DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

RESEARCH ARTICLES

Theoretical Data Science: bridging the gap between domain–general and domain–specific studies

Chaolemen Borjigin*, Chen Zhang, Zhizong Sun, Ni Yi

Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China
School of Information Resource Management, Renmin University of China, Beijing, China

ABSTRACT

The entering into big data era gives rise to a novel discipline called Data Science. Data Science is interdisciplinary in its nature, and the existing relevant studies can be categorized into domain-independent studies and domain-dependent studies. The domain-dependent studies and domain-independent ones are evolving into Domain-general Data Science and Domain-specific Data Science. Domain-general Data Science emphasizes Data Science in a general sense, involving concepts, theories, methods, technologies, and tools. Domain-specific Data Science is a variant of Domain-general Data Science and varies from one domain to another. The most popular Domain-specific Data Science includes Data journalism, Industrial Data Science, Business Data Science, Health Data Science, Biological Data Science, Social Data Science, and Agile Data Science.

The difference between Domain-general Data Science and Domain-specific Data Science roots in their thinking paradigms: DGDS conforms to data-centered thinking, while DSDS is in line with knowledge-centered thinking. As a result, DGDS focuses on the theoretical studies, while DSDS is centered on applied ones. However, DSDS and DGDS possess complementary advantages. Theoretical Data Science (TDS) is a new branch of Data Science that employs mathematical models and abstractions of data objects and systems to rationalize, explain and predict big data phenomena. TDS will bridge the gap between DGDS and DSDS. TDS contrasts with DSDS, which uses casual analysis, as well as DGDS, which employs data-centered thinking to deal with big data problems in that it balances the usability and the interpretability of Data Science practices.

The main concerns of TDS are concentrated on integrating the data-centered thinking with the knowledge-centered thinking as well as transforming a correlation analysis into the casual analysis. Hence, TDS can bridge the gaps between DGDS and DSDS, and balance the usability and the interpretability of big data solutions.

The studies of TDS should be focused on the following research purpose: to develop theoretical studies of TDS, to take advantages of active property of big data, to embrace design of experiments, to enhance causality analysis, and to develop data products.

KEYWORDS

Data Science; Big Data; Theoretical Data Science; Domain-general Data Science; Domain-specific Data Science

* Corresponding author: chaolemen@ruc.edu.cn

1 Introduction

The bottlenecks in human being's data capabilities that capture or create, store, manage, compute, analyze as well as utilize data have been eliminated because of widespread applications of novel technologies. For instance, Internet of Things extends makes it possible for us to capture or digitalize the information of total populations instead of samples; Cloud Computing virtualizes computing resources and provides scalable on-demand services so that we can store, manage, compute, and analyze data at a low cost; Mobile Computing records the thoughts, feelings, and behaviors of the individuals as well as the social networks between them. As a result, we are entering into a data enriched-offerings era that is distinct from any previous era in human history.

Big data is shifting today's scientific paradigm, and giving rise to a novel discipline called Data Science. How to take advantages of big data in order to survive in data enriched-offerings era is one of the hot topics for most disciplines from basic science such as Statistics and Computer Science to applied sciences, including Social Sciences. As a result, the research on big data from different disciplines begins to converge on an emerging discipline called Data Science. Data Science deems data-centered thinking as an alternative paradigm for data-related tasks, which is different from the knowledge-centered thinking in traditional research. However, the studies of Data Science are spread across a variety of disciplines, and we need to conduct the in-depth research on its core theories, main methods, typical techniques, and best practices.

The rest of this paper is structured as follows: Section 2 discusses the brief history, interdisciplinarity, and taxonomy of Data Science, and categorizes the existing studies into two basic subgroups: Domain-general DS and Domain-specific DS. Then, Section 3 proposes the current research topics of Domain-general Data Science, including data wrangling, data computing, data management, data analysis, and data product development. In addition, the states of arts of typical Domain-specific Data Science are described in Section 4: Data Journalism, Industrial Data Science, Business Data Science, Health Data Science, Biological Data Science, Social Data Science, and Agile Data Science. Furthermore, Section 5 provides a comparative study between Domain-general Data Science and Domain-specific Data Science, and a comprehensive solution for integrating those two distinct branches of Data Science theories by introducing Theoretical Data Science. Finally, in Section 6, the critical topics for Theoretical Data Science studies are proposed.

2 Data Science: The Science of Big Data

Data Science is a new emerging discipline that is termed to address challenges that we are facing and going to face in data-enriched offerings era. It provides new theories, methods, models, technologies, platforms, tools, applications, and best practices of big data. And one of the main purposes of Data Science research is to reveal the new challenges and opportunities brought by big data.

2.1 A Brief History of Data Science

Peter Naur, the Turing Award winner, coined the term of Data Science in his book entitled *Concise Survey of Computer Methods* in 1974. He defined Data Science as the science of dealing with data, and further proposed that it is different from Datalogy, which is the sci-

ence of data and of data processes and its place in education (Naur, 1974). In 2001, William S. Cleveland published the paper, *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*, proposing that Data Science is an emerging branch of Statistics. In 2013, Nature published the article, *Computing: A Vision for Data Science* (Mattmann, 2013), and Communications of the ACM published the paper, *Data Science and Prediction* (Dhar, 2013). Both of those two articles discussed the Data Science from the perspective of Computer Science. Then, Data Science was also identified as a branch of Computer Science. Data Science has begun to get much more public attention since 2010s. Patil DJ and Davenport T H published the article entitled *Data Scientist: The Sexiest Job of the 21st Century* in Harvard Business Review in 2012. Barack Obama won the presidency by implementing and using big data strategies in the 2012 US presidential election (Kitchin, 2013). The White House announced Patil DJ the first U.S. Chief Data Scientist in 2015.

Data Science was on an one-way trip to the peak of inflated expectations and would enter plateau of productivity in 2 - 5 years according to Gartner's 2014 Hype Cycle for Emerging Technologies. Gartner's 2016 Hype Cycle for Data Science is a growth curve that shows the breadth and depth of excitement about Data Science, with new technologies and some significant movements from last year. Gartner's 2016 Hype Cycle for Data Science shows: R entered the plateau of productivity; Simulation, Ensemble Learning, Video or Image Analytics and Text Analysis were climbing the slope of enlightenment; Hadoop-Based Data Discovery was obsolete before plateau; Speech Analytics, Model Management and Natural-Language Question Answering have passed the peak of inflated expectations and slided into the trough of disillusionment; Citizen Data Science, Model Factory, Algorithm Marketplaces and Prescriptive Analytics have recently come to light.

2.2 The interdisciplinarity of Data Science

In 2010, Drew Conway proposed his Data Science Venn Diagram (Figure 1) to reveal the interdisciplinary of Data Science. The Venn Diagram shows that Data Science is a combination of hacking skills, math & statistics knowledge, and substantive expertise. Now, there are many variations of his Venn diagram such as Jerry Overton's Data Science Venn Diagram (2016), but all of them are less influential than Drew Conway's Venn Diagram.

Data Science is interdisciplinary, it has three basic components: knowledge, expertise, and skills. The knowledge in Data Science is domain-independent, while the expertise and the skills are domain-dependent. The knowledge in Data Science is evolving into Domain-general Data Science and the expertise and the skills from application fields is the source of Domain-specific Data Science.

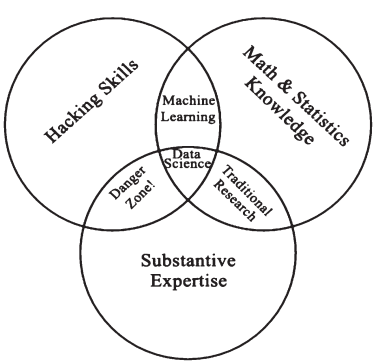

Figure 1 Venn Diagram by Drew Conway (2010)

2.3 Taxonomy of Data Science

Data Science is an emerging discipline that incorporates theories with domain-independent knowledge and domain-dependent business practices and skills. As a result, there are two types of Data Science: Domain-general Data Science and Domain-specific Data Science.

Domain-general Data Science regards Data Science as an independent discipline, while the Domain-specific Data Science argues that Data Science heavily depends on a specific application domain. The main research topics on Domain-specific Data Science include Data Journalism, Materials Data Science, Big Data Finance, Big Data Society, Big Data Ethics, and Big Data Education.

Domain-general Data Science is a theoretical foundation for Domain-specific Data Science. Domain-general Data Science involves the general ideas, theories, methods, concepts and tools of Data Science, Domain-specific Data Science is commonly restricted within a specific application discipline.

3 Domain–general Data Science

Domain-general Data Science is devoted to issues on domain-independent Data Science, involving concepts, theories, methods, technologies, tools and so on. The counterpart is Domain-specific Data Science, which are two different terms. Domain-general Data Science aims to solve theoretical challenges related to Data Science itself, including the core theories of Data Science and Data Wrangling, Data Computing, Data Management, Data Analysis and Data Products Development. It is worth noting that the basic theories are within the research scope of Data Science, while the theoretical bases are outside the scope.

3.1 Core theories

Core theories of DGDS involve new concepts, theories, methods, technologies, and tools applied in Data Science.

Big data and Data Science. Data Science is a science about big data, which covers a whole set of knowledge system of it. Big data is also one of the research objects of Data Science. It is broken by IBM (2013)data scientists into four dimensions: volume, variety, velocity and veracity. Big data analytics is the application of advanced analytic techniques to very big data sets (Russom, 2011). Many organizations have invested in developing products using Big Data Analytics to address the monitoring, experimentation, data analysis, simulations, large and disperse datasets. Data-Driven: Data-Driven refers to the process of doing things based on big data rather than purely on experience or intuition, such as Data-Driven Decision Making and Data-Driven Modelling. Jim Gray (2007)proposed that the scientific paradigm is shifting from Experimental Science, Theoretical Science and Computational Science to the fourth scientific paradigm—Data-Intensive Science. More efficient data-intensive techniques are required, such as cloud computing, social computing, and biological computing. Datafication is the transformation of social action into online quantified data, thus allowing for real-time tracking and predictive analysis(Mayer-Schönberger & Cukier, 2013). Along with the Internet of Things and Sensors, Quantified Self is also a hot topic of datafication.

Life Cycle of Data Science. Theories of Data Science involves potential guidance for the process of Data Wrangling, Data Computing, Data Management, Data Analysis, Data Products Development. In general, models are implementations of theory of Data Science (Das, 2017). But Data Science has grown up rapidly in the model applications of the big data

era. It has achieved successful practice in many fields, but its theoretical research is still lagging far behind its practical application. Theory-guided Data Science was proposed as an emerging paradigm for scientific discovery from data to the effectiveness improvement of Data Science models (Karpatne et al., 2017).

Methodologies of Data Science. Method guides the direction of problem solving and can promote the development of technology. Technology is used to solve the problem by performing an operation. Data Science is an interdisciplinary field that requires methods related to statistics, machine learning, deep learning, data analysis, data visualization, data processing, cloud computing, data engineering and so on. But Data Science is not to cover all aspects of these fields, but to unify some of their methods to gain insights from data. Technologies of Data Science are the specific executions of these methods.

Technologies of Data Science. Data scientists need to master the popular technologies include: Linear Regression, Resampling Methods, Nonlinear Models, Variance Analysis and Time Series Analysis, Decision Tree, Support Vector Machines, Random Forest, Principal Component Analysis, Classification, Clustering, Semi-supervised Learning and Reinforcement Learning, Deep Learning, Data Collection, Data Storage, Data Preprocessing, Data Mining, Natural Language Processing, and computer vision.

Tools of Data Science. Data Science tools that most of the data scientists used include: open-source Data Science programming languages, such as Python, R and Julia; big data computing tools, especially Hadoop MapReduce and Spark; big data storage tools, including HDFS and GFS; big data management systems, such as NoSQL, new SQL, and cloud RDB; Data mining tools, such as RapidMiner, Data Melt; Data visualization tools, including Tableau, D3.js and PowerBI; Machine learning or Deep Learning tools such as TensorFlow, PyTorch and Keras.

3.2 Data Wrangling

Data Wrangling is the initial process of transforming raw data into another form for improving data quality (Kandel et al., 2011). It is the first phase of most data-driven projects and known as data wrangling, data munging or janitorial work (Endel & Piringer, 2015). It also concerns that how to apply data scientists' skills of creative design, critical thinking, and curious questioning to the data wrangling activities and how to avoid Garbage in and Garbage out (GIGO).

Data preparation accounts for about 80% of the time of a Data Science project (Patil, 2012). Data wrangling is the preparation for analysis, which involves the following operations: data auditing, data cleaning, data conversion, data integration, data masking, data reduction, data annotation.

Data Auditing. Data auditing uses pre-determined evaluation methods to check data quality and identify its problems (Abdallah et al., 2017), the problems include: missing data, abnormal data, data contradicting each other, tampered data that cannot be traced back to its source. Taleb et al. (2018) proposed an across-the-board quality management framework describing the key quality evaluation practices to be conducted through the different Big Data stages.

Data Cleaning. Data cleaning is the process of altering messy data to tidy data by filtering duplicate data, identifying incorrect data, and processing missing values. However, big data processing is different from the small-scale data preprocessing in that the former has high robustness with low data quality. Big data cleaning aims to promote the quality of data form, whether the data is Tidy Data. Hadley Wickham (2014) put forward the concepts of Tidy Data

as well as Data Tidying, and proposed that Tidy Data should follow three basic principles: each variable must have its own column; each observation must have its own row; each value must have its own cell. In most cases, completely tidy data cannot be obtained after once data cleaning operation. Therefore, these middle data that may contain messy data need to be audited again and then cleaning is continuing. Müller and Freytag (2005) pointed that data cleaning is an iterative and normally never finished process that consists of the four consecutive steps: data auditing, workflow specification, workflow execution and post-processing and controlling.

Data Conversion. When the form of the original data does not meet the requirements of the analysis algorithm, it is necessary to perform Data Conversion on the original data from one data format or one big data environment to another (Gao et al., 2016). The usual techniques for data conversion include: smoothing data by binning regression and clustering to remove noise from the data (Han et al., 2011), constructing new features based on other features, and performing data standardization such as Min-max normalization and zero-mean normalization.

Data Integration. Data integration is the practice of combining data from different sources, and then providing the user with a unified view of these data (Lenzerini, 2002). Representative tools and techniques for data integration include Manual Integration, Common User Interface, Integration by Applications, Integration by Middleware, Uniform Data Access, and Common Data Storage (Amghar et al., 2019). But there are many challenges for the process of data integration, such as the entity identification problem, some attributes of the data set are redundant, different data sources have different measurement scales.

Data Masking. The main purpose of data masking is to protect sensitive data (Kuacharoen , 2014), such as someone's address or phone number. Specifically, data masking is the process of transforming the individual (or organization) sensitive data to reduce the sensitivity of information on the premise of not affecting the accuracy of data analysis. Typical techniques for data masking include: substitution, shuffling, number and date variance, data encryption, deleting sensitive data and replacing with NULL values (Sarada et al., 2015; Mansfield-Devine, 2014).

Data Reduction. Miles and Huberman (1994) explained that Data reduction is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that "final" conclusions can be drawn and verified. Data reduction hardly affects the results of data analysis subsequently. There are two methods used commonly for data reduction: dimensionality reduction and numerosity reduction (Ghojogh & Crowley, 2019). The former usually uses linear algebra methods such as PCA, SVD, FLDA and DWT, the typical method of the latter is SOS (Kalegele et al., 2013).

Data Annotation. Data Annotation is the process of adding metadata to the data enabling modelling (Nagowah et al., 2019) which is adding necessary contexts of color, texture, shape, keywords, or semantic information. Data Annotation involves coding, rating, grading, tagging, and labeling of data (Carpenter, 2008).

Data Wrangling tools include Excel, SQL, Python, R, and Trifecta. In traditional data warehousing, Data Wrangling was carried out using Extract-Transform-Load (ETL) platforms, with significant manual involvement in specifying, configuring or tuning many of them (Koehler et al., 2017). It is needed to adopt adaptive, pay-as-you-go solutions that automatically tune the wrangling process, which means that users are able to contribute effort to the process of data wrangling in whatever form and at whatever moment they choose (Furche et al., 2016).

3.3 Data Computing

The next step of data preparation is to gain an insight into the value from big data and solve various problems of big data. Big data computing is an effective way that combines large scale compute, new data intensive techniques and mathematical models to build data analytics (Kune et al., 2016).

The four dimensions (volume, variety, velocity, and veracity) of big data bring challenges to traditional methods of data computing. Firstly, the volume of big data is exploding. Data computing can no longer be done by a single computer, but can only be shared by multiple machines. Secondly, the types of big data is various. Christie Schneider, a Watson marketing lead of IBM, claimed that more than 80% of today's data is unstructured (Schneider, 2016). It is hard to present in columns and rows and calculated in a structured database. The unstructured data is a big hurdle in computing and analysis part as they do not have a common format (Prasad & Agarwal, 2016). Thirdly, the data is growing at a high speed, and about 1.7 megabytes of fresh information will be created every second by 2020 (Dadheech et al., 2019). Fourthly, veracity of the data will directly affect the results of data calculation and data analysis. Veracity is probably the toughest nut to crack, and one of the biggest problems with big data is the tendency for errors to snowball (Tee, 2013). There are a lot of false, noise, or dirty data mixed in with valuable data. How to calculate these data is a big problem for traditional data computing technology.

Therefore, the traditional methods of data computing are not suitable for big data. A solution dividing data into small data units and calculating in the storage place where the data is located was found. Cloud computing is a distributed computing paradigm and differs from traditional ones such as centralized computing and grid computing. The emergence of MapReduce promoted the development of big data computing, and it has quickly become the current mainstream tool of the big data computing models. There are also some representative tools of cloud computing: Google GFS, Google BigTable, Spark and YARN.

Batch computing and stream computing are two important forms of big data computing (Sun et al., 2015). Batch computing is a big data computing method in which data is collected uniformly, stored in a database, and then processed in batches. Streaming computing is to process the data stream, which is a method of requiring real-time computing. Nowadays, there are many researches and applications related to batch computing. The basic MapReduce model and its implementations like Hadoop, is completely focused on batch processing (Shahrivari, 2014). Hadoop provides a distributed file system and off-line batch computing framework (Lin et al., 2013). The traditional batch computing process is carried out after a certain amount of data is accumulated; stream computing can achieve real-time processing and effectively reduce processing delay. While big data is becoming ubiquitous, the demand for large-scale processing of data streams is becoming increasingly urgent, which leads to the sprout of many distributed stream computing systems (Lu et al., 2014). The current widely used stream computing frameworks are Spark Streaming, Flink, Storm, S4 and Kafka.

3.4 Data Management

Data management refers to management activities, including acquiring, validating, storing, and processing required data to ensure the accessibility, reliability, and timeliness of the data to users (Myers, 2019). Data Management Maturity (DMM) model is a comprehensive framework designed in 2014 by CMMI of data management practices in six key categories include: Data Strategy, Data Governance, Data Quality, Data Operations, Platform & Architecture, Supporting Processes. According to the CMMI's introduction (2019), it is used to "provide

the best practices to help organizations build, improve, and measure their enterprise data management capability allowing for timely, accurate and accessible data across your entire organization". While in the age of big data, data management is the practice of organizing and maintaining data processes to meet ongoing big data lifecycle needs. Managing the ubiquitous big data is a challenge for data scientists. A database is a structured collection of data stored in a computer (Bai & Bhalla, 2020). Users store the data of transactions to be managed in the database, which helps to organize, maintain, process, and utilize data more conveniently. Data scientists must be proficient in not only traditional relational database, but also some emerging technologies such as NoSQL, NewSQL and relational cloud for data management.

Database, Data Warehouse and Data Lake are different data storage approaches for data management. Table 1 shows the studies that related to database, data warehouse, data lake.

Table 1 Comparison between database, data warehouse and data lake

Comparing Dimension	Database	Data Warehouse	Data Lake
oriented	application –oriented (Velicanu & Matei, 2007; Bontempo & Zagelow, 1998; Warners & Randriatoamanana, 2016)	subject –oriented (Velicanu & Matei, 2007; Bontempo & Zagelow, 1998; Warners & Randriatoamanana, 2016; Meredith et al., 2008)	operation –oriented (John & Misra, 2017)
data	Structured (Bontempo & Zagelow, 1998)	Structured (Bontempo & Zagelow, 1998)	Structured, Semi –structured, Unstructured (original forms) (Khine & Wang, 2018; John & Misra, 2017)
objective	support the operational system (Velicanu & Matei, 2007; Meredith et al., 2008; Vaisman & Esteban, 2014; Lechtenbörger & Vossen, 2003)	support the decision –making system (Velicanu & Matei, 2007; Meredith et al., 2008; Vaisman & Esteban,2014; Mallach, 2000; Lechtenbörger & Vossen, 2003)	Support dynamic analytical applications (for query) (Khine & Wang,2018; John & Misra,2017)
integration	Limited integration (Bontempo & Zagelow,1998)	Integrated (Bontempo & Zagelow,1998; Warners & Randriatoamanana,2016; Meredith et al.,2008; Lechtenbörger & Vossen,2003)	Integrated (Miloslavskaya & Tolstoy,2016)
Update frequency	High (Bontempo & Zagelow, 1998; Meredith et al., 2008; Vaisman & Esteban,2014)	Low (Bontempo & Zagelow, 1998; Warners & Randriatoamanana,2016; Meredith et al., 2008; Vaisman & Esteban, 2014; Lechtenbörger & Vossen, 2003)	High (Khine & Wang,2018; Aftab & Siddiqui,2018)
Usage	Predictable retrieval (Bontempo & Zagelow,1998; Meredith et al., 2008; Vaisman & Esteban,2014)	Ad hoc retrieval (Bontempo & Zagelow,1998; Meredith et al., 2008;Vaisman & Esteban,2014; Lechtenbörger & Vossen,2003)	Real time analytics (Madera & Laurent,2016)

Comparing Dimension	Database	Data Warehouse	Data Lake
Data modeling	UML, ER model (Meredith et al., 2008;Vaisman & Esteban, 2014;Lechtenbörger & Vossen, 2003)	multidimensional model (Meredith et al., 2008;Vaisman & Esteban, 2014; Lechtenbörger & Vossen, 2003; Mallach, 2000)	No specific data model (John & Misra,2017;Aftab & Siddiqui, 2018)
User type	Operators,office employees (Vaisman & Esteban,2014)	Managers,executives (Vaisman & Esteban, 2014), analysts (Meredith et al., 2008)	Data Scientists (especially those familiar with domain) (Madera & Laurent,2016)
Access frequency	High (Meredith et al., 2008; Vaisman & Esteban,2014)	From medium to low (Meredith et al., 2008;Vaisman & Esteban,2014)	accessible as soon as it is created (Khine & Wang,2018; Miloslavskaya & Tolstoy,2016)
Access type	Read,insert,update,delete (Meredith et al., 2008;Vaisman & Esteban,2014)	Read,append only (Vaisman & Esteban,2014),select (Meredith et al., 2008)	Read and write (Khine & Wang,2018;Miloslavskaya & Tolstoy,2016)
Response time	Short (Vaisman & Esteban, 2014;Lechtenbörger & Vossen, 2003)	Can be long (Vaisman & Esteban, 2014)	Short (Madera & Laurent, 2016)
normalized level	normalized tables (Warners & Randriatoamanana,2016; Meredith et al., 2008;Vaisman & Esteban,2014)	non–normalized(Warners & Randriatoamanana,2016; Meredith et al., 2008;Vaisman & Esteban,2014)	non–normalized(Southwick, et. al., 2015)

3.5 Data Analysis

Exploratory Data Analysis (EDA) was proposed by American statistician John Tukey in the 1970s (Tukey, 1977). Any activity of a Data Science project starts with EDA (Putatunda et al., 2019). When data scientists are faced with a variety of messy "dirty data" and do not know how to understand the data immediately. Exploratory data analysis is an effective way to help them achieve the purpose of data understanding and lay the foundation for subsequent data analysis.

According to purposes, the data analysis can be divided into descriptive analysis, predictive analysis and prescriptive analysis (Sivarajah et al., 2017). Descriptive analysis is most used in business analysis and it mainly solves the problem of "what has happened" by analyzing the collected data to obtain various quantitative characteristics reflecting objective phenomena. It includes data dispersion analysis, concentration analysis and frequency analysis. Predictive analysis is based on history and facing the future and it mainly focuses on what will happen in the future by the means of data mining and statistical modeling tools to analyze historical data, so as to predict what will happen in the future or the probability that something will happen. A typical method of predictive analysis is time series analysis. Finally, prescriptive analysis is a practice-oriented method mainly to solve the problem of "what should be done"

by analyzing what has happened, the causes of events and various possibilities in order to help users determine and choose the best actions and measures.

Traditional data analysis is deeply influenced by the formal theories of statistics (Tukey & Wilk, 1966). It mainly uses sampling data to infer the real situation, which means that traditional data analysis needs to extract useful information when the amount of data is limited. With the rapid growth of data volume in modern society, traditional data analysis is gradually turning to big data analysis. Big data analysis is mainly to acquire insights from all data (not sampling data) to support decision making, without considering the distribution status of data and without hypothesis testing. Traditional data analysis tools are not enough to manage big data, so some open-source tools for big data analysis are indispensable to Data Science. The most popular open-source tools for big data analysis are R and Python.

3.6 Data Products Development

Data Product is a kind of product that facilitates an end goal through the use of data (Patil, 2012). Data product development is indispensable for Data Science. Data product development activities are rarely undertaken in a traditional product development sequence that involves identifying the need, developing the product. On the contrary, data product development activities often take place in a continuous, iterative fashion, with the important activities conducted in parallel (Davenport & Kudyba, 2016). And the ability to develop data products is becoming increasingly critical to every business in big data era. Therefore, one of the missions of Data Science project is to develop data products.

Unlike traditional industrial products, data products can be entities or invisible objects. Cao (2017) defined data product as an output of Data Science which is "from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system". Data product refers to anything that can help others use data to achieve goals. For example, Google Glass is a data product enabled by Google big data. Data products include data set products, information products, knowledge products, and intelligence products.

Data product development involves all activities of the Data Science project process, including datafication, data munging, data tidying, exploratory data analysis, data analysis, data product development. Not only is the result of a Data Science project a data product, but the intermediate product created by each activity is also a data product. Data Jujitsu is the art of turning data into products (Patil, 2012). It focuses on that the process of data product development must be highly artistic and centered on target users.

4 Domain–specific Data Science

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

4.1 Data Journalism

As one of the new research directions of Journalism, Data Journalism is a way of seeing journalism as interpolated through the conceptual and methodological approaches of computation and quantification in the era of big data (Parasie & Dagiral, 2013; Lewis, 2015). The development of data journalism has roughly gone through the following three stages. At the

beginning, the typical event is the report titled Investigation of the Education System for Juvenile in The Guardian in 1982. This report breaks the narrative mode of traditional news (Timetoast, 2021). However, data journalism in that period lacked the necessary technical means, and there was no systematic theoretical support and was only scattered attempts. In the period of precision journalism, the typical event is Meyer's "Precision journalism: A reporter's introduction to social science methods (Meyer, 2002). Statistical and social survey research methods are introduced into news practice to collect data scientifically and improve the accuracy and objectivity of data in news reports. Although the data at this stage was paid attention to, news reports were still mainly narrative. Figures and charts were also only auxiliary to news reports. In the period of data journalism, the typical event is the establishment of Data Journalism Awards in 2012. This marks the beginning of data journalism as a new form of news that has received widespread attention.

Research hotspots of data journalism include the following aspects: Talent training of data journalism, presentation of data journalism, combination of data journalism and artificial intelligence. The talents required for data journalism have interdisciplinary characteristics. They not only need to understand traditional news acquisition and editing methods, but also need to master the theories and techniques of Data Science. Therefore, how to establish a new talent training model that meets the needs of data journalism has become a hot issue discussed by academia and journalism. For example, "The Data Journalism Handbook" co-written by journalists from various countries provides practical operating procedures and classic cases for talent training (Gray et al., 2012). Computational Journalism courses are available at Columbia Journalism School and Stanford University (Jstray, 2012; Nguyen, 2016). The presentation of data journalism has changed from the original table and visualization to the present storytelling and gamification. This has always been a hotspot in this field. Storytelling of data is different from visualization. It can provide data journalism with stronger narrative and better user experience. Gamification is the use of games to tell data stories, which can capture readers' interest more than traditional news. The combination of data journalism and artificial intelligence has become a new trend in the development of data journalism. For example, Reuters began to use artificial intelligence technology to track social media networks such as Twitter to obtain newsworthy events and discussions in 2017 (Matthews, 2019). This approach can save a lot of labor and time, so it has been imitated by many news media. The Post uses artificial intelligence technology to combine data with story templates to develop software for automatic news writing (Underwood, 2019).

The research challenges of data journalism mainly include two levels: data and technical. The difficulty at the data level lies in how to judge the reliability of the data source and how to ensure that the data is true during data processing and analysis. Common data sources currently include publications, traditional media, the Internet, government public data, and public geospatial data. It is generally believed that government public data and publications have high credibility, while the data authenticity on the internet is low. Data authenticity is the life of data journalism, but it is not easy to achieve data authenticity in the practical applications. For example, operations such as cleaning, conversion, merging or reshaping data during data processing may inadvertently cause data errors. Thus, the data is not true. Technical difficulties, from the perspective of practitioners of data journalism, are mainly the difficulty of mastering new technologies. For example, it is not easy to master the technologies commonly used in the field of data journalism, such as data analysis models, artificial intelligence technologies, Java, HTML, and Python. From the reader's point of view, technology in-

evitably increases the cognitive burden when it provides an intuitive presentation effect. For example, Java-based interactive data journalism often requires the installation of corresponding software to run normally.

There are many typical applications of data news, which mainly focus on health news and financial news. Take health news as an example, "Battling Infectious Diseases in the 20th Century: The Impact of Vaccines" from The Wall Street Journal (DeBold & Friedman, 2015). The background of the work was that many families in the United States believed that the children were too young to have an inoculation, which would affect the health of the children. Using a calendar heat map, based on 70 years of publicly available government data in the United States, the team created the work, which shows that vaccination has led to a significant decline in the number of people falling ill from the epidemic. This work reassured fearful parents with real data.

The main breakthrough of data news is the construction of professional data news teams and the positive usage of new technologies. The production of data news is a process in which multiple departments cooperate and multi-types of work full participate. It is necessary to break the internal departmental restrictions of traditional news organizations and establish a professional data news team that meets the new needs of the development of data news. Team members should cover all types of work of data news production, from journalists and editors to designers and technicians. At present, data news mainly uses data visualization technology to visualize data. However, data visualization is not the only choice. The rise of artificial intelligence technology provides a new opportunity for the development of data news. The automatic generation of data news has become a new trend in data news utilization technology.

4.2 Industrial Data Science

Industrial big data mainly studies how to apply big data in the field of industrial manufacturing, so as to realize the innovation of industrial manufacturing. Different from the previous focus on internally structured data, industrial big data needs to focus on the entire life cycle data of products and services in the industrial field, including structured and unstructured data (Ministry of Industry and Information Technology of the People's Republic of China, 2020).

The research hotspots in the field of industrial big data can be divided into two aspects: How to establish industrial big data and how to use industrial big data. Different countries have put forward different plans on how to establish industrial big data, the most representative of which are German Industries 4.0, Industrial Internet of the United States and Made in China 2025. In the practice of industrial big data, research hotspots can be subdivided into: attribute data extraction (Ma et al., 2014), data management philosophy and standards (Zhou et al., 2016), data-centric business operations, Physical deployment, cloud storage and supporting software platform deployment of the Internet of Things (Raptis et al., 2019). How to use industrial big data? There are four types of common products, respectively is process visualization, process optimization, decision support, fault detection. Process visualization and fault detection are mainly oriented to the production field, and visualization 3D modeling algorithm (Yandun et al., 2020) and fault identification algorithm (Dahbura & Masson, 1984) are the research hotspots at present. Process optimization and decision support are oriented to the activities in multiple fields of production and management. At present, the

research focus is the establishment of process optimization model, decision algorithm and relevant management formulation (Kruzhilko & Maystrenko, 2019; Hollowell et al., 2019).

The research challenges of industrial big data are divided into: in the industrial scenario, the data format of multiple data sources is not uniform. There are structured data, unstructured data, and non-digital data (hand-drawn charts). How to digitize and unify the format of these data is one of the difficult problems faced by industrial big data. In addition, data is generated in real time in industrial scenarios. How to store and analyze these data is also difficult. Moreover, the high utilization of data requires the establishment of supporting data collection, storage, processing, analysis, utilization standards and long-term mechanisms. But the reality is that the establishment of utilization mechanism of industrial big data requires a lot of manpower and financial resources. Finally, it is difficult to find new insights from massive amounts of data. This requires professional domain knowledge, keen insight and Data Science capabilities, but there is currently a lack of such compound talents (Davenport & Patil, 2012).

Typical applications of industrial big data: In the automobile manufacturing industry, automobile manufacturers collect vehicle conditions and operating habits of the owner through the sensors that come with the automobile, and then analyze the returned data to improve services and quality of products. In addition, the use of industrial big data can also help factories quickly discover machine failures and deal with them in time to reduce losses.

The main breakthrough of industrial big data lies in the guidance of government's policy. Because companies that develop industrial big data require high upfront investment, they are currently dominated by large companies. The participation of small and medium-sized enterprises is not enough, and if things go on like this, they will lose their competitiveness. Therefore, the guidance of government's policy can help small and medium-sized enterprises to participate in industrial big data. At the same time, it can also guide sharing and co-construction of data between enterprises to save social costs.

4.3 Business Data Science

Business Big Data was used to support business decision or produce products via precision marketing, user profiling and advertising. Consumption big data comes from the links related to product sales, such as customer registration data, order data, browsing record, purchase record, evaluation, consultation, feedback, complaint, suggestion. Research in this field focuses on how to use consumption big data. According to the analysis purpose, it can be divided into descriptive analysis, predictive analysis, diagnostic analysis and prescriptive analysis. Descriptive analysis mainly uses the descriptive statistical information of the data, such as median, mean value, standard deviation, to understand the distribution and characteristics of the data, so as to help the merchants understand the current situation of the goods as a whole. Predictive analysis mainly studies how to use models to predict unknown situations, such as establishing linear regression models using economic and population variables to predict electricity consumption (Bianco et al., 2009), using multiple random forests to predict urban water consumption (Chen et al., 2017), and using neural network technology to predict the online buying behavior of Indian buyers (Prashar et al., 2016). Diagnostic analysis mainly looks for the reasons that influence buying behavior, such as discussing the influence of advertising, social media (Zhang & Pennacchiotti, 2013) and website functions (Zhao et al. , 2016) on buying behavior. Normative analysis mainly studies how to make plans to increase

product sales, such as bundle sales (Kaserman, 2007), bonus incentive policies (Chung et al., 2014), comprehensive sales strategies (Leigh & Marshall, 2001).

The research challenge of consuming big data is not the technology of data collection and utilization, but in the legitimacy of data collection and utilization. In 2018, the California Consumer Privacy Act of 2018 (CCPA) was issued by the California Government of the United States, which restricts some of the rights of enterprises to collect and use information, and increases the right to know and Opt-Out right of users. Cambridge Analytica obtained data of as many as 87 million people from Facebook (including sensitive data such as personal accounts, personality tests and social networks of users), and sold it to the Trump presidential campaign to accurately display customized messages for specific groups of people (Grothaus, 2018). As a result, Facebook was fined $643,000 and Cambridge Analytica went bankrupt (Zialcita, 2019). Therefore, how to reasonably collect and use data under legal circumstances is an important problem faced by the consumption big data.

Recommendation system is a typical application in consumption big data. It predicts users' shopping tendency based on user-related consumption big data, so as to select products similar to users' buying tendency for recommendation. For example, Companies like Taobao, Youtube use recommender systems to help their users to identify the correct product or movies. According to a McKinsey survey, Netflix saves the company about $1 billion a year. Amazon owes 35% of its annual revenue to the recommendation system (Sigmoidal, 2017).

The main breakthrough in consumption big data is how to establish long-term relationship with users, which can be divided into two levels. For the first level, as privacy protection is getting more and more attention, it is needed to obtain user authorization if you want to collect and use data, and establishing trust relationship with users is helpful to obtain user authorization. As for second level, enterprises need to establish user loyalty programs to maintain user loyalty through organizing activities and giving small gifts regularly.

4.4 Health Data Science

With the gradual popularization of cordless medical treatment, electronic medical records, and online consultation, the work process in the medical and health field tends to be digitized, resulting in health big data. It mainly focuses on the wide application of big data in health and medical fields including life logging (Gurrin et al., 2014), medical diagnosis, pharmaceutical production, and health care (Raghupathi & Raghupathi, 2014).

The research hotspots of health big data mainly include precision medicine, disease identification and monitoring, and data privacy. For precision medicine, most of the existing research is to explore its feasibility and racial bias (Gurrin et al., 2014). For disease identification and monitoring, it involves the application of machine learning, natural language processing and other technologies in the health field. For example, Automated Identification of Surveillance Colonoscopy in Inflammatory Bowel Disease Using Natural Language Processing (Hou et al., 2013), Prediction of fatty liver disease using machine learning algorithms (Wu et al., 2019). For data privacy, the focus is on emphasizing the importance of privacy and proposing feasible ways to protect privacy. For example, the privacy issues were explained in health big data from a legal and technical perspective (Mounia & Habiba, 2015). A security life cycle model for health big data was proposed (Abouelmehdi et al., 2018). And a security framework and algorithm were built (Chandra et al., 2017).

The difficulty in health big data research is that health data involves important private data

of patients. Therefore, how to ensure the safety of health data in the process of transmission, storage and analysis is very necessary (Mooney & Pejaver, 2018). But at present, security in health big data has not been paid enough attention. The reason is that there is currently no relevant law clarifying the responsibilities of owners of health data. In addition, the protection of data security requires a lot of manpower and financial resources. For example, Community Health Systems was exploited by hackers to obtain social security numbers, dates of birth, phone numbers and actual addresses of 4.5 million patients in 2014. In 2015, Medical Informatics Engineering, an electronic Medical record software company, leaked the data of 3.9 million patients. The leaked content included names, social security numbers, phone numbers, mailing addresses, dates of birth, diagnosis and other sensitive information (Lord, 2020).

The successful application of Google Flu Trends (GFT) is a typical application that utilizes big data on health. In 2009, Jeremy Ginsberg, Matthew H. Mohebbi and Rajan S. Patel published a paper titled "Detecting Influenza Using Search Engine Query Data Based on Search Engine Data" in Nature. This paper introduces GFT, a flu prediction tool launched by Google in 2008, which can predict the nationwide spread of H1N1 in real time, overcoming the lag of official data release. The successful application of GET plays an important role in promoting the application of health big data.

The main breakthrough of health big data is that data masking can be used to protect the privacy of patients. Common technologies of data masking include data encryption, data randomization and data replacement technologies. Among them, data encryption is a reversible method of data masking. This may be cracked through the ciphertext, and the data needs to be decrypted before being used. Data randomization means that when collecting customer information on the server side, if only interested in the attributes of the information in the overall statistical sense, the client can use random algorithms to interfere with data privacy. For example, some real information is randomly deleted, and some false information is introduced to protect personal privacy. This technology can meet the needs of aggregated attribute while desensitizing data. Data replacement includes data pseudonymization, shuffling, and synthetic data.

4.5 Biological Data Science

Harnessing powerful computers and numerous tools for data analysis is crucial in drug discovery and other areas of big-data biology (Marx, 2013). The principles, theories, methods, technologies, and tools of big data are widely adopted to biology, and biological research paradigm is transferring from knowledge-centered paradigm to data-centered paradigm.

Its research problems include three aspects: (1) the development of gene analysis towards "de-sampling", scientists manage to apply big data technologies to efficiently analyze all data of DNA and RNA, instead of sampling analysis. (2) The traditional methods of biological research are to examine and determine the structure and characteristics of the subject using a variety of experimental techniques, such as NUCLEAR magnetic resonance and X-ray crystallography, and new methods such as cryo-electron microscopy, but these methods rely on a wide range of trial and error. The development of big data makes it possible to make strong predictions about complex structures through deep learning. The AlphaFold (2020) used by CASP14, for example, creates an attention-based neural network system that treats protein residues as nodes, connecting neighboring residues together. (3) The transformation of drug discovery to "precision". The example is the analysis of HIV drug resistance. Stanford

University in the United States established a special database, Hivdb. By sequencing HIV from patients in the database and comparing it with standard sequences, drug-resistant mutations can be found to know which drugs are no longer effective for that particular patient, and the remaining drugs can be combined to suppress HIV (Stanford University, 2021).

Big data research focus in the biology mainly includes: "gene sequencing + artificial intelligence" and "deep learning + medical image" and "big data + health records" (1) "gene sequencing + artificial intelligence" refers to the use of machine learning methods, prediction on the genome will change the characteristics of human body/disease/how to impact on phenotype. The implementation method is divided into two steps. First, identify the gene susceptibility locus associated with a characteristic/disease/phenotype. Second, use machine learning to simulate changes in characteristics/diseases/phenotypes. (2) "Deep learning + Medical imaging" refers to the direct analysis of medical images by deep learning algorithm. The existing image processing method is to treat each layer of 3d medical image as 2d image separately, and there are also methods to directly process 3D image after reducing complexity. The detection methods can also be divided into the method of locating and then classifying and the method of directly predicting the target location. (3) "Big data + Health archive" refers to the information about personal lifelong health status and health care behaviors managed electronically, which involves all the process information of patient information collection, storage, transmission, processing, utilization, and integrates information into a huge database. For example, the data volume of PubMed, the internationally famous biomedical database, reaches nearly 20 million records, increasing at a rate of 600,000 to 700,000 each year. The biomedical and pharmacological literature database Embase has more than 11 million records, with 500,000 more records added every year.

Medical Ethics and data security in the era of Big data. On the one hand, the development of science and technology is increasingly dependent on big data, and open source and data sharing have become an important driving force for biological research. But as concerns grow about privacy, particularly genomic privacy, access to important information, such as personal genome data, may be restricted in the future. On the other hand, the more involved the patient, the more likely the biomedical research project is to succeed. However, how to benefit the patients and how to share the benefits is a problem that people face.

Typical applications of big data biology: (1) clinical effect testing. For example, Germany's RWTH Aachen university (RWTH helmholtz-institute for biomedical engineering), German cancer research center (DKFZ), German cancer research association (DKTK) and Heidelberg (NCT) national center for tumor disease scientists have developed a kind of adaptive algorithm, can be directly according to the tumor HE staining tissue slice image prediction of microsatellite instability (MSI), which helps to identify potential can benefit from the immune therapy of gastrointestinal cancer patients. (2) Establish a library of genomics pre-training models. 23andme, an emerging technology company in Silicon Valley, takes the lead in the commercialization of precise SEQUENCING of DNA sequence to deal with diseases caused by genetic code. Based on Apriori algorithm and linear recursive model, Goran Hrovat utilizes big data visual analysis technology to explore patient data and serve for hospital management and decision-making.

4.6 Social Data Science

Social big data comes from joining the efforts of the two previous domains: social media and big data (Bello-Orgaz et al., 2016). Applications of social big data can be extended to a

wide number of domains such as health and political trending and forecasting, hobbies, e-business, cyber-crime, counterterrorism, time-evolving opinion mining, social network analysis, and human-machine interactions. Its research problems mainly include two aspects: big data based on content and big data based on opportunity network. The former focuses on extracting insight from user-generated content across a variety of social media platforms, while the latter focuses on extracting knowledge from interactions between online users by analyzing the web (Zhang et al., 2019).

The research hotspots of social big data mainly include the development and improvement of data mining and data analysis technologies used in social big data and the research on the methods of applying big data to different social fields, such as e-commerce, marketing, journals and public policies.

The research challenges of social big data include knowledge representation, data management, data processing, data analysis, data visualization and other aspects for mass data (Kaisler et al., 2013). Specific examples include accessing a large amount of unstructured data, determining how much data is sufficient to have a large amount of high-quality data, dealing with dynamically changing data streams, or implementing sufficient privacy (ownership and security). One of the most challenging problems is to identify valuable data from large heterogeneous datasets from social media, and analyze that data to discover useful knowledge and improve decisions for individual users and businesses. In order to correctly analyze social media data, traditional analysis techniques and methods need to adapt to and integrate the new big data paradigm to form structured data processing.

Typical applications of social big data include myriads of applications related to marketing, crime analysis and user experience. Marketing applications include advertising on social platforms. Maurer and Wiegmann (2011) analyzed the effectiveness of advertising on social networks. The experiment found that when the social network ads were placed in front of unscreened subjects alone, most of them thought the Facebook ads were annoying, and that placing the same ads on social interactions generated by Facebook tools and applications increased the number of visits and purchases by Jigar consumers. Applications of crime analysis include identifying patterns of crime through big data, allowing the detection and discovery of crimes and their relationships with criminals. Crime hotspots can be identified using a variety of mapping techniques, such as point mapping, geographic area thematic mapping, spatial ellipse, grid thematic mapping, and kernel density estimation (KDE). For User Experience-based Applications, Big data from social media needs to be visualized for better user experiences and services. For example, large amounts of digital data (usually in tabular form) can be converted to different formats. Thus, user intelligibility can be improved. The ability to visualize such big data to support timely decision-making is critical in areas as diverse as business success, drug therapy, network and national security, and disaster management (Keim et al., 2013). Therefore, user experience-based visualization is recognized as an important tool to support decision making. Visualization is also recognized as an important data analysis tool for social media (Kotval & Burns, 2013). It is important to understand what users want from social networking services. There are many visual ways to gather (and validate) the user experience. One of the most famous ways is interactive activity data analysis.

The main breakthrough of social big data lies in the increasingly advanced analysis technology and the increasing risk of privacy leakage. Therefore, many researches on privacy protection have been put forward to solve the problems related to privacy. We can note that there are two well-known methods. The first is to take advantage of "k-anonymity," which is

an attribute of some anonymous data (Sweeney, 2002). Given private data and a specific set of fields, the system (or service) must make the data useful without identifying the body of the data. The second approach is "differential privacy,, which can provide an effective way to maximize the accuracy of statistical database queries while minimizing the opportunity to identify their records (Dwork, 2008).

4.7 Agile Data Science

Agile Big Data is a development methodology that copes with the unpredictable realities of creating analytics applications from data at scale (Jurney, 2017). It is helpful to develop agile software, manage agile projects and establish agile organizations. The philosophy and principles of agile big data include four aspects: componentization and platformization, unification and openness, standardization and interface, self-service and intelligence, and engine driving. Componentization and platformization refer to the modularization abstraction of big data processing links to form a componentized platform with high cohesion of multiple functions. Componentized platforms can be used independently with existing platform components or combined to solve more problems on different links. Unification and openness refers to achieving a balance between simplifying system complexity, improving management and control ability and enhancing fitness, and improving flexibility. Standardization and interface refers to the formation of a series of standardized protocols in big data processing links, including data namespace protocol/metadata and data type specification protocol/data Access Interface protocol/Query language protocol/data transmission protocol/data security protocol. Intersystem interactions are provided in the form of service interfaces and queue interfaces. Self-service and intelligent routine operations including self-service can be better supported in an automated manner; Self-service insight analysis can be better supported in an intelligent way. Engine-driven includes the introduction of advanced engine-driven capabilities to enable agile big data applications to reach external audiences more quickly and actively. At this time, big data applications themselves have become a powerful business-driven engine. Operations including self-service can be better supported in an automated manner; Self-service insight analysis can be better supported in an intelligent way.

The main research content of Agile big data includes three aspects: feature extraction, fusion encapsulation and service interface. 1) Feature extraction: the structured and unstructured data and semi-structured data for data integration and feature extraction, extracted the data of all kinds of different characteristics, including time, space, characteristics or other global features, implementation of data related to the location of the associated attributes, time, space and other observation attribute such as the characteristics of the description. 2) Fusion encapsulation: All kinds of extracted data features or preliminarily preprocessed data are encapsulated into data processing units with unified structure and format according to data processing characteristics and requirements of different computing models, forming standard analysis data sets and providing fast data adaptation for the upper mining and computing services. Metadata definition method and XML/JSON and other technologies can be used to realize the unified definition of different types of data units, and basic information and various attributes definition and description can be carried out for each type of unified data unit, including identification ID, basic attribute, semantic attribute, structural attribute, and other contents. 3) Service interface: encapsulated unified data unit data sets, according to different computing service model to realize fast data adapter, uniform data unit call interface design, through the interface definition and parameter setting unit to encapsu-

late data parsing, and the data sets of various attributes, such as structure information are extracted (Bello-Orgaz et al., 2016).

The key breakthrough of Agile Big Data is how to achieve a unified, standardized, modular and configurable big data architecture to solve the problem of difficult integration between different types of heterogeneous subsystems. Application functions can be combined with existing functional components, and the cost can be reduced through service reuse. The form of data exchanged between components should be standardized and interlaced. Components can be combined with minimal programming or configuration, standardized integration of common models and tools, and simplified usage to provide out-of-the-box data mining and analysis capabilities to non-programmers; Big data application whole process (collection, storage, analysis, management) visualization operation. Based on the iterative nature of the scientific data and using efficient componentized tools, for big data each function subsystem (modules) modular, standardized design model, and according to the actual demand fast quick selection, configuration, structures, large data prototype system, the rapid iteration big data analysis results, and adapt to changing needs, the prototype as soon as possible into a production system. In the process of rapid iteration, rapid feedback and closed-loop verification, customers can gradually complete the reform of system thinking and management thinking of big data analysis, and the principle of quick proof and lean design is the core goal of agile big data application.

5 Integrative studies of Domain–general Data Science and Domain–specific Data Science

There is a wide range of disciplines which have developed Domain-specific Data Science as discussed in Section 4. However, DSDS varies from one domain to another, and different DSDS domains have their own unique research perspectives and interests on Data Science. At the same time, there are also some studies, which focus on Data Science itself and are intent on building Domain-general Data Science.

5.1 Nexus between DGDS and DSDS

There are subtle nexuses between Domain-general Data Science (DGDS) and Domain-specific Data Science (DSDS).

First, their fundamental difference roots in the thinking paradigms: DGDS conforms to data-centered thinking, while DSDS is in line with knowledge-centered thinking. Knowledge-centered thinking pattern believes that data will be utilized effectively only when the causality in data is identified. Hence, data analysis in the past dedicates to find, validate and take advantage of causalities. That conventional thinking pattern is very effective in DSDS but inefficient in DGDS since it is hard to identify and validate a causality from large-scale data sets. As a result, the aims of data analysis have shifted from causal analysis to correlation analysis, which put more emphasis on the correlation analysis. In contrast with causality analysis, correlation analysis is time-saving and easy to put into practices. This separation of causality analytics and correlation analytics also triggers collaboration between data scientists and domain experts, and provides a new analysis pattern for DSDS data analysis. For example, applying data analysis to the banking industry can make it more agile. Bank of America has developed a virtual assistant called Erica, which uses predictive analysis and natural language processing to showcase information about bank transaction history or upcoming

bills for customers.

Second, DGDS focuses on the theoretical studies, while DSDS is centered on applied ones. DGDS involves basic theories and activities of Data Science, while DSDS focuses on the applications of Data Science in a specific domain. What DGDS has in common with DSDS is that Statistics, Machine Learning, Data Visualization, and Domain Knowledge are their theoretical bases, and the research in DGDS and DSDS drives the development of the Data Science. Applying DGDS to other specific domains is one of the popular topics in recent studies. Those specific domains include life science, medical care, social governance, education, and business management. As a result, some new research topics such as quantitative self, data journalism and big data analysis gained widely attention of data scientists.

Third, DSDS is domain-dependent, but DGDS is domain-independent. DSDS incorporates theories with domain knowledge and business practice, which was termed to address challenges that we are facing in data enriched offerings era. DGDS provides theoretical guidance for the practical application of DSDS, which involves knowledge that data scientists in every field should master. The core theories of DGDS include concepts, theories, methods, technologies, and tools focus on solving the problem of discipline construction. DGDS puts forward some methods, techniques, and tools at the macro level, which can be used in a few DSDS. It means that DSDS aims to use the core theories of Data Science to solve other disciplines' own problems.

Fourth, DSDS and DGDS possess complementary advantages. Table 2 shows the gaps between DGDS and DSDS in some dimensions of Data Science. Data Science projects are completed via collaborative efforts of domain-general data scientists as well as domain-specific data scientists. Data wrangling, for instance, is a value-added process that needs efforts from not only domain-general data scientists who are good at data-related tasks but also domain-specific data scientists who are familiar with specific business of application domains.

Table 2 the Comparisons of DGDS and DSDS

	DGDS	DSDS
Theory	*	
Practice		*
Data Wrangling		*
Data Computing	*	
Data Management	*	
Data Analysis		*
Data Products Development		*

5.2 Integrating DGDS and DSDS

Theoretical Data Science (TDS) is supposed to bridge the gap between Domain-general Data Science (DGDS) and Domain-specific Data Science (DSDS). Theoretical Data Science is a branch of Data Science that employs mathematical models and abstractions of data objects and systems to rationalize, explain and predict big data phenomena. This contrasts with DSDS, which uses casual analysis, as well as DGDS, which employs data-centered thinking to deal with big data problems in that it balances the usability and the interpretability of Data Science practices (Figure 2).

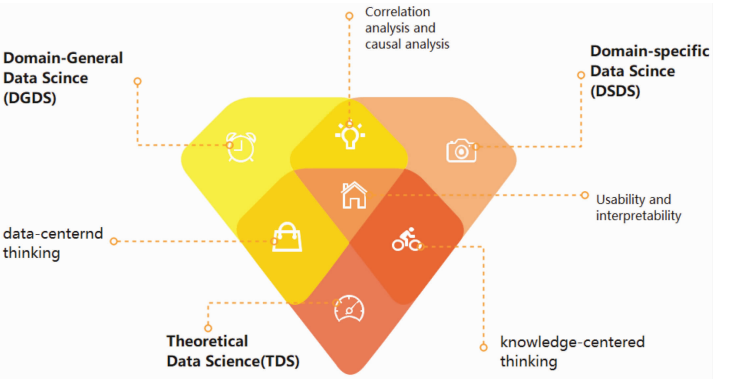

Figure 2 Three Types of Data Science

The main concerns of TDS are concentrated on the following topics:

(1) To integrate the data-centered thinking with the knowledge-centered thinking. Data-centered thinking is the unique thinking pattern of DSDS, while knowledge-centered thinking is the typical thinking pattern of DGDS. TDS integrates them by two different ways: data-centered thinking triggers knowledge-centered thinking, or vice versa. In practical Data Science projects, the integration of DSDS and DSDG is mainly implemented via the collaboration between professional data scientists and experts from other specific business domains.

(2) To transform correlation analysis into casual analysis. TGS believes that correlation analysis is insufficient to address big data problems, and the Data Science projects should convert correlation analysis into casual analysis. Further, TDS regards correlation analysis as the pre-requirements of causal analysis. Correlation analysis is conducted by employing machine learning or statistical methods. However, the causal analysis heavily depends on the related domain knowledge.

(3) To balance the usability and the interpretability. Contradictions between the usability and the interpretability of big data solutions are the trickiest challenges in Data Science studies. TGS balances them by introducing interpretable Machine Learning or explainable Artificial Intelligence. Interpretable methods of TGS can be classified into global interpretation and local interpretation. Global interpretability implies knowing what patterns are present in general, while local interpretability implies knowing the reasons for a specific decision (Doshi-Velez & Kim, 2017).

6 Conclusions

Theoretical Data Science (TDS) is an integrated study of Domain-general Data Science (DGDS) and Domain-specific Data Science (DSDS) in order to bridge the gaps between them. In contrast with DSDS as well as DGDS, TDS adopts the data-centered thinking pattern, recognizes that the property of data is more active than passive, manages to convert data into intelligence, solves data-intensive tasks, conducts data wrangling or munging, enhances user experiences of big data systems, introduces data intensive scientific discovery, as well as educates data scientists. TDS is unique in its scientific objectives as well as research paradigm, and does not replicate directly the experiences from DGDS and DSDS. The following topics are essential for further research on TDS.

1) To conduct in-depth theoretical research on Data Science. There are no shared understandings on Data Science yet. Some of the researchers insist that Data Science is merely interdisciplinary applications of Statistics and Machine Learning, and it does not need its own new theories. They argue that application of Statistics and Machine Learning is crucial for Data Science. They fail to admit the unique theories of Data Science. In fact, Statistics and Machine Learning are the theoretical foundation of Data Science, not its core components. Data Science is an independent discipline like Statistics and Machine Learning. TDS is unique in its scientific mission, research perspective, thinking pattern, underlying principles, and theoretical framework, which are distinct from other disciplines.

2) To take advantage of active property of big data. One of the main contributions of Data Science is that it shifts our thinking pattern and views big data as active beings. People have seen data as passive or dead thing to date, and how to input human intelligence into data is the main concern of the related studies. For instance, traditional data prepressing theories try to convert complex data into simple data through defining schema, data cleansing, and filling missing values. However, TDS highlights the active property of data and begins to discuss how to take advantage of data. As a result, some novel terms, such as data-driven applications, data-centric design, data insights, and big data ecosystem, are widely accepted. TDS regards complexity as a natural attribute of big data and does not conduct traditional data preprocessing. Admitting that data is active rather than passive is the basic starting point of studying TDS.

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

4) To shift Data Science´s research focus from correlation analysis into causality inference. There is a misconception that Data Science only concentrates on correlation analysis, and causality inference is outside the scope of it. However, correlation inference can only be used to identify the correlations in big data, but cannot guide how to optimize and intervene in the identified correlations. Where the correlation changes, the causation relation in big data is required to be analyzed. Hence, to shift the research focus from correlation analysis into causality inference is one of the unique purposes of TDS. In a TDS project, the data scientists are responsible not only to discover possible correlations in big data, but also to reveal the causality behind the correlations with the collaboration of domain experts. To embrace causality analysis is becoming one of the most discussed topics in Data Science. For instance, the course titled Experiments and Causality Analysis or the Causal Inference for Data Science are listed in DS courses at University of California, Berkeley, and Columbia University as well.

5) To take data product development as one of the main tasks of Data Science Projects. Developing data products is one of the distinct objectives of TDS studies. Data products in TDS are not limited to products in data form. All products that utilize data to provide new services should be regarded as a data product. Data can be used to promote product innovation, and traditional products will be transformed into data products by application of DS theories. Google Glasses, for instance, is a data product in that its novel features are derived

from data. Data-centered thinking is the fundamental difference between data products and traditional ones. Data products will be the most common applications of TDS.

Acknowledgements

This work was supported by the Ministry of education of Humanities and Social Science project (Project No.20YJA870003).

Conflicts of Interest

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

Reference

Abdallah, Z. S., Du, L. & Webb, G. I.(2017). "Data preparation" *in Encyclopedia of Machine Learning and Data Mining*. Springer Publishing Company, Boston.

Abouelmehdi, K., Beni–Hessane, A., & Khaloufi, H.(2018). Big healthcare data: preserving security and privacy. *Journal of Big Data, 5* (1), 1. doi: https://doi.org/10.1186/s40537–017–0110–7

Aftab, U., & Siddiqui, G. F.(2018). Big data augmentation with data ware–house: A survey. In *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, 10 –13 December, 2785 –2794. doi: https://doi.org/10.1109/BigData.2018.8622206

Amghar, S., Cherdal, S., & Mouline, S.(2019). Data Integration and NoSQL Systems: A State of the Art. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, New York, October 23–24, 1–6. doi: https://doi.org/10.1145/3372938.3372954

Bai, Y., & Bhalla S.(2020). Introduction to Databases. In Practical database *programming with Visual Basic*.NET, John Wiley & Sons. Hoboken.

Bello–Orgaz, G., Jung, J. J., & Camacho, D.(2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59. doi: https://doi.org/10.1016/j.inffus.2015.08.005

Bianco, V., Manca, O., & Nardini, S.(2009). Electricity consumption forecasting in Italy using linear regression models. *Energy, 34*(9), 1413–1421. doi: https://doi.org/10.1016/j.energy.2009.06.034

Bontempo, C., & Zagelow, G.(1998). The IBM data warehouse architecture. *Communications of the ACM, 41*(9), 38–48. doi: https://doi.org/10.1145/285070.285078

Cao, L.(2017). Data science: a comprehensive overview. *ACM Computing Sur–veys (CSUR), 50* (3), 1–42. doi: https://doi.org/10.1145/3076253

Chandra, S., Ray, S., & Goswami, R. T.(2017, January). Big data security in healthcare: survey on frameworks and algorithms. In *2017 IEEE 7th International Advance Computing Conference(IACC)*(pp. 89–94). IEEE. doi: https://doi.org/10.1109 / IACC.2017.0033

Chen, G., Long, T., Xiong, J., & Bai, Y.(2017). Multiple random forests modelling for urban water consumption forecasting. *Water Resources Management, 31* (15), 4715–4729. doi: https://doi.org/10.1007/s11269–017–1774–7

Chung, D. J., Steenburgh, T., & Sudhir, K.(2014). Do bonuses enhance sales productivity? A dynamic structural analysis of bonus –based compensation plans. *Marketing Science, 33* (2), 165 –187. doi: https://doi.org/10.1287/mksc.2013.0815

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/10.1111/j.1751–5823.2001.tb00477.x

CMMI.(2019). Data Management Maturity (DMM). Retrieved from https://cmmiinstitute.com/data–management–maturity

Dadheech, P., Goyal, D., & Srivastava, S.(2019). Information Management and Ma–chine Intelligence. *Proceedings of International Conference on Information Management& Machine Intelligence. Jaipur*, 14–15 December, 85–100. doi: https://doi.org/10.1007/978–981–15–4936–6_9

Dahbura, A. T., & Masson, G. M.(1984). An 0(n2. 5) fault identification algorithm for diagnosable systems. *IEEE Computer Architecture Letters,33* (06),486–492.doi:https://doi.org/486–492.10.1109/TC.1984.1676472

Das, S.(2021). *Data Science: Theories, Models, Algorithms, and Analytics*. Srdas.github.io. Retrieved 1 March 2021, from https://srdas.github.io/MLBook/index.html

Davenport, T. H., &Patil, D. J.(2012). Data scientist. *Harvard Business Review, 90*(5), 70–76. doi: https://doi.org/10.1007/s11213–012–9233–0

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

DeBold, T., & Friedman, D.(2015). *Battling Infectious Diseases in the 20th Century: The Impact of Vaccines*. WSJ. Retrieved 1 March 2021, from http://graphics.wsj.com/infectious–diseases–and–vaccines/

Dhar, V.(2013). Data science and prediction. *Communications of the ACM, 56* (12), 64–73. doi: https://doi.org/10.1145/2500499

Doshi–Velez, F., & Kim, B.(2017).Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Dwork, C.(2008, April). Differential privacy: A survey of results. In *Interna–tional conference on theory and applications of models of computation* (pp. 1–19). Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978–3–540–79228–4_1

Earley, S.(2014). Agile analytics in the age of big data. *IT Professional,16*(4), 18–20. doi: https://doi.org/10.1109/MITP.2014.44

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

Furche, T., Gottlob, G., Libkin, L., Orsi, G., & Paton, N. W.(2016). Data Wrangling for Big Data: Challenges and Opportunities. *19th International Conference on Extending Database Technology (EDBT)*. Bordeaux, 15–18 March, 473–478. doi: https://doi.org/10.5441/002/edbt.2016.44

Gao, J., Xie, C., & Tao, C.(2016). Big Data Validation and Quality As–sur–ance—Issuses, Challenges, and Needs. *2016 IEEE symposium on ser–vice–oriented system engineering (SOSE)*, Oxford, March 29–April 2, 433–441. doi: https://doi.org/10.1109/SOSE.2016.63

Ghojogh B., & Crowley M.(2019) Instance Ranking and Numerosity Reduction Using Matrix Decomposition and Subspace Learning. In *Canadian Conference on Artificial Intelligence*, Kingston, ON, 28–31 May, 160–172. doi: https://doi.org/10.1007/978–3–030–18305–9_13

Gray, J., Chambers, L., & Bounegru, L.(2012). *The data journalism handbook: How journalists can use data to improve the news*. O´Reilly Media, Inc.

Grothaus, M.(2018). *How our data got hacked, scandalized, and abused in 2018*. Fast Company. Retrieved 1 March 2021, from https://www.fastcompany.com/90272858/how–our–data–got–hacked–scandalized–and–abused–in–2018.

Gurrin, C., Smeaton, A. F., & Doherty, A. R.(2014). Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval, 8* (1), 1–125. doi: http://dx.doi.org/10.1561/1500000033

Han, J., Pei, J., & Kamber, M.(2011). *Data mining: concepts and techniques*. Elsevier, Morgan Kaufmann, Waltham.

Hollowell, J. C., Kollar, B., Vrbka, J., & Kovalova, E.(2019). Cognitive decision–making algorithms for sustainable manufacturing processes in Industry 4.0: Networked, smart, and responsive devices. *Economics, Management and Financial Markets, 14* (4), 9–15. doi: https://doi.org/10.22381/EMFM14420191

Hou, J. K., Chang, M., Nguyen, T., Kramer, J. R., Richardson, P., Sansgiry, S., ...& El–Serag, H. B.(2013). Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Digestive dis–eases and sciences, 58*(4), 936–941. doi: https://doi.org/10.1007/s10620–012–2433–8

John, T., & Misra, P.(2017). *Data lake for enterprises*. Packt Publishing Ltd, Birmingham.

Jurney, R. (2017) . *Agile data science 2.0: Building full–stack data analytics ap–plications with spark*. O´Reilly Media, Inc.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W.(2013, January). Big data: Issues and challenges moving forward. In *2013 46th Hawaii international con–ference on system sciences* (pp. 995–1004). IEEE. doi: https:

//doi.org/10.1109/HICSS.2013.645

Kalegele, K., Takahashi, H., Sveholm, J., Sasai, K., Kitagata, G., & Kinoshita, T.(2013). Numerosity reduction for resource constrained learning. *Journal of Information Processing, 21* (2), 329–341. doi: https://doi.org/10.2197/ipsjjip.21.329

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ...& Buono, P.(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

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ...& Kumar, V.(2017). Theory–guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering, 29* (10), 2318–2331. doi: https://doi.org/10.1109/TKDE.2017.2720168

Kaserman, D. L.(2007). Efficient Durable Good Pricing And Aftermarket Tie - In Sales. *Economic Inquiry, 45*(3), 533–537. doi: https://doi.org/10.1111/j.1465–7295.2007.00022.x

Keim, D., Qu, H., & Ma, K. L.(2013). Big–data visualization. *IEEE Computer Graphics and Applications, 33*(4), 20–21. doi: https://doi.org/10.1109/MCG.2013.54

Khine, P. P., & Wang, Z. S.(2018). Data lake: a new ideology in big data era. *4th Annual International Conference on Wireless Communication and Sensor Net–work (WCSN 2017)*, Wuhan, 15–17 December, 2017, 17. doi: https://doi.org/10.1051/itmconf/20181703025

Kitchin, R.(2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography, 3*(3), 262–267. doi: http://www.wired.com/science/discoveries/magazine/16–07/pb_theory

Koehler, M., Bogatu, A., Civili, C., Konstantinou, N., Abel, E., Fernandes, A. A., ...& Paton, N. W.(2017). Data context informed data wrangling. In *2017 IEEE Inter–national Conference on Big Data. Boston*,11–14 December, 956–963. doi: https://doi.org/10.1109/BigData.2017.8258015

Konkel, L. (2020, March). *Who Will Benefit From Precision Medicine?. Who Will Bene–fit from Precision Medicine?* UC San Francisco. Retrieved from https://www.ucsf.edu/magazine/benefit–precision–medicine

Kotval, X. P., & Burns, M. J.(2013). Visualization of entities within social media: Toward understanding users´ needs. *Bell Labs Technical Journal, 17*(4), 77–102. doi: https://doi.org/10.1002/bltj.21576

Kruzhilko, O., & Maystrenko, V.(2019). Management decision–making algorithm development for planning activities that reduce the production risk level. *Journal of Achievements in Materials and Manufacturing Engineering, 93* (1–2). doi: https://doi.org/10.5604/01.3001.0013.4141

Kuacharoen, P.(2014). Combination of data masking and data encryption for cloud database. *Applied Mechanics and Materials*, 571–572, 617–620. doi: https://doi.org/10.4028/www.scientific.net/amm.571–572.617

Kune, R., Konugurthi, P., Agarwal, A., Chillarige, R., & Buyya, R.(2015). The anatomy of big data computing. *Software: Practice and Experience, 46* (1), 79–105. doi: https://doi.org/10.1002/spe.2374

Lechtenbörger, J., & Vossen, G.(2003). Multidimensional normal forms for data warehouse design. *Information Systems, 28* (5), 415–434. doi: https://doi.org/10.1016/S0306–4379(02)00024–8

Leigh, T. W., & Marshall, G. W.(2001). Research priorities in sales strategy and performance. *Journal of Personal Selling & Sales Management, 21*(2), 83–93. doi: https://doi.org/10.2307/20832582

Lenzerini, M.(2002). Data integration: a theoretical perspective. *Proceedings of the twenty–first ACM SIGMOD–SIGACT–SIGART symposium on Principles of data–base systems*. New York, 3–5 June, 233–246. doi: https://doi.org/10.1145/543613.543644

Lewis, S. C.(2015). Journalism in an era of big data: Cases, concepts, and critiques. *Digital Journalism, 3*(3), 321–330. doi: https://doi.org/10.1080/21670811.2014.976399

Lin, X., Wang, P., & Wu, B.(2013). Log analysis in cloud computing environment with Hadoop and Spark. *5th IEEE International Conference on Broad–band Network & Multimedia Technology, Guilin*, 17–19 November, 273–276. doi: https://doi.org/10.1109/ICBNMT.2013.6823956.

Lord, N.(2020, March). *Top 10 Biggest Healthcare Data Breaches of All Time*. Digital Guardian. Retrieved from https://digitalguardian.com/blog/top–10–biggest–healthcare–data–breaches–all–time

Lu, R., Wu, G., Xie, B., & Hu, J.(2014). Stream bench: Towards benchmarking modern distributed stream computing frameworks. In *20z14 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, London, 8–11 December, 69–78, doi: https://doi.org/10.1109/UCC.2014.15

Ma, C., Zhang, H. H., & Wang, X.(2014). Machine learning for Big Data analytics in plants. *Trends in plant science, 19* (12), 798–808. doi: https://doi.org/10.1016/j.tplants.2014.08.004

Madera, C., & Laurent, A.(2016, November). The next information architecture evolution: the data lake wave. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, New York,1 –4 November, 174–180. doi: https://doi.org/10.1145/3012071.3012077

Mallach, E.(2000). *Decision support and data warehouse systems*. Irwin/McGraw–Hill.

Mansfield–Devine, S.(2014). Masking sensitive data. *Network Security, 10*, 17–20. doi: https://doi.org/10.1016/S1353–4858(14)70104–7

Marx, V.(2013). The big challenges of big data. *Nature, 498* (7453), 255–260. doi: https://doi.org/10.1038/498255a

Matthews, K.(2019). *AI in Data Journalism: Pros and Cons*. Datafloq.com. Retrieved 1 March 2021, from https://datafloq.com/read/ai–data–journalism–pros–cons/7116.

Mattmann, C. A.(2013). A vision for data science. *Nature, 493* (7433), 473–475. doi: https://doi.org/10.1038/493473a

Maurer, C., & Wiegmann, R.(2011). Effectiveness of Advertising on Social Network Sites: A Case Study on Facebook. In: Law R., Fuchs M., Ricci F.(eds) *Information and Communication Technologies in Tourism 2011*. Springer, Vienna.

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

Meredith, R., O´Donnell, P., & Arnott, D.(2008). Databases and data ware–houses for decision support. In *Handbook on Decision Support Systems 1*(pp. 207–230). Springer, Berlin, Heidelberg.

Meyer, P.(2002). *Precision journalism: A reporter´s introduction to social science methods*. Rowman& Littlefield Publishers.

Miles, M. B., & Huberman, A. M.(1994). *Qualitative data analysis: An expanded sourcebook*. SAGE Publications.

Miloslavskaya, N., & Tolstoy, A.(2016). Application of big data, fast data, and data lake concepts to information security issues. In *2016 IEEE 4th Internation–al Conference on Future Internet of Things and Cloud Workshops(FiCloudW)*, Vienna, 22–24 August, 148–153. doi: https://doi.org/10.1109/W–FiCloud.2016.41

Ministry of Industry and Information Technology of the People´s Republic of China.(2020, May). *Guiding opinions of the development of industrial big data from the Ministry of Industry and Information Technology*. Retrieved from https://www.miit.gov.cn/xwdt/gxdt/sjdt/art/2020/art_a61849ebec144ebdb91fa9bc5474554c.html

Mooney, S. J., & Pejaver, V.(2018). Big data in public health: terminology, machine learning, and privacy. *Annual Review of Public Health, 39*, 95–112. doi: https://doi.org/10.1146/annurev–publhealth–040617–014208

Mounia, B., & Habiba, C.(2015). Big data privacy in healthcare Moroccan context. *Procedia Computer Science, 63*, 575–580. doi: https://doi.org/10.1016/j.procs.2015.08.387

Müller, H., & Freytag, J. C.(2005). *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. FürInformatik.

Myers, R.(2019). D*ata Management and Statistical Analysis Techniques.* Scientific e–Resources. Waltham.

Nagowah, S. D., Sta, H. B., & Gobin–Rahimbux, B. A.(2019). Towards Achieving Semantic Interoperability in an IoT–enabled Smart Campus. In *2019 IEEE International Smart Cities Conference (ISC2)*, Casablanca, Morocco, 14–17 October, 593–598. doi: https://doi.org/10.1109/ISC246665.2019.9071694

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.

Nguyen, D.(2016, March). *Computational Journalism at Stanford University | Computational Journalism, Spring 2016.* Retrieved from http://www.compjour.org/

Overton, J.(2016). *Going Pro in Data Science: What it Takes to Succeed as a Professional Data Scientist*. O´Reilly Media.

Parasie, S., & Dagiral, E.(2013). Data–driven journalism and the public good:"Computer–assisted–reporters" and

"programmer–journalists" in Chicago. *New Media & Society, 15* (6), 853–871.doi: https://doi.org/10.1177/1461444812463345

Patil D. J.(2012). *Data jujitsu: The art of turning data into data product.* O´Reilly Media, Inc., Sebastopol.

Prasad, B. R., & Agarwal, S.(2016). Comparative Study of Big Data Computing and Storage Tools. *International Journal of Database Theory and Application, 9*(1), 45–66. doi: https://doi.org/10.14257/ijdta.2016.9.1.05

Prashar, S., Vijay, T. S., & Parsad, C.(2016). Predicting online buying behavior among Indian shoppers using a neural network technique. *International Journal of Business and Information, 11* (2), 175. doi: https://doi.org/10.6702/ijbi.2016.11.2.3

Putatunda, S., Rama, K., Ubrangala, D., & Kondapalli, R.(2019). *SmartEDA: An R Package for Automated Exploratory Data Analysis*. arXiv preprint arXiv:1903.04754.

Raghupathi, W., & Raghupathi, V.(2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2* (1), 3. doi: https://doi.org/10.1186/2047–2501–2–3

Raptis, T. P., Passarella, A., & Conti, M.(2019). Data management in industry 4.0: State of the art and open challenges. *IEEE Access, 7*, 97052–97093. doi: https://doi.org/10.1109/ACCESS.2019.2929296

Russom, P.(2011). Big data analytics. *TDWI best practices report, fourth quarter, 19* (4), 1–34.

Sarada, G., Abitha, N., Manikandan, G., &Sairam, N.(2015). A few new approaches for data masking. In *2015 International Conference on Circuits, Power and Computing Technologies*, Nagercoil, 19–20 March, 1–4. doi: https://doi.org/10.1109/ICCPCT.2015.7159301

Schneider, C.(2016). *The biggest data challenges that you might not even know you have–Watson Blog. Watson Blog.* Retrieved 1 March 2021, from https://www.ibm.com/blogs/watson/2016/05/biggest –data –challenges–might–not–even–know

Shahrivari, S.(2014). Beyond batch processing: towards real–time and streaming big data. *Computers, 3*(4), 117–129. doi: https://doi.org/10.3390/computers3040117

Sigmoidal.(2017). *Recommendation Systems – How Companies are Making Money – Sigmoidal*. Sigmoidal. Retrieved 1 March 2021, from https://sigmoidal.io/recommender–systems–recommendation–engine

Sivarajah, U., Kamal, M. M., Irani, Z., &Weerakkody, V.(2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263 –286. doi: https://doi.org/10.1016/j.jbusres.2016.08.001

Southwick, S. B., Lampert, C. K., & Southwick, R.(2015). Preparing controlled vocabularies for linked data: benefits and challenges. *Journal of Library Metadata, 15* (3 –4), 177 –190. doi: https://doi.org/10.1080/19386389.2015.1099983

Stanford University.(2021, March). *HIV drug resistance database*. Retrieved from https://hivdb.stanford.edu.

Sun, D., Zhang, G., Zheng, W., & Li, K.(2015). Key Technologies for Big Data Stream Computing. In *Big Data: Algorithms, Analytics, and Applications,* CRC Press, Boca Raton.

Sweeney, L.(2002). k–anonymity: A model for protecting privacy. International *Journal of Uncertainty, Fuzziness and Knowledge–Based Systems, 10*(05), 557–570. doi: https://doi.org/10.1142/S0218488502001648

Taleb, I., Serhani, M. A., & Dssouli, R.(2018, July). Big data quality: A survey. In *2018 IEEE International Congress on Big Data(BigData Congress)* (pp. 166–173). IEEE.doi: https://doi.org/10.1109/BigDataCongress.2018.00029

Tansley, S., & Tolle, K.(2009). *The fourth paradigm: data–intensive scientific discovery (Vol. 1)*. T. Hey(Ed.). Redmond, WA: Microsoft research.

Tee, J.(2013). *Four ´V´s of big data: volume velocity variety veracity*. TheServ–erSide.com. Retrieved 1 March 2021, from https://www.theserverside.com/feature/Handling–the–four–Vs–of–big–data–volume–velocity–variety–and–veracity

The AlphaFold team.(2020). AlphaFold: a solution to a 50–year–old grand challenge in biology. Deepmind. Retrieved 1 March 2021, from https://deepmind.com/blog/article/alphafold–a–solution–to–a–50–year–old–grand–challenge–in–biology

The Four V´s of Big Data. *IBM Big Data & Analytics Hub.*(2021, March). Retrieved from http://www.ibmbigdatahub.com/infographic/four–vs–big–data

Timetoast.(2021, March). *The history of data journalism timeline*. Retrieved from https://www.timetoast.com/time-

lines/the–history–of–data–journalism

Tukey, J. W.(1977). *Exploratory Data Analysis*. Pearson.

Tukey, J. W., & Wilk, M. B.(1966). Data analysis and statistics: an expository overview. In *Proceedings of the November 7–10, 1966, fall joint computer conference*, New York, 7–10 November, pp. 695–709. doi: https://doi.org/10.1145/1464291.1464366

Underwood, C.(2019). *Automated Journalism – AI Applications at New York Times, Reuters, and Other Media Giants | Emerj.* Emerj. Retrieved 1 March 2021, from https://emerj.com/ai–sector–overviews/automated–journalism–applications

Vaisman, A, & Esteban Z.(2014). *Data Warehouse Concepts in Data Ware–house Systems* (pp.75). Springer, Berlin Heidelberg.

Velicanu, M., & Matei, G.(2007). Database versus Data Warehouse. *Revista Informatica Economică*, 91–95.

Warners, H. L. H. S., & Randriatoamanana, R.(2016). Datawarehouser: A Data Warehouse artist who have ability to understand data warehouse schema pictures. In *2016 IEEE Region 10 Conference (TENCON),* Singapore, 22–25 November, 2205–2208. doi: https://doi.org/10.1109/TENCON.2016.7848419

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

Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J.(2019). Prediction of fatty liver disease using machine learning algorithms.*Computer Methods and Programs in Biomedicine, 170*, 23–29. doi: https://doi.org/10.1016/j.cmpb.2018.12.032

Yandun, F., Silwal, A., & Kantor, G.(2020). Visual 3D Reconstruction and Dynamic Simulation of Fruit Trees for Robotic Manipulation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi: https://doi.org/10.1109/cvprw50498.2020.00035

Zhang, X., Wang, S., Cong, G., & Cuzzocrea, A.(2019). Social Big Data: Mining, Applications, and Beyond. *Hindaw, 12*, 3, 1749–1772. https://doi.org/10.1155/2019/2059075

Zhang, Y., & Pennacchiotti, M.(2013, May). Predicting purchase behaviors from social media. In *Proceedings of the 22nd international conference on World Wide Web* (pp.1521 –1532). doi: https://doi.org/10.1145/2488388.2488521

Zhao, Y., Yao, L., & Zhang, Y.(2016). Purchase prediction using Tmall - specific features. *Concurrency and Computation: Practice and Experience, 28* (14), 3879–3894. doi: https://doi.org/10.1002/cpe.3720

Zhou, K., Fu, C., & Yang, S.(2016). Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews, 56*, 215–225. doi: https://doi.org/10.1016/j.rser.2015.11.050

Zialcita, P.(2019, October). Facebook pays $643,000 fine for role in Cambridge Analytica Scandal. Retrieved from https://www.npr.org/2019/10/30/774749376/facebook–pays–643–000–fine–for–role–in–cambridge–analytica–scandal