Information Sciences 367–368 (2016) 747–765

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Towards felicitous decision making: An overview on challenges and trends of Big Data

Hai Wang[a], Zeshui Xu[a,b,*], Hamido Fujita[c], Shousheng Liu[d]

[a] School of Economics and Management, Southeast University, Nanjing, Jiangsu 211189, China
[b] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[c] Faculty of Software and Information Science, Iwate Prefectural University, 020-0193 Iwate, Japan
[d] College of Sciences, PLA University of Science and Technology, Nanjing, Jiangsu 210007, China

ARTICLE INFO

Article history:
Received 11 January 2016
Revised 5 May 2016
Accepted 2 July 2016
Available online 5 July 2016

Keyword:
Big Data
Data deluge
Decision making
Data analysis
Data-intensive applications
Computational social science

ABSTRACT

The era of Big Data has arrived along with large volume, complex and growing data generated by many distinct sources. Nowadays, nearly every aspect of the modern society is impacted by Big Data, involving medical, health care, business, management and government. It has been receiving growing attention of researches from many disciplines including natural sciences, life sciences, engineering and even art & humanities. It also leads to new research paradigms and ways of thinking on the path of development. Lots of developed and under-developing tools improve our ability to make more felicitous decisions than what we have made ever before. This paper presents an overview on Big Data including four issues, namely: (i) concepts, characteristics and processing paradigms of Big Data; (ii) the state-of-the-art techniques for decision making in Big Data; (iii) felicitous decision making applications of Big Data in social science; and (iv) the current challenges of Big Data as well as possible future directions.

1. Introduction

Nowadays, an exponential growth of data may come from every imaginable source such as sensors, purchase transactions and social media networks. The speed of data growth has already exceeded Moore's law [30]. Everyday 2.5 quintillion bytes of data are created in 2011 [56]. IBM has reported that 90% of the data created in the world has been produced in the last two years.[1] More than 267 million transactions are produced per day over 6 thousands of stores of Wal-Mart. Till 2011, almost 3 terabytes of data are collected by the US Library of Congress. Over 30 trillion bytes of image data are recorded by the Large Synoptic Survey Telescope in a single day [30]. Especially in China, Baidu conducts dozens of petabytes of data caused by users' queries everyday; Alibaba generates almost 20 terabytes of data by over 880 million online transactions. Fig. 1 shows the prediction results of global data volume provided by International Data Corporation (IDC) [132]. We can conclude without doubt that the era of Big Data has arrived [101].

Besides the huge volume, Big Data also refer to the complex structure of data, the complexity of capturing and managing data [26]. Since it was introduced, Big Data have become one of the most popular issues in both scientific and engineering area. The recent upsurge began with the special issue entitled "Big Data" published by Nature [58]. After the Big Data

* Corresponding author.
E-mail addresses: wanghai17@sina.com (H. Wang), xuzeshui@263.com (Z. Xu), HFujita-799@acm.org (H. Fujita), ssliunuaa@sina.com (S. Liu).

[1] http://www-01.ibm.com/software/data/bigdata/.

http://dx.doi.org/10.1016/j.ins.2016.07.007
0020-0255/

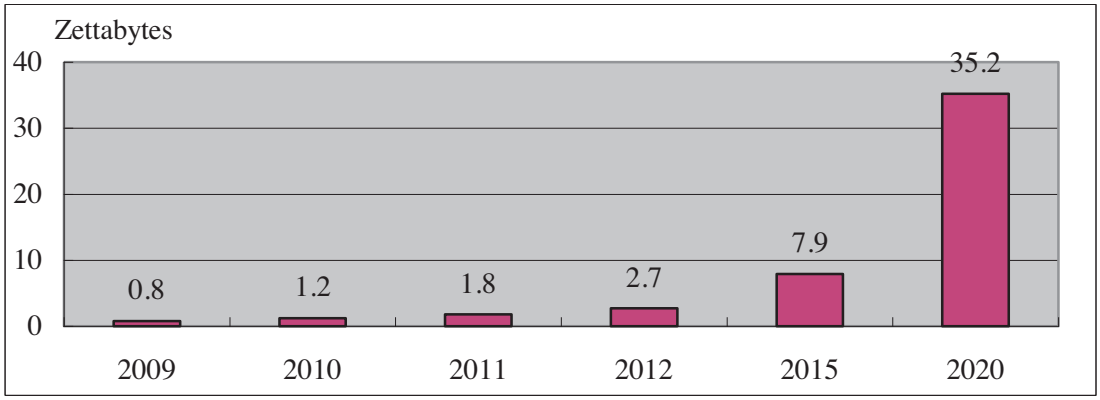

Fig. 1. Global data volume predicted by IDC.

initiative presented by Obama Administration in 2012, Gartner listed Big Data in both the "Top 10 Strategic Technology Trends for 2013" and "Top 10 Critical Tech Trends for the Next Five Years" . Many other projects and solicitations, such as the US National Science Foundation and National Science Foundation of China, have announced to investigate and tackle the challenges of Big Data.

Big Data bring big value. The value takes the form of a value chain and is created through the processes of data discovery, integration and exploitation [102]. Regardless of the specific challenges, technologies and tools have been developed to support decision making in each phase of processing and applying Big Data. Till now, Big Data have been playing a central role in many decision making and forecasting domains such as business analysis, product development, loyalty, healthcare, clinicians, tourism marketing, transportation, etc. For example, as reported by McKinsey institute [96], over 50% of 560 enterprises insist that Big Data can help them in increasing operational efficiency, selecting informing strategic direction, supplying better customer service and so on. The use of knowledge exploited from Wal-Mart's large volume of transaction data has significantly benefitted its pricing strategies and advertising campaigns [30]. Moreover, benefitting from processing of Big Data, over 300 billion dollars and 250 billion euros potential annual values are produced to US health care and European public administration, respectively [96]. It is obvious that Big Data can support intelligent and felicitous decisions for organizations.

Big Data need decision support. Generally, decision science (or theory of choice) in economics, computer science, statistics and mathematics is referred to as identifying the values, uncertainties, rationalities, resultant optimal decision and other relative issues. Normative decision theory focuses on finding methodologies, technologies and tools (software) to identify the best decision to make based on the assumption that the decision maker is fully rational or bounded rational. Under this perspective, decision making, in general, exists in each procedure of Big Data, such as data acquisition/storage, data cleaning/integration, data analysis, data visualization and predicting by derived knowledge. Although it is far from achieving prefect solutions in each procedure, there are several useful techniques and technologies that have been applied for decision making in Big Data. For example, some decision making techniques involved with multiple disciplines are optimization methods, statistics, data mining, machine learning, visualization approaches and social network analysis. In addition, popular Big Data tools include three categories, which are batch processing, stream processing and hybrid processing tools [30]. Based on the paradigm of producing Big Data, the relationship between decision science and Big Data can be explained by Fig. 2. For one thing, Decision theory supports decisions in each phase of processing Big Data; for another, the solutions of Big Data enrich the content and scope of decision sciences. In this sense, one can make more intelligent and felicitous decisions by utilizing better prediction. For instance, we can analyze the preferences of consumers' purchase and the correlation of two classes of products so that more efficient sales promotion can be designed; we can mine the social community of users so that targeted advertising would be more accurate; we can analyze the mood and sentiment of user so that public opinions, even criminal activities, can be predicted; we can also forecast the trends of epidemics so that reasonable emergency plans can be prepared; we can also predict the long-term and/or short term traffic flow to shorten the averaging driving and waiting time [136].

In order to figure out the existing developments, trends and challenges of both decision supporting technologies of processing Big Data and decision making based on Big Data, this paper presents an overview of both aspects. It is, of course, difficult to separate processing Big Data from Big Data applications. We try to do this in this paper, in despite of some inevitable overlaps, to exploit how decision sciences support the development of handling Big Data and how much felicitous DM is essential to be provided in the context of Big Data. Roughly, the first aspect focuses on the emerging techniques and technologies that are elaborately designed for processing Big Data based on decision sciences, whereas the second aspect concentrates on specific applications which process special data sets to support decision making in specific fields such as business and management.

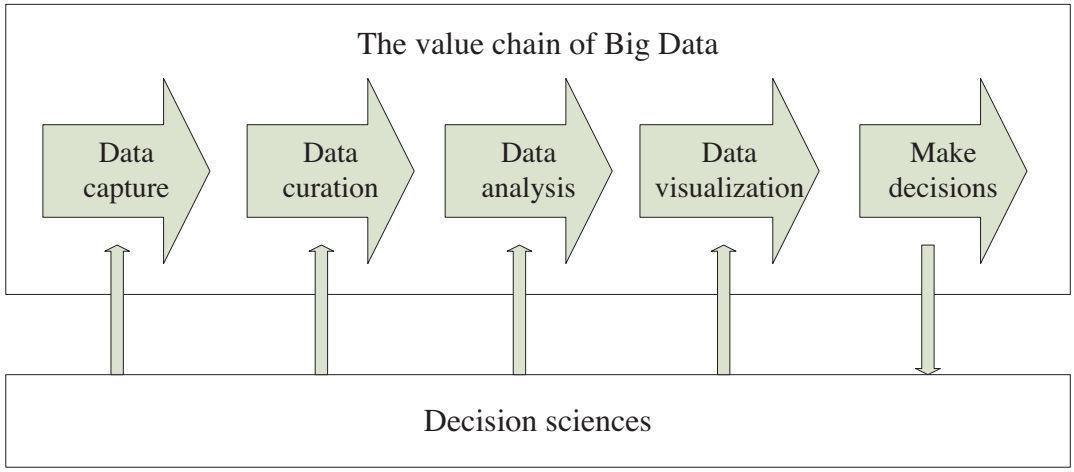

Fig. 2. The relationship between Big Data and decision sciences.

The rest of the paper is organized as follows: Section 2 reviews some basic aspects of Big Data, such as concepts, characteristics, paradigms and related contributions. The existing decision making techniques for processing Big Data are summarized in Section 3. Then a brief review of Big Data applications in social science are presented in Section 4. Challenges and possible directions are depicted in Section 5 and some conclusions are drawn in Section 6.

2. A bird's eye on Big Data

Before embarking on the discussion of decision making in Big Data, we need to specify the scopes, concepts and characteristics of Big Data in this section.

2.1. Concepts and characteristics of Big Data

Many different definitions concerning what constitutes Big Data exist based on distinct perspectives, such as the product-oriented perspective, the process-oriented perspective, the cognition-oriented perspective and the social movement perspective [41].

The product-oriented perspective emphasizes the attributes of data regarding to their sizes, speeds and structures. The motivation of this perspective is, based on a historical view, to compare the quantities of data presented with the volume in the recent past. In this sense, a National Science Foundation solicitation [64] refers to Big Data as: *Large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, emails, videos, click streams, and/or all other digital sources available today and in the future.* In addition, bigness is not just about the size. Gobble [47] states that *data is big because there is too much of it, because it is moving too fast, or because it is not structured in a usable way.*

The process-oriented perspective highlights the novelty of processes required and involved in storing, managing, aggregating, searching and analyzing Big Data. In order to underscore the challenges of processing Big Data, the requisite technological infrastructure, especially the technical tools, programming techniques, computational, statistical and technical advances, is also emphasized in this perspective. Therefore, Kraska [75] defined Big Data as: *When the normal application of current technology does not enable users to obtain timely, cost-effective and quality answers to data-driven questions.* Similarly, Jacobs [67] referred to Big Data as: *Data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time.*

The cognition-based perspective focuses on the challenges caused by Big Data with respect to their cognitive capacities and limitations. One most frequently cited definition fallen in this category is [2]: *Big Data exceed the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population.* A seminal report of McKinsey Global Institute emphasizes the similar thing [96]: *Data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.* Therefore, this perspective conceptualizes Big Data by the fact of exceeding human ability to comprehend which results to the necessity of mediation to enhance interpretability.

Finally, the social movement perspective draws our attention to the gap between vision and reality, especially the socioeconomic, cultural and political shifts that underlie the presence of Big Data [41]. This perspective emphasizes the vision of Big Data for scientific discovery, environmental and biomedical research, education and national security [109] and refers to Big Data as: *The potential to quantify and change various aspects of contemporary lift, to revolutionize the art of management [61], or to be part of a major transformation that requires national effort.*

Table 1
Four scientific research paradigms.

Paradigm	Time	Methodology
Experimental science	Thousand years ago	Description of natural phenomena
Theoretical science	Last few hundred years	Newton's laws, Maxwell's equations, etc.
Computational science	Last few decades	Simulation of complex phenomena
Data-intensive science	Today	Data are captured and generated by many different sources; eScience is used to support data collaboration, analysis, visualization, exploration

The four perspectives provide useful insights for conceptualizing Big Data and then several characteristics can be summarized. Three terms, which are volume, velocity and variety, are commonly accepted and discussed [80]:

Volume focuses on the size of the data set. The volume of Big Data may reach the level of terabytes or even petabytes, which is far beyond the conventional limits of megabytes or gigabytes.

Velocity indicates the speed of data in and out. It refers to the frequency of the data generation, the dynamic feature of the data, and the necessity of generating results in real-time.

Variety describes the range of data types and sources, which highlights the different sources of information and the distinct data schemas of each source. For example, traffic dataset includes numeric information about vehicular traffic on roads, textual information about active events and scheduled events (e.g., sporting events, music events) [147].

Besides, other Vs, such as value, veracity, variability and virtual, are also mentioned in literature to serve as complement features of Big Data [13,59,154]:

Value refers to the monetary worth that an organization can derive from processing Big Data. It includes two aspects: the big potential value and the extremely low density of value.

Veracity corresponds to what extent the data can be trusted, given the reliability of its source. For instance, when receiving data from sensors, some devices may even be compromised.

Rather than representation by several Vs, Wu et al. [146] also highlighted the characteristics of Big Data by the so-called HACE theorem: *Big Data start with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seek to explore complex and evolving relationships among data.*

2.2. The paradigms of Big Data

The characteristics of Big Data have resulted to a novel paradigm of scientific research, i.e., the fourth paradigm [54]. Table 1 shows a historical view of the four science paradigms.

This novel paradigm has led to a data-intensive science. The central issue of data-intensive science has moved from computation to data and results to thinking with data. Three basic activities are usually considered in data-intensive science, which are capture, curation, and analysis, according to Jim Gray [54]. But the purpose of processing Big Data is to exploit knowledge from data to support intelligent decision. It is necessary to underline data visualization/interpretation at the same time. As suggested by Chen and Zhang [30], data visualization/interpretation is apart from the process of analysis. Thus, from the perspective of decision making, the process of data-intensive science can be interpreted by Fig. 3.

Data may come in all scales and shapes of sources (even including individuals' lives). Curation covers a wide range of procedures, including data cleaning, integration and representation. It contains schemas and necessary metadata for longevity and integration. Data analysis includes the whole range of activities throughout the workflow pipeline, including using the databases, analysis and modeling. When designing a database for a given discipline, Jim Gray [54] argued that it should be able to answer the key twenty questions that are concerned by scientists.

When social science meets Big Data, a novel discipline named computational social science (or e-social science) emerges. Generally, computational social science involves many inter-disciplines which analyze and use data with an unprecedented breath, depth and scale with the aid of Big Data techniques and technologies [29]. For example, the researchers' perspective of computational organization science has been broadened on social, organizational and policy systems, leveraging computational models that combine social science, computer science and network science. Similarly, in e-business, new approaches have been designed for the research context such as human and managerial decision making, consumer behavior, operational processes and market interactions based on more advantageous costs of data collection in the context of social networks, blogs, mobile telephony and digital entertainment and the new capabilities that are hard to implement before. Based on that, Chang et al. [29] stated that a paradigm shift in scientific research is coming now, speeded by new business practices and organizational environments. The data spectrum available in this new paradigm, which spans the macro-level, meso-level to the micro-level, brings new advanced technologies to support high value decisions.

As shown in Fig. 2, decision making plays an important role in processing Big Data. On the one hand, decision science supports decisions in the procedures of analysis. On the other hand, the overarching purpose and reason of Big Data is about decision making. The latter aspect can result to intelligent decisions based on raw data. The decision making paradigm of this aspect is summarized in Fig. 4 [132], in which decisions are made by deriving information from data, obtaining knowl-

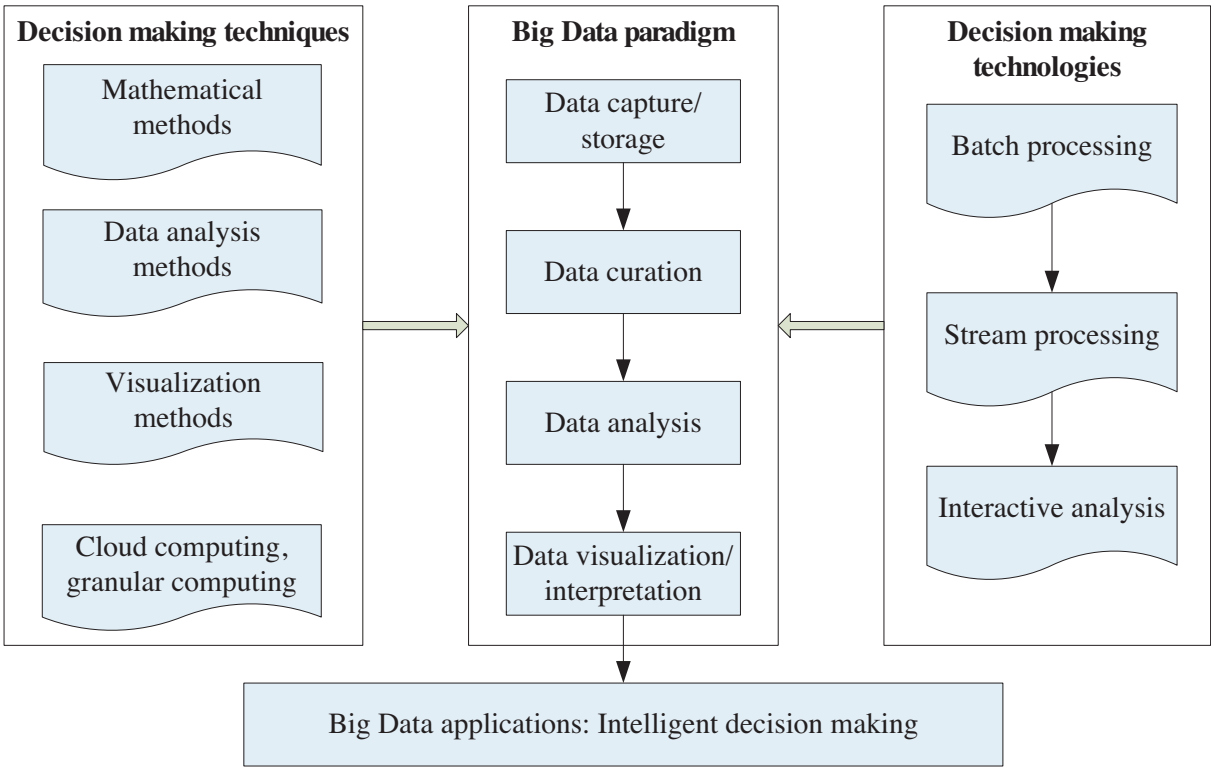

Fig. 3. Paradigms of Big Data processing.

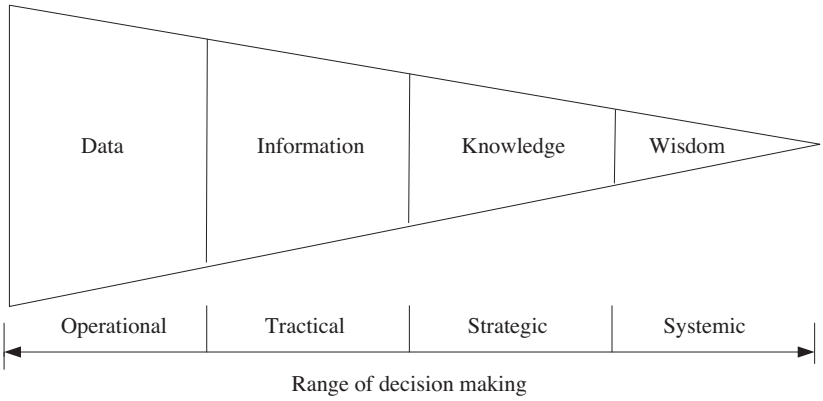

Fig. 4. The framework of Big Data decision making.

edge from information and then achieving wisdom from knowledge. Thus, Big Data have become a competitive advantage although there remains a mass of technical problems. In practice, there are two natural strategies to analyze data, i.e., the scientific strategy and the engineering strategy. The scientific strategy investigates natural phenomena, acquires new knowledge, integrates and/or corrects the existing knowledge and interprets the laws of nature from the obtained multiple sources of data. The engineering strategy, also called decision informatics, pays more attention on the requirement of real-time decision making in the presence of Big Data. It is supported by information technologies and decision science, and underpinned by data fusion/analysis, decision modeling and systems engineering.

2.3. A bibliometric perspective of Big Data

Since the first special issue named "Big Data" was presented by Nature in 2008, Big Data have attracted many research activities and become one of the hottest research topics. Several special issues have been held, such as "Dealing with data" proposed by Science in 2011 [43], "Big Data" proposed by NEC Technical Journal named in 2012 [129], "Big Data" proposed by

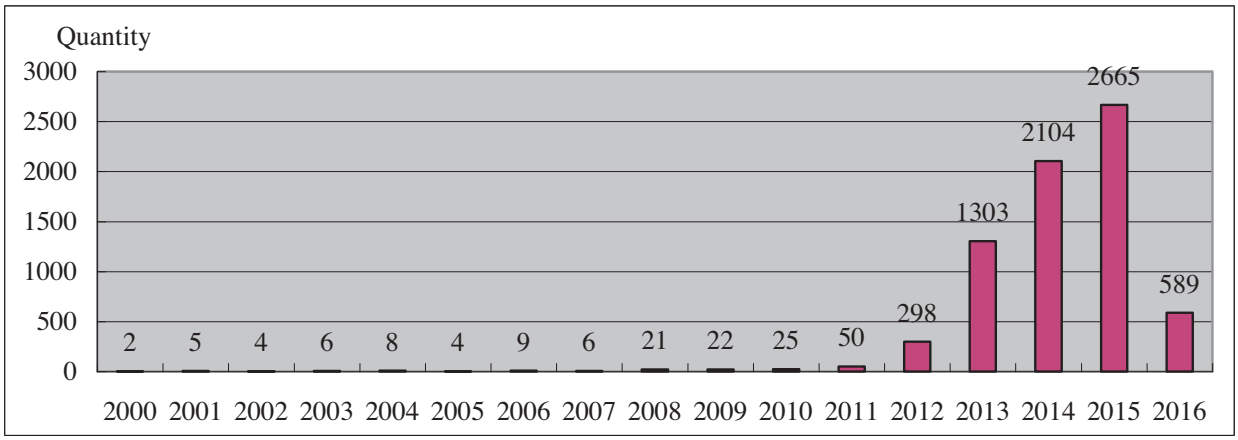

Fig. 5. Numbers of publications in Web of Science.

Table 2
Journals publishing the top 10 quantities of articles of Big Data.

Ranking	Journal	Quantity
1	Plus One	48
2	Big Data	45
3	Neurocomputing	33
4	Future Generation Computer Systems The International Journal of Grid Computing and eScience	31
5	Cluster Computing The Journal of Networks Software Tools and Applications	28
6	IEEE Network	26
6	International Journal of Distributed Sensor Networks	26
7	BMC Bioinformatics	25
8	IEEE Access	24
9	Computer	23
10	Information Sciences	22

Table 3
Top 10 research areas of the existing Big Data contributions.

Ranking	Research area	Quantity
1	Computer science	3678
2	Engineering	1680
3	Telecommunications	486
4	Business economics	349
5	Science technology other topics	229
6	Information science library science	205
7	Mathematics	177
8	Health care sciences services	174
9	Physics	147
10	Biochemistry molecular biology	146

Entropy in 2014 [48], “Special Issue on Cloud Services Meet Big Data” proposed by IEEE Transactions on Services Computing in 2015 [1], “Big Data Meets Multimedia Analytics” proposed by Signal Processing in 2016 [32] and so on. According to Web of Science, there are totally 7121 publications from 2000 to April 30, 2016, as shown in Fig. 5, if we search the database by the term of “Big Data” and “data deluge” . Among which, there are 2924 articles, 3144 proceeding papers and 258 reviews. The 2924 articles are published in more than 1000 distinct journals.

Table 2 shows the journals which have published the top 10 quantities of contributions among totally 2924 articles. Note that journals that usually publish proceeding papers are not included in this table. Moreover, disciplines involved in these contributions cover computer science, engineering, economics and so on, as shown in Table 3. Most of the contributions focus on the developments of technologies of Big Data. At the same time, the applications in economics, health care and medicine are also popular issues. Till now, these contributions have won 16,704 citations (including 3896 self-citations). The top 10 cited publications are summarized in Table 4.

Table 4
Top 10 cited papers among 7121 publications.

Year	Title	Journal /Proceedings	Research area	Number of citations	Number of citations per year
2003	DnaSP, DNA polymorphism analyses by the coalescent and other methods [116]	Bioinformatics	Biochemistry and molecular biology	4157	296.93
2014	Mass-spectrometry-based draft of the human proteome [143]	Nature	Science and technology – other topics	289	96.33
2011	A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry [39]	American Journal Of Psychiatry	Psychiatry	279	46.50
2008	Big data: The future of biocuration [58]	Nature	Science and technology – other topics	224	24.89
2012	Critical questions for big data provocations for a cultural, technological, and scholarly phenomenon [22]	Information Communication & Society	Communication; sociology	189	37.80
2012	Business intelligence and analytics: From big data to big impact [31]	MIS Quarterly	Computer science; information science; business and economics	177	35.40
2010	Computational solutions to large-scale data management and analysis [120]	Nature Reviews Genetics	Genetics and heredity	168	24.00
2014	The parable of Google flu: Traps in big data analysis [81]	Science	Science and technology – other topics	154	51.33
2009	Beyond the data deluge [14]	Science	Science and technology – other topics	120	15.00
2008	Big data: How do your data grow? [92]	Nature	Science and technology – other topics	114	12.67

3. Decision making tools for processing Big Data

Scientists have developed a wide variety of tools to capture the value of Big Data along with the value chain. Although it is far from meeting various needs, the existing decision making tools which cross multiple disciplines, have been applied to many data-intensive applications and shown their excited effectiveness for capturing, curating, analyzing and visualizing Big Data. In this section, we review some developments and current trends of decision making techniques and technologies in this area following the paradigm described in Fig. 3.

3.1. Decision making techniques

Because of the necessity of processing huge volume of data within acceptable times, the decision making techniques for Big Data are developed and under developing to facilitate decision making in each phase of processing. These techniques are usually driven by specific requirements of applications. There are a number of techniques involving many disciplines and usually overlapping with each other, which can be roughly classified by several categories: mathematical tools, data analysis tools, visualization tools and other higher-level techniques.

3.1.1. Mathematical techniques

Statistics and optimization methods usually emerge in studies of Big Data, based on specific fundamental mathematics. Statistics usually support decision in the phase data curation and analysis. The core issues of decision making by statistical techniques are the exploitation of causal relationships and co-relationships among objectives and the derivation of numerical descriptions of samples. To overcome the limitation of traditional techniques, some new fields have been developed to suit the necessity of managing huge volume of data, such as parallel statistics [94], statistical computing [144] and statistical learning [51]. Especially, the parallel implementation of statistical techniques is intuitively reasonable. Pébay et al. [110] showed that moment-based statistics, such as principal component analysis, scale nearly linearly with the size of data set and number of processes, whereas entropy-based statistics (including order and contingency statistics) only scale well if the data are discrete. Some other extensions or new statistical techniques can be found in Sysoev et al. [128]. A famous example of this field is that Wal-Mart supports its decision involving pricing strategy and advertising campaigns by exploiting patterns from transaction data using statistical techniques, associated with machine learning. The mentioned techniques improve the ability of processing huge volume of data.

Optimization methods have also been clarified as efficient techniques due to their nature of quantitative implementation. In order to solve their usual high costs in memory and time consumption, they are married with data reduction [148] or parallelism [117] naturally. Their application in large-scale optimization can be found in several studies based on cooperative co-evolutionary algorithms [86]. Real-time optimization [122] is another hot topic of this field, whose capability has

been demonstrated by decision making problems in large-scale wireless sensor networks [124] and intelligent transportation systems [150].

3.1.2. Data analysis techniques

Data analysis techniques include a wide range of disciplines such as data mining, machine learning, artificial neural networks and signal processing, most of which have shown their capabilities in processing Big Data.

Data mining, including classification, regression and clustering, is the collection of artificial intelligent techniques that mine hiding knowledge and patterns from given data. To mining valuable information from Big Data, it is natural to improve the existing algorithms. Parallel implementation [52] and dimensionality reduction [137] are also natural to be considered. Several classification and clustering algorithms have been proposed for large-scale samples, such as linguistic fuzzy rule based classification systems, clustering large application algorithm and so on. Because of the complex uncertainties of Big Data, more and more theoretical tools, such as fuzzy reasoning, are taken for modeling uncertainties of both raw data and outputs of algorithms.

Machine learning, including supervised learning and unsupervised learning, is another main field of artificial intelligence. It can derive valuable knowledge automatically once a designed algorithm learns behaviors from empirical data. Traditional machine learning algorithms, such as support vector machine, have improved their effectiveness based on large-scale parallel framework like Map/Reduce. Scale machine learning algorithms, for example, the boosting based sparse approximation algorithm proposed by Sun and Yao [127], are presented to suit large-scale problems. Especially, multi-classification is another hot spot in machine learning which has been used in various applications including big biological data [142] and sensors data [139]. Multi-classification on multi-dimensional data, such as the stocks market prediction [153], becomes one challenge in Big Data analysis for prediction using the extended techniques of multi-classification in single data stream.

Artificial neural network has been successfully employed for pattern recognition, adaptive control and so on. However, there is a big challenge if it is employed for analyzing Big Data because of the natural contradiction between the necessity of more hidden layers and nodes for higher performance and the memory and time consuming in a neural network. Two approaches have been adopted to ease the contradiction. The first one is to reduce the sizes of data by sampling techniques, and the second one is to put neural networks in a parallel and distributed setting [106].

Opinion mining (OM), or sentiment analysis, implemented by specific machine learning techniques and lexicon-based techniques, plays an important role of processing Big Data that include (word of mouth) texts [115]. This ongoing field has been discussed by a large number of contributions because it enables the users to understand sentiment polarities from huge volume of texts. Regardless of the preprocessing, the goals of OM in Big Data may be quite different due to the aims of applications, including subjectivity classification [24], polarity determination [7], opinion spam detection [55], review usefulness measurement [108], aspect extraction [42], Lexica and corpora creation [141] and so on. The granularity of OM varies from document level, sentence level, and word level to concept level and clause level. To fit for the huge volume of Big Data and scale up OM, these algorithms can be implemented based on some Big Data platforms like MapReduce and Storm [3].

Feature extraction (FE) is a kind of common techniques for reducing high-dimensionality in machine learning problems. According to the report of UCI Machine Learning Repository, the maximum dimensionality of data was about 100 in 1980s, but had increased to more than 3 million by 2009 [21]. In the context of Big Data, new methods for FE are constantly being developed. Filter methods are the most frequently used. Embedded methods have also increased in popularity recently because they enable FE and classification simultaneously [95]. The combination of algorithms, in the form of either hybrid methods [135] or ensemble method [20], is also a popular issue. In addition, multi-classification is another hot spot in machine learning which has been used in various applications including big biological data [142] and sensors data [139]. some specific FE techniques have been developed for multi-classification, where most of them use or extend state-of-the-art FE techniques mentioned above, and others present some new idea, such as the multi-class centroid FE method [142].

As a kind of representation learning, deep learning techniques compose simple but non-linear modules into a representation at a higher and more abstract level. These techniques can start with the raw inputs, comparing with the traditional machine learning algorithms [82]. To meet the requirement of analyzing Big Data, specific deep learning techniques have been developed by many researchers [6,16]. Also, to improve the performances of artificial neural networks in processing Big Data, Schmidhuber came up with deep artificial neural networks [121]. Till now, it is interesting that deep learning techniques have played important role in many Big Data application including drug discovery [45], genomic medicine [83] and text mining [62].

3.1.3. Visualization techniques

Visualization techniques interpret data by intuitive display ways like tables, images and diagrams. For instance, the Facebook Timeline is a visualization tool for manipulating and organizing the data in its database. Big Data visualization would make data meaningful but is more challenging than the traditional case due to the complexity in data [126]. Most of the existing extensions focus only on the huge volume of Big Data, especially large-scale data, try to find the proper data representation [131], reduce the sizes of data by feature extraction, or run batch-model software rendering in a parallel way [4]. For example, Tompson et al. [131] introduced a data representation for scalar data and then formed a compact yet information rich approximation of large-scale data.

Although the existing tools are far from enough, powerful information visualization tools are realizing famed statistician's a 50-year-old prediction: "The graphical potentialities of the computer are going to be the data analyst's greatest single resource" [134]. It can be expected that more effective approaches will be able to put human users in control because they can frequently identify patterns that machines cannot. Scientists have been increasingly realizing that visual strategies for exploiting Big Data would lead to more potent and meaningful insights [125]. As stated by Gan [44], innovations in data visualization demonstrate that a good user interface is worth a thousand petabytes.

3.1.4. Cloud computing

Cloud computing enables organizations to pay their attention only on the resources and services they use, revolutionizes the way that information science is consumed. Clouds provide platform, infrastructure and even decision supporting software as services though they often vary significantly. When it comes to Big Data, the model based on cloud computing is exciting. Solutions of analyzing data are hosted on the cloud and consumed by users in a pay-as-you-go fashion. The important issues of cloud based decision supporting are data management, tuning of models, data quality and data currency [8]. To consider the multiple cloud deployment models to prepare data, clouds for different organizations might be private, public or hybrid. Compared to other mentioned techniques, cloud computing seems to be the one customized for processing Big Data, and this technique are of few disadvantages. Moreover, by putting analytics and Big Data in cloud, Demirkan and Delen [37] stated that data, information and analytics all can be defined as services in a service-oriented decision support system.

3.1.5. Fuzzy sets and systems

As an important component of granular computing (GrC), techniques related to fuzzy sets and systems have been popular and effective solutions for some Big Data challenges. We prefer to list them as an individual kind of techniques because they are helpful for decision making in many issues of Big Data.

The use of fuzzy sets and fuzzy logic can model vagueness and uncertainties which are natural for Big Data. For knowledge extraction, concepts, ontologies, connections and communities may be fuzzy and thus can be modeled by fuzzy sets [105]. These techniques become more efficient when extracting knowledge from incomplete Big Data [88]. Based on the architecture of storage, parallel sampling is an important issue. He et al. [52] developed a fuzzy sets-based approach for this issue with uncertain distribution. Uncertainties can also be handled by defining distinct knowledge granular [77]. Besides, the traditional fuzzy sets are usually applied associated with other similar techniques. For instance, fuzzy rough sets have been used in knowledge extraction and feature selection [78]. Type-1 and type-2 fuzzy systems may be interesting for deep learning in Big Data application as well [71].

This kind of techniques can also be used to improve the traditional data analysis techniques in the presence of Big Data. For example, evolving fuzzy system [60], neural fuzzy classifier [28], linguistic fuzzy rule-based classifier [79] have been developed for classification of Big Data; while fuzzy c-means algorithm might be the most popular one for clustering of Big Data [91]. Other pattern recognition algorithms, such as fuzzy inference systems, fuzzy Bayesian process [114], fuzzy query system [90], have been investigated for distinct Big Data applications. Some algorithms are usually presented along with specific ensemble learning techniques. Neural fuzzy classifiers are also served as a solution of dimensionality reduction [10]. Moreover, the proper use of fuzzy sets can speed up Big Data analyzing process by common data analysis techniques [100].

3.2. Decision making technologies

Except for the techniques mentioned in the above subsection, new technologies, such as platforms and infrastructures, are required for the problems arisen by characteristics of Big Data. A historical perspective of the frameworks of the technologies is shown in Fig. 6 [26]. The milestone of batch processing is the MapReduce framework [36] proposed by Google in 2003. However, organizations only focus on large volume of data rather than Big Data in that age. When another tool, Hadoop, was born in 2006, Big Data processing comes into the first generation. As there are no new developments for batch processing, the first generation has been closed. Stream processing is regarded as the second generation whose milestone is the S4 developed by Yahoo! in 2010. Both big static data and big streaming data are dealt with in the second generation. The hybrid processing enables the possibility of coming into the third generation. Although there are some promising developments of hybrid processing, it is not enough to say that we have arrived in the third generation. Table 5 shows a list of the three generations of technologies classified by specific processing paradigms. Comprehensive reviews of these technologies can be found in Refs. [26,30].

Serving as a solution of processing large volume of static data, batch processing only takes into account the data that are already in the data storages. The main advantages of batch processing are scalability and reliability. To achieve scalability, a parallel distributed framework like the infrastructure implemented in MapReduce is usually considered. However, batch processing technologies usually suffer from high throughout latency in its implementations and thus do not meet the requirement of real-time response for processing large amount of stream data such as log files. Some existing tools and platforms of batch processing are listed in Table 5.

To handle stream Big Data with high volume, high velocity and complex data types, the stream processing (namely real-time processing) paradigm is developed for real-time analytics. Big Data are generally stored in a distributed environment. Therefore, stream processing is conducted based on distribution and parallelism as well, like those of batch processing. As

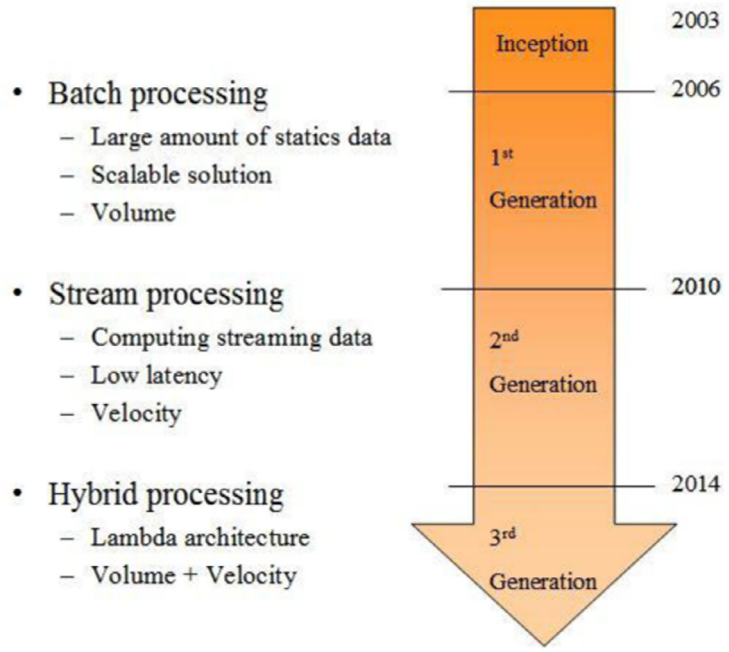

Fig. 6. The three generations of processing paradigms.

Table 5
Big Data technologies classified by distinct processing paradigms.

Paradigm	Technology	Reference/URL
Batch processing	MapReduce	[36]
	Hadoop (HDFS, Hive, HBase)	http://hadoop.apache.org
	Flume	http://flume.apache.org
	Scribe	https://github.com/facebook/scribe
	Dryad	[65]
	Apache Mahout	[63]
	Jaspersoft BI Suite	[140]
	Pentaho	http://www.pentaho.com/explore/pentaho-business-analytics
	Skytree Server	[119]
	Cascading	http://www.cascading.org
	Spark	http://spark.incubator.apache.org
	Tableau	[25]
	Karmasphere	http://www.karmasphere.com
	Pig	http://pig.apache.org
	Sqoop	http://sqoop.apache.org
Stream processing	Kafka	https://kafka.apache.org
	Flume	http://flume.apache.org
	Kestrel	https://github.com/twitter/kestrel
	Strom	http://storm-project.net/
	S4	[107]
	SQLstream	http://www.sqlstream.com/products/server
	Splunk	[118]
	SAP Hana	[74]
	Spark Streaming	http://spark.incubator.apache.org
Hybrid processing	Lambdoop	http://www.lambdoop.com
	SummingBird	http://www.infoq.com/news/2014/01/twitter-summingbird

there are not too much data at the time dimension, this paradigm uses the diskless approach to achieve low latency. That is, stream processing can be seen as an infinite sequence of small batch processing that processes small sets of data stored in memory. Famous existing tools and platforms of stream processing can be found in Table 5.

In practice, hybrid processing is necessary for many domains. This paradigm synthesizes both batch processing and stream processing paradigms based on the Lambda Architecture [98]. A high-level architecture of this paradigm contains three layers. Batch layer manages the master dataset that has been stored in a distributed system and is not changeable. Serving layer load and exposes the views of batch layer in a data store for query, and speed layer only deals with new data

Table 6
Recent contributions related to social Big Data.

Contribution	Data source	Technique and/or technologies	Focus
Trattner and Kappe [133]	Facebook	Facebook tools and applications; real-time measures	Targeted advertising on Facebook
Jansen et al. [68]	Twitter	Summarize tool; classification; statistics	Twitter as eWOM advertising mechanism
Asur et al. [9]	Twitter	Classification; statistics	Forecast box-office revenues for movies
Ma et al. [93]	Epinions[a]	Social network analysis, information diffusion models	Viral marketing in social networks
Dodds et al. [38]	Twitter	Hedonometrics	Uncover temporal variations in happiness and information levels
Guo and Vargo [49]	Twitter during the 2012 US presidential election	Semantic network analysis	Empirically test "issue ownership network"
Jin et al. [69]	LiveJournal[b], ASKitter[c], Orkut[d], 3 artificial networks	MapReduce, map equation	Mine community structure
Durahim and Coşkun [40]	Twitter	Sentiment analysis model	Calculate the Gross National Happiness of Turkey
Bohlouli et at. [19]	Wikipedia data sources	MapReduce, sentiment analysis, NoSQL databases	Discover knowledge from social media
Wu et al. [53]	Twitter	Apache, Hadoop, MySQL	Glean industry-specific marketing intelligence with sentiment benchmarks
Almaatouq et al. [5]	Twitter	Gaussian mixture models, graph	Observe trends in social microblogging spam
Bettencourt [17]	Data involved with Tokyo	Feedback control theory, self-organization	Better understand and manage cities
Zhang et al. [151]	Online reviews in e-commerce	A novel analytics framework with parallel co-evolution genetic algorithm	Detect deceptive reviews
Ku and Leroy [76]	Crime video	Natural language processing; similarity measure; classification	Analyze crime trends
Phillips and Lee [113]	38 Crime datasets	Statistics; network analysis	Discover geospatial co-distribution relations among crime incidents
Gerber [46]	Twitter	Linguistic analysis; statistical topic modeling	Identify discussion topics to predict crimes
Chainey et al. [27]	Geocoded crime point data	Spatial analysis, mapping methods	Comparative evaluation of mapping techniques to predict crimes
Li et al. [85]	Data of a bank in Taiwan	Neural network, association rules	Identify the signs and patterns of fraudulent accounts
Liao et al. [87]	Youtube commentaries	Epidemic models	Predict negative behavior spreading dynamics
Lin and Utz [89]	Facebook	Multi-method approach	Explore emotional responses of browsing Facebook
Crampton [34]	US intelligent community	Not focused	Identify the impacts on national security

[a] http://www.epinions.com/.
[b] http://snap.stanford.edu/data/com-LiveJournal.html.
[c] http://snap.stanford.edu/data/as-skitter.html.
[d] http://snap.stanford.edu/data/com-Orkut.html.

with low latency. At last, a complete result is merged by the combination of batch and real-time views. Some current hybrid processing technologies are shown in Table 5 as well.

4. Intelligent decision making based on Big Data: the evidence from social Big Data

As have mentioned hereinabove, Big Data can be applied in various disciplines due to their power of felicitous decision making based on large, diverse and complex data. In this section, we only focus on some recent applications in social science, such as marketing, e-commerce, and social management. In this circumstance, Big Data come from multiple social media sources and can be referred to as social Big Data. Applications in other area, such as health care, medical, bioinformatics can be found in some recent reviews such as Refs. [132,138].

Intelligent decision making based on social Big Data includes techniques, technologies, systems and platforms that provide better understanding of data for organizations to support decisions. For example, researchers and managers can derive knowledge from the customers' opinions to realize the market transformation and improve their business strategies; Agencies can identify the features and the patterns of crimes and criminals from environmental and situational factors to support law enforcement; Service providers could visualize social media data to enable better user experience and service [15]. Table 6 summarizes some recent contributions in this field.

Most of the existing contributions focus on the use of social Big Data in e-business and marketing. Trattner and Kappe [133] demonstrated that Big Data generated by social network streams, such as Facebook, can increase the number of visitors and the profit of a web-based platform (called VirWoX) and detect the most valuable users (of Facebook). It is the first contribution that provides the detailed results of social stream marketing campaign. In social network media, word of mouth expressed by users plays a major role for customers' potential buying decisions, thus mining information and knowledge such as comments and sentiments is helpful for enterprises. Jansen et al. [68] investigated microblogging of Twitter as a form of electronic word of mouth for sharing the consumers' sentiments about brands. Technically, automated approaches of sentiment classification and manual coding were compared. Simultaneously, the range, frequency, timing and content of tweets in a corporate account were analyzed. They reported that microblogging plays an important role for costumer to communicate and discuss the implications for corporations online. Bohlouli et at. [19] also considered OM for better understanding of customer feedback so that the next generation of products can be improved. Wu et al. [53] presented another OM model guided by the competitive analytics framework. Almaatouq et al. [5] investigated spam in online social networks through the lens of their behavioral characteristics. This would be helpful for advertisers and potential investors, as well as negatively affecting users' engagement. Asur et al. [9] investigated social Big Data based on Twitter to predict box-office revenues of movies. They stated that social media can be effective to indicate real-world performance. Based on simple statistics model, movie box-office revenue can be predicted very well, and the power of the model can be even stronger if sentiment classification is associated. Ma et al. [93] proposed three models to diffuse both positive and negative comments on products or brands, in order for the selection of the best individuals to receive marketing samples based on heat diffusion processes. The models were claimed to be scalable to large social networks. The community structure mining is sometimes challenging for recommendation systems and network marketing. Jin et al. [69] proposed a distributed mining framework for this challenge based on the map equation of information theory.

Another popular topic of social Big Data is related to decision making in management, including its application to the management problems of society and enterprises. Li et al. [84] summarized some existing applications of Big Data in product lifecycle management and exploited the way of employing Big Data to enhance the intelligent decisions related to design, production and service process. Dodds et al. [38] uncovered and explained temporal variations in happiness and information level by building a tunable, real-time, remote-sensing, noninvasive and text-based hedonometer. Similarly, the gross national happiness of Turkey was investigated by Durahim and Coşkun [40] by adopting a sentiment analysis model. They found strong correlations between the users' happiness levels and Twitter characteristics. Guo and Vargo [49] examined the social Big Data on Twitter during the 2012 US presidential election involving its power of determining public's identification of a political candidate. They showed that new media is more powerful than the traditional news media. The workshop of Bettencourt [17] illustrated the use of social Big Data for effective urban planning. However, when using data from online social networks sites, deceptive reviews may be inevitable. Thus, Zhang et al. [151] presented a novel parallel co-evolution genetic algorithm for adaptive detection of deceptive reviews with respect to different social media contexts.

Besides, social Big Data can be used for social and national securities. For instance, Ku and Leroy [76] developed an intelligent decision support system to automate and facilitate crime analysis based on the combination of natural language processing, similarity measures and classification approaches. Phillips and Lee [113] developed a crime data analysis system which enables discovering co-distribution patterns between large, aggregated and heterogeneous data sources to help the detection involving where, when and why particular crimes are likely to occur. Recently, Gerber's approach [46] showed the proper use of Twitter can result to automatically identify the discussion topics of an area and the effective crime prediction, associated with a linguistic analysis and statistical topic model. Similar studies focusing on crime analysis can be found in Refs. [27,85]. A Big Data case study presented by Crampton [34] highlights two ways of social Big Data can make big impacts on national security: a reconceptualization of geoprivacy and algorithmic security. Liao et al. [87] took use of epidemic models to explain and predict negative behavior that spreads dynamics in online social networks based on the empirical analysis on Youtube commentaries. Lin and Utz [89] explored the emotions of users, such as happiness and envy, of reading a post on Facebook. They demonstrated that the positive emotions are more prevalent than the negative emotions while browsing social media.

It is also interesting that intelligent transportation systems have been developed based on social big data although most of these are based on vehicle trajectories, human mobility, etc. For instance, Bao et al. [12] predicted transportation by the location-based social networks. Zheng et al. [152] reported that this type of developments would be a new and effective path.

Finally, there are some contributions focusing on the challenges and limitations of decision making based on social Big Data. Tan et al. [130] addressed several challenges of leveraging social network paradigm to derive knowledge. Zúñiga [35] summarized pressing issues when employing social Big Data for political communication research. Hargittai [50] stated that potential biases may exist when using Big Data that rely on specific sites and social network platforms. But this argument is contradictory with the opinion of Kimble and Milolidakis [73]. Phillips-Wren et al. [112] presented a Big Data analytics framework, which proposes a process view of components for data analytics in organizations, to increase the relevance of academic research to practice. Cowls and Schroeder [33] provided insight into considerations of causal versus correlational research, the utility of theory as well as the use of inductive methods in the presence of social Big Data.

5. Big challenges and possible directions

Big Data remain big challenges. Till now, it is too early to say that we have reached the standard theory for handling Big Data. Thus, the challenges are usually related to the application fields, including challenges in Big Data management and analysis, semantic challenges and other non-technical challenges [8,97]. In addition, more challenges will arise along with the continuous development of new technologies and techniques. This section summarizes some general challenges of Big Data and figures out some possible alternative solutions.

5.1. Challenges

We first discuss challenges based on the processing paradigms shown in Fig. 3, and then some other challenges such as system challenges and non-technical challenges are involved.

5.1.1. Challenges in data capture/storage and curation

The way that we capture and store data should be significantly changed along with the appearance of Big Data. However, the accessibility of Big Data is restrained by the system imbalance of CPU-heavy but I/O-poor [54]. This limitation should be broken, or partially broken, to ensure easy and prompt assess for further analysis. Some relative technologies, such as solid-state drive, phase-change memory and optimizing data access [66], may be helpful to alleviate this challenge. Moreover, the network bandwidth capacity is also a bottleneck because data are usually designed to store in distributed centers and cloud. In addition, SQL-based database systems are not suitable for Big Data curation any more. Although the NoSQL database technology is under developing, it is far from enough.

Data security is another challenge in these phases. Big Data applications related to sensitive information, such as medical records and banking transactions, may be not suitable for simple data transmissions. The privacy concerns should be resolved before defining the strategy and protocol of information sharing, for instance, designing certification or access control and anonymization. However, the development of secured certification mechanisms is still challenging, while anonymization approaches may lead to more challenges to data analysis because it may increase the uncertainties of data.

5.1.2. Challenges in data analysis and visualization

Challenges involved in data analysis are caused by data complexity and computational complexity. The inherent data complexity of Big Data comes from complex types, complex structures and complex patterns of them as well as complex uncertainties in these aspects. For instance, there is no acknowledged effective and efficient model to handle heterogeneous data types of Big Data; The description of semantic features and the construction of semantic association models in some applications are challenging as well. Traditional data analysis techniques have shown their difficulties (or even disabilities) for handling Big Data. This is mainly because we cannot understand the laws of distribution and association relationship, the inherent relationship of data complexity and computational complexity and the domain-oriented processing methods of Big Data [70]. Thus, we arrive at a great challenge involving how to formulate and depict the complexity of Big Data quantitatively.

Data complexity can also be caused by spare, uncertain, incomplete and dynamic data. In some Big Data applications, the number of samples may be quite few and the dimension of them may be very high. One cannot mine clear trends or distributions for deriving reliable conclusions. The challenge for uncertain data is that the data field may be subjected to some random/error distributions rather than deterministic ones. Most existing techniques cannot be adopted directly. If incomplete data appear in some samples, the existing models which ignore data fields with missing values or predict possible values possess limitations when applied to Big Data. These would become more challenging if various heterogeneous and distributed data sources are involved.

When it comes to the computational complexity of Big Data, computability should be mentioned at first. Because of the key characteristics of Big Data, the traditional computing approaches are not capable for supporting the decision making problems with multi-sources, huge volume and fast-changing datasets. New techniques should be presented to break away from assumptions of the traditional approaches to re-investigate the computability (and then computational complexity) of Big Data. In order to do that, new features of Big Data processing, such as insufficient samples, uncertain data relationships and unbalance (or even uncertain) distributions of value density, should be fully considered.

Scalability and timeliness are two issues with high priorities with regard to Big Data. Although increment algorithms have good scalability, they cannot solve this issue fundamentally. For real-time applications, such as intelligent transport systems and internet of thing, the existing solutions for stream processing paradigm are far from enough. It is sure that this challenge would lead to the swerve of developments of hardware and software to cloud computing. In addition, non-deterministic algorithm theory may be more suitable for Big Data analysis [70].

The challenges of Big Data visualization come from the large sizes and high dimensions of data. Current visualization techniques mostly suffer from poor performances in functionalities, scalability and response time [30]. We may need to reconsider the way adopted for visualization. Moreover, the effectiveness of visualization may be challenged by uncertainties of data sources.

5.1.3. Systematic challenges

The development of proper system architecture is vital to support decisions in handling a diversity of complex data and conduct complex computation of Big Data. The challenges raised by this requirement include the design of system architecture, computing frameworks, processing modes, as well as high energy-efficient processing platforms. One possible solution may rely on cluster computers with a high performance computing platform. However, this challenges both hardware and software system architectures. Their final solutions will form a significant foundation for the development of system architectures. Besides, the evaluation and optimization of such energy-efficient processing systems is also of great challenge.

5.1.4. Non-Technical challenges

Non-technical challenges refer to challenges which are arisen by management problems of service suppliers and users, rather than technical challenges related to Big Data processing.

Human expertise still plays an important role for decision making and cannot be easily replaced by Big Data analysis in business and management models [99]. Lazer et al. [81] stated that human analysts are necessary to remain in the loop in certain scenarios. Thus, there is another challenge concerning how to support human analysts and managers to make quicker decisions. Technologies for Big Data should enhance their functions of interacting with users.

A series of other non-technical challenges have been discussed in Assunção et al. [8]. For instance, proper tools should be developed to estimate the costs and risks of performing data analytics for users and suppliers; some analytics services, such as analytics as a service and Big Data as a service, lack well defined contracts because of the difficulty of measuring quality and reliability of input data and output results, providing promises on execution latency and etc.

Besides, semantic challenges of Big Data, which refer to locating and meaningfully integrating the data that is relevant to users' benefit, have been discussed and reviewed in [18,72].

5.2. Principles for developing Big Data techniques

It is doomed that Big Data processing is more complicated than the traditional data analysis. New techniques and technologies, or even new thinking ways, are necessary to be developed for exploitation of Big Data. The key points of data-intensive applications are the capability of supporting in-memory processing in real-time and the satisfactory scalability. In what follows, we present some principles as the guideline of introducing new techniques in Big Data:

Principle 1: Possess powerful ability to handle uncertainties. It is obvious that uncertainties exist in almost every phase of Big Data processing. For instance, the raw data themselves may contain various categories of uncertainties; the outputs produced by specific platforms and algorithms also generate uncertainties due to their nature. Thus, we expect that the adopted techniques can model more uncertainties and make rational decisions. In addition, it is more interesting if the algorithms can be convergent with these uncertainties. Typical granular computing (GrC) techniques, such as fuzzy sets and rough sets, are popular tools for handling uncertainties [111].

Principle 2: Possess satisfactory scalability [30]. Scalability is one of the most significant properties that Big Data techniques should be satisfied for dealing with large-scale datasets. For example, one of the most famous machine learning frameworks, ensemble learning (EnL), can work well with many specific pattern recognition algorithms.

Principle 3: Enable implementation in-memory systems. In other words, good techniques are simple [57]. For one thing, in order for real-time processing, complex algorithms may be not appropriate. For another, it has been demonstrated that the simple algorithms usually perform not worse than the complex ones.

Principle 4: No size fits all. Every tool owns its advantages as well as limitations. Thus, no one size can fit all solutions [104]. We need to choose proper tools for different data-intensive applications to achieve more benefits unless we reach common theory for Big Data. But we cannot image if that level of common theory can be reached.

5.3. Potential decision making techniques and future researches

To facilitate Big Data processing, a number of techniques and technologies have been developed and adopted to benefit scientific investigations and economical applications. The ultimate aims of Big Data would drive to develop the techniques that are more sophisticated and scientific than ever before. In this subsection, we will discuss some ongoing and underlying decision making techniques to harness Big Data, except for some commonly acknowledged tools such as clouding computing.

5.3.1. Granular computing

GrC, based on techniques including fuzzy sets, rough sets, computing with words, etc., is a relatively new area that plays an important role in designing decision making models with acceptable performance. To meet the needs and challenges from several distinct domains of applications, GrC has been developed by various researchers like the ones reviewed in Section 3.1 and its capability and advantages have been exhibited in intelligent data analysis, pattern recognition, machine learning and uncertain reasoning for noticeable sizes of data [111].

The computing paradigm of GrC is based on the concept of information granulation and abstraction, and the concept of granulation is inherent in GrC techniques such as the two most successful and considered tools: fuzzy sets and rough sets. The fuzzy set theory employs the concept of membership function to produce the fuzzy granulation of feature space; while the rough set theory allows us to capture knowledge from an information system by the upper and lower approximations

and to make decisions according to the predefined indistinguishability relation and attribute reduction. When processing data with GrC, the task is, actually, to find a mapping from the original finest-grained data to the knowledge behind the set of optimized coarser and more abstract information granules associated with techniques like fuzzy sets and rough sets [111]. Different features and patterns emerge if data are represented by different granularity.

Based on these features, GrC can provide powerful support for multi-granularity and multi-view data analysis which may be towards better understanding of the complexity of Big Data. Analyzing Big Data at different granularity levels and/or viewpoints will be helpful to understand the data for different uses' requirements. Moreover, GrC techniques can find simple approximate solutions and provide the improved description of intelligent systems based on the process of large-scale data [111]. Besides, the proper use of GrC techniques would help to enhance the privacy and security of special Big Data applications if different granularities of information are provided for distinct roles of users. It is exciting that some relative researches have gone deep into heterogeneous, complex and large-scale data analysis recently. Sengoz and Ramanna [123] proposed a granular model to structure categorical noun phrase samples and semantically related noun phrase pairs from large number of unlabeled data. Kundu and Pal [78] developed a novel technique for fuzzy rough community detection in a social network. Their focused data of the online social networking sites are dynamic, large-scale, diverse and complex.

Note that GrC techniques are not available for all Big Data applications. Information hidden in data may be partially lost if the data are reduced to a coarser version. Thus, the decision may be not acceptable if high confidence and accuracy are required in the applications.

5.3.2. Information fusion

Information fusion (InF) refers to the merging of information (or data) from heterogeneous sources into a new set of information towards consistent, accurate and useful representation as well as reducing uncertainty. Techniques related to InF usually provide the textual representations of knowledge that is mined and consolidated from structured, semi-structured or unstructured data. Depending on the processing stage to take place, InF processes can be categorized as low, intermediate and high. For example, low level of InF combines a set of sources of raw data to result in new raw data. Other levels produce new information and knowledge at different degrees.

Although partially overlapped by GrC, we would like to highlight the role of InF in processing Big Data. GrC is formulated based on information representational models, whereas InF focuses on the integration, combination and synthesis of data. Proper InF techniques would benefit the analysis of Big Data from some aspects: First of all, InF techniques would improve the performances of data integrations and data storages. When data are captured from distinct and heterogeneous sources, a certain strategy should be determined to fuse and then store them. InF techniques in the low level or the intermediate level would help to handle it. In addition, InF techniques, especially ones in the intermediate level or the high level, would present powerful intelligent decision supporting for data analysis and semantic understanding. Considering a special application, we may need to synthesize information from all distributed data storages to derive new knowledge at an abstract level. In fact, the marriage to GrC techniques may be necessary to achieve this synthesis. Till now, the first aspect has been resolved to a certain extent; while the second aspect remains a big gap between myth and reality.

5.3.3. Ensemble learning

Ensemble learning (EnL) is a category of techniques that is hardly an exhaustive list. In practical terms, it is trick to say which single machine learning algorithm performs the best. One may assert that none of them, or all of them. Thus, the EnL techniques provide a framework to obtain better decision from any of the constituent learning algorithms. It is acknowledged that EnL has been a popular resolution for mining patterns from data and has been widely applied in real life.

However, the EnL techniques that we highlight here are not the traditional ones. When applied to Big Data, two important issues of EnL should be worked out. Firstly, current observed phenomena of different performances of algorithms are based on small datasets, comparing with Big Data. Along with the volume of the dataset grows, the performances may converge asymptotically to the same level of predictive accuracy [11]. When dealing with Big Data, the properties of specific algorithms as well as their activities in the EnL frameworks should be reevaluated. Secondly, over-fitting is another inevitable issue of machine learning whenever algorithms approach the noise floor of a given dataset. Also affected by the huge volume, this issue is a new challenge when using the EnL techniques in Big Data. The solutions of these two issues would bring EnL to a second generation, and new versions of techniques of the second generation will improve the capability of processing large-scale datasets.

5.3.4. Feature extraction and sampling

In the era of Big Data, two classes of techniques, i.e., FE and sampling, are vital for machine learning techniques to deal with the unprecedented scale of data with "big dimensionality", and their aims and functions are clear. Below we address some further directions that need to pay more attention:

The first challenge of the existing FE techniques is its negative repercussions on performance when millions of dimensions are confronted [149]. Moreover, most existing algorithms have been designed when the sizes of datasets are relative small. This fact causes a new problem of scalability of learning. Large-scale problems cannot be designed by an in-memory style like the small-scale problems do. Specifically, we need to find a trade-off to obtain good enough solutions as fast as possible and as efficiently as possible [21]. Finally, the FE techniques should suit the setting of Big Data storage and analysis.

If data are distributed, then the FE techniques should take advantage of processing multiple subsets of data in sequence or concurrently [23]. It would be more desirable if the techniques can meet the requirement of real-time processing.

Compared to the other issues of Big Data, sampling has been paid very little attention. Due to the space and time complexities, it is impossible to process the entire Big Data set currently. Hence, sampling techniques are necessary. The traditional sampling methods (maybe associate with the parallel algorithms), such as the statistic method, are commonly used whenever machine learning algorithms are considered. However, various kinds of uncertainties (including missing values) may be involved in Big Data sets. The distribution of patterns may be extremely unbalanced. Thus, more effective sampling techniques should be figured out for the purposes of both accurate prediction and real-time processing. These techniques, indeed, may be data-intensive or application-intensive. Although there are some studies that focus on sampling [94,103,145], it is far from enough.

6. Conclusions

Along with the accumulation of ubiquitous and incessantly generated data, Big Data have become a new popular and booming discipline based on techniques and technologies from many other disciplines. More and more initiatives have been presented by different organizations and governments. A large amount of literature has been published, which facilitate and accelerate the development of Big Data. The concepts, aims and processing paradigms of Big Data are becoming more and more explicit and distinct. A number of techniques and technologies focusing on processing Big Data have been presented and have brought big value for organizations and users. We believe that the developments of Big Data would result to the following achievements:

Techniques will make the processing of Big Data more intelligent. The existing challenges of processing Big Data will not only develop the current status, but also bring new thinking and idea into this field. More and more elaborate techniques have been introduced or under-developing to focus on the characteristic of Big Data. It will enable to make more intelligent decisions in each phase of processing.

Developments of Big Data will enrich current decision science. Big Data will produce bigger value along with the resolutions of current challenges. It is no doubt that the value will be created by intelligent decision making based on the analytical results of raw data. Especially, in social science, decisions can be made by not only analytical approaches but also computational ones. Computational approaches may be even more powerful and effective.

Theory of Big Data will be towards systematic standardization. The standards cover many specific points of Big Data from conceptualizing Big Data to applications. For example, the aims and scopes, as well as processing paradigms, should be standardized at first. Then standards are necessary to ensure the storages and transformations of data. The issues involved with privacy and security can be settled by this kind of standardization. The design and development of software platforms require standards as well so that the developed technologies can be easily reused and extended. In summary, the standardized theory for Big Data will be systematically formed.

Big Data will change the paradigms of investigation in social science. The presence of Big Data alters the research style such as the questions we can ask and the methods we can apply. The constantly depressed cost of data capturing and new techniques enable us to achieve frequent, controlled and meaningful observations of real-world business and economic phenomena. Techniques related to computational social science, such as ensemble learning and penalized regression, have much broader applications than the traditional techniques of social science, such as standard regression analyses.

Acknowledgments

The authors would like to thank the Editor-in-Chief, the associated editor and three anonymous reviewers for their insightful and constructive commendations that have led to an improved version of this paper. The work was supported by the National Natural Science Foundation of China (Nos. 61273209, 71571123), the Scientific Research Foundation of Graduate School of Southeast University (No. YBJJ1528).

References

[1] W. van der Aalst, E. Damiani, Processes meet big data: connecting data science with process science, IEEE Trans. Serv. Comput. 8 (2015) 810–819.
[2] M. Adrian, Big Data, Teradata Magazine. http://www.teradatamagazine.com/v11n01/Features/Big-Data/ (accessed December 2015).
[3] R. Agerri, X. Artola, Z. Beloki, G. Rigau, A. Soroa, Big data for natural language processing: a streaming approach, Knowl. Based Syst. 79 (2015) 36–42.
[4] J. Ahrens, K. Brislawn, K. Martin, B. Geveci, C.C. Law, M. Papka, Large-scale data visualization using parallel data streaming, IEEE Comput. Graph. 21 (2001) 34–41.
[5] A. Almaatouq, A. Alabdulkareem, M. Nouh, E. Shmueli, M. Alsaleh, V.K. Singh, A. Alarifi, A. Alfaris, A.S. Pentland, Twitter: who gets caught? observed trends in social micro-blogging spam, in: Proceedings of the 2014 ACM conference on Web science, ACM, 2014, pp. 33–41.
[6] I. Arel, D.C. Rose, T.P. Karnowski, Deep machine learning-a new frontier in artificial intelligence research, IEEE Comput. Intell. Mag. 5 (2010) 13–18.
[7] M.Z. Asghar, A. Khan, S. Ahmad, I.A. Khan, F.M. Kundi, A unified framework for creating domain dependent polarity lexicons from user generated reviews, PLoS One 10 (2015) 1–19 Document number e0140204.
[8] M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, Big Data computing and clouds: trends and future directions, J. Parallel Distrb. Comput. 79 (2015) 3–15.
[9] S. Asur, B. Huberman, Predicting the future with social media, in: Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2010, pp. 492–499.
[10] A.T. Azar, A.E. Hassanien, Dimensionality reduction of medical big data using neural-fuzzy classifier, Soft Comput. 19 (2014) 1115–1127.

[11] M. Banko, E. Brill, Scaling to very very large corpora for natural language disambiguation, in: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2001, pp. 26–33.
[12] J. Bao, Y. Zheng, M.F. Mokbel, Location-based and preferenceaware recommendation using sparse geo-social networking data, in: Proceedings of the 20th International Conference of Advanced Geographic Information Systems, 2012, pp. 199–208.
[13] H. Barwick, The "four Vs" of Big Data. Implementing Information Infrastructure Symposium, 2012. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/ (accessed December 2015).
[14] G. Bell, T. Hey, A. Szalay, Beyond the data deluge, Science 323 (2009) 1297–1298.
[15] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Inf. Fusion 28 (2016) 45–59.
[16] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. 35 (2013) 1798–1828.
[17] L.M. Bettencourt, The uses of big data in cities, Big Data 2 (2014) 12–22.
[18] C. Bizer, P. Boncz, M.L. Brodie, O. Erling, The meaningful use of big data: four perspectives–four challenges, ACM SIGMOD Rec. 40 (2012) 56–60.
[19] M. Bohlouli, J. Dalter, M. Dornhöfer, J. Zenkert, M. Fathi, Knowledge discovery from social media using big data-provided sentiment analysis (SoMABiT), J. Inf. Sci. 41 (2015) 779–798.
[20] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Data classification using an ensemble of filters, Neurocomputing 135 (2014) 13–20.
[21] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Recent advances and emerging challenges of feature selection in the context of big data, Knowl. Based Syst. 86 (2015) 33–45.
[22] D. Boyd, K. Crawford, Critical questions for big data provocations for a cultural, technological, and scholarly phenomenon, Inf. Commun. Soc. 15 (2012) 662–679.
[23] M. Bramer, Principles of Data Mining, Springer, 2007.
[24] F. Bravo-Marquez, M. Mendoza, B. Poblete, Meta-level sentiment models for big social data analysis, Knowl. Based Syst. 69 (2014) 86–99.
[25] J. Brooks, Review: Talend open studio makes quick etl work of large data sets, 2009. http://www.eweek.com/c/a/Database/REVIEW-Talend-Open-Studio-Makes-Quick-ETL-Work-of-Large-Data-Sets-281473/ (accessed December 2015).
[26] R. Casado, M. Younas, Emerging trends and technologies in big data processing, Concurr. Comp-Pract. E. 27 (2015) 2078–2091.
[27] S. Chainey, L. Tompson, S. Uhlig, The utility of hotspot mapping for predicting spatial patterns of crime, Secur. J. 21 (2008) 4–28.
[28] H.-T. Chang, N. Mishra, C.-C. Lin, IoT big-data centred knowledge granule analytic and cluster framework for BI applications: a case base analysis, PLoS One 10 (2015) e0141980.
[29] R.M. Chang, R.J. Kauffman, Y. Kwon, Understanding the paradigm shift to computational social science in the presence of big data, Decis. Support Syst. 63 (2014) 67–80.
[30] C.P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on Big Data, Inf. Sci. 275 (2014) 314–347.
[31] H.C. Chen, R.H.L. Chiang, V.C. Storey, Business intelligence and analytics: From big data to big impact, MIS Q. 36 (2012) 1165–1188.
[32] T.-S. Chua, X. He, W. Liu, M. Piccardi, Y. Wen, D. Tao, Big data meets multimedia analytics, Signal Process. 124 (2016) 1–4.
[33] J. Cowls, R. Schroeder, Causation, correlation, and big data in social science research, Policy Intern. 7 (2015) 447–472.
[34] J.W. Crampton, Collect it all: national security, Big Data and governance, GeoJournal 80 (2015) 519–531.
[35] H.G. de Zúñiga, Citizenship, social media, and big data current and future research in the social sciences, Soc. Sci. Comput. Rev. (2015) 0894439315619589, doi:10.1177/0894439315619589.
[36] J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, Commun. ACM 51 (2008) 107–113.
[37] H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud, Decis. Support Syst. 55 (2013) 412–421.
[38] P.S. Dodds, K.D. Harris, I.M. Kloumann, C.A. Bliss, C.M. Danforth, Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter, PLoS One 6 (2011) e26752.
[39] L.E. Duncan, M.C. Keller, A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry, Am. J Psychiat. 168 (2011) 1041–1049.
[40] A.O. Durahim, M. Coşkun, # iamhappybecause: gross national happiness through Twitter analysis and big data, Technol. Forecast Soc. 99 (2015) 92–105.
[41] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V.R. Suri, A. Tsou, S. Weingart, C.R. Sugimoto, Big data, bigger dilemmas: a critical review, J. Assoc. Inf. Sci. Technol. 66 (2015) 1523–1545.
[42] Q. Fang, C.S. Xu, J.T. Sang, M.S. Hossain, G. Muhammad, Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media, IEEE Trans. Multimed. 17 (2015) 2281–2296.
[43] A.E.T. Finlayson, Dealing with data: fostering fidelity, Science 331 (2011) 1515-1515.
[44] J. Gan, C. Norman, 2012 visualization challenge, Science 339 (2013) 509.
[45] E. Gawehn, J.A. Hiss, G. Schneider, Deep learning in drug discovery, Mol. Inform. 35 (2016) 3–14.
[46] M.S. Gerber, Predicting crime using Twitter and kernel density estimation, Decis. Support Syst. 61 (2014) 115–125.
[47] M.M. Gobble, Big Data: the next big thing in innovation, Res. Technol. Manag. 56 (2013) 64–66.
[48] J. Grzymala-Busse, Discretization based on entropy and multiple scanning, Entropy 15 (2013) 1486–1502.
[49] L. Guo, C. Vargo, The power of message networks: A big-data analysis of the network agenda setting model and issue ownership, Mass Commun. Soc. 18 (2015) 557–576.
[50] E. Hargittai, Is bigger always better? Potential biases of big data derived from social network sites, Ann. Am. Acad. Polit. Soc. Sci. 659 (2015) 63–76.
[51] T.J. Hastie, R.J. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
[52] Q. He, H. Wang, F. Zhuang, T. Shang, Z. Shi, Parallel sampling from big data with uncertainty distribution, Fuzzy Sets Syst. 258 (2015) 117–133.
[53] W. He, H. Wu, G.J. Yan, V. Akula, J.C. Shen, A novel social media competitive analytics framework with sentiment benchmarks, Inform. Manage-Amster. 52 (2015) 801–812.
[54] A.J. Hey, S. Tansley, K.M. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, Redmond, WA, 2009.
[55] A. Heydari, M.A. Tavakoli, N. Salim, Z. Heydari, Detection of review spam: a survey, Expert Syst. Appl. 42 (2015) 3634–3642.
[56] M. Hilbert, P. López, The world's technological capacity to store, communicate, and compute information, Science 332 (2011) 60–65.
[57] M. Hindman, Building Better Models Prediction, Replication, and Machine Learning in the Social Sciences, Ann. Am. Acad. Polit. Soc. Sci. 659 (2015) 48–62.
[58] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D.P. Hill, R. Kania, M. Schaeffer, S. St Pierre, Big data: the future of biocuration, Nature 455 (2008) 47–50.
[59] IBM, What is big data? Bringing big data to the enterprise, 2012. http://www-01.ibm.com/software/data/bigdata/ (accessed December 2015).
[60] J.A. Iglesias, A. Tiemblo, A. Ledezma, A. Sanchis, Web news mining in an evolving framework, Inf. Fusion 28 (2016) 90–98.
[61] A. Ignatius, From the editor: big data for skeptics, Harv. Bus. Rev. 10 (2012) 12-12.
[62] N. Indurkhya, Emerging directions in predictive text mining, WIREs Data Min. Knowl. 5 (2015) 155–164.
[63] G. Ingersoll, Introducing Apache Mahout Scalable, Commercial-Friendly Machine Learning for Building Intelligent Applications, IBM Corporation, 2009.
[64] N.N.I. Initiative, Core techniques and technologies for advancing big data science and engineering (BIGDATA), 2012. http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf12499 (accessed December 2015).
[65] M. Isard, M. Budiu, Y. Yu, A. Birrell, D. Fetterly, Dryad: distributed data-parallel programs from sequential building blocks, in: Proceedings of ACM SIGOPS Operating Systems Review, ACM, 2007, pp. 59–72.
[66] R.P. Ishii, R.F. De Mello, An online data access prediction and optimization approach for distributed systems, IEEE Trans. Parallel DistrB. 23 (2012) 1017–1029.

[67] A. Jacobs, The pathologies of big data, Commun. ACM 52 (2009) 36–44.
[68] B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: Tweets as electronic word of mouth, J. Am. Soc. Inf. Sci. Tecnol. 60 (2009) 2169–2188.
[69] S. Jin, W. Lin, H. Yin, S. Yang, A. Li, B. Deng, Community structure mining in big data social media networks with MapReduce, Cluster Comput. 69 (2015) 1–12.
[70] X. Jin, B.W. Wah, X. Cheng, Y. Wang, Significance and challenges of big data research, Big Data Res. 2 (2015) 59–64.
[71] V.G. Kaburlasos, G.A. Papakostas, Learning distributions of image features by interactive fuzzy lattice reasoning in pattern recognition applications, IEEE Comput. Intell. Mag. 10 (2015) 42–51.
[72] C. Kacfah Emani, N. Cullot, C. Nicolle, Understandable big data, Comput. Sci. Rev. 17 (2015) 70–81.
[73] C. Kimble, G. Milolidakis, Big data and business intelligence: debunking the myths, Global Bus. Organ. Excell. 35 (2015) 23–34.
[74] S. Kraft, G. Casale, A. Jula, P. Kilpatrick, D. Greer, Wiq: work-intensive query scheduling for in-memory database systems, in: Proceeding of 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), IEEE, 2012, pp. 33–40.
[75] T. Kraska, Finding the needle in the big data systems haystack, IEEE Intern. Comput. 17 (2013) 84–86.
[76] C.-H. Ku, G. Leroy, A decision support system: Automated crime report analysis and classification for e-government, Gov. Inf. Q. 31 (2014) 534–544.
[77] S. Kundu, S.K. Pal, FGSN: fuzzy granular social networks – model and applications, Inf. Sci. 314 (2015) 100–117.
[78] S. Kundu, S.K. Pal, Fuzzy-rough community in social networks, Pattern Recognit. Lett. 67 (2015) 145–152.
[79] V. López, S. del Río, J.M. Benítez, F. Herrera, Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data, Fuzzy Sets Syst. 258 (2015) 5–38.
[80] D. Laney, 3D Data Management: Controlling Data Volume, Velocity and Variety, Research Note 6, META Group, 2001.
[81] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: traps in big data analysis, Science 343 (2014) 1203–1205.
[82] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[83] M.K.K. Leung, A. Delong, B. Alipanahi, B.J. Frey, Machine learning in genomic medicine: a review of computational problems and data sets, Proc. IEEE 104 (2016) 176–197.
[84] J. Li, F. Tao, Y. Cheng, L. Zhao, Big Data in product lifecycle management, Int. J. Adv. Manuf. Technol. 81 (2015) 1–18.
[85] S.H. Li, D.C. Yen, W.H. Lu, C. Wang, Identifying the signs of fraudulent accounts using data mining techniques, Comput. Hum. Behav. 28 (2012) 1002–1013.
[86] X. Li, X. Yao, Cooperatively coevolving particle swarms for large scale optimization, IEEE Trans. Evol. Comput. 16 (2012) 210–224.
[87] C. Liao, A. Squicciarini, C. Griffin, Epidemic behavior of negative users in online social sites, in: Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, ACM, 2015, pp. 143–145.
[88] C.W. Lin, T.P. Hong, A survey of fuzzy web mining, Wires. Data Min. Knowl. 3 (2013) 190–199.
[89] R. Lin, S. Utz, The emotional responses of browsing Facebook: Happiness, envy, and the role of tie strength, Comput. Hum. Behav. 52 (2015) 29–38.
[90] Z.L. Liu, J.W. Li, J. Li, C.F. Jia, J. Yang, K. Yuan, SQL-based fuzzy query mechanism over encrypted database, Int. J. Data Wareh. 10 (2014) 71–87.
[91] H.P. Lu, Z.Y. Sun, W.C. Qu, Big data-driven based real-time rraffic flow state identification and prediction, Discrete Dyn. Nat. Soc. 2015 (2015) 284906.
[92] C. Lynch, Big data: how do your data grow? Nature 455 (2008) 28–29.
[93] H. Ma, H. Yang, M.R. Lyu, I. King, Mining social networks using heat diffusion processes for marketing candidates selection, in: Proceedings of the 17th ACM conference on Information and knowledge management, ACM, 2008, pp. 233–242.
[94] A.S. Mahani, M.T.A. Sharabiani, SIMD parallel MCMC sampling with applications for big-data Bayesian analytics, Comput. Stat. Data Anal. 88 (2015) 75–99.
[95] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Inf. Sci. 181 (2011) 115–128.
[96] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers, Big Data: The Next Frontier For Innovation, Competition, and Productivity. Report, McKinsey Global Institute, 2012.
[97] V. Marx, The big challenges of big data, Nature 498 (2013) 255–260.
[98] N. Marz, J. Warren, Big Data: Principles and Best Practices of Scalable Realtime Data Systems, Manning Publications Co., 2012.
[99] A. McAfee, E. Brynjolfsson, Big data: the management revolution, Harv. Bus. Rev. 90 (2012) 60–68.
[100] J.M. Mendel, M.M. Korjani, On establishing nonlinear combinations of variables from small to big data for use in later processing, Inf. Sci. 280 (2014) 98–110.
[101] E. Miller, Community cleverness required, Nature 455 (2008) 1.
[102] H.G. Miller, P. Mork, From data to decisions: a value chain for big data, IT Prof. 15 (2013) 57–59.
[103] S. Molavipour, A. Gohari, Recovery from random samples in a big data set, IEEE Commun. Lett. 19 (2015) 1929–1932.
[104] C. Molinari, No one size fits all strategy for big data, says IBM, 2012. http://www.bnamericas.com/news/technology/no-one-sizefits-all-strategy-for-big-data-says-ibm (accessed December 2015).
[105] J.A. Morente-Molinera, I.J. Perez, M.R. Urena, E. Herrera-Viedma, Creating knowledge databases for storing and sharing people knowledge automatically using group decision making and fuzzy ontologies, Inf. Sci. 328 (2016) 418–434.
[106] N. Nedjah, F.P.d. Silva, A.O.d. Sá, L.M. Mourelle, D.A. Bonilla, A massively parallel pipelined reconfigurable design for M-PLN based neural networks for efficient image classification, Neurocomputing 183 (2016) 39–55.
[107] L. Neumeyer, B. Robbins, A. Nair, A. Kesari, S4: Distributed stream computing platform, in: Proceedings of 2010 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2010, pp. 170–177.
[108] T.L. Ngo-Ye, A.P. Sinha, The influence of reviewer engagement characteristics on online review helpfulness: a text regression model, Decis. Support Syst. 61 (2014) 47–58.
[109] OSP, Obama administration unveils “big data” initiative: Announces $200 million in new R&D investments, 2013. http://www.whitehouse.gov/sites/efault/files/microsites/ostp/big_data_press_release_final_2.pdf (accessed December 2015).
[110] P. Pébay, D. Thompson, J. Bennett, A. Mascarenhas, Design and performance of a scalable, parallel statistics toolkit, in: Proceedings of 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), IEEE, 2011, pp. 1475–1484.
[111] S.K. Pal, S.K. Meher, A. Skowron, Data science, big data and granular mining, Pattern Recognit. Lett. 67 (2015) 109–112.
[112] G. Phillips-Wren, L.S. Iyer, U. Kulkarni, T. Ariyachandra, Business analytics in the context of big data: a roadmap for research, Commun. Assoc. Inf. Syst. 34 (2015) 448–472.
[113] P. Phillips, I. Lee, Mining co-distribution patterns for large crime datasets, Expert Syst. Appl. 39 (2012) 11556–11563.
[114] S. Ramachandramurthy, S. Subramaniam, C. Ramasamy, Distilling big data: refining quality information in the era of yottabytes, Sci. World J. 2015 (2015) 1–9.
[115] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl. Based Syst. 89 (2015) 14–46.
[116] J. Rozas, J.C. Sanchez-DelBarrio, X. Messeguer, R. Rozas, DnaSP, DNA polymorphism analyses by the coalescent and other methods, Bioinformatics 19 (2003) 2496–2497.
[117] M. Sahimi, H. Hamzehpour, Efficient computational strategies for solving global optimization problems, Comput. Sci. Eng. 12 (2010) 74–83.
[118] T. Samson, Splunk storm brings log management to the cloud, 2012. http://www.infoworld.com/t/managed-services/splunk-storm-brings-logmanagement-the-cloud-201098?source=footer (accessed December 2015).
[119] D. Samuels, Skytree: machine learning meets big data, 2012. http://www.bizjournals.com/sanjose/blog/2012/02/skytree-machinelearning-meets-big-data.html?page=all (accessed December 2015).
[120] E.E. Schadt, M.D. Linderman, J. Sorenson, L. Lee, G.P. Nolan, Computational solutions to large-scale data management and analysis, Nat. Rev. Genet. 11 (2010) 647–657.

[121] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[122] G. Seenumani, J. Sun, H. Peng, Real-time power management of integrated power systems in all electric ships leveraging multi time scale property, IEEE Trans. Contr. Syst. Technol. 20 (2012) 232–240.
[123] C. Sengoz, S. Ramanna, Learning relational facts from the web: a tolerance rough set approach, Pattern Recogn. Lett. 67 (2015) 130–137.
[124] H. Shen, L. Zhao, Z. Li, A distributed spatial-temporal similarity data storage scheme in wireless sensor networks, IEEE Trans. Mob. Comput. 10 (2011) 982–996.
[125] B. Shneiderman, The big picture for big data: visualization, Science 343 (2014) 730-730.
[126] C. Staff, Visualizations make big data meaningful, Commun. ACM 57 (2014) 19–21.
[127] P. Sun, X. Yao, Sparse approximation through boosting for learning large scale kernel machines, IEEE Trans. Neural Netw. 21 (2010) 883–894.
[128] O. Sysoev, O. Burdakov, A. Grimvall, A segmentation-based algorithm for large-scale partially ordered monotonic regression, Comput. Stat. Data Anal. 55 (2011) 2463–2476.
[129] H. Takemi, Remarks for special issue on big data, NEC Tech. J. 7 (2012) 8–10.
[130] W. Tan, M.B. Blake, I. Saleh, S. Dustdar, Social-network-sourced big data analytics, IEEE Intern. Comput. 17 (2013) 62–69.
[131] D. Thompson, J.A. Levine, J.C. Bennett, P.-T. Bremer, A. Gyulassy, V. Pascucci, P.P. Pébay, Analysis of large-scale scalar data using hixels, in: Proceedings of 2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV), IEEE, 2011, pp. 23–30.
[132] J.M. Tien, Big data: unleashing information, J. Syst. Sci. Syst. Eng. 22 (2013) 127–151.
[133] C. Trattner, F. Kappe, Social stream marketing on Facebook: a case study, Int. J. Soc. Humanist. Comput. 2 (2013) 86–103.
[134] J.W. Tukey, The technical tools of statistics, Am. Stat. 19 (1965) 23–28.
[135] H. Uğuz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowl. Based Syst. 24 (2011) 1024–1032.
[136] C. Wang, X. Li, X.H. Zhou, A.L. Wang, N. Nedjah, Soft computing in big data intelligent transportation systems, Appl. Soft Comput. 38 (2016) 1099–1108.
[137] R. Wang, Y.-L. He, C.-Y. Chow, F.-F. Ou, J. Zhang, Learning ELM-Tree from big data based on uncertainty reduction, Fuzzy Sets Syst. 258 (2015) 79–100.
[138] W. Wang, E. Krishnan, Big data and clinicians: a review on the state of the science, JMIR 2 (2014) e1.
[139] Y. Wang, X.L. Jiang, R.Y. Cao, X.Y. Wang, Robust indoor human activity recognition using wireless signals, Sensors 15 (2015) 17195–17208.
[140] P. Wayner, 7 top tools for taming big data, 2012. http://www.networkworld.com/reviews/2012/041812-7-top-tools-for-taming-258398.html (accessed December 2015).
[141] A. Weichselbraun, A. Gindl, A. Scharl, Enriching semantic knowledge bases for opinion mining in big data applications, Knowl. Based Syst. 69 (2014) 78–85.
[142] Z.S. Wen, W.W. Zhang, T. Zeng, L.N. Chen, MCentridFS: a tool for identifying module biomarkers for multi-phenotypes from high-throughput data, Mol. Biosyst. 10 (2014) 2870–2875.
[143] M. Wilhelm, J. Schlegl, H. Hahne, A.M. Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, Mass-spectrometry-based draft of the human proteome, Nature 509 (2014) 582–587.
[144] L. Wilkinson, The future of statistical computing, Technometrics 50 (2008) 418–435.
[145] X. Wu, W. Fan, J. Peng, K. Zhang, Y. Yu, Iterative sampling based frequent itemset mining for big data, Int. J. Mach. Learn. Cybern. 6 (2015) 875–882.
[146] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, IEEE Trans. Knowl. Data Eng. 26 (2014) 97–107.
[147] Y.J. Xia, J.L. Chen, C.H. Wang, Formalizing computational intensity of big traffic data understanding and analysis for parallel computing, Neurocomputing 169 (2015) 158–168.
[148] J. Yan, N. Liu, S. Yan, Q. Yang, W. Fan, W. Wei, Z. Chen, Trace-oriented feature analysis for large-scale text data dimension reduction, IEEE Trans. Knowl. Data Eng. 23 (2011) 1103–1117.
[149] Y. Zhai, Y.S. Ong, I.W. Tsang, The emerging "big dimensionality", IEEE Comput. Intell. Mag. 9 (2014) 14–26.
[150] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, C. Chen, Data-driven intelligent transportation systems: a survey, IEEE Trans. Intell. Transp. 12 (2011) 1624–1639.
[151] W. Zhang, R. Lau, C. Li, Adaptive big data analytics for deceptive review detection in online social media, in: Proceedings of 2014 Proceedings of International Conference on Information Systems (ICIS), 2014, pp. 1–19.
[152] X.H. Zheng, W. Chen, P. Wang, D.Y. Shen, S.H. Chen, X. Wang, Q.P. Zhang, L.Q. Yang, Big data for social transportation, IEEE Trans. Intell. Transp. 17 (2016) 620–630.
[153] L. Zhou, K.P. Tam, H. Fujita, Predicting the listing status of Chinese listed companies with multi-class classification models, Inf. Sci. 328 (2016) 222–236.
[154] P. Zikopoulos, C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2011.