of big data differ from what managers might generate from traditional analytics</td>
<td>Conceptual</td>
<td>n/a</td>
<td>32</td>
</tr>
<tr>
<td>Ross <em>et al.</em> (2013)</td>
<td>Provides four practices that a company should adopt in order to improve their operations in ways that rivals cannot easily replicate</td>
<td>Case study</td>
<td>Interviews</td>
<td>3</td>
</tr>
<tr>
<td>Brown <em>et al.</em> (2013)</td>
<td>Clarifies the most important tasks for executives and then sets out some critical questions whose answers will inform any reconfiguration of the C-suite</td>
<td>Conceptual</td>
<td>n/a</td>
<td>4</td>
</tr>
<tr>
<td>Biesdorf <em>et al.</em> (2013)</td>
<td>Identifies the challenges of implementing crowdsourcing platforms and shows how CIOs and other organizational leaders can build the absorptive capacity necessary to extract business value from crowd sourced data</td>
<td>Conceptual</td>
<td>n/a</td>
<td>4</td>
</tr>
<tr>
<td>Biesdorf <em>et al.</em> (2013)</td>
<td>Discusses strategic plan and its core elements to start with big data</td>
<td>Conceptual</td>
<td>n/a</td>
<td>5</td>
</tr>
<tr>
<td>Waller and Fawcett (2013a)</td>
<td>Develops a 2 × 2 model to explain the role of predictive analysis in the theory development process</td>
<td>Conceptual (theoretical model)</td>
<td>n/a</td>
<td>5</td>
</tr>
<tr>
<td>Waller and Fawcett (2013b)</td>
<td>Illuminates the myriad of research opportunities where SCM intersects with DPB</td>
<td>Conceptual</td>
<td>n/a</td>
<td>25</td>
</tr>
<tr>
<td>Lee <em>et al.</em> (2014)</td>
<td>Provides a three dimensional framework to describe the role of chief data officer</td>
<td>Conceptual</td>
<td>n/a</td>
<td>6</td>
</tr>
<tr>
<td>Hazen <em>et al.</em> (2014)</td>
<td>Introduces the data quality problem in the context of SCM and proposes methods for monitoring and controlling data quality</td>
<td>Conceptual</td>
<td>n/a</td>
<td>8</td>
</tr>
</tbody>
</table>
Table AII. Most frequently co-cited articles (2011-2015)

<table>
<thead>
<tr>
<th>Publications</th>
<th>Maximum co-cited value</th>
<th>Publication most co-cited with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacobs (2009)</td>
<td>2</td>
<td>Manyika et al. (2011)</td>
</tr>
<tr>
<td>Jacobs (2009)</td>
<td>2</td>
<td>Lavalle et al. (2011)</td>
</tr>
<tr>
<td>Beath et al. (2012)</td>
<td>2</td>
<td>Davenport et al. (2012)</td>
</tr>
<tr>
<td>Kiron (2011), Shockley (2011)</td>
<td>2</td>
<td>Lavalle et al. (2011)</td>
</tr>
<tr>
<td>Davenport (2006)</td>
<td>2</td>
<td>Lavalle et al. (2011)</td>
</tr>
<tr>
<td>Hazen et al. (2014)</td>
<td>2</td>
<td>Manyika et al. (2011)</td>
</tr>
<tr>
<td>Barton (2012), Court (2012)</td>
<td>2</td>
<td>Lavalle et al. (2011)</td>
</tr>
<tr>
<td>Lavalle et al. (2011)</td>
<td>2</td>
<td>Waller (2013), Fawcett (2013)</td>
</tr>
<tr>
<td>Manyika et al. (2011)</td>
<td>2</td>
<td>McAfee (2012), Brynjolfsson (2012)</td>
</tr>
<tr>
<td>Manyika et al. (2011)</td>
<td>2</td>
<td>Syed et al. (2013)</td>
</tr>
</tbody>
</table>

Corresponding author
Zongwei Luo can be contacted at: luozw@sustc.edu.cn

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com

A bibliographic study on big data

573
Past, present and future of contact centers: a literature review

Morteza Saberi, Omar Khadeer Hussain and Elizabeth Chang
University of New South Wales Canberra, Canberra, Australia

Abstract

Purpose – Contact centers (CCs) are one of the main touch points of customers in an organization. They form one of the inputs to customer relationship management (CRM) to enable an organization to efficiently resolve customer queries. CCs have an important impact on customer satisfaction and are a strategic asset for CRM systems. The purpose of this paper is to review the current literature on CCs and identify their shortcomings to be addressed in the current digital age.

Design/methodology/approach – The current literature on CCs can be classified into the analytical and the managerial aspects of CCs. In the former, data mining, text mining, and voice recognition techniques are discussed, and in the latter, staff training, CC performance, and outsourced CCs are discussed.

Findings – With the growth of information and communication technologies, the information that CCs must handle both in terms of type and volume, has changed. To deal with such changes, CCs need to evolve in terms of their operation and public relations. The authors present a state-of-the-art review of the challenges in identifying the gaps in order to have the next generation of CCs. Lack of an interactive CC and lack of data integrity for CCs are highlighted as important issues that need to be dealt with properly by CCs.

Originality/value – As far as the authors know, this is the first paper that reviews CCs’ literature by providing the comprehensive survey, critical evaluation, and future research.

Keywords Decision making, Customer service, Customer service management, Customer information, Database management

Paper type Literature review

1. Introduction

Customer relationship management (CRM) is a framework for managing a company’s interactions with current and future customers that will assist them in the better management of customer queries (Christen, 2012). In today’s highly competitive business world, it is essential that every organization has an efficient and smart CRM system. In the past two decades, providing efficient service has become essential for organizations, especially service delivered through customer contact centers (CCs). CCs, as the organization’s touch point, have a considerable effect on customer experience and retention. In a CRM framework, CCs form the main interaction point between a customer and the CSR. It has been shown that 70 percent of all business interactions are handled in CCs (Dhesi et al., 2011). Having a customer-focused CC is thus crucial for any organization. A study by Bain & Company found that, for many companies, an increase of 5 percent in customer retention can increase profit by 25-95 percent (Reichheld, 2001). Another report mentions that “Every single step needs to be done well – it is not sufficient just to have a wonderful set of people in the call center, you need to sustain that experience at every point of contact and address the total customer experience (TCE)” (Millard, 2006). In fact, CCs are of prime importance for all types of CRM.

With the emerging information and communication technologies (ICT) paradigms, the basic operation of a CC too has evolved. With the high speed and volume of information...
now acquired by companies, new types of complaints and drawbacks exist that cannot be
dealt with by traditional CCs. Today’s customers, with their access to powerful ICT
technologies, have a varied range of expectations and needs which companies have
to consider and respect. Issues, such as enabling customers with 24-hour access to
communication channels, and the unstructured relationship between customers and
organizations, show that it is critical to develop an intelligent CC. In other words, it is both
important and advantageous to the target company that it constructs appropriate
infrastructure to capture the customers’ voices and ensure their expectations are met.
However, in order to achieve this, companies need to make better sense of the information
available to them. Bucher et al. (2009) state that: “It cannot be disputed that information has
become a major competitive factor in today’s business world”. This is further complicated
by the fact that an increasing use of ICT has led to the production of a huge amount of data
as well as diverse communication channels, both of which affect the performance of CCs
(LaValle et al., 2011). According to Stringfellow et al., customer access to rich channels of
communication, such as the telephone, lead to data complexity which, in turn, can produce
inconsistencies on the CC side. Moreover, sometimes the customers’ unfamiliarity with
various ICT technologies also produces inconsistency on the CC side (Meuter et al., 2000).

Furthermore, apart from simply considering the huge volume of information, CCs these
days also have to consider the various types of information being generated. Currently,
about 80 percent of data are semi-structured or unstructured (Herschel and Jones, 2005).
However, traditional CRM systems mostly deal with structured data and provide little,
if any, support for semi-structured or unstructured data. In order for the current CCs to deal
with such a change in the type of information, they need an efficient framework that enables
them to understand the diverse and huge volume of information for better decision making.
In this paper, we present the state-of-the-art of the existing literature and highlight the gaps
that need to be addressed that will enable CCs to deal with the aforementioned issues.

In summary, the contributions of this paper are as follows:

- We explain how call centers have evolved into CCs and discuss their various dimensions.
  We compile 18 definitions of CCs from the current literature, and we propose a
  comprehensive definition which we believe is best suited in the current age (Section 2).

- We present the first study to review the CC literature, which comprises 90 papers
  (Sections 3-5).

- We classify the surveyed literature into two groups, namely, analytical and
  managerial studies. The analytical studies incorporate call monitoring and additional
  CC administration tasks (Section 3), while the managerial studies incorporate
  CC performance (Section 4.1), customer service representatives (Section 4.2),
  and outsourcing the CC (Section 4.3).

- We discuss two existing research gaps based on the reviewed papers: a lack of
  an interactive CC in the current literature is highlighted as the first research gap
  (Section 5.1). We also demonstrate the lack of data integrity in CCs and elaborate its
  negative effect on CC performance as the second research gap (Section 5.2).

- We propose a research agenda for CCs by using big data techniques. Big data
  analytical techniques open a new direction for CC research and have the potential to
  consequently enhance their performance (Section 5.3).

2. CC
The origins of the modern call center can be found in the 1973 Continental Airlines booking
system developed by Rockwell Galaxy. According to Duggirala et al. (2011), a call center
may be defined as: “a centralized, specialized, and dedicated operation for both inbound and
outbound communication handling, wherein employees using computers receive inbound or make outbound telephone calls”.

Since 1973, new developments in communication technology have resulted in call centers growing into “contact centers.” Some companies have chosen to retain call centers rather than diversify into CCs. Most current literature defines CCs as the physical station that allows an organization’s customers to contact them by different communication channels, such as telephone, touch-point telephone, fax, letter, e-mail, web, online live chat, and social networks. These communication channels have been studied in either a conceptual or an analytical way, for example: e-mail (Legros et al., 2013; Gupta et al., 2012; Li et al., 2012; Jolai et al., 2008; Malik et al., 2007; Nenkova and Bagga, 2003; Rodenstein, 2004), telephone (Murphy and Cerqua, 2012; Ali, 2011; Byrd et al., 2008; Millard and Hole, 2008; Balakrishnan and Munisamy, 2007; Koole, 2004; Armony and Maglaras, 2004; Lewis et al., 2002), instant message (Luo and Zhang, 2013; Sparks, 2012), live chat (Sparks, 2012; Mehrotra, 2003; Padmanabhan and Kummamuru, 2007; Steul, 2000), and social networks (Acharya et al., 2013; Burns and Friedman, 2012; Bordoloi et al., 2011; Schuster et al., 2011).

Table I lists 18 definitions of CCs from the current literature. Most of these definitions refer to the same concept when defining CCs. We, based on current trends in data science, propose the following definition for a CC:

Contact centers are worksites where staff interact with customers over available multi-communication channels such as: telephone, email, touch-point telephone, fax, letter, web, online live chat and social networks. This center is equipped with customized intelligent tools that enable the center to have clean and integrated data as well as be empowered with customized knowledge.

Table II lists nine different communication channels which can be divided into two types: one-way and two-way communication. These channels can be coupled with voice or text data or both forms of data. This table demonstrates that unstructured data spans across CC databases and is therefore it is extremely important to deal with it effectively. For communication tools that only receive voice data, this matter becomes even harder as another step is needed to transform the data to unstructured text. It should be noted that although some communication channels provide two-way communication, their degree of interaction may differ. Each of these communication tools has advantages and disadvantages. For example, although e-mail is less interactive than the telephone, it has a better capability for saving and producing clean data. In the next sections, we demonstrate that there is a research gap in the literature in relation to examining the advantages and disadvantages of various communication channels and providing efficient frameworks to deal with their disadvantages.

CCs have been studied in various industries and countries which demonstrates their importance. For example, previous studies have examined a diversity of international and national industries including: health (Guillot and Fung, 2010; Rohleder et al., 2013; Malm et al., 2013; Liebow et al., 2012; Bowers and Fish, 2013), pharmaceutical and medical, mobile telecommunications (Anaman et al., 2008), water utilities (Olstein, 2009), academic libraries (Murphy and Cerqua, 2012), car rental (Takeuchi et al., 2007, 2009), Amazon.com (Keblis and Chen, 2006), banking (van Dun et al., 2012; Hakan Ozkan, 2012; Surana and Singh, 2012), telecommunications (Tate and van der Valk, 2008), internal IT helpdesks (Padmanabhan and Kummamuru, 2007), telecom service providers in Europe (Visweswariah et al., 2010), the service industry (van der Aa et al., 2012) and call centers in Bangalore (Ghosh, 2013), India (Surana and Singh, 2012; Poster, 2013; Taylor et al., 2013; Das et al., 2013; Aneesh, 2012; Nandialath et al., 2012), the Philippines (Hechanova, 2013), Germany (Gerpott, 2012), South Africa (Hunter and Hachimi, 2012), Hungary (Dezső et al., 2010), Australia (Owens, 2014), Malaysia (Abdullateef and Salleh, 2013), Egypt (Kamel and Hussein, 2008), and the Netherlands (van der Aa et al., 2012).
<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Contact center definitions in the literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Millard et al. (2004)</td>
<td>“Contact Centers are regarded as a primary mechanism for businesses to talk to their customers” (Millard et al., 2004)</td>
</tr>
<tr>
<td>Armony and Maglaras (2004)</td>
<td>“Many organizations use customer contact centers as an important channel of communication with their customers” (Armony and Maglaras, 2004)</td>
</tr>
<tr>
<td>Koole (2004)</td>
<td>“A call center is a collection of resources (typically agents and ICT equipment) capable of handling customer contacts by telephone. If the call center handles not only telephone contacts but also contacts by fax, email, and so forth, then it is usually called a customer contact center” (Koole, 2004)</td>
</tr>
<tr>
<td>Whitt (2006)</td>
<td>“A contact center is a collection of resources providing an interface between a service provider and its customers” (Whitt, 2006b)</td>
</tr>
<tr>
<td>Mendes et al. (2006)</td>
<td>“A call center handles by phone the customer contacts of several customer companies. If the call center also uses other means of communication, such as email or post, it is called a customer contact center (CCC) or outsourcer” (Mendes et al., 2006)</td>
</tr>
<tr>
<td>Park (2007)</td>
<td>“Most medium- to large-size businesses operate customer contact centers or call centers to provide services to customers. With the advancement of Internet technology, many modern contact centers support various channels of customer interactions including telephony, emails, web-page fill-in forms and instant messaging” (Park, 2007)</td>
</tr>
<tr>
<td>Padmanabhan and Kummamuru (2007)</td>
<td>“Contact centers (or Call Centers) is a general term for help desks and customer service centers” (Padmanabhan and Kummamuru, 2007)</td>
</tr>
<tr>
<td>Helber and Henken (2010)</td>
<td>“Contact centers are the multi-channel successors of phone-based call centers. Customers can use phone, fax, e-mail, etc., to reach the agents working in an inbound contact center in order to receive some kind of service” (Helber and Henken, 2010)</td>
</tr>
<tr>
<td>Jolai et al. (2008)</td>
<td>“A contact center is a collection of resources providing an interface between the service provider and its remote customers. The interface can be through any type of media – telephone, email, fax, paper, chat sessions and the Web” (Jolai et al., 2008)</td>
</tr>
<tr>
<td>Takeuchi et al. (2009)</td>
<td>“A contact center is a general term for customer service centers, help desks, and information phone lines” (Takeuchi et al., 2009)</td>
</tr>
<tr>
<td>Visweswariah et al. (2010)</td>
<td>“Customer contact centers are, in general, one of the primary, and in many cases the only, interface between an enterprise and the customer base that it caters to” (Visweswariah et al., 2010)</td>
</tr>
<tr>
<td>Soujanya and Kumar (2010)</td>
<td>“A telephony call center, now called a Contact Center, is composed of a set of resources (personnel or agents, computers and telecommunication equipment), which enables the delivery of services via telephone lines” (Soujanya and Kumar, 2010)</td>
</tr>
<tr>
<td>Duggirala et al. (2011)</td>
<td>“A contact center can handle customer communications over many channels and modalities (e.g. email, chat, telephone calls, etc.)” (Duggirala et al., 2011)</td>
</tr>
<tr>
<td>Rijo et al. (2012)</td>
<td>“Call centers became ‘contact centers’ with the use of interaction channels such as e-mail, fax, short message service, chat or web” (Rijo et al., 2012)</td>
</tr>
<tr>
<td>Qin et al. (2012)</td>
<td>“Contact centers are modern versions of call centers, capable of managing all client contacts through a variety of mediums such as telephone, fax, letter, email and increasingly, online live chat, etc. (Qin et al., 2012)</td>
</tr>
<tr>
<td>Dudin et al. (2013)</td>
<td>“A contact center is a centralized office used by companies for receiving and servicing its clients’ requests through a variety of media” (Dudin et al., 2013)</td>
</tr>
<tr>
<td>Owens (2014)</td>
<td>“Call centers are understood to be worksites where staff provide services over the telephone to customers at remote locations. Contact centers is the more recent terminology for similar work where staff interact with customers, not only over the telephone, but also over the Internet and by fax, accessing and entering customer data as they are communicating” (Owens, 2014)</td>
</tr>
</tbody>
</table>

There are two sections in a CC, as described in the literature, which perform quite different roles, customer support and the help desk. A help desk is the support point for internal employees in a large organization and provides IT support to employees. Customer support provides support to customers when they have issues relating to the use of company products. Customer support is an important part of customer service and helps customers use company products efficiently. As evidence of its importance, various classes of customer...
support have been used in organizations and companies. Frequently asked questions and online help libraries are often a suitable and efficient way of helping and supporting customers in their use of company products. In some situations, a higher level of support to customers is required. Companies define three types of customer support to meet the different expectations of customers: proactive support automation, self-support automation, and assisted support automation.

In summary, there are four similar types of customer touch points: the CC, call center, help desk, and customer support. The differences between each are described in Table III. Customers may deal with either a CC, call center or customer support departments and sections. In terms of structure, call centers and customer support departments are subdivisions of a CC although it is possible that, in some companies, this relationship is not considered.

2.1 CC complexity

With the rapid development and adoption of ICT, the ways by which customers communicate with various companies have changed. Recent research has shown that, these days, the customers’ preferred method for contacting a company is via the internet with an ongoing decline in preference for in-person contact (Christen, 2012). CCs’ operations are too complex. These centers require a combination of technology, human talent and task procedures in order to deliver an appropriate and efficient performance (Miciak and Desmarais, 2001). The high number of calls is one of the complexities of these centers, specifically for large organizations. As an example, Amazon receives millions of e-mail messages and voice calls annually (Keblis and Chen, 2006). Figure 1 depicts the different types of customers who are able to access some or all communication channels. This figure also demonstrates the various parts of the CC along with typical analytical and managerial tasks. The next section discusses these two important aspects of CCs.

<table>
<thead>
<tr>
<th>Table II. Data format of communication channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Telephone</td>
</tr>
<tr>
<td>Touch-tone phone</td>
</tr>
<tr>
<td>E-mail</td>
</tr>
<tr>
<td>Fax</td>
</tr>
<tr>
<td>Letter</td>
</tr>
<tr>
<td>Online live chat</td>
</tr>
<tr>
<td>Instant feedback (community blogs and forums)</td>
</tr>
<tr>
<td>Automated online assistant</td>
</tr>
<tr>
<td>Online help</td>
</tr>
</tbody>
</table>

Notes: a, the presence of various data formats in each communication channel. b, customers do not expect a direct reply from the organization.

<table>
<thead>
<tr>
<th>Table III. Differences between the four types of customer support in the contact center literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call center</td>
</tr>
<tr>
<td>Help desk</td>
</tr>
<tr>
<td>Customer support</td>
</tr>
<tr>
<td>Contact center</td>
</tr>
</tbody>
</table>

Note: ✓, the presence of various data formats in each communication channel
3. Analytical techniques in the CC: state-of-the-art

Text mining, data mining, and queuing models are three common analytical techniques that have been studied and utilized in the literature on CCs. Queuing modeling applies statistical tools to determine the optimum number of CSRs and servers for a customized CC. Queuing modeling has been extensively researched for call centers (Koole and Mandelbaum, 2002) despite the fact that little research has recently focused on CCs (Jolai et al., 2008; Balakrishnan and Munisamy, 2007; Luo and Zhang, 2013; Whitt, 2006b; Dudin et al., 2013; Greasley et al., 2013; Chromy et al., 2011; Dhesi et al., 2011; Mišuth et al., 2009, 2010). These studies are not included here as they do not use customers’ communications data and information during their contact. In the following subsection, data mining and text mining techniques are discussed.

3.1 Data mining

Data mining (or knowledge discovery) is an interdisciplinary technique combining aspects of artificial intelligence, machine learning, statistics, and database systems (Kantardzic, 2011). The main aim of the data mining technique is to extract useful knowledge from a large amount of raw data. This knowledge is represented in the following forms: classification, estimation prediction clustering affinity grouping (associative rules). One main branch of data mining is text mining which has been used widely in CCs and is discussed in the following section.

3.1.1 Text mining

CCs, as the touch point of customers, provide various communication channels for customers and receive both a high amount and different types of data. The majority of the received data are text data which are mostly produced through e-mails and social media communication channels. In some CCs, voice messages are saved as text which allows text mining techniques to be utilized. Text mining is an important type of data mining which, as the name suggests, has the main usage of mining and harvesting knowledge from the text data. In fact, text mining finds new patterns in a large volume of unstructured texts via natural language processing, and statistical and machine learning techniques. In other words, patterns are extracted from unstructured texts rather than rational databases. The main difference between text mining and information retrieval is that in the latter, new information and knowledge are not extracted.
3.2 Text mining usage in CCs

An examination of the current literature demonstrates that, compared with other ordinary data mining techniques, text mining is the most prevalent. The current literature uses text mining to assist analytical CRM, whereas the CC is the primary interface of operational CRM. Specifically, the operational aspect of the CC has not been thoroughly studied. As mentioned earlier, the focus of the current literature has been on analytical techniques for the following purposes: call monitoring and additional CC administration tasks. We briefly explain these tasks in the following section.

3.2.1 Call monitoring. Call monitoring is one of the main tasks of quality analysts (QAs) in CCs (Zweig et al., 2006). QAs monitor inbound and outbound calls to observe and measure the performance of CSRs to ensure that CSRs adhere to a given CC’s regulations, policies and strategic objectives. Recently, researchers have focused on developing automatic call monitoring to address four common problems with manual call monitoring: inconsistencies in the QA’s analysis, only analyzing a small part of the voice of the customer (VOC) data, the time consuming nature of call monitoring and failure to find the exact reason for customer dissatisfaction (Godbole and Roy, 2008). The four terms most prominently used in relation to automatic call monitoring are being automatic customer satisfaction (C-Sat) analysis, emotion detection (ED), problematic call detection, and call segmentation. Each is discussed as follows:

1. Automated C-Sat analysis: C-Sat analysis has traditionally been implemented via QAs to find the reasons for customer dissatisfaction. VOC is an essential source for the information analysis procedure, which is captured via various communication channels. The organization can understand the requirements of its customers by analyzing VOC. VOC helps organizations identify the gaps between expected service quality and current quality. C-Sat analysis is one important part of the VOC analysis procedure. Unstructured data that is produced through feedback surveys is another valuable source of information that can illustrate the degree of C-Sat. Researchers have mainly used text mining and machine learning techniques to build an automated C-Sat analysis system since the measurement of C-Sat via a manual survey is not very accurate and is quite costly. Godbole and Roy proposed an integrated approach to this end, utilizing text classification, business intelligence, and interactive text labeling in their proposed system. Prak and Gates (2009) proposed a model to calculate C-Sat during a customer’s call to a given organization. They utilized the following techniques in order to estimate C-Sat using a five-point classification scheme: decision tree, naive bayes, logistic regression, and support vector machines (SVMs). In their study, C-Sat is categorized into three groups: satisfied, neutral, and dissatisfied.

2. ED: ED in relation to customers has been studied widely by researchers in related academic communities (Devillers and Vidrascu, 2006). Systems have been developed to detect the emotion of customers based on what customers have said in their call-based communication (Devillers and Vidrascu, 2006) or e-mail-based communication (Gupta et al., 2012). ED of customers in their call-based communication is a useful technique that assists QAs in their task of call monitoring. They can take over a call when an ED system shows a given customer is unhappy to solve the customer’s issue(s). QAs can also analyze those calls during which customers display negative emotions to find the obstacles in the CC procedures (Ming and Yi, 2011). Also, using an ED system allows QAs to identify good CSRs and utilize their expertise to train weak CSRs (Pandharipande and Koppapuru, 2012). Having a suitable customer ED system increases customer loyalty as well as enhances the performance of the CC. It should be noted that ED should be executed before C-Sat analysis since, for example, a given
customer with the emotion of dissatisfaction may be satisfied at the end of the call. In fact, as ED enhances the performance of the CC, this has a direct impact on the degree of C-Sat. The emotion of the customer is categorized by various researches into a disparate number of categories: 3 (Pandharipande and Kopparapu, 2012), 4 (Devillers and Vidrascu, 2006), 5 (Ming and Yi, 2011), and 6 (Abdulateef et al., 2011).

(3) Problematic call detection: identifying problematic or abnormal calls is another research area that falls under the call monitoring aspects of CCs. Unlike automated C-Sat analysis and ED, which provide information about a given conversation, problematic call detection uses a binary classifier to determine whether the conversation between the customer and the CSR is normal or not. Pandharipande and Kopparapu (2012) developed a model that identifies abnormal calls by using the speaking rate of a given conversation and model it as a directed graph. Zweig et al. (2006) proposed a maximum entropy classifier to identify a bad call. Their maximum entropy classifier works based on set of ASR-derived features by testing 676 calls from IBM’s North American call centers.

(4) Call segmentation: researchers believe that calls in CCs generally follow the same pattern and that their segments are useful for QAs (Park, 2007). These calls have common components, such as greeting, question, refine, research, resolution, closing, and out-of-topic (Park, 2007). Call segmentation is useful and valuable to CCs in several ways. For example, Park mentions that segmentation may result in CSRs spending less time based on similar contact that it reduces the time of responding as well as better resolution. Also, it may help CCs to identify suitable calls for self-servicing. Calls with a short Question and Resolution section are good candidates for self-servicing and result in reducing CC costs. Park provided the CSRs and customers with a list of keywords and based on this, the recognition of speaker will be done in his approach. The quality of the call between the customer and the organization’s CSR is based on two factors (Park, 2007). Utilizing the advanced text analysis approach, Park classifies the call as either good or bad based on the customer’s sentiment and the CSR’s attitude. The output of his system helps organizations to identify customer calls for human monitoring. Padmanabhan and Kummamuru also mentioned working on sub-procedure text segments (SPTS) is also useful. They stated that “there are some typical exchanges of information which would be characteristic of the topic or the procedure employed in the call” (Padmanabhan and Kummamuru, 2007). Padmanabhan and Kummamuru provided five applications of SPTS: using labeled information, identifying inefficiencies, aiding faster problem resolution, agent prompting, and identifying agent experience/innovations. Padmanabhan and Kummamuru utilized text clustering, the frequent sequences detection algorithm and the AprioriAll algorithm in their proposed model. The transcripts of the customers’ calls are obtained by using an automatic speech recognition (ASR) system. They presented the customer calls as a sequence of cluster labels and finally, by utilizing the AprioriAll algorithm, found frequent and distinct patterns in a given call. They showed the extracted sequences are useful for assisting CSRs in their jobs.

3.2.2 Additional CC administration tasks. Several studies focus on other CC administration tasks, although the majority of research in relation to the analytical aspects of CCs falls under the call monitoring category. In this section, we review four of these applications and briefly explain them:

- Logging telephone calls: Byrd et al. (2008) proposed a model to log telephone calls (semi-automatic approach) to reduce CC costs by decreasing the time spent on call
logging in comparison to a manual approach. The audio capture of a telephone conversation, ASR, text analysis, and log generation is utilized in their solution to provide a candidate call log for CSRs. As the final decision is made by the CSR, the model is referred to as a semi-automated approach.

- **E-mail routing**: Nenkova and Bagga (2003) proposed a system for e-mail routing and classification. The main reason for proposing such a system is that, unlike call communication, there is a lack of e-mail-based communication system routing. For call-based communication, interactive voice response (IVR) and a specialized backend system are two tools for customer call phone handseling. Nenkova and Bagga's work tries to address this gap. Their proposed system classifies e-mails into three types: root (immediate customer query), inner (middle of communication), and leaf (end of communication). Rainbow and svmLight were the two tools used to perform the classification. Their proposed system is useful for managing customers' e-mails and directing them to the appropriate operator.

- **Application on car rental CC**: CCs generally analyze about 1-2 percent of their customer call transcripts. Thus, utilizing automatic or semi-automatic approaches can increase this percentage and would be beneficial to the center. Takeuchi et al. (2007) proposed a three-step model to analyze the conversations between customers and CSRs in a car rental CC. The analysis is based on the conversation transcript data and the results show that the productivity of the CC is improved. They first determined notable expressions in the call transcripts and then utilized text mining techniques to gain a better insight into their business processes. They used SVMs to classify customers as "picked-up" or "not picked-up." They also classified customers as either a booking customer or a rates customer. Their two-dimensional association analysis showed that "cars booked by rates customers tend to be 'not picked-up'." Thus, they concluded if CSRs act properly after recognizing the type of customer, such as by offering a discount, they can change the decision of the customer to pick up the car.

4. **Managerial aspects of CCs: state-of-the-art**

The managerial aspects of CCs are another important feature that is studied in the literature. We have focused on three managerial aspects of CCs in this study: the performance of CCs, CSRs, and outsourced CCSs. There are other topics that are not discussed in this paper although they are briefly discussed in the aforementioned topics: presence-based open CC (Acharya et al., 2013), e-CCs (Ming and Yi, 2011), inbound CCs (Helber and Henken, 2010; Abdullateef et al., 2011), virtual CCs (Brian, 2003), knowledge management (KM) technologies in CCs (Stieger and Aleksy, 2009a, b; Parikh and Walton, 2012; Hart et al., 2009), contact collector tools (Visweswariah et al., 2010), CC costs (Byrd et al., 2008; Balakrishnan and Munisamy, 2007; Chromy et al., 2012; Kim and Ha, 2012; Nambiar et al., 2011; Smith, 2009; Gilbert et al., 2005), ergonomic issues (Ramesh Babu et al., 2012; Rod and Ashill, 2013), and job quality (van Dun et al., 2012).

4.1 **CC performance**

Measuring the performance of complex systems is an imperative and important task in order to have better control, monitoring and management of their activities. Performance, as an important aspect of each organization, has also been taken into consideration by researchers in the CC area. An examination of the literature demonstrates that there are some good descriptive studies in this area which provide comprehensive results that highlight the important and influential factors in relation to CC performance.
However, there is a lack of practical and empirical studies which show the process of measuring a CC’s performance. A summary of the existing literature is discussed in the following:

- Impact of four CRM dimensions on the perceived service quality (PSQ) of the CC: Abdullateef and Salleh studied the impact of four important factors (dimensions of CRM) on PSQ in the contact center industry: customer orientation, CRM organization, KM, and technology-based CRM (TCRM) (Abdullateef and Salleh, 2013). The degree to which various products and services meet customers’ expectations is PSQ. Data from questionnaires completed by 173 managers of Malaysian CCs are collected to test the study’s hypothesis. The authors concluded that KM and TCRM are positively correlated with PCQ.

- Better planning for contact centers: According to Koole (2004), personnel costs account for the highest proportion of CC costs, at approximately 60-70 percent, and they considered waiting time and the number of abandoned calls as the service level (SL) metrics (Koole, 2004). He explained the weaknesses of the current methods and discussed a solution to ensure better planning. In his view, one-way to increase the performance of CCs is to increase the flexibility in CSR’s contracts and task assignments and to employ cross-trained CSRs.

- Employee turnover rate impact on CC performance: A low turnover rate is positively correlated with CC performance. This is because when CSRs stay in a CC for a long period of time, the CC will benefit from more experienced staff which directly influences the center’s performance. Whitt et al. proposed a mathematical model that describes the transition costs of staff turnover and the performance benefits of employee retention. They used the following variables that are traditionally used to measure a CC’s performance: the number of calls answered per hour, the revenue earned per hour, and the number of problems successfully resolved per hour (Whitt, 2006a).

- Outsourced CC performance: Tate and Valk studied the impact of external supplier service performance on a buying firm’s output. They stated that both the efficiency and effectiveness of an outsourced CC should be considered to ensure suitable performance. They felt that if only efficiency is taken into consideration, this will have a negative effect on performance and quality in both process and outcome. Tate and Valk stated that low performance on the side of the service supplier has a direct (low end C-Sat) and indirect (reduced access to customer intelligence) impact on the buying company’s performance (Duggirala et al., 2011). They identified the following 12 features from the literature by which to measure an offshoring CC’s performance: “average speed of answer, queue time, first call resolution, abandonment rate, average talk time, total calls, callers who receive a busy signal, time before abandoning a call, inbound calls per customer service agent per 8-h shift, agent turnover, SLs, adherence to schedule” (Tate and van der Valk, 2008).

- The impact of 11 dimensions of service quality on CC performance: Duggirala et al. studied the impact of 11 dimensions of service quality on CC outcomes and performance. They utilized traditional and stepwise regression to find the most effective variables, these being top management commitment and leadership, human resource management, process management, continuous improvement, benchmarking, employee focus, customer focus, facilities, service culture, error management, and measuring outcomes. They showed that benchmarking and error management are the two most effective variables of the 11 dimensions (Duggirala et al., 2011). “Customers of an enterprise
contact an agent in a call center via email or chat or telephone call. Each such interaction results in a ‘ticket’ being generated. Agents determine the problem being faced by the customer and try to resolve it based on their knowledge or by searching knowledge bases. A few senior agents periodically enter information about frequent problems faced by customers and their resolutions into these knowledge bases. QAs analyze agent-customer conversations to check if the prescribed call handling procedures are being followed and to verify that calls are being handled efficiently. A sample of the customers is contacted to fill in customer survey forms to indicate their satisfaction with the proposed resolution and with the calls. These survey forms result in C-Sat scores: the (estimated) percentage of calls that resulted in satisfactory resolution of problems as rated by the customers. The CSAT scores are a key metric for measuring the service quality of CCs” (Duggirala et al., 2011).

4.2 Customer service representative
A CSR is an employee of a CC, also known as an agent. The role of CSRs is to maintain a suitable relationship between the center and the customers. CSRs are largely responsible for the customer service experience, through responding to customers’ enquiries, empathizing with their needs, solving their problems, and handling their complaints (Owens, 2014):

CSRs provide an intermediary role in the contact center between the customer and the system. They must not only maintain a coherent conversation with the person on the other end of the phone line but must also cope with a number of interfaces and interaction styles contained within the (often) multiple databases that they need to get information from and to (Millard et al., 2004).

CSRs’ skills, training and performance have been studied in the CC literature. The CSRs’ skills are highly important to the success of the CC and have a positive correlation with the CC’s performance. Millard et al. stated that CSRs must have two different types of skills and knowledge when they respond to enquiries on the phone or offline. First, they must be able to access knowledge efficiently and quickly; and second, they must be able to engage in knowledge sharing, personalization, and exploration (Millard et al., 2004). In turn, CC managers must consider the quality of the CSR’s performance to ensure employee satisfaction and avoid CSR turnover (van Dun et al., 2012).

One of the issues in CCs is the high rate of employee attrition (Owens, 2014), referred to as CSR turnover (van der Aa et al., 2012; Whitt, 2006a). The cost of employee turnover is classified as transition and productivity costs in Whitt’s (2006a) study. “Transition costs account for the per-agent costs of terminating the departing agent; recruiting and training a new agent; disruption costs associated with the change, such as the cost of hiring a temporary employee; and the cost of managers coping with the change, such as the cost of performing exit interviews, the administrative costs of stopping benefit deductions and performing benefit enrollments, and so forth” (Whitt, 2006a).

CSR training, CSR performance, CSR cross-skill, and helping CSRs to provide semi-automatic CC operations are other aspects of CSRs that are beyond the scope of the current paper.

4.3 Outsourcing CC
Procurement, implementation, operations and maintenance, technical support staff, and operations staff are the five main costs of CCs (Mišuth et al., 2010). Labor costs contribute to about 60-80 percent of CC costs (Armony and Maglaras, 2004). These significant costs motivate organizations to outsource their CCs. As a result, since the 1980s, many industries
and organizations have made the decision to move their staff to countries with lower wages and costs, which was the beginning of offshoring in the business world (Meuter et al., 2000):

Outsourcing is an accepted solution for budget-constrained operations, and today, most governments and many private-sector call centers are outsourced (Bernett, 2005).

Outsourcing is an agreement between a business and a third party for the management and operations of a corporate activity, such as human resources, marketing, or customer service. The third-party company assumes part or all of the business function (Bernett, 2005).

The rate of employee attrition is high in CCs, especially in developed countries (Owens, 2014; van der Aa et al., 2012). In addition, hiring suitable CSRs is not an easy task in developed countries (Owens, 2014). These two problems result in organizations looking toward outsourcing. Two types of CC outsourcing are described in the following section.

4.3.1 Two types of outsourcing: offshoring and hosting. Two types of outsourcing are proposed in the literature: offshoring and hosting. In offshoring, the CC is outsourced to companies in foreign countries, such as India, China, Mexico, and the Philippines (Poster, 2013; Tate and Ellram, 2009; Tate et al., 2009; Derakhshani and Hart, 2010a, b). Another type of outsourcing is hosting (Bernett, 2005), where a third party provides technological infrastructure and software applications and the CC is managed via the organization’s staff.

Outsourcing is associated with several problems such as “cultural and language differences, political and legal uncertainties and geographical and time zone issues” (Owens, 2014). Studies have shown that “the most common reason for negative customer assessment of overseas CCs is ‘difficulty in understanding accents’ (Cards International, 2007) and this can and has led to customer defection on the basis of CC interaction alone” (Owens, 2014). Thus, offshoring CCs focus their staff training on “accent training” or “neutralisation” (Meuter et al., 2000).

5. Shortcomings and future research in CCs: critical evaluation

Most of the literature on CCs is focused on analytical and managerial aspects, similar to the trend in the CRM literature (Awasthi and Sangle, 2012). In the analytical studies, text mining is predominantly utilized to monitor calls by calculating the customers’ satisfaction during the calls, detecting the customers’ intent and emotion, segmenting calls and identifying abnormal calls. In studies that focus on the managerial aspects of CCs, researchers have identified the factors that influence a CC’s performance. However, there is a lack of practical and empirical studies that describe the process of measuring a CC’s performance. In this paper, we focus on the two weaknesses in the current literature from analytical and operational perspectives. First, we present an example to help the reader follow these two shortcomings:

Example: let us consider the example of a customer of an organization, John Smith, in four different scenarios: first, John has sent a complaint to a given CC by e-mail and follows up on it two days later with a telephone call. The CSR, before responding to John, should be able to identify and determine the nature of the complaint, and any subsequent follow-up by the organization rather than asking John to repeat his complaint and placing him on hold. Current techniques in CCs can help to trace the feedback of the customer only if it was communicated to a person and was entered and linked manually to the appropriate record in the relational database system. Second, John dislikes a particular type of product and this dislike was made known to the organization through a telephone call. The CC’s organization should ensure that John is not sent any information on this type of product in their marketing promotions. But there is another customer named John Smith who would like to receive information about this type of product. Third, the CC’s organization wants to
perform a predictive business analysis on a group of valued customers like John Smith to
determine the chance of them leaving the organization’s services and to develop new
products depending upon the feedback received from that group. Fourth, John is dissatisfied
with the complaint handling process of the CC and this may affect the CC organization’s
chances of having John accept new products, leading to a common mode type of failure in
their relationship. How does the organization utilize this type of information for customer
management and marketing?

These examples raise two main issues which have not been studied in the literature:
the reactive behavior of CCs with their customers and the poor management of unstructured
data in the current state of CCs. We elaborate on these issues in the following sections.

5.1 Lack of an interactive CC: an analytical perspective
Despite the availability of a vast array of communication channels, such as face-to-face,
telephone, e-mail, instant chat, social networks, and fax, telephone calls remain the primary
form of communication. Many customers who make phone calls tend to be impatient and
require a quick response from CCs compared to customers who use e-mail messages,
social networks, or instant chats (Aghhari and Balcioğlu, 2009). This demonstrates the
importance of establishing an appropriate interactive system in response to customer
enquiries. Customers prefer to interact with humans rather than machines in this interactive
system. However, CCs prefer to use technology instead of humans, as humans account
for approximately 60-80 percent of their costs. Solutions such as self-help, IVR and
outsourcing have their own drawbacks and establishing an interactive CC using advanced
technology and analytical mining is imperative for organizations to enhance their
customers’ experience.

As previously explained, the current CCs are far from being interactive. For example,
it can be seen from the aforementioned example that information about the current state of
John’s relationship with the organization (such as complaints, preferences, interaction
experiences, and follow-ups) is mostly obtained from sources such as records, e-mails, and
transcribed calls, which are semi-structured in format. Such information presents an
incomplete picture of John Smith and should be linked dynamically to the corresponding
fields in the structured databases in order to obtain more comprehensive information that
will assist in answering queries appropriately and consequently, build interactive CCs.
Unfortunately, existing CCs consider these two sources of information separately and
essentially only analyze the structured data (Maimon and Rokach, 2005) for which many
techniques exist (Agrawal et al., 1996; Han and Fu, 1995). But the results obtained from such
analysis may not be complete and do not capture the current state of the customer. Part A of
the example demonstrates how two different communication channels threaten the
effectiveness of CCs due to poor interactive capabilities.

There are several reasons for the paucity of interactive CCs in the literature, the main
being the massive amount of unstructured data as well as the limited use of data mining
tools in the analytical studies. Some studies used text mining tools in an interactive manner,
such as C-Sat analysis, but the authors believe that ordinary data mining tools as well as
text mining techniques can be used to help CCs become more interactive. Let us look at the
example from this perspective: have appropriate techniques been employed in current CCs
to deal with the issue of the recognition of customers with the same name? A survey of the
existing CC literature highlights that the issue of customer ID matching from the viewpoint
of the operational process has not been fully studied. The current literature on CCs and
CRMs is focused on the behavior of customer recognition and the primary task of customer
ID matching has been ignored (Shy and Stenbacka, 2013; Fudenberg and Villas-Boas, 2006).
Although identifying customers using their primary ID is an easy task, the problems
created by dirty data or when the customer forgets their ID are not straightforward to solve.
Much future research should be undertaken to facilitate the development of interactive CCs. It is not possible to investigate all these areas deeply in this paper, hence, the areas that will be covered are as follows: online customer data mining, identifying the right offer for customers, customer query optimization, efficient customer inquiry tracking, etc.

5.2 Lack of data integrity for CCs
Two sections are provided in the appendices which elaborate the importance of dealing with unstructured data and the weaknesses of current CRM systems in handling them. In fact, there is the lack of techniques to handle unstructured data and data integration in the current CC literature. Currently, 80 percent of all data are unstructured, showing the importance of developing tools that fully comprehend unstructured data. Researchers believe related scientific communities are at least several decades away from this goal (Parameswaran, 2013). Awasthi and Sangle (2012) referred to the lack of studies which discuss the technical and nontechnical aspects of multichannel CRM which have the support of empirical studies, in their literature review on multichannel CRM (Awasthi and Sangle, 2012). It is clear that CCs, as an important part of CRM that is in direct contact with multichannel communication tools, should be studied in this regard. There is a paucity of studies that provide empirical or theoretical results. The current literature provides useful methods to mine unstructured data on some occasions but the issues of heterogeneity and dirty data are not addressed. We list two main obstacles and research gaps which have prevented CCs from having an integrated database, based on the given example:

(1) Difficulty in having conjoint data: having conjoint data is complicated when associations have to be formed between information that may be spread across different heterogeneous sources, each having its own format and its own level of nesting. In such cases, how is the relevant data across the different heterogeneous sources linked with appropriate records? Existing approaches do not propose efficient and effective techniques for identifying and linking the relevant data with the relevant fields, possibly resulting in mismatched information.

(2) Complex structural relationships: semi-structured information may be in different levels of nested or embedded elements and can have two-dimensional relationships among data entities manifested through structural relationships among attribute nodes. XML schema is used to understand the information and represent it as a hierarchical structure for a more semantic representation (Feng et al., 2002). But such schema might have different nested sub-trees and challenges arise when patterns of an occurrence of a scenario or association rules within these levels of nested hierarchies of data have to be determined. Due to complex structural relationships, mining semi-structured information poses additional challenges and is quite different from the well-researched field of structured data mining. Efficient techniques are needed for analyzing the XML document and its different nested hierarchies before determining the occurrence of patterns in them.

Thus, CCs are in need of novel techniques for data standardization and data linkage so that conjoint mining can occur. Also, it is essential to develop a visualization framework to identify the relevant data for a customer from different business processes, conceptually understand it and provide adaptive business intelligence recommendations to improve C-Sat, management, and retention.

5.3 Future research direction: harnessing big data power
In addition to the elaborated two research gaps which should be addressed by researchers in the future, the utilization of big data analytics in CCs is another important research direction for the future. This section emphasizes the importance of using big data techniques in CCs.
Big data is an evolving term which refers to the rapid increase in the velocity, volume, and variety of data. In recent years, sophisticated algorithms, techniques and technologies have been proposed in this area in the literature to enhance data analytics capabilities (Zikopoulos and Eaton, 2011). In general, the proposed algorithms and techniques complemented rather than replaced the ordinal data mining techniques. We have demonstrated in the above sections that there is a paucity of usage of big data techniques in CCs. Research areas such as in-memory analytics (Acker et al., 2011), big data visualization (Keim et al., 2013), cloud-based BI (Khan et al., 2015), data governance (Khatri and Brown, 2010), big data profiling (Papenbrock et al., 2015; Abedjan et al., 2015), etc. can make excellent contributions to the improvement of a CC’s performance. Zikopoulos et al. (2012) stated that employing big data capabilities in CCs can provide the following benefits to customers: it can help CSRs to have a better quality answer to a query, the call can be escalated to the next level, it can offer specific incentives to the customer, and it can motivate CSRs to reply more politely.

As mentioned in Section 3.2.1, call monitoring is one of the important tasks undertaken in CCs and the quality of call monitoring can be enhanced more rapidly by using big data techniques. Although some papers in the literature focused on capturing VOC and implemented call monitoring systems on top of it, these systems remain in early stages of development (Lohr, 2012). We believe that in particular, the quality of customer sentiment analysis can be considerably improved by incorporating big data capabilities (Zikopoulos et al., 2012). This improvement would occur by developing the following systems: customer exasperation identifier system, sentiment analysis system by establishing a real-time correlation between the CC trends and the rest of the business operations, etc. We elaborate the benefits of integrating offline techniques and stream data analysis techniques to CCs in the following section.

One important way big data can be used in CCs is to integrate the capability of offline techniques and stream data analysis techniques to provide better sentiment analysis. Harnessing the power of social networks’ streaming data, such as a live Twitter stream, in CCs by transparently integrating and correlating it with CC trends is the next step in enhancing the quality of service of CCs (Hao et al., 2011). Leveraging stream analysis allows CCs to monitor their procedures to determine whether it adheres to the guidelines. Moreover, this integration can provide a more complete picture of the customers by accessing the complete VOC and consequently assisting CCs in their duties and procedures. One of main issues in analyzing VOC is that most CCs listen to what customers are saying, but they do not have the ability to analyze the information and respond accordingly. In fact, there is a considerable difference between the captured VOC and why the person said it or thought it. We borrow an example from the internet service provider (ISP) industry in Zikopoulos et al.’s (2012) book to explain how big data can be employed to better analyze VOC in CCs.

Example: assume we have a CC for an ISP. One of their customers, Andrew J., is not happy about the speed of the provided service. Andrew J. consequently makes a complaint to one of the CRSs in the CC; however, he felt that the CSR was only doing what was required of her and that she did not feel Andrew’s complaint was important. Andrew felt that the CSR wanted to hang up the phone quickly, leaving Andrew wondering if his issue of the slow speed of the service had really been captured. So how can big data assist the CC to tackle this issue? The CC can develop an online monitoring system to improve its customer service. This system should check the quality of the network and alert CCs whenever there is a quality issue in a certain part of its coverage (Zikopoulos et al., 2012).

6. Implications and conclusion
CCs play a pivotal role in organizations and form an important part of CRM operations. In the literature, the importance of CCs has been studied in different domains, but with the
changes in technology, their operation and structure has to evolve to meet the ongoing challenges. We presented a review of the state-of-the-art literature and the challenges in identifying the gaps in order to build intelligent CCs. We classified the current literature on CCs into two categories: analytical and managerial studies. From the discussion in Section 5, it is clear that current CCs suffer from two main issues: lack of interactive CCs and high amounts of unstructured data. Given the production of massive amounts of data in CCs, especially unstructured data, we discussed the benefit of using big data in CCs and present it as a new research agenda for CCs. Based on the provided critical literature review on CCs, CCs have the potential to receive more attention in the next decade from data scientists.

We would like to emphasize the importance of publishing big data studies with use cases of CCs. It is likely that some practitioners have applied big data techniques in CCs but they have possibly not published their findings. The publication of case study papers that elaborate the benefits of big data usage in CCs provides an opportunity for mutual collaboration between scholars and practitioners. In this highly competitive world, empowering organizations with the next generation of CCs is the competitive edge that can be gained with the mutual cooperation of both academics and practitioners.

References


Digital Nation Expanding Internet Usage (2011), National Telecommunications and Information Administration, US Department of Commerce, Armonk, NY.


Ponemon, L. (2008), Survey on the Governance of Unstructured Data, Ponemon Institute LLC.


Rodenstein, R. (2004), “How online e-mail deflection can benefit your bottom line”, E-Content, pp. 4-5.


**Further reading**

Appendix 1. The importance of unstructured data

This section highlights the quotes of several famous business analysts and the results of surveys to demonstrate the importance of unstructured data in the business context:

- William H. Inmon: “A wealth of invaluable information exists in unstructured textual form, but organizations have found it difficult or impossible to access and utilize it” (Inmon and Nesavich, 2007).

- Dr Feldman Ronen: “Unstructured data provides a situational context around an event or set of events that answers the questions ‘why’ and ‘how,’ essentially filling in the ‘cause, complain, correction’ knowledge cycle. Knowing the ‘why’ and ‘how’ empowers organizations to uncover hidden relationships, evaluate events, discover unforeseen patterns and facilitate problem identification for rapid resolution. Utilizing intelligence extracted from unstructured data enables organizations to avoid loss of profit margins due to preventable write-offs, customer churn, legal settlements, warranty claims, or inefficient product development cycles” (Feldman, 2005).

- J. Kuo: “Having access to both unstructured and structured information from the same application means businesses can finally get a complete view of their organizations” (Kuo, 2007).

- Nancy Scott: in writing for TechRepublic, Mary Shacklett describes a scenario that encapsulates the value of unstructured data. "For years, non-profit aid organizations have been sending in field workers to advise local farmers on best agricultural practices. These workers file progress reports and keep tabs on agricultural projects to see if crop yields improve [...] The difficulty has been in collecting all of these reports, which come in many different forms – and then trying to glean insights into them after they become a monolithic body of unstructured and semi-structured data. By using big data collection, grooming and analytics techniques, humanitarian aid organizations are now able to compile all of these unstructured reports of field farming activity into databases – and then to mine these databases for information about which farming projects are succeeding, which are not, and why” (Scott, 2013).

- According to Ventana’s survey, 87 percent of respondents indicated that unstructured and semi-structured data types are very important (Research, 2007).

- Ponemon Institute LLC: “Unstructured data is at risk in most organizations” (Ponemon, 2008).

- Ponemon Institute LLC: “Data access privileges are too permissive” (Ponemon, 2008).

- To protect unstructured data, IT professionals need automated solutions.

- Adrian Bridgwater: “The digital universe of western Europe will double every two and a half years; ‘growing 3 times faster than structured data’.”

- Coveo survey: “According to a recent Coveo survey of 100 customer service executives, 87% believe it is ‘very important’ to share near real-time integrated data – if they had it – between product management, customer service and sales in their organizations” (www.coveo.com/en/news-releases/Coveo-survey-shows-organizations-falling-short-in-generating-insight-from-unstructured-content).

- The internet presents new challenges for contact centers in maintaining the quality of their data, as more data comes from e-customers than the centers. According to a report from the US Department of Commerce, released in 2011, it is estimated that there were $10 trillion online transactions annually. This has dramatically increased from 2000, when more than 58 million US consumers engaged in online transactions. The growth of e-commerce makes it crucial that there is an efficient system in place to handle unstructured data (Digital Nation Expanding Internet Usage, 2011).

- “In a survey of more than 100 executives conducted on December 15 at the Argyle Executive 2011 Customer Care Leadership Forum, 84% said their companies' management of unstructured content will determine how effectively and efficiently they’ll be able to serve customers” (www.coveo.com/en/news-releases/Coveo-survey-shows-organizations-falling-short-in-generating-insight-from-unstructured-content).
Appendix 2. Weaknesses of CRM systems in handling unstructured data

Unstructured data handling is difficult due to the lack of data repeatability and predictability. Also, unstructured data are only one dimension of big data, and big data constantly changes (hence, it is difficult to keep up with) and makes data dirty which is hard to handle:

- “According to a recent Coveo survey of 100 customer service executives (CSEs), 66% of those interviewed said their organizations either could not bring together customer, product and project data from all sources and share it or are trying to share but are facing challenges” (Derakhshani and Hart, 2010a).

- Organizations are in need of automated or semi-automated ways to handle unstructured data. “Today, 90% of unstructured data workload is handled manually.” One of the challenges facing current systems is to make their information more accessible to their customers and employees.

- “79% of interviewed CSEs said that they can only sometimes or almost never get the information they need about their businesses and their customers to make informed business decisions quickly” (www.coveo.com/en/news-releases/Coveo-survey-shows organizations-falling-short-in-generating-insight-from-unstructured-content).


- Only 22 percent of executives said that employees can easily access the information they need to do their jobs. The remaining participants reported that their employees likely do not (29 percent) or “maybe” (41 percent) have access to the information they need. An alarming 8 percent said they did not believe employees could access the information.

- “Data Management was traditionally about managing structured data. This focus needs to change.” (McIsaac, 2007).

- “While the digital revolution has made creating information easier, our ability to find this information when we need it decreases exponentially over time” (Blumberg and Atre, 2003).

Corresponding author
Morteza Saberi can be contacted at: m.saberi.ie@gmail.com

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Let’s stop trying to be “sexy” – preparing managers for the (big) data-driven business era

Kevin Daniel André Carillo
Department of Information Management, Toulouse Business School, University of Toulouse, Toulouse, France

Abstract

Purpose – The purpose of this paper is to analyze the inadequacies of current business education in the tackling of the educational challenges inherent to the advent of a data-driven business world. It presents an analysis of the implications of digitization and more specifically big data analytics (BDA) and data science (DS) on organizations with a special emphasis on decision-making processes and the function of managers. It argues that business schools and other educational institutions have well responded to the need to train future data scientists but have rather disregarded the question of effectively preparing future managers for the new data-driven business era.

Design/methodology/approach – The approach involves analysis and review of the literature.

Findings – The development of analytics skills shall not pertain to data scientists only, it must rather become an organizational cultural component shared among all employees and more specifically among decision makers: managers. In the data-driven business era, managers turn into manager-scientists who shall possess skills at the crossroad of data management, analytical/modeling techniques and tools, and business. However, the multidisciplinary nature of big data analytics and data science (BDADS) seems to collide with the dominant “functional silo design” that characterizes business schools. The scope and breadth of the radical digitally enabled change, the author are facing, may necessitate a global questioning about the nature and structure of business education.

Research limitations/implications – For the sake of transparency and clarity, academia and the industry must join forces to standardize the meaning of the terms surrounding big data. BDA/DS training programs, courses, and curricula shall be organized in such a way that students shall interact with an array of specialists providing them a broad enough picture of the big data landscape. The multidisciplinary nature of analytics and DS necessitates to revisit pedagogical models by developing experiential learning and implementing a spiral-shaped pedagogical approach. The attention of scholars is needed as there exists an array of unexplored research territories. This investigation will help bridge the gap between education and the industry.

Practical implications – The findings will help practitioners understand the educational challenges triggered by the advent of the data-driven business era. The implications will also help develop effective trainings and pedagogical strategies that are better suited to prepare future professionals for the new data-driven business world.

Originality/value – By demonstrating how the advent of a data-driven business era is impacting the function and role of managers, the paper initiates a debate revolving around the question about how business schools and higher education shall evolve to better tackle the educational challenges associated with BDADS training. Elements of response and recommendations are then provided.

Keywords Big data, Business schools, Data science, Analytics, Curricula, Data-driven business

Paper type Research paper

Introduction

The amount of turmoil caused by digital disruption is unprecedented, destroying most companies’ long-successful business models and reshuffing the competition cards for the vast majority of sectors and industries (Weill and Woerner, 2015). After a few years since the “big data” alarm bell started heralding big data to be the “next frontier for innovation, competition, and productivity” (Manyika et al., 2011), there is now no doubt that digitization and more specifically big data analytics (BDA) and data science (DS) will deeply impact industries, institutions, and most jobs (Loebbecke and Picot, 2015). Facing the urgent need to develop an analytics capability allowing to analyze big data and create value through the generation of actionable insights, a new breed emerged within firms, data scientists,
in charge of capitalizing on the analytical opportunities associated with the big data phenomenon (Lee et al., 2014). Depicted as the “sexiest” job of the twenty-first century, a data scientist has been often described as “a high-ranking professional with the training and curiosity to make discoveries in the world of big data” (Davenport and Patil, 2012, p. 72). Companies have recognized the need to “grow, nurture and retain data scientists” in order to be able to proactively grasp the opportunities enabled by big data and analytics (Fosso Wamba et al., 2015). As of 2016, the data scientist “frenzy” has not faded as organizations are still “recruiting like crazy” (Ross et al., 2013, p. 90) such high-level profiles in a variety of industries and sectors such as information technology, marketing/finance services, consulting, healthcare, retail, or governments (Press, 2016; Burtch, 2016). Also, educational institutions have been no exception. They have quickly responded to the exponentially increasing demand for data scientists in the job market, a fact clearly demonstrated by the blossoming of hundreds of freshly minted Master’s programs, certifications, Massive Open Online Courses, or even bootcamps (Waller and Fawcett, 2013a). However, the rigid discipline-based functioning of schools and universities is seriously challenged as the data scientist skills lie somewhere at the crossroad of a multitude of areas such as mathematics, machine learning, artificial intelligence, statistics, database management, and optimization, along with solid foundations in business and management (Chatfield et al., 2014).

Furthermore, the disruptive technologies associated with digitization and advanced BDA have triggered the transformation of our entire society (Loebbecke and Picot, 2015), engendering the advent of a new data-driven business era (Carillo, 2015). This paradigm shift from a business- to a data-driven perspective (Lavalle et al., 2011) strongly impacts most companies’ strategies, business models, and processes (Chen et al., 2015), while it creates new ways of working, communicating, and interacting (Loebbecke and Picot, 2015). Consequently, analytics competencies shall not pertain to data scientists only, they must rather become part of an organizational cultural component shared among all employees and more specifically among decision makers: managers (Provost and Fawcett, 2013; Davenport, 2013). The worldwide lack of managers with such skills has been quite salient in expert’s reports (e.g. Deloitte, 2016; Pettey, 2012). Business schools have clearly positioned themselves as privileged actors for equipping professionals with the necessary analytics skillset (Chiang et al., 2012; Gupta et al., 2015). They have also joined forces with other neighboring university departments to take on the mission to train the people who are considered by some as the most important people of the twenty-first century: data scientists. However, the multidisciplinary nature of big data analytics and data science (BDADS) seems to collide with the dominant “functional silo design” that characterizes business schools which have suffered for years from being criticized for their lack of multidisciplinary integration and experiential components (Navarro, 2008). This claim first arose when practitioners and academics realized the inadequacies of business education when training cross-functional domains such as business process management or information technology (Ahmad et al., 2007; Seethamraju, 2012). Ever since, efforts have been deployed to better adapt curricula to the business reality; a concern materialized in the evolution of the recommendations provided by accreditation bodies such as the Association to Advance Collegiate Schools of Business (AACSB).

Nevertheless, it seems that the scope and breadth of the radical change we are facing, necessitate a global questioning about the nature and structure of business education. Business schools answered to the need to reinforce cross-functional skills to better address business process management training by incorporating in their curricula pedagogical strategies such as integrated case studies or capstone projects (Seethamraju, 2012). The same reductive approach cannot suffice when it comes to address an entire business paradigm shift impacting the very functioning of organizations. Analytics capabilities are
gradually pervading every aspect of organizations (Fosso Wamba et al., 2017), basically changing the way “we do business.” Engraining an analytical DNA not only within the brain of managers but also data scientists and other analytics experts cannot be limited to adding a few analytics-based tools within a professional’s toolkit. Does hiring a team of data scientists for the marketing department of a company makes it a data-driven marketing department? Is not it rather that the very nature and function of managers is changing in the data-driven business era? This reflection also shakes the disciplinary walls that have for long characterized business schools and universities to some extent. Since analytics, DS but also machine learning and artificial intelligence, are gradually permeating every business function and job, employees’ skill domains must become more and more impregnated with notions from domains such as mathematics or data management. The legitimacy of the disciplinary pedagogical approach is severely at stake. It is perhaps time to rethink about the pertinence and boundaries of our knowledge domains. As the “datification” phenomenon is irremediably changing our entire society, it becomes urgent to initiate a thorough reflection about the role, nature, and function of higher education and more specifically business education.

This paper does not have the pretense to provide, through a rigorous scientific investigation, the keys to all the enigmas mentioned above. Instead, it is an opinion paper that aims at shedding some light on the implications of the advent of the data-driven business era on higher education and more specifically business education. It is an attempt to provide a stepping stone in the initiation of a global and collective reflection gathering practitioners, academics, and educational institutions. With that purpose in mind, the paper strives to answer the following research questions:

- **RQ1.** What are the implications of the advent of the data-driven business era on business education and the overall educational system?
- **RQ2.** How shall business schools prepare managers to tackle the challenges of the coming data-driven business era?

This paper first analyses how digitalization and more precisely how BDADS together impact the nature and functioning of organizations, with a particular focus on decision-making processes and the function of managers. It then presents how educational institutions have responded to the lack of analytical talent. The paper then argues that managers must develop an analytics mindset turning them into data-driven manager-scientists. Finally, we conclude by proposing a set of implications to research and practice with the purpose of initiating and nurturing a debate about how business schools and higher education shall evolve to better tackle the educational challenges of the data-driven business era.

**Literature review: BDA – definitions and impact on organizations**

Over the last five years, the term “big data” has gathered tremendous attention in the media as well as in both the academic and practitioners’ spheres. Despite the fact that professionals and researchers have rapidly apprehended the major digitally enabled societal transformation that was looming, the term “big data” has always remained highly ubiquitous without any shared understanding or definition (Ward and Barker, 2013).

**BDADS: definitions and viewpoints**

The most common approach has been to define “big data” through characteristics that pertain to data, also called Vs. While the initial definition involved three Vs that are volume, variety, and velocity (Russom, 2011; McAfee and Brynjolfsson, 2012); the most recent version includes four additional Vs: validity, veracity, value, and visibility (Livingstone, 2013; Chartier, 2016). There also exist a number of definition variants
(often with diverging viewpoints and understanding) focusing on the technological aspects that surround big data as illustrated in IDC’s definition:

Technologies as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis (Gantz and Reinsel, 2012).

Others, such as Fosso Wamba et al. (2015) have striven to provide more encompassing and comprehensive definitions bridging the technological layers of the big data phenomenon:

A holistic approach to manage, process and analyze 5 Vs (i.e. volume, variety, velocity, veracity and value) in order to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantages (p. 235).

Shortly after the first uses of the term “big data” in the press, expert reports, and research outlets, the term “analytics” started being often attached, probably due to the growing awareness that the business value derived from the treatment of big data was a more crucial challenge (Gantz and Reinsel, 2012). The term “analytics” thus has been commonly used to address the analysis techniques and tools allowing to transform big data into actionable insights (Lavalle et al., 2011; Porter and Heppelmann, 2015). However, the differences between the terms “big data,” “analytics,” and “BDA” have remained fuzzy, while the distinction with business intelligence (BI) is rather unclear (Gupta et al., 2015). Some scholars have preferred to use “BDA” in a broad and encompassing way such as Loebbecke and Picot (2015, p. 150): “a means to analyze and interpret any kind of digital information”; while others speak of business analytics (BA) as the evolution of BI in the “big data era” (Chiang et al., 2012; Holsapple et al., 2014). The “BA” definition provided by the Institute for Operations Research and Management Science is among the most acknowledged and cited ones:

Analytics facilitates realization of business objectives through reporting of data to analyze trends, creating predictive models for forecasting, and optimizing business processes for enhanced performance[1].

Confusion and ambiguity have kept propagating with the growing attention surrounding “DS.” While some reduced it to the use of data-mining algorithms or statistics (Provost and Fawcett, 2013; Dhar, 2012), others have defined it as a domain including analytics (Waller and Fawcett, 2013b). Big data/BA has also been defined as DS applied to business (Chiang et al., 2012) while DS definitions tend to converge toward the one provided by Waller and Fawcett (2013b): “the use of quantitative and qualitative methods and strategies on data to solve relevant problems to predict outcomes.” Debating on the knowledge domains covered by notions, such as “big data,” “big data/BA,” “DS” and other related terms that have been used repeatedly, is beyond the scope of this paper. Such task may not even converge toward a unique taxonomy as it could go on and on as academia, the business world, and the media have continuously fueled the discussion with diverging views and understandings. Unfortunately, the lack of consensus about the scope and breadth of the reality hiding behind the big data bubble, along its overall impact on our society, has created a chasm between the needs on the ground and the responses provided by educational institutions. Each view and opinion corresponds to different knowledge domains and distinct associated skillsets. We argue and call for cutting short the terminology debate. At the intersection of all viewpoints lies a common awareness: the advent of a data-driven business era. Organizations must develop the capacity to derive value-creating actionable insights from the maelstrom of data that surround them (Sivarajah et al., 2017; Fosso Wamba et al., 2015). With that focus in mind, matters of volume, speed, variety, and other such characteristics become relative and somehow secondary (Ross et al., 2013). Behind the ill-chosen term “big data” lies a societal change
triggered not by the nature of the data that is produced and processed, but rather by the role it now plays and the strategic importance it has gained. Data now drives business and has become the new digital oil that runs through the veins of organizations …

**Impact of BDADS on organizations**

Digitization and more precisely BDA have a disruptive effect on business strategies and business models for nearly all industries and sectors (Weill and Woerner, 2015). There exists a substantial amount of anecdotal evidence about how insights drawn from big data have deeply impacted and improved core business functions (such as marketing, HR, operations, etc.) through the transformation of companies’ strategy and business model (Chen et al., 2015). Organizations that engage the data-transformation path were found to shift toward a more service-centric strategy (Zolnowski et al., 2016). Furthermore, the introduction of smart connected products combined with the implementation of BDA represent new competitive opportunities and threats for organizations by shifting industry boundaries or even by creating new ones (Porter and Heppelmann, 2015). The case of Tesla is particularly illustrative. The company has managed to disrupt the well-established automotive industry by eliminating the involvement of third-parties for repairs. The analysis of the data collected from their cars allows to identify the cars that are due for repairs. Customers are then automatically notified and have the possibility to request a valet to deliver the car to a Tesla facility. This strategic shift has helped Tesla to continuously improve customer experience (the firm regularly transmits software upgrades to its cars) and has placed the company among to most highly ranked companies in terms of customer satisfaction (Porter and Heppelmann, 2015). However, transforming an organization’s strategy and business model through BDA is particularly challenging as companies often fail to incorporate the associated economic features and underlying mechanisms that characterize the digitization phenomenon (Loebbecke and Picot, 2015).

Academia is also strongly impacted by the advent of BDADS, which have engendered a wealth of new means to conduct research even though theoretical and methodological implications are still being debated (Agarwal and Dhar, 2014). Among the most recent outlets relying on the analysis of big data, Papadopoulos et al. (2017) for instance, used unstructured big data to investigate supply chain networks resilience in the context of disaster recovery. They found that the critical enablers of supply chain networks resilience were swift trust, information sharing, and public-private partnership.

Nonetheless, academia is still at a very nascent understanding of the mechanisms through which BDA can impact firm performance for value creation (Fosso Wamba et al., 2017). Big data and predictive analytics (BDPA) adoption has, for instance, been conceptualized as a threefold process involving acceptance, routinization, and assimilation (Gunasekaran et al., 2016). Based on a sample of 205 Indian companies, the authors found that BDPA acceptance is positively related to BDPA assimilation under the mediation effect of BDPA routinization, which in turn positively affects both supply chain performance and organizational performance (Gunasekaran et al., 2016).

**Impact of BDADS on organizational roles and positions**

Internally, BDADS and digitization in general are reshuffling the cards of companies’ organization and structure while new cards are being introduced in the card deck. A number of leading companies have decided to create a new breed of executives, chief data officers (CDOs), in charge of managing big data at the executive level and to ensure its alignment with business strategy. A recent survey involving 600 global executives concluded that companies with a top executive in charge of big data had higher firm performance than other companies (Lee et al., 2014). According to a recent Gartner (2016) report, around 90 percent of large organizations will have a CDO by 2019. Nonetheless, the critical role
played by technology in strategic business decisions has also engendered a new executive role, the chief technology officer (CTO), whose mission is to identify profitable applications of technology to products, services, and processes. Meanwhile, marketing has become totally dependent to technology up to a level that another type of executive such as the chief marketing technologist has also emerged intersecting with stakeholders such as the chief marketing officer, the chief information officer (CIO), and the overall marketing function (Brinker and McLellan, 2014). The intrusion of CDOs within organizations has also modified the power relationships among the various functions of organizations. Until today, there is no clear consensus about the power and authority entailed to CDOs, ranging from data policy to business strategy (Lee et al., 2014). CDOs within large firms were found to report to either CEOs directly, chief operation officers, CFOs, CIOs, or even CTOs (Lee et al., 2014).

The “sexiest job of the 21st century” (Davenport and Patil, 2012), data scientists, is the most obvious example of how BDADS is impacting organizations internally. As companies are rushing to capitalize on the potential of big data, data scientists are among the most wanted (and well-paid) profiles on the job market. In short, data scientists are those who can make data speak. Using Davenport and Patil’s (2012) own words, they are the “people who understand how to fish out answers to important business questions from today’s tsunami of unstructured information” (p. 73). Organizations are now hiring data scientists for a majority of their business functions as analytics is now pervading areas such as finance, operations and production, sales and marketing, research & development, or human resource management (Lavalle et al., 2011; Waller and Fawcett, 2013a).

**Impact of BDADS on organizational processes**

While BDA impacts the strategy, structure, and organization of companies, it also transforms their processes. First, BDA pervades most existing processes (Lavalle et al., 2011) allowing to boost business value by increasing their speed (Gartner, 2013). For instance, the adoption of BDA into supply chain management enhances significantly its processes by allowing to “gain visibility into expenditures, identify trends in costs and performance, and support process control, inventory monitoring, production optimization, and process improvement efforts” (Hazen et al., 2014, p. 72). Moreover, embedding BDA into business processes goes further than merely supporting and improving them. Its adoption requires to completely revamp business processes and work flows if a company wants to fully benefit from its use (Lavalle et al., 2011; Zolnowski et al., 2016). Walmart, for example, processes transactions by the million every hour. The data generated from mobile devices (through mobile apps or through store cards) fuel the company’s customer offer and service processes. The identification of buying behaviors and patterns allows to push personalized offers in real time (Gandomi and Haider, 2015). Eventually, embedding sensor networks, fast processing, and analytics into business processes impacts their very nature. While the generation and storage of real-time data equip processes with a nervous system that is capable of sensing all activities that surround them, the use of real-time and predictive analytics make business processes alive. They are acquiring the capability to constantly learn, adapt, and evolve through the results derived from both machine learning and artificial intelligence algorithms. Such capacity boosts the dynamic capabilities of firms (Braganza et al., 2017) by providing powerful means to continuously reconfigure processes in the sake of achieving more beneficial outcomes (Braganza et al., 2017).

**Impact of BDADS on decision-making processes**

Among the many processes that are impacted and often transformed by BDA, decision-making processes are particularly concerned (Picciano, 2012; Brynjolfsson et al., 2015). According to Provost and Fawcett (2013) “the ultimate goal of data science is improving decision making” (p. 53). IT-supported decision making has gained popularity in the 1980s
with the advent of information technologies that helped managers “optimize” their decisions (Picciano, 2012). Group support systems were implemented in a large number of firms in the 1990s. They consist of software tools used in group discussions which are used to optimize the outcome of decision processes, thanks to the enhancement of participant communications through features such as parallelism or anonymity (Dennis et al., 2001).

Today, data-driven or “algorithmic” decision making goes much further. Data-driven decision making (DDD) refers to “the practice of basing decisions on the analysis of data rather than purely on intuition” (Provost and Fawcett, 2013, p. 53). Through the real-time collection and analysis of large quantities of data, managers have access to conclusions provided by algorithms when making strategic decisions (Newell and Marabelli, 2015). As Jeff Stanton, from Syracuse University puts it simply: “data science is really about building things that can lead to better decision-making” (Dumbill et al., 2013). A recent study involving 179 large publicly traded firms found that the adoption of DDD has an impact on productivity, asset utilization, return on equity, and market value (Brynjolfsson et al., 2015).

However, introducing DDD within organizations is impossible without setting dedicated processes that cover the whole data lifecycle span from its generation to actual decision making (Janssen et al., 2017). For instance, Bizer et al. (2012) identify the following six necessary steps: data capturing, data storage, data searching, data sharing, data analysis, and data visualization. Other scholars, such as Zhou et al. (2014), proposed closely related stages: data collection, data storage, data management, data manipulation, data cleansing, and data transformation. In spite of the identification of the necessary steps to enable DDD within firms, less attention has been paid to the individuals in charge of executing them as well as to the interdependencies between all those steps (Janssen et al., 2017). As a result, there is no clear consensus within both the academic and practitioners’ spheres about the role and function of decision makers, managers, within each of the data lifecycle stages as well as how the interactions among all steps ultimately affect decision efficiency and quality. The only certainty is that DDD is about to redefine the job and function of managers.

Organizations’ lack of managers with analytics skills

It is now clearly acknowledged that the global gap between the demand for big data and analytics talent and its supply is one of the key challenges that seriously hamper BDA implementations across organizations. The often quoted 2012 Gartner report predicted that, by 2015, the demand for big data-related jobs would reach 4.4 million jobs throughout the world while barely one-third of those jobs would be filled (Petley, 2012). Predictions for the following years are not any better as the gap is forecast to significantly widen … IDC predicts a need by 2018 for 181,000 people with deep analytical skills, and five times more jobs with the need for data management and interpretation skills (Deloitte, 2016). Such predictions rang the alarm bell for all universities and business schools, urging them to develop new programs and adapt older ones to urgently answer the job market’s demand. Business schools may be tempted to think that this drastic skill shortage only concerns IT-related jobs. Such a view is absolutely erroneous as it would simply omit the essential human and management layers that surround the big data revolution.

Peter Sondergaard’s words (Senior Vice President at Gartner and Global Head of Research), during his 2012 keynote speech of the Gartner Symposium/ITxpo, illustrate how the boundaries of the skill shortage bridge IT-related domains:

In addition, every big data-related role in the US will create employment for three people outside of IT, so over the next four years a total of 6 million jobs in the US will be generated by the information economy (Sondergaard, 2013).

Meanwhile, the 2012 Gartner report foresaw a lack of 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data.
The report also anticipated a lack of an additional 1.5 million managers and analysts in the USA alone for skills related to the effective identification and formulation of big data opportunities, and the consumption of their associated analysis.

Throughout the world, there is now a dearth of academic programs that train future data scientists (Asamoah et al., 2015; Provost and Fawcett, 2013; Waller and Fawcett, 2013a). The question about whether data scientist profiles actually exist in the job market or if they are simply “mythical” profiles is outside the scope of this paper. What organizations can be sure about is that embedding BA and DS necessitates much more than just hiring a squad of data scientists. While the attention remains on the “sexy” side, that is to say the persistent search for gifted data scientists, what about managers and decision makers? Experts’ reports are quite clear about the lack of managers with analytical talent and understanding. While data scientists are still among the sexiest and most wanted jobs, they do not have the power to transform alone an entire organization. Engaging on the path of data transformation signifies that organizations must overall become “scientific” or “analytical” in its nature and functioning. In turn, the function of managers in the data-driven business era must also evolve and embed a “scientific” or analytical component as decision-making processes are becoming data driven. In other words, it seems that managers are about to turn into DDD scientists.

Methodology
This paper is an opinion paper that aims at positioning itself as helping to tackle a challenge that is controversial, institutional, and disciplinary in nature (Rowe, 2012). It addresses the complex and multi-faceted problem of adapting business education to the new data-driven business paradigm shift; a critical issue that deserves the urgent attention of both practitioners and researchers but that cannot be addressed through a typical scientific investigation (Te’eni et al., 2015). The paper develops a discourse with arguments that are supported by both the academic and practitioners’ literatures and provides implications to research and practice. As a result, the literature that is mentioned throughout the argumentation is used as supportive evidence of the claims that are being expressed. The literature review on which this paper is built has been performed in a manner that was as systematic and exhaustive as possible using the conventional bibliographic databases, with the specific purpose of capturing the complexity and multiple facets of the data-driven business transformation along the associated challenges faced by education. The main objective is not to investigate the issue through a scientific analysis but rather to initiate an open debate involving academics, practitioners, and educational institutions, questioning and revisiting the role, function, and nature of business education in a business world that is irremediably becoming “data driven.”

Beyond BDA, the advent of a data-driven business era
Whether we call it big data, analytics, DS, or even if we use any other trendy term, there is no doubt that businesses are facing a reshuffling of the cards across all industries and sectors. Companies are struggling to survive with more or less success, in a maelstrom of digitally enabled changes that impact their very nature that is to say: their strategy, business model, business processes, and internal structure. While this phenomenon was originally perceived as a technological disruption (Agarwal et al., 2008), it seems that we are witnessing a change of greater scope: the advent of a new business paradigm.

The big picture: the new data-driven business paradigm
Digitization and BDADS have been engendering a transformation of the business world as well as our entire society (Loebbecke and Picot, 2015). While some rather label this phenomenon
“datification” (Galliers et al., 2015; Newell and Marabelli, 2015), there is now no doubt that it permeates all aspects of life and has given birth to “new ways of working, communicating and cooperating” (Loebbecke and Picot, 2015, p. 149). The gradual shift from initially understanding such phenomenon as a mere technological innovation to eventually perceiving it as an entire societal change is well illustrated in the disappearance of “big data” from 2015 Gartner’s hype curve for emerging technologies (Burton and Walker, 2015); while it was still in the peak of inflated expectations in 2014. A few years back, Gartner’s hype curve had predicted that big data would start making a deep transformational impact within two to five years (Heudecker, 2013). To justify the absence of big data from the 2015 Hype Curve, Betsy Burton explains:

What’s happening is that big data has quickly moved over the Peak of Inflated Expectations […] and has become prevalent in our lives across many hype cycles. So big data has become a part of many hype cycles.

In other words, it can be argued that the reality that looms behind the big data phenomenon has bridged technological boundaries for becoming a much broader societal change supported by a wealth of technological innovations such as machine learning, natural language processing, or the Internet of Things; all located on Gartner’s 2015 and 2016 hype curves.

By engaging the path leading to data-driven business, there is no way back for companies as it is a transformative journey that changes the nature and character of products, processes as well as marketplaces and competitive environments (Loebbecke and Picot, 2015). This justifies the use of the term “business paradigm shift” when qualifying this overall change in the business world. The data transformation consists of turning traditionally business-driven companies into data-driven businesses that is to say into entities which overall functioning rotates around the collection, storage, and analysis of data in both a real-time and predictive manner (Lavalle et al., 2011). It is however an evolutionary process along which a company will gradually understand the potential of BDA and integrate such capability through the routinization of processes (Janssen et al., 2017).

A data-driven company is “smart” in the sense that analytics gets embedded within every organizational component allowing to turn data into actionable insights and this at all organizational levels (Lavalle et al., 2011). While the main motto of IT processes and systems used to be about “automation,” they now are specifically designed for generating insights in the data-driven world (Davenport et al., 2012). Besides, data-driven organizations are characterized by a certain analytics mindset and culture that are quite scientific in nature. It is about constantly scrutinizing a company’s successes and failures through the analysis of data (Gino and Staats, 2015). Companies such as Pixar are known for having such analytics spirit as each project, whether it is a success or not, is an analyzed post-mortem through a data-based investigation. Ed Catmull, the President of Pixar and Disney Animation Studios explains that “data can show things in a neutral way, which can stimulate discussion and challenge assumptions arising from personal impressions” (Catmull, 2008).

Whereas failure is used to be severely condemned in our nowadays highly competitive world, it becomes a positive mechanism in the new data-driven business reality. A data-driven mindset is characterized by both an experiential and learning cultural component in which one learns from failures and gradually improves through several cycles of experiments and actions (Carillo, 2015). For instance, sentiment analysis modeling is an ever evolving task as language keeps changing: new words keep appearing and our society also keeps giving birth to new ways of expression. For instance, hundreds of distinct emoticons and emojis are used in digital communication channels to express a broad range of emotions. Developing social media data models is thus an incremental process during which human actions and corrections allow to refine models up to satisfactory levels of accuracy and performance. In short, the data-driven business “way of working” is not about success but rather about experimental failures that eventually lead to success.
In the data-driven reality, analytics pervades every aspect of organizations. Data becomes the new digital oil that runs through the veins of companies and its analysis in a real-time and predictive manner, nurtures strategy, business models, operations and processes and particularly decision making. Today, analytics experts, especially data scientists, are the most wanted profiles (Davenport and Patil, 2012). All business functions are in desperate need to develop their analytics capability and hire professionals. This includes finance, marketing, production and supply chain management, and obviously IT. Job variants have appeared in job postings, with the term “analyst” being juxtaposed to specialization domains (or sub-domains). The job market is inundated with offers for HR/people analysts, marketing analysts (including CRM, advertising analysts), financial analysts, product analysts, or business process analysts (Jain, 2015). The job descriptions that are associated with such offers are all strongly tainted with “big data” and analytics.

However, while companies are striving to engrain an analytics capability within each business function and department, it seems that the main point is being missed. Analytics is part of the data-driven reality as it embodies the sensing system that characterizes data-driven organizations, providing companies with the capacity to derive actionable insights from big data. In other words, once an organization has toppled into the data-driven business paradigm, there is no marketing analytics function within marketing. Rather, the marketing function operates in a data-driven mode which relies on the daily use of analytics. As a result, companies shall not put too much emphasis on implementing analytics but shall rather concentrate their effort on their data transformation which involves adopting analytics among other important related aspects. This argument has important implications regarding how human resources are to be managed throughout the data-transformation journey. Hiring a team of data scientists for the marketing department of a company does not necessarily make it a data-driven marketing department. Rather, the nascence of the data-driven business era is engendering a mutation in the nature and function of managers.

Moving away from the data scientist frenzy …

The 2011 McKinsey’s report was among the first signals to herald the “big data” revolution (Manyika et al., 2011). When companies started realizing not only the need but also the desperate talent shortage in terms of analytical skills, educating professionals capable of mastering the challenges of analytics became of utmost importance (Schoenherr and Speier-Pero, 2015). The data scientist profile appeared as the missing piece of the big data puzzle. Companies concluded that they had to simply inject data scientists within organizations’ departments and to partially rethink organizational structure so that data scientists could be incorporated (Davenport and Patil, 2012). Thus, there has been a tendency to narrow down the difficult and all-encompassing challenge of the data transformation to simply hiring data scientists. In a survey involving 600 global executives, Lee et al. (2014) concluded that data scientists are usually placed in operational business units without any execute-level leadership, preventing to efficiently harness the full value of BDA. The results of this study confirm that companies have overall suffered from a lack of global awareness of the scope, depth, and long-term implications of the data-driven business transformation.

Data scientists have often been called “five-legged sheep” (Carillo, 2015) or “superheroes” (a term used by Pascal Clement, the Head of Amadeus Travel Intelligence) as their skillset is at the convergence of three expertise domains that are mathematics, computer science, and business:

Data scientists [...] understand analytics, but they also are well versed in IT, often having advanced degrees in computer science, computational physics or biology- or network-oriented social sciences. Their upgraded data management skill set – including programming, mathematical and statistical skills, as well as business acumen and the ability to communicate effectively with decision-makers – goes well beyond what was necessary for data analysts in the past. This combination of skills, valuable as it is, is in very short supply (Davenport and Patil, 2012, p. 23).
The idea that a data scientist can “do it all” has often proven its limits once on the field. First of all, the definition of the data scientist species renders the number of potential legitimate candidates very scarce. Indeed, it seems reasonable to assume that the brain of an individual with advanced degrees in both computer science and mathematics may not be wired the same way as the one with high-level business education. The necessity to have a computer science/mathematics brain cohabiting with a business brain within the same skull has been seen as particularly challenging and even unrealistic by both academia and practitioners. Often, talented individuals in scientific disciplines lack a deep understanding about the business domain in which he or she works.

Dumbill et al. (2013) report the following words from Professor Jeff Stanton from Syracuse University:

From a teaching perspective, as a faculty member I can teach someone how to do a t-test in 10 min, and I can teach them how to write a Python program in half an hour, but what I cannot teach them very easily is the domain knowledge. In other words, in a given area, if you are from healthcare, what you need to know in order to be effective at analysis is very different than if you are in retail. That underlying domain knowledge, to be able to have a student come up to speed on that is very hard (p. 22).

In the same outlet, Shelly Farnham, Executive Director and Research Scientist at Third Place Technologies (formerly at Microsoft Research and Yahoo) reiterates the key importance of mastering domain knowledge:

One of the challenges is that data science is not agnostic of domain. For example, when we are looking for people, interns or full-time people on our team, we definitely look for people who have experience analyzing data, but they also should be deeply engaged with the topic […]. I think that the domain knowledge is a very important aspect of what we are looking for (p. 25).

Second, there has been no real consensus to date on the necessary skillset associated with data scientist profiles (Waller and Fawcett, 2013a; Schoenherr and Speier-Pero, 2015). As a consequence, anyone can claim to belong to the data scientist species. One of the reasons is perhaps that there has been confusion about what BDADS really consists of (Provost and Fawcett, 2013). Added to the very attractive salaries that are attached to data scientist job posts, it is understandable that individuals with skills in BDADS have been tempted to market themselves as data scientists (Press, 2016).

How shall we prepare the future managers of the data-driven business era?

The business education chess game

Since the beginning of the twentieth century, the main mission of business schools has been to educate professional managers and to align curricula and teaching models to the ever-changing business environment (Seethamraju, 2012). Since the early 2000s, academics and practitioners have regularly voiced their concern about the broad-scope crisis faced by management education (Mintzberg, 2004; Ethie, 2003). Some have criticized the dramatic cultural shift of business schools which have been placing more and more importance to the rigor and quality of their scientific research at the expense of the academic competence of their graduates (Bennis and O’Toole, 2005). Others have raised their concern about the structural aspects that characterize business schools. The pedagogical model of business education originates from the beginning of the twentieth century and has been organized in a discipline-based manner typically including departments such as finance, marketing, logistics, accounting, human resources, or strategy (Seethamraju, 2012). This has resulted in a criticized “functional silo-based approach to teaching” lacking multidisciplinary integration (Navarro, 2008) and the overusing of “cookie cutter” curriculum designs (Porter and McKibbin, 1988, p. 314).
Practitioners from knowledge domains that are inherently cross-functional such as information systems (IS) (Ethie, 2003) or business process management (Seethamraju, 2012) have raised serious concern about the critical need to develop process-oriented curricula. The measures taken toward that direction are reflected in the requirements of accreditation bodies such as the AACSB (Seethamraju, 2012). Business schools have also relied on an array of pedagogical strategies to tackle this problem such as integrated case studies, capstone subject projects, simulation games, or team teaching.

(Big data) analytics/DS trainings and programs
Back in 2012, the nascent awareness about the talent shortage in terms of analytics skills made academics and practitioners realize the overall absence of university programs offering degrees in analytics/DS (Davenport and Patil, 2012). The reality is now quite different as we have witnessed a surge of academic initiatives and programs that deliver big data/analytics/DS-related training (Asamoah et al., 2015; Schoenherr and Speier-Pero, 2015). As of 2016, they can be counted by the hundreds worldwide. North Carolina State University[2] reports more than 110 Master degree programs in analytics and DS in the USA only. Program formats vary along the range of programs being offered. While certain institutions provide comprehensive degree programs in DS or BA (Anderson et al., 2014; Schoenherr and Speier-Pero, 2015), others offer individual or packaged courses that are embedded within broader academic programs (Asamoah et al., 2015). In terms of academic legitimacy, there exist different viewpoints according to whether academic institutions perceive the big data phenomenon to be a technical/technological topic, a business one, or else somewhere lying in between. Consequently, analytics/DS trainings have been delivered by either the computer science or the IS department, or by a combination of business school departments (Asamoah et al., 2015; Chiang et al., 2012; Gupta et al., 2015).

Available trainings in big data/analytics/DS can be categorized into three overall groups. First, specifically dedicated Master of Science programs have been launched by an array of educational institutions worldwide. Master of Science in Analytics/BA/DS programs are run as university-wide collaborations (such as in dedicated institutes), or in colleges of engineering or science. Master of Science in BA is however often based in business schools, sometimes in collaboration with other university departments. What is rather puzzling about the various offerings is that the labels “analytics,” “BA,” and “DS” are often used interchangeably. Second, some Master’s programs offer a specific analytics track. This particularly includes Master of Science programs in IS offering a concentration in analytics/DS (Chen and Storey, 2012) but also MBA programs proposing a special focus on big data/BABA (Warner, 2013; Gupta et al., 2015).

Finally, on-site or online certification programs are also offered by either educational institutions or by firms. These programs target IT professionals who want to acquire skills to extend their business or IT expertise, as well as business professionals from non-IT domains who wish to gain some technical and analytical skills (Chen and Storey, 2012). For instance, Harvard University proposes a DS Certificate program while Boston University and the University of Maryland deliver their own Data Analytics graduate certificates. Partly in response to the absence of adequate profiles on the job market, a number of companies have launched their own certification programs. EMC, for example, launched their own program (Davenport and Patil, 2012) while IBM started the “Big Data University’ initiative back in 2011 and has ever since registered over 400,000 students. The company has opted for working toward “democratizing” access to big data education (courses are offered for free) with the intention to grow a broad community of future potential buyers, users, and customers.
**IS coming to the rescue of management education**

The main quest of the IS discipline is to tackle challenges and identify opportunities that can have a long lasting scientific and societal impact (Chen, 2011). Shortly after the publication of the 2011 McKinsey report, the IS field realized the urgent need to revisit existing curricula and launched action plans to provide BI and BA education programs that would address the new generation of data/analytics savvy and business students/professionals (Chen and Storey, 2012). In 2012, the AIS Special Interest Group on Decision Support, Knowledge and Data Management Systems (SIGDSS) and the Teradata University Network conducted surveys to assess academia’s response to the growing market need for students with BI and BA skillsets with an emphasis on “big data” (Wixom et al., 2014). The report concluded that the IS field was particularly well positioned to train the next-generation BI/BA workforce.

The most challenging curricular aspects when reflecting on the most efficient means to prepare future managers are the breadth and depth of skillsets that are needed to become a highly capable professionals (Schoenherr and Speier-Pero, 2015). Since its early beginnings, some 30 to 40 years back, IS has been an interdisciplinary field in nature mastering trans-disciplinary dialogs (Galliers, 2003). If one discipline could claim legitimacy in being the most appropriate candidate to deliver education programs based on a new knowledge domain that is at the junction of computer science (data/database management), statistics (analytical and modeling tools), and business (business processes and decision making), IS is the first discipline that immediately comes to mind. The advent of the data-driven business era is an unprecedented opportunity for IS departments to play a central role in leading the education of next-generation professionals. However, considering the implications and subtleties of the advent of the data-driven business era, it is still not consensually decided whether IS departments alone will be able to take on such responsibility.

**Turning managers into manager-scientists**

Data scientists cannot replace decision makers. In data-driven businesses, algorithmic/DDD consists of involving the insights generated by algorithms (that often compute large amounts of data) into decision-making processes (Newell and Marabelli, 2015). However, without proper management and executive-level leadership, the recommendations provided by algorithms are often ignored by decision makers (Mayer-Schonberger and Cukier, 2013). Becoming a successful manager in the data-driven business era goes way beyond acquiring a few additional skills. It is rather about shifting mindset and adopting a new way of thinking:

In the era of Analytics [...] managers must drive efforts on at least 10 fronts, from creatively combining data management approaches to shaping new analytics-focused roles to setting guidelines for responding to “digital smoke signals” (Davenport, 2013, p. 67).

Analytics and DS are the support of data-analytic thinking. However, skill at thinking data analytically shall not be reserved to data scientists. It must rather become an organizational cultural component that shall spread to the entire organization (Provost and Fawcett, 2013). Managers can only get the maximum value that can be extracted from analytics/DS resources only if they have a minimum understanding of the core principles (Provost and Fawcett, 2013). Even though algorithms make very fast and accurate predictions, “they also create risks of their own, especially if we do not understand them” (Luca et al., 2016, p. 98). This statement also extends to operational and line employees. For instance, managers shall be familiar with the fact that the results provided by BDA generally involve correlation, not causation, and could also be due to mere chance (Davenport, 2013). Expertise in experimental design, for instance, can help crossing the gap between correlation and causation (McAfee and Brynjolfsson, 2012).
The job of managers often involves making predictions. BDA and algorithms provide new means to process data with incredible speed and scale, generating predictions with accuracy levels that are not comparable to those a human brain can produce. However, algorithms can suffer from myopia (Luca et al., 2016). Because the data that is used to parameter algorithms tends to focus on short-term outcomes, there is often a misalignment between short-term success and longer-term profits and corporate objectives (Luca et al., 2016). From the managers' perspective, the problem does not come from analytics itself but rather from the way they interact with it. In other words, managers shall develop an understanding about what questions analytics-based algorithms answer and what questions they do not.

Managers shall also develop the ability to understand BDA outcomes (Janssen et al., 2017). Early research about “data-based” decision making showed that decision quality is strongly increased whenever a decision-maker has a clear understanding about the relationships among the problem variables (Raghunathan, 1999). This result implies that a decision-maker shall interact with the actors in charge of collecting and processing the data (Janssen et al., 2017). Assuring that the main objective of DS is improving decision making (Provost and Fawcett, 2013, p. 53), interactions among all direct and indirect actors of the BDA chain are indispensable (Janssen et al., 2017). In other words, with no common view and understanding, DDD cannot lead to “better” decisions.

Data come at managers with such velocity and volume that some layer of visual abstraction has become unavoidable to comprehend the reality that hides behind the data (Berinato, 2016). Data visualization has become a real art and research field. A plethora of data visualization tools is now available in the market, providing techniques that go way beyond the use of spreadsheets and charts. Organizations engaging the data-transformation path have understood the business value that can be derived from visualization tools and techniques (McAfee and Brynjolfsson, 2012). The company eBay has developed a marketplace with hundreds of millions of active users and selling billion of goods every year. The company adopted Tableau’s visualization tool which allows employees to access petabytes of user behavior data and monitor in real-time issues such as search relevance and quality, customer feedback, or conduct sentiment analysis (Sivarajah et al., 2017). The main purpose of using visualization tools and techniques for managers is to convey knowledge hidden in complex and large-scale data sets in a clear and aesthetic fashion (Chen and Zhang, 2014). Furthermore, visual communication thus becomes an essential skill for managers (Berinato, 2016). Traditional presentation tools that rely on the presentation of fixed conventional charts and graphs are not appropriate to present complex analytics-based results that often require to be presented in a dynamic way. For instance, Berinato (2016) identifies four types of visual communication depending on the nature of the question being asked: idea illustration, idea generation, visual discovery, and “everyday dataviz.” As a consequence, presentation and communication skills get more and more tainted with a strong analytics flavor.

Engraining an analytics-based DNA into data-driven businesses consists of developing a common understanding between analytics specialists but also managers (in terms of core analytics and DS concepts) added to a shared “analytics way of thinking.” This consists of a mix of skills combining conventional knowledge in business and management as well as data management and analytical/modeling techniques and tools; all this coupled with an awareness of the strategic value of data (see Figure 1). Without such shared vision and knowledge, implanting data scientists into organizations will lead to no benefits but rather to friction among employees due to the brutal empowerment of this new breed of high-level profiles. Research efforts aiming at determining the skillset data scientists shall be equipped with, highlight such combination of skills from all three domains added to a peculiar analytics mindset often characterized by an inquisitive and passionate attitude toward analytics-based business problems.
(Chatfield et al., 2014; Waller and Fawcett, 2013b; Schoenherr and Speier-Pero, 2015; Dubey and Gunasekaran, 2015). For instance, Schoenherr and Speier-Pero (2015) report the results of interviews conducted at Michigan State University with leading analytics experts from firms such as IBM, PriceWaterhouseCoopers, and Kellog’s. The interviewees were asked to share their understanding and express their needs when it comes to hiring data scientist profiles. The responses (reported in Table I)

<table>
<thead>
<tr>
<th>Expert and title</th>
<th>Company</th>
<th>Desired skillsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tim Rey, Director of Advanced Analytics</td>
<td>Steelcase</td>
<td>Being able to convert data into business gain; being inquisitive about problems;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>creativity; having a mathematical slant; statistics; machine learning; operations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>research</td>
</tr>
<tr>
<td>Philip Lear, Manager of Trade Analytics</td>
<td>Kellogg’s</td>
<td>Critical thinking; mathematics; programming; however, it is not really only about</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crunching numbers and getting statistics, but to develop insights from numbers;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>understanding the business behind it is thus important; passion</td>
</tr>
<tr>
<td>David Dorleans, Manager, Advanced Risk and</td>
<td>PriceWaterhouseCoopers</td>
<td>Ability to analyze the data, but then also to convey usable results and implications</td>
</tr>
<tr>
<td>Compliance Analytics</td>
<td></td>
<td>to executives (communication skills)</td>
</tr>
<tr>
<td>Mike Marshall, Director of Marketing and</td>
<td>J.D. Power &amp; Associates</td>
<td>Quantitative skillsets, ability to find and see patterns, passion for discovering</td>
</tr>
<tr>
<td>Statistical Science</td>
<td></td>
<td>things; inquisitive mindset; technical capabilities and skillsets</td>
</tr>
<tr>
<td>Richard Rodts, Manager of Global Academic</td>
<td>IBM</td>
<td>Understanding what questions to ask (not necessarily with a big technology background);</td>
</tr>
<tr>
<td>Programs for Data Analytics</td>
<td></td>
<td>being able to address business needs; leverage technology to look further into data to facilitate better decisions; communication skills (need to tell a story about why the data matter); mathematics; sociology</td>
</tr>
<tr>
<td>Jeremie Juban, Chief Data Scientist, Statistics, Data Mining, Machine Learning</td>
<td>The Weather Company</td>
<td>Being able to spend time with the data, coupled with the desire to understand what is behind the data</td>
</tr>
</tbody>
</table>

**Table I.** Desired data scientist skillsets

*Source: Adapted from Schoenherr and Speier-Pero (2015)*
are particularly illustrative of the need for a knowledge domain blend encapsulated within a specific analytics mindset. Managers cannot be obviously expected to possess an equivalent level of sophistication when it comes to aspects such as unstructured database management, Python/R programming, or clustering algorithms. They shall nonetheless gain basic skills in such domains as it is a sine qua non condition to developing an analytics way of thinking and DNA.

Finally, practitioners and academics have recurrently emphasized the need for data scientists along with BDADS professionals to develop certain soft skills (including presentation and communication) so as to bridge the knowledge gap that exists within the rest of the organization. Reciprocally, employees and more precisely managers shall also move toward the world of data scientists by adopting an analytics-based vision and “way of thinking” reflected in the development of such soft skills. Indeed, such competences are core elements that define the analytics mindset and cultural component that organizations shall aim at embedding within their employees’ mind. Recent research efforts have attempted to identify the critical soft skills, future big data and BA professionals shall be equipped with (Dubey and Gunasekaran, 2015). Based on the analysis of qualitative data collected by interviewing ten heads of BA units, the authors identified the following key soft skills: leadership, team skills, listening, positive attitude, communication, interpersonal skills, patience, and passion. Developing and strengthening such analytics-based inquisitive and passionate attitude shall not be neglected as it is the essence of the manager-scientists’ DNA.

Discussion and implications

Our business world will have completed its data-mutation when the word “analytics” will stop being attached at the end of each business function and domain name. Irremediably, organizations are turning into entities dotted with a nervous system sensitive to organizations’ internal and external stimuli. While electric signals and chemicals transit in reaction to stimuli along the nervous system of living beings, organizations have grown a nervous systems based on the generation, transfer, process, and storage of data. To carry on with the metaphor, nervous communications play a central role in determining and triggering sensations, feelings, thoughts, motor/emotional responses, learning, or memory. While instant reactions such as feelings or actions (motor responses) correspond to real-time analytics processes, those triggered through learning and memory pertain to the domain of predictive analytics. The brain (top management) and the spinal cord (middle management) thus play a central role in determining the reactions and responses to organizational and external stimuli.

Implications for practice

*Standardization and consensus.* For the sake of transparency and clarity, academia and the industry must join forces to standardize the meaning and scope of the terms surrounding “big data.” This starts with “big data” itself which emphasis on size and volume have been misleading. The big data phenomenon is not a matter of size. It is about the central and critical role that data now play in businesses and in our entire society. As a consequence, the term “data driven” such as in data-driven business, DDD, or data-driven strategy seems more suitable as it better addresses the reality hiding behind the cloud consisting of all the fuzzy terms having been recurrently used. As a matter of fact, the term “BDA” is also evasive, “BA,” understood as the evolution of BI in the data-driven business era, seems more appropriate. Again, at the end of the data transformation of our society, the word analytics will naturally disappear since it will become inherently embedded within all aspects of businesses. Analytics will become an essential gene that belongs to organizations’ DNA. Caution shall also be employed with the use of the term DS.
Even though it seems that the existence of both terms is legitimate, scholars and industry experts must still reach a consensus about the domain characterizing each term. Educational institutions shall be fully concerned as it is critical that they rely on a shared taxonomy in their program offerings. Overall, both the academia and business spheres must retake control on the significance of terms and notions as the big data debate has overly drifted away, pushed by the waves engendered by the constant attention the media have paid to the big data buzz over the last few years.

**Intense collaboration with the big data ecosystem.** BDADS are inherently complex. This has resulted in the formation of an evolving and interconnected network of actors that interact (and often collaborate) with each other, covering a wide spectrum of specialization domains such as applications (vertical, log data, ad/media), BI, analytics, visualization, data/infrastructure as a service, analytics infrastructure, and even traditional structured database specialists. As a result, BDADS projects usually require the expertise and services from several companies belonging to a broad ecosystem that includes big data-related actors (including the many big data startups), as well as companies specialized in IT integration and services (usually provided by consulting firms), and open source software projects such as Hadoop, Apache HBase, or MongoDB to name but a few.

From a pedagogical perspective, teaching analytics and DS is thus particularly challenging. The complexity and many techniques and tools can only be apprehended through direct involvement with the overall BDADS ecosystem. In simple terms, BDADS training programs, courses, and curricula shall be organized in such a way that students shall interact with an array of specialists and experts providing them a broad enough picture of the BDADS landscape. Furthermore, the BDADS ecosystem evolves at an extremely fast pace. For instance, until the end of 2013, Hadoop was seen as the breakthrough technology of the big data phenomenon, allowing to manage and transform large quantities of data with performance levels that had never been reached before. In 2014, Apache Spark dethroned Hadoop by allowing to work 100 times faster than Hadoop 1.0, and the pattern will continue as Hadoop 2.0 and Spark 2.0 are now in the spotlight …

The BDADS sphere is a bubbling ecosystem in which startups keep emerging and growing, and in which sub-domain leaders are a mix of technology dinosaurs and mere infants with only a few years of existence.

As a consequence, higher education institutions have no choice but to try to keep up with the fast-paced evolution of BDADS techniques and tools. The only way to tackle such challenge is to close the gap between educational institutions and the world of practitioners and to engage into active and long-term collaboration. Teaching analytics or big data cannot be efficiently done without being directly “plugged” into the ecosystem. In other words, it is not sufficient to work toward reducing the dramatic analytics-related skill shortage that is foreseen for the next decade. BDADS are evolving at such a fast pace that new expertise, skills, and technologies keep emerging. It will not be astounding to hear from industry experts about another dramatic big data 2.0 skill shortage within the next few years …

**Breaking the walls of academic disciplines.** If the “next frontier for innovation, competition, and productivity” (Manyika et al., 2011) has heralded the transformation of the business world and our entire society (Loebbecke and Picot, 2015), it seems logical to presume that our educational system shall be severely impacted. It shall also engage a transformational path aiming at tackling the opportunities and challenges of the data-driven business era. The current function/disciplinary-based pedagogical model of business education was built at the beginning of the twentieth century when the business reality was a different one (Seethamraju, 2012). In the beginning of the twenty-first century, analytics is getting engrained within businesses’ DNA, providing a scientific component to the functioning of organizations and redefining the job of all employees from the top, middle, to the operational level.
The disciplinary walls of business schools started loosening with the need to develop cross-functional skills that are crucial in domains such as business process management and information technology/systems (Seethamraju, 2012; Ahmad et al., 2007). These walls are crumbling and about to collapse with the advent of the data-driven business era. The legitimacy of teaching DS cannot be only owned by computer science departments or BA training cannot belong to business schools or IS departments only. BDADS is basically the result of merging business, computer science, and mathematics together. It is the synergistic effect of all three fields together that defines BDADS, and not their mere juxtaposition. Such reductive view is rigorously equivalent to equate the result of a group task to the sum of its associated individual contributions. Moreover, a knowledge domain such as BDADS cannot be defined in a “cooking recipe” manner that is to say a mix of different disciplinary knowledge. Instead, knowledge domains shall influence and determine disciplinary boundaries. It is not satisfactory to design analytics/DS programs built on the delivery of courses provided by different departments and sprinkling a few capstone projects on top of them. This cannot be the most effective way to develop analytics genes within future managers’ and professionals’ DNA. In short, we may have come to the end of the well-established function-based model of business schools and perhaps universities. It is time to initiate a broad reflection on the mutation of higher education in the age of the data-driven business era. Accreditation bodies such as the AACSB or the European Quality Improvement System may certainly have an important role to play to speed up the process. Having realized that business education curricula had a tendency to fail at developing a sound understanding of emerging IT-enabled processes, products and services (Seethamraju, 2012), their requirements keep asking for incorporating multidisciplinary/cross-functional concepts (Seethamraju, 2012). Perhaps, it is under their responsibility that an open debate shall be launched, involving both the academic and professional worlds, and discussing about how the nature, structure, and functioning of business schools shall evolve.

**Developing experiential learning.** The multidisciplinary nature of analytics and DS has inherent pedagogical implications. It seems critical that the skill development process shall incorporate an experiential learning component allowing to provide the necessary practice to integrate the different areas at a granular level (Schoenherr and Speier-Pero, 2015). The data-driven business paradigm engenders a cultural shift in organizations. Whereas failure is used to be severely condemned in our nowadays highly competitive world, it becomes a positive mechanism in the big data world. Indeed, an analytics mindset is characterized by an experiential culture in which one learns from failures and gradually improves through several cycles of experiments and actions. Engendering an experiential culture within the mind of managers is a difficult task as it contradicts the basic nature of a manager’s job. Time is money and failing costs even more money. To convince future managers of the benefits of experimentation in the data-driven business context, the use of serious games or other types of simulation can be effective pedagogical strategies. Indeed, it is only through direct experimentation that one can seize that incremental learning allows to reach higher ends (at least in the BDADS context). To plant such seed in the mind of managers, well-designed analytics simulations shall accompany students through several loops of a fail-and-learn processes and demonstrate the overall benefits of the analytical approach.

**Implementing spiral-shaped pedagogical models.** Manager-scientists shall possess solid skills at the crossroad of data management, analytical/modeling techniques and tools, and business. In order to engrain an analytics mindset (and develop the associated skillset) within the mind of future professionals, these three knowledge domains cannot be taught in isolation from each other. When it comes to developing a mindset or a DNA, a more granular pedagogical approach seems more appropriate. In other words, BDADS programs shall not consist of the delivery of courses pertaining to each domain. Instead, analytics curricula
shall be built around the principle that each individual course or intervention shall systematically mix all three domains (even though “dosage” may vary). This seems to be the first necessary step to erase the mental perception of disciplinary walls separating all three domains. It is thus a means to develop a unique skill domain, “analytics,” within the mind of trainees. Nonetheless, this pedagogical approach is extreme in the sense that it demands a brutal mindset shift for trainees. Implementing a more gradual, “spiral-shaped” pedagogical strategy may be a more effective and reasonable solution. The underlying logic is to design curricula that accompany students along a learning path which steadily leads to the development of an analytics mindset and associated skills. As an individual progresses throughout the program, the perception of boundaries among domains gradually become more and more blurry to eventually disappear during the final stage of the program. The learning experience is spiral-shaped and not simply linear as the sequencing of courses and interventions is equally important. It assumes that the overall pedagogical approach shall be designed in a cyclical fashion, each loop following a predetermined sequence of domains that are business/management – analytical/modeling techniques and tools – data management. Each cycle differs from the previous one as the three knowledge domains get more and more blended with each other, up to the final stage (the center of the spiral shape) where individuals get totally immersed into the analytics world (see Figure 2). Program duration and program objectives would help determine the number of cycles throughout each program. Nonetheless, independently of program duration, a minimum of three cycles is necessary: one cycle during which the concepts are being taught in relative isolation. For instance, this could consist of providing introductory courses in management, in business statistics (including modeling techniques such as regressions), a programming course (R or Python being the most recurrent programming languages being used for DS), and finally an introduction to Relational Database (RDBMS) management (using SQL for instance). The second would involve the blending of two domains at a time for each course or intervention while the final one would be based on the integration and fusion of all three. Such high-level pedagogical strategy could help design curricula that more effectively prepare future professionals to the data-driven business reality.

Implications for research

Identifying key skills for managers of the data-driven business era. Academic research has an important role to play as there exists a number of unanswered questions for educational institutions and practitioners. First, this paper aims at being a stepping stone toward the identification of the skills that are critical for the new breed of managers. The provided

Figure 2.
The “analytics” spiral-shaped pedagogical model
insights have emerged from the extent literature but a more rigorous and scientific investigation is needed in order to determine an extensive list of skills and competencies that must be mastered, and this for each type of hierarchical position, job type, and business function. This is a sine qua non condition for the effective development of BDADS curricula that effectively match the skills that are needed in the different industries and sectors. In turn, this shall put an end to the monolithic “data scientists can do it all” view that has resulted in the proliferation of educational programs and trainings aiming at forming future data scientists. A variety of profiles and jobs are necessary for a company to enter the data-driven business world. Acknowledging such view, schools and universities shall thus work toward diversifying their offerings.

Gaining a deep understanding of the data-driven business transformation mechanisms. Various research efforts have started investigating the potential superiority of data-driven companies over traditional business-driven ones. For instance, Brynjolfsson et al. (2015) found that firms that adopt DDD have higher output and productivity while Zolnowski et al. (2016) determined that the implementation of data-driven innovation within organizations had an impact on their business model along four distinct transformational patterns that are: cooperative value innovation, customer-centric value innovation, cooperative productivity improvement, and company-centric productivity improvement. A lot more research is needed to fully comprehend the implications of engaging the path of data transformation. Meanwhile, assuming the overall superiority of the data-driven business mode is highly reductive as data-driven companies will face obstacles and pitfalls that are inherent to the data-driven way of functioning. We anticipate critical legal, ethical, privacy- and security-related considerations to play an important role along this line of inquiry. The attention of scholars is definitely needed to prevent organizations from falling into the data-driven business traps …

Business-driven vs DDD processes. As DDD processes are about to redefine the job of managers, research is needed to determine how this will impact managers’ cognitive and emotional processes. For instance, Mayer-Schonberger and Cukier (2013) indicate that suggestions provided to decision makers by algorithms that process large amounts of data are most often ignored. Research efforts are needed to further explore this phenomenon. Luca et al. (2016) emphasizes the need for managers to develop an understanding of algorithms but also of how to interact with them; the rationale behind such reasoning is that the short-sighted nature of algorithms may lead to decisions in contradiction with the long-term vision of organizations. Recent research works on the importance of humanizing algorithms have been providing insightful preliminary results (Luca et al., 2016). Psychology research can also help better understand how cognitive processes are impacted in data-driven business decisions while education research can in turn develop pedagogical strategies and models that are specifically tailored to build an analytics “way of thinking.” At a more societal level, the diffusion of algorithm-based decisions in many aspects of people’s lives is raising important ethical issues. Proprietary decision-based mathematical models are being more and more infused in our daily lives. Such algorithms can become “weapons of math destruction” through their tendency to encode and perpetuate inequality and discrimination (O’Neil, 2016). Academia shall dedicate research efforts to pursue along this critical line of inquiry.

Revisiting well-established theories and theoretical models. The business paradigm shift from a business- to a data-driven world shall engender a broad investigation of the validity and applicability of well-established theories in disciplines such as management, finance, economics, or the social sciences. For instance, Herbert Simon was awarded the Nobel Prize in economics in 1978 for his research on decision making in organizations. His theory develops the notion of “bounded rationality” based on the principle that organizations operate along a continuum of rational and social behaviors because the knowledge necessary to strictly function in a rational way is beyond their reach. The theory has
resisted the test of time and is still widely used in the understanding of organizational processes such as decision making (Picciano, 2012). There is no certainty that Simon’s theory of “bounded rationality” will withstand to the turning of organizations into data-driven analytics-based entities. Overall, research domains and streams such as cognitive psychology or decision theories (such as cognitive decision theory) shall be encouraged to dedicate research efforts to apprehend the specifics of DDD.

Limitations
This paper aims at initiating an open discussion involving researchers, practitioners, and educational institutions. However, this investigation inherits the limitations that pertain to opinion papers. Even though scholarly evidence has been extensively used to support the arguments having been developed, the provided recommendations and implications are not based on conventional scientific investigation. Rather, they have emerged from the synthesis of the extent literature around BDA, DS, and data-driven business combined with the opinions of leading industry experts and academics. This paper has identified a number of research avenues that can help further explore the implications of the advent of the data-driven business paradigm as well as its impact on our society, organizations, and individuals with a specific focus on decision-makers. It is hoped this paper will trigger such research efforts.

Conclusion
Moving away from the myriad of terms and interpretations that surround the advent of a business world that is becoming more and more data driven, we can no more question the fact that organizations are now caught into a gigantic data “storm” and that their survival will depend on the development of their capacity to analyze data and to turn it into actionable value-creating insights. The changes that are necessary are far reaching as it is not simply about adding an extra string to the bow of organizations. It is rather about engaging on a transformational path that will affect all organizational aspects ranging from strategy, business models, business processes, management, or structure. In other words, data scientists cannot alone engrain a data-driven business DNA within organizations. Besides, the data “wind gusts” are so strong that the walls of the educational system and more specifically business education, are shaking. Educational institutions shall also engage the path of data transformation while business schools and universities need to rethink the pertinence of their silo-based functioning as well as the effectiveness of the implemented pedagogical models. Irremediably, academic thinking has no choice but to be gradually infused with analytics and become “data driven.” Education may otherwise miss the data-transformation train and thus fail at tackling the contemporary educational challenges that pertain to the datification of our society and overall mankind.

Notes

References


Gartner (2013), “Gartner says CIOs will need to manage both technology and business innovation to gain competitive advantage with big data”, Gartner, Stamford, CT, available at: www.gartner.com/newsroom/id/2476715 (accessed September 1, 2016).


Managers for the (big) data-driven business era


Corresponding author
Kevin Daniel André Carillo can be contacted at: k.carillo@tbs-education.fr

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Does big data analytics influence frontline employees in services marketing?

Saradhi Motamarri, Shahriar Akter and Venkat Yanamandram

School of Management, Operations and Marketing, University of Wollongong, Wollongong, Australia

Abstract

Purpose – Big data analytics (BDA) helps service providers with customer insights and competitive information. It also empowers customers with insights about the relative merits of competing services. The purpose of this paper is to address the research question, “How does big data analytics enable frontline employees (FLEs) in effective service delivery?”

Design/methodology/approach – The research develops schemas to visualise service contexts that potentially benefit from BDA, based on the literature drawn from BDA and FLEs streams.

Findings – The business drivers for BDA and its level of maturity vary across firms. The primary thrust for BDA is to gain customer insights, resource optimisation and efficient operations. Innovative FLEs operating in knowledge intensive and customisable settings may realise greater value co-creation.

Practical implications – There exists a considerable knowledge gap in enabling the FLEs with BDA tools. Managers need to train, orient and empower FLEs to collaborate and create value with customer interactions. Service-dominant logic posits that skill asymmetry is the reason for service. So, providers need to enhance skill levels of FLEs continually. Providers also need to focus on market sensing and customer linking abilities of FLEs.

Social implications – Both firms and customers need to be aware of privacy and ethical concerns associated with BDA.

Originality/value – Knitting the BDA and FLEs research streams, the paper analyses the impact of BDA on service. The research by developing service typology portrays its interplay with the typologies of FLEs and BDA. The framework portrays the service contexts in which BD has major impact. Looking further into the future, the discussion raises prominent questions for the discipline.

Keywords Big data analytics (BDA), Frontline employee (FLE), Service typology, Frontline employee typology, Big data maturity level, Service adaptation, Service customization, Service delivery, Co-creation

1. Introduction

In service organisations, frontline employees (FLEs) are rightfully referred to as boundary spanners (Bettencourt and Brown, 2003; Bettencourt et al, 2001) because they represent the face of the provider organisation to the customers. The FLEs play a vital role not only in echoing the “voice of the firm” but also communicating up the managerial layers the “voice of the customer” (Hauser and Clausing, 1988; Bettencourt et al., 2005). The FLEs are also instrumental in adapting their service to suit the individual customer needs so as to enhance the customer’s service experience (Sony and Mekoth, 2014; Di Mascio, 2010; Lai et al., 2014). In high contact services, like financial, healthcare and airlines, FLEs need to deal with every other customer differently as the interactions are highly personal and variable in nature. Detailed information about customers and their path to service facilitate FLEs to adapt the service in an optimal fashion. Recent developments in information and communications technology (ICT) opened up multiple data channels (e.g. internet, peer networks, social media, twitter etc.). These channels capture not only transactional data but also a host of associated environmental factors (internet searches,
click stream, Facebook chats, Twitter exchanges, peer networking sites, etc.) in a granular mode. This data explosion (Demirkan and Delen, 2013; Wielki, 2015a), coupled with the Internet of Things (IOT), is exponentially multiplying data, leading to the phenomenon called big data (BD). Big data analytics (BDA) is the capture of data and derivation of insights that act as decisional aids. In the context of services marketing, BDA may imply derivation of insights about customers' preferences and market conditions, which facilitate FLEs to adapt a service to suit the individual customer (Ostrom et al., 2015; Rust and Huang, 2014). For example, for a financial consultant to have a prior idea (Wilder et al., 2014) about a customer beyond transactional data (Cambra-Fierro et al., 2014) will be of excellent value to kick start the discussion with the customer and propose an effective solution; a retailer may well offer better discount to a loyal customer; and an airline may propose a flexible itinerary for a frequent traveller. Aply, Ostrom et al. (2015) have identified BD as one of the 12-key research priorities for services marketing.

In a way, BDA re-enables the firms to strike a balance between standardisation vs customisation, namely, effectiveness over efficiency, and deliver the service in a much more personalised manner than hitherto possible (Kiron and Shockley, 2011; Vargo and Lusch, 2004). Demirkan and Delen (2013) highlight that in order for BD consisting of customer's online searches or his/her association in discussion groups and other exchanges in social media to be useful, organisations need to overcome the challenges of extracting actionable insights. Axiomatically, BDA can help FLEs in service delivery (Kiron and Shockley, 2011; Kiron et al., 2012), as it lessens service adaptation challenges (Di Mascio, 2010; Sony and Mekoth, 2014). These views about BDA (Dubey et al., 2015; Fosso Wamba et al., 2015; MGI, 2011) and its promise to FLEs (Kiron and Shockley, 2011; Kiron et al., 2012) enabled the authors to undertake a literature search to address the research question:

RQ1. How does BDA enable FLEs in effective service delivery?

This research aims to tackle the key challenges encountered by FLEs in BDA, such as how to: enhance service delivery (Ostrom et al., 2015; Rust and Huang, 2014; Lai et al., 2014); facilitate critical support for frontline to adopt and apply BDA (Brown et al., 2014; Barton and Court, 2012); improve frontline to enable mass customisation (Lai et al., 2014; Rust and Huang, 2014); and build a deeper and lasting relationships with the customers.

The rest of the paper is organised as follows: Section 2 outlines what BDA is; Section 3 details research approach, service typology, FLEs and models of FLEs; Section 4 draws insights from the literature, presents a schema to visualise the impact of BDA in services context and privacy and ethical issues of BD; Section 5 analyses the impact of BDA on FLEs; Section 6 discusses managerial, theoretical and practical implications of the research; and finally, Section 7 looks into the limitation of the present work, opportunities for future research and summarises the contributions of this research.

2. What is BDA?
Over the last few years, both practitioners and academics have placed enormous emphasis on BD. According to market analyst firm IDC, the BD and technology services market is growing at the rate of 27 per cent year on year and in 2017, it will reach US$32.4 billion (Wielki, 2015a). Mahr and Wetzel (2015) predict that investment in BD will explode to US $125 billion in 2015. Researchers attribute varying characteristics to BD, such as volume, velocity, variety, value and veracity. The amount of data captured, stored and processed is ever multiplying. In tune, the definition of BD itself is evolving, from 3Vs to 4Vs to 5Vs, as summarised below (Fosso Wamba et al., 2015):

- 5Vs: volume+velocity+variety+value+veracity (White, 2012);
- 4Vs: volume+velocity+variety+value (IDC, 2012; Oracle, 2012; Forrester, 2012);
3Vs: volume + velocity + variety (Gartner, 2012; Kwon and Sim, 2013; McAfee and Brynjolfsson, 2012).

Synthesising the BD discipline thus far, Fosso Wamba et al. (2015, p. 235) define BD “as a holistic approach to manage, process and analyse 5Vs (i.e. volume, variety, velocity, veracity and value) of data”. In defining BDA, Kiron et al. (2012) develop an integrated focus on statistical, contextual, quantitative, predictive, cognitive and other aspects of analytics in dealing with BD. This study defines BDA as the collection, analysis, visualisation, use and interpretation of data for various functional divisions with a view to gaining actionable insights, creating business value and establishing competitive advantage. The advancements in virtualisation, in-memory grid computation, large storage arrays and cheaper computational power (Demirkan and Delen, 2013) have enabled the crunching of this voluminous, heterogeneous, high velocity collection of data in real-time and generating actionable results to the end-users.

While leveraging of BD naturally commences at the top of the command in hierarchy, senior executives are deeply concerned about the little support at the frontline (Brown et al., 2014). Firms need to clearly identify their business drivers for customer relationship management objectives and then strive to cascade them down to the frontline effectively (Beaujean et al., 2006). Those objectives may also influence a firm’s service portfolio, i.e., the breadth and depth of service offerings. As the initial framework for this current research was in progress, the authors recognised that as services vary broadly from one context to another, the need for, as well as the impact of BDA may vary too. The notion of service typology surfaced after discussion with some services marketing scholars. The services marketing literature widely recognises the vast heterogeneity associated both with services and service delivery (Lai et al., 2014). BDA may be more useful, depending on the context. For example, in a financial advice context, a banker has more opportunity to engage with the customer than in a simple teller service. Owing to these, it is highly relevant to concisely view service typology in this discussion. Thus, the study schema is devised to incorporate service typology as one of the determinants in assessing the impact of BDA on FLEs. This also prompted the possible existence of different types of FLEs. As the personal behavioural traits of FLEs are unique in nature, it is also relevant to consider a typology of FLEs (Di Mascio, 2010; Bowen, 2016). In summary, the review aims to blend these relevant streams to draw critical insights to serve as a foundation for future endeavours.

To start with, Figure 1 presents the key focal elements of this research: FLEs, BDA, service typology and customer interactions. The diagram conveys that a firm’s customer relationship objectives are reflected in two ways: one, the firm’s service portfolio; and second, how FLEs are oriented towards customers. Analytic innovator firms (Kiron et al., 2012) cascade the...
insights of BDA to the FLEs so that they can better serve the customers. The customers too are becoming more insightful through discussion groups, peer networks and publicly shared information on internet (Kilcourse and Rosenblum, 2014). So, a FLE needs to deal with an informed customer, by blending firm objectives, service offerings, differing characteristics of a service (service typology), customer intelligence and customer needs enumerated through the service encounters (Joseph, 1996).

3. Research approach

The study is motivated by the impact of BDA on FLEs (Fosso Wamba et al., 2015; Shibata and Kurachi, 2015). In high-interaction-oriented service contexts, the FLEs have a significant challenge in facing the customers who are increasingly aware of the market through the multiple information channels (Demirkan and Delen, 2013; Wielki, 2015a). Analogous to typical ICT adoption by users (Delone and Mclean, 2003; Venkatesh et al., 2011), cascading BD insights to the frontline, and ensuring FLEs adopt these emerging tools and technologies effectively is a major challenge (Brown et al., 2014). This generic desire to understand how well the new breed of analytical tools enables FLEs in service delivery, has prompted the authors to initiate a literature search. The literature stream pertaining to BD is multiplying every day and a typical search string “big data” gets over 11,000 hits in Scopus and 149,000 hits in Google Scholar. In comparison, a search for “FLEs” in Scopus produces over a 1,000 hits and in Google Scholar presents over 9,640 results. A systematic approach is needed to scour this voluminous section of articles and identify relevant literature to address our research question.

As the research domain comprises the intersection of multiple literature streams of BD and FLEs, it is logical to look for FLEs within BD results or vice versa to narrow down the search to a restricted set of articles. The different search strings and the resultant hits are summarised in Appendix 1. As noted above, both BD and FLEs are important areas of research within services marketing literature. However, it is evident from the search results listed in Appendix 1 that the literature seems to be very sparse with respect to BD in frontline enablement. The results may not be coincidental when it is reckoned with the senior executive concerns about little support at the frontline from BD (Brown et al., 2014) and the formidable challenges of extracting actionable insights from BD (Demirkan and Delen, 2013).

3.1 Service typology

It is well known that services are generally complex to define and have many dimensions that differentiate one service from the other. A particular service like financial advice delivered to a customer by a financial professional is both time and resource intensive in comparison to withdrawing money from a teller counter. While the former one requires collation and presentation of data in a manner that is relevant to the customer and may require long customer contact time and keen participation of the customer, the withdrawal of money is a less time intensive and requires minimal inputs from the customer. Thus, one can extrapolate that the need of analytics is not a one size fits all (Lai et al., 2014) situation from a service delivery point of view. From the perspective of FLEs, it is further possible to assert that the role and necessity of analytics differ from one service type to another. To differentiate better the service contexts, a service typology is necessary and is developed as shown in Figure 2.

While preparing this typology, the seminal ideas of Schmenner (1986) and other studies (Verma and Boyer, 2000; Glückler and Hammer, 2011; Mills and Margulies, 1980/1986; Bullinger et al., 2003; Prajogo, 2006) are considered. In his paper, Schmenner has used labour intensity and interaction and customisation dimensions to arrive at a 2 × 2 service process matrix, forming in all, four service categories. Schmenner termed those services that are low in interaction/labour and customisation as service factory. Highly labour intensive, low interaction and low customisation varieties are termed as mass service.
The services that are low in interaction and customisation are referred to as service shops. And finally, at the extreme end of the spectrum where both interaction and customisation are high, are referred to as professional services.

Synthesising the service categorisations of Schmenner (1986) the authors enumerated the dimensions along which a service can be categorised. Figure 2a and b portray the identified 13 dimensions for categorising services and they are: channel, customer contact, needs, customisation, service script, relationship, customer input, customer involvement, knowledge intensity, labour intensity, technology intensity, delivery medium and pricing. The advent of BD and multi-dimensional visualisation techniques help in relatively positioning various services along the dimensional space characterised by the 13 parameters.

The first dimension, channel, differentiates a service whether it is a B2B or B2C. The second dimension, customer contact, categorises whether a service involves low or high customer contact. The third dimension, needs, highlights whether the requirements for service are a priori or need to be gathered during the interaction by the FLE. Customisation differentiates whether a service is pre-defined or needs to assemble prior defined service elements. The next dimension, service script, essentially indicates whether the FLE orchestrates routine delivery script or whether the FLE has the flexibility to interact with the customer as s/he tries to understand the customer needs. The relationship dimension helps to define whether a provider treats a service as a mere transaction or wants to build a long-term relationship with the customer. This could as well be interpreted from the customer perspective too in a similar manner. Basing on Schmenner’s (1986) work, Danaher et al. (2008) segment customers to identify which customers normally like a relationship with the provider.
The customer input dimension differentiates services based on how much information the customer is expected to share with the provider prior to the delivery of a service. The customer involvement dimension differentiates services based on the customer’s active involvement, low or high, during service consumption. Knowledge intensity is an important dimension, which highlights whether a service is routine or requires significant professional knowledge from the provider, like a lawyer, engineer or doctor. Similarly, labour intensity signifies how laborious a service is for an FLE to deliver. Following the lines of knowledge and labour, technology intensity signifies how dependent a FLE is on technology to deliver the service. The delivery medium, i.e., face-to-face, phone, over internet chat, or asynchronous modes like e-mail, is also an important dimension to be kept in mind. The last dimension, pricing, indirectly indicates how much a FLE is empowered to dynamically alter the pricing for a particular customer in context.

3.2 FLEs

The divergence in the variety of services delivered in the marketplace is a contributing factor to the diversity of names associated with boundary spanners. For example, FLEs are also referred to as frontline staff, frontline personnel, frontline service employees, frontline service staff, customer support, customer support staff, customer service executives, help desk operators, receptionists, service desk employees, service desk consultants and in many more different ways, depending on the country, firm and industry. Our view of FLEs is consistent with those of Zeithaml et al. (2012), who define FLEs as those at the boundary of a firm, interacting with the customers on behalf of representing the firm. Firms typically maintain several touchpoints as a form of interaction with their customers. FLEs acting as sensors at several touchpoints, naturally acquire different names as noted above. In the context of this current study, FLEs are treated synonymous with all the titles listed above or implied thereof. And in essence, a FLE interacts with a customer at one or several touchpoints.

3.3 Typology of FLEs

The above operational definition of FLEs fits a great number of roles. However, to better understand the generic group of FLEs, it is necessary to categorise them. Di Mascio (2010) premises that each FLE is unique in terms of how they deliver their service. This might mean there could be as many service models as there are FLEs. Through an empirical enumeration, Di Mascio (2010) concludes that three distinct service models exist among retail FLEs, namely, efficiency, efficacy, means and win-win (Di Mascio, 2010 p. 67). First, efficiency means the act of giving customers what they ask for, efficiently and courteously; second, efficacy means the act of giving customers what they ask for, efficiently and courteously; second, means conveys a method to accomplishing immediate objectives, such as sales quotas; and finally, win-win reflects the formation of mutually beneficial relationships with customers through problem solving. In contrast to this classification, based on the creative discretions of FLEs during service encounters, Kelley et al. (1996) categorise FLEs into three categories as: creative, routine and deviant discretion.

Coming from a different end of analysis, Bowen (2016) explicitly addresses the evolving position of FLEs, suggests four roles, namely: innovators: (technology cannot substitute for human creativity as the source of new ideas for services and their delivery); differentiators: (the non-substitutable personal touch avoids the commoditization of service); enablers: (employees, including and understudied back office, ensuring that both customers and technology are able to perform their own roles in coproduction and value creation, overall); and coordinators: (integration of resources and collaboration across multiple actors in the service system).

4. Visualising the impact of BDA in services context

Scholarly research on BDA can be classified into five themes (Fosso Wamba et al., 2015; Dubey et al., 2015; Ji-fan Ren et al., 2016): decision making and performance improvement,
needs identification, creating infrastructure and transparency, new product/business model innovation (Cadwallader et al., 2010; Santos-Vijande et al., 2015) and market segmentation (Fosso Wamba et al., 2015; Shibata and Kurachi, 2015). Fosso Wamba et al. (2015) also observe that a third of the 132 publications they reviewed focus on decision making and performance improvement, revealing the greater interest placed by scholars. The extant literature is predominantly centred on return on investment or monetary value of BDA to the firms rather than on the firm’s creation of value with BDA. There exists a wider gap between the importance and the knowledge about BDA (Ostrom et al., 2015), and between the importance and frontline application (Brown et al., 2014; Beaujean et al., 2006). Thus, there is significant opportunity to study how a firm’s frontline leverages BDA in adapting service to meet the individual needs of the customers. One of the important insight that emerged from the literature review is that firms need to align their BD programs to strategic goals like improved service delivery and value creation (Wilder et al., 2014; Lavalle, 2009). The strategic decision making and successful cascading of a defined strategy to the frontline determines the success/failure of a firm as it reflects on the firm’s service offerings to capture a profitable segment of the market. BD of a typical firm may consist of several streams of data elements, and thus, its analysis and impact to the firm also has many dimensions. For the success of BDA, firms also require multi-pronged tools like statistical, contextual, quantitative, predictive, cognitive and other models (Kiron et al., 2012). There has been considerable focus on how to go about analysing the BD or understanding the value/utility of BD (Fosso Wamba et al., 2015). However, there has been lesser emphasis on assessing how BDA enables FLEs and its value creation potential.

Like any other ICT deployments, BD programs cannot happen overnight. Firms need to go through several iterations, before fruitful outcomes are derived and integrated into their service delivery. The starting point to the BD journey may be unique to each firm, but as they progress, they need to develop their programs with the end in mind (Schmarzo, 2012). The storage vendor, EMC provides an interesting framework for the firms to self-evaluate their level of maturity in terms of BD phenomenon as shown in Table I (Schmarzo, 2013; Schmarzo, 2012). In contrast to this classification, Kiron et al. (2012) have classified organisations into five distinct levels based on the competitive advantage with analytics and using analytics to innovate. Kiron et al. (2012) referred organisations that are least effective at creating competitive advantage and driving innovation as Level 1, and those organisations that exhibit mastery in the use of analytics both in gaining competitive advantage and for innovation as Level 5. These Level 5 organisations are termed as analytical innovators.

The value of BD comes from the insights it provides (or services it renders) to both service providers and consumers alike (Demirkan and Delen, 2013). The level of insight derived by individual users is not only a function of the BDA climate, but also depends on

<table>
<thead>
<tr>
<th>Business monitoring</th>
<th>Business insights</th>
<th>Business optimisation</th>
<th>Data monetisation</th>
<th>Business metamorphosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring existing business performance using BI to identify under- and over-performing business areas</td>
<td>Users statistics, predictive analytics, and data mining to integrate insights into existing business processes</td>
<td>Embedding advanced analytics to automatically optimise certain business operations</td>
<td>Net new revenue opportunities (1) selling data with analytics (2) creating “intelligent” products (3) transforming customer relationships</td>
<td>Transform insights about customers, products, and market trends to create new services and/or new markets</td>
</tr>
</tbody>
</table>

**Source:** Schmarzo (2012)
the structural empowerment of the firm user (Wilder et al., 2014; Bowen and Lawler III, 1992). The authors posit that the users need to co-create value through their interactions with the information system (IS), and possibly with peers, before those insights contribute to improved processing abilities or decision capabilities (Beaujean et al., 2006). While social media platforms like Twitter or Facebook are serving the hedonic needs of consumers, they too are holding valuable information that is of interest to services marketing and FLE’s service delivery (Cambra-Fierro et al., 2014). Thus, there has been a surge in both industry and academia to harvest BD to gain a competitive advantage.

Making sense of BD and visualising connectedness of the underlying business processes through the vast pools of data is a non-trivial task, and goes beyond the mechanical emphasis of tools (Demirkan and Delen, 2013). Mining of data for knowledge extraction has earlier been a dominant concern to artificial intelligence research (Motamarri, 2014). Value derivation from BD is essentially tacit and hermeneutic in nature. While tools of BD are important functional aids, the extraction of knowledge/insights (Kiron, 2015; Chang et al., 2014) is far more complex and hitherto less explored especially from the individual FLE’s perspective. Such an exploration contributes new knowledge and understanding of the imperatives of BDA. Furthermore, the exercise can transform BDA as a real value contributor to a firm’s sustainable operation.

To map the transient linkages between BDA and FLEs, it is important to answer three distinct questions, namely: what business challenges are driving business intelligence/business analytics? What are the important uses of the data within a firm context? And how much of the various functional departments of a firm are data driven? The literature search has identified a few interesting publications dealing with some of these aspects (Wielki, 2015a; Kilcouse and Rosenblum, 2014; Mckinsey, 2011; NewVantage Partners, 2012; NewVantage Partners, 2014). In all, these works postulate 23 aspects of business drivers/impacts of BD. Consolidation and summarisation of these drivers emerged to 12 themes as portrayed in Figure 3. The spanning blue boxes reflect the themes emerged out of consolidating business drivers derived from these publications. The labels refer to the research publication as noted in the legend.

From Figure 3, it is apparent that firms are still grappling with the emerging phenomena of BD and that the business drivers or impact of BD varies from firm to firm. The most prominent themes are: business/customer insights; intelligent application of resources and efficient operations; and followed by faster answers to business questions and creation of transparency and data-driven culture for decision making. As noted by Kiron et al. (2012), Ransbotham et al. (2015) while BD has enormous potential to shift organisational power structure to the frontline, a good number of firms are yet to recognise these benefits and equip and empower their FLEs with BDA tools.

![Figure 3. Primary business drivers/important uses of big data](image-url)

**Notes:** A, NewVantage Partners (2012); B, NewVantage Partners (2014); C, Kilcouse and Resenblum (2014); D, McKinsey (2011)
4.1 The interplay of service typology, FLE typology and BDA maturity

The previous sections have proposed a service typology, FLE typology and BDA maturity. These typologies play a vital role in determining the extent to which BDA can impact a service firm. The combination of a service and FLE typologies with varying degrees of BDA maturity create myriad service delivery contexts. Some combinations may have highest possible outcomes both for the provider and the customer, while other contexts may not work out to be differentiators in terms of service outcomes. Figure 4 illustrates such an attempt to depict the most effective combination where these three typologies intersect and create the highest level of opportunity for the BDA. Using a technique called “parallel coordinates” for multi-dimensional visualisation (essentially, BD is multi-dimensional), Figure 4 portrays this most rewarding combination. For the sake of clarity, the diagram does not include all the service dimensions. Furthermore, for clarity, the diagram does not indicate all the permutations of connections.

Figure 4, for example, conveys that in the service contexts, where knowledge intensity and customer contact are high, customisation is variable and relationship leans towards collaboration; and FLEs belong to innovators and where BDA in a firm aims for metamorphosis (or Level 5), the resultant value to the firm and customer is exceptional. In parallel coordinates, the connection line thickness is an indicator of the strength of the relationship. For example, four links are culminating at innovators, and thus the link between innovators and business metamorphosis is depicted as four times thicker. The authors hypothesise that in order to assess the impact of BDA in service contexts, the typologies of service, FLE and BDA play significant role. Furthermore, firms attempting to leverage BDA need to enhance empowerment and analytic climate (Wilder et al., 2014; Schmarzo, 2013; Mckinsey, 2013; Beaujean et al., 2006).

4.2 BD: paradoxes, privacy and ethical issues

It is also important to recognise some of the downsides of BD revolution. One of the major concerns to BD is privacy of individuals (Brown et al., 2014; Wielki, 2015a; Bollier, 2010) and
ethical aspects (Richards and King, 2013, 2014; Davis and Patterson, 2012). Researchers like Bollier and Brown et al. (2014) caution that while BD can benefit consumers, firms and governments with benefits such as improved healthcare outcomes, new services, that reflect the consumer preferences and result in valuable digital experience, there are some ethical concerns. Apart from privacy, BD also cuts into the individual’s safety to his/her online identity and raises a multitude of ethical questions on who owns the data, who controls it and ultimately whether individuals have any say about their personal information (Brown et al., 2014; Bollier, 2010). Another subtle aspect with technology including BDA is that while it assists service delivery it can also manifest as a barrier during service encounters (Giebelhausen et al., 2014).

The aforementioned discussion drives home that in the current business environment, BD is ubiquitous and its potential for value is very much dependent on an organisation’s harvesting abilities (EIU, TEIU, 2015; MGI, 2011; LCIA, 2011). Benefits apart, there are some formidable ethical and legal challenges with the use of BD. While the collection of BD per se is ethically neutral, the usage is not (Davis and Patterson, 2012; Wielki, 2015b). Fortunately, things are changing with the acceleration of BD, and corporates and end-users started recognising the need for a better data governance model. Data are now recognised as a critical organisational asset as international bodies like World Economic Forum describe “personal data” as “new economic asset class” (Davis and Patterson, 2012; EIU, TEIU, 2015). That being the case, the importance of securing data asset is also increasingly becoming a necessity. Organisations while preparing their BD strategy need to duly recognise the risks associated with the ongoing collection, securing and usage of BD so that they do not end up in ethical and costly legal battles (Mateosian, 2013).

Coming from a legal perspective, Richards and King (2013) argue that the evangelists’ claims of the benefits of data-driven culture from the minutest tasks of human endeavour to that of medical care are might become true in future but they are anecdotal at the moment. They highlight three paradoxes of BD and caution the public of their consequences. In the view Richards and King (2013) the first one is transparency paradox, for BD consists of pervasive collection of private information and its operation trespasses into legal and commercial secrecy. The second one is identity paradox as BD promises miraculous outcomes at the expense of individual and collective identity. And the third one is power paradox as BD claims power for the society to transform while in reality only governments and large corporations have the muscle at the expense of ordinary individuals. Richards and King (2013) advise that adopters of BD may gain true value only when they examine these paradoxes in their service contexts and address the issues adequately.

Davis and Patterson (2012) infer that BD’s volume, variety and velocity is creating a “forcing function”, whereby business operations are deeply pushed into the personal lives of people and is altering the common meanings and value implications. Even if the BD removes identification data (de-identify), it is said the simple characteristics of gender, birthdate and zip code allow identification of a person with 87 per cent of certainty. Thus, Davis and Patterson (2012) warn that though BD creates enormous opportunity for innovation, deeper insights, broader outlooks and better customer engagement, one cannot simply ignore the associated risks, like: security, privacy, legal compliance and customer engagement. Creating meaningful vocabulary and extending these realisations, Davis and Patterson (2012) formulate the four aspects associated with BD ethics as: identify, privacy, ownership and reputation. They also identify certain ethical incoherence within the value framework of the Fortune 50 companies itself, and suggests that it is time to reconcile and create a coherent policy. They further propose a continuous cycle for ethical decision points: inquiry, understanding, articulation and action. Davis and Patterson (2012) in conclusion urges organisations to adopt transparent BD governance policies to clearly address the questions whether the current practices align with their own core values and
customer’s values. Inspired by these valuable arguments, the authors urge academic scholars, organisations and practitioners not to undervalue the associated risk factors of BD in advocating BD for economic leverage. Mateosian (2013, p. 61) warns that “[b]ig data is coming like a runaway freight train, and we must methodically examine the issues and take appropriate actions, or we won’t like the mess it makes when it crashes”.

5. Impact of BDA on FLEs: challenges and opportunities
Prior to analysing how BD may have a positive impact on the roles of FLEs, it is worthwhile to look into the demands of FLEs succumb in facing customers and delivering the marketing promises. Zeithaml et al. (2012) argue that irrespective of the skill or pay, the boundary spanning roles cause high-stress due to the demands on emotional labour and mental and physical needs. FLEs are frequently confronted with irate customers and need to deal with both intra-organisational as well as inter-organisational or organisation vs customer conflicts (Browning, 2008; Gruber, 2011; Kashif and Zarkada, 2015; Reynolds and Harris, 2006). FLE’s job performance may also be impacted and become stressful due to organisational factors, training, inadequate resources, lack of recognition, pressure to exceed targets and inadequate tools to perform the job meeting the quality expectations (Oh et al., 2014; Devi and Sharma, 2013; Dean and Rainnie, 2009). FLEs may need to react to situations in real-time and require to trade-off between quality and productivity of the work. Zeithaml et al. (2012) further reflect that these stresses may contribute to service delivery failures and thus contribute to performance gap.

FLEs are not always be able to accommodate and adjust the system to meet the requirements of internal as well as external customers alike, causing dissatisfaction at both ends (Zeithaml et al., 2012). Sony and Mekoth (2012) observe that FLEs need to adopt to situational factors and actively draw information from situational cues and adapt their service on the basis of the information drawn. In such situations, the better the FLEs can gauge the customers’ needs and behavioural profiles, the better they can serve them.

5.1 BDA: business challenges to service industry
Retail Systems Research (RSR) organisation has conducted extensive study with its constituent retail industry members on BD. The survey identified the significant business challenges for BD adoption. The top five drivers identified are (Kilcourse and Rosenblum, 2014, p. 6): one, consumers expect instantaneous access to information; second, consumers “path to purchase” (a dynamic, consumer journey that moves from customer’s initial awareness to pre-purchase, purchase and advocacy stages in a multi-channel environment); three, information – empowered customers are demanding more; four, firms need to react to sudden changes in customer demands; and five, the necessity for constructing alternative scenarios to cope up with market dynamism.

RSR researchers have also identified that the retail winners are leading in the first four items over the laggards. However, for item five, the laggards have a higher score. With this they conclude: “[r]etail winners understand that consumers are demanding, expect full access to information about products and services wherever they may be, and that these same consumers can be fickle, rapid and radical changes in trends and demands are expected and lag-time to action must be decreased. Laggards on the other hand are perennially focussed on their competitors. […] Winners look within or to the customers, laggards look over their shoulders” (Kilcourse and Rosenblum, 2014, pp. 6-7).

From these assessments, it is apparent that FLEs constantly face more informed customers who demand a quicker and optimal solution tailored to them. While some of the successful companies are leveraging BD, there are far more firms that are yet to recognise the value of customer relationship over imitating the competition (Kiron et al., 2012). Reinforcing these ideas, Harvard Business Analytics Services team reflects that frontline managers do not get
adequate representation and they are not equipped with the right BD tools to enable them serve better \((Harvard \ Business \ Review (HBR), 2014)\). The HBR study asserts that while organisations recognise frontline managers as linchpin for organisational success, that recognition however, does not translate into giving them the requisite resources, adversely affecting firm performance. At again, the firm-customer relationship must happen through the frontline as sometimes FLEs are themselves equated to the service they deliver \((Zeithaml \ et \ al., 2012)\). So, it is imperative for the firms to percolate/grant access to the strategic customer insights from BD, so that FLEs serve the customers in an effective manner.

In high contact services, standardisation may not work as each customer context is different \((Lai \ et \ al., 2014)\). Though, technology enhances service delivery, it can also become a barrier for rapport building, meaning BD tools shall not hinder the creative process of the FLEs on one hand and the comfort of the customers on the other hand \((Giebelhausen \ et \ al., 2014)\). FLEs need to innovate during the service encounter, so they better serve the customer \((Lai \ et \ al., 2014; Moosa \ and \ Panurach, 2008)\). Moosa and Panurach also contend that centralised innovation is not only insufficient but also ineffective. Being at the forefront of service delivery, FLEs have much more experience, but their voice is not accounted for in a firm’s strategic roadmap \((Moosa \ and \ Panurach, 2008)\). FLEs need tools and better information. BDA has a great potential to serve such innovation needs of FLEs by virtue of providing deeper insights on what the customer actually needs and his/her set of preferences \((Cambra-Fierro \ et \ al., 2014)\). Building on these insights, the next sub-section enumerates how BDA can be an enabler to services marketing.

5.2 BDA: an enabler for services marketing

The following enumeration will look into how some service marketing imperatives can be addressed by BDA. The challenge however is to strike a balance between customers’ desire for privacy and personalised services. BDA has a big promise; however, there are formidable implementation challenges. Shibata and Kurachi \((2015)\) categorise them as: difficulty of introducing new technologies; difficulty of defining system requirements; and difficulty of estimating effects of implementation. Service firms need to be aware of these issues and make a systematic contingency plan prior to embarking on BD programs. The following scenarios highlight the ways in which BDA helps in dealing with changing service situations:

1. Dynamically manage customer value over time: BDA provides firm-customer relationship, customer satisfaction information, as well as customer pain points with the firm and firm’s service offerings. Service providers are able to link customer complaints, word of mouth campaign (blogs, peer networks, social media exchanges etc.) so that the FLEs are aware of the customer concerns. This enables them to empathise with the customer situation and offer amenable solutions \((Wilder \ et \ al., 2014)\) to increase customer value over time.

2. Improve the customer experience and customer-firm relationships: FLEs are notified through BDA the kind of products/services the customer is looking for and customer’s past purchasing history. Shibata and Kurachi \((2015, p. 39)\) describes a case where “store management increased sales due to optimising the floor layout and enhancing customer interaction through customer behaviour analysis”.

3. Personalise service dynamically and in real-time: FLEs ability to personalise the product/service to make them develop brand loyalty. The authors came across that some of the financial institutions are pre-alerting their planners the internet searches and click stream data associated with a prospective customer. The information enables the planner to be aware of what the customer had already viewed and gauge what the customer may be looking for.
(4) Uncover opportunities for service innovation and create new service offerings: FLEs ability to assess strategic gaps between existing service offerings and customer needs, and dynamically ensemble service elements to customise the service to the customer. As noted in the beginning of this paper FLEs are also critical in reflecting the voice of the customer (Hauser and Clausing, 1988) to the strategic planning teams, enabling them to devise new service offerings.

(5) Assist in real-time decision making: up to date information about customer’s online searches for specific services; customer’s feedback on services consumed. Similar to points 2 and 8. For example, a hotel amending the breakfast choices to guests based on the feedback received.

(6) Customise: customising a service offering as per the needs of the customer. Shibata and Kurachi (2015, p. 35) describe that “[…] this company has been able to increase sales by switching product promotion to low-calorie drinks popular among women on the day of a concert that a large number of women are expected to attend. This is a case study of a frontline department that has achieved micromarketing and expanded business by introducing a big data analysis environment”.

(7) Undertake predictive modelling: ability to forecast future purchases of the customer. Shibata and Kurachi (2015, p. 35) describe that “sales staff at a beverage manufacturer can now perform detailed promotions at each sales outlet in their area through demand forecasting that takes into account not just detailed sales data from those retail stores but also external data such as regional weather and events”.

(8) Deliver dynamic pricing: ability of FLEs to visualise customer’s historical spending, competitive choices and fine tuning pricing. The authors came across solutions based on BD presentation tools like Tableau targeted to help sales personnel to offer special discounts for repeat customers as well as to win a share of the customer wallet by not letting them switch to a competitor.

(9) Segmentation: categorising customers into groups that enable FLEs to visualise best possible solution for a customer. Shibata and Kurachi (2015) describe that customer behaviour analysis model combines conventional purchasing data with external information from FLEs and social networks.

(10) Campaign management: advertisement campaign management based on customer’s preferences about location, medium, timing and gender. Shibata and Kurachi (2015) describe that while conventional mass marketing is based on wholesale shipping data, BDA solution supports micromarketing approach in accordance to demand forecasting.

(11) Supply chain or operational effectiveness: Shibata and Kurachi (2015, p. 39) describe that “optimised inventory due to increasing the accuracy of demand forecasting (reduce out-of-stock losses and decrease waste)”.

(12) Manage service design: continual customer-firm interactions across the touchpoints that help to move away from static design of service to one that is evolving, iterative and personalised (Erickson, 2009; Ostrom et al., 2015; Rust and Huang, 2014).

6. Discussion

6.1 Managerial implications
Rust and Huang (2014) conclude that information revolution and service revolution are two sides of the same coin. Interestingly, analysing the socio-technical systems and service-dominant logic (SDL), Motamarri (2015) comes to a similar conclusion. Ostrom et al. (2015, p. 127) basing
on the propositions of Rust and Huang (2014) summarise that: “[…] rapidly evolving information technologies (e.g. IOT, social network technology, mobile technology and cloud computing) enable ubiquitous customer communication and the acquisition, storage and analysis of BD, presenting opportunities for more personalised, higher quality service, and deeper customer relationships”. In their detailed assessment on service research priorities in the rapidly changing context, Ostrom et al. (2015) stress that BD has a significant role in advancing service. In essence, BD to advance service is one of the 12 research priorities identified by them. Considering the massive drive towards BD in the industry (Fosso Wamba et al., 2015; Brown et al., 2014), it is quite natural for the service research to focus on BD so as to be relevant both for theory and practice.

Ostrom et al. (2015) summarise that the theme using BD to advance service, has the widest gap between the importance and the knowledge ratings. This implies a fertile service research opportunity and calls for building new knowledge pertaining to this domain. Ostrom et al. (2015, p. 137) have identified seven subtopics, out of which four of them centre around BDA and FLEs as noted below: one, using BD to dynamically manage customer value over time; two, understanding conflicts between customer’s desire for privacy and their desire for personalised service; three, using BD to improve customer experience and customer-firm relationships; and four, developing analytic and recommendation models for dynamic and real-time service personalisation.

The McKinsey team, Brown et al. (2014), probed a group of frontline executives responsible for data-analytics revolution across different industries. They conclude that BDA is delivering value at strategic level, but not to the FLEs. Decision makers perceive that making the frontline adopt and use BDA tools effectively as a major challenge (Brown et al., 2014). The HBR (2014) team asserts that lack of adequate training, resources and BD tools for frontline managers is hindering firm performance. To bridge this implementation gap, the Harvard team stresses that senior managers need to eliminate barriers that hinder progress as well as impact performance of frontline. This is an important input to managers as it demonstrates time and again that recognition of the importance of frontline is not adequate; firms need to move forward and train and equip the frontline so that they can serve the firms in a better way.

It is worth recalling another major research theme, enhancing the service experience, illustrated by Ostrom et al. (2015). They have deduced seven major issues. BDA has the ability to address all these issues. It enables the FLEs to gauge customer service experience attributes inferred through the service encounters across the touchpoints and channels. FLEs will be able to dynamically structure service delivery to match the expectations of the customers. Rust and Huang (2014) contend that mass customisation of service by FLEs is feasible as BD provides granular customer information which facilitates segmentation of customers as well as helps to fine tune the service. Similar to the notion of standardised service may not fit all customers; standardised training in BDA may not be sufficient to all FLEs. The managers need to focus on the service contexts at the interplay of service typology and FLE typology (Di Mascio, 2010) to design customised training programs to equip the FLEs to achieve firm objectives.

6.2 Implications to theory and practice

This study provides the foundation for extending the SDL which views that technological progression results in greater skill-differentiation (Vargo and Lusch, 2004). The skill asymmetry between the provider and customer is the fundamental reason for the service economy. It transiently conveys a retrograde from service standardisation (an aftermath of industrial revolution) to service differentiation. Though Vargo and Lusch (2004, 2008) have not directly indicated it, the skill asymmetry also implies that service firms need to enhance their knowledge and skill-sets of their resources (put together, operant resources)
continually, in order to compete for their pie in the market. As service contexts become dynamic and the customers take greater role in value creation, it is imperative that firms leverage BDA to ensure that their FLEs adapt services and serve customers well. An empowered FLE with due access to the insights about the prospective customers, can potentially adapt the service delivery with a priori information. In essence, the current research focus on FLEs asserts that there is tremendous opportunity to understand and value the contributions of the frontline. Furthermore, FLEs service adaptability enables to assess not only their value creation through service encounters, but also the generic enhancement of the firm’s operant resources, thus competitiveness.

The findings of our study are also relevant to service adaptation theory. For example, Wilder et al. (2014) discuss that customers are expecting individual customisation in contrast to conventional efficiency drives. Wilder et al. further hypothesise that a FLE’s capacity to anticipate customers’ feelings, motives and concerns shall aid in service customisation and thus enhance customer experience. As eventually service is delivered by the FLEs, firms have an implicit challenge to strive for skill asymmetry between their frontline and customers. Firms can achieve this by keeping their frontline abreast of the information that enables them to serve the customers better. Thus, BDA has a significant role in preparing FLEs prior to/ during the service encounters with significant information about the customers they are servicing. The authors imply that the skill asymmetry is the counter force for the customisation demands of the customers. Thus, organisations setting aside the personality traits of FLEs, can promote adaptability capacities of FLEs (Wilder et al., 2014) by providing them training as well as granting access to insightful information (Fang et al., 2014). The insights help FLEs in two ways: first, identification of customisation opportunities for the specific customer they are servicing; and second, determination of how they can adapt the service to suit to this customer (Wilder et al., 2014).

Through their empirical research, Wilder et al. (2014) validated that while empathy and anticipation helps in customisation opportunity, FLEs creativity demonstrates their actual ability to adapt a service. This argument is also supported by Bowen and Lawler (1992). While discussing empowerment of FLEs, they address the basic questions of empowerment, like: what, why, how and when. In essence, a firm’s service climate and empowerment found to be positively related to the FLEs ability to adapt service to the needs of the customers (Bowen and Lawler, 1992; Wilder et al., 2014). In addition to SDL and service adaptation theories, our research provides necessary platform for advancing market orientation theory. For example, firms that concentrate on market orientation structure their activities to create superior value for their customers (Fang et al., 2014). Market orientation can be understood in two parts: internal market orientation (IMO) and external marketing capability (EMC) (Fang et al., 2014). Fang et al. (2014, p. 172) refer IMO as “the process of generating and disseminating intelligence about internal market needs and then responding to and satisfying these needs”. Market-oriented organisations concentrate on market sensing and customer linking to signify their EMC. Fang et al. (2014, p. 173) define market sensing as “an organisation’s capability of learning market knowledge and utilising this knowledge in forecasting the future market development”. Fang et al. (2014, p. 173) further define customer linking capability as “an organisation’s capability of effectively establishing and maintaining appropriate customer relationships after the organisation identifies and understands its target customer’s needs”. BDA by virtue of capturing and synthesising granular information about customers they significantly contribute to EMC of a firm.

From the relational paradigm perspective, continual firm-to-customer interactions and customer-to-customer co-learning affect non-transactional behaviour of customers (Cambra-Fierro et al., 2014). Some of these interactions are voluntary, and may not translate to immediate sales, but in the long run, improves company image and future purchase decisions of consumers (Cambra-Fierro et al., 2014). BDA provides an opportunity to
capture, process and present such interactional information from social media to frontline, so that they can fine-tune delivery patterns to influence the customer engagement.

Following these observations, the authors deduce that BDA climate, empowerment of the FLEs in combination with the FLEs skills of market sensing, and customer linking abilities will positively influence service delivery. As firms derive market/customer insights from BDA (Figure 3), it is imperative that firms focus on creating a transparent data-driven culture, be innovators in BDA, and empower the FLEs with right information and tools, so that FLEs will succeed in delivering the marketing promise of the firms.

From the perspective of practice, BDA may yield greater benefits in high contact service contexts. A win-win style of FLEs over other types (Di Mascio, 2010), may leverage BDA to deliver value both to the customers and firms. Managers also need to recognise that a firm’s BDA maturity and strategy (Schmarzo, 2012, 2013; Kiron and Shockley, 2011) will have significant impact on the frontline so that greater value is derived from each service encounter. Managers need to focus on training FLEs with market orientation and customer linking and improve the BDA climate to generate greater benefits for the firm, the customer and the FLEs (Wilder et al., 2014).

7. Conclusions and future agenda

7.1 Limitations
The aim of this research is to understand how well firms are leveraging BDA in enabling the frontline to serve customers better. As an auxiliary, the research also intended to uncover how BDA enables FLEs in service customisation and adaptation to suit individual customers. However, the search within the extant literature has not pointed to any relevant scholarly work, except the tools being developed and marketed by Fujitsu (Shibata and Kurachi, 2015).

This does not seem to be surprising as evinced from the panel discussion (consisting of experts both from academia and industry) on BDA’s relevance to marketing at the Australia New Zealand Marketing Academy Conference (ANZMAC2015) held on 30 November 2015 in Sydney. The panel is of the opinion that there is tremendous opportunity for marketing to leverage BD but the progress seems to be rather slow. It is also opined that the chances of new product/service development out of BD is very difficult for the fact that human innovative spirit cannot easily be mimicked by sophisticated ISs, i.e., Artificial Intelligence tools (Demirkan and Delen, 2013).

7.2 Future research agenda
The authors are of the opinion that the service firms may be at different stages along the BD adoption cycle (Table I), and it is plausible that firms may be placing significant focus on building the infrastructure rather than deeply looking into the programme to ensure that the investments have a direct impact on the frontline (Mckinsey, 2011; Kilcourse and Rosenblum, 2014). Similar opinion is expressed by the panel of experts at ANZMAC2015.

The BDA is touted to revolutionise the business world. The extant literature too is predominantly anecdotal. Given this context, the authors are confronted with the following broad questions to drive the future research agenda: what specific insights are derived from BD by firms about customer preferences and service delivery? How do firms utilise those insights? (e.g. improving service/service portfolio, what information is cascaded to frontline so that they can serve better; and does frontline staff distinguish cascaded information as BD)? Do FLEs have a direct access to BD? (If so, what sort of information? And do frontline staff run analytic queries, test models or just use information pushed down to them?); and do BD insights empower FLEs in linking customers and improving firm performance?

As there is a significant knowledge gap on these broad issues in the extant literature, the authors intend to take up a further study with major service firms like banks, airlines or
healthcare providers. It is anticipated that a qualitative field study will help to portray the maturity of BDA and its influencing role on the frontline. Having synthesised the services context, the authors intend to develop a survey instrument to collect data from different industries and build a quantitative model to serve predictive purposes. Such empirical work while producing quantitative evidence to the premise would also benefit both management and practitioners alike. The future work shall focus on develop a robust plan, so that firms are set to realise benefits from BD in enhancing their service delivery, and creating meaningful positive impact to the client organisation.

7.3 Conclusions
Service, which is more of personal in nature (Vargo and Lusch, 2004), has succumbed to standardisation. Coupled with the transformative changes of service revolution (Rust and Huang, 2014), BDA is re-enabling the firms to balance standardisation vs customisation challenges, namely, effectiveness vs efficiency (Kiron and Shockley, 2011). BDA contributes to not only transparency and data-driven decision making (NewVantage Partners, 2014) but also shifting the power to downstream, i.e., FLEs (Kiron et al., 2012). Following the empirical works of Wilder et al. (2014) and Sony and Mekoth (2014), the authors are motivated to find how well BDA is enabling the FLEs in service adaptation. Our literature search pointed out that there is a significant gap in literature about FLEs within BDA or vice versa. This research fills that knowledge gap by bridging the streams of FLEs and BDA.

References


Oracle (2012), “Big data for the enterprise”.


(The Appendix follows overleaf.)
### Appendix 1

<table>
<thead>
<tr>
<th>Search theme</th>
<th>Frontline employees and big data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Search</strong></td>
<td><strong>Date</strong></td>
</tr>
<tr>
<td>1</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.1</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.2</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.3</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.4</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.5</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.6</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.7</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.8</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>1.9</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>2</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>3</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>4</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>5</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>5.1</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>5.2</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>5.3</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>6</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>7</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>7.1</td>
<td>11 November 2015</td>
</tr>
<tr>
<td>8</td>
<td>13 November 2015</td>
</tr>
<tr>
<td>8.1</td>
<td>13 November 2015</td>
</tr>
<tr>
<td>1</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>1.1</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>1.2</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>1.3</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>1.4</td>
<td>12 November 2015</td>
</tr>
<tr>
<td>21</td>
<td>23 November 2015</td>
</tr>
<tr>
<td>21.1</td>
<td>23 November 2015</td>
</tr>
<tr>
<td>21.2</td>
<td>23 November 2015</td>
</tr>
<tr>
<td>21.3</td>
<td>23 November 2015</td>
</tr>
</tbody>
</table>

### Table AI.

<table>
<thead>
<tr>
<th>Literature search summary for big data and frontline employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>644 BPMJ 23,3</td>
</tr>
</tbody>
</table>

**Corresponding author**

Saradhi Motamarri can be contacted at: msaradhi@yahoo.com

---

For instructions on how to order reprints of this article, please visit our website: [www.emeraldgrouppublishing.com/licensing/reprints.htm](http://www.emeraldgrouppublishing.com/licensing/reprints.htm)

Or contact us for further details: permissions@emeraldinsight.com
Big data in the Danish industry: application and value creation

Sune Dueholm Müller
Department of Management, Aarhus University, Aarhus, Denmark, and
Preben Jensen
Aarhus University, Aarhus, Denmark and
Santander Consumer Bank, Aarhus, Denmark

Abstract

Purpose – The development within storage and processing technologies combined with the growing collection of data has created opportunities for companies to create value through the application of big data. The purpose of this paper is to focus on how small and medium-sized companies in Denmark are using big data to create value.

Design/methodology/approach – The research is based on a literature review and on data collected from 457 Danish companies through an online survey. The paper looks at big data from the perspective of SMEs in order to answer the following research question: to what extent does the application of big data create value for small and medium-sized companies.

Findings – The findings show clear links between the application of big data and value creation. The analysis also shows that the value created through big data does not arise from data or technology alone but is dependent on the organizational context and managerial action. A holistic perspective on big data is advocated, not only focusing on the capture, storage, and analysis of data, but also leadership through goal setting and alignment of business strategies and goals, IT capabilities, and analytical skills. Managers are advised to communicate the business value of big data, adapt business processes to data-driven business opportunities, and in general act on the basis of data.

Originality/value – The paper provides researchers and practitioners with empirically based insights into how the application of big data creates value for SMEs.

Keywords Big data, Process management, Value analysis

Paper type Research paper

1. Introduction

The concept of big data is attracting a lot of attention in both the mass media and the academic literature, and data are seen as a competitive resource and new means of creating value for organizations. However, in a report from 2013 by the Danish Business Authority (Erhvervsstyrelsen, 2013), big data was identified as a hard-to-grasp and unwieldy concept that many companies lacked experience with. The report categorized the companies actively working with big data into two groups. The first group consists of large, often multinational companies with long-standing experience in business intelligence and analytics. The second group consists of relatively small and young companies, i.e. startups, focusing on business opportunities with regard to big data. The two groups of companies represent only a small part of the total number of companies in Denmark. The large majority of companies are small and medium-sized companies, which have limited experience with big data. As a consequence, there is a lack of knowledge and therefore value in studying this large and diverse group of companies. This is also emphasized by extant literature. SMEs often find access to extensive consumer data prohibitively expensive (Donnelly and Simmons, 2013). “Small and medium-size businesses are often intimidated by the cost and complexity of handling large amounts of digital information” (Simon, 2013), which put them “at a severe disadvantage to big competitors that had the financial muscle” (Donnelly and Simmons, 2013) to collect, analyze, and act upon data on, e.g., customer behaviors and market trends. Donnelly and Simmons (2013) call for more research focusing on SMEs, which is echoed by Simon (2013) who wants “attractive alternatives for companies that can’t
afford to – or simply don’t want to – hire their own data scientists”. Against this backdrop, it is the aim of this paper to investigate the extent to which the application of big data creates value for small and medium-sized companies. Specifically, the paper addresses the following research question:

**RQ1.** To what extent does the application of big data create value for small and medium-sized companies?

In focusing on value, we follow McAfee and Brynjolfsson (2012) in asking how big data will help companies improve business performance. In other words, what is the business value of being data driven. Our research is based on an in-depth literature review combined with empirical data from an online survey. The literature review describes state-of-the-art knowledge on big data. This knowledge forms the basis for the survey, which was used to collect data from a sample of small and medium-sized companies. The data were collected through an online survey, which was designed specifically for the purpose of this paper. The survey yielded responses from 457 small and medium-sized companies, which in turn form the basis for our analysis of whether and how big data is creating value for small and medium-sized companies. Based on the literature review and our analysis of the empirical data, we discuss our findings and the implications for researchers and practitioners.

The term “big data” implies that size is a defining characteristic. However, other characteristics are also mentioned in the literature. Laney (2001) suggests that “volume,” “variety,” and “velocity” (sometimes referred to as the three V’s of big data) are key data management challenges, and according to Gandomi and Haider (2015) “the three V’s have emerged as a common framework to describe big data” (Gandomi and Haider, 2015, p. 138). Volume and variety refer, respectively, to the magnitude and heterogeneity of data, whereas velocity refers to the speed at which data are generated, analyzed, and acted upon. More recently, IBM has added “veracity” as a fourth “V” (see, e.g. www.ibm.com/bigdatahub/com/infographic/four-vs-big-data), which refers to the uncertainty of data. For an extensive account of big data definitions, including the additional characteristics of “variability” and “complexity” proposed by SAS as well as “value” introduced by Oracle, please see Gandomi and Haider (2015).

The paper is structured as follows. First, we introduce our choice of analytical framework. Second, we present our approach to reviewing the literature, followed by an account of state-of-the-art knowledge on big data. Third, we analyze the empirical data by applying statistics and qualitative content analysis. Last, but not least, we discuss our findings, the implications for practitioners, and avenues for future research.

### 2. Analytical framework

For the purpose of studying the complex concept of big data, we decided that an analytical framework was needed to guide and structure our research efforts. Such a framework provides us with structure and overview, and it guide us in interpreting and understanding the concept of big data from all relevant perspectives. For these reasons we have chosen the DELTTA model by Davenport (2014) as our analytical framework. The DELTTA model is established specifically for the purpose of analyzing and understanding the concept of big data. Thus, the DELTTA model defines big data by dividing the concept into six elements. Each element of the DELTTA model is clearly defined, and each element adds insights into the big data concept. The six elements of the DELTTA model are summarized in Table I.

As an analytical framework, the DELTTA model helps us structure our research, compartmentalizing the analysis into manageable parts. Dividing the analysis into six parts enable us to look at each element of the DELTTA model individually and the relationships between elements.
In this paper, the DELTTA model serves a number of purposes. First, it helps us structure the literature review. The literature review describes state-of-the-art knowledge on big data and is structured according to the six elements of the DELTTA model. Second, the survey, which is distributed to a sample of companies for the purpose of empirical data collection, is also structured around the DELTTA model. This structuring ensures that all relevant aspects of big data are covered in the survey. Third, the DELTTA model supports our research by structuring the statistical analysis of the data. In line with the survey, the analysis is divided into the six elements of the DELTTA model. The analysis looks into each element as well as the relationship between them. Finally, the DELTTA model is used in the discussion of the results.

3. Review methodology

The literature review is based on the guidelines and recommendations by Webster and Watson (2002) and Okoli and Schabram (2010). According to Fink (2005), a quality literature review must be systematic in following a methodological approach, explicit in explaining the procedure by which it was conducted, comprehensive in its scope by including all relevant material, and reproducible by other researchers following the same procedure in reviewing the topic. Acknowledging that “the quality of literature reviews is particularly determined by the literature search process” (Vom Brocke et al., 2009, p. 2206), in this section we describe in detail how the literature was identified and analyzed.

The purpose of our literature review is to identify papers, which can contribute to an understanding of how big data can be applied to create value from an organizational and business perspective. The level of analysis in this literature review is the organization. Papers written at another unit of analysis are excluded unless they contribute to an understanding of the business application of big data within organizations. We focus on

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Capture, storage, and analysis of high-quality data characterized by high volume, velocity, and variety. This element includes the processes of preparing data for analysis. Capturing, processing, and structuring data for analysis is an integral part of any big data project</td>
</tr>
<tr>
<td>Enterprise</td>
<td>The entire organization understands the opportunities created by big data and is willing to make the appropriate changes to take advantage of these opportunities. Big data is not seen as a technical IT project, but as an integrated part of all relevant business processes of the company</td>
</tr>
<tr>
<td>Leader</td>
<td>The leader and enterprise elements are closely related. It is the responsibility of the leader to ensure that big data is integrated into every part of the company. This responsibility includes the willingness to act on the basis of data. Senior management is actively looking into opportunities created by big data</td>
</tr>
<tr>
<td>Target</td>
<td>Target is closely related to the leader element in the sense that the leader should decide the direction and goals of big data projects. The company should strategize the use of big data in accordance with the goals of the company and the level of IT and data maturity of the organization</td>
</tr>
<tr>
<td>Technology</td>
<td>The technology and data elements are closely related. Technology in combination with Data forms the foundation for the other elements. Big data is mainly driven by developments in storage and processing technologies. Without technologies to capture, store, and retrieve large amounts of data, a company cannot realistically hope to create value through big data</td>
</tr>
<tr>
<td>Analysts</td>
<td>The analysts element represents the people aspect of big data. In relation to big data, the analysts element focuses on the skill set that a company needs in order to successfully execute big data projects. The Analysts element covers all the technical roles, which are needed in big data projects. The Analysts element is closely related to the Data and Technology elements in the sense that analysts must be able to understand and handle data using the available big data technology</td>
</tr>
</tbody>
</table>

Table I. The six elements of the DELTTA model
companies, but papers reporting on other types of organizations are included to the extent that they offer relevant insights. The goal of the literature review is to identify relevant papers as a solid foundation for the survey. Our paper seeks to address the research question at a general and not domain- or industry-specific level. The literature is selected in support of this goal, i.e. we are focusing on general business perspective papers on big data. Papers focusing on industry- or domain-specific applications of big data are only included to the extent that lessons learned are distilled and generalized across industries and domains.

Based on the taxonomy by Vom Brocke et al. (2009) and following Cooper (1988), we characterize our literature review as illustrated in Table II. We focused on the research outcome described in the analyzed articles with the goal of integrating the findings of existing studies based on the key concepts of the DELTTA model. We strive for a neutral representation with general scholars as target audience. Finally, we limit our review to literature that is considered pivotal or central to the topic of big data.

The Scopus and Web of Science (WoS) citation databases were used in identifying the relevant literature. WoS and Scopus are often used in combination for bibliometric analyses, because their coverage differs substantially, for example with regard to the arts and humanities field (Mongeon and Paul-Hus, 2016). Scopus covers more journals and includes most journals indexed in WoS, but WoS has more exclusive journals in the field of natural sciences and engineering. Although using these citation databases introduces biases (favoring, e.g. English-language journals), there is no “suitable alternative to WoS and Scopus when it comes to performing multidisciplinary and international bibliometric analyses” (Mongeon and Paul-Hus, 2016, p. 226). For example, Google Scholar’s “suitability for research evaluation and other bibliometric analyses has been highly questioned because of the sporadic coverage of non-English literature, various inconsistencies (e.g. indexation of non-existing journals) in the data, and a lack of transparency of the coverage” (Mongeon and Paul-Hus, 2016, p. 226). Moreover, both Scopus and WoS provide access to leading IS journal articles and conference papers (Vom Brocke et al., 2009). Since research on big data is still embryonic in nature, most research is expected to be reported in conference papers. The literature review therefore includes both journal articles and conference papers. Books and other non-scientific material is excluded to ensure that only peer-reviewed articles and papers are included, which is in line with Vom Brocke et al. (2009) emphasizing that “it is commonly recommended to focus on articles published in scholarly journals” (p. 2213). Furthermore, by thus explicating our choice of literature sources, we also adhere to the recommendations by Vom Brocke et al. (2009) that “the process of excluding sources (and including, respectively) has to be made as transparent as possible in order for the review to proof credibility” (p. 2207).

**Table II.** Taxonomy of literature reviews

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) focus</td>
<td>research outcomes, research methods, theories, applications</td>
</tr>
<tr>
<td>(2) goal</td>
<td>integration, criticism, central issues</td>
</tr>
<tr>
<td>(3) organisation</td>
<td>historical, conceptual</td>
</tr>
<tr>
<td>(4) perspective</td>
<td>neutral representation, espousal of position</td>
</tr>
<tr>
<td>(5) audience</td>
<td>specialised scholars, general scholars, practitioners/politicians, general public</td>
</tr>
<tr>
<td>(6) coverage</td>
<td>exhaustive, exhaustive and selective, representative, central/pivotal</td>
</tr>
</tbody>
</table>

**Source:** Vom Brocke et al. (2009); following Cooper (1988)
3.1 Search criteria

Since the survey focuses on value creation through business application of big data, the literature search criteria should reflect this choice of subject. To ensure that all relevant papers are identified and included in the review, synonymous words were included owing to the fact that there is no established terminology within emerging research fields like big data. The search is limited to papers written in the English language, assuming that most of the peer-reviewed literature on the topic is written for academic journals or conferences.

The main search criterion is that papers contain the words “big data.” There is no generally accepted abbreviation of this concept, and it is assumed that these search words are only spelled, sequenced, and combined in this particular way (i.e. “big data”). The search is, however, not case sensitive. Since, our research focuses on the application of and value creation through big data within companies, the search also includes “value,” “application,” “company,” and synonymous words. The value creation aspect is covered by “value” and “valuable” as well as the synonyms words “benefit,” “beneficial,” “profit,” and “profitable.” The words “benefit,” “profit,” and “value” are used in combination with a wildcard operator (e.g. “benefit*”) in order to include compound words and plural nouns. The “application” aspect is more difficult to capture. The word “application” has several meanings (e.g. a request or petition and the act of applying something). In order to capture the use of big data for business purposes, we included “business case” in our search for relevant literature. In searching for the literature, the chosen unit of analysis is the organizational level as previously mentioned. The words “organization,” “company,” and “corporation” were included, and a wildcard was again added to account for plural and compound words. Accounting for our choice of keywords, conforms to the guidelines by Vom Brocke et al. (2009) who stress that “particularly the applied keywords have to be documented precisely, so that other scholars can evaluate whether they sufficiently match the topic under investigation” (p. 2214).

The concept of big data is relatively new and most papers about big data are from 2005 and up until today. As a pragmatic means of reducing the number of papers and focusing on up-to-date research, we decided to limit the search to papers from 2010 and onward. This selection criterion ensures a focus on the most recent research with perspectives on the technological possibilities of the 2010s. Forward and backward searches (see Subsection 3.3) minimized the risk of excluding older but highly relevant papers. Papers in the two literature databases are extracted by using the querying tools provided for each database. The results of the database queries are presented in Table III.

The total number of papers selected from the two databases for further screening was 629. This number includes duplicate papers that are found in both databases. The next step of the literature review was screening the selected papers.

3.2 Screening the selected papers

Searching through the two databases resulted in the identification of 629 potentially relevant papers. The next step was to screen the papers and determine their relevance in relation to our research. For that purpose we established three selection criteria. First, the paper should provide insight into at least one of the six elements of the DELTTA model. Second, the paper should have an organizational level of analysis and not be limited to any particular domain or industry. Third, the paper should not focus purely on the technical aspects (hardware and software) of big data. Big data is enabled by new technologies for the capture, storage, and analysis of data. In the context of the DELTTA model, storage and processing technologies are, however, treated in terms of their possibilities and role in big data. We therefore decided to exclude papers that describe, compare, or review the specific types or brands of hardware or software. Many papers include the subject of big data only as a minor part, and they were discarded. A lot of
papers describe the potential of big data to specific companies or kinds of businesses. A few papers focus on big data implications for schools and educational systems. These papers were also not included. A number of papers were selected at first only to be discarded after a more thorough review of their content. In cases where the relevance of the papers was unclear, the papers were included and later subjected to a second reading. During the review, the contribution of each paper was categorized under one or more elements of the DELTTA model. A number of articles contribute to more than one element. All contributions were placed in a concept matrix as advocated by Webster and Watson (2002). In the end, titles and abstracts of all 629 papers were read, resulting in 26 papers being selected for inclusion.

3.3 Backward and forward searches
Backward and forward searches were conducted. The backward search involved looking through the reference lists of all selected papers for the purpose of finding additional relevant papers, which had not been discovered in the initial searches in Scopus and WoS. This search involved browsing the titles of the referenced papers in order to decide whether any of the papers might be relevant to include. Only titles, which were obviously not relevant, were discarded. All other papers were selected for further study and evaluated according to the same selection criteria as the other papers. The backward search yielded another two papers. The forward search was performed by identifying and evaluating the papers which reference the previously selected papers. The Scopus and WoS databases were used to locate these papers. The evaluation followed the same procedure as the backward search. The forward search resulted in another two papers being selected. The literature search process is illustrated in Figure A1.

4. Literature review
In the following, state-of-the-art knowledge in the literature is summarized for each element of the DELTTA model. All selected papers have been categorized according to the DELTTA model, and each paper is contributing to at least one element of the DELTTA model. The contribution of each paper is presented with regard to the particular element of the DELTTA model.
4.1 First element: data

The extant literature looks at data from different perspectives. Data, including sourcing of data and data quality, are key to value creation. Data have no inherent value to businesses and become valuable only when they are placed in relevant contexts. This is no more apparent than in the article by Miller and Mork (2013), which presents a value chain perspective on data. From this perspective, data travel through a value chain from its source through a process of quality assurance to the end receiver who uses it as a basis for decision marking.

The paper by Debortoli et al. (2014) focuses on the differences between business intelligence and big data. From a data perspective, business intelligence uses structured data residing in company-internal databases, whereas big data seeks to extract value from semi-structured or unstructured data originating from sources outside the organization. Chen et al. (2014) describe a value chain for big data divided into four phases: data generation, data acquisition, data storage, and data analysis. Similarly, Miller and Mork (2013) present a similar value chain with an emphasis on how data move through the value chain to become a basis for making informed decisions. At the input end of the value chain, the paper by Barton and Court (2012) encourages creative sourcing of data, internally from other departments and externally from public databases. The paper by Joseph and Johnson (2013) introduces the notion of overproduction and underconsumption of data. The paper suggests that overproduction of data is analyzed and either reduced or consumed.

The security aspect of big data is the focus of the paper by Sagiroglu and Sinanc (2013). The paper points out the weaknesses of storing data centrally and stresses the importance of controlling data access both physically and electronically.

In terms of value creation, Power (2014) stresses that value from data is not created as a function of size but through context and presentation. In a similar vein, the paper by Boyd and Crawford (2012) argues that data taken out of context lose its meaning and value, and that big data has no extra value due to sheer size compared to small data. In order to create value from unstructured data, the paper by Beath et al. (2012) point out the importance of documenting the workflows that create and use unstructured data. Another and more direct way of creating value from data is discussed by Najjar and Kettinger (2013) who propose selling data to other organizations.

The quality of data is the primary concern of Hazen et al. (2014). The intrinsic quality of data is described along four dimensions (accuracy, timeliness, consistency, and completeness). The paper introduces methods for monitoring and controlling data quality. In the paper by O’Leary (2013), the challenges of securing reliable data are described through a case study of mobile device, sensor-based apps. Data reliability is challenged, on the one hand, by variations in user incentives and behavior and, on the other hand, by variations in data depending on the type of mobile device.

Finally, two papers take a critical stance by questioning the value of big data. The paper by Lavalle et al. (2011) discusses the notion of too much data. In a survey by Intel (2012), 200 IT managers were asked to rank the top three sources of data in terms of value, and traditional data came out on top despite all the hype about unstructured data.

4.2 Second element: enterprise

Big data and value creation is not only about data and technology. For a company to truly gain value from big data, use of data for decision making and other purposes must be part of the organizational culture. All employees need to understand and trust data, and they should become accustomed to asking questions like “what does the data say?” and trust the answer when the data are not in line with commonly held beliefs.

Phillips-Wren and Hoskisson (2015) quote a paper by Weill and Ross saying that the alignment of technology, people, and organizational resources in becoming a data-driven
company is difficult. The paper by McAfee and Brynjolfsson (2012) emphasize that the most important question of any data-driven organization is not “what do we think?” but “what do we know?” According to Beath et al. (2012), IT departments are, however, unable to cope with the proliferation of information by themselves. The challenge of interpreting and using data to improve the organizational flexibility and business performance of an organization necessitates close cooperation between IT and business managers. Rajpurohit (2013) expresses this sentiment by saying that “business domain understanding and technology solutions need to work hand in hand to deliver effective analytics solutions.” This is furthermore echoed by Wamba et al. (2015) who state that reaping the benefits of big data requires an alignment of the organizational culture and capabilities across the organization. Referencing Barton and Court (2012), they stress that a key challenge is making big data trustworthy and comprehensible to all employees. Barton and Court (2012) state that “the lead concern expressed to us by senior executives is that their managers do not understand or trust big data-based models.” Organizations are, however, not equally mature in terms of analytical capability. The paper by Lavalle et al. (2011) categorize organizations based on their analytical capabilities. They identify three levels of analytical capability: aspirational, experienced, and transformed.

4.3 Third element: leader
The leader and enterprise elements are closely related in the sense that management needs to show their trust in data ahead of any organization-wide business process changes. A manager who trusts data more than intuition sends a powerful message to the rest of the organization, and it paves the way for changing the culture of a company into one that relies on data for its internal processes.

The paper by McAfee and Brynjolfsson (2012) points to the specific actions that leaders may take in order to lead the big data transformation of companies. The first thing is to ask questions like “what does the data say?” when faced with difficult decisions, followed by questions like “where do the data come from?” and “what kind of analyses have been performed?”. The second thing is to allow themselves to be overruled by data; “few things are more powerful for changing a decision-making culture than seeing a senior executive concede when data have disproved a hunch” (McAfee and Brynjolfsson, 2012). Another approach to big data adoption is described by Gopalkrishnan and Steier (2012) who suggest that leaders ask three questions: “What is the business problem?”, “is the available data suitable for problem solving?”, and “what is the ROI of big data?” The paper by Rajpurohit (2013) emphasizes the value in learning from failures and the need to analyze the gap between potential and realized value. Phillips-Wren and Hoskisson (2015) discuss the organizational and managerial challenges in transitioning from using data to relying on analytics and integrating big data into organizational decision making. Despite challenges, Tallon (2013) points out that data governance and information management are of increasing strategic importance to organizations. Although the use of big data may offer companies strategic advantages, the case study by Najjar and Kettinger (2013) highlight the importance of balancing the advantages of information transparency across business partners against the loss of power from information sharing with customers, suppliers, and competitors. Similarly, Barton and Court (2012) elaborate on these advantages “in a deliberate effort to weave big data into the fabric of daily operations”. Meanwhile, McNeely (2014) cautions leaders not to focus on technology and be aware that “there is a huge gap between our ability to acquire data and our ability to make effective use of data to advance discovery”. Instead, Ebner et al. (2014) suggest that senior management asks three questions in order to determine how to deal with big data: “Do we have a big data or an IT infrastructure problem?”, “are we lacking critical information that the use of a big data solution will help us acquire?”, and “what are our analytical requirements?” Power (2014)

652
also warns leaders of the dangers of putting too much faith in systems and big data models, and points to the 2008 financial crisis as an example of the failure of data-driven models to accurately factor in financial risks. Instead, Beath et al. (2012) recommend senior managers to commit to three practices: identify your sacred data, define the workflows relying on unstructured data, and use data to improve business processes. Furthermore, based on a large-scale survey, Lavalle et al. (2011) conclude that "the leading obstacle to widespread analytics adoption is lack of understanding of how to use analytics to improve the business," and that "lack of management bandwidth due to competing priorities" is an obstacle to adoption of analytics.

4.4 Fourth element: target

Any company needs to start by asking what the goal or purpose of applying big data is. Big data is a solution to which business problem? Start with the question before asking about possible solutions and how data can help.

Gopalkrishnan and Steier (2012) stress the importance of having one or more organizational goals as a basis for establishing and continually monitoring the business case for any big data investment. Likewise, Barton and Court (2012) emphasize that the desired business impact should drive decisions regarding data sourcing, model building, and organizational transformation. The paper by Lavalle et al. (2011) advocate starting with the problem (i.e. the question) rather than the solution (i.e. the data). McNeely (2014) warns of the dangers in basing decisions on correlations instead of an in-depth understanding of big data. Phillips-Wren and Hoskisson (2015) mention description, prediction, and prescription as three ways in which analytics supports target management and provides value to an organization. According to Joseph and Johnson (2013), big data analytics also facilitates business process redesign. More generally, Rajpurohit (2013) points out the importance of seeing analytics as a means of transforming data into valuable insights. To that end, Tallon (2013) argues that establishing data management practices, which balance value creation and risk exposure, is a new organizational imperative for achieving competitive advantage and maximizing value from big data. The relationship between competitive advantage and the application of big data is studied by Kamioka and Tapanainen (2014). Their paper concludes that a positive impact on competitive advantage depends on extensive and systematic big data usage. The Hospitals and Health Networks' (2014) paper elaborates by suggesting that data are used in a structured manner in pursuit of relevant questions (i.e. business targets).

4.5 Fifth element: technology

New technologies for capturing, storing, and analyzing data must be combined with more traditional technologies. Big data technologies should be used alongside the existing legacy systems. The combination of traditional and big data storage technologies help reduce costs while creating value.

Philip Chen and Zhang (2014) call for new techniques and technologies to be developed. Multidisciplinary approaches (computer science, economics, mathematics, and statistics) are required for the purpose of discovering valuable information in big data. With regard to technologies, Barton and Court (2012) discuss how legacy systems may challenge the application of big data. They problematize whether existing systems are able to handle the data required for real-time decision support. Ebner et al. (2014) conclude that a hybrid strategy combining relational database structures and the MapReduce programming model (framework for large-scale data processing) is preferable and most likely to create value. In terms of storage, Beath et al. (2012) argue that a company’s IT department should take the lead in creating a reliable and cost-effective solution. The paper suggests a three-tier data storage solution, each tier having different configurations and applications.
Intel (2012) predicts that the current mix of batch and real-time delivery of data will change, and that more and more data will be delivered in real-time. The Hospitals and Health Networks’ (2014) paper is even more specific and describes the goal as being a move from retrospective to real-time analytics and eventually predictive analytics.

### 4.6 Sixth element: analysts

There is general agreement that the skills of analysts are needed in order to create value out of big data. These analysts need to work together with managers and domain experts to realize this value. A company needs to carefully consider the need for particular skills before hiring analysts.

Sagioglu and Sinanc (2013) reason that companies should not only employ managers and analysts with insights into applications of big data; they also need to invest in education and training of key personnel. The different requirements for job positions within business intelligence and big data jobs are analyzed by Debortoli et al. (2014) who look at the wording of job advertisements. They conclude that big data requires skills within software engineering and statistics. Similarly, skills and job descriptions of data scientists are discussed by Davenport (2012). The paper describes how to attract and retain data scientists with the skill set required to create value from big data. The skills of data scientists are also the topic of the paper by Davenport et al. (2012). A data scientist needs to understand analytics, have skills in statistics and mathematics, understand the business, and possess good communication skills.

The new organizational role of chief data officer is described by Ebbage (2014). The chief data officer is characterized as someone who knows how to use data across an organization and is able to chart a course for the data scientist to follow. Power (2014) argues that managers need to understand what data scientists can do and not do for a company before hiring any. Gerhardt et al. (2012) describe the role of the so-called data infomediary who is viewed as an employee who does the matchmaking between data originators and data beneficiaries. Finally, Viaene (2013) explains the need for domain experts and data scientists to work together, leveraging their different competencies in order to create value.

### 4.7 Mapping state-of-the-art literature

The six elements of the DELTTA model enable us to look at the different aspects of big data individually. Looking at each aspect individually helps us understand how each element contributes to value creation. Hence, the literature review creates an overview of state-of-the-art knowledge of value creation from the perspective of the DELTTA model.

Table IV shows the concept matrix resulting from our literature review, which provides a map of the literature on big data with the DELTTA model as “compass.” Although the DELTTA model allows us to focus on each element in turn, it does not help us understand how the six elements influence each other. A large part of the papers in the literature review contribute to an understanding of two or more elements of the DELTTA model. This observation reveals a close relationship between some elements in the DELTTA model. Looking at Table IV, there is however no clear pattern between the six elements of the DELTTA model. Each paper included in the literature review looks at big data from the perspective of companies, but they focus on different aspects of big data. In order to view big data from a more holistic company perspective, it is necessary to include empirical observations from the collected data. The empirical data were collected for the purpose of understanding the relationship between applications of big data and value creation from the viewpoint of Danish companies. This data have been analyzed and the findings are presented in the following section.
5. Empirical research
In order to better understand how big data is used in practice to create value, an online survey was created and distributed to a sample of Danish companies. The unit of analysis is small and medium-sized private sector companies (SMEs). Small and medium-sized companies are defined as companies with 249 employees or less. This definition corresponds to that of Statistics Denmark (www.smvportalen.dk/om-smvportalen/definition-af-smv), i.e. the central authority on Danish statistics. The definition of small and medium-sized companies is solely based on the number of employees. Turnover or any other accounting-based number is not part of the definition. The Danish definition may vary from that of other EU countries. All organizations are for-profit companies. By focusing on for-profit companies, the definition of value is limited to economic profit.

5.1 Constructing the survey
The survey is structured around the DELTTA model, covering all the six elements of big data described by the framework. The survey focuses on both the value creation and application aspects of big data. Thus, the survey contains two questions for each element of the DELTTA model. The questions are phrased as statements, and respondents are asked to express their level of agreement on a five-point Likert scale. The scale ranges from strongly agree to strongly disagree and includes a neutral response. Strongly agree is attributed the value “1” and strongly disagree the value “5.” All questions are phrased as

<table>
<thead>
<tr>
<th>No.</th>
<th>Reference</th>
<th>Data</th>
<th>Enterprise</th>
<th>Leader</th>
<th>Target</th>
<th>Technology</th>
<th>Analysts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Barton and Court (2012)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Beath et al. (2012)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Boyd and Crawford (2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Chen et al. (2014)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Davenport (2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Davenport (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Debertoli et al. (2014)</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Ebner et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Ebbage (2014)</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Gerhardt et al. (2012)</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Gopalkrishnan and Steier (2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Hazen et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Hospitals and Health Networks (2014)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Intel (2012)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Joseph and Johnson (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Kamioka and Tapanainen (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Lavalle et al. (2011)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>McAfee and Brynjolfsson (2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>McNeely (2014)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Miller and Mork (2013)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Najar and Kettinger (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>O’Leary (2013)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Philip Chen and Zhang (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Power (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Rajpurohit (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Sagiroglu and Sinanc (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Tallon (2013)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Viaene (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Wamba et al. (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table IV. Concept matrix
positive statements, which makes it easy for respondents to understand each question and provide an accurate response to the statement.

Because each question is specific to a particular element of the DELTTA model, the responses to the statements are also narrowly focused, which strengthens the external validity of the survey. In the introduction to the survey, it is emphasized that no prior knowledge of big data is assumed or needed. Therefore, companies with no or limited experience with big data were also included in the survey. The survey questions have been carefully phrased taking the broad spectrum of potential respondents into consideration. The questions have to be specific enough to ensure valid answers but without assuming an understanding of or prior experience with big data. As a consequence, we made it possible to answer the questions in such a way that the responses would reveal the level of experience with big data. Respondents were also encouraged to elaborate their answers through qualitative comments.

All respondents received the same questions in the sense that their responses to each statement did not affect the sequence or content of subsequent questions. Asking all respondents to answer the same questions also supports the external validity of the survey. The survey contains questions regarding the company, e.g. number of employees and the company’s general use of data. These questions serve to confirm a random selection of companies, since it is important for generalization purposes that the companies vary in terms of age, number of employees, and types of businesses.

Due to the survey being aimed at different types of businesses and not requiring any prior knowledge of big data, the concept of big data needed to be explained in such a way that most respondents would be able to understand it. The explanation of big data used in the survey is based on the various definitions in the literature identified during the review. Big data was explained as “Data of very high volume and complexity which compared to ordinary data requires special skills, technologies, and tools to capture and use.” Furthermore, the distinction between data and big data was characterized as blurry, and the respondents were asked to rely on their judgment and knowledge of their companies in deciding whether to characterize their data as big data or ordinary data.

The survey includes two questions for each element of the DELTTA model yielding 12 questions in total. This was a conscious choice given the complexity of the subject and that we set out to collect responses from a heterogeneous sample of companies in terms of prior experiences with and knowledge of big data. The respondents have answered these questions based on the perceived application and value of big data, which introduces the risk of misunderstanding one or more questions, resulting in misleading answers. The risk is, however, reduced by the sheer number of respondents. More importantly, special attention was paid to the wording of questions and the terminology used.

Prior to distributing the survey, it was reviewed by the peers with particular attention to the wording and use of terminology in the survey. Subsequent changes were made, particularly with respect to the explanation of big data. Balancing the need to make the survey easily comprehensible to the group of diverse respondents and ensuring the external validity of the survey through the use of appropriate big data terminology were our main concerns. Ongoing discussions among the authors and survey adjustments helped us achieve this goal.

5.2 The respondents

The respondents were selected from the Danish Business Register (http://datacvr.virk.dk/data/). The companies were chosen based on their number of employees, which is registered in the database. In addition, companies were selected based on their geographical location. Companies are categorized according to the region in which they are situated, and we decided to include both the Danish capital and rural regions due to differences in
demography as well as industry composition and density across regions. Some companies were deselected because they had not registered valid e-mail addresses in the database. A valid e-mail address is required for survey distribution purposes. All companies are registered in the Danish Business Register by company type, which allows for the identification of privately owned companies.

Two selection criteria were applied. First, the sample was limited to companies with fewer than 250 employees. Second, only for-profit private companies were selected. This inclusive approach to sampling provides a basis for generalizing the survey results across industries and types of private businesses. The selection process resulted in a sample of 4,043 companies.

The survey was sent by e-mail to the company address registered in the Danish Business register. This is typically a general purpose contact address. Therefore, the e-mail encouraged the recipient to forward the message to the employee best qualified to participate in the survey. This person was described as a management-level employee, preferably with an understanding of both the IT and business side of the company. Due to the broad variety of companies, role descriptions and job titles were not mentioned.

5.3 Distribution of the survey
The survey was distributed using SurveyXact, which is a web-based tool for developing and sending out surveys. Out of the 4,043 e-mails, approximately 400 were returned because of delivery failures. A number of respondents declined to participate in the survey. In total, 471 companies selected for participation in the survey were later discarded, primarily due to invalid e-mail addresses. Reminders were sent out within a week, and the survey remained active over the following week. At the end of the two-week period, 457 responses had been received. The responses were subsequently exported to the statistical analysis software SPSS for analytical purposes.

The number of survey invites arriving at the intended e-mail addresses was 3,572, and the number of completed responses is 457, which translates into a response rate of 12 percent. This response rate is considered normal taking into account that the survey was aimed at the management level (Baruch, 1999). A total of 109 responses were only partially completed. The majority of partial responses did not include answers to the 12 questions about big data. This is in line with a report by the Danish Business Authority (Erhvervsstyrelsen, 2013), finding that many small and medium-sized companies decline to answer surveys about big data due to limited understanding of and experience with big data.

The answers to the six questions about the application of big data and the six questions about value creation form the basis for our data analysis. The responses to the 12 statements are used as 12 separate variables in SPSS. The 12 statements are found in Appendix 1. Statistics regarding the responding companies can be found in Table AII.

5.4 Data analysis
An exploratory factor analysis was performed in SPSS in order to identify the factors (combination of variables) explaining big data application and value creation. The 12 variables derived from the DELTTA model were subjected to principal axis factoring with the purpose of identifying the latent variables in the data. The SPSS correlation matrix shows that all variables have correlations greater than 0.6 with all other variables. All details of the correlation matrix can be found in Table AIII. The result of the Kaiser-Meyer-Okin measure of sampling adequacy is 0.941, which is well above the threshold of 0.6 for factor analyses (Tabachnick and Fidell, 2014). In addition, the ratio between the number of variables and the number of observations is also above the recommended ratio of 1.5 (12 variables and 457 observations), which means that the Bartlett’s test is irrelevant and might be misleading. Cronbach’s $\alpha$ for the factor analysis is 0.971. This reflects the close relationship between the included variables.
The six elements of the DELTTA model were subjected to linear regression. The purpose of performing the regression analysis is to analyze the close relationship between the elements of the DELTTA model identified by the factor analysis. The linear regression shows the degree to which one element of the DELTTA model can predict the other five elements. Each element includes one variable for big data application and one for value creation. Each element was tested against the other five elements. The results show predictability ranges from 0.37 to 0.73. Finally, the 12 variables were subjected to mean value calculations. The results are elaborated in the following.

In order to compare and align the results of the literature review with the survey, we adopt triangulation as it allows for cross-checking and validation from several sources (Miles et al., 2014). Thus, triangulation is applied as a technique to show how our use of multiple data sources (qualitative survey data, quantitative survey data, and extant literature) produces a rich understanding of big data.

5.5 Survey results

In the following, we present our findings from the survey. These findings reflect the perceived rather than the actual value created through the application of big data. This is a consequence of our research design and asking respondents to express their opinions rather than facts (which would be methodologically infeasible). A word of caution: when interpreting the survey results, associations should be interpreted as correlations rather than causations. A high degree of predictability among variables does not necessarily imply that A causes B (or vice versa) but only that they correlate.

5.5.1 Factor analysis. The principal factor analysis resulted in high correlations between all 12 variables. With a high loading on all 12 variables, the factor big data application and value creation was identified. A total of 75.6 percent of the variance is explained by this factor. The high degree of correlation between all 12 variables suggests one or more strong latent variables behind the big data application and value creation factor. The latent variable(s) influence(s) the 12 variables, which manifests itself in the dependence among them. The high degree of explanation is a testament to the strong relationship between the six elements of the DELTTA model. This relationship between the elements of the DELTTA model is further examined by applying the regression analysis.

5.5.2 Regression analysis. The regression analysis confirms the close relationship between the six elements of the DELTTA model. The regression shows the degree to which responses in relation to one element of the DELTTA model predict responses to another element of the DELTTA model. The strong mutual influence among the six elements of the DELTTA model suggests that a company must focus on all the elements in order to maximize the value from the application of big data. For instance, the enterprise element (the need for big data acceptance by, and application throughout, the whole organization) predicts 64 percent of the variance of the data element (capture, storage, and analysis of high volume, high velocity, and high variety data). The strong predictive power suggests that quality data and data handling must be combined with an understanding of big data throughout the company in order to create real value. Other examples of strong predictive power among the six elements of the DELTTA model are found in Tables V and VI. The regression calculations are divided into big data application and value creation.

Application. The lowest degree of prediction is between the target and leader elements. The response to target predicts 37 percent of the response to leader. In contrast, the highest degree of prediction is between analysts and technology. The response to analysts predicts 68 percent of the response to technology. The strong predictive power can be interpreted as a high degree of similarity in the responses. Seen from an application perspective, this means that the use of big data is viewed similarly across companies. This in turn
indicates relatively low variance in the combinations of responses. The similarity in response patterns fits with the assumption of latent variables identified in the factor analysis.

Value creation. The lowest degree of prediction is between the target and data elements. The response to target predicts 40 percent of the responses to data. The highest degree of prediction is between enterprise and data. The response to enterprise predicts 73 percent of the response to data. As with application, the combinations of responses to the statements related to value creation are relatively similar. From the perspective of value creation, it indicates a relatively high level of agreement among respondents with regard to value creation. This is also in agreement with the assumption of latent variables identified in the factor analysis.

5.6 Mean calculations

The mean calculations show the distribution in the responses to the 12 statements. All questions are phrased using positive statement, e.g. “Big data is being actively applied.” Using a negative statement, the same question would be “Big data is not being actively applied.” If a company is actively applying big data, the response to the positive statement is “agree” with an associated value of two (see below). With regard to the negative statement, the response would be “disagree” with an associated value of four. This would, however, result in the mean values not being comparable across statements. Because all questions are phrased as positive statements, the responses are comparable across all 12 statements.

The results are based on 457 responses. The response value should be interpreted as follows: A value of one translates into strongly agree, two equals agree, three is neutral, four means disagree, and five corresponds to strongly disagree. Given these values, low values suggest agreement and high values imply disagreement. For example, the value of 3.43 for big data application with regard to the target element of the DELTTA model (see Table VII) is between three (meaning neutral) and four (being disagree). The mean score of 3.43 can therefore be interpreted as respondents tending to slightly disagree with big data being applied for business strategizing and goal setting.

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Enterprise</th>
<th>Leader</th>
<th>Target</th>
<th>Technology</th>
<th>Analysts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Enterprise</td>
<td>0.64</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Leader</td>
<td>0.60</td>
<td>0.52</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Target</td>
<td>0.39</td>
<td>0.44</td>
<td>0.37</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Technology</td>
<td>0.57</td>
<td>0.45</td>
<td>0.47</td>
<td>0.41</td>
<td>1</td>
<td>n/a</td>
</tr>
<tr>
<td>Analysts</td>
<td>0.57</td>
<td>0.48</td>
<td>0.50</td>
<td>0.48</td>
<td>0.68</td>
<td>1</td>
</tr>
</tbody>
</table>

Table V. Results of regression analysis of big data application

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Enterprise</th>
<th>Leader</th>
<th>Target</th>
<th>Technology</th>
<th>Analysts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Enterprise</td>
<td>0.73</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Leader</td>
<td>0.61</td>
<td>0.64</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Target</td>
<td>0.40</td>
<td>0.45</td>
<td>0.49</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Technology</td>
<td>0.61</td>
<td>0.64</td>
<td>0.58</td>
<td>0.49</td>
<td>1</td>
<td>n/a</td>
</tr>
<tr>
<td>Analysts</td>
<td>0.51</td>
<td>0.51</td>
<td>0.47</td>
<td>0.47</td>
<td>0.64</td>
<td>1</td>
</tr>
</tbody>
</table>

Table VI. Results of regression analysis of big data value creation
6. Survey analysis

6.1 The big data application and value creation factor

The one factor resulting from the principal factor analysis suggests that to the extent that big data is applied, value is created in roughly 75 percent of all cases. An important characteristic of a good factor analysis is that it makes sense. As previously mentioned, the high degree of explanation by this factor suggests the existence of one or more latent variables. A latent variable is found in the framework. The strong relationship and mutual influence between variables is revealed by the regression analysis, displaying a high degree of predictability among the six elements of the DELTTA model. A latent variable is also found in the sample of survey respondents. The companies taking part in the survey are similar in some respects. All respondents are from small and medium-sized companies. In general, these companies have limited resources (competencies, money, etc.). They are limited in terms of their big data investment capabilities and will carefully weigh the costs and benefits by establishing business cases and calculating ROI. The companies responding to the survey may be considered adept at turning the application of big data into value.

6.2 The DELTTA model

Having investigated the overall ability of small and medium-sized companies to turn application of big data into actual value, the next step is looking at each element of the DELTTA model. The statements of the survey can be found in Appendix 1. For each element of the DELTTA model, respondents were given two questions asking them to respond to statements about big data application vis-à-vis value creation. By zooming in on the differences (mean response value) between application and value creation, new insights are generated.

6.2.1 The data element of the DELTTA model. The data element of the DELTTA model is predicted by the enterprise and leader elements. The responses to the enterprise element predict 64 and 73 percent of the responses to the data element (big data application and value creation), and responses to the Leader element predict 60 and 61 percent of the responses. This suggests that data does not create value by itself. In other words, having large amounts of high-quality data is not enough. big data initiatives require management involvement and big data must be accepted and used throughout the company in connection with various business processes. The answers to the questions concerning Data have a mean response value of 2.58 and 2.50 (see Table VII), which is close to neutral. The difference between the application and value creation scores might indicate that data are being used in the companies but that the respondents are unsure about the extent to which it creates value.

6.2.2 The enterprise element of the DELTTA model. Enterprise is predicted by leader and technology (both 64 percent). Leadership involvement is a prerequisite for big data application across the different parts of a company. big data technology facilitates access to data across business processes in a company. In terms of the Enterprise element of the DELTTA model, the mean responses are 2.94 and 2.69 (see Table VII) with the application

<table>
<thead>
<tr>
<th>DELTTA model element</th>
<th>Application</th>
<th>Value creation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>2.58</td>
<td>2.50</td>
</tr>
<tr>
<td>Enterprise</td>
<td>2.94</td>
<td>2.69</td>
</tr>
<tr>
<td>Leader</td>
<td>2.64</td>
<td>2.62</td>
</tr>
<tr>
<td>Target</td>
<td>3.43</td>
<td>3.25</td>
</tr>
<tr>
<td>Technology</td>
<td>2.78</td>
<td>2.79</td>
</tr>
<tr>
<td>Analysts</td>
<td>2.96</td>
<td>2.93</td>
</tr>
</tbody>
</table>

Table VII. Mean calculations of big data applications and value creation
value being lower than the value creation aspect. This implies that the respondents do not see big data being applied across their companies, and that they are unsure as to whether big data creates value to their companies as a whole.

6.2.3 The leader element of the DELTTA model. The leader element of the DELTTA model is mainly predicted by technology (58 percent). This is interpreted as the company leaders allowing big data initiatives to be influenced if not controlled by technological possibilities. Big data is of course also about technology, and the basic premise is that data management, i.e. data storage and analysis, is enabled and facilitated by IT. The mean values of the responses are 2.64 and 2.62, which are almost identical. The fact that respondents answer in the negative with regard to management involvement suggests that leaders of small and medium-sized companies should be more involved in big data initiatives.

6.2.4 The target element of the DELTTA model. The target element of the DELTTA model is predicted by all the other elements. This suggests that goal achievement in terms of big data application and value creation requires an interplay between technology, data, and the involvement of both management and the organization as a whole. The mean values of the responses are 3.43 and 3.25. In other words, the respondents answer in the negative, which suggests that small and medium-sized companies are not adept at setting targets for big data initiatives aligned with their business strategies and goals.

6.2.5 The technology element of the DELTTA model. The analysts element of the DELTTA model predicts 64 percent of the technology element, which indicates a close relationship between the use of technology and analytical skills in big data. By implication, the target and leader elements do not seem to influence the application and value creation through Technology in any noteworthy degree. The mean values of the responses are 2.78 and 2.79. This indicates lack of technology use and consequently a low degree of value creation from its application.

6.2.6 The analysts element of the DELTTA model. The responses to the analysts element are mainly predicted by technology, showing a close relationship between analytical skills and technological use in the application of and value creation from big data. This also indicates that the leader and target elements do not influence this people factor of big data to the same extent. The mean values of the responses are 2.96 and 2.93. The answers are almost neutral, implying uncertainty regarding the application of and value creation from big data competencies.

7. Discussion and conclusion
This research, seeking to investigate the degree to which the application of big data creates value, has yielded a number of findings. First of all, our research shows a correlation between the application of big data and value creation. The close relationship between application and value creation is highlighted by the principal factor analysis, which points to one factor that describes 75.6 percent of the variance in the 12 variables.

Second, our research shows that the six elements of the DELTTA model affect each other. The fact that each of the six elements predicts the variance in other elements points to a high degree of interdependency among the elements of the DELTTA model. From the perspective of private companies, this means that creation of real value through the application of big data depends on all six elements of the DELTTA model being addressed.

Third, our research reveals important insights into the close relationship between application and value creation by analyzing respondents’ survey responses to 12 statements regarding big data.

The responses to the Data element show that companies currently apply data and that it is perceived as creating value. The similarity in responses indicates that application and
value creation (mean response values of 2.58 and 2.50, respectively) go hand in hand. This is in line with Power (2014) who emphasizes that data have no value in itself but becomes valuable through its use in particular organizational and business contexts. Likewise, Boyd and Crawford (2012) stress that data without context are of no value. The results confirm this in the sense that respondents see data as creating value in the specific contexts of their companies. Our qualitative content analysis of survey responses shows that companies use both structured and unstructured data from internal as well as external sources. Internal data include financial, sales, CRM, ERP, and usage statistics data as well as mails and use cases. External data include those from social media, industry reports, and public databases (Eurostat, ECB, etc.) as well as GIS and EDI data. There is tendency to use structured data more. As one respondent says: “We use unstructured data to a lesser extent because the validity, completeness, etc. is not good enough.”

The neutral responses to the statements related to the enterprise element reveal that respondents are not sure how big data is used across their companies and how it creates value to different business processes and parts of the organization. According to Lavalle et al. (2011), the average company would be placed at the aspirational level of analytical capability. Our research shows that the companies are not yet at the level where they continually ask themselves “what do we know” as suggested by McAfee and Brynjolfsson (2012). Meanwhile, our analysis of the qualitative survey comments reveals that companies use data for many different purposes across the enterprise. Generally, data are used for, e.g., production planning, daily operations, KPI management, as well as purchasing and logistics decisions. More specifically, data are used to better understand and cater to customer needs. Data are thus used to analyze customer needs and behaviors, offer product recommendations, improve customer and after-sales service, measure degrees of customer satisfaction, understand user experiences, perform web analytics, personalize marketing campaigns, tailor products to customer preferences, and distill learning from customer complaints and product returns. The focus is on customer loyalty, retention, and resale. One respondent comments: “With accurate information about our customers, we are able to spot potential problems and act proactively based on the information.”

In terms of the leader element, management in an average company is aware of the possibilities for big data application and value creation. However, considering the importance of management involvement, the mean response value is low. This may be interpreted as managers not communicating the importance of big data to the rest of the organization, which goes against the recommendations by Gopalkrishnan and Steier (2012) and McAfee and Brynjolfsson (2012). Nevertheless, the qualitative survey comments indicate that managers use data for decision support, including marketing efforts, competitor analyses, business strategy adaptation, investment decisions, quotations, human resource management, as well as budgeting and forecasting. One respondent says: “Data help identify focus areas, which enable management to make better decisions and establish action plans for a specific area.”

The responses to the statements about the target element reveal that goals have not been defined in many of the 457 companies participating in the survey. With reference to the papers by Gopalkrishnan and Steier (2012) and Barton and Court (2012), which emphasize the importance of clear goals in guiding the use of big data, the companies still face the challenge of connecting big data to business strategies and business processes. However, our qualitative content analysis shows that the use of customer data enables companies to detect and react to new or changing patterns in general markets trends as well as specific customer behaviors. One respondent asserts: “When we see a negative tendency, we often react to it even before the customer is aware of it.”

In connection with the data element, technology is identified as the backbone of big data supporting the other four elements. Technology supports or drives the pursuit of big data
targets, but technology is not an end in itself. Data storage and processing are important aspects of technological use, and the extant literature describes different data storage strategies. Whereas Ebner et al. (2014) recommend a hybrid strategy combined with the use of both traditional databases and new types of data storage, Beath et al. (2012) suggest a three-tier data storage strategy to reduce costs. The qualitative survey comments clearly confirm the link between the data and technology elements. Data from, e.g., ERP, CRM, and case handling systems, are used internally for monitoring and improvement of business processes. According to one respondent: “The business logic and processes are trimmed continuously with greater service orientation and more efficient operations in mind.” Lean and process innovation are made possible by the use of data. One respondent stresses that “by dividing a work process into smaller activities and analyzing each activity, we are able to optimize the individual activities and the entire process.”

Last, but not least, responses to the statements concerning analysts, indicate uncertainty with regard to how big data competencies are utilized and whether they create value. This is problematic considering the extant literature. Previous research suggests different kinds of job roles in relation to big data. First, the paper by Davenport et al. (2012) advocates employing data scientists as a means of creating value. Second, the paper by Sagiroglu and Sinanc (2013) underscores the importance of focusing on educating and training key personnel. Third, Viaene (2013) emphasizes the need for data scientists to work together with domain experts in order to create value out of big data for any company. Yet, the qualitative survey comments reveal that companies use data to provide employees with an “Analyst’s” overview of projects, sales, and more. This provides structure and enhanced understanding of employees’ contributions to business goals, which in turn improves employee satisfaction. In the words of one respondent: “More structure leads to greater job satisfaction.” It also helps management improve the work environment, allocate employees to work activities depending on business needs, and improve business processes.

This study has several implications for researchers and practitioners. For one, our research provides insights into how and to which extent the application of big data creates value to small and medium-sized companies. The empirical data allow us to address the question from the perspective of companies that are working with big data on many different levels. In response to our research question “To what extent does the application of big data create value for small and medium-sized companies?”, we are able to conclude that big data is perceived as creating value to the extent that the six elements of the DETTTA model are addressed. This in turn leads us to recommend that managers pay attention not only to capture, storage, and analysis of data (the Data element), but that they demonstrate leadership through explicit and clear goal setting (the target element), aligning business strategies and goals with IT capabilities (the technology element) and analytical skills (the analysts element). This managerial responsibility also extends to communicating the importance and value of big data in supporting and driving the business, adapting business processes to take advantage of identified opportunities (the enterprise element), and acting on the basis of data (the leadership element). Out study reveals the importance of not only communicating but also showing employees how big data is or should be used, for what purpose, and with which benefits in mind. This is a prerequisite for their being able to support big data initiatives and help realize planned benefits. With regard to other practical implications, managers are advised to take an active role in strategizing, implementing, and using big data. Big data is a powerful tool for both management and business innovation. Widespread big data adoption requires, however, that managers acquire greater understanding of potential applications and benefits of business analytics (Lavalle et al., 2011). Meanwhile, we lack knowledge of the particular competencies and skillsets needed by managers and employees in coping with the challenges of big data application. This paves the way for future studies. Speaking of research implications, additional empirical studies of
big data are needed to extend the insights of this paper. In this paper, we have relied on the DELTTA model for analytical purposes. The DELTTA model has proven to be a useful analytical framework when investigating the complex concept of big data. Future studies may provide additional knowledge by extending the empirical data collection to include large companies, and by relying on other conceptualizations of big data. Moreover, our research has included all types of businesses and companies from all industries. Future research may investigate the relationship between big data application and value creation at a more detailed level by looking at specific organizations or company types. The bird’s eye perspective in this paper carries with it the advantage of yielding knowledge that can be generalized across companies, but this comes at the expense of a more detailed understanding of the hows and whys of big data. Our study also reveals limitations in a survey-based, quantitative study of big data. Thus, the qualitative survey comments reveal that respondents struggle with the distinction between data and big data. One respondent remarks: “The blurry line between big and small data makes it difficult to answer the questions precisely.” Future research may address this limitation and close the knowledge gap by focusing on qualitative case studies of concrete big data initiatives.

References


Corresponding author
Sune Dueholm Müller can be contacted at: sdm@mgmt.au.dk
Appendix 1. Survey questions

The survey, which has been developed for the purpose of collecting empirical data for this paper, is based on the DELTTA model. The survey includes two questions in the form of statements for each element of the DELTTA model. The first statement concerns the application of big data (see the “application” column in Table AI). The second type of statement focuses on value creation through the application of big data (see the “value creation” column in Table AI). All statements were answered with one of the following responses: strongly agree, agree, neutral, disagree, and strongly disagree.

<table>
<thead>
<tr>
<th>DELTTA model element</th>
<th>Application</th>
<th>Value creation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Big data (our own) is being actively applied</td>
<td>Big data (our own) is creating value to our organization</td>
</tr>
<tr>
<td>Enterprise</td>
<td>We apply big data across the entire company</td>
<td>Big data creates value to the whole organization</td>
</tr>
<tr>
<td>Leader</td>
<td>The management understands how big data is being applied in our organization</td>
<td>The management knows how the application of big data creates value to our organization</td>
</tr>
<tr>
<td>Target</td>
<td>Our organization has specific targets for the application of big data</td>
<td>Our organization monitors the value creation and ensures that the goals for big data application are being realized</td>
</tr>
<tr>
<td>Technology</td>
<td>We have technologies that enable the application of big data</td>
<td>Big data technologies support value creation in our organization</td>
</tr>
<tr>
<td>Analysts</td>
<td>Our organization applies relevant big data skills</td>
<td>Big data competencies are prerequisites for the success of our business</td>
</tr>
</tbody>
</table>

Table AI. Statements of selected survey questions
Appendix 2

Step 1: Search in citation databases

<table>
<thead>
<tr>
<th>Web of Science</th>
<th>Scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>339</td>
<td>290</td>
</tr>
</tbody>
</table>

Step 2: Screening for duplicate papers and selection of papers

26

Step 3: Backward and forward searches

4

30

Figure A1.
The paper search process
### Appendix 3

<table>
<thead>
<tr>
<th>Description</th>
<th>Options</th>
<th>Number of respondents</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of company</td>
<td>Less than 5 years</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Between 5 and 15 years</td>
<td>121</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Older than 15 years</td>
<td>313</td>
<td>69</td>
</tr>
<tr>
<td>Number of employees</td>
<td>Fewer than 10</td>
<td>96</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Between 10 and 49</td>
<td>237</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Between 50 and 250</td>
<td>124</td>
<td>27</td>
</tr>
<tr>
<td>Type of industry</td>
<td>Farming, forestry, and fishing</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Production and utilities</td>
<td>118</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Building and construction</td>
<td>67</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Trade and transportation</td>
<td>95</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Information and communication</td>
<td>46</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Finance and insurance</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Property trading and rental</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Business services</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>97</td>
<td>21</td>
</tr>
</tbody>
</table>

Table AII. Respondent statistics

Big data in the Danish industry
## Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correlation</strong></td>
<td>1.000</td>
<td>0.890</td>
<td>0.801</td>
<td>0.835</td>
<td>0.777</td>
<td>0.297</td>
<td>0.629</td>
<td>0.659</td>
<td>0.753</td>
<td>0.782</td>
<td>0.753</td>
<td>0.733</td>
<td>0.733</td>
</tr>
<tr>
<td>S_appl_BD</td>
<td>0.890</td>
<td>1.000</td>
<td>0.787</td>
<td>0.857</td>
<td>0.764</td>
<td>0.783</td>
<td>0.575</td>
<td>0.633</td>
<td>0.711</td>
<td>0.781</td>
<td>0.725</td>
<td>0.715</td>
<td>0.670</td>
</tr>
<tr>
<td>S_BD_val</td>
<td>0.801</td>
<td>0.787</td>
<td>1.000</td>
<td>0.843</td>
<td>0.721</td>
<td>0.769</td>
<td>0.665</td>
<td>0.673</td>
<td>0.671</td>
<td>0.717</td>
<td>0.694</td>
<td>0.670</td>
<td>0.670</td>
</tr>
<tr>
<td>S_appl_BD_org</td>
<td>0.835</td>
<td>0.857</td>
<td>0.843</td>
<td>1.000</td>
<td>0.775</td>
<td>0.802</td>
<td>0.610</td>
<td>0.675</td>
<td>0.733</td>
<td>0.799</td>
<td>0.730</td>
<td>0.716</td>
<td>0.648</td>
</tr>
<tr>
<td>S_BD_val_org</td>
<td>0.777</td>
<td>0.764</td>
<td>0.721</td>
<td>0.775</td>
<td>1.000</td>
<td>0.909</td>
<td>0.607</td>
<td>0.671</td>
<td>0.688</td>
<td>0.745</td>
<td>0.705</td>
<td>0.650</td>
<td>0.648</td>
</tr>
<tr>
<td>S_man_appl_BD</td>
<td>0.797</td>
<td>0.783</td>
<td>0.769</td>
<td>0.802</td>
<td>0.909</td>
<td>1.000</td>
<td>0.635</td>
<td>0.507</td>
<td>0.736</td>
<td>0.765</td>
<td>0.730</td>
<td>0.669</td>
<td>0.669</td>
</tr>
<tr>
<td>S_man_BD_val</td>
<td>0.629</td>
<td>0.675</td>
<td>0.685</td>
<td>0.610</td>
<td>0.607</td>
<td>0.635</td>
<td>1.000</td>
<td>0.675</td>
<td>0.688</td>
<td>0.660</td>
<td>0.696</td>
<td>0.696</td>
<td>0.682</td>
</tr>
<tr>
<td>BD targ real</td>
<td>0.629</td>
<td>0.673</td>
<td>0.675</td>
<td>0.671</td>
<td>0.763</td>
<td>0.875</td>
<td>1.000</td>
<td>0.660</td>
<td>0.703</td>
<td>0.703</td>
<td>0.730</td>
<td>0.669</td>
<td>0.682</td>
</tr>
<tr>
<td>S_tech_BD</td>
<td>0.753</td>
<td>0.711</td>
<td>0.671</td>
<td>0.733</td>
<td>0.688</td>
<td>0.726</td>
<td>0.638</td>
<td>1.000</td>
<td>0.848</td>
<td>0.848</td>
<td>0.833</td>
<td>0.803</td>
<td>0.754</td>
</tr>
<tr>
<td>S_BD_tech_val</td>
<td>0.782</td>
<td>0.781</td>
<td>0.717</td>
<td>0.799</td>
<td>0.745</td>
<td>0.765</td>
<td>0.660</td>
<td>0.703</td>
<td>1.000</td>
<td>0.833</td>
<td>1.000</td>
<td>0.833</td>
<td>0.803</td>
</tr>
<tr>
<td>S_appl_BD_comp</td>
<td>0.753</td>
<td>0.725</td>
<td>0.694</td>
<td>0.730</td>
<td>0.706</td>
<td>0.729</td>
<td>0.696</td>
<td>0.736</td>
<td>0.823</td>
<td>0.833</td>
<td>1.000</td>
<td>0.794</td>
<td>1.000</td>
</tr>
<tr>
<td>S_BD_comp_succ</td>
<td>0.713</td>
<td>0.715</td>
<td>0.670</td>
<td>0.716</td>
<td>0.650</td>
<td>0.682</td>
<td>0.648</td>
<td>0.689</td>
<td>0.745</td>
<td>0.803</td>
<td>0.794</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Sig. (1-tailed)</strong></th>
<th>S_appl_BD</th>
<th>S_BD_val</th>
<th>S_appl_BD_val</th>
<th>S_appl_BD_org</th>
<th>S_BD_val_org</th>
<th>S_man_appl_BD</th>
<th>S_man_BD_val</th>
<th>S_targ_BD</th>
<th>BD targ real</th>
<th>S_tech_BD</th>
<th>S_BD_tech_val</th>
<th>S_appl_BD_comp</th>
<th>S_BD_comp_succ</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_appl_BD</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_BD_val</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_appl_BD_org</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_BD_val_org</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_man_appl_BD</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_man_BD_val</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BD targ real</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_tech_BD</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_BD_tech_val</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_appl_BD_comp</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>S_BD_comp_succ</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Employees’ reactions to IT-enabled process innovations in the age of data analytics in healthcare

Hillol Bala
Department of Operations and Decision Technologies, Indiana University, Bloomington, Indiana, USA, and Viswanath Venkatesh
Department of Information Systems, University of Arkansas, Fayetteville, Arkansas, USA

Abstract
Purpose – Interorganizational business process standards (IBPS) are IT-enabled process specifications that standardize, streamline, and improve business processes related to interorganizational relationships. There has been much interest in IBPS as organizations from different industries implement these process standards that lead to successful organizational outcomes by integrating and standardizing intra- and interorganizational business processes. These process standards enable data analytics capabilities by facilitating new sources of interorganizational process data. The purpose of this paper is to unearth employees’ reactions to a new type of supply chain process innovations that involved an implementation of new IBPS, a supply chain management (SCM) system, and associated analytics capabilities.

Design/methodology/approach – The authors gathered and analyzed qualitative data for a year from the employees of a healthcare supplier, a high-tech manufacturing organization, during the implementation of a SCM system and RosettaNet-based IBPS.

Findings – In what the authors termed the initiation stage, there was quite a bit of confusion and unrest among employees regarding the relevance of the new process standards and associated analytics capabilities. With the passage of time, in the institutionalization stage, although the situation improved slightly, employees found workarounds that allowed them to appropriate just part of specific processes and the analytics capabilities. Finally, once routinized, employees felt comfortable in the situation but still did not appropriate the new supply chain processes faithfully. Overall, employees’ reactions toward the SCM system and associated analytics capabilities were different from their reactions toward the new business processes.

Originality/value – The paper contributes to the literature by offering novel insights on how employees react to and appropriate process innovations that change their work processes.

Keywords Data analytics, Supply chain management, Business process change, Interorganizational business process standards, Interorganizational systems, RosettaNet

Paper type Research paper

Introduction
In today’s hypercompetitive global economy, efficient and effective execution of interorganizational business processes (e.g. supply chain processes) is a key to firm performance (Tang and Rai, 2014; Rai et al., 2006; Venkatesh and Bala, 2012). Organizations implement information technologies (ITs) to make these processes more effective and efficient (Fosso Wamba, 2012; Fosso Wamba, Akter, Edwards, Chopin, and Gnanzou, 2015; Rai and Tang, 2010; Rai et al., 2006). In recent years, organizations have started building analytics capabilities to further improve their operations and business processes (Chen et al., 2012; Fosso Wamba, Akter, Colman, and Ngai, 2015; McAfee and Brynjolfsson, 2012). Healthcare, the largest sector in the USA, lags in terms of leveraging IT to improve organizational processes, including supply chain processes (Fosso Wamba and Ngai, 2013;
Min, 2014; O’Neill, 2007; Sinha and Kohnke, 2009). In fact, ineffective supply chain processes is a major concern for healthcare administrators because of its adverse effect on the quality of care and healthcare costs (Brody, 2007; Fosso Wamba, 2012; Min, 2014; Sinha and Kohnke, 2009). Given that healthcare costs (about $3 trillion) represent about 18 percent of the US GDP and is expected to reach about 20 percent of the GDP by 2022 (approx. $5 trillion) and 25 percent of these costs are supply chain related (CMS, 2015; Wettstein, 2014; World Bank, 2015), ineffective supply chain management (SCM) can have substantial financial implications. Hence, improving supply chain processes in healthcare is of great importance to various stakeholders in the healthcare industry (e.g. professionals, administrators, policy makers, and customers). It has been suggested that implementation and use of IT-enabled SCM systems, process innovations, and analytics capabilities are keys to improving healthcare supply chain (Brody, 2007; Burns, 2002; Fosso Wamba, 2011; Min, 2014). Therefore, successful implementation of IT-enabled SCM systems and processes in healthcare is of importance to researchers and practitioners alike.

Researchers have suggested many different designs for supply chain and means to improve supply chain processes in organizations in general (e.g. Cachon and Fisher, 2000; Lambert and Cooper, 2000; Vonderembse et al., 2006) and healthcare organizations (HCOs) in particular (e.g. Burns, 2002; Min, 2014; Neumann, 2003; Sinha and Kohnke, 2009). Interorganizational business process standards (IBPS) – open specifications for integrating and automating collaborative business processes using IT – have been suggested as a new way to improve supply chain performance and analytics capabilities (e.g. Bala and Venkatesh, 2007; Gosain et al., 2003; Markus et al., 2006; Nelson et al., 2005; Zhao et al., 2005; Venkatesh and Bala, 2007, 2012). Further, organizations implement enterprise systems (ES) to support supply chain activities (e.g. Bala, 2013; Mabert et al., 2003) that further facilitate big data capabilities by supporting data sourcing, storing, and use. Implementation of IBPS and ES often requires changes in existing business processes. Employees typically have strong emotional reactions toward process change initiatives, including process standardization (Long, 2004; Melone, 1995). Due to the complexity of the healthcare industry because of the presence of many different stakeholders and professional groups, we expect that implementation of IBPS will be critical as well as a significant challenge for HCOs (Kaplan and Robeznieks, 2014). We suggest that IBPS and associated analytics capabilities can lead to greater supply chain performance only if employees who are responsible for executing interorganizational processes and performing analytics favorably react to these changes. Therefore, understanding employees’ reactions during IBPS and ES implementation is critical for the successful implementation and use of these process standards and systems.

Much prior research on SCM has been at the organizational level (e.g. Malhotra et al., 2005; Rai et al., 2006; Tang and Rai, 2014), including various supply chain issues, such as optimization, effectiveness, or success of supply chain and its determinants, supply chain relationship, and the role of IT in supply chains (see Venkatesh, 2006). Notwithstanding such prior work, there is little or no research that examined the effect of new supply chain process standards and analytics capabilities on employees who are responsible for executing supply chain processes in the healthcare industry. Similarly, prior work on business process changes has not focused much on employees’ reactions to process change. This research has focused primarily on two areas (see Sarker and Lee, 2002): process design research (e.g. Basu and Blanning, 2003; Malone et al., 1999; Pentland and Feldman, 2008; Pentland et al., 2011) and process implementation and management (e.g. Ravichandran and Rai, 2000; Tang and Rai, 2014; Venkatesh and Bala, 2012; Wang and Tai, 2013). Although both streams provide rich insights on the nature and design characteristics of business processes and how organizations may (re)design and implement them, there has not been much research on how employees react to
business process changes and adapt new business processes in HCOs. Our research questions thus are:

RQ1. How do employees react to supply chain process standardization in the context of healthcare supply chain?

RQ2. How do these reactions unfold over time?

To answer these questions, we conducted a year-long qualitative study of the implementation of an ES and RosettaNet-based IBPS in a healthcare manufacturer and supplier organization. As these IBPS were independent of the ES, we understand the unique reactions to the new business processes vis-à-vis the ES that supports these processes. We build on structuration theory (Giddens, 1984; Orlikowski, 1992; Walsham, 2002) to understand the relationships among the processes, ES, and employees in a healthcare supplier.

Conceptual background
In this section, first, we review the literature on business processes and its relationship with ES followed by a discussion of the importance of process standardization in healthcare. Next, we discuss structuration theory in the context of ES implementations and business process changes.

Business process and ES
A business process is defined as a specific ordering of work activities across time and place, with a beginning, an end, and clearly identified inputs and outputs (Basu and Blanning, 2003; Davenport, 1993). Business processes are the fundamental building blocks for organizational value chain activities, such as SCM, logistics, and marketing. Examples of enterprise-level business process are: order fulfillment, application processing (e.g. loan applications), new product development, customer services, inventory management, and financial planning (Davenport, 1993; Smith and Fingar, 2003[1]. Business processes are important for several reasons. First, business processes are the foundation for corporate strategy and the basic unit of competitive advantage (Grant, 2002; Ray et al., 2004; Stalk et al., 1992). They are the mechanisms by which organizational resources and capabilities (e.g. IT initiatives, data analytics capabilities) provide ultimate value, such as positive economic outcomes (Davenport, 1993; Ray et al., 2004). Second, clearly structured and routinized business processes can improve organizational reliability, generate customer value, and improve interorganizational relationships (Davenport, 1993; Gosain et al., 2003; Hannan and Freeman, 1984; Rai et al., 2006). Third, the effectiveness of business processes is believed to be a more accurate measure of corporate performance as opposed to a global firm-level measure because an organization may possess competitive advantage at the level of business processes that may not be reflected in the organization’s overall performance (Ray et al., 2004). Process metrics are even more important for HCOs as these organizations are typically organized around processes, such as clinical, administrative, financial, and supply chain. Finally, various forms of business process innovations, such as process improvement, process reengineering, process redesign, process optimization, process integration, and process standardization, are keys to the success of much organizational activities (Davenport, 1993; Gosain et al., 2003; Hammer and Stanton, 1999). In recent years, these process innovations are complemented by analytics initiatives due to the abundance of data being generated and captured, and availability of analytics tools and capabilities (Chen et al., 2012; McAfee and Brynjolfsson, 2012). Although process innovation failures are often reported, a careful and systematic approach to IT-enabled process innovation could help obtain various positive outcomes, particularly for HCOs as they attempt to reduce variation, uncertainty, and ambiguity in their processes (Burns, 2002; Davenport, 2000).
ES and business processes are closely related (Davenport, 1993, 1998; Mabert et al., 2003; Markus and Tanis, 2000) because business processes are enabled or constrained by ES (Broadbent et al., 1999; Davenport, 2000; Hammer and Stanton, 1999). Although other types of IT (e.g. word processor) can affect how employees execute their tasks, HCOs implement ES to automate and support different clinical and administrative processes (Devaraj and Kohli, 2000; Soh et al., 2000). However, implementation of these systems typically involves substantial changes in existing IT infrastructure and an extensive redesign of or changes to existing processes (Davenport, 1998, 2000; Gattiker and Goodhue, 2002). For example, a SCM system can provide a structure for the flow of cross-functional and interorganizational activities (e.g. sourcing, procurement, shipping) and enables collaboration, planning, execution, and coordination of the entire supply chain for HCOs (Davenport, 2000), thus requiring radical changes in the existing business processes associated with these activities. Due to such changes, the ramifications of an ES for HCOs are more complex and far-reaching than that of simpler applications, such as productivity tools or transaction processing systems (Bingi et al., 1999; Nandhakumara et al., 2005).

**Process standardization in healthcare**

The importance of process standardization for HCOs has been underscored in the academic and practitioner literatures (e.g. Brody, 2007; GS1 US, 2015; Kaplan and Robeznieks, 2014; PricewaterhouseCoopers, 1999). Despite the availability of technologies needed to standardize business processes, the healthcare industry lags in terms of leveraging these technologies to standardize business processes. Healthcare is one of the sectors that can dramatically improve the quality of service and efficiency of service delivery through process standardization. Process standardization can help HCOs reduce variation, uncertainty, and ambiguity during process execution (Kaplan and Robeznieks, 2014). It will offer a clear guideline for the orchestration of sequential activities performed within a process (Davenport, 2005). It will create a single face of the organization for the external stakeholders, such as patients, suppliers, insurance provides, and other agencies. A standardized process requires accurate and timely information and clear guidelines on the flow of events. IT can provide required information and guide through the flow of events by giving notification and maintaining audit trail of events. Thus, when integrated with IT infrastructure and capabilities, the standardized and integrated processes can ensure consistent and efficient continued healthcare services.

Although process standardization is critical for HCOs, these organizations face several unique challenges in adopting and implementing process standards, such as IBPS for SCM (Ramanujam and Rousseau, 2006). First, the healthcare industry is a complex economic sector comprising organizations with multiple and often conflicting missions. Although the core mission of these organizations is to provide safe, effective, timely, and equitable patient care, they differ significantly in terms of their emphasis on clinical care, community service and outreach, teaching, research, profits and in some cases, religious values. These factors dictate organizational culture and propensity to invest and embrace innovations. For example, a non-profit HCO may not be interested in spending millions to implement IT-enabled IBPS. Second, HCOs typically consist of multiple professionals who socialized in significantly different settings (Ramanujam and Rousseau, 2006; Venkatesh et al., 2011). These professionals possess qualitatively distinct set of goals and professional values (Garman et al., 2006). For example, healthcare administrators who have a degree in business may not share the same values that physicians and other clinical stakeholders share (Garman et al., 2006). Even when physicians become administrator, they do not necessarily share the same values that other administrators in and outside the organization may share due to different organizational and professional identification that physicians possess (Hekman et al., 2009). Similarly, healthcare providers may not have the same disposition toward adopting and implementing IBPS that healthcare suppliers may have. Third, HCOs face a complex industry environment comprising multiple
external stakeholders who influence them in myriad ways (Ramanujam and Rousseau, 2006). For example, healthcare providers are influenced by various governmental agencies, insurance providers, accreditation agencies, and professional associations. Therefore, adoption and implementation of IBPS that spans multiple stakeholders are daunting tasks and require complex decision making by different stakeholders (e.g. Hess et al., 2006).

Finally, the task environment in the healthcare industry is complex, ambiguous, and dynamic (Ramanujam and Rousseau, 2006). Although the clinical practices are indeed complex, the supporting processes, such as financial, administrative, and supply chain, are also complex and ambiguous due to the presence of multiple stakeholders that may or may not share the same value system and culture. The issue of a complex task environment is even more pertinent to the implementation of IBPS as these standards dramatically change the existing task environment related to SCM. For example, after the implementation of IT-based IBPS, employees in a healthcare supplier who used to receive orders from healthcare providers through phone or fax will receive orders real time and have to process the orders within a certain time period dictated by IBPS. Given that the quality and accuracy of order fulfillment is of immense importance in the healthcare supply chain, employees may feel that their workloads have increased significantly following the implementation of IBPS. Consequently, they may develop negative reactions toward these process standards. Our focus in this research is to understand the employees’ reactions to process standardization in HCOs.

**Structuration theory**

The structuration[2] and related theoretical perspectives (e.g. human agency, organizational learning, and practice lens)[3] help us understand the dynamics of the relationships among technology, human agents, and social contexts (Black et al., 2004; Orlikowski, 1992; Walsham, 2002). These perspectives provide explanations for the recursive and dynamic relationships between human agents (e.g. employees) and social structures (e.g. technology and business processes) and provide insights on how one influences the other through the process of appropriation and enactment. Giddens’s (1984) structuration theory has been employed as an appropriate grounding and lens to study the interactions of human agents and various social systems, such as technology and organizations. It offers a solution to the dilemma of choosing between agency and structure, subjective and objective, and micro and macro conceptions of any phenomenon and allows researchers to embrace both (Giddens, 1984). It posits that there is a reciprocal interaction of human agents and organizational structure or structural properties – a set of rules and resources that human agents produce or reproduce in their daily activities (Giddens, 1984; Orlikowski, 1992). It suggests that human actions are enabled and constrained by structural properties of social systems (e.g. organizations), and these structural properties are the result of previous actions by human agents (see Jones, 1997; Jones and Karsten, 2008; Pozzebon and Pinsoneault, 2005 for reviews). Human agents are purposeful, knowledgeable, reflexive, and active, and have the ability to transform the structural properties of a social system (e.g. organization) by altering the rules or the distribution of resources (Giddens, 1984). Such actions will produce various (un)anticipated consequences that influence human agents’ subsequent actions (Giddens, 1984).

Although the original structuration theory did not explicitly incorporate the notion of technology as a social system, subsequent research has done so (see Figure A1). Barley (1986) studied the changes in social interactions among human agents (e.g. radiologists, doctors) are directly affected by the implementation of a new technology – CT scanner – in HCOs. His work was an important step in understanding the recursive interactions among human agents, technology, and institutional properties in HCOs. Subsequently, Orlikowski (1992) explained how human agents (e.g. technology designers, users, and decision makers), technology, and institutional properties (e.g. structural arrangements, business strategies, culture, control mechanisms, and standard operating procedures) influence each other.
Following Gidden’s notion of the duality of structure, Orlikowski (1992) proposed the duality of technology to capture the recursive relationship between human agents and technology, which was viewed as interpretively flexible – i.e. “there is flexibility in the design, use, and interpretation of technology” (p. 409).

Understanding business processes using a structuration framework

The structuration perspectives help us understand and explain the dynamic interplay among human agents, business processes, technology (e.g. ES), and institutional properties, and how this interplay unfolds over time in organizations in general and HCOs in particular. Figure 1 presents a graphical view of business processes’ influence on human agents, technology, and institutional properties. Building on prior work (e.g. Orlikowski, 1992), it presents our position on the recursive relationships among business processes, technology, human agents, and institutional properties in the healthcare setting. We argue that this reciprocal interaction between human agents and business processes can be independent of the interaction between human agents and technology. Further, we argue that business processes will have a recursive relationship with institutional properties in contrast to technology which is suggested in prior research to have a one-way influencing path with the institutional properties (see Orlikowski, 1992).

The thick lines in Figure 1 are the focus of this paper – how business processes related to SCM can influence and be influenced by human agents and institutional properties. The thin continuous lines represent structurational model of technology as described in Orlikowski (1992). As those aspects of structuration theory are well-understood from prior research, we do not delve into those details. As noted earlier, in the context of ES (e.g. ERP), business processes and technology are tightly coupled via the dotted line between the two in the picture – we discuss this briefly here. In an ERP context, the technology is designed to support specific business processes that are designed by the vendor as best practices and are fairly standardized across implementations of the particular vendor. Further, the technology and the business processes enable and/or constrain each other. For example, an ES may reduce complexity of a process by automating certain components of the process. In contrast, a business process that requires coordinated activities performed by different employees can increase the perceived complexity of an ES as employees may have to use some features of the system that are not easy to use (Venkatesh et al., 2003). This is particularly true for healthcare employees who have to coordinate with many internal and external stakeholders. Although this specific recursive relationship is important, this is not

---

**Figure 1.**
Business process in the structuration framework
the focus of this research. Our primary focus is to understand the recursive relationships between business processes and human agents.

Notwithstanding the tight coupling between ES and business processes as discussed earlier, we suggest that human agents (i.e. healthcare employees) can assign different meanings to business processes (e.g. SCM processes) and ES and they may enact different aspects of process and system. For example, when organizations implement new business processes along with an ES, it is possible for employees to appropriate the system faithfully and resist the new process or vice versa. For example, in the context of an IT implementation in HCOs, Lapointe and Rivard (2005) found that healthcare employees developed different levels of resistance to the new technology and business processes. Malhotra et al. (2001) found that virtual team members exerted differential efforts to processes and technology in R&D projects. In the context of SCM, an employee may feel the need for a SCM system and may find the system useful in his or her job. However, he or she may be reluctant to process a customer order within three hours (assuming it is a business process requirement) because he or she may find it difficult (or unfair from a workload perspective) to process an order in such a short time. In this case, although the employee may develop favorable perceptions toward the ES as it is making his or her job easier in many respects, he or she may resist the new order management business process. This is a likely situation for most ES that supports cross-functional processes and IBPS that are standardized by vendors or industry consortia – e.g. RosettaNet (Bala and Venkatesh, 2007; Venkatesh and Bala, 2012). Bala (2013) studied SCM processes using the socio-technical systems theory and discussed how employees (i.e. human agents) develop perceptions of SCM process characteristics or structures. Therefore, our objective is to understand the recursive interactions of supply chain processes with human agents and institutional properties in a healthcare context.

Although an ES facilitates (or hinders) the execution of business processes, they are conceptually distinct. We explain this distinction from three different theoretical perspectives. First, the business process literature suggests that organizations may first design business processes (for existing processes, the design connotes design for innovation or improvement) and then implement an ES to automate/support these processes (e.g. Davenport, 1993; Smith and Fingar, 2003). This temporal separation suggests that business processes even if they are enabled by an ES are conceptually distinct from the system[4]. Second, prior research on work processes and organizational routines (e.g. Feldman, 2000; Malone et al., 1999; Pentland, 2003a, b) has suggested that work processes may have various types of variations (e.g. improvisation, ad hoc combination, substitution, shortcuts, unexpected delays) depending on how human agents attempt to expand, repair, or strive these processes to achieve certain outcomes. Although prior research (e.g. Boudreau and Robey, 2005; Orlikowski, 2000; Orlikowski, 1992) has suggested that individuals appropriate the technology, such appropriation is usually at the feature level (Griffith, 1999; Jasperson et al., 2005). Therefore, technology and business processes are distinct in terms of how human agents appropriate them and the implications of such appropriation (see Bala and Venkatesh, 2013). Finally, resource-based view suggests that a technology (e.g. an ERP system) may not be a unique resource for an organization, but a business process can be a unique resource or capability that may not be easily imitated (Ray et al., 2004). For example, many organizations can implement the same ERP (e.g. SAP) system but, as reported in much prior academic and trade press articles, these organizations may not achieve similar benefits (e.g. competitive advantage) because their business processes are different.

We suggest that there is a need to understand the unique relationship between business processes and human agents. Although prior work has incorporated business processes as a part of the structural features of organizations (e.g. rules), a type of institutional property (e.g. operating procedures), and an aspect of norms (e.g. rules that define the organizationally sanctioned way of executing a work), that research has not explicitly separated business
processes from technology structures. Recent research (e.g. Bala and Venkatesh, 2013; Beaudry and Pinsonneault, 2005; Boudreau and Robey, 2005) provides examples of how human agents reacted toward various changes in their work methods (i.e. business processes) due to technology structures (e.g. complexity of the system). Robey et al. (2002) discussed how users reacted differently to changes in business processes and implementation of ERP systems. As noted earlier, Lapointe and Rivard (2005) found that individuals resisted changes in the business processes over and above their resistance to the technology itself in three HCOs. This suggests that business processes are separable from the technology that supports them and human agents can influence, appropriate, and enact each differently. As noted earlier, although an employee may appropriate a technology faithfully, he or she may not execute a business process that he or she finds cognitively burdensome. Another reason for understanding the unique role of business processes over and above the technology that supports them is the emergence of IBPS in today’s IT and business environments (see Davenport, 2005 for a discussion of process standardization). There have been several IBPS, advanced by organizational consortia – that have garnered great interest among researchers and practitioners alike – e.g. RosettaNet (Bala and Venkatesh, 2007; Gosain et al., 2003, 2004/2005; Malhotra et al., 2005; Venkatesh and Bala, 2007, 2012). With the sustained importance of understanding business process change for the past several years and the growth of process standards as a phenomenon, there is a need to separate business processes from technologies and to understand the interaction of business processes with the other components in the structuration theory framework.

There are a few reasons prior research on structuration theory will help in studying technology-driven business process changes (i.e. standardization) in organizations, especially in HCOs. First, like any other social system, a business process has rules and resources that constitute the structure of the process (see Feldman and Pentland, 2003). By definition, a business process consists of a sequence of activities that need to be accomplished in a specific way – i.e. rules (Davenport, 2000; Pentland, 2003a, b). In addition, a business process requires certain resources (technology, people, and skills) for proper execution. Second, like technology, business processes have duality and are interpretively flexible. The duality of business processes is manifested in the following way: on one hand, business processes are designed and executed by human agents in organizations; and, on the other hand, once institutionalized business processes can impose constraints on human agents’ interaction with the processes. The flexibility in business processes is underscored in prior research that has suggested that even though a business process consists of a sequence of activities, these sequences are not necessarily structurally fixed (Feldman, 2000; Feldman and Pentland, 2003; Pentland, 2003a, b). Pentland (2003a, p. 858) noted that business processes “are better conceptualized as generative structures that can produce a wide variety of different patterns or sequence of events.” Further, business processes consist of task sequences, with occasional variations, exceptions, and/or shortcuts, and there is typically an opportunity for improvisation or ad hoc combinations. Third, business processes and human agents can have a recursive relationship similar to the relationship between human agents and other social systems. Similar to the way human agents build certain interpretive schemes, facilities, and norms into technologies (Orlikowski, 1992), it is possible that they will build the same into business processes in their day-to-day execution of such processes[5].

Method
We adopted an inductive exploratory approach to understand how business processes related to SCM play a role in the dynamics of human agents, technology, and institutional properties suggested in Figure 1. Pozzebon and Pinsonneault (2005) suggested several methodological strategies for empirically applying structuration theory, such as the use of a grounded theory approach, provide a narrative, and develop visual mapping
(e.g. graphically tracking and comparing sequence of interactions and actions), and
temporal bracketing (e.g. temporal sequence of actions in organizations), for dealing with
the duality of structure and interplay between the micro and macro. Our empirical
approach is generally consistent with these strategies as we employed a grounded theory
approach and used visual mappings and temporal bracketing to understand the
employees’ reactions to business process changes over time. Using a grounded theory
approach (Strauss and Corbin, 1998), we collected qualitative data using semi-structured
interviews of employees of one organization that implemented a new SCM system and
analytics tools along with a set of IBPS. Our approach was similar to prior exemplars of IT
implementation research employing a grounded theory approach (e.g. Boudreau and
Robey, 2005; Maznevski and Chudoba, 2000; Orlikowski, 1993). Following the tradition of
this approach, we incorporated various theoretical perspectives from structuration theory
that are pertinent to the ideas emerging from the data (Boudreau and Robey, 2005; Strauss
and Corbin, 1998). Given the novelty of the context (i.e. concurrent implementations of
IBPS and an SCM system), a qualitative approach helped us develop rich insights on the
context and employee reactions (Venkatesh et al., 2010, 2013, 2016).

Research site
Our study site was a hardware manufacturer – HealthSup, Inc. (fictitious name) – that
implemented a new supply chain module in its ERP system and RosettaNet-based IBPS – i.e.
partner interface processes (PIPs) for SCM. HealthSup is a US-based manufacturer that
produces and supplies high-tech equipment for HCOs. Its customers include distributors of
high-tech equipment and large HCOs in the USA and other countries. It had a fairly flat
organizational structure with a large number of employees being engineers and designers of
hardware circuitries. Based on the many visits of one of the authors to the organization, we
gathered that the organization operated in a highly competitive environment and realized
the need to constantly respond to the changes in the market. Further, there was great
pressure on the organization to meet customers’ demands with very fast turnaround times.
All these pressures pushed the organization to streamline its supply chain processes and
other interorganizational business activities. The general organizational climate was geared
toward innovation in products and market responsiveness.

Technology
Prior to the implementation of the new SCM system, HealthSup implemented an integrated
ERP system to automate some of their internal business functions, such as finance and human
resources. However, the top management realized that there were certain inefficiencies in their
interorganizational business processes. As noted earlier, many HCOs do not have IT-enabled
supply chain processes and HealthSup had to maintain different parallel processes to meet the
requirements of each of its customers. For example, HealthSup’s order management process
included various modes of communication such as websites, fax, e-mails, telephone calls, and a
proprietary electronic data interchange (EDI) system to support the need of different HCOs.
The customers were able to order via any of these channels and HealthSup employees
responded to the customer inquiries using these different communication channels. Moreover,
the transactions were handled through the EDI and legacy database systems. The processes
were inefficient and the technology was not very responsive to the environmental demands.
In order to make HealthSup’s order management functions more efficient and market oriented,
top management decided to implement an integrated SCM module in their ERP system along
with RosettaNet-based IBPS to develop seamless supply chain business processes. The top
management also decided to include analytics tools and capabilities offered by the vendor to
capture high volume and variety of data (both from the SCM system and externally) and
perform real-time analytics.
Supply chain business processes

The supply chain module was implemented by an external IT solution provider who was affiliated with the ERP vendor. The RosettaNet PIPs were also implemented by the same IT solution provider. RosettaNet is an industry consortium of major technology-related and logistics organizations. It develops industry-wide, open business process standards (i.e. PIPs) for interorganizational relationships. Given that our purpose was to understand the interplay among business processes, technology, human agents, and institutional properties, studying an organization that had adopted RosettaNet-based IBPS provided an appropriate context to isolate the implementation of technology from the business processes. RosettaNet PIPs define business processes between trading partners by specifying the activities, decisions, and roles for each partner involved in a particular business activity (GS1 US, 2015). Each PIP includes a business document with the vocabulary and a business process with the choreography of the message dialog (GS1 US, 2015). A detailed description of PIPs is available at GS1 US (2015). GS1 US, is a neutral, not-for-profit, international organization that develops and maintains standards for supply and demand chains across multiple sectors, supports the ongoing maintenance and implementation of the RosettaNet standards through membership in the GS1 US Partner Connections Program (GS1 US, 2015). Figure A2 presents an example PIP.

RosettaNet PIPs provide specifications for interorganizational or public business processes. Public business processes involve interactions with trading partners (e.g. exchange of business messages), whereas private business processes are internal to the organizations (e.g. interaction with internal back-end systems). It is important to note that standardization of these processes requires substantial changes to the internal private processes in order to respond to the needs of the trading partners efficiently. Moreover, interorganizational processes are executed by employees from the respective trading partners. Therefore, implementation of these processes will impact not only the interorganizational relationships, but also the employees responsible for the execution of the interorganizational and internal processes. In the case of HealthSup, implementation of the new process standards significantly altered the previous order management process – both public and private portions of the process. Employees who were responsible for managing HealthSup’s order management process were greatly affected by the implementation of both the new system and processes. Our focus was on the employees who went through this change.

During the time of our data collection, HealthSup implemented three PIPs to streamline its order management processes with its major distributors and customers. Because several of the other distributors as well as customers had already implemented some of these PIPs, HealthSup’s plan was to utilize these PIPs to improve the supply chain processes internally and for customer interactions. The specific PIPs are described in Table I. These three PIPs were closely related as they were responsible for distinct aspects of HealthSup’s SCM function.

<table>
<thead>
<tr>
<th>PIP</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIP3A4: request purchase order</td>
<td>The purpose of this PIP is to enable a buyer to issue a purchase order and obtain an immediate response from the supplier that acknowledges the status of the order (e.g. which of the purchase order product line items are accepted, rejected, or pending)</td>
</tr>
<tr>
<td>PIP3A6: distribute order status</td>
<td>The purpose of this PIP is to enable a seller to send the status of a product order to a buyer. For example, a product line item may be backordered, shipped, or canceled. Order status is distributed when an open purchase order exists</td>
</tr>
<tr>
<td>PIP3B2: notify of advance shipment</td>
<td>This PIP allows a shipper to notify a receiver that a shipment has been assigned. This notification contains detailed product-level information about a shipment (e.g. when a shipment is expected to arrive)</td>
</tr>
</tbody>
</table>

Table I. PIPs implemented by HealthSup
Implementation of these PIPs required substantial changes to HealthSup’s overall SCM function. For example, HealthSup originally updated the order status on a specific web-based system designed for its customers and later transferred the data to its ERP system. HealthSup anticipated that it would be able to eliminate this redundant web-based system as the customers would have access to the order status information through the new PIPs. With the implementation of the new SCM module and RosettaNet-based IBPS, HealthSup was able to eliminate many redundant steps from its previous supply chain processes.

Data collection
We collected data from HealthSup over a period of 12 months. Our source company – the independent solution provider that implemented the RosettaNet standards at HealthSup – allowed one of the authors to closely observe the implementation of the PIPs and interact with HealthSup employees who were directly affected by the changes. Before and during the implementation phase, the author spent extended periods of time in the business unit where the changes were being implemented. Therefore, we gained substantial knowledge regarding the old processes and system and understood how the new system and processes would alter HealthSup’s SCM. Immediately after the deployment of the IBPS (which took about four months after the deployment of the SCM module), series of interviews of HealthSup employees who were responsible for executing various aspects of SCM were conducted. The employees interviewed were from multiple levels of the organizational hierarchy, with a few being responsible for operational work (e.g. receive and validate purchase orders), others were supervisors (e.g. approve purchase orders), and two were executives in the business unit. Several additional interviews were conducted over the next eight months in order to understand employees’ continued interactions with the new system and business processes. In addition to the interviews, we collected project documents, organizational memos, and press releases related to supply chain module and RosettaNet PIPs implementations.

A total of 25 employees were interviewed over the 12-month period. We chose a small set of employees so we could interview them multiple times and develop a rapport with them over the course of the study. Getting to know the interviewees well was important to ensure their candor. On average, each employee was interviewed three times during our data collection. Interviews were semi-structured in nature with more open-ended questions in the early interviews to understand employees’ general reactions toward the new system and processes, and lasted from 30 to 60 minutes. The interview script was consistent with Bala and Venkatesh (2007). The early interviews started with general questions regarding employees’ interactions with the new system and processes, their general assessment, and evaluations of the system and processes. Subsequent questions were dictated by their early responses. The later interviews had more specific questions regarding the use of the new system, business processes, and changes in day-to-day activities.

Data analysis
Following the guidelines of Strauss and Corbin (1998), we analyzed the data using three coding procedures: open, axial, and selective. Open coding is the process of breaking down, comparing, conceptualizing, and categorizing the qualitative data from the interview transcripts (Boudreau and Robey, 2005). The important step in open coding is to compare various incidents, events, quotes, and instances to find similarities and dissimilarities (Strauss and Corbin, 1998). We compared the responses from the interviews to identify similar text segments. We coded these similar text segments into meaningful categories. We used axial coding to further group these categories. This grouping was primarily based on the conceptual similarities of the categories. Finally, selective coding, the process of integrating and refining the theory (Strauss and Corbin, 1998), was used to formulate a
coherent story line from the findings (Boudreau and Robey, 2005). In selective coding, we integrated all the major categories identified in axial coding to form a larger theoretical scheme. We iterated between the theory and data and incorporated various ideas, concepts, and theoretical perspectives from prior literature to understand and explain the employees’ reactions. We continued the data analysis until theoretical saturation, the point at which diminishing returns are obtained from new data analysis or refinement of coding categories (Strauss and Corbin, 1998), was reached. We stopped the analysis when no new categories were emerging and new text segments could be placed into the existing codes and categories. In addition to analyzing the interview data, we reviewed internal corporate documents, memos, and press releases that we gathered during our data collection. This review provided greater insights on the context and helped us interpret and understand the interview data thoroughly.

Results

In this section, we discuss how the interaction of human agents, business processes, and institutional properties unfolded over time. We conceptualize these temporal dynamics using the following three stages: initiation, institutionalization, and routinization. These stages are consistent with the prior research on technology implementation in organizations (e.g. Cooper and Zmud, 1990; Orlikowski, 1992). These stages capture the temporal dynamics of the human agents’ recursive relationships with technology and business processes. Table II summarizes the key categories and subcategories across the stages.

Initiation

The initiation stage began right after the deployment of the new system and RosettaNet PIPs. As we mentioned earlier, it took about four months for HealthSup to implement the PIPs after the implementation of the SCM module. During this time, HealthSup offered formal training to the employees who would be using the system and the accompanying business processes. The training was conducted by a contracted training provider appointed by the solution provider that implemented the SCM module and the RosettaNet PIPs. In addition to the formal training, the IT department offered documentation and tutorials to make the users comfortable with the new system and business processes. Even though the IT department was particularly happy about the quality and breadth of the training, the reactions from the employees were mixed. Some employees were overwhelmed by the complexity of the new processes and technology. Each process required an understanding of the complex flow of activities and the orchestration of these activities. For example, for some activities within the new processes, employees had to respond within a certain timeframe. These restrictions added a significant burden on the employees. A few employees commented that due to the implementation of new processes, their job had transformed to “a whole new job” and the training was not adequate to perform their tasks. The training did not help improve individual and collective efficacy with respect to the new processes and technology. In sum, although some employees found the training to be helpful, others found the training to be inadequate:

Training was very helpful. I needed it. Some others are better at learning things on the fly. That’s not me.

How can we do a whole new job with so little training?

I think the training is insufficient. Sure there are change management people around but they don’t know my job as well as I do. Not to mention, my job has changed.

From a learning perspective, we found two alternative mindsets on the employees’ part: exploration and exploitation. Some employees were exploring the new technology (including

...
the analytics tools and capabilities) and processes, whereas the others were reluctant to explore. We found that the employees who did not explore had limited themselves to a set of basic functionalities of the new system, i.e. only exploitation. Organizational learning and process management literatures suggest that both exploration and exploitation are important to achieve greater productivity (Benner and Tushman, 2002; March, 1999). We noticed that both groups of employees expressed certain interests in the new processes for different reasons. On the one hand, the employees who were involved in exploration were excited about the new processes as they believed that they would be able to discover new aspects of the processes that would enhance their job performance. On the other hand, the employees who were in favor of the routine aspects of the new processes believed that the new processes would reduce their “guess work.” They believed that the new process will reduce performance variations and increase efficiency:

I am a dabbler. I have experimented with the technology quite a bit. I am coming to terms with how the process will flow and am discovering that it’s gonna be great.

The new processes have many inefficiencies in my view. I find it better to pick and choose certain elements of the process. The rest I just ignore.

<table>
<thead>
<tr>
<th>Influence</th>
<th>Initiation</th>
<th>Institutionalization</th>
<th>Routinization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human agents and business processes</td>
<td><em>Enactment</em> Resistance Unwillingness to follow the new process Avoidance Shadow system Process-technology compatibility</td>
<td>Workarounds Shortcuts Exploration Improvisation</td>
<td>Reinvention Commitment Extended use Integrative use Emergent use</td>
</tr>
<tr>
<td>Job changes</td>
<td>New skills Deskilling responsibility Autonomy Reengineering</td>
<td>Task variety Interdependence Deskill Reskill</td>
<td>Empowerment Task identity Job transformation</td>
</tr>
<tr>
<td>Learning</td>
<td>Training Complexity Exploration Exploitation</td>
<td>Adaptable Exploitation Knowledge Flexibility</td>
<td>Readaptability Exploitation Innovation Expertise</td>
</tr>
<tr>
<td>Business processes and institutional properties</td>
<td>Relational boundary Changes in organizational structure Information flow</td>
<td>Social interaction Formal interaction Trust</td>
<td>New social relationship Independence Relational separation</td>
</tr>
<tr>
<td>Organizational culture</td>
<td>Receptiveness to innovation Risk taking Resistance to change</td>
<td>Control Innovation Stability</td>
<td>Responsiveness Inflexibility Outcome orientation</td>
</tr>
<tr>
<td>Institutional pressure</td>
<td>Norms Relationship with external stakeholders</td>
<td>Relationship with external stakeholders Organizational mandate</td>
<td>New relationship with external stakeholders Coordination</td>
</tr>
</tbody>
</table>

Table II. Key findings
It reduces the guess work and unnecessary thinking on my part. Now, I can focus on doing the [my] best.

As noted earlier, employees who were responsible for order management functions anticipated that their day-to-day activities were going to change. Some employees were under the impression that the new system and associated process changes would drastically alter the way they managed orders and interorganizational relationships. Although some users were looking forward to the changes, others were quite anxious. The primary characteristic of these initial reactions toward the process changes was resistance. We found that employees were unwilling to follow the steps in the new processes as they were reluctant to accept the changes – “I don’t see any reason to change the way we did things.” Given that the old system was still available, many employees were still using the old system. Other employees had started to use the new system, but were still following the old processes. This suggested that although employees were willing to accept the new system, they were reluctant to execute the new processes (e.g. Malhotra et al., 2001):

For as long as I have been here, we have always done it the old way. I am having a hard time really following the new rules. They are so regimented.

There are several aspects of the process that are easy to avoid. I use the system but I skip various steps in the other activities.

I have a choice really. I can use the old system for many things still since we have both systems.

One of the key reasons for such strong resistance to the new processes was probably employees’ perceptions about their job changes due to the new technology and processes (e.g. Bala and Venkatesh, 2013; Venkatesh et al., 2010). The work practices had changed significantly. Employees had to respond quickly which required the ability to make quick decisions, greater job responsibility, and autonomy. We found that although some employees had to acquire new skills to be able to follow the processes, some were concerned about the potential deskilling. Similarly, job responsibility and autonomy for some employees increased, although others anticipated a potential decrease:

It is like having two jobs. One old job which helps me get stuff done. One new job with all the new flow, process, and technology. It’s overwhelming.

The expanded list of duties is good. I was starting to feel bored with my old job.

I am not sure I can handle all the new things I have to do.

We found that employees’ interactions with the new business processes were influenced by HealthSup’s institutional properties, such as relational boundaries (e.g. organizational structure, informational flow, and social interactions), organizational culture (e.g. innovativeness) of the unit responsible for SCM, and various institutional pressures (e.g. norms, relationship with external stakeholders). We noticed that employees were deeply concerned about the potential changes in the organizational structure. Although this change may not be a formal overhaul of the existing hierarchical structure (even though some employees anticipated such changes), it clearly represented a change in the decision making processes, degree of supervision, and more importantly, the information flow:

Many redundant and bureaucratic steps are eliminated. Of course, I am sure we have added new ones. I suspect the org chart may look different. It should.

My job is different but so is the organization. I have told my wife that I have changed my job and my employer without changing my job and employer, if that makes any sense.

The people I have to talk to now are different. The information is coming from new places and it is new information. It is going to new places and it is also new type of information. I am at a loss.
We found that employees’ perceptions regarding the existing organizational culture also influenced the way they acted upon the new processes. Some employees perceived that HealthSup was not very receptive to changes and, therefore, the changes associated with new processes might not work favorably in the long run. Other employees found the implementation of new processes to be risky but they thought the reward could potentially be high:

Our organization is a large beast. I would be quite surprised if all these changes work in our case.

Our products are usually built on an element of high risk and high reward. I see our process changes here with all new technology to be quite the same.

Although the top management of HealthSup was quite clear in terms of their intention to make the new technology and processes mandatory, some employees had the perception that the new processes were not really mandatory. Because of the perceived efficiency of the old processes, they were more inclined to follow the old processes using the new technology. As employees started to execute the new processes about four months after the deployment of the new system, some employees thought that the new processes were a recommended, but not a required, way of managing orders. A few employees were even surprised that the management wanted them to abandon the old ways of communicating with the representatives of their major customers. Further, given that many external stakeholders (e.g. boundary spanners at HCOs and distributors) preferred the old manual processes (according to some HealthSup employees), these employees were concerned about whether the new processes would adversely affect their relationships with these external stakeholders. Because of all these reasons, the initiation stage was ridden with confusion in terms of new processes that affected the relationship with external stakeholders:

I see the technology use as required. However, I see the process as quite a different thing. It is a recommendation.

My rapport with our external stakeholders will deteriorate because we don’t have to interact as much anymore.

**Institutionalization**

About two months following the implementation of new processes, the top management realized that contrary to expectations and goals, the order management processes had become inefficient and problematic. Employee morale was low and the number of undesired incidents (e.g. miscommunication, failure to deliver the product on time) had increased. This prompted the top management to strictly enforce the use of the new technology and the processes. In an official memo circulated to all employees responsible for order management activities, the management called for a strict following of the new order management processes. We refer to this stage as institutionalization as the new technology and the processes were institutionalized (e.g. widespread adoption and use) during this stage.

Following the circulation of the memo, most employees started to use the new technology and follow the new processes more than they did before. Employees knew that technology use could be easily monitored, but the execution of processes could not as easily be monitored. Therefore, the employees used the new system, but attempted to workaround the processes (Bala and Venkatesh, 2016; Boudreau and Robey, 2005). For example, many employees were uncomfortable accepting a purchase order without getting verbal approval from their supervisors, even though supervisors could approve a purchase order electronically in the new system. At the same time, some employees were exploring the new...
system and processes to find shortcuts so that they could avoid certain steps and directly input data into the system. Many employees improvised and found alternative ways to accomplish the same objectives of the new processes:

I have had some time to figure out how to circumvent some aspects of the process. It’s good and no one notices, which is better.

I expected there to be inefficiencies [in the process] and there are. I have found ways to get around the problems and now it’s all better and quicker.

I follow the process to the tee and then some. I have been able to find some ways to do things even more effectively. Give it time. So will others.

During this stage, many employees learned various aspects of the technology and processes and became proficient in both. They discovered various ways to use features (e.g. various reports and predictive analytics) of the new system to improve their effectiveness. However, some employees were still struggling to master the system, new analytics tools, and processes. These employees were concerned about the rigidity of the process and blamed the process for their apparent inefficiency and inflexibility:

I think there are so many new and good reports that are available on the system. I have started to find them and use them to improve my job performance. I hope it is noticed at review time.

I have mastered the ins and outs of my job now. I can do things quite well by following the structure.

I am still fairly clueless about the process. There have been so many changes and so many new things and touch points for my job.

Now that the employees figured out their own ways of doing things using the system and processes, some employees discovered that their job had become more interesting and they have acquired some useful skills. In contrast, other employees were concerned that their job had become less important as they felt that their responsibility and autonomy had reduced:

I have learned some new and marketable skills with being familiar with the business processes of [vendor]. I have also learned some new technologies.

I like my new job. I am doing new things.

I thought the job would get better. It has gotten worse, a lot worse.

During this stage, the institutional properties were heavily influenced by the employees’ interactions with the new processes. Given that the employees were forced to use the new processes, they had to stop interacting with many employees with whom they used to interact. However, they were also establishing new social networks as they started to interact with various other groups of people (e.g. IT support people). The employees who mastered the new processes and technology found that they had become somewhat central to the social network as others had started to come to them to get help (e.g. Sykes, 2015; Sykes and Venkatesh, 2016; Sykes et al., 2014). Some employees noticed that unexpectedly the interaction with the boundary spanners from the customers had improved as both parties had a clear understanding of the status of a given order or shipment:

The new processes have crippled my social circles.

I have gotten to know some new and powerful people in the org.

I think I understand the workflow and process better than most. So, I am able to help a lot of people now.

I think we are doing better by our external stakeholders with our new and improved processes.
Routinization

With their ongoing interaction with the new technology and processes, the employees became very familiar with both. The technology and processes were no longer new to the employees and had become parts of the organizational work systems (Tyre and Orlikowski, 1994). We called this stage routinization that started approximately six months after the deployment of new processes. During this phase, some employees realized the improvement and expressed their commitment to the new order management processes. Some employees had started to explore the processes to accomplish the objectives that were not originally intended. However, some employees noted that they were not able to explore the process due to feature constraints of the ES. From a learning perspective, many employees became experts on the new processes as they were attempting to use the system, analytics tools, and the processes for more integrative and emergent tasks (Cooper and Zmud, 1990):

I was a skeptic. I have come around now. I think everything is working well. I like it.

I think the processes are great. The technology works well with the process most of the time. However, I don’t think we can innovate. It takes the ideas out of the hands of the people.

Due to their mastery of the new processes and technology, we found that some employees felt that they became powerful. One key reason for this empowerment is their ability to exploit the system and processes to get accurate and timely information regarding the status of an order (Bala, 2013). At the same time, the employees felt that their jobs had finally transformed into something more meaningful. They felt a greater cognitive fit with the new system and processes. They developed a different sense of task identity as they had a clear understanding of their task boundaries and job duties (Venkatesh et al., 2010):

I like being as powerful as I am now because of the central role I occupy in the process and the information I have access to.

While I was a skeptic early, I like the fact that my job now has a much clearer delineation. I am not responsible for screw-ups before or after the work gets on my desk.

I have some skills now. Previously, I was pushing paper. Now, I have some [vendor] system experience. What I do now is much much better and far more important.

From the perspective of institutional properties, the routinization stage was characterized by the emergence of new intraorganizational and interorganizational relationships and significant changes in the organizational culture in terms of customer orientation. As we found in the previous two stages, the relational boundary changed significantly due to the implementation of new processes. In the routinization phase, we found that the relational separation increased and employees became more independent, whereas the relationship and the degree of coordination with the external stakeholders had increased substantially:

I am rather happy that I have greater wiggle room. I can do as I see fit.

I am disappointed that I don’t see many people that frequently. The social fabric has changed dramatically.

I think we coordinate better with our touch points on practically everything due to tighter ties and the time freed up for interactions exclusively with them.

Organizational culture had become more customer oriented but inflexible. Given that the ES coupled with the new processes and analytics capabilities made it possible to respond to customers’ needs quickly and effectively (e.g. accurate decision making due to real-time analytics capabilities), employees gradually became used to such responses. As the process standards diffused, the entire unit became more outcome oriented. Supervisors were no longer recognizing the subordinates’ successful execution of the processes. Over time, this
led to the unit becoming inflexible and closed to innovation and improvement of processes. Also, employees were reluctant to work with customers who did not use the RosettaNet standards:

I have said all along, if we can serve customers better, I am game. After a year, I can unequivocally say that the answer is yes.

To me, it has boiled down to a simple tradeoff. Standardization or innovation. We have gone with standardization at the expense of potential for innovation. We have to do it one way, we are not going to be nimble and flexible.

It’s all about outcomes now. I mean, it is about the new process but the new process has so many metrics that it is really about outcomes.

Figure 2 presents a process view of how employees act upon new business processes and technology and how these actions unfold over a period of time. The figure shows that the human agents’ enactment of business processes and technology (including analytics tools and capabilities) changes over time. Starting with strong resistance, the enactment reaches a point of mixed yet greater reinvention and commitment. However, this does not indicate a sign of stability as human agents continue their enactment through exploration. They appropriate the business processes by continuously creating and recreating the structural properties of business processes. At the same time, business processes mediate the human agents’ action by enabling or constraining their job. Further, human agents’ enactment of business processes is influenced by the institutional properties and at the same time influences various aspects of the institutional properties. In sum, our findings clearly suggest the recursive relationships between technology-enabled (or constrained) business processes and human agents and institutional properties, respectively. More discussion of human agents’ enactment of business processes is provided in the next section.
Discussion

Our objectives were to understand how employees react to changes in supply chain processes and how these reactions unfold over time. We studied process changes associated with RosettaNet-based IBPS and an ES module implementation at HealthSup Corporation—a healthcare equipment supplier. Our key finding was that employees reacted differentially toward business processes from technology (i.e. ES), suggesting that although employees may enact technology in accordance with the spirit of the technology, they may not enact a new business process faithfully or vice versa. We also found that employees attempted to improvise the new business processes (e.g. workarounds, shortcuts, avoidance). Although process variation is common in HCOs and other organizations (e.g. Pentland, 2003a, b), we suggest that the employees’ enactment of business processes related to supply chain in various unanticipated ways may potentially result in unintended consequence with respect to the successful execution of healthcare supply chain processes.

Human agency and enactment of processes and analytics capabilities

Our results illustrate that healthcare employees’ enactment of supply chain business processes is different from that of the technology (i.e. ES and analytics tools and capabilities) that enables (or constrains) these processes. Drawing on Giddens’ (1984) structuration theory, we discussed a basic conjecture that business processes are the social entities with structural properties (rules and resources) that are continuously produced or reproduced by the recursive interaction of human agents. We found this recursive interaction between HealthSup’s employees and the new supply chain process standards. In particular, we found that the human agents’ enactment of these processes (re)created new aspects of work practices characterized by workarounds and improvisations in SCM activities. Such work practices enabled (and/or constrained) the employees by changing the meaning of their job. Prior research attributed misalignments between technology functionality and work processes as a source of workarounds (Markus and Tanis, 2000). However, we found that in a healthcare context, human agents were engaged in workarounds and improvisation (at least initially) even when the technology was closely supporting the processes, suggesting that healthcare employees have differential levels of resistance to business process changes and IT implementations. These findings are consistent with the prior research on organizational routines (organizational routines are “repeated patterns of behavior that are bound by rules and customs and that do not change very much from one iteration to another” (Feldman, 2000, p. 622). It emphasized the role of agency and collective learning in explaining the situated changes in the routines (e.g. Feldman, 2000). Drawing on Feldman (2000) and Feldman and Pentland (2003, p. 105), we suggest that business processes have both ostensive (i.e. structural properties) and performance (i.e. agency) aspects, “with the performance creating and recreating the ostensive aspect and the ostensive aspect constraining and enabling the performances.” Our results extend this theoretical perspective by studying complex IT-enabled business process changes.

From a temporal perspective, we found that initial resistance toward the process changes evolved over time as healthcare employees were continuously involved in various (re)creations of enactment of the new processes. Although the institutionalization phase was dominated by various workarounds and improvisation, the routinization was characterized as reinvention and continued commitment. This general pattern of enactment was consistent with the prior technology implementation research (Beaudry and Pinsonneault, 2005; Boudreau and Robey, 2005; Lapointe and Rivard, 2005; Tyre and Orlikowski, 1994). Yet, our findings were quite contrary in terms of how healthcare employees reacted to the new technology vs the new processes. More specifically, prior research suggested that users were reluctant to use the technology immediately after the deployment (e.g. Lapointe and Rivard, 2005; Sykes et al., 2011), whereas we found that employees were strongly opposed to
the new processes, although not as opposed to the technology (including analytics tools and capabilities). We suggest that due to the complexity in the healthcare industry as discussed earlier, employees have a greater degree of resistance to new processes. Given that employees in this industry have to deal with different heterogeneous stakeholders and professional groups, and operate in an environment where mistakes and errors can be fatal, they are more likely to persist in routines that they have developed over time and have been successful. They, however, are receptive to the new analytics tools and capabilities because they believe these tools and capabilities can help them avoid mistakes and make effective decisions. The changes in enactment over time can be explained through Feldman’s (2000) suggestion that human agents could alter the routines based on the outcomes of previous iterations. She suggested three responses to organizational routine changes: repairing the routine to produce the intended outcomes, expanding the routine to take advantage of the new possibilities, and striving to attain something that is difficult. We found that initially employees attempted to repair the processes, and with time, they attempted more striving and expanding (e.g. willingness to maximize benefits from the new analytics capabilities).

We noticed significant variations of enactment by different employees at different stages. For example, some employees were involved in exploration, whereas others were more exploitive when appropriating the new supply chain processes. This can be explained using the practice perspective suggested by Orlikowski (2000) who found that different groups of individuals enact a structure of practice differently depending on their role, job responsibility, experience, and other contextual factors. Moreover, Bourdieu’s (1977) theory of practice suggested an improvisational nature of practice, indicating that practices are influenced by rules and expectations and a course of action within a given practice is to some extent novel no matter how constraining the situation is. HealthSup’s employees were engaged in different enactment moves that were novel and distinct from each other. Thus, our results complement prior IT research employing practice and human agency perspectives (e.g. Boudreau and Robey, 2005; Orlikowski, 2000; Schultze and Orlikowski, 2004) by illustrating how the “notion of enactment” is different for business processes that are enabled (or constrained) by technology (e.g. new systems and analytics capabilities).

Business processes and institutional properties
Consistent with the idea that technology cannot directly change institutional properties (e.g. Orlikowski, 1992), we argue that business processes and associated analytics capabilities cannot causally influence institutional properties. It is the human actions on business processes and appropriations of new analytics capabilities that act on institutional properties and (re)create new structures for organizations. In the case of HealthSup, various enactments of business processes and analytics capabilities produced changes in relational boundaries and other institutional properties. With the human agents’ continued interaction with business processes and associated changes in enactments, the structuration of institutional properties also changed. This is consistent with Orlikowski (1996) who suggested that, over a period of time, organizational change occurred through the ongoing, gradual, and reciprocal adjustments, accommodations, and improvisations enacted by the human agents on their work practices. Barley (1986, 1990) and Black et al. (2004) discussed how technology implementation influenced expertise and relational boundaries. We noticed similar patterns in social interactions and relational boundaries in HealthSup. Although human actions on business processes enact upon institutional properties, other aspects of institutional properties influence these actions. For example, employees’ perceptions about the receptiveness to innovation and risk taking constrained their enactment of business processes and use of analytics tools and capabilities.
Theoretical contributions

Our study makes several key contributions. First, we contribute to the research on SCM in healthcare by highlighting the importance of examining employees’ reactions to supply chain process changes. As noted earlier, prior SCM work has primarily focused on organizational-level phenomena. Although there has been a vast body of work on supply chain designs, there is little or no work on how employees react to new supply chain processes, particularly in healthcare contexts. Our findings suggest that regardless of how efficient a supply chain process is, employees may not execute it faithfully if they feel that the process changes their jobs and routines. Even if a supply chain process is mandated, employees may find workarounds and shortcuts to bypass the steps in the process. SCM research has seen substantial theoretical advances in recent years (e.g. Ketchen and Hult, 2007; Miles and Snow, 2007). However, this research has traditionally overlooked the employees’ reactions to SCM practices and designs. We believe that there is a need to incorporate the relationships between human agents (e.g. employees) and supply chain processes, and between institutional properties and supply chain processes into theories of SCM to capture employees’ reactions toward these processes. Healthcare is a unique context to understand these relationships due to the complex nature of this industry (Garman et al., 2006; Ramanujam and Rousseau, 2006). Further, given that supply chains are becoming more global and complex, examining these relationships will help researchers develop a holistic theoretical understanding of supply chain design and arrangements.

Second, we incorporated business processes in the complex dynamics suggested in the structurational models of technology (e.g. Orlikowski, 1992). Although much prior research has incorporated the notion of business processes in technology or institutional properties, we argued that business processes, when enabled or constrained by an ES, are distinct social entities upon which human agents can enact different from how they enact the system. We found that human agents faithfully appropriated an ES but did not faithfully appropriate new processes as a part of their enactment. We thus extend prior research by providing a rich understanding of the agency perspective on technology-initiated business process changes. Our study revealed reciprocal interactions between business processes and human agents, and pointed to the changes in institutional properties resulting from human agents’ actions on business processes. Further, we advance our understanding of how human agents recursively produce and reproduce structures in social entities (e.g. business processes) through their day-to-day activities.

Our research has three important implications for research on business processes. First, prior research in this area has primarily focused on process reengineering best practices. It does not provide a rich understanding of how process changes influence employees and how employees’ execution of certain aspects of a process can change certain properties of an organization. We found that employees’ jobs changed substantially from initiation to routinization with significant empowerment and job transformation occurring (or vice versa) as time elapsed. Thus, our research addresses the call for a wider view of complex technology implementation and business process reengineering that includes communication, people, and organizational structure (Grover et al., 1995). Second, our study focused on the implementation of process standards in healthcare contexts. Implementation of process standards can be important drivers for improved productivity and quality, reduced errors, successful interorganizational relationships, and better coordination in the healthcare industry (Burns, 2002; Davenport, 2005; PricewaterhouseCoopers, 1999). In the context of software development, prior research has found that coding standards have a positive impact on coordination (e.g. Maruping et al., 2009). By studying healthcare employees’ recursive interactions with process standards, we have provided additional theoretical depth in the area of process management. Similarly, this research offers important implications for organizational change literature.
(e.g. Van de Ven and Poole, 1995) by focusing on both micro (i.e. individual) and macro (organizational) aspects of process changes. It offers insights on managing organizational and technological changes in healthcare contexts. Finally, this research furthers our understanding of emergent structures of the norms, organizational structure, and formal and informal relationships in healthcare contexts. Although IT has been suggested as a driver for such emergent structures in HCOs, our results suggest that employees’ actions on new business processes (re)produce such structures recursively. Thus, organizational theorists could incorporate IT-enabled business process changes in their view of new forms of organizational structure.

Practical implications
RosettaNet PIPs represent best practices in interorganizational processes in the high-tech industry. Our research, however, suggests that implementation of best practice-based supply chain processes may not be well-received by employees in the healthcare industry because they may feel radical changes in their jobs and may not want to give up their existing routines and habits. In addition, they may feel changes in their relationships with their coworkers and external stakeholders. Our research suggests that managers responsible for implementing new supply chain design should be mindful about potential negative impact of process changes on employees. Given the recent call for IT-enabled process innovations in healthcare SCM (e.g. Brody, 2007; Long, 2005; Neumann, 2003), we suggest that careful consideration should be given by HCO managers when changing processes as employees may not faithfully execute new processes, thus resulting in inefficiencies and other unintended negative consequences. Compared to other organizations, this is a critical issue for HCOs that represent pluralistic settings characterized by the presence of multiple powerful internal and external stakeholders (e.g. physicians, nurses, administrators, pharmacists, government agencies, insurance companies, accreditation agencies) with divergent and even competing objectives. Although we are not suggesting that HCOs should not consider implementing IT-enabled supply chain process standards to reduce process variations and improve operational efficiency, we are suggesting that HCO managers should be more mindful during the implementation about potential negative reactions by its stakeholders and develop interventions and change management strategies to minimize negative reactions. Examples of such interventions and change management strategies include: gaining sponsorship and support from different stakeholder groups; creating multidisciplinary implementation team with representatives from different stakeholder groups, particularly employees who will execute the new processes; training for new processes; and aligning and integrating new processes with internal processes (e.g. clinical and administrative processes) so that employees and other stakeholders develop a clear understanding of how new processes fit the overall business model (Bala and Venkatesh, 2013, 2016; Sykes and Venkatesh, 2016).

Recent research has suggested that the effectiveness of business processes is the appropriate gauge of organizational performance (Ray et al., 2004). It is thus important that HCOs carefully manage their business processes to achieve efficiency and effectiveness. Managing process change initiatives will be crucial for HCOs in the future (Sykes et al., 2011). Our findings suggest that employees may initially react to process standards negatively which may reduce efficiency. However, as time goes by, the negative outlook may disappear. Still, some employees will continue to find ways to avoid new processes. Therefore, managerial tracking of business processes is very important. Although tracking is difficult to implement, it is possible to adopt standards such as TQM or six HealthSup to track process execution. Finally, our research provides practical implications for assimilation of complex technology (e.g. ES) research. Prior research has suggested the process misfit is responsible for numerous complex technology implementation failures in
organizations in general and HCOs in particular (Bala and Venkatesh, 2013; Barley, 1986; Lapointe and Rivard, 2005; Robey et al., 2002; Soh et al., 2000). Healthcare managers need to be more proactive in terms of assessing the implications of IT implementation from a process management perspective. In order to foster employees’ acceptance of process changes, managers need to develop interventions, such as training programs, tailored toward emphasizing the benefits associated with these changes (Sykes, 2015).

Limitations and future research
We studied only one organization, a manufacturer in the healthcare industry, with a few specific process changes associated with SCM. It is possible that the employees’ reaction toward process changes will be significantly different in other HCOs and/or in other processes. In particular, employees may react differently to IBPS and SCM systems in a more traditional HCO, such as a hospital (e.g. Venkatesh et al., 2011). Research is thus needed on how individuals react toward and interact with new processes and how organizational structure is (re)created by the implementation of new processes in different types of organizations, including HCOs. Also, process characteristics (inter- vs intra-organizational, core vs sub processes) play a key role in the complex dynamics, as shown in Figure 1. Work is needed to understand how different aspects of a process can influence employees and organizational structures differently.

This work focused on the intraorganizational aspects of process changes. Although the new processes implemented by HealthSup Corporation were interorganizational in nature, we did not study the trading partners of HealthSup to understand how their employees reacted toward the new processes. This limits the scope of our understanding. Therefore, future research should study the adoption and impacts of interorganizational process standards. Further, more research is needed to understand the process changes associated with different types of IT. For example, process changes associated with an ERP system could be considerably different from that of a CRM or SCM system. Furthermore, we only studied the employees of a particular unit of HealthSup. We believe that in order for us to understand the overall impacts of process changes, it is important to study other key stakeholders, such as top management, customers, suppliers or buyers (Rai et al., 2010; Setia et al., 2013). Therefore, future research should incorporate different stakeholders’ perspectives to understand business process changes.

Another important future research direction would be to study the different types of interventions that can potentially influence employees and other stakeholders to have favorable view toward process standards. Although prior research has suggested various interventions associated with IT adoption and use (e.g. Venkatesh and Bala, 2008), future research can examine interventions to foster adoption of new business processes or process changes in HCOs. We believe that carefully designed and orchestrated interventions can improve employees’ reactions toward process changes. These interventions can help increase employees’ trust in new business processes and technology. Finally, we did not study the outcomes of the deployment of new processes – this will be important next step because if the outcomes are not positive, it will be difficult to justify process standardization efforts, such as IBPS implementations (Rai et al., 2010).

Conclusions
Our objective was to understand the employees’ reactions to IBPS and a new SCM system implementation in healthcare contexts that enabled data analytics capabilities. We found that employees of a healthcare manufacturer differentially reacted toward new supply chain business processes, new SCM system, and data analytics capabilities. Our findings regarding the dynamic unfolding process of the relationship between human agents and material aspects of new technology structures (e.g. new system, processes, and analytics
BPMJ 23,3

Note 1. Although process concepts have long been used by researchers to understand the characteristics of organization structure, work role behavior, and resource interdependence, it was only in the 1990s that the concept of a process gained significant attention due to an increased focus on business process reengineering (BPR) initiated by large organizations (Grover et al., 1995). Since then, a process view of business has gained prominence as process innovation was recognized as a key driver of performance improvement in the face of intense competition, globalization, and demand idiosyncrasies (Davenport, 1993; Ray et al., 2004).

Note 2. Three basic perspectives have been employed to understand the impact of technology on individuals and organizations (see Black et al., 2004; Boudreau and Robey, 2005). First, the objectivist stance has taken an imperative or deterministic view of technology and proposed relatively static models of human behavior to study the influence of technology characteristics on human action or vice versa (e.g. Davis et al., 1989; Venkatesh, 2006; Venkatesh et al., 2003). Second, a subjective view has studied how the characteristics of “new technologies evolve as they are used and modified by people in the course of day-to-day activity” (Black et al., 2004, p. 573). Finally, suggesting that neither approach (i.e. objectivist or subjectivist) can provide a comprehensive account of the impact of technology on individuals and organizations, some have taken an integrative perspective that suggests that the causality runs in both directions: “technology influences the patterns of human activity, and the technology changes as it is modified in the course of day-to-day activity” (Black et al., 2004, p. 573). Drawing primarily on Giddens’ (1984) structuration theory, this stream has focused on the social context surrounding technology, human agents, and organizations (e.g. Barrett and Walsham, 1999; Jones, 1997; Jones and Karsten, 2008; Majchrzak et al., 2000; Orlikowski, 1992, 1993, 2000; Pozzebon and Pinsonneault, 2005; Walsham, 2002).

Note 3. See Appendix 1 for more about different theoretical perspectives.

Note 4. For example, a healthcare supplier may design a business process for order management and then find a system that supports this newly designed process. In this case, the system may not support all aspects of the newly designed process (i.e. sequence of activities by which an organization receives, processes, and fulfills customer orders) that may require some manual tasks or tasks supported by a different system. This suggests that a business process has distinct attributes (e.g. rules, resources, and constraints) beyond the systems that support the process.

Note 5. For example, human agents involved in an interorganizational process (e.g. supply chain process) will accumulate specific knowledge regarding the trading partners (i.e. interpretive schemes), utilize various interorganizational systems technologies (i.e. facilities), and follow a specific communication mode (i.e. norms) while executing the process.

References


**Further reading**


**Corresponding author**

Viswanath Venkatesh can be contacted at: vvenkatesh@vvenkatesh.us
Appendix 1. Theoretical perspectives related to the structuration theory

Poole and DeSanctis (1990) and DeSanctis and Poole (1994) proposed the adaptive structuration theory (AST) that focused primarily on the structures embedded into technologies and how human agents interact with such structures. AST suggests that human agents appropriate the structure inscribed in technologies either faithfully or unfaithfully depending on the degree to which their use of technology corresponds to the structures embedded in a technology. Orlikowski (2000) suggested a practice perspective to understand technology in organizations emphasizing the emergent structures of technology as opposed to stable, predictable, or embedded structures. The perspective she advanced emphasizes enactment of as opposed to appropriation of technology. The essential thesis of this perspective is that human agents will enact a particular technology “in particular ways in particular conditions” (Orlikowski, 2000, p. 407) and continuously create and recreate the structural properties of the technology. The human agency perspective suggests that humans are relatively free to enact technologies in different ways (Boudreau and Robey, 2005; Emirbayer and Mische, 1998). Like the practice perspective, the human agency perspective posits that technology structures are not stable or embedded. Humans can use technology minimally, invoke it individually or collaboratively, and improvise in ways that produce novel and unanticipated consequences. This perspective posits against treating technology as a determinant of social change. Rather, technology is a vehicle of social change at the discretion of human agents.

Appendix 2

Notes: (1) Human agents (e.g. developers, users) build into technologies certain structures (e.g. rules, resources) and appropriate technology by assigning shared meaning (arrow a); (2) technology can mediate human agents’ action by enabling or constraining their performance (arrow b); (3) human agents are influenced by various institutional properties when they act (e.g. designing, appropriating, modifying, resisting) on technology (arrow c); and (4) when human agents use technology, they can reinforce or transform institutional properties of an organization (arrow d)
Appendix 3

The process of issuing a purchase order typically occurs after checking for price and availability (PIP3A2), requesting quotes (PIP3A1) or transferring shopping carts (PIP3A3). The process of issuing a purchase order may be followed by changing the purchase order (PIP 3A8), canceling the purchase order (PIP3A9), querying for purchase order status (PIP3A5) or disturbing purchase order status (PIP3A6).

**Figure A2.**
RosettaNet business process standard (PIP3A4)

*Source: Reproduced from: www.rosettanet.org*
Big data analytics: transforming data to action

Daniel Bumblauskas
Department of Management, University of Northern Iowa, Cedar Falls, Iowa, USA
Herb Nold
Polk State College, Winter Haven, Florida, USA
Paul Bumblauskas
PFC Services, Inc., Marietta, Georgia, USA, and
Amy Igou
Department of Accounting, University of Northern Iowa, Cedar Falls, Iowa, USA

Abstract
Purpose – The purpose of this paper is to provide a conceptual model for the transformation of big data sets into actionable knowledge. The model introduces a framework for converting data to actionable knowledge and mitigating potential risk to the organization. A case utilizing a dashboard provides a practical application for analysis of big data.

Design/methodology/approach – The model can be used both by scholars and practitioners in business process management. This paper builds and extends theories in the discipline, specifically related to taking action using big data analytics with tools such as dashboards.

Findings – The authors’ model made use of industry experience and network resources to gain valuable insights into effective business process management related to big data analytics. Cases have been provided to highlight the use of dashboards as a visual tool within the conceptual framework.

Practical implications – The literature review cites articles that have used big data analytics in practice. The transitions required to reach the actionable knowledge state and dashboard visualization tools can all be deployed by practitioners. A specific case example from ESP International is provided to illustrate the applicability of the model.

Social implications – Information assurance, security, and the risk of large-scale data breaches are a contemporary problem in society today. These topics have been considered and addressed within the model framework.

Originality/value – The paper presents a unique and novel approach for parsing data into actionable knowledge items, identification of viruses, an application of visual dashboards for identification of problems, and a formal discussion of risk inherent with big data.

Keywords Big data, Decision making, Actionable knowledge, Dashboards

Paper type Conceptual paper

Introduction
Vast quantities of data often inundate modern organizations that use business processes developed in a past industrial era. Current technologies, such as expanded storage capability at low cost, allow organizations to produce and collect vast amounts of data, creating what we call the potential for a data binge as data is collected due to simplicity and not thoroughly analyzed. More is not necessarily better as vast data pools make it difficult to convert data into information in a timely fashion. This is particularly important now, as product and service life cycles have become shorter. The ability to analyze meaningful and relevant data and convert data to information, knowledge, and ultimately action in time to favorably influence an organization is a key competitive differentiator. In this paper,
we explore reasons for this condition and offer practical suggestions for agile companies in
the twenty-first century business environment by developing a conceptual framework to
convert big data to actionable knowledge. This model provides both a foundation
for scholarly research to build upon and provides industry practitioners with a set of tools
for analyzing big data to make better decisions. When considering the impact of amassing,
accessing, and storing large data sets we must consider the business opportunities,
but we must also consider necessary resources for the assessment of risks, including but not
limited to determining the presence of risk, virus identification and resolution, and internal
control requirements.

The primary contribution of this paper is in providing a conceptual model framework for
the transformation of large data sets into actionable knowledge while a secondary
contribution is a specific industry case where such methods are being utilized.
The theoretical model introduces the transitions from data to information to actionable
knowledge as well as the introduction of viruses across the organization. In this model,
a virus is defined as any outside influence that can lead to performance degradation,
i.e., consider a virus afflicting a patient and being able to remedy such a virus with
medications, precautions, time, etc. Dashboards are practical tools to help analyze large data
constructs within the model framework. Using a business process management approach,
a model has been developed for both strategic and practical application in operations and
production management and provides a perspective on how managers can address what we
refer to as a data binge and others have called “information overload” (Jacoby et al., 1974;
Power, 2013; Sela and Berger, 2011), and poor “data quality,” or the quality of the data
analysis process (Hazen et al., 2014). By working with an industry partner, an example of a
dashboard and a case study illustrate a visual tool within the model framework. This paper
provides a unique theoretical contribution for operations and business process managers
responsible for collecting, analyzing, and acting upon the information collected and
contained in large data sets.

The transitions required to reach the actionable knowledge state, virus identification
process, and dashboard visualization tools can all be further developed empirically by
academic researchers and immediately deployed by practitioners in industry. The focus
should be on taking action and continuously improving rather than amassing large
amounts of data solely because we have the resources and technologies available to
do so (i.e. quality over quantity). Information assurance, security, and the risk
of large-scale data breaches present contemporary challenges and these topics have been
addressed within the model framework. This paper presents a unique and novel
approach for parsing data into actionable knowledge items, identification of viruses,
which inhibit the organization’s ability to use large data effectively, an application
of visual dashboards for identification of problems and a formal discussion of risk
inherent with big data.

The historical context of big data and decision making
A chronological literature review traces the historical contexts of big data. The evolution of
big data began with Fredrick Winslow Taylor and the scientific management techniques
of the early 1900s with his world famous work “The Principles of Scientific Management”
(Winslow, 1911). Applying scientific management techniques required the accumulation and
analysis of detailed work-related data but was limited by the technology of the time. Willard
Brinton (1917) introduced data visualization, which serves as a precursor to more modern
data analytics and dashboards. During the Second World War and post-war era, the work of
Deming, statistical quality control, and the 14-point management method was followed by
total quality management during the 1940s and 1950s, which spawned the need for data to
support management decisions, and this need grew rapidly. Anderson et al. (1994) discuss
this in their paper in the *Academy of Management Review*. The introduction of digital computers in the 1930s and 1940s (Burks, 1989) later led to mainstream deployment of computing technology during the 1970s to mid-1980s enabling the collection of vast quantities of data with limited ability to get information and actionable knowledge from the systems. During this period, information overload was a problem in decision making (Jacoby *et al.*, 1974). During the mid-1980s to mid-1990s, the advance of distributed computing made data available across the organization and lead to the evolution of MRP and ERP systems. The internet made data available instantaneously in real-time across organizations, countries, and nations from the mid-1990s to present, driven by increased use of cloud-based systems and applications (e.g. Amazon Cloud, Microsoft SharePoint and SkyDrive/OneDrive, Google Drive, etc.). Over the past two decades, the field of business intelligence and analytics (BI&A) and special techniques for big data analytics (BDA) have become increasingly important and challenging with technological advances that allow the accumulation of larger and larger data sets from multiple sources, many of which did not exist 10-20 years ago (Chen *et al.*, 2012). Big data and business analytics has also prompted the need for specific analytical skillsets in terms of training and for educational purposes (Dubey and Gunasekaran, 2015).

Founded on well-established database management techniques, the initial stages of BI&A took a data-centric approach relying heavily on data collection, extraction, and analysis technologies to help make sense of large data sets (Chaurhuri *et al.*, 2011; Turban *et al.*, 2008; Watson and Wixom, 2007). One common approach for data collection by organizations has been the use of a variety of legacy systems with data archived in commercial relational database management systems (RDBMS). The techniques for analyzing data from RDBMS were developed using statistical methods in the 1970s and data mining techniques developed in the 1980s and popularized in the 1990s (Chen *et al.*, 2012; Wu *et al.*, 2006). Using statistical analysis techniques, major IT vendors including Microsoft, IBM, Oracle, and SAP incorporated a wide variety of business management tools into offerings (Sallam *et al.*, 2011), and more contemporary tools such as the statistical software R and visualization tool Tableau. RDBMS allowed the introduction of business performance management tools, such as scorecards and dashboards, as well as a range of reporting tools for statistical analysis, association analysis, regression analysis, predictive modeling, and more.

The widespread deployment and acceptance of the internet and World Wide Web in the early 2000s created massive amounts of readily available data. Building on traditional RDBMS-based product information databases, companies began collecting IP-specific user information in the background using cookies and server logs to identify customer needs and wants to identify new business opportunities (Chen *et al.*, 2012). Focusing on social media and crowd-sourcing systems companies developed web intelligence and analytics to help identify opportunities and drive general decision making (Doan *et al.*, 2011), supply chain and logistics decision making (Michaelides, 2016), and aviation business analytics (Huang *et al.*, 2016).

One of the main reasons that BDA has expanded in the last few years is the speed of data creation. Gartner (2015) estimates the number of connected devices will reach 20.8 billion by 2020. Techniques to analyze the massive, continuous stream of mobile, location-aware, person-centered, and context-relevant data from internet-enabled devices represents a still underutilized target area for BI&A (Chen *et al.*, 2012).

**So what exactly is big data and how can it be defined? A contemporary literature review**

This section addresses a review of the articles published on big data to help better define the terminology. Big data is an ambiguous and flexibly defined term often associated with the collection and analysis of “large” data sets. An article in Forbes lists a variety
of different definitions commonly used to describe what “big data” really means (Arthur, 2013). There is now a common use of the terms big data and BDA to describe huge data sets requiring advanced and unique data storage, management, analysis, visualization technologies (Chen et al., 2012) as well as statistical analysis. Originally, researchers defined the Three V’s of big data (Chern et al., 2015; Manyika et al., 2011; McAfee et al., 2012; Sun et al., 2015) and we now frequently describe the Seven Pillars or Four V’s of big data (Buhl et al., 2013; Dong and Srivastava, 2013; Hitzler and Janowicz, 2013; IBM, 2014; Kulkarni and Tulasidas, 2015), and Five V’s (5V’s) (Fosso Wamba et al., 2015; White, 2012) which are:

1. volume as data sets that are at least a petabyte in size;
2. velocity as the pace at which data flow in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc.;
3. variety as sources and types of big data are both structured and unstructured (e.g. free-from text, sensor data, graphics, audio, and video files);
4. veracity as uncertainty of data (IBM, 2014); and
5. value of the “[…] economic benefits from the available big data” (Fosso Wamba et al., 2015).

For the value component of the 5V’s, Fosso Wamba et al. (2015) provide a comprehensive review of the literature, types, and examples of value as well as issues related to big data cited by researchers. They also define BDA “as a holistic approach to manage, process and analyze the ‘5 Vs’ data-related dimensions (i.e. volume, variety, velocity, veracity, and value) in order to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantages” (Fosso Wamba et al., 2015; Ji-fan Ren et al., 2016). Buhl et al. (2013) built upon these big data taxonomies by expanding the importance of models to analyze big data sets while Dubey et al. (2016) look specifically at the impact of big data on sustainable manufacturing. Watson (2014) focuses on the improved decision making and insight that are possible with the evolution to BDA. With the sudden popularity of big data and BDA, models for high-level analyses of big data are becoming increasingly important.

Many of the newer machine learning techniques have increased the usefulness of acquiring large amounts of data. Machine learning allows connections to be discovered amongst many data points (Agarwal and Dhar, 2014). Without preconceived theory restricting the analysis, different correlations are found using computing power. Many times these connections would have been overlooked by researchers that are biased by their own hypotheses and theories. The computer systems analyzing big data are able to look past existing knowledge to make new discoveries (Waller and Fawcett, 2013b; George et al., 2014). It is this predictive nature achieved through the analysis of big data that is changing decision making in organizations. The additional value from big data is the use of these new insights in order to make better decisions and take action within the organization. A review of BDA in e-commerce analyzed recent literature on the business value in organizations using BDA (Akter and Fosso Wamba, 2016). The literature showed that BDA was used for personalization of services, improved customer service, and predictive analytics to demonstrate a few functions.

The literature has shown many potential uses and applications in various areas of organizations. Schoenherr and Speier-Pero (2015) found nearly 50 percent of supply chain professionals were using BDA or had plans to use BDA in the near future.
These professionals saw better decision making as one of the primary benefits of using BDA as a predictive analytics tool. In an examination of non-traditional data sources, Chae (2015) reviewed the Twitter feed for supply chain mentions and determined potential uses for supply chain professionals. Demand driven BDA has looked at the customer driving the logistics and supply chain or channel decision making of the organization (Bumblauskas et al., 2016; Michaelides, 2016). Financial audits can be more timely and more accurate when entire populations of transactions can be reviewed (Moffit and Vasarhelyi, 2013). These are just a few of the ways that BDA is already changing business decision making. Despite all of the opportunities available, many organizations still struggle with the best way to find value with DBA.

While the size, scope, and scale of data are difficult to limit in defining big data, the definition of big data must revolve around the analysis of the data rather than the actual size of the database or spreadsheet (i.e. large data sets or databases) since that still seems to be rather subjective or in the “eye of the beholder.” In fact, one of the major challenges in practice is how to limit the size and scope of the data set. Analysis of data and action based upon that information is the key to the process of effectively defining and utilizing big data. The ultimate objective of accumulating and analyzing data is to drive decision making and action while creating value across all levels of the organization. However, organizations often lag in utilizing the data that have been acquired. Sommer (2015) estimates that organizations have only analyzed 0.5 percent of existing data. This is because organizations consider time to be one of our most valuable resources and, hence, time to decision, with proper mitigation of risk (viruses), is an important element within BDA. With such large volumes of data virtually untapped, organizations that can succeed in creating time effective actionable knowledge can gain a distinct advantage over their competitors. This paper presents a model and framework for transformation from raw data to actionable knowledge within an organization.

There are many different general types of business decisions impacted by BDA. For example, to illustrate one particular example in business process management, analytics can impact accurate forecasting for sales, revenues, and production of goods and services which are imperative in business decision making. The Aberdeen Group (2004) discussed some of the challenges faced in this respect, one of which is having multiple or “no single” demand forecast or our data binge state. One of the problems with forecasting is information overload and another is groupthink (Heizer and Render, 2013). Power (2013) referenced the work of Herbert Simon (1974) in saying: “[…] the central problem will not be how to organize to produce efficiently, but how to organize to make decisions – that is, to process information. Big data means more processing of information and a greater need to organize to use the information in decision making” (Power, 2013). Power (2002) provides seven reasons why managers do not maximize decision support systems, one of which is “information overload,” which supports his reference to the work of Janis and Mann (1977) and O’Reilly (1980) stating, “[…] when the degree of complexity of an issue exceeds the limits of a person’s cognitive abilities, there is a marked decrease in the adequacy of human information processing that is a direct effect of information overload and ensuing fatigue” (Power, 2002). Sela and Berger (2011) also discuss information overload in the context of getting weighted down by trivial decisions (Sela and Berger, 2011) and reference Jacoby et al’s (1974) work which documented information overload back in the mid-1970s. Information overload, or possible data binges, paralyzes decision making and action. The rapid increase in the amount of data acquired and available for further analysis increases the problem.

In working with industry practitioners, we have identified that large industry databases have evolved into what is now “big data.” The ability to assemble huge data sets can drive companies to make sweeping statements or over-generalizations, such as “we can analyze
anything,” and often technology, particularly computer applications, allow us to have advanced analytical capabilities. This collection of data from all sources presents another challenge. Most think of data in columns, rows, and tables, such as in spreadsheets and databases. However, the breadth of data has expanded to include text data, such as social media, video and audio, requiring a change in mindset when analyzing data. However, having the ability to analyze in an unstructured ad hoc basis can cause paralysis by analysis and stifle action. This is particularly the case when one person or department creates the data and another must actually take action with the information provided. Access to more data does not necessarily lead to better decision making. One of the challenges of big data is the veracity that is apparent from combining this many sources. This uncertainty leads to possible lack of trust in the data, leading to further paralysis in decision making.

Chris Argyris (1996) introduced the term actionable knowledge and suggested that knowledge without causing action has limited value to organizations. Distilling value from the vastness of big data is a major challenge since only humans are capable of interpreting, integrating, and assimilating data to form new meanings that shape decisions. Many challenges exist making it difficult for humans to convert massive amounts of raw data into focused action. Industry analysts frequently observe that the volume of data is not the only challenge, but also the variety and velocity and that focusing primarily on volume leads to underutilization of the data value (Gartner Group, 2011) which is further explored in the next section. Jagadish et al. (2014) suggested that generating value from big data is a multi-step process consisting of data acquisition, information extraction and cleaning, data integration, modeling and analysis, and interpretation and deployment. Furthermore, companies tend to focus on one or two steps, at the expense of others, which degrades the value of BDA. Analysis, interpretation, and deployment are uniquely human steps in the process requiring humans to absorb data then combine it with other sets of data to visualize new meaning (Jagadish et al., 2014). Maximizing the value of big data therefore requires human interaction to create actionable knowledge. Russom (2011) observed that there is a shortage of professionals with the knowledge and skills needed to effectively manage the volume, velocity, and variety of big data. Additionally, Chen et al. (2012) observed that demand for individuals is knowledge of key perspectives for decision making: descriptive, predictive, and prescriptive analytics are increasing. Focusing attention on the human intervention or skills needed to convert data to action becomes increasingly important. There are also major implications for predictive implications in supply chain management as detailed by Waller and Fawcett (2013a).

The research objective of this work is to provide a theoretical framework to translate big data into useable information that will lead to improved decision making, action, and positive change. We have also provided industry cases for the use of the method and associated recommended tools.

**Methodology: the actionable knowledge and performance triangle model**

Data with no objective analysis, and knowledge without action, have relatively marginal value to organizations. Davenport and Prusak (1998) offered a useful description of the differences between data, information, and knowledge suggesting that increased knowledge has the potential to improve decision making. Understanding the differences between these three constructs and the transformational process of changing meaningless raw data into knowledge that drives action, as shown in Figure 1, is essential for success or failure in BDA.

Figure 1 provides an excellent and practical reference point for techniques that create action, mitigate risk, and create efficiency with the ultimate goal of identifying impactful ways to have a positive influence on the organization. Therefore, we suggest that at each interface point, responsible parties should consider the controls, timeliness of movement...
between each conversion step and who and how impactful ideas can be prioritized and implemented. Through this process, the organization can optimize knowledge conversion.

Data consists of facts about some event with little direct relevance or purpose. Data without context or reference point analysis has no meaning, but is essential for the creation of information. Humans give data meaning by adding context and reference points that are relevant and purposeful then communicating new information to a receiver. Interpretations by the receiver allow for a decision on whether the information has value (Davenport and Prusak, 1998). Individuals then combine and synthesize multiple pieces of information to create a higher level of understanding that adds value through action. Davenport and Prusak defined knowledge as follows:

A fluid mix of framed experience, values, contextual information, and expert’s insight that provides a framework for evaluating and incorporating new experience and information. It originates and is applied in the mind of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms (p. 5).

The primary difference between data and information remains that data is a collection and interpretation process while information generates knowledge to make decisions that drive action. Bellinger et al. (2011) defined the differences of data, information, knowledge, understanding, and wisdom. The annual review of the Editors of International Journal of Knowledge, Culture and Change in Organizations (2012) expressed the relationship between data, information, and knowledge as follows:

Knowledge is the process of connecting the stuff of the mind and the stuff of the world. It is not a recorded thing (data, information), or at least, it is not just that. Knowledge is a form of action (Editors of International Journal of Knowledge, Culture and Change in Organizations, 2012).

The concept of actionable knowledge is not new, having been discussed extensively in various contexts including behavioral science (Argyris, 1996), business management (Argyris, 1993), organization science (Cross and Sproull, 2004), and other-related disciplines. However, the concept of actionable knowledge has taken on additional meaning and importance in recent years with the vast expansion of data availability and the need for quick and effective decision making (Cao, 2012). Unlike academics who generate unique literary contributions focused on generating knowledge for the sake of knowledge, the judgment of performance for business executives are results that are outcomes of management decisions. Management decisions translate into results only if they generate action with positive outcomes that add value to the organization (Nold, 2013).

Chris Argyris (1995), who has been a key figure in popularizing the concept of actionable knowledge, defined actionable knowledge as follows: “information that actors could use, for example, to craft conversations that communicate the meanings they intend.
Actionable knowledge has to specify how to produce meanings but leave actors free to select the specific words" (p. 2). What this means is that “actors” who are decision makers must be able to derive meaning from data or information driving decision making that can translate into specific action and communication to others. Confronted with vast amounts of data, twenty-first century leaders must find those bits of data that provide information leading to actionable knowledge. This is no easy task because the environment is constantly changing. Essential elements needed to generate actionable knowledge include:

- having valid and timely information;
- the ability to make informed choices; and
- vigilant monitoring of both the validity of input information and implementation of decisions (Argyris, 1995).

In a world of rapid change, having valid and timely information is not necessarily an easy task. Information that may be valid, meaningful, and useful today may not be so a year or two years in the future. Changes in the internal or external environment or results from management decisions may, and probably will, make valid, meaningful, and useful information today of questionable value in the future. Therefore, it becomes critically important to constantly monitor both data and information input as well as output in the form of results and make adjustments as needs change. The result of not doing so could lead to decision making, and taking action, based on flawed information. Actions based on invalid or flawed information may lead to unintended negative outcomes.

Because of the continuous addition of new data generated throughout the organization, additional data are generated in a recursive manner. As firms acquire new data, the previous knowledge gained, may become obsolete. The existing knowledge needs to change based on the new data and the actions of the organizations can change accordingly. With the variety and volume of sources generated from big data, the possibilities of continually reanalyzing the data and taking action based on the new knowledge is endless. In order to provide value to management, the data must be continuously reevaluated with new perspectives in the changing business world. The ability of organizations to accelerate the rate of feedback loops accelerates the rate of knowledge creation that directly drives performance (Nold, 2012, 2013).

The process of interpreting big data in context requires transporting or transforming relevant information to key decision makers to take action based on the knowledge gained. Insightful and timely interpretation of data is critical to the success of the organization. One method for transforming big data to actionable knowledge is the performance triangle, which Lukas Michel describes as an intricate, dynamic system consisting of culture, leadership, and systems. As shown in Figure 2, the system is powered by people through relationships, collaboration, and purpose to drive success of the organization.

The process of generating actionable knowledge from big data becomes dependent on intricate and complex interactions of people working within the performance triangle (Michel, 2013; Bumblauskas et al., 2015). Nold and Michel (2016) demonstrated significant positive relationships between the performance triangle constructs and success across multiple industries, national cultures, size, and legal structure from a sample of 50 organizations. Simply gathering masses of data and distributing the data throughout the organization with systems is not enough to maximize performance. Developing a culture of trust where individuals are able to focus attention and use internet creative talents along with leaders who are able to interpret meaning in the data and communicate effectively are essential elements needed to maximize BDA (Nold and Michel, 2016).
Using the analogy of a living organism, Michel (2013) suggested that a virus infecting the organization at any place could inhibit the flow of data and ultimately knowledge creation and decision making. Viruses also cause interference with the ability to optimize the decision-making process and, if left unattended, can cause less than optimal or even inaccurate decisions to occur. Unseen viruses creep into an organization through an infinite number of ways such as obsolete data gathering systems, capture and display of irrelevant data, a culture where people lack trust so do not share what they know, or leaders who use industrial age management practices with knowledge workers to name a few (Michel, 2013). Viruses that disrupt flow are similar to non-value added activities in traditional lean context. Identification and remediation of viruses help to prevent disruptions to the organization by allowing the analysis of big data and conversion to actionable knowledge. One example of an internal virus problem is the story profiling Blockbuster LLC in which a CEO and activist shareholder’s battle led to the unraveling of the organization (Antioco, 2011). The results illustrated in Figure 3 are deteriorating performance due in part to ineffective decisions based on irrelevant, untimely or lack of information. Actions taken based on perceived but inherently flawed knowledge will rarely yield expected results.

One of the issues identified in BDA is a lack of vision of what “questions” need to be answered by the data. This leads to data collection without analysis. As such, there are pressures placed on the organization due to five specific elements for consideration in any BDA activity or project:

1. the optimization of revenue and gross profit;
2. the optimization of working capital and the investment in tangible and intangible capital or assets;
3. the optimization of the expense structure, or the expense side of the profitably model;
4. the possible opportunity, or opportunity costs, associated with the activity or project; and
5. the risk associated in terms of both new risks or the mitigation or modification of existing internal controls.
The manifestation of these risks are a loss of efficiency and/or functionality. We suggest that the recent popularity surrounding big data has been associated with bullet 4 above and then bullets 1-3. However, due to the speed and low cost of implementation, there could be huge exposures associated with bullet 5 above.

There is an inherent risk associated with the availability of big data. Prior paper-based systems allowed for a physical internal control systems. Today, automated data creates access risk inside and outside of organizations. Take, for example, the recent breach of controls associated with Apple’s iCloud (Wakabayashi, 2014) and credit card information at retailers The Home Depot (Perlroth, 2014) and Target illustrating that this is “only the beginning” (Miller, 2016), of such cybersecurity concerns with big databases. The problem extends beyond the breach of extremely sensitive information, such as credit cards. With the wealth of information that is shared by individuals online, the possibility or a breach of private information is increased. It is nature for users to share information in order to gain discounts or other benefits, but there is an expectation the data will be used according to any privacy notice (Akter and Fosso Wamba, 2016; Martin, 2015). The ability to anonymize and protect private information is a concern for most organizations undertaking a BDA project. Inadequate big data training of executives is a common problem in organizations. We are not sure yet whether a weak control environment caused these types of breaches, but there is no question that the accumulation of such large data sets exacerbated the situation.

**The dashboard framework and ESP International case**

With the explosion of data analytics over the past 30 years, and big data more recently, capturing the most critical data, formatting and visualizing the data, and getting it in front of key decision makers has become increasingly difficult. Dashboards have become a popular way to make key data sets available to overloaded executives and managers at a glance. Each dashboard is unique for the customer and manager typically containing flashy gauges, charts, tables, meters, and graphics intended to draw the viewer’s attention to key elements of the organization that might demand action. Figure 4 is an example of a dashboard from ESP International, Cedar Rapids, Iowa. Different definitions exist for dashboards but after a lengthy search for a definition,
Stephen Few (2006) developed a definition that seems to capture the unique essence of dashboards:

A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance (p. 34).

ESP International is using their dashboards to give business process managers a tool from which they can make critical decisions in working with suppliers. The input data streams consist of traditional metrics such as quality, on-time deliveries, etc. and the analysis of these data input streams provides the information that leads to actionable knowledge. While the concept of having all critical data displayed on one page so the manager can appraise the performance triangle of the organization at a glance is appealing, most dashboards fail to communicate the right data efficiently and effectively (Few, 2006). This failure is not primarily due to inadequate technology but rather because the dashboard is poorly designed thereby not communicating essential information as effectively as the unsuspecting manager believes. In some cases, the individuals charged with viewing the dashboards have not been trained in how to interpret the data, how to identify high vs low risk, and specifically what action to take in the event that the dashboard reveals an issue. Software designers, often times, become enamored with creating glitzy, flashy displays while failing to recognize the basic purpose of the dashboard as a tool for making actionable decisions. Once deployed, many cute displays lose their luster in a few days becoming annoying and forgotten.

Essential to the dashboard concept is idea of key performance indicators (KPIs) that drive business within the performance triangle. KPI proponents advocate identifying data that indicates operational effectiveness then cascading those or related performance indicators to succeeding lower levels in the organization. Fundamental to the effectiveness of the KPI approach is the belief that there is a cause and effect relationship with the KPIs and financial performance, which is not necessarily true in many cases. Choosing KPIs that have a cause and effect relationship, particularly for lower levels of the organization is difficult. Additionally, surveys indicate that on the average, organizations track nine times more KPIs than are actually needed. The reason for capturing so much data is that traditional approaches to KPIs follow a “more-is-better” philosophy (Battista and Shea, 2007).

As if choosing relevant KPIs was not difficult enough, consider that the world is changing at an ever-faster pace. Identification of relevant KPIs, and presenting them in an appealing visual format that stimulates management action, exposes the challenge that
those same KPIs may not be relevant in the future. Relevant KPIs today may become irrelevant due to the management action that resulted or simply because of changes in internal or external environment. Either way, the need exists to constantly monitor and question the cause, effect, and relevance of KPIs on a continuous basis to avoid the trap of making good decisions with bad data.

The use of a cautious approach in deploying dashboards can help with distilling data and help to create actionable knowledge. Upon launch, dashboards are often coveted and embraced with excitement, with the exception of those resistant to change, but lose momentum as users start to become immune to warnings (e.g. flashing red indicators all the time, so there is no ability to prioritize or determine the degree of the problem, which often-times leads to a lax attitude toward the significance of warning signs). A validation process for both presentation and relevance is periodically required to verify that valuable information leads to communication and action. As Rick Warren (2002) observed “familiarity breeds complacency” which can be costly. Questions to ask in the dashboard design process include:

- Dashboards are great in theory, but do they work for our organization or unit?
- Is the return on investment (ROI) worth the effort to compile the data?
- How will data be collected and how accurate is the data?
- Is any of the historical data flawed or poor quality?
- Are the people that need to see the data seeing it, processing the information, and making informed decisions?
- Can we trace and track that positive change is occurring with each item contained in the dashboard?

Dashboards provide oversight at all levels of the organization, including the task level. Dashboards are only one tool used in analyzing data to build information and knowledge, but are an important component to the actionable knowledge framework presented herein. While there are some reports and papers on dashboard design (Few, 2006), specific applications such as software team productivity (Biehl et al., 2007), etc. indicate this is a rapidly growing and transforming field. Amassing and visualizing data leads to a need for storage (e.g. data centers) and leads to the information security and risk concerns. Big data may not increase the likelihood of an actual internal control breach, but the potential increase in magnitude of a breach with big data can reach the point where the underlying internal control system must be re-considered. This is because a potentially exponentially higher risk of loss may not be tolerable to the organization. A good example is that if a sales clerk manually copies or actually steals a credit card, the risk is limited to one credit card number. Consumers’ diligence can minimize this risk. However, the database containing the data might be accessible and transportable with a single unanticipated breach. This can compromise millions of credit card numbers with one simple transfer, often then sold to others by “mules.” Therefore, design must consider risk in development of data intensive tools, determine whether there is a need for dashboards, and then perform a general risk assessment.

**Conclusions and future work**

The future of big data is very uncertain – will it be a fad, short-lived buzzword, or have impact on research and practice for generations to come? Buhl et al. (2013) say this very well in their article on this subject:

Big data – besides all hype and cherished expectations as “the next big thing” – above all is a multidisciplinary and evolutionary fusion of new technologies in combination with new dimensions.
in data storage and processing (volume and velocity), a new era of data source variety (variety) and the challenge of managing data quality adequately (veracity). However, to render big data a worthwhile innovation rather than merely a gadget, companies need well-founded and innovative business models that create value for the customer and thus the company while simultaneously considering privacy restraints. Hence, both from the research and practice perspective, big data needs to be taken as the basis rather than a guarantor of success. For long-term success, IT infrastructure, business processes, applications as well as the business model focusing on the customer need to be completely aligned (p. 68).

Actionable knowledge, the performance triangle, and dashboards are important theoretical constructs and frameworks for harnessing the power of big data. A successful organization must be able to efficiently convert data to information to actionable knowledge. The actionable knowledge model and framework detailed require further testing and validation in the field to prove statistical significance of this theory. This process can vary from industry to industry for practical execution on a case-by-case basis. In future work, we intend to explore the use of balanced scorecards and management by objective for applicability. In addition, we will address ways to prioritize alternative big data initiatives, including incorporating ROI concepts.

Final recommendations include ensuring that the data collected, acted upon knowledgably, and conveyed in dashboards are relevant, timely, and informative. Important considerations include when to refresh, what story the data tells, and whether the data can be used predictively as a forward looking indicator (e.g. to perform maintenance, etc.). The top two “Best Business Jobs,” in 2014 as noted by US News and World Report (2014) were market research analyst and operations research analyst both of which require the translation of data to information to actionable knowledge (using a dashboard or some other visualization tool).

If we assume that the availability of data drives our ability to interpret and consequently make decisions, we must further consider how the availability of massive amounts of data has evolved to facilitate more effective decision making. In future work, we plan to introduce the concept of relevant range theory which considers the changes which an individual or organization must under-go when operating outside of normal operating conditions. This would include changes required to operate in high growth or rapid decline, i.e., relevant range theory, scenarios such as the contingency planning undergone at organizations such as Caterpillar Inc. since the start of the Great Recession or Global Financial Crisis in 2008. When applied to data, the relevant range theory would indicate that when an organization or individual are below or above a “normalized” quantity or quality of data we are unable to properly interpret that data we must channel that data to the place where the data can be correctly interpreted. There is an underlying assumption in research and practice that the recipient of data has the time and ability to convert the data into information and actionable knowledge. With the absence of an appropriate filtering and directional system, training and skill, and when the effect of timing enters the equation, once outside the relevant range, the data diminishes in effectiveness to the point that it is useless.

Another area of interest for future research is the impact of big data on the legal and social responsibility associated with accountability. That is, when big data exists, the organization must do something to analyze the data and report output results which entities, e.g., attorneys’, activists, etc. could argue were obvious results later. The strategic retention and destruction of data should also be further explored as it is an important element of managerial oversight. The application and transfer of knowledge addresses data conversion to information but is as of yet unapplied in people, computers, and systems. Data that is not acted upon is trapped in the human mind (or computer) and is wiped clean, either at death (deleting the data on a computer) or by some other brain trauma (computer trauma). It is therefore critical that actionable knowledge be
disseminated to all decision makers we anticipate being involved today and in the future at our organizations. As data sets become larger, the scope and scale of big data will constantly grow and evolve. An actionable knowledge method with a dashboard framework is one solution to manage this big data challenge.

References


**Further reading**

About the authors

Daniel Bumblauskas is an Assistant Professor of Management and the Hamilton/ESP International Fellow of Supply Chain and Logistics Management at the University of Northern Iowa. Previous to this appointment, Daniel was an Assistant Teaching Professor of Management at the University of Missouri where he still holds a courtesy appointment. Daniel conducts research, teaches, and consults in the area of operations and supply chain management globally and has published upwards of 35 peer-reviewed journal articles and conference proceedings, including publications in journals such as *IEEE Transactions on Industry Applications* and the *International Journal of Quality and Reliability Management*. He earned a BS Degree in Industrial Engineering and Economics from the Iowa State University, a Master of Liberal Arts in General Management from the Harvard University, and received his MS and PhD Degrees, both from the Iowa State University, in Industrial Engineering. Prior to his faculty appointments, Dr Bumblauskas was previously employed in industry by ABB Inc. and Sears Holding Corporation. Daniel Bumblauskas is the corresponding author and can be contacted at: daniel.bumblauskas@uni.edu

Dr Herb Nold is a Professor of Business Administration at the Polk State College in Lakeland, Florida, USA. Dr Nold earned a Doctor of Management in Organizational Leadership from the School of Advanced Studies, University of Phoenix, holds Master and Bachelor Degrees in Education from the Northern Illinois University and is a Certified Public Accountant. Prior to entering academia Dr Nold logged over 30 years of leadership experience as an educator and business leader ranging from Teacher and Coach to Manufacturing Engineer to CFO in various industries.

Paul Bumblauskas is a Business Advisor, Facilitator and International Speaker and is highly regarded by business owners and executives, especially in building material and home décor-related industries. Paul has over 30 years of experience in retail, wholesale and light manufacturing and has amassed unique sets of skills ranging from sales, staffing and store presentation to accounting, human resources, information technology, operations and administration. Paul is currently the President of PFC Services, Inc., a consulting and facilitating firm with a dominant client base in building materials, retail, and distribution. PFC is known for high-impact recommendations that can be immediately implemented. Notably, Paul developed the proprietary Demand-Based Series, as well as the Sales Optimization Program, which has over 20 approaches – all designed to optimize sales and profit. He also has developed the acclaimed seminars on increasing average ticket, comfort zones, tasking techniques and practical ways to sell green.

Amy Igou is an Assistant Professor in Account at the University of Northern Iowa. She received her PhD Degree from the Southern Illinois University in Carbondale in Management with an emphasis in Information Systems. Prior to starting her academic career, Amy worked at Deere & Company in a variety of financial systems positions. She has published in the *Journal of Accounting Education* as well as several conferences and proceedings. Amy’s research areas include the use of analytics in business applications, nonprofit information technology, and accounting information systems pedagogy. She is also a Certified Management Accountant.

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Digital competences of the workforce – a research topic?

Matthias Murawski and Markus Bick

Department of Business Information Systems,
ESCP Europe Business School Berlin, Berlin, Germany

Abstract

Purpose – Considering working in the digital age, questions on the consequences for the individual workers are, so far, often neglected. The purpose of this paper is to deal with the question of whether the digital competences of the workforce is a research topic. The authors argue for the thesis that it is indeed a research topic.

Design/methodology/approach – In addition to a literature analysis of the top IS, HR, and learning publications, non-scientific sources, as well as the opinions of the authors, are included. The authors’ thesis is challenged through a debate of corresponding pros and cons.

Findings – The definition of digital competences lacks scientific depth. Focussing on the workforce is valid, as a “lifelong” perspective is not mandatory for research. Digital competence research is a multidisciplinary task to which the IS field can make a valuable contribution.

Research limitations/implications – Although relevant references are included, some aspects are mainly driven by the opinions of the authors. The theoretical implications encompass a call for a scientific definition of digital competences. Furthermore, scholars should focus on the competences of the workforce, including occupations, roles, or industries. The authors conclude by providing a first proposal of a research agenda.

Practical implications – The practical implications include the alignment of multiple stakeholders for the design of “digital” curricula and the integration by HR departments of the construct of digital competences, e.g. for compensation matters and job requirements.

Originality/value – This paper is one of very few contributions in the area of the digital competences of the workforce, and it presents a starting point for future research activities.

Keywords Big data, Research agenda, Workforce, Curriculum design, Digital competences

1. Introduction

In 2017, the main challenges for companies are neither technological trends, disruptive innovations, nor new customer behaviour. Their main duty is to adapt their culture, mindset, and competences to the new, digital way of working (Accenture, 2016; Maedche, 2016). Results matter the most, and place and time have become less relevant, especially for knowledge workers, who are sometimes even called “digital nomads” (Makimoto and Manners, 1997). Although there are boundaries for digital nomadism, such as submission deadlines, update calls, or the availability of colleagues, the degree of freedom for employees working in the digital age is technically unlimited. Collaboration in the digital age occurs across geographical, functional, and hierarchical boarders, while creating interactive content using social media (Lepofsky, 2016; Malone, 2004). In addition to knowledge intensive jobs, industrial jobs are affected by digitisation as well. Just think of the impact due to automation, such as 3D printing, drones, the Internet of Things, and robots.

One thing that all digital technologies have in common is that they produce data. While the underlying effects need further investigation (Lycett, 2013; Sharma et al., 2014), big data, and respectively, data analytics, are potentially the foundation for creating value through digitisation (Chen et al., 2012; LaValle et al., 2011). Big data analytics might enable new insights that were not previously available. Algorithms provide very specific information about customers, and tailored solutions can be offered. In addition, companies applying big data analytics to themselves could learn more about their internal processes and initiate respective improvements. That is, why data are sometimes called the oil of the digital age (Brustein, 2012).
But what does this kind of revolution mean for the workforce? The current debate is mainly driven by researchers who aim at elaborating on the consequences of digitisation, especially concerning unemployment rates. The work that has generated the most public attention is the one by Frey and Osborne (2013), who estimated that 47 per cent of current US employment is at risk due to computerisation. Triggered by such alarming results, other reports dealing with this topic have recently been published (Autor, 2015; Miroudot et al., 2016; OECD, 2016; World Economic Forum, 2016b). Although the numerical results and conclusions of these reports differ, they all agree that the concept of work will change significantly in coming decades.

While the macroeconomic scientific debate is still ongoing, the consequences for the individual workers are, so far, often neglected (Wang and Haggerty, 2011). However, there are very relevant questions, and one of the most potentially important ones is “Which digital competences are required for employability in the digital age?” In relation to big data, the specific digital competences that are necessary to make use of the huge opportunities related to this new field of work might be elaborated. There have been first attempts to answer this question for big data competences in general (Debortoli et al., 2014; Dubey and Gunasekaran, 2015), as well as for upcoming data-driven occupations, such as those of “mobile analyst” (Brauer and Wimmer, 2016) and “data scientist” (Davenport and Patil, 2012; Schumann et al., 2016). Other scholars have focussed on big data skills, which are required especially for supply chain management (Schoenherr and Speier-Pero, 2015; Waller and Fawcett, 2013b).

However, most of the existing publications on the topic of big data competences have followed rather exploratory approaches, for instance, Debortoli et al. (2014) stated that specific big data job profiles are often only anecdotally described. Generally, rigorous academic investigations of digital competences, as well as fields of work other than big data, are rare.

Based on this background, we believe that it is time to establish a research stream, in addition to the macroeconomic one, that deals with the consequences and requirements of digitisation for the workforce, wherein various perspectives, such as different IT generations, industries, occupations, roles, or individuals, might be the focus. The existing approaches mentioned above mark the first step, but they need to be extended and enriched, making use of rigorous scientific practice. Thus, we formulate the thesis: researchers must focus on the topic of the digital competences of the workforce.

After presenting our research design (Section 2) and the current definitions and meanings of digital competences (Section 3), we conduct a thorough debate of the pros and cons regarding our thesis in Section 4. Section 5 contains a discussion of the practical and theoretical implications. Finally, we conclude by further asserting our thesis and presenting a first proposal for a research agenda for the topic at hand (Section 6).

2. Research design

While arguing for a new research stream which, thus far, is not existent, we have developed and challenged our thesis based on the experiences and findings gathered from a literature analysis. Regarding the literature review, we started with checking business and IS databases such as EBSCO and JSTOR with the search term “digital competenc*”, which resulted in only a few results. None of the hits had any relationship to the workforce. One reason for this result might be that the term “digital competences” has not yet been established outside of the political and pedagogy fields (see Section 3). Another issue that arose when applying the search term “digital competenc*” without quotes was that the single components, “digital” and “competenc*”, led to unmanageably extensive result lists. Thus, a common systematic literature review on the term digital competences was not a suitable method. Therefore, our approach was to manually check the Senior Scholars’ Basket of Journals (AIS, 2011) from January 2011 until July 2016 regarding the topic of the digital competences of the workforce. This timeframe was chosen with respect to the general novelty of our topic and covers the emergence of the most popular publications dealing with
the effects of digitisation on the workforce, e.g. Brynjolfsson and McAfee (2014) and Frey and Osborne (2013). In addition to the IS journals and their technology foci, and as we are focussed on the workforce, we included suitable HR journals (Human Resource Management, Human Resource Management Journal, Leadership Quarterly, Industrial and Commercial Training) in our search. With respect to the aspects regarding curriculum design, we considered scientific learning journals (International Journal of Innovation and Learning, Journal of Education and Learning, International Journal of Learning Technology) as well. We carefully screened the titles, keywords, and abstracts of all volumes for any relation to our topic. Subsequently, we checked the reference lists of selected publications for further relevant sources. However, while confirming our database search mentioned above, within this set of journals and the selected timeframe, we were unable to find any literature focussing specifically on the digital competences of the workforce. We only identified a few articles that touch upon this topic, e.g. when dealing with virtual team leadership. These results indicate that there are a very limited number of scientific articles on the topic of the digital competences of the workforce. We therefore decided to also consider non-reviewed references, e.g. the publications of political institutions and consulting agencies, thereby ensuring a high level of relevance.

3. Current definitions and meanings of digital competences

When focussing on firms and other work-related organisations, competences can be understood as a combination of abilities, (work-related) knowledge, and skills held by an individual (Nordhaug, 1993). Following Gorbacheva et al. (2016) and Müller et al. (2014), abilities innately belong to an individual (like the ability to engage in logical reasoning). Knowledge means a theoretical understanding of a concept, while skills are the practical application of that knowledge. Other definitions of competences emphasise the aspect of attitudes, instead of knowledge, as well as the target of “problem solving” (Holtkamp et al., 2015; Winterton, 2009). Ha (2016) distinguished between two competence dimensions: epistemic and heuristic. While epistemic competence is the formal or knowledge-related aspect of competence, heuristic competence comprises the experiences and capacity to act which can be developed once the knowledge is constantly applied. Generally, competences should be understood within a specific context, e.g. existing systems or work practices (Hoel and Holtkamp, 2012).

According to Ala-Mutka (2011), digital competences encompass instrumental knowledge and the skills for tool and media usage; advanced skills and knowledge for communication and collaboration, information management, learning and problem solving; and meaningful participation; and attitudes towards strategic skill usage in intercultural, critical, creative, responsible and autonomous ways. Ala-Mutka (2011) herself stated that this concept has a generic character (e.g. not focussed on the workforce) and needs to be tailored to specific target groups, which corresponds to the postulation of Hoel and Holtkamp (2012) to consider a specific context. A recent IS-oriented definition of digital competences that considered an organisational context is the one developed by Vieru (2015, p. 6718): “Digital competence consists in the ability to adopt and use new or existing information technology to analyse, select and critically evaluate digital information in order to investigate and solve work-related problems and develop a collaborative knowledge body while engaging in organizational practices within a specific organizational context”. He developed a conceptual model covering the technological, cognitive, and organisational culture dimensions, and the integration of these three dimensions. Ferrari (2012) presented a definition of digital competences based on an analysis of 15 cases (e.g. school curricula projects, certification schemes). Additionally, she distinguished different “building blocks” of the construct, which are mentioned on the right side in Figure 1.

Besides these rather generic definitions, a number of publications have mentioned some more concrete digital competences, which have often been derived from a practical perspective. Although these concepts are not always digital competences in a narrow sense,
they provide valuable insights into the dimensions of digital competences which need to be specified. Table I provides an overview of selected digital competences.

The construct of digital competences is not yet consistently used (Ala-Mutka, 2011; Ferrari, 2012; Vieru, 2015). Instead, there are several concepts meaning more or less the same thing, leading experts to speak about a “jargon jungle” (Ferrari, 2012, p. 11). Furthermore, the existing definitions have a rather generic character; they do not consider any specific context, although this has been highlighted as being essential for competence research (Hoel and Holtkamp, 2012). More concrete digital competences (cf. Table I) often have a practical background and tend to be overlapping, e.g. communication and digital teamwork. Thus far, rigorously developed scientific definitions or concepts, as well as those considering specific contexts, e.g. occupations or industries, are missing.

4. Pros and Cons

We challenge our thesis, Researchers must focus on the topic of the digital competences of the workforce, with regard to the three arguments. First, we deal with the opinion that the relatively new phenomenon of digitisation is simply part of permanent technical progress. Related technical skills are already targeted by current research. Thus, focussing on digital competences is redundant. The second insertion is that developing digital competences is a “lifelong” learning process. Consequently, it is not valid to limit research to the workforce. Third, because of other established research fields regarding competence research, it is argued that digital competences research is not a task for IS.

Focussing on digital competences is redundant

Pro. Technical progress has always existed, and digitisation is part of the technical progress. For instance, banks and insurance companies have digitised many of their number-intensive processes since the 1960s. Police and intelligence services have made use of databases since the 1970s, and hospitals have supervised patients with computers and have stored the resulting digital data since the 1970s. The digitisation of texts (1980s/1990s), music (1980s/1990s), traffic lights (early 1990s), and telecommunication (late 1990s) followed soon thereafter (see e.g. Passig (2016) for further examples). Interestingly, the term digitisation, with its “innovative” connotation, became popular around the year 2000.
However, the content behind it is not new; it is simply technical progress. Consequently, digital competences, as a research topic, is redundant, as research on the competences for dealing with technical progress, often called ICT skills, has already existed for a long time (Autor et al., 2003; Bassellier et al., 2001; Lopez-Bassols, 2002). Additionally, the term digitisation is criticised for being too unspecific, and therefore, not applicable, as mentioned by Passig (2016) during a speech at the well-known IT conference, CeBIT, in March 2016. Based on mainly philosophical debates, she argued that the two concepts “analog” and “digital” cannot be clearly distinguished. Therefore, digitisation is not a proper concept and should be replaced by a description of the particular change in a detailed and unambiguous way. This statement can be transferred to digital competences. Instead of calling certain competences “digital”, one should describe them in more detail.

Contrary to the previous paragraph, the objections mentioned therein are mainly grounded in definitional challenges. From our perspective, there is a difference between technical progress and digitisation. We agree that technical progress and the use of computers are not new at all; computers have been used for more than half a century. However, they have improved exponentially in recent years (e.g. “Moore’s Law”). Brynjolfsson and McAfee (2014, p. 9) stated: “But just as it took generations to improve the steam engine to the point it could power the Industrial Revolution, it’s also taken time to refine our digital engines”.

Within this context, digitisation does not only mean the change from analog to digital. Instead, it means that an enormous and impressively rapid transformation has just begun, which is

<table>
<thead>
<tr>
<th>Digital competence</th>
<th>Example</th>
<th>Source(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information processing</td>
<td>Using search engines/search strategies</td>
<td>European Union (2015)</td>
</tr>
<tr>
<td>Communication</td>
<td>Assessing the reliability of information from the internet</td>
<td>European Union (2015)</td>
</tr>
<tr>
<td></td>
<td>Using a wide range of online communication tools</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Creating and managing content with collaboration tools (e.g. Google Drive)</td>
<td></td>
</tr>
<tr>
<td>Content creation</td>
<td>Using a programming language for content creation</td>
<td>European Union (2015)</td>
</tr>
<tr>
<td></td>
<td>Using advanced formatting functions (e.g. mail merge, macros)</td>
<td></td>
</tr>
<tr>
<td>Safety</td>
<td>Monitoring the security settings of the devices used</td>
<td>European Union (2015)</td>
</tr>
<tr>
<td></td>
<td>Knowing how to encrypt e-mails or files</td>
<td></td>
</tr>
<tr>
<td>Problem solving</td>
<td>Choosing the right tool, device, etc. to solve (non-technical) problems</td>
<td>European Union (2015)</td>
</tr>
<tr>
<td>Digital rights</td>
<td>Understanding and upholding personal and legal rights</td>
<td>World Economic Forum (2016a)</td>
</tr>
<tr>
<td></td>
<td>(rights to privacy, intellectual property, freedom of speech and protection from hate speech)</td>
<td></td>
</tr>
<tr>
<td>Digital emotional</td>
<td>Being empathic online</td>
<td>World Economic Forum (2016a)</td>
</tr>
<tr>
<td>intelligence</td>
<td>Building good relationships with others online (e.g. in social networks)</td>
<td></td>
</tr>
<tr>
<td>Digital teamwork</td>
<td>Working across time zones and language barriers</td>
<td>E.g. Accenture (2016),</td>
</tr>
<tr>
<td></td>
<td>Being available almost any time, any place via mobile or wearable devices</td>
<td>Lepofsky (2016)</td>
</tr>
<tr>
<td>Making use of big data</td>
<td>Accessing, analysing data, and reporting insights</td>
<td>E.g. McAfee and</td>
</tr>
<tr>
<td>Self-disruption</td>
<td>Data-based decision making</td>
<td>Brynjolfsson (2012)</td>
</tr>
<tr>
<td>Making use of</td>
<td>Being open to even radical change regarding one’s own role</td>
<td>E.g. Johnson (2015),</td>
</tr>
<tr>
<td>artificial intelligence</td>
<td>Seeking to align with digitisation and its impact</td>
<td>Accenture (2016)</td>
</tr>
<tr>
<td>Virtual leadership</td>
<td>Collaborating with e.g. robots, avatars, and chatbots</td>
<td>E.g. Wilson and</td>
</tr>
<tr>
<td></td>
<td>Being able to design AI artefacts</td>
<td>Bataller (2015)</td>
</tr>
<tr>
<td></td>
<td>Motivating rather than controlling workers</td>
<td>E.g. Mortensen (2015),</td>
</tr>
<tr>
<td></td>
<td>Establishing personal ties even through impersonal technical channels</td>
<td>Serban et al. (2015)</td>
</tr>
</tbody>
</table>

Table I. Selected digital competences of the workforce.
having much more of an impact than the preceding general technical progress. We believe, although ambiguously, that digitisation is the proper key term to encompass such a complex situation. Following our understanding of digital, there is a difference between ICT skills and digital competences. ICT skills have a narrow focus and cover “the ability to use the software and hardware of an information technology-based device such as a personal computer, laptop, or a tablet” (Vieru, 2015, p. 6725). We interpret ICT skills as a part of digital competences, as they can be assigned to instrumental knowledge and skills for tool and media usage which is one out of three basic pillars of the digital competences construct according to Ala-Mutka (2011), or one out of the main dimensions, namely technological, according to Vieru (2015).

Developing digital competences is a “lifelong” learning process

Pro. “Lifelong learning strategies need to answer to the growing need for advanced digital competence for all jobs and for all learners” (Ala-Mutka et al., 2008, p. 5). Focussing on the workforce is therefore not a valid approach, as it is critical to consider any part of a life separately. If digital competences are not taught as early as possible, e.g. in primary school, this deficiency will be difficult to remedy later. Furthermore, the younger generations are all so-called digital natives (Palfrey and Gasser, 2008; Prensky, 2001) who possess the digital competences required today. They are both generally tech-savvy and used to developing future digital competences. Thus, one should focus on this group instead of the workforce.

Contra. The concept of lifelong learning is not a mandatory approach for research. We believe that this concept is mainly a political one and might be too broad to be researched properly. This opinion is supported by Jarvis (2006), who concluded that there is no single comprehensive learning theory. We agree that the basis of the digital competences of the workforce should ideally be developed prior to the working phase, e.g. during primary school. But this is not possible for everybody, e.g. digital immigrants, meaning older people who did not grow up surrounded by IT (Wang et al., 2013). Furthermore, the concept of digital natives can be criticised. First, they are often discussed on a popular science level. Rigorous scientific work, especially empirical work, such as that recently done by Hoffmann et al. (2015), is still rather rare. Second, there is strong evidence that the competences within the digital native group vary, raising the question of whether digital natives or digital naives is the more suitable term (Hargittai, 2010). Recommendations regarding curricula design or professional competence management (Baladi, 1999) must be given carefully for such a heterogeneous group. Therefore, we think that focussing on the workforce is a valid approach, although a clear differentiation of subgroups must be made.

Digital competences research is not a task for IS

Pro. Competence development is mainly researched by disciplines such as psychology, human resources, and educational science (Salganik and Rychen, 2003; Shippmann et al., 2000; Stevens, 2013). Additionally, these fields mark the basis for related research on curricula design, competence management and job requirements analysis, thereby developing a number of relevant contributions to both practice and theory. Especially because digital competences have a strong link to practical aspects such as curricula design and employability, they should consequently be investigated by these disciplines. The IS field might contribute by formulating research requests, e.g. regarding the digital competences of big data analysts, but should not engage in actual digital competences research.

Contra. We partly agree to this argument. Digital competence is a multidisciplinary topic which is impossible to assign to only one research field. Thus, each of the aforementioned research disciplines could certainly add value to digital competence research. And this holds true for IS, as well, for several reasons. First, digital competences have a strong link to technology, which is usually not covered by psychology, human resources, and educational
science. Second, digital competences’ basis, digitisation, is a topic traditionally occupied by IS (Yoo et al., 2010), and it thereby focusses on various digitisation sub-domains like big data (Chen et al., 2012; Newell and Marabelli, 2015). Third, although not the core of the discipline, IS already has some experience in competence research, e.g. for specific occupations like business process managers (Gorbacheva et al., 2016; Müller et al., 2014), requirements analysts (Klendauer et al., 2012), or IT infrastructure consultants (Ha, 2016), for specific processes like the global software development process (Holtkamp et al., 2015; Holtkamp and Pawlowski, 2015), or for analysing the IT competences of business managers (Bassellier et al., 2001), as well as non-IS workers (Davis, 2013). Therefore, we believe that IS can valuably contribute to digital competences research (Table II).

5. Practical and theoretical implications
In the following, we aim to elaborate both the theoretical and the practical implications related to our thesis that the digital competence of the workforce is a research topic. We emphasise the need for a comprehensive scientific definition of the term, relevant actions in the HR field, and the link between research and curriculum design.

As discussed, the term digital competences is rather imprecise and context dependent. Thus, it is not surprising that it is often dismissed as just another buzzword. However, a common understanding of digital competences, among different fields such as politics or research, is the basis for its useful application. The identified practical and scientific relevance, in line with the lack of theoretical developments regarding the digital competences of the workforce, is the basis for the theoretical implication for the development of a consistent scientific definition. This implication is supported, for instance, by a workshop at this year’s European Conference of Information Systems (2016), which dealt with very basic aspects regarding digital competences. Conceptual issues covering fundamental questions like “What is digital competence?” have been targeted, thereby highlighting the need for an improved understanding of the construct of digital competences. Interestingly, the workshop was aimed particularly at working practice, instead of school settings, which is also in line with our argument in favour of focussing on the workforce.

Scientific work dealing with competences traditionally considers the entire organisation (Chakravarty et al., 2013; Kalampokis et al., 2012; Pastuszak et al., 2013; Peppard and Ward, 2004; Prahalad and Hamel, 1990). Other levels, such as the individual or the individual’s role, are often neglected (Wang and Haggerty, 2011). However, they are probably the most interesting levels for practical use. From our point of view, focussing on the workforce and further sub-levels (e.g. occupations, roles, or industries) provides significant value. This life phase is interesting, as the working phase is the one in which people can benefit the most from their set of competences in terms of compensation. However, so far, the relation between digital competences and compensation has not been investigated rigorously. Similar approaches have dealt with other constructs instead of digital competences. For instance, Peng and Eunni (2011) elaborated that employees are rewarded for both the

<table>
<thead>
<tr>
<th>Area</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>The term “digital competence”</td>
<td>Digital competence is a single construct, different from e.g. IT skills</td>
</tr>
<tr>
<td>Research focus: workforce</td>
<td>Digital competence definitions lack scientific depth</td>
</tr>
<tr>
<td></td>
<td>Focussing on the workforce is valid, as the “lifelong” perspective is not mandatory for research</td>
</tr>
<tr>
<td></td>
<td>Digital native research does not replace investigations of the workforce</td>
</tr>
<tr>
<td>Research discipline</td>
<td>Digital competence research is a multidisciplinary task</td>
</tr>
<tr>
<td></td>
<td>The IS field can contribute valuably to digital competence research</td>
</tr>
</tbody>
</table>

Table II. Main findings of the pros and cons discussion
depth and the breadth of their computer skills. Schulz et al. (2013) found that human capital is associated with higher employee compensation. Another upcoming HR topic related to digital competences is the identification of job requirements. One established approach for this purpose is traditional job analysis, which aims at elaborating job-specific requirement profiles (e.g. Shippmann et al., 2000). Another essential approach is competency modelling. Here, desired competences are derived from a firm’s strategy. These competences are not assigned to one specific profile, but are relevant for all job positions within the organisation (Redmond, 2013). As yet, neither approach integrates digital competences. Due to the increasing relevance of digital competences, we believe that both HR professionals and scholars are asked to integrate them into their work.

Digital competences are in the centre of the current debate, especially regarding employability and university curricula design. Headlines like Curriculum experts say coding is essential in a digital economy (Bird, 2016) emerge almost daily. Very often, this postulation is driven by the employer side and describes the demand for specific digital competences, which are a prerequisite, e.g. for big data analytics. Also, scholars have emphasised the need for curricula adjustments and have called for related research (Dubey and Gunasekaran, 2015; Karimi et al., 2012; Provost and Fawcett, 2013; Waller and Fawcett, 2013a). There are even single cases in which scholars have developed curricula focussed on big data analytics (Schoenherr and Speier-Pero, 2015). However, most existing curricula fail to fulfil the aspect of employability in the digital age. Of course, employability is not the only task educational institutions have, but it is one core objective next to personality development and scientific foundation, and it should be integrated into curricula design (Sumanasiri et al., 2015). Due to the fast-changing nature of specific digital competences, universities ought to emphasise general or basic digital competences, such as collaborating virtually, selecting tools and applications according to a task and not according to a dogma, and being open for innovation, while maintaining a critical and reflective perspective, e.g. regarding the security aspects. These digital competences could be taught by integrated learning approaches. More domain- or occupation-specific digital competences, such as coding with specific programming languages, could be offered in specialised courses or even study programmes like the one presented by Schoenherr and Speier-Pero (2015). They constitute a positive example of how research and education can be aligned, which is a precondition for successful curricula design. Adjusting curricula requires an understanding of the relevant digital competences. Thus, a strong exchange between educational and research institutions should be established, while also including other stakeholders like employers (Table III).

6. Conclusion and research agenda
When we take a critical and honest look at today’s working environment, and when we ask ourselves how our current job will be designed in 20 years (if it still exists), we must recognise...
that we have to radically transform our thinking about the way of working and the respective working requirements. But what exactly are the consequences on an individual level? Digital competences, whether or not one likes the term, are part of the answer, as they encompass the knowledge, abilities, skills, and attitudes we need for working in the digital age. Digital competences on an individual level is an umbrella term covering both the general digital competences (e.g. virtual collaboration) which are needed for nearly every knowledge occupation, and the specific role- or task-related digital competences which are different for every occupation. Coming back to big data analytics, completely new “digital” jobs have recently emerged, e.g. the “data scientist”, but it is rather unclear which competences these jobs require. However, both career opportunities and potential earnings in the digital age will depend on developing suitable digital competences. But what these digital competences are, how they can be measured, and how they can be developed are only some of the questions for the research community to answer. This paper is a call to start shedding light on these issues.

The main limitation of this paper lies in its subjective character. Although we aimed at including relevant references, some aspects are very new, and thus, cannot be supported by publications, but are driven by the opinions of the authors. Another limitation is the focus on IS, and to some extent, HR and learning literature. We see various opportunities for research. Thus, as a research outlook and with respect to the novelty of the topic, we provide a first step towards a research agenda. We summarise those research areas, as well as the corresponding research questions we have identified as the most important ones in Table IV, thereby following the structure of Vom Brocke et al. (2011).

<table>
<thead>
<tr>
<th>Research area</th>
<th>Selected research problems/questions</th>
<th>Selected research strategies/methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding the construct</td>
<td>What is digital competence? What constitutes it?</td>
<td>Conceptual research; interdisciplinary research; qualitative research such as</td>
</tr>
<tr>
<td>digital competences</td>
<td>To what extent is digital competence a generic competence and to what extent is its role- or domain-specific resp. context dependent?</td>
<td>domain expert interviews; case-study research</td>
</tr>
<tr>
<td></td>
<td>How can digital competences be measured?</td>
<td>Qualitative research such as content analysis; text-mining; case-study research;</td>
</tr>
<tr>
<td>HRM – supply of digital</td>
<td>Which digital competences are actually offered, e.g. by members of a specific occupation?</td>
<td>quantitative research, e.g. surveys</td>
</tr>
<tr>
<td>competences</td>
<td></td>
<td>Qualitative research such as content analysis; interviews with domain experts</td>
</tr>
<tr>
<td></td>
<td>Which digital competences are actually required, e.g. for a specific occupation or role?</td>
<td>and Delphi studies; case-study research; conceptual research; theory development</td>
</tr>
<tr>
<td>HRM – demand for digital</td>
<td>How will job profiles and required digital competences change in the future?</td>
<td></td>
</tr>
<tr>
<td>competences</td>
<td>What is the relation between digital competences and compensation?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>How can digital competences be integrated into job requirement models?</td>
<td></td>
</tr>
<tr>
<td>Different IT generations</td>
<td>Which digital competences do digital natives/digital immigrants possess?</td>
<td>Quantitative research, especially surveys; case-study research</td>
</tr>
<tr>
<td></td>
<td>How can digital natives and digital immigrants collaborate successfully? Which competences are</td>
<td></td>
</tr>
<tr>
<td></td>
<td>required?</td>
<td></td>
</tr>
<tr>
<td>Curricula design</td>
<td>How should curricula be designed in order to impart digital competences?</td>
<td>Conceptual research; interdisciplinary research; case-study research</td>
</tr>
<tr>
<td></td>
<td>Which tools are needed to support the development of digital competences?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Which digital competences (e.g. generic vs role focus) should be imparted?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>How can curricula remain “up-to-date” regarding fast-changing digital competences? Is this necessary?</td>
<td></td>
</tr>
</tbody>
</table>

Table IV. Towards a research agenda for the topic of the digital competences of the workforce
The construct of digital competences still requires substantial scientific effort to make it applicable, e.g. for designing curricula and training. It seems to be necessary to develop a comprehensive definition of the term. Also, there are various options for researching the digital competences of the workforce related to HR management. To provide a structured approach, we suggest differentiating between the “supply” and “demand” of digital competences. Regarding the supply side, upcoming approaches, such as analysing social media profiles for specific jobs and making use of text-mining techniques (Gorbacheva et al., 2016), could be applied. The demand side might include the investigation of the required competences on a job level, e.g. by analysing job ads (Debortoli et al., 2014; Muller et al., 2014), but it should also deal with compensation and theory development matters. We think that the IS field, with its strong links to technology-driven industries and occupations, offers interesting and relevant research opportunities related to these points. Furthermore, besides individuals, analysing different groups of the workforce, such as digital natives and digital immigrants, promises to be both a relevant and interesting field of study (Wang et al., 2013). One topic concerning these groups was formulated by Becker et al. (2015), who asked how different IT generations can work together. The question of “how” enables various ways to tackle this question, but we are sure that one important aspect is looking at the required digital competences of all of the involved parties. Another important research area is curricula design. We follow existing calls for research (e.g. Waller and Fawcett, 2013b) and propose some relevant questions that should be investigated. From our point of view, the curricula-related approaches having an interdisciplinary research design are the most promising.

References


**Corresponding author**
Matthias Murawski can be contacted at: mmurawski@escpeurope.eu
Automated competitor analysis using big data analytics
Evidence from the fitness mobile app business

Liang Guo
Neoma Business School, Mont Saint Aignan, France

Ruchi Sharma
School of Strategy and Leadership, Neoma Business School, Mont Saint Aignan, France

Lei Yin
Neoma Business School, Mont Saint Aignan, France

Ruodan Lu
University of Cambridge, Cambridge, UK, and

Ke Rong
Business School, Bournemouth University, Bournemouth, UK

Abstract

Purpose – Competitor analysis is a key component in operations management. Most business decisions are rooted in the analysis of rival products inferred from market structure. Relative to more traditional competitor analysis methods, the purpose of this paper is to provide operations managers with an innovative tool to monitor a firm’s market position and competitors in real time at higher resolution and lower cost than more traditional competitor analysis methods.

Design/methodology/approach – The authors combine the techniques of Web Crawler, Natural Language Processing and Machine Learning algorithms with data visualization to develop a big data competitor-analysis system that informs operations managers about competitors and meaningful relationships among them. The authors illustrate the approach using the fitness mobile app business.

Findings – The study shows that the system supports operational decision making both descriptively and prescriptively. In particular, the innovative probabilistic topic modeling algorithm combined with conventional multidimensional scaling, product feature comparison and market structure analyses reveal an app’s position in relation to its peers. The authors also develop a user segment overlapping index based on user’s social media data. The authors combine this new index with the product functionality similarity index to map indirect and direct competitors with and without user lock-in.

Originality/value – The approach improves on previous approaches by fully automating information extraction from multiple online sources. The authors believe this is the first system of its kind. With limited human intervention, the methodology can easily be adapted to different settings, giving quicker, more reliable real-time results. The approach is also cost effective for market analysis projects covering different data sources.

Keywords Mobile apps, Big data, Naïve Bayes, Operational strategy, Probabilistic topic modelling, User segment overlapping

Paper type Research paper

Introduction

Firms today face constant pressure to maintain sustainable growth, stay ahead of their competitors, and present superior customer-centric products. It is impossible for any firm to adequately survive, without developing a thorough market perspective. One of the tools for gaining the market insight is by developing the right competitive intelligence that can have a

This research is supported by “the Innovation Method Fund of China’ from Ministry of Science and Technology of China (MOST) under Project No. 2016IM010200”, the National Natural Science Foundation of China (Grant No. 71402051), and “Cyrus Tang Young Scholar Awards”. All errors are of the authors.
far reaching strategic impact on a firm's operations strategy and business process management. Following Amoako-Gyampah and Boye (2001, p. 59) in this paper, we examine the role of competitive acumen in designing operational-level strategy and business processes for business sustainability. Competitor analysis, a set of methods to assess the strengths and weaknesses of current and potential competitors, is a key task for operations managers as they scan their competitive terrain, attempt to understand their market structure, shore up their defenses against likely competitive incursions, improve their business process of core activities, and plan competitive attack and response strategies (Aho and Uden, 2013; Allen and Helms, 2006; Bergen and Peteraf, 2002; Espino-Rodriguez and Rodriguez-Díaz, 2014). These methods to collect competitor information and draw inferences have been the lifeblood of operations managers and the focus of much academic research in the management literature (e.g. Grossler and Grübner, 2006; Hamel and Prahalad, 2005; Porac and Thomas, 1990; Porter, 1980, 1985; Zajac and Bazerman, 1991). It informs operations managers with product design strategies for the price sensitive audience, rich or simple product assortments or the differentiated mix of both. These have direct implications performance on a firm's operations in terms of better quality, lower cost and flexibility in adapting to changing market trends (de Waal and Batenburg, 2014; Shamsuzzoha, 2011). Finally, operations managers will have to constantly refer to the updated competitive intelligence to re-engineer their product strategies according to changing market perspectives (Wieland et al., 2015). Previous research indicates that competitor analysis helps firms appreciate interactive market behavior, understand firm rivalry, strategize for superior competitive gains (Caves, 1984; Porter, 1980; Scherer and Ross, 1990), and improve their assessment of competitors' competencies and the threats these represent (Zajac and Bazerman, 1991).

Competitor analysis is a multi-disciplinary function affecting sales, marketing, product development, operations strategy, and product re-engineering. It requires diverse information spanning these departments. In the past, with fewer available sources of information, conducting competitor analysis was a difficult activity, limiting the power of advanced business analytics, such as conjoint analysis (Green and Srinivasan, 1978), multidimensional scaling (MDS) mapping (Elrod, 1988, 1991; Elrod et al., 2002), market clustering (DeSarbo et al., 1991), and “voice of consumer” analysis (Griffin and Hauser, 1993).

The advent of the internet has led to superior methods of information collection and analysis (Lee and Bradlow, 2011). However, conducting such analyses to monitor competitors can be a time-consuming process, with a vast amount of information available. Even though the content on the web is enormously helpful, it does present difficulties. With overwhelmingly large quantities of data, the task of constantly tracking and detecting new sources of information and then assimilating the knowledge from various online sources would be a big challenge. Further more severe constraint is imposed by the unstructured set of online data that is primarily qualitative in nature and full of noise. Godes et al. (2005) note that one of the challenges in using online content is the impossibility of automatically analyzing textual information. Very often, expert systems, based on the human interpretation of knowledge, are used to extract wisdom (Jayaraman and Srivastava, 1996). However, these methods are expensive, and due to a lack of automated tools, such data sources often remain unused.

Recently, with the increasing capabilities of big data analytics, managers are looking to collect market information from vast pools of data and automatically analyze it to search for meaningful knowledge (Feldman et al., 2010). This is especially useful for the extraction and coding of vast amounts of unstructured data in the form of text content, such as product descriptions, expert reviews, blogs, customer reviews, employee testimonials, investor reports, and media news (Doan et al., 2011), that have previously needed a great deal of human intervention, such as hand-coded rules and keyword-based searches, to extract information from web-based text (Shi and Yu, 2013). In this scenario, modern machine learning methods like statistical natural language processing (NLP) with fewer context-specific rules tailored to
specific domains can be useful to improve knowledge relationships (Netzer *et al*., 2012). Some terminologies similar to NLP include content analytics, text mining, ontology induction, concept hierarchy, and so on, all of which use the same methodology of grouping terms into concepts and of identifying different types of relationships between concepts.

Our objective is to harness the growing body of free online content for automated competitor analysis that does not rely on a predefined set of language rules or ex-post interpretation of derived dimensions from consumer surveys. We intend to provide operations managers with a cost-efficient tool to monitor a firm’s market position in real time with higher resolution and at lower cost than traditional methods. We combine the techniques of Web Crawler, Naïve Bayes, Latent Dirichlet Allocation (LDA) Topic Modeling, MDS, K-Nearest Neighbors (k-NN) Clustering with data visualization to develop a big data competitor analysis system that can inform operations managers about competitors and the meaningful relationships among them. We illustrate our approach, using the fitness mobile app business. We develop a big data system to collect structured as well as unstructured data from multiple sources, to analyze online content automatically, and to discover competitor knowledge.

We contribute to the operations management literature by presenting a novel big data methodology with management theories to expand the traditional scope of competitor analysis. Our work complements previous studies in at least four ways. First, we present a simpler, automatic and easily replicable method when compared to existing methods. Our method analyses producers’ self-provided product descriptions and users’ social media data while requiring minimal human intervention. Second, our method is truly “big data” based – large volume (i.e. 1,381 mobile apps; 100,892 user comments; and 95,705 Google Plus user profiles), various formats (i.e. numerical and textual data) and from various sources, which provides more information than traditional market structure analysis and advances existing competitor analysis methods. Third, our approach can automatically map out market positions of similar products, visualize product attributes and reveal customer segmentation. Thus, our method not only facilitates direct competitor analysis, but also identifies indirect as well potential competitors. Finally, our method not only captures the attributes that a producer should emphasize in a product’s description, but also tracks subtle differences in vocabulary that may separate brands or identify unique submarkets. These include attributes that are not highlighted by the use of more traditional text mining methods to elicit attributes and dimensions (Lee and Bradlow, 2011; West *et al*., 1996).

In the following section, we describe the current state of research with respect to competitor analysis, big data analytics and NLP. Thereafter, we describe our methodology and apply it to the fitness mobile app business. We conclude with a discussion of the potential of our big data competitive analysis system, its limitations, and directions for future research.

**Literature review**

**Competitor analysis**

In order to remain competitive, it is essential for operations managers to have a clear understanding of their firm’s competitors (Calori *et al*., 1994; Hodgkinson and Johnson, 1994; Porter, 1980) and make right operations strategies that reflect the planning, design and implementation of strategic decisions that span across business processes of a firm (Barnes, 2001; Englyst, 2003; Minarro-Viseras *et al*., 2005; Miller and Roth, 1994; Paiva *et al*., 2008; Riis *et al*., 2006; Slack and Lewis, 2002; Rytter *et al*., 2007; Ward *et al*., 1996) and strengthen a firm’s competitiveness in the market through improved quality, better delivery, lower cost and enhanced market adaption flexibility (Alegre-Vidal *et al*., 2004; Amoako-Gyampah and Meredith, 2007; Boyer, 1998; Boyer and Pagell, 2000; Boyer and McDermott, 1999; Christiansen *et al*., 2003; Dangayach and Deshmukh, 2001; Diaz *et al*., 2005; Flynn and Flynn, 2004;
In essence, competition between firms refers to the rivalry between their respective business lines at both product- and firm-level. Competitor analysis, also called "competitors' acumen" (Tsai et al., 2011), provides valuable competitive intelligence by creating an accurate, strategic method of understanding competitor's operations (Bloodgood and Bauerschmidt, 2002) and thus becomes a driver of competitive success (Lamb, 1984). Porter (1980) argues that competition between the firms can be classified based on customer offerings that differ in terms of specific functions and associated ease of use, technology, the raw materials used in producing the product and the market segment being catered for. Czepiel and Kerin (2012) further suggest that competitors can be classified into three categories, namely, direct, indirect, and potential. Knowledge about direct competition is a "must have" for any firm when building its competitive intelligence (Ulrich and Eppinger, 2003). It is also the fiercest form of competition that exists between firms with hardly any differentiation between their product offerings. Firms can become indirect competitors in a given business domain, if they serve the same customer needs but with different resources. Finally, there are the potential competitors who do not serve the same customer base but use the same resource base or have equivalent capability (Czepiel and Kerin, 2012). In short, operations managers need to clearly understand the extent of competition that exists in their domain and benchmark the various types of competitors in the market.

Big data analytics and NLP
Following a data-centric approach to the development of business intelligence, big data analytics can significantly facilitate competitor analysis. From a business-function perspective, big data analytics can be described as a set of analytical techniques used on relatively large data sets and involving complicated digital data-gathering sources (Chen et al., 2012). From an information systems perspective, big data analytics can be viewed as a system with extraordinary data integration and warehousing capabilities, used for information extraction, online analytical processing, and reporting, and based on intuitive data mining, statistical analysis and predictive analytics (Goes, 2014; Schlegel, 2014). Recent related research and application topics in big data analytics include text analytics, web analytics, network analytics, mobile analytics, social media analytics, and sentiment analysis on qualitative data (Chen et al., 2012; Pang and Lee, 2008). All of these are based on NLP techniques to process semi-structured text data, to extract meaningful, relevant and non-trivial information and to discover business knowledge from huge amounts of online content (Dörre et al., 1999; Feldman and Sanger, 2006; Lee, 2007).

NLP is a computational approach to text analysis that originated in the 1950s with the convergence of artificial intelligence and linguistics (Jurafsky and Martin, 2009). Creating fundamental algorithms and mathematical models for human language processing, NLP can be considered as an advanced type of text mining technique used for content analysis, that relies more on complex language processing and less on hand-coded parsing rules, hence requiring minimal human intervention only in the analysis phase (Hu and Liu, 2004; Lee, 2005). NLP forms part of the computational linguistics stream of text analytics that can perform lexical acquisition, word sense disambiguation, part-of-speech-tagging, and probabilistic context-free grammars in a given unstructured text context (Manning and Schutze, 1999). Current approaches to NLP are based on probabilistic machine learning and knowledge engineering methods that examine and use patterns in data to improve a program's own understanding (Krauthammer and Nenadic, 2004).
Therefore, many statistical NLP models are able to deliver relatively accurate analytics results and close to human interpretation (Jurafsky and Martin, 2009).

Some applications of text mining and NLP in the management literature include analysis of the relationship between product attributes and sales (Archak et al., 2011); hotel-room demand estimation based on text mining (Ghose et al., 2012); corporate stock performance, by mining the text of sentiment and star ratings of product reviews (Seshadri and Tellis, 2012); ascertaining consumer preferences for products through user comments (Decker and Trusov, 2010); the determination of relationships and predictive analytics of demand in response to price changes (Archak et al., 2007; Ghose and Ipeirotis, 2008; Kamakura and Russell, 1989); marketing strategy (Erdem and Keane, 1996); and new product design (Srivastava et al., 1984). In the area of competitor analysis, Pant and Sheng (2009) mine corporate-level text data, using network linkages between web pages and online news to identify market competitors. Feldman et al. (2007, 2008) apply NLP to extract and visualize relationships between product brands from online blogs. Lusch et al. (2010) extract market information by analyzing the “conversations” between firms and customers. More recently, Lee and Bradlow (2011) and Netzer et al. (2012) improve understanding of market structure by text mining semi-structured product attributes. However, most previous studies still use manual rules and human tagging to understand complex linguistic patterns. Also it is noteworthy that most prior studies consider only “the voice of the customer”, which may not be reliable as the sole reference point for product- or firm-level competitor analysis. In our study, we improve on prior work by fully automating information extraction from multiple online sources. We believe this is the first system of its kind. Our methodology can easily be adapted to multiple settings and scenarios, while yielding quicker, reliable, real-time results. By limiting the need for human intervention, our system is cost effective for market analysis projects covering different data sources. In addition, the availability of various types of free online content, such as the continuous stream of expert reviews, product descriptions, user comments, and social media user profiles, provides a practical reason to augment traditional methods (such as surveys and focus groups) of conducting competitor analysis, which can be used continuously, automatically, inexpensively, and in real time. Thus, this study intends to introduce a fast, inexpensive big data analytics system that can have a significant impact on operational decisions. In the remainder of this paper, we describe our methodology, apply our approach to 1,381 health and fitness mobile Apps, and demonstrate how our approach helps operation mangers in conducting competitor analysis.

Methodology
In this study, we classify competitors by applying binary Naïve Bayes, LDA topic modeling and k-NN Clustering, and conduct competitor analysis by applying various data tabulation and visualization techniques. To do so, we develop a big data competitor analysis system (see Figure 1) that can specifically deal with the difficulties involved in collecting data from multiple sources and mining unstructured textual information. The system is divided into three parts: data collection, data classification and data analyses. We first developed three web crawler programs that collected the data from three websites – Google Play app store for the producer self-provided descriptions of the applications within the health-fitness category and the user comments of each application; Google Plus for the demographic information of each user that has commented at least one application within the health-fitness category; and Factiva for the expert reviews of each application in our sample. Then we conducted NLP-based classification analyses to eliminate the non-fitness applications from our sample and extracted ten categories of application functionalities as well as four clusters of applications. Finally, a series of competitor analysis metrics were developed and applied to examine the fitness mobile application market.
We do not focus on the fitness mobile app business by chance, but rather due to its popularity and its fit with big data analytics of freely available online content. The “anything, anytime, anywhere” mobile sports app gold rush is on, with an unprecedented number of firms engaged in methods to best monetize the high-value touch points between their apps and users (Flurry, 2014). However, the fitness app’s pick-and-shovel business is not thriving. Although a few perceptive mobile application companies are reaping significant gains, the vast majority of firms find themselves with little revenue. Therefore, competitor analysis is a key step in designing and developing new mobile apps, as well as in
repositioning existing ones in the mobile app business. It can help managers understand
the substitution and complementary relationships between the brands and alternatives
that define the market, predict marketplace responses to changes, and make appropriate
operational decisions.

Data collection
We used a set of web crawler programs to collect data from the following websites:
Google Play app store and developer console (play.google.com, under the category “health and
fitness”) from which we collect each app’s name, developer firm’s name, number of downloads,
user evaluation (from 1 to 5), self-provided description, price, and in-app purchase items;
Google Plus from which we collect each user’s recommendation and/or comments for each app
and profile information; and The Factiva database (www.factiva.com, a database by Dow
Jones and Company), that aggregates content from more than 32,000 sources such as
newspapers, journals, magazines, television and radio transcripts, and so on from which we
collect fitness app expert reports. Up to December 31, 2014, we collected data about 1,381
health and fitness apps available on Google’s Play Store (US market) and 26 fitness app expert
reports. We then deleted HTML tags and non-textual information such as photos, icons, and
videos. We also removed punctuation marks and stop words based on NLTK’s Stop words
Corpus (Bird et al., 2009). Finally, all capitalized words were converted to lowercase.

Data classification
Naïve bayes classification to eliminate non-fitness apps. The first task of competitor analysis
is to identify rivals that offer similar products/services and compete head-on in the
marketplace. However, Google Play’s “health and fitness” category lists apps for nutrition,
sleep, and healthy lifestyle along with those for sports, fitness and workout. Therefore, we
selected only fitness related apps from our sample, using a binary naïve Bayes algorithm.

Naive Bayes is based on probabilistic models using Bayes’ theorem to create
classification tasks, which makes it possible to predict for uncertain situations
(probabilities) based on prior knowledge (probability values) (Wang et al., 2011). Being
the simplest of the Bayesian classifiers, Naive Bayes follows a structural model assuming
conditional independence for all instance attributes in a given class. It is applied to the
learning of linear functions alone using binary values (Duda and Hart, 1973). It assumes that
every independent attribute in the given class is equally important. The advantage of the
above assumption is that the model has the flexibility to estimate each attribute separately.
Prior research has found that this simplified assumption works well for textual information
classification, having “text” as the unit of analysis (Chen et al., 2009; Koller and
Sahami, 1997; Li, 2010; McCallum and Nigam, 1998; Rennie et al., 2003; Sahami, 1996;
Yu et al., 2013). Friedman (1997) and Domingos and Pazzani (1997) argue that Naive Bayes is
an excellent classification method because it has high-classification accuracy, is easy to
implement and is relatively effective in text classification tasks. Therefore, we chose the
binary Naive Bayes as our classifier, as our sample data have a large number of attributes
proportional to large vocabularies. The classifier works by transforming an app’s
self-provided description into a list of strings and then to a feature (i.e. word) vector level.
The classifier then calculates the prior probability of each class, which is determined by
checking the frequency of each class in the training set. Every feature helps determine
which class should be assigned to a given input value. The contribution from each feature is
then combined with this prior probability to arrive at a likelihood estimate for each class and
to choose the class that has the highest value.

In our binary classification tasks, the binary set of classes was defined in advance – 1 for
fitness apps and 0 for other types. The objective of this binary classification was to choose
the correct class label (1 or 0) for a given app. The success of this machine learning method lies in selecting relevant features and encoding them. Typically, feature extractors are built through a trial-and-error process, guided by intuition based on relevant information (Bird et al., 2009). We established a panel of five IT and mobile business experts to develop a binary feature classifier. First, the panel carefully read the self-provided descriptions of five popular fitness apps featured on the app intelligence website AppAnnie (www.appannie.com), including Runkeeper, Nike+ Runner, Endomondo Sports Tracker, Noom Coach, and Runtastic Six Pack Abs Workout. The panel then used these five apps for three weeks. They selected 92 keywords from the descriptions of these five apps. Next, the panel read 26 expert reports on the fitness apps business from leading online publications specializing in IT and e-Business such as CNET, PCMag, TechRadar, ZDNet, and TechCrunch. They identified a set of 179 keywords (121 unigrams, 39 bigrams and 19 trigrams) and merged them with the previous 92 keywords to form a set of relevant product features of a typical fitness app.

Once an initial set of features was chosen, the panel manually annotated and classified 100 randomly selected apps from the 1,381 sample apps into two categories – fitness app (i.e. class 1) or non-fitness app (i.e. class 0). These 100 apps were randomly split into three sets: the training set (20 apps) that was used to train the classifier model with the trial-and-error approach by classifying the training sample into the two classes with the initial set of features; the development test set (30 apps) to identify errors and rebuild the set of features; and the test set (50 apps) to validate the classifier model. Our binary Naïve Bayes classifier began by calculating the prior probability of each category (i.e. the frequency of each class in the training set). The contribution from each feature was then combined with this prior probability in order to arrive at a likelihood estimate for each class. An app was then classified as a fitness app if the class 1 likelihood estimate was higher than that of the class 0.

We compared the results of the automatic classifier model with those of the manual classification performed by the panel. We then revised and retested the set of features with the development test set. This generated a list of the errors that the classifier made when predicting whether an app focuses on fitness or not. We examined individual error cases where the model predicted the wrong category to determine what additional pieces of information would enable the classifier to make the right decision, or which existing pieces of information lead to a wrong decision. We then adjusted the product feature set accordingly.

We used the following metrics (Jurafsky and Martin, 2009; Salton and McGill, 1983) to evaluate the binary Naïve Bayes classifier in the test set by comparing the classes that it generated with the correct classes manually classified by the panel:

(1) Accuracy, the overall correctness of the classifier, was calculated as the sum of correct classifications divided by the total number of classifications. The accuracy value of our classifier was 0.801.

(2) Precision, the number of correct classifications made for each class divided by the total number of classifications predicted by the specific class was estimated as 0.863.

(3) Recall (also called sensitivity), a measure of the ability of a prediction model to select instances of a certain class from a data set, was calculated as the number of correct classifications made for each class divided by the total number of test examples of the considered class. The recall measure of our classifier was 0.791.

(4) The F-measure (or F-score), which combines precision and recall to give a single score, is defined as the harmonic mean of precision and recall (i.e. \(2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall})\)). The F-value of our classifier was 0.825.

These four metrics indicated that the performance of our binary Naïve Bayes classifier was quite satisfactory. We then employed this classifier to select fitness apps from the whole
sample, resulting in 846 non-fitness apps being eliminated. Finally, the panel revised the list of keywords and followed the approach of Ulrich and Eppinger (2003) and Lee and Bradlow (2011) to group them into ten categories of product functionality representing the constituent dimensions of a fitness app (keywords in brackets): walking tracker (e.g., walk, pedometer, step counter), running tracker (e.g., run, jogging, hill run), biking tracker (e.g., ride, off-road, speedometer), gym workout and aerobic dance (e.g., coaching, abdominals, Pilates), health metrics (e.g., heart rate, blood pressure, calorie counter), challenger and motivator (e.g., cheering, chase, leader board), navigation (e.g., GPS, maps, turn-by-turn), artificial intelligence algorithms (e.g., auto stop, automatic calibration, voiceover feedback), wearable accessory (e.g., smartwatch, wristband, compatibilities), and yoga and weight control (e.g., yoga, fat loss, weight recorder). These ten functional categories of the product will be used to examine within-cluster rivalry in the subsequent sections.

LDA topic modeling, MDS and k-NN clustering to identify potential competitors. For the remaining 535 fitness apps, the goal of competitor analysis was to identify three types of competitors (Czepiel and Kerin, 2012): direct (i.e. directly competing with other firms to serve the same customer needs using the same resources), indirect (i.e. serve similar customer needs but with different resources) and potential (i.e. serve the similar or same customer base but have different resources or capability). To do so, we plotted the market positions of the apps based on the scales of differing purchase needs and place them in different clusters based on their functionalities.

We first applied probabilistic topic modeling (PCM) and MDS to conduct a market position mapping analysis. PCM is a type of machine learning method that makes it possible to explore documents automatically based on themes that run through a collection or corpus (Blei, 2012). PCM makes use of probability-based algorithms to detect “thematic structure” based on the topic distribution in large pools of online documents and depicts documents as a bundle of these topics (Blei et al., 2003; Griffiths and Steyvers, 2002, 2003, 2004; Hofmann, 1999, 2001). A topic is derived from a probability distribution over a fixed vocabulary from a particular subject (Blei, 2012). PCM is a powerful method for automatically organizing, understanding, searching, and summarizing extensive electronic textual data. It offers an alternative method of conducting content analysis and retrieval that is often based on the keyword-based search method. It can also detect the relationship between different documents in a given collection based on the topics or themes that run through these documents (Vulić et al., 2015).

We adopted the LDA PCM algorithm, which uses a statistical approach based on probabilities to detect the themes that run through a collection of documents, assuming that one document can have multiple topics (Blei, 2012). LDA is similar to principal component analysis for discrete data (Buntine, 2002; Buntine and Jakulin, 2005) based on statistical assumptions about the corpus and topics. LDA relaxes the sequence of words condition by considering a document as only a “bag of words” (Blei, 2012) and uses the order of words to generate topics (Wallach, 2006). In addition, LDA also uses a generative probabilistic model based on the hidden structures for generating words and thus does not require pre-annotated documents (Blei, 2012).

We followed the commonly used “kitchen sink” approach (Bird et al., 2009) of treating each app’s self-provided description as a topic to create a wide assortment of product features. The distribution of words in the description forms a topic’s vocabulary of words (Blei, 2012; Steyvers and Griffiths, 2007). Each app’s description was considered as an observable element. The topic distribution per description and the distribution of words per topic per distribution were considered as hidden structures. Using probability distributions, we constructed the hidden thematic structures that were assumed to resemble the thematic structure of the collection and hence, we calculated a bivariate-topic (i.e. between two apps)
functionality similarity index (i.e. an indicator measuring how similar two apps were) using
the algorithm of Blei et al. (2003). Assuming that each app represents a topic, a similarity
matrix of $535 \times 535$ was generated.

We then employed a traditional market structure analysis and visualization tool – MDS
to generate product-level market position mapping of the 535 fitness apps. MDS is a means
of visualizing the level of similarity of individual products, by plotting the products on a
2D map. To do so, we followed Netzer et al. (2012), and applied principal component analysis
to reduce the dimensions from 535 to 2. The result of eigenvalue analysis indicated that
31.69 percent of total variance of the similarity matrix was explained by the two factors.
We plotted the market positions of the 535 fitness apps in a MDS in Figure 2, which mapped
their relative positioning differentiated on their features. This figure was instrumental to
our understanding of the classification of these apps in terms of various physical activities
and smartphone sensors. The $y$-axis measured physical activity intensity ranging from light
to vigorous exercise-oriented workouts. Physical activity intensity refers to the amount of
physical power that the body uses to perform a physical activity. There are different ways
to measure physical activity intensity using various electronic sensors. With the specialized
electronic sensors available, mobile apps can now monitor body movement, heart rate,
respiratory rate, blood pressure, blood sugar, and so on. Hence, the $x$-axis represents the
ability and means adopted by an app to record vital data using built-in or external sensors
such as wristband and smartwatch.

We observed a gradual increment in the intensity of physical activity as we move in an
anti-clockwise direction starting from the third quadrant. The apps in this quadrant act as
training guides for low intensity yoga exercises. These can also be synchronized with
external wearable sensors (e.g. heart rate monitors and wristbands), to collect health metrics
and track body positions. The apps in the lower-left portion of the fourth quadrant deal with
moderately intensive aerobic gymnastics and aerobic dancing activities for women users.
The apps in the upper-right portion of the fourth quadrant and the lower-right portion of the
first quadrant focus on unisex intensive indoor physical workout activities. Finally, the apps
in the second quadrant and the upper-left portion of the first quadrant are the outdoor sports
trackers (e.g. walking, running, hiking, and biking) with built-in smartphone sensors or
wearable widgets or both.

To group these 535 fitness apps accurately in such a way that apps in the same cluster
are more similar to each other than to those in other clusters, we followed Netzer et al. (2012)
and used the two-dimension coordinates of the MDS to run a k-NN cluster analysis. k-NN is
one of the most popular supervised machine learning algorithms (Adeniyi et al., 2014;
Chen et al., 2012; Przewozniczek et al., 2010; Weiss et al., 2010; Wu et al., 2007). It is based on the assumption that things that look alike must be alike (Cover and Hart, 1967). In general, k-NN offers high-accuracy classification with no prior assumptions about the data and is not sensitive to outliers (Cheng et al., 2008; Han et al., 2001; Kacur et al., 2011; Ogbonaya, 2008; Xiao-peng and Xiao-gao, 2007). We found that four clusters could best fit the data (see the dashed ovals in Figure 2):

1. **Cluster A** (156 apps): the apps in this cluster are used for tracking sport activities such as running, biking, and walking, and for providing users with vital data such as heart rate, steps, and distance. These apps predominantly target users who engage in physical activities (especially outdoor) on a regular basis and rely not only on internal smartphone sensors (e.g. GPS antenna, accelerometers and gyroscopes), but also on external widgets (e.g. wristband, smartwatch, heart rate monitor) for performance measurement. Many apps in this cluster are interoperable with a wide variety of external third-party widgets and permit activity performance metrics data sharing with social media websites.

2. **Cluster B** (36 apps): the apps in this cluster act more like personal yoga training guides. The uniqueness of this cluster lies in the way the apps offer an alternative fitness paradigm based on very low intensity yoga exercises. These apps can be connected with external wearable widgets and help users to adjust their body movement. Most of these apps provide animation or video guidance modules.

3. **Cluster C** (73 apps): the apps in this cluster target the health and fitness of women through aerobics, gymnastics, and dancing (e.g. Belly, Pilates, Salsa, and Zimba) training tutorials. Their main purpose is to assist with exercises through videos, with or without the aid of exercise equipment based on varying level of workout intensities. Unlike those in cluster B, these apps are self-supported fitness assistants needing no connection with external wearable widgets and do not require any geo-positional data. They are specifically suited for women’s indoor physical training programs.

4. **Cluster D** (270 apps): the apps in this cluster target the health and fitness segment by providing challenging and motivational exercise video tutorials and a timer/logger for general unisex vigorous physical activities (e.g. calisthenics, strength, rowing machine, and basketball). The apps act as video tutorials and usually track a user’s body movement using the smartphone’s built-in gyroscopes, accelerometers, and magnetometers. These apps usually do not use GPS sensors to record geo-location information as they mainly deal with indoor physical training programs. Finally, most apps in this cluster synchronize fitness performance data with social media websites.

Following Czepiel and Kerin’s (2012) classification of competitors, we regard that apps within the same cluster are direct and indirect competitors as they address the same or similar customer needs with the same or similar resources. The apps in other clusters are usually potential competitors as they address the same customer base (i.e. those who engage in fitness activities) but with different resources or capabilities.

**Direct and indirect competitor analyses**

**Product-level competitor analyses.** In order to identify market leaders who usually are the most powerful rivals for many products in the same cluster, we first conducted within-cluster ranking with three different metrics:

1. **Number of downloads**, which reflects the total of user downloads as obtained from Google Play app store (US market). However, this number does not comprehensively reflect the actual or perceived (by users) performance of the app, especially for free
apps that can be easily discarded by users. This metric can be used to calculate market share (i.e. number of downloads of an app divided by the total number of downloads of all apps in the same cluster).

(2) Google Plus recommendation, derived from the number of users recommending an app on their Google Plus webpage. This metric infers the overall popularity of an app, but cannot be considered to reflect the general user experience, as it only indicates satisfied users.

(3) User satisfaction, which factors in both the positive and negative feedback of an app from users on a scale of 1 (very unsatisfied) to 5 (very satisfied). We weighted the average score by the number of downloads of each app to overcome the disparity in the number of downloads across popular/old and unpopular/new apps. This measure provides a better, unbiased performance indicator based on users’ positive and negative feedback.

Table I summarizes the rankings using these three metrics. In cluster A, we observed that even though the free app ANT + Plugins, which offers a robust system integrator using WiFi, NFC, and Bluetooth connection to access external widgets from a smartphone, scored highly on the number of downloads, it was not among the top-five apps in terms of Google Plus recommendations and user satisfaction. Perhaps this was because the app was far too generic, without dedicated features for fitness activities. The other top-five apps (both free and paid) had similar scores for the three ranking metrics. The ranking in cluster B followed a predictable pattern as the same apps scored similarly on the three metrics. The app Daily Yoga – Fitness On-the-Go emerged as the top free app with a wide variety of training videos and live video assistance. Similarly in the paid segment of this cluster, Pocket Yoga was the best seller. Likewise, the top-five free and paid apps in cluster C were almost identical across the three metrics. We observed that stretch exercises were among the top-five downloads but it failed to expand its reach through Google Plus recommendations and user satisfaction. Finally, in cluster D, the rankings of both free and paid apps were consistent across the three metrics. In the paid segment of this cluster, Runtastic Sit-Ups PRO was among the top-five most downloaded apps but it failed to register among the top recommended or satisfactory apps.

Product functionality comparison through radar charts. The top app league table calls for a more detailed study of these apps to identify and compare their specific features. In this section, we map the functionality dimensions of these top-five apps (in terms of user satisfaction) on radar charts by calculating each app’s semantic similarity against the ten functionality components as in the previous section of the naïve Bayes classification using the Log-Hyponym algorithm (see Miao et al., 2010; Pirró and Seco, 2008). The higher score of one functionality component represents a higher proportion of words in the app’s self-provided description relevant to this functionality component. Even though, we chose these top-five apps for illustration purpose, the radar chart analysis can be applied to other apps in our sample.

From Figure 3(a)-(d), we find that in cluster A, RunKeeper emerged as the best overall performer, excelling on functional aspects such as tracking and logging for biking, gymnastic workouts, yoga, and weight loss. Runkeeper also outperformed the other four apps in terms of interoperability with a wide variety of wearable widgets. Runtastic running and fitness, however, excelled on navigation. It is worth noting that all apps, except cardiograph, performed poorly in their ability to measure internal body statistics such as heartbeat beyond physical activity workout statistics. Focusing on the health metrics dimension, cardiograph differentiated itself in cluster A by offering scope for complementarities with other apps. In cluster B, all top-five apps focused on a specific market niche with very similar functionalities on the yoga/weight control feature. Daily Yoga – Fitness On-the-Go emerged as
<table>
<thead>
<tr>
<th>Download</th>
<th>G+Recommendation</th>
<th>User satisfaction</th>
<th>Download</th>
<th>G+Recommendation</th>
<th>User satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Free app</strong></td>
<td></td>
<td></td>
<td><strong>Cluster B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. 1</td>
<td>ANT + Plugins</td>
<td>RunKeeper – GPS Track</td>
<td>RunKeeper – GPS Track</td>
<td>Daily Yoga – Fitness On-the-Go</td>
<td>Daily Yoga – Fitness On-the-Go</td>
</tr>
<tr>
<td>No. 2</td>
<td>RunKeeper – GPS Track</td>
<td>Run Walk</td>
<td>Run Walk</td>
<td>Simply Yoga FREE</td>
<td>Simply Yoga FREE</td>
</tr>
<tr>
<td>No. 3</td>
<td>Runtastic Running &amp; Fitness</td>
<td>Cardiograph</td>
<td>Endomondo Sports Tracker</td>
<td>Daily Yoga for Abs</td>
<td>Daily Yoga for Abs</td>
</tr>
<tr>
<td>No. 4</td>
<td>Cardiograph</td>
<td>Endomondo Sports</td>
<td>Cardiograph</td>
<td>Yoga Breathing for Beginners</td>
<td>Yoga Breathing for Beginners</td>
</tr>
<tr>
<td>No. 5</td>
<td>Endomondo Sports Tracker</td>
<td>Nike + Running</td>
<td>Nike + Running</td>
<td>Yoga for Body Toning</td>
<td>Yoga for Body Toning</td>
</tr>
<tr>
<td><strong>Paid app</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. 1</td>
<td>Runtastic PRO</td>
<td>Runtastic PRO</td>
<td>Pocket Yoga</td>
<td>Pocket Yoga</td>
<td>Pocket Yoga</td>
</tr>
<tr>
<td>No. 2</td>
<td>Endomondo Sports Tracker PRO</td>
<td>Endomondo Sports Tracker PRO</td>
<td>Simply Yoga</td>
<td>Simply Yoga</td>
<td>Simply Yoga</td>
</tr>
<tr>
<td>No. 3</td>
<td>Zombies, Run!</td>
<td>Zombies, Run!</td>
<td>A Facial Yoga &amp; Facelift</td>
<td>A Facial Yoga &amp; Facelift</td>
<td>A Facial Yoga &amp; Facelift</td>
</tr>
<tr>
<td>No. 4</td>
<td>Runtastic Pedometer PRO</td>
<td>Runtastic Pedometer PRO</td>
<td>Complete Yoga For Beginners</td>
<td>Complete Yoga For Beginners</td>
<td>Complete Yoga For Beginners</td>
</tr>
<tr>
<td>No. 5</td>
<td>Instant Heart Rate – Pro</td>
<td>Instant Heart Rate – Pro</td>
<td>Yoga Training</td>
<td>Yoga for Weight Loss Pro</td>
<td>Yoga for Weight Loss Pro</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>No.</th>
<th>Free app</th>
<th>Paid app</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ladies' Ab Workout</td>
<td>Gym Book: training notebook &amp; Log</td>
</tr>
<tr>
<td>2</td>
<td>Ladies' Butt Workout</td>
<td>Circuit Training Assistant Pro</td>
</tr>
<tr>
<td>3</td>
<td>Ladies' Waist Workout</td>
<td>Tabata Pro – Tabata Timer</td>
</tr>
<tr>
<td>4</td>
<td>Butt Workout</td>
<td>Ladies’ Butt Workout Circuit Training Assistant Pro</td>
</tr>
<tr>
<td>5</td>
<td>Stretch Exercises</td>
<td>Ladies’ Chest Workout Tabata Pro – Tabata Timer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Download</th>
<th>G+Recommendation</th>
<th>User satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ladies' Ab Workout</td>
<td>Ladies' Ab Workout FREE</td>
<td>Noom Coach: Weight Loss</td>
</tr>
<tr>
<td>2</td>
<td>Ladies' Butt Workout</td>
<td>Ladies' Butt Workout FREE</td>
<td>Abs workout</td>
</tr>
<tr>
<td>3</td>
<td>Ladies' Waist Workout</td>
<td>Ladies' Waist Workout FREE</td>
<td>Workout Trainer</td>
</tr>
<tr>
<td>4</td>
<td>Butt Workout</td>
<td>Circuit Training Assistant FREE</td>
<td>Daily Ab Workout FREE</td>
</tr>
<tr>
<td>5</td>
<td>Stretch Exercises</td>
<td>Butt Workout</td>
<td>8 Minutes Abs Workout</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Download</th>
<th>G+Recommendation</th>
<th>User satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gymrat: Workout Tracker &amp; Log</td>
<td>Gymrat: Workout Tracker</td>
<td>Just 6 Weeks</td>
</tr>
<tr>
<td>2</td>
<td>Zumba Dance</td>
<td>Circuit Training Assistant</td>
<td>Runtastic Push-Ups PRO</td>
</tr>
<tr>
<td>3</td>
<td>Circuit Training Assistant</td>
<td>Tabata Pro – Tabata Gym Book: training notebook</td>
<td>Fitness Buddy: 1,700 Exercises</td>
</tr>
<tr>
<td>4</td>
<td>Tabata Pro – Tabata</td>
<td>Gym Book: training notebook</td>
<td>JEFIT Pro – Workout &amp; Fitness</td>
</tr>
<tr>
<td>5</td>
<td>Gym Book: training</td>
<td>Tabata Pro – Tabata Timer</td>
<td>Runtastic Sit-Ups PRO</td>
</tr>
</tbody>
</table>

Table I.
Notes: (a) The functionality component comparison of cluster A; (b) the functionality component comparison of cluster B; (c) the functionality component comparison of cluster C; (d) the functionality component comparison of cluster D
the best overall performer as it provided a wide variety of yoga exercises and live voice guide. Simply Yoga FREE scored better than the other four apps on the gym dimension because it specializes in simple yoga tutorials. In cluster C, although circuit training assistant did not score highly in any of the three rankings, it managed to distinguish itself by providing a huge variety of workout videos targeted at women and by allowing a high level of customization in the workout regimes. Finally, unlike most apps in cluster D, Noom Coach: Weight Loss emerged as the top performer by integrating built-in fitness programs, diet tracking and workout coaching in one app. It relies heavily on wearable widgets and AI algorithms to control weight. All top-five apps in this cluster are comparable on the coach/motivator and gymnastics/dance dimension with a wide variety of workouts and exercise tutorials.

Revenue model and pricing. Table II shows the results on the revenue models followed by the fitness apps. Based on the free or paid and with or without in-app purchases business model components, the numbers of mobile apps are tabulated for each cluster. In cluster A, the free revenue model dominated the cluster with approximately 54 percent apps 68 percent of free apps and 86 percent of paid apps opted for without in-app purchases. In cluster B, rather more than two-thirds (69 percent) of the apps followed the free revenue model and only 3 out of 36 chose the within-app purchase model. In cluster C, the free revenue model dominated the cluster with 59 percent free apps. Just 4 out of 73 apps adopted for the within-app purchase model. Finally, 54 percent of the apps in cluster D followed the free revenue model whereas just 4 percent of the paid apps chose the within-app purchase model. In general, free (55.7 percent) was the dominant revenue model in the fitness app business and firms seemed to be reluctant to choose the within-app purchase model (15.14 percent).

Table III summarizes the pricing trend for paid, free and in-app purchases of apps in general. The price of a paid app ranges from $0.50 to $19.90 while the prices of the paid in-app purchase items vary from as low as $0.50 to as high as $199. However, prices for the free in-app purchase items varied less, from $0.71 to $55.

Firm-level competitor analyses

Market structure analysis. As a mobile app firm can own multiple apps in different clusters, in this section, we aggregate app-level to firm-level and conduct firm-level competitor analysis. Table IV shows the trend in the number of apps of which a particular mobile app

<table>
<thead>
<tr>
<th></th>
<th>Cluster A</th>
<th>Cluster B</th>
<th>Cluster C</th>
<th>Cluster D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paid</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without in app purchase</td>
<td>63</td>
<td>10</td>
<td>29</td>
<td>118</td>
<td>220</td>
</tr>
<tr>
<td>With in app purchase</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Paid total</td>
<td>73</td>
<td>11</td>
<td>30</td>
<td>123</td>
<td>237</td>
</tr>
<tr>
<td><strong>Free</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without in app purchase</td>
<td>57</td>
<td>23</td>
<td>40</td>
<td>114</td>
<td>234</td>
</tr>
<tr>
<td>With in app purchase</td>
<td>26</td>
<td>2</td>
<td>3</td>
<td>33</td>
<td>64</td>
</tr>
<tr>
<td>Free total</td>
<td>83</td>
<td>25</td>
<td>43</td>
<td>147</td>
<td>298</td>
</tr>
</tbody>
</table>

Table II.

<table>
<thead>
<tr>
<th>Pricing model</th>
<th>Price</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paid</td>
<td>19.9</td>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>Paid with in app purchase</td>
<td>199</td>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>Free with in app purchase</td>
<td>54.99</td>
<td></td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table III.
A firm may own. We identified a total of 269 mobile app firms from the sample of 535 mobile apps. The majority of the firms (68.02 percent) owned just one app, while nine firms had between five and nine mobile apps, seven firms had 10 to 20 mobile apps and two firms had more than 20 mobile apps. Table II shows the distribution of firms across the four clusters. The majority of the firms (92.94 percent) operated in a single cluster – 104 and 123 firms in clusters A and D, respectively, whereas clusters B (19 firms) and C (44 firms) were less crowded. A small number of firms offered apps across two clusters and three firms offered apps across three clusters.

We then calculated a firm’s market share by cluster as the cumulative sum of the number of downloads of all its apps in a cluster divided by the total number of downloads for that cluster. We analyzed market structure with the concentration ratio, which is the measure of the percentage market share in an industry held by the largest firms within that industry (Bain, 1951). We adopted the most common concentration ratio, CR4, namely the market share of the four largest firms. The results in Table V show that the four largest firms share more than 50 percent of the market in clusters A and D, and more than 80 percent in clusters B and C, suggesting an oligopoly in all four clusters. However, the concentration ratio does not use the market share of all the firms in each cluster. In order to understand market share distribution, we estimated the location and scale parameters for each cluster. The results in Table V suggest that all four location parameters were close to zero and all four scale parameters were much larger than 1, indicating that the market share in all four clusters was not normally distributed. Finally, to gain a complete picture of industry concentration, we calculated the Herfindahl index (HHI). The results of HHI indicated that clusters A and D were less concentrated than clusters B and C.

**Firm-level direct and indirect competitor identification.** We visualized firm-level within-cluster competition with a brand-new 2D chart. The point of origin represents the reference firm.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>No. of firms</th>
<th>Market share</th>
<th>Cluster</th>
<th>No. of firms</th>
<th>Market share</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>ANT+</td>
<td>104</td>
<td>Fitness</td>
<td>CR4</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>Runtastic</td>
<td>19</td>
<td>Keeper, Inc.</td>
<td>HHI</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
<td>Daily Yoga Inc.</td>
<td>44</td>
<td>Endomondo.com</td>
<td>Location</td>
</tr>
<tr>
<td>D</td>
<td>68</td>
<td>IMOBILIFE Co.</td>
<td>123</td>
<td>IMOBILIFE INC.</td>
<td>Scale</td>
</tr>
<tr>
<td>Cross 2 clusters</td>
<td>183</td>
<td>Daily workout apps, LLC</td>
<td>16</td>
<td>Skimble Inc.</td>
<td>CR4</td>
</tr>
<tr>
<td>Cross 3 clusters</td>
<td>269</td>
<td>Sally Tam</td>
<td>3</td>
<td>Daily workout apps, LLC</td>
<td>HHI</td>
</tr>
</tbody>
</table>

Table IV. Number of firms by cluster and number of app

Table V. Market structure comparison
The x-axis measures the relative functionality similarity of other firms in the cluster to the reference firm using the LDA functionality similarity index matrix. A greater distance from the origin to a particular firm along the x-axis denotes a greater similarity in functionality between that firm and the reference firm. The y-axis measures the level at which a particular firm competes with the reference firm in terms of user segments. To do so, we first identified users who have recommended or commented on the apps of the reference firm and those of the other firms in the same cluster by mining data from every user’s Google Plus webpage. Then, we calculated a User Segment Overlap (USO) Index using the following equation:

\[
\text{User Segment Overlap Index}_{i,j} = \frac{\text{User}_{\text{Firm}_i} \cap \text{User}_{\text{Firm}_j}}{\text{User}_{\text{Firm}_i} \cup \text{User}_{\text{Firm}_j} - \text{User}_{\text{Firm}_i} \cap \text{User}_{\text{Firm}_j}}
\]

where \(\text{User}_{\text{Firm}_{ij}}\) is the number of users who have recommended or commented on the apps of Firm \(i\) (the reference firm) or those of Firm \(j\) (any particular firm in the same cluster) on Google Plus.

For the firms that have multiple apps in a cluster, we took the averages of LDA similarity indices and USO indices as measures. To provide a better understanding of the customer influence on product needs and thus help an operations manager with more detailed information, we split users into two groups based on their gender information on Google Plus and calculated the male/female USO indices, respectively. The values along the negative y-axis indicated the female USO index (i.e. the red line in Figure 4) and the positive y-axis represented male USO index (i.e. the blue line in Figure 4). We illustrate our 2D firm-level competition graph with Runtastic GmbH (hereafter, Runtastic) as the reference firm. We chose this firm because its apps scored highly in the rankings of clusters A and D.

Figure 4(a)-(b) map Runtastic’s competing firms in clusters A and D, respectively.

We divided each chart into three segments:

1. Area of indirect competitors, which includes a set of firms whose apps are similar to Runtastic in terms of functionality (i.e. close to the origin along the x-axis). For example, firms like Z2 Software in cluster A and Get Fit in cluster D, with some degree of product differentiation, are indirect competitors of Runtastic.

2. Area of direct competition with low user lock-in, which refers to a set of firms whose apps are very similar to Runtastic in terms of their product functionalities and with highly overlapping user segments. Hence, firms in this area can be regarded as direct or immediate competitors of Runtastic. Given that these firms have a high proportion of the same user segments, this means that a large number of their users have used both their apps and those of Runtastic, indicating a high possibility that users will switch to an alternate app. Consequently, changes in Runtastic’s pricing or product features will have direct implications for the other firms. In cluster D, NorthPark is one such firm. Because it operates with only four free apps, user switching to or from Runtastic is highly probable. The trend in cluster A is similar, where firms like MapMyFitness, Endomondo and Garmin are direct competitors of Runtastic and run the risk of user switch over in the free app market.

3. Area of direct competition with high user lock-in. This is a set of firms whose apps are very similar to those of Runtastic. However, very few users have recommended or commented on both Runtastic’s apps and those of the other firm in this set. That is to say, users are thus more loyal to their firm without trying Runtastic’s apps. We infer that the firms in this area will not be easily defeated by Runtastic. For example, in cluster D, Fitness World provides highly similar apps and hence is a
direct competitor of Runtastic. Yet the low value on the y-axis suggests that Fitness World successfully locks its users via its paid apps. We may infer that any incremental change in Runtastic’s pricing or product offers will not generate any substantial impact on Fitness World. A similar trend is also visible in cluster A, where a firm like Sportstracklive operates with two paid apps and exhibits high user lock-in. FitnessKeeper, in contrast also maintains high user lock-in but provides only one free app (Runkeeper). This might well be due to Runkeeper’s significantly superior quality.

In short, our analyses show that market structure analysis can assist managers in designing sustainable operations strategies based on the mapping of their products vis-à-vis competitors offerings. Different product and process re-engineering strategies can be adapted depending on the competitive segment that an app belongs to. Our analyses

Notes: (a) Firm-level comparison in cluster A; (b) firm-level comparison in cluster D

Figure 4.
Firm level comparison in clusters A and D
suggest that one operational competitiveness indicator that firms need to include is that based on service innovation in product design. This can be well illustrated for apps in segment 2 that exhibit poor user lock-in in comparison to other segments. In this segment, customers are constantly trying new products, and hence operations managers may have to be extra vigilant to competitors’ offerings. Competitor intelligence information through our automated market structure analyses can be a new tool in business process re-engineering of mobile apps. Product functionality competitor analyses, revenue model and pricing analyses can help in faster product redesign to adapt to the constantly changing market needs. Moving from the traditional performance metrics of an operational strategy like quality, flexibility, delivery, and costs (Aboelmaged, 2012), we propose innovation in product design as a valuable indicator to create new winning strategies in the marketplace, as explained in the market structure analyses.

**Conclusion**

With the increasingly complex competitive environment, operational strategists are investing increased time and effort in revising sustainability strategies of their firms. In this paper, we investigate competitor intelligence as means to determine the operational/functional strategy for business sustainability. We present through our analyses the competitive performance metrics that can guide operations managers in better design of the strategic objectives. Operational strategies aimed at various pricing models, product differentiation, faster adaptation to competitor offerings may be investigated by firms depending on the segment that a firm falls into, as depicted in the market structure analysis. Meanwhile the recent revolution in information and communication technologies has created new means to collect and examine data in radically different ways. In this paper, we have presented a big data analytics approach to process online text content automatically and to conduct competitor analysis and thus propose operational strategy recommendations. Notably, our approach extends previous work by developing a fully automatic big data competitor analysis system that integrates state-of-the-art NLP, machine learning algorithms and data visualization techniques. We have incorporated multiple data sources to overcome the bias that may exist in consumer feedback data. As we are aware that there is no way of validating the authenticity of customer feedback, we employ sources such as expert reviews and product descriptions offered by firms along with user evaluations and social media profiles to better understand the market competition. Thus, the credibility of our study is substantially enhanced by employing multiple data sources and large amounts of structured as well as unstructured data, subjected to a fully automated approach, which is able to provide both a predictive and a defensive mechanism to identify opportunities and threats, coalesce all relevant sources of competitor analyses into a single framework, and support efficient and effective strategy formulation, operations implementation and performance monitoring.

Broadly, our work contributes to the operations management literature by providing significant cross-fertilization between market structure theories and information technology to improve the efficiency and effectiveness of extracting competitor information. This big data approach is systematic and rigorous enough to inject a large dose of objectivity into the competitor analysis process. It is versatile and flexible enough to handle unstructured text information, which can create substantial economic gain for firms. In particular, we contribute to the study of competition analysis with our new functionality-similarity – user-segment overlapping chart (i.e. Figure 4(a)-(b), which harnesses the power of social media data and becomes a key basis of competition metrics to classify a firm’s rivals into
three different categories. Valuable insights can be gained from all three categories, prompting a swift response to those firms that offer highly similar products and are competing for the attention of the same user segments. The product-level competitor and functionality analyses as well as the market structure analysis can be instrumental in guiding operational managers in gaining a wider market perspective based on customer needs and competitor performance. Finally, these can be employed in designing performance metrics like cost efficiency, product detailing, flexibility in business functions renovations/changes, etc.

In terms of managerial implications, our sonar-like big data competitor analysis system promises to support operational decision making both descriptively and prescriptively. The Naïve Bayes classifier and k-NN clustering separate potential competitors from direct and indirect ones. Our novel LDA topic modeling algorithm, combined with conventional MDS, reveals a product’s market position in comparison with its peers. The league table enables managers to identify top performers, and the product feature radar chart highlights salient app attributes and value propositions. The analyses of the revenue model, pricing policies and market structure enables managers to detect tomorrow’s opportunities and to predict better courses of action regarding all aspects of a firm’s operations. Finally, the rise of hyper-competition makes it necessary to emphasize exploration over exploitation. Operations managers can use the overlapping user segments and functionality similarity indices to map indirect and direct competitors with and without user lock-in. Thus, our system offers firms multiple benefits. With better, explicit information on rival firms in terms of their product features, new product development can be enhanced by delivering superior or unique functionality. Our system also reduces the cost of human intervention while leveraging the vast data marts of free data to obtain competitive intelligence. In summary, our analyses of fitness mobile app firms demonstrated the ability of NLP and machine learning-based big data analytics to assess competitive market structure, gain information on competitor perceptions, and gauge the future moves of rivals using different data sources in a timely and cost effective manner. Our big data competitor analysis system would be helpful for firms in emerging industries in which competitor data are not readily available through conventional marketing studies.

Several limitations should be noted. First, our study does not include historical versions of the descriptions of each app, so we cannot track brand position evolution and evaluate the performance of a firm’s operations in the past. Future research could perform such dynamic analyses by using historical data points either from mobile app firms or from Google Play app store. In addition, using Google Plus recommendations and user comments may cause a self-selection bias. Those individuals who are highly motivated to recommend or comment on an app, typically individuals who have strong opinions or substantial consumer knowledge, are over-represented. Thus, individuals who are indifferent or apathetic are less likely to respond and are therefore being under-represented. This often leads to a polarization of responses, with extreme perspectives receiving disproportionate weight in our app rankings. Future studies should aim at developing sophisticated algorithms to overcome this sample bias issue. Finally, we demonstrated the value of the proposed system in the fitness mobile app business setting. Future research could also explore applications of our approach in other domains in order to test its external validity. In short, we hope that our big data competitor analytics approach provides a first step in exploring the enormous, rich, useful body of online data readily and freely available on the internet. This is just the beginning of our big data journey underpinning new waves of productivity growth, innovation, and consumer surplus as long as the right policies and enablers are in place (McKinsey, 2015).
References


Han, E., Karypis, G. and Kumar, V. (2001), “Text categorization using weight adjusted K-nearest neighbor classification”, Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Hong Kong, pp. 53-65.


Jurafsky, D. and Martin, J.H. (2009), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ.


About the authors
Dr Liang Guo is an Associate Professor of big data and the Deputy Head of the BNP-KPMG Endowed Centre of Entrepreneurship & Innovation at NEOMA Business School, France. He completed his PhD at the Cambridge University.

Dr Ruchi Sharma works as a Lecturer in e-business at the School of Strategy and Leadership, Coventry University, UK. She completed her PhD in Management from ECRICOME PhD Universa in 2016 and worked as a Research Assistant at BNP-KPMG Endowed Centre of Business Models & Entrepreneurial Innovation, at NEOMA Business School, Rouen. Her research interests are business model business model innovation and its impact on value creation and value appropriation for firms operating in a coopetitive ecosystem. She holds a MBA and a BTech Degree in Mechanical Engineering.

Lei Yin is a PhD Candidate in Big Data Analytics at NEOMA Business School, France. He is under the supervision of Dr Liang Guo.

Ruodan Lu is a PhD Candidate in Applied Machine Learning at the Cambridge University, UK. Ruodan Lu is the corresponding author and can be contacted at: rl508@cam.ac.uk

Dr Ke Rong is an Associate Professor in the Institute of Economics, School of Social Science at Tsinghua University in China. Ke has received the PhD Degree from the University of Cambridge and obtained Bachelor Degree in Tsinghua University. Before joining Tsinghua, he was a Senior Lecturer in the University of Exeter and Bournemouth University in UK and a Visiting Scholar in the Harvard Business School. His research interests include business/innovation ecosystems, data driven ecosystem, and platform market and network effects in sharing economy. His research has been published in International Journal of Production Economics, Journal of International Management, Group and Organization Management and Technological Forecasting and Social Change.