Data Science as an Interdiscipline: Historical Parallels from Information Science

MATTHEW S. MAYERNIK [iD]

ABSTRACT

Considerable debate exists today on almost every facet of what *data science* entails. Almost all commentators agree, however, that data science must be characterized as having an interdisciplinary or metadisciplinary nature. There is interest from many stakeholders in formalizing the emerging discipline of data science by defining boundaries and core concepts for the field. This paper presents a comparison between the data science of today and the development and evolution of information science over the past century.

Data science and information science present a number of similarities: diverse participants and institutions, contested disciplinary boundaries, and diffuse core concepts. This comparison is used to discuss three questions about data science going forward: (1) What will be the focal points around which data science and its stakeholders coalesce? (2) Can data science stakeholders use the lack of disciplinary clarity as a strength? (3) Can data science feed into an "empowering profession"? The historical comparison to information science suggests that the boundaries of data science will be a source of contestation and debate for the foreseeable future. Stakeholders face many questions as data science evolves with the inevitable societal and technological changes of the next few decades.

CORRESPONDING AUTHOR:

Matthew S. Mayernik

NCAR Library, National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), P.O. Box 3000, Boulder, CO 80307, US

mayernik@ucar.edu

INTRODUCTION

Data science has been called the "sexiest job of the 21st century" (Davenport & Patil 2012). Data science has also been extensively critiqued by scholars across numerous fields. One particularly vivid critique labels data science as "machinic Neoplatonism," stating that data science techniques encourage and enable thoughtlessness in the context of decision-making and societal analysis (McQuillan 2018a). Other commentary on the nature of data science is similarly divergent. Data science has been characterized as being little more than statistics relabeled (*Statistics Views* 2013) while also being characterized as encompassing almost every kind of science (Fowler 2015).

Within this bubbling commentary, considerable debate can be found on almost every facet of data science. The one thing that most commentators agree on, however, is that data science must be characterized as having an interdisciplinary and/or metadisciplinary nature. To be doing data science, according to almost every description, one must be pulling tools, skills, algorithms, concepts, or data from multiple disciplinary or methodological frameworks. As one of the above quoted scholars phrased it, "The imaginary ideal data scientist is a Renaissance figure with a mastery of all these arts," referring to programming, statistics, mathematics, and data visualization, among other skills (McQuillan 2018a: 255). Looking across various academic and popular press descriptions of data science, significant differences can be found in the characterizations of the appropriate mélange of skills and tools that constitute a "data scientist." But the fundamental interdisciplinary nature of data science—the fact that people who do data science cross or transcend traditional disciplinary boundaries, tools, and methods—seems to be a consensus view.

This paper aims to build a lens for understanding the diversity, complexity, and interdisciplinarity or metadisciplinarity of data science by drawing lessons from the history of information science and its precursors. This analysis highlights the historical parallels between the emergence of data science in the 21st century and the emergence and evolution of information science over the past 100 years to provide insight into interdisciplinary challenges facing data science as a professional and academic endeavor.

This comparison is particularly timely. Debates about the disciplinary status of data science are growing within government, corporate, and higher education institutions. A number of recent consensus reports have been written to help shape the present and future of data science (Berman et al. 2016; National Academies of Sciences, Engineering, and Medicine [NASEM] 2017; NASEM 2018).

Disciplines are social, organizational, and institutional constructs that often emerge around nascent problems or topics where resources like funding and students are in a growth period (Jacobs 2013; Hammarfelt 2019). Such is certainly the case for data science.

The term *interdisciplinarity* can be a linguistic stand-in for modern, creative, and/or progressive ways of working (Madsen 2017). As such, interdisciplinarity can be a way of working and/or a way of talking. This tension is one manifestation of how interdisciplinarity can encompass many different things (Huutoniemi et al. 2010). Because of this variation in meanings, "good interdisciplinary work requires a strong degree of epistemological reflexivity" (Klein 1996: 214). Epistemological reflexivity may have value in moving data science toward becoming a "critical technical practice" (Agre 1997), that is, an area of work that actively examines and engages with its own limitations and inherent challenges. People working within and around information science have repeatedly debated the relative merits and drawbacks of interdisciplinarity and metadisciplinarity throughout the past century and up to the present (Borko 1968; Bates 1999; Buckland 2012; Madsen 2016).

This paper begins by characterizing data science as an inter- and metadiscipline by highlighting a number of key features of recent research and professional work in the area. I then depict similar characteristics of information science and finish with a discussion of the following set of questions related to the interdisciplinary pros and cons of current data science:

1. What will be the focal points around which "data science" and its stakeholders coalesce?

2. Can data science stakeholders use the lack of disciplinary clarity as a strength?

3. Can data science feed into an "empowering profession," namely, a profession that promotes the growth, competence, and autonomy of the people that it serves?

METHODOLOGICAL APPROACH

This paper is based on a review of the literature in the information and data sciences related to interdisciplinarity. Many of the sources used in the characterizations below are personal narratives that present the perspective of a single individual. In the case of data science, some are white papers, blog posts, or opinion papers. In the case of information science, many relevant sources are papers published in peer-reviewed journals by prominent people in the field, including scholars, educators, and administrators. Any single perspective among these voices may have particular limitations or biases. Taken together, however, such sources prove extremely valuable for tracking the evolution of interdisciplinary research areas (Klein 1996). Personal narratives serve as indicators of the ways that particular issues were discussed at different points in time. These personal narratives have been compared and contrasted with relevant research papers appearing in peer-reviewed journals discussing the nature of information science and data science as disciplines and professions.

The method for gathering relevant peer-reviewed materials for this paper included systematic queries of article databases such as the Web of Science and Google Scholar for articles related to "information science," "data science," "interdisciplin*," and "metadisciplin*." These sources are useful to find relevant materials related to information science given the long history of the topic area, but they are less useful for tracing longer-term developments relevant to data science given its nascent development as a named entity. For example, as of February 10, 2023, the Web of Science Core Collection returns 14,448 total results when searching for the phrase "data science," of which only 26 date from 2009 or earlier. Of these, 19 are spurious hits, and 5 are book reviews or news items. Only 2 peer-reviewed articles discuss data science in a way that is close to current understandings: Cleveland (2001), discussed below, is a foundational article for the statistical aspects of data science. Mezey et al. (2001: 375) discuss a number of aspects of database analysis that "provide challenging tasks and opportunities for data science" but otherwise do not directly discuss data science itself. As a point of comparison, the Google NGram viewer, which quantifies usage of particular words or phrases across the Google Books corpus, shows almost zero use of the term "data science" through 2008, but there is significant year-over-year growth in use of the phrase since 2009 (https://books.google.com/ngrams/).

Notably, as of February 2023, the Web of Science does not index any journals that include the term "data science" in the title. Thus, an additional method for finding relevant articles was to directly investigate journals that focus on data science but are not indexed by the Web of Science, either by using Google Scholar or by visiting the journals' websites and examining issues. Some specific journals that were investigated in this fashion are described further in the section that follows. Another method for finding materials relevant to this article's discussion, perhaps the most valuable, was citation chaining. Once a relevant article was found, following citation networks both forward and backward in time frequently resulted in the discovery of more relevant articles.

DATA SCIENCE

At the time of this writing, historical literature related to data science is scant. The best chronological depictions of the development of data science are found in Cao (2017) and Press (2013). Both articles illustrate how the trajectory of what we now call *data science* can be dated back at least 50 years, encompassing developments in data analytics and visualization, statistics, database design, and other topics. Phillips (2019) points to other trends related to data gathering and analytics that extend back a century or more. The phrase *data science*, however, has only been in use for about 20 years. In this section, I highlight particular developments in the past two decades related to the emergence of data science, focusing on characterizations of data science's boundaries and participants.

CONCEPTIONS OF DATA SCIENCE

The recent growth of data science has been stimulated by the large volumes and varieties of data being made public on the internet via the explosive growth in digital technologies, such as personal computers, cell phones, social media, smart devices, and sensor networks. As the generation of data by these technologies has increased, the need for methods of storing,

accessing, analyzing, and presenting data has also increased. *Data science* has emerged as a panoply of techniques, tools, and skills that can be applied to derive value (economical or intellectual) out of the growing piles of data. The concomitant need for people with skills to work in these areas within the commercial and public sectors has also been a significant driver of the growth of data science (Manyika et al. 2011).

From different points of view, data science can be viewed as (1) a proto-discipline, (2) a toolkit of analytical pipelines and platforms, (3) a bundle of transformative forces at work inside and outside the academy (Carson et al. 2016), or even (4) "a community of practice of data-driven scientists of whatever scientific discipline they ask questions about" (Scheider et al. 2020: 8). Statistician David Donoho's (2017) recent paper, "50 Years of Data Science," provides a useful starting point for this discussion about the scope of data science. This paper has been highly cited since its initial publication, and Donoho is a prominent figure in many discussions of data science. Donoho describes his view of six divisions of "data science" activity:

1. Data gathering, preparation, and exploration

2. Data representation and transformation

3. Computing with data

4. Data modeling

5. Data visualization and presentation

6. Science about data science

Donoho explicitly excludes this engineering component from his typology, namely, the activities involved in building systems to effectively deal with data, move data, and distribute data at different scales. Other commentators, however, call out infrastructure development as a core component of data science and computational work in general (Blanchette 2012; Fox & Hendler 2014; Gray, Gerlitz and Bounegru 2018).

As Donoho acknowledges, his typology builds on earlier works by the prominent statisticians Chambers (1993) and Cleveland (2001), who stimulated the idea of a "data science" by urging the field of statistics to broaden its focus beyond its traditional emphasis on theoretical analyses. Cleveland's paper, for example, lays out a precursor to Donoho's data science typology and depicts how statistics curricula could be expanded to better train people as "data scientists." The *Journal of Data Science*, launched in 2003 by two statisticians, provides a publication venue for statistically focused data science research very much in line with Cleveland's call, focusing on the applications of statistical methods in a variety of contexts. Many prominent articles about data science, however, including Donoho's, have been published in core statistics journals.

At around the same time as Cleveland's call for data science within the field of statistics, data science was also becoming a named entity in other sectors. CODATA, the Committee on Data of the International Council for Science (ICSU), published the first issue of its *Data Science Journal* in 2002.[1] CODATA was established in 1966 "to promote throughout the world the evaluation, compilation and dissemination of data for science and technology and to foster international collaboration in this field" (CODATA 2012). The six founders of CODATA included chemists, physicists, and an engineer. The *Data Science Journal* was formed to facilitate the dissemination of scholarly work on topics related to the committee. The launch of the journal was also specifically motivated by disciplinary aspirations. As stated in a retrospective on the 45th anniversary of CODATA, "*A journal gives identity to a discipline*" (Lide & Wood 2012: 55, italics in original). The first editor of the journal, F. Jack Smith, outlined his view of the key topics of interest within the new discipline and journal:

> ... the study of the capture of data, their analysis, metadata, fast retrieval, archiving, exchange, mining to find unexpected knowledge and data relationships, visualization in two and three dimensions including movement, and management. Also included are intellectual property rights and other legal issues. (Smith 2006: 163)

As of 2023, the scope of the *Data Science Journal* had not varied significantly from Smith's initial focus (Rumble 2023; Smith 2023).

---

[1]   Since 2015, I have been a member of the *Data Science Journal*'s editorial board, and I became joint editor-in-chief of the *Data Science Journal* in September 2021.

The *Data Science Journal*'s emphases were largely disjointed with Donoho's typology of data science and the goals of the aforementioned *Journal of Data Science*. Some of the *Data Science Journal*'s areas of emphasis fall into the engineering category that Donoho acknowledges but does not include in his typology, but some others are much further afield, such as the *Data Science Journal*'s mention of legal issues related to data.

Looking at the *Journal of Data Science* and the *Data Science Journal* in parallel, we clearly see two distinct notions of what data science encompasses, both generally exclusive of the other. Numerous other journals and conferences related to data science expand the boundaries of the topic even further, including titles launched since 2019, such as the *Harvard Data Science Review*, *Data Intelligence*, and *Patterns* (Meng 2019; Hendler, Ding & Mons 2019; Callaghan & Darbyshire 2020).

Data science is often depicted as a nexus of certain kinds of skills. Drew Conway's (2010) data science Venn diagram is a commonly referenced visualization for this view, in which data science is depicted as the amalgamation between (1) math and statistics knowledge, (2) "hacking skills," and (3) "substantive knowledge," referring to knowledge within a particular disciplinary specialization. Conway is careful to note that this Venn diagram is intended to apply to data science broadly, not necessarily any specific data scientist (Conway 2018). But others, such as Davenport and Patil (2012), take this view further by stating that having computer science and statistical expertise are the defining features in distinguishing a data scientist from a traditional scientist. Blei and Smyth (2017) also note this distinction between data scientists and "domain scientists," but they emphasize that the two groups should be partners (or integrated) whenever possible:

> Crucially, the data scientist solves the problem iteratively and collaboratively with the domain expert. (We note they do not need to be two different people; the data scientist and domain expert could simply be two "hats" for the same person). (Blei & Smyth 2017: 8691)

This conceptual separation between regular (or domain) science from data science is in fact necessary for the data scientist to exist as a distinct type of person (Ribes 2018). There would be no need to create a new label like *data scientist* if there was no conceptual or practical distinction between what a data scientist does and what a typical researcher would be doing within chemistry, astronomy, or meteorology. While the tools and methods used are one notable distinction, another could be that data scientists are expected to be able to apply their skills to data regardless of the disciplinary focus of those data. In other words, in the characterization of the above authors, data scientists are expected to be able to work with data for which they have no specific disciplinary training (Feder 2016), while domain scientists are only expected to be able to work with data from within their own discipline, such as chemistry, astronomy, or meteorology.

KEY CHARACTERISTICS OF DATA SCIENCE AS AN INTER- AND METADISCIPLINE

In looking at recent discussions of the trajectory of data science, three key issues related to interdisciplinarity repeatedly manifest: (1) the diversity in participants and communities, (2) the diffuse and contested boundaries of data science, and (3) the debated disciplinary status of data science. This section expands on these points.

The diversity in participants and communities

It is clear that data science, however bounded, is a topic area that encompasses many participants and communities and involves people with a multiplicity of skills and backgrounds. The statistics-centric view emphasizes the need for data scientists to be knowledgeable about data representation, transformation, modeling, and visualization. The data management and engineering conception of data science spans computational infrastructure building, metadata development, data retrieval and archiving, and intellectual property regimes for software and data products. Some discussions of data science include components of both views (Fox & Hendler 2014; Blei & Smyth 2017; Stanton et al. 2012), but this is less common. It is also clear that there is an evolving spectrum of skills and expertise that data scientists hold in practice (Fayyad & Hamutcu 2022). People with data science job titles or responsibilities work in nearly every

societal sector, including government, industry, nonprofit organizations, and higher education (Cao 2017; Carter & Sholler 2016). Many people who could be characterized as data scientists, however, do not fully identify as such, as noted by a recent survey of data scientists in academia, in which many respondents "somewhat" identified as a data scientist (Geiger et al. 2018).

The diffuse and contested boundaries of data science

With this diversity of people involved, few individuals follow the same path into the field. A former editor for the *Data Science Journal* hoped that the journal could serve as "a saloon for data scientists and experts in other fields" (Iwata 2008). As such, the boundaries between data science and other fields are porous. Commentators have drawn parallels between data science and numerous other disciplines, ranging from statistics and information science to computer science (Mattmann 2013) and journalism (Keegan 2016).

The diffuse and contested boundaries of data science manifest clearly as departments and schools jockey for position to own data science within academic institutions. Educational programs for data science are blooming, albeit in highly heterogeneous ways, which makes identifying any broad trends in curriculum development problematic (Wing et al. 2018). The US National Academies of Science report on data science undergraduate curricula provides little closure around what should or should not be part of data science education (NASEM 2017: 33). The summary lists nine central conceptual areas within the scope of data science and asks, "Which key components should be included in data science curriculum, both now and in the future? How could these components be prioritized or best conveyed for differing types of data science programs?" The report does not attempt to answer these questions directly.

De Veaux et al. (2017), on the other hand, define an undergraduate curriculum for data science in great detail, encompassing mathematical and statistical components, as well as data modeling, description, and curation. Their proposed curricula also includes a significant emphasis on communication, reproducibility, and data ethics. The EDISON Framework likewise breaks data science curricula into a number of competency areas, specifically (1) data analytics, (2) data engineering, (3) data management, (4) research methods and project management, and (5) domain-related competencies (Demchenko et al. 2016; EDISON Project 2017). Numerous other curricula can be found, both undergraduate and graduate, each covering various topic areas (Finzer 2013; Tang & Sae-Lim 2016; Ortiz-Repiso, Greenberg & Calzada-Prado 2018; Zhang et al. forthcoming).

An ongoing question for many universities is where to situate data science programs within the ecosystem of existing schools and departments (Berman et al. 2016; Wing & Banks 2019). Many of the curriculum topics listed in the previous paragraph are already being taught within statistics, engineering, computer science, and information science programs. Data science students and instructors alike have diverse backgrounds, and it is common for instructors to be active practitioners, not tenured faculty (Kross & Guo 2019). One model is to create data science institutes as distinct entities while drawing faculty from multiple existing campus departments (Moore-Sloan Data Science Environments 2018). These institutes provide forums for building coalitions of faculty, student interest, and financial investments and provide testing grounds for broader data science undertakings across a campus (Carson et al. 2016).

Significant diversity exists, however, in how data science has been instituted within university structures. An intensive study conducted by the University of California, Berkeley, assessed 16 different options for providing organizational support for data science, including forming new schools or colleges, creating new divisions within existing schools, creating programs that are spread across multiple schools, and creating new research units or centers (Carson et al. 2016). Many universities, for-profit companies, and nonprofit organizations have also started online data science courses and certification programs (Fox et al. 2015; Pournaras 2017; Cao 2018b; Bezuidenhout et al. 2021). These online programs have been able to reach much larger numbers of students, including populations beyond the typical undergraduate student (Kross et al. 2019). These programs are responding to the need to scale up the number of graduates to meet employment demands in the private and public sectors (Miller & Hughes 2017).

The debated disciplinary status of data science

All these factors contribute to the contested disciplinary status of data science (Stodden 2020). This debate is rooted in the variable understandings of what the central concept, data, actually

means. Defining *data* is itself an area of active scholarly research, though mostly by philosophers and information scientists (Floridi 2005; Borgman 2015; Leonelli 2015; Furner 2017; Hjørland 2018). Many discussions of data science that are otherwise very comprehensive, such as Donoho (2017), Cao (2017), and the EDISON Project (2017), do not engage in the fundamental question of defining the core concept of the emerging field. Nonetheless, numerous definitions of *data* can be found, ranging from disciplinary or technology-centric perspectives to abstract conceptualizations (Furner 2016). The ubiquity of the concept of *data*, combined with its elusiveness, frame the ongoing debates about the formalization of data science as a discipline.

Here it is important to note the distinction between (a) a formally defined discipline and (b) sets of people who are interested in, working on, or conducting research related to a particular topic or phenomena. The latter kinds of groups, which might be characterized from different points of view as "invisible colleges," "epistemic communities," or "communities of practice" (De Solla Price & Beaver 1966; Knorr-Cetina 1991; Wenger 1998), encompass groups of people who are connected via social and/or intellectual networks but may have different formal disciplinary affiliations. This distinction is important in relation to the question about whether the goals of data science should be to develop a "science with data" or a "science of data" (Fox & Hendler 2014). In the next section, I return to the question of the degree to which data science as a discipline will encompass broader areas of research that focus on data as a phenomenon of interest.

DISCUSSION

This section presents a discussion of the literature review and works through the three central questions of the paper in detail, focusing on the comparison between data science and information science. Table 1 presents high-level parallels between data science today and current and past information science. As shown in the table, the notion that there is an explosion of information and data that is outpacing our ability to manage, use, and understand them is not new to the "big data" or "data science" era. Rhetoric of "information overload" has been used to motivate new developments in information and data management techniques at least as far back as the early 20th century (Day 2009).

Table 1 Comparison between data science and information science.

POINTS OF COMPARISON	INFORMATION SCIENCE PRECURSORS, 1920S THROUGH 1950S	INFORMATION SCIENCE, 1960S THROUGH 1990S	DATA SCIENCE AND INFORMATION SCIENCE, 1990S THROUGH 2010S
Explosion of information/data resources	• Growth of US government in 1930s • Technical reports classified during World War II made public afterward • Seized documents from Axis power countries	• Cold War–driven expansion of research and research outputs • Digital resources distributed through electronic media (magnetic and optical disc formats) • Emergence of the web	• Digital technologies, such as personal computers, cell phones, social media, and sensor networks, that enable faster generation of information and data • Large volumes and varieties of data being made public on the internet
New technologies that promise to improve capacity	• Microfilm and microfiche • Punch card–based document sorting and selection tools • Early computing technologies	• Digital computing technologies • Personal computers • Internet and web technologies	• High-bandwidth cellular networks • Artificial intelligence and machine learning tools • Advanced data mining • Cloud computing • App development on social media platforms
Diversity of participants	• Engineers • Librarians • Mathematicians • Scientists	• Computer scientists • Economists • Engineers • Information scientists • Librarians and archivists • Psychologists • Scientists	• Computer scientists and engineers • Information scientists • Librarians and archivists • Philosophers • Scientists • Social scientists • Statisticians

Information science became a distinct disciplinary and professional label in the 1960s. The prehistory of information science, however, centers on international efforts in the first half of the 20th century that focused on "documentation," the initial predominant name for the topic (Farkas-Conn 1990). After World War II ended, the interest and activity related to information and documentation increased dramatically. The governments of many countries, particularly the United States, wanted to leverage the research conducted during the war to facilitate growth of public knowledge (Bush 1945). This resulted in an explosion of technical reports into

the public domain after the war. In addition, the victorious Allied forces seized a huge number of government documents from Nazi Germany and other Axis countries (Richards 1994). The challenge of organizing these documents stimulated interest and activity in documentation, information organization, and information retrieval. Information and intelligence work related to the growing Cold War with the Soviet Union likewise stimulated growth in information research and professionalization (Johnson 2017; Burke 2018). Many organizations undertook information-related work during this time, and the number of information workers grew rapidly, including many scientists who encountered information work during the war (Burke 2007).

The information science educational ecosystem expanded through the 1970s, often (though not exclusively) through programs based in library schools. The library and information sciences coalesced enough during this period for a number of specializations to become prominent, if somewhat disconnected (White & McCain 1998). The following decades saw a serious retrenchment of the educational landscape, as over 20 library and information science schools closed or went through administrative realignments in the 1980s and 1990s (Ostler, Dahlin & Willardson 1995; Hildreth & Koenig 2002). This retrenchment slowed in the 2000s as the internet emerged as a social and technological phenomenon, causing renewed interest in information within governments, universities, and the private sector. In 2005, in the midst of the dot-com period, the "iSchool" caucus was formed by a group of nine library and information science schools (Larsen 2008). As of mid-2020, it contained 114 members across six tiers of membership. The iSchool membership is diverse, intellectually and programmatically. Some schools retained strong connections to the earlier information science focus areas, but many differ substantially from what an information science school looked like in prior decades (Wiggins & Sawyer 2012; Craig Finlay, Ni & Sugimoto 2018).

Throughout the past century, information science has demonstrated the same characteristics analyzed above for data science: (1) diversity in participants and communities, (2) diffuse and contested boundaries, and (3) debated disciplinary status. As shown in Table 1, as long as information science and its precursors have existed, there has been a diverse and clumpish mix of participants. This diversity of participants and intellectual approaches has provided a constant source of new ideas and contributors within information science, but it has inevitably engendered boundary arguments about what the discipline should (or should not) include. During the 1960s, information science emerged as a contested space, and it has continued to face boundary disputes to this day (Kline 2004; Burnett & Bonnici 2013).

Articulating and negotiating the unique value and niche of information science within the ecosystem of constitutive and related disciplines has been an ongoing challenge (Van House & Sutton 1996; Cox et al. 2012; Bates 2015). The diverse and evolving sets of participants and ongoing boundary challenges have repeatedly engendered debates about the disciplinary status of information science. Many commentators within these debates have noted that interdisciplines face continuous struggles to achieve power and legitimacy inside academic and government institutions that favor (implicitly or explicitly) traditional disciplines. Some question the wisdom of arguing for the field explicitly by championing its interdisciplinary nature (Buckland 2012; King 2006). In part, ongoing challenges in articulating the common thread(s) within information science stem from the elusiveness of *information* as a topic. Definitions and characterizations of the concept of *information* abound by individuals inside and outside of information science (Cornelius 2002; Capurro & Hjørland 2003).

What can be drawn from the parallels between the ongoing evolution of information science and the emergence of data science? The disciplinary ecosystems between the two fields are not identical. Marchionini (2023) argues that information science can be considered an academic discipline because it has developed distinct "principles, key research questions, and communities of practice that have given rise to subspecialties, professional standards, curricula and degrees; whereas data science at present consists of a set of techniques that have arisen out of allied fields such as statistics, computer science, and information science and is driven by applications and problems from a variety of endeavors of modern life."

To build insight from this comparison, the following sections discuss three key questions regarding the future of data science. The intention behind these questions is to identify issues that are either already important in the data science landscape or will be likely to be important in the near future. For the stakeholders involved in data science, there is benefit in discussing how these debates can be turned into productive discussions rather than having them manifest as impediments going forward.

1. WHAT WILL BE THE FOCAL POINTS AROUND WHICH "DATA SCIENCE" AND ITS STAKEHOLDERS COALESCE?

Given the general vagueness described above around the conceptualization of *data* within data science, why has the term *data* emerged as the focal point for this conglomeration of activity? Here, the historical comparison to information science may shed light. *Information* superseded *documentation* as the central concept of the field in the 1950s, but a considerable body of work since that time has argued that other concepts provide more theoretically robust entities, including "documents" (Frohmann 2004), "literatures" (White & McCain 1998), "relevance" (Furner 2004), or, more recently, people and their use of networked computers (Shaw 2019).

What then holds information as the central concept of the field? Certainly, the formalization of information science in the 1950s and 1960s was related in large part due to the success of Shannon's "information theory" within the fields of mathematics and electrical signal processing (Shannon & Weaver 1949). Information theory, as developed by Shannon and many others (Aspray 1985), provided conceptual metaphors of information "senders," "channels," and "receivers" that persist within information science research and education to this day (Day 2000; Gorichanaz 2017). Perhaps equally important, the success of information theory brought attention and resources to the study of information. Governments, private foundations, and for-profit companies invested in a wide range of information-focused research during the postwar period (Geoghegan 2008). As such, it is tempting to attribute the movement from documentation to information science as being one of status seeking, that is, adopting the term *information* to align preexisting bodies of work under the *documentation* label with emergent and highly prestigious research focused on *information* (Spang-Hanssen 1970/2001).

Such alignments are inevitable and are certainly happening today in the movement to data science. But the *information* concept provides more than just status. The vagueness of the term provides affordances in how it can be used and understood. As Agre (1995) illustrated, "information" provides a neutral term that enables research and professional communities to make broad intellectual territorial claims without overaligning to any particular technology, institution, or knowledge area.

Many of these same characteristics can be seen in the centering of *data* within *data science*, namely, foundational metaphors, pragmatic alignment with trending topics, and a vagueness that enables broad territorial claims. *Data* certainly comes with a foundational metaphor, detailed by Rosenberg (2013) and Frické (2019), that is at the root of work in many fields: data being that which underlies facts, evidence, truth, and information. Roseberg, Frické, and others (c.f. Borgman 2015) point out conceptual problems with this metaphor, but it undoubtedly remains strong in most sectors of academia and society. On a practical level, the label *data* also serves as a pragmatic sign of alignment with emergent and resource-heavy research areas, such as big data, the Internet of Things, and social media analytics. Finally, like *information*, the term *data* lacks conceptual baggage that would tie it to any specific technology, institution, or knowledge area. This characteristic is at the root of the "data science as metadiscipline" commentary, namely, that its techniques (whether data organization, processing, analytics, or visualization) can have application in almost any setting.

As such, *information* and *data* serve a number of functions, even if they lack unifying conceptual clarity. Their conceptual vagueness presents both benefits and drawbacks. As Hjørland (2013) describes, both centripetal and centrifugal forces exist with regard to the formation of a coherent discipline based on such a diffuse topic. Even as this tension has caused ongoing practical and institutional challenges for people involved in the study of information, it has stimulated considerable intellectual advancement related to the understanding of *information* as a conceptual and theoretical entity. Whether the development of *data science* stimulates similar advances in the understanding of the concept of *data* remains a question for future research.

In its current formative period, data science is perhaps most coherent as a platter of methods and tools, not as a grouping of research or professional areas. Data science projects tend to be distinguished by the kinds of tools and methods used, not the disciplinary topic on which they are working (Saltz, Shamshurin & Connors 2017). As vividly shown in the "Periodic Table of Data Science" (Willems 2017), data scientists may engage with a variety of tools for data collection, cleaning, processing, analysis, archiving, and distribution, including programming languages like Python and R, frameworks for analyzing large data sets like Apache Hadoop

and Pangeo, and machine learning and artificial intelligence approaches like decision trees and neural networks (Gil 2017; Ma 2023). As with any discipline, some specializations within data science will have little crossover with each other. It remains to be seen, however, whether data science specializations will continue to be structured around particular tool sets and methods, or whether particular theoretical developments, topical interests, or social problems will become more prominent focal points.

2. CAN DATA SCIENCE STAKEHOLDERS USE THE LACK OF DISCIPLINARY CLARITY AS A STRENGTH?

This leads to the next point of discussion. Conceptual advances in the understanding of *data* as a foundational concept are taking place, but as noted above, this work is largely being conducted by non–data scientists. This is one demonstration of the blurriness of the boundaries around data science. Boundaries between disciplines are always blurry. Scholars tend to interact most closely with people who work on similar topics and/or with similar methods, regardless of their disciplinary affiliation. Porous boundaries mean that new participants with diverse backgrounds will move into or across the field. This will inevitably lead to rediscovery or reinvention of particular ideas or approaches and periodic circularity in the topics of current interest. This trend has been noted by many prominent information scientists (Herner 1984; Soergel 1999; Bates 2004).

Such reinvention and circularity can stem from a lack of knowledge of historical predecessors, but it is also reflective of the ongoing nature of many information and data challenges. Some problems reemerge repeatedly, despite the best efforts of many experts. Within information science, it has long been known that information organization and retrieval methods that once worked well will break down if not regularly revisited due to changes in how language is used across space and time (Shera 1970). Data scientists encounter such circularity as they attempt to standardize data within and across organizations, leading to the well-documented fact that a significant portion of recurring data science work involves data wrangling and cleaning (Beaton et al. 2017; Kross & Guo 2019; Keller et al. 2020).

For information science, these characteristics have been viewed as problems that limit the field from gaining status within the broader ecosystem of academic disciplines (Hjørland 2013). In contrast, sociologist Jerry Jacobs (2013) has argued that innovation in the face of diffuse boundaries is what ensures the vitality of disciplines over time. Attempting to demark disciplinary boundaries is counterproductive when the grounds to claim such boundaries are uneven, as is the case with information and data science (Burnett & Bonnici 2013). Disciplinary boundaries are important to demark educational, professional, and funding institutions, but focusing too much on the need to form and define disciplinary boundaries implies a discourse of "weakness," which can cause unnecessary and repetitious debates about how to make the discipline stronger (Madsen 2016).

For data science, embracing porous boundaries by evincing openness to new ideas and people could be a means for continually refreshing the field and for broadening the diversity of data science participants generally. There are positive movements in this direction already. The Academic Data Science Alliance, for example, was created in part to "advocates for justice, diversity, equity, and inclusion of all backgrounds and lived experiences in data science and more broadly in academia" (Academic Data Science Alliance 2023). In another example, the development of the CARE principles has been an important motivator and signpost in bringing Indigenous voices into data-focused discussions (Carroll et al. 2020). This set of principles outlines key approaches to working with any data related to Indigenous peoples or communities, namely, that there should be *collective* benefit for the relevant Indigenous communities, the Indigenous communities should have *authority* to control their data, there is a *responsibility* to engage respectfully with Indigenous communities regarding any data collection or use, and that *ethics* (of the researchers and Indigenous people) should inform data use (Carroll et al. 2021). This has been an important addition to the discourse around data science over the past decade.

The extensive debates on these boundary issues have not "solved" the problems of interdisciplinarity within information science over the past 50-plus years. Views on extant or desired boundaries are inevitably dependent on one's viewpoint and will thus evolve in concert with the participants involved. But as noted in the discussion of question 1 above, these debates have been highly generative intellectually within information science. Studies

focused on data and data science may be able to take a lesson from this duality, namely, that though the recurrence of such debates will cause frustration and occasional points of circular argumentation, new voices adding to these discussions have the potential to significantly advance understanding of the nature of *data* as a foundational concept.

Finally, understanding data in all of its facets requires coupling (or at least embracing) multiple kinds of research, development, and analytical methods. Buckland (2012) has argued that the ability to be "methodologically versatile" must be a calling card for studies of data and information. Some data phenomena can only be studied via statistical methods, while other phenomena can best be studied via engineering, bibliometric, survey, or ethnographic research methods. Qualitative and quantitative methods used in complementary ways may be more effective in solving data-related problems than either type of method individually (Aragon et al. 2022). Versatility to shift between or combine these methods allows those who work in interdisciplinary areas to be flexible in the face of new societal and technical developments (Szostak 2013). Because of this need for methodological flexibility and versatility, information or data science programs that emphasize only a single methodological approach may be less resilient over time.

The openness to the new voices and methodological versatility described in this section is particularly critical as society becomes ever more data driven. As of this writing, in mid-2020, questions about data are at the center of national and international politics (use of social media data for targeted election advertising), public health (COVID-19 disease data gathering, sharing, and analysis), and global environmental change (measurements and projections of climate change). Twenty years ago, Saracevic (1997: 20) noted that "contemporary information problems are too important to be left to any one discipline." To paraphrase this for today, contemporary data problems may be too important to be able to be gathered under any one discipline or professional group.

3. CAN DATA SCIENCE FEED INTO AN "EMPOWERING PROFESSION"?

Given the broad importance of data within a range of societal sectors, there are many calls for data science stakeholders to embrace ethics as a core competency (Floridi & Taddeo 2016; Bowne-Anderson 2018; Cao 2018a). Machine learning, in particular, is under considerable scrutiny as a tool that can be used for ethically questionable purposes (McQuillan 2018b). These techniques may also produce unethical results, even with good intentions (Poirier 2021). If, as noted above, data scientists often work with "domain experts" or stakeholders in a client-like relationship, this connection with ethics will manifest on a day-to-day basis through questions about data bias, reliability, integrity, and quality. Instead of trying to deal with this issue indirectly by building better data products (e.g., visualizations or representations), data scientists have an opportunity to embrace the idea of becoming an "empowering profession" (Maack 1997). "Empowering professions" promote their client's growth and competence. They do not withhold information or stand behind a bulwark of "expertise" in limiting what is shared with a client.

This does not mean that data scientists should be trying to train everybody else to be data scientists. Instead, it suggests that data scientists could promote data literacy as a means toward the personal empowerment of the people that they work with (Monroe-White 2022). "Data literacy" in this context refers to enabling clients to understand that the products of a data science project (whether machine learning outputs or data infrastructure developments) come with certain embedded assumptions, limitations, and ethical concerns. It also refers to helping clients understand that data are embedded within particular situational and relational contexts, both when collected and when analyzed (Wilkerson & Polman 2019). This also would involve finding ways to ensure transparency and interpretability of the outcomes of data science workflows, particularly when they are used for decision-making (Stoyanovich & Lewis 2019).

A move toward empowerment would seek to understand information and data in relation to concepts like vulnerability, trust, autonomy, and agency and would work to support people in approaching the use of technology, documents, information, and data from their own cultural viewpoints, personal interests, and social settings (Srinivasan 2017). As an example, Pierre (2019) studied social media use by children and showed how digital technologies serve as sources of social support, self-expression, and self-assurance, as much as (or more so than) tools for information, data, or knowledge creation.

Data scientists are inevitably political and ethical actors, even if they do not intend or desire to be (Green 2018). Open research questions remain about the extent to which the data science profession embraces political, ethical, and empowerment-focused research agendas and professional norms. This may be critical to the future development of the field, as empowering clients—that is, enabling them to better understand their own data, and what can (or cannot) be done with them, without needing the data scientist to shepherd every step—will help data science to build a reputation for trustworthiness and social responsibility.

CONCLUSION

Information and data work must draw from multiple conceptual and practical domains. Understanding and using information and data involves articulations between people, their societies and institutions, and the technologies they create and use (Rayward 1996). Because of the vague and contested nature of *data* as a central organizing concept, new people, institutions, and technologies will continually enter the ecosystem of *data science*, engendering continual discussion of its disciplinary status. One response to this dynamic is to push for formalization of a discipline and profession with agreed-upon curricula, skills, and professional responsibilities. This may result in periodic stabilizations of disciplinary characteristics, but the discussion within this paper suggests that the definition of *data science* and its boundaries will be a source of contestation and debate for the foreseeable future. Similar issues have manifested in information science for a century or more and continue to be points of debate today.

Understanding the potential pitfalls in focusing too much effort on disciplinary formalization is critical for data science moving forward. Creating a discipline involves significant institutional work of establishing social and organizational support structures (Lenoir 1997). The new technologies and analytical methods depicted in Table 1, such as microfilm photography, early digital computers, the internet, and social media, emerged out of, and into, an interconnected web of social institutions (Agre 2003). If data science is to continue to grow as a distinct entity, attendant institutions must likewise be developed. These may include formally organized entities, such as professional associations or caucuses of academic programs. But institutional development also encompasses the emergence of professions and professional norms of conduct, processes for governing standards and tools, and the development of consortia that mediate institutional interactions (Mayernik 2016; Cutcher-Gershenfeld et al. 2017).

Many current initiatives assume that solidifying the boundaries around data science is possible or desirable. Examining these kinds of assumptions is central to building data science to be a "critical technical practice" (Agre 1997). The stakeholders who are developing the present and future of data science will need to examine the relative merits of embracing porous boundaries and methodological versatility, and they will have to deal with reinvention and circularity of central topics. Leaders in the many data science communities will also have to address whether ethics and empowerment are central to strengthening the foundation of the emerging field and associated set of professional roles. Finally, over time, funders, universities, and professional leaders will need to identify the kinds of institutional developments that will make data science more robust when it encounters the inevitable societal and technological changes of the next few decades.

ACKNOWLEDGEMENTS

FUNDING INFORMATION

COMPETING INTERESTS

Matthew Mayernik is Joint Editor-in-Chief of the Data Science Journal. This paper was submitted before he started in that role, and the editorial oversight of this article was managed by the other Joint Editor-in-Chief, Mark Parsons.

AUTHOR AFFILIATION

Matthew S. Mayernik ⓘ orcid.org/0000-0002-4122-0910

NCAR Library, National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), US

REFERENCES

Academic Data Science Alliance. 2023. ADSA vision, mission, and values. Available at https://academicdatascience.org/data-science/mission/.

Agre, PE. 1995. Institutional circuitry: Thinking about the forms and uses of information. *Information Technology and Libraries*, 14(4): 225–230.

Agre, PE. 1997. Toward a critical technical practice: Lessons learned in trying to reform AI. In: Bowker, G, Star, SL and Turner, B (eds.), *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*. Mahwah, NJ: Erlbaum. pp. 130–157.

Agre, PE. 2003. Information and institutional change: The case of digital libraries. In: Bishop, AP, Van House, NA and Buttenfield, BP (eds.), *Digital Library Use: Social Practice in Design and Evaluation*. Cambridge, MA: MIT Press. pp. 219–240.

Aragon, C, Guha, S, Kogan, M, Muller, M and **Neff, G.** 2022. *Human-Centered Data Science: An Introduction*. Cambridge, MA: MIT Press.

Aspray, WF. 1985. The scientific conceptualization of information: A survey. *IEEE Annals of the History of Computing*, 7(2): 117–140. DOI: https://doi.org/10.1109/MAHC.1985.10018

Bates, MJ. 1999. The invisible substrate of information science. *Journal of the American Society for Information Science*, 50(12), 1043–1050. DOI: https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1043::AID-ASI1>3.0.CO;2-X

Bates, MJ. 2004. Information science at the University of California at Berkeley in the 1960s: A memoir of student days. *Library Trends*, 52(4): 683–701. http://hdl.handle.net/2142/1693.

Bates, MJ. 2015. The information professions: Knowledge, memory, heritage. *Information Research*, 20(1): paper 655. Available at http://InformationR.net/ir/20-1/paper655.html.

Beaton, B, Acker, A, Di Monte, L, Setlur, S, Sutherland, T and **Tracy, SE.** 2017. Debating data science: A roundtable. *Radical History Review*, 2017(127): 133–148. DOI: https://doi.org/10.1215/01636545-3690918

Berman, F, Rutenbar, R, Christensen, H, Davidson, S, Estrin, D, Franklin, M, Hailpern, B, Martonosi, M, Raghavan, P, Stodden, V and **Szalay, A.** 2016. *Realizing the Potential of Data Science: Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group*. National Science Foundation Computer and Information Science and Engineering Advisory Committee Report. Available at https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf.

Bezuidenhout, L, Drummond-Curtis, S, Walker, B, Shanahan, H and **Alfaro-Córdoba, M.** 2021. A school and a network: CODATA-RDA data science summer schools alumni survey. *Data Science Journal*, 20(1): 1–10. DOI: https://doi.org/10.5334/dsj-2021-010

Blanchette, JF. 2012. Computing as if infrastructure mattered. *Communications of the ACM*, 55(10): 32–34. DOI: https://doi.org/10.1145/2347736.2347748

Blei, DM and **Smyth, P.** 2017. Science and data science. *Proceedings of the National Academy of Sciences*, 114(33): 8689–8692. DOI: https://doi.org/10.1073/pnas.1702076114

Borgman, CL. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/9963.001.0001

Borko, H. 1968. Information science: What is it? *American Documentation*, 19(1): 3–5. DOI: https://doi.org/10.1002/asi.5090190103

Bowne-Anderson, H. 2018. What data scientists really do, according to 35 data scientists. *Harvard Business Review*, August 15, 2018. Available at https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists.

Buckland, M. 2012. What kind of science can information science be? *Journal of the American Society for Information Science and Technology*, 63(1): 1–7. DOI: https://doi.org/10.1002/asi.21656

Burke, C. 2007. History of information science. *Annual Review of Information Science and Technology*, 41(1): 3–53. DOI: https://doi.org/10.1002/aris.2007.1440410108

Burke, CB. 2018. *America's Information Wars: The Untold Story of Information Systems in America's Conflicts and Politics from World War II to the Internet Age*. Lanham, MD: Rowman & Littlefield.

Burnett, K and **Bonnici, LJ.** 2013. Rhizomes in the iField: What does it mean to be an iSchool? *Knowledge Organization*, 40(6): 408–413. DOI: https://doi.org/10.5771/0943-7444-2013-6-408

Bush, V. 1945. *Science, the Endless Frontier: A Report to the President*. Washington, DC: US Government Printing Office.

Callaghan, S and **Darbyshire, T.** 2020. The first piece of the pattern. *Patterns*, 1(1): 100020. DOI: https://doi.org/10.1016/j.patter.2020.100020

Cao, L. 2017. Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3): 1–42. DOI: https://doi.org/10.1145/3076253

Cao, L. 2018a. Data science challenges. In: *Data Science Thinking*. Dordrecht: Springer International Publishing. pp. 93–128. DOI: https://doi.org/10.1007/978-3-319-95092-1_4

Cao, L. 2018b. Data science education. In: *Data Science Thinking*. Dordrecht: Springer International Publishing. pp. 329–348. DO: https://doi.org/10.1007/978-3-319-95092-1_11

Capurro, R and **Hjørland, B.** 2003. The concept of information. *Annual Review of Information Science and Technology*, 37(1): 343–411. DOI: https://doi.org/10.1002/aris.1440370109

Carroll, SR, Garba, I, Figueroa-Rodríguez, O, et al. 2020. The CARE principles for Indigenous data governance. *Data Science Journal*, 19(1): 1–12. DOI: https://doi.org/10.5334/dsj-2020-043

Carroll, SR, Herczog, E, Hudson, M, Russell, K and **Stall, S.** 2021. Operationalizing the CARE and FAIR principles for Indigenous data futures. *Scientific Data*, 8(1). DOI: https://doi.org/10.1038/s41597-021-00892-0

Carson, C, et al. 2016. *Data Science Planning Initiative Faculty Advisory Board Report*. University of California, Berkeley. Available at https://drive.google.com/open?id=0B8gpOw0SuKG4cGR1NTZpTzBQRGM.

Carter, D and **Sholler, D.** 2016. Data science on the ground: Hype, criticism, and everyday work. *Journal of the Association for Information Science and Technology*, 67(10): 2309–2319. DOI: https://doi.org/10.1002/asi.23563

Chambers, JM. 1993. Greater or lesser statistics: A choice for future research. *Statistics and Computing*, 3: 182–184. DOI: https://doi.org/10.1007/BF00141776

Cleveland, WS. 2001. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1): 21–26. DOI: https://doi.org/10.2307/1403527

CODATA. 2012. *CODATA Constitution (Statutes and By-Laws)*. Paris: CODATA. Available at http://www.codata.org/uploads/Constitution%202012%20Revised%20Final%20(2).pdf.

Conway, D. 2010. The data science Venn diagram. Available at http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Conway, D. 2018. *DataFramed*. #15 Building Data Science Teams. March 25, 2018. Available at https://podcastaddict.com/episode/49574942.

Cornelius, I. 2002. Theorizing information for information science. *Annual Review of Information Science and Technology*, 36(1): 392–425. DOI: https://doi.org/10.1002/aris.1440360110

Cox, RJ, Mattern, E, Mattock, L, Rodriguez, R and **Sutherland, T.** 2012. *Assessing iSchools. Journal of Education for Library and Information Science*, 53(4): 303–316. https://www.jstor.org/stable/43686923.

Craig Finlay, S, Ni, C and **Sugimoto, C.** 2018. Different mysteries, different lore: An examination of inherited referencing behaviors in academic mentoring. *Library & Information Science Research*, 40(3–4): 277–284. DOI: https://doi.org/10.1016/j.lisr.2018.09.010

Cutcher-Gershenfeld, J, et al. 2017. Five ways consortia can catalyse open science. *Nature*, 543(7647): 615–617. DOI: https://doi.org/10.1038/543615a

Davenport, TH and **Patil, DJ.** 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, October 2012. Available at https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Day, RE. 2000. The "conduit metaphor" and the nature and politics of information studies. *Journal of the American Society for Information Science*, 51(9): 805–811. DOI: https://doi.org/10.1002/(SICI)1097-4571(2000)51:9<805::AID-ASI30>3.0.CO;2-C

Day, RE. 2009. Information explosion. In: Bates, MJ and Niles Maack, M (eds.), *Encyclopedia of Library and Information Sciences*, 3rd ed. Boca Raton, FL: CRC Press. pp. 2416–2420. DOI: https://doi.org/10.1081/E-ELIS3-120044391

De Solla Price, DJ and **Beaver, D.** 1966. Collaboration in an invisible college. *American Psychologist*, 21(11): 1011–1018. DOI: https://doi.org/10.1037/h0024051

De Veaux, RD, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4: 15–30. DOI: https://doi.org/10.1146/annurev-statistics-060116-053930

Demchenko, Y, et al. 2016. EDISON Data Science Framework: A foundation for building data science profession for research and industry. In: *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. December 12–15, 2016, Luxembourg. Piscataway, NJ: IEEE. pp. 620–626. DOI: https://doi.org/10.1109/CloudCom.2016.0107

Donoho, D. 2017. 50 Years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766. DOI: https://doi.org/10.1080/10618600.2017.1384734

EDISON Project. 2017. EDISON Data Science Framework (EDSF). Available at http://edison-project.eu/edison/edison-data-science-framework-edsf.

Farkas-Conn, IS. 1990. *From Documentation to Information Science: The Beginnings and Early Development of the American Documentation Institute–American Society for Information Science.* New York: Greenwood Press.

Fayyad, U and **Hamutcu, H.** 2022. From unicorn data scientist to key roles in data science: Standardizing roles. *Harvard Data Science Review*, 4(3). DOI: https://doi.org/10.1162/99608f92.008b5006

Feder, T. 2016. Data science can be an attractive career for physicists. *Physics Today*, 69(8): 20–22. DOI: https://doi.org/10.1063/PT.3.3261

Finzer, W. 2013. The data science education dilemma. *Technology Innovations in Statistics Education*, 7(2): 1–9. https://escholarship.org/uc/item/7gv0q9dc. DOI: https://doi.org/10.5070/T572013891

Floridi, L. 2005. Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2): 351–370. DOI: https://doi.org/10.1111/j.1933-1592.2005.tb00531.x

Floridi, L and **Taddeo, M.** 2016. What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083): 20160360. DOI: https://doi.org/10.1098/rsta.2016.0360

Fowler, R. 2015. Visualizing data science. *GeoSpace* blog. December 22, 2015. American Geophysical Union. Available at https://blogs.agu.org/geospace/2015/12/22/visualizing-data-science/.

Fox, G, Maini, S, Rosenbaum, H and **Wild, D.** 2015. Data science and online education. In: *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom).* November 30, 2015–December 3, 2015, Vancouver, BC, Canada. Piscataway, NJ: IEEE. DOI: https://doi.org/10.1109/CloudCom.2015.82

Fox, P and **Hendler, J.** 2014. The science of data science. *Big Data*, 2(2): 68–70. DOI: https://doi.org/10.1089/big.2014.0011

Frické, M. 2019. The knowledge pyramid: The DIKW hierarchy. *Knowledge Organization* 49(1): 33–46. Also available in Hjørland, B and Gnoli, C (eds.), *ISKO Encyclopedia of Knowledge Organization*. Available at http://www.isko.org/cyclo/dikw. DOI: https://doi.org/10.5771/0943-7444-2019-1-33

Frohmann, B. 2004. *Deflating Information: From Science Studies to Documentation.* Toronto: University of Toronto Press. DOI: https://doi.org/10.3138/9781442673779

Furner, J. 2004. Information studies without information. *Library Trends*, 52(3): 427–446. http://hdl.handle.net/2142/1684.

Furner, J. 2016. "Data": The data. In: Kelly, M and Bielby, J (eds.), *Information Cultures in the Digital Age: A Festschrift in Honor of Rafael Capurro.* Dordrecht: Springer. pp. 287–306. DOI: https://doi.org/10.1007/978-3-658-14681-8

Furner, J. 2017. Philosophy of data: Why? *Education for Information*, 33(1): 55–70. DOI: https://doi.org/10.3233/EFI-170986

Geiger, RS, Cabasse, C, Cullens, CY, Norén, L, Fiore-Gartland, B, Das, D and **Brady, H.** 2018. *Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey.* Berkeley, CA: UC-Berkeley Institute for Data Science. DOI: https://doi.org/10.31235/osf.io/xe823

Geoghegan, BD. 2008. Historiographic conceptualization of information: A critical survey. *IEEE Annals of the History of Computing*, 30(1): 66–81. DOI: https://doi.org/10.1109/MAHC.2008.9

Gil, Y. 2017. Thoughtful artificial intelligence: Forging a new partnership for data science and scientific discovery. *Data Science*, 1(1–2): 119–129. DOI: https://doi.org/10.3233/DS-170011

Gorichanaz, T. 2017. Applied epistemology and understanding in information studies. *Information Research*, 22(4): paper 776. Available at http://InformationR.net/ir/22-4/paper776.html.

Gray, J, Gerlitz, C and **Bounegru, L.** 2018. Data infrastructure literacy. *Big Data & Society*, 5(2): 205395171878631. DOI: https://doi.org/10.1177/2053951718786316

Green, B. 2018. *Data Science as Political Action: Grounding Data Science in a Politics of Justice.* arXiv:1811.03435. Ithaca, NY: Cornell University. https://arxiv.org/abs/1811.03435.

Hammarfelt, B. 2019. Discipline. In: Hjørlan B and Gnoli, C (eds.), *Encyclopedia of Knowledge Organization*. International Society of Knowledge Organization. Available at https://www.isko.org/cyclo/discipline.

Hendler, J, Ding, Y and **Mons, B.** 2019. A journal for human and machine. *Data Intelligence*, 1(1): 1–5. DOI: https://doi.org/10.1162/dint_e_00001

Herner, S. 1984. Brief history of information science. *Journal of the American Society for Information Science*, 35(3): 157–163. DOI: https://doi.org/10.1002/asi.4630350308

Hildreth, CR and **Koenig, M.** 2002. Organizational realignment of LIS programs in academia: From independent standalone units to incorporated programs. *Journal of Education for Library and Information Science*, 43(2): 126–133. DOI: https://doi.org/10.2307/40323973

Hjørland, B. 2013. Information science and its core concepts: Levels of disagreement. In: lbekwe-SanJuan, F and Dousa, T (eds.), *Theories of Information, Communication and Knowledge*. Studies in History and Philosophy of Science, vol. 34. Dordrecht: Springer. pp. 205–235. DOI: https://doi.org/10.1007/978-94-007-6973-1_9

Hjørland, B. 2018. Data (with big data and database semantics). *Knowledge Organization*, 45(8): 685–708. Also available in Hjørland, B and Gnoli, C (eds.), *ISKO Encyclopedia of Knowledge Organization*. Available at http://www.isko.org/cyclo/data. DOI: https://doi.org/10.5771/0943-7444-2018-8-685

Huutoniemi, K, Klein, JT, Bruun, H and **Hukkinen, J.** 2010. Analyzing interdisciplinary: Typology and indicators. *Research Policy*, 39(1): 79–88. DOI: https://doi.org/10.1016/j.respol.2009.09.011

Iwata, S. 2008. Editor's note: Scientific "agenda" of data science. *Data Science Journal*, 7: 54–56. DOI: https://doi.org/10.2481/dsj.7.54

Jacobs, JA. 2013. *In Defense of Disciplines: Interdisciplinarity and Specialization in the Research University*. Chicago, IL: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226069463.001.0001

Johnson, NR. 2017. Rhetoric and the Cold War politics of information science. *Journal of the Association for Information Science and Technology*, 68(6): 1375–1384. DOI: https://doi.org/10.1002/asi.23866

Keegan, B. 2016. Journalism as a professional model for data science. Blog post, February 9, 2016. Available at https://www.brianckeegan.com/2016/02/journalism-as-a-professional-model-for-data-science/.

Keller, SA, Shipp, SS, Schroeder, AD and **Korkmaz, G.** 2020. Doing data science: A framework and case study. *Harvard Data Science Review*, 2(1). DOI: https://doi.org/10.1162/99608f92.2d83f7f5

King, JL. 2006. Identity in the I-School movement. *Bulletin of the American Society for Information Science and Technology*, 32(4). Available at http://asis.org/Bulletin/Apr-06/king.html. DOI: https://doi.org/10.1002/bult.2006.1720320406.

Klein, JT. 1996. *Crossing Boundaries: Knowledge, Disciplinarities, and Interdisciplinarities*. Charlottesville: University Press of Virginia.

Kline, R. 2004. What is information theory a theory of? Boundary work among information theorists and information scientists in the United States and Britain during the Cold War. In: Rayward, WB and Bowden, ME (eds.), *Conference on the History and Heritage of Scientific and Technological Information Systems*. Medford, NJ: Information Today. pp. 15–28.

Knorr-Cetina, KD. 1991. Epistemic cultures: Forms of reason in science. *History of Political Economy*, 23: 105–122. DOI: https://doi.org/10.1215/00182702-23-1-105

Kross, S and **Guo, PJ.** 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. May 4–9, 2019, Glasgow, Scotland. New York: ACM. DOI: https://doi.org/10.1145/3290605.3300493

Kross, S, Peng, RD, Caffo, BS, Gooding, I and **Leek, JT.** 2019. The democratization of data science education. *American Statistician*, 74(1): 1–7. DOI: https://doi.org/10.1080/00031305.2019.1668849

Larsen, RL. 2008. *History of the iSchools*. Available at https://ischoolsinc.wildapricot.org/resources/Documents/Member_Info_Resources/Coorporate%20Documents/History-of-the-iSchools-2009.pdf.

Lenoir, T. 1997. *Instituting Science: The Cultural Production of Scientific Disciplines*. Stanford, CA: Stanford University Press. DOI: https://doi.org/10.1515/9781503616059

Leonelli, S. 2015. What counts as scientific data? A relational framework. *Philosophy of Science*, 82(5): 810–821. DOI: https://doi.org/10.1086/684083

Lide, DR and **Wood, GH.** 2012. *CODATA @ 45 Years: The Story of the ICSU Committee on Data for Science and Technology (CODATA) from 1966 to 2010*. Paris: CODATA. Available at http://www.codata.org/publications/codata-history.

Ma, X. 2023. Data science for geoscience: Recent progress and future trends from the perspective of a data life cycle. In: Ma, X, Mookerjee, M, Hsu, L and Hills, D (eds.), *Recent Advancement in Geoinformatics and Data Science*. Boulder, CO: Geological Society of America. DOI: https://doi.org/10.1130/2022.2558(05)

Maack, MN. 1997. Toward a new model of the information professions: Embracing empowerment. *Journal of Education for Library and Information Science*, 38(4): 283–302. DOI: https://doi.org/10.2307/40324190

Madsen, D. 2016. Liberating interdisciplinary from myth: An exploration of the discursive construction of identities in information studies. *Journal of the Association for Information Science and Technology*, 67(11): 2697–2709. DOI: https://doi.org/10.1002/asi.23622

Madsen, D. 2017. Conspicuous by presence: The empty signifier "interdisciplinary" and the representation of absence. In: Schröter, M and Taylor, C (eds.), *Exploring Silence and Absence in Discourse*. Cham, Switzerland: Springer International Publishing. pp. 359–390. DOI: https://doi.org/10.1007/978-3-319-64580-3_13

Manyika, J, Chui, M, Brown, B, Bughin, J, Dobbs, R, Roxburgh, C and **Byers, AH.** 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 1, 2011. Available at http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

Marchionini, G. 2023. Information and data sciences: Context, units of analysis, meaning, and human impact. *Data and Information Management*, 7(1). DOI: https://doi.org/10.1016/j.dim.2023.100031

Mattmann, CA. 2013. Computing: A vision for data science. *Nature*, 493: 473–475. DOI: https://doi.org/10.1038/493473a

Mayernik, MS. 2016. Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4): 973–993. DOI: https://doi.org/10.1002/asi.23425

McQuillan, D. 2018a. Data science as machinic Neoplatonism. *Philosophy & Technology*, 31: 253–272. DOI: https://doi.org/10.1007/s13347-017-0273-3

McQuillan, D. 2018b. People's councils for ethical machine learning. *Social Media + Society*, 4(2). DOI: https://doi.org/10.1177/2056305118768303

Meng, XL. 2019. Data science: An artificial ecosystem. *Harvard Data Science Review*, 1(1). DOI: https://doi.org/10.1162/99608f92.ba20f892

Mezey, PG, Warburton, P, Jako, E and **Szekeres, Z.** 2001. Dimension concepts and reduced dimensions in toxicological QShAR databases as tools for data quality assessment. *Journal of Mathematical Chemistry*, 30(4): 375–387. DOI: https://doi.org/10.1023/A:1015138426162

Miller, S and **Hughes, D.** 2017. *The Quant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market*. Boston: Burning Glass Technologies. http://www.bhef.com/sites/default/files/bhef_2017_quant_crunch.pdf.

Monroe-White, T. 2022. Emancipating data science for Black and Indigenous students via liberatory datasets and curricula. *IASSIST Quarterly*, 46(4): 1–8. DOI: https://doi.org/10.29173/iq1007

Moore-Sloan Data Science Environments. 2018. *Creating Institutional Change in Data Science*. Available at http://msdse.org/files/Creating_Institutional_Change.pdf.

NASEM. 2017. *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: National Academies Press. DOI: https://doi.org/10.17226/24886

NASEM. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: National Academies Press. DOI: https://doi.org/10.17226/25104

Ortiz-Repiso, V, Greenberg, J and **Calzada-Prado, J.** 2018. A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. *Journal of Information Science*, 44(6): 768–784. DOI: https://doi.org/10.1177/0165551517748149

Ostler, LJ, Dahlin, TC and **Willardson, JD.** 1995. *The Closing of American Library Schools: Problems and Opportunities*. Westport, CT: Greenwood Press.

Phillips, CJ. 2019. The bases of data. *Harvard Data Science Review*, 1(2). DOI: https://doi.org/10.1162/99608f92.5c483119

Pierre, J. 2019. Building a digital family: Examining social media and social support in the development of youth "at-risk." Unpublished thesis (PhD), University of California, Los Angeles, CA. Available at https://escholarship.org/uc/item/6c70b07v.

Poirier, L. 2021. Reading datasets: Strategies for interpreting the politics of data signification. *Big Data & Society*, 8(2). DOI: https://doi.org/10.1177/20539517211029322

Pournaras, E. 2017. Cross-disciplinary higher education of data science—beyond the computer science student. *Data Science*, 1(1/2): 1–17. DOI: https://doi.org/10.3233/DS-170005

Press, G. 2013. A very short history of data science. *Forbes*, May 28, 2013. Available at https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/.

Rayward, WB. 1996. The history and historiography of information science: Some reflections. *Information Processing & Management*, 32(1): 3–17. DOI: https://doi.org/10.1016/0306-4573(95)00046-J

Ribes, D. 2018. STS, meet data science, once again. *Science, Technology, & Human Values*, 44(3): 514–539. DOI: https://doi.org/10.1177/0162243918798899

Richards, PS. 1994. *Scientific Information in Wartime: The Allied-German Rivalry, 1939–1945*. Westport, CT: Greenwood Press.

Rosenberg, D. 2013. Data before the fact. In: Gitelman, L (ed.), *"Raw Data" Is an Oxymoron*. Cambridge, MA: MIT Press. pp. 15–40.

Rumble, J. 2023. Thoughts on Starting the CODATA Data Science Journal. *Data Science Journal*, 22: 13, 1–4. DOI: https://doi.org/10.5334/dsj 2023-013

Saltz, J, Shamshurin, I and **Connors, C.** 2017. Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology*, 68(12): 2720–2728. DOI: https://doi.org/10.1002/asi.23873

Saracevic, T. 1997. Users lost: Reflections on the past, future, and limits of information science. *ACM SIGIR Forum*, 31(2): 16–27. DOI: https://doi.org/10.1145/270886.270889

Scheider, S, Nyamsuren, E, Kruiger, H and **Xu, H.** 2020. Why geographic data science is not a science. *Geography Compass*, 14(11): 1–15. DOI: https://doi.org/10.1111/gec3.12537

Shannon, CE and **Weaver, W.** 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Shaw, R. 2019. The missing profession: Towards an institution of critical technical practice. *Information Research*, 24(4): paper colis1904. http://InformationR.net/ir/24-4/colis/colis1904.html.

Shera, JH. 1970. *Sociological Foundations of Librarianship*. Bombay, India: Asia Publishing House.

Smith, FJ. 2006. Data science as an academic discipline. *Data Science Journal*, 5: 163–164. DOI: https://doi.org/10.2481/dsj.5.163

Smith, FJ. 2023. The Launch of the Data Science Journal in 2002. *Data Science Journal*, 22: 11, 1–5. DOI: https://doi.org/10.5334/dsj-2023-11

Soergel, D. 1999. The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, 50(12): 1119–1120. DOI: https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1119::AID-ASI12>3.0.CO;2-I

Spang-Hanssen, H. 1970/2001. How to teach about information as related to documentation? *HumanIT*, 5(1). https://humanit.hb.se/article/view/168.

Srinivasan, R. 2017. *Whose Global Village? Rethinking How Technology Shapes Our World*. New York: New York University Press.

Stanton, J, Palmer, CL, Blake, C and **Allard, S.** 2012. Interdisciplinary data science education. In: Xiao, N and McEwen, LR (eds.), *Special Issues in Data Management*. ACS Symposium Series 1110. Washington, DC: American Chemical Society. pp. 97–113. DOI: https://doi.org/10.1021/bk-2012-1110.ch006

Statistics Views. 2013. Nate Silver: What I need from statisticians. *Statistics Views*, August 23, 2013. Available at https://www.statisticsviews.com/article/nate-silver-what-i-need-from-statisticians/.

Stodden, V. 2020. The data science life cycle. *Communications of the ACM*, 63(7): 58–66. DOI: https://doi.org/10.1145/3360646

Stoyanovich, J and **Lewis, A.** 2019. *Teaching Responsible Data Science: Charting New Pedagogical Territory*. arXiv: 1912.10564. DOI: https://doi.org/10.1145/3360646

Szostak, R. 2013. The state of the field: Interdisciplinary research. *Issues in Interdisciplinary Studies*, 31: 44–65. http://hdl.handle.net/10323/4479.

Tang, R and **Sae-Lim, W.** 2016. Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, 32(3): 269–290. DOI: https://doi.org/10.3233/EFI-160977

Van House, N and **Sutton, SA.** 1996. The panda syndrome: An ecology of LIS education. *Journal of Education for Library and Information Science*, 37(2): 131–147. DOI: https://doi.org/10.2307/40324268

Wenger, E. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511803932

White, HD and **McCain, KW.** 1998. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4): 327–355. DOI: https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4

Wiggins, A and **Sawyer, S.** 2012. Intellectual diversity and the faculty composition of iSchools. *Journal of the American Society for Information Science and Technology*, 63(1): 8–21. DOI: https://doi.org/10.1002/asi.21619

Wilkerson, MH and **Polman, JL.** 2019. Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*, 29(1): 1–10. DOI: https://doi.org/10.1080/10508406.2019.1705664

Willems, K. 2017. The periodic table of data science. *DataCamp Official Blog*, April 12, 2017. https://web.archive.org/web/20220116104729/https://www.datacamp.com/community/blog/data-science-periodic-table.

Wing, JM and **Banks, D.** 2019. Highlights of the inaugural data science leadership summit. *Harvard Data Science Review*, 1(2). DOI: https://doi.org/10.1162/99608f92.e45fcb79

Wing, JM, Janeja, VP, Kloefkorn, T and **Erickson, LC.** 2018. *Data Science Leadership Summit: Summary Report*. New York: Columbia University. DOI: https://doi.org/10.13140/RG.2.2.13710.61764

Zhang, Y, et al. Forthcoming. Data science curriculum in the iField. *Journal of the Association for Information Science and Technology*. DOI: https://doi.org/10.1002/asi.24701

Ju[