2015-10-08

A survey of big data research

Hua (Julia) Fang
University of Massachusetts Medical School

Zhaoyang Zhang
University of Massachusetts Medical School

Chanpaul Jin Wang
University of Massachusetts Medical School

See next page for additional authors

A survey of big data research

Authors
Hua (Julia) Fang, Zhaoyang Zhang, Chanpaul Jin Wang, Mahmoud Daneshmand, Chonggang Wang, and Honggang Wang

Keywords
Big data, Bioinformatics, Business, Data visualization, Genomics, Medical services, Research and development, UMCCTS funding

A SURVEY ON BIG DATA RESEARCH

Hua Fang, Zhaoyang Zhang, Chanpaul Jin Wang, Mahmoud Daneshmand, Chonggang Wang, Honggang Wang

[†]Corresponding Author's Address:
Honggang Wang
Department of Electrical and Computer Engineering
University of Massachusetts Dartmouth
North Dartmouth, USA
Tel: +001(508) 999-8469
Email: hwang1@umassd.edu

Hua Fang (e-mail: hua.fang@umassmed.edu) is with the University of Massachusetts Medical School, MA, USA. Zhaoyang Zhang (e-mail: Zhaoyang.Zhang@umassmed.edu) and Chanpaul Jin Wang (email: Chanpaul.Wang@umassmed.edu) are with the University of Massachusetts Medical School& University of Massachusetts Dartmouth. Mahmoud Daneshmand (e-mail: daneshmand@ieee.org) is with Stevens Institute of Technology. Chonggang Wang (e-mail: cgwang@ieee.org) is with Interdigital Communications. Honggang Wang (e-mail: hwang1@umassd.edu) is with the Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, USA.

The paper was submitted October 29, 2015.

Abstract

Big data create values for business and research, but pose significant challenges in terms of networking, storage, management, analytics and ethics. Multidisciplinary collaborations from engineers, computer scientists, statisticians and social scientists are needed to tackle, discover and understand big data. This survey presents an overview of big data initiatives, technologies and research in industries and academia, and discusses challenges and potential solutions.

I. INTRODUCTION

Big data are defined in various ways but the three "V" features are their common characteristics: Volume (large datasets), variety (different types of data from myriad sources); and velocity (data collected in real time). Variability and complexity are considered as two other features especially by analytic areas. Big data require new forms of processing to enable enhanced decision making, insight discovery and process optimization. These large-scale data can be produced on the web, by sensors or monitoring systems [1]. For example, 2.7 Zetabytes data exist in the digital universe; 235 Terabytes data have been collected by the U.S. Library of Congress in April 2011; business transactions on the internet, business-to-business and business-to-consumer by 2020, will reach 450 billion per day. The term of big data is also used to capture the opportunities and challenges facing all researchers in accessing, managing, analyzing, and integrating datasets of diverse data types. The rapid growth in data size and scope created a need for multi-disciplinary collaboration and joint efforts from industries, academics and governments to develop novel methods, disciplines and workforce that can blend data networking, management, computational and statistical sciences. This multi-disciplinary collaboration initiative has been launched at the prestigious 2014 Joint Statistical Meetings, American Statistical Association, where top computer scientists, engineers and statisticians share respective approaches to Big Data models, algorithms and network [2].

This paper presents a survey of the state of the art in the big data area, discusses the challenges and solutions in industries and academics from the perspectives of engineers, computer scientists and statisticians. The rest of the paper is organized as follows: Section 2 surveys technologies in big data networking, storage and management; Section 3 introduces analytic research; and Section 4 presents the future trends and challenges in the big data area.

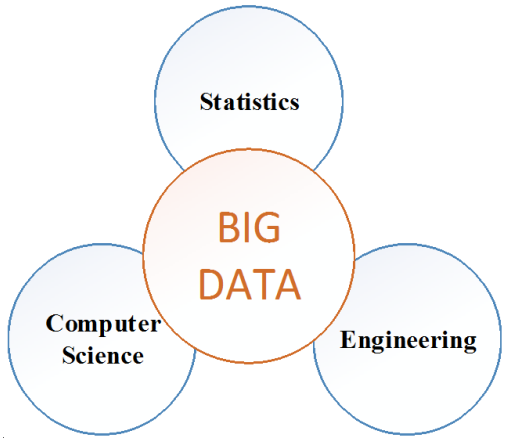


Fig. 1: A Multidisciplinary Approach to Big Data.

II. Big Data in Computer Science and Engineering

This section discusses the big data research and applications from the work of computer scientists and engineers in academia and industries. It is primarily based on publications from the Association for Computer Machinery (ACM), IEEE Xplore Digital Library and Google Scholar using keywords such as big data, large-scale or high-dimensional data.

A. Big Data in Networking

Recent big data networking studies focus on two areas, networking architecture and network optimization, mostly involving Software-Defined Networking (SDN) and cloud computing. SDN architecture is directly programmable, agile, centrally managed, programmatically configured, open standards-based and vendor-neutral. It is dynamic, manageable, cost-effective, and adaptable, suitable for the high-bandwidth and dynamic nature of today's applications. For example, Monga et al. used SDN to construct big-data networking architectural models from campus to WAN. Wang et al. studied the run-time SDN configuration to jointly optimize application performance and network utilization. Das et al. proposed a network management framework (FlowComb) to achieve high utilization and low data processing time for big data. FlowComb detects network transfers, and adapts the network by changing the paths in response to these transfers. Ferguson et al. proposed a unified architecture, PANE, which built API interfaces for the applications atop of the SDN network, allowing these applications to directly operate the SDN network. For network optimization, MapReduce Scheduling was one of methods under

study. An comprehensive study on the existing works of big data is reported in [3]. A general distortion model for big video data was also developed to tackle a convex optimization problem according to the network transmission mechanisms.

Advances in cloud computing provide an elastic and cost-efficient exploration of voluminous data sets. However, there are many challenges. Costa et al. introduced a Network-as-a-service framework (NaaS) to integrate current cloud computing techniques with network infrastructure, and proposed an in-network aggregation method (Camdoop) for big data applications. Instead of increasing the network bandwidth, Camdoop was developed to decrease the traffic by pushing aggregation from the edge into the network.

Researchers considered Internet of Things and Cloud computing to address the Big Data issues and proposed a prototype model for providing sensing as a service on cloud. Others utilize the social structure and spectral characteristics of the network topology to enhance the dissemination of big data information and to reduce the processing overheads. A holistic architecture was presented to provide a set of abstractions for the different types of sensors and services, leveraging Big Data and cloud technologies as it caters for the data flow from sensors through to services. Vizzly, a middleware for interactive browsing of large sensor network data sets, is designed to provide map and line plot widget to visualize structured data from network sensors. Smart Traffic Cloud, an infrastructure for traffic applications, was proposed to enable traffic data acquisition, manage, analyze and present the results in a flexible, scalable and secure manner using a cloud platform. To analyze the big data gathered from social networks, the researcher explored personal ad hoc clouds comprising individuals in social networks. A 3G coverage analysis method is proposed to make use of the vast amount of network data and big data processing schemes to enhance network coverage.

B. Big data in Computer Science and Engineering

The AMPLAB in UC Berkeley was built to tackle big data [4]. They have developed an open source software stack, the Berkeley data analytics stack (BDAS), to use and understand big data by integrating software components. Goethe University Frankfurt also established a big data analytics research lab to unify the research activities in data analytics from the perspective of information system and computer science. Their approaches are based on the interdisciplinary binding between data management technologies and analytics. The ASTERIX project at UC

TABLE I: Big Data Management Systems

Company	System
IBM	Apache Hadoop, InfoSphere
Cloudera	CDH, Cloudera Standard, Cloudera Enterprise
Oracle	Oracle Big Data Appliance
Google	BigTable
Yahoo!	Sherpa
Amazon	SimpleDB
Microsoft	Dryad
Facebook	Apache Cassandra
Hypertable	HyperTable
ASF	Apache CouchDB

Irvin has developed a platform to answer the question of the right components and the right set of layers for taming the big data.

The MIT big data laboratory, CSAIL, was built to identify and develop technologies to solve the next generation data challenges [5]. They are developing platforms that are reusable, scalable and easy to deploy across multiple application domains for people to truly leverage big data. Their approaches are to: (1) collaborate closely with industries to provide real-world applications; and (2) view the big data problem as fundamentally multi-disciplinary. Their on-going big data projects involve using big data for a better life in the Trento smart city, execution Migration Machine (EM2) that interprets big data for Healthcare, Criminology, and Natural Language interface for Big data, etc..

Big Data needs efficient database management platforms. Researchers tackled questions such as what query classes can be considered tractable and how to facilitate query-response on big data. Ongoing research projects include the static and just-in-time compilation of analytics programs and database systems, and the automatic synthesis of out-of-core algorithms that efficiently exploit the memory hierarchy.

Enterprises have also been developing big data management systems, as shown in Table 1. Among them, the Apache Hadoop, composed of Hadoop Common, Distributed file system, YARN and MapReduce modules, is the most widely used. It is a framework supporting reliable, scalable, and distributed computing, specially designed to scale up from a single server to thousands of computers, offering local storage and computing, and allowing for distributed

processing of large data sets.

III. BIG DATA IN STATISTICS

In the big data era, a greater use of analytics is required to uncover hidden patterns and relationships among big data [6-7]. New tools are needed for big data exploration and visualization to support fast or even real-time decision making, and find information that were unable to be discovered in the past [8-9]. Big data need computational statisticians in collaboration with networking engineers and computer scientists. For this section, materials were obtained from publications including Journal of American Statistical Association (JASA), Journal of the Royal Statistical Society, Pattern Recognition, Biostatistics, Biometrics, Biometrika, Statistica Sinica, Bioinformatics, and Statistics in Medicine.

A. Statistical Analytics for Big Data in Academia

Big data were termed differently, most often as large-scale or high-dimensional data in statistics. A data structure, called X-tree, was proposed to store high dimensional data for various big data applications. Many emerging methods have been developed to analyze and understand big data. Here we focused on analytics in Genome, Web-based and Mobile Health, Behavioral and Administrative data studies.

The large scale data produced in the procedure of sequencing, mapping, and analyzing make genomics fall into the realm of big data. Hundreds of petabytes of data may be easily generated by sequencing multiple human genomes, and the analysis of the gene will further create more data. There are platforms for genomics analysis. For example, The ENCODE consortium is an encyclopedia of functional DNA elements to be used for scientific community. NextBio offers a platform which sits on top of existing systems to aggregate and analyze genomic data. Bina Technologies has built a system to enable users to access and analyze genomic sequence data, and its hybrid architecture keeps parts of data on the premises and parts in the cloud, which is used to speed up the sequencing time and facilitate the data transfer. The Portable Genomics uses a mobile visualization platform for genomics. The visualization concept has brought genomics to professionals and consumers to understand and use personalized and preventive medicine. The analytics in this area are mostly based on either Frequentist' or Bayesian approaches, although classical data mining techniques are being increasingly utilized. The joint work of computational

biostatisticians, geneticists, computer scientists and engineers are needed to advance this big genetic data area.

Web-based and Mobile Health intervention studies have the advantages of combining tailored approaches of face-to-face interventions with the scalability of public health interventions via the internet with lower cost [10]. It is a promising solution for healthcare due to its accessibility, time and cost savings. They have been developed for the following clinical areas: 1) chronic conditions, such as heart diseases, arthritis, and asthma; 2) health promotion, such as alcohol reduction, smoking cessation, diet and exercise; 3) mental health, such as anxiety and depression. For example, MAPIT, a web-based intervention system targeting substance abuse treatment in the criminal justice system, has been developed. It includes the extended parallel process model, motivational interviewing, and social cognitive theory. A web-based personally controlled health management system, Health.me, integrates an untethered personal health record with consumer care pathways and social forums to support healthcare. Caring Web is introduced to support for family caregivers of persons with stroke residing in home settings. An innovative web-based system is developed to allow patient-reported outcome measures to be easily administered. It also can be used for any medical interventions. The performance of a web-based intervention for mild to moderate depression, namely MoodGYM program, is evaluated. The usability of Tobacco Tractics, a website for reducing tobacco usage, is studied. However, the analytics for unstructured big data from these web-based and Mobile Health interventions are underdeveloped, because they are usually spatially- and temporally-varied data with missing values. Trajectory pattern recognition approaches are being initiated, developed and verified for such studies at University of Massachusetts Medical School, such as the DISC project funded by National Institute of Health (NIH) [11].

Big Data has the potential to improve behavioral medicine research and outcomes. Big Data to Knowledge (BD2K) is an initiative which aims to develop new approaches, standards, methods, tools, software, and competencies that will enhance the use of biomedical Big Data. National Library of Medicine is also promoting the use of common data elements to support sharing of Big Data. Integrative data analysis (the pooling of independent data sets that can be analyzed as one) was promoted as one of the techniques in behavioral medicine. At the 2014 Annual meeting of Society of Behavioral Medicine, pattern recognition for big data was emphasized. The aforementioned DISC project proposed the trajectory pattern recognition approach to the

behavioral interventions.

Linking big administrative data sets, such as the National Death Index, Medicare and Medicaid enrollment claims, and Social Security Administration Retirement and Disability data, to national health surveys [e.g. National Health Interview Survey (NHIS),National Health and Nutrition Examination Survey (NHANES), The Second Longitudinal Study of Aging (LSOA II), National Nursing Home Survey (NNHA)] is also challenging. The linkage methodologies are called on to examine issues such as health status, health conditions, health care, and health behaviors.

B. Statistical Analytics for Big Data in Industry

Big data includes unstructured data coming from sensors, devices, third parties, web applications, and social media, in real time and on a large scale. SAS provides four kinds of big data solutions: data management, high performance analytics, high performance data visualization, and flexible deployment options. A comprehensive data management approach is offered in these solutions to allow any amount of data to be managed, analyzed, and visualized effectively. EMC insists that vision, talent, and technology are required to make big data a success, providing solutions to big data management and analysis. GigaSpaces builds and deploys a large-scale real-time analytics system using big data technologies where customers are able to handle the scalability, performance, and database integration seamlessly by providing a simple event processing business logic.

IBM is using big data technologies to harness individual's data resources in healthcare. Their approaches include building sustainable healthcare systems; collaborating to improve care and outcomes; and increasing access to healthcare. Infosys Labs are building (1) Big Data Medical Engine in the Cloud (BDMEiC), a new Health Doctor, which uses the method of diagnosing, customizing and administering health care on real time using BDMEiC; (2) Big Data Powered Extreme Content Hub, which uses the approach of taming big content explosion and providing contextual and relevant information; and (3) Nature Inspired Visualization of Unstructured Big Data, which reconstructs self-organizing maps as spider graphs for better visual interpretation of large unstructured datasets. GNS is discovering healthcare methods through big data. Its analytics solutions are being applied across the healthcare industry from pharmaceutical and biotechnology companies to integrated delivery systems and Accountable Care Organizations (ACOs).

IV. CONCLUSION

Big data has the advantages of dramatic cost reductions, substantial improvement in the time to perform a computing task, and in offering new services. With the advance of big data, we could answer questions that are beyond research in the past, extract knowledge and insight from data, and can even improve the productivity of business and create substantial values for the world economy. However, it should be noted that the primary values of big data come not from its raw form, but from its processing and analysis. The sweeping changes in big data technologies and management will result in the multidisciplinary collaborations to support decision making and service innovation.

Two trends are making big data increasingly attractive: the ubiquity of mobile phones and advances in physical instrumentation [12]. While big data can yield extremely useful information and value, it presents different kinds of challenges. For example, we need to understand how much data to store, how much it costs, whether the data will be secured, and how long it must be maintained [9]. McKinsey Global Institute identified big data challenges in five domains: healthcare in the United States, the public sector in Europe, retail in the United States, manufacturing, and personal-location data globally [13].

Technical challenges for big data include strong real-time analysis requirements, new storage models, and parallel/distributed operators for data with new n-dimensional array-based data structures. These data need to be server- or cloud-managed, compared and visualized with joint efforts from hardware/software engineers, computer scientists and statisticians.

The privacy and publicity is an ethical challenge. Big data reflect a cultural, technological, and scholarly phenomenon [14]. For example, legislation has already been proposed to curb the collection and retention of data due to privacy concerns (e.g. the US Do Not Track Online Act of 2011).

Interpretation is critical and challenging for big data analyses. Misinterpretation always results from data limitation and bias regardless of the size of a data. Big data analyses are most effective when the complex methodological processes for that data are considered [14].

Big data visualization is another challenge but an urgent need. New database technologies, coupled with emerging web-based technologies, hold the key to lower the cost of visualization generation and allow it to become a more integral part of the scientific process. For example,

TASUKE, a web application is used to visualize large scale sequencing data generated by next generation sequencing technologies. In general, the graphic artists, communicators and visualization scientists should be brought into conversation with theorists and experimenters before all the data have been gathered.

Big data, while generating values, brings challenges to both research and business communities. These big data require novel and effective data networking, storage, management, access and analytics. It needs a collaborative vision and dialogs from various disciplines including engineers, computer scientists, statisticians, and social scientists [15].

V. ACKNOWLEDGEMENT

This research was supported by NIH/NIDA grant R01 DA033323-01A1 and the pilot project program award to Dr. Hua Fang from the National Center for Research Resources UL1RR031982.

REFERENCES

[1] D. Bollier and C. M. Firestone, *The promise and peril of big data*. Aspen Institute, Communications and Society Program, Washington, DC, USA, 2010.

[2] [online]. H. Fang, M. Franklin, U. O'Reilly, G.-C. Yuan and N. Chawla, http://www.amstat.org/meetings/jsm/2014/onlineprogram/ActivityDetails.cfm?SessionID=209882.

[3] M. Chen, S. Mao, Y. Liu, "Big Data: A Survey", *ACM/Springer Mobile Networks and Applications*, Vol. 19, No. 2, pp. 171-209, April 2014.

[4] Big Data [Online]. Available: https://amplab.cs.berkeley.edu/

[5] Big Data [Online]. Available: http://bigdata.csail.mit.edu/

[6] E. Birney, "The making of encode: Lessons for big–data projects," *Nature*, vol. 489, no. 7414, pp. 49–51, 2012.

[7] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The meaningful use of big data: four perspectives–four challenges," *ACM SIGMOD Record*, vol. 40, no. 4, pp. 56–60, 2012.

[8] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, "Mad skills: new analysis practices for big data," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1481–1492, 2009.

[9] K. Michael and K. W. Miller, "Big data: New opportunities and new challenges [guest editors' introduction]," *Computer*, vol. 46, no. 6, pp. 22–24, 2013.

[10] [Online]. Available: http://www.medicine20.com/2012/2/e3/

[11] [Online]. Available: http://projectreporter.nih.gov/project_info_description.cfm?icde=0&aid=8505922.

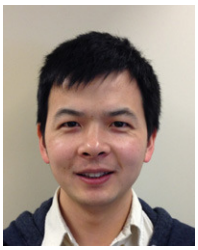
[12] E. Dumbill, "Making sense of big data," *Big Data*, vol. 1, no. 1, pp. 1-2, 2013.

[13] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," Technical report, McKinsey Global Institute, Tech. Rep., 2011.

[14] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," Information, Communication & Society, vol. 15, no. 5, pp. 662-679, 2012.

[15] [Online]. For other related references, please refer to: http://www.faculty.umassd.edu/honggang.wang/journal/bigdata/references.pdf

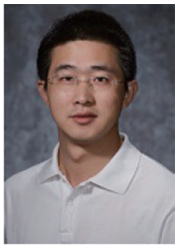
Hua Fang is an Associate Professor in the Department of Quantitative Health Sciences' Division of Biostatistics and Health Services Research at the University of Massachusetts Medical School. Dr. Fangs research interests include computational statistics, research design, statistical modeling and analyses in clinical and translational research. She is interested in developing novel methods and applying emerging robust techniques to enable and improve the studies that can have broad impact on the treatment or prevention of human disease.



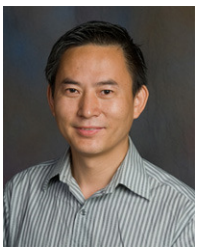
Zhaoyang Zhang received the B.S. degree in science and the M. S. degree in electrical engineering from Xidian University, Xian, China, in 2007 and 2010, respectively. He is currently pursuing his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Massachusetts, Dartmouth, MA, USA. His current research interests include wireless healthcare, wireless body area networks, and cyber-physical systems.

Chanpaul Jin Wang received the MS and PhD in communication engineering from the University of Electronic Science and Technology (UESTC), China, in 2006 and 2013, respectively. He is currently an assistant researcher of the University of Massachusetts Medical School& University of Massachusetts Dartmouth. His primary research interests are in big data and machine learning.

Mahmoud Daneshmand Mahmoud Daneshmand has a PhD and MA in Statistics from the University of California, Berkeley and MS and BS in Mathematics from the University of Tehran. He is a Distinguished Member of Technical Staff, AT&T Labs Research; Executive Director of University Collaborations Program and Assistant Chief Scientist of AT&T Labs; Affiliate Professor of the School of Technology Management and Computer Science, Stevens Institute of Technology. He has more than 30 years of teaching, research, and management experience in academia and Industry. His current areas of research are Sensor Networks and RFID Systems including reliability, performance and data mining of sensor and RFID data.



Chonggang Wang received the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China in 2002. He is currently a member of technical staff at InterDigital Communications, focusing on Internet of Things (IoT) R&D activities including technology development and standardization. His current research interests include IoT, mobile communication and computing, and big data management and analytics. He is the founding Editor-in-Chief of IEEE Internet of Things Journal and an IEEE ComSoc Distinguished Lecturer (2015-2016).



Honggang Wang received the Ph.D. degree in Computer Engineering at the University of Nebraska-Lincoln in 2009. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Massachusetts Dartmouth, USA.

His research interests include wireless communication and networking, sensor networks, multimedia communications, social networks and wireless healthcare. He has published more than 90 papers in his research areas. He serves as a chair/co-chair for several international conferences and on the editorial board for several journals. He is a Lead Guest Editor of IEEE Journal of Biomedical and Health Informatics (J-BHI) (previous IEEE Transactions on Information Technology in Biomedicine (TITB)) special issue on "Emerging Wireless Body Area Networks (WBANs) for Ubiquitous Healthcare" in 2013.