Introduction to the Issue on Signal Processing for Big Data

WITH the Internet, social media, wireless mobile devices, and pervasive sensors continuously collecting massive amounts of data, we undoubtedly live in an era of "data deluge." Learning from such huge volumes of data however, promises ground-breaking advances in science and engineering along with consequent improvements in quality of life. Indeed, mining information from big data could limit the spread of epidemics and diseases, identify trends in financial and e-markets, unveil topologies and dynamics of emergent social-computational systems, accelerate brain imaging, neuroscience and systems biology models, and also protect critical infrastructure including the power grid and the Internet's backbone network.

While Big Data can be definitely perceived as a big blessing, big challenges also arise with large-scale datasets. The sheer volume of data makes it often impossible to run analytics using a central processor and storage, and distributed processing with parallelized multi-processors is preferred while the data itself is stored in the cloud. As many sources continuously generate data in real time, analytics must often be performed "on-the-fly" and without an opportunity to revisit past entries. Due to their disparate origins, the resultant datasets are often incomplete and include a sizable portion of missing entries. In addition, massive datasets are noisy, prone to outliers, and vulnerable to cyber-attacks. These effects are amplified if the acquisition and transportation cost per datum is driven to a minimum. Overall, Big Data present challenges in which resources such as time, space, computing power, and energy, are intertwined in complex ways with data resources.

Given these challenges, ample signal processing opportunities arise. This special issue of JSTSP contains novel modeling approaches, algorithmic advances along with their performance analysis, as well as representative applications of Big Data analytics to address practical challenges, while revealing fundamental limits and insights on the analytical trade-offs involved.

The contribution by Sorber, Van Barel and De Lathauwer offers a unifying framework for *modeling and inference* from big data under the umbrella of what is termed structured data fusion (SDF)—a notion originating from the integration of multiple data sources, and generalizing tensor decompositions. SDF describes big datasets as a superposition of possibly coupled factor models with low-dimensional structure that emerges in various SP applications. Implementation of SDF is facilitated by Tensorlab, which is a software package developed to reveal the underlying data structure. Also on a modeling framework for big data, the paper by Braun, Pokutta, and Xie, tames the growth of data sizes with an information-theoretic approach to adaptive compressive sensing. The outcome is an efficient greedy algorithm that features approximate compressive recovery guarantees with near-optimal sample complexity for sparse, low-rank, and mixture of Gaussian signal models. Addressing key issues of big data analytics, the contribution by Bruer, Tropp, Cevher, and Becker, delineates rigorous tradeoffs among computational time, sample complexity and accuracy of estimators derived from convex criteria. It further unveils performance-complexity "sweet spots" by tuning the degree of smoothing in pertinent optimization tasks.

Further on *big data tools*, the two-part paper by Ravishankar, Wen, and Bresler introduces schemes for online learning of sparsifying transforms, which are tailored for signal denoising and compressed sensing. The first part demonstrates that these schemes can scale to extremely large data sets, while the second part establishes that they exhibit faster convergence than conventional approaches based on batch learning. Along related lines of large data streams, the contribution by Hall and Willett deals with optimization schemes implementing online convex programs for learning tasks. The resultant algorithms are computationally efficient, can adapt to dynamic environments, and their performance is assessed using associated regret bounds.

On the *algorithmic side*, the paper by Wang, Dong, Zhang, and Tan, shows how diverse observations of susceptible-infectious graph models can be leveraged for online identification of a single or multiple rumor sources in large-scale social networks. As increased diversity enhances detection probability, their approach can be also useful for network forensics to mitigate epidemic spreading of, e.g., fraudulent email spams. Challenges emerging with big data clustering applications is the subject of the paper by Traganitis, Slavakis, and Giannakis, where either the dimensionality and/or the sheer volume of the datasets are prohibitively large for conventional approaches. Inspired by random sampling and consensus tools advocated in computer vision for robust regression, a host of randomized algorithms are developed for clustering settings that scale much more favorably than the state-of-the-art random projection based clustering alternatives. By leveraging the attributes of sparsity and low rank that are often present in big datasets, Patel, Nguyen, and Vidal put forth computationally efficient algorithms for joint dimensionality reduction and clustering of data living in a union of subspaces. Online learning and prediction of big spatiotemporal traffic data is the subject of the paper by Xu, Deng, Demiryurek, Shahabi, and van der Schaar. In addition to adaptive partitioning of the traffic context space, the ability to cope with delayed data and impute missing data, short- and long-term bounds are provided to assess prediction performance.

Digital Object Identifier 10.1109/JSTSP.2015.2418393

Applications related to recommender systems require real-time learning from a variety of sources. Context—the information available to the decision-maker or learner—is typically very high dimensional. But the relevant dimensions are usually small, time-varying, and dependent upon the desired action. Based on contextual multi-armed bandit tools Tekin and van der Schaar develop an approach to learning the relevant dimension for a given action, and further establish a regret bound that only depends upon the maximum number of relevant dimensions across all actions. The resultant algorithm is tested on datasets arising from network security and online news article recommendations. To cope with missing entries in recommender systems, the contribution by Wang, Wang, and Gu, deals with graph-based signals over large-scale networks of (e.g., sensor) nodes. It offers provably convergent algorithms and associated performance bounds for distributed least-squares reconstruction of bandlimited time-varying signals from a limited number of samples collected by a subset of selected nodes.

Context is also the main attribute leveraged in the paper by Krishnan and Baron for efficient data compression. With streaming big data, parallel per-block processing is well motivated for fast compression, but processing each block independently is agnostic to the context correlation, which sacrifices compression quality for a given throughput. To reduce the compression loss due to parallel processing, this paper develops a work-efficient algorithm that coordinates compression across multiple blocks through a two-pass minimum description length algorithm. Such a context-cognizant parallel compression approach attains an improved compression-throughput tradeoff at low computational complexity.

Finally, the paper by Gonzalez-Dominguez, Eustis, Lopez-Moreno, Senior, Beaufays, and Moreno, puts forth a multi-language speech recognition platform that can cope with arbitrary combinations of spoken languages. It has a scalable deep neutral network architecture that can handle mixtures of 34 languages, and is used to serve 200 million real time translations per day. The use of efficient and scalable algorithms allows it to be trained on large data sets leading to state-of-the-art word error rate performance.

As a closing note, we would like to express the appreciation of the Guest Editorial team to the Editorial Board and the Staff supporting IEEE JSTSP for encouraging, reviewing, welcoming, and facilitating the processing of this special issue. And of course, this issue would have not been possible without the high-quality feedback received from the conscientious reviewers whom we wish to thank for their volunteer effort and timely responses.

GEORGIOS B. GIANNAKIS, *Lead Guest Editor*
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455 USA

RAPHAEL CENDRILLON, *Guest Editor*
Google
Mountain View, CA 94043 USA

VOLKAN CEVHER, *Guest Editor*
École Polytechnique Fédérale de Lausanne
CH-1015 Lausanne, Switzerland

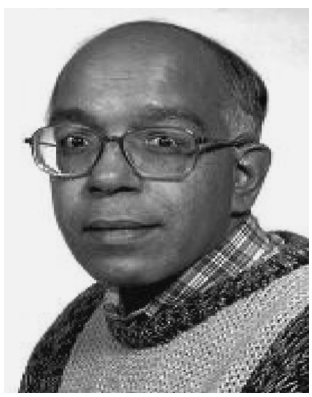
ANANTHRAM SWAMI, *Guest Editor*
Army Research Laboratory
Adelphi, MD 20783 USA

ZHI TIAN, *Guest Editor*
Electrical and Computer Engineering Department
George Mason University
Fairfax, VA 22030 USA

Georgios B. Giannakis (F'97) received his Ph.D. degree from the University of Southern California in 1986. Since 1999, he has been with the University of Minnesota, where he holds the ADC chair in wireless telecommunications in the Department of Electrical and Computer Engineering and serves as director of the Digital Technology Center. His interests are in the areas of communications, networking, and statistical signal processing—subjects on which he has published more than 375 journal and 635 conference papers, 21 book chapters, two edited books, and two research monographs (h-index 112). His current research focuses on learning from Big Data, wireless cognitive radios, and network science with applications to social, brain, and power networks with renewables. He is the (co) inventor of 23 patents and the (co)recipient of eight best paper awards from the IEEE Communications and Signal Processing Societies. He is a Fellow of the IEEE and EURASIP and has also received technical achievement awards from the IEEE Signal Processing Society and EURASIP, including the IEEE Fourier Technical Field Award (2015).

Raphael Cendrillon received his BSEE (Highest Honors) from the U. of Queensland, Australia, in 1999, and the Ph.D. in EE (Summa Cum Laude) at the Katholieke Universiteit Leuven, Belgium, in 2004. He has held senior research positions in leading companies including ASSIA Inc. and at Huawei Technologies, and was a Visiting Researcher at Princeton and at Stanford. He is currently a Software Engineer at Google. He has been awarded 35 patents and has published over 30 papers in internationally recognized journals and conferences. His research work has over 1,300 citations and his algorithms are now part of the ITU and ANSI standards on broadband networks. Dr. Cendrillon was awarded the University of Queensland Early Career Researcher Grant in 2005, the UniQuest Trailblazer Prize for Commercialization in 2005, the K.U. Leuven Bursary for Advanced Foreign Scholars in 2004, the Alcatel Scientific Prize in 2004, and IEEE Grants in 2003, 2004 and 2005.

Volkan Cevher received the B.S. (valedictorian) degree in electrical engineering in 1999 from Bilkent University in Ankara, Turkey, and he received the Ph.D. degree in electrical and computer engineering in 2005 from the Georgia Institute of Technology in Atlanta. He held research scientist positions at the University of Maryland, College Park from 2006 to 2007 and at Rice University in Houston, Texas, from 2008 to 2009. Currently, he is an Assistant Professor at the Swiss Federal Institute of Technology Lausanne with a complimentary appointment at the Electrical and Computer Engineering Department at Rice University. His research interests include signal processing, optimization, machine learning, and information theory. He received a Best Paper Award at SPARS in 2009 and an ERC StG in 2011.



Ananthram Swami (S'79–M'79–SM'96–F'08) is with the U.S. Army Research Laboratory (ARL) as the Army's ST (Senior Research Scientist) for Network Science. He is an ARL Fellow and Fellow of the IEEE. He has held positions with Unocal Corporation, the University of Southern California (USC), CS-3 and Malgudi Systems. He was a Statistical Consultant to the California Lottery, developed a MATLAB-based toolbox for non-Gaussian signal processing, and has held visiting faculty positions at INP, Toulouse. He received the B.Tech. degree from IIT-Bombay, the M.S. degree from Rice University, and the Ph.D. degree from the University of Southern California (USC), all in electrical engineering. His research interests are in the broad area of network science: the study of interactions and co-evolution, prediction and control of inter-dependent networks, with applications in composite tactical networks.

Zhi (Gerry) Tian (M'98–SM'06–F'13) is a Professor in the Electrical and Computer Engineering Department of George Mason University, Fairfax, VA, as of January 2015. Prior to that, she was on the faculty of Michigan Technological University from 2000 to 2014. She served as a Program Director at the US National Science Foundation from 2012 to 2014. Her research interests lie in statistical signal processing and data analytics, wireless communications and wireless sensor networks. She is an IEEE Fellow. She served as Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2002–2008) and IEEE TRANSACTIONS ON SIGNAL PROCESSING (2006–2009). She is a Distinguished Lecturer of the IEEE Vehicular Technology Society (2013–2015) and the IEEE Communications Society (2015–2016).