What is Data Science?

Mimmo Parisi, Jonathan Barlow, and Merrill Warkentin

Mississippi State University takes a unique approach to data science as the field that advances methods to improve the use of data for human progress. Specifically, these methods allow humans to:

- Represent the world with virtual data objects through a process of datafication;
- Extract insights and facilitate new discoveries about the world by studying these data objects;
- Create artificially intelligent systems that align harmoniously with human intelligence to perform tasks; and
- Increase the performance (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 (see diagram below) 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.

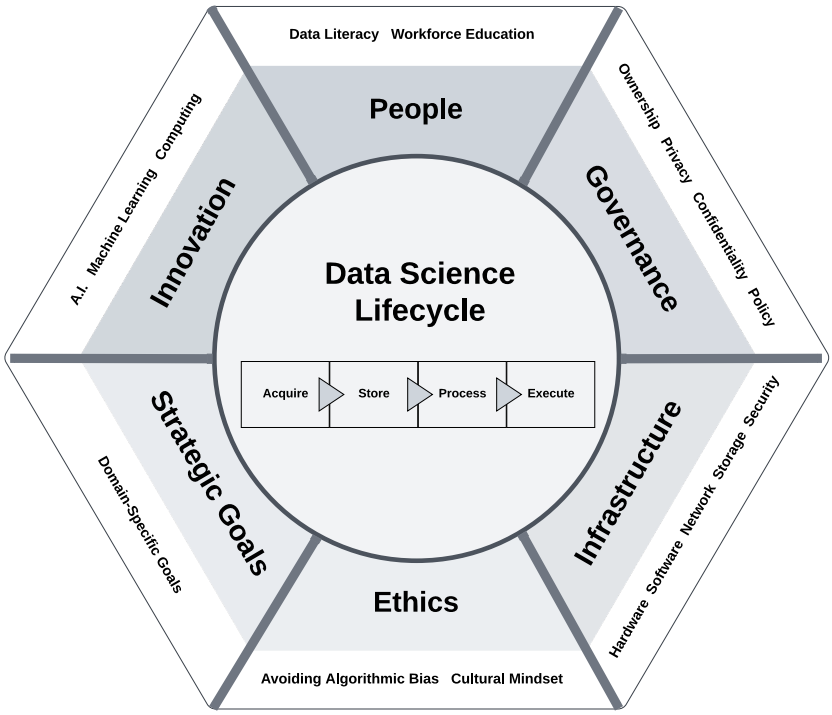

Figure 1: A Contextual Framework for the Data Lifecycle

The Data Lifecycle

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. 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.

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 a business strategy is a theory about the likely future in which the organization will compete, and effective data storage strategy requires 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.

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. 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.

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

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 and management in line with the organization's strategy.

Putting Data Science in Context

The tangible steps of acquisition, storage, and processing of data that enables the execution of a digitally transformed, artificially intelligent operating model gains its 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. Digital transformation proceeds at first by instincts about the value of data, graduates into projects in which these data are formally understood and harmonized in a social context, and finally data are 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.

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.

Infrastructure. Data and AI require storage and processing power, thus proper infrastructure is necessary to unlock the value of an organization's intellectual capital. Increasingly, the infrastructure necessary to support AI consists of GPU computing and sufficient storage for large datasets.

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. 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. With regard to AI, data science also considers ways to solve the so-called "alignment problem," ensuring that any artificially intelligent system's goals and definitions for success align with human goals and values.

Strategic Goals. An organization's theory of what the future will be like and how it will bring value amid that situation. While strategy differs from organization to organization, the presence of a strategy drives the way people within the organization 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.

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 is 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. With this context in mind, Mississippi State University is committed to preparing its data science students to participate in and lead the change required for positive digital transformation.

June 6, 2023

Domenico Parisi
Professor and Director, Data Science Programs at Mississippi State University

Jonathan Barlow
Assistant Teaching Professor and Associate Director, Data Science Programs at Mississippi State University

Merrill Warkentin
James J. Rouse Endowed Professor of Information Systems, Mississippi State University

Credits: Portions of this article appeared in Parisi, D., Barlow, J., & Warkentin, M. (2023). Editorial: Where the data meets the road in the Industry 4.0 economy. *Journal of Intellectual Capital*, 24(3), 601-609. The "Contextual Framework for the Data Science Lifecycle" diagram was originally developed by Mimmo Parisi and Jonathan Barlow, 2021.