Editorial: Where the data meets the road in the Industry 4.0 economy

Introduction

In the emerging Industry 4.0-based economy, businesses are transitioning toward digitally operated models (Del Giudice *et al.*, 2021) relying heavily on large volumes of data for their success. A digitally operated business model is one in which data not only drives decisions but also becomes an integral part of building AI systems that transform the creation of products or the delivery of services. Businesses that fail to transition will be left behind, unable to promote, expand and sustain growth. For example, digitally operated models such as those of Amazon and Tesla significantly outperform traditional operating models by leveraging detailed data on customer behavior to improve their products and services, often customizing experiences and leveraging data produced by a growing customer base to improve the overall value of their offerings (Iansiti and Lakhani, 2020). In the health care and distribution sectors, digitally operated models have been found to increase the value of latent intellectual capital, and therefore the performance of companies (Gashenko *et al.*, 2020; Gravili *et al.*, 2021). Furthermore, these models allow business to increase efficiency and productivity enabling automation of between 25 and 35% of their decision-making processes, especially in management and administrative functions (Hatzius *et al.*, 2023). As Industry 4.0 matures and expands, the level of automation in decision-making is expected to grow substantially, emphasizing the need to tap effectively into large amount of data.

In the context of Industry 4.0, countless publications liken data to the new oil. Yet no consensus exists on the question of how to best tap into this high-value resource. Here, we present a framework based on more than 20 years of experience in the design, development and implementation of digitally operated models aimed improving the performance of government programs in education, workforce and economic development. This framework is illustrated in the figure below (Figure 1). It includes two major components. The first, in center of the diagram, outlines the major activities along the data life cycle labeled by four verbs: (1) Acquire, (2) Store, (3) Process and (4) Execute. This process describes the way data becomes "electricity" to power a digitally operated business or organization. The second part of the diagram, surrounding the core, presents the contextual components that give cultural meaning to the data lifecycle, contributing intangibly to the operation and management of an organization.

In short, this model describes the productive context in which data turn into useble intellectual capital. The next few paragraphs provide a brief description of each part of the two components and how they relate to our experience in helping Mississippi become one of the most digitally operated states in the USA.

The data business cycle

The practice of data science is often conceptualized in terms of the lifecycle of data as it moves from acquisition through its use in training sophisticated AI systems. For convenience, we divide this lifecycle into four parts that, while broadly sequential, overlap significantly in an iterative process.

Data acquisition

Acquisition entails the datafication of an organization and is not only the first step in the data lifecycle but also the first step toward a complete digital transformation. The datafication process represents every aspect of an organization's operations and processes as data objects.
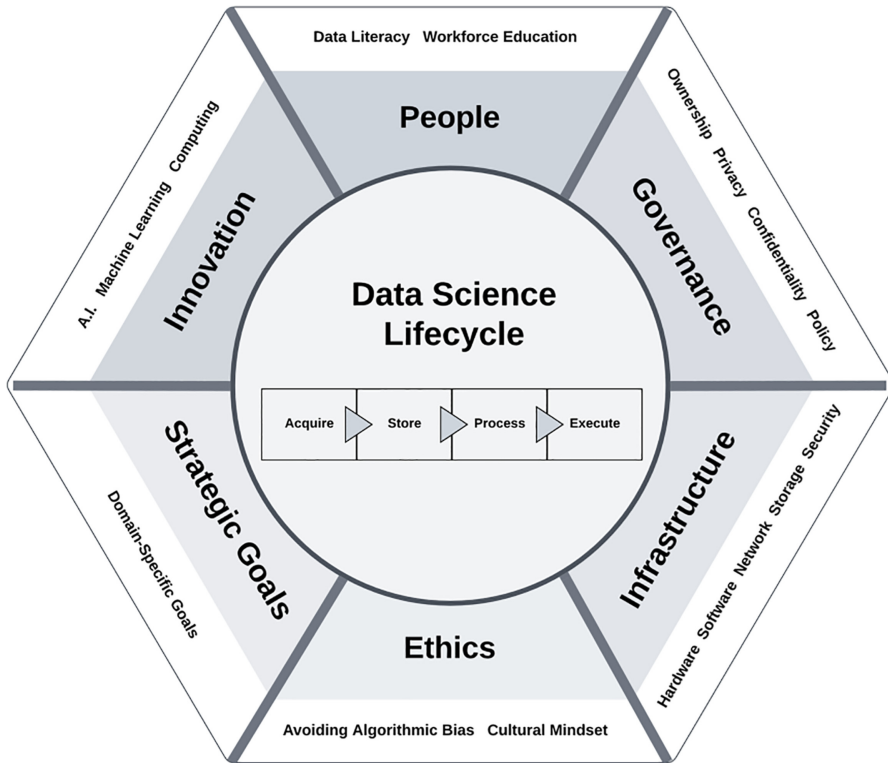
This translates an organization's goods and services into digital, virtual representations. In traditional manufacturing, datafication may result in the creation of digital twins of manufacturing facilities that enable nearly infinite exploration and experimentation with production methods prior to expensive, physical capital investment. Often datafication supports the disruption of an industry. For example, Uber introduced a completely digitally operated model for taxi services. Uber datafies drivers, riders, places, destinations, fare models and vehicles. While avoiding the costs of owning and operating a fleet of cars or employing an army of human drivers and centralized dispatchers worldwide, Uber operates a peer-to-peer system that matches people or groups of people with rides from pickup to drop-off points in adequately sized vehicles driven by car owners incentivized by variable pricing.

The authors used the process of datafication to transform state workforce development services in Mississippi into digitally operated services. Workforce development includes activities designed to train and prepare a region's workforce along with the process of matching qualified workers with employer needs. This traditionally involves an inefficient, organic process of workforce training and recruitment. In Mississippi, datafication of workforce development entailed representing job seekers, businesses, training and job-matching services and job vacancies as data objects. For example, we datafied job seekers in terms of relevant properties such as the timeline of their experience in specific occupations (e.g., welder), their expertise in using specific tools and technologies (e.g., Microsoft Office), their occupational aspirations, their desired salary, desired commuting distance, level of education and specific degrees or credentials. By representing job openings using

complementary properties (occupation of the vacancy, experience required, tools and technologies involved, salary, distance from the job to the applicant's residence) we were able to automate matching. This allowed a participant or a business to seek job matching services online or through mobile applications, reducing the need for traditional physical job center services and freeing workforce agency staff to focus on training and helping hard-to-serve job seekers overcome barriers to employment. Over time, as job seekers created and maintained profiles (resumes) in the new labor exchange system, the state used occupational transition data to train machine-learning based matching algorithms that further improved the quality of customer service.

Data storage
Based on an organization's competitive strategy, the next consideration in the data lifecycle is the development of a clear strategy for secure storage. In the same way that an enterprise strategy encompasses a conceptualization or theory about the likely future in which the organization will compete, the data storage strategy involves a theory about how data will be used. Elements of an organization's storage strategy may include the level of data accessibility (real-time vs. periodic), standards for data quality, statutory or administrative laws governing the secure handling of data, privacy and confidentiality policies and cybersecurity considerations. Based on an organization's strategy, data will be stored either in a local or data center environment, and the data center may be owned and operated by the organization or by a third party (cloud computing). The volume of data, overall size requirements and the projected rate of data growth will also factor into decisions about storage. Development of a storage strategy usually coincides with acquisition because decisions about datafication generally imply the format of storage, strategies for archiving and backing up data, needed size and the business processes that involve data transfers that are integral to the organization's business model. For example, a business relying heavily on sensors to gather data will choose a storage strategy amenable to receiving streaming data from many internet-of-things devices. Finally, even in an organization oriented around a real-time business purpose, the organization may warehouse valuable data for interactive research and analytical purposes. Often these warehouses or data lakes are optimized for efficient long-term storage and not for real-time business operations.

In Mississippi, three ways of using data in the context of workforce development drove decisions about storage. The first was the use of data for case management. This use of data required the collection of structured data in real time as caseworkers provided workforce services to job seekers and employers, as training providers delivered educational and training programs and as job seekers and employers interacted with online and mobile systems. To support this use of data, a relational database tied to several case management systems was used to store the datafied objects (records of services given, employer data, job vacancy data, resumes, etc.). Because these data-intensive systems operated 24/7, the state leveraged two geographically distant data centers, one managed by the state's information technology agency and the other managed by a public university. This combination kept public data in trustworthy, in-state data centers while providing a high-degree of resiliency to power outages, natural disasters and other possible disruptions to service. Replication of relational databases ensured the consistency of data across data centers and made possible rapid recovery in the event of disaster. The second type of data use was the retrospective use of programmatic data to enable federal reporting. This required periodic, time-stamped exports of data from case management systems to ensure archiving and reproducibility of reporting results. Again, the same data centers were leveraged, using secure, distributed file systems to store tabular data in large files. The third type of data use was for research and program improvement. Often, this type of research required aligning records across agency

case management systems while ensuring the privacy of individual citizens. The state developed a strategy for de-identification of records that resulted in data cleaned of personally identifiable information while still maintaining the link between person-level data across datasets. This required two storage environments. The first was a clean-room environment similar to a sensitive compartmented information facility (SCIF) accessible only by trained and qualified staff able to deidentify raw datasets and affix a cross-dataset identifier. The second was a research environment in which the deidentified data were made available to researchers using analytical tools such as Python or SAS. In some cases, data transfer strategies were created that deidentified data while still on the premises of the agency that collected it. In other cases, agencies relied on memoranda of understanding with a university-based research center tasked with stewarding these valuable data and helping to analyze the data with an eye toward programmatic improvement, economic development and strategic planning.

Process

As we have seen, elements of processing data (cleaning and managing data) are closely aligned with storage of data. Based on the data structures established during datafication, and the storage strategy that flows from a business strategy, data pipelines are created to process raw data and populate business data structures from many sources including business applications. Data processing often requires augmenting business data to fulfill the goals of an organization. For example, Uber leverages GPS and map data as it geocodes the location of drivers and riders to find best matches or estimate travel time. Airlines correlate third-party weather data with plane and airport location data to predict delays and route traffic. Online car auction sites leverage VIN (vehicle identification number) databases to augment seller data with manufacturer information on vehicle color, engine type, year and installed options. Government vehicle registration agencies further combine this VIN data with sales data and insurance data to complete the registration (Warkentin and Orgeron, 2020). This pattern of extract-transform-load exists in most organizations as they augment business data with third-party data either in real time through web services or as datasets are loaded into systems that support operations.

In Mississippi, processing of economic, workforce, social services and K-20 education data entailed the deidentification process described above along with augmentation of location data through geocoding, association of programmatic data with geographical entities (e.g., school to county) and integrating federally produced taxonomies for occupations, industries and instructional programs. This data came from all levels of the state public educational system and all other state agencies involved in workforce and economic development. The resulting data warehouse stored data in conformity with data dictionaries that were shared with stakeholders to enable requesting longitudinal reports or analyses. Data dictionaries document all nominal and operational definitions of each data object's fields or properties. In addition, the state entrusted a university research center with stewardship of the data resulting in a mature understanding of state administrative data that informed additional digital transformation efforts.

Execute

In the execution stage of the data lifecycle an organization leverages data in two ways: **analytically** to understand the world and drive human decision making and **operationally** to automate business process management and operations using artificial intelligence. Based on the previous stages of the data lifecycle, the organization is in the position to leverage its valuable data through technology or tools that enable operations to align with the organization's strategy. Additionally, data provide the linkages to translate the

enterprise strategy into mid-level tactics and day-to-day process operations, thus facilitating the effective execution of the strategy through data-driven operationalization.

To support analytical use of data to improve education, workforce and economic development outcomes, Mississippi created the State Longitudinal Data System (SLDS [1], Miss. Code Ann. § 37–154–1, 2013). Through a unique common identifier, state data can be used to examine how individuals fair as they transition from various parts of the K-20 educational system into the workforce. The state developed an online, one-stop portal to provide secure access to the data. It developed the hardware and software capacity for building and hosting the data and it built the appropriate infrastructure and technology for data collection, storage and use. Mississippi also developed and adopted a statewide, comprehensive policy on data quality assurance and it trained state and local personnel on data entry and use to facilitate full adoption and effective use of system. A data inventory and dictionary were developed to identify the necessary data to link pre-school through K-20 and into the workforce.

To support operational use of its data resources, the state created systems that could leverage linked public information in real time to help automate the decision-making processes required to deliver education and workforce services. This system, a case management *hub*, connected agency case management systems to each other in real time, thereby supporting electronic referrals and alignment of all customer data at the point of service while allowing staff and self-service customers to leverage existing agency systems.

Putting data science in context

The tangible steps of acquisition, storage and processing of data that enable the execution of a digitally transformed operating model gain their full meaning through an intangible context in which stakeholders share a cultural mindset around the value of data for promoting the growth and sustainability of the organization (Padua, 2021). Experts "know more than they can tell" (Polanyi, 1966). While the steps in the data science lifecycle are as well-known as the neatly described steps of the Baconian scientific method, as with waltzing, knowing the steps is not the same thing as dancing. Mississippi's public sector digital transformation proceeded at first by instincts about the value of data, graduated into a federally funded project in which these data were formally understood and harmonized in a social context, and finally these data were incorporated into smart systems. In the era of digital transformation, every organization must walk this path from big data to AI. To make explicit the tacit knowledge about how to leverage data as the core intellectual capital of an organization that allows it to innovate and maintain a competitive edge requires a clear understanding of the context in which the data operate. Here, we identify six factors that contribute to create a cultural environment that champions data as a key asset of an organization.

People. Members of the organization who embody the cultural mindset and pair it with technical knowledge required to create, maintain, employ and work alongside artificially intelligent systems. A healthy context for digital transformation includes data science literate individuals who have been trained or re-trained with the skills necessary to work in all parts of the data lifecycle. In the Mississippi story, individuals in state agencies worked with experts in the state's public research universities to assemble intellectual capital required to understand the value of public administrative data and leverage it for analytics and AI.

Governance. A framework for operating based on a clear understanding of the value of data and its social meaning, especially with reference to laws, administrative policies, regulatory rules, privacy policies and other normative constraints on the use of data. In Mississippi, governance for the SLDS consisted of representatives from all data contributors to ensure that each agency's mission and rules were honored as data began to be used for public good. While legislation helped to cement the public legitimacy of the SLDS, partnerships between data contributors were formed through specific memoranda of

understanding that affirmed a unified approach to the goals and methods that would be used within the project to leverage data for public good.

Infrastructure. Data and AI require storage and processing power, thus proper infrastructure is necessary to unlock the value of an organization's intellectual capital. In the case of Mississippi, necessary infrastructure included facilities for secure data handling along with redundant data centers. Increasingly the infrastructure necessary to support AI consists of GPU computing.

Ethics and culture. A cultural mindset within the organization that understands the value of data-driven decision making and supports the use of data and AI for addressing challenges is critical. While innovations such as generative AI (e.g. ChatGPT) have raised the stakes for articulating policies designed to ensure that AI remains salutary for human flourishing, a culture that will succeed in digital transformation must encourage a positive approach to the use of data. The decision *not* to use data to solve human challenges is just as subject to moral scrutiny as decisions to use data. A clearly articulated and shared culture also contributes to making decisions about AI that avoid algorithmic bias in the way machine learning or statistical models are used to automate decision making. As the agency partners collaborated in creating a longitudinal data system and building a case management hub in Mississippi, trust was built positively as data were used to improve life in Mississippi. In addition, trust was built as, year by year, data were used in ways that avoided damaging partnerships.

Strategic goals. Strategic goals encompass an organization's theory of what the future will be like and how it will bring value in the midst of that situation. Each organization's strategy will be unique as it is grounded on the organization's unique external and internal drivers, including its goals, strengths, weaknesses, opportunities, threats and overall capabilities (such as its intellectual capital). A properly designed and implemented strategy can drive the way internal stakeholders place the data lifecycle in the context of the other five contextual factors. An organization builds a team (people), acquires infrastructure, recognizes or establishes governance, clarifies its ethical/cultural values and pursues innovations that it believes will support its strategy. In the case of Mississippi, a driving theory of the future was that, as a smaller state, unless Mississippi could begin to tell its own story about labor availability using data, we would be passed over by site location consultants in favor of areas with so many potential workers that the risk of not reaching hiring goals would be less likely, even if Mississippi was a better climate for success for that particular business. Another theory was that in order to produce an effective workforce, we would need to use data to understand whether and how training connected individuals to the labor market. These goals allowed the state to focus its efforts in a unified and coherent way to achieve a high level of digital transformation.

Innovation. The ability to create or adopt new advances in AI, machine learning and computing. Innovations in computing, AI, networking, sensor design, materials and many other areas make digital transformation of organizations possible. An organization's ability to recognize promising innovations and leverage their value is key to turning data into an asset. For example, ChatGPT is trained on a large corpus of text that likely does not include an organization's internal intellectual property. The ability to leverage the innovation of generative AI outside of the general knowledge/search context where the technology has begun its life will depend upon an organization's learning to expose its own material to a large language model capable of answering questions in terms of the organization's strategy.

By addressing each of the six contextual elements of data science, an organization creates an environment in which every team member understands the value of data-driven smart systems, possesses a clear understanding of how high-quality data are required to build these AI systems, and has access to the infrastructure necessary to leverage data. An organization in which this mindset prevails can pair a business strategy with an appropriate strategy for the entire lifecycle of data: acquisition, processing and storage and incorporation of insights from data into the execution of increasingly smarter systems.

Ensuring successful digital transformation through education

Given the importance of human capital as one of the six contextual elements of data science, the authors realized that a response to the growing need for data literacy required changes to the education system. The workforce demand for data science skills is expected to be ubiquitous in the next ten years in every field. To address this need, with support from the Office of the Provost, the authors led an effort at Mississippi State University (MSU) that culminated in 2022 with the launch of an intercollege Bachelor of Science in Data Science.

The MSU data science program [2] eschews the conception of data science as mere data analytics, recognizing that data science is a field that supports and expands the conduct of science within every field. Specifically, the program defines data science as the field that advances methods to improve the use of data for human progress. Specifically, these methods allow humans to (1) Represent the world with virtual data objects through a process of datafication; (2) Extract insights and facilitate new discoveries about the world by studying these data objects; (3) Create smart systems to perform tasks that have ordinarily required human intelligence; and (4) Increase the performance (measured in terms of scale, scope and speed) of organizations as they produce or deliver virtual and tangible goods and services. This definition of data science places the data lifecycle within a contextual framework described above that emphasizes the role of scientific innovation (A.I. and Computing), people (workforce education and data science literacy), governance (ownership, privacy and confidentiality and policy), infrastructure (hardware, software, network, storage and security), ethics (the avoidance of algorithmic bias and a cultural mindset to use data to promote human flourishing) and the strategic goals that guide organizations in specific domains.

Students enrolled in the program receive, in addition to grounding in the fields that contribute methods to data science, three key advantages that address the contextual framework described above. First, students receive a traditional liberal arts education emphasizing strengths in communication, analytical thinking and imagination. This gives students the opportunity to reflect on how their learning impacts their lives and the lives of fellow humans. Second, opportunities for experiential learning allow students to deepen their understanding of diversity and teamwork while gaining technical fluencies that enable success in an increasingly digitally connected and operated world. Third, a personalized success plan allows us to work with students in a coherent way from the time they arrive on campus to the moment they leave the university and enter the workforce. In this way the student's experience becomes the organizing principle of all support, advising and academic counseling activities. These three ingredients—an emphasis on liberal arts, experiential and shared learning and a personalized success plan—provide students with the confidence and tools they need to reach their educational and career goals and contribute to organizations in which they will carry a mindset capable of understanding data as the core intellectual capital of an organization that allows it to innovate and maintain a competitive edge.

Conclusion

In digitally operated business models, data constitute the core value of the operational and intellectual capital of an organization. The ability to use and capitalize on large amounts of data to transform operations in line with strategy gives an organization a competitive edge. We have presented a framework, developed in Mississippi through digital transformation in the public sector, for guiding an organization's digitalization and measuring an organization's maturity along the path to total digital transformation. Organizations that use this model to assess their current level of maturity and direct their efforts toward developing a cultural mindset around the value of data for promoting the growth and sustainability of the organization will be in the position to succeed in the context of Industry 4.0. The framework produces a self-reinforcing, virtuous circle as the mindset motivates an attention to data quality and the use of data to create

smart systems that, themselves, generate more data relevant to the organization's successful execution of a sound business strategy. The framework also serves as a useful guide in designing educational programs designed to prepare a workforce with the desired cultural mindset around the value of data. Finally, the framework offers an opportunity to guide future research on assessing the ability of an organization to turn data into intellectual capital in the context of Industry 4.0.

Domenico Parisi, Jonathan Barlow and Merrill Warkentin

Notes

1. https://law.justia.com/codes/mississippi/2013/title-37/chapter-154/section-37-154-1

2. https://dsci.msstate.edu/ (for more information, visit https://www.datascience.msstate.edu/)

References

Del Giudice, M., Scuotto, V., Papa, A., Tarba, S.Y., Bresciani, S. and Warkentin, M. (2021), "A self-tuning model for smart manufacturing SMEs: effects on digital innovation", *Journal of Product Innovation Management*, Vol. 38 No. 1, pp. 68-89.

Gashenko, I.V., Khakhonova, N.N., Orobinskaya, I.V. and Zima, Y.S. (2020), "Competition between human and artificial intellectual capital in production and distribution in industry 4.0", *Journal of Intellectual Capital*, Vol. 21 No. 4, pp. 531-547.

Gravili, G., Manta, F., Cristofaro, C.L., Reina, R. and Toma, P. (2021), "Value that matters: intellectual capital and big data to assess performance in healthcare. An empirical analysis on the European context", *Journal of Intellectual Capital*, Vol. 22 No. 2, pp. 260-289.

Hatzius, J., Briggs, J., Kodnani, D. and Pierdomenico, G. (2023), "The potentially large effects of artificial intelligence on economic growth", *Goldman Sachs Economic Research*, available at: https://www.scribd.com/document/635924425/The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth

Iansiti, M. and Lakhani, K.R. (2020), "Competing in the age of AI", *Harvard Business Review*, Vol. 98 No. 1, pp. 60-67.

Padua, D. (2021), *Digital Cultural Transformation: Building Strategic Mindsets via Digital Sociology*, Springer Cham.

Polanyi, M. (1966), *The Tacit Dimension*, University of Chicago Press, Chicago, IL.

Warkentin, M. and Orgeron, C. (2020), "Using the security triad to assess blockchain technology in public sector applications", *International Journal of Information Management*, Vol. 52, 102090.

About the authors

Domenico Parisi is the Executive Director of the Data Science programs at Mississippi State University and a Professor of Sociology. Dr. Parisi has secured more than $150M to promote Data Science initiatives at MSU and beyond. He is nationally and internationally recognized for his work on creating digitally operated smart systems designed to improve the performance of government programs in education, workforce and economic development.

Jonathan Barlow is Associate Director of the Data Science Program and an Assistant Teaching Professor at Mississippi State University. Previously, Barlow was an associate director at NSPARC, a research center at MSU. Before joining Mississippi State in 2012, he worked for 15 years in the private sector. With a background in industry and university research, Barlow has more than 25 years of experience in software development, data modeling, data-intensive applications and data analysis. Barlow received his Ph.D. from Saint Louis University. His data science research interests involve natural language processing and the ethics of artificial intelligence.

Merrill Warkentin is a William L. Giles Distinguished Professor at Mississippi State University, where he serves as the James J. Rouse Endowed Professor of Information Systems in the College of Business. He was named an ACM Distinguished Scientist and has been identified (in a study by Stanford University) as one of the top 2% of global scientists for his academic impact. Dr. Warkentin is also among the fifty most published global scholars in IS over the last decade as ranked by the Association for Information Systems. He is the Editor-in-Chief of the *Journal of Intellectual Capital*.

609