Twinning Data Science with Information Science
In I-Schools
Lin Wang
School of Management
Tianjin Normal University
393th Binshuixidao Tianjin, China
86-022-23766058
wanglinpku@163.com

ABSTRACT
The mission, task and nature of data science are consistent with those of information science, and they are closely interrelated and together form the components of "information chain" research. From the process perspective, data science also shares similar concerns with information science. Ischool should integrate both sciences and develop organizational ambidexterity. Information science can make unique contributions to data science research.

CCS Concepts
- Social and professional topics → Professional topics → Computing education → Computing education programs → Information science education.

Keywords
Information Science, Data science, ischool, Document, Relations

1. INTRODUCTION
Accompanied with Big Data and its applications, Data Science has emerged as a new discipline in recent years. Some ischools begin to launch masters programs of data science or add more courses of data science in the current curriculum development. By doing these, they tend to address the problem of a talent shortage in data science professionals, which is estimated at as many as 140,000 to 190,000 people with deep knowledge of data analytics in U.S in 2018 [1]. The Ischool in UC Berkeley has successfully run a Master program of Information and Data Science, which pioneers the graduate education of data science in ischools. Ischools at Indiana University in Bloomington and University of Illinois at Urbana Champaign have offered a data science specialization within Master of Library Science and Master of Information Science degrees. The Ischool in University of Washington includes courses such as Introduction to Data Science, Programming for Data Science and Visualization, and Data Curation in its course catalogs. The Ischool in Syracuse University also offers a graduate-level program of data science. All of these indicate that data science has become a new growth point of ischool education. However, big data does not mean a big potential for ischool automatically. Only if we theoretically clarify the relationship and overlaps between data science and information science, can data science become an integral part of information sciences and produce the most effective education in ischools, especially those with historical background of library and information science.

2. RELATIONSHIP BETWEEN DATA SCIENCE AND INFORMATION SCIENCE
The basic idea is that the mission and nature of data science are consistent with those of information science and that they are closely interrelated and together form the components of "information chain" research. The coherent set of metatheoretical assumptions applies to both of sciences. [2] claim that data science and information science have the same core principles and share the same theory logic, methods and technique. The mission of data science is “to transform raw, messy data into actionable knowledge” which supports decision-making [3]. The mission of information science, as V. Bush said in 1945, is “making more accessible a bewildering store of knowledge” [4]. In other words, its mission is to effectively and efficiently organize and utilize knowledge [5]. Therefore, the two disciplines greatly overlap in their mission.

From the process perspective, data science also shares similar concerns with information science. What data scientists do are data wrangling, transformation, analysis, visualization and curation. Information scientists mainly conduct research on generation, collection, organization, interpretation, storage, retrieval, dissemination, transformation and use of information [6, 7]. They converge on many areas. This is not to say that they allocate the same research efforts to these areas or they have the same research scope. In fact, information science features a few core ideas, such as information retrieval, relevance research, and interaction [8]. IS researchers also attach great importance to human information behavior, information ecology, knowledge management, bibliometrics etc. These are not or are only partially covered by data science. Besides convergence, data science has some fields that IS researchers find somewhat unfamiliar or pay less attention to. They should also work to apply IS principles to these fields and acquire new methods and technical tools of data science. Stanton [3] argues that librarians are good at understanding information users’ needs and curating data that are at the start and finish ends of big data problems respectively; they further need to exercise their skills to occupy the central ground between the two ends. It is predictable that data science and information science will interact and complement each other, the ideal consequence of which is an ischool organically integrates them and develops organizational ambidexterity [9] to fully both explore the academic resources of library and information science and exploit new potentials of data science, meeting the challenges of big data era.

The relationship between data science and information science is rooted in the relationship between data and information, two
fundamental concepts of information science. The more basic a concept, the more complex it is [10]. The reason is that it contains all genes (memes) of a discipline. The relation of data and information prescribes the relation of disciplines named by them. In his Critical Delphi study of 57 leading IS scholars, Zins [11] found that the mainstream conceptual approach for defining data, information and knowledge is human-centered, cognitive-based and propositional; furthermore, the most common model for concept definition in IS treats data and information as external phenomena and knowledge as internal. Being in the same category, data and information are research foci of information science, founding the rationale of the discipline name. Using the textual data of Zins, Badia [12] applied standard information retrieval techniques and information extraction techniques to explore aggregate opinion of the experts. The result showed the consensus that data, information and knowledge must be viewed as related but different concepts. Most experts agreed that information is derived from data, and knowledge from information, which indicates the existence of a conceptual hierarchy in information science. The common definitive terms between data and information are human, use and organize or organization.

The popular names of this conceptual hierarchy mentioned above are DIKW Hierarchy (data-information-knowledge-wisdom), “Information Hierarchy” or “Knowledge Hierarchy”, which is a fundamental model of information science. In the hierarchical view, data, information and knowledge are defined in terms of one another: information is organized or structured data, knowledge may be a mix of information, understanding, capability, experience, skills and values; the processes of endowing meaning, relevance value and purpose transform data into information [13]. Oppenheim, Stenson and Wilson maintain that data is discrete and factual; when being contextualized, categorized, calculated, corrected and condensed, data becomes information [14]. In summary, concept data is inherently inter connected with concept information. Data is the basis for defining information and information is materialized by processing data. Following this logic, data science and information science are twin disciplines by nature. Theories in data science are comparable to those in information science. It is highly possible and beneficial to bridge them, which may lead to emerging a general theory of information sciences.

When we look at the evolution of the scope of research in information science, we can find a tendency of moving upwards in the DIKW hierarchy. Since 1980’s, information science research has gradually extended its scope from information issues to knowledge and intelligence ones. Knowledge management, knowledge discovering, knowledge mapping, knowledge organization become important topics in the research agenda. Some scholars explore the possibility of establishing a knowledge science discipline and a knowledge behavior field [15, 16], both of which are a logical extension or refinement of information science. These new discipline or field parallel traditional fields like epistemology and cognitive science. Intelligence issues have been unfading subjects in information science for more than twenty years, such as intelligent retrieval, AI-based interface design, business intelligence, competitive intelligence etc. This kind of word use change also makes our professionals sound more impressive. The trend seems to demonstrate that wisdom research should become a research front in these years, but it has not happened. The discussion of a wisdom concept and its nature is still scarce in information management and knowledge management literature [13]. On the contrary, data instead of wisdom has become a hot theme in information science. The newest research object is data rather than wisdom, which draws the attention of IS to the bottom of DIKW Hierarchy. Data science seems to be a step backwards from the position where library and information science already gets [3]. Does it really mean that the upward evolving direction of hierarchy is wrong? The answer is not. Lower level is not the synonym of ‘stupid’. The goal of data visualization, data analytics, data mining comes across concept information and directly refers to organizing and discovering knowledge, obtaining intelligence to support business, public and scholarly decision-making. The value of data science will be greatly undermined in the absence of collaborating with knowledge theory and intelligence theory. The development of ‘smart data’ in sciences, humanities, and business areas illustrates this point clearly. Therefore, the emergence of data science does not alter the developmental trajectory of information science indicated by DIKW Hierarchy.

3. UNIQUE CONTRIBUTIONS OF INFORMATION SCIENCE TO DATA SCIENCE

In the intellectual history of the digital world, information science theory has been ignored not a few times. Bates’ argument is still relevant in the big data era: “Our (information scientists) expertise is ignored while newcomers to information questions stumble through tens of millions of dollars of research and startup money to rediscover what information science know in 1960” [17]. What we information scientists and professionals should do right now is to bring our tools, techniques and theories into big data and data science research and best leverage them in these new areas [18]. By doing this, the positioning of information science can be reinforced and improved.

3.1 Conception of Data

The concept of data is at the heart of the entire discipline of data science. Data scientists have subscribed to the belief that data is objective and neutral. This belief is challenged by information studies demonstrating data has bias inside. Scholars of the social science of science also have similar findings. The term data is understood to mean “statistical observations and other recordings or collections of evidence”, or “a series of disconnected facts and observations” [11]. However, according to historicism and pragmatism that are the most fruitful epistemologies for information science, observations are theory-laden and influenced by culture and social contexts. The positivist view that observations are unmediated, neutral and objective facts is not true. It is on the background of theoretical presuppositions that observations are made [19]. Therefore, data is theory dependent and may be full of biases. When accepting this view of data, data scientists may be aware of the fact that data resources being analyzed and transformed are heterogeneous. They should design data access service and conduct data analysis work on the basis of distinguishing theories or biases hidden in the data. They should also apply the outputs of data mining and analytics cautiously due to the inherent theory heterogeneity of data.

3.2 Data Quality Control

Another field of data science that information science can contribute to is data quality. Not all data are created equal. Sources of big data are various: recordings of human response, physiological response, eye movement, large-scale text corpora, the long tail of “small data,” etc. [20]. Some of them are valuable,
some less, some even noise. Problems of data accuracy and credibility exist when data scientists process big data.

Until now, data science is mainly concerned with data volume, and data quality is relatively ignored. If this situation continues, the old saying “rubbish in, rubbish out” in 1980’s will survive in the big data age. It is necessary to perform data quality selection and control in order to develop mature data science. Information science can play an important role in this aspect. Marchionini points out data quality (good data) is one of the main interests of information science [21]. Information quality has been well studied in LIS since the information explosion on the Internet. Recent study [22] discusses this topic in depth and attempts to situate information quality within a sound information philosophy.

As for data credibility, information science has a long research history of information credibility that can be used for reference. Wilson’s ideas on cognitive authority [23] figured prominently in this kind of research. Hence, information science theories qualify as an outstanding candidate to substantiate the data quality field in data science. Armed with theories of information quality such as cognitive authority, data engineers and scientists are able to select up-to-date, relevant, accurate, reliable, useful and understandable data and produce high-quality data-centered products and services.

3.3 Theory Dualism of Data Science

Data science research is increasingly bifurcated between science and humanities. As for the former branch, systems, tools, scientific models and methods of data collection, analysis, and curation are research fronts. As for the latter branch, data ethics, data literacy, data privacy and policymaking are at the center of research interests. Essentially, it is the cultural neo-dualism [24]: one emphasizing pattern, syntax, and quantitative methods, the other preferring meaning, semantics and qualitative methods. This neo-dualism creates a divides that we should overcome. The academic split of young data science is not strange to information science since information science experienced a similar split during its development. Technology oriented research parallels socio-culture oriented research in information science, but they seldom cross.

The attempts to build a bridge over the divide of IS never stop. Buckland highlights the complexity of information science [25]: it deals with human understanding and belief, IT and social policies. He thinks that the disciplinary landscape is so complex that information scientists must learn to master both formal techniques and humanistic analyses. Obviously, it is not an easy task. Wilson demonstrated that information science research involves a wide range of social sciences and some highly specialized engineering [26]. Bates [17] states that the two most important methodological traditions of information science are social sciences and engineering sciences. Information scientist must be at least comfortable with both traditions in order to perform a good job. Is it possible to integrate different traditions and methodologies and to provide a coherent underlying rationale of information science? The answer is yes, but it needs time. A promising theory is document theory. Document refers to evidence in support of a fact; it denotes anything regarded as signifying something and it is equivalent to information-as-thing [27, 28]. As a verb, document means to make evident. Echoing Buckland’s view [25] that documents are the anchor of LIS, Robinson [29] argued that the nature of LIS is the study of documents and the realm of LIS is actually the realm of the document. The future of LIS is inextricably tied with the future of document. She proposed that the creation, dissemination, management, organization and retrieval, and use of document are essentials of LIS. Buckland distinguished two fundamental traditions in information science [30]: document tradition and computational tradition. Document tradition “entails acknowledging twin foundations, building not only on technique and technology (tools), but also human beings (tool-users) as individuals, groups, and society, concerned with meaning and values” [30]. From this, we can infer that document tradition fully covers the social and engineering dimensions of information science. The theory of neo-documentation, which comprehensively comprises all aspects of documents study with cognitive, technological and social dimensions as the core [31], is probably the most appropriate and practical solution to tackle the issue of disciplinary divide and establish a unified foundation of information science.

The implication of the lengthy discussion above is that document theory should to be introduced in data science. What data science means here is neither data for science nor practical tools for data processing, but the science of data. Data and document are interchangeable entities in some condition. Buckland [32] states that the difference between data retrieval and document retrieval is part-whole distinction. In his seminal article, Blair [33] pointed out the differences between data retrieval and document retrieval can be tracked down to the fundamental problem of representational indeterminacy. Data retrieval and document retrieval models exemplify the information systems with lower and higher degrees of representational indeterminacy respectively. From this viewpoint, data and document are not separate entities, but located on different positions of a spectrum of representation indeterminacy. They lie on a conceptual continuum with different levels of meaning, structure and action ability [14, 33]. “A dataset is made up of documents, and the dataset is a species of document” [34]. Theory in data science can be best described as a special kind of document theory. It is feasible to apply the general document theory to a special genre field like data in order to solve the problem of divide of data science.

4. CONCLUDING REMARKS

Brookes [35] recalled that in the beginning of information science education in University College London in 1966, the Library School was belonged to the faculty of Arts and had no discourse power in the interdisciplinary conversations of information sciences. He said it was very hard to image how to develop a respectable information science in the Library School at that time. Half a century passed, and things changed greatly. Today, ischools in the US, UK and Asian countries have played indispensable and sometimes leading role in the interdisciplinary study of information sciences. The information science research and education in ischools are flourishing and may be in the best of times. Facing the challenges and opportunities of data science, we have no reason to doubt information science will make unique contributions to data science and benefit from the interaction with the great potential partner—data science.

5. ACKNOWLEDGMENTS

This paper is an achievement of youth project of Chinese National Social Science Fund “Research on influencing factors and search patterns of SME managers' information behaviour” (Number 13CTQ040).

6. REFERENCES

functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation


