Information Science Roles in the Emerging Field of Data Science

Gary Marchionini†

School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3360, USA

Citation: Gary Marchionini (2016). Information Science Roles in the Emerging Field of Data Science.

Received: Mar. 10, 2016
Accepted: Mar. 22, 2016

Gary Marchionini is the Dean and Cary C. Boshamer Professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill. He has published over 200 articles, chapters and reports in a variety of books and journals. Professor Marchionini was Editor-in-Chief for the *ACM Transaction on Information Systems* (2002–2008) and is the editor for the Morgan-Claypool Lecture Series on Information Concepts, Retrieval, and Services. He has been program chair for ACM SIGIR (2005) and ACM/IEEE JCDL (2002) as well as general chair of ACM DL 96 and JCDL 2006. He serves or has served on a dozen editorial boards and is a former president of the American Society of Information Science and Technology. He has had funding from the US National Science Foundation and other agencies and from Microsoft, Google, and IBM. His current interests and projects are related to interfaces that support information seeking and information retrieval, usability of personal health records, alternative representations for electronic documents, multimedia browsing strategies, digital libraries, and evaluation of interactive media, especially for learning and teaching.

There has long been discussion about the distinctions of library science, information science, and informatics, and how these areas differ and overlap with computer science. Today the term data science is emerging that generates excitement and questions about how it relates to and differs from these other areas of study. For our purposes here, I consider information science to be the general term that subsumes library science and informatics and focuses on distinctions and similarities among these disciplines that each informs data science. At the most general levels, information science deals with the genesis, flow, use, and preservation of information; computer science deals with algorithms and techniques for computational processes. Data science as a concept emerges from the applications of existing studies of measurement, representation, interpretation, and management to problems in

† Corresponding author: Gary Marchionini (E-mail: gary@ils.unc.edu).

JDIS
Journal of Data and Information Science
Vol. 1 No. 2, 2016
pp 1–6
DOI: 10.20309/jdis.201609
http://www.jdis.org

commerce, health, environment, government, and other domains. Each area of human endeavor increasingly leverages our rapidly improving abilities to capture, share, and analyze bit streams from natural and engineered activity. Ultimately, data science matters because people are able to make better decisions about the world—that is, how people in medicine, business, government, or research apply its principles and techniques to domain-specific problems.

Many of the ideas that drive data science were presented in the book *The fourth paradigm: Data-intensive scientific discovery* (Hey, Tansley, & Tolle, 2009). This book includes chapters organized into four different application areas (environment, health, science infrastructure, and scholarly communication) and gives examples of how large datasets are being used to make discoveries in those different knowledge domains. The title is based on Jim Gray's argument that three existing scientific paradigms (empirical, theoretical, and computational) are now being augmented by the new paradigm of data exploration. Data exploration is fundamentally dependent on capture, curation, and analysis of data streams. The curatorial function falls squarely within the domain of information science and information schools have much to contribute to data science as it develops in the years ahead.

It has become common to characterize data science by a series of 'v' words: volume, velocity, variety, veracity, value, and sometimes visualization (Zikopoulos & Eaton, 2011; McAfee & Brynjolfsson, 2012). Today we must manage enormous volumes of data (e.g. petabytes and exabytes) that come at high rates (e.g. gigabits/second) from a host of sources (e.g. sensors, social media, transactions, human or artificial entities) and in many different forms (e.g. numbers, text, audio, graphic, and video). Early efforts to cope with these first three "v's" stimulated work in distributed storage (e.g. data grids and clouds) and computation (e.g. parallel processing, MapReduce[①] and similar techniques across data lakes[②]) as well as mixed computational methods to integrate structured and unstructured data. Increasingly, there is recognition that our techniques only matter if we are working with GOOD data (rather than simply big data) and if the analytic techniques are appropriate to the data and the problems to be solved. For instance, see Tufekci (2014) for a critique of big data hype and methodological challenges, and Borgman (2015) or Johnson (2016) for examples of how data are embedded in social processes. Information science strongly informs data science by considering the entire data life

[①] MapReduce is a software framework for easily writing applications which process vast amounts of data in-parallel on large clusters of commodity hardware. (See details at http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html).

[②] A data lake is a massive, easily accessible, centralized repository of large volumes of structured and unstructured data. (See more at https://www.techopedia.com/definition/30172/data-lake).

cycle rather than the storage and analytics alone. This end-to-end focus is especially important for the veracity and value components of data science.

At this early stage of development, data science, like any emerging field, draws theory and practice from existing knowledge bases. Figure 1 illustrates how data science emerges from four key sectors. These sectors include three distinct but related fields (information science, statistics, and computer science) and a general sector representing different knowledge domains. It suggests that the three fields are more distinct than they are in practice. For example, a quarter of the faculty at the information school at Chapel Hill have PhDs in computer science and others have degrees in domains such as sociology, education, physics, and management, as well as information or library science fields. To explain the different contributions, each sector is briefly explained with more emphasis on information science.

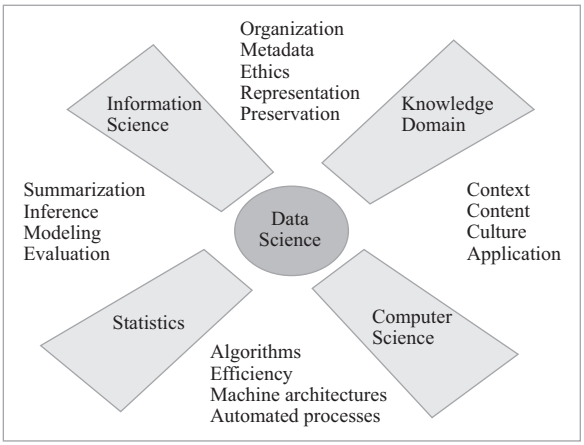

Figure 1. Data science foundations.

The knowledge domain sector will surely consume the most data scientist practitioners and each domain will strongly influence the kinds of data that are collected and the decisions made based on analysis. Chemists, political scientists, physicians, journalists, accountants, and scholars from all fields will adopt and adapt the tools and techniques created by data scientists and many of these professionals will want to specialize in data science in their domains. Disciplinary context strongly determines what data are collected, what metadata is most essential to understanding the data, how data quality is determined, and the value of analytic outcomes. Any data science training program will have to consider how domain expertise is incorporated into curricula, capstone projects, and student portfolios.

Statistics first emerged as an applied subfield of mathematics and over the past 150 years leveraged principles of probability to establish theories for sampling, estimation and error, and significance. The first statistics departments and first

dedicated journal were founded at the turn of the 20th century and statisticians today create new techniques for summarizing datasets, making and evaluating inferences from data, and invent new models for analyzing meaning and making predictions based on data. Statistical theory and techniques contribute to data science by informing decisions about what data to collect or include in analysis (e.g. sampling), providing statistical procedures that may be automated, and assessing efficacy of results (e.g. error estimation).

Computer science first emerged in the first part of the 20th century as mathematicians and physicists engineered machines that counted and compared data at speeds well beyond human abilities. Key ideas such as storing instructions (programs) as well as data and estimating computational complexity led to computational theory that blossomed in the middle of the 20th century into the new field that gave form to academic departments, journals, and professional associations. Computer scientists create or adapt mathematical and statistical algorithms that make high-speed and high-volume data processing practical. Increasingly, computer scientists work with domain experts to model data distributions, patterns, and dynamics using empirical methods (e.g. machine learning). Computer science contributes to data science by providing the machine and algorithmic techniques that domain experts can apply to ask and answer questions.

The field we know as 'information science' has roots going back to the end of the 19th century when Paul Otlet and Henri La Fontaine founded the documentation movement in Europe. This movement shifted the focus of librarians from collecting historical knowledge sources to systematic management of those sources and reflected the development of scientific management practices (Taylorism) in industry. Shiyali Ramamrita Ranganathan and others began using the term 'library science' to indicate systematic library management techniques and education. The development of information theory by scientists like Alan Turing and Norbert Wiener led to analog and digital computers in the 1940s. World War II drove advances in communications and intelligence gathering, management, and interpretation (cryptography and statistical modeling that preluded data mining) as well as information control techniques (operations research) to manage huge numbers of people and supplies moving around the entire planet. Vannevar Bush's vision of the Memex (1945) based on microform technology gave prelude to hypertext and the WWW. Claude Shannon's seminal work on information was published in 1948 and served as the base for information theory. Shannon's work (Shannon & Weaver, 1948) defined information as the amount of uncertainty reduced by a message, the bit (binary unit) as the fundamental unit of information, and the maximum capacity (bandwidth) of communication channels. Today, information science investigates architectures and ontologies for data collections;

curatorial and management processes for digital assets; information ethics and policies; knowledge genesis, flow, and preservation; human information interactions; and a variety of applications of information science principles and practices to health (health informatics), humanities (digital humanities), commerce (financial informatics), law and government (legal informatics), and environment (environmental informatics).

As noted above information scientists are concerned with the entire data life cycle and also with the socio-cultural issues associated with data collection and use. Some examples include:

- Ethical and legal conditions associated with data collection (e.g. informed consent, privacy, and legal regulations);
- Appraisal, data quality, and cleaning;
- Metadata assignment in documents context (e.g. units of measure, conditions of sensors, and vocabulary controls) both for machine efficiency and human interpretation;
- Storage and preservation of data (e.g. replication and authentication policies) and associated traces of processing and use (e.g. documentation of data cleaning, analysis algorithms applied, and overall workflows);
- Evaluation of conclusions based on data exploration and analysis and making the data and workflows findable and reusable.

It could be argued that data science is a subset of information science and some data science training programs may be housed in information schools, however, it is more strategic to view information science as an essential component of data science so that the emerging field can benefit from the diversity of perspectives that interdisciplinary collaborations bring. Information science programs can participate as key partners to ensure that students are prepared to take the lead on the socio-cultural issues noted above. Additionally, information schools that have strong technical faculty can help students develop technical skills to manage distributed data (e.g. implement relational and NoSQL solutions, and create policy-based rules for replication, security, and access); apply natural language processing, data mining, and machine learning suites for analysis; create customized scripts for management and reporting functions (e.g. using R or Python programming language); develop appropriate query systems, reports, and visualizations for data and outputs; and audit metadata assignments using domain-specific standards or ontologies (e.g. HIVE toolkit). As the field of data science continues to develop, the specific skills that data scientists and practitioners need will continue to evolve. Information science programs should be active participants in the interdisciplinary teams that will shape the field.

Perspective

Information is itself a fundamental phenomenon of interest and stands alone. Schools of information stand as meaningful and substantive entities that are critical to the education of scholars and practitioners who work across a wide range of enterprises. Data science is but one emerging field that will benefit from information school engagement.

References

Borgman, C. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge, MA: MIT Press.

Bush, V. (1945). As we may think. The Atlantic, 7. Retrieved from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Johnson, J. (2016). The question of information justice. Communications of the ACM, 59(3), 27–29.

Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf.

McAfee, A. & Brynjolfsson, E. (2012). Big data: The management revolution. Harvard Business Review, 90(10), 60–68.

Shannon, C.E., & Weaver, W. (1948). The mathematical theory of communication. Bell System Technical Journal, 27 (3), 379–423.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM '14) (pp. 505–514). Palo Alto, CA: AAAI Press.

Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data (1st ed.). New York, NY: McGraw-Hill Osborne Media.