ASA
AMERICAN STATISTICAL
ASSOCIATION
Promoting the Practice and Profession of Statistics

Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society

July 2, 2014

A Working Group of the American Statistical Association[1]

Summary:

The Big Data Research and Development Initiative is now in its third year and making great strides to address the challenges of Big Data. To further advance this initiative, we describe how statistical thinking can help tackle the many Big Data challenges, emphasizing that often the most productive approach will involve multidisciplinary teams with statistical, computational, mathematical, and scientific domain expertise.

With a major Big Data objective of turning data into knowledge, statistics is an essential scientific discipline because of its sophisticated methods for statistical inference, prediction, quantification of uncertainty, and experimental design. Such methods have helped and will continue to enable researchers to make discoveries in science, government, and industry.

The paper discusses the statistical components of scientific challenges facing many broad areas being transformed by Big Data—including healthcare, social sciences, civic infrastructure, and the physical sciences—and describes how statistical advances made in collaboration with other scientists can address these challenges. We recommend more ambitious efforts to incentivize researchers of various disciplines to work together on national research priorities in order to achieve better science more quickly. Finally, we emphasize the need to attract, train, and retain the next generation of statisticians necessary to address the research challenges outlined here.

[1] Authors: Cynthia Rudin, MIT (Chair); David Dunson, Duke University; Rafael Irizarry, Harvard University; Hongkai Ji, Johns Hopkins University; Eric Laber, North Carolina State University; Jeffrey Leek, Johns Hopkins University; Tyler McCormick, University of Washington; Sherri Rose, Harvard University; Chad Schafer, Carnegie Mellon University; Mark van der Laan, University of California, Berkeley; Larry Wasserman, Carnegie Mellon University; Lingzhou Xue, Pennsylvania State University. Affiliations are for identification purposes only and do not imply an institution's endorsement of this document.

Data touch almost every aspect of our lives, from the way we transact commerce on the web, to how we measure our fitness and safety, to the way doctors treat our illnesses, to economic decisions that affect entire nations. The age of Big Data will be a golden era for statistics. Scientific fields are transitioning from data-poor to data-rich and—across industries, science, and government—methods for making decisions are becoming more data-driven as large amounts of data are being harvested and stored. However, alone, data are not useful for knowledge discovery. Insight is required to distinguish meaningful signals from noise. The ability to explore data with skepticism is required to determine when systematic error is masquerading as a pattern of interest. The keys to such skeptical insight are rigorous data exploration, statistical inference, and the understanding of variability and uncertainty. These keys are the heart of statistics and remain to be used to their full potential in Big Data research. Indeed, the National Academies' report, *Frontiers in Massive Data Analysis*,[2] states that "the challenges for massive data … hinge on the ambitious goal of inference."

Statistics—the science of learning from data, and of measuring, controlling, and communicating uncertainty—is the most mature of the data sciences. Over the last two centuries, and particularly the last 30 years with the ability to do large-scale computing, this discipline has been an essential part of the social, natural, biomedical, and physical sciences, engineering, and business analytics, among others.[3] Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility and general uncertainty associated with these discoveries. Because one can easily be fooled by complicated biases and patterns arising by chance, and because statistics has matured around making discoveries from data, statistical thinking will be integral to Big Data challenges.

The Big Data Research and Development Initiative[4] was announced in March, 2012 "to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning." Since the launch of the initiative, the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Defense Advanced Research Projects Agency (DARPA) have launched major Big Data programs. The initiative and the agency programs have benefitted enormously from the Computing Community Consortium (CCC) white papers.[5] To further advance the initiative and programs, we outline some of the current challenges for Big Data that would benefit from statistical thinking. Due to the computational challenges involved, we expect innovation to occur at the interface with computer science.

This document takes a very broad view of statistics as a big tent and uses the term statistician to include a wide variety of people, allowing interdisciplinary areas to be contained in multiple fields. For instance, machine learning incorporates concepts from *both* statistics and computer

[2] http://www.nap.edu/catalog.php?record_id=18374

[3] See for example, the American Statistical Association's *Statistical Significance* documents: http://www.amstat.org/policy/statsig.cfm.

[4] http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

[5] http://cra.org/ccc/visioning/ccc-led-white-papers and, specifically, "Challenges and Opportunities with Big Data," "From Data to Knowledge to Action: A Global Enabler for the 21st Century," and "Big-Data Computing."

science, and it is possible for an individual working in this area to be both a statistician and a computer scientist. By our definition, many data scientists are statisticians. Further, statistical research overlaps with computational disciplines such as machine learning, data mining, and optimization, and is conducted by individuals immersed in specific domains such as bioinformatics, sociology, econometrics, and the physical sciences. Similarly, we use the term, "Big Data," very generally to include the “Four-V” dimensions—volume, variety, velocity and veracity—and many aspects of the knowledge discovery process.

This paper is organized around several societal and scientific areas—biological sciences and bioinformatics; healthcare; civic infrastructure, government and living; business analytics, web searches, and information retrieval; social sciences; and physical sciences—being transformed by Big Data and statistics. These areas are only a sampling of statistically-driven areas that are important to society, not an exhaustive list. Within each section below, we highlight a few of the methodological advances that are important for that application area and generalize to other areas. The last section discusses workforce issues and multidisciplinary teams. Whereas the CCC white papers focused on the full knowledge discovery process (data creation, storage, processing and access, privacy, architectures), this document focuses only on the statistical aspects (e.g., learning, inference, uncertainty quantification). This document is broader in scope than the influential 2010 white paper, “Data-Enabled Science in the Mathematical and Physical Sciences,”[6] touching areas outside the physical sciences but more narrow in the sense of its focus on statistics.[7] Note also that we do not intend this document to be comprehensive of all the ways that statisticians could help advance the Administration’s Big Data Initiative. For example, there are many important advances to be made in statistical theory that will inform new methodology for Big Data.[8]

Our choice of organization for the paper highlights that the impact of methodological advances can best be judged in real world contexts. Data intensive discovery is application-specific and so specific knowledge of the domain is required to be most effective in applying and developing statistical methods. For example, the winning team (led by a statistician) of the Netflix Prize incorporated knowledge of how people rate movies, and heavily relied on the use of descriptive statistics for careful data exploration; this helped them tailor their model to the domain. A similar approach to knowledge discovery for personalized recommendations is taken by large stores like Amazon, search engine companies like Google, and health informaticists who now create personalized recommendations for healthcare patients. Companies like Dstillery use massive amounts of data to provide personalized internet advertisements.[9,10] The technology developed for personalized recommendations is only a hint of how modern statistics, tailored to the domain, will revolutionize many different aspects of society.

[6] https://www.nsf.gov/mps/dms/documents/Data-EnabledScience.pdf

[7] See also, B. Yu. Embracing Statistical Challenges in the Information. *Technometrics 49* (2007), 237-248.

[8] See, for example, *Frontiers in Massive Data Analysis*. http://www.nap.edu/catalog.php?record_id=18374 (footnote 2), and J. Fan, F. Han, and H. Liu. Challenges in Big Data. *National Science Review 00* (2014) 1-22.

[9] R. Kohavi, L. Mason, R. Parekh, Z. Zheng. Lessons and challenges from mining retail e-commerce data. *Machine Learning 57* (2004) 83-113.

[10] C. Perlich, B. Dalessandro, O. Stitelman, T. Raeder, F. Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning* (to appear).

We emphasize that advances will occur most productively when statisticians are working in multidisciplinary teams with computer scientists, mathematicians, and domain scientists. Further, scientific teams working with large data sets should include those with statistical expertise.

Biological Sciences/Bioinformatics

The rapid development of high-throughput technologies has had a profound impact on biological research and biomedical applications. Biology has changed from a data-poor discipline to a data-intensive one. For example, before high-throughput technologies came about, the analysis of gene expression data amounted to spotting black dots on a piece of paper or extracting a few numbers from standard curves. Today biologists regularly sift through large data sets. Furthermore, many outcomes of interest, such as gene expression, are dynamic quantities: different tissues express different genes at different levels and at different times. The complexity is further exacerbated by cutting edge yet unpolished technologies producing measurements found to be noisier than anticipated. This complexity and level of variability makes statistical thinking an indispensable aspect of the analysis. The biologists who used to say "if I need statistics, the experiment went wrong" are now seeking statisticians as collaborators. The results of these collaborations have led to, among other things, the development of breast cancer recurrence gene expression assays that identify patients at risk of distant recurrence following surgery.[11] In this section, we will focus on genomics as one example in the biological sciences where statistics has had a profound impact and will continue to be critical to further scientific advances. For a similar discussion in neuroscience, see the American Statistical Association white paper, "Statistical Research and Training under the BRAIN Initiative."[12]

With the capacity to measure millions to billions of genomic features in a single experiment, scientists are better equipped than ever to investigate fundamental questions in basic biology and develop tools that improve medicine and, broadly, human health. Data collected at the population level are being used to decipher how DNA sequences influence and predict an individual's disease risk. Pharmacogenetics—the study of predicting how patients will respond to medication, and what dose they should receive—heavily depends on analysis of large amounts of patient data. The amount of research in biology and medicine that rely on complex data sets is growing and with it, the need for further development of statistical and computational methods.

As an example of the potential for more engagement of statisticians in the biological sciences, consider the now mature Bioconductor project (www.bioconductor.org), an effort of the statistical community tackling challenges across a broad range of genomic technologies. The paper describing the project[13] has been cited over 6,000 times and the most popular software packages are downloaded more than 100,000 times a year. Bioconductor centralizes and standardizes statistical packages for the analysis of large data sets in genomics. One of the

[11] L.J. van 't Leer, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415* (2002) 530-536.

[12] http://www.amstat.org/policy/pdfs/StatisticsBRAIN_April2014.pdf.

[13] R.C. Gentleman, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 5* (2004).

fundamental insights of the Bioconductor project is that uncertainty matters at multiple levels. The raw data are produced by complicated measurement technologies that introduce bias and variation that must be removed before the measurements represent biological quantities of interest. Similarly, to produce results that will replicate in future studies, it is critical to account for the biological variability between individuals and even between cells within an individual. Examples of the advances made possible by implementing the Bioconductor methods include improved methods of preprocessing and normalization,[14,15] solutions for the multiple testing problem,[16] and performing statistical inference to identify differentially expressed genes.[17] (Each of the papers cited above has been cited more than 2,000 times.) More recently, statisticians have developed some of the most widely used methods for the latest measurement technology.[18]

The scale of genomics data has increased substantially between the introduction of the first microarrays and the availability of the next generation sequencing, aided in part by the open access movement. To separate signal from noise at this scale, there is a need to normalize and standardize these data and then to apply statistical inference. The abrupt increase in the size of data sets and the continued success of projects such as Bioconductor will require integrating statistics with computer science at the interface with biology. We expect this approach to produce new statistical methodologies that can be generally applied. Below, we highlight several methodological areas where further advances will contribute to research objectives in biological sciences, namely *data visualization*, *clustering*, *variance decomposition*, and *multiple hypothesis testing*. (Note that these methodological areas are also important to other sections.)

Data visualization is important for complicated data exploration problems and the knowledge discovery process (e.g., knowledge discovery processes defined by KDD[19] and CRISP-DM[20]). Visualization alone can play an important role in discovery with data. For example, scientists used visualization techniques to discover 18 genes that play a role in muscular dystrophy.[21] A visualization of the critical RBP1 protein—a genomic carrier of Vitamin A necessary for reproduction and vision—revealed the previously unknown involvement of RBP1 in cell cycle control.[22] Visualization is also used to identify bias and systematic errors and was used

[14] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2003) 249-264.

[15] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30* (2002).

[16] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Statistical Methodology 64* (2002) 479–498.

[17] G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3* (2004).

[18] S. Anders, W. Huber. Differential expression analysis for sequence count data. *Genome Biology 11* (2010) R106.

[19] G. Piateski and William Frawley. *Knowledge Discovery in Databases*. MIT Press, Cambridge, 1991.

[20] Cross-Industry Standard Process for Data Mining. The CRISP Process Model, 1999.

[21] P. Zhao, J. Seo, Z. Wang, Y. Wang, B. Shneiderman, and E.P. Hoffman. In vivo filtering of in vitro MyoD target data: An approach for identification of biologically relevant novel downstream targets of transcription factors. *Comptes Rendus Biologies 326* (2003) 1049-1065.

[22] S. Stubbs, and N. Thomas. Dynamic green fluorescent protein sensors for high-content analysis of the cell

development of the widely used preprocessing and normalization techniques in genomics.[14]

The statistics community has a long history of developing data visualization techniques—not just histograms, boxplots, scatterplots, but also techniques such as trellis plots and dynamic graphs. More recently, visualization software has become popular, including GGobi and ggplot2 (downloaded 250,000 times in 2013 alone.) Software stemming from the human-computer interaction community is popular for data visualization (e.g., TIBCO Spotfire with 500 employees and Tableau with a $5 Billion market capitalization as of this writing.) The daily use of visualizations in the NY Times, Wall Street Journal, USA Today, Bloomberg, and other media sources speak to the power and value of visualization. (the blog FlowingData provides hundreds of public information examples). Modern visualization techniques such as treemap[23,24] and other techniques for visualizing network data[25,26] are going to be heavily in demand, and new ways of visualizing complex data with specific properties will need to be developed. Combining data mining with visualization has the potential to exceed the power of either field alone[27]

Clustering involves the separation of data into meaningful subsets to facilitate scientific insight. It has been used, for example, to discover new tumor subclasses which in turn resulted in the development of better prognosis in the clinic.[28] Myriad clustering methods have been designed over the last two decades, including methods that cluster nodes of graphs, methods that use low-dimensional structure of the data space, methods that look for dense core sets of variables and observations, approximate methods that scale well, and methods that have other special properties. Bioinformatics problems, including clustering genes, cells and samples from patients heavily depend on methods that produce high quality clusters.

A key problem in clustering is to develop methods with clearly defined statistical goals. Clustering methods are difficult to evaluate computationally since ground truth is not well defined, particularly for Big Data problems where experts cannot manually determine the quality of the clusters. Examples of clustering methods based on a statistical model are mean-shift clustering (originally developed in the computer vision literature), where clusters are defined by way of modes of a density function, and the k-means algorithm. Unfortunately, these methods either do not scale well—either in terms of sample size or in terms of the number of features—or do not produce high quality clusters. Even simpler statistical models, like the maximum likelihood method, turn out to be NP-hard (computationally intractable) and so are less than optimal for large data sets. Thus, an important open problem is to find ways to scale these

cycle.

[22] *Methods in Enzymology 414* (2006) 1–21.

[23] B. Shneiderman, C Plaisant. Treemaps for space-constrained visualization of hierarchies. http://www.cs.umd.edu/hcil/treemap-history/.

[24] http://en.wikipedia.org/wiki/Treemapping/

[25] M. Hansen, B. Shneiderman, and M.A. Smith. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann Publishers, 2011.

[26] D. Keim, J. Kohlhammer, G. Ellis, and G. Mansmann, G. (Editors). *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, Goslar, Germany, 2010.

[27] Ben Shneiderman. *Inventing discovery tools: combining information visualization with data mining.* Information Visualization (2002) 1, 5-12.

[28] T Sørlie, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences 98* (2001) 10869-74.

methods for Big Data.

For some problems, it is necessary to cluster on several variables simultaneously. As an example, consider a matrix of gene expression values on a collection of individuals where each row of the matrix represents a subject and each column a gene. Clustering genes and people simultaneously, for example, is known as bi-clustering. The current best practical bi-clustering algorithms do not achieve the optimal statistical performance. That is, there is a gap between computational tractability and statistical optimality. Filling this gap is an important and challenging problem.

Another Big Data challenge is to incorporate clustering into supervised learning methods. For example, credit card companies first segment customers and then build predictive models of risk separately for each segment. Ideally, they should cluster with the predictive outcome in mind, but it is not clear how to do this, particularly considering the computational tractability problems inherent in clustering.

Variance decomposition methods are fundamental statistical tools (e.g., ANOVA, factor analysis, PCA, multi-level modeling). Understanding the sources of variation and the relative contribution of each source to the total data variance is crucial to discovery. For instance, with biological high-throughput technology data, even the same experiment conducted by different people or at different times can give very different results.[29] The discovery of such "batch effects" illustrates how variance decomposition provides important insights for the design and development of data analysis strategies. Although many methods for analyzing data variance exist, variance decomposition for Big Data is a new challenge. Methods that are computationally efficient and capable of handling heterogeneous data types (e.g., continuous values, discrete counts, etc.) and complex data structure (e.g., trees, networks, or domain-specific structures such as exon-intron structures in genes) are greatly needed. Statisticians have a fruitful history of dissecting data variance and will continue to play a leading role in this area.

Multiple hypothesis testing is a fundamental problem in statistical inference where the primary goal is to consider a set of hypothesis tests simultaneously. The classical methods (such as the Holm–Bonferroni method[30]) are intended to control the probability of making at least one false discovery (i.e., the familywise error rate). However, with high-throughput technology advances, it is now common to conduct large-scale hypothesis testing, with tens of thousands or even millions of tests performed simultaneously. In such testing, classical methods have low power, and controlling the familywise error rate is no longer suitable, especially in the presence of rare and weak signals, demonstrating the need for new statistical methods. In response to this, statisticians have been working on a more appropriate approach for large scale hypothesis testing. The false discovery rate concept was introduced as a promising alternative, and there are now methods to control the false discovery rate when signals are rare and relatively strong. The

[29] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, R.A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews: Genetics 11* (2010) 733-739.

[30] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6 (1979) 65-70.

seminal paper[31] has been cited more than 22,000 times, and controlling the false discovery rate has gained broad recognition in scientific fields such as clinical trials, microarray experiments, brain imaging studies, and astronomical surveys. In the more challenging setting where the signals are rare and weak, some progress has been made with the higher criticism approach and Bayesian approaches to incorporate certain prior information. However, with the increasing dimensionality and complexity of massive data sets, modern applications face an exploding number of tests, often with complex dependencies between tests (e.g., spatial or temporal dependence). Thus, it is important to develop further statistical methods to solve large-scale simultaneous hypothesis testing problems.

Healthcare and Public Health

A McKinsey report estimates that data analytics could save $300 billion to $450 billion annually in U.S. healthcare costs.[32] Statisticians can play a large role in this cost savings through the improvement of quality of care, evaluation of care, and personalized predictions of disease risk. Statisticians can also play a role in addressing challenges for the use and functionality of analytics in practice. [32]

Healthcare providers have an obligation to provide the best possible care. Characterizing the best care requires principled (i.e., evidence-based) comparative evaluation of existing guidelines, procedures, and protocols associated with every clinical intervention. Many healthcare systems have built or are building the infrastructure to collect, store, manipulate, and access records of all patient contacts with the healthcare system. These records include billing and diagnosis codes as well as adverse event reports, unstructured text, genetic data, and hospital operations data. There is enormous potential to use these records to improve healthcare quality. Examples include: (i) evaluating existing procedures across a range of domains including efficacy, side-effect burden, cost, and quality of life; (ii) constructing individualized patient predictions for health outcomes; (iii) identifying high-risk patients as candidates for a preventative intervention; (iv) monitoring diseases for outbreak detection and prevention; and (v) informing the design of and recruitment for clinical trials. However, drawing inference from non-experimental data like electronic health records requires careful statistical thinking to account for potential sampling bias, treatment bias, non-standardized phenotype and diagnosis definitions, and evolving clinical guidelines and standard of care. Healthcare data are massive, can vary from patient to patient in a very significant way, and often omit important information such as a definitive diagnosis. Over the course of the past century, statisticians working with observational data to draw causal inferences have developed a rigorous mathematical framework for analyzing non-experimental data. Statisticians have contributed in demonstrating how data can be used to improve healthcare, often working with medical domain experts.

Program evaluation for quality of care is a high profile research area in healthcare. For instance, consider the new Accountable Care Organizations (ACOs), created under the federal Patient

[31] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Statistical Methodology 57* (1995), 289-300.

[32] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken. *The 'big data' revolution in healthcare: Accelerating value and innovation.* McKinsey and Company, 2013.

Protection and Affordable Care Act, which coordinate care for Medicare patients. With a system of healthcare providers working together, the Centers for Medicare and Medicaid Services foresee improved patient outcomes and substantial cost savings. However, measuring quality of care is challenging, because it is difficult to control for heterogeneous patient populations in a way that allows accurate estimates of key quality-of-care statistics. These statistics are essential in order to compare the quality of care provided by ACOs to the quality of care provided by, for example, a fee-for-service model, where services are provided and billed separately, and physicians may receive incentives to order a larger number of services.

Personalized predictions of disease risk, as well as time to onset, progression, or disease-related adverse events, has the potential to revolutionize clinical practice. Such predictions are important in the context of mental health, cancer, autoimmune disorders, diabetes, inflammatory bowel diseases, stroke, and organ transplants, among others. Statisticians are using huge amounts of medical data to make personalized predictions of disease risk, to understand the effects (benefits and harms) of drugs and other treatments in addition to environmental factors, to analyze the quality of care, and to understand changing trends in health. A major challenge in developing personalized health predictions is that the data are usually observational time series data and there is a myriad of different factors that could possibly be considered together as risk factors for a disease. Statisticians are working to create models with huge amounts of longitudinal observational data to create fine-grained personalized predictions; this involves identifying important patient biomarkers and deriving measures of statistical uncertainty. Statisticians are also using machine learning techniques to identify possible adverse effects of drugs, and to determine which populations are at the highest risk for adverse events.

A main challenge will be how to share information across individuals in a way that yields high quality predictions from data. Statisticians have made great strides in hierarchical Bayesian modeling in order to borrow strength across individuals. Now the challenge for these models is scalability: we must build these models on a very large scale yet still preserve predictive accuracy.

The major advances in public health benefitting from statistics, both from the past and with Big Data, speak to the importance of further engaging statisticians going forward. For instance, statisticians and statistics played a key role in studies on the effects of air pollution on health,[33] on the case for water fluoridation,[34] and the dangers of cellular telephones while driving.[35] A recent example where Big Data were critical is the withdrawal of the drug Vioxx after increased levels of heart attacks were observed after the start of its widespread use.[36] There is also an IBM research team that is using machine learning methods to classify text messages collected by

[33] S.L. Zeger, F. Dominici, A. McDermott, J.M. Samet. Mortality in the Medicare Population and Chronic Exposure to Fine Particulate Air Pollution in Urban Centers (2000-2005). *Environmental Health Perspectives 116* (2008) 1614-1619.

[34] A. Chadwick, T. Lamia, and J. Zaro-Moore. *Building capacity to Fluoridate.* Department of Health and Human Services, Centers for Disease Control and Prevention. June 2003.

[35] D. Redelmeier and R. Tibshirani. Association between Cellular-Telephone Calls and Motor Vehicle Collisions. *New England Journal of Medicine 336* (1997) 453-458.

[36] D. Madigan, D.W. Sigelman, J.W. Mayer, C.D. Furberg, & J. Avorn, Under-reporting of cardiovascular events in the rofecoxib Alzheimer disease studies. *American Heart Journal 164* (2012) 186-193.

UNICEF from across Uganda to determine where humanitarian aid is most needed.[37] Statisticians have also been proved invaluable in cases where research reproducibility is an issue (e.g., where clinical trials were halted based on a statistical scrutiny of the study.[38]) Hospital administrators are also beginning to explore benefits derived from Big Data. For instance, Atlanta-based Grady Health System implemented a decision support system based on machine learning for decision support in emergency departments in hospitals.[39] The potential for data analytics to transform neonatal care has been recognized, and has led to platforms such as Artemis, for which pilot studies are being conducted.[40]

Below we highlight three methodological areas where further advances will contribute to research objectives for healthcare and public health, namely *causal inference*, *bias reduction*, *pattern mining*, and *model tradeoffs*.

Causal inference is often a core goal in medical, healthcare, and public health research. Understanding the true causal impact and comparative effectiveness of treatments or care in terms of outcomes, satisfaction, and cost may involve the analysis of electronic medical records, claims databases, and quality surveys from thousands, and eventually, millions of people. However, these records are observational data rather than data generated from controlled clinical trials, and determining causal effects can be challenging. Even determining causal effects in randomized controlled trials is not straightforward due to imperfect randomization and informative drop-out, among other complications.

Traditional statistical methods for estimation of causal effects rely heavily on the use of parametric models in order to provide measures of effect and statistical significance. These are models that assume that the entire data-generating distribution (i.e., the underlying mechanism that created the data) can be defined by relatively few parameters, and therefore make highly unrealistic assumptions in complex healthcare settings. Accordingly, these methods may be unreliable due to bias introduced by misspecified parametric models and are generally not flexible enough to handle the large number of variables available in the era of Big Data. In response to these challenges, methods have been developed to obtain valid inference when relying only on realistic assumptions about the probability distribution of the data as described by semiparametric and nonparametric models. These semiparametric and nonparametric models only incorporate real subject-matter knowledge about the data experiment and avoid convenient, yet untrue, assumptions. Estimation and inference for big data within these realistic models is then the goal. Statisticians have been working in the area of targeted learning over the last decade, which is a field that integrates the state-of-the-art in machine learning methods for causal

[37] P. Melville, V. Chenthamarakshan, R. Lawrence, J. Powell, and M. Mugisha. Amplifying the Voice of Youth in Africa via Text Analytics. Proceedings of the 19th Conference on Knowledge Discovery and Data Mining (KDD-13) , Chicago, IL, August, 2013.

[38] E.S. Reich. Cancer Trial Errors Revealed. *Nature 469* (2011) 239-140. G. Kolata. How Bright Promise in Cancer Testing Fell Apart. *New York Times*. July 7, 2011 (http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=0).

[39] H.Y. Atallah, E.K. Lee. Modeling and Optimizing Emergency Department Workflow. 2013. https://client.blueskybroadcast.com/Informs/2013/wagner/pdf/Modeling%20ED%20Throughput.pdf.

[40] C. McGregor, Big Data in Neonatal Intensive Care. *Computer 46* (2013) 54-59.

and statistical inference. Notably, high-dimensional propensity score[41] adjustment methods have been used to study the causal effects of treatment in healthcare claims data.[42] Targeted learning methods have also been used at the FDA to improve causal inference in clinical trial analysis by controlling for baseline confounders.[43]

Advances in causal inference will continue to rely on deep statistical theory, and statistical innovations are being tailored to the new challenge of accurately estimating the uncertainty of highly data-adaptive estimators of causal effects. The theory of causal graphs[44] developed in both the computer science and statistics literature is complementary to this estimation framework, and will likewise continue to evolve in both communities. Thus, progress in causal inference for big data builds on a rich foundation in the fields of statistics and computer science.

Methods to handle bias, confounding, and missing data are central to big-data applications. Typically the data to which we have access are not from the population we aim to model; and, even when the populations overlap, available data will not generally be anything like a random or representative sample of the population. In particular, this means we can have sparse data for important sub-populations. Predictions under the presence of bias can be considerably challenging when there are a large number of variables: the stratification cells become smaller and smaller, requiring sophisticated modeling when simple reweighting no longer suffices. For instance, when conducting surveys using non-random samples (e.g., data that are opportunistically available), bias correction can be facilitated by techniques such as multilevel regression that allow adjustment for poststratification cells representing small slices of the population.

One important set of applications of these methods is in healthcare research: If we are able to account for the effects of bias, very large amounts of data can be (and have been) used broadly for prediction of adverse events (e.g., stroke), predicting adverse drug reactions, drug efficacy, dose-response effects, and patients’ perception of quality of care.

Censored and missing data issues are endemic in medical databases. Consider, for example, patients who change service providers but whose medical records are only partially transferred (or not at all.) Further, not all important medical information is contained within the service-provider’s database. For instance, a patient not complying with medical practitioners’ instructions may not be reported. Further, patients’ dietary information and many other aspects of patients’ general health are often not reported, or reported inaccurately.

Pattern mining is the problem of finding all “interesting” correlations or patterns between

[41] P.R. Rosenbaum, D.B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1983), 41-55.

[42] S. Schneeweiss, J.A. Rassen, R.J. Glynn, J. Avorn, H. Mogun, and M.A. Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology 20* (2009) 512–22.

[43] K. Moore and M. van der Laan. Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. *Statistics in Medicine 28* (2009) 39-64.

[44] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.

entities in a database, whether causal or simply correlative. This becomes increasingly difficult as the size of the data grows, for example in healthcare problems. Data mining methods that scale up to large data sets in order to detect useful correlations and patterns are needed. This can be most successfully (and efficiently) accomplished with statisticians working both with data mining experts to design measures of “interestingness” and with computer scientists to build algorithmic techniques for finding these patterns. Much has been accomplished on this topic over the past three decades but the quantity of data has changed substantially in that time, and the types of data (e.g., networks, text, images, video) has also changed. Finding patterns in structured Big Data (for instance, time series) or patterns with specific properties is challenging and requires new algorithmic techniques and true multidisciplinary efforts.

Models that trade off competing goals, such as interpretability and accuracy, are important for decision making in healthcare. There is a long tradition of close engagement of statisticians with clinicians in improving healthcare and public health. For instance, statisticians have been engaged as leaders with clinical scientists, policy makers, healthcare informaticists, and patient advocacy groups to embed predictive models in healthcare. With the flood of new streams of data now becoming available, this engagement has the potential to reach an entirely new level. A high-quality predictive model in healthcare must balance several potentially competing objectives. For example, to be useful, the model must be sufficiently flexible so as to produce accurate predictions. Yet, to be trusted and widely adopted, the model must also be parsimonious, transparent, and interpretable. Similarly, predictive accuracy may increase with the number of patient biomarkers collected but these biomarkers are potentially expensive and burdensome to collect. Statisticians are actively working on building sequences of predictive models of increasing complexity and requiring an increasing set of patient biomarkers where a subsequent predictive model is used only if preceding models cannot produce a sufficiently precise estimate.

Many other important statistical areas are relevant to healthcare including natural language processing, which is particularly important for health record analysis, and, as discussed below, integration of heterogeneous data sources such as doctor’s notes, test results, results of medical studies, and data from other sources.

Civic Infrastructure, Governance, Demography, and Living

The root of the word “statistics” means state, owing to statistics initially being concerned with affairs of the state. Statistics continues to inform a multitude of societal topics from transportation, the Internet, energy, and communication to government, law enforcement, and education. With large amounts of data now being collected from government and the private sector—aided by the open data movement—we are in a position to use and develop statistical techniques in a way that deeply assists with both policy and way of life. Indeed, the recently established Center for Urban Science and Progress at New York University to help cities become more productive, livable, equitable, and resilient exemplifies this potential as does the City of Chicago Data Portal. The federal statistical system is also working to leverage administrative records—records collected for administrative purposes (e.g, social security records)—with its survey work to provide timely, accurate statistical data.

Data and statistics have long been used to inform policy around a multitude of issues listed above. The influential studies of Riis on New York tenements in the 1880's exemplify this.[45] Modern success stories include, for example, Paris21's efforts[46] to develop poverty reduction policies in Africa and Global Forest Watch's efforts[47] to combine heterogeneous data sources (satellite, open data, crowdsourced data) to identify regions of deforestation around the world. The Gapminder foundation[48] uses statistics in order to combat ignorance of important world trends in many areas, including poverty and life expectancy. Statistics and data mining for crime analysis and predictive policing have also had a major impact on such areas as the way police patrol in major cities,[49,50] the way police respond to domestic violence cases,[51,52,53,54] sentencing,[55] and crime policy.[56] Statistics was used to demonstrate improved efficiency and better performance for the electric utility grid.[57] Statisticians have also made important contributions on the challenges of measuring traffic[58,59] and civic infrastructure maintenance.[60] Finally, statisticians have made great progress towards public use of government microdata with synthetic data techniques to protect respondents' privacy.[61]

[45] J. Riis. *How the Other Half Lives: Studies among the Tenements of New York* (Kessinger Publishing, 2004). Originally published in 1890. Available at http://www.authentichistory.com/1898-1913/2-progressivism/2-riis/index.html.

[46] http://paris21.org/sites/default/files/Success%20stories%20in%20stats%202012%20low%20res_0.pdf

[47] http://www.wri.org/our-work/project/global-forest-watch

[48] http://www.gapminder.org/

[49] For a summary, see http://www.nij.gov/topics/law-enforcement/strategies/hot-spot-policing/Pages/welcome.aspx.

[50] D. Weisburd and C.W. Telep. Hot Spots Policing: What We Know and What We Need to Know. *Journal of Contemporary Criminal Justice 30* (2014) 200–220.

[51] L. Sherman and R. Berk. The Specific Deterrent Effects of Arrest for Domestic Assault. *American Sociological Review 49* (1984) 261-271.

[52]R. Berk and L Sherman. Police Responses to Family Violence Incidents: An Analysis of an Experimental Design with Incomplete Randomization." *Journal of the American Statistical Association 83* (1988) 70-76.

[53] R. Berk, A. Campbell, R. Klap, and B. Western. The Differential Deterrent Effects of An Arrest in Incidents of Domestic Violence: A Bayesian Analysis of Four Randomized Field Experiments. *American Sociological Review 5* (1992) 689-708.

[54] Sherman, L. and Cohn, L.1989 "The Impact of Research on Legal Policy: The Minneapolis Domestic Violence Experiment" Law and Society Review, Vol. 23, No. 1, 117-144, 1989.

[55] The Growth of Incarceration in the United States: Exploring Causes and Consequences, National Research Council (2014) http://www.nap.edu/catalog.php?record_id=18613.

[56] R.A. Berk. *Criminal Justice forecasts of Risk: A Machine Learning Approach*. New York, Springer, 2012.

[57] R.N. Anderson, A. Boulanger, W.B. Powell, and W. Scott. Adaptive Stochastic Control for the Smart Grid. *Proceedings of the IEEE 99* (2011) 1098-1115.

[58] P.J. Bickel, C. Chen, J. Kwon, J. Rice, E. van Zwet and P. Varaiya. Measuring Traffic *Statistical Science 22* (2007) 581-597.

[59] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, Expectation, and Surprise: Methods, Designs,

[59]and Study of a Deployed Traffic Forecasting Service. http://arxiv.org/ftp/arxiv/papers/1207/1207.1352.pdf.

[60] C. Rudin et al. Analytics for Power Grid Distribution Reliability in New York City. *Interfaces*, (2014).

[61] S.K. Kinney, J.P. Reiter, A.P. Reznek, J. Miranda, R.S. Jarmin, and J.M. Abowd. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review 79* (2011) 362–384.

New sources of data—public and private—hold exciting potential for statistics to have an impact in society. Predictive policing is one area where significant advances are possible since criminals' patterns are, in some sense, repetitive and thus predictable. Massive online courses collect detailed data from tens if not hundreds of thousands of people across the world that take a single course simultaneously, allowing the study of learning on the micro-scale. Traffic flow in cities should use statistics to adjust traffic signals in real-time based on the predicted level of traffic and its uncertainty. Aging energy grids could be made substantially more sustainable with the use of reliability models. Emergency services could be transformed with analytics to help determine ambulance placement in order to minimize response time. Better estimates for the impact of technology on environmental resources and biodiversity are also possible. Further, we will be able to analyze the costs and benefits of capital project expenditures that will allow us to make better investments for the future of our cities.

Some combination of new and existing statistical methods for conducting observational studies will be essential to determine infrastructure vulnerabilities based on data from sensors and other monitors. Our infrastructure is often composed of networks, including transportation, Internet, energy grids, and communication networks (e.g., social networks), and statistical techniques that are specifically designed for sampling network data and modeling network growth will be important, as also discussed later. Many pieces of our infrastructure (e.g., transportation networks) will rely on control systems that also benefit from statistical expertise.

Integrating heterogeneous data sources remains a major challenge in Big Data, including for infrastructure and governance. For a single project it may be necessary to combine sensor data, textual data, presentation slides, time series, and other forms of structured and unstructured data. Data integration can help one to use complementary information in different data sources to make discoveries. It may also allow one to better detect weak signals by leveraging the correlation structure among different data sets. For instance, when performing demographic studies, we might integrate several data sources in order to more accurately track people's residences and employment status over time. Integrating heterogeneous data sources is challenging for many reasons including that unique identifiers between records of two different data sets often do not exist. For example, people's names are not unique, are misspelled, or can change. Further, determining which data should be merged may not be clear at the outset. Often, working with heterogeneous data is an iterative process in which the value of data is discovered along the way and the most valuable data are then integrated more carefully. Integration of heterogeneous data is a well-studied problem in the data management literature as is information extraction from heterogeneous sources. Combining statistical thinking with computer science input will lead to new scalable and automated (or semi-automated) tools for Big Data integration. Integration is also a major challenge in the healthcare domain.[62]

Anomaly and changepoint detection is an important statistical area whose goal is to find patterns that are out of the ordinary and do not conform to expected behavior. Such patterns often provide critical actionable information. Similarly, changepoint detection methods sense changes in a signal's distribution. Security/cybersecurity monitoring, infrastructure resilience, fraud detection,

[62] G.M. Weber, K.D. Mandl, I.S. Kohane. Finding the Missing Link for Big Biomedical Data. *JAMA* (2014) E1-E2.

spam detection, biosurveillance, and quality control in manufacturing are important application areas for anomaly detection and changepoint detection. Modern applications have significantly challenged our ability to detect anomalies within massive data of increasing complexity and dimension. However, traditional methods usually require that the number of features be smaller than the sample size, and do not provide an effective control for false positive discoveries. Most existing methods do not scale well for Big Data problems. Further, we require methods that can measure the extent to which an object or value is an anomaly. Recently, statisticians working with hospital data used an unusual variety of anomaly detection techniques to pinpoint inappropriate access detection to medical records. As a result, some offenders identified by the statistical method were faced with termination of employment.[63,64]

Business Analytics, Internet Search Engines, and Recommendation Systems

Business analytics, retrieval systems, and recommendation systems are three business areas where statistics has been and will continue to integral.

Business analytics is one of the fastest changing areas in the Data Science world, with new business school programs, new departments, and new positions (e.g., "Chief Data Officer"). Companies are now collecting data from all aspects of their businesses: speed, quality, and reliability of their product manufacturing and assembly processes; customer service; marketing; customer perceptions of their products; surveys about potential successes of future products; product ratings; and competitors' products. Realizing the value of data has not reached its full potential, many companies are hiring data scientists to harness some of that potential to improve many aspects of business practices.[65] It is essential these data scientists possess a statistical perspective, which emphasises the quality of the data, the power of sampling, the characteristics of the sample, the validity of generalization, and the balance of humans and machines.[66]

Because businesses are starting to realize the power of data for tracking customers and understanding their preferences, they have started to collect massive amounts of data about different aspects of customer behavior, leading to the current discussions around data privacy. Web browsing, web search histories, and online purchase histories of customers can be very powerful information. Protecting individual privacy is an important legal topic, but also an important statistical topic. Statisticians have long been leaders on privacy and confidentiality issues and, working with computer scientists and others, will help to address these very important issues for Big Data.[67]

[63] A. Boxwala, J. Kim, J. Grillo, and L. Ohno-Machado, Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association 18* (2011) 498–505.

[64] A.K. Menon, X. Jiang, J. Kim, J. Vaidya, L. Ohno-Machado. Detecting inappropriate access to electronic health records using collaborative filtering. *Machine Learning* 95 (2014) 87-101.

[65] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Eamon Dolan/Houghton Mifflin Harcourt (2013).

[66] K. Fung. The pending marriage of big data and statistics. *Significance 10* (2013) 20-25.

[67] J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Editors). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press, 2014.

For some businesses, statistics are the backbone of their products; search engine companies and internet advertising companies are two examples.

Current search engines generally tell us where to find information whereas the next generation of search engines, now starting to emerge, will do much more than that. New search engines will return information, organize it in a meaningful way, and personalize it to the user. These new search engines will have a core comprised of new statistical methodology in order to be customizable and personalized for different types of tasks. For instance a search engine customized for, say, military strategy might aim to return a full overview of information about the local area, local militants, neighboring regions, and a summary of local social media. In contrast, a search engine customized for set completion aims to return more examples similar to those given by the user (e.g., "aquamarine, sky blue" - find more like these). In that case, the search engine must determine in what contexts these examples appear (colors) and be able to find more (e.g., turquoise, robin egg blue). To be personalized to each user, a search engine must condition their queries on information that includes the user's browsing history, the browsing history of similar users, and general trends in browsing over time. Another expectation is for the search engine to deliver results instantaneously for a particular user and query. Simple statistical models that have the right modeling assumptions will allow us to scale up to the point where enormous calculations can be done in parallel and instantly. For instance, for the set completion problem mentioned earlier, there is a simple statistical technique that scales extremely well.[68,69]

Recommendation systems will undoubtedly grow in importance and extend to an even wider swath of application areas as the amount of data grows. The goal is to generate ever more accurate recommendations, made all the more challenging by the increasing amount of data and its expanding heterogeneity, including a diverse user population that changes rapidly over time. For advertisers, their goal is to present advertisements that use so many relevant characteristics of the user that the user might actually want to receive almost every ad. Or, for health-related outcomes as discussed above, the goal would be that each patient has an individualized risk or recovery recommendation created from data about other patients. Using heterogeneous data sources like videos and images, search engine companies will be able to use the latest technologies in machine vision and speech recognition to allow searches that retrieve the right information rapidly and accurately.

Tradeoffs between speed and accuracy are important towards the practical building of recommender systems and to large scale statistical modeling in general, and broadly to decision-making problems in business analytics. Recent work at the intersection of statistics and computer science has focused on the idea of tradeoffs, where prediction quality, statistical estimation, or quality of inference trades off with computation time. Taking this one step further, prediction/inference quality should also trade off with other important characteristics of modeling including the interpretability of solutions, computational memory, and the ability to

[68] Z. Ghahramani and K.A. Heller. Bayesian Sets. *Proceedings of Advances in Neural Information Processing Systems* (2005).

[69] B. Letham, C. Rudin and K.A. Heller. Growing a List. *Data Mining and Knowledge Discovery 27* (2013) 372-395.

make decisions. Formalizations of these types of ideas would help us discover these tradeoffs explicitly. We are now resource-limited in several possible ways (computation, memory, interpretability, etc.) and these need to be taken into account in domain-specific ways.

This will require redefining what optimization means in statistical modeling because there are now multiple practical problems to consider. A method that is asymptotically optimal but in practice runs out of memory or does not produce good solutions for finite data is not practical. A method that creates more interpretable solutions can be more computationally difficult. In general, methods and analysis must adapt based on the most important aspects over which to optimize, and what the resources allow. This may involve dealing with a partial ordering of estimators induced by asymptotic efficiency, computational efficiency, and memory requirements; thus requiring notions of Pareto optimality for estimators, or something similar. Collaboration between statistical and computer scientists will be essential, as trade offs often occur between statistical and computational properties.

Social Sciences

A fundamental and ongoing question in social science research concerns the relationship between individuals' environments and their thoughts, opinions, and actions. While there is little doubt that a person's environment matters, neither existing data nor statistical tools have been able to adequately capture these relationships. Emerging data sources mean that we now have a better picture than ever before of a person's environment. Data from social media outlets, for example, demonstrate real-time reactions of millions of individuals to both local (e.g., a friend makes a racist comment) and global (e.g., a major news story) events. Even more possibilities arise from recent work leveraging multiple forms of complementary data, such as social media or other communications data, to understand individual situations while overlaying high-resolution location data to capture characteristics of individuals' neighborhood environments. Administrative records, mentioned above, also offer opportunities to contribute to the understanding of broad changes in population composition and the nuanced dynamics of specific groups or individuals.

Leveraging new forms of data provides opportunities to take another look at some of the most fundamental questions in the social sciences. Social science has long pondered the role of an individual's social network in her/his decision making process. The notion of homophily, or the tendency for similar individuals to interact, has been observed in both close personal networks (e.g., people one would trust with sensitive information) and weaker tie association networks (day-to-day connections). For the first time we can observe on a large scale not just the current state of a community but also the formation of communities. The temporal dimension offers new opportunities to tease apart, for example, whether individuals tend to seek out others who are similar, or whether they first interact and then become more similar through mutual influence. Identifying the role of other factors (e.g., a person's physical environment or the political climate) in influencing behavior is another opportunity.

Despite great potential, substantial challenges prevent these new data sources from fully penetrating mainstream social science. Though there are some computing challenges (e.g., appropriately managing and curating data that are structurally very different from surveys commonly used by social scientists), the largest barriers are statistical. First, these new forms of data are completely unsolicited. That is, information is only available when a respondent chooses to take an action or make a statement. Respondents who choose to reveal information about, for example, their political opinions on Twitter likely differ in substantively important ways from those who do not. We must be careful to ensure that algorithms that learn from historical decisions made by humans do not also learn human biases and prejudices. Techniques from survey statistics, epidemiology, and missing data methods provide opportunities to recalibrate results or assess sensitivity, as discussed earlier in this document. Further, social network data arising through social media or other observational sources are fundamentally different from data typically used by social scientists to study networks. Typical social science work on social networks elicits the network directly by asking respondents, for example, to nominate their friends from a population of potential friends. New forms of data, however, require social scientists to infer these deeper relationships. These data instead present only continuous-time manifestations of underlying relationships (e.g., the frequency and duration of calls between two individuals). Statistical techniques are then required to extract underlying patterns in temporal and social network structure.

To illustrate the potential for statisticians to be an important part of the advances in the social sciences in the Big Data era, consider a few examples. Statistics was central to new UN population and fertility probabilistic projections obtained by combining demographic models with Bayesian methods.[70,71] These methods were central to producing projections released by the United Nations which influence national policies worldwide. The large scale field experiments that revolutionized political campaigns are another example where statistical methods to adjust for non-compliance were critical.[72] Finally, statistics was central to research deducing the behavior of individual group members, work that has subsequently been used in litigation by both sides in every U.S. state over the Voting Rights Act.[73]

Network science is an emerging interdisciplinary field that studies "network representations of physical, biological, and social phenomena leading to predictive models of these phenomena."[74] Statistics has been used to make key contributions to network inference and analysis (e.g., the p^* model and the stochastic block model) and will be central to further advances. Many currently important sources of data have a natural network structure, as exemplified by social networking sites, transportation routes, energy grids, biological systems and communication networks. There are specific challenges in working with network data, such as how to model the spread of information through a network, how to infer the structure of networks, how to sample on networks, how to analyze networks over time, and so on. Classical maximum likelihood

[70] A.E. Raftery, N. Li, H. Ševčíková, P. Gerland, and G.K. Heilig. Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences 109* (2012) 13915–13921.

[71] L. Alkema, A.E. Raftery, P. Gerland, S.J. Clark, F. Pelletier, T. Buettner, G.K. Heilig. Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography 48* (2011)815-839.

[72] A.S. Gerber, D.P. Green, and C.W. Larimer. Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review 102* (2008) 33-48.

[73] G. King, *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press, 1997.

[74] *Network Science*. National Research Council. Washington, DC: The National Academies Press, 2005.

estimation breaks down in these settings due to the curse of dimensionality and network dependence, requiring new estimators to be developed.

A major challenge for network analysis is to address the fact that the individual is not the causal unit, due to the underlying dependence structure of the data; people's actions are interdependent. Substantial challenges also arise from the need to present results at multiple scales. Characterizing the forces that drive a large network of blogs to become more or less polarized, for example, likely differs from describing friendship formation in a small group of individuals. Further, much network data increasingly come with unstructured covariates (e.g., text in a social media network). Differentiating, for example, an interaction on a social media site that arises from an existing relationship from one that occurs out of mutual interest remains an open but important problem in the study of network communication. Causal models from statistics and network analysis tools from computer science are currently being expanded to address some of these challenges but much more remains to be done.

Data collection, large-scale experimental design, and crowdsourcing are core topics of interest to many different social sciences. We now have the ability to crowdsource large amounts of survey data. Some of these data can be obtained quickly and at a relatively low cost. We may even be able to alter the survey while it is being taken in some cases. For instance, crowdsourcing has been useful for studying human visual perception, gathering relevance assessments of books and videos, affect recognition, policy adoption (e.g., for climate change), public opinion polling, and targeted marketing. The ability to rapidly obtain such large quantities of data means new challenges when designing surveys and experiments (keeping in mind that a million empirically observed pieces of data are often far less informative than a thousand well-chosen ones.)

Crowds of humans are also useful for solving problems that cannot be solved well by computers. For example, it may be difficult to find teaching assistants to grade tens of thousands of homework assignments in a massive online open course, but crowdsourcing the grading tasks among students themselves may offer a feasible solution. Crowds have been useful for many different kinds of tasks, and have made product suggestions for companies, have cast votes to express their opinions on various ideas, and have completed software testing to find bugs in code. Crowdsourcing produces data with complex bias and variance structure. For example, data provided by different people may have different quality. Some tasks are intrinsically harder than others and therefore require more people and resource to provide input. How to design a crowdsourcing project, how to evaluate and control the quality of data contributed by different people, and how to optimally allocate the available resources and integrate the data are all important questions that remain to be answered. Many of these problems are statistical in nature, requiring one to use existing principles and methods of experimental design, survey sampling, and data integration, and to develop new ones that are applicable to large-scale crowdsourcing projects.

Innovative statistical models: Whether we are modeling an individual's food product preferences at a grocery store in a similar way as we would model molecular interactions, or whether we model gang violence in the same way as aftershocks of an earthquake, we often tailor statistical models to the domain in order for inference to be meaningful. Models must capture domain-specific features, and capture domain-specific constraints, such as monotonicity of certain types of effects, or preservation of certain orderings (e.g., one event always follows another in time). For instance, if a statistical model is able to produce predictions that cannot physically happen, this can harm the quality of its results, and it can undermine the believability of the model. Some areas where novel modeling is happening include: (i) Time series and point processes: social network data are all time stamped, allowing us to model their evolution through innovative time series models. In epidemiology, diseases can be modeled as a point process. (ii) Network connectivity models for understanding the structure of social, biological, and physical networks. (iii) Hierarchical clustering models that embed structure among the clusters. (iv) Dynamical systems models with uncertainty, for instance for atmospheric science (as discussed in the section below on physical sciences) or systems biology. Dynamical systems with uncertainty bring together the realms of physics, applied mathematics, and statistics. This type of approach brings domain constraints into the models through the dynamics, but allows the flexibility of fitting to data with a quantification of uncertainty.

Innovative models are also needed to approximate more complex models that cannot be fit to massive amounts of high-dimensional data. These new kinds of models should harness low-dimensional structure, exploit conditional independences, and other simple structures that enable efficient computations and inference. Innovative modeling is relevant for all of the domains listed here and beyond.

Physical Sciences/Geosciences

There are many data challenges and opportunities for engagement of statisticians in the physical sciences. The 2010 white paper of a NSF-sponsored workshop, "Data-Enabled Science in the Mathematical and Physical Sciences,"[6] provides a solid introduction of the data issues for the mathematical and physical sciences. Here we focus on the statistical perspective for two areas, astrostatistics and climate science.

Ongoing astronomical surveys such as the Sloan Digital Sky Survey (SDSS) gather measurements on millions of galaxies, quasars, stars, and other celestial objects. The Large Synoptic Survey Telescope (LSST), currently under construction in Chile, will push these counts into the billions. Despite the wealth of information, reaching the goals of precise estimates of key cosmological parameters will require the development of novel and sophisticated analysis methods. For example, the Planck satellite measured the Cosmic Microwave Background (CMB) to unprecedented precision, but translating these data into constraints on cosmological parameters required the careful implementation of Markov Chain Monte Carlo methods adapted to the complexities of the situation.[75] Further recent analysis of the CMB revealed the presence of gravitational waves, providing long-sought confirmation of the inflationary hypothesis regarding the formation of the Universe. Such a discovery was only possible after the use of statistical methods that adjust for contamination and other observational limitations.[76] This is

[75] The Planck Collaboration, et al. Planck 2013 Results. XVI. Cosmological Parameters. To appear in *Astronomy and Astrophysics,* (2014).

[76] The BICEP2 Collaboration. Detection of B-Mode Polarization at Degree Angular Scales by BICEP2. *Physical Review Letters,* (2014).

typical of astronomical data sources: they are noisy and high-dimensional; other examples include thousand-dimensional emission spectra, blurry images, and irregularly-spaced time series. Finding low-dimensional structure in these massive data collections is of particular importance, as this structure can be related to important object-specific properties. These properties must then be used to estimate a population-level summary statistic, with an objective of not discarding useful scientific information while adjusting for the biases that result from our Earth-centered observations. Finally, the relationships between the cosmological parameters of interest and these summaries are complex and in some cases the only sufficiently accurate picture of the likelihood function comes from computationally expensive simulation models.

Similarly, statisticians have much to contribute in climate science and meteorology. Statisticians have long used data to develop scientific knowledge, convey uncertainty, and inform decision and policy makers. Achieving these goals in the context of predicting weather and understanding climate and the impacts of climate change requires enhanced collaborations among statisticians and scientists from diverse disciplines, ranging from the climate and environmental sciences to agriculture, biology, economics, sociology, and human health and wellness. The most fundamental challenges faced by climate scientists involve making predictions regarding future behavior of aspects of the climate system, and attaching useful measures of uncertainty to those predictions. Examples include forecasts for the number of hurricanes in the Atlantic during the upcoming season and longer-range forecasts for the global average temperature decades into the future. As was the case in astrophysics, the complexity of the systems under study requires the use of simulation models to gain insight into their natural trends and variability. Two further challenges where statisticians could be very helpful are in combining information from different models (the multi-model ensemble problem), particularly when the models give conflicting results, and in incorporating observational data into the climate models (data assimilation).

Numerical models for complex phenomena such as the climate system rely on uncertain inputs and other quantities. The development of these quantities is increasingly reliant on the extraction of information from massive data sets. Though these computer models are usually based on well-established science, they do involve approximations and are subject to a variety of uncertainties such as model errors, numerical errors, etc. Assessment of the value of these models requires special attention in that their outputs are massive data sets and the data used to judge and correct them is contained in massive data sets. The goal of understanding, assessing, and modeling the effects of uncertainties in input data and models in large scale problems is now known as uncertainty quantification. As suggested by a new collaborative initiative[77] between the Society for Industrial and Applied Mathematics and the American Statistical Association, this area is recognized as critical for applications and ripe for new growth.

Climate scientists also work with historical records, but face significant challenges in the analysis of this information as rich, reliable data are often only available for the past half-century. The available observations, combined with the output of simulation models, yields a collection of massive and diverse sources of information, linking numerous processes in a variety of spatial-temporal scales. In the context of weather forecasting, the needs for data analyses and their rapid assimilation into forecast models are spawning new challenges to statistical modeling.

[77] SIAM/ASA Journal on Uncertainty Quantification, https://www.siam.org/journals/juq.php.

In the context of climate, massive data sets collected by NASA and other providers are to be merged with climate system models, namely large-scale computer models that also produce massive data sets. All of these problems require solutions that quantify and manage the uncertainties in our data and our science, reinforcing the importance of engaging statisticians.

The construction of data sets that combine observations over large spatial and temporal scales, such as temperature averages over large grid boxes or the full globe, is another challenge. Single data sets may contain jumps due to instrument changes or spurious trends due to heat island effects. Statistical methods are needed to identify and remove these effects. The spatial distribution of meteorological stations is very far from uniform, and simple averaging over all stations may not be adequate to capture large-scale spatial variability. Recently, it has been suggested that the widely reported leveling off of global warming after 1998 may be an artifact of the sampling network, which would be corrected with more sophisticated methods based on spatial statistics.

To illustrate the promise of further engagement of statisticians with Big Data challenges in climate science, we provide a few examples of contributions from statisticians. Statisticians have developed treatment for uncertainty in paleoclimate reconstructions that were pivotal to establishing that temperature increases over recent decades are rapidly overwhelming previous maxima.[78] Statisticians have also made important contributions to the North American Regional Climate Change Assessment Program,[79] an international program that produces high resolution climate change simulations in order to investigate uncertainties in regional scale projections of future climate and generate climate change scenarios for use in impacts research. Statistical work was also critical to the development of the PRISM data product, the standard for monthly U.S. surface temperature and rainfall.[80] For more on further opportunities in climate science, see the parallel white paper from the ASA Advisory Committee on Climate Change Policy, "Statistical Science: Contributions to the Administration's Research Priority on Climate Change."[81]

Astrophysics data, climate change data, and meteorological data are so enormous that they cannot be handled in traditional ways. In what follows we discuss two topics related to statistical modeling on the massive scale: optimization and estimation.

Optimization for statistical modeling techniques or machine learning algorithms generally becomes more difficult as the size of the data set increases. Handling huge amounts of data using new platforms (e.g., cloud databases) is addressed elsewhere and here we will briefly mention some challenges related to statistics:

[78] B. Li., W.D. Nychka, and C.M. Ammann. The "Hockey Stick" and the 1990s: A statistical perspective on reconstructing hemispheric temperatures, *Tellus 59* (2007) 591-598.

[79] L.O. Mearns, R. Arritt, S. Biner, M.S. Bukovsky, S. McGinnis, S. Sain, D. Caya, J.Correia, D. Flory, W. Gutowski, E.S. Takle, R. Jones, R. Leung, W. Moufouma-Okia, L. McDaniel, A.M.B. Nunes, Y. Qian, J. Roads, L. Sloan, and M. Snyder. The North American Regional Climate Change Assessment Program: Overview of Phase I Results. *Bulletin of the American Meteorological Society 93* (2012) 1337-1362.

[80] C. Johns, D. Nychka, T. Kittel, and C. Daly. "Infilling Sparse Records of Precipitation Fields." *Journal of the American Statistical Association 98* (2003) 796–806.

[81] "Statistical Science: Contributions to the Administration's Research Priority on Climate Change," April 2014, http://www.amstat.org/policy/pdfs/ClimateStatisticsApril2014.pdf.

1. Scaling up basic statistical modeling techniques, such as logistic regression, is extremely important. For instance, there has been work done on logistic regression in distributed environments and on parallel architectures. There has also been work on distributed versions of the bootstrap and related resampling methods for assessing the quality of statistical estimators.

2. Optimization should be taken into account within the design of a statistical model as discussed earlier in the context of speed and accuracy trade offs. One approach here is to take advantage of independence assumptions between variables, or of natural convex approximations, to enable inference on larger scales. A classic example of this is the Naive Bayes algorithm, which makes strong independence assumptions that reduce optimization to simple counting. Greedy approximations (e.g., decision tree algorithms) can also be extremely useful in practice. The machine learning community has been pursuing online learning approaches, subsampling methods, and algorithms designed to handle larger data sets more efficiently than they handle smaller ones. They sometimes use submodular functions over subsets, which are easier to optimize than arbitrary functions over subsets. The statistics community has been pursuing convex approximations to hard problems (e.g., the lasso for producing sparse models), and bootstrapping approaches.

3. Hard problems for small to moderately sized datasets should also be addressed. There are some problems that are intrinsically hard (non-convex, non-smooth), and for which no known reasonable approximation algorithm exists. Decision tree algorithms are an example of this—they solve hard problems using greedy methods. Problems of moderately-sized data become Big Data problems when they are significantly complex. For instance, sometimes we are required to create millions of variables in order to solve a problem with a much smaller data set (e.g., discrete optimization problems). In the era of Big Data we should not neglect these important problems. For instance, how would one compute an optimal decision tree in a reasonable amount of time (within hours)? Could an approximation other than a greedy approximation be useful? There are many other relevant questions in this category that have been overlooked by the machine learning community for years; that community has almost exclusively favored methods that are extremely computationally efficient on large data sets, at the expense of other aspects of modeling.

Markov Chain Monte Carlo (MCMC) techniques for Bayesian analysis must be extended to handle learning in large-scale data sets. These methods incorporate prior information while characterizing uncertainty in a principled manner. MCMC remains the gold standard but tends to scale poorly, particularly as the number of variables increases. There has been a recent focus on MCMC algorithms that can be implemented in a distributed manner where data are split into “shards” that are stored on different computers. For instance, the consensus Monte Carlo algorithm simply runs MCMC in a parallel manner for each data set, and then averages draws from each subset. (When subset posteriors are approximately Gaussian, this has some justification but otherwise substantial errors may arise.)

In many modern scientific applications ranging from neuroimaging to genomics, the sample size is small to moderate but the number of variables measured on each study subject is massive. There is a critical need for developing divide-and-conquer methods for scaling MCMC using massive computing clusters to distribute the computational burden arising from a massive numbers of parameters. These types of methods currently become less practical as the variables depend on each other more heavily.

Multidisciplinary Teams and Next Generation Statisticians

The importance of multidisciplinary teams is a theme throughout this document and, in this section, we expand on it and also discuss the great need to attract, train, and retain the next generation of statisticians to contribute to the research challenges outlined here.

Increasingly over the years, the National Science Foundation and other federal agencies have recognized the key role that multidisciplinary work will play in handling problems of importance to our society. Multidisciplinary teams—with each discipline having much to learn from other disciplines—will ensure the best science is brought to bear, help to avoid reinvention of existing techniques from the contributing data science disciplines, and spur development of new theories and approaches.

The history of statistics shows how statisticians have engaged in interdisciplinary research, and how that engagement advanced domain sciences, informed policy, and provided new insights. For example, John Graunt, who invented many of our modern descriptive statistics and the study of populations, was studying the plague in England, and Siméon Denis Poisson developed his eponymous distribution in his studies of wrongful convictions. R.A. Fisher, who spent a portion of his career at a leading agricultural research center and later became a professor of genetics at Cambridge, is another excellent example of someone making fundamental contributions to statistics while primarily working in another field. His work inspired the modern versions of the p-value, analysis of variance, and maximum likelihood inference. Survival analysis, regression, and many other areas of statistics are other examples of statistical advances arising from domain-specific problems.

Today, statisticians continue to contribute expertise to a growing range of scientific and social problems. Their training and experience in collaborative research make them the natural leaders of interdisciplinary teams. As discussed throughout this paper, however, the Big Data era is marked by an ever-growing class of truly important and hard interdisciplinary problems where more engagement would be widely productive.[82]

Despite the laudable federal efforts to encourage multidisciplinary research, more should be done to have researchers of various disciplines working on issues so important to our country and society to achieve better science more quickly. Implicit in the recommendation for more ambitious efforts to encourage multidisciplinary teams is the need for statisticians doing interdisciplinary work to be acknowledged and supported. There should be incentives for

[82] See for instance, S.H. Jain, M. Rosenblatt, J. Duke. Is Big Data the New Frontier for Academic-Industry Collaboration? *JAMA 311 (*2014) 2171-2172.

statisticians doing such research, where more emphasis is placed on the impacts in domain science research or on societal challenges rather than only theoretical or methodological advances. Having NSF recognition of the value of such interdisciplinary research—through funding for problems centered around Big Data or discovery from data—would help promote such work more broadly. In faculty evaluation and promotion considerations, universities should recognize statistical contributions to research advances in other disciplines (e.g., joint publications, contributions to grants on which they are not the principal investigator, publications in domain-specific journals). Impactful work based on statistical discoveries from data, without necessarily advancing statistical methods or theory, should have recognized and appreciated publishing venues.

The demand for data scientists—many of whom will be statisticians and all of whom will require significant knowledge of statistics—is booming. While steadily growing (doubling from 2003 to 2012), the number of PhDs in statistics and biostatistics granted in the U.S. annually has only just surpassed 500.[83] In the work of OSTP, NSF, and other federal agencies to address STEM workforce issues, we strongly encourage attention to attracting and retaining the next generation of statisticians, especially those who can work seamlessly across disciplines. The statisticians engaged in interdisciplinary research involving Big Data will need to be computationally savvy, possessing expertise in statistical principles and an understanding of algorithmic complexity, computational cost, basic computer architecture, and the basics of both software engineering principles and handling/management of large-scale data. Just as critical—for both the training and recruitment—as the computational skills are interpersonal skills for the next generation of statisticians to be effective communicators, leaders, and team members.

Conclusion

In this paper, we have described many Big Data challenges at the interface of statistics and computer science, highlighting those where statistical thinking is required and multidisciplinary teams involving statisticians important. Statistical thinking fuels the cross-fertilization of ideas between scientific fields (biological, physical, and social sciences), industry, and government. Throughout, we have made the case that further engagement of statisticians and cutting-edge statistics (as one of the core data science disciplines) will help advance the aims of the Administration's Big Data Initiative. We have also emphasized the need to attract and train the next generation of statisticians.

Reviewers

The following have reviewed this document and agreed to have their names listed. Affiliations are for identification purposes only and do not imply endorsement of this document.

Genevera Allen, Rice University
Mohsen Bayati, Stanford University
Kristin Bennett, Rensselaer Polytechnic Institute
James Berger, Duke University

[83] http://magazine.amstat.org/blog/2013/10/01/undergrad-women/

David Banks, Duke University
Brian Caffo, Johns Hopkins University
Marie Davidian, North Carolina State University
Richard D. De Veaux, Williams College
Finale Doshi, MIT
Jianqing Fan, Princeton University
Andrew Gelman, Columbia University
Robert Gentleman, Genentech
Joydeep Ghosh, University of Texas, Austin
Lauren Hannah, Columbia University
David Higdon, Los Alamos National Laboratory
Peter Hoff, University of Washington
David Jensen, University of Massachusetts, Amherst
Michael Jordan, University of California, Berkeley
Sallie Keller, Virginia Tech
Gary King, Harvard University
Vipin Kumar, University of Minnesota
Edward Lazowska, University of Washington
Andrew W. Lo, MIT
Thomas Lumley, University of Auckland
David Madigan, Columbia University
Katherine Pollard, University of California, San Francisco
Foster Provost, New York University
Ben Shneiderman, University of Maryland
Richard L. Smith, Statistical and Applied Mathematical Sciences Institute
Terry Speed, Walter and Eliza Hall Institute of Medical Research
Hal S. Stern, University of California, Irvine
Galit Shmueli, Indian School of Business
Christopher Winship, Harvard University
Wing H. Wong, Stanford University
Bin Yu, University of California, Berkeley

To be added as a reviewer, email ASA Director of Science Policy Steve Pierson: pierson@amstat.org.