Big Data: Theoretical Aspects

By SIMON HAYKIN, *Fellow, IEEE*
Guest Editor

STEPHEN WRIGHT
Guest Editor

YOSHUA BENGIO
Guest Editor

I. THEORETICAL ASPECTS OF BIG DATA

Big data has burst into public awareness over the past few years as people have become more and more aware of the massive amount of data being produced by social and scientific activities, and its potential utilization for good or harm. On the research front, big data has spurred new activity across a range of fields, including statistics, machine learning, and computer systems. Many areas have been profoundly altered by the big data revolution, including wireless communications, speech processing, social networking, online commerce, medical informatics, and finance. In these areas, and in many others, analysis of the data yields valuable information that deepens understanding, improves decision making, and enhances performance of predictive models.

This special issue highlights a number of algorithmic approaches that are fundamental to data analysis, both in formulating and solving problems. These methods form part of the core of the field, a set of tools that can be applied to many specific application areas.

This special issue highlights a number of algorithmic approaches that are fundamental to data analysis, both in formulating and solving problems.

The issue consists of nine papers covering a variety of topics in formulation and algorithms. We summarize each of them briefly.

"**A Review of Relational Machine Learning for Knowledge Graphs**" by Nickel *et al.* Relational machine learning studies methods for statistical analysis of relational or graph-structured data. This paper reviews how such statistical models can be trained on large knowledge graphs, and then used to predict new facts, such as prediction of new edges in the graph. Two fundamentally different kinds of statistical relational models are addressed. The first is based on latent feature models (for example, tensor factorization and multiway neural networks), while the second is based on mining observable patterns in the graph. The paper shows how to combine the latent and observable models to improve modeling power while decreasing computational cost. It is shown how such statistical models of graphs can be combined with text-based information extraction methods for automatically constructing knowledge graphs from the Web. Google's Knowledge Vault project provides an example of such combinations.

"**Learning to Hash for Indexing Big Data—A Survey**" by Wang *et al.* The explosive growth in big data has created a great deal of demand for efficient indexing and searching procedures. In many critical applications, including large-scale search and pattern matching, finding the nearest neighbors to a query is a difficult proposition. The paper looks to the approximate nearest neighbor (ANN) search based on hashing techniques, which has become popular due to its promising efficiency and accuracy. The paper describes new approaches that incorporate data-driven learning methods in the development of advanced hash functions. Such learning-to-hash methods exploit information such as data distributions or class labels when optimizing hash codes or functions. Most importantly, the hash

Digital Object Identifier: 10.1109/JPROC.2015.2507658

codes so learned preserve the proximity of neighboring data in the original feature spaces to the hash-code spaces. The paper provides a systematic understanding of insights and the pros and cons of emerging techniques. It also provides a comprehensive survey of the learning-to-hash framework and representative techniques of various types that include unsupervised, semisupervised, and supervised procedures.

"**Implementing Randomized Matrix Algorithms in Parallel and Distributed Environments**" by Yang *et al.* Distributed systems built on top of clusters of commodity hardware provide inexpensive and reliable storage and scalable processing of massive data. Because inexpensive storage is so readily available, it is common to store as much data as possible (not just currently relevant data), in the hope that its value can be extracted at a later time. Exabytes of data are being created on a daily basis. Extraction of value from such data requires scalable implementations of advanced analytical algorithms, including statistical regression methods, linear algebra, and optimization algorithms. Traditionally, such methods are designed to minimize floating-point operations, which is the dominant cost of in-memory computation on a single machine. Parallel and distributed environments give rise to more complex measures of performance that take into account load balancing and communication, including disk and network I/O. These factors greatly increase the complexity of algorithm design and challenge traditional ways of thinking about the design of parallel and distributed algorithms. This paper reviews recent work on developing and implementing randomized matrix algorithms in large-scale parallel and distributed environments, with a focus on the theory and practical implementation of random projection and random sampling algorithms for very large and very overdetermined l_1- and l_2-regression problems. Theoretical results demonstrate that in near-input-sparsity time and with only a few passes through the data, strong relative-error approximate solutions can be obtained with high probability. Under various tradeoffs, empirical results demonstrate that l_1- and l_2-regression problems can be solved to low, medium, or high precision in existing distributed systems on up to terabyte-sized data.

"**Foundational Principles for Large-Scale Inference: Illustrations Through Correlation Mining**" by Hero and Rajaratnam. This paper presents a correlation-mining framework for large-scale inference. In such applications as genomics, connectomics, and eco-informatics, the data set is often variable rich but sample starved—the number of acquired statistical replicates is far fewer than the number of observed variables. This paper develops a unified statistical framework that explicitly quantifies the sample complexity of various inferential tasks. Sampling regimes can be divided into three categories: 1) the classical asymptotic regime, where the variable dimension is fixed and the sample size goes to infinity; 2) the mixed asymptotic regime, where both variable dimension and sample size go to infinity at comparable rates; and 3) the purely high-dimensional asymptotic regime, where the variable dimension goes to infinity and the sample size is fixed. Each regime has its niche, but only the latter regime applies to exascale data dimension. For this high-dimensional framework with correlation mining as the basis for illustrations, the matrix of pairwise and partial correlations is of particular interest. Correlation mining arises in numerous applications and subsumes the regression context as a special case. The paper concludes with demonstrations of correlation mining in various regimes, from the unifying perspective of high-dimensional learning rates and sample complexity for different structured covariance models and different inference tasks.

"**Resource Allocation for Statistical Estimation**" by Berthet and Chandrasekaran. In various situations that involve acquisition, analysis, and aggregation of data sets from multiple sources, statistical estimation could have significant differences in character as well as value. Consequently, the effectiveness of employing a given resource depends on the nature of that source, and the appropriate division and assignment of resources among a set of data sources can have a strong impact on the overall performance of an inferential strategy. This paper adopts a general view of the notion of a resource and its effect on the quality of the corresponding data source. With statistical efficiency as the objective, several stylized examples involving inferential tasks such as parameter estimation and hypothesis testing based on heterogeneous data sources are discussed. Accordingly, optimal allocations can be computed either in closed form, or via efficient numerical procedures based on convex optimization.

"**Magging: Maximin Aggregation for Inhomogeneous Large-Scale Data**" by Bühlmann and Meinshausen. Analysis of large-scale data poses statistical and computational problems that need to be addressed simultaneously. A solution is often straightforward if the data are homogeneous: Classical ideas of subsampling and mean aggregation can be used to get a computationally efficient solution with acceptable statistical accuracy. The aggregation step simply averages the results obtained on distinct subsets of the data. However, for the more typical case of inhomogeneous data, this simple approach is inadequate—it is influenced too much by effects that are not persistent across all the data (outliers or time-varying effects, for example). This paper shows that a tweak to the aggregation step can produce an estimator whose influences are common to all the data. Thus, the procedure often results in a better prediction than would be the case with pooled effects.

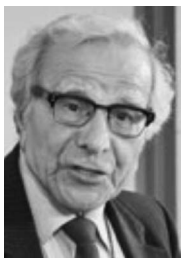
"**Learning Reductions That Really Work**" by Beygelzimer *et al.* This paper presents a summary of the mathematical and computational techniques that have enabled learning reductions to address a wide class of tasks effectively.

This suite of approaches is shown to be broadly useful in solving machine-learning problems. The techniques are instantiated and tested in a machine-learning library called Vowpal Wabbit to prove their practical viability.

"**Taking the Human Out of the Loop: A Review of Bayesian Optimization**" by Shahriari *et al.* Big data applications are typically associated with systems that involve large numbers of users, massive complex software systems, and large-scale heterogeneous computing and storage architectures. The construction of such systems involves many distributed design choices, and thus may involve many tunable configuration parameters, which are often specified and hard-coded into the software by various developers or teams. Joint optimization of these parameters may result in significant improvements. Bayesian optimization is a powerful tool for performing this joint optimization. It promises greater automation so as to increase both product quality and human productivity. This paper introduces Bayesian optimization, highlights some of its methodological aspects, and showcases a wide range of applications.

"**Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets**" by Leung *et al.* This last paper provides an introduction to machine-learning tasks in genomic medicine. One of the objectives of genomic medicine is to determine how variations in the DNA of individuals can affect the risk of different diseases, and to find causal explanations so that targeted therapies can be designed. The objective is to show how machine learning can help to model the relationship between DNA and the quantities of key molecules in the cell, with the premise that these quantities, called cell variables, may be associated with disease risks. Modern biology allows high-throughput measurement of many cell variables, including gene expression, splicing, and protein binding to nucleic acids, all of which can be treated as training targets for predictive models. With the growing availability of large-scale data sets and advanced computational techniques such as deep learning, researchers in machine learning can help to usher in a new era of effective genomic medicine. ■

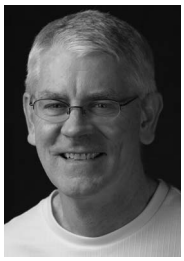
ABOUT THE GUEST EDITORS



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

He is 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 to learn from the human brain and apply 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.



Stephen J. Wright is the Amar and Balinder Sohi Professor of Computer Sciences at the University of Wisconsin—Madison, Madison, WI, USA. His research is on computational optimization and its applications to many areas of science and engineering. Prior to joining UW—Madison in 2001, he was a Senior Computer Scientist at Argonne National Laboratory (1990–2001) and a Professor of Computer Science at the University of Chicago (2000–2001). He is the author or coauthor of widely used text/reference books in optimization including *Primal Dual Interior-Point Methods* (Philadelphia, PA, USA: SIAM, 1997) and *Numerical Optimization* (New York, NY, USA: Springer-Verlag, 2006, 2nd ed., with J. Nocedal). He has published widely on optimization theory, algorithms, software, and applications.

Prof. Wright has served as Chair of the Mathematical Optimization Society and as a Trustee of the Society for Industrial and Applied Mathematics (SIAM). He is a Fellow of SIAM. In 2014, he won the W.R.G. Baker award from the IEEE. He is the Editor-in-Chief of the *SIAM Journal on Optimization* and has served as the Editor-in-Chief or an Associate Editor of *Mathematical Programming (Series A)*, *Mathematical Programming (Series B)*, *SIAM Review*, *SIAM Journal on Scientific Computing*, and several other journals and book series.

Yoshua Bengio received the Ph.D. degree in computer science from McGill University, Montréal, QC, Canada, in 1991.

After two postdoctoral years, one at the Massachusetts Institute of Technology (MIT) with Michael Jordan and one at AT&T Bell Laboratories with Yann LeCun and Vladimir Vapnik, he became Professor at the Department of Computer Science and Operations Research, Université de Montréal, Montréal, QC, Canada. He is the author of two books and more than 200 publications, the most cited being in the areas of deep learning, recurrent neural networks, probabilistic learning algorithms, natural language processing, and manifold learning. His current interests are centered around a quest for AI through machine learning, and include fundamental questions on deep learning and representation learning, the geometry of generalization in high-dimensional spaces, manifold learning, biologically inspired learning algorithms, and challenging applications of statistical machine learning.

Prof. Bengio is among the most cited Canadian computer scientists and is or has been associate editor of the top journals in machine learning and neural networks. He has been holding a Canada Research Chair in Statistical Learning Algorithms since 2000 and the NSERC Industrial Chair since 2006. Since 2005, he has been a Senior Fellow of the Canadian Institute for Advanced Research, and since 2014 he has codirected its program focused on deep learning. He is on the board of the NIPS foundation and has been Program Chair and General Chair for NIPS. He has coorganized the Learning Workshop for 14 years and cocreated the new International Conference on Learning Representations.