Big data and data science: What should we teach?

Article *in* Expert Systems · October 2015

DOI: 10.1111/exsy.12130

CITATIONS
145

READS
9,120

2 authors:

Il-Yeol Song
Drexel University
318 PUBLICATIONS **5,073** CITATIONS
SEE PROFILE

Yongjun Zhu
Weill Cornell Medical College
50 PUBLICATIONS **686** CITATIONS
SEE PROFILE

Expert Systems

Article

DOI: 10.1111/exsy.12130

Big data and data science: what should we teach?

Il-Yeol Song and Yongjun Zhu

College of Computing and Informatics, Drexel University, Philadelphia, PA, 19104, USA
E-mail: song@drexel.edu

Abstract: *The era of big data has arrived. Big data bring us the data-driven paradigm and enlighten us to challenge new classes of problems we were not able to solve in the past. We are beginning to see the impacts of big data in every aspect of our lives and society. We need a science that can address these big data problems. Data science is a new emerging discipline that was termed to address challenges that we are facing and going to face in the big data era. Thus, education in data science is the key to success, and we need concrete strategies and approaches to better educate future data scientists. In this paper, we discuss general concepts on big data, data science, and data scientists and show the results of an extensive survey on current data science education in United States. Finally, we propose various approaches that data science education should aim to accomplish.*

Keywords: big data, data science, data scientist, chief data officer, data science education

1. Introduction

In information science, a model called 'knowledge pyramid' (Figure 1) is widely used to represent the relationships among data, information, and knowledge (e.g. Rowley, 2007; Zins, 2007). Nowadays, because of tremendous amount of data that are being produced at an unprecedented rate, these data are not being effectively processed into information, which delays the extraction and production of knowledge. Thus, our society is facing even more challenging problems in transforming data into information and/or knowledge. Extracting value from raw data needs a systematic and well-defined approach. The era of big data has arrived. Big data bring us the data-driven paradigm and enlighten us to challenge new classes of problems we were not able to solve in the past. In order to solve these emerging real-world big data problems, a new multi-disciplinary study is needed.

Data science represents a set of disciplines necessary to solve big data challenges. Data, technologies, and people are the three pillars of data science, and it is obvious that these three components are not at the equal line at this moment. Data are everywhere, and technologies are being actively developed in order to cope with an increasing number of big data problems. What lagged far behind these two components is people. There is a strong indication of a shortage of people who can critically think about big data problems and who possess the necessary skills and knowledge to solve big data problems using big data technologies. Education should play its role on equipping people with the appropriate knowledge and technologies. By taking a close look at education in big data and data science, we should try to fill the gap among data, technologies, and people and build an efficient big data ecosystem where the right people use the right technologies to solve big data problems correctly and reliably. While academia and industries are trying to define what data science is, it is time for us to take a step further to see what the important characteristics of data science are as well as how we should approach data science.

The rest of this paper is organized as follows. Section 2 discusses characteristics of big data, various views on data science, and the roles of data scientists; Section 3 surveys on the current state of data science education in the United States; Section 4 presents our recommendation to data science education; and Section 5 concludes our paper.

2. What is big data, data science, and a data scientist?

2.1. Big data

'Big data' are no longer a mysterious term. It describes the data challenges that we are facing and going to face. A report from a research firm states that the global big data market is expected to grow to $46.34 billion by 2018 (MarketsandMarkets, 2013). It is also known that 90% of the data in the world today have been created in the last two years alone (Miller, 2014). But what is more challenging is that many companies have not successfully exploited big data. Gartner reports that through 2015, 85% of Fortune 500 organizations will fail to effectively exploit big data for competitive advantage (Gartner, 2012a).

How and why did the era of big data come about? Two of the major contributing factors to the emergence of the big data era include rapid advances in computing technologies and the resulting explosion of data; the former including hardware technologies such as CPU speeds and network

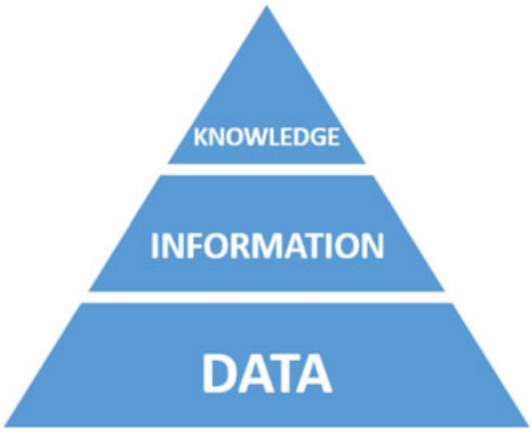

Figure 1: *The knowledge pyramid.*

bandwidths, as well as software technologies such as advent of distributed parallel processing frameworks (e.g. MapReduce and Hadoop); the latter including the increasing popularity of web-based software (e.g. search engines, social media networks, and e-commerce systems) as well as widespread usages of various sensors. These factors have collectively brought sudden explosion of data and contributed to the emergence of the big data era.

How do people define big data? Gartner popularly defined it as 3Vs: 'high-*volume*, high-*velocity*, and high-*variety* information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making' (Gartner, 2012b). Based on this definition, 4Vs were coined by adding the *veracity* dimension to the 3Vs, and 5Vs are also frequently mentioned by further adding the *value* dimension to the 4Vs. In the next few sentences, we are going to explain each V briefly based on our own thoughts and experiences. (1) *Volume* means the size of data that scales to terabytes, petabytes, or even more. We view volume as a technology solution as we can easily buy those technologies. (2) *Velocity* means the speed of creating/processing/analyzing/storing data. We view velocity as a semi-technology solution, as we can buy some solutions, but we still need to develop creative software to handle them. (3) *Variety* means different data types, sources, and modes to handle. We view variety as a software solution, as there are still many remaining challenging software issues that need to be addressed. (4) *Veracity* means quality, reliability, and uncertainty in data. We view veracity as a challenging research dimension, as it is an area that still needs to be more thoroughly researched, especially on the impacts of veracity to data integration and analytics. (5) *Value* means the discovery of actionable knowledge, high return on investment, increased relevancy to customers or products, or innovations in business operations/processes. We consider value as the most important V in the big data era. Without extracting value from big data, big data projects would not be meaningful. While we view variety and veracity as challenging dimensions, value is by far the most challenging dimension. If we are able to address these challenges and extract value from big data, then big data projects will give us opportunities for innovative solutions and chances to make an impact on technology, society, and business. We are beginning to see profound impacts of big data in every aspect of our lives and society. Figure 2 explains the concept of the 5Vs of big data, where value is at the heart of the diagram and intersects with the other 4Vs.

2.2. *Data science*

In 2011, McKinsey reported that 'By 2018, the United States alone could face a shortage of 140000 to 190000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions' (McKinsey, 2011). In order to effectively deal with today's many big data problems, a discipline called data science has emerged. Data science is referred to as the discipline that educates people who are capable of addressing challenges in the big data era.

Figure 3 shows the emergence of data science. As previously mentioned, advances in computing technologies and the sudden explosion of data have contributed to the big data era. Big data give us chances to think about which data can be used/integrated to solve business questions, new ways of solving problems, and new classes of problems we were previously not able to solve. This new way of utilizing big data in an innovative way is called the data-driven paradigm (Mayer-Schönberger & Cukier, 2013). To solve big data problems in the era of big data and the data-driven paradigm, data science should incorporate the following factors: big data infrastructure, a big data analytics lifecycle, data management skills, and behavioural disciplines. Big data infrastructures include big data technologies such as Hadoop ecosystems, NoSQL databases, in-memory computing, as well as big data enabling technologies such as cloud computing. The big data analytics life cycle includes all the stages of data analysis including business analysis, data understanding, data preparation and

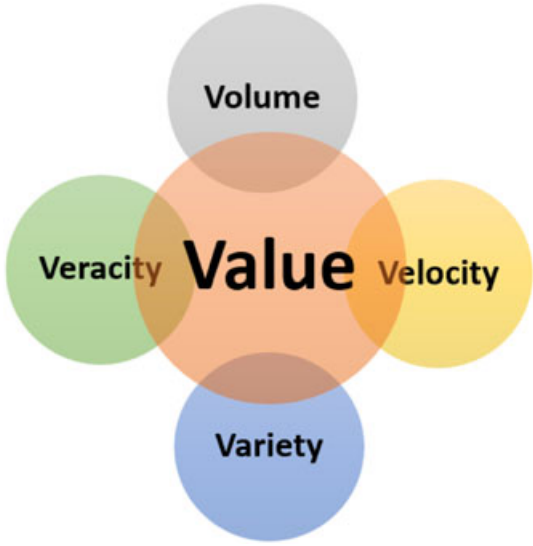

Figure 2: *5Vs of big data.*

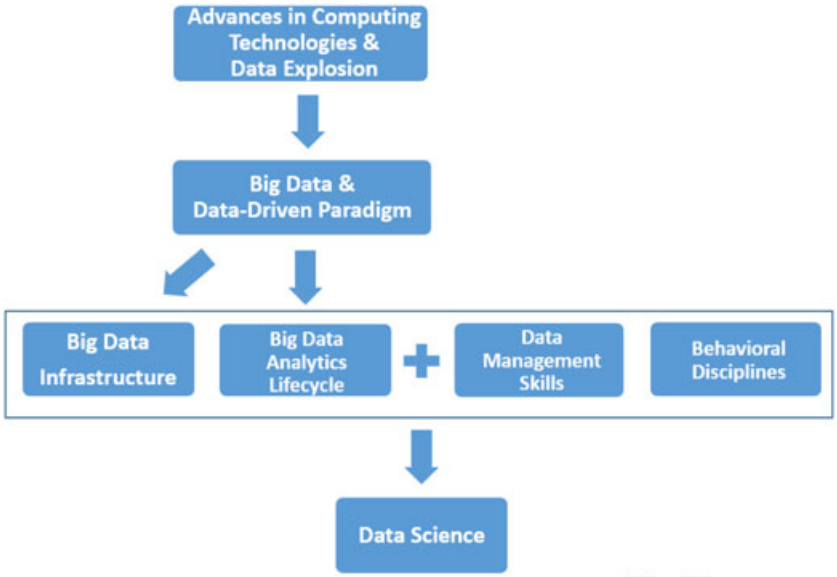

Figure 3: *The emergence of data science.*

integration, model-building, evaluation, deployment, and monitoring. Data management skills include traditional data modelling and relational database knowledge. Behavioural disciplines include soft skills related to people and business such as abilities to think critically, to ask creative questions, to communicate with domain experts (who may have no or little knowledge of data management), and to make project outcomes relevant to business.

Many scholars have given their own varying definitions of data science. In our view, Dhar (2013)'s definition – 'data science is the study of the generalizable extraction of knowledge from data' – correctly addresses the heart of data science, but in a narrow sense. Stanton (2012) defined data science as 'an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information.' As this definition encompasses all aspects of big data lifecycle in a broader sense, we think that this definition is closer to our views of data science as discussed in the preceding texts.

Provost and Fawcett (2013) stated that 'data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.' We view this characterization as containing important components of data science that should be included in data science education.

Figure 4 shows three pillars of data science (i.e. data, technologies, and people); 'data' refers to domain areas, such as relational data, non-relational data (e.g. unstructured and semi-structured data such as social media data and web data), and sensor data; 'technologies' includes Hadoop ecosystems, NoSQL, in-memory computing, data mining, machine learning, and cloud computing; and 'people' include computer scientists, statistician, domain experts, data scientists, and business analyzers.

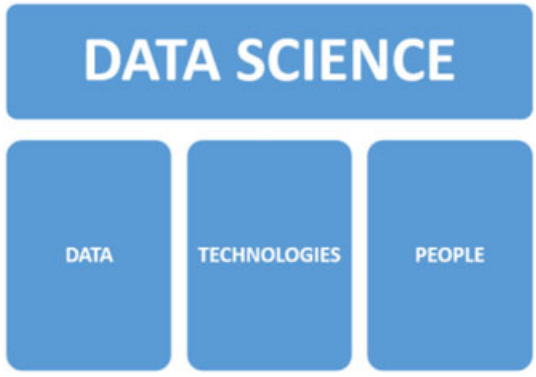

Figure 4: *Three pillars of data science.*

Based on Figure 4, data science encompasses big data and is a multi-disciplinary field that needs iterative analysis and interaction with domain experts. Data science also has strong human-side (e.g. requirements analysis, user interface design, validation of models, and communicating with domain experts), which are closely related to behavioural disciplines. Among the three pillars, the most important one is people. We can buy more computers, storages, and tools to effectively process big data, but the human ability does not scale; educating people, called data scientists, is key to addressing the challenges of the big data era.

2.3. Data scientist

It is known that the term data scientist is coined by D.J. Patil and Jeff Hammerbacher in 2008 (Davenport & Patil, 2012). Data scientist is a newly emerging job title used to refer people who are able to tackle big data problems. Major tasks that data scientists do include the following:

(1) They focus on extracting actionable knowledge from data to solve business problems. (2) They ask the right questions, with business goals and metrics, to evaluate the results against the questions in mind. (3) They get the right requirements. (4) They identify relevant data and use/reuse/merge data. (5) They select the right technologies and tools. (6) They explore solution spaces iteratively without a predetermined end in mind. (7) They work with domain experts. (8) They perform analytics, evaluate, and visualize. (9) They automate data-driven decision-making.

Who can be data scientists? People who have training in computer science, statistics, and mathematics can be data scientists as long as they possess the necessary knowledge and expertise to work on some of the tasks listed in the preceding texts. In the future, many data scientists will be educated in formal data science degree programmes at universities – a number of which are currently offering data science programmes, with an even a larger number of universities planning to offer such programmes. These formal educational programmes will be the main source of data scientists in the future. Our next section surveys the current data science educational offerings provided by universities and presents our recommendations on what should be taught in those programmes.

3. The current state of data science education in the United States

3.1. Different types of programmes

At the time of writing this paper, data science education in the United States falls into four categories: bachelor's programmes, certificate programmes, master's programmes, and specializations/concentrations in doctoral programmes. Bachelor's programmes are not very common yet, and only a few universities offer these kinds of programmes, such as Ohio State University's[1] bachelor's programme on data analytics. Certificate programmes are mostly offered online, are usually completed in less than one year, and are mostly at the graduate level, such as Columbia University's[2] certification of professional achievement in data sciences. Master's programmes are the most popular type in the data science education in the United States. The duration of master's programmes is flexible, ranging from one year to two years. Master's programmes are offered as comprehensive programmes as well as specializations/concentrations of existing related programmes. Some programmes are offered online but are largely offered as face-to-face programmes, such as New York University's[3] master's in data science. Specializations/concentrations in data science among doctoral programmes are the rarest among the four types of programmes. We were only able to find one available doctoral programme, offered by University of Washington:[4] a big data doctoral degree with tracks such as astronomy, engineering, and computer science.

We further investigated bachelor's and master's programmes to understand these programmes in detail. Our survey included programmes with the name of data science, data analytics, and analytics, and intentionally excluded programmes in areas such as business analytics and business intelligence, which are more oriented to business rather than general data science. We found a total number of 42 programmes as of August 2014. Table 1 shows the data science programmes and the types of departments that offer such programmes.

Table 1 shows that bachelor's programmes are mostly offered by joint departments and departments of computer science. Unlike master's programmes, two programmes, offered by College of Charleston[5] and University of Mary Washington,[6] are housed in departments of data science which are newly established departments. Among the existing master's programmes, more than half of the programmes are offered by joint departments; this aspect reflects that the multi-disciplinary characteristics of data science are more apparent as the level of education increases. As an independent unit, information science offers the largest number of programmes (7) followed by computer science (3) and statistics (3). This fact is not surprising because information science programmes typically focus on people, information, and technologies, which are closely related to the three pillars of data science as shown in Figure 4.

It is apparent that most academic institutes are offering data science programmes based on the collaborated effort of multiple colleges/schools/departments. By doing so, each department can contribute with its traditional strengths and can exclude complex issues such as faculty employment, logistics of developing new courses, dispersion of educational capacity, and avoiding contention among departments. In learning from each department, students have diverse opportunities to develop multiple perspectives to view and solve problems.

3.2. Core courses in bachelor's and master's programmes

We investigated core courses that are being taught in bachelor's programmes. Courses that are offered by more than one university are considered as core courses, and courses that are offered by only one university were excluded. We investigated seven universities that provide course information on their websites: Ohio State

[1]https://data-analytics.osu.edu/
[2]http://datascience.columbia.edu/certification
[3]http://cds.nyu.edu/academics/ms-in-data-science/
[4]http://data.washington.edu/education/IGERT/index.html
[5]http://www.cofc.edu/academics/majorsandminors/data-science.php
[6]http://publications.umw.edu/undergraduatecatalog/courses-of-study/minors/data/

Table 1: *Bachelor's and master's programmes in the United States (as of August 2014)*

Degree	*College/school/department offering the programme*	*No. of programmes*
Bachelor's	University/joint departments	3
	Computer Science	3
	Data Science	2
	Business	1
Master's	University/joint departments	17
	Information Science	7
	Computer Science	3
	Statistics	3
	Information Technology	1
	Operational Research	1
	Professional Studies	1

Table 2: *Core courses in bachelor's programmes (as of August 2014)*

Course	*No. of universities offering the course*
Probability and Statistics	7
Data Mining	7
Programming	5
Discrete Mathematics	4
Data Structures and Algorithms	4
Database	4
Machine Learning	4
Statistical Modelling	3
Data Visualization	3
Introduction to Data Science	2
Artificial Intelligence	2
Computer Security	2

University,[7] University of Rochester,[8] Illinois Institute of Technology,[9] Northern Kentucky University,[10] College of Charleston,[11] Ottawa University,[12] and University of Mary Washington.[13] Table 2 shows the title of the courses and the number of universities offering each course. The courses are ranked in the descending order based on the number of universities offering the course. Since courses with similar content can be offered in different titles, we used our own judgment to avoid confusion and to obtain meaningful results.

Among the courses listed in Table 2, 'Probability and Statistics' and 'Data Mining' are the most popular courses and are offered by all seven universities. In data science, knowledge on probability and statistics is fundamental and essential, which is proved by the fact that all seven universities offer the course. Data Mining has been a popular course even before the emergence of data science, although the course is important within the data science programme. Other courses offered by more than half of all seven universities include 'Programming'. 'Discrete Mathematics', 'Data Structures and Algorithms', 'Database', and 'Machine Learning'. Except for Machine Learning, all other courses are fundamental courses. Most notably, almost all of the courses listed in Table 2 are being offered by traditional computer science programmes and have similar curricula to courses taught in computer science programmes; this fact informs us that the knowledge of computer science is an important component of data science.

Using the same criteria as the bachelor's programmes, we also investigated core courses in master's programmes. A total of 15 universities that show curricula on their websites were investigated, including New York University,[14] Columbia University,[15] Worcester Polytechnic Institute,[16] University of Virginia,[17] North Carolina State University,[18] Northeastern University,[19] Texas A&M University,[20] Louisiana State University,[21] UC Berkeley,[22] Carnegie Mellon University,[23] University of Illinois at Urbana-Champaign,[24] University of Washington,[25] University of Pittsburgh,[26] University of Maryland, College Park,[27] and Indiana University Bloomington.[28] Table 3 shows the titles of the courses and the number of universities that are offering the courses.

Unlike bachelor's programmes, many diverse, advanced courses are taught in master's programmes such as 'Information Retrieval', 'Information and Social Network Analysis', and 'Text Mining'. Some courses appear in both bachelor's programmes and master's programmes, including Data Mining, Database, Machine Learning, 'Data Visualization', 'Statistical Modelling', 'Algorithms', and 'Introduction to Data Science'. Among the five courses that offered by more than half of all 15 universities, two most

[7]https://data-analytics.osu.edu/
[8]http://www.rochester.edu/data-science/degrees/BSdetails.html
[9]http://science.iit.edu/computer-science/programs/undergraduate/undergraduate-specializations#spec_dsci
[10]http://informatics.nku.edu/departments/computer-science/programs/datascience.html
[11]http://www.cofc.edu/academics/majorsandminors/data-science.php
[12]http://www.ottawa.edu/Adult-Education/Degree-Programs/Business-and-Management/Undergraduate/Data-Science-and-Technology/Data-Science-and-Technology
[13]http://publications.umw.edu/undergraduatecatalog/courses-of-study/minors/data/

[14]http://cds.nyu.edu/academics/ms-in-data-science/
[15]http://datascience.columbia.edu/master-science-data-science-0
[16]http://www.wpi.edu/academics/datascience/degree-requirements.html
[17]https://dsi.virginia.edu/academics
[18]http://analytics.ncsu.edu/?page_id=4184
[19]http://www.analytics.northwestern.edu/
[20]http://analytics.stat.tamu.edu/
[21]http://business.lsu.edu/Information-Systems-Decision-Sciences/Pages/MS-Analytics.aspx
[22]http://datascience.berkeley.edu/
[23]http://mcds.cs.cmu.edu/
[24]http://www.lis.illinois.edu/academics/degrees/ms
[25]https://ischool.uw.edu/academics/msim/curriculum/specializations
[26]http://www.ischool.pitt.edu/ist/degrees/specializations/big-data.php
[27]http://ischool.umd.edu/content/mim-specializations
[28]http://www.soic.indiana.edu/graduate/degrees/information-library-science/dual-degrees/data-science-mis.html

Table 3: *Core courses in master's programmes (as of August 2014)*

Course	*No. of universities offering the course*
Exploratory Data Analysis	10
Database	10
Data Mining	9
Data Visualization	8
Statistical Modelling	8
Machine Learning	6
Information Retrieval	5
Information and Social Network Analysis	4
Data Warehouse	4
Introduction to Data Science	3
Research Methods	3
Social Aspects of Data Science	3
Algorithms	2
Data Cleaning	2
Text Mining	2
Healthcare Analytics	2

prevalent courses are statistics-related: 'Exploratory Data Analysis' and Database. The two statistics-related courses teach advanced statistics – an advanced version of Probability and Statistics offered at an undergraduate level – which shows us that statistics is another core component of data science education along with computer science.

3.3. Observations

Based on previously mentioned investigations, we observed following things: (1) Data science bachelor's programmes are at the beginning step. (2) Most well-known universities we investigated offer data science programmes at the graduate level instead of the undergraduate level. (3) Most data science master's programmes are offered at university level or by joint departments. (4) Data Mining, Machine Learning, and Data Visualizations are the most popular core courses, whereas statistics and databases are two essential background courses in both bachelor's and master's programmes. (5) 'Explorative Data Analysis' is the most common core course at the graduate level, indicating the importance of exploring solution spaces iteratively without a predetermined end in mind.

4. Approaches to data science education

Based on our survey on data science education in the United States and our studies on data science disciplines, we suggest the following approaches to data science education. We do not intend to imply that each data science programme should teach all these disciplines we discuss in the succeeding texts; each data science programme should focus on what they do best.

4.1. Teach CDO disciplines

CDO stands for chief data officer (IBM, 2014). We define a CDO as a senior data scientist who manages big data projects with vision and leadership for data-driven business and projects. A CDO communicates with business leaders, data analysts, and users. Their primary strengths include having strong leadership skills, communications skills, an eye for business and business values, project management skills, systems-thinking skills, and technical knowledge on data. They also need to have a good understanding of big data technologies and the solution space, big data analytics lifecycle, and data management. A CDO needs to have the strong behavioural disciplines we mentioned in Section 2, as shown in Figure 3. We summarized these disciplines in Figure 5. It is not easy to create a curriculum that exactly satisfies all these requirements. However, the courses offered in data science programmes should have clear connection to

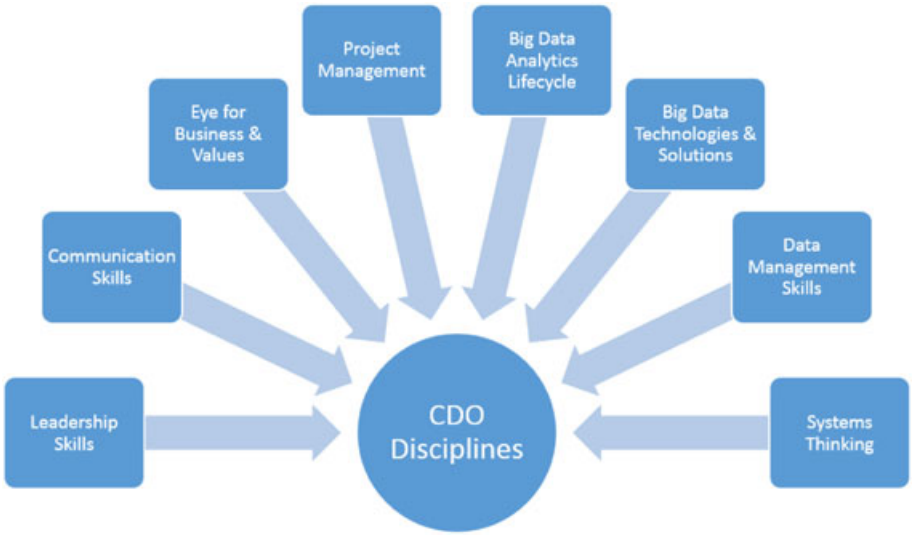

Figure 5: *CDO disciplines.*

these requirements and should try their best to meet a significant part of these requirements. In other words, individual courses should be well connected to each other to create a synergy effect.

Data science programmes offered by information science programmes have more advantages in teaching these disciplines than those offered by other programmes , as most of the previously mentioned schools teach a balanced perspective of people, information, and technologies. Some examples of courses that are offered by many information science programmes that foster CDO disciplines include 'Requirements Modelling and Management', 'Information Systems Analysis and Design', 'Information Systems Evaluation', 'Human-Computer Interaction', 'Software Project Management', 'Information Systems Management', and 'Social Aspects of Information Systems'. Many subjects of the CDO disciplines derive from the behavioural disciplines mentioned in Figure 3.

We understand that it is not easy to find an individual with the complete knowledge and skillset mentioned previously. However, it is our conviction that a leader of a big data project should be equipped with as much knowledge as possible in those areas.

4.2. *Teach with the data analytics lifecycle in mind*

The most well-known and widely used data mining process model is Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman *et al.*, 2000). This model consists of the following six steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. We recommend a data analytics lifecycle model (DALM), which modified the six steps of the CRISP-DM into eight steps, as shown in Figure 6.

In this new DALM, we split Modelling step into Modelling Planning and Model Building. We view the distinction between Model Planning and Model Building as critical in data science education and practice. By Model Planning, we mean the capability of a data scientist who knows how to select right algorithms/techniques that fit the current data set, select parameters of algorithms, establish evaluation criteria, interpret the results from the model, and change the algorithms/techniques to explore different analyses, if necessary. By Model Building, we mean the process of actually developing customized algorithms for a given question and data sets. This capacity is important in developing a new class of problems when existing techniques or tools are not effective, or when improving the effectiveness or efficiency of existing techniques or tools.

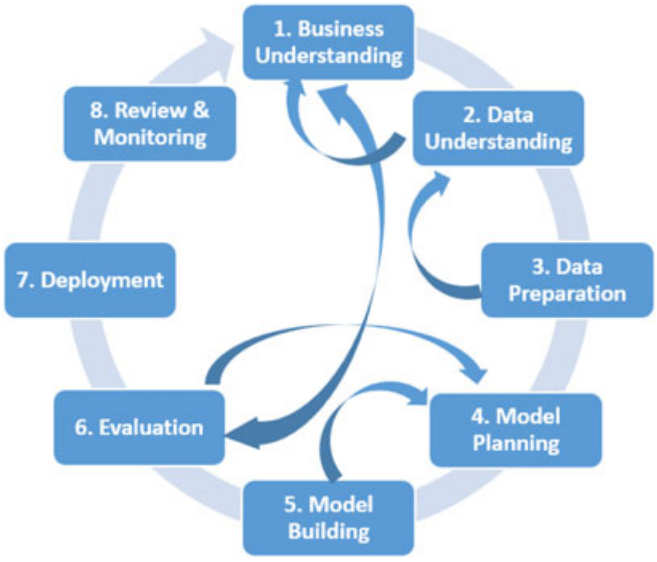

Figure 6: *The eight-step data analytics lifecycle model.*

A data scientist who works in the Model Planning step would be able to select right tools and use them for a given data set, interpret the results, and adopt different algorithms, if the first approach is not satisfactory. We believe that not every data scientist needs to know how to develop customized algorithms, which is a step included in Model Building. There are many classes of problems existing techniques or tools that can be employed to solve problems; this is especially true as more and more analytics tools become intelligent and powerful and are able to provide the users with efficient algorithms and even suggest variables to analyze. Users of these tools need to know how to use them, which existing algorithms to use, and how to interpret the results. An excellent example of a sophisticated analytics tool is IBM Watson Analytics,[29] a free, powerful tool that contains features such as data exploration, prediction, and visualization. Watson Analytics hides complex parts of data analytics but provides users with an easy-to-use interface and robust analytics power for them to perform data analytics easily. In short, data scientists who work in the model-planning mode can perform an automated analysis that uses a black-box approach using a tool such as Watson Analytics. We note that the user of those tools should still be familiar with the algorithms/methods employed in those tools. The user should be able to select the right model that fits to the data set and interpret the results properly. The user should also have enough critical thinking and reasoning ability to explore the solution space in the tool and to determine whether the tool can indeed provide a satisfactory outcome.

Another addition to CRISP-DM is step 8 on Review and Monitoring. Problem-solving big data problem needs to be constantly reviewed to see if there is any way we can improve the current solution without being constrained by the existing solution. Some ideas we can employ are adding new data sets, extending the scope of the solutions, adopting of newly developed techniques or tools, and adopting new visualization methods to improve usability.

We briefly introduce each step of the eight-step DALM: (1) The Business Understanding step identifies the business problems we try to improve/solve. In this step, a specific

[29]http://www.ibm.com/analytics/watson-analytics/

research question that could deliver benefits to a business should be identified to begin with. (2) The Data Understating step identifies the right data sources and data sets (as well as requirements) that are necessary in analyzing the business' problems. Any new data sources or merging multiple data sources could be considered. (3) The Data Preparation step performs actual data preparation and integration by using techniques such as extract, transform, and load; extract load transform; and data virtualization to integrate/generate the data set for the analysis. (4) The Model Planning step identifies parameters, methods, and techniques that will be used in Model Building step. (5) The Model Building step builds models that execute the methods and techniques planned in the previous step. (6) The Evaluation step validates the models and results and determines the success or failure against the business question to be solved. (7) The Deployment step implements models in a production environment. (8) The Review and Monitoring step monitors performance of the deployed solution and identifies the parts that need to be improved.

From the eight-step data analytics lifecycle shown in Figure 6, each department/school could find its own strengths in different steps. For example, business schools have strengths in Business Understanding, information schools have strengths in Data Understanding, departments of statistics have strengths in Model Planning, and departments of computer science have strengths in Model Building. Of course, this may not be possible as the departments/schools at different universities, who may share the same name, may also have different characteristics. Nevertheless, each department/school can choose several steps where they can excel and focus on those steps.

4.3. Teach big data technologies and Model Building techniques

Big data technologies and Model Building techniques are the two most technical components of data science programmes. These two components should be emphasized in data science programmes with a strong focus on computing.

Some of the most important big data technologies include Hadoop and its ecosystems (i.e. HBase, Hive, Pig, Mahout, Storm, and Spark) and distributed parallel processing framework such as MapReduce. Both technologies are widely used in big data applications to process clickstream data, social network data, geolocation stream data from sensors, and web log data. Other important big data technologies include NoSQL databases, in-memory computing, cloud computing, big data warehousing technologies, and data virtualization. These technologies are rapidly evolving and consolidating. Data science education should cover core principles and techniques because big data infrastructures are complex and need to be taught systematically to cope with any new evolving big data technologies.

As we outlined in the data analytics lifecycle section, Model Building includes the process of actually developing new algorithms, tools, and systems when existing techniques or tools cannot handle or when we need to improve effectiveness or efficiency of existing techniques or tools. In this regard, data scientists working in the Model Building mode should be able to write programmes using widely used programming languages (such as Java or Python) and statistical languages (such as R).

There are many good existing models/algorithms, but are they also applicable to big data in terms of the 4Vs: volume (scalability), velocity, variety, and veracity? Many of the models/algorithms need to be modified in order to meet the characteristics of the 4Vs; if they are not modifiable, we have to develop alternatives. All of these requirements need customized development when using a programming language in the Model Building mode. Challenges in Model Building include dealing with real-time stream data, large graph data, scalable machine learning algorithms, and making sense of visual analytics from big data. It is important to note that the role of machine learning is becoming bigger and bigger as the size of data gets larger and larger. Machine learning allows the paradigm of learning-by-data and provides effective ways of discovering knowledge from large data sets. Machine learning should be an important component of any data science education that has a focus in computing. We need an innovative way, within data science education, to effectively integrate machine learning with big data analytics to cope with the evolving challenges of big data era.

4.4. Incorporate research methods in data analysis

Advanced data science students should be trained for scientific thinking, reasoning, and analysis methods. Unlike undergraduate programmes, graduate programmes should emphasize the importance of research methods because data science is a multi-disciplinary discipline (and usually deals with more complex issues than traditional data analysis problems). From our survey on core courses among data science master's programmes, we found a lack of research method-related courses, except for the exploratory data analysis courses which contain research methods. Even though big data problems are discovery-based and learning-based by nature, students should learn how to come up with a research question, how to approach the question, and how to validate the outcomes. Students should be able to distinguish discovery-based research questions from traditional hypothesis-driven research questions. Research method-related courses can help students improve the ability to think critically, incorporate knowledge from various disciplines, solve problems using scientific methods, and evaluate the outcomes. It is recommended that research methods should be incorporated into data analysis courses.

4.5. Teach small data analytics as well

As data science has been popularized with the emergence of big data, most data science programmes focus on dealing with big data. We argue that data science programmes

should also teach how to deal with 'small data' – meaning those data that do not necessarily possess all the first 4Vs of big data but still have value. Hence, small data are not a concept that describes the volume but is a relative concept to big data. Similarly, by 'small data analytics', we mean data analytics that does not necessarily involve big data specific technologies (i.e. Hadoop and NoSQL), but involve general techniques (i.e. statistics, data mining, machine learning, and visualization). With small data analytics, it is simpler for students to practice research methods – create meaningful questions, assemble data, create hypotheses, build models, visualize the output, interpret results, validate them, and iterate until the outcome is satisfactory. Students will be able to learn scientific disciplines quickly with critical thinking ability.

Even though there is unimaginable value in big data, many companies have continuously innovated their ways of using small data – and are actually extracting big value from it (Ross *et al*., 2013). Thus, data science programmes should also teach how we can make better use of small data. Small data do not mean the data that are easy to process. On the other hand, small data can also be processed in the way big data are analyzed. Data science education does not necessarily need to emphasize 'big' and instead should focus on 'analysis', no matter what kind of data they are.

4.6. *Provide students with real-world project experience*

Data science addresses real-world problems by using real-world data. This means that a traditional way of education based on textbooks only is not appropriate for data science education. On the other hand, learning by doing and engaging in real-world problems (via projects or case studies) is a critical component of data science education. That is, students should have opportunities to tackle real-world problems in various ways and study in a learn-by-doing environment, where students can get experience on how data science addresses real-world problems using big data technologies – this is a mandatory component of data science education.

4.7. *Collaborate with multiple departments*

Data science is a multi-disciplinary study. It is very challenging for a single department/school to teach the whole spectrum of data science. The data in Table 1 also show that most data science programmes are offered at a university or at a joint-department level. For example, data science programmes can be offered jointly by department of computer science, department of statistics, or school of business. Jointly offering data science programmes is not the only way of collaboration; collaboration can also be achieved by sharing faculty resources. For example, even though a data science education programme is offered by a department of computer science, some courses could be taught by faculty from other departments/schools (e.g. exploratory data analysis by faculties from a department of statistics).

4.8. *Collaborate with industry/government*

Industry and government are good sources of real-world data and problems. These two sources are the most important components of data science education not only for problems and data but also for other resources such as computing resources, internships, training, educational programme support (e.g. certificates), and jobs for students. Industry/university cooperative research is an established model that moves education forward. Students can also get hired after graduation by the companies that they have worked with. Companies will be happy to recruit these students because the students have already been involved in the projects and know their business' problems well. Universities can also receive funding through their collaboration with industry/government to foster data science education.

4.9. *Actively use MOOCs*

Under the umbrella of data science, there are many high quality massive open online courses (MOOCs) available. Many universities provide data science and related courses in Coursera.[30] In addition, many universities are operating their own websites to offer MOOCs. Even though MOOCs cannot replace formal courses offered offline by universities, they can be used as supplementary courses. Many MOOCs are taught by well-known faculties; this means taking MOOCs is a good opportunity to learn about ideas and experiences from those people.

5. Conclusions

In this paper, we first investigated the background of big data, data science, data scientist, and the current state of data science education in the United States. We then proposed several approaches that data science education could take. The biggest bottleneck in the big data era is the production of capable data scientists. Tools and languages can be learned, but people who can manage real-world data science projects and who own the necessary big data analytic skills and knowledge are rare – producing such capable people takes time. Hence, we need multiple approaches within data science education.

We proposed various approaches for data science education based on an extensive survey of current data science education programmes as well as domain knowledge in the field. To summarize, data science education should try to (1) teach CDO disciplines; (2) teach with the eight-step data analytics lifecycle in mind; (3) teach big data technologies and model-building techniques; (4) incorporate research methods in data analysis; (5) teach big data analytics as well as small data analytics; (6) provide students with real-world project experience; (7) collaborate with

[30]https://www.coursera.org/

many departments; (8) collaborate with industry/government for data, projects, resources, and practicums; and (9) actively use MOOCs. We do not think that any single data science programme can adopt all of these approaches, and no data scientists could master all of these skills and knowledge that we have recommended. Each programme should focus on what they do best, and each data scientist should focus on what they do best. However, a certain level of broad coverage of topics is strongly recommended.

We note the emergence of big data analytics with the usage of automated tools such as IBM Watson Analytics. Using automated tools or dashboards that use a black-box approach would be an important solution in training data scientists. However, the users of those tools should still be familiar with the methods implemented in the systems to choose a right method that fits the given data set and to interpret the outcomes properly. Those users should have critical thinking and reasoning ability to explore the solution space provided in the tool and to determine whether the tool can indeed provide a satisfactory outcome.

We believe that data science will get more and more attention, and it is time to do more research on how data science should be taught. Data science educators should keep working on how to re-train people, process, and technologies around big data to change for the better. As we get more experience, we will be able to improve data science programmes and educate successful data scientists.

References

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth (2000) CRISP-DM 1.0 step-by-step data mining guide, SPSS Inc.

Davenport, T.H. and D.J. Patil (2012) Data scientist: the sexiest job of the 21st century, at https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/. (Accessed on 5 August 2014)

Gartner (2012a) Big data, at http://www.gartner.com/technology/topics/big-data.jsp. (Accessed on 5 August 2014).

Gartner (2012b) Big data, at http://www.gartner.com/it-glossary/big-data/. (Accessed on 5 August 2014)

IBM (2014) Leadership and innovation, at http://www-935.ibm.com/services/c-suite/cdo/ (Accessed on 5 August 2014).

MarketsandMarkets (2013) Big data market worth $46.34 billion by 2018, at http://www.marketsandmarkets.com/PressReleases/big-data.asp. (Accessed on 5 August 2014).

Mayer-Schönberger, V. and K. Cukier (2013) Big data: a revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt.

McKinsey (2011) Big data: the next frontier for innovation, competition, and productivity, at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. (Accessed on 5 December 2014).

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

Miller R. (2014) If you think big data's big now, just wait, at http://techcrunch.com/2014/08/10/big-data-bound-to-get-really-really-big-with-the-internet-of-things/. (Accessed on 5 August 2014).

Ross, J. W., C. M. Beath and A. Quaadgras (2013) You may not need big data after all. *Harvard Business Review*, **91**, 90–98.

Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy, *Journal of Information Science*, **33**, 163–180.

Stanton, J. (2012) An introduction to data science, at http://ischool.syr.edu/media/documents/2012/3/datasciencebook1_1.pdf. (Accessed on 5 August 2014)

Dhar, V. 2013 Data science and prediction. *CACM*, **56**(December 2013), 64–73.

Zins, C. (2007) Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, **58**, 479–493.

The authors

Il-Yeol Song

Il-Yeol Song is professor in the College of Computing and Informatics at Drexel University and Director of the PhD Program in Information Studies in his college. He is also an affiliated professor of the Computer Science Department at KAIST, Korea. He is an ACM Distinguished Scientist and an ER Fellow. He has received the Peter P. Chen Award in Conceptual Modeling in 2015. His research interests include conceptual modeling, data warehousing, big data management, CRM, and smart health. Dr. Song published over 200 peer-reviewed papers and co-edited 22 proceedings. He is a co-Editor-in-Chief of Journal of Computing Science and Engineering (JCSE) and is an editorial board member of DKE, JDM, IJEBR, and JDFSL. He won the Best Paper Award in the IEEE CIBCB 2004. In addition, he had also won 14 research awards from competitions like the annual Drexel Research Days. He also won four teaching awards from Drexel, including the most prestigious Lindback Distinguished Teaching Award. Dr. Song served as the Steering Committee chair of the ER conference between 2010 and 2012. He is also a steering committee member of DOLAP, BigComp, and ADFSL conferences. He has served as a program/general chair of over 20 international conferences/workshops including DOLAP'98-'15, CIKM'99, ER'03, DaWaK'07-'08, DESRIST'09, CIKM'09, and MoBiD'13-'15.

Yongjun Zhu

Yongjun Zhu is a PhD candidate in the College of Computing & Informatics at Drexel University. His research interests are in the general area of data science. In particular, he is interested in information retrieval, text mining, and network science. He has authored/co-authored papers published in journals such as Decision Support Systems, Scientometrics, and Journal of Informetrics.