Big Data: Practical Applications

By SIMON HAYKIN, *Fellow IEEE*
Guest Editor

VOLKER TRESP
Guest Editor

JON ATLI BENEDIKTSSON, *Fellow IEEE*
Guest Editor

The papers in the first part of the Special Issue on "Big Data" were devoted to theoretical aspects of big data and were published in January 2016. This second Special Issue on Big Data is devoted to practical applications; in a way, these two parts of the Special Issue complement each other nicely. It is of interest to note that big data and the Internet of Things (IoT), representing the two sides of the same coin, reinforce each other, all the more recognizing that we are well and truly in the fourth industrial revolution. Thus, publication of the two parts of the special Issue on Big Data is highly timely.

This second part of the Special Issue on Big Data embodies seven papers.

1) "Big Data Analysis for Media Production"
by Blat et al.

A typical high-end film production generates several terabytes of data per day, either as footage from multiple cameras or as background information regarding the set: laser scans, spherical captures, etc. This first paper presents solutions to improve the integration of multiple data sources and understand their quality and content, which are useful both to support creative decisions on set (or near it) and enhance the postproduction process. The main cinema-specific contributions, tested on a multisource production data set made publicly available for research purposes, are the monitoring and quality assurance of multicamera setups, multisource registration and acceleration of 3-D reconstruction, anthropocentric visual analysis techniques for semantic content annotation, and integrated 2-D/3-D web visualization tools. The discussion includes improvements carried out in basic techniques for acceleration, clustering, and visualization necessary to deal with the very large multisource data, and the application of other big data problems in diverse fields.

The papers in this special issue focus on the practical applications of big data.

2) "Parallel Processing Systems for Big Data: A Survey"
by Zhang et al.

The volume, variety, and velocity properties of big data with valuable information contents have motivated the investigation of many new parallel data processing systems in addition to the approaches using traditional database management systems (DBMS). MapReduce pioneered this paradigm change, rapidly becoming the primary big data processing system for simplicity, scalability, and fine-grain fault tolerance. However, compared to DBMSs, MapReduce also arouses controversy in processing efficiency, low-level abstraction, and rigid dataflow. Inspired by MapReduce, novel big data systems are blooming. Some of them follow MapReduce's idea, but with more flexible models for general purpose usage; some of them absorb the advantages of DBMSs with higher levels of abstraction. There are also specific systems for special applications, such as machine learning and stream data processing. To explore new research opportunities and assist users in selecting suitable processing systems for specific applications, this survey paper provides a high-level overview of the existing parallel data processing systems categorized by data input as batch processing, stream

Digital Object Identifier: 10.1109/JPROC.2016.2614849

processing, graph processing, and machine learning, and introduces representative projects in each of the categories mentioned. System benchmarks and open issues for big data processing are also studied in the survey.

3) "Classification of Big Data With Application to Imaging Genetics"
by Ulfarsson et al.

Typically, big data applications, such as medical imaging and genetics, generate data sets that consist of few observations n on many more variables p, which provide a scenario denoted as $p >> n$. Traditional data processing methods are often insufficient for extracting information out of big data. There is, therefore, the need for developing new algorithms that can deal with the size, complexity, and the special structure of such data sets. In this paper, the problem of classifying $p >> n$ data is presented, as well as that of proposing a classification procedure based on linear discriminant analysis (LDA). Traditional LDA depends on covariance estimate of the data, but with $p >> n$, sample covariance estimate is singular. The proposed method presented herein estimates the covariance by using a sparse version of noisy principal component analysis (nPCA). The use of sparsity in this setting aims at automatically selecting variables that are relevant for classification. In performing experiments, the new method is compared to state-of-the art methods for big data problems, using both simulated data sets and imaging genetics data sets.

4) "'Big Data Assimilation' Toward Post-Petascale Severe Weather Prediction: An Overview and Progress"
by Miyoshi et al.

Following the inventions of telegraphy, electronic computer, and remote sensing, big data is the source of yet another revolution to weather prediction: as sensor and computer technologies advance, orders of magnitude bigger data are produced by new sensors and high-precision computer simulation or "big simulations." Data assimilation (DA) is a key to numerical weather prediction (NWP) by integrating the real-world sensor data into simulation. However, the current DA and NWP systems are not designed to handle the big data from next-generation sensors and big simulations. Accordingly, it has been proposed that the innovative "big data assimilation" (BDA) be used to fully utilize big data. Since October 2013, Japan's BDA project has been exploring the revolutionary NWP at 100-m mesh refreshed every 30 s, orders of magnitude finer and faster than the current typical NWP systems; this has been achieved by taking advantage of the combination of next-generation technologies: the 10-petaflops K computer, phased-array weather radar, and geostationary satellite Himawari-8. Thus far, a BDA prototype system has been tested with real-world retrospective local rainstorm scenarios. This paper summarizes activities and progress of the BDA project, and concludes with perspectives toward the post-petascale supercomputing era.

5) "Going Digital: A Survey on Digitalization and Large-Scale Data Analytics in Healthcare"
by Tresp et al.

This paper provides an overview of the recent trends toward digitalization and large-scale data analytics in healthcare, both of which are instrumental in the dramatic changes in the way healthcare will be organized in the future. Recent political initiatives have been designed to shift the care delivery processes from paper to electronics, with more effective treatments and better outcomes as major goals. Under newly developed networks of healthcare providers, research organizations and commercial vendors have stated to work jointly to analyze data for the development of decision-support systems. This paper addresses the trend toward continuous healthcare, where health is being monitored by wearable and stationary devices. In a related development, patients increasingly take responsibility for their own health-related data. Finally, recent initiatives toward personalized medicine are based on advances in molecular medicine, data management, and data analytics.

6) "Big Data for Remote Sensing: Challenges and Opportunities"
by Chi et al.

Every day a large number of Earth observation (EO) spaceborne and airborne sensors, from many different countries, provide massive amounts of remotely sensed data. These data are being used for different applications, such as natural hazard monitoring, global climate change, urban planning, and so on. The applications are data driven and mostly interdisciplinary. Based on this scenario, it can be stated that we are now living in big data in a remote sensing environment. Furthermore, the data are becoming an economic asset and a new important resource in many applications. This paper discusses the challenges and opportunities that big data bring along the essence of remote sensing applications. The focus herein is to analyze what big data exactly means in remote sensing applications and how it can provide added value. Furthermore, this paper describes the most challenging issues in managing, processing, and efficient exploitation of big data for remote sensing problems. In particular, two case studies are discussed: in the first case, big data is used to automatically detect marine-oil spills using a large archive of remote sensing data; in the second test case, content-based information retrieval is performed using high-performance computing to extract information from a large database of remote sensing images.

7) "Some Comments on the Analysis of 'Big' Scientific Time Series"
by Thomson and Vernon.

Experience, gained from long-time series from space, climate, seismology, and engineering, has demonstrated

the need for even longer data series with better precision, timing, and larger instrument arrays. For example, almost all the data that have been examined in the paper, i.e., atmospheric, seismic data, and dropped calls in cellular phone networks, contain evidence for solar mode oscillations that couple into Earth systems through magnetic fields. The two examples examined suggest that robustness assumptions are often not justified and that many of the extremes in geomagnetic and space physics data may be the result of superposition of numerous modes. Evidence is also presented to show that the evolution of turbulence in interplanetary space may be controlled by similar modes. Returning to the theme of "big data," experience has shown that theoretical predictions, i.e., that spectra would be asymptotically unbiased, have turned out to be largely irrelevant with the very long time series examined. Data that were considered to have excessively variable spectra appear to evolve into processes with dense sets of modes. ■

ABOUT THE GUEST EDITORS

Simon Haykin (Fellow, IEEE) received the B.Sc. (first class honors), Ph.D., and D.Sc. degrees, all in electrical engineering, from the University of Birmingham, Birmingham, U.K.

He is a Distinguished University Professor in the Faculty of Engineering, McMaster University, Hamilton, ON, Canada. For much of the past 15 years, he has focused his entire research program on learning from the human brain and applying it to a new generation of cognitive dynamic systems, exemplified by cognitive radar, cognitive control, and cognitive radio. He is the author/coauthor of over 50 books on communication systems (analog and digital), adaptive filter theory, neural networks and learning machines, and cognitive dynamic systems.

Prof. Haykin is a Fellow of the Royal Society of Canada; the recipient of the Honorary Doctor of Technical Sciences, ETH, Zurich, Switzerland; the recipient of the Booker Gold Medal from URSI for his outstanding contributions to radar and wireless communications; as well as many other medals and awards.

Volker Tresp received a Diploma degree from the University of Goettingen, Goettingen, Germany, in 1984 and the M.Sc. and Ph.D. degrees from Yale University, New Haven, CT, USA, in 1986 and 1989, respectively.

Since 1989, he has been the head of various research teams in machine learning at Siemens, Research and Technology. He filed more than 70 patent applications and was the Inventor of the Year of Siemens in 1996. He has published more than 150 scientific articles and administered over 20 Ph.D. dissertations. The company Panoratio is a spinoff out of his team. His research focus in recent years has been on machine learning in information networks for modeling knowledge graphs, medical decision processes, and sensor networks. He is the coordinator of one of the first nationally funded big data projects for the realization of precision medicine. In 2011, he became an Honorary Professor at the Ludwig Maximilian University of Munich, Munich, Germany, where he teaches an annual course on machine learning.

Jon Atli Benediktsson (Fellow, IEEE) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He joined the Department of Electrical Engineering, University of Iceland, in 1991 as an Assistant Professor and was promoted to Professor in 1996. On July 1, 2015, he became the President/Rector of the University of Iceland. From 2009 to 2015, he was the Pro Rector of Science and Academic Affairs at the University of Iceland. He is a cofounder of the biomedical startup company Oxymap (www.oxymap.com). His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields.

Prof. Benediktsson was the 2011-2012 President of the IEEE Geoscience and Remote Sensing Society. He was the Editor-in-Chief of the IEEE Transactions on Geoscience and Remote Sensing (TGRS) from 2003 to 2008 and has served on numerous editorial boards. He is a Fellow of the International Society for Optics and Photonics (SPIE) and has received many awards for his research and professional activities.