Ann. Data. Sci. (2015) 2(4):363–372
DOI 10.1007/s40745-015-0063-7

# Exploring Big Data Analysis: Fundamental Scientific Problems

**Zongben Xu**[1] · **Yong Shi**[2]

Received: 20 April 2015 / Revised: 8 December 2015 / Accepted: 10 December 2015 /
Published online: 9 January 2016


**Abstract** Although Big Data has been one of most popular topics since last several years, how to effectively conduct Big Data analysis is a big challenge for every field. This paper tries to address some fundamental scientific problems in Big Data analysis, such as opportunities, challenges, and difficulties encountered in the analysis. The challenges rise from multiple domains that include how Management Science influences data acquisition and data management, Information Science for data access and processing, Mathematics and Statistics for data understanding and Engineering for data applications. The paper outlines six open research problems on Big Data. It also reports some advances on current Big Data research, particularly in high-dimensional data and non-structured data processing. Finally, remarks on how to develop a Big Data algorithm are provided.

**Keywords** Big Data analysis · Big Data algorithm · Open challenges

## 1 Introduction

The issue of Big Data has been gradually discussing across all of fields since 2001 when Gartner Co. released its "3Vs" (volume, velocity and variety) description of Big Data [1]. It became a hot topic for last 4 years [2]. However, there is not yet a unified defin-

✉ Yong Shi
yshi@ucas.ac.cn

Zongben Xu
zbxu@mail.xjtu.edu.cn

1 Department of Mathematics, Xi'an Jiaotong University, Xi'an, China

2 The Key Research Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

ition of Big Data; its interpretation varies from academic and business communities. The US National Science Foundation describes Big Data as "large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future" [3], while Wikipedia says, "Big Data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process using traditional data processing applications [4]. It is worth to mention that a group of international scholars brainstormed two definitions of Big Data in a session on Data Science and Big Data at the Xiangshan Science Conference [5] in Beijing. The first definition, for academic and business communities, is "a collection of data with complexity, diversity, heterogeneity, and high potential value that are difficult to process and analyze in reasonable time," and the second, for policy makers, is "a new type of strategic resource in the digital era and the key factor to drive innovation, which is changing the way of humans' current production and living." [6] Now most of researchers and professionals commonly agree to use "4Vs"—volume, velocity, variety, and veracity—to describe the main characteristics of Big Data.

Big Data allows us to take advantage of full scale of data, structured, semi-structured and/or non-structured data to make decision. Since it embodies great values that might not be explored in small sized data, Big Data provides a big opportunity of utilizing the rapid upgraded information technologies to generate expectations in all fields. For example, scientific researches in high-energy physics, astronomy, life science, geosciences and remote sensing can improve their ability with Big Data to discover unknowns. Policy makers can use Big Data in a systematic approach to assessing public policies and strategies, while commercial companies explore Big Data for benefit/income, valuable customer finding, and marketing shares.

Unfortunately, the current information technologies still lack of capability in dealing with the "4Vs" characteristics of Big Data. For "volume", when data reaches to Petabyte to Zettabyte in scale, the distributed storage and processing are necessary for many data owners (either academia or corporations) can be a costly investment. For "velocity", due to Big Data flow is growing tremendously in a short time, the device, equipment and software collecting Big Data need flexibility of responses, which may go beyond the existing capability for many data owners. For "variety", the different types of Big Data, such as multisource, correlated, heterogeneous, non-structured, unreliable, and inconsistent data demand high efficiency and capacity of database management environment. Lastly, for "value", because most existing data mining or knowledge discovering algorithms (tools) are based on how to handling structured data, it is difficult to adopt them for finding the value of Big Data.

In this paper, we will present some fundamental scientific problems in Big Data analysis, such as challenges and difficulties encountered in the analysis. Section 2 of the paper will describe challenges rise from multiple domains that include how Management Science influences data acquisition and data management, Information Science for data access and processing, Mathematics and Statistics for data understanding and Engineering for data applications. Section 3 will propose six open research problems on Big Data. Section 4 will report some advances on current Big Data research, particularly in high-dimensional data and non-structured data processing. Finally, Sect. 5 will provide remarks and future expectation on Big Data algorithm and analysis.

## 2 Challenges of Big Data Analysis

The process of Big Data analysis can be described by a general data analysis, which consists of several steps, including data acquisition and management, data access and processing, data mining and interpretation, and data applications (Fig. 1). However, due to the "4Vs" characteristics of Big Data, the activities of each step in the process face fundamental challenges. The techniques of multidisciplinary fields need to apply in addressing such challenges.

The first fundamental challenge is how to effectively acquire, store, and document Big Data in data acquisition and management. Majority of Big Data are represented as semi-structured and non-structured formats. Even though the technologies of MapReduce (Hadoop) can be used to acquire Big Data, the traditional data acquisition and management of Computer Science should be reinforced by the knowledge of Management Science. For example, the organizational strategy of using Big Data must be considered before performing the Big Data acquisition. The basic design of Big Data base and management should be built up in terms of data capabilities, value, ethic, ownership, policy, quality assurance etc. [7]. With help of Management Science, Big Data can play as an important role for us to make effective decision.

The second fundamental challenge is related to Big Data access and processing. The complex formats and features of Big Data lead the difficulty of assessing, especially processing the data for data mining and interpretation. Many existing techniques of

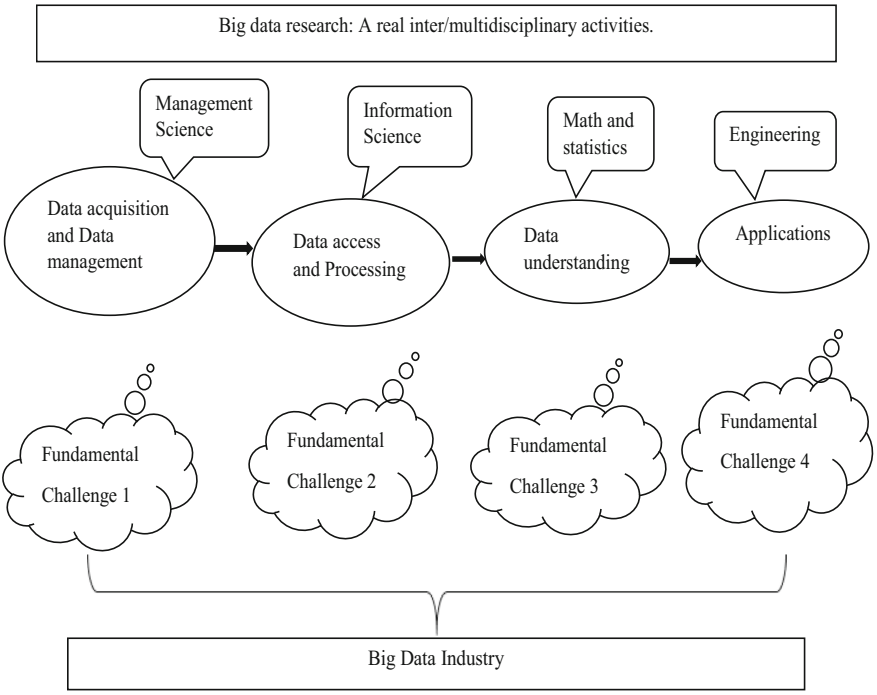


**Fig. 1** Big Data: Fundamental challenges

Information Science are ready to respond this challenge. Since most data mining or machine leaning algorithms are constructed to handle structured data, they cannot be used directly analyze a large-scale of semi-structured and non-structured data. Note that the current information technology still lacks the ability of computing big volume of semi-structured and non-structured data, such as clustering millions of text files, images or both in a reasonable time. To do this, we have to find a way to transform the semi-structured and non-structured data to structured data or pseudo structured formats which can be analyzed by many known data mining or machine leaning algorithms [6]. This transformation process can be done by using the existing information retrieval algorithms in documents such as for information within documents and metadata about documents and web page. For a given objective of the transformation, some information retrieval algorithm can be applied to turn each text file into a single record with a number of attributes into a "structured or pseudo structured format". Similarly, an image can be transformed by using a known pattern recognition algorithm as a record of the transformed format. It can be observed that whenever the transformation objective is changed, the structured or pseudo structured format will vary. Therefore, the knowledge of Information Science can be effectively applied to treat the Big Data access and processing problem.

The third fundamental challenge is how to utilize rules and principles of Mathematics and Statistics in Big Data mining and interpretation. With the analyzable Big Data formats, all possible methods in Mathematics and Statistics may be used to conduct Big Data analysis. For instance, the modeling methods can include parent space identification and sampling; clustering, classification, regression, prediction and variable selection in data mining methods; relevance analysis, latent variable analytics and statistical inference in analytical methods; and sub sampling, complexity and distributed computation in computation methods. The challenge reflects when and which method is appropriate to be used in a particular Big Data mining case. Because the transformation of Big Data is subject to the pre-determined objective, it can be useful to choose a method for data mining or knowledge discovery. Like traditional data mining procedure, experimental design for method choice should be conducted in such Big Data mining for most of cases. However, the results of Big Data mining have to be interacted with the user's judgment for the reason that knowledge changes with the individual and situation [8]. In order to let the user have a better understanding of knowledge from Big Data mining, different representation or visualization methods, like uniform scheme can be employed to show the simple versions of Big Data complexity.

The forth fundamental challenge is how to use knowledge from Big Data analysis in the real-life applications. This perhaps turns to an Engineering problem. Engineering is generally defined as "the application of scientific, economic, social, and practical knowledge in order to invent, design, build, maintain, research, and improve structures, machines, devices, systems, materials and processes" [9]. Use of Big Data knowledge in most of situations has to do with enhancing the current stages of either scientific, economic, or social conditions. Nowadays every corner and event of our human society depends on Big Data. Data-driven decision is eventually becomes the most reliable approach to any problem. A good engineering design for Big Data application will naturally yield the better way to achieve scientific, social and/or economic benefits.

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

## 3 Open Problems on Big Data Research

This section outlines some of scientific problems in Big Data analysis and processing as part of efforts of dealing with Big Data. It is reminder that these problems are what authors believe and they are urgent to be solved. There could be many remaining problems depend on the different view of Big Data.

### 3.1 Problem 1: High Dimensionality

Given a database, when the number of features (p) is far larger than the sample size ($n$), and $n$ varies with $p(n = n(p))$, the situation is called high dimensionality (HD) problem. When the problem occurs at Big Data, $p \gg n(p)$. HD frequently appears in medical science, such as DNA scanning. In the linear case a basic solution can be shown as:

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

An asymptotical normality of this is $\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \frac{1}{n}(X'X)^{-1}\sigma^2) \xrightarrow{d} N(0, \sigma^2 I_{p\times p})$.

There are a number of recent approaches that may categorized as sparse modeling, including compressed sensing, low rank decomposition of matrix and sparse learning to deal with HD problems (for instance, see [10–13]). Some of these developed algorithms are available to be used to handle HD problems in Big Data. The open research questions for HD problems are how to add priors so that a HD problem can be well defined; and how to find effective sparse modeling and etc. Eventually, systematically solving HD problems need a build-up on the theory and methodology of either HD statistics or HD data mining.

### 3.2 Problem 2: Sub-sampling

The current technologies, like Hadoop system of processing Big Data are some types of divide-and-conquer'schemes, where sub-sampling techniques have been employed. For example, in Map Reduce, Map is designed as random sub-sampling of sub datasets with intermediate solutions from given large database, where Reduce is an aggregation process of intermediate solutions for the final estimation of a given database [14]. Although sub-sample is one of the key concepts in Big Data processing, there are many open questions that need to address so that the more advance technology on

Big Data can be developed. For example, how to sub-sampling/aggregate so that the final estimation model of the given original database is properly representing the database? Is the distributed processing feasible? How about traditional sub-sampling/re-sampling technologies work? Are there sub-sampling axioms, such as similarity and transitivity?

### 3.3 Problem 3: Computational Complexity

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

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

### 3.4 Problem 4: Real and Distributed Computation

Parallel and distributed processing become necessary, perhaps is a unique way of processing for Big Data [15]. The main challenges of such a real and distributed (R/D) computation in handling Big Data come from the relationship of three components of Hadoop system: the Hadoop Distributed File System (HDFS) which is a distributed file system designed to run on commodity hardware, HBase which is an open source, non-relational, distributed database, and MapReduce. The quality measurements of Hadoop system for a real and distributed computation include real time, feasibility, efficiency, scalability, etc. It should be noted that some of the measures are conflicted each other and a compromised standard among them is a way to look for a good computational result. There are some open questions in this area. For example, does the R/D computation support fast storage/reading/ranking? For problem decomposability, can a data modeling problem be decomposed into a series of sub-data set dependent problems? For solution assemblies, how can the solution of a problem be assembled with its sub-solution (component solutions)? When the distributed process is conducted, can the forward and incremental steps be performed by on-line computation?

### 3.5 Problem 5: Nonstructured Processing

It has been commonly recognized that structured data are those that can be represented with finite number of rules and can be processed within acceptable time. Otherwise, the data are nonstructured (some of them are also called semistructured), which are difficulty to process (for example, thousands of images or text files). The main challenge of processing nonstructured data is that they are multi-sourced and heterogeneous. In most of cases, the understanding of the data is cognition dependent. In this area, the

core open questions are how to build a uniform platform on which different types of nonstructured data (e.g., mixture of images, text, video and audio) can be processed simultaneously? How to develop the cognition consistent approaches for nonstructured data modeling?

### 3.6 Problem 6: Visualization

Using visual-consistent figures or graphics to exhibit the intrinsic structure and patterns in HD Big Data is challenge visualization analysis. This requires building a basic tool for human–machine interface and expanding applications. For example, by using feature extraction, a HD data space can be transformed into feature space with low dimension (LD), and then by using to visualization techniques, the latter can be turned into visualized space with 2-dimension or 3-dimension). The key concept of judging a good visualization tool is that the end user can easily understand the meaning of Big Data results without knowing any technical analysis behind. Some current visualization techniques used in showcases, such as The Second Life (http://secondlife.com/) and video games, can be effectively applied to Big Data visualization. The core open questions are: is there essential feature extraction of HD data (say, dimension-reduction)? What is structured representation of imaginable thinking? How to construct appropriate visualized space? How to map a problem in feature space (or data space) to a representation problem in visualized space? [16]

## 4 Some Advances on Big Data Research

In HD problems, a progress of the sparse modeling has been made. The sparsity (of $x$) problem can be described as: There exists a characteristic quantity $q(x)$ such that $q(x)$ is of singularity (i.e., smaller than the normal). Three orders to deal with HD problems are illustrated. 1st order is to find the norm format of $q(x)$ as Card $(x)$; 2nd order is to look for all $q(x)$ of {Rank $(\mathbf{X})$, Trace $(\mathbf{X})$, Card$(\mathbf{X})$} such as the minimal Rank $(L)$ + Card $(E)$ s.t. $Y = A(L + E)$. 3rd order is to identify $\mathbf{X}$ where $q(\mathbf{X}) =$ {Trank$(\mathbf{X})$, Card$(\mathbf{X})$}. Among several theories developed for the sparse modeling, the thresholding representation theory based on norm $L1/2$ regularization has now been accepted as a useful method [17–19]. The extensions of such modeling techniques are done as from linear to nonlinear cases, from 1st order problems to higher order ones and from unconstrained to constrained situations.

There are also the progresses of clustering stability made in the HD problems. Clustering analysis is to categorize a data set into subgroups according to data similarity. It can be viewed as the basis of pattern recognition. The fundamental clustering is K-means approach.

If a traditional K-means method is $C = K(D) = \arg\min_S \sum \sum_i \sum_{x \in S_i} d(x - \mu_i)$, then given a data flow $D_t$, a HD setting can be $C_t = K(D_t)$, $D_t \subset R^{n(p_t)}$.

Then, the new challenges are related to variable dimension $(p_t)$, variable sample size n$(p_t)$ and $C_t \to C^*$ (which means consistency and stability).

Chang and Xu [20] have proposed a new modeling with feature decomposability as an optimal clustering. If the data flow is mixture Gaussian distributed, then the sparse K-Means is consistent, i.e., $n(p) \to \infty$ and the optimal solution is stable, or $p \to \infty$.

In real and distributed (R/D) computation problems, there is a finding for feasibility of Hadoop-based regression:

Let $S$ be a global machine, $S_j$ be local machines. Then, a Big Data setting can be that $S$ is too big to process in a central computer. A Hydoop-based regression can be set up by two steps. Step one is to find all regression results of local machines as $f_j$, then the Hadoop-based regression is $f = \frac{1}{m}\sum_j^m f_j$. However, the key is how to estimate $\varepsilon_s(f_s) \leq \varepsilon_s(f_0)$. This problem, in fact, can be addressed by using the random sampling inequality to estimate the hypothesis error. Under certain conditions, a feasibility theory of [20] has showed that the Hydoop-based regression algorithm is feasible in the sense of consistency, i.e., $\varepsilon_s(f^*) - \varepsilon_s(f_\rho) \to 0$.

In nonstructured data processing problems, visual clustering machine becomes a new recognized approach to Big Data. The basic concept is if a data modeling problem is viewed as a cognition problem, then the problem can be solved by simulating visual psychology principles. Leung et al. [21] developed the HL model through visual intuition and transmit it to HD situation by mathematical induction. This finding follows a basic visual principle, in which the distribution of light strength reaching at retina is controlled by the distance between the object and retina, or the curvature of crystalline lens. This method has two phases. The first one is called scale space representation that views the distance or curvature of lens as the scale, the image, i.e., the light strength, of an object can be represented in multiple scales [22,23]. The second phase is called scale space clustering that views a data as a light point and the data set as an image, then one observes the clustering structures from the multi-scale representation of the data image [21]. Note that in this approach, the light blob is a cluster. It corresponds to a set of data, starting from which the same local maximum is reached. A hierarchical clustering procedure was developed to address three basic problems for how to discretize scale? What is the real clustering? Do clusters monotonically evolve? Through defining the lifetime of a cluster, [21] also provided a cognition based solution for "what is real clustering". In addition, the paper proved that the number of cluster centers is monotonically and regularly decreasing. Visual clustering machine has potential in many application areas, such as image segmentation, geographic data analysis, image processing, and protein analysis.

## 5 Remarks

Big Data analysis is still a very pre-mature field at this point. Fundamentally speaking, in order to conduct an applicable Big Data analysis, one should think about how to design Big Data algorithm structure. Here is some ideas open to be discussed. First, a Big Data algorithm should be an algorithm that can process and analyze Big Data under available computational resource and complete in a reasonable time. The Big Data can be handled by it has at least one of following characteristics: large-size, heterogeneous, distributed, multi-sources, data steam, high-dimension, and high-uncertainty. The algorithm can be performed at appropriate degree of time, storage and communi-

cation complexity. It also has some unique properties, such as highly fault toleration, solution integration and assembled capability. Second, the key ideas of designing a Big Data algorithm could include maintaining the proper ratio of data sample and population; simple modeling and simple procedure; inferior preciseness, complex inherence and theory-based. Finally, in addition to well-known statistics or data mining methods, other computational methods, such as set-based processing, stochastic computing, online computing, distributed/parallel computing, cloud computing may be employed to construct a high-efficient Big Data algorithm.

**Acknowledgments** The contents of this work are mainly based on the keynote speeches of the first author at a series of Shuangqing Forums, the research seminars organized by the National Nature Science Foundation of China (NSFC) on Big Data at Beijing, China, 2013–2014 and additional opinions of the second author in his related activities. The work was partially supported by NSFC (Grant Nos. 70921061, 71331005).

## References

1. Laney D (2001) 3D data management: controlling data volume, velocity, and variety. MetaGroup, Stamford
2. Gantz J, Reinsel D (2012) The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the far east. International Data Corporation (IDC), Framingham. www.emc.com/leadership/digital-universe/index.htm
3. National Science Foundation (NSF) (2012) Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), Washington. www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm
4. Wikipedia (2015) Big Data. http://en.wikipedia.org/wiki/Big_data
5. XSSC (2013) Report on the 462nd session: data science and Big Data. In: Xingshan science conferences, May 29–31. Chinese Academy of Sciences, Beijing
6. Shi Y (2014) Big Data: history, current status, and challenges going forward. Bridge 44(4):6–11
7. Laudon KC, Laudon JP (2012) Management information systems. Pearson, Upper Saddle River
8. Shi Y, Zhang LL, Tain YJ, Li XS (2015) Intelligent knowledge: a study beyond data mining. Springer, New York
9. Wikipedia (2015) Engineering. http://en.wikipedia.org/wiki/Engineering
10. Kriegel HP, Kröger P, Renz M, Wurst S (2005) A generic framework for efficient subspace clustering of high-dimensional data. In: IEEE international conference on data mining (ICDM), Houston, pp. 205–257
11. Donoho DL (2006) For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution. Commun Pure Appl Math 56(6):797–829
12. Wang Y, Chang X, Li R, Xu Z (2013) Sparse K-means with Lq($0 \leq q < 1$) constrain for high-dimensional data clustering. In: IEEE international conference on data mining, Dallas
13. Chang X, Wang Y, Li R, Xu Z (2014) Sparse K-means with L∞/Lo penalty for high-dimensional data clustering. arXiv:1403.7890
14. Kleiner A, Talwalkar A, Sarkar P (2012) The Big Data bootstrap. In: The 29th international conference on machine learning, Edinburgh, Scotland, UK
15. Xu Z, Leung KS, Liang Y, Leung Y (2003) Efficiency speed-up strategies for evolutionary computation: fundamentals and fast-Gas. Appl Math Comput 142(2,3):341–388
16. Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y (2010) T-Drive: driving directions based on taxi trajectories. In: ACM SIGSPATIAL GIS
17. Xu ZB (2010) Data modeling: visual psychology approach and L(1/2) regularization theory. In: Proceedings of the international congress of mathematicians, Hyderabad, India
18. Xu ZB, Chang X, Xu F, Zhang H (2012) L(1/2) regularization: a thresholding representation theory and a fast solver. IEEE Neural Netw Learn Syst 23(7):1013–1027
19. Zeng J, Lin S, Wang Y, Xu ZB (2014) L(1/2) regularization: convergence of iterative half thresholding algorithm. IEEE Trans Signal Process 62(9):2317–2329

20. Chang X, Xu C (2014) Feasibility of distributed regression for BigData. In: Workshop on new learning frameworks and models for Big Data, ICML 2014
21. Leung Y, Zhang JS, Xu Z (2000) Clustering by scale-space filtering. IEEE Trans Pattern Anal Mach Intell 22(12):1396–1410
22. Witkin A (1983) Scale space filtering. In: Proceedings of the international joint conference on artificial intelligence
23. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell 12(7):629–639