Frontiers of Engineering Management
DOI 10.15302/J-FEM-2015042

ENGINEERING MANAGEMENT REPORTS

Yong Shi

Challenges to Engineering Management in the Big Data Era

Abstract This paper presents a review of the challenges to engineering management in the Big Data Era as well as the Big Data applications. First, it outlines the definitions of big data, data science and intelligent knowledge and the history of big data. Second, the paper reviews the academic activities about big data in China. Then, it elaborates a number of challenging big data problems, including transforming semi-structured and non-structured data into "structured format" and explores the relationship of data heterogeneity, knowledge heterogeneity and decision heterogeneity. Furthermore, the paper reports various real-life applications of big data, such as financial and petroleum engineering and internet business.

Keywords: big data, data science, intelligent knowledge, engineering management, real-life applications

1 Introduction

As the computing technology has been rapidly advancing since the 1950s, our human culture has developed the ability to generate masses of data. Internet connections make it possible to share data in real time on a global basis. The Big Data Era has surely arrived. According to an IDC report by Gantz and Reinsel (2012) the US has 32% of the digital universe (big data market), Western Europe has 19%, China has 13%, India has 4%, and the rest of combined nations have 32%. By 2020, the emerging markets will have 62% of the share of the digital universe while China alone will generate 21% of the big data in the world. This prediction could be true since China has 1.3 billion populations with 641 million internet users and 700 million cellular phone users. More people generate more data. Due to its population, China may soon become the country generating the most data in the world. Big data becomes the most influential force for daily life in China. The purpose of this paper is to address the challenges of management science and engineering in the Big Data Era, especially from the point of engineering management view. The paper proceeds as follows. Section 2 outlines the definition of big data, data science and intelligent knowledge and the history of big data. Section 3 reviews the academic activities about big data in China. Section 4 elaborates a number of challenging big data problems. Section 5 provides various real-life applications of big data, such as financial and petroleum engineering and internet business. Finally, Section 6 concludes the paper with some reflective remarks.

Manuscript received June 19, 2015; accepted September 22, 2015

Yong Shi (✉)
Research Center of Fictitious Economy & Data Science, University of the Chinese Academy of Sciences and Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
Email: yshi@ucas.ac.cn

2 Big data, data science and intelligent knowledge

2.1 Definition of big data

There are various definitions of big data. For example, the National Science Foundation of the USA (NSF, 2012) refers to big data as "large, diverse, complex, longitudinal, and/or distributed data sets 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, the 462nd Session of Xiangshan Science Conferences[1)] on "Data Science and Big Data," co-chaired by the author of this paper, provided two definitions for big data. The first for academic and business communities states that big data are "a collection of data with complexity, diversity, heterogeneity and high potential value, which are difficult to process and analyze in reasonable time," while the second says that big data are "a new type of strategic resource in digital era and the key factor to drive innovation, which is changing the way of human beings' current production and living." Commonly, wherever big data are mentioned, the "4V's" — volume,

1) The 462nd session: data science and big data. [Previous Meetings]. Xiangshan Science Conferences. Retrieved from http://www.xssc.ac.cn/xs/showconf.asp?tid = 4&pid = 342

velocity, variety and veracity — are used to capture its main characteristics (Laney, 2012; Villanova University, 2014). It is necessary to point out that the key to big data applications is to use data mining for discovering knowledge from big data so as to create powerful productivity.

2.2 History of big data and big data mining

The history of data analysis can be viewed in three stages. The first one was that from 1700 to 1950, in which statistical analysis played a key role for 250 years. This kind of analysis is descriptive. The second stage was from 1950 to 2012, in which machine learning and artificial intelligence dominated data-mining methods were used in addition to statistical analysis. These results are analytic. The third stage is just starting; now, big data analysis can be conducted.

One of founding fathers in statistical analysis is Richard Price (1723–1791). He collected observations on reversionary payments for calculating the values of assurances on lives and the British national debt (Price, 1783). In his 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. For a quite long time, data analysis could not widely affect our daily life until the creation of computers in the 1950s. Using the computational and storage power of computers, people have gradually developed databases, database management systems, and data warehouses in past 60 years. These computing environments provided the bases for us to efficiently use a number of mathematical tools, such as statistics, artificial intelligence, machine learning, and so on to solve large-scale data analysis problems. It is noted that almost all of data mining methods, including statistical analysis, are used to handle the structured data. Most of this data are digital data. These methods are not effective in analyzing semi and/or unstructured data, such as image, voice and sound. The evolution of data analysis or now big data analysis tells that there are no data mining methods or algorithms that can directly and effectively analyze big data. Although the computing technology has exponentially been developing, the big data mining methods that directly deal with all kinds of big data may not occur for a while due to the huge computational complexity.

2.3 Big data, intelligent knowledge and data science

Presently, big data as notion has two different meanings. To some people, big data refers to its applications. To others, big data means both theory and applications. If big data are a phenomenon, data science is behind big data. The definition of data science is not commonly stated. However, one can roughly say that data science is the science for data collection, management, transfer, analysis and application, in which the key is the research in obtaining knowledge from data. Data science may differ from other sciences of subjects. It has no universal rules or laws for all kinds of big data. Instead, the rules and laws of data science will change with big data from different fields. For example, the contexts of data science in finance could be different from that of petroleum engineering.

In discovering knowledge from big data, the concept of intelligent knowledge will be important (Zhang, Li, Shi, & Liu, 2009; Shi, Zhang, Tian, & Li, 2015). In big data mining, although rough knowledge (known as hidden patterns) in the "first-order mining" (known as from data to hidden patterns by data mining) is derived from heterogeneous data, it can still be 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" (the process of combing human knowledge with hidden patterns) is used, the structured knowledge is combined with human knowledge of decision makers that are semi-structured or unstructured and upgraded into intelligent knowledge. In other words, intelligent knowledge is influenced by the combination of rough knowledge and human knowledge through the second-order mining process, which is the representation of unstructured knowledge.

3 Some activities about big data in China and around the world

3.1 Big data in China

To prepare the academic environment for big data research and applications, Chinese institutions have organized a number of conferences and meetings for the past 11 years. In 2004, Chinese Academy of Sciences (CAS) held an international symposium on data mining and knowledge management which gathered 40 attendees (Shi, Xu, & Chen, 2005). This symposium initiated the academic exchange between computer scientists who work on data mining and management scientists who concentrate on knowledge management. The promotion of two fields became a driving force to support data analysis, later being called "big data analysis." In 2006, the 278th session of Xiangshan Sciences Conferences, called "The Frontier Studies on Data Technology and Knowledge Economy," focusing on the theory and technology in data technology and knowledge economy was held in Beijing. The aim of this conference was to provide a platform for researchers, practitioners, and the data mining user communities to propose novel ideas, share their research and experience, exchange techniques and tools, and explore cutting-edge research. There were 35 Chinese scholars, 16 well-known international scholars, and 16 Chinese government representatives, CEOs and managers who participated in the two-day conference. During the conference, these participants freely shared their fresh ideas, challenging thoughts,

and new research directions in dealing with various data technology and knowledge economy problems. All of them believed that based on the diversity of existing fields, known as data mining, intelligent knowledge, wisdom mining, and data technology, the academic community should and need to build a new and fresh field called "data science." As long as it is known, this was the first time that the academic world announced the notion of data science based on data mining and knowledge management (Cheng, Dai, Xu, & Shi, 2006). Following this event, CAS formally established its Research Center on Fictitious Economy and Data Science in 2007. This was five years before a similar data science unit built by Columbia University, USA.

From 2010 to 2013, a group of researchers organized annual International Workshops on Data Science to discuss and explore the phenomena and rules of data nature as well as some fundamental theoretic issues in data science. The annual workshops attracted more than 300 scholars and industrial professionals from Australia, Canada, China, Japan, the UK and the US. To further expand the preliminary findings and exchanges on this platform, the First International Conference on Data Science (ICDS 2014) was held at the CAS in Beijing in 2014. The second International Conference on Data Science (ICDS 2015) will be held on August 8–9, 2015, by the QCIS Research Centre, University of Technology Sydney. The theme of ICDS 2015 will be "Global Perspectives on Data Science." In addition, two more sessions of the Xiangshan Sciences Conferences were held in 2012 and 2013, in Beijing, China. The 424th session was called "Network Data Science and Engineering" organized by CAS, Chinese University of Hong Kong and Tsinghua University in 2012. The 462th session was called the "Data Science and Scientific Foundation of Big Data and its Perspectives," organized by the CAS, Fudan University and the University of Illinois in Chicago, USA in 2013. The attendees at these conferences discussed a broad range of problems in big data, including the challenges of big data science and engineering; social, economic and IT problems in Big Data Era; common theoretic bases in network data science; ecology of network big data; and the foundations of data science and big data mining techniques. In 2013, the Research Center on Fictitious Economy and Data Science, and the CAS organized "Top-level Forum Computer and Economy Development under big data" in Beijing. This forum invited 41 experts, including 10 members of the CAS and the Chinese Academy of Engineering (CAE), some officers from various governmental branches, as well as a CEO of corporations. The topics involved how the government can begin a big data source, how corporations can share big data with each other, and how academia can build an alliance to use big data across different fields.

Based on the above academic activities in big data, CAS announced its decision to build the Key Laboratory of Big Data Mining and Knowledge Management in order to reinforce its research strength on big data research and applications. With active collaborations, four units — the Research Center on Fictitious Economy and Data Science, the CAS, School of Mathematical Sciences, School of Computer and Control Engineering and School of Management of the University of CAS are united as the major components of the laboratory. Other institutes of the CAS, including the Institute of Automation and the Institute of Policy and Management also jointed its research projects. This laboratory is concentrated on the scientific base of big data, big data modeling, mining algorithms, big data techniques and applications. The goal of the laboratory is to use empirical methods and big data mining to explore the characteristics and rules of finance and industrial development, management decision, scientific innovation, and internet and social computing. It will also seek the data science theory, and build foundations of intelligent knowledge management. Big data analysis can be classified as the process of big data collection, storage, analysis/mining, knowledge creation, applications and policy advice (see *Figure 1*). This laboratory is mainly involving with all aspects, but storage which is handled by computing technology. It has a unique reputation of dealing with social, economic and internet big data.

3.2 Big data in the world

In 2013 and 2014, the author was invited by "The Bridge," the official magazine of the National Academy of Engineering of the US, to be the Guest Editor of the special issue on "A Global View of Big Data" (Winter 2014), where seven articles contributed by 14 authors and co-authors from 9 countries and regions — Australia, Brazil, China, Japan, Romania, Spain, the UK, the US, and Hong Kong of China (Shi, 2014a).

The first paper of this issue, by the author (Shi, 2014b), outlined the trends of big data development from a science and engineering point of view. The second paper, Tien (2014) reviewed the current stages of big data development in the US, focusing on big data challenges in the 14 related areas. The third paper, by Li, Zhang, Wu and Zhang (2014), showed how to use big data to improve the performance of the Chinese financial industry. The fourth paper, by Tsumoto (2014), summarized the achievements of big data in various areas of Japan. The fifth paper, by Filip and Herrera-Viedma (2014), provided the strategies of the European Union in its big data movement. The sixth paper, by He, Liu, Huang, Blumenstein and Leung (2014), reported how commonwealth countries such as the UK, Canada, India, Australia and South Africa, have taken different efforts to deal with big data. The seventh paper, by Gomes (2014), provided the highlights of big data applications in Latin American countries including Brazil, Mexico, Chile, Peru, Colombia and Argentina. Big data now is quickly changing the world.

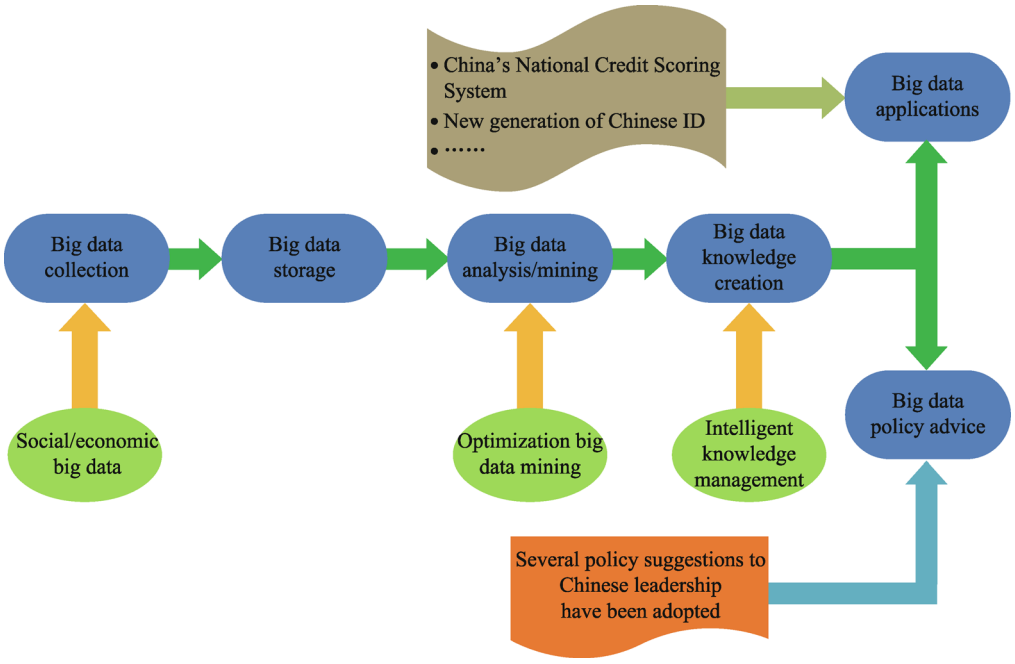


Figure 1. Research functions at the Key Laboratory of Big Data Mining and Knowledge Management, CAS.

4 Challenges of big data

There are many challenges in utilizing big data. Discussions of these issues are made from two different perspectives. The first is about challenges cross a broad range of fields when big data are applied. The second is about challenges to the field of engineering management. To begin, it is necessary to restate three major challenges that require urgent solution in order to gain benefits from big data in science, engineering and business applications (Shi, 2014a&b).

4.1 How to transform semi-structured and unstructured data into "structured format"

As it was pointed out before, the key objective of applying big data are to discover knowledge (or value) from data. However, many known data mining methods are designed to analyze the structured data. Due to the limitation of current computing technology, these methods cannot analyze a large amount of semi-structured and unstructured data, collected by Hadoop and Map Reduce, in a reasonable amount of time. The challenge is effectively analyzing this data and discovering knowledge from them in an expedient timeframe. The answer could be to 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 unstructured data into structured format should be subject-oriented.

4.2 Exploring the complexity, uncertainty and systematic modeling of big data

Data as the representation of a given object is only a partial picture of the facts. The complexity of big data comes from the interpretation of data representation while the uncertainty of big data comes from the changes of the objects in the nature as well as the variety of data representations due 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. A breakthrough to a systematic modeling on complexity and uncertainty of big data analysis and mining is needed for gaining knowledge from big data. 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. If the engineers can find 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 a huge value and added effects to social development and economic changes.

4.3 Exploring the relationship of data heterogeneity, knowledge heterogeneity and decision heterogeneity

Decision makers face three heterogeneous problems with big data environment — data heterogeneity, knowledge heterogeneity and decision heterogeneity. In the Big Data Era, discovering knowledge now is more based on data analysis and data mining. Under a theory of management information system (Laudon, K., & Laudon, J., 2012), decision making can be classified as structured decision, semi-structured decision and unstructured decision according to different levels of the responsibilities of individuals in an organization. The low-rank operational staff members produce structured decisions, while the managers create semi-structured decisions by adding their own judgments to subordinates' structured reports. Finally, the executive officer (CEO) makes a final decision, which is unstructured by remarks or voice. Big data disruptively changes the decision making process. A system of big data analysis or mining can combine all functions of business operations (structured decisions), managers (semi-structured decision), and the CEO (non-structured decision) into a single decision without mistakes. The challenging problem will be determining if knowledge from data that is demanded to make a decision should be heterogeneous, (such as structured knowledge), semi-structure knowledge or unstructured knowledge. Once this is determined, it is necessary to decide how to continue studying them.

In addition, more challenges include exploring the laws and rules for big data collection, exploring different impacts of close database and open data source on knowledge discovery, exploring the logical patterns of semi-structured and unstructured data and the rules of multidimensional tables based on semi-structure and unstructured data, exploring the global and local solutions in big data mining, and exploring the change of decision making structure based on data structure.

Considering the field of engineering management, the author observes the following challenges:

4.4 Decision must be formed by data and knowledge

Normally human beings make decisions based on their own experiences and added information which gathered from outside experience. If the experience is reviewed as known knowledge, then the added information can be advice or suggestions from others. With the rapid development of IT and the complex world, such added information is also represented by data. According to the theory of intelligent knowledge (Zhang, Li, Shi, & Liu, 2009; Shi, Zhang, Tian, & Li, 2015), an automatic process of producing effective decisions based on big data should be done in two stages. First, analyzing the related data via the first-order of data mining provides rough knowledge. Second, combing rough knowledge with human knowledge by second-order mining obtains intelligent knowledge. Here, data implying rough knowledge as the added information in coordination with decision makers' knowledge, will result in effective decision.

4.5 Data-driven decision making dominates engineering management

Traditional engineering and decision scientists make decisions based on either case-driven or model-driven approaches. In the Big Data Era, none of these approaches can take fully advantages of value from data. On one hand, although the case bank can provide the decision makers with good examples, it has a limitation of offering various environments where big data can be supported. Data-driven approaches based on big data can explore more variations than case-driven approaches. On the other hand, the model-driven approach requires the basic elements of modeling, which has a limitation of handling non-normative events. However, many real-world applications cannot be solved by normative methods, such as model-driven approaches. Data-driven approaches based on big data have no pre-requirements and are more loose than model-driven approaches which can be flexibly used in the search for a solution.

4.6 Cross-industrial standards as a research map of engineering management

If big data can offer a better solution than other known methods, a better solution using big data should be explored. This challenge can be answered by the CRISP-DM (Cross Industry Standard Process for Data Mining), which was recommended by the European Union with the support of several well-known companies[1)] (Olson & Shi, 2007). The traditional decision making process of establishing decision objectives, informational gatherings, building decision alternatives, choosing optimal solutions, and executing the solutions, can be replaced by six steps of CRISP-DM in big data. The steps include business understanding, data understanding, data preparation, modeling, evaluation and deployment. The process of the CRISP-DM as a road map for engineering management is the preferable road map for transferring data to knowledge.

Given the current stage of big data applications, some challenges need to be overcome for different management issues. For example, what is the rule of business behavior under big data? How can big data change enterprises' operations? What marketing strategies should be used with big data? What is the pattern of business intelligence in big data? How to build a big data computing and analytical platform for various applications? Any advancement toward answering these challenges will bring the benefit

1) CRISP-DM. Retrieved from https://en.wikipedia.org/wiki/Cross_Industry_ Standard_Process_for_Data_Mining

of enhancing the social and economic development of our society.

4.7 Research progress on big data problems

Based on the big data challenges listed above, researchers across different fields are now actively working on the related problems and trying to make breakthroughs. To transform semi-structured and unstructured data into a "structured format," a strategy of ancient Chinese method of "Paodingjieniu" (Watson, 1964) can be used. For example, given a hundreds and thousands of images, one cannot classify them into different categories with a reasonable time due to the computational complexity. If one builds a classification task or target, he can set up a number of attributes or variables, such as indoor photo vs. outdoor photo, human vs. non human subjects, and so on, to map each image by attributes as a record onto a multi-dimensional table. As a result, the images become digital values on the table. Here the author treats the image as a cow in the "Paodingjieniu" story and its abstract form as a record with attributes is a bone structure of the cow. With the abstract multi-dimensional table, one can easily conduct analysis by employing data mining algorithms. Once the rough knowledge results from data mining process, one can transform the meaning back to the original identity of the image. In this way, the large-scale images can successfully be classified. Similarly, when one handles a large amount of documents or texts, he can apply a known topic model (Blei, 2012) as a tool to transfer them into another abstract multi-dimensional table so as to perform the data mining algorithm. It is important to point out that the research difficulty in this problem is to find an appropriate method to transform semi-structured and unstructured data into a multi-dimensional table. The method changes with the subjects and nature of big data and may cost time to figure out. Furthermore, researchers have made progress on a number of problems, such as reduction of high dimensions, sub-sampling, computational complexity, real and distributed computation, unstructured processing and visualization (Xu & Shi, 2015).

5 Real-life big data projects

In the past ten years, a number of real-life big data projects have been completed. Now this paper shares some meaningful projects, which range from credit scoring systems, credit card issuing, insurance, petroleum engineering, internet business, financial exchanges and online business.

5.1 China's National Credit Scoring System

From 2006 to 2009, the author led a research and technical group at the People's Bank of China to use the National Personal Credit Database (Petabyte Big Data), which is the world's largest database of its kind and contains all Chinese banking records, and developed China's National Credit Scoring System, called "China Score"[1]. This system differs from the "FICO Credit Scoring" that is widely used by Credit Bureaus in the US, Australia and some European countries. China Score was designed by our theory and methods in optimization based data mining and intelligent knowledge (Shi, Tian, Kou, Peng, & Li, 2011).

Figure 2 shows the comparison of four data mining results — Logistic Regression (LR), Multiple Criteria Linear Programming (MCLP), Multiple Criteria Quadratic Programming (MCQP) and Support Vector Machine (SVM) — over a big credit data. Among them, LR is a known method in the field while SVM is a recent popular method. MCLP and MCQP are our unique and new data mining methods. As one can see, in the classification of good vs. bad accounts, the Kolmogorov-Smirnov (K-S) score (used for odds) of MCLP and MCQP were slightly lower than that of LR and SVM, and the Gini score of these are much better than LR and SVM. *Figure 3* shows that the comparison of the average score of four alternative models in China Score, the American credit score (682), and the average Australia credit score (666). In the US, one of three major criteria for personal subprime is that the credit score is below 620. The US credit scoring system has played an important role to save the US economy from the past subprime crisis. Here, China's scores illustrated the similar pattern with that of the US and Australia since the major part of China's personal expenditure is mortgage, car loan, etc. The accuracy rates, K-S score and Gini score of four China Scores are higher than those of known approaches used in FICO scores, including American and Australian scores. Dr. Min Zhu, the former vice governor of the People's Bank of China and the vice president of International Monetary Fund (IMF) evaluated "China Score outperforms other international scoring systems"[2]. From 2008, all Chinese commercial banks are using China Score in their banking and financial activities, such as mortgage and business loans, and have saved billions of CNY each year. For example, seven major banks saved 150×10^9 CNY (about 24×10^9 US Dollars) in 2008. In 2009, a proposal was initiated to the Chinese Government to establish a "National Fair Credit Act" to guarantee the implementation of China Score in all commercial banks in China. Since China Score is used for all of 1.3 billion Chinese to handle daily commercial life, it will be an

1) Credit scoring system. Retrieved from http://baike.baidu.com/link?url= 13IyeLcfD4 Frg MQ1gqPtbQdDdJtun1Ya77LonOxtA4-1wsCZhBq7h99-zvYqqVp94se7SZURCJ4mBbds9QbnCq

2) China News. (February 25, 2010). China's credit scoring system. Retrieved from http://www.chinanews.com/gn/ news/2010/02-25/2137072.shtml

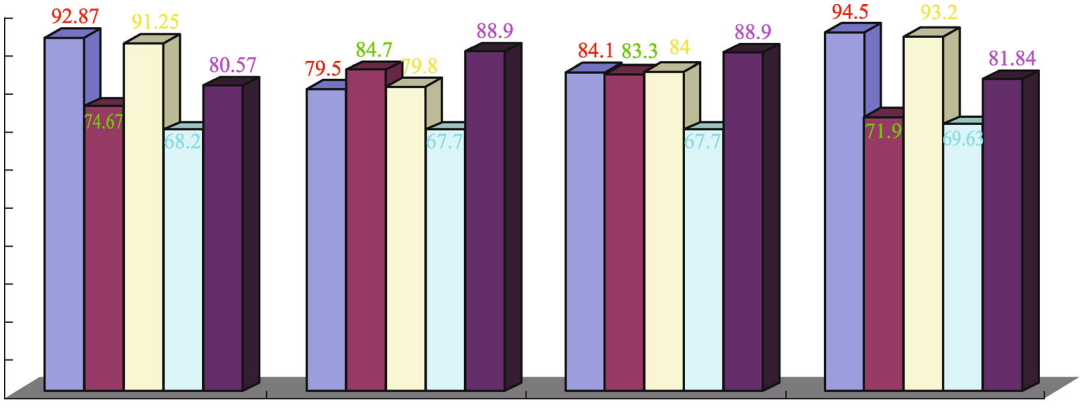
important tool for China to prevent the possible economic crisis in the future and could be one of the most influential big data applications in history.

5.2 Real-time credit card approval system

In 2008, the author supervised a team building a real-time credit card approval system for the Nebraska Furniture Mart (NFM), USA, which is one of Warren Buffett's firms. Normally, a bank or credit issuer (which may not be a bank) in the US will take two or three days to process and issue a credit card to applicant. There is a risk for the issuers when the cardholder will soon become bankrupt. NFM has more than 2×10^6 credit cardholders and their records formed big data. The previous credit card approval system has some shortcomings. First is that it rejected a certain number of applicants who deserve a credit card. This is called a false negative or type I error in statistics. Second is that NFM issued a card to an applicant who soon be bankrupted, which is called a false positive or Type II error in statistics. To reduce both error rates, the author and his fellow researchers developed the new real-time approval system, called OverScore™, by using data mining tools including Decision Tree, SVM, Naive Bayes and Multiple Criteria Mathematical Programming (Shi, Tian, Kou, Peng, & Li, 2011). When the applicant submits the application form including personal history and related data, the OverScore™ system will take two seconds to calculate the applicant's credit scores from three credit bureaus (Equifax, Experian and TransUnion) and other information. Then, it will perform the analytic procedure that first classifies the applicant into his or her group among 2×10^6 cardholders, then searches 4 cardholders who are similar to the applicant. The clerk of NFM will make a decision on issuing a credit card with a correct limit based on the credit limits of these 4 customers. The OverScore™ system is easy for clerks to operate. The clerk's decision is final and no team discussion behind the service desk is needed. The system has been running for seven years and no mistakes or errors have occurred and the risk of misjudging applicants is nonexistent. It does not only eliminate Type II error, but also prevents Type I error. As it is known, a Chinese commercial bank currently needs a few days to a week to issue a credit card. It is reasonable to believe that this real-time big data technique can help Chinese banks to improve the credit card service since China soon will be the number one nation to use credit cards for daily events.

5.3 Petroleum exploration engineering via big data

In petroleum exploration engineering, the spatial database generated from seismic tests and well log data collection is big data with a high degree of complexity. In the traditional approach, analysts first construct differential equations to model the spatial database. Then the equations are either reduced into pair-wise linear equations solved by software or are simulated for approximate solutions by simulation software. The solutions are finally fed back to the database to adjust the parameters of the equations. The process will be repetitive until the better solution is found. Instead of using this approach, the author tried a data mining approach for petroleum exploration engineering (see *Figure 4*). The project was a joint research with the well-known resource company, BHP Billion, in Australia between 2004 and 2010. Given its Kipper 3D database, where each record or point with impedance (attribute) and



	Good	Bad	Total	K-S	Gini
Logistic regression	92.87	74.67	91.25	68.2	80.57
MCLP	79.5	84.7	79.8	67.7	88.9
MCQP	84.1	83.3	84	67.7	88.9
SVM	94.5	71.9	93.2	69.63	81.84

Figure 2. Models comparison in China's National Credit Scoring System (China Score).

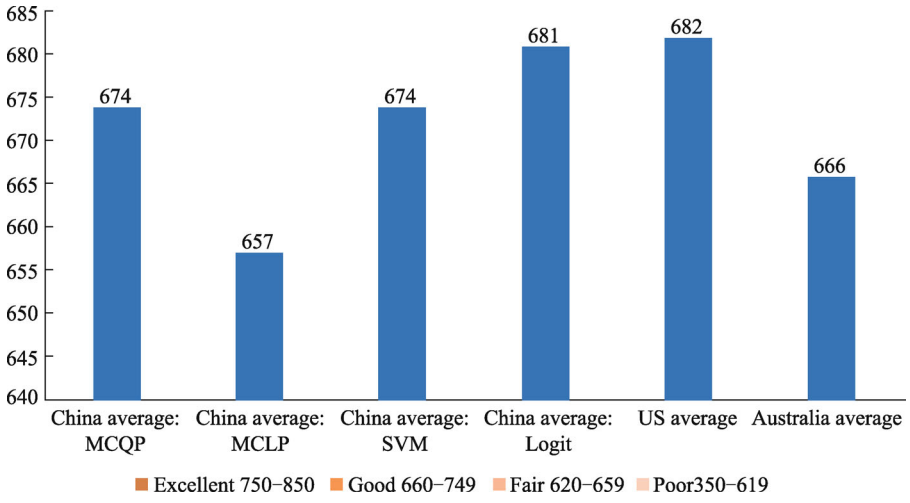


Figure 3. China score vs. US score vs. Australia score.

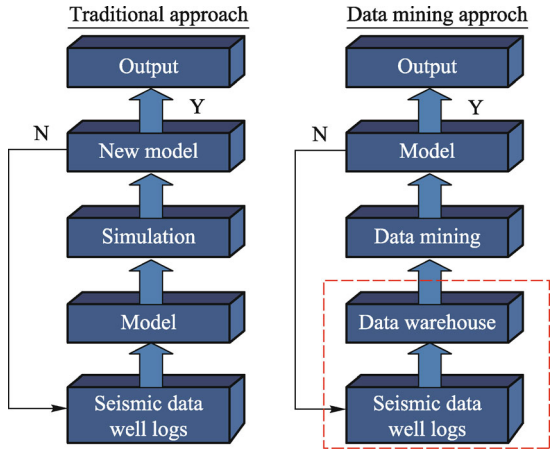


Figure 4. Traditional approach vs. data mining approach in petroleum exploration engineering.

three coordinates (x, y, z), the nonlinear patterns are changeable via different dimension and angles due to its complexity. The author built a multi-dimensional table for Kipper 3D database, and then performed the clustering analysis on seismic data and classification on well log data when a surface of data are fixed. When the surface was changed, the author obtained different data mining results. The core contribution of the data mining approach was to derive the intelligent knowledge based on data mining results from all surfaces (Ouyang & Shi, 2011). The result of data mining approach to this problem of petroleum exploration engineering not only proved to be more accurate than the traditional approach, but also demonstrated the great potential of big data mining in the field.

5.4 Netease VIP churn analysis

Netease is one of China's major portal sites. The company has three major businesses. The first one is the online games, which is not related to data analysis since the players are not required to reveal their true identity. If data mining process is conducted, it will be "garbage in and garbage out." The second business is advertisement. Although there are hundreds of companies using the Netease's platform for their ads, there are a few companies paying big fees to Netease. This does not require data mining. However, the third business, which is VIP churn analysis, can be accomplished by data mining. The problem that the company faced is that each day there are hundreds or thousands of customers joining its VIP service, but also many of VIP customers leave the service because of the intense competition and other objective reasons. It is necessary to simplify the rules so customers are not turned away by the complex nature of having so many rules. For this reason, Netease asked our team to investigate its real-life big data in 2006. A large amount of time was spent on transforming its log files which was data streamed into a multi-dimensional table. Using a Decision Tree analysis, 245 rules were found which could be thought as too many. When they were presented to the marketing manager, she immediately pointed out several from the list, which could be used for her decision support. This observation triggered the author to discover the need for intelligent knowledge which is the combination of rough knowledge (245 rules from data mining) and human knowledge (expertise of the manager) (Shi, Zhang, Tian, & Li, 2015). Intelligent knowledge showed the importance of human knowledge in big data analysis and applications.

5.5 Insurance claim fraud detection

In 2004–2006, the author worked with Mutual of Omaha Corporation, USA, which is one of Fortune 500 companies for its insurance claim fraud detections. Mutual of Omaha is one of largest insurance providers in USA. Each day, it

issues hundreds of insurance claims to customers with a large amount of money involved. The company has about 5×10^6 rows of health claims data at the Service Line Level with more than 100 variables. The project required understanding of the characteristics of insurance claims data, identifying abnormal insurance claim groupings by clustering, and reducing the workload of investigators. Given the case of claims in the state of Texas, the author used data mining techniques to find the largest distance to Nearest Cluster value with 76,921 records for the full evaluations of customers' claims. The significant number of abnormal insurance claims was correctly identified and the project was successful (Shi, Tian, Kou, Peng, & Li, 2011).

5.6 Online advertisement with big data

In 2012–2013, the author worked with Sojern Company on how to use its big data of air tickets information to form online marketing strategies. Sojern, as the leading company in US processing travelers' online boarding passes, now has more than 200×10^6 traveler profiles, and billions of intent data points as a big data driven travel marketing platform. It took the team 14 months to figure out and convert Sojern's big data stream into several multi-dimensional tables, including all converted users, turn users, airports in the US, and users with 162 attributes. The data mining techniques were used to perform analyses for several groups of attributes depending on the nature of online advertisements. The Scoring Cookie IDs were finally identified as the product for the clients to form an online marketing strategy for the travelers before their trips. In addition to this project, online advertisement problems can be captured as advertisement clicking predictions, which can further be viewed as click-through rate (CTR) prediction and identifying clicked ads in a set of ads. Some well-known data mining methods, such as SVM and multiple criteria mathematical programming can be effectively applied in handling these online ad big data problems (Lee, Shi, Wang, Lee, & Kim, 2015).

5.7 Traders monitoring in financial futures market

In 2012, our team had a joint project with China Financial Futures Exchange (CFFEX) to use CFFEX's trading big data to prevent trading risk. CFFEX was established in 2006 under the authorization of China Securities Regulatory Commission (CSRC). CFFEX is the only platform of China's financial futures market for trading of financial derivatives. It generates 40×10^6 CNY in service fees each day. CFFEX has 60 clearing members and each member has a number of supporting trade members. To maintain a normal trading stability, our project built a real-time risk prediction model and alternate day risk prediction model for each clearing member. However, the key was how to access the behaviors of trading members who are associated with a clearing member. Among their trading data stream, the author and his fellow researchers elicited the multi-dimensional tables with different time windows for data mining processes. The rough knowledge from data mining procedures was used to construct the trading behavior patterns to monitor the current day's trading and predict future trade. This project is the first kind of big data applications in financial future markets in China.

5.8 Big data warehousing at Xinhua News Agency

Since 2012, the author has been working with Xinhua News Agency to develop its big text data into a super data warehouse. The goal is to utilize such a data warehouse for various analyses similar to what Reuters and Bloomberg do, which can generate different indices and reports to support China's social and economic development. In the first stage, the author and his fellow researchers built hundreds of attributes from a variety of topics as the data structure for the big data warehouse. Then, they established several data marts, such as real estates, bio-technology, etc. These data marts will serve as the simulators of the data warehouse. Using new big data mining techniques, they transformed the big text data into the multi-dimensional tables with the "Paodingjieniu" method mentioned above. The author and his fellow researchers first examined the pattern of change in China's real estates and China's bio-technology innovation, respectively. Then, they were further able to compare the influence of these two fields for its impact on China's sustainable economic development. This project is ongoing and will be another significant big data application in China.

5.9 Yihaodian Online Credit Scoring System

Shanghai Yi, an Agel E-commerce Ltd. known as Yihaodian, was founded in 2008 as an online grocery store in China. It has now become one of the top online businesses and is growing rapidly. The business chain of Yihaodian consists of three components — the supply, online business clients and consumers. In 2013–2014, the author worked with the company on forming the credit scoring model for its online business clients and both the credit scoring model and the value model for consumers. Since the supply market of Yihaodian is a buyer's market, its credit scoring model was not considered at this point. With big online data in Yihaodian, the credit scoring for business clients was made of bank indices, the Yihaodian index and derived index of business clients, while the credit scoring model for consumers is similar to the structure of China's National Credit Scoring System. The online data has already been added. The uniqueness of the project is found in the value model for consumers, which derives a special value of particular consumers to promote a long-

term business with them. For example, from the spending behaviors, if it is found that a consumer has an upper level house or a luxury car, milk will be freely delivered to him or her to promote online shopping opportunity. After these credit systems were implemented on Yihaodian's platform, the call rate from consumers to the company declined quickly, which showed the effectiveness of the system in online business.

6 Conclusive remarks

This paper has outlined the major concepts in the Big Data Era including big data, data science and intelligent knowledge, although these definitions are still changeable. The author has recalled the history of data analysis so as to understand the process of how our human beings use data to improve our life. The author has reviewed the known academic activities of China in big data from the last decade. The key concentration of this paper is to state a number of challenging big data problems for engineering management, such as semi-structured and non-structured data transformation, data heterogeneity, knowledge heterogeneity, decision heterogeneity, decision components, and so on. In addition to the progress of big data research, the author has also reported several real-life applications of big data. These cases have demonstrated that big data can change all aspects of our society. However, it is necessary to observe several problems. The first is that our current technology is not adequate to handle big data. Although people of today have certain computing power to process and analyze big data, they still lack efficient technology that allows them find the value of big data as they wish. Given the present computing technology, it could take a long time for us to fully utilize the potential value of big data. The second is that big data means big opportunity. Since human beings must rely on computing devices such as cellular phones, one can constantly generate data from our daily activities. The generated big data in turn provides us unlimited data expansion to upgrade our lives. A person with any discipline will find a chance of using big data to create profit or benefit. From this point, if any challenge mentioned in this paper has been solved, engineering management as a field will be advanced, our social and economic life will be improved.

Acknowledgements Part of this paper has been presented at the 9th China Engineering Management Forum organized by Division of Engineering Management, Chinese Academy of Engineering, May 16–17, 2015. This work was partially supported by the key research grant "Innovative Research on Management Decision Making under Big Data Environment" (Grant No. 71331005), "Non-structured Data Analysis Methods and Key Technologies for Management Decision Making" (Grant No. 71501175) and the key international collaboration grant "Business Intelligence Methods Based on Optimization Data Mining with Applications of Financial and Banking Management" (Grant No. 71110107026) from the National Natural Science Foundation of China.

References

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55, 77–84

Cheng, S., Dai, R., Xu, W., & Shi, Y. (2006). Research on data mining and knowledge management and its applications in China's economic development: significance and trend. *International Journal of Information Technology & Decision Making*, 5, 585–596

Filip, F.G., & Herrera-Viedma, E. (2014). Big data in the European Union. *The Bridge*, 44, 33–37

Gantz, J., & Reinsel, D. (2012). Big data, bigger digital shadows, and biggest growth in the far east. *An ICD report.* Retrieved from www.emc.com/leadership/digital-universe/index.htm

Gomes, L.F.A.M. (2014). Snapshot of big data trends in Latin America. *The Bridge*, 44, 46–49

He, J., Liu, X., Huang, G., Blumenstein, M., & Leung, C. (2014). Current and future development of big data in Commonwealth countries. *The Bridge*, 44, 38–45

Laney, D. (2012). *The importance of "big data": A definition.* (Report, no number). (No location): Gartner Co

Laudon, K.C., & Laudon, J.P. (2012). *Management information systems.* Upper Saddle River, NJ: Pearson Education, Inc

Lee, J., Shi, Y., Wang, F., Lee, H., & Kim, H. (2015). Advertisement clicking prediction by using multiple criteria mathematical programming. *World Wide Web Journal*, (forthcoming)

Li, J., Zhang, Y., Wu, D., & Zhang, W. (2014). Impacts of big data in the Chinese financial industry. *The Bridge*, 44, 20–26

NSF. (2012). *Core techniques and technologies for advancing big data science & engineering (BIGDATA).* National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2012 /nsf12499/nsf12499.htm

Olson, D., & Shi, Y. (2007). *Introduction to Business Data Mining.* Boston: McGraw-Hill

Ouyang, Z.B., & Shi, Y. (2011). 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*, Volume 03, 233–236

Price, R. (1783). Observations on reversionary payments: on schemes for providing annuities for widows, and for persons in old age: on the method of calculating the values of assurances on lives: and on the national debt (Vols. 1–2). (4th ed.). London: T Cadell

Shi, Y. (2014a). A global view of big data. *The Bridge*, 44, 4–5

Shi, Y. (2014b). Big data: History, current status, and challenges going forward. *The Bridge*, 44, 6–11

Shi, Y., Tain, Y., Kou, G., Peng, Y., & Li, J. (2011). *Optimization based Data Mining: Theory and Applications.* New York: Springer

Shi, Y., Xu, W., & Chen, Z. (2005). *Chinese Academy of Sciences symposium on data mining and knowledge management (CASDMKM 2004), LNAI 3327.* New York: Springer-Verlag

Shi, Y., Zhang, L., Tain, Y., & Li, X. (2015). *Intelligent Knowledge: A Study beyond Data Mining.* New York: Springer

Tien, J. (2014). Overview of big data: A US perspective. *The Bridge*, 44, 12–19

Tsumoto, S. (2014). Big data education and research in Japan. *The Bridge*, 44, 27–32

Villanova University. (2014). What is big data? Retrieved from http://

www.villanovau.com/university-online-programs/what-is-big-data/

Watson, B. (1964). *Chuang Tzu*: *Basic Writings*. New York: Columbia University Press

Xu, Z., & Shi, Y. (2015). Exploring big data analysis: Fundamental scientific problems. *Annals of Data Science*, (forthcoming)

Zhang, L., Li, J., Shi, Y., & Liu, X. (2009). Foundations of intelligent knowledge management. *Journal of Human Systems Management*, 28, 145–161