Data Science: Trends, Perspectives, and Prospects

Chaolemen Borjigin (✉ chaolemen@ruc.edu.cn)

Renmin university of China

Chen Zhang

Renmin university of China

Research

Keywords: data science, big data, data products, data-driven management, The DIKW pyramid

Posted Date: November 1 st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1014621/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Data Science is one of today's most rapidly growing academic fields and has significant implications for all conventional scientific studies. However, most of the relevant studies so far have been limited to one or several facets of Data Science from a specific application domain perspective and fail to discuss its theoretical framework. Data Science is a novel science in that its research goals, perspectives, and body of knowledge is distinct from other sciences. The core theories of Data Science are the DIKW pyramid, data-intensive scientific discovery, data science lifecycle, data wrangling or munging, big data analytics, data management and governance, data products development, and big data visualization. Six main trends characterize the recent theoretical studies on Data Science: growing significance of DataOps, the rise of citizen data scientists, enabling augmented data science, diversity of domain-specific data science, and implementing data stories as data products. The further development of Data Science should prioritize four ways to turning challenges into opportunities: accelerating theoretical studies of data science, the trade-off between explainability and performance, achieving data ethics, privacy and trust, and aligning academic curricula to industrial needs.

1. Introduction

Data Science is gaining lots of momentum across a wide 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 Data Science first occurred as a scientific concept in Computer Science. In 2001, the statistician William S. Cleveland further proposed an action plan for expanding the technical areas of the field of statistics (Cleveland, 2001), and statistics is the second discipline that delineated Data Science. Hence, computer science and statistics are widely accepted as the two main theoretical foundations of Data Science. Drew Conway, the founder of Alluvium, in 2010, proposed Data Science Venn Diagram and first discussed the interdisciplinary of Data Science. He argued that Data Science locates at the core of the intersection of hacking skills, math and stats knowledge, and substantive expertise. His Venn diagram further embarked on more discussions with the interdisciplinarity of Data Science, and there are many variations of this Venn diagram (Ullman, 2020; Taylor, 2016). Now, Data Science turns to a hot topic of a variety of disciplines and nurtured new branches of traditional sciences, such as Geographic Data Science (Singleton & Daniel, 2021), Materials Data Science (Kalidindi & De Graef, 2015), Health Data Science (Peek & Rodrigues, 2018), Business Data Science (Provost & Fawcett, 2013), Environmental Data Science (Gibert et al., 2018), Surgical Data Science (Maier-Hein et al., 2017), and Cybersecurity Data Science (Sarker et al., 2020).

However, to start with, most related studies by far are dedicated to one or several facets of Data Science from their distinct domain perspective and fail to discuss its theoretical framework. This paper carried out an in-depth study of the Data Science theoretical framework based on comprehensive literature research and presented trends, perspectives, and prospective of Data Science. The remainder of this paper is organized as follows: Section 2 discussed its main research motivations, unique thinking pattern, the body of knowledge, as well as best practices. Then, section 3 described the core theories of Data Science and their recent progress, and Section 4 proposed the emerging trends of Data Science studies. Section 5 further provided some recommendations for the academic research or industrial application of Data Science. Finally, Section 6 presented its conclusion.

2. Data Science

Data Science is a new cross-disciplinary science that dealt with big data drawing on machine learning, statistics, and data visualization as its primary theoretical basis. Data Science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data at a high level. Data Science mainly focuses on

processing, computing, managing, analyzing big data, as well as providing data products. Data Science is a novel science in that its research goals, perspectives, and body of knowledge is distinct from the conventional sciences.

2.1 The Essential Research goal of Data Science

The essential goal of Data Science is to accelerate the inter-transformation between materials, energy and data, notably to reduce the consumption of materials and(or) energy or improve the effectiveness and(or) efficiency of exploiting them via taking advantage of data. The motivations of Data Science studies can be further 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 DevOps;
- (7) To cultivate and maintain big data ecosystems.

2.2 The Unique Research Perspective of Data Science

With the advent of the era of big data, our main concerns for data are undergoing a significant shift from "what can we do for data" to "what can data do for us"(Figure1). The shift (or diversification) of research perspectives is the main difference between Data Science and traditional data-related studies. Many new terms frequently occurred in the era of big data, such as data-intensive scientific discovery, data-driven decision making, data-centric architecture, as well as data jiu-jitsu, and most of them are in line with that new shift of research perspectives.

The concern of traditional data-related theories mainly concentrates on "what can I do for data." Traditional data engineering, data structure, database, data warehouse, data mining, and other data-related theories mainly focus on cleaning, labeling, extracting, transforming, and loading data. Those conventional theories place a high value on how to process data by the human effort to make sure data to be more valuable or ready for the subsequent process and future usage.

However, Data Science conforms to the alternative research perspective that "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? What business opportunities or new target markets can be identified by harvesting big data? What uncertainty can be reduced by big data? 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, in the era of big data, 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." The latter is more emphasized in Data Science, that is to say, "What can data do for me."

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 (Figure 2). Data Science is enabled by statistics, machine learning, and data visualization, and those three distinct disciplines are the theoretical foundation of Data Science; The research topics of Data Science can be categorized into six main branches: fundamental concepts and principles for Data Science, data processing, data computing, data management, data analysis, and data product DevOps; putting Data Science theories into practices is commonly domain-dependent so that domain expertise is essential for those applications. In addition, the theory of Data Science involves humanities and social science issues, especially big data ethics, privacy, and trust.

(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. It is worth noting 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 one of the novel terms coined for Data Science. Data Wrangling refers to a series of data processing activities for the purpose of enhancing data quality, reducing the complexity of data computing, and improving the accuracy of data processing. Data science projects need to perform a series of preprocessing activities on raw data, including data audit, data cleaning, data ETL, data integration, data reduction, and data 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, the computing model has undergone a significant shift 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 those big data computing technologies are GFS, BigTable, MapReduce, Spark, and YARN. The main changes in data computing theories involve the primary bottlenecks, research motivations, main contradictions, and thinking patterns for data computing, which are discussed later in this paper.

(4) Data management: Big data needs to be effectively managed for the purpose of carrying out data analysis, data reuse, and long-term storage. Data Science needs not only relational databases but also emerging big data management technologies such as NoSQL, NewSQL, and relational cloud.

(5) Data analysis: The data analysis in Data Science is mainly focus on prescriptive analysis and predictive analysis rather than descriptive or diagnostic ones. Prescriptive models involve large-scale testing and optimization and are a means of embedding analytics into key processes and employee behaviors (Davenport, 2013). Data scientists prefer to choose open-source tools, which are significantly different from commercial software. Consequently, Python and R language are by far the most popular data analysis tools for data scientists.

(6) DevOps for Data products: data product refers to a special meaning in Data Science. Data product development is one of the critical research tasks of Data Science projects since it represents the unique research task to distinguish data science from other sciences. Unlike traditional product development, data product development is data-centric, diverse, hierarchical, and value-added. Data product development capabilities are also 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. The representative practical applications by far are Google Flu Trends (Ginsberg et al., 2009), Target pregnancy prediction (Hill, 2012), MetroMile insurance, IBM Workbench, Databricks,

3. Core Theories And Recent Progress

The core theories of Data Science to date are the DIKW pyramid, data-intensive scientific discovery, data science lifecycle, 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 hierarchical 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, the processing to improve its usefulness; Information is contained in descriptions, answers to questions that begin with such words as who, what, when, where, and how many; Knowledge is conveyed by instructions, answers to how-to questions; Wisdom deals with values and involves the exercise of judgment (Ackoff, 1989).

The DIKW pyramid is one of the widely discussed topics in Data Science since the pyramid represents the underlying motivation for Data Science Studies: turning big data into big wisdom. For instance, John D. Kelleher and Brendan Tierney (2018) proposed the Data Science Pyramid based upon the DIKW pyramid in order 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, one of the most notable differences between the discussion on the DIKW pyramid from a Data Science perspective and that of the conventional ones stems from the former seeks an integrated solution for turning data into information, knowledge, or wisdom instead of isolated solutions (Figure 3). Datafication refers to recording real world into data; Data Wrangling is employed to turn messy data into tidy data; Data Analytics are utilized to acquire information from data; Data Insights are applied to obtain knowledge from data directly. Data Product DevOps are developed and operationalized via turning data into wisdom. As a result, Data Scientist tends to regard information, knowledge, and wisdom as analyzed data, valuable insights and the capability to turn data into products, respectively.

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, 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). The conventional data-related studies, however, 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. As a result, Data Science is becoming one of the hot research topics of traditional data-related studies. Zhu and Xiong (2015) argued that data researchers tend to study data in cyberspace, different from natural science and social science. Chen and Zhang (2014) discussed applications and tools to address big data challenges and put forward 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 are 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. The computational sciences, by contrast, tend to employ hypothesis first data later approach, that is to say, putting forward hypotheses in advance of 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. Data Science further enables enterprises to make data-driven decisions by capturing, mining, and analyzing massive data and measuring and verifying it with statistical models or machine learning algorithms. The introduction of that novel scientific paradigm motivated a shift from computing-centered thinking towards data-centered thinking.

3.3 Data Science Lifecycle

Data Science Lifecycle is one of the basic theories of Data Science and reveals the conceptional workflow of Data Science projects. Although it is a widely accepted convention that the lifecycle model is the typical means to describe Data Science projects, the researchers have not yet reached a consensus on the stages included in Data Science Lifecycle. 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, validate, deployment, as well as support/feedback; Boehm et al. (2020) proposed an open-source machine learning system for the end-to-end data science lifecycle, 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 data science life cycle, including data ingestion, data scrubbing, data visualization, data modeling and data analysis, and further discuss common biases at each stages; Song and Zhu (2017) proposed that the data science lifecycle has eight main stages: business understanding, data understanding, data preparation, model planning, model building, evaluation, deployment, and review and monitoring; Wang et al. (2021) described a data science lifecycle which contains ten distinct stages: requirement gathering & problem formulation, data acquisition & governance, data readiness & data preprocess & data cleaning, feature engineering, model building & model training, model presentation & stakeholder verification, model deployment, runtime monitoring, model refinement (Post-Deployment), decision making & optimization.

The main purpose of Data Science projects is to obtain valuable insights from big data to make better decisions. With the maturity of machine learning, cloud computing, and artificial intelligence, more and more jobs are auto-completed by machines. Humans, however, 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. Allowing for man-machine collaborative data science, we propose a new data science lifecycle model with nine main steps: business understanding, datafication, data wrangling, data analysis, data understanding, data insights, visualizing/storytelling/communicating, data products DevOps, decision -support or automated decision making (Figure 4).

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 main concerns of data preprocessing. Wickham (2014) demonstrated how to transform messy data into tidy data using a set of tools with R programming. Endel and Piringier (2015) proposed that Data Wrangling is not only about transforming and cleaning procedures, and many other aspects like data quality, merging of different sources, reproducible processes, and managing data provenance have to 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) utilized 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' skills of creative design, critical thinking, and curious asking to the data processing tasks. Data Wrangling is a new kind of data preprocessing in that it involves data cleansing as well as data tidying. Data cleansing is the process of converting dirty data into clean data by enhancing the quality of data. Alternatively, data tidying refers to transforming messy data into tidy data by reshaping or reformatting data.

It is worth noting that Data Wrangling usually leads to information loss or even information distortion. Some valuable information may be lost when transforming unstructured data into structured data, where it cannot be directly stored in a structured form. It is also possible that the original real meaning of data is distorted when the data is transferred from one format to another. Therefore, data scientists have to strike a balance between data wrangling and its information loss.

3.5 Big Data Analytics

Big Data Analytics has been the most widely discussed as well as the most advanced topic in Data Science. As a result, a few of researchers are under the impression that Data Science is only the new alternative name of 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 Analysis is merely one of the stages of a Data Science Lifecycle, and data analysis systems must provide effective mechanisms to design and complete analysis tasks (Elshaw, 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 project is a challenge in Data Science, and Big Data Analytics can provide new insights into QS and other biological issues.

One of the hottest topics on big data analytics is developing tools and technologies for Data Science projects. There have been some well-established tools to help data scientists and data analysts carry out 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 now generally deemed to be composed of the related projects, including HDFS, YARN, MapReduce, Pig, Hive, and HBase (Bappalige, 2014). There is also a vast range of companies are 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 objectives to make data analysis faster. It is currently one of the most popular technological solutions for big data analytics.

The studies on Big Data Analytics mainly 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 may not be the bottleneck of future big data analysis, while the bottleneck lies in the input/output hardware that transfers data between storage and the processor. Business applications need real-time big data analytics to implement dynamic auto-decisions, which requires 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 as shown in Figure 5. Data Management Maturity (DMM) lists the 25 Key Processes (KP) required for organizational data management and further categorizes them into 6 Key Process Areas (KPA): data management strategy, data governance, data quality, platform & architecture, data operations, and supporting processes. The Data Governance defined by DMM involves three key processes: governance management, business glossary, and metadata management. In addition, Information Technology Services—Governance Part 5: Data Governance Specification (GB/T 34960.5-2018) issued by the China National Information Technology Standardization Network, Data Management is defined 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.

3.7 Data Products DevOps

Providing data products is one of the ultimate research motivations of Data Science. Data Product in Data Science refers to all kinds of 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 (some) of their goals (Patil, 2012). As a result, 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: providing data-based(fact-based) solutions rather than knowledge-based solutions, dedicating to address data-intensive problems in preference to computing-intensive tasks, being driven by data instead of hypothesis, conforming to data-analytic thinking in place of intuition-based thinking, adopting data-centered architectures as a substitute for app-centered architectures, and creating value from data in favor of creating data for value(Figure 6). Therefore, data products are fact-based and go beyond the limit 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).

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 turning data into products. He further provided thirteen underlying principles of Data Jujitsu. In industry, Data products development is one of the required skills of data scientists and big data analysts. Data Science is located at the intersection of statistics, machine learning, and domain knowledge (Schutt & O'Neil, 2013). Li, Roy and Saltz (2019) discussed how engineers and data scientists could effectively collaborate on new product development in a hybrid team with data-driven features. Online data products, such as search engines developed by companies such as Yahoo and Google, can be used as mobile applications, forming the so-called Application Economy (Davenport & Kudyba, 2016). To generate full value from 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 widely adopted as one of most common tools 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 explicit the hidden information behind big data. However, the diversity of big data 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 bring by the limitations of presenting big data, while data reduction and reducing latency are the 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 the 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 format, to import/export data or share visualization results with other tools, to provide collaborative working space, and to make sure better user experiences.

Visual Analytics was proposed in 2004 by a working group at the National Visualization and Analytics Center (NVAC). 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: growing significance of DataOps, the rise of citizen data scientists, enabling augmented data science, diversity of domain-specific data science, and 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, Distefano and Mazzara (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 more reliably.

One of the significant trends in Data Science is to integrate DataOps with MLOps. As one of the building blocks of Data Science, Machine learning provides big data analysis with mythological foundations. As a result, Data Science usually takes advantage of MLOps to deploy machine learning models in Data Science Projects reliably and efficiently. MLOps enables data scientists to monitor, validate, and govern machine learning 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 one of the new topics of Data Science. Citizen Data Scientist is a 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). 2016 is the year of the citizen data scientist

because users throughout the business want a more democratized approach to big data and analytics (Shacklett, 2016). Citizen data scientists possess more expertise than professional ones with regard to particular application domains, as shown in Table 1.

Table 1
Citizen data scientist vs 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 the practices of Data Science are often heavily dependent on domain expertise. The selection, interpretability, and evaluations of models in Data Science projects require knowledge or skills from the corresponding fields. Typically, Citizen Data Scientists mainly focus on utilizing Data Science tools but usually lack the ability to understand the underlying principles of these tools. However, understanding those principles are crucial for selecting algorithms, optimizing models, and tuning their hyperparameters. The roles of citizen data scientists and professional data scientists complement each other, and the collaboration between them is a new trend of Data Science practices.

4.3 Enabling Augmented Data Science

Augmented Data Science is a data-driven method in which software tools automatically carry out data exploration and processing stages 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 machine learning 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: augmented data preparation, augmented analytics as part of analytics and business intelligence, and augmented data science or machine learning (Gartner, 2018). Augmented analytics can implement automatic analysis, reduce the difficulty of data analysis for non-professional users, and help data scientists carried out data analysis tasks more efficiently and effectively.

Augmented Data Science is redefining the roles of man and machine in the relevant practices. Augmented Data Science simultaneously enhances the ROI (Return on Investment) of data science investments, reduces time to value, and expands the footprint of machine learning. Experts become more efficient and productive, and a broader population of quantitative professionals can succeed in data science (Gartner, 2019). As a result, Augmented Data Science will promote the collaboration between data scientists and its application domain scientists; thus, the human-machine collaborative working pattern will 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 with traditional Data Warehouse technologies, Data Lake supports 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 key to Data Science projects.

Most Data Science platforms will be built on the top of Data Lakehouse that combines data warehouses and data lakes. Data Lakehouse is a new generation of open platforms that unify data warehousing and big data analytics (Armbrust et al., 2021). Databricks Lakehouse, for instance, implements unifying data, analytics, and AI and provides a collaborative working platform for data science projects (Databricks, 2021). As a consequence, 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 is one of the hot topics in recent studies. Those studies can be categorized into two groups: domain-general Data Science and domain-specific Data Science. The former regards Data Science as an independent discipline and mainly focuses on nurturing Data Science as an independent new discipline. However, the latter tends to discuss Data Science from a specific application discipline. As a result, more traditional disciplines show a new research trend that conducts a comprehensive study with Data Science, and hence domain-specific Data Science turns out to be an emerging topic of application disciplines. The most active application domains to date are life science, healthcare, government, education, and business management. Some new research topics, in turn, enter from those application areas to Data Science studies, 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 bridge the gaps between distinct domain-specific studies. Theoretical Data Science is a new branch of Data Science, which employs mathematical model and abstraction 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 more 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 one of the emerging research directions 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 story requires visualization to make key observations and detail 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 (Table 2): 1) motivation: data story is only designed to meet a given business requirement, while the literacy story is often created for general purposes, such as entertainment, education, recreation all the audience. As a result, data story only works for the target users in a specific business life cycle. 2) content: the content of a data story must source from actual business data, but that of a literacy story may stem from imagination, life experience, or hearsay. 3)

creator: data stories are automatically created by algorithms, whereas conventional literacy story is directly written by human beings. 4) lifespan: the lifespan of a data story is shorter than that of a literacy story in that the former is strictly restricted with the corresponding business lifecycle. As a result, an excellent data story will be expired when tasks of the business process are completed, while an excellent literary story can 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 further development of Data Science should prioritize turning the four most acute challenges into opportunities: accelerating theoretical studies of data science, the trade-off between explainability and performance, achieving data ethics, privacy as well as trust, and 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 of itself. Despite the fact that Data Science is one of the hottest topics of recent academic studies, the in-depth study of its theoretical framework is overlooked. As a result, there are no shared understandings of the theoretical system of Data Science and its essential components. Furthermore, a few of researchers tend to misuse Data Science as a new name for some old approaches of data analysis or data processing, such as machine learning, statistics, data engineering, or business intelligence. This weakness is becoming a new bottleneck of the further development of Data Science.

The primary task of accelerating theoretical studies of Data Science is to integrate domain-general Data Science with a diverse range of domain-specific Data Science. Chaolemen Borjigin et al.(2021) proposed a new term Theoretical Data Science in order to bridge the gap between the domain-general and domain-specific studies and provide its five essential topics: to conduct in-depth theoretical research on Data Science, to take advantage of the active property of big data, to introduce Design of Experiments into Data Science studies, to shift Data Science' research focus from correlation analysis into causality inference, and to take data product development as one of the main tasks of Data Science projects.

In addition, expanding the technical areas of today's consensus Data Science is also crucial to theoretical studies of Data Science. Donoho (2017) proposed a new field called Greater Data Science(GDS) that is a better academic enlargement of statistics and machine learning than today's Data Science initiatives, while being able to accommodate the same short-term goals.

5.2 The trade-off between explainability and performance

The most critical challenge facing a Data Science practice is to balance its interpretability with performance. Explainability and effectiveness are also goals that must be weighed in model design for a 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 needs to take account of the trade-off between accuracy and fidelity and strike a balance among 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. At the earliest stages of Data Science, the researchers tended to focus on correlation instead of causation. However, the ignorance of causal analysis results in lower trust in Data Science solutions.

Expertise in the design of experiments can help cross the gap between correlation and causation (McAfee et al., 2012). Besides, Explainable Artificial Intelligence (XAI) provides a new solution for balancing interpretability and performance. The existing XAI studies can be mainly divided into various groups from two distinct dimensions (Rai, 2019): 1) whether the technique is model-specific or model-agnostic. Model-specific techniques only work with a given machine learning model. In contrast, model-agnostic techniques can be employed in various machine learning models. For instance, a model-agnostic technique called LIME is mainly used to perturb the input samples to observe the impact on the output results. 2) whether the technique provides a global explanation or a local explanation. The global explanation mainly 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. In addition, inferring causality in Data Science also needs to integrate domain-general Data Science with diverse domain expertise.

5.3 Achieving data ethics, privacy, and trust

Data ethics, privacy, and trust are the potential risks of Data Science practices. The threat of data security comes from a diverse range of factors, including confidentiality, integrity, availability, and privacy (Talha, Abou El Kalam, & Elmarzouqi, 2019). Furthermore, an ethical expert should be included in a data science project to avoid BIBO (Bias In, Bias Out) (Ho & Beyan, 2020). Data bias, such as survivorship bias, Simpson's paradox, and Bergson's paradox, probably occurs in any stage of a Data Science lifecycle. Explainable Artificial Intelligence is a way to check whether there is algorithmic bias (Sen, Dasgupta, & Gupta, 2020). Besides, User authentication and consent on personnel data is critical for protecting ethics and privacy.

Data masking and data auditing are two essential approaches to achieving data ethics, privacy, and trust. Data masking is implemented via the replacement or deletion of original personal (or organizational) sensitive data without affecting the accuracy of the data analysis results to avoid security risks and personal privacy issues. Data auditing can help data scientists ensure data integrity, control data quality, and prevent data leakage. Data masking and data auditing are effective ways to achieve data ethics, privacy, and trust in data science projects and are essential for data scientists to procure insights from big data in accordance with 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, Thomas, & Patil, 2012). The main challenges of higher education in cultivating qualified data scientists root in three factors: 1) the curriculum is loosely coupled with Data Science practices so that Data Science major is merely an alternative title of traditional ones, notably statistics or machine learning; 2) some of the required course of Data Science majors are missing, including Exploratory Data Analysis, Design of Experiment, Causality, and Data Products Design; 3) student's poor capability to address the real-world challenges. At present, there is no single model of which department, school, or cross-unit collaboration within higher education institutions should have the responsibility for data science education (Berman et al, 2018).

A study conducted by Chaolemen 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 and Hadoop MapReduce /HBase /Hive, Spark/Storm, Visual Analysis with Tableau, ETL, Data Warehouse /Data Lake /BI,Statistics,Machine learning (including deep learning),Natural Language Processing,Text Analysis and Computer Vision. General purpose-oriented qualifications mainly involve the candidate's readiness for communication and cooperation,problem-solving,3C characteristic of data scientists, independent learning,attention to detail,stress management and leadership skills (Bojigin C, Xiao J, & Wang X, 2021)). Therefore, the top-level design of Data Science curriculums in higher education should meet those industry needs as well as their further evolutions.

6. Conclusions

The gap between data and knowledge is at its highest level in the early stages of big data era, and traditional knowledge cannot match the new data created by or stored in Cloud Computing, The Internet of Things, Mobile Internet as well as emerging scientific instruments or manufacturing equipment. As a result, the contradiction between new data and old knowledge is the biggest challenge faced by most of the traditional sciences. Computer Science and Statistics first perceived that challenge and proposed a new science called Data Science from their distinct perspectives, respectively. Then, more and more disciplines noticed that gap and conducted interdisciplinary studies. Consequently, Data Science is one of today's most rapidly growing academic fields and has tremendous implications for all traditional studies. However, the relevant studies fail to conduct in-depth research on the theoretical system of Data Science, and some doubt whether Data Science is an independent science or not.

One of the more significant findings from this study is that Data Science possesses 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 of Data Science. Further work should be carried out to establish theoretical systems of Data Science, especially to accelerate theoretical studies of Data Science, to address the paradox between usability and interpretability of big data solutions, to achieve big data ethics, privacy, and trust, and to align data science curriculums with industrial needs.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China [No. 72074214].

Authors' contributions

CB designed the research work, and wrote the first draft of the manuscript, and revised and reviewed the manuscript. CZ wrote the first draft of the manuscript and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Ackoff RL. From data to wisdom. *Journal of applied systems analysis*. 1989;16,1:3-9.
2. Agrawal R, Kadadi A, Dai X, Andres F. Challenges and opportunities with big data visualization. In *Proceedings of the 7th International Conference on Management of computational and collective intelligence in Digital EcoSystems*. 2015;169-173.
3. Al SM, Gupta N, Nayak GK, Lenka RK. Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. 2016;656-660.
4. An introduction to Apache Hadoop for big data. Retrieved from <https://opensource.com/life/14/8/intro-apache-hadoop-big-data>. Accessed 17 September 2021.
5. Augmented Analytics Is the Future of Data and Analytics. <https://www.gartner.com/en/documents/3773164>. Accessed 17 September 2021.
6. Augmented Data Science: Towards Industrialization and Democratization of Data Science. <https://arxiv.org/abs/1909.05682>. Accessed 25 October 2021.
7. Azeroual O. Data Wrangling in Database Systems: Purging of Dirty Data. *Data*. 2020; doi:10.3390/data5020050.
8. Battle of the data science Venn diagrams. <https://deeplearning.lipingyang.org/wp-content/uploads/2017/10/Battle-of-the-Data-Science-Venn-Diagrams.pdf>. Accessed 17 September 2021.
9. Beck DAC, Carothers JM, Subramanian VR, Pfaendtner J. Data science: Accelerating innovation and discovery in chemical engineering. *AIChE Journal*. 2016; doi:10.1002/aic.15192.
10. Berman F, Rutenbar R, Hailpern B, Christensen H, Davidson S, Estrin D, et al. Realizing the potential of data science. *Communications of the ACM*. 2018; doi:10.1145/3188721.
11. Biases in Data Science Lifecycle. <https://arxiv.org/abs/2009.09795>. Accessed 25 October 2021.
12. Borjigin C, Xiao J, Wang X. Typical Responsibilities, Key Qualifications and Higher Education for Data Scientist. *Journal of Library Science in China*. 2021; doi:10.13530/j.cnki.jlis.2021023.
13. Borjigin C, Zhang C, Sun Z, et al. Theoretical Data Science: bridging the gap between domain-general and domain-specific studies. *Data Science and Informatics*. 2021;1(1):1-28.
14. Cao L. Data Science. *ACM Computing Surveys*. 2017; doi:10.1145/3076253.
15. Capizzi A, Distefano S, Mazzara M. 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. 2019; 52-62.

16. Citizen Data Science Augments Data Discovery and Simplifies Data Science.
<https://www.gartner.com/en/documents/3534848>. Accessed 17 September 2021.
17. Cleveland WS. Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*. 2001; <https://doi.org/10.1002/sam.11239>.
18. Data Visualization and Data Storytelling: A Visual Revolution.
<https://www.dbta.com/BigDataQuarterly/Articles/Data-Visualization-and-Data-Storytelling-A-Visual-Revolution-124077.aspx>. Accessed 17 September 2021.
19. Data Visualization: How to Tell A Story With Data. <https://www.forbes.com/sites/nicolemartin1/2018/11/01/data-visualization-how-to-tell-a-story-with-data/>. Accessed 17 September 2021.
20. DataOps (data operations). <https://searchdatamanagement.techtarget.com/definition/DataOps>. Accessed 17 September 2021.
21. Davenport TH. Analytics 3.0. *Harvard business review*. 2013; 91(12):64-72.
22. Davenport TH, Kudyba S. Designing and developing analytics-based data products. *MIT Sloan Management Review*. 2016; 58(1), 83.
23. Davenport TH, Patil DJ. Data scientist. *Harvard business review*. 2012;90:70-76.
24. DGI Data Governance Framework. <https://opensource.com/life/14/8/intro-apache-hadoop-big-data>. Accessed 17 September 2021.
25. Donoho D. 50 years of data science. *Journal of Computational and Graphical Statistics*. 2017; doi:10.1080/10618600.2017.1384734.
26. Dykes B. Effective data storytelling: how to drive change with data, narrative and visuals. John Wiley & Sons; 2019.
27. Elshawi R, Sakr S, Talia D, Trunfio P. Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service. *Big Data Research*. 2018; doi:10.1016/j.bdr.2018.04.004.
28. Endel F, Piringer H. Data Wrangling: Making data useful again. *IFAC-PapersOnLine*. 2015; doi:10.1016/j.ifacol.2015.05.197.
29. Explainable recommendation: A survey and new perspectives. <https://arxiv.org/abs/1804.11192>. Accessed 25 October 2021.
30. Gibert K, Horsburgh JS, Athanasiadis IN, Holmes G. Environmental data science. *Environmental Modelling & Software*. 2018; doi:10.1016/j.envsoft.2018.04.005.
31. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009; doi:10.1038/nature07634.
32. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. *Science Robotics*. 2019; doi:10.1126/scirobotics.aay7120.
33. Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research; 2009.
34. How Augmented Machine Learning Is Democratizing Data Science.
<https://www.gartner.com/en/documents/3956825-how-augmented-machine-learning-is-democratizing-data-sci>. Accessed 17 September 2021.
35. How target figured out a teen girl was pregnant before her father did.
http://elearning.algonquincollege.com/coursemat/haugs/1_F_19_BUS2303/SECURITY%20ARTICLE%20%20Target-Teen-Pregnancy-Forbes.pdf. Accessed 17 September 2021.
36. Hüttermann M. DevOps for developers. Apress; 2012.

37. Jiang S, Kahn J. Data wrangling practices and collaborative interactions with aggregated data. *International Journal of Computer-Supported Collaborative Learning*. 2020; doi:10.1007/s11412-020-09327-1.
38. Jin X, Wah BW, Cheng X, Wang Y. Significance and Challenges of Big Data Research. *Big Data Research*. 2015; doi:10.1016/j.bdr.2015.01.006.
39. Joshi MP, Su N, Austin RD, Sundaram AK. Why So Many Data Science Projects Fail to Deliver. *MIT Sloan Management Review*. 2021;62(3), 85-89.
40. Kalidindi SR, De Graef M. Materials Data Science: Current Status and Future Outlook. *Annual Review of Materials Research*. 2015; doi:10.1146/annurev-matsci-070214-020844.
41. Kambatla K, Kollias G, Kumar V, Grama A. Trends in big data analytics. *Journal of parallel and distributed computing*. 2014; doi:10.1016/j.jpdc.2014.01.003.
42. Kandel S, Heer J, Plaisant C, Kennedy J, Ham F, Riche NH, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*. 2011; doi:10.1177/1473871611415994.
43. Kelleher J D, Tierney B. *Data science*. MIT Press; 2018.
44. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. https://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf. Accessed 17 September 2021.
45. Larson D, Chang V. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*. 2016; doi:10.1016/j.ijinfomgt.2016.04.013.
46. Li Y, Roy U, Saltz JS. Towards an integrated process model for new product development with data-driven features (NPD3). *Research in Engineering Design*. 2019; doi:10.1007/s00163-019-00308-6.
47. Machine Learning Operations. In: *Data Science Solutions on Azure*. Apress, Berkeley, CA. doi:10.1007/978-1-4842-6405-8_8. Accessed 17 September 2021.
48. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*. 2017; doi:10.1038/s41551-017-0132-7.
49. Mathur N, Purohit R. Issues and Challenges in Convergence of Big Data, Cloud and Data Science. *International Journal of Computer Applications*. 2017; doi:10.5120/ijca2017913082.
50. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data: the management revolution. *Harvard business review*. 2012; 90(10): 60-68.
51. Nadikattu RR. Research on data science, data analytics and big data. *INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND*. 2020; <http://dx.doi.org/10.2139/ssrn.3622844>.
52. Nakamura H. Big data science at AMED-BINDS. *Biophysical Reviews*. 2020; doi:10.1007/s12551-020-00628-1.
53. Nargesian F, Zhu E, Miller RJ, Pu KQ, Arocena PC. Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*. 2019; doi:10.14778/3352063.3352116.
54. Naur P. *Concise survey of computer methods*. Petrocelli Books; 1974.
55. Pantelis K, Aija L. Understanding the value of (big) data. 2013 IEEE International Conference on Big Data. 2013; doi:10.1109/bigdata.2013.6691691.
56. Patil DJ. *Data Jujitsu*. " O'Reilly Media, Inc."; 2012.
57. Peek N, Rodrigues PP. Three controversies in health data science. *International Journal of Data Science and Analytics*. 2018; doi:10.1007/s41060-018-0109-y.
58. Provost F, Fawcett T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc."; 2013.

59. Rai A. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*. 2019; doi:10.1007/s11747-019-00710-5.
60. Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*. 2020; doi:10.1186/s40537-020-00318-5.
61. Schutt R, O'Neil C. *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reilly; 2013.
62. Sen S, Dasgupta D, Gupta KD. An Empirical Study on Algorithmic Bias. 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC). 2020; doi:10.1109/compsac48688.2020.00-95.
63. Singleton A, Arribas-Bel D. Geographic data science. *Geographical Analysis*. 2021; doi:10.1111/gean.12194.
64. Sokol K, Flach P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020; doi:10.1145/3351095.3372870.
65. Song IY, Zhu Y. Big Data and Data Science: Opportunities and Challenges of iSchools. *Journal of Data and Information Science*. 2017; doi:10.1515/jdis-2017-0011.
66. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? *PLOS Biology*. 2015; doi:10.1371/journal.pbio.1002195.
67. Swan M. *The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery*. Big Data. 2013; doi:10.1089/big.2012.0002.
68. SystemDS: A declarative machine learning system for the end-to-end data science lifecycle. <https://arxiv.org/abs/1909.02976>. Accessed 25 October 2021.
69. Talha M, Abou El, Kalam A, Elmarzouqi N. Big Data: Trade-off between Data Quality and Data Security. *Procedia Computer Science*. 2019; doi:10.1016/j.procs.2019.04.127.
70. The Battle for Data Science. <http://sites.computer.org/debull/A20june/p8.pdf>. Accessed 17 September 2021.
71. The Databricks Lakehouse Platform. Retrieved from <https://databricks.com/product/data-lakehouse>. Accessed 17 September 2021.
72. Top 10 Strategic Technology Trends for 2019. <https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/3891569-top-10-strategic-technology-trends-for-2019.pdf>. Accessed 17 September 2021.
73. Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. *Journal of Big data*. 2015; doi:10.1186/s40537-015-0030-3.
74. Wang D, Andres J, Weisz JD, Oduor E, Dugan C. AutoDS: Towards Human-Centered Automation of Data Science. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021; doi:10.1145/3411764.3445526.
75. Why 2016 might be the year of citizen data scientists. Retrieved from <https://www.techrepublic.com/article/why-2016-may-be-the-year-of-the-citizen-data-scientist/>. Accessed 17 September 2021.
76. Wickham H. Tidy Data. *Journal of Statistical Software*, 59(10). 2014; doi:10.18637/jss.v059.i10.
77. Zhu Y, Xiong Y. Towards Data Science. *Data Science Journal*. 2015; doi:10.5334/dsj-2015-008.

Figures

Figure 1

A new unique perspective of Data Science

Figure 2

Data Science Body of Knowledge

Figure 3

The DIKW pyramid in Data Science perspectives

Figure 4

Data Science Lifecycle (adapted from O'Neil, Cathy, and Rachel Schutt. Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.", 2013.)

Figure 5

Data Governance and Data Management

Figure 6

Six main features of Data Products