DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.3, AUG. 2022

RESEARCH ARTICLES

Data science: Trends, perspectives, and prospects

Chaolemen Borjigin, Chen Zhang

a. Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China

b. Information Resource Management School, Renmin University of China, Beijing, China

ABSTRACT

Data science is a rapidly growing academic field with significant implications for all conventional scientific studies. However, most relevant studies have been limited to one or several facets of data science from a specific application domain perspective and less to discuss its theoretical framework. Data science is unique in that its research goals, perspectives, and body of knowledge are distinct from other sciences. The core theories of data science are the DIKW pyramid, data-intensive scientific discovery, data science life cycle, data wrangling or munging, big data analytics, data management, and governance, data products DevOps, and big data visualization. Six main trends characterize the recent theoretical studies on data science are: (1) the growing significance of DataOps, (2) the rise of citizen data scientists, (3) enabling augmented data science, (4) integrating data warehouse with data lake, (5) diversity of domain-specific data science, and (6) implementing data stories as data products. Further development of data science should prioritize four ways to turn challenges into opportunities: (1) accelerating theoretical studies of data science, (2) the trade-off between explainability and performance, (3) achieving data ethics, privacy and trust, and (4) aligning academic curricula with industrial needs.

KEYWORDS

CCS concepts; General and reference; Surveys and overviews; Data science; Big data; Data products; Data-driven management; The DIKW pyramid

1 INTRODUCTION

Data science is gaining momentum across a range of disciplines. The term "data science" can be traced back to 1974 when the computer scientist Peter Naur coined and defined it as the science of dealing with data (Naur, 1974), and then data science first occurred as a scientific idea in computer science. In 2001, William S. Cleveland, a statistician, proposed an action plan for expanding the technical areas of the field of statistics (Cleveland, 2001), and statistics was the second discipline that delineated data science. Hence, computer science and statistics are the two main theoretical foundations of data science. In 2010, Drew Conway, the founder of Alluvium, published a data science Venn diagram and first discussed the interdisciplinary of data science. He argued that data science is located at the intersection of hacking skills, math and statistics knowledge, and substantive expertise. Further, this Venn

diagram has many variations (Ullman, 2020; Taylor, 2016). Currently, data science is a hot topic in a variety of disciplines and has nurtured new data science branches in traditional sciences, such as Geography (Singleton & Arribas-Bel, 2021), Materials Science (Kalidindi & De Graef, 2015), Health Science (Peek & Rodrigues, 2018), Business Data Science (Provost & Fawcett, 2013), Environmental Science (Gibert et al., 2018), Surgery(Maier-Hein et al., 2017), and Cybersecurity(Sarker et al., 2020; Zhang et al., 2021).

However, most related studies are dedicated to discussing one or several practical facets of data science from their distinct domain perspective and less to discussing its theoretical framework. This study carries out an in-depth analysis of the data science theoretical framework based on comprehensive literature research and presents trends, perspectives, and prospects of data science. This paper is organized as follows: Section 2 discusses the main research motivations, unique thinking patterns, and the body of knowledge. Section 3 describes the core theories of data science and their recent progress, and Section 4 proposes the emerging trends of data science studies. Further, Section 5 provides some recommendations for the academic research or industrial application of data science. Finally, Section 6 presents the conclusion.

2 DATA SCIENCE

Data science is a new cross-disciplinary science dealing with big data drawing on machine learning (ML), statistics, and data visualization as its primary theoretical basis. Data science concludes a set of fundamental principles that guide the extraction of information and knowledge from data. Data science focuses on processing, computing, managing, analyzing big data, and providing data products. Data science is novel in that its research goals, perspectives, and body of knowledge are distinct from the traditional sciences.

2.1 The Essential Research Goal of Data Science

Data science aims to accelerate the inter-transformation between materials, energy, and data, notably to reduce the consumption of materials and energy or improve the effectiveness and efficiency of exploiting them by taking advantage of data. Further, the essential research goals of data science studies can be categorized into the following subjects:

1. To reveal the underlying mechanism of big data;
2. To turn data into knowledge, understandings, or wisdom;
3. To gain insights from big data;
4. To convert big data into business value;
5. To enable data-driven decision-making or data-driven decision support;
6. To implement data product development and operations (data product DevOps);
7. To cultivate and maintain big data ecosystems.

2.2 The Unique Research Perspective of Data Science

With the raising of the big data era, our main concerns for data have undergone a significant shift from "what can we do for data?" to "what can data do for us?" (see Figure 1). This shift (or diversification) in research perspectives is the main difference between data science and traditional data-related studies. Many new terms have been coined in big data era, such as "data-intensive scientific discovery," "data-driven decision-making," "data-centric architecture," and "data jiu-jitsu", most of them are in line with this new shift in research perspec-

tives.

The concern of traditional data-related theories concentrates on "what can I do for data?" Traditional data engineering, data structure, database, data warehouse, data mining, and other data-related theories focus on cleaning, labeling, extracting, transforming, and loading data. Traditional theories place a high value on ways to manually process data to make sure they are more valuable or ready for the subsequent process and future usage. However, data science conforms to the alternative research perspective of "what can data do for me?" The main concerns of data science include:

- What automatic decision-making or decision support can be enabled by taking advantage of big data?
- Which business opportunities or new target markets can be identified by harvesting big data?
- What are the uncertainties that big data can reduce?
- What predictive or prescriptive analysis can be conducted based on big data?
- Are there any potential, valuable, and usable hidden patterns or models within big data?

In short, we deal with the relationship between humans and data from two distinct perspectives "What can I do for data?" and "What can data do for me?" in the era of big data. The latter is emphasized in data science.

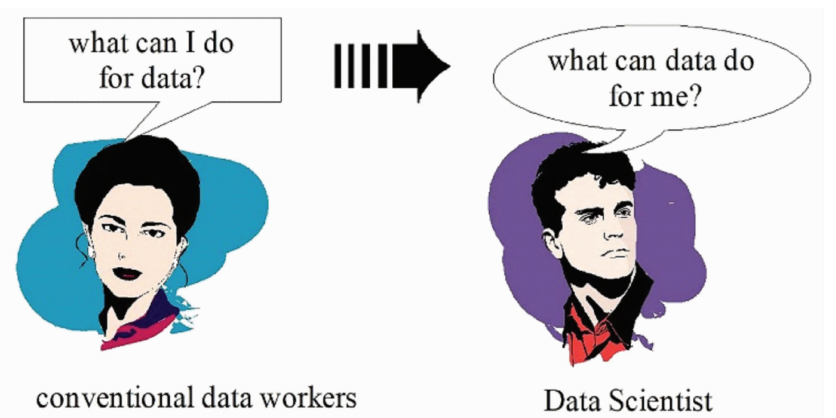

Figure 1 New and unique perspective of data science

2.3 Data Science Body of Knowledge

The body of knowledge for data science involves its theoretical foundations, main branches, domain expertise, as well as issues from humanities and social sciences (see Figure 2). Data science is enabled by statistics, ML, and data visualization, and these three distinct disciplines are the theoretical foundation of data science. The research topics of data science are categorized into six main branches: fundamental concepts and principles of data science, data wrangling, data computing, data management, data analysis, and data products DevOps. Also, placing data science theories into practice is commonly domain-dependent; domain expertise is essential for these applications. The data science theory involves humanities and social science issues, especially big data ethics, privacy, and trust.

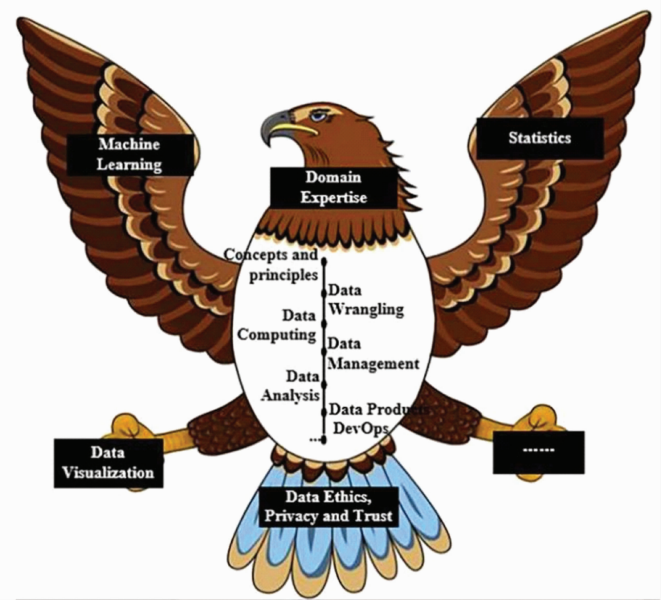

Figure 2 Data science body of knowledge

1. Fundamental concepts and principles: the basic theories of data science include its core concepts, research motivations, research areas, life cycle, main principles, typical applications, and project management. Note that the basic theories are distinct from the theoretical basis. The former is within the research boundary of data science, while the latter is outside that scope.
2. Data wrangling: Data wrangling (or data munging) is a novel term coined for data science. "Data wrangling" refers to a series of data preprocessing activities to enhance data quality, reduce the complexity of data computing, and improve the accuracy of data processing. Data science projects must perform a series of preprocessing activities on raw data, including data audit, cleaning, ETL, integration, reduction, and labeling. Unlike traditional data preprocessing, data wrangling (or data munging) in data science highlights value-added processes through integrating the creative design, critical thinking, and curiosity of data scientists into data preprocessing.
3. Data computing: In data science, computing models have significantly shifted from traditional computing technologies such as centralized computing, distributed computing, and grid computing to emerging new technologies like cloud computing, edge computing, and mobile computing. Examples of big data computing technologies are GFS, BigTable, MapReduce, Spark, and YARN. Changes in data computing theories involve the primary bottlenecks, research motivations, main contradictions, and thinking patterns for data computing, which will be discussed later.
4. Data management: Big data needs to be effectively managed to conduct data analysis, data reuse, and long-term storage. Also, data science needs relational databases and emerging big data management technologies such as NoSQL, NewSQL, and relational cloud.
5. Data analysis: In data science, data analysis focuses on prescriptive and predictive analysis rather than descriptive or diagnostic. The prescriptive model involves large-scale

testing and optimization and it is a means of embedding analytics into key processes (KP) and employee behaviors (Davenport, 2013). Data scientists prefer to choose open-source tools, which are different from commercial software. Consequently, Python and R are popular data analysis tools for data scientists.

6. Data products DevOps: "Data product" has a special meaning in data science. Data products development is a critical research task in data science projects because it represents a unique research topic that distinguishes data science from other sciences. Unlike traditional products development, data products development is data-centric, diverse, hierarchical, and value-added. Also, data products development capabilities are the primary source of competitiveness for data scientists. Therefore, one of the specific purposes of data science studies is to provide a wide range of data products.

Data science has various domain applications. Representative practical applications by far are Google Flu Trends (Ginsberg et al., 2009), Target pregnancy prediction (Hill, 2012), MetroMile insurance, IBM Workbench, Databircks, London Olympics data news, Google Translate, and the Climate FieldView.

3 CORE THEORIES

The core theories of data science to date are the DIKW pyramid, data-intensive scientific discovery, data science life cycle, data wrangling or munging, big data analytics, data management, data governance, data products development, and big data visualization.

3.1 The DIKW Pyramid

The DIKW pyramid is a hieratical framework that describes functional relationships between data, information, knowledge, and wisdom. Data are symbols that represent the properties of objects and events. Information consists of processed data; data are processed to improve their usefulness. Information is contained in descriptions, and in answers to questions that begin with such words as "who," "what," "when," "where," and "how many." Instructions and answers convey knowledge to how-to questions. Wisdom deals with values and involves the exercise of judgment (Ackoff, 1989).

The DIKW pyramid is a widely discussed topic in data science because the pyramid represents the underlying motivation for data science studies: converting big data into big wisdom. For instance, John D. Kelleher and Brendan Tierney (2018) proposed the data science pyramid based on the DIKW pyramid to show a hierarchy of data science activities from data capture and generation to decision support using data-driven models deployed in the business context.

However, there is a notable difference between the discussion on the DIKW pyramid from a data science perspective and the conventional one that stems from the fact. The former seeks an integrated solution for converting data into information, knowledge, or wisdom instead of isolated solutions (see Figure 3). Datafication refers to recording the real world into data; data wrangling is employed to turn messy data into tidy data; data analytics are used to acquire information from data; and data insights are applied to obtain knowledge from data directly. Data products DevOps are operationalized by converting data into wisdom. Data scientists tend to regard information, knowledge, and wisdom respectively as analyzed data, valuable insights, and the capability to convert data into products.

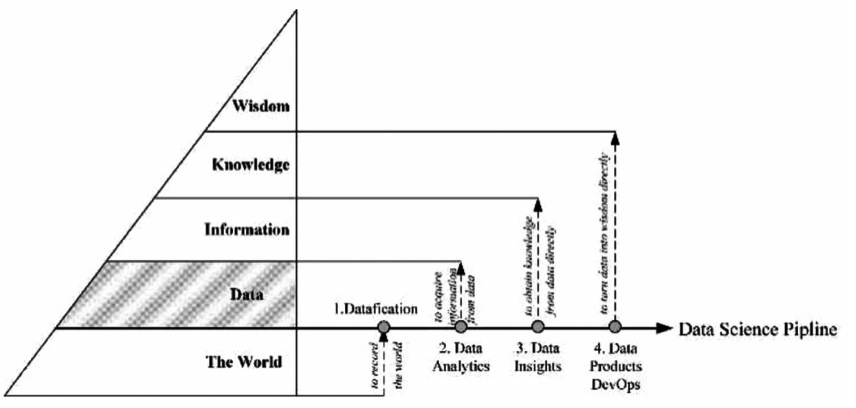

Figure 3 DIKW pyramid from data science perspectives

3.2 Data–Intensive Scientific Discovery

Data-intensive scientific discovery is the unique thinking paradigm of data science in that it is distinct from conventional data-related studies, including data engineering, data analysis, data retrieval, and data preprocessing. Jim Gray (2009) proposed that our society is turning to the fourth scientific paradigm, namely the data-intensive scientific discovery paradigm, a new expansion of established scientific methods (Tansley & Tolle, 2009). However, conventional data-related studies conform to alternative research paradigms, such as empirical evidence, scientific theory, and computational science.

Introducing the novel research paradigm into data science enables it to obtain previously unknown patterns, insights, and knowledge from big data. Data science has become one of the hot research topics in traditional data-related studies. Zhu and Xiong (2015) argued that data researchers tended to study data in cyberspace, which is different from natural science and social science. Chen and Zhang (2014) discussed applications and tools to address big data challenges and suggested some principles for designing effective data systems. Cao (2017) discussed the significance of data DNA and conducted a comprehensive investigation of fundamental aspects of data science. Beck (2016) proposed that data scientists were equipped to seamlessly process, analyze, and communicate in a data-intensive context.

Data science primarily adopts the "data first, hypothesis later or never" approach to dealing with big data. By contrast, the computational sciences tend to employ the "hypothesis first, data later" approach, i.e., putting forward hypotheses before collecting or analyzing data. As for the data-intensive paradigm, the researchers first collect data as much as possible and conduct predictive or prescriptive analysis to identify unknown insights or hidden patterns. Furthermore, data scientists enable enterprises to make data-driven decisions by capturing, mining, and analyzing massive amounts of data and measuring and verifying data with statistical models or ML algorithms. The introduction of this novel scientific paradigm motivated a shift from computing-centered thinking toward data-centered thinking.

3.3 Data Science Life cycle

The data science life cycle is one of the basic theories of data science and reveals the conceptional workflow of data science projects. Although it is an accepted convention that the life cycle model is the typical means to describe data science projects, researchers have not yet reached a consensus on the stages of the data science life cycle. Larson and Chang

(2016) contrast business intelligence life cycle with data science life cycle in terms of scope, data acquisition/discovery, analyze/visualize, model/design/development, validation, deployment, as well as support/feedback. Boehm et al. (2020) proposed an open-source ML system for the end-to-end data science life cycle, involving such activities as data integration, cleaning, and preparation, over local, distributed, and federated model training, debugging, and serving. Ho and Beyan (2020) described phases in the data science life cycle, including data ingestion, scrubbing, visualization, modeling, and analysis, and further discussed common biases at each stage. Song and Zhu (2017) proposed that the data science life cycle has eight main stages: (1) business understanding, (2) data understanding, (3) data preparation, (4) model planning, (5) model building, (6) evaluation, (7) deployment, and (8) review and monitoring. Wang et al. (2021) described a data science life cycle that contained ten distinct stages: (1) requirement gathering and problem formulation, (2) data acquisition and governance, (3) data readiness, data preprocessing, and data cleaning, (4) feature engineering, (5) model building and model training, (6) model presentation and stakeholder verification, (7) model deployment, (8) runtime monitoring, (9) model refinement (post-deployment), and (10) decision-making and optimization.

Data science projects aim to obtain valuable insights from big data to make better decisions. With the maturity of ML, cloud computing, and artificial intelligence, more jobs are auto-completed by machines. However, humans still play an irreplaceable role in data science projects. While data scientists are responsible for transforming raw data into data products, domain experts are also required to validate, explain and implement those products. Enabling man-machine collaborative data science, we propose a new data science life cycle model with nine steps:

1. business understanding,
2. datafication,
3. data wrangling or munging,
4. data analysis,
5. data understanding,
6. data insights,
7. visualizing/storytelling/communicating,
8. data products DevOps, and
9. decision-support or automated decision-making (Figure 4).

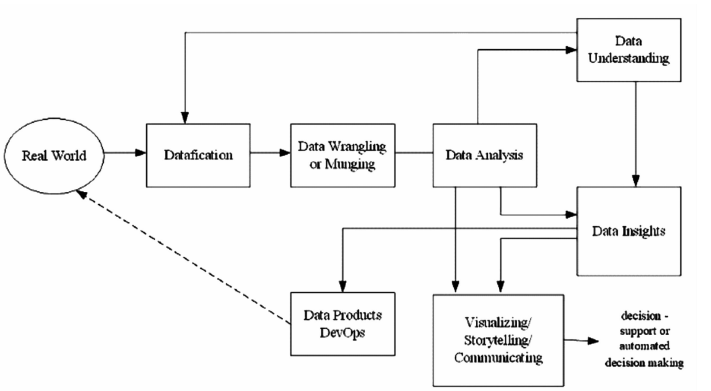

Figure 4 Data science life cycle adapted from O'Neil, Cathy, and Rachel Schutt. Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.", 2013.

3.4 Data Wrangling or Munging

Data wrangling (or data munging) is one of the novel concepts commonly employed by data scientists since it reflects the shift in the main concerns in data preprocessing. Wickham (2014) demonstrated how to transform messy data into tidy data using a set of tools with R. Endel and Piringer (2015) proposed that data wrangling is not only about transforming and cleaning procedures, and other aspects like data quality, merging of different sources, reproducible processes, and managing data provenance must be considered. Jiang and Kahn (2020) insisted that data wrangling is a strategy for selecting, managing, and aggregating datasets to produce a model and story. Azeroual (2020) discussed the main steps for data wrangling: exploring, structuring, cleaning, enriching, validating, and publishing. Kandel et al. (2011) used visualization methods such as graphics and charts to identify data quality problems and data wrangling.

In contrast to conventional data preprocessing, data wrangling is supposed to be a value-adding process. It concentrates on applying data scientists' creative design skills, critical thinking, and curiosity to data processing tasks. Data wrangling is a new type of data preprocessing involving data cleansing and tidying. Data cleansing is converting dirty data into clean data by enhancing data quality. Alternatively, data tidying refers to transforming messy data into tidy data by reshaping or reformatting data.

Note that data wrangling usually causes information loss or information distortion. Some valuable information may be lost when transforming unstructured data into structured data, when it cannot be directly stored in a structured form. Also, it is possible that the original meaning of the data is distorted when the data are converted from one format into another. Therefore, data scientists have to find a trade-off between data wrangling and information loss.

3.5 Big Data Analytics

Big data analytics has been widely discussed and is the most advanced topic in data science. A few researchers were under the impression that data science was only a new alternative name for big data analytics. For instance, Nakamura (2020) regarded big data analytics as data science. In practice, data science provides broader insights and focuses on what questions should be asked, while big data analysis emphasizes finding answers to the questions asked (Nadikattu, 2020). Big data analytics is one of the stages in a data science life cycle, and data analysis systems must provide effective mechanisms to design and complete analysis tasks (Elshawi, 2018). Tsai et al. (2015) discussed the development of a high-performance big data analytics platform and appropriate mining algorithms in the entire process of knowledge discovery in databases. Kambatla et al. (2013) described the application prospects of big data analytics and suggested that some computing work should be transferred to the data source itself in the future system. Swan (2013) argued that the Quantified Self (QS) is a challenge in data science and that big data analytics can provide new insights into QS and other biological issues.

One of the hottest topics in big data analytics is the development of tools and technologies for data science projects. There have been some established tools to help data scientists and data analysts perform tasks related to big data analytics. Most big data analysis projects adopt Hadoop and Hadoop-related technologies to provide novel solutions. The Apache Hadoop platform is deemed to be composed of related projects, including HDFS, YARN,

MapReduce, Pig, Hive, and HBase (Bappalige, 2014). Also, there is a vast range of companies devoted to advance technologies for big data analytics. Databricks, for instance, provides a Spark-based unified analytics engine for large-scale data analytics. Spark is an open-source cluster computing system based on in-memory computing that aims to make data analysis faster. Currently, it is one of the most popular technological solutions for big data analytics.

Studies on big data analytics face the following challenges: difficulty in storing vast volumes of data and lack of professionalized analytics tools. Stephens et al. (2015) proposed that CPU capacity might not be the bottleneck of future big data analysis; the bottleneck lies in the input/output hardware that transfers data between storage and processor. Business applications need real-time big data analytics to implement dynamic auto-decisions, which require big data analytics tools to process more data in less time.

3.6 Data Management and Data Governance

Data management and data governance represent the management facets of data science. With big data playing an increasingly important role in governments, enterprises, and institutions, big data management or big data governance is becoming one of the main concerns of relevant studies.

Typically, data management possesses a broader scope than data governance (see Figure 5). Data management maturity (DMM) lists 25 KPs required for organizational data management. Further, it categorizes them into six key process areas: (1) data management strategy, (2) data governance, (3) data quality, (4) platform and architecture,(5)data operations, and (6) supporting processes. Data governance, as defined by DMM, involves three KPs: (1) governance management,(2) business glossary, and (3) metadata management. Also, the standard entitled "Information Technology Services—Governance Part 5: Data Governance Specification" issued by the China National Information Technology Standardization Network (Standards China, 2018) defines data management as the collection of activities in which data resources are acquired, controlled, and promoted value. In that document, "data governance" refers to the collection of related governance activities, performance, and risk management in data resources and their applications.

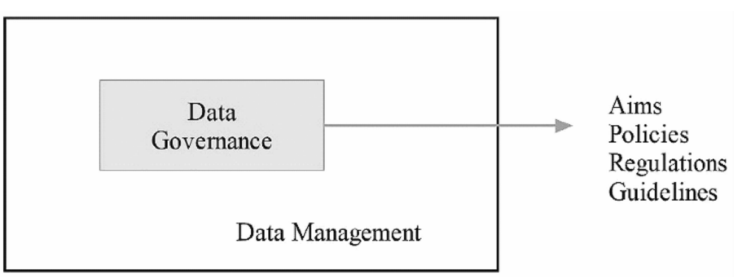

Figure 5 Data governance and data management

Data governance provides the aims, policies, regulations, guidelines, tools, and solutions for ensuring successful data management activities. Mathur and Purohit (2017) argued that it is necessary to deal with the main problems of access, metadata, utilization, update, governance, and reference. Ranjan et al. (2018) designed data management components, including data governance, data analysis, and data warehousing from the perspective of the Internet of Things (IoT). Bakken and Koleck (2019) summarized the benefits as well as challenges

of data governance, data science infrastructure, and data science pipelines from a nursing perspective.

The DGI Data Governance Framework (2020) proposed by the Data Governance Institute is adopted as one of the best data governance practices. It is a logical framework for classifying, organizing, and transmitting complex enterprise data. Data governance tasks usually have three stages:

1. Define or order rules of engagement;
2. Identify the relevant people and administrative bodies, especially the data stakeholders, a data governance office, and data stewards;
3. Divide and implement specific governance processes (Thomas, 2020).

3.7 Data Products DevOps

Providing data products is one of the ultimate research motivations of data science. In data science, data products refer to all products developed based on data. In other words, data products involve not only products in the form of data but also products that use data to help users achieve one or some of their goals (Patil, 2012). Data products include datasets, documents, databases, software, hardware, services, insights, decisions, and their various combinations.

Developing data products is the hallmark of data science as a distinct new discipline. Data products typically incorporate six main features (see Figure 6):

(1) providing data-based (fact-based) solutions rather than knowledge-based solutions,

(2) addressing data-intensive problems in preference to computing-intensive tasks,

(3) being driven by data instead of hypothesis,

(4) conforming to data-analytic thinking in place of intuition-based thinking,

(5) adopting data-centered architectures as a substitute for app-centered architectures,

(6) creating value from data in favor of creating data for value. Therefore, data products are fact-based and are beyond the limits of intuition. Key aspects that define the types of data products or services include intellectual property rights, licensing terms, and type of owner (Pantelis & Aija, 2013).

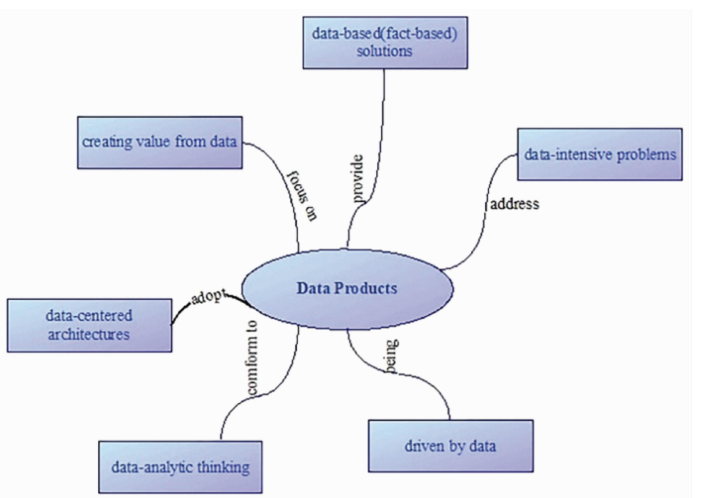

Figure 6 Six main features of data products

The methodology for developing data products is one of the hot topics in data science studies. Patil (2012) coined a new term "data jujitsu" that refers to the art of converting data into products. Further, he provided 13 underlying principles of data jujitsu. In the industry, data products development is a required skill for data scientists and big data analysts. Data science is located at the intersection of statistics, ML, and domain knowledge (Schutt & O'Neil, 2013). Li et al.(2019) discussed how engineers and data scientists could effectively collaborate on new products development in a hybrid team with data-driven features. Online data products, such as search engines developed by companies such as Yahoo and Google, could be used as mobile applications, forming the so-called "App Economy" (Davenport & Kudyba, 2016). To leverage the full potential of data science, user experience analysis should be included in the design process, and user testing should be part of the project life cycle (Joshi, 2021). Consequently, A/B testing is adopted as a common tool to evaluate and improve user experience of data products.

3.8 Big Data Visualization

Big data visualization is a fundamental building block of data science in that visualization is one of the most effective ways to reveal the hidden information behind big data explicitly. However, big data's diversity brings challenges to traditional data visualization methodologies since semi-structured and unstructured data are challenging to process (Jin et al., 2015). Real-time scalability and interactive scalability are the main challenges that limit the presentation of big data, while data reduction and reducing latency are better ways to present big data (Agrawal et al., 2015). Ali et al. (2016) argued that choosing the dimensions of data to be visualized, low performance, visual noise, information loss, large image perception, high rate of image change, and high-performance requirements are substantial challenges faced for big data visualization tools.

The industry has provided a variety of big data visualization tools, including Tableau, D3.js, Power BI, Infogram, and Google charts. Big data visualization tools should have new capabilities to handle various data formats, to import/export data or share visualization results with other tools, to provide collaborative working space, and to improve user experiences.

Visual analytics was proposed in 2004 by a working group at the National Visualization and Analytics Center (NVAC) (Cook & Thomas, 2005). It aims to combine the flexibility, creativity, and background knowledge of humans with the vast storage and fast processing power of computers to gain big data insights to address complex problems. Visual analytics is a promising field for data visualization in data science studies.

4 EMERGING TRENDS

Six main trends characterize the recent theoretical studies on data science:

1. the growing significance of DataOps,
2. the rise of citizen data scientists,
3. enabling augmented data science,
4. integrating data warehouse with data lake,
5. diversity of domain-specific data science, and
6. implementing data stories as data products.

4.1 Growing Significance of DataOps

The motivation of DataOps is to combine DevOps and Agile methodologies to manage

data in alignment with business goals (Vaughan, 2019). DevOps is a blend of development (representing software developers, including programmers, testers, and quality assurance personnel) and operations (representing the experts who put software into production and manage the production infrastructure, including system administrators, database administrators, and network technicians) (Hüttermann, 2012). Capizzi et al. (2019) proposed that DataOps aims to deploy data flow pipelines and toolchains in a cloud environment for real-time adjustment of pipelines to meet actual operational needs. In contrast to traditional software development methodologies, DevOps improves communication or collaboration among those in charge of the software deployment process and aims to produce higher-quality products faster and reliably.

One of the significant trends in data science is integrating DataOps with MLOps. As one of the building blocks of data science, ML provides big data analysis with mythological foundations. Data science usually leverages MLOps to deploy ML models in data science projects reliably and efficiently. MLOps enables data scientists to monitor, validate, and govern ML models throughout the process; collaborate with other business people; and enhance the speed and quality of delivery for model development (Soh & Singh, 2020).

4.2 The Rise of Citizen Data Scientist

Citizen data scientist is a new topic in data science. It is the kind of person who creates or generates models that use advanced diagnostic analytics or predictive and prescriptive capabilities but whose primary job function is outside the field of statistics and analytics (Gartner, 2016). In 2016, citizen data scientists came to prominence because users throughout the business world wanted a democratized approach to big data and analytics (Shacklett, 2016). Citizen data scientists possess more expertise than professional ones with regard to particular application domains (see Table 1).

Table 1 Citizen data scientist versus expert data scientist

	Citizen data scientist	Professional data scientist
Primary job function	Outside the field of data science and big data analytics	The field of data science and big data analytics
Ability to understand business requirements	Higher	Lower
Domain–specific expertise	More	Less
Readiness of data science knowledge or skills	Good	Bad
Data science coding capability	Higher	Lower
Roles in data science projects	To select, interpret, and evaluate the candidate solutions proposed by professional data scientist	To provide candidate solutions for citizen data scientists

The rise of citizen data scientists indicates that data science practices are dependent on domain expertise. The selection, interpretability, and evaluation of models in data science projects require knowledge or skills from the corresponding fields. Typically, citizen data scientists focus on using data science tools but usually lack the ability to understand the underlying principles of these tools. However, understanding these principles is crucial for

selecting algorithms, optimizing models, and tuning their hyperparameters. The roles of citizen data scientists and professional data scientists complement each other, and collaboration between them is a new trend in data science practices.

4.3 Enabling Augmented Data Science

Augmented data science is a data-driven method in which software tools automatically conduct data exploration and processing to assist data scientists in making decisions (Uzunalioglu et al., 2019). Augmented data science stems from augmented analytics. Augmented analytics is a next-generation data analytics paradigm that uses ML to automate data preparation, insight discovery, and insight sharing for a broad range of business users, operational workers, and citizen data scientists (Gartner, 2017). There are three main trends for augmented data analytics:

1. augmented data preparation,
2. augmented analytics as part of analytics and business intelligence,
3. augmented data science or ML (Gartner, 2018).

Augmented analytics could implement automatic analysis, reduce the difficulty of data analysis for non-professional users, and help data scientists carry out data analysis tasks efficiently and effectively.

Augmented data science is redefining the roles of man and machine in relevant practices. Augmented data science simultaneously enhances the return on data science investments, and reduces time to value, and expands the ML footprint. Experts become efficient and productive, and a broader population of quantitative professionals could succeed in data science (Gartner, 2019). Augmented data science would promote the collaboration between data scientists and scientists of specific application domain; thus, the human-machine collaborative working pattern would be the first choice of data science solutions.

4.4 Integrating Data Warehouse with Data Lake

Recent trends in data science have led to a proliferation of studies that intend to integrate traditional data warehouses with data lakes. There are complementary advantages between data warehouses and data lakes from a data science perspective. For data science, data lakes provide a convenient storage layer for experimental data, both the input and output of data analysis and learning tasks (Nargesian, 2019). In sharp contrast to traditional data warehouse technologies, data lakes support all data types, load all data from their source system, and retain them in an untransformed or nearly untransformed state. Therefore, integrating data warehouses with data lakes is the key to data science projects.

Most data science platforms will be built on a data lakehouse that combines data warehouses and data lakes. A data lakehouse is a new generation of an open platform that unifies data warehousing and big data analytics (Armbrust et al., 2021). Databricks lakehouse, for instance, unifies data, analytics, and AI to provide a collaborative working platform for data science projects (Databricks, 2021). Consequently, data lakehouse is becoming one of the most commonly used solutions for data storage layers in data science.

4.5 Diversity of Domain–Specific Data Science

Introducing data science to other specific application domains has been a hot topic in recent studies. These studies could be categorized into two groups: domain-general data science and domain-specific data science. The former regards and nurtures data science as

an independent new discipline. However, the latter discusses data science from a specific application discipline. A new research trend of using data science for comprehensive studies can be seen in traditional disciplines; hence, domain-specific data science has become an emerging topic in application disciplines. Data science is applied in life science, health care, government, education, and business management. Some new research topics, in turn, emerge from these application areas, such as quantitative self, data journalism, and big data analysis.

The diversity in domain-specific studies will advance a new research direction called theoretical data science that bridges the gap between distinct domain-specific studies. Theoretical data science is a new branch of data science which employs mathematical model abstractions of data objects and systems to rationalize, explain, and predict big data phenomena (Borjigin et al., 2021). Consequently, theoretical data science will further boost the development of domain-general data science. The interdisciplinary research on data science will not only provide efficient data science tools but also facilitate the communication between data scientists and domain experts.

4.6 Implementing Data Story as Data Products

Data storytelling is an emerging research direction in data science. Essentially, data storytelling is a form of persuasion that employs data, narrative, and visuals to help an audience see something in a new light and convince them to act (Dykes, 2019). Storytelling and visualization are complementary approaches for presenting big data in data science studies. Data visualization is widely adopted in data storytelling in that a story needs to be visualized to make key observations and details to build a picture in someone's mind (Martin, 2018). Data visualization is a literary device to tell stories with data, and they are two halves of the same coin (Ryan, 2018).

Data stories will be an alternative type of data product. However, data story and literacy story differ in the following aspects (see Table 2):

- Motivation: Data stories are only designed to meet a given business requirement, while literary stories are created for general purposes, such as entertainment, education, and recreation for all the audience. Data story only works for the target users in a specific business life cycle.
- Content: The content of a data story has to be sourced from actual business data, but that of a literacy story can stem from imagination, life experience, or hearsay.
- Creator: Data stories are automatically created by algorithms, whereas human beings directly write a conventional literacy story.
- Lifespan: The lifespan of a data story is shorter than that of a literacy story in that the former is strictly restricted to the corresponding business life cycle. An excellent data story will expire when tasks of the business process are completed, while an excellent literary story could be passed from generation to generation.

Table 2 The difference between the data story and the literary story

	Data Story	Literary Story
Motivation	specific	general
Content	real	fiction
Creator	machine	man
Lifespan	short	long

5 OPPORTUNITIES AND CHALLENGES

The future development of data science should prioritize turning the four most acute challenges into opportunities: (1) accelerating theoretical studies of data science, (2) the trade-off between explainability and performance, (3) achieving data ethics, privacy, and trust, and (4) aligning academic curricula with industrial needs.

5.1 Accelerating Theoretical Studies of Data Science

The most significant weakness of data science to date is the lack of systematic theoretical studies. Despite the fact that data science is one of the hottest topics in recent academic studies, the in-depth study of its theoretical framework is overlooked. There are no shared understandings of the theoretical data science system and its essential components. Furthermore, a few researchers tend to misuse data science as a new name for some old approaches to data analysis or data processing, such as ML, statistics, data engineering, or business intelligence. This weakness is becoming a new bottleneck for the future development of data science.

Theoretical studies on data science can be promoted by integrating domain-general data science with a diverse range of domain-specific data science. Borjigin et al. (2021) proposed a new term "theoretical data science" to bridge the gap between the domain-general and domain-specific studies and provide its five essential topics: (1) to conduct in-depth theoretical research on data science, (2) to take advantage of the active property of big data, (3) to introduce design of experiments into data science studies, (4) to shift data science' research focus from correlation analysis into causality inference, and (5) to consider data products development as one of the main tasks of data science projects. Also, expanding the technical areas of today's consensus data science is crucial to theoretical studies of data science. Donoho (2017) proposed a new field called greater data science that is a better academic enlargement of statistics and ML than today's data science initiatives, while accommodating the same short-term goals.

5.2 The Trade–off Between Explainability and Performance

The most critical challenge in data science practice is balancing its interpretability with performance. Explainability and effectiveness are goals that have to be considered for designing models for data science practice (Zhang & Chen, 2018). By default, simple models should be used as much as possible unless the explainer explicitly asks for more complex ones (Sokol & Flach, 2020). Explainability must consider the trade-off between accuracy and fidelity and strike a balance between accuracy, explainability, and ease of processing (Gunning et al., 2019).

The motivations of data science projects should be shifted from identifying correlation to inferring causation. There has been a common mistake that data science focuses merely on correlation rather than causation. During the earliest stages of data science, researchers tend to focus on correlation instead of causation. However, ignoring causal analysis results in lower trust in data science solutions.

Expertise in experimental design can help address the gap between correlation and causation (McAfee & Brynjolfsson 2012). Besides, explainable artificial intelligence (XAI) provides a new solution for balancing interpretability and performance. Existing XAI studies could be divided into various groups from two distinct dimensions (Rai, 2019):

Whether the technique was model-specific or model-agnostic, model-specific techniques only work with a given ML model. By contrast, model-agnostic techniques can be employed in various ML models. For instance, a model-agnostic technique called LIME was used to perturb input samples to observe the impact on the output results, whether the technique provided a global explanation or a local explanation. The global explanation demonstrates how to explain the model as a whole, involving the algorithm selection, training process, and trained results. Alternatively, the local explanation aims to help people understand the decision-making process of the trained model for a given input sample. Also, inferring causality in data science needs to integrate domain-general data science with diverse domain expertise.

5.3 Achieving Data Ethics, Privacy, and Trust

Data ethics, privacy, and trust problems are the potential risks for data science practices. Data security threats come from a diverse range of factors, including confidentiality, integrity, availability, and privacy (Talha et al., 2019). Furthermore, an ethical expert should be included in a data science project to avoid "Bias In, Bias Out (BIBO)" (Ho & Beyan, 2020). Data bias, such as survivorship bias and Simpson's and Bergson's paradoxes, probably occurs at any stage in the data science life cycle. Explainable artificial intelligence is an approach to verifying the presence of algorithmic bias (Sen et al., 2020). Besides, user authentication and consent regarding the use of personal data are critical for protecting ethics and privacy.

Data masking and data auditing are essential to achieving data ethics, privacy, and trust. Data masking is implemented by replacing or deleting original personal (or organizational) sensitive data without affecting the accuracy of the data analysis results to avoid security risks and privacy issues. Data auditing can help data scientists ensure data integrity, control data quality, and prevent data leakage. Data masking and auditing are effective ways to achieve data ethics, privacy, and trust in data science projects. They are essential for data scientists to gain insights from big data following the user's preference.

5.4 Aligning Data Science Curricula with Industrial Needs

The shortage of data scientists is becoming a serious constraint in some sectors (Davenport et al., 2012). The main challenges of higher education in cultivating qualified data scientists are rooted in three factors:

1. the curriculum is loosely coupled with data science practices; therefore, data science major is merely an alternative title for traditional majors, notably statistics or ML;
2. some essential courses such as exploratory data analysis, design of experiment, causality, and data product design are missing in data science majors;
3. student's poor capability to address real-world challenges.

At present, there is no single model in terms of the department, school, or cross-unit collaboration within higher education institutions that should take responsibility for data science education.

A study conducted by Bojigin et al. (2011) found that the qualifications for data scientists could be divided into two categories: data science-specific qualifications and general purpose-oriented ones. Data science-specific qualifications include SQL programming, Python/R/SAS, Hadoop MapReduce/HBase/Hive, Spark/Storm, Visual Analysis with Tableau, ETL, Data Warehouse/Data Lake/BI, Statistics, ML (including deep learning), natural language processing, text analysis, and computer vision. General purpose-oriented qualifications

involve the candidate's readiness for communication and cooperation, problem-solving, 3C characteristics of data scientists, independent learning, attention to detail, stress management, and leadership skills (Bojigin et al., 2021). Therefore, the top-level design of data science curriculums in higher education should meet these industry needs as well as their further evolutions.

6 CONCLUSIONS

The gap between data and knowledge is at the highest level in the early stages of the big data era. Traditional knowledge cannot match the new data created by or stored in cloud computing, the IoT, mobile internet, and emerging scientific instruments or manufacturing equipment. The contradiction between new data and traditional knowledge is the biggest challenge faced by most traditional sciences. Computer science and statistics first perceived this challenge and proposed a new science called data science from their distinct perspectives. Then, other disciplines noticed that gap and conducted interdisciplinary studies. Consequently, data science is a rapidly growing academic field and has tremendous implications for all traditional studies today. However, the relevant studies failed to conduct in-depth research on the theoretical system of data science, and there is doubt whether data science can be considered an independent science.

A significant finding from this study is that data science includes unique research goals, distinct perspectives, and an independent body of knowledge. The study contributes to our understanding of core theories, recent developments, and emerging trends in data science. Future work should focus on establishing theoretical systems of data science, especially to accelerate theoretical studies of data science, address the paradox between usability and interpretability of big data solutions, achieve big data ethics, privacy, and trust, and align data science curriculums with industrial needs.

REFERENCES

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16* (1), 3–9.

Agrawal, R., Kadadi, A., Dai, X., & Andres, F. (2015). Challenges and opportunities with big data visualization. In *Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems (MEDES ´15) , October 25, 2015, Caraguatatuba, Brazil.* Association for Computing Machinery, New York, NY, USA, 169–173. DOI:https://doi.org/10.1145/2857218.2857256

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016). Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), December 14–17, 2016,* Noida, India. IEEE Inc., Piscataway, N J, 656–660. https://doi.org/10.1109/IC3I.2016.7918044

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). *Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics.* Retrieved September 17, 2021 from https://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

Azeroual, O. (2020). Data wrangling in database systems: Purging of dirty data. *Data, 5* (2), 50. DOI: https://doi.org/10.3390/data5020050

Bappalige, S. P. (2014). *An introduction to Apache Hadoop for big data*. Retrieved September 17, 2021 from https://opensource.com/life/14/8/intro–apache–hadoop–big–data

Beck, D. A. C., Carothers, J. M., Subramanian, V. R., & Pfaendtner, J. (2016). Data science: Accelerating innovation and discovery in chemical engineering. *AIChE Journal, 62* (5), 1402–1416. DOI: https://doi.org/10.1002/aic.15192

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. S. (2018). Realizing the potential of data science. *Communications*

of the ACM, 61 (4), 67–72. DOI:https://doi.org/10.1145/3188721

Boehm, M., Antonov, I., Baunsgaard, S., et al. (2019). SystemDS: A declarative machine learning system for the end–to–end data science lifecycle. arXiv: 1909.02976. Retrieved from https://arxiv.org/abs/1909.02976

Borjigin, C., Xiao, J., & Wang, X. (2021). Typical responsibilities, key qualifications and higher education for data scientist. *Journal of Library Science in China, 47* (3),100–112. DOI:https://doi.org/10.13530/j.cnki.jlis.2021023

Borjigin, C., Zhang, C., Sun, Z & Yi, N. (2021). Theoretical data science: Bridging the gap between domain–general and domain–specific studies. *Data Science and Informetrics, 1* (1), 1–28. DOI:https://doi.org/CNKI:SUN:DSIR.0.2021–01–002.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys, 50* (3), 1–42. DOI:https://doi.org/10.1145/3076253

Capizzi, A., Distefano, S., & Mazzara, M. (2019). From DevOps to DevDataOps: Data management in DevOps processes. In *International Workshop on Software Engineering Aspects of Continuous Development and New Paradigms of Software Production and Deployment (DEVOPS 2019), January 19, 2020, Villebrumier, France.* Springer, Cham, Switzerland AG, 52–62. https://doi.org/10.1007/978–3–030–39306–9_4

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International statistical review, 69* (1), 21–26. DOI: https://doi.org/10.1002/sam.11239

Cook, K. A., & Thomas, J. J.(2005). Illuminating the path: The research and development agenda for visual analytics. *Retrieved November 28,* 2021 from https://www.osti.gov/biblio/912515

Databricks. (2021). *The Databricks Lakehouse Platform*. Retrieved from https://databricks.com/product/data–lakehouse.

Davenport, T. H., & Kudyba, S.(2016). Designing and developing analytics–based data products. *MIT Sloan Management Review, 58* (1), 83–89. Retrieved from https://www.proquest.com/scholarly–journals/designing–developing–analytics–based–data/docview/1831862457/se–2?accountid=13625.

Davenport, T. H. (2013). Analytics 3.0. *Harvard Business Review, 91* (12), 64–72.

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review, 90* (5), 70–76.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26* (4), 745–766. DOI:https://doi.org/10.1080/10618600.2017.1384734

Dykes, B. (2019). *Effective data storytelling: How to drive change with data, narrative and visuals* (1nd ed.). John Wiley & Sons, New York, NY.

Elshawiab, R., Sakrbc, S., Taliad, D., & Trunfio, P. (2018). Big data systems meet machine learning challenges: Towards big data science as a service. *Big Data Research, 14, 1–11*. DOI:https://doi.org/10.1016/j.bdr.2018.04.004

Endel, F., & Piringer, H. (2015). Data Wrangling: Making data useful again. *IFAC–PapersOnLine, 48* (1), 111–112. DOI:https://doi.org/10.1016/j.ifacol.2015.05.197

Gartner. (2016). *Citizen data science augments data discovery and simplifies data science.* Retrieved from https://www.gartner.com/en/documents/3534848

Gartner. (2017). *Augmented analytics is the future of data and analytics.* Retrieved from https://www.gartner.com/en/documents/3773164

Gartner. (2018). *Top 10 strategic technology trends for 2019*. Retrieved from https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/3891569–top–10–strategic–technology–trends–for–2019.pdf

Gartner. (2019). *How augmented machine learning is democratizing data science*. Retrieved from https://www.gartner.com/en/documents/3956825–how–augmented–machine–learning–is–democratizing–data–sci

Gibert, K., Horsburgh, J. S., Athanasiadis, L. N., & Holmes, G. (2018). Environmental data science. *Environmental Modelling & Software, 106, 4–12*. DOI:https://doi.org/10.1016/j.envsoft.2018.04.005

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L.(2009). Detecting influenza epidemics using search engine query data. *Nature, 457* (7232), 1012–1014. DOI:https://doi.org/10.1038/nature07634

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf,S., & Yang, G. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4* (37). DOI:https://doi.org/10.1126/scirobotics.aay7120

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data–Intensive scientific discovery*. Microsoft Re–

search, Redmond, Washington, WA.

Hill, K. (2012). *How target figured out a teen girl was pregnant before her father did*. Retrieved September 17, 2021 from http://elearning.algonquincollege.com/coursemat/haugs/1_F_19_BUS2303/SECURITY% 20ARTI–CLE%20%20Target–Teen–Pregancy–Forbes.pdf.

Ho, D., & Beyan, O. (2020). *Biases in data science lifecycle*. arXiv: 2009.09795. Retrieved from https: //arxiv. org/abs/2009.09795

Hüttermann, M. (2012). *DevOps for developers.* Apress, New York, NY.

Jiang, S., & Kahn, J. (2020). Data wrangling practices and collaborative interactions with aggregated data. *International Journal of Computer–Supported Collaborative Learning, 15* (3), 257–281. DOI: https: //doi.org/10.1007/s11412–020–09327–1

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research, 2* (2), 59–64. DOI:https://doi.org/10.1016/j.bdr.2015.01.006

Joshi, M. P., Su, N., Austin, R. D., & Sundaram, A. K. (2021). Why so many data science projects fail to deliver. *MIT Sloan Management Review, 62* (3), 85–89. Retrieved from https://www.proquest.com/scholarly–journals/why–so–many–data–science–projects–fail–deliver/docview/2516954047/se–2?accountid=13625

Kalidindi, S. R., & De Graef, M. (2015). Materials data science: Current status and future outlook. *Annual Review of Materials Research, 45* (1), 171–193. DOI:https://doi.org/10.1146/annurev–matsci–070214–020844

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of parallel and distributed computing, 74* (7), 2561–2573. DOI:https://doi.org/10.1016/j.jpdc.2014.01.003

Kandel, S., Heer, J., Plaisant, C., et al. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization, 10* (4), 271 –288. DOI:https://doi.org/10.1177/1473871611415994

Kelleher, J. D., & Tierney, B. (2018). *Data science.* MIT Press, Cambridge, MA.

Larsona, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36* (5), 700– 710. DOI:https://doi.org/10.1016/j.ijinfomgt.2016.04.013

Li, Y., Roy, U., & Saltz, J. S.(2019). Towards an integrated process model for new product development with data–driven features (NPD3). *Research in Engineering Design, 30* (2), 271–289. DOI:https://doi.org/10.1007/s00163–019–00308–6

Martin, N. (2018). *Data visualization: How to tell a story with data.* Retrieved September 17, 2021 from https:// www. forbes.com/sites /nicolemartin1/2018/11/01/data–visualization–how–to–tell–a–story–with–data/

Mathur, & Purohit, R. (2017). Issues and challenges in convergence of big data, cloud and data science. *International Journal of Computer Applications, 160* (9), 7–12. DOI:https://doi.org/10.5120/ijca2017913082

McAfee, A., & Brynjolfsson, E.(2012). Big data: The management revolution. *Harvard Business Review, 90* (10), 60 –68. Retrieved from https://wiki.uib.no/info310/images/4/4c/McAfeeBrynjolfsson2012 –BigData –TheManagementRevolution–HBR.pdf.

Nadikattu, R. R. (2020). Research on data science, data analytics and big data. *International Journal Of Engineering, Science And, 9* (5), 99–105. DOI: http://dx.doi.org/10.2139/ssrn.3622844

Nakamura, H. (2020). Big data science at AMED–BINDS. *Biophysical Reviews, 12* (2), 221–224. DOI:https://doi.org/10.1007/s12551–020–00628–1

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment, 12* (12), 1986–1989. DOI:https://doi.org/10.14778/3352063.3352116

Naur, P. (1974). *Concise survey of computer methods.* Petrocelli Books, New York, NY.

O´Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline.* O´Reilly, Sebastopol, CA.

Pantelis K., & Aija, L. (2013). Understanding the value of (big) data. In *Proceedings of the 2013 IEEE International Conference on Big Data, October 6–9, 2013, Silicon Valley, CA. IEEE, Piscataway, NJ, 38–42.* https://doi.org/10.1109/bigdata.2013.6691691

Patil, D. J. (2012). *Data Jujitsu.* O´Reilly Media, Inc. Sebastopol, CA.

Peek, N., & Rodrigues, P. P. (2018). Three controversies in health data science. *International Journal of Data*

Science and Analytics, 6 (3), 261–269. DOI:https://doi.org/10.1007/s41060–018–0109–y

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data–analytic thinking.* O´Reilly Media, Inc. Sebastopol, CA.

Rai, A. (2019). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48* (1), 137–141. DOI:https://doi.org/10.1007/s11747–019–00710–5

Ranjan, R., Rana, O., Nepal, S., et al.. (2018). The next grand challenges: Integrating the Internet of Things and data science. *IEEE Cloud Computing, 5* (3), 12–26. DOI:https://doi.org/10.1109/MCC.2018.032591612

Ryan, L. (2018). *Data visualization and data storytelling: A visual revolution.* Retrieved September 17, 2021 from https://www.dbta.com/BigDataQuarterly/Articles/Data–Visualization–and–Data––Storytelling–A–Visual–Revo–lution–124077.aspx

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters P., & Ng, A. (2020). Cybersecurity data sci–ence: An overview from machine learning perspective. *Journal of Big Data, 7* (1). DOI:https://doi.org/10.1186/s40537–020–00318–5

Sen, S., Dasgupta, D., Gupta, K. D. (2020). An empirical study on algorithmic bias. In Proceedings of the *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), July 13–17, 2020. IEEE, Piscataway, NJ, 1189–1194*. https://doi.org/10.1109/compsac48688.2020.00–95

Shacklett, M. (2016). Why 2016 might be the year of citizen data scientists. *Retrieved September 17,* 2021 from https://www.techrepublic.com/article/why–2016–may–be–the–year–of–the–citizen–data–scientist/

Singleton, A., & Arribas–Bel, D. (2021). Geographic data science. *Geographical Analysis, 53* (1) , 61–75. DOI: https://doi.org/10.1111/gean.12194

Soh, J., & Singh, P. (2020). Machine learning operations. In: Data Science Solutions on Azure. Apress, Berke–ley, CA. Retrieved September 17, 2021 from https://doi.org/10.1007/978–1–4842–6405–8_8

Sokol, K., & Flach, P. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ´20)*. Association for Computing Machinery, New York, NY, USA, 56 – 67. DOI:https://doi.org/10.1145/3351095.3372870

Song, I., & Zhu, Y. (2017). Big Data and Data Science: Opportunities and Challenges of iSchools. *Journal of Data and Information Science 2, 3* (April 2017), 1– 18. DOI:https://doi.org/10.1515/jdis–2017–0011

Standards China. (2018). Information technology services—Governance part 5: Data governance specification (GB/T 34960.5–2018). SAC/TC 28.

Stephens, Z. D., Lee, S. Y., Faghri, F., et al. (2015). Big data: Astronomical or genomical?. *PLOS Biology, 13* (7), e1002195. DOI:https://doi.org/10.1371/journal.pbio.1002195

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data, 1* (2), 85–99. DOI:https://doi.org/10.1089/big.2012.0002

Talha, M., Kalam, A. A. E., & Elmarzouqi, N. (2019). Big Data: Trade–off between data quality and data securi–ty. *Procedia Computer Science, 151,* 916–922. DOI:https://doi.org/10.1016/j.procs.2019.04.127

Taylor, D. (2017). *Battle of the data science Venn diagrams.* Retrieved September 17, 2021 from https://deeplearning.lipingyang.org/wp–content/uploads/2017/10/Battle–of–the–Data–Science–Venn–Diagrams.pdf.

Thomas, G. (2020). *DGI data governance framework.* Retrieved September 17, 2021 from https://opensource.com/life/14/8/intro–apache–hadoop–big–data

Tsai, C., Lai, C., Chao, H., & Vasilakos, A. V. (2015). Big data analytics: A survey. *Journal of Big data, 2* (1), 1–32. DOI:https://doi.org/10.1186/s40537–015–0030–3

Ullman, J. D. (2020). *The battle for data science.* Retrieved September 17, 2021 from http://sites.computer.org/debull/A20june/p8.pdf

Uzunalioglu, H., Cao, J., Phadke, C., Lehmann, G., Akyamac, A., He, R., Lee, J., & Able, M. (2019). Augment–ed data science: Towards industrialization and democratization of data science. arXiv:1909.05682. Retrieved from https://arxiv.org/abs/1909.05682.

Vaughan, J. (2019). *DataOps (data operations).* Retrieved September 17, 2021 from https://searchdatamanage–ment.techtarget.com/definition/DataOps

Wang, D., Andres, J., Weisz, J. D., Oduor, E., & Dugan, C.(2021). AutoDS: Towards human–centered automa–

tion of data science. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery*, New York, NY, USA, Article 79, 1–12. DOI:https://doi.org/10.1145/3411764.3445526

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59* (10), 1–23. DOI: https://doi.org/10.18637/jss.v059.i10

Zhang, Y., & Chen, X.(2018). *Explainable recommendation: A survey and new perspectives.* arXiv:1804.11192. Retrieved from https://arxiv.org/abs/1804.11192

Zhang, Z., Chang, R., Dai, Y., & Zhao, R. (2021). Research on the evaluation model and demonstration of patentee′ s discourse power: A case study of cyber security. *Data Science and Informetrics, 1* (3), 93–108.

Zhu, Y., & Xiong, Y.(2015). Towards data science. *Data Science Journal, 14*, 8. DOI:https://doi.org/10.5334/dsj–2015–008