Scientific Research Publishing

Data Science and Informetrics, 2020, 1, 22-49
https://www.scirp.org/journal/dsi
ISSN Online: 2694-6106
ISSN Print: 2694-6114

Data Science: State of the Art and Trends

Lemen Chao[1,2], Chunxiao Xing[3,4,5], Yong Zhang[3,4,5], Chen Zhang[1,2]

[1]Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China
[2]School of Information Resource Management, Renmin University of China, Beijing, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]Reasearch Institute of Information Technology, Tsinghua University, Beijing, China
[5]Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, China

Email: chaolemen@ruc.edu.cn

How to cite this paper: Chao, L.M., Xing, C.X., Zhang, Y. and Zhang, C. (2020) Data Science: State of the Art and Trends. *Data Science and Informetrics*, **1**, 22-49.
https://doi.org/10.4236/dsi.2020.11002

Received: September 23, 2020
Accepted: October 20, 2020
Published: October 23, 2020

Abstract

The entering into big data era gives rise to a novel discipline called Data Science. To start with, a very brief history, interdisciplinarity, theoretical framework, and taxonomy of Data Science are discussed. Then, the differences between domain-general Data Science and domain-specific Data Science are proposed based upon conducting literature reviews on hot topics in big data-related studies. In addition, ten common debates in Data Science are described, including debates on thinking pattern, properties of big data, enablers of intelligence, bottlenecks in data products development, data preparation, quality of services, big data analysis, evaluation of big data algorithms, the fourth paradigm and big data skills shortage. Moreover, the emerging trends in Data Science are presented: shifts in data analysis methodologies, adoption of model integration and meta-analysis, introducing data first, schema later or never paradigm, rethinking data consistency in big data systems, recognizing data replication and data locality, growth in integrated data applications, changes in the complexity of data computing, the advent of data products, the rise of pro-ams and citizen data science, as well as the increasing demand for data scientists. In conclusion, some suggestions for further studies are also proposed: to avoid misconstruing Data Science, to take advantages of active property of big data, to balance the three dimensions of Data Science, to introduce Design of Experiments, to embrace causality analysis, and to develop data products.

Keywords

Data Science, Big Data, Data Products, Data Wrangling, Data-Driven

1. Introduction

Big data is revolutionizing the ways how we live, work, and think [1], and lead to

provocations for cultural, technological, and scholarly phenomena [2]. As a consequence, shifts in epistemologies and paradigm occur in a wide range of disciplines [3]. Data-centered thinking started to be an alternative paradigm for data-related tasks, which is different from the Knowledge-centered thinking in traditional research. It is a significant change in modern science to take advantage of data-centered thinking. Furthermore, the acquisition, storage and computing of data are no longer the biggest bottlenecks so that the contradiction between traditional knowledge and big data is increasingly prominent in various disciplines. Traditional theories fail to deal with big data problems; thus, big data is highly concerned by experts from various disciplines. How to employ big data is one of the hot topics for most disciplines from Computer Science to Statistics. As a result, the research on big data from different professional fields began to converge on an emerging discipline called Data Science (DS). However, as Big Data inexorably draws attention from every segment of society, it has also suffered from many characterizations that are incorrect: size is all that matters; the central challenge with Big Data is that of devising new computing architectures and algorithms; analytics is the central problem with Big Data; data reuse is low hanging fruit; Data Science is the same as Big Data; and Big Data is all hype [4]. Therefore, Data Science requires to conduct in-depth research on the new phenomena, ideas, theories, methods, techniques, tools and practices of big data.

The rest of this paper is structured as follows: Section 2 describes a brief history, interdisciplinarity, its theoretical framework as well as taxonomy of Data Science (DS), and categories the existing studies into two basic types: domain-general DS and domain-specific DS. Then, Section 3 proposes ten main debates in DS-related research, including the shifts in thinking pattern, properties of big data, enablers of intelligence, bottlenecks in data products development, data preparation, quality of services, big data analysis, evaluation of big data algorithms, the fourth paradigm and big data skills shortage. Ten emerging trends in Data Science studies are provided in Section 4: shifts in data analysis methodologies, adoption of model integration and meta-analysis, introducing data first, schema later or never paradigm, rethinking data consistency in big data systems, recognizing data replication and data locality, growth in integrated data applications, changes to the complexity of data computing, the advent of data products, the rise of pro-ams and citizen data science, as well as the increasing demand for data scientist. Finally, we summarize the study and put forward some suggestions for data science researchers.

2. Data Science: The Science of Big Data

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 [5]. It also provides new theories, methods, models, technologies, platforms, tools, applications and best practices of big data. And one of the goals of Data Science research is to reveal the new challenges and opportunities brought by big data.

2.1. A Brief History of Data Science

Peter Naur, a Turing Award winner, coined the term of Data Science his book entitled *Concise Survey of Computer Methods* in 1974. He defined data science as the science of dealing with data, and further proposed that it is different from Datalogy which is the science of data and of data processes and its place in education [6]. In 2001, William S. Cleveland published the paper, D*ata Science: An Action Plan for Expanding the Technical Areas of the Field of Statistic*s, proposed that Data Science is an emerging branch of Statistics [7]. In 2013, Nature published the article, Computing: A Vision for Data Science [8], and *Communications of the ACM published* the paper, Data Science and Prediction [9]. Both of those two articles discussed the Data Science from the perspective of Computer Science. Then, Data Science was also identified as a branch of Computer Science. Data Science has begun to get much more public attention since 2010s. Patil DJ and Davenport T H published the article entitled Data Scientist: The Sexiest Job of the 21st Century in *Harvard Business Review* in 2012 [10]. Barack Obama won the presidency by implementing and using big data strategies in the 2012 US presidential election [11]. The White House announced Patil DJ the first U.S. Chief Data Scientist in 2015 [12].

According to Gartner's 2014 Hype Cycle for Emerging Technologies [13], Data Science was on a one-way trip to the "peak of inflated expectations" and would enter "plateau of productivity" in 2 - 5 years. Gartner's 2016 Hype Cycle for Data Science (**Figure 1**) [14] is a growth curve that shows the breadth and depth of excitement about data science, with new technologies and some significant movements from last year. Gartner's 2016 Hype Cycle for Data Science shows: R entered the plateau of productivity; Simulation, Ensemble Learning, Video/Image Analytics and Text Analysis were climbing the slope of enlightenment; Hadoop-Based Data Discovery was obsolete before plateau; Speech

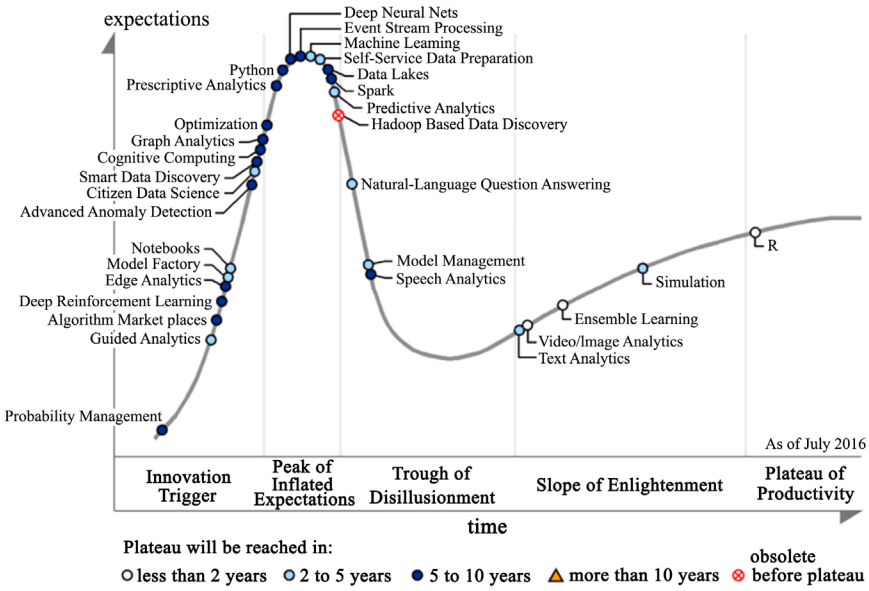

Figure 1. Gartner's 2016 hype cycle for data science.

Analytics, Model Management and Natural-Language Question Answering have passed the peak of inflated expectations and sliding into the trough of disillusionment; Citizen Data Science, Model Factory, Algorithm Marketplaces and Prescriptive Analytics were developing rapidly.

2.2. Interdisciplinarity of Data Science

In 2010, Drew Conway came up with and shared his Data Science Venn Diagram (**Figure 2**) to reveal the interdisciplinary of Data Science [15]. The Venn Diagram shows that Data Science is a combination of hacking skills, math & statistics knowledge, and substantive expertise. Now, there are many variations of his Venn diagram such as Jerry Overton's Data Science Venn Diagram (**Figure 3**) [16], but all of them are less influential than Drew Conway's Venn Diagram.

Drew Conway's Venn Diagram shows that Data Science sits at the intersection of machine learning, traditional research, and danger zone. From the periphery of the graph, it can be seen that not only Math & Statistics and substantive expertise are necessary for Data Science, but also hacking skills are involved in it. In other words, Data Science has three basic components: theories (Math & Statistics), practices (Substantive Expertise) and skills (Hacking Skills).

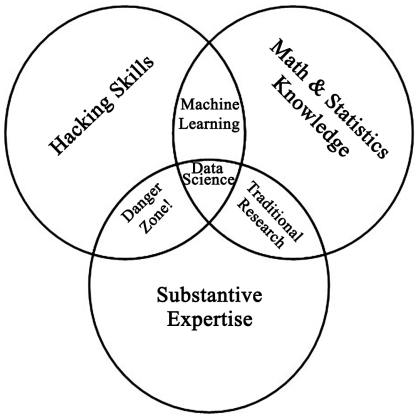

Figure 2. Venn diagram by drew Conway (2010).

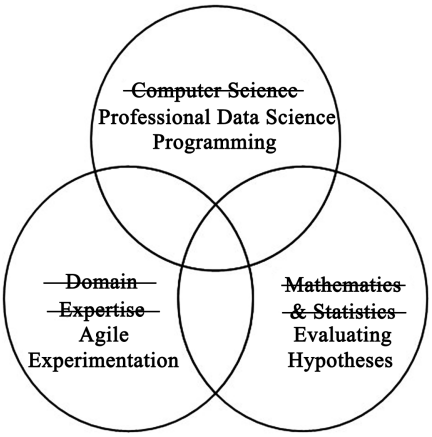

Figure 3. Venn diagram by jerry Overton (2016).

2.3. Theoretical Framework of Data Science

Data Science is a discipline based on Statistics, Machine Learning, Data Visualization and Domain Knowledge, and its research contents include: basic theory of data science, data wrangling, data computing, data management, data analysis and data product development. **Figure 4** shows the theoretical framework of Data Science [17].

Basic theories of Data Science include new concepts, theories, methods, technologies and tools appeared in Data Science, as well as research purpose, basic processes, main principles, typical applications, talent cultivation, data project management. It is worth noting that basic theories and theoretical bases are two different terms. The basic theories are within the research scope of Data Science, while the theoretical bases are outside the scope.

Data Wrangling (or Data Munging) is the initial process of transforming raw data into another form for the principal purpose of improving data quality. Data Wrangling involves the following operations: data auditing, data cleansing, data conversion, data integration, data masking, data reduction, data annotation. It also concerns that how to apply data scientists' skills of creative design, critical thinking and curious questioning to the data wrangling activities.

Cloud computing is a distributed computing paradigm and differs from traditional ones such as centralized computing and grid computing. There are some representative technologies or products of cloud computing: Google GFS, Google BigTable, Google MapReduce, Hadoop MapReduce, Spark and YARN.

Data Management is the practice of organizing and maintaining data processes to meet ongoing big data lifecycle needs [18]. Data scientists have to harness not only traditional relational database, but also some emerging technologies such as NoSQL, NewSQL and relational cloud for data management.

Traditional data analysis tools are not enough to manage big data, so some open source tools for big data analysis are indispensible to Data Science. The most popular open source tools for big data analysis are R and Python.

Data Product is a product that facilitates an end goal through the use of data [19]. Data product development is indispensible for Data Science. Data product development activities are rarely undertaken in a traditional product development sequence that involves identifying the need, developing the product. On

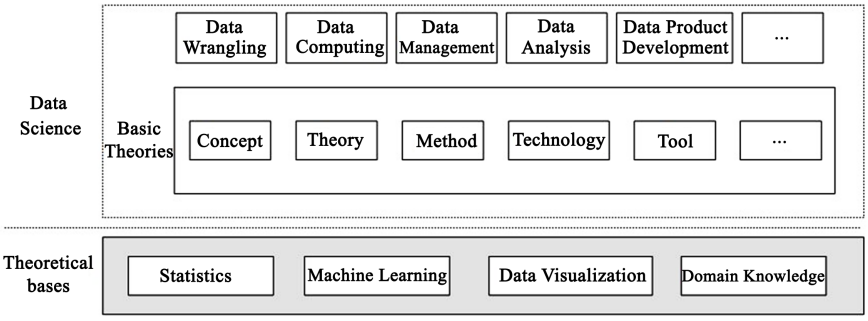

Figure 4. The theoretical framework of data science.

the contrary, data product development activities often take place in a continuous, iterative fashion, with the important activities conducted in parallel. [20] And the ability to develop data products is becoming increasingly critical to every business in big data era. Therefore, one of the missions of Data Science project is to develop data products.

2.4. Taxonomy of Data Science

Data Science is an emerging discipline that incorporates theories with domain knowledge and business practice. There are two types of Data Science: domain-general Data Science and domain-specific Data Science. Domain-general Data Science is an independent discipline, while the domain-specific Data Science refers to the big data research that depends on a specific domain. Some research topics in domain-specific Data Science include Data Journalism [21], Materials Data Science [22], Big Data Finance [23], Big Data Society, Big Data Ethics [24], and Big Data Education [25].

Domain-general Data Science is a theoretical foundation for Domain-general Data Science. Specifically, domain-general Data Science involves the general ideas, theories, methods, concepts and tools of Data Science, domain-specific Data Science focuses on the applications of data science in a specific domain.

3. Research Progress in Data Science

The existing topics on Data Science can be categorized into two types: the core issues and the periphery issues. The core issues discuss the basic questions of Data Science, including its unique principles, theories, methods, techniques, tools, applications and best practices. At the same time, the periphery issues are mainly focused on the relevant topics, such as theoretical basis, applications, and related research areas of Data Science.

3.1. Hot topics in Data Science

The most discussed topics in the relevant literature have mainly concentrated on the periphery issues instead of the core ones. There are five typical hot topics in Data Science:

1) Big data challenges and the introduction to Data Science. There are not only challenges, but also opportunities with big data [26]. The characteristics of Big Data are commonly referred to as the four V's: Volume, Variety, Velocity, and Value [27] [28]. Four V's are critical to identifying big data challenges. Companies across most industries have realized that they need to hire more data scientists. Academic institutions are scrambling to put together programs to train data scientists. Publications are touting data science as a hot career choice [10] [29].

2) The impact of data science on statistics [7] and computer science [8]. Data science was first born in Statistics and Computer Science. While it has inherited some of their methodologies, it also seeks to blend, refocus, and develop them to address the needs of modern scientific data analysis [15] [30].

3) New technologies for Data Science. New technologies, such as cloud computing, Internet of Things, mobile computing, are employed to improve data scientists' efficiency at data acquisition, data storage and data computing. Some new technologies allow professionals to perform big data analytics, even if they don't have a background in programming. Some new tools such as Spark [31], Hadoop [32], and NoSQL [33] are employed by data scientists to solve data intensive problems. Data First - Model Later or Never paradigm, CAP theorem, BASE and ACID [34] are becoming underlying principles of Data Science.

4) The impact of Data Science on specific domains. Applying Data Science to other specific domains is one of the popular topics in recent studies. Those specific domains include life science [35], medical care [36], government governance [37], education [38], and business management [39]. As a result, some new research topics such as quantitative self [40], data journalism [41] and big data analysis [42] gained widely attention of data scientists.

5) Student programs in Data Science. One of the principal purposes of student programs in Data Science is to enhance students' ability to think structurally about data [43]. There are four related trending topics: building data science curriculum, reforming the teaching program [44], cultivating interdisciplinary talents [45] and training female data scientists [46]. Students are educated as data engineers or data scientists who with the three skills (3C's): Creative design, Critical thinking, and Curious questioning [47].

3.2. Domain-General Data Science

From the perspective of domain-general Data Science, Data Science is a novel independent discipline. However, Data Science is also a relatively emerging branch of many traditional disciplines from the perspective of domain-specific Data Science. There are five hot topics in domain-general Data Science research.

1) DIKW model. The DIKW model reflects that the growth of the Internet is changing the traditional hierarchies of "experts" and changing ways of disseminating information [48]. The research task for Data Science is to turn raw data into useful information, and then into knowledge, and finally wisdom [49]. The transformation from data to wisdom is a value-added process, which involved in the transformation of an entity at a lower level in the hierarchy to the one at a higher level in the hierarchy. Its implicit assumption is that data can be used to create information; information can be used to create knowledge, and knowledge can be used to create wisdom [50].

2) Data Analytics. Data analytics is the science of collecting, storing, extracting, cleansing, transforming, aggregating and analyzing data, with the purpose of discovering information and knowledge [51]. Data Analysis and Data Analytics are two different terms. The former emphasizes the activity of Data Analysis itself, while the latter lays more emphasis on the methods, techniques and tools used in the activities of Data Analysis. Development of algorithms and tools for big data analysis is one of the popular topics. Domain-oriented analysis is also

discussed in different domains, such as logistics and supply chain management [52], network security [53], and health care [36]. Jeffery T. Leek, Roger D. Peng (2015) discussed the main types and common errors in big data analysis [53].

3) Datafication. Datafication is a new paradigm in science and society [55]. It is the transformation of social action into online quantified data, thus allowing for real-time tracking and predictive analysis [1]. Along with the Internet of Things and Sensors, Quantified Self [40] [56] is also a hot topic of datafication. The conventional research focus shifts from the datafication of business to the businessization of data with the development of big data. The businessization of data is meant that data is taken advantage to optimize existing business or define new business.

4) Data Governance. Data governance is the process by which a company manages the quantity, consistency, usability, security and availability of data [57]. Current researches mainly focus on data governance design, implementation methods [58], reference framework [59] and platform ecosystems [60]. In addition, as a key part of Data Maturity Model, data governance includes three processes: Governance Management, Business Glossary, and Metadata Management [61].

5) Data Quality. The internal nexus between data quality and data availability is also a hot topic for Data Science. Different types of big data have emerged so that studies that were difficult to conduct in the past time due to data availability can now be carried out. How to mitigate or rectify big-errors brought by big data [62] and how to improve the low quality and usability of big data are most challenging problems that need to be addressed [63]. Traditional data management mainly focuses on the quality of data source, whether the data source contains Clean Data or Dirty Data [64]; Data Science mainly focuses on the quality of data form, whether the data is Tidy Data or Messy Data. Hadley Wickham (2014) put forward the concepts of Tidy Data as well as Data Tidying, and proposed that Tidy Data should follow three basic principles: each variable must have its own column, each observation must have its own row, each value must have its own cell [65].

Apart from the above topics, big data security [66], data privacy protection [67], project management and building Data Science teams [68], citizen data science [69] are also frequently discussed in domain-general Data Science.

3.3. Domain-Specific Data Science

Researchers from different disciplines have shown their own distinct concerns and perspectives on Data Science. The new term of Data Science and its variant concepts are widely used in domain-specific Data Science. There are nine hot topics in domain- specific Data Science literature.

1) Data Journalism. As one of the new areas of Journalism, Data Journalism is a way of seeing journalism as interpolated through the conceptual and methodological approaches of computation and quantification in the era of big data

[41] [70].

2) Industrial Big Data. It mainly studies how to apply big data to the practices of Industrial Manufacturing. The representative application cases are Germany's Industry 4.0, American Industrial Internet and Made in China 2025.

3) Consumption Big Data. Consumption Big Data was used to promote more products to consumers through precision marketing [71], user profiling [72] and advertising push [73].

4) Health Big Data. It mainly focuses on the wide application of big data in health and medical fields including life logging [74], medical diagnosis, pharmaceutical production, and health care [36].

5) Biological Big Data. Harnessing powerful computers and numerous tools for data analysis is crucial in drug discovery and other areas of big-data biology [35]. The principles, theories, methods, technologies and tools of big data are widely adopted to biology, and biological research paradigm is transferring from knowledge-centered paradigm to data-centered paradigm.

6) Social Big Data. Social big data comes from joining the efforts of the two previous domains: social media and big data. Applications of social big data can be extended to a wide number of domains such as health and political trending and forecasting, hobbies, e-business, cyber-crime, counterterrorism, time-evolving opinion mining, social network analysis, and human-machine interactions. [75]

7) Big Data in Organizations. Big data can be used in enterprises [76], governments [77] and public welfare departments [78] to improve effectiveness and efficiency of these organizations.

8) Smart applications. Big data also play an important role in smart applications such as smart cities, smart medical care, smart elderly care, smart transportation and smart education.

9) Agile Big Data. Agile Big Data is a development methodology that copes with the unpredictable realities of creating analytics applications from data at scale. [79] It is helpful to develop agile software, manage agile projects and establish agile organizations.

3.4. Big Data Ecosystem

A big data ecosystem is usually defined as a system of different components that allow the storage, processing, preparation, visualization and delivery of useful information to target applications or end-users [80]. For example, Big Data Landscape [81] shows the major institutions and products in a Big Data ecosystem. Existing relevant literature mainly discuss the components of big data ecosystem and their interrelationships. There are five hot topics:

1) Infrastructures. Infrastructures in big data ecosystem include Cloud Computing, Internet of Things, mobile computing and social media. The relevant literature mainly focuses on the impact of those infrastructures on Data Science and how to make full use of them in Data Science.

2) Supporting Technologies. Researches mainly discuss the application of supporting technologies in Data Science, such as machine learning, statistics,

batch processing, flow computing, graph computing, interactive computing, NoSQL, NewSQL and cloud SQL.

3) Tools and Platforms: There are some popular tools or platforms that include R, Python, Hadoop, Spark, MongoDB, HBase, Memcached, MongoDB, CouchDB and Redis.

4) Project Management: It involves the management of data science project's scope, duration, cost, quality, risk, human resources, communication, procurement and system.

5) External Environment. Laws, policies, systems, cultures, morals and ethics need to be updated with the advent of big data era. New legislation on data ownership should take into account the public interest in maintaining competition in the market [82]. Big data is becoming a kind of asset, and legislation on data ownership is a necessary condition for data utilization.

4. Contemporary Debates in Data Science

Most disciplines have identified big data challenges over the last few years. Academics journals such as Nature and Science have published special issues dedicated to discuss the opportunities and challenges brought by big data [83]. From the perspective of Computer Science, the volume of data just beyond technology's capability to store, manage and process efficiently [27]. From the perspective of Statistics, the first step of statistical analysis is to determine whether the data dealing with is a population or a sample when sample proportion is close to population [1]. From the perspective of Machine Learning, current intelligent machine-learning systems are not inherently efficient enough which ends up, in many cases, a growing fraction of big data unexplored and underexploited [84]. From the perspective of Data Analysis, the challenge is to gain data insight and realize the transformation from data to wisdom through analyzing data [85]. There are ten debates and challenges faced by Data Science research.

4.1. Thinking Pattern: Knowledge-Centered Thinking or Data-Centered Thinking

Traditional scientific studies have to adopt knowledge-centered thinking due to its limited capability of collecting, storing and computing data. It is the basic feature of knowledge-centered thinking that knowledge is enablers of practical solutions, and utilization of data implemented via extraction knowledge from data. However, a shift in thinking pattern occurs in big data era, and peoples manage to solve problems by using data directly. Data-centered thinking, by contrast, prefers to take advantage of data without extracting knowledge from it. Traditional machine translation methods, for instance, are in line with knowledge-centered thinking since those are based on theories called Natural Language Understanding, which is built on the top of linguistic and statistical knowledge. As a result, knowledge-centered thinking became the bottleneck of traditional machine translation, and traditional machine translation fail to make breakthrough in its theoretical system. Recent machine translation studies

change their thinking pattern and turn to an alternative thinking pattern called data-centered thinking. Data-centered thinking pattern for machine translation is implemented not by traditional knowledge but by data such as parallel or multilingual corpora, and other large-scale domain data.

IBM machine translation in 1950s and Google Translate in 2000s are two typical thinking patterns for knowledge-centered thinking and data-centered thinking, respectively. It is one of the common beliefs in traditional thinking pattern that knowledge is power. However, we have to admit that data is also an alternative power in the era of data-enriched offerings. The capability to taking advantages of big data is becoming one of the core competitiveness of modern organizations. The main challenges to adopt data-centered thinking patterns are to implement data-centered design, to make data-driven decisions [86] and to develop data-intensive applications [87].

4.2. Property of Big Data: Passive or Active

Conventional thinking pattern used to regarding data as a passive being and focuses on taking advantages of human initiative to deal with data problems. The schema of data, for instance, should be defined prior to writing their content into a relational database (RDB), and its main functions focus on processing, miming, and analyzing data. In other words, it is the underlying beliefs of RDB that the property of data is a passive rather than active. A critical change regarding the property of data occurs in big data era, which is people begin to realize that data is more active than passive. Furthermore, data is no longer regarded as a dead and passive thing, and the related studies are paying more attention to active roles of data. As a result, a variety of novel terms are coined, including data-driven decision support, data business, data insights, data intensive application, and data first, data locality, schema later or never approaches. Those emerging terms share the same beliefs in the property of data, which data is more active than passives in current business scenarios.

Therefore, the ultimate research aim of Data Science is to change the traditional perception and cognition about data, and to exploit the active property of data.

4.3. Enabler of Intelligence: AI-Based Intelligence or Big Data-Based Intelligence

The implementation of intelligence mainly relies on algorithms in past studies, especially complex algorithms. The more complex the algorithm, the higher degree of intelligence. For example, KNN is a commonly used classification algorithm in Machine Learning, and its underlying principle is quite simple. Over the years, variants of KNN have been proposed in order to address various applications. However, the complexity of KNN is increased, as the levels of intelligence are improved [88]. The data-centered thinking paradigm implies that data can also be used directly to solve problems, and further leads to a debate on

More Data or Better Model [89]. The debate is ended in a conclusion that the best model can be implemented by more data and simple algorithm. Therefore, how to develop simple and efficient algorithms is one of the main objectives of Data Science (DS). The challenges of DS studies root in how to design novel algorithms, integrate multiple models, solve the curse of dimensionality, and introduce deep learning to big data applications.

4.4. Bottlenecks in Data Products Development: Compute-Intensive or Data intensive

Software development and algorithm design was mainly responsible for solving compute-intensive problems. Computing is the bottleneck in the design and development of traditional products. However, with the advent of large-scale distributed computing technologies, especially cloud computing, computing is no longer the primary bottleneck of product design and development. As a result, the main challenges in software development and algorithm design shifted from computation to data, and data-intensive tasks are the next bottleneck. Data is the biggest bottleneck of Data intensive applications [90], and the research on data-intensive problems will be indispensable for the data-centered research paradigm. The main topics of data-intensive applications concentrate on data replication, materialized views, CAP theorem, BASE principles and Lambda architecture.

4.5. Data Preparation: Clean Data or Tidy Data

Data preparation in traditional applications devoted to data cleansing and avoiding Garbage in and Garbage out (GIGO). Data cleansing involves filtering duplicate data, identifying incorrect data, and processing missing values. It can be seen that data cleansing mainly focuses on data quality. However, big data processing is different from the small-scale data preprocessing in that the former has high robustness with low data quality. Big data preprocessing values data format rather than data quality, and tends to distinct types of data preparation called data wrangling or data munging. Data wrangling/data munging is the creative and value-added process of data preparation. In contrast to data cleansing, data wrangling/data munging is more concerned on how to integrate the knowledge or wisdom of data scientists into data processing so that enhance the value of data. Therefore, data wrangling or munging not only requires technical qualifications, but also involves creative value creation activities, particularly data jujitsu and data tidying.

4.6. Quality of Services: Performance of Services or User Experiences

Recall ratio and precision ratio are two widely used criteria for evaluating service quality in traditional data applications. However, it is difficult to calculate exact recall rate and precision rate where the population is unknown, the amount of data increases rapidly, the types of data are constantly changing, or the require-

ment for data processing speed is high. Therefore, big data applications concern more on user experience rather than recall rate as well as precision rate. Response time or response delay is one of the most prominent criteria to influence the quality of user experience. Aberdeen Group's research found that a 1-second delay in page load time equals 11% fewer page views, a 16% decrease in customer satisfaction, and 7% loss in conversions ; Google found an extra 0.5 seconds in search page generation time dropped traffic by 20%; Amazon also found that every 100ms of latency cost them 1% in sales [91]. The challenges of enhancing user experience involve to reduce response delay, to design human-computer interaction interface, to implement service virtualization, and to provide on-demand services.

4.7. Big Data Analysis: Causality or Correlation

Knowledge-centered thinking pattern believes that data will be utilized effectively only when the causality in data is identified. Hence, data analysis in the past dedicates to find, validate and taking advantage of causalities. That conventional thinking pattern is very effective in small-scale data sets but inefficient in big data sets since it is hard to identify and validate a causality from large-scale data sets. As a result, the aims of data analysis have shifted from causal analysis to correlation analysis, which put more emphasis on the correlation analysis. In contrast with causality analysis, correlation analysis is time-saving and easy to put into practices. This separation of causality analytics and correlation analytics also triggers collaboration between data scientists and domain experts. The challenges of big data analysis often stem from data complexity, noises data, and data dependence [92]. Proposing new methods, technologies and tools for big data analysis, especially enabling the dynamic evolution of data analysis methods, real-time computing and resilient computing, is critical to big data analytics.

4.8. Evaluation of Big Data Algorithms: Complexity or Scalability

Complexity, especially time complexity and space complexity, were two common evaluation criteria of traditional algorithms [93]. However, the introduction to big data technologies, such as Spark, Hadoop, and New SQL, provide a new solution for overcoming those complexity challenges. As a result, algorithms take into consideration providing real-time analyzing, on-demand services and supporting data-driven applications. For instance, Google launched Google Flu Trends (GFT) in 2008, which in real time predicted the nationwide spread of H1N1 [94], but then it was estimated that the number of infected people was twice as high as the actual number by January 2013. The decrease in precision of GFT mainly stems from big data hubris, algorithm dynamics, as well as transparency, granularity, and all-data [95]. In the era of big data, the scalability of algorithms mainly reflects the ability of a system to handle increasing data or dynamic access request loads. Research challenges are applying low-dimensional algorithms in high-dimensional data, data reduction, and working with da-

ta-intensive applications.

4.9. Research Paradigm: Conventional Paradigm or the Fourth Paradigm

Jim Gray proposed scientific research has experienced four different paradigms: experimental science, theoretical science, computational science, and data-intensive science. Data-intensive science, also called the fourth paradigm, is that scientific researchers can find and mine the needed information and knowledge from big data without directly facing the physical objects they are studying. For example, big data has changed the way that astronomers do their research. Their main task is to find pictures of objects or phenomena from vast databases, rather than having to take pictures of themselves in space [96]. In most studies, researchers often directly access the data which they need, and they can also achieve the purpose that understanding physical world via its historical records. The historical records in many fields are already enough to support researchers to conduct a scientific research with the advent of data-enriched offerings. Data scientists do not have to collect or investigate data in the physical world by means of conventional ways such as questionnaire and interview. Historical data are more objective and credible, compared with investigative data. The challenges in related studies are to distinguish the Third Paradigm from the Fourth Paradigm, to put the Fourth Paradigm into practices, to conduct in-depth theoretical studies and to apply it to more specific domains.

4.10. Big Data Skills Shortage: Data Engineer or Data Scientist

The goal of traditional education for data professionals is to train data engineers who are able to design, build, secure, and monitor, backup and recovery of data processing systems. However, those data engineers are not qualified for the challenges of Data Science (DS), and the DS has to find alternative data professionals called Data Scientists. Data Engineers and Data Scientists differ in their job responsibilities: the former is responsible for data management, while the latter is good at data-based management, such as data-based decision-making, data product development, and business definition. The challenges of theoretical research about data scientists are to exactly define their job responsibilities, to improve their unique capabilities, to manage data science projects and to conduct their career development planning.

5. Emerging Trends in Data Science

There are 10 trends in Data Science (DS) research, and they share four common characteristics as follows:

1) Shifts in thinking paradigm will be the underlying trend. The Shifts in thinking paradigm refers to the transfer of methodologies from knowledge-centered thinking to data-centered thinking discussed in Section 4. The introduction to data-centered thinking will further accelerate the adoption of

Data-intensive Scientific Discovery Paradigm. Dualistic cognition of the world will be replaced by ternary cognition, and the aim of relevant research is to understand the physical world via studying the data world. Diversification of thinking patterns and shift in research paradigms has a profound influence on the research of DS, which will change the motivations, methodologies and objectives of data projects.

2) Domain-specific Data Science will be the hot topic. One of the main purposes of Data Science (DS) is to bridge the gap between big data and traditional knowledge. There are revolutionary changes to the volume, variety, velocity, and values of data, but the knowledge has not been updated to keep up with them. Consequently, traditional knowledge fails to deal with big data problems. Taking advantages of big data is the crucial research questions in most disciplines, and domain-specific DS will be the popular topics in the future.

3) Domain-general Data Science will be the research bottleneck. Domain-specific Data Science is heavily depended on domain knowledge, and its studies limited only within the specific disciplines. The differences between Domain-general Data Science and Domain-specific Data Science lie in their basic theories, methods, techniques, tools, and typical practices. The bottleneck of DS studies is to build Domain-general Data Science, to answer the shared question from different disciplines, and to provide a new theoretical basis for the research in various disciplines.

4) Nurturing a data ecosystem is the ultimate purpose. Data Science (DS) is a practical discipline, and the research should not be isolated from its application in the specific domains. Along with technological issues, the DS studies also involve development strategies, infrastructure, human resources, ethics, policies, laws of big data. Therefore, the ultimate purpose of DS practices is to utilize big data for nurturing a new ecosystem. Putting DS researches into the big data ecosystem needs a holistic approach for data problems and further promote the mutual transformation between data, energy, and materials.

5.1. Shifts in Data Analysis Methodologies

In this data-intensive world, predictive models are more important than ever in order to make sense out of what is around us and to estimate, assess, or plan for what might happen in the future [97]. Predictive modeling rather than interpretive modeling (which are used for attribution) is an import part of data science projects. Predictive modeling is forward-looking, and constructed for predicting new observations. In contrast, explanatory modeling is retrospective and used to test an already existing set of hypotheses [98]. Besides, Predictive modeling follows Data Paradigm, while explanatory modeling follows Knowledge Paradigms.

One of the reasons why data scientists should emphasize predictive modeling is that "patterns often emerge before the reasons for them are apparent." [99] Predictive modeling is based on Hypothetico-Deductive Method [100]. A hypo-

thetico-deductive method works in two steps: in the first step, hypotheses (models) are proposed and in the second step the consequences of the hypotheses (models) are deduced using the available facts [100]. A good research hypothesis requires data scientists to have creativity, critical thinking and curiosity.

Complex models often do a bad job of predicting new data, though they can be made to fit the old data quite well [101]. The prediction model should be simpler than the interpretation model for two reasons. The first reason is that prediction model is required operating in real-time in some cases and the simpler model take less computing time. The second reason is that the simplest solution is almost always the best (Occam's Razor [102]).

Correlation is simply a relationship between events, causation indicates that one event is the result of the occurrence of the other events. The concept of correlation underlies a vast number of techniques for predictive modeling, statistical analysis, and other data mining [29].

5.2. Adoption of Model Integration and Meta-Analysis

Single model is often used to in traditional data analysis. The proliferation of new big data technologies in recent years and requirement to improve accuracy of data analysis have made data model increasingly complex. In other words, the model used in traditional data analysis has two characteristics: single and complex.

However, it is difficult to find a single theoretical model can fully apply to dynamic and heterogeneous data. So people try to integrate multiple and simple models to address this issue. In other words, the model in big data analysis has two characteristics: multiple and simple.

Adoption of model integration is a new trend for Data Science research. Data scientists usually split a data analysis project into multiple small tasks and perform them with model integration. For example, deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [103].

The use of data model will become more and more important as the need grows for data analysis. Meta-analysis is as important as model integration. Primary analysis is the original analysis of raw data in a research study. It includes descriptive statistics, parameter estimation and hypothesis testing. Secondary analysis is the re-analysis of data for the purpose of answering the original question with better statistical techniques, or answering new questions with old data. While meta-analysis refers to the analysis of analyses, it is used to analyze a large collection of analysis results from individual studies for the purpose of integrating the findings [104].

5.3. Introducing Data First, Schema Later or Never Paradigm

The traditional relational data model uses the "schema first, data later" approach. Before users can load and store any data, they must design a schema

firstly [105]. When the schema is modified, they need to not only redefine the data structure, but also adjust the data application. But the data schema may be constantly changing or it cannot be design for big data. Besides, it is easy to cause data loss when loading and storing large amount of data. So the "schema first, data later" approach is not applicable to big data management.

Instead, data first, schema later or never is a trend in Data Science. For example, the technology of NoSQL uses key-value data model and schema later or never approach to ensure the agility of data management system. Despite the fact that data first, schema later or never approach is effective in the era of data-enriched offerings, it also brings some problems: it limits the performance of data management systems and increases the difficulty of application development.

The rise of data first, schema later or never paradigm is based on the change that from Pay-before-you-go to Pay-as-you-go. Pay-before-you-go pattern in information system construction requires people to define an information system firstly. The information system will be relatively stable within a short period of time after it is developed. The disadvantage of Pay-before-you-go pattern is that it cannot adapt to the complex and applications' dynamic changes.

5.4. Rethinking Data Consistency in Big Data Systems

Traditional data managers want to achieve strong and perfect data consistency, which means that data values in one data set should be consistent with values in another data set at the same point in any time. To have strong consistency, developers apply Two-phase Locking Protocol and Two-phase Commit protocols to relational databases. Strong consistency is not only helpful to ensure data quality, but also reduce the cost of computing subsequently. But big data management requires high scalability, high performance, fault tolerance, high scalability and high economic benefits. Strong consistency does not apply in the age of big data.

NoSQL and other emerging data management technologies fundamentally change people's traditional understanding to data consistency. People put forward some new data management theories such as CAP theorem and BASE principle, and concepts such as weak consistency and eventual consistency. Session consistency, update consistency and read-write consistency are used to optimize data management. In Data Science, different consistency needs are proposed according to diverse application scenarios.

The shift in people's need of data consistency reflects that the primary goal of data management is from perfectionism to realism. According to the CAP theorem [106], a distributed system can deliver only two of three desired characteristics: Consistency, Availability, and Partition tolerance. For example, Cassadra and Dynamo abandoned Consistency for Availability and Partition Tolerance.

5.5. Recognizing Data Replication and Data Locality

Data redundancy leads to high cost for achieving data consistency, so the negative impact of it should be eliminate in traditional relational database. But re-

dundant data is useful for load balancing, disaster recovery, and integrity inspection. The redundant data takes up more storage space through adding materialized views and adopting multi-copy technology, so that database can reduce the response time of user requests and ensures a better user experience. For example, Google Chrome saves images and html in a cache folder.

In addition, Data Locality has gained increasing attention in the development of computing application system. Data locality refers to the ability to move the computation close to actual data, instead of moving large data to computation. For example, RDD's get Preferred Locations method makes data can be read locally. And MapReduce schedules Map tasks to the machine that stores copy data. The rising of Multi-copy technology and Data Locality reflects that the traditional "compute-centric" product deployment pattern is changing to the "data-centric" one. The rising of Multi-copy technology and Data Locality reflects that the product deployment pattern is changing from compute-centric to data-centric.

5.6. Growth in Integrated Data Applications

Although there are many traditional relational databases, each of them adopts single highly standardized technology (such as the relational model or SQL). While the emerging NoSQL database adopts a variety of data management technologies based on different query interfaces and data models (such as Key-Value, Key-Document and Key-Column, and graph storage model). Diversification and high specialization are trends in emerging technologies. For example, MapReduce focus on Map/Reduce process splitting and combination, Tez focus on distributed batch processing, Storm focus on real-time processing and Druid focus on OLAP-oriented column storage. Some common techniques, such as Spark and YARN, are also becoming more specialized.

In traditional data computing/management, data product development uses single model that is relational, hierarchical or network. However, big data leads to integration of diverse computing/management technologies. Some integrated products combine multiple technologies into a single package. Hardware-software integration and embedded applications are emerging as key solutions to today's data computing/management problems. For example, Oracle Big Data [107] offers an integrated portfolio of products including HDFS, Oracle NoSQL, Cloudera CDH, data warehouse, memory computing and analytical applications.

Developing integrated applications is an important trend at the product level. The implementation of a product relies on integrating multiple technologies. Besides, the specialization trend is worth to be paid attention to at the technology level. A new technology addresses a relatively single problem.

5.7. Changes in the Complexity of Data Computing

Data scientists should adopt the simplest techniques to deal with complex data problems and to face constantly changing application scenarios. In contrast, the

techniques adopted in traditional data management are often complex. For example, traditional relational database uses Join operation to complete complex multi-table query. But Join operation requires that data cannot distributed among multiple nodes, so it is a bottleneck for relational databases to improve their data management capabilities. NoSQL abandoned complex Join operation, and instead simpler technologies.

Data Science is a highly practical subject, and its researches mainly focus on how to solve practical problems in the current society, rather than realize complex computing. Reduction in complexity of data computing is an emerging trend in Data Science.

5.8. The Advent of Data Products

Data product is a product that is featured by the use of data. Data scientists should think of a data product purely as a data problem [19]. Specifically, data product is all product that involves ingesting, processing, or presenting data. A data product can be used by people, computers, and other software and hardware systems to meet their needs. Data products include data sets, documents, knowledge bases, application systems, hardware systems, data service, data insights, data-driven decisions. For example, Google Glass should be regarded as a data product since it is enabled by Google big data.

As a key part of Data Science, data product development involves building out models, as well as running algorithms and technical deployment into production systems [108]. Combining data science with traditional product development is an emerging trend. The boundary between data product development and traditional product development will be increasingly blurred in the future. Data scientists should apply data-centric architecture when designing data product, and take user experience as the main evaluation indicator after they use the product. Embedding data science into a specific domain is a necessary part of product development, and it is beneficial to optimize traditional products and improve traditional products' competitiveness.

5.9. The Rise of Pro-Ams and Citizen Data Science

Knowledge is usually learned from domain experts/specialists in a traditional data analysis team. For example, ontologies need to be constructed by domain experts and an expert system contains the knowledge of human experts in a specific domain. But Pro-Am [109] plays an important role in the big data era. Pro-Am refers to quasi-expert people between expert and business people. Crowdsourcing has become increasingly prominent as a method of data processing in recent years. Most participants in a crowdsourcing task are Pro-Ams. Wikipedia is different from traditional encyclopedias in that its main contributions are completed by Pro-Ams instead of professionals only.

Traditional knowledge base is either a highly structured collection with insufficient data or a lowly structured collection with sufficient data. Crowdsourcing

is one of the best practices for solving this contradiction. Large-scale collaboration in crowdsourcing is classified into three types: machine collaboration, human-machine collaboration, and interpersonal collaboration. Among them, human-machine collaboration is the preferred means of conducting Data Science practices. For example, hybrid intelligence combines machine and human intelligence to overcome the shortcomings of existing AI systems [110]. The Semantic Web technology also supports for human-machine collaboration.

Citizen Data Science refers to an area that typically non-technical people participate in. It is the result of applying Pro-Am and large-scale collaboration in Data Science. Citizen Data Scientists are mainly non-professional hobbyists and volunteers who are different from Professional data scientists. In other words, Citizen Data Science is a quasi-data science based on crowdsourcing and Pro-Ams.

5.10. The Increasing Demand for Data Scientists

Data science is one of the hottest professions in recent years, data scientists are needed urgently to realize the opportunities presented by big data [111]. According to Drew Convey's data science Venn diagram [112], there are three basic components: theory (statistics and mathematics), practice (domain practices) and skill (hacking skills). A great data scientist not only masters the theory and practice of traditional science, but also has skills of creative design, critical thinking, and curiously questioning.

Therefore, cultivating great data scientists who have degree in data science is an important topic. There are four questions to be solved: 1) How to cultivate data science students through undergraduate [113], master [114] and PhD [115] programs? Current data science programs are mostly aimed at undergraduate and master students, programs for PhD students are not enough. 2) Is it necessary to set up a data science major? In China, the major of data science is named as "data science and big data technology". 3) How to build data science on knowledge from other disciplines? 4) How to create a tailored curriculum for learning data science? [113]

6. Conclusions

Data Science (DS) is an interdisciplinary discipline that employs statistical and machine learning methods in order to convert big data into actional insights, and its development remains at the innovation trigger stage. In contrast with well-developed disciplines, DS adopts the data-centered thinking pattern, recognizes that the property of data is more active than passive, converts big data into intelligence, solves data-intensive tasks, conducts data wrangling or munging, enhances user experiences of big data systems, introduces data intensive scientific discovery, as well as educates data scientists. DS is unique in its scientific objectives and research paradigm, and does not replicate directly the experiences from traditional disciplines. Some novel topics, principles, and paradigms are essential for further research on Data Science.

1) To avoid misconstruing Data Science. It is central to build the theoretical system of Data Science that most of the researchers should understand the meaning of this new discipline correctly. However, there is no shared understanding on Data Science yet. Some of them insist that data science is merely interdisciplinary applications of Statistics and Machine Learning, and it does not need its own new theories. They argue that application of Statistics and Machine Learning is crucial for DS, and overlook the unique theories of Data Science. In fact, Statistics and Machine Learning are the theoretical foundation of DS, not its core components. DS is an independent discipline like Statistics and Machine Learning. DS is unique in its scientific mission, research perspective, thinking pattern, underlying principles and theoretical framework, which are distinct from other disciplines. Therefore, to avoid misinterpreting Data Science is a prerequisite for building its theoretical system.

2) To take advantages of active property of big data. One of the main contributions of Data Science is that it shifts our thinking pattern and views big data as active beings. People have been thought of data as passive or dead thing to date, and how to input human intelligence into data is the main concern of the related studies. For instance, traditional data prepressing theories try to convert complex data into simple data through defining schema, data cleansing, and filling missing values. However, Data Science highlights the active property of data and begins to discuss how to take advantage of data. As a result, some novel terms, such as data-driven applications, data-centric design, data insights, and big data ecosystem, are widely accepted. Data Science regards complexity as a natural attribute of big data and does not conduct traditional data preprocessing no longer. Admitting that data is active rather than passive is the basic starting point of studying Data Science.

3) To balance the three dimensions of Data Science. Data science (DS) not only involves theory and practice, but also requires skills such as creative design, critical thinking and curious asking. DS differs from traditional disciplines in that it has three main dimensions: theory, practice and skills. The research on data science should not only bridge the gap between theory and practice, but also avoid overlooking its skill dimension. To balance those three dimensions is one of the main challenges in Data Science studies.

4) To introduce Design of Experiments. Design of Experiments（DOE）is one of the essential activities of data science projects. Data scientists should creatively propose research hypotheses according to the objectives of data science projects, and design corresponding experiments, and conduct the data experiments and test the hypothesis. Taking the student programs of Data Science majors in the University of Washington as well as the University of California, Berkeley as examples, courses titled Applied Statistics & Experimental Design or Experiments and Causality were provided, respectively. The both courses focus on improving students' ability in DOE as well as hypothesis testing.

5) To embrace causality analysis. There is a misconception that Data Science

(DS) only concerns correlation analysis, and causality analysis is outside the scope of it. However, correlation analysis can only be used to identify the correlations in big data, but cannot guide how to optimize and intervene in the identified correlations. Where the correlation changes, or human intervention is needed, the causation relation in big data is required to be analyzed. In a DS project, the data scientists are responsible not only to discover possible correlations in big data, but also to reveal the causality behind the correlations with the collaboration of domain experts. To embrace causality analysis is becoming one of the most discussed topics in DS. For instance, the Experiments and Causality Analysis or the Causal Inference for Data Science are listed in DS courses at University of California, Berkeley and Columbia University, respectively.

6) To develop data products. Developing data products is one of the distinct objectives of Data Science (DS) studies. Data products in DS are not limited to products in data form. All products that utilize data to provide new services should be regarded as a data product. Data can be used to promote product innovation, and traditional products will be transformed into data products by application of DS theories. For example, Google Glasses is a data product in that its novel features are derived from data. Data-centered thinking is the fundamental difference between data products and traditional ones. Data products will be the most common applications of Data Science.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mayer-Schönberger, V. and Kenneth, C. (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston.

[2] Boyd, D. and Crawford, K. (2012) Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society*, **15**, 662-679. https://doi.org/10.1080/1369118X.2012.678878

[3] Kitchin, R. (2014) Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, **1**, 1-12. https://doi.org/10.1177/2053951714528481

[4] Jagadish, H.V. (2015) Big Data and Science: Myths and Reality. *Big Data Research*, **2**, 49-52. https://doi.org/10.1016/j.bdr.2015.01.005

[5] Song, I. and Zhu, Y.J. (2016) Big Data and Data Science: What Should We Teach. *Expert Systems*, **33**, 364-373. https://doi.org/10.1111/exsy.12130

[6] Naur, P. (1974) Concise Survey of Computer Methods. Petrocelli Books, New York.

[7] Cleveland, W.S. (2001) Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review*, **69**, 21-26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

[8] Mattmann, C.A. (2013) Computing: A Vision for Data Science. *Nature*, **493**, 473-475. https://doi.org/10.1038/493473a

[9] Dhar, V. (2013) Data Science and Prediction. *Communications of the ACM*, **56**, 64-73. https://doi.org/10.1145/2500499

[10] Davenport, T.H. and Patil, D.J. (2012) Data Scientist. *Harvard Business Review*, **90**, 70-76.

[11] Kitchin, R. (2013) Big Data and Human Geography: Opportunities, Challenges and Risks. *Dialogues in Human Geography*, **3**, 262-267. https://doi.org/10.1177/2043820613513388

[12] Smith, M. (2015) The White House Names Dr. DJ Patil as the First US Chief Data Scientist. https://obamawhitehouse.archives.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist

[13] Rivera, J. and Van der Meulen, R. (2014) Gartner's 2014 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business. Connecticut, EEUU: Gartner Group.

[14] Gartner, J. (2016) Hype Cycle for Data Science, 2016. https://www.gartner.com/doc/3388917/hype-cycle-data-science

[15] O'Neil, C. and Schutt, R. (2013) Doing Data Science: Straight Talk from the Frontline. O'Reilly Media Inc., Newton, 7.

[16] Overton, J. (2016) Going Pro in Data Science. O'Reilly Media Inc., Newton, 12.

[17] Chao, L. (2017) Data Science Theory and Practice. Tsinghua University Press, Beijing, 15.

[18] Myers, R. (2019) Data Management and Statistical Analysis Techniques. Scientific e-Resources, 2.

[19] Patil, D.J. (2012) Data Jujitsu. O'Reilly Media Inc., Newton.

[20] Davenport, T.H. and Kudyba, S. (2016) Designing and Developing Analytics-Based Data Products. *MIT Sloan Management Review*, **58**, 83.

[21] Gray, J., Chambers, L. and Bounegru, L. (2012) The Data Journalism Handbook: How Journalists Can Use Data to Improve the News. O'Reilly Media Inc., Newton.

[22] Kalidindi, S.R. and De Graef, M. (2015) Materials Data Science: Current Status and Future Outlook. *Annual Review of Materials Research*, **45**, 171-193. https://doi.org/10.1146/annurev-matsci-070214-020844

[23] Fang, B. and Zhang, P. (2016) Big Data in Finance. In: *Big Data Concepts, Theories, and Applications*, Springer, Cham, 391-412. https://doi.org/10.1007/978-3-319-27763-9_11

[24] Davis, K. (2012) Ethics of Big Data: Balancing Risk and Innovation. O'Reilly Media Inc., Newton.

[25] West, D.M. (2012) Big Data for Education: Data Mining, Data Analytics, and Web Dashboards. *Governance Studies at Brookings*, **4**, 1-10.

[26] Labrinidis, A. and Jagadish, H.V. (2012) Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment*, **5**, 2032-2033. https://doi.org/10.14778/2367502.2367572

[27] Kaisler, S., *et al.* (2013) Big Data: Issues and Challenges Moving Forward. 2013 46*th Hawaii International Conference on System Sciences IEEE*, Wailea, 7-10 January 2013, 995-1004. https://doi.org/10.1109/HICSS.2013.645

[28] Chen, H., Chiang, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, **36**, 1165-1188. https://doi.org/10.2307/41703503

[29] Provost, F. and Fawcett, T. (2013) Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, **1**, 51-59.

https://doi.org/10.1089/big.2013.1508

[30] Blei, D.M. and Smyth, P. (2017) Science and Data Science. *Proceedings of the National Academy of Sciences*, **114**, 8689-8692. https://doi.org/10.1073/pnas.1702076114

[31] Shanahan, J.G. and Dai, L. (2015) Large Scale Distributed Data Science Using Apache Spark. *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, 10-13 August 2015, 2323-2324. https://doi.org/10.1145/2783258.2789993

[32] Holmes, A. (2012) Hadoop in Practice. Manning Publications Co., New York.

[33] Sharma, S., *et al.* (2016) Leading NoSQL Models for Handling Big Data: A Brief Review. *International Journal of Business Information Systems*, **22**, 1-25. https://doi.org/10.1504/IJBIS.2016.075714

[34] Sadalage, P.J. and Fowler, M. (2013) NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Pearson Education, London.

[35] Marx, V. (2013) Biology: The Big Challenges of Big Data. *Nature*, **498**, 255-260. https://doi.org/10.1038/498255a

[36] Raghupathi, W. and Raghupathi, V. (2014) Big Data Analytics in Healthcare: Promise and Potential. *Health Information Science and Systems*, **2**, 3. https://doi.org/10.1186/2047-2501-2-3

[37] Kim, G.-H., Trimi, S. and Chung, J.-H. (2014) Big-Data Applications in the Government Sector. *Communications of the ACM*, **57**, 78-85. https://doi.org/10.1145/2500873

[38] Daniel, B. (2015) Big Data and Analytics in Higher Education: Opportunities and Challenges. *British Journal of Educational Technology*, **46**, 904-920. https://doi.org/10.1111/bjet.12230

[39] George, G., Haas, M.R. and Pentland, A. (2014) Big Data and Management. *Academy of Management Journal*, **57**, 321-326. https://doi.org/10.5465/amj.2014.4002

[40] Swan, M. (2013) The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data*, **1**, 85-99. https://doi.org/10.1089/big.2012.0002

[41] Lewis, S.C. (2015) Journalism in an Era of Big Data: Cases, Concepts, and Critiques. Taylor & Francis, Abingdon-on-Thames, 321-330. https://doi.org/10.1080/21670811.2014.976399

[42] Rahm, E. (2016) Big Data Analytics. *It—Information Technology*, **58**, 155-156. https://doi.org/10.1515/itit-2016-0024

[43] Baumer, B. (2015) A Data Science Course for Undergraduates: Thinking with Data. *The American Statistician*, **69**, 334-342. https://doi.org/10.1080/00031305.2015.1081105

[44] Hardin, J., *et al.* (2015) Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*, **69**, 343-353. https://doi.org/10.1080/00031305.2015.1077729

[45] Cassel, L.N., *et al.* (2017) Advancing Data Science for Students of All Majors. *Proceedings of the* 2017 *ACM SIGCSE Technical Symposium on Computer Science Education*, Seattle, 8-11 March 2017, 722. https://doi.org/10.1145/3017680.3022362

[46] Berman, F.D. and Bourne, P.E. (2015) Let's Make Gender Diversity in Data Science a Priority Right from the Start. *PLoS Biology*, **13**, e1002206. https://doi.org/10.1371/journal.pbio.1002206

[47] Chao, L.M. (2016) Data Science. Tsinghua University Press, Beijing.

[48] Cooper, P. (2014) Data, Information, Knowledge and Wisdom. *Anaesthesia & Intensive Care Medicine*, **15**, 44-45. https://doi.org/10.1016/j.mpaic.2013.11.009

[49] Erl, T., Khattak, W. and Buhler, P. (2016) Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall Press, Upper Saddle River.

[50] Rowley, J. (2007) The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, **33**, 163-180. https://doi.org/10.1177/0165551506070706

[51] Riofrio, G., *et al.* (2015) Business Intelligence Applied to Learning Analytics in Student-Centered Learning Processes. 2015 *Latin American Computing Conference* (*CLEI*) *IEEE*, Arequipa, 19-23 October 2015, 1-10.

[52] Wang, G., *et al.* (2016) Big Data Analytics in Logistics and Supply Chain Management: Certain Investigations for Research and Applications. *International Journal of Production Economics*, **176**, 98-110. https://doi.org/10.1016/j.ijpe.2016.03.014

[53] Cárdenas, A.A., Manadhata, P.K. and Rajan, S.P. (2013) Big Data Analytics for Security. *IEEE Security & Privacy*, **11**, 74-76. https://doi.org/10.1109/MSP.2013.138

[54] Leek, J.T. and Peng, R. (2015) What Is the Question? Mistaking the Type of Question Being Considered Is the Most Common Error in Data Analysis. *Science*, **347**, 1314-1315. https://doi.org/10.1126/science.aaa6146

[55] Van Dijck, J. (2014) Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology. *Surveillance & Society*, **12**, 197-208. https://doi.org/10.24908/ss.v12i2.4776

[56] Ruckenstein, M. and Pantzar, M. (2017) Beyond the Quantified Self: Thematic Exploration of a Dataistic Paradigm. *New Media & Society*, **19**, 401-418. https://doi.org/10.1177/1461444815609081

[57] Cheong, L.K. and Chang, V. (2007) The Need for Data Governance: A Case Study. *ACIS* 2007 *Proceedings*, Toowoomba, 5-7 December 2007, 100.

[58] Khatri, V. and Brown, C.V. (2010) Designing Data Governance. *Communications of the ACM*, **53**, 148-152. https://doi.org/10.1145/1629175.1629210

[59] Thomas, G. (2006) The DGI Data Governance Framework. The Data Governance Institute, Orlando, 20.

[60] Lee, S.U., Zhu, L.M. and Jeffery, R. (2017) Design Choices for Data Governance in Platform Ecosystems: A Contingency Model.

[61] CMMI Institute. Data Management Maturity (DMM)℠ Model. http://cmmiinstitute.com/data-management-maturity

[62] Liu, J.Z., *et al.* (2016) Rethinking Big Data: A Review on the Data Quality and Usage Issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, **115**, 134-142. https://doi.org/10.1016/j.isprsjprs.2015.11.006

[63] Lee, J.Z., Wang, Z.H. and Gao, H. (2016) State-of-the-Art of Research on Big Data Usability. *Journal of Software*, **27**, 1605-1625.

[64] Rahm, E. and Do, H.H. (2000) Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, **23**, 3-13.

[65] Wickham, H. (2014) Tidy Data. *Journal of Statistical Software*, **59**, 1-23. https://doi.org/10.18637/jss.v059.i10

[66] Lafuente, G. (2015) The Big Data Security Challenge. *Network Security*, **2015**, 12-14. https://doi.org/10.1016/S1353-4858(15)70009-7

[67] Perera, C., *et al.* (2015) Big Data Privacy in the Internet of Things Era. *IT Professional*, **17**, 32-39. https://doi.org/10.1109/MITP.2015.34

[68] Patil, D. and Noren, A. (2011) Building Data Science Teams: The Skills, Tools and

Perspectives behind Great Data Science Groups. O'Reilly Media Inc., Newton.

[69] Banerjee, S. (2015) Citizen Data Science for Social Good: Case Studies and Vignettes from Recent Projects. https://www.researchgate.net/publication/283119007_Citizen_Data_Science_for_Social_Good_Case_Studies_and_Vignettes_from_Recent_Projects

[70] Parasie, S. and Dagiral, E. (2013) Data-Driven Journalism and the Public Good: "Computer-Assisted-Reporters" and "Programmer-Journalists" in Chicago. *New Media & Society*, **15**, 853-871. https://doi.org/10.1177/1461444812463345

[71] Du, D.Y., Li, A.H. and Zhang, L.L. (2014) Survey on the Applications of Big Data in Chinese Real Estate Enterprise. *Procedia Computer Science*, **30**, 24-33. https://doi.org/10.1016/j.procs.2014.05.377

[72] Middleton, S.E., Shadbolt, N.R. and De Roure, D.C. (2004) Ontological User Profiling in Recommender Systems. *ACM Transactions on Information Systems*, **22**, 54-88. https://doi.org/10.1145/963770.963773

[73] Marshall, P., Rhodes, M. and Todd, B. (2014) Ultimate Guide to Google AdWords. Entrepreneur Press, Irvine.

[74] Gurrin, C., Smeaton, A.F. and Doherty, A.R. (2014) Life Logging: Personal Big Data. *Foundations and Trends in Information Retrieval*, **8**, 1-125. https://doi.org/10.1561/1500000033

[75] Bello-Orgaz, G., Jung, J.J. and Camacho, D. (2016) Social Big Data: Recent Achievements and New Challenges. *Information Fusion*, **28**, 45-59. https://doi.org/10.1016/j.inffus.2015.08.005

[76] Mohanty, S., Jagadeesh, M. and Srivatsa, H. (2013) Big Data Imperatives: Enterprise "Big Data" Warehouse BI Implementations and Analytics. Apress, New York. https://doi.org/10.1007/978-1-4302-4873-6

[77] Bertot, J.C., *et al.* (2014) Big Data, Open Government and e-Government: Issues, Policies and Recommendations. *Information Polity*, **19**, 5-16. https://doi.org/10.3233/IP-140328

[78] Aggarwal, A.K. (2019) Opportunities and Challenges of Big Data in Public Sector. In: *Web Services*: *Concepts*, *Methodologies*, *Tools*, *and Applications*, IGI Global, Hershey, 1749-1761. https://doi.org/10.4018/978-1-5225-7501-6.ch090

[79] Jurney, R. (2017) Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark. O'Reilly Media Inc., Newton.

[80] Moreno, J., *et al.* (2020) Improving Incident Response in Big Data Ecosystems by Using Blockchain Technologies. *Applied Sciences*, **10**, 724. https://doi.org/10.3390/app10020724

[81] Matt Turck. Big Data Landscape 2016 v18 FINAL. http://mattturck.com/big-data-landscape-2016-v18-final

[82] Drexl, J. (2016) Designing Competitive Markets for Industrial Data-Between Propertization and Access. Max Planck Institute for Innovation & Competition Research Paper 16-13. https://doi.org/10.2139/ssrn.2862975

[83] Jin, X.L., *et al.* (2015) Significance and Challenges of Big Data Research. *Big Data Research*, **2**, 59-64. https://doi.org/10.1016/j.bdr.2015.01.006

[84] Al-Jarrah, O.Y., *et al.* (2015) Efficient Machine Learning for Big Data: A Review. *Big Data Research*, **2**, 87-93. https://doi.org/10.1016/j.bdr.2015.04.001

[85] Batra, S. (2014) Big Data Analytics and Its Reflections on DIKW Hierarchy. *Review of Management*, **4**, 5.

[86] Donhost, M.J. and Anfara Jr., V.A. (2010) Data-Driven Decision Making. *Middle*

School Journal, **42**, 56-63. https://doi.org/10.1080/00940771.2010.11461758

[87] Chen, C.L. and Zhang, C.-Y. (2014) Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Information Sciences*, **275**, 314-347. https://doi.org/10.1016/j.ins.2014.01.015

[88] Voulgaris, Z. and Magoulas, G.D. (2008) Extensions of the k nearest Neighbour Methods for Classification Problems. *Proceedings of the* 26*th IASTED International Conference on Artificial Intelligence and Applications*, *AIA*, Vol. 8, 23-28.

[89] Rajaraman, A. (2008) More Data Usually Beats Better Algorithms. Datawocky Blog.

[90] Kleppmann, M. (2017) Designing Data-Intensive Applications: The Big Ideas behind Reliable, Scalable, and Maintainable Systems. O'Reilly Media Inc., Newton.

[91] Brewer, E. (2013) Parallelism in the Cloud. Workshop on Hot Topics in Parallelism. Keynote Talk.

[92] Fan, J.Q., Han, F. and Liu, H. (2014) Challenges of Big Data Analysis. *National Science Review*, **1**, 293-314. https://doi.org/10.1093/nsr/nwt032

[93] Edgar, R.C. (2004) MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics*, **5**, Article No. 113.

[94] Ginsberg, J., *et al.* (2009) Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, **457**, 1012-1014. https://doi.org/10.1038/nature07634

[95] Lazer, D., *et al.* (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science*, **343**, 1203-1205. https://doi.org/10.1126/science.1248506

[96] Tansley, S. and Tolle, K. (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Vol. 1, Microsoft Research, Redmond.

[97] Kalechofsky, H. (2016) A Simple Framework for Building Predictive Models. A Little Data Science Business Guide. 1-18.

[98] Shmueli, G. (2010) To Explain or to Predict? *Statistical Science*, **25**, 289-310. https://doi.org/10.1214/10-STS330

[99] Dhar, V. and Chou, D. (2001) A Comparison of Nonlinear Models for Financial Prediction. *IEEE Transactions on Neural Networks*, **12**, 907-921. https://doi.org/10.1109/72.935099

[100] Føllesdal, D. (1979) Hermeneutics and the Hypothetico-Deductive Method. *Dialectica*, **33**, 319-336. https://doi.org/10.1111/j.1746-8361.1979.tb00759.x

[101] Sober, E. (2002) Instrumentalism, Parsimony, and the Akaike Framework. *Philosophy of Science*, **69**, S112-S123. https://doi.org/10.1086/341839

[102] Rasmussen, C.E. and Ghahramani, Z. (2001) Occam's Razor. In: *Advances in Neural Information Processing Systems* 13 (*NIPS* 2000), MIT Press, Cambridge, 276-282.

[103] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[104] Glass, G.V. (1976) Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, **5**, 3-8. https://doi.org/10.3102/0013189X005010003

[105] Liu, Z.H., Hammerschmidt, B. and McMahon, D. (2014) JSON Data Management: Supporting Schema-Less Development in RDBMS. *Proceedings of the* 2014 *ACM SIGMOD International Conference on Management of Data*, Snowbird, 22-27 June 2014, 1247-1258. https://doi.org/10.1145/2588555.2595628

[106] Brewer, E. (2012) CAP Twelve Years Later: How the "Rules" Have Changed. *Computer*, **45**, 23-29. https://doi.org/10.1109/MC.2012.37

[107] Plunkett, T., *et al.* (2013) Oracle Big Data Handbook. McGraw Hill Professional, New York.

[108] Chawla, S., Hartline, J. and Nekipelov, D. (2014) Mechanism Design for Data Science. *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, Palo Alto, 8-12 June 2014, 711-712. https://doi.org/10.1145/2600057.2602881

[109] Leadbeater, C. and Miller, P. (2004) The Pro-Am Revolution: How Enthusiasts Are Changing Our Society and Economy. Demos, London.

[110] Kamar, E. (2016) Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (*IJCAI*-16), New York, 9-15 July 2016, 4070-4073.

[111] Power, D.J. (2016) Data Science: Supporting Decision-Making. *Journal of Decision Systems*, **25**, 345-356. https://doi.org/10.1080/12460125.2016.1171610

[112] Conway, D. (2011) Data Science in the US Intelligence Community. *IQT Quarterly*, **2**, 24-27.

[113] Anderson, P., McGuffee, J. and Uminsky, D. (2014) Data Science as an Undergraduate Degree. *Proceedings of the* 45*th ACM Technical Symposium on Computer Science Education*, Atlanta, 5-8 March 2014, 705-706. https://doi.org/10.1145/2538862.2538868

[114] Marshall, L. and Eloff, J.H.P. (2016) Towards an Interdisciplinary Master's Degree Programme in Big Data and Data Science: A South African Perspective. In: *Annual Conference of the Southern African Computer Lecturers' Association*, Springer, Cham, 131-139. https://doi.org/10.1007/978-3-319-47680-3_13

[115] West, J.D. and Portenoy, J. (2016) Chapter 10: The Data Gold Rush in Higher Education. In: *Big Data Is Not a Monolith*, The MIT Press, Cambridge, 129.