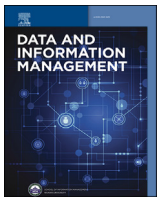
Data and Information Management 7 (2023) 100028

Contents lists available at ScienceDirect

Data and Information Management

journal homepage: www.journals.elsevier.com/data-and-information-management



The past, the present, and the future of information and data sciences: A pragmatic view

Chirag Shah

Information School, University of Washington, Seattle, 98195, WA, USA

A B S T R A C T

While data science and information science emerged as two separate disciplines with different roots, in the recent past, they have been getting integrated and intertwined in interesting and impactful ways. The traditional distinction between data and information does not easily explain the differences and overlaps between the two sciences named after them. If one claims, for instance, that information is 'meaningful data' then it is important to note that a main objective of data science is indeed to derive meaningful information out of data. Information science is not necessarily a superset or a higher level of data science. Both of these disciplines have earned their place in sciences through different pasts, paths, and possibilities. Keeping that in mind, they are discussed here while tracing their origins and understanding their positionalities in the current context. More than the past and the present, what becomes then important is where they are heading next. Several suggestions are provided to keep data science a meaningful offering within information science – as a uniqueness for the former with the strengths of the latter.

1. Introduction

On the surface, defining and differentiating data science and information science seem like definitional issues. All we need to do is look up meanings of 'data', 'information', and 'science' to declare what these two branches of science are and how they relate to each other and distinguished from one another. Alas, if it were that simple, we would not be asking these basic questions about their meanings. I believe in practice, data science and information science have more to do with how they are presented, perceived, and practiced than how they are defined or differentiated. Here, I will present my views based on my own experiences as well as that of many of my colleagues and collaborators working in these fields. I will draw these experiences from not only academic, but also industry. And while my presentation is US-centric, there are bits and pieces here that are incorporated from and should have implications for other parts of the world as well.

I will start with the easy part — trying to explain data science and information science as their historical conceptions and evolutions (Section 2). Then I will dwell into how these sciences are currently taught and practiced (Section 3). Finally, I will present a few thoughts (Section 4) on where they are going (projections) and where I believe they should go (perspectives).

2. The past

While this article is devoted to connecting and contrasting data science and information science, these two started from different places with different objectives and structures. Data science has roots in statistics, whereas information science stemmed out of library science or various interdisciplinary programs. In this section, we will try to trace back their origins in an effort to understand why they were formed and later question how they are living up to those goals.

2.1. Data science

The 1962 work "The Future of Data Analysis" by John Tukey, an American mathematician and statistician, is considered a starting point for what is now considered data science (Donoho, 2017; University of Virginia, 2022). In this piece, Tukey discusses the emerging study of "data analysis" which can be compared to contemporary data science according to David Donoho in "50 years of Data Science" (Donoho, 2017). However, it was not until 1974 when Peter Naur became the first to use the term "data science" in his work "Concise Survey of Computer Methods", in which he defined data science as "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences" (Cao, 2017). Subsequently, the field of data science evolved from related studies such as computer science, mathematics, and statistics as a way to analyze and organize fast growing databases.

The statistician Chien-Fu Jeff Wu suggested data science as a surrogate term to statistics in a 1985 lecture at the Chinese Academy of Sciences and later again in a 1998 lecture, "Statistics = Data Science?" (Donoho, 2017). The term became more popular throughout the 1990s and early 2000s and in 2001 another statistician, William Cleveland,

E-mail address: chirags@uw.edu.

https://doi.org/10.1016/j.dim.2023.100028
Received 25 December 2022; Accepted 26 January 2023

initiated an outline for the emerging field of data science as "a new discipline, broader than statistics" which would assist professionals with data computing (Brady, 2019; Cleveland, 2001; Yan & Davis, 2019). Data Science Journal was first published by The International Council for Science: Committee on Data for Science and Technology in 2002, followed soon after by Columbia University's The Journal of Data Science in 2003 (Yan & Davis, 2019).

Rachel Schutt, a Managing Director at BlackRock,[1] led one of the first data science courses in the U.S. in 2012 at Columbia University. She combined skills learned from her academic background in statistics, mathematics, and Engineering-Economic Systems to design a class "for people who are quantitatively minded and interested in putting their brains to work to solve the world's problems" (Murphy, 2022; Schutt & O'Neil, 2013). Right around that time in 2012, the Data Science Institute was launched at Columbia University. Many other academic institutions helped solidify the field of data science soon after that with the creation of their own such institutions such as the Data Science Institute at the University of Virginia in 2013, and the Data Science Initiative at the University of Michigan in 2015 (Donoho, 2017; University of Virginia, 2022). Northern Kentucky University was one of the first schools in the US to start a data science program and one of the first to be officially accredited in 2015 (Northern Kentucky University, 2022).

Since mid 2010s, many new programs and even departments have sprung up around the US and indeed around the world that specifically cater to data science. In contrast, some universities created data science as a concentration or a specialization in their existing computer science, information science, business, or statistics masters programs, which was often a quick way to have data science education on campus. This ended up creating several offerings of data science on the same campus, which often create some confusion for students. For example, my own institution – University of Washington – provides multiple data science concentrations and specializations through Information School, Department of Computer Science and Engineering, as well as Applied and Computational Math, while also offering a standalone masters in data science program. Students (and sometimes faculty) are often left figuring out how these options compare and differ.

My own version of a definition for data science refers to it as "a field of study and practice that involves collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon" (Shah, 2020). But what does this really mean in practice, especially within information science? We will come back to it after looking at what information science is and how it came to be.

2.2. Information science

It is difficult to trace a history of information science, because it is a discipline that simultaneously draws on many related disciplines (library science, computer science, communications; among others), and because the foundational principles and definitions of its core components have been debated since it emerged as a distinct discipline in the 1970–1980s, and remain unresolved (Bawden, 2008; Pawley, 2005, pp. 223–238; Rayward, 1996). Lerner (Lerner, 1994) notes that the American Documentation Institute (ADI) was founded on March 13, 1937 to improve documentation and circulation of scientific publishing using microfilm, amid advances in science and communication technologies. Schultz and Garwig (Schultz & Garwig, 1969) covers a nice history of ADI. As the spread of scientific and technological advances from World War II resulted in mass microfilm projects, the ADI was expanded, restructured, and renamed ASIS (American Society of Information Science) in 1968. Later, ASIS became ASIST, with 'T' added for 'Technology' in 2000. While one could see the evolution of ASIST and information science as parallel threads, the discipline that we came to recognize as 'information science' has another related story line about its foundation.

[1] https://www.blackrock.com/corporate/about-us.

Rayward (Rayward, 1996) examines the advances in threads that eventually contributed to information science and credits the phrase "information science" as being a product of the postwar computer revolution. He concludes that a history of information science must "draw on a range of related historical studies such as the history of science and technology, the history of printing and publishing, and the history of information institutions such as libraries, archives and museums." Bawden (Bawden, 2008) discusses key debates about the philosophical and practical foundations of information science in the early issues of Journal of Information Science, beginning in 1979, which remain unresolved. This includes controversies over the definition of Information Science, its philosophical principles by Bertie Brookes, who Bawden says is considered the founder of the "cognitive approach" to information science, and the "exact nature of the discipline." He concludes that there remain debates around the foundational aspects of this science and questions of where information science should be multidisciplinary and applied to other areas or distinct, what education in the field should entail, and what the core theories behind the application of information science are or should be.

While the discussion (and perhaps a debate) around what constitutes as information science will continue, no better example can be found for a clear recognition of information science than iSchools movement.

2.3. iSchools

The creation of iSchools and the iSchool movement evolved from a debate about the relationship between the fields of information science and library science. The interdisciplinary connections between these fields became more convoluted with the evolution of technology (Chakrabarti & Mandal, 2017; Shu & Mongeon, 2016). Professionals in the field agree that the iSchool movement started in 1988 with the 'Gang of Three,' a group of academic professionals with varying backgrounds in information, technology, and library studies from the University of Pittsburgh, Syracuse University, and Drexel University (Chakrabarti & Mandal, 2017; Dillon, 2012, pp. 267–273; Shu & Mongeon, 2016). "The initial purpose of this small group was to share information and facilitate interaction when facing the new intellectual and professional challenges in the field of information science" (King, 2006). It became the Gang of Four in 1990 with the addition of Rutgers University, and the Gang of Five in 2001 with the addition of the universities of Washington and Michigan (Rutgers was not in the Gang by 2001). Later, by 2003, Florida State University and the universities of Illinois, North Carolina, Indiana, and Texas joined, making it the Gang of Ten (Shu & Mongeon, 2016).

Between 2003 and 2005 the group started to meet twice a year to continue discussions on the intersections and evolution of the information, technology, and library fields. It became the "iCaucus" during this time, and in 2005 approximately eighty schools around the world had contributed to the movement (Shu & Mongeon, 2016). Shu and Mongeon (Shu & Mongeon, 2016) emphasized the importance of 14 out of the original 19 members of the group demonstrating the iCaucus' roots in library and information science (LIS) as they offered American Library Association (ALA) accredited graduate degree programs. Many of today's members still have this accreditation. In 2008 the iCaucus began the 'iConference,' a "forum in which iSchools' deans share their collective interests through the websites of iSchools" (Chakrabarti & Mandal, 2017). The organization was officially incorporated as the 'iSchool' in 2015 and "received 501(c)(3) non-profit status in 2016" (iSchools, 2022). The iSchool movement developed to help bridge the interdisciplinary relations and intersections of Library and Information Studies, technological developments, and societal needs.

3. The present

Today, iSchools are offering their own flavor of data science. In this section, we will examine what that flavor is, which will allow us to better understand how data science and information science co-exist and offer

interesting solutions to the same problems.

Currently, there are easily over a hundred data science and related programs offered throughout the country in both undergraduate and graduate degrees, many with either capstone or portfolio graduation requirements to give students an integrative experience in the field. There are also over 120 colleges and universities from all over the world associated with the iSchool organization. Not all LIS programs are necessarily part of the iSchool, and as of early 2022, there are over 60 American Library Association-accredited such programs in North America (ALA, 2022; iSchools, 2022; Zhang et al., 2022).

Over the years as data science gained popularity and iSchools and LIS programs found a natural way to integrate data science offerings into information science curricula, many data science concentrations, specializations, certificates, and degrees stemmed out. Often these offerings were in response to market demands and opportunities. Other times, it was a way for iSchools to claim a segment of data science that is more human-centric and context dependent, much like information science itself (Shah et al., 2021).

But by 2018, it was clear that we cannot (or should not) simply keep creating and running such data science programs without understanding the larger world of data science. I believe there were three main reasons for this. First, the field of data science was maturing. It was no longer the case that one could put anything and everything under that umbrella based on what courses or expertise they had available (e.g., databases, interface design). There are clear trends emerging from industry and various parts of academia that were shaping specific expectations from someone doing data science. Second, the students were asking how one data science program on campus differs from the other. Often, a student who could not get into one program came to another program thinking that they must all the be same/similar, but would later be disappointed to learn that not to be the case. How is a data science offered by an information science program different from the one offered by computer science or business school? Finally, as campus resources are constrained, upper-level administrators started asking if these many data science curricula are really needed and can be sustained. After all, when they are making pitch to potential donors, which data science should they be touting? For federal fundings related to data science such as the Big Data Regional Innovation Hubs (BD Hubs) (National Science Foundation (NSF), 2022a) and Harnessing the Data Revolution (National Science Foundation (NSF), 2022b) programs by the US National Science Foundation (NSF), there are often limits to how many applications can come from a single university. In that case, multiple programs across the campus may end up competing instead of collaborating, if that university does not have a way to foster collaborations across the campus. As we saw before, information science programs are still asking the fundamental questions about their identity, and therefore, it should not be surprising that they found it to be important to ask what is data science in the context of information science.

The iSchools organization created the iSchool Data Science Curriculum Committee (iDSCC) in 2019 to better understand how data science is structured in undergraduate and graduate programs. Through a collaboration across several information science programs and scholars from across the globe, they conducted surveys and studies and came up with their proposals. In an article published in 2022 Zhang et al. (Zhang et al., 2022), the authors from this committee argue that the most substantial considerations for data science curriculum include "combining technical skills with social good to solve important data and information problems, " developing a proficient and adaptable staff to provide instruction in an ever-evolving field, and assessing institutional values within the field. Other works also focus on the importance of providing students with practical experience outside the classroom as well (Elkhatib, 2017; Wang, 2018). Data science is viewed as a very interdisciplinary field as it overlaps with other fields such as library and information science, mathematics, and statistics to name a few (Brady, 2019; Donoho, 2017; Zhang et al., 2022). In particular, data science focuses on building data management and programming skills in a manner that helps students understand that information's ethical implications and translate that knowledge into more human-centered applications (Zhang et al., 2022).

Among other things, library science and information science both deal in the ways knowledge is collected, organized, retrieved, and preserved. They each uphold the value of connecting people with information and knowledge (Pawley, 2005, pp. 223–238). It is not difficult to see then how data science and information science work in tandem and even complement the other, both in how they are taught and practiced. Wang (Wang, 2018) argues that Information Science education puts a greater emphasis on "human information behavior, information ecology, knowledge management, and bibliometrics." These values are reflected in the above mentioned value of human-information connectedness and therefore in the structure of LIS programs. However, information science education and practices will have to improve their approach to technology, as they are not as proficient in computation skills as data scientists (Pawley, 2005, pp. 223–238; Wang, 2018).

Overall, the fields of data science and information science must develop curricula that emphasize the importance of doing empirical work, understanding social demands and ethical concerns that arise out of evolving technology, expanding computational and other technological innovations, and increasing information literacy in order to solve real world problems. In the next section, we look at more specifically what this means for the future of both of these fields.

4. The future

It should be clear from the previous two sections that both data science and information science have established themselves as independent and firm disciplines. While traditionally we have talked about information as a higher-level construct than data with the typical Data-Information-Knowledge-Wisdom (DIKW) hierarchy (Wikipedia, 2022), it is perhaps not appropriate to think of data science and information science with the same lens. This is important to understand as we think about where they are heading and perhaps where they should go.

Both the disciplines, like many other disciplines, came out of necessities – the necessities to solve certain kinds of problems that other areas at the time were not addressing. As these disciplines became more successful and established at doing that, we started asking questions about what they really are and what they really should be. Of course, certain practicalities of education and employment also demand that we find their strengths and distinctions, as we saw in the previous section. These discussions and questions are not going to cease anytime soon, and perhaps it is a good sign because continuing to ask such questions will keep us honest and constantly looking at the horizon.

With that in mind, let us consider the integration of data science and information science – specifically, what makes the offering of the former by the latter unique and compelling. I argue that when a data science curriculum is taught through an information science program, there are four unique characteristics that are brought out. The educators should ensure them; the students should consider them while choosing a data science program; and the employers should pay attention to them to meet their data science hiring needs.

1. **Transdisciplinary.** Information science is inherently interdisciplinary. But we can go a step further. Rather than simply working *across* disciplines, information science fosters collaboration *between* disciplines, providing a transdisciplinary framework (Gibbs, 2017). This is not an easy thing to do and while many programs and fields talk about it, information science has done it from its very inception. There, you will find many people, projects, and opportunities that really embrace transdisciplinarity in a meaningful way. Many important data science problems cut across multiple disciplines and can be more effectively solved if one were to work not just across, but with those different disciplines. Information science can provide a solid training for that.

2. **Human-centric.** A core tenet of information science is the focus on humans. This is not to say that other disciplines do not care about humans, but it is a defining characteristic of information science. This becomes very important for many of the data science processes. It is when we look at data with the right context (next point) and solve problems with humans (or users) in mind, data science becomes more than statistical or computational problem-solving. Often, people talk about *data science for social good* or *data and society*, but when it comes to data science as seen and studied through the lens of information science, there are no separate data subfields like these; caring for social good and society is integral.
3. **Context dependent.** It is almost impossible to talk about deriving information from data without talking about the context in which the data is collected, stored, analyzed, and used. Information science pays attention to context in almost everything it does. Generations of information scientists have been trained with context as a core idea. Applying that to data science problems to derive insightful information comes natural to them. Any education of data science within an information science program must emphasis this.
4. **Focus on ethics and responsibility.** The issues of fairness, equity, accountability, and ethics have been central to all forms of investigations stemming from information science – a long time before scholars started pointing out computational and data biases. It would be a shame not to leverage that strength in studying data science.

Now that we have seen what information science could, would, and should offer to data science, let us see how some of these things should be put in practice. Note that I am referring to information science schools and programs collectively as 'iSchools' below.

- First and foremost, any data science program – whether situated in iSchools or not – needs to teach certain basic skills related to data science. Before the students engage in a discussion about bias and fairness from social perspective (often included under 'data science for social good' or 'responsible AI'), they should be fluent in statistical analysis and know things like overfitting, spurious correlations, and Simpson's paradox (Wagner, 1982).
- Data science, and for that matter most computational sciences, are seen as only for those with 'hacker mentality' and can do lots of coding. Yes, coding is necessary and may not be avoided, but data science is a lot more than coding. Before diving into coding and perhaps turning some people off, especially women and certain minorities, from ever continuing with data science, we could do a better on-ramp that explores many fundamental ideas in the field without getting bogged down by challenges of datasets and programming. For example, the 1929 paper by Edwin Hubble (Hubble, 1929) that reports his observations about how the galaxies are moving away from each other and the universe is expanding, essentially giving us the first empirical proof of The Big Bang, is a very easy read. It has a simple table with 24 observations and a simple scatterplot that very clearly and effectively shows the relationship between a distant object's distance from the Earth and how fast it is moving away. No coding or database skills are needed to understand this regression problem, and it is based on a phenomenon that any high school student would know.
- While the on-ramp to data science should be an easy slope, do not let that hinder how far the actual road could or should go. Classroom realities and infrastructure limitations make it hard for us to often work with large-scale datasets or real-life problems, but we should not count them out. While we may not be launching new products from our classes, can we still give enough exposure to A/B tests to our students? Perhaps we can read some studies and invite a guest speaker from industry. Better yet, see if the class can have a *field trip* of some kind to see how things are done out there.
- Most data science programs in iSchools are professional in nature. That means the students are looking for good jobs and career advancements. Sure, some may decide to go for a PhD, but they are rare in my experience. We have to, therefore, pay extra attention to what are some of the trends in the industry (trends, not fads) and how best to incorporate them into our curriculum. Of course, one way to allow students to experience this first-hand is for them to go for internships.

5. Conclusion

Data science and information science have both earned their rightful places in the realm of scientific inquiry with the rigor that is needed. They each are able to address many important problems of our time. Rather than trying to grasp which one sits where in a seemingly unhelpful inquiry of chicken-or-egg, we should be asking what happens at the intersection of these two. While many data science offerings in information science schools or departments happened through responding to market demands and identifying right opportunities in a given time, we have come to find a better justification for why information science should be teaching data science. There are many strengths and uniqueness of information science that are ideal for data science. They include focus on human-centric values and ethics with strong methodological roots addressing tough problems at the intersection of people, information, and technology. Information science can teach the technical rigor needed for practical data science, while also providing a strong platform that has embraced transdisciplinary and context-driven work since its inception. It is time these strengths and uniqueness are recognized by educators, students, and employers.

Declaration of competing interest

The authors have no competing interests to declare.

References

iSchools. (2022). *About: iSchools*. https://www.ischools.org/about.

ALA. (2022). *Directory of ala-accredited and candidate programs in library and information studies*. american library association. https://www.ala.org/educationcareers/accreditedprograms/directory.

Bawden, D. (2008). Smoother pebbles and the shoulders of giants: The developing foundations of information science. *Journal of Information Science, 34*(4), 415–426.

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science, 22*, 297–323.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys, 50*(3), 1–42.

Chakrabarti, A., & Mandal, S. (2017). *The ischools: A study*. Library Philosophy & Practice.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69*(1), 21–26.

Dillon, A. (2012). *What it means to be an ischool*. Journal of education for library and information science.

Donoho, D. (2017). 50 years of data science. *Journal of Computational & Graphical Statistics, 26*(4), 745–766.

Elkhatib, Y. (2017). Navigating diverse data science learning: Critical reflections towards future practice. In *2017 IEEE international conference on cloud computing technology and science (CloudCom)* (pp. 357–362). IEEE.

Gibbs, P. (2017). *Transdisciplinary higher education: A theoretical basis revealed in practice*. Springer.

Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences, 15*(3), 168–173.

King, J. L. (2006). Identity in the iSchool movement. *Bulletin of the American Society for Information Science, 13-15*.

Lerner, R. G. (1994). *From documentation to information science: The beginnings and early development of the american documentation institute–american society for information science*.

Murphy, M. (2022). *Rachel schutt*. https://magazine.amstat.org/blog/2018/03/01/rachel-schutt/.

National Science Foundation (NSF). (2022a). *Big data regional innovation Hubs (BD Hubs)*. https://beta.nsf.gov/funding/opportunities/big-data-regional-innovation-hubs-bd-hubs.

National Science Foundation (NSF). (2022b). *Harnessing the data revolution*. https://beta.nsf.gov/funding/opportunities/harnessing-data-revolution-data-science-corps-dsc.

Northern Kentucky University. (2022). *Northern Kentucky university among first in u.s. to receive accreditation for data science program*. https://www.nkytribune.com/2022/09/northern-kentucky-university-among-first-in-u-s-to-receive-accreditation-for-data-science-program/.

Pawley, C. (2005). *History in the library and information science curriculum: Outline of a debate*. Libraries & culture.

Rayward, W. B. (1996). The history and historiography of information science: Some reflections. *Information Processing & Management, 32*(1), 3–17.
Schultz, C. K., & Garwig, P. L. (1969). History of the american documentation institute—a sketch. *American Documentation, 20*(2), 152–160.
Schutt, R., & O'Neil, C. (2013). *Doing data science: Straight talk from the frontline*. CA: O'Reilly Sebastopol.
Shah, C. (2020). *A hands-on introduction to data science*. Cambridge University Press.
Shah, C., Anderson, T., Hagen, L., & Zhang, Y. (2021). An ischool approach to data science: Human-centered, socially responsible, and context-driven. *Journal of the Association for Information Science and Technology, 72*(6), 793–796.
Shu, F., & Mongeon, P. (2016). The evolution of ischool movement (1988-2013): A bibliometric view. *Education for Information, 32*(4), 359–373.
University of Virginia. (2022). *The school of data science celebrates one year*. https://datascience.virginia.edu/pages/school-data-science-celebrates-one-year.
Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician, 36*(1), 46–48.
Wang, L. (2018). Twinning data science with information science in schools of library and information science. *Journal of Documentation*. Vol. 74 No. 6, pp. 1243-1257 https://doi.org/10.1108/JD-02-2018-0036.
Wikipedia. (2022). *DIKW pyramid*. https://en.wikipedia.org/wiki/DIKW_pyramid.
Yan, D., & Davis, G. E. (2019). A first course in data science. *Journal of Statistics Education, 27*(2), 99–109.
Zhang, Y., Wu, D., Hagen, L., Song, I.-Y., Mostafa, J., Oh, S., Anderson, T., Shah, C., Bishop, B. W., Hopfgartner, F., Eckert, K., Federer, L., & Saltz, J. S. (2022). Data science curriculum in the iField. *Journal of the Association for Information Science and Technology*, 1–22. https://doi.org/10.1002/asi.24701