Chapter 1
Big Data and Big Data Analytics

Big data now is a common term. However, the evolution of big data comes from twofold. The creation of the computer in the 1940s gradually provides tools for human beings to collect massive data, while the term "big data" becomes a popular slogan to represent the collection, processing, and analysis of various data [1]. The data has been exponentially growing for the last 70 decades. EMC2 [2] estimated that the world generated 1.8 zettabytes of data (1.8 multiple 21 zeros) by 2011. In fact, this figure has grown to 44 zettabytes, about 24 times in 2020. Big Data Analytics has arisen as the technical means dealing with both theory and application of big data. This chapter elaborates on the understanding of big data and its analytics. Section 1.1 briefly describes big data evolution and challenges. Section 1.2 is about big data's current status, including its development in the world as well as in China. Section 1.3 explores big data analysis and data science problems.

1.1 Big Data Evolution and Challenges

Nowadays, in human society, big data is the environment that we cannot ignore in our daily activities. Big data occurs as a phenomenon. Whenever we make a decision, the impact of big data must be considered. Big data is a "buzz" word that is a better capture about the name of data collection and analysis which went through the stages of database management in 1960s, data warehouse in 1970s, knowledge discovery in databases (KDD) in 1980s, enterprise resource planning (ERP) and data mining in 1990s, customer relationship management (CRM) and business analytics (BA) in 2000s. Big data, as a good term, unifies all of the above concepts so that the majority of people know what it means. Big data is also a big challenge for those people who analyze the data due to its complex structure and lack of available technology. Furthermore, big data provides a big opportunity to the business world for increasing productivity [3].

Y. Shi, *Advances in Big Data Analytics*,
https://doi.org/10.1007/978-981-16-3607-3_1

The common concept of big data contains the applications, engineering, and scientific issues of big data. The definition of big data varies from academic and business communities and there is no unified definition about big data. In some professional communities, the terms of business intelligence and business analytics are also used to represent big data analytics [4].

In 2012, the National Science Foundation of USA [5] provided the following definition:

Definition 1.1 Big data is "large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, internet transactions, email, video, click streams, and/or all other digital sources available today and in the future".

In May 2013, a group of international scholars has brain-stormed two versions of big data definitions at the 462nd Session: Data Science and Big Data in Xiangshan Science Conferences [6] at Beijing, China, where the author served as one of co-chairs. The first version of big data was given for academic and business communities as:

Definition 1.2 Big data is a collection of data with complexity, diversity, heterogeneity and high potential value, which are difficult to process and analyze in reasonable time.

The second version is for organizations and governmental policy making as:

Definition 1.3 Big data is a new type of strategic resource in digital era and the key factor to drive innovation, which is changing the way of human being's current production and living.

In addition, "4V's" have been commonly used to capture the main characteristics of big data: Volume, Velocity, Variety and Veracity [7, 8].

The history of data analytics can be traced back to more than 200 years ago when people used statistics to solve real-life problems. In the area of statistics, Bayes' Theorem has been playing a key role in developing probability theory and statistical applications. However, it was Richard Price (1723–1791), a famous statistician, edited Bayes' Theorem after Thomas Bayes' death [9]. Richard Price was also one of the scientists who initiated the use of statistics in analyzing social and economic datasets. In 1783, Price published "Northampton table", which collected observations for calculating of the probability of the duration of human life in England. In this work, Price showed the observations via tables with rows for records and columns for attributes as the basis of statistical analysis. Such tables now are commonly used in data mining as multi-dimensional tables. Therefore, from the historical point of view, the multi-dimensional table should be called as "Richard Price Table" and Price should be honored as the father of data analytics, later called data mining. Since the 1950s, as computing technology has been gradually used in commercial applications, many corporations have developed databases to store and analyze collected datasets. Mathematical tools employed to handle datasets revolute from statistics to methods of artificial intelligence, including neural networks and decision trees. In the 1990s, the database community started using the

term "data mining", which is interchangeable with the term "knowledge discovery in databases (KDD)" [10]. Now data mining becomes the common technology of data analytics over the intersection of human intervention, machine learning, mathematical modeling, and databases.

In recent years, many authors published their opinions about how big data is deeply impacting the evolution of science and engineering as well as the development of society. One of the popular books, written by Mayer-Schönberger and Cukier [11] showed three advantages of big data: (1) access to all data undermines the sampling; (2) rough measurements for big data replace the requirement of high-quality data preparation; and (3) decision making is based on correlations of big data, instead of the reasoning. Although such advantages represent the importance of big data analytics, they cannot change the fundamentals of data analytics, which are still sampling, accuracy and reasoning (see Sect. 1.3). Big data does not mean the entire data. It is impossible to collect the entire data, which is a relative concept. However, big data, comparing with small data, can provide a very large sample. The larger the sample is, the more robust the results are. Big data may lead a better learning result, but the sampling process is needed to test and predict. With a rough data preparation, big data may produce a quick response or rough knowledge for people to make decision. However, such a decision could be good for a short run, not for a long run. The long run decision requires the solid and high-quality data preparation. In addition, for decision makers, seeking the reasoning of big data is more important than finding the correlations. Using of big data in engineering practice or business actions is not only for what we can do, but more on what we should do for the future. Therefore, big data needs data mining techniques to discover knowledge and predict the future. Based on known data mining methods, big data analytics should consider the large sample from all available data (structure or non-structured data); look for the precise solution based on the rough solutions; and identify the reasoning from the correlations. Big data analysis does not remove the fundamentals of data analysis or data mining. Instead, it improves the analytic methodologies since all data are supposedly are available.

Among many challenges of the big data problems, it is believed that the following three problems are urgent to solve in order to gain benefits from big data in science, engineering and business applications:

Challenge 1.1 Transforming Semi-structured and Non-structured Data into "Structured Format"

In the academic field of big data, it is not clear about the principle, basic rules and properties of data, especially semi-structured and non-structured data due to complexity of such data. The data complexity reflects not only the variety of the objects that data represents, but also a partial image that each dataset can present for a given object. The relationship of data representation and a real object just likes that of "the blind men and an elephant" [12]. Even though each data set truly represents an angle of the object, it cannot be its whole picture. The investigation of theoretical components of big data, which can be viewed as "data science" deserves the interdisciplinary efforts from mathematics, sociology,

economics, computational science and management science. However, the term of data science is still under discussion among the research communities. Thanks to the advancement of information technology in recent years, the techniques, such Hadoop and MapReduce allow us to collect a large amount of semi-structured and non-structured data in a reasonable amount of time. Now, the key engineering challenge is how to effectively analyze these data and discover knowledge from them in an expected time. The answer could be that first transform the given semi-structured and/or non-structured data into a structured data-like format (or pseudo multi-dimensional table), and then conduct a data mining process by taking advantage of the existing data mining algorithms that are mainly developed for the structured data. Note that, the transformation from semi-structured and non-structured data into structured format should be subject-oriented. Once the structured data-like format is built up, the "first-order mining" by using data mining tools can result "rough knowledge" (called hidden patterns in data mining). To upgrade such knowledge into the "intelligent knowledge" that can be used for decision support, the analysts should combine some sort of human knowledge, such as experience, common sense, domain preference with the rough knowledge. This is viewed as the "second-order mining" [13].

Since most big data is based on semi-structured and/or non-structured representations, the "structured rough knowledge" from big data may reflect new properties, which can be captured by decision makers when it is upgraded as intelligent knowledge. The key value of big data analytics or data mining is to obtain intelligent knowledge.

Challenge 1.2 Exploring the Complexity, Uncertainty and Systematic Modeling of Big Data

As mentioned as in the above, any data representation of a given object is its partial picture of the facts. The complexity of big data comes from the coherent of data representation while the uncertainty of big data causes from the changes of the objects in the nature as well as the variety of data representations duo to measurements. Although a certain data analytic method is applied on big data, the knowledge discovered from the analysis is just knowledge from that particular angle of the real object. Once the angle is changed by the way of collecting or viewing the data from the object, the knowledge is no longer to be useful. For example, in petroleum exploration engineering, which can be viewed as a big data problem, the data mining has been done on spatial database generated from seismic tests and well log data collection. The underground geological structure itself is complicated. The non-linear patterns of data are changeable via different dimension and angles. Any results of data mining or data analysis could be knowledge that is only true for the given surface. If the surface is changed, the result is changes as well. Therefore, how can one derive the intelligent knowledge from knowledge from a surface and knowledge from a surface turning around 90° is challengeable [14]. The breakthrough to a systematic modeling on complexity and uncertainty of big data analysis and mining is needed for gaining knowledge from big data. Form a long-run point of view, it could be not easy for us to establish a comprehensive mathematic

system design about big data as a whole. However, through the understanding of particular complexity or uncertainty in given subjects or domain of fields, it is possible to build a domain-based systematic modeling for the specific big data. As long as a series of such modeling structures are founded, the collection of them can be viewed as a systematic modeling of the big data. From a short-run point of view, if the engineers can find out some general approaches to deal with complexity and uncertainty of big data in a certain field, say in financial market (with data stream and media news) or internet retails (images and media evaluations), it will bring added value to social and economic development. In addition, the formats of complexity and uncertainty of big data result in the measurement and evaluation on the rough knowledge from big data mining. Many known techniques in engineering, such as optimization, utility theorem, expectation analysis, can be used to measure how the rough knowledge gaining from big data should be better combined with human judgment into the "second-order mining" process to effectively elicit the intelligent knowledge for decision support. Note that since the knowledge changes with individual and situation, the machine-man (big data mining vs. human knowledge) is still playing a key role in big data modeling.

Challenge 1.3 Exploring the Relationship of Data Heterogeneity, Knowledge Heterogeneity and Decision Heterogeneity

At the big data environment, decision makers face three heterogeneous problems: data heterogeneity, knowledge heterogeneity and decision heterogeneity. Traditionally speaking, decision making depends on the learning of knowledge from others and accumulation of experience. Learning of knowledge now is more based on the data analysis and data mining. In a theory of management information system, decision making can be classified as three levels: structured decision, semi-structured decision and non-structured decision depending on the responsibilities of individuals in an organization [15]. The operational staffs handling routine works relate to structured decision. The managers' decision is based on subordinates' reports (almost of them are structured) and their own judgments and refers as semi-structured. The top-managers or chief executive officer (CEO) make a final decision is non-structured, which is most likely text or voice. The demand of decision makers for data or information (quantitative forms) and knowledge (qualitative forms) are different according to different levels of the responsibilities. However, big data is disruptively changing the decision-making process. Based on big data analysis or mining, the functions of business operation (structured decision), manager (semi-structured decision), and CEO (non-structured decision) can be combined as a whole picture for decision making. For instance, a marketing person may use a real-time credit card approval system based on big data mining technology to quickly approve a credit limit to a customer without reporting to a supervisor. Such a decision has almost zero risk. He or she is a final decision maker, representing both manager and CEO.

In a data mining process using structured data, the rough knowledge normally is structured knowledge due to its numerical formats. In big data mining, although rough knowledge in the "first-order mining" is derived from heterogeneous data,

it can be still reviewed as structure knowledge since the data mining is carried out on structured data-like format or pseudo multi-dimensional table. When the "second-order mining" is used, the structured knowledge is combined with domain knowledge of managers or CEO that are semi-structured or non-structured and gradually upgraded into intelligent knowledge [16]. Therefore, intelligent knowledge may be the representation of non-structured knowledge. Note that if the business operations only involve with semi-structured data and/or non-structured data, either it results in non-structured knowledge without data analysis (mining) or structured knowledge which is from data mining. Such structured or non-structured knowledge can impact semi-structured decision or non-structured decision depending on the levels of management involvements. Big data, nevertheless, creates a challenge to traditional decision-making process. Research on how the impact of big data on decision making is complicate and perhaps philosophy oriented. An observation is that no matter which kind of data heterogeneity is presented by big data, rough knowledge is in the domain of "first-order mining" and searching intelligent knowledge by the "second-order mining" is a key to study the relationship between data heterogeneity, knowledge heterogeneity and decision heterogeneity. Exploring how decision making can be changed in big data environment is equivalent to investigating the relationships of processing heterogeneous data, big data mining, domain knowledge of decision makers and involvement in decision making.

It can be predicted that any of theoretical contribution and engineering technologic breakthrough on the above three challenges can enhance the applications of big data in our society. It will start from the field of information technology, and then widely spreads to multi-media, finance, insurance, education, etc. for the formulation of new business models, boosting investment, driving the consumption, improving production, increasing productivity. In a word, it generates the big data revolution.

1.2 Big Data Development

It is not easy to describe how big data deeply and quickly influences the world. However, four big data events in academic community should be first mentioned. They are the big data associations, big data conferences, big data journals and big data sources opened by governments.

1.2.1 Big Data in Academic Community

Recent years, both academic and professional communities built various big data related non-profit organizations to exchange and disseminate theoretical findings, practical experience and case studies about big data as well as data science. Some of them are the National Consortium for Data Science (NCDS), the Big Data Institute

(BDI), Data Science Association (DSA), Institute for Big Data Analytics (IBDA), Institute for Data Science, Institute for Data Sciences & Engineering (IDSE), Data & Society Research Institute (DSRI), the Data Warehousing Institute (TDWI), Global Association for Research Methods and Data Science, ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), SNIA—Analytics and Big Data Committee (ABDC), Association of BIG DATA professionals (aBIGDATAp), The Big Data Alliance (BDA), Digital Analytics Association (DAA), and Data Science Consortium. It can be observed that, in some academic communities, the term "big data" means the information technology business applications of dealing with massive data problems while the scientific components or research aspects of big data is called data science. This is why data science somehow is interchangeable with big data. In some professional communities, terms "business intelligent and business analytics" are used to describe big data analysis or big data mining [4].

In 2013–2020, for example, numerous big data conferences have been held around the world. Some of them are International Conference on Big Data and Cloud Computing, IEEE International Conference on Big Data, ISC Big Data Conference, Big Data Technology Conference, CCF Big Data, IEEE International Congress on Big Data, the series of Big Data Conferences (Stanford, Beijing and Cambridge), International Conference on Algorithms for Big Data, International Conference on Data Science (ICDS, which was founded by the author and his institutions), the Conference on Nonparametric Statistics for Big Data and INFORMS Conference on the Business of Big Data. Most of the above conferences have been held annually since 2014. And these conferences have attracted thousands of scholars, engineers and practitioners for their common interests in big data problems.

There two categories of big data related academic journals. One is under the name of big data and another is under names of data science. The big data journals are Journal of Big Data, Big Data Research, International Journal of Big Data Intelligence, International Journal of Big Data, Big Data Journal and Big Data & Society Journal. The data science journals are Annals of Data Science, The Data Science Journal, Journal of Data Science, EPJ Data Science, Data Science Journal, and International Journal of Data Science. Most of these journals are newly established in recent years and need to demonstrate the reputations by publish cutting-edge research findings and technological advances in big data related areas.

1.2.2 Big Data in the World

It has been recognized that most data of "big data" come from three sources: The large amount of data generated from social and economic activities are controlled by governments. The enterprises, especially the well-known big data companies such as Google in U.S., Baidu in China and Yandex in Russia, own their business data as the important assets. The rest of open data accesses from online are free for anyone to download or use. Therefore, governments play a key role in making

policies and promoting big data applications. Governmental actions on big data can be categorized as two stages.

1.2.2.1 Stage 1 (2009–2012): Start-Up

In 2009 the U.S. government launched its Data.gov website to offer governmental datasets to the public, and later in 2011 the Open Government Partnership (OGP) initiated by a United Nation General Assembly meeting. The United States and the United Kingdom, both are founders of OGP, delivered their national action plans for the first time, in order to open the government data as their main priority and promise to accomplish this goal. In 2011 the McKinsey Global Institute (MGI) published a special report called "Big Data: The next frontier for innovation, competition, and productivity", which was for the first-time a thorough introduction and prospection of big data released from a distinguished institute. Around that time, big data was widely discussed among both academic circles and economic circles. The Obama Administration launched the "Big Data Research and Development Initiative" on the White House official website in 2012, demonstrating that big data technologies evolved from early commercial operations to national scientific and technological strategies. Meanwhile, as the ongoing development of Internet and mobile communication technologies are increasing exponentially, (including intelligent terminal devices, semi structures and unstructured data), mathematical tools used to process datasets were shifted from statistics to artificial intelligence. Hence, the Age of Big data began.

1.2.2.2 Stage 2 (2013–Today): High-Speed Developing

Along with matured big data fundamental technologies and techniques, academic and business domains steered to applications research accordingly.

Big data technologies started to infiltrate into all society sectors, such as government administration, finance, science and technology, health care, education, transportation, industry. These sectors formed a complete big data industrial chain and developed amounts of applications in diverse fields: smart government, smart city, intelligent manufacturing, new retailing, etc. And this is when Hereupon big data entered its high-speed development stage. In 2013 at the G8 Summit, eight G8 members signed an Open Data Charter, which established the basic principles and standards for members to improve transparency of government information [17]. It encourages the governments open their data to public on five principles: Open Data by Default, Quality and Quantity, Usable by All, Releasing Data for Improved Governance, and Releasing Data for Innovation. So far, a number of countries have set up their data.gov style websites. The released big data covers broad categories, including Agriculture, Climate & Weather, Infrastructure, Energy, Finance & Economy, Environment, Health, Crime & Justice, Government &

Policy, Law, Job & Employment, Public Safety & Security, Science & Technology, Education, Society & Culture, Tourism and Transportations [18].

In terms of the subject of "Open Data", Europe is at the forefront. In 2014, the European Commission adopted the "Towards a Thriving Data-driven Economy" strategy and advocated European countries to seize the opportunities in data economy development. In March 2018, news of Facebook data leakage scandal started spreading and it is now still heating up. In the Age of Big Data (data sharing and data safety), individual privacy balancing and protection became a worldwide problem. In 2016, General Data Protection Regulation (GDPR) was approved by the European Parliament and came into effect on May 25 2018. On April 25 the same year, the European Commission released the policy document "Towards a Common European Data Space", addressing principles on how public sectors open datasets, retain and collect research data, and how private companies are processing and opening data. On October 4 2018, the European Parliament voted through the Regulation on "The Free Flow of Non-personal Data". Henceforth, the European Union has built a systematical legal system for individuals' privacy protection, as well as data opening and sharing.

In 2019, the Tianfu Institute of International Big Data Strategy and Technology (TIBD, which was founded by the author and local government and institutions), Chengdu Government Service Management & Network Administration Office, the Research Center on Fictitious Economy and Data Science Chinese Academy of Science, and the Key Laboratory of Big Data Mining and Knowledge Management Chinese Academy of Sciences jointly released the first "Annual Big Data World Report", called "Global Big Data Development Analysis Report 2018" [19]. This report was based on the source data of 79 OGP-membership countries as well as that of China. It produced many interesting findings regarding how big data developed in the world. For example, the proportion of OGP membership countries by continents making commitments to open government data in November 2018 is distributed as the following: European countries account for 36.5%, African 19.2%, North American 13.5%, Asian 13.5%, Latin American 13.5%, and Oceania 3.8% (see Fig. 1.1).

Another example is in terms of correlation between per capita GDP growth and government's efforts in opening-up data in major countries, the former is proportional to the latter with an exception of India (see Fig. 1.2). Though the per capita GDP growth rate in India is comparatively low, India is also a pioneering country in government open data efforts.

In 2021, the second "Annual Big Data World Report" named "Global Big Data Development Analysis Report 2020" was released by the TIBD and these agencies mentioned above. This report introduces the COVID-19 pandemic and its acceleration of big data development. It also incorporates the current status of data opening efforts around the world with its promotional effect on digitalization and "High-quality Development".

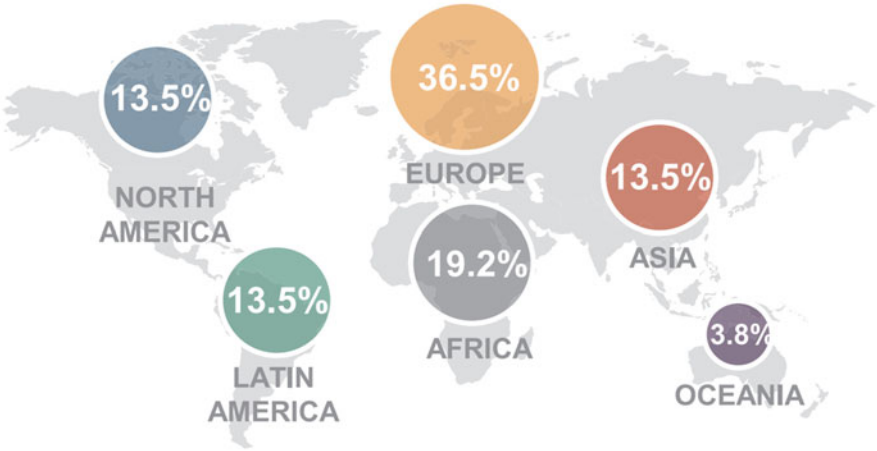

Fig. 1.1 OGP participating countries making commitments of open data (2018)

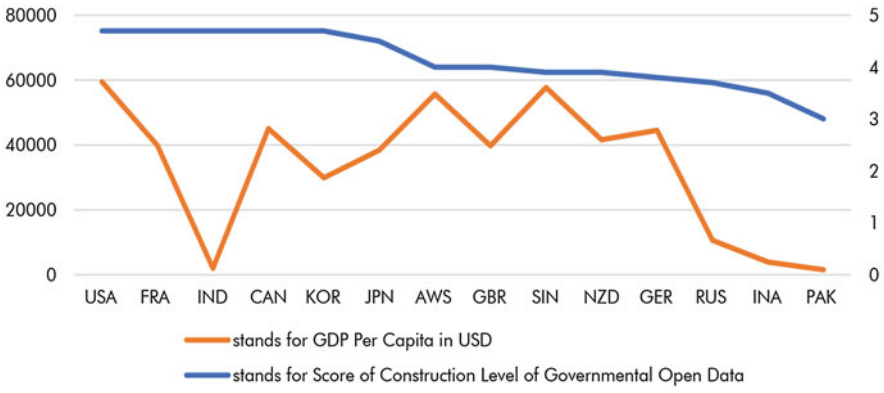

Fig. 1.2 The Relationship between Per Capita GDP and the score of construction level of Governmental Big Data Open in world's main countries (2018)

1.2.3 Big Data in China

In 2015, the State Council of the People's Republic of China issued the "Action Plan for Promoting the Development of Big Data" with the purpose of comprehensively promoting the development and application of big data in China and accelerating the construction of a powerful data nation [20]. Big data industry in China comes into flourish in all fields. Later, Big data becomes one of China's national strategies of economic development in it's the 13th Five-year plan (2016–2020).

The action framework can be interpreted as a Top-Level Design, which consists of three national platforms: National Data Opening Platform, Trans-Departmental Data Sharing Platform and Internet based National Data Service Platform (see Fig.

Fig. 1.3 China's big data top-down design

1.3). By 2020, the Chinese government commits to complete ten key big data projects for three platforms so as to provide big data applications in a number of public areas, including credit, transportation, healthcare, employment, social security, geography, culture, education, science and technology, agriculture, environment, safety and security, product quality, statistics, meteorology and ocean service. These projects will finally assemble a numbers of big data systems across a wide range of governmental departments, industries, academic and education institutes. After the announcement of the Action Framework, the Chinese government has published a planning program to further nail down key tasks for related departments to carry out its respective potions of Fig. 1.3 in terms of the responsibilities, road maps and target dates into 2020. This program offers a strong support for the completion of the Action Framework [21].

Importantly, the Chinese government tries to use the implementation of the framework to influence the Chinese culture and social value towards data and

builds social data awareness. In May 2013, a group of the Chinese and international scholars, including the author of this book, brain-stormed the meanings of big data and provided Definition 1.2 and Definition 1.3. These definitions had a positive influence on the Chinese leadership in the following events to build China's big data strategy. A Chinese version of big data can be regarded as the large-scale data set produced and being utilized from China's modern informatization process, the totality of data source in the current information society and the whole set of data, not only internet data, but also governmental and commercial data [22]. The framework calls for the entire Chinese society to have big data thinking and emphases social data awareness. The traditional Chinese decision making relies on qualitative thinking, not quantitative thinking. Such a cultural behavior has burdened nation's science and technology development. Thus, the Action Framework aims to change the current Chinese culture by enhancing data awareness and promoting data spirit.

Since the release of the Action Framework, the importance of big data has been highly recognized by leaders. An integrated national big data center was proposed as the country seeks to enhance governance capability. Not only has the national level paid attention to big data development, local governments also attach importance to big data increasingly. Up to June 2017, more than 40 provinces and cities issued nearly 100 big data development policies and big data industrial plans. An expert committee is composed of academicians, scholars from scientific research institutes, and representatives of industrial circles. The innovation alliance is made up of more than 70 related entities of the 14 Big Data National Engineering Laboratory. Through these two mechanisms, big data scientists and companies would be gathered contributing to policy making, technology consulting and technology transformation. In addition, a series of open data competitions will be hold in eight National Big Data Comprehensive Test Areas, to promote pubic data opening and encourage innovative applications.

1.3 Big Data Analytics

This subsection presents some fundamental scientific problems in big data analytics. Section 1.3.1 describes an overview of big data analytics based on multiple domains. They include the influences of Management Science on data acquisition and data management, Information Science on data access and processing, Mathematics and Statistics on data understanding, and Engineering on data applications. Section 1.3.2 outlines six open research problems in big data analytics.

1.3.1 Overview of Big Data Analytics

Although there are many different interpretations of big data, big data analytics and data science, one can view that big data is consist of both data science and applications, where big data analytics is an intersection of both. To distinguish these three concepts, the following definition of data science is used in the book:

Definition 1.4 Data Science is mathematical means and algorithm to extract knowledge from big data.

The definition above is very rough, not precise at this point due to the complexity of big data. The boundary of data science for a given filed can be change because the nature of big data in the field differs from others.

With Definitions 1.1–1.4, it can be viewed that if the common-known big data is represented as a set, then data science and application are two subsets while big data analytics is also a subset of big data, across both data science and application since big data analytics is used data science to deal with some specific application problems. A relationship of big data, big data analysis, data science and application can be shown as Fig. 1.4.

In general, the process of big data analytics, as a subset of big data can be shown in Fig. 1.5. It is consisting of several steps, including data acquisition and management, data access and processing, data mining and interpretation, and data applications [23]. However, due to the "4Vs" characteristics of big data, the activities of each step in the process also face Challenges 1.1–1.3. The techniques of multidisciplinary fields need to apply in addressing such challenges.

For Challenge 1.1, majority of big data are represented as semi-structured and non-structured formats. Even though the technologies of MapReduce (Hadoop) can be used to acquire big data, the traditional data acquisition and management of Computer Science should be reinforced by the knowledge of Management Science. For example, the organizational strategy of using big data must be considered before performing the big data acquisition. The basic design of big data base and management should be built up in terms of data capabilities, value, ethic, ownership, policy, quality assurance etc. [15] With help of Management Science, big data can play as an important role for us to make effective decision.

Fig. 1.4 Relationship of big data, big data analytics, data science and application

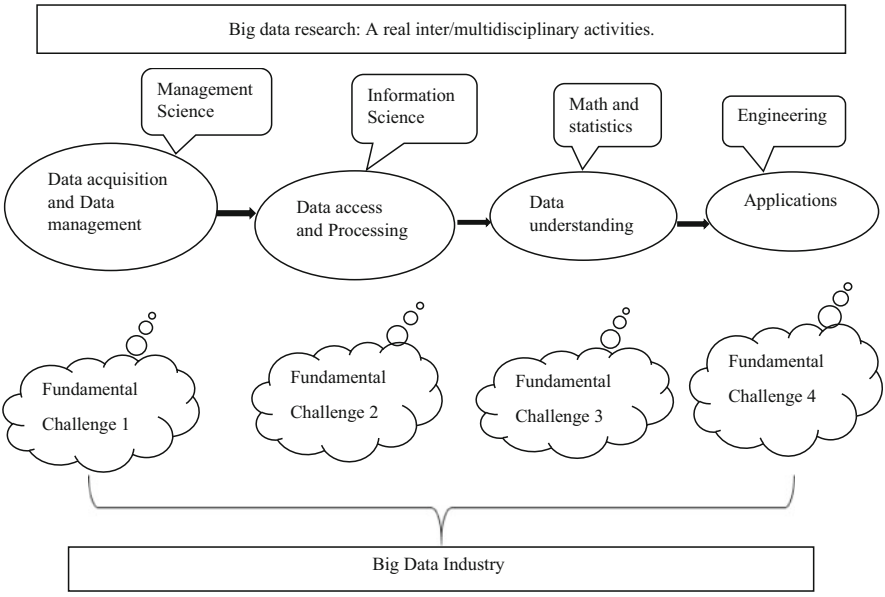

Fig. 1.5 An overview of big data analytics

For Challenge 1.2, the complex formats and features of big data lead the difficulty of assessing, especially processing the data for data mining and interpretation. Many existing techniques of Information Science are ready to respond this challenge. Since most data mining or machine leaning algorithms are constructed to handle structured data, they cannot be used directly analyze a large-scale of semi-structured and non-structured data. Note that the current information technology still lacks the ability of computing big volume of semi-structured and non-structured data, such as clustering millions of text files, images or both in a reasonable time. To do this, we must find a way to transform the semi-structured and non-structured data to structured data or pseudo structured formats which can be analyzed by many known data mining or machine leaning algorithms [18]. This transformation process can be done by using the existing information retrieval algorithms in documents such as for information within documents and metadata about documents and web page. For a given objective of the transformation, some information retrieval algorithm can be applied to turn each text file into a single record with many attributes into a "structured or pseudo structured format". Similarly, an image can be transformed by using a known pattern recognition algorithm as a record of the transformed format. It can be observed that whenever the transformation objective is changed, the structured or pseudo structured format will vary. Therefore, the knowledge of Information Science can be effectively applied to treat the big data access and processing problem.

For Challenge 1.3, it is necessary to utilize rules and principles of Mathematics and Statistics in big data analytics and its interpretation. With the analyzable big data

formats, all possible methods in Mathematics and Statistics may be used to conduct Big data analysis. For instance, the modeling methods can include parent space identification and sampling; clustering, classification, regression, prediction and variable selection in data mining methods; relevance analysis, latent variable analytics and statistical inference in analytical methods; and sub sampling, complexity and distributed computation in computation methods. The challenge reflects when and which method is appropriate to be used in a particular case of big data analysis. Because the transformation of Big data is subject to the pre-determined objective, it can be useful to choose a method for data mining or knowledge discovery. Like traditional data mining procedure, experimental design for method choice should be conducted in such Big data mining for most of cases. However, the results of big data mining should be interacted with the user's judgment for the reason which knowledge changes with the individual and situation [16]. In order to let the user has a better understanding of knowledge from big data analysis, different representation or visualization methods, like uniform scheme can be employed to show the simple versions of big data complexity.

In addition, how to use knowledge from big data analysis in the real-life applications is not easy. This perhaps turns to an Engineering problem. Engineering is generally defined as "the application of scientific, economic, social, and practical knowledge in order to invent, design, build, maintain, research, and improve structures, machines, devices, systems, materials and processes" [24]. Use of big data knowledge in most of situations has to do with enhancing the current stages of either scientific, economic, or social conditions. Nowadays every corner and event of our human society depends on big data. Data-driven decision eventually becomes the most reliable approach to any problem. A good engineering design for Big data application will naturally yield the better way to achieve scientific, social and/or economic benefits.

Variety of big data applications can form a new industry, which can be called big data Industry. In such an industry, big data is the input, through big data analytic process mentioned in the above, the output will be data generated knowledge that can be easily turned into products, such as value chain management, business pattern, etc., to create a remarkable productivity.

1.3.2 Some Open Big Data Research Problems

In the notion of the theoretical and technical components of big data, big data analytics has the following open problems.

Problem 1 High Dimensionality
Given a database, when the number of features (p) is far larger than the sample size (n), and n varies with p ($n = n(p)$), the situation is called high dimensionality (HD) problem. When the problem occurs at Big data, $p \gg n(p)$. HD frequently appears in

medical science, such as DNA scanning. In the linear case a basic solution can be shown as:

Consider a linear model as $y = \beta_1 x_1 + \beta_2 x_2 + , \cdots, \beta_p x_p$ for dataset, $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. Then, the matrix format can be represented as $Y = X_{n \times p} \beta_{p \times 1}$ and the solution is $\hat{\beta} = (X'X)^{-1} X' Y$.

An asymptotical normality of this is:

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \sim N\left(0, \frac{1}{n}(X'X)^{-1}\sigma^2\right) \xrightarrow{d} N\left(0, \sigma^2 I_{p\times p}\right)$$

There are a number of recent approaches that may categorized as sparse modeling, including compressed sensing, low rank decomposition of matrix and sparse learning to deal with HD problems (for instance, see [25–28]). Some of these developed algorithms are available to be used to handle HD problems in big data. The open research questions for HD problems are how to add priors so that a HD problem can be well defined; and how to find effective sparse modeling, etc. Eventually, systematically solving HD problems need a build-up on the theory and methodology of either HD statistics or HD data mining.

Problem 2 Sub-sampling

The current technologies, like Hadoop system of processing Big data are some types of "divide-and-conquer" schemes, where sub-sampling techniques have been employed. For example, in MapReduce, Map is designed as random sub-sampling of sub datasets with intermediate solutions from given large database, where Reduce is an aggregation process of intermediate solutions for the final estimation of a given database [29]. Although sub-sample is one of the key concepts in big data processing, there are many open questions that need to address so that the more advanced technologies on big data can be developed. For example, how to sub-sampling/aggregate so that the final estimation model of the given original database is properly representing the database? Is the distributed processing feasible? How about traditional sub-sampling/re-sampling technologies working? Are there sub-sampling axioms, such as similarity and transitivity?

Problem 3 Computational Complexity

Traditionally, computational complexity concerns with how difficult a problem can be solved, or how much computation cost must be paid if an algorithm is used to solve a problem.

As an illustration, if a traditional setting can be $R = A(P) := A(D)$, where D is database, A is computation and R is the complexity. Then a Big data setting should be $R_t = A_t(D_t)$, where all D, A and R are changed with time associated with the cost. In this case, the core open questions are how to properly define complexity in big data setting? Is the complexity easy or difficult to measure for a given big data problem? How to establish complexity theory for some specific types of big data problems?

Problem 4 Real and Distributed Computation

Parallel and distributed processing become necessary, perhaps is a unique way of processing for Big data [30]. The main challenges of such a real and distributed

(R/D) computation in handling Big data come from the relationship of three components of Hadoop system: the Hadoop Distributed File System (HDFS) which is a distributed file system designed to run on commodity hardware, HBase which is an open source, non-relational, distributed database, and MapReduce. The quality measurements of Hadoop system for a real and distributed computation include real time, feasibility, efficiency, scalability, etc. It should be noted that some of the measures are conflicted each other and a compromised standard among them is a way to look for a good computational result. There are some open questions in this area. For example, does the R/D computation support fast storage/reading/ranking? For problem of decomposability, can a data modeling problem be decomposed into a series of sub-data set dependent problems? For solution assemblies, how can the solution of a problem be assembled with its sub-solution (component solutions)? When the distributed process is conducted, can the forward and incremental steps be performed by on-line computation?

Problem 5 Unstructured Processing
It has been commonly recognized that structured data are those that can be represented with finite number of rules and can be processed within acceptable time. Otherwise, the data are unstructured (some of them are also called semi-structured), which are difficulty to process (for example, thousands of images or text files). The main challenge of processing unstructured data is that they are multi-sourced and heterogeneous. In most of cases, the understanding of the data is cognition dependent. In this area, the core open questions are how to build a uniform platform on which different types of unstructured data (e.g., mixture of images, text, video and audio) can be processed simultaneously? How to develop the cognition consistent approaches for unstructured data modeling?

Problem 6 Visualization
Using visual-consistent figures or graphics to exhibit the intrinsic structure and patterns in HD big data is challenge visualization analysis. This requires building a basic tool for human-machine interface and expanding applications. For example, by using feature extraction, a HD data space can be transformed into feature space with low dimension (LD), and then by using to visualization techniques, the latter can be turned into visualized space with 2-dimension or 3-dimension. The key concept of judging a good visualization tool is that the end user can easily understand the meaning of big data results without knowing any technical analysis behind. Some current visualization techniques used in showcases, such as The Second Life (http://secondlife.com/) and video games, can be effectively applied to Big data visualization. The core open questions are: is there essential feature extraction of HD data (say, dimension-reduction)? What is structured representation of imaginable thinking? How to construct appropriate visualized space? How to map a problem in feature space (or data space) to a representation problem in visualized space? [31].

Big data analytics is still a very pre-mature field at this point. Fundamentally speaking, in order to conduct an applicable big data analysis, one should think about how to design big data analytic algorithm structure. Here are some ideas

open to be discussed. First, a big data analytic algorithm should be an algorithm that can process and analyze big data under available computational resource and complete in a reasonable time. The big data can be handled by it has at least one of following characteristics: large-size, heterogeneous, distributed, multi-sources, data steam, high-dimension, and high-uncertainty. The algorithm can be performed at appropriate degree of time, storage and communication complexity. It also has some unique properties, such as highly fault toleration, solution integration and assembled capability. Second, the key ideas of designing a big data analytic algorithm could include maintaining the proper ratio of data sample and population; simple modeling and simple procedure; inferior preciseness, complex inherence and theory based. Finally, in addition to well-known statistics or data mining methods, other computational methods, such as set-based processing, stochastic computing, online computing, distributed/parallel computing, cloud computing may be employed to construct a high-efficient big data analytic algorithm. These concepts and discussions about big data and big data analytics have been implemented in the following chapters of this book.

Looking around the world, big data development is just at the beginning. Big data is treasure created by the people and should be used to benefit the people. Even the precise meaning of big data analytics is not clear yet, the data scientists and engineers should figure out the fundamental issues of big data which may lead a context of data science. The advancement of data science will provide more theoretical findings and creative or innovative techniques to support the big data development into the future.

References

1. Tuitt, D.: A history of big data. HCL Technologies Blogs. http://www.hcltech.com/blogs/transformation-through-technology/history-big-data (2012)
2. EMC2: Digital universe study: extracting value from chaos. http://www.emc.com/leadership/programs/digital-universe.htm (2011)
3. Shaw, J.: Why "big data" is a big deal. Harvard Business Review, March–April. http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal (2014)
4. Chen, H., Chiang, R.H.L., Storey, V.: Business intelligence and analytics: from big data to big import. MIS Q. **36**(4), 1165–1188 (2012)
5. NSF: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA). National Science Foundation. http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm (2012)
6. XSSC The 462nd Session: Data Science and Big Data of Xiangshan Science Conferences. http://www.xssc.ac.cn/xs/showconf.asp?tid=4&pid=342 (2013)
7. Laney, D.: The Importance of "Big Data": A Definition. A Gartner Co. Report (2012)
8. Villanova University: What is big data? http://www.villanovau.com/university-online-programs/what-is-big-data/ (2014)
9. Bayes, T., Price, R.: An essay towards solving a problem in the doctrine of chances. Philos. Trans. R. Soc. Lond. **53**, 370–418 (1763)
10. Fayyad, U.M., Piatetsky, S.G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Fayyad, U.M., Piatetsky, S.G., Smyth, P., Uthurusamy, R. (eds.) Advances in

Knowledge Discovery and Data Mining, pp. 1–34. AAAI Press/The MIT Press, Menlo Park (1996)
11. Mayer-Schönberger, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, New York, NY (2013)
12. Blind men and an elephant. http://en.wikipedia.org/wiki/Blind_men_and_an_elephant (2014)
13. Zhang, L., Li, J., Shi, Y., Liu, X.: Foundations of intelligent knowledge management. J. Human Syst. Manag. **28**(4), 145–161 (2009)
14. Ouyang, Z.B., Shi, Y.: A fuzzy clustering algorithm for petroleum data. In: WI-IAT '11 Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 03, pp. 233–236 (2011)
15. Laudon, K.C., Laudon, J.P.: Management Information Systems. Pearson, Upper Saddle River, NJ (2012)
16. Shi, Y., Zhang, L.L., Tian, Y.J., Li, X.S.: Intelligent Knowledge: A Study beyond Data Mining. Springer, New York (2015)
17. G8 report on open data charter and technical annex. https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex (2014)
18. Shi, Y.: Big data: history, current status, and challenges going forward. The Bridge. **44**(4), 6–11 (2014)
19. Zhong, Y., Shi, Y., Jing, X.: Global big data development analysis report 2018. A report by Tianfu Institute of International Strategy and Technology, Chengdu, China (2019)
20. Action Framework for Promoting the Development of Big Data (AFPDBD). www.gov.cn (in Chinese) (September 2015)
21. Shi, Y., Shan, Z., Li, J., Fang, Y.: How China deals with big data. Ann. Data Sci. **4**, 433–440 (2017)
22. Shan, Z.: Interpretation on action framework for promoting big data. http://news.xinhuanet.com/info/2015-09/17/c_134632375.htm (2015) (in Chinese)
23. Xu, Z., Shi, Y.: Exploring big data analysis: fundamental scientific problems. Ann. Data Sci. **2**, 363–372 (2015)
24. Wikipedia: Engineering. http://en.wikipedia.org/wiki/Engineering (2020)
25. Chang, X., Wang, Y., Li, R., Xu, Z.: Sparse K-means with L∞/Lo penalty for high-dimensional data clustering. arxiv.org/pdf/1403.7890 (2014)
26. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution. Commun Pure Appl Math. **56**(6), 797–829 (2006)
27. Kriegel, H.P., Kröger, P., Renz, M., Wurst, S.: A generic framework for efficient subspace clustering of high-dimensional data. In: IEEE International Conference on Data Mining (ICDM), Houston, Texas, USA, pp. 205–257 (2005)
28. Wang, Y., Chang, X., Li, R., Xu, Z.: Sparse K-means with Lq (0<=q<1) constrain for high-dimensional data clustering. In: IEEE International Conference on Data Mining, Dallas, USA (2013)
29. Kleiner, A., Talwalkar, A., Sarkar, P.: The big data bootstrap. In: The 29th International Conference on Machine Learning, Edinburgh, Scotland, UK (2012)
30. Xu, Z., Leung, K.S., Liang, Y., Leung, Y.: Efficiency speed-up strategies for evolutionary computation: fundamentals and fast-Gas. Appl. Math. Comput. **142**(2, 3), 341–388 (2003)
31. Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y.: T-Drive: driving directions based on taxi trajectories. In: ACM SIGSPATIAL GIS (2010)