Georgios B. Giannakis, Francis Bach,
Raphael Cendrillon, Michael Mahoney,
and Jennifer Neville

Signal Processing for Big Data

The information explosion propelled by the advent of online social media, the Internet, and global-scale communications has rendered learning from data increasingly important. At any given time around the globe, large volumes of data are generated by today's ubiquitous communication, imaging, and mobile devices such as cell phones, surveillance cameras, medical and e-commerce platforms, as well as social networking sites. While many find this intrusive and raise legitimately "Big Brother" concerns, there is no denying that tremendous economic growth and improvement in quality of life hinge upon harnessing the potential benefits of analyzing massive data.

The term *big data* was coined to describe this information deluge, and signal processing (SP) tools and applications are clearly well seasoned to play a major role in this data science endeavor. Quoting a recent article published in *The Economist,* "The effect (of Big Data) is being felt everywhere, from business to science, and from government to the arts" [1]. Mining information from unprecedented volumes of data promises to prevent or limit the spread of epidemics and diseases, identifying trends in financial markets, learning the dynamics of emergent social-computational systems, and also protect critical infrastructure including the smart grid and the Internet's backbone network. But great promises come with formidable research challenges; as Google's chief economist explains in the same article, "Data are widely available, what is scarce is the ability to extract wisdom from them." While significant progress has been made in the last decade toward achieving the ultimate goal of "making sense of it all," the consensus is that we are still not quite there.

In this context, this special issue (SI) of *IEEE Signal Processing Magazine* (*SPM*) aims to 1) delineate the theoretical and algorithmic underpinnings along with the relevance of SP tools to the emerging field of big data and 2) introduce readers to the challenges and opportunities for SP research on (massive-scale) data analytics. The latter entails an extended and continuously refined technological wish list, which is envisioned to encompass high-dimensional, decentralized, parallel, online, and robust statistical SP, as well as large, distributed, fault-tolerant, and intelligent systems engineering. The goal of this SI is to selectively sample a diverse gamut of big data challenges and opportunities through surveys of methodological advances, as well as more focused- and application-oriented contributions chosen on the basis of timeliness, importance, and relevance to SP.

The interest in big data-related research from the SP community is evident from the increasing number of papers submitted on this topic to SP-oriented publications, workshops, and conferences. In terms of funding programs, the importance of big data research is also apparent. The White House Office of Science and Technology Policy in concert with several federal departments and agencies announced the Big Data Research and Development Initiative in 2012 [2]. The launch included generous funding in new commitments through the National Science Foundation, National Institutes of Health, Defense Advanced Research Projects Agency, and U.S. Department of Defense (DoD) at large, U.S. Department of Energy (DoE), and the U.S. Geological Survey. The DoD is placing a "Big Bet on Big Data," with two dozen open solicitations. Likewise, the European Union Commission shows increasing interest in big data analytics, e.g., under the Seventh Framework Programme for Research. All these provide ample testament that the theme of this SI is timely, and we hope that it offers something from which the SP readership will benefit.

Our opening article by Slavakis, Giannakis, and Mateos begins with a fairly rich family of models capturing a wide range of SP-relevant data analytic tasks. These include principal component analysis, nonnegative matrix factorization, dictionary learning, compressive sampling, and subspace clustering. Building on these models, the article further offers scalable inference and optimization algorithms for decentralized and online learning problems, while revealing fundamental insights into the various analytic and implementation tradeoffs involved. Generalizations of these encompassing models to timely data-sketching and tensor- and kernel-based learning tasks are also provided. The contribution finally demonstrates how the presented framework applies to several big data tasks, such as network visualization, decentralized and dynamic estimation, prediction, and imputation of network link load traffic, as well as imputation in tensor-based magnetic resonance imaging.

The second article, by Cevher, Becker, and Schmidt, places particular emphasis on recent advances in convex optimization algorithms tailored for big data, having as ultimate goal to markedly reduce the computational, storage, and communication bottlenecks. The valuable overview of this emerging field comprises contemporary approximation techniques such as first-order methods and randomization for

Digital Object Identifier 10.1109/MSP.2014.2330054
Date of publication: 19 August 2014

scalability, as well as parallel and distributed schemes that play an increasingly instrumental role in large-scale computation. The new big data algorithms outlined are based on surprisingly simple principles and attain impressive accelerations even on classical optimization tasks.

As the size of data grows, so does the chance to involve outlying observations. This in turn motivates the need for outlier-resilient learning algorithms scaling to large-scale application settings. In this context, the article by Tajer, Veeravalli, and Poor deals with robust, sequential detection schemes for big data. Outlying sequence detection is particularly important in health, the Internet, energy, telecommunications, and related large-scale problems. The article demonstrates how outlying sequence detection algorithms can be analyzed by viewing them as strategies for hypothesis testing with different outlying recovery objectives. Using this approach allows the effectiveness of outlying sequence detection strategies to be evaluated in the big data regime.

The acquisition modality, information processing, and inference from observations often dictates the need to deal with tensors—often big arrays of data collected in (hyper)cubes, thus generalizing the notion of data matrices. The growth of big data platforms makes it possible to solve large-scale tensor problems, which are encountered in various applications ranging from multiantenna communication transceivers to speech and audio, as well as machine learning from Internet data, to name a few. Sidiropoulos, Papalexakis, and Faloutsos introduce, in their article, interesting identifiability results and a parallel decomposition approach for tensors having low rank. This allows the resultant algorithms to scale nicely to sizes growing inversely proportional to the tensor rank.

High-order tensors and their decompositions are abundantly present in domains such as statistical SP (e.g., high-order moments and sensor arrays), scientific computing (e.g., discretized multivariate functions), and quantum information theory (e.g., for quantum many-body states). Representing the full tensor quickly becomes impractical for modern practical problems as the tensor's order increases. The article by Vervliet, Debals, Sorber, and De Lathauwer focuses on compact multilinear models that enable computational manipulation and estimation of such models from incomplete information.

After overviewing pertinent models and algorithms, two case studies are presented in multidimensional harmonic retrieval and material science to illustrate the potential of these approaches. In addition to matrices and tensors, big data emerge often from large-scale networks and generally graphs that are abundant in SP-relevant applications. The article by Sandryhaila and Moura highlights recent work on developing a paradigm for the analysis of graph-based data based on the so-called discrete signal processing on graphs (DSPG) approach—an effort to extend classical SP notions and techniques to data indexed by general graphs. The motivation should be clear: large data sets that are naturally modeled as graphs are generated and analyzed in a wide range of applications, and extracting valuable information from these data requires innovative approaches. Not surprisingly, some DSPG methods result from a straightforward mapping of time series to spectral graphs, which allows for drawing parallels from the former to the latter in notions as classical as filtering, spectral analysis, and transform theory. Interestingly, this is just the tip of the iceberg, since there are many subtle and fundamental issues that arise in DSPG, as articulated in this article. The discrete Fourier transform (DFT) is one of SP's "workhorses," and its popular implementation relies on the celebrated fast Fourier transform (FFT). The article by Gilbert, Indyk, Iwen, and Schmidt describes recent developments in an alternative, so-called sparse Fourier transform (SFT) implementation, which offers promises in certain large-scale data tasks involving sparse signals. The SFT can compute a compressed Fourier transform using only a subset of the input data in time, considerably shorter than the original data set size. SFT can thus be faster than the FFT when it is hard in large-scale applications to acquire enough data to run the FFT, and/or it is desirable to run DFT in time sublinear in the input size—a welcome attribute in medical imaging, when it is important to reduce the time that the patient spends in the magnetic resonance imaging machine. In addition to an overview of SFT, the article outlines the basic techniques and tradeoffs involved, as well as the connections between the SFT and related methods such as streaming algorithms and compressive sampling.

Given the deluge we experience from video, audio, medical imagery, spectroscopic, geophysical, and seismic data, the models and SP-related tools exposed in this SI promise a significant impact on many traditional but also in various emerging large-scale applications. One such innovative application wraps up this SI and deals with collaborative bike sensing for automatic geographic enrichment. Verstockt, Slavkovikj, De Potter, and Van de Walle put forth in their article a system for automatic annotation of geographical data from cyclists' smartphones. The article describes the effectiveness of this system with large-scale data sets in real-world conditions.

In closing, we would like to express our appreciation to the Editorial Board and staff of *IEEE SPM* (especially SI Area Editor Fulvio Gini) for encouraging, reviewing, welcoming, and facilitating the processing of this SI. 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 efforts and timely responses.

REFERENCES

[1] K. Cukier. (2010, Feb. 25). "Data, data everywhere," *The Economist*. [Online]. Available: http://www.economist.com/node/15557443

[2] Office of Science and Technology Policy, "Big data research and development initiative," Executive Office of the President, Mar. 29, 2012. [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

[SP]