Data Mining with Big Data

Xindong Wu[1,2], Xingquan Zhu[3], Gong-Qing Wu[2], Wei Ding[4]

[1] School of Computer Science and Information Engineering, Hefei University of Technology, China

[2] Department of Computer Science, University of Vermont, USA

[3] QCIS Center, Faculty of Engineering & Information Technology, University of Technology, Sydney, Australia

[4] Department of Computer Science, University of Massachusetts Boston, USA

Abstract: Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences. This article presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

1. Introduction

Dr. Yan Mo won the 2012 Nobel Prize in Literature. This is probably the most controversial Nobel prize of this category, as Mo speaks Chinese, lives in a socialist country, and has the Chinese government's support. Searching on Google with "Yan Mo Nobel Prize", we get 1,050,000 web pointers on the Internet (as of January 3, 2013). "For all praises as well as criticisms," said Mo recently, "I am grateful." What types of praises and criticisms has Mo actually received over his 31-year writing career? As comments keep coming on the Internet and in various news media, can we summarize all types of opinions in different media in a real-time fashion, including updated, cross-referenced discussions by critics? This type of summarization program is an excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing.

Along with the above example, the era of Big Data has arrived (Nature Editorial 2008; Mervis J. 2012; Labrinidis and Jagadish 2012). Every day, 2.5 quintillion bytes of data are created and 90% of the

data in the world today were produced within the past two years *(IBM 2012).* Our capability for data generation has never been so powerful and enormous ever since the invention of the Information Technology in the early 19th century. As another example, on October 4, 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within two hours (Twitter Blog 2012). Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about Medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 (Michel F. 2012). Assuming the size of each photo is 2 megabytes (MB), this resulted in 3.6 terabytes (TB) storage every single day. As "a picture is worth a thousand words", the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters etc., only if we have the power to harness the enormous amount of data.

The above examples demonstrate the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time". The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions (Rajaraman and Ullman, 2011). In many situations, the knowledge extraction process has to be very efficient and close to real-time because storing all observed data is nearly infeasible. For example, the Square Kilometer Array (SKA) (Dewdney *et al.* 2009) in Radio Astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes(GB)/second data volume, the data generated from the SKA is exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies (Reed *et al.* 2011) can be discovered from the SKA data, existing methods are incapable of handling this Big Data. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast-response and real-time classification for such Big Data.

The remainder of the paper is structured as follows. In Section 2, we propose a HACE theorem to model Big Data characteristics. Section 3 summarizes the key challenges for Big Data mining. Some key research initiatives and the authors' national research projects in this field are outlined in Section 4. Related work is discussed in Section 5, and we conclude the paper in Section 6.

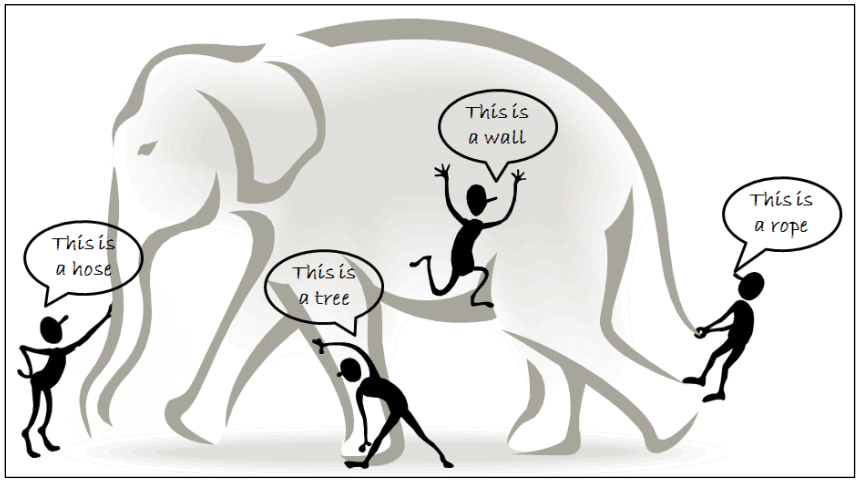

Figure 1: The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

2. Big Data Characteristics: HACE Theorem

HACE Theorem: Big Data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data.

These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Figure 1), which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant according to the part of information he collected during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let's assume that (a) the elephant is growing rapidly and its pose also changes constantly, and (b) the blind men also learn from each other while exchanging information on their respective feelings on the elephant. Exploring the Big Data in this

scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

2.1 Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors use their own schemata for data recording, and the nature of different applications also results in diverse representations of the data. For example, each single human being in a bio-medical world can be represented by using simple demographic information such as gender, age, family disease history etc. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic related test, microarray expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

2.2 Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function

without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data related applications, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/regions. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors. More specifically, the local government regulations also impact on the wholesale management process and eventually result in data representations and data warehouses for local markets.

2.3 Complex and Evolving Relationships

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections which is one of the most important factors of the human society. People form friend circles based on their common hobbies or connections by biological relationships. Such social connections commonly exist in not only our daily activities, but also are very popular in virtual worlds. For example, major social network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process. In the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship representation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big

Data applications, where the key is to take the complex (non-linear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

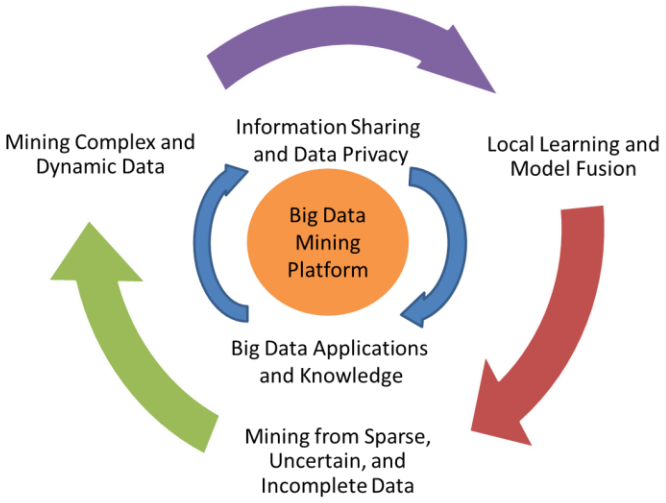

Figure 2: A Big Data processing framework: The research challenges form a three tier structure and center around the "Big Data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms.

3. Data Mining Challenges with Big Data

For an intelligent learning database system (Wu 2000) to handle Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem. Figure 2 shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

The challenges at Tier I focus on data accessing and actual computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For

example, while typical data mining algorithms require all data to be loaded into the main memory, this is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. Sharing sensor network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing mobile users' location information is clearly not acceptable for majority, if not all, applications. In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs. For example, in market basket transactions data, each transaction is considered independent and the discovered knowledge is typically represented by finding highly correlated items, possibly with respect to different temporal and/or spatial restrictions. In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modeling etc. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high level mining algorithm designs.

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics. The circle at Tier III contains three stages. Firstly, sparse, heterogeneous, uncertain, incomplete, and multi-source data are preprocessed by data fusion techniques. Secondly, complex and dynamic data are mined after pre-processing. Thirdly, the global knowledge that is obtained by local learning and model fusion is tested and relevant information is fed back to the pre-processing stage. Then the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

In the following, we elaborate challenges with respect the three tier framework in Figure 2.

3.1 Tier I: Big Data Mining Platform

In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is therefore needed to have efficient access to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the data mining goals. Indeed, many data mining algorithm are designed to handle this type of problem settings. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory. Common solutions are to rely on parallel computing (Shafer et al. 1996; Luo et al. 2012) or collective mining (Chen et al. 2004) to sample and aggregate data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process.

For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with a high performance computing platform, where a data mining task is deployed by running some parallel programming tools, such as MapReduce or ECL (Enterprise Control Language), on a large number of computing nodes (*i.e.*, clusters). The role of the software component is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of samples, is split into many small tasks each of which is running on one or multiple computing nodes. For example, as of this writing, the world most powerful super computer Titan, which is deployed at Oak Ridge National Laboratory in Tennessee, USA, contains 18,688 nodes each with a 16-core CPU.

Such a Big Data system, which blends both hardware and software components, is hardly available without key industrial stockholders' support. In fact, for decades, companies have been making business decisions based on transactional data stored in relational databases. Big Data mining offers opportunities to go beyond their relational databases to rely on less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information. Major business intelligence companies, such IBM, Oracle, Teradata etc., have all featured their own products to help customers acquire and

organize these diverse data sources and coordinate with customers' existing data to find new insights and capitalize on hidden relationships.

3.2 Tier II: Big Data Semantics and Application Knowledge

Semantics and application knowledge in Big Data refer to numerous aspects related to the regulations, policies, user knowledge, and domain information. The two most important issues at this tier include (1) data sharing and privacy; and (2) domain and application knowledge. The former provides answers to resolve concerns on how data are maintained, accessed, and shared; whereas the latter focuses on answering questions like "what are the underlying applications ?" and "what are the knowledge or patterns users intend to discover from the data ?".

3.2.1 Information Sharing and Data Privacy

Information sharing is an ultimate goal for all systems involving multiple parties (Howe et al. 2008). While the motivation for sharing is clear, a real-world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records, and so simple data exchanges or transmissions do not resolve privacy concerns (Duncan 2007, Huberman 2012, Schadt 2012). For example, knowing people's locations and their preferences, one can enable a variety of useful location-based services, but public disclosure of an individual's movements over time can have serious consequences for privacy. To protect privacy, two common approaches are to (1) restrict access to the data, such as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and (2) anonymize data fields such that sensitive information cannot be pinpointed to an individual record (Cormode and Srivastava 2009). For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconducted by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. For example, the most common k-anonymity privacy measure is to ensure that each individual in the database must be indistinguishable from k−1 others. Common anonymization approaches are to use suppression, generalization, perturbation, and permutation to generate an altered version of the data, which is, in fact, some uncertain data.

One of the major benefits of the data annomization based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrict access controls. This naturally leads to another research area namely privacy preserving data mining (Lindell and Pinkas 2000), where multiple parties, each holding some sensitive data, are trying to achieve a data mining goal without sharing any sensitive information inside the data. This privacy preserving mining goal, in practice, can be solved through two types of approaches including (1) using some communication protocols, such as Yao's protocol (Yao 1986), to request the distributions of the whole dataset, rather than requesting the actual values of each record, or (2) to design some special data mining methods to derive knowledge from anonymized data (this is inherently similar to the uncertain data mining methods).

3.2.2 Domain and Application Knowledge

Domain and application knowledge (Kopanas et al. 2002) provides essential information for designing Big Data mining algorithms and systems. In a simple case, domain knowledge can help identify right features for modeling the underlying data (e.g., blood glucose level is clearly a better feature than body mass in diagnosing Type II diabetes). The domain and application knowledge can also help design achievable business objectives by using Big Data analytical techniques. For example, stock market data are a typical domain which constantly generates a large quantity of information, such as bids, buys, and puts, in every single second. The market continuously evolves and is impacted by different factors, such as domestic and international news, government reports, and natural disasters etc. An appealing Big Data mining task is to design a Big Data mining system to predict the movement of the market in the next one or two minutes. Such systems, even if the prediction accuracy is just slightly better than random guess, will bring significant business values to the developers (Bughin et al. 2010). Without correct domain knowledge, it is a clear challenge to find effective matrices/measures to characterize the market movement, and such knowledge is often beyond the mind of the data miners, although some recent research has shown that using social networks, such as Twitter, it is possible to predict the stock market upward/downward trends (Bollen et al. 2011) with good accuracies.

3.3 Tier III: Big Data Mining Algorithms

3.3.1 Local Learning and Model Fusion for Multiple Information Sources

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each different site often leads to biased decisions or models, just like the elephant and blind men case. Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two-step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthetized by aggregating patterns across all sites (Wu and Zhang 2003). At the knowledge level, model correlation analysis investigates the relevance between models generated from different data sources to determine how relevant the data sources are correlated to each other, and how to form accurate decisions based on models built from autonomous sources.

3.3.2 Mining from Sparse, Uncertain, and Incomplete Data

Spare, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues, where data in a high dimensional space (such as more than 1000 dimensions) does not show clear trends or distributions. For most machine learning and data mining algorithms, high dimensional spare data significantly deteriorate the difficulty and the reliability of the models derived from the data. Common approaches are to employ dimension reduction or feature selection (Wu et al. 2012) to reduce the data dimensions or to carefully include additional samples to decrease the data

scarcity, such as generic unsupervised learning methods in data mining.

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain specific applications with inaccurate data readings and collections. For example, data produced from GPS equipment is inherently uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels (such as 1 meter). As a result, each recording location is represented by a mean value plus a variance to indicate expected errors. For data privacy related applications (Mitchell 2009), users may intentionally inject randomness/errors into the data in order to remain anonymous. This is similar to the situation that an individual may not feel comfortable to let you know his/her exact income, but will be fine to provide a rough range like [120k, 160k]. For uncertain data, the major challenge is that each data item is represented as some sample distributions but not as a single value, so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters. For example, error aware data mining (Wu and Zhu 2008) utilizes the mean and the variance values with respect to each single data item to build a Naïve Bayes model for classification. Similar approaches have also been applied for decision trees or database queries.

Incomplete data refers to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (*e.g.*, dropping some sensor node readings to save power for transmission). While most modern data mining algorithms have inbuilt solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field which seeks to impute missing values in order to produce improved models (compared to the ones built from the original data). Many imputation methods (Efron 1994) exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance.

3.3.3 Mining Complex and Dynamic Data

The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature (Birney 2012). Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks etc. are all featured with complex data. While

complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving. For example, researchers have successfully used Twitter, a well-known social networking facility, to detect events such as earthquakes and major social activities, with nearly online speed and very high accuracy. In addition, the knowledge of people's queries to search engines also enables a new early warning system for detecting fast spreading flu outbreaks (Helft 2008). Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. Such a connection is quadratic with respect to the number of nodes in the network, so a million node network may be subject to one trillion connections. For a large social network site, like Facebook, the number of active users has already reached 1 billion, and analyzing such an enormous network is a big challenge for Big Data mining. If we take daily user actions/interactions into consideration, the scale of difficulty will be even more astonishing.

Inspired by the above challenges, many data mining methods have been developed to find interesting knowledge from Big Data with complex relationships and dynamically changing volumes. For example, finding communities and tracing their dynamically evolving relationships are essential for understanding and managing complex systems (Aral and Walker 2012, Centola 2010). Discovering outliers in a social network (Borgatti et al. 2009) is the first step to identify spammers and provide safe networking environments to our society.

If only facing with huge amounts of structured data, users can solve the problem simply by purchasing more storage or improving storage efficiency. However, Big Data complexity is represented in many aspects, including complex heterogeneous data types, complex intrinsic semantic associations in data, and complex relationship networks among data. That is to say, the value of Big Data is in its complexity.

Complex heterogeneous data types: In Big Data, data types include structured data, unstructured data, and semi-structured data etc. Specifically, there are tabular data (relational databases), text, hyper-text, image, audio and video data etc. The existing data models include key-value stores, bigtable clones, document databases, and graph database, which are listed in an ascending order of the complexity of these data models. Traditional data models are incapable of handling complex data in the context of Big Data. Currently, there is no acknowledged effective and efficient data model to handle Big Data.

Complex intrinsic semantic associations in data: news on the Web, comments on Twitter, pictures on Flicker, and clips of video on YouTube may discuss about an academic award-winning event at the same time. There is no doubt that there are strong semantic associations in these data. Mining complex semantic associations from "text-image-video" data will significantly help improve application system performance such as search engines or recommendation systems. However, in the context of Big Data, it is a great challenge to efficiently describe semantic features and to build semantic association models to bridge the semantic gap of various heterogeneous data sources.

Complex relationship networks in data: In the context of Big Data, there exist relationships between individuals. On the Internet, individuals are webpages and the pages linking to each other via hyperlinks form a complex network. There also exist social relationships between individuals forming complex social networks, such as big relationship data from Facebook, Twitter, LinkedIn and other social media [Banerjee and Agarwal 2012, Chen et al. 2012, Zhao et al. 2012], including call detail records (CDR), devices and sensors information [Ahmed and Karypis 2012, Silva et al. 2012], GPS and geo-coded map data, massive image files transferred by the Manage File Transfer protocol, Web text and click-stream data [Alam et al. 2012], scientific information, e-mail [Liu and Wang 2012], etc. To deal with complex relationship networks, emerging research efforts have begun to address the issues of structure-and-evolution, crowds-and-interaction, and information-and-communication.

The emergence of Big Data has also spawned new computer architectures for real-time data-intensive processing, such as the open source project Apache Hadoop which runs on high-performance clusters. The size or complexity of the Big Data, including transaction and interaction data sets, exceeds a regular technical capability in capturing, managing, and processing these data within reasonable cost and time limits. In the context of Big Data, real-time processing for complex data is a very challenging task.

4. Research Initiatives and Projects

In order to tackle the Big Data challenges and "seize the opportunities afforded by the new, data driven resolution", the US National Science Foundation (NSF), under President Obama Administration's Big Data initiative, announced the BIGDATA solicitation in 2012. Such a federal initiative has resulted in a number of winning projects to investigate the foundations for Big Data management (led by the

University of Washington), analytical approaches for genomics based massive data computation (led by Brown University), large scale machine learning techniques for high-dimensional datasets which may be as large as 500,000 dimensions (led by Carnegie Mellon University), social analytics for large-scale scientific literatures (led by Rutgers University), and several others. These projects seek to develop methods, algorithms, frameworks, and research infrastructures which allow us to bring the massive amounts of data down to a human manageable and interpretable scale. Other countries such as the National Natural Science Foundation of China (NSFC) are also catching up with national grants on Big Data research.

Meanwhile, since 2009, the authors have taken the lead in the following national projects that all involve Big Data components:

- ✧ Integrating and Mining Bio-Data from Multiple Sources in Biological Networks, sponsored by the U.S. National Science Foundation (NSF), Medium Grant No. CCF-0905337, October 1, 2009 - September 30, 2013.

  Issues and significance: We have integrated and mined bio-data from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission and processing for information networks. We have developed methods for semantic-based data integration, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

- ✧ Big Data Fast Response: Real-time Classification of Big Data Stream, sponsored by the Australian Research Council (ARC), Grant No. DP130102748, January 1, 2013 – Dec. 31 2015.

  Issues and significance: We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making. The key challenges and research issues include: (1) designing Big Data sampling mechanisms to reduce Big Data volumes to manageable size for processing; (2) building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data, as well as

accurately predict the trend of the data in the future; and (3) a knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

- ✧ Pattern Matching and Mining with Wildcards and Length Constraints, sponsored by the National Natural Science Foundation of China (NSFC), Grant Nos. 60828005 (Phase 1, January 1, 2009 - December 31, 2010) and 61229301 (Phase 2, January 1, 2013 - December 31, 2016).

  Issues and significance: We perform a systematic investigation on pattern matching, pattern mining with wildcards, and application problems as follows: (1) exploration of the NP-hard complexity of the matching and mining problems, (2) multiple pattern matching with wildcards, (3) approximate pattern matching and mining, and (4) application of our research onto ubiquitous personalized information processing and bioinformatics.

- ✧ Key Technologies for Integration and Mining of Multiple, Heterogeneous Data Sources, sponsored by the National High Technology Research and Development Program (863 Program) of China, Grant No. 2012AA011005, January 1, 2012 - December 31, 2014.

  Issues and significance: We have performed an investigation on the availability and statistical regularities of multi-source, massive and dynamic information, including cross-media search based on information extraction, sampling, uncertain information querying, and cross-domain and cross-platform information polymerization. In order to break through the limitations of traditional data mining methods, we have studied heterogeneous information discovery and mining in complex inline data, mining in data streams, multi-granularity knowledge discovery from massive multi-source data, distribution regularities of massive knowledge, quality fusion of massive knowledge.

- ✧ Group Influence and Interactions in Social Networks, sponsored by the National Basic Research 973 Program of China, Grant No. 2013CB329604, January 1, 2013 - December 31, 2017.

  Issues and significance: We have studied group influence and interactions in social networks, including (1) employing group influence and information diffusion models, and deliberating group interaction rules in social networks using dynamic game theory, (2) studying interactive individual selection and effect evaluations under social networks affected by group emotion, and analyzing emotional interactions and influence among

individuals and groups, and (3) establishing an interactive influence model and its computing methods for social network groups, in order to reveal the interactive influence effects and evolution of social networks.

5. Related Work

Big Data Mining Platforms (Tier I): Due to the multi-source, massive, heterogeneous and dynamic characteristics of application data involved in a distributed environment, one of the important characteristics of Big Data is computing tasks on the petabytes (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computer infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed PB, even EB-level data are the critical goal for Big Data processing to change from "quantity" to "quality".

Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. How to improve the performance of MapReduce and enhance the real-time nature of large-scale data processing is a hot topic in research. The MapReduce parallel programming model has been applied in many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for getting the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. In order to improve the efficiency of algorithms, Chu et al. proposed a general-purpose parallel programming method which is applicable to a large number of machine learning algorithms based on the simple MapReduce programming model on multi-core processors. 10 classic data mining algorithms are realized in the framework, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear support vector machines, the independent variable analysis, Gaussian discriminant analysis, expectation maximization and back-propagation neural networks [Chu et al., 2006]. With the analysis of these classical machine learning algorithms, we argue that the computational operations in the algorithm learning process could be transformed into a summation operation on a number of training data sets. Summation operations could

be performed on different subsets independently and achieve penalization executed easily on the MapReduce programming platform. Therefore, a large-scale data set could be divided into several subsets and assigned to multiple Mapper nodes. Then various summation operations could be performed on these Mapper nodes to get intermediate results. Finally, learning algorithms are parallel executed through merging summation of Reduce nodes. Ranger et al. [2007] proposed a MapReduce-based application programming interface Phoenix, which supports parallel programming in the environment of multi-core and multi-processor systems, and realized three data mining algorithms including k-Means, principal component analysis, and linear regression. Gillick et al. [2006] improved the MapReduce's implementation mechanism in Hadoop, evaluated the algorithms' performance of single-pass learning, iterative learning and query-based learning in the MapReduce framework, studied how to share data between computing nodes involved in parallel learning algorithms, how to deal with distributed storage data, and then showed that the MapReduce mechanisms suitable for large-scale data mining by testing series of standard data mining tasks on medium-size clusters. Papadimitriou and Sun [2008] proposed a distributed collaborative aggregation (DisCo) framework using practical distributed data preprocessing and collaborative aggregation techniques. The implementation on Hadoop in an open source MapReduce project showed that DisCo has perfect scalability and can process and analyze massive data sets (with hundreds of GB).

For the weak scalability of traditional analysis software and the poor analysis capabilities of Hadoop, Das et al. [2010] conducted a study of the integration of R (open source statistical analysis software) and Hadoop. The in-depth integration pushes data computation to parallel processing, which makes Hadoop obtain powerful deep analysis capabilities. Wegener et al. [2009] achieved the integration of Weka (an open-source machine learning and data mining software tool) and MapReduce. Standard Weka tools can only run on a single machine, and cannot go beyond the limit of 1GB of memory. After algorithm parallelization, Weka breaks through the limitations and improves performance by taking the advantage of parallel computing, and it can handle more than 100GB data on MapReduce clusters. Ghoting et al. [2009] proposed Hadoop-ML, on which developers can easily build task-parallel or data-parallel machine learning and data mining algorithms on program blocks under the language run-time environment.

Big Data Semantics and Application Knowledge (Tier II): In privacy protection of massive data, Ye, et al, (2013) proposed a multi-layer rough set model, which can accurately describe the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criteria in the anonymization process, and designed a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization / refinement order for classification. A recent paper on confidentiality protection in Big Data (Machanavajjhala and Reiter 2012) summarizes a number of methods for protecting public release data, including aggregation (such as k-anonymity, I-diversity etc.), suppression (*i.e.*, deleting sensitive values), data swapping (*i.e.*, switching values of sensitive data records to prevent users from matching), adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions.

For applications involving Big Data and tremendous data volumes, it is often the case that data are physically distributed at different locations, which means that users no longer physically possess the storage of their data. To carry out Big Data mining, having an efficient and effective data access mechanism is vital, especially for users who intend to hire a third party (such as data miners or data auditors) to process their data. Under such a circumstance, users' privacy concerns may include (1) no local data copies or downloading, (2) all analysis must be deployed based on the existing data storage systems without violating existing privacy settings, and many others. In Wang et al. (2013), a privacy-preserving public auditing mechanism for large scale data storage (such as cloud computing systems) has been proposed. The public key based mechanism is used to enable Third Party Auditing (TPA), so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy.

For most Big Data applications, privacy concerns focuse on excluding the third party (such as data miners) from directly accessing the original data. Common solutions are to rely on some privacy-preserving approaches or encryption mechanisms to protect the data. A recent effort by Lorch et al (2013) indicates that users' "data access patterns" can also have severe data privacy concerns and lead to disclosures of geographically co-located users or users with common interests (e.g., two users searching for the same map locations are likely to be geographically co-located). In their system, namely Shround, it hides data access patterns from the servers by using virtual disks. As a result, it can support a variety of

Big Data applications, such as microblog search and social network queries, without compromising the user privacy.

Big Data Mining Algorithms (Tier III): In order to adapt to the multi-source, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods [Chang et al., 2009], designing a data mining mechanism from a multi-source perspective [Wu and Zhang, 2003; Wu et al., 2005; Zhang et al., 2005], as well as the study of dynamic data mining methods and the analysis of convection data [Domingos and Hulten, 2000; Chen et al, 2005]. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data. Due to massive data typically coming from different data sources, the knowledge discovery of the massive data must be performed using a multi-source mining mechanism. As real-world data often come as a data stream or a characteristic flow, a well-established mechanism is needed to discover knowledge and master the evolution of knowledge in the dynamic data source. Therefore, the massive, heterogeneous and real-time characteristics of multi-source data provide essential differences between single-source knowledge discovery and multi-source data mining.

Wu et al. [Wu and Zhang, 2003; Wu, et al, 2005; Su et al, 2006] proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multi-source data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find. Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing.

Data streams are widely used in financial analysis, online trading, medical testing, and so on. Static knowledge discovery methods cannot adapt to the characteristic of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream mining [Domingos and Hulten, 2000].

Knowledge evolution is a common phenomenon in real-world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution of cardiovascular and other chronic epidemiological changes with the passage of time. In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by context changes in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple feature, and streaming features [Wu et al., 2013].

6. Conclusions

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are (1) huge with **h**eterogeneous and diverse data sources, (2) **a**utonomous with distributed and decentralized control, and (3) **c**omplex and **e**volving in data and knowledge associations. Such combined characteristics suggest that Big Data requires a "big mind" to consolidate data for maximum values (Jacobs 2009).

In order to explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high performance computing platforms are required which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and

data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

Acknowledgements

This work is supported by the National 863 Program of China (2012AA011005), the National 973 Program of China (2013CB329604), the National Natural Science Foundation of China (NSFC 61229301, 61273297, and 61273292), the US National Science Foundation (NSF CCF-0905337), and the Australian Research Council (ARC) Future Fellowship (FT100100971). The authors would like to thank the anonymous reviewers for their valuable and constructive comments on improving the paper.

References

1) Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 603-630
2) Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, SangKeun Lee, Novel approaches to crawling important pages early, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 707-734
3) Aral S. and Walker D. 2012, Identifying influential and susceptible members of social networks, *Science*, vol.337, pp.337-341.
4) Machanavajjhala and Reiter 2012, Ashwin Machanavajjhala, Jerome P. Reiter: Big privacy: protecting confidentiality in big data. *ACM Crossroads*, 19(1): 20-23, 2012.
5) Banerjee and Agarwal 2012, Soumya Banerjee, Nitin Agarwal, Analyzing collective behavior from blogs using swarm intelligence, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 523-547
6) Birney E. 2012, The making of ENCODE: Lessons for big-data projects, *Nature*, vol.489, pp.49-51.
7) Bollen et al. 2011, J. Bollen, H. Mao, and X. Zeng, Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, 2(1):1-8, 2011.

8) Borgatti S., Mehra A., Brass D., and Labianca G. 2009, Network analysis in the social sciences, *Science*, vol. 323, pp.892-895.
9) Bughin et al. 2010, J Bughin, M Chui, J Manyika, *Clouds, big data, and smart assets: Ten tech-enabled business trends to watch*, McKinSey Quarterly, 2010.
10) Centola D. 2010, The spread of behavior in an online social network experiment, *Science*, vol.329, pp.1194-1197.
11) *Chang et al., 2009*, Chang E.Y., Bai H., and Zhu K., Parallel algorithms for mining large-scale rich-media data, In: *Proceedings of the 17th ACM International Conference on Multimedia* (*MM '09*), New York, NY, USA, 2009, pp. 917-918.
12) Chen et al. 2004, R. Chen, K. Sivakumar, and H. Kargupta, Collective Mining of Bayesian Networks from Distributed Heterogeneous Data, *Knowledge and Information Systems*, 6(2):164-187, 2004.
13) Chen et al. 2012, Yi-Cheng Chen, Wen-Chih Peng, Suh-Yin Lee, Efficient algorithms for influence maximization in social networks, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 577-601
14) Chu et al., 2006, Chu C.T., Kim S.K., Lin Y.A., Yu Y., Bradski G.R., Ng A.Y., Olukotun K., Map-reduce for machine learning on multicore, In: *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS '06)*, MIT Press, 2006, pp. 281-288.
15) Cormode G. and Srivastava D. 2009, Anonymized Data: Generation, Models, Usage, in Proc. of *SIGMOD*, 2009. pp. 1015-1018.
16) *Das et al., 2010*, Das S., Sismanis Y., Beyer K.S., Gemulla R., Haas P.J., McPherson J., Ricardo: Integrating R and Hadoop, In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10)*, 2010, pp. 987-998.
17) Dewdney P., Hall P., Schilizzi R., and Lazio J. 2009, The square kilometre Array, *Proc. of IEEE*, vol.97, no.8.
18) *Domingos and Hulten, 2000*, Domingos P. and Hulten G., Mining high-speed data streams, In: *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD΄00)*, 2000, pp. 71-80.
19) Duncan G. 2007, Privacy by design, *Science*, vol. 317, pp.1178-1179.
20) Efron B. 1994, Missing data, imputation, and the Bootstrap, *Journal of the American Statistical Association*, vol.89, no.426, pp.463-475.
21) *Ghoting et al., 2009*, Ghoting A., Pednault E., Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics, In: *Proceedinds of the Large-Scale Machine Learning: Parallelism and Massive Datasets Workshop (NIPS-2009)*.
22) *Gillick et al., 2006*, Gillick D., Faria A., DeNero J., *MapReduce: Distributed Computing for Machine Learning*, Berkley, December 18, 2006.
23) Helft M. 2008, Google uses searches to track Flu's spread, The New York Times, http://www.nytimes.com/2008/11/12/technology/internet/12flu.html.
24) Howe D. et al. 2008, Big data: the future of biocuration, *Nature*, 455, pp.47-50, Sept. 2008.
25) Huberman B. 2012, Sociology of science: Big data deserve a bigger audience, *Nature*, vol. 482, pp.308.
26) IBM 2012, What is big data: Bring big data to the enterprise, http://www-01.ibm.com/software/data/bigdata/, IBM.
27) Jacobs A. 2009, The pathologies of big data, *Communication of the ACM*, vol.52, no.8, pp.36-44.
28) Kopanas et al. 2002, I. Kopanas, N. Avouris, and S. Daskalaki, The Role of Domain Knowledge in a Large Scale Data Mining Project, in I.P Vlahavas, C.D. Spyropoulos (eds), *Methods and Applications*

of Artificial Intelligence, Lecture Notes in AI, LNAI no. 2308, pp. 288-299, Springer-Verlag, Berlin, 2002.
29) Labrinidis and Jagadish 2012, A. Labrinidis and H. Jagadish, Challenges and Opportunities with Big Data, In *Proc. of the VLDB Endowment*, 5(12):2032-2033, 2012.
30) Lindell Y. and Pinkas B. 2000, Privacy Preserving Data Mining, *Journal of Cryptology*, pp.36-54.
31) Liu and Wang 2012, Wuying Liu, Ting Wang, Online active multi-field learning for efficient email spam filtering, *Knowledge and Information Systems*, October 2012, Volume 33, Issue 1, pp 117-136
32) Lorch et al, 2013, J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center, In: *Proc. of the 11th USENIX Conference on File and Storage Technologies (FAST'13)*, San Jose, CA, 2013.
33) Luo et al. 2012, Dijun Luo, Chris Ding, Heng Huang, Parallelization with Multiplicative Algorithms for Big Data Mining, In: Proc. of *IEEE 12th International Conference on Data Mining*, pp.489-498, 2012
34) Mervis J. 2012, U.S. SCIENCE POLICY: Agencies Rally to Tackle Big Data, *Science*, vol.336, no.6077, pp.22.
35) Michel F. 2012, How many photos are uploaded to Flickr every day and month? http://www.flickr.com/photos/franckmichel/6855169886/.
36) Mitchell T. 2009, Mining our reality, *Science*, vol.326, pp.1644-1645.
37) Nature Editorial 2008, Community cleverness required, *Nature*, Vol.455, no.7209, Sept. 4, 2008.
38) *Papadimitriou and Sun, 2008*, Papadimitriou S., Sun J., Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining. In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*, 2008, pp. 512-521.
39) *Ranger et al., 2007*, Ranger C., Raghuraman R., Penmetsa A., Bradski, G., and Kozyrakis C., Evaluating MapReduce for multi-core and multiprocessor systems, In: *Proceedings of the 13th IEEE International Symposium on High Performance Computer Architecture (HPCA '07)*, 2007, pp. 13-24.
40) Rajaraman and Ullman, 2011, A. Rajaraman and J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2011.
41) Reed C., Thompson D., Majid W., and Wagstaff K. 2011, Real time machine learning to find fast transient radio anomalies: A semi-supervised approach combining detection and RFI excision, *Int'l Astronomical Union Sym. on Time Domain Astronomy*, UK. Sept. 2011
42) Schadt E. 2012, The changing privacy landscape in the era of big data, *Molecular Systems*, 8, Article number 612.
43) Shafer et al. 1996, J. Shafer, R. Agrawal, and M. Mehta, SPRINT: A Scalable Parallel Classifier for Data Mining, In: *Proc. of the 22nd VLDB Conference*, Mumbai, India, 1996.
44) Silva et al. 2012, Alzennyr da Silva, Raja Chiky, Georges Hébrail, A clustering approach for sampling data streams in sensor networks, *Knowledge and Information Systems*, July 2012, Volume 32, Issue 1, pp 1-23
45) *Su et al., 2006*, Su K., Huang H., Wu X., and Zhang S., A logical framework for identifying quality knowledge from different data sources, *Decision Support Systems*, 2006, 42(3): 1673-1683.
46) Twitter Blog 2012, Dispatch from the Denver debate, http://blog.twitter.com/2012/10/dispatch-from-denver-debate.html, October 2012.
47) *Wegener et al., 2009*, Wegener D., Mock M., Adranale D., Wrobel S., Toolkit-Based high-performance data mining of large data on MapReduce clusters, In: *Proceedings of the ICDM Workshop*, 2009, pp. 296-301.

48) Wang et al. 2013, Qian Wang; Kui Ren; Wenjing Lou, Privacy-Preserving Public Auditing for Data Storage Security in Could Computing, *IEEE Transactions on Computers*, 62(2):362-375, 2013.
49) Wu X. and Zhu X. 2008, Mining with Noise Knowledge: Error-Aware Data Mining, *IEEE Transactions on Systems, Man and Cybernetics*, Part A, vol.38, no.4, pp.917-932.
50) Wu X. and Zhang S. 2003, Synthesizing High-Frequency Rules from Different Data Sources, *IEEE Transactions on Knowledge and Data Engineering*, vol.15, no.2, pp.353-367.
51) *Wu et al., 2005*, Wu X., Zhang C., and Zhang S., Database classification for multi-database mining, *Information Systems*, 2005, 30(1): 71-88.
52) Wu X. 2000, Building Intelligent Learning Database Systems, *AI Magazine*, vol.21, no.3, pp.61-67.
53) *Wu et al., 2013*, Wu X., Yu K., Ding W., Wang H., and Zhu X., Online feature selection with streaming features, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(5):1178-1192, 2013.
54) Yao A. 1986, How to generate and exchange secretes, in *Proc. Of 27th FOCS Conference*, pp.162-167.
55) *Ye et al., 2013*, Ye M., Wu X., Hu X., Hu D., Anonymizing classification data using rough set theory, *Knowledge-Based Systems*, 43: 82-94, 2013,.
56) Zhao et al. 2012, Jichang Zhao, Junjie Wu, Xu Feng, Hui Xiong, Ke Xu, Information propagation in online social networks: a tie-strength perspective, *Knowledge and Information Systems*, September 2012, Volume 32, Issue 3, pp 589-608.

Xindong Wu is a Yangtze River Scholar in the School of Computer Science and Information Engineering at the Hefei University of Technology (China), a Professor of Computer Science at the University of Vermont (USA), and a Fellow of the IEEE and AAAS. He received his Bachelor's and Master's degrees in Computer Science from the Hefei University of Technology, China, and his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, Britain. His research interests include data mining, knowledge-based systems, and Web information exploration.

Dr. Wu is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), and a Series Editor of the Springer Book Series on Advanced Information and Knowledge Processing (AI&KP). He was the Editor-in-Chief of the IEEE Transactions on Knowledge and Data Engineering (TKDE, by the IEEE Computer Society) between 2005 and 2008. He served as Program Committee Chair/Co-Chair for ICDM '03 (the 2003 IEEE International Conference on Data Mining), KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), and CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management).

Xingquan Zhu received his Ph.D. degree in Computer Science from Fudan University, Shanghai China. He is a recipient of the Australia ARC Future Fellowship and a Professor of the Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS), Australia. Dr. Zhu's research mainly focuses on data mining, machine learning, and

multimedia systems. Since 2000, he has published more than 160 refereed journal and conference proceedings papers in these areas. Dr. Zhu was an Associate Editor of the IEEE Transactions on Knowledge and Data Engineering (2008-2012), and a Program Committee Co-Chair for the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2011) and the 9th International Conference on Machine Learning and Applications (ICMLA 2010). He also served as Conference Co-Chair for ICMLA 2012.

Gongqing Wu is an Associate Professor of Computer Science at the Hefei University of Technology (China). He received his Bachelor's degree from Anhui Normal University (China), his Master's degree from the University of Science and Technology of China (USTC), and his Ph.D. degree from the Hefei University of Technology (China), all in Computer Science. His research interests include data mining and Web intelligence. He has published over 20 refereed research papers. He is the recipient of a Best Paper Award at the 2011 IEEE International Conference on Tools with Artificial Intelligence (ICTAI) and a Best Paper Award at the 2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI). His research is currently sponsored by the National 863 Program of China and the National Natural Science Foundation of China (NSFC).

Wei Ding received her Ph.D. degree in Computer Science from the University of Houston in 2008. She has been an Assistant Professor of Computer Science in the University of Massachusetts Boston since 2008. Her main research interests include Data Mining, Machine Learning, Artificial Intelligence, Computational Semantics, and with applications to astronomy, geosciences, and environmental sciences. She has published more than 70 refereed research papers, 1 book, and has 1 patent. She is an Associate Editor of Knowledge and Information Systems (KAIS) and an editorial board member of the Journal of System Education (JISE). She is the recipient of a Best Paper Award at IEEE International Conference on Tools with Artificial Intelligence (ICTAI) 2011, a Best Paper Award at IEEE International Conference on Cognitive Informatics (ICCI) 2010, a Best Poster Presentation award at ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPAITAL GIS) 2008, and a Best Ph.D. Work Award between 2007 and 2010 from the University of Houston. Her research projects are currently sponsored by NASA and DOE.