Frontiers in Massive Data Analysis

Committee on the Analysis of Massive Data

Committee on Applied and Theoretical Statistics

Board on Mathematical Sciences and Their Applications

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

PREPUBLICATION DRAFT – Subject to Further Editorial Correction

COMMITTEE ON THE ANALYSIS OF MASSIVE DATA

MICHAEL I. JORDAN, University of California, Berkeley, *Chair*
KATHLEEN M. CARLEY, Carnegie Mellon University
RONALD R. COIFMAN, Yale University
DANIEL J. CRICHTON, Jet Propulsion Laboratory
MICHAEL J. FRANKLIN, University of California, Berkeley
ANNA C. GILBERT, University of Michigan
ALEX G. GRAY, Georgia Institute of Technology
TREVOR J. HASTIE, Stanford University
PIOTR INDYK, Massachusetts Institute of Technology
THEODORE JOHNSON, AT&T Labs Research
DIANE LAMBERT, Google, Inc.
DAVID MADIGAN, Columbia University
MICHAEL W. MAHONEY, Stanford University
F. MILLER MALEY, Institute for Defense Analyses
CHRISTOPHER OLSTON, Google, Inc.
YORAM SINGER, Google, Inc.
ALEXANDER SANDOR SZALAY, Johns Hopkins University
TONG ZHANG, Rutgers, The State University of New Jersey

Staff

SUBHASH KUVELKER, Study Director (until October 17, 2011)
SCOTT WEIDMAN, Study Director (after October 17, 2011)
BARBARA WRIGHT, Administrative Assistant

1
Introduction

THE CHALLENGE

Although humans have gathered data since the beginning of recorded history—indeed, data gathered by ancestral humans provides much of the raw material for the reconstruction of human history—the rate of acquisition of data has surged in recent years, with no end in sight. Expectations have surged as well, with hopes for new scientific discoveries pinned on emerging massive collections of biological, physical, and social data, and with major areas of the economy focused on the commercial implications of massive data.

Although it is difficult to characterize all of the diverse reasons for the rapid growth in data, a few factors are worth noting. First, many areas of science are in possession of mature theories that explain a wide range of phenomena, such that further testing and elaboration of these theories requires probing extreme phenomena. These probes often generate very large data sets. An example is the world of particle physics, where massive data (e.g., petabytes per year for the Large Hadron Collider; 1 petabyte is 10^{15} bytes) arises from the new accelerators designed to test aspects of the Standard Model of particle physics. Second, many areas of science and engineering have become increasingly exploratory, with large data sets being gathered outside the context of any particular theory in the hope that new phenomena will emerge. Examples include the massive data arising from genome sequencing projects (which can accumulate terabytes (10^{12} bytes) of data for each project) as well as the massive data expected to arise from the Large Synoptic Survey Telescope, which will be measured in petabytes. Rapid advances in cost-effective sensing mean that engineers can readily collect massive amounts of data about complex systems, such as those for communication networks, the electric grid, and transportation and financial systems, and use that data for management and control. Third, much human activity now takes place on the Internet, and this activity generates data that has substantial commercial and scientific value. In particular, many commercial enterprises are aiming to provide personalized services that adapt to individual behaviors and preferences as revealed by data associated with the individual. Fourth, connecting these other trends is the significant growth in the deployment of sensor networks that record biological, physical, and social phenomena at ever-increasing scale, and these sensor networks are increasingly interconnected.

In general, the hope is that if massive data could be exploited effectively, science would extend its reach, and technology would become more adaptive, personalized, and robust. It is appealing to imagine, for example, a health-care system in which increasingly detailed data are maintained for each individual—including genomic, cellular, and environmental data—and in which such data can be combined with data from other individuals and with results from fundamental biological and medical research, so that optimized treatments can be designed for each individual. One can also envision numerous microeconomic consequences of massive data analysis where preferences and needs at the level of single individuals are combined with fine-grained descriptions of goods, skills, and services to create new markets. In general, what is particularly notable about the recent rise in the prevalence of "big data" is not merely the size of modern data sets, but rather that their fine-grained nature permits inferences and decisions at the level of single individuals.

It is natural to be optimistic about the prospects. Several decades of research and development in databases and search engines has yielded a wealth of relevant experience in the design of scalable data-centric technology. In particular, these fields have fueled the advent of cloud computing and other parallel

and distributed platforms that seem well suited to massive data analysis. Moreover, innovations in the fields of machine learning, data mining, statistics, and the theory of algorithms have yielded data-analysis methods that can be applied to ever-larger data sets. When combined with arguments that simple algorithms can work better than more sophisticated algorithms on large-scale data (see, e.g., Halevy et al., 2009), it is natural to be bullish on big data.[1]

While not entirely unwarranted, such optimism overlooks a number of major difficulties that arise in attempting to achieve the goals that are envisioned in discussions of massive data. In part these difficulties are those familiar from implementations of large-scale databases—involving finding and mitigating bottlenecks, achieving simplicity and generality of the programming interface, propagating metadata, designing a system that is robust to hardware failure, and exploiting parallel and distributed hardware—all at an unprecedented scale. But the goals for massive data go beyond the storage, indexing, and querying that have been the province of classical database systems (and classical search engines), instead focusing on the ambitious goal of *inference*. Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of variables (e.g., a patient's general state of health, or a shopper's tendency to buy) that are not present in the data per se, but are present in models that one uses to interpret the data. Statistical principles are needed to justify the inferential leap from data to knowledge, and many difficulties arise in attempting to bring these principles to bear on massive data. Operating in the absence of these principles may yield results that are not useful at best or harmful at worst. In any discussion of massive data and inference, it is essential to be aware that it is quite possible to turn data into something resembling knowledge but which actually is not. Moreover, it can be quite difficult to know that this has happened.

Consider a database where the rows correspond to people and the columns correspond to "features" that are used to describe people. If the database contains data on only a thousand people, it may suffice to measure only a few dozen features (e.g., age, gender, years of education, city of residence) to make the kinds of distinctions that may be needed to support assertions of "knowledge." If the database contains data on several billion people, however, we are likely to have heightened expectations for the data, and we will want to measure many more features (e.g., latest magazine read, culinary preferences, genomic markers, travel patterns) to support the wider range of inferences that we wish to make on the basis of the data. We might roughly imagine the number of features scaling linearly in the number of individuals. Now, the knowledge we wish to obtain from such data is often expressed in terms of combinations of the features. For example, if one lives in Memphis, is a male, enjoys reading about gardening, and often travels to Japan, what is the probability that the person will click on an ad about life insurance? The problem is that there are exponential numbers of such combinations of features and, in any given data set, a vast number of these combinations will appear to be highly predictive of any given outcome by chance alone.

As this scenario suggests, a naive appeal to a "law of large numbers" for massive data is unlikely to be justified. If anything, we should expect the perils associated with statistical fluctuations to *increase* as data sets grow in size. Of course, if we do not ask new questions as the data grow in size, but are content to more precisely answer old questions, then statistical error rates may not grow as the data scale. But that is not the perspective that underlies the current interest in massive data.

The field of statistics aims to provide a mathematical understanding of inference, quantifying the degree of support that data offer for assertions of knowledge as well as providing a basis for evaluating actions that are proposed on the basis of these assertions. The field has developed tools not only for computing estimates and evaluating hypotheses, but also for assessing the error rates of such procedures. One example of such a tool is "cross-validation," whereby one holds out a certain fraction of the data (the "held-out data"), runs the estimation procedure on the rest of the data (the "training data"), and tests on the held-out data. This accords well with the intuitive notion that an assertion can be viewed as "knowledge" if it applies not merely to the data at hand, but also to additional data. A difficulty, however, is that if one runs such a procedure many times on a fixed set of held-out data, for example, with many

[1] This report uses the terms "big data" and "massive data" interchangeably to refer to data at massive scale.

combinations of features, then many combinations will appear to be highly supported by both the training data and the held-out data, again by chance alone.

There are many additional issues that impinge on the quality of inference. A major issue is that of sampling bias. Data may have been collected according to a certain criterion (for example, in a way that favors "larger" items over "smaller" items), but the inferences and decisions we wish to make may refer to a different sampling criterion. This issue seems likely to be particularly severe in many massive data sets, which often consist of many subcollections of data, each collected according to a particular choice of sampling criterion and with little control over the overall composition. Another major issue is provenance. Many systems involve layers of inference, where "data" are not the original observations but are the products of an inferential procedure of some kind. This often occurs, for example, when there are missing entries in the original data. In a large system involving interconnected inferences, it can be difficult to avoid circularity. Circularity can introduce additional biases and amplify noise.

Many of these issues can be addressed in principle. For example, there are sophisticated statistical tools that can assess the errors in error-assessment procedures. But in the context of massive data, care must be taken with all such tools, for two main reasons:

1. These tools are all based on assumptions and they can fail when the assumptions are not met. Such assumptions include various assertions regarding sampling, stationarity, independence, and so on. Unfortunately, massive data sets are often collected in ways that seem most likely to break these assumptions.
2. Tools for assessing errors of procedures and for diagnostics are themselves computational procedures that make demands on the computing infrastructure. These demands may be infeasible for massive data sets, even when the underlying procedure is feasible.

Having aimed to temper some of the optimism that is often found in contemporary discussions of massive data, the committee does not want to align itself with an unduly pessimistic stance. The committee believes that many of the issues involved in performing inference on massive data can be confronted usefully, giving rise to an engineering discipline that is based solidly on both inferential and computational principles. But this will necessitate a major, sustained research effort that will require due attention to both the opportunities and the perils of massive data. It is necessary to develop scalable computational infrastructures that embody inferential principles that themselves are based on considerations of scale. Researchers will need to worry about real-time decision cycles and the management of trade-offs between speed and accuracy. While inference is increasingly used to power "data products" that are generated by machines—advanced search engines, movie recommender systems, news story and advertisement selection, and so on—there is also a need to develop tools for bringing humans into the data-analysis loop at all stages, because knowledge is often subjective and context-dependent, and there are aspects of human intelligence that (for the foreseeable future) are beyond the capability of machines.

This effort goes well beyond the province of a single discipline, and one of the main conclusions of this report is the need for a thoroughgoing interdisciplinarity in approaching problems of massive data. The major roles that computer scientists and statisticians have to play has already been alluded to above, and the committee emphasizes that the computer scientists involved in building big data systems must develop a deeper awareness of inferential issues, while statisticians must concern themselves with scalability, algorithmic issues, and real-time decision-making. Mathematicians also have important roles to play, with areas such as applied linear algebra already contributing to large-scale data analysis and likely to continue to grow in importance. But while the focus in much applied mathematical work has historically been on the control of small numerical errors, in massive data analysis, small numerical errors are likely to be dominated by statistical fluctuations and biases, and new paradigms need to be considered. This report also highlights the transdisciplinary domain of optimization theory, which already plays a major role in modern data analysis, but which also needs further shaping so as to meet the particular context of massive data. Also, as mentioned above, the role of human judgment in massive data analysis

is essential, and contributions are needed from social scientists and psychologists as well as experts in visualization. Finally, domain scientists and users of technology also have an essential role to play in the design of any system for data analysis, and particularly so in the realm of massive data, with the explosion of design decisions and possible directions that analyses can follow.

The focus in this report is on the technical issues—computational and inferential—that surround massive data. The committee recognizes that this focus is restrictive; there are also major issues to consider in areas such as public policy, law, and ethics. In particular, issues of privacy and the ownership of data are of major concern, and there are few established cultural or legal frameworks in place to deal with such issues. Privacy is especially important in the context of massive data because of the potential for finding associations across sets of data. Given this potential, should it be acceptable for a cell-phone company to make available tracking data for a large number of customers for academic research without restrictions or controls? Would it be acceptable for law-enforcement purposes? For map-makers or the news media? What if the phone data were correlated with other information to give a picture of the owners' patterns of activities?

The companion to data privacy is data ownership. If Google were to go out of business, who owns all the stored email data? If Google's email service were sold off as a separate business to an overseas entity, who owns that data, and what country's laws apply? If InstaBook were to decide to sell all of the user-posted pictures in their system and also declare copyright ownership of them, is that acceptable, and do the people who posted the data have any recourse? If a government pays for plot maps of all properties in its jurisdiction, are these maps public or private? Can mapping companies use them for free? Can they be kept from taxpayers of the jurisdiction? Many transit agencies now track their buses in real time. Does the public own that data? Can services access it for free to show arrival times for future buses and other useful information?

Such thorny issues of privacy and ownership will need to be resolved as society continues to collect data on individuals and their activities. It is easy to see why such topics merit a full study of their own by a committee with a broad set of expertise.

While these issues will not be addressed in this report, the committee does hope to see them addressed in complementary studies. Two comments may help to connect this report to future reports that focus on privacy and other issues of public policy. First, the committee believes that it is impossible to achieve absolute levels of privacy while exploiting the data that arise from human activity. There will necessarily have to be a trade-off, one which is based on an assessment of the relative value of privacy when compared with the possible gains from data analysis. For society to agree on the terms of this trade-off, it will be necessary to understand exactly what are the possible gains from data analysis. This latter is part of the focus of this report. Second, the focus of this report on computation and inference not only aims to understand what can be achieved from data analysis, but what cannot be achieved (cf. the earlier discussion of statistical errors above). In the context of privacy considerations, it may be desirable that certain inferences cannot be obtained reliably; thus, a clear understanding of computation and inference will help feed an understanding of mechanisms for achieving privacy.

WHAT HAS CHANGED IN RECENT YEARS?

In 1995 the National Research Council's Committee on Applied and Theoretical Statistics held a workshop to examine the challenges and promises of the ability to process massive data. The workshop was documented in *Massive Data Sets: Proceedings of a Workshop* (NRC, 1996). Comparing the situation depicted in that report with the current situation allows three areas to be highlighted where changes have been particularly noteworthy.

First, there has been a qualitative leap in the amount of data regarding human interests and activities, much of it generated voluntarily via human participation in social media. Crowdsourcing is also a new phenomenon, as are massive multiplayer online games. With the rise of such human-oriented data sources comes a number of technical challenges. For example, social data are often relational, taking the

form of networks that link people and objects along a variety of dimensions. Moreover, such data are often fragmentary and subject to a variety of sampling biases. There are also issues of willful misrepresentation and governmental restrictions on access. Finally, human-oriented data often involves natural language and other representations with a rich underlying semantics, and the inferential problems of interest often involve reasoning about underlying causes and human intentions.

Second, distributed computing systems have become a reality, with major implications for the collection and processing of massive data. The 1995 workshop clearly recognized that existing algorithms for data analysis would not scale to the kinds of data set sizes that were beginning to accrue, even accounting for the ongoing increase in central processing unit speed, and solutions were sought in parallel and distributed processing systems. Such systems have begun to emerge, driven by a variety of technological and economic factors, and have opened up new vistas and new challenges. In particular, cloud computing now allows access to very large computing infrastructures through a network on an as-needed basis. This has led to new trade-offs involving storage, networking, and processing. It is also important to emphasize that the issue is not solely that of distributed computing, but also of distributed data. Mobile platforms have proliferated, and data often originate on small devices that have bandwidth limitations that limit or preclude data movement. Moreover, the inferential goals for data analysis often involve bringing together multiple data sources that may have been collected independently. For example, as already noted, social media often provide partial and fragmentary perspectives on individuals, and many questions of interest can only be answered if these perspectives are brought into register. In general, many new challenges have arisen involving computational frameworks that are capable of integrating data across spatial, representational, and administrative domains.

Third, many issues involving the geo-temporal nature of data have come to the forefront. For example, a significant fraction of the data on the Internet is in the form of video streams, a trend that is accelerating. Moreover, a growing number of social media and mobile technologies are generating geo- and time-tagged data. Computer networks generate massive data streams. Scientific data often take the form of time series. In such cases, even if an individual time frame does not involve a massive data set, the temporal sequence can quickly overwhelm storage and computing resources. Indeed, it is common in such cases to develop streaming algorithms that attempt to process the data on the fly, avoiding storage. However, the inferential goals associated with such data often involve the discovery and indexing of temporally extended behaviors, and this generally requires some form of storage. It is also the case that many instances of streaming data require real-time or near-real-time processing; examples include the online auctions run for ad placement in search engines and early alert systems for disease outbreaks. This requirement creates new algorithmic challenges where answer *quality* needs to be traded off against answer *timeliness*. Finally, many data sets are also indexed by spatial coordinates (an issue emphasized in the 1996 NRC report). This creates new algorithmic challenges where answer quality and timeliness needs to be traded off against the geographic granularity of the answer. The overall issue is often that of coping with massive spatio-temporal and geo-temporal data.

Another way to contrast the situation in 1995 with the current situation is to compare the areas of science and technology that were thought to be impacted by massive data issues. Table 1.1 provides a partial listing of these areas, focusing on scientific and engineering fields. Many of the differences depicted in this table can be attributed to the rapid growth in social media, mobile devices, and sensor networks during the past decade and a half.

TABLE 1.1 Scientific and Engineering Fields Impacted by Massive Data

Area Affected in 1995	Area Affected in 2012	Noteworthy Use Cases
Physical sciences	Physical sciences	Astronomy, particle physics
Climatology	Climatology	
Signal processing	Signal processing	
Medicine	Medicine	Imaging, medical records
Artificial intelligence	Artificial intelligence	Natural language processing, computer vision
Marketing	Marketing	Internet advertising, corporate loyalty programs
N/A	Political science	Agent-based modeling of regime change
N/A	Forensics	Fraud detection, drug/human/CBRNe trafficking
N/A	Cultural studies	Human terrain assessment, land use, cultural geography
N/A	Sociology	Comparative sociology, social networks, demography, belief and information diffusion
N/A	Biology	Genomics, proteomics, ecology
N/A	Neuroscience	fMRI, multi-electrode recordings
N/A	Psychology	Social psychology

NOTE: CBRNe, chemical, biological, radiological, nuclear, enhanced improvised explosive devices; fMRI, functional magnetic resonance imaging; N/A, not applicable.

ORGANIZATION OF THIS REPORT

The statement of task for the study that led to this report reads as follows:

The study will carry out the following tasks:

- Assess the current state of data analysis for mining of massive sets and streams of data,
- Identify gaps in current practice and theory, and
- Propose a research agenda to fill those gaps.

A primary audience for this report is the community of researchers who need to be adept at analyzing massive data. Because, as will be seen, this is an inherently multidisciplinary subject, the report assumes the reader has (or is willing to develop) an understanding of topics in computer science (including databases and distributed systems), statistics, and optimization. Another important audience consists of the research organizations, especially federal funding agencies, which are building capabilities for the analysis of massive data. The report's identification of research challenges should help those organizations target their programs.

Chapter 2 provides an overview of some of ways in which massive data are currently arising in various scientific and technological fields. Focusing on systems and computer architecture issues, it discusses general trends and then turns to several examples: Earth and planetary science, astronomy, biological and medical research, large numerical simulations, telecommunications and networking, social

network analysis, and national security. Chapter 3 pursues the systems perspective further, discussing recent developments in parallel and distributed systems, databases, and streaming architectures.

Chapter 4 addresses issues surrounding the temporal nature of data, serving to highlight the fact that many massive data sets arise as temporal streams and that many interesting inferential questions revolve around the detection of temporal trends, changes, and patterns. Moreover, it is often the case that real-time responses are needed.

In Chapter 5, a more general discussion of data representation is provided, including some of the ways in which massive data arrive in raw form and the transformations that are often applied to data, particularly transformations that attempt to reduce the representational complexity of the data.

Chapter 6 turns to a formal treatment of some of the computational complexity issues that arise in the setting of massive data analysis. The discussion focuses on computational resources and the theoretical characterization of trade-offs among these resources.

Chapter 7 and 8 focus on inferential issues. Chapter 7 addresses statistical model-building in the massive data setting, discussing several of the stages in the inferential pipeline, including data cleansing and validation. In Chapter 8 sampling is discussed, focusing on the data-gathering process but also making links to Chapter 5, where sampling is a key methodology for data reduction.

Chapter 9 treats some of the issues that arise when humans are included in the data-analysis loop. This includes crowdsourcing, where humans are used as a source of training data for learning algorithms, as well as visualization, which not only helps humans to understand the output of an analysis, but also allows human input into model revision.

Chapter 10 attempts to bring several of the strands of the report together into a proposal for a taxonomy of some of the major algorithmic problems arising in massive data analysis. The committee hopes that the ideas in this section will serve to organize the research landscape and also provide a point of departure for the design of "middleware" that links high-level inferential goals to the algorithms and hardware needed to achieve those goals.

In accordance with the study's statement of task, Chapters 2 through 10 identify gaps in current theory and practice, and Chapters 3 through 10 propose a number of elements of a research agenda. Finally, Chapter 11 presents the committee's primary conclusions.

REFERENCES

Halevy, A., P. Norvig, and F. Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 2:8-12.

NRC (National Research Council). 1996. *Massive Data Sets: Proceedings of a Workshop.* National Academy Press, Washington, D.C.