DE GRUYTER
OPEN

Perspective

Big Data and Data Science: Opportunities and Challenges of iSchools

Il-Yeol Song† & Yongjun Zhu

College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA

Citation: Il-Yeol Song & Yongjun Zhu (2017). Big Data and Data Science: Opportunities and Challenges of iSchools.

Vol. 2 No. 3, 2017
pp 1–18
DOI: 10.1515/jdis-2017-0011

Received: Mar. 25, 2017
Revised: Apr. 5, 2017
Accepted: Apr. 18, 2017

Il-Yeol Song is a Professor in the College of Computing and Informatics of Drexel University. He served as Deputy Director of NSF① -sponsored Center for Visual and Decision Informatics (CVDI) from 2012 to 2014. He is an ACM Distinguished Scientist and an ER② Fellow. He is the recipient of 2015 Peter P. Chen Award in Conceptual Modeling. His research areas include conceptual modeling, data warehousing, big data and analytics, and smart aging. Dr. Song published over 200 peer-reviewed papers in data management areas. He is a co-Editor-in-Chief of *Journal of Computing Science and Engineering* (*JCSE*) and also an Area Editor for *Data & Knowledge Engineering*. He won the Best Paper Award in the IEEE CIBCB 2004. He won four teaching awards from Drexel University, including the most prestigious Lindback Distinguished Teaching Award. Dr. Song served as the Steering Committee Chair of the ER Conference from 2010 to 2012. He delivered a keynote speech at the First Asia-Pacific iSchool Conference in 2014, ACM SAC 2015 Conference, ER2015 Conference, EDB 2016 Conference, and A-LIEP 2016 Conference.

Abstract: Due to the recent explosion of big data, our society has been rapidly going through digital transformation and entering a new world with numerous eye-opening developments. These new trends impact the society and future jobs, and thus student careers. At the heart of this digital transformation is data science, the discipline that makes sense of big data. With many rapidly emerging digital challenges ahead of us, this article discusses perspectives on iSchools' opportunities and suggestions in data science education. We argue that iSchools should empower their students with "information computing" disciplines, which we define as the ability to solve problems and create values, information, and knowledge using tools in application domains. As specific approaches to enforcing information computing disciplines

† Corresponding author: Il-Yeol Song (E-mail: song@drexel.edu).

① National Science Foundation (https://www.nsf.gov/)

② "ER" refers to "Entity-Relationship," the original root of the International Conference on Conceptual Modeling. For ER Fellow, refer to http://www.er.byu.edu/ConfGuide.html#ERfellows.

in data science education, we suggest the three foci of user-based, tool-based, and application-based. These three foci will serve to differentiate the data science education of iSchools from that of computer science or business schools. We present a layered Data Science Education Framework (DSEF) with building blocks that include the three pillars of data science (people, technology, and data), computational thinking, data-driven paradigms, and data science lifecycles. Data science courses built on the top of this framework should thus be executed with user-based, tool-based, and application-based approaches. This framework will help our students think about data science problems from the big picture perspective and foster appropriate problem-solving skills in conjunction with broad perspectives of data science lifecycles. We hope the DSEF discussed in this article will help fellow iSchools in their design of new data science curricula.

Keywords Big data; Data science; Information computing; The fourth Industrial Revolution; iSchool; Computational thinking; Data-driven paradigm; Data science lifecycle

1 Big Data, the Industrial Revolution 4.0, and Data Science

Big data has been with us for a while. During the past decade, we have witnessed the exponential growth of data and rapid advances in computing technologies. With these new trends, our society has been rapidly going through digital transformation and entering a new world with numerous eye-opening developments. There are many new concepts, technologies, tools, and applications our students need to learn, such as MapReduce/Hadoop, Spark, NoSQL, NewSQL, in-memory computing, data virtualizations, big data warehousing, data lake, cloud computing, Internet of Things (IoT), artificial intelligence, robots communicating with human, virtual reality, augmented reality, machine learning, deep learning, cognitive computing, and big data analytics. These technologies contributed to the digital revolution with the development of many innovative applications in all sectors of industry and society. Even though iSchool students do not have to learn technical aspects of all these technologies, they need to understand the concepts as well as their roles, strengths, and limitations. They also need to learn how to create new applications based on existing methods and learn tools that support those technologies, allowing them to engage the applications.

A noticeable movement in these innovative applications is the fusion of multiple technologies in physical, mechanical, and biological applications. For example, think about a scenario using smart aging technologies that are effective for the rapidly growing elderly population. Various types of health and event data are now collected on the elderly by wearable sensors and IoT sensors installed in houses and hospitals. The collected raw data can be stored in the Cloud in the form of a data lake. From the raw data, semantic metadata needs to be semi-automatically extracted, where interesting events and scenarios are identified and lifelog data schema needs

to be built. During this process, machine-learning algorithms are used to learn about behavior of the elderly. A personal profile of the elderly person is built by integrating the learned data and electronic healthcare data. An intelligent human-care robot may communicate with the elderly by learning from their personal profiles and integrated data from various sources. Any abnormal symptoms or behaviors of the elderly can be identified from the integrated data merged from the learned data, sentiment data mined from social media, and real-time data fed from wearable sensors and IoT sensors. Healthcare professionals use a dashboard built with big data analytic tools and communicate with the elderly for any important messages via a smart phone. As illustrated in this example, an innovative application utilizes a fusion of multiple technologies. It is possible to create a new service or business by collecting data through IoT sensors, storing it in the Cloud, and analyzing it through artificial intelligence and big data analytics.

A recent paradigm that explains the super-connected industry that fuses multiple technologies to manufacturing is termed the fourth Industrial Revolution, also called "IR 4.0" or "Industry 4.0." The German government's Federal Ministry of Education and Research used the term "Industry 4.0" to describe its future project in the context of planning its high-tech strategy (Lasi et al., 2014). Industry 4.0 emphasizes a smart and networked world by introducing the concept of Cyber-Physical Systems (CPS), which comprises "smart machines, storage systems, and production facilities capable of autonomously exchanging information, triggering actions and controlling each other independently" (Kagermann et al., 2013, p. 5).

Schwab (2016) discussed Industry 4.0's broad impacts on business, government, and people. To briefly summarize: (1) customers will be served with products and services with increased values due to new technologies, data, and analytics (business side); (2) citizens will increasingly be enabled to engage with governments while governments will have more tools to increase their control over populations (government side); and (3) people's notions of privacy, ownership, consumption pattern, and careers and skills development will be affected. Overall, all the aforementioned impacts are closely related to connecting people with values. The potential global economic consequences are characterized by extreme automation and connectivity (Baweja et al., 2016). It is expected that low-skill and middle-skill jobs will continue to be automated and may eventually disappear. Driverless cars and delivery drones represent this trend. IoT and related Cybersecurity issues will be another important consequence of Industry 4.0. It is highly possible that more complex problems would reside in the cyber side of Cyber-Physical Systems rather than in the traditional physical side, which requires us to be equipped with a new mindset on how we view the world and manage our resources.

In Industry 4.0, big data and its technologies, notably artificial intelligence that includes machine learning and robots, are playing important roles. Experts believe

that they will have significant impacts on our future lives and help us deal with new challenges. These new challenges and opportunities are increasingly imminent, and we should understand as well as get prepared for them in order to take full advantage of their benefits. There are several ways we can prepare for IR 4.0, such as merging offline businesses with online businesses, and integrating virtual reality and the real world to develop human-centered services and industries.

These recent trends influence virtually all societies and future jobs, including the careers of students, who must prepare for new digital challenges and data science that incorporates big data. In order to create talent for these new digital environments and future educational needs, it would be best to teach iSchool students under the newly burgeoning data science programs. Data science could be the main discipline supporting key aspects of these new developments. Data science has become one of the most demanding disciplines in recent years, yet it provides many opportunities for talented students and scientists. Industries are seeking skilled data scientists who are talented in big data technologies and capable of applying data science disciplines to applications using appropriate problem-solving tools. Many universities have begun to offer data science educational programs at all levels of undergraduate, Master's, and doctoral study. At the heart of this information field, the iSchools have been successfully producing information scientists who are proficient in doing studies/work related to data, information, and knowledge acquisition. Now is the time for iSchools to foster talented data scientists by leveraging strengths of information science disciplines.

With many rapidly emerging digital challenges ahead of us, this article discusses perspectives on iSchools' opportunities and suggestions for data science education. We argue that the iSchools should empower their students with "information computing" disciplines, which we define as the ability to solve problems and create values, information, and knowledge using computing in application domains. As specific approaches to enforcing information computing disciplines in data science education, the use of three foci we propose (user-based, tool-based, and application-based) will distinguish data science education of iSchools as presenting advantages over that of computer science or business schools. We present a layered Data Science Education Framework with building blocks that include the three pillars of data science (people, technology, and data), computational thinking, data-driven paradigms (DDPs), and the data science lifecycle. Data science courses that are built on top of this framework should thus be executed in user-based, tool-based, and application-based ways. This framework will help our students view data science problems from the big picture perspective and foster appropriate problem-solving skills alongside a data science lifecycle.

2 Data Science Education and iSchools

2.1 Data Science and Data Scientist

Davenport and Patil (2012) state that the term "data scientist" was coined in 2008 by D. J. Patil and J. Hammerbacher. Some popular definitions of data science include "*the study of the generalizable extraction of knowledge from data*" (Dhar, 2013) and "*an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information*" (Stanton, 2012). Provost and Fawcett (2013) stated that "*data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.*" Davenport and Patil (2012) discuss that "*data scientists are high-ranking professionals with the training and curiosity to make discoveries in the world of big data; they make discoveries while swimming in data; they communicate what they've learned and suggest its implications for new business directions.*" Based on these definitions, there are three important dimensions of data scientists. Data scientists should be able to (1) understand roles of big data and big data technologies, (2) work with all the steps of a data science lifecycle—discover problems, solve problems, and communicate solutions, and (3) use a range of tools to solve big data problems. We discuss how iSchools should approach the data science education in the following sections.

2.2 Challenges in Data Science Education

Industries are struggling to hire talented data scientists to handle increasingly complex data-processing demands and related challenges. Although many universities have begun to offer data science education programs at various levels (Song & Zhu, 2016), big data technologies are relatively new in traditional disciplines' curricula. Most data scientists have acquired big data skills outside the university. Due to its broad scope, incorporating big data technologies seamlessly into curriculum is a huge challenge to traditional educational departments. Broadly, we summarize those challenges as follows: (1) how to seamlessly incorporate big data revolution in data science education, (2) how to educate students to create tools/value/information/knowledge from big data by utilizing data science skills, (3) how to expand curriculum to newer topics such as intelligent augmentation and cognitive computing, and (4) how to put these emerging data management topics into curriculum.

2.3 iSchools as the Hub of Data Science Education

According to our survey (Song & Zhu, 2016), iSchools are one of pioneering communities in data science education. For over 100 years, iSchools have been

dedicated to advancing the information field, beyond library science, connecting the two ends of data and knowledge. Modernized iSchools have trained their students as versatile researchers and professionals who are skilled in information management and data analysis. What made this possible is the multidisciplinary characteristics of iSchools whose faculty members come from diverse educational backgrounds, such as including Library and Information Science, Computer Science, Education, Communication, Management, History, Information Systems, English, Psychology, Philosophy, and Industrial Engineering (Zhu, Yan, & Song, 2016). This multidisciplinary human resource asset would enable iSchools to educate successful data scientists equipped with diverse skills with broad perspectives, compared to other departments that have faculty members with focused expertise. In this regard, iSchools are probably the most ideal institutions for teaching user-based and application-focused data science education.

In summary, big data drives the fourth Industrial Revolution, and iSchools should educate successful data scientists who are prepared for this revolution (Figure 1).

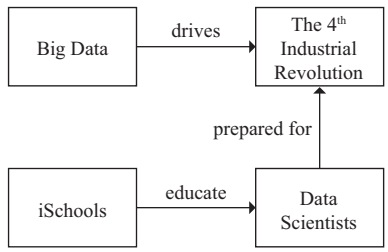

Figure 1. iSchools educate data scientists to prepare for the fourth Industrial Revolution.

3 Data Science Education Foci in iSchools

Every department has its own vision and ways of teaching data science. For iSchools, we propose the concept of *information computing*, which we define as the ability to solve problems and create values, information, and knowledge using tools in application domains. Information computing in iSchools is different from traditional computing in Computer Science. Computer Science teaches computing with more emphasis on fundamental elements such as data structures, algorithms, computational theory, and computing models. On the other hand, information computing in iSchools should emphasize users, tools, and applications. In order to implement information computing for data science education at iSchools, we recommend three foci of user-based, tool-based, and application-based (Figure 2), which should be emphasized with equal credits to produce balanced and full-fledged data scientists.

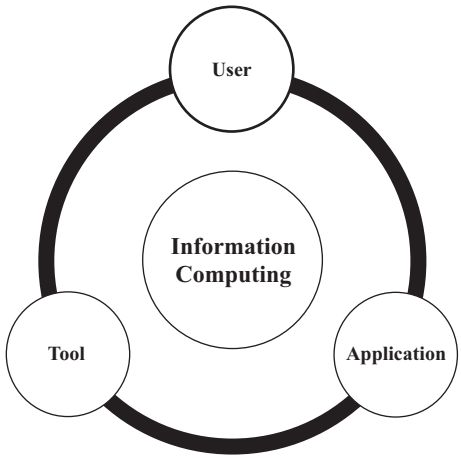

Figure 2. Three foci of information computing for iSchools.

3.1 User-based Data Science Education

User-based data science education implies the training of a data scientist who solves data science problems in some application domains or who can use data science products effectively, rather than those who develop data science products. We argue that iSchools should focus on educating data science problem-solvers at application levels, rather than developers. There are several different types of data scientists, such as those with hardcore programming backgrounds, in-depth statistical knowledge, or extensive backgrounds in business issues. While all these types of data scientists have the same goal of creating values from data, their foci are somewhat different. Data scientists with hardcore programming skills are good at developing and implementing new algorithms, tools, and systems; they also try to make programs/systems effective and efficient. The data science education offered by computer science departments will do an appropriate job in this direction. Data scientists with strong statistical backgrounds are good at interpreting/analyzing data with advanced statistical techniques. Data scientists with business backgrounds try to validate business models from data with a focus on business analytics. Data science education or its cousin business analytics offered by many business schools focus on the latter two directions.

Compared with the aforementioned types, the suggested user-based data science education tries to produce data scientists who can think from the perspective of end users of data science products or problem-solvers in applications with broad perspectives on data science lifecycles from requirements to model-building, evaluation, deployments, and data curation. Most iSchool curricula emphasize concepts such as requirement modeling, metadata management, data and software

lifecycles, human-centered computing and design, software or system evaluation, social aspects of information such as ethics of big data and data science, security and privacy issues of big data, visualization, and data curation. The coverage of these topics in computer science or business school curricula is not as extensive as those of iSchools, however. The iSchool curriculum has strong advantages in the user-based data science education by training students who understand the importance of requirement modeling, know the roles of metadata and utilize them, design and develop systems with human-centered usability in mind, consider security and privacy of data in all the stages of the data science lifecycle, validate analysis outcomes, perform appropriate story-telling for stakeholders, manage projects with keen insights on data science lifecycles, protect data and generated insights, utilize data and outcome ethically, and know how to set up strategies for data archiving and curation.

Examples of questions these data scientists address from the user-perspectives are (1) what are the major problems that online marketers want to solve using a data science approach and what are specific questions they need to ask to solve the problems? (2) what aspects of a specific data science product do the end users feel comfortable and uncomfortable with? and (3) what kinds of final reports/insights/visualizations are most helpful in delivering key information to each stakeholder, and how should they be presented to them as a story? These questions can be effectively tackled only if we focus on those users' needs, requirements, validations, and effectiveness of visualizations and presentations.

Based on these considerations, iSchools' methods of teaching data science education should broadly emphasize actual problem-solvers or users who are going to use data science products. From the practical point of view, however, with the limited number of courses in any data science degree programs, there will be no luxury of having separate courses for the subjects listed above. Hence, the main ideas of those subjects should be strategically injected into several different hard-core data science courses such as Introduction to Data Science, Data Science Project Management, Communicating with Data, Visualization, the Capstone Project, and so forth.

One pitfall we need to guard against is to be careful not to turn our students into generalists who have no strong expertise of their own in specific areas. Even though students should have a broad understanding on various issues they will meet while they work on data science projects, it is imperative that each student develops his/her own expertise in some aspects or areas in a data science lifecycle. In order to cultivate one's own expertise, students could work with established data scientists and engage in real-world projects led by capable mentors. As the Capstone Project course will help achieve this goal, it should be a mandatory core course in any data science curriculum.

3.2 Tool-based Data Science Education

The purpose of the tool-based data science education at iSchools is to emphasize the importance of automated tools and utilization of available libraries. This tool-based data science education does not mean students just need to learn tools without learning programming languages. On the contrary, we claim that iSchool students should learn both high-level analytical languages and data science tools. Learning a programming language helps students think logically and computationally (which we discuss further in Section 4.2), which is a critical skill in any IT education. We recommend high-level scripting languages such as Python and R, instead of Java or C++. Compared to Java and C++, both Python and R are relatively easy to learn, and they have a wide collection of libraries and packages, which can be easily mixed with big data analytics frameworks such as Hadoop and Spark. The key point is to develop applications by utilizing existing libraries or Web services, not to implement algorithms, which is done in computer science education. Students should be exposed to a wide collection of available libraries and services and learn how to effectively use them, without having to "reinvent the wheel." This approach will give students ample experience in developing new applications with existing libraries and help them focus on problem-solving rather than low-level coding.

In addition to high-level languages, students should also learn one or more major data science tools. There are three important reasons we emphasize the use of automated tools in data science education. First, students will more quickly understand the flow of data science projects by just clicking, without writing a code. Automated tools will handle data input, some data pre-processing, executing algorithms, and visualizing output. With automated tools, students can focus on business problems, the nature of input data, strengths and weaknesses of the algorithms executed, interpreting the output, and validating the results. These tools would allow them to quickly appreciate the value and importance of data science approaches, and support the idea of learning by doing.

Second, once students graduate from college, most will use tools or high-level data science languages in industry to work on real-world data science problems in some application domains. Hence, it makes sense to let them gain experience with popular automated data science tools and understand their strengths and limitations.

Third, automated tools are getting more powerful and diverse, with features such as data pre-processing, model-building, machine learning, validation, visualization, and report generation. Examples are IBM Watson Analytics[3], PurePredictive[4],

[3] http://www.ibm.com/analytics/watson-analytics/
[4] www.PurePredictive.com

BigML[5], RapidMiner[6], and DataRobot[7]. Strengths of these tools include a relatively shallow learning curve, partial or full automation of data mining or machine learning algorithms, and proven efficiency and efficacy. These automated tools will be progressively getting more mature, robust, and reliable. Keeping up with features of newly developed tools, experimenting with them with small to medium sized data sets for learning, understanding their pros and cons, and collecting successful use cases of the tools will give tremendous power to user-oriented data scientists. When developing an application, automated tools should be considered first before using languages. If the tools are not satisfactory, languages should be considered next. From this point of view, iSchools should not neglect teaching existing tools at multiple levels (e.g. algorithm-level, library-level, and software-level).

The shortage of qualified data scientists is already widely known as a challenging problem in industry and education[8]. The tool-based approach could help reduce this shortage by training users with limited data science knowledge to work on data science problems. The trend of empowering users with automated tools was recently termed as the "citizen data science" by Gartner (2016). Citizen data scientists are end users who solve data science problems using automated tools without coding. We see this trend in other sectors such as the auto industry, which has a much larger number of drivers than mechanics.

3.3 Application-based Data Science Education

The key points of application-based data science education at iSchools are three-fold, developing (a) project-based education, (b) the ability to work with domain experts, and (c) the expertise in one application domain. First, students should be involved in data science projects in diverse forms and levels. In earlier courses, they should understand big data use cases to which data science disciplines are applied for digital transformation. Students should experience several complete case studies with exposure to each step of a data science lifecycle using an automated tool in the very early stage of the degree program. This will give them a big picture of a data science project without coding, allow them to think about a domain they want to work on further, and help them choose a specific area they want to develop their own expertise during the rest of their degree program. Then, there should be a capstone course that students work on as a team to address real-world data science projects, perhaps with a domain expert.

[5] www.BigML.com
[6] www.rapidminer.com
[7] www.DataRobot.com
[8] www.mckinsey.com/features/big_data/

Second, students should learn how to work with domain experts and students of other disciplines as a team. Team work is generally important in any discipline, but this requirement cannot be emphasized too much in data science education as most data science problems are complex and highly interdisciplinary, and require a team effort. Data scientists should frequently communicate with each other, try to understand each other's domain specific knowledge, work to make the data and data science methods understood by all parties, and try to visualize and present the results in a clear and jargon-free manner. Therefore, the ability to collaborate with domain experts or other members of a team is an essential skill.

Third, it is desirable for students to develop their own aptitude or expertise in one domain area. Completely and fully understanding a single domain requires tremendous time and effort, and it is thus impossible for a single data scientist to have in-depth knowledge of many domains. Students should be encouraged to develop expertise and in-depth knowledge in one domain area, as most data science problems are complex and require extensive experience. Students should also be encouraged to work on non-traditional domains to create innovative applications. For example, it is reported that IBM Watson can create recipes by merging multiple food ingredients, create a piece of music after learning an artist's style, produce paintings, and create movie previews and scenarios. These applications are beyond traditional experiential avenues such as sales, marketing, healthcare, traveling, and so forth. Having an aptitude for innovation will give students opportunities to make contributions to the domain, and possibly open a new field.

Overall, iSchools' data science educational strategies should focus on teaching user-based perspectives as a problem-solver in application domains, and tending to specific needs of end users of data science products. iSchools can teach both coding using a high-level scripting language and automated data science tools. The curriculum should provide several project-based experiences so that students can explore their own options and find a preferred domain.

4 The Proposed Data Science Education Framework

In terms of data science education frameworks, iSchools have been teaching multidisciplinary subjects such as human, technological, and societal aspects of information. A typical modern iSchool curriculum already includes several core courses for data science education (Song & Zhu, 2016) such as programming, basic statistics, data management, and research methods. iSchools can strengthen data science education by incorporating a few advanced topics in curriculum, such as data science programming, big data and cloud computing, data analytics, visualization, machine learning, capstone project, and others depending on the focus of the institution. Naturally, data science curriculum design and production should be subject to each institution's strengths and preferences.

When designing a data science curriculum, we suggest implementing a consistent and coherent framework, creating courses that cover the framework while teaching students the principles that support the framework. Figure 3 shows the Data Science Education Framework (DSEF) for iSchools proposed in this paper, where we lay out the three pillars (people, technology, and data) of data science education as its foundation. As building blocks of data science courses, we then place computational thinking, data-driven paradigm, and data science lifecycle on top of the three pillars. We view computational thinking as a way of cultivating problem-solving skills, the data-driven paradigm as the underlying principle of data science, and the data science lifecycle as the basis of creating data science courses. We propose that data science courses of iSchools should be taught with the three foci of information computing—user-based, tool-based, and application-based ways.

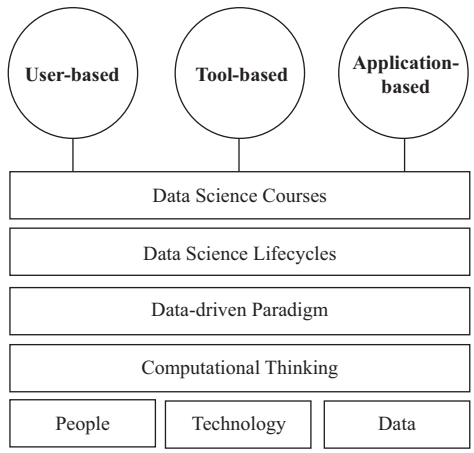

Figure 3. The proposed Data Science Education Framework (DSEF) for iSchools.

4.1 The Three Pillars of Data Science

Just as Information Science schools and their recent successors, iSchools, have focused on people, technology, and information, we also recommend people, technology, and data as the baseline of data science education of iSchools (Song & Zhu, 2016). The only difference in data science education is that the focus is on data instead of information in order to emphasize the data-driven paradigm we are currently working under. Based on these three pillars of data science, we propose to enforce computational thinking as a basic problem-solving framework. Data science courses should then be designed to broadly cover issues and techniques present in a data science lifecycle. The ways that data science courses are taught should accommodate the information computing principles in the forms of user-based, tool-based, and application-based, as noted above.

4.2 Problem Solving Framework: Computational Thinking

After being prepared with necessary background knowledge, students need to know how to solve big data problems with the use of computational thinking. Wing (2006) defines computational thinking as "*a way of solving problems, designing systems, and understanding human behavior that draws on concepts fundamental to computer science. It represents a universally applicable attitude and skill set everyone, not just computer scientists, would be eager to learn and use*" (Wing, 2006, p.33). Computational thinking is a set of thought processes necessary for solving problems using computers, used to develop all software products. As there are many software products to be applied by end users, it is regarded as a fundamental skill for everyone, especially as software becomes widespread and more important than hardware. Computational thinking is known as analytical thought that shares in many ways with mathematical thinking (solving problems), engineering thinking (designing systems), and scientific thinking (understanding human behavior) (Wing, 2008). With core concepts such as abstraction, decomposition, and patterns, Wing (2006) propose that a large, complex, and difficult problem can be reformulated into problems that we know, and these problems can be further solved using algorithms and data. Bundy (2007) stated that "*computational thinking is influencing research in nearly all disciplines, both in the sciences and the humanities*" (Bundy, 2007, p.67). For example, Wing (2008) noted several disciplines such as statistics, biology, economics, chemistry, and physics that have been influenced by computational thinking. As coding is the best way to learn computational thinking, students must learn coding as a means for solving problems systematically and thinking logically.

4.3 Data-driven Paradigm

The fundamental principle below big data and data science is the data-driven paradigm (DDP). It is a belief that data is an asset and data tells the story, leading to the data-driven decision-making. iSchools have a unique position to motivate students to understand the DDP. Students should try to understand where big data leads us, how data is used in data science methods, what the impacts of the data-driven paradigm are to science, engineering, industry, and the society. Changes introduced by big data to the society are extensively discussed by Mayer-Schonberger and Cukier (2013).

One important point discussed by Mayer-Schonberger and Cukier (2013) is the notion of datatification in DDP. Datatification implies the creation of machine-processable data by digitizing the thoughts or behaviors of humans, machines, or objects. Some important activities here could include quantification and measuring of granular data, and transforming a phenomenon into quantified data for tabulation and analysis. An example could be datafication of personal lives, i.e. datafy user

behaviors such as which music they listen to and how often and when, how often a user turns heaters on, and how often people go food shopping and when, etc. Creating a new data set via datatification will be a basis for innovative applications in many industries and disciplines.

Students should clearly understand how DDP enables new opportunities that were not possible in the past. Google is the most prominent example of a company that implemented data as a service model. Along with maintaining its position as the premiere search engine, Google has datafied our search keywords into Google Trends, and used collected data to translate one language to another. Big data enables us to ask new business questions and derive supporting decisions. As every new data source appears, we have to think about how they can be used and with which data they can be integrated to create synergy. For example, social network data, which has datafied users' behaviors, is widely used for many different purposes —social marketing, mobile sales, opinion mining, election analysis, identifying most influential users, etc. Traditional points of sale data are analyzed together with weather data to see the impacts of weather on sales. And Twitter data could be combined with weather data and electronic healthcare data to predict the number of asthma patients coming to the emergency rooms, etc. When a data source is combined with another data source, new insights can be generated.

The DDP also leads to data-driven scientific discovery and data-driven economy. Evidence-based medicine has long been practiced in medical fields, which continue to invest heavily in data science management. In the science, applications of data science techniques to big scientific data are now a new paradigm in astronomy, physics, and biology. Data science is a fast-changing field and one needs to pay attention to the state-of-the-art scientific research to be on top of new developments. This helps data science projects be equipped with the best available and effective technologies based on past experiences and cutting-edge trends.

In addition to data-driven scientific discovery, understanding of data-driven markets, data-driven services, and data-driven economy is also necessary for student learning. Major IT companies such as Google, Facebook, IBM, and Amazon are rapidly expanding their data assets by acquiring companies that have accumulated data on markets, professional relationships (e.g. LinkedIn acquired by Microsoft), weather, news, and other topics, while collecting and analyzing massive amounts of the data to find new opportunities for innovative applications. Nowadays, many customers give more value to services that are delivered by products than the products themselves, and data science helps create high quality data-driven services. New types of businesses such as Uber and Airbnb are good examples of the data-driven economy, which utilizes data to sell business services, seriously cutting into the business of taxis and hotels.

Students should also learn many big data use cases, including how the DDP is applied in those use cases. Understanding existing cases would help them create a new big data use case and datatification ideas. When students know how to identify an important big data problem by combining multiple big data sources and applying systematic thinking to create new knowledge, products, or services, they will be able to join the field of big data specialists.

4.4 Data Science Lifecycles

The specific steps used to solve data problems can be summarized as a data science lifecycle or data analytics lifecycle. We proposed a data analytics lifecycle in our previous study (Figure 6 in Song & Zhu (2016)), which we call it a data science lifecycle in this paper. As shown in Figure 4, the eight important steps include business understanding, data understanding, data preparation, model planning, model building, evaluation, deployment, and review and monitoring (Table 1).

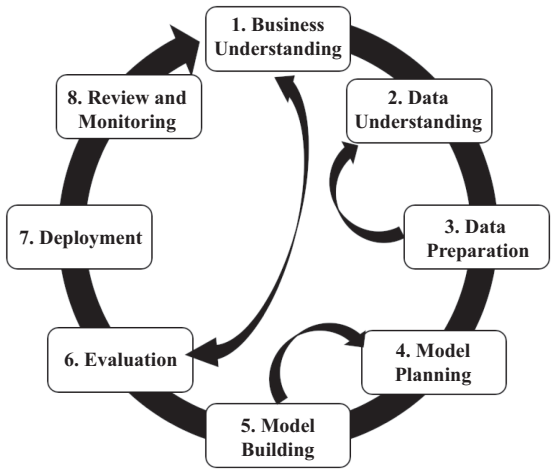

Figure 4. A data science lifecycle proposed in our previous study (Song & Zhu, 2016).

Table 1. Eight steps of a data science lifecycle.

Step	Sub-steps
1. Business understanding	What is the question to solve and what metrics are to be evaluated? Generate hypothesis; Assess resources (people, data, and tools).
2. Data understanding	Identify data resources, data reuse and integration plan, datatification, and decision on tools.
3. Data preparation	Acquire data; Perform data profiling, cleanse, and transform; Explore data and verify quality.
4. Model planning	Determine the methods, techniques, and workflow; Select key variables and determine correlation between them.
5. Model building	Build models; Perform analysis and iterate.

Step	Sub-steps
6. Evaluation	Perform evaluation against metrics; Communicate results and recommendations.
7. Deployment	Integrate analytics procedures into management dashboards and operational systems.
8. Review and monitoring	Monitor performance; Identify parts that need to be improved.

iSchools can embed all the steps of the data science lifecycle into a data science curriculum and structure courses based on it. The data science capstone project course can be a good platform for students to practise all the steps of the lifecycle.

4.5 iSchool-specific Data Science Topics

In this section, we discuss two areas where iSchools have strengths over other disciplines—data science project management and big data curation (For modeling aspects of big data, we refer to Storey & Song, 2017).

4.5.1 Data Science Project Management

One of the most important skills in data science is data science project management, which is closely related to the data science lifecycle. To successfully manage big data and data science projects, one should have overall knowledge in big data technologies, data science life cycles, data science techniques, communication skills, and application domains. Specifically, one should have the ability to identify business problems, ask specific questions on the business problems, identify appropriate data sources, choose the right data science tools and analytical platforms, evaluate the outcomes, and communicate with stakeholders. Data science project managers may also consider the datafication of things to create digital footprints. A senior officer who leads and manages big data and data science project is called a Chief Data Officer (CDO). If a data science curriculum is built on the top of data science life cycle, as we discussed above, the iSchool students have an excellent potential to learn desired skills and characteristics of a Chief Data Officer (CDO) (We refer to our previous work on this topic and related issues, Song & Zhu, 2016).

4.5.2 Big Data Curation

Big data includes both structured and unstructured data. Hence, data curation in big data is much more complex than cases dealing with only structured data. The narrow sense of data curation is archiving data to a permanent storage for the future retrieval. In the context of big data, the scope of data curation may be expanded to include the processes of collecting, integrating, organizing, annotating, and publishing data from various sources in order to keep track of provenance of data. These activities are closely related to the maintenance of data quality and metadata

management. Data quality is an important requirement for successful data science projects, and metadata management is important for data curation. Challenges to big data curation may include determining: (1) how much data to store, (2) how much the stored data will cost, (3) how data will be made secure, (4) how long data must be maintained, (5) how the curation process is automated, (6) how to remove redundancy in the curated data, (7) how to curate software and applications, (8) what is the necessary scope of data curation, (9) how to automatically capture metadata of the curated data in different data sets, and (10) how to annotate the curated data for easy extraction and management. Addressing these and other challenges will have a great impact on big data and data science ecology developments.

5 Discussion and Conclusion

As data science has now become one of the most challenging fields, there is a great demand for more talented data scientists. Many colleges and departments rapidly responded to the demand by teaching courses on data science. Among the organizations, iSchools are becoming a leading community that actively promotes data science education across many disciplines. Because of disciplinary similarities between information and data science, iSchools have strengths in data science education. By effectively utilizing these strengths, we believe iSchools will be able to educate many successful data scientists.

In this paper, we acknowledged the needs of data science education to cope with recent developments in big data and Industrial Revolution 4.0. In order to cope with the latest trends, we argue that iSchools should empower students with information computing by educating them to create values, information, and knowledge. We presented a layered Data Science Education Framework with three building blocks that form the foundation of data science—computational thinking, data-driven paradigm, and data science lifecycle. We claimed that iSchools could teach data science courses in user-based, tool-based, and application-based ways. There are many challenges iSchools need to overcome for effective data science education, however. Yet every challenge is also an opportunity to innovate. We hope the data science education framework discussed in this article will help fellow iSchools design new data science curricula that meet the needs of users in industry and society.

References

Baweja, B., Donovan, P., Haefele, M., Siddiqi, L., & Smiles, S. (2016). Extreme automation and connectivity: The global, regional, and investment implications of the Fourth Industrial Revolution. UBS White Paper for the World Economic Forum Annual Meeting 2016.

Retrieved on October, 1, 2016, from https://www.ubs.com/global/en/about_ubs/follow_ubs/highlights/davos-2016.html.

Bundy, A. (2007). Computational thinking is pervasive. Journal of Scientific and Practical Computing, 1(2), 67–69.

Davenport, T.H., & Patil, D.J. (2012). Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90(10), 70–76.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73.

Gartner, Inc. (2016). Organizing for big data through better process and governance. Retrieved on September 15, 2016, from https://www.gartner.com/doc/3002918?ref=SiteSearch&sthkw=citizen%20data%20scientist&fnl=search&srcId=1-3478922254.

Kagermann, H., Helbig, J., Hellinger, A., & Wahlster, W. (2013). Recommendations for implementing the strategic initiative Industrie 4.0: Securing the future of the German manufacturing industry; final report of the Industrie 4.0 Working Group. Retrieved on September 15, 2016, from http://www.manufacturing-policy.eng.cam.ac.uk/documents-folder/policies/germany-recommendations-for-implementing-the-strategic-initiative-industrie-4-0-bmbf-aquatic/view.

Lasi, H., Fettke, P., Kemper, H., Feld, T., & Hoffmann, M. (2014). Industry 4.0. Business & Information Systems Engineering, 6(4), 239–242.

Mayer-Schonberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51–59.

Schwab, K. (2016). The fourth industrial revolution: What it means, how to respond. World Economic Forum. Retrieved on October 20, 2016, from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond.

Stanton, J. (2012). An introduction to data science. Retrieved on September 15, 2016, from http://ischool.syr.edu/media/documents/2012/3/datasciencebook1_1.pdf.

Song, I.-Y., & Zhu, Y. (2016). Big data and data science: What should we teach? Expert Systems, 33(4), 364–373.

Storey, V., & Song, I.-Y. (2017). Big data technologies and management: What conceptual modeling can do? Data & Knowledge Engineering, 108, 50–67.

Wing, J.M. (2006). Computational thinking. Communications of the ACM, 49(3), 33–35.

Wing, J.M. (2008). Computational thinking and thinking about computing. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366(1881), 3717–3725.

Zhu, Y., Yan, E., & Song, M. (2016). Understanding the evolving academic landscape of library and information science through faculty hiring data. Scientometrics, 108(3), 1461–1478.