Embracing Statistical Challenges in the Information Technology Age

Bin Yu

Department of Statistics
University of California at Berkeley
Berkeley, CA 94720
(*binyu@stat.berkeley.edu*,
www.stat.berkeley.edu/users/binyu)

This article examines the role of statistics in the age of information technology (IT). It begins by examining the current state of IT and of the cyberinfrastructure initiative aimed at integrating the technologies into science, engineering, and education to convert massive amounts of data into useful information. Selected applications from science and text processing are introduced to provide concrete examples of massive data sets and the statistical challenges that they pose. The thriving field of machine learning is reviewed as an example of current achievements driven by computations and IT. Ongoing challenges that we face in the IT revolution are also highlighted. The paper concludes that for the healthy future of our field, computer technologies have to be integrated into statistics, and statistical thinking in turn must be integrated into computer technologies.

1. INTRODUCTION

"*Information technology (IT) is a broad subject concerned with technology and other aspects of managing and processing information, especially in large organizations. In particular, IT deals with the use of electronic computers and computer software to convert, store, protect, process, transmit, and retrieve information*" (Wikipedia).

The roots of information technology (IT) can be traced back at least to the invention in 1946 of the electronic numerical integrator and computer (ENIAC), the first device able to solve a large range of computing problems. ENIAC weighed 27 tons and was a very different creature from what we know today as a computing device: a compact desktop in offices and homes, a book-sized laptop at airports and cafes, and a tiny hand-held device like a cell phone, MP3, and iPod. There have also been major advances in network and measurement technologies that allow us to collect, store, analyze, and transport massive amounts of data. These data come in many forms: numerical, text, image, video, audio, multimedia, and so on. Images and videos are generated by optical sensors on satellites, medical scanners such as position emission tomography and magnetic resonance imaging; biological imaging tools with the aim of understanding macroscale, microscale, and nanoscale activities of cells and molecules; digital sky surveys; security surveillance cameras; personal digital cameras and videos; and more. Audio waveform data, such as those created by human speech, radio broadcasts, movie sound tracks, and concerts, are stored in digital form on computers. Multimedia data consist of text, image, and audio all at once and are the norm for TV programs, movie DVDs, and websites of major newspapers.

The explosion in both the amount and complexity of data has been enabled by the exponential rate of increase in computing and measurement capabilities. Gordon Earle Moore, a cofounder of Intel, predicted in 1964 that the number of macroscale, microscale, and nanoscale components on a chip would increase annually by a factor of 1.5 or 2. Moore's law has been true not only for component density on chips, but also for memory and storage technology. Developments in optical fiber technology have brought the internet into our homes and offices, leading to massive movement and retrieval of data across the networks. Although we may be meeting physical limitations that mean the end to Moore's law, advances in computing (such as parallel, grid, and virtual computing) will ensure that we continue to be faced with a data deluge for years to come.

In 2003, a National Science Foundation blue-ribbon advisory panel on cyberinfrastructure noted in its report (Atkins et al. 2003) that "we now have the opportunity and responsibility to integrate and extend the products of the digital revolution to serve the next generation of science and engineering research and education." The report "Towards 2020 Science" (Emmott et al. 2005), written by a dozen prominent scientists invited by Microsoft, concluded in 2005 that "an important development in science is occurring at the intersection of computer science and the sciences that has the potential to have a profound impact on science. It is a leap from the application of computing to support scientists to do science (i.e., computational science) to the integration of computer science concepts, tools, and theorems into the very fabric of science." This report lists some specific areas of science: earth's life-support systems, biology (cell, immune system, brain), the origin of life, global epidemics, revolutionized medicine, and future energy. It is interesting to note that the types of data in these areas encompass data from large-scale simulation and computational models, as well as experiments and observations.

Both reports send the same unmistakable message: The IT revolution has progressed to a tipping point where a cyberinfrastructure is needed that is well integrated into the very fabric of science, engineering, and education and this cyberinfrastructure is used to convert the massive amounts of data into useful information. According to Wikipedia, cyberinfrastructure "describes the new research environments that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization, and other computing and information processing services over the Internet. In

TECHNOMETRICS, AUGUST 2007, VOL. 49, NO. 3
DOI 10.1198/004017007000000254

scientific usage, cyberinfrastructure is a technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge." As we can see, the term "data" plays a central role in this definition of cyberinfrastructure, suggesting that statistics—the science of dealing with data—has an indispensable role in these new developments. So, it is natural to ask: What is the role of statistics? Are we prepared to play a fundamental role in meeting the challenges of the IT age, or are we content with making incremental changes to our existing paradigms? It is obvious that for the healthy existence of our field, the changes and developments within statistics should be sufficiently fundamental, commensurate with developments in computer science and IT, to meet the changing environment.

The field of statistics indeed has been undergoing major changes over the last few decades. There has been considerable discussion and introspection within the statistics community regarding the challenges and the future of the discipline (see, e.g., Lindsay, Kettenring, and Siegmund 2004). In this article, we attempt to answer the foregoing questions by reviewing our achievements and speculating our future endeavors. Admittedly, the coverage on existing works is selective, and the speculations are limited by our imagination. Nevertheless, it is hoped, that some degree of synthesis is achieved and that the challenges we face are distilled. At a minimum, the article might stimulate more exchanges in the statistics community as well as among the different disciplines to help statistics position itself properly in the future development of cyberinfrastructure.

The rest of the article is organized as follows. Section 2 describes three areas of science in which massive data sets arise. A particular project of ours—arctic cloud detection—is covered in detail and used to suggest a multicomponent framework for interdisciplinary statistical investigations. Section 3 reviews machine learning, a frontier in the field of statistics driven by computation, a key consideration in cyberinfrastructure. Text processing and sensor networks are the focus of Section 4. The discussions give a glimpse of new opportunities brought by IT. Section 5 provides personal views on where exciting statistics research is likely to occur and why. The article concludes with a short discussion section.

2. INTERDISCIPLINARY STATISTICAL RESEARCH: A MULTICOMPONENT ENDEAVOR

Enormous amounts of data are being collected in many fields of science, providing numerous opportunities in interdisciplinary research for statisticians. Even though subject matter varies from one field to another, one statistical challenge is common: how to extract useful information from massive data. This challenge demands new ways of data management, visualization, statistical investigation, and validation. Meeting this challenge undoubtedly requires the integration of statistics (thinking and methodology) into the cyberinfrastructure.

2.1 Tales From Science

2.1.1 Digital Sky Surveys. Data are flooding astronomers from the next generation of sky surveys such as the 2 Micro All Sky survey and the Sloan Digital Sky Survey (cf. Welling and Dearthick 2001; Jacob and Husman 2001). From the SDSS website (*www.sdss.org*),

> "Simply put, the Sloan Digital Sky Survey (SDSS) is the most ambitious astronomical survey ever undertaken. When completed, it will provide detailed optical images covering more than a quarter of the sky, and a 3-dimensional map of about a million galaxies and quasars. As the survey progresses, the data are released to the scientific community and the general public in annual increments."

In a 5-year period, the 2 Micron All Sky Survey and Sloan Digital Sky Survey produced 550 gigabytes of reduced data and 1 terabyte of cutout images around each detected object. These volumes surpass humans' ability to study them in detail, leaving us to rely on computing power to sift through them in real time. This real time processing or streaming data analysis must extract useful information, with the possibility of discarding data on the fly due to storage limitations. Visualization human interaction could be part of this processing, and clustering, classification, and multiple testing seem to be useful inference frameworks within which to address questions raised by these sky surveys.

2.1.2 Particle Matters. In particle physics, gigantic experiments are undertaken to understand the most elementary ingredients of matter and their interactions. One such experiment, the Compact Muon Solenoid at CERN in Geneva, generates about 40 terabytes per second, which must be reduced to about 10 terabytes per day in real time for subsequent analysis. This is another example of streaming data. (For more details, see Knuteson and Padley 2003.)

2.1.3 Arctic Cloud Detection. Arctic cloud detection is a problem very familiar to us, and we describe it here in detail to motivate the steps involved in interdisciplinary collaborations. (See also Speed 2005 for more discussions on interdisciplinary research.)

Arctic cloud detection falls in the increasingly important realm of atmospheric science. Much of the remotely sensed observation data in atmospheric science are publicly available and provide a good resource for statistical research (Braverman et al. 2006). Data are also generated by large-scale computational models, such as Mesoscale Model version 5, a joint effort of the National Center for Atmospheric Research (NCAR) and Penn State University (Berk et al. 2002), and the Weather Research and Forecast Model (*http://www.wrf-model.org/index.php*), a joint effort of five agencies, including NCAR. Both models use atmospheric observations as initial values and solve partial differential equations regarding physical thermodynamic and microphysical processes on a three-dimensional grid. Complex computational and simulation models are also common in many other areas, including meteorology, wildfire control, transportation planning, and immune system function, as evident in the workshop on this topic (Berk et al. 2002). There are also important issues related to model validation, prediction, and other topics. (See Berk et al. 2002 for a discussion of these topics.)

Shi, Yu, Clothiaux, and Braverman (2006b) dealt with the arctic cloud detection problem. This investigation was motivated by the fact that "global climate models predict that the strongest dependence of surface temperatures on increasing atmospheric carbon dioxide levels will occur in the Arctic, and this regional temperature increment can lead to global temperature increase. A systematic study of this relationship requires accurate global scale measurements, especially the cloud coverage, in the Arctic regions. Ascertaining the properties of clouds in the Arctic is a challenging problem, because liquid and ice water cloud particles often have similar properties to the snow and ice particles that compose snow- and ice-covered surfaces. As a result, the amount of visible and infrared electromagnetic radiation emanating from clouds and snow- and ice-covered surfaces is often similar, which leads to problems in the detection of clouds over these surface types. Without accurate characterization of clouds over the Arctic we will not be able to assess their impact on the flow of solar and terrestrial electromagnetic radiation through the Arctic atmosphere and we will not be able to ascertain whether they are changing in ways that enhance or ameliorate future warming in the Arctic" (Shi et al. 2006b).

The Multiangle imaging spectroradiometer (MISR) is a sensor aboard NASA's EOS satellite Terra launched in 1999. MISR takes novel electromagnetic radiation measurements at nine different viewing angles at wavelengths (red, green, blue, and near-infrared) collected initially at 275 × 275 m resolution (leading to about 1 million pixels per image). Due to the high data rate, all other bands except the red are aggregated to the coarser 1.1 × 1.1 km resolution before transmission to the base station on Earth.

It was known to the MISR team that the MISR operational Arctic detection algorithms had not worked very well in the Arctic (or the polar regions). We were invited to work on this problem by Dr. Braverman from JPL and later were fortunated to have an atmospheric scientist Eugene Clothiaux join us. Our goal in this project was to provide better cloud labeling for each pixel based on MISR's red band nine-viewing angle data. (Other bands have a coarser resolution and do not seem to offer more information for the cloud label.) For MISR operational purposes, we would like to have an online algorithm that outputs a label while data come in from the MISR sensor. The first step in obtaining data from NASA data center turned out to be lengthy (more than 3 months). The data format could not be read directly into Matlab at that time, so that we borrowed a special program from the MISR team to convert the data into a suitable form to allow Matlab. Because the data volume was too large to allow Matlab to carry out computations in a timely manner, we programmed a graphical user interface just to obtain some very simple summary statistics. Thanks to the advancing computer technology, stereo-visualization of MISR images is now available through the Leica photogrammetry suite (LPS) by Leica Geosystems (*www.leica-geosystems.com*), with a customized interface for MISR data. This provides the necessary means to obtain validation data from experts to estimate the cloud heights, the next goal of our project.

The MISR red-band data is nine-dimensional per pixel corresponding to nine angles, and there are about three million pixels per image block (which consists of three original images). The MISR operational algorithm is called stereo-derived cloud mask (SDCM), which uses the red-band data to first retrieve cloud height based on matching of clouds in images of different angles. The cloud mask, or estimated cloud-or-not label, is obtained by thresholding the cloud height based on the terrain height. The matching step is computationally expensive and error-prone in the polar regions.

Our first breakthrough came 6 months after we embarked on the project. We realized that we could look for "snow/ice" pixels, bypassing the error-prone cloud height retrieval underlying the MISR operational algorithm. The next 3 years witnessed our interactions with the MISR team: presentations at MISR science meetings, e-mail exchanges, and visits. As a result, a simple and effective cloud detection algorithm called enhanced linear matching clustering (ELCMC) (Shi et al. 2006b) has been devised and tested. The cornerstone of our algorithm is three physically meaningful features from the MISR red-band measurements. We then fused the MISR–ELCMC labels with MODIS cloud labels (MODIS is another sensor on Terra that is hyperspectral but one angle) to get the training data to apply quadratic discriminant analysis (QDA) for a probability label of a pixel (Shi, Clothiaux, Yu, Braverman, and Groff 2006a). When compared with the best "ground truth" data (expert labels), our algorithm gives an average of 94% accuracy on labeled pixels of 60 blocks of data of millions of pixels (whereas MISR–SDCM gave only 80%). Moreover, our algorithm is able to give a label for every pixel, whereas SDCM has a label for only 27% of the pixels. Finally, it is noteworthy that two expert labels typically differ by about 5%, so our method is basically reproducing expert labels automatically.

2.2 A Multicomponent Framework for Interdisciplinary Research

The cloud detection experience reveals several considerations that arise in interdisciplinary research involving large amounts of data:

1. Access to good scientific or subject problems and expertise
2. Collection and management of large data sets (including effective transmission and storage and possibly data reduction or feature selection)
3. EDA (visualization and descriptive statistics and possibly also data reduction or feature selection)
4. Processing mode: offline or online (streaming data)
5. Formal modeling with computation and accuracy considerations (estimation and uncertainty assessment)
6. Data fusion from various sources
7. Validation using information from outside statistics (quantitative test data or qualitative validation based on subject matter).

The first step could be the most challenging for mathematically trained people. Good problems rarely fall from the sky. Finding them takes open-mindedness, interpersonal skills, and good luck, and solving them requires teamwork. Yet not all collaborations, as not all relationships, end well. We feel blessed to have assembled an excellent multidisciplinary team for the cloud problem.

Another example of successful multidisciplinary collaboration that we should mention is the probabilistic weather forecasting project at the University of Washington with a mixed team of statisticians (Adrian Raftery and Tilmann Gneiting) and meterologists (Susan Joslyn and Earl Hunt) (Ban, Andrew, Brown, and Changnon 2006). In particular, it is worth noting their use of cognitive science to decide on how to best display uncertainty information.

3. DRIVEN BY COMPUTATION: MACHINE LEARNING

Problems from fields of science were the driving force behind paradigm developments in Statistics. In 1922 Fisher published his foundational work (Fisher 1922). As a scientist himself working on problems from genetics and agriculture, Fisher identified three "types" of statistical investigation: (1) problems of specification; (2) problems of estimation, and (3) problems of distribution. Fisher then builds a mathematical framework based on the assumption that data are random samples from an underlying population. Consequently, he relies on the available mathematical tools, probability theory and calculus, to define the concepts of consistency and efficiency and prove results about them. Comparing Fisher's list with ours described earlier, it is easy to see that the differences come from the impact of computer technology broadly and the consideration of computation narrowly.

Machine learning is at the frontier of statistics because of its serious utilization of computation in statistical inference. In a most interesting and thought-provoking work, Breiman (2001) called it algorithmic modeling and argued that we have to aim at solving real-data problems and consider more diverse tools often driven by computation than those dependent on data models.

Retrospectively, we might view the development of computation in statistics in three phases. The first phase was precomputer, where we depended on closed-form solutions. The second phase used computers, but not in an integrated manner; we would design a statistical method and then worry about how to compute it later. Frequently calling a numerical optimization routine was the solution, and we relied on the routine to determine how numerical convergence would be achieved; that is, convergence parameters were tuned for numerical reasons, and the optimization routine was used as a black box by statisticians. The third phase is the IT phase, where the data volume is so gigantic that procedures designed without computational considerations might not be implementable. This is also the cyberinfrastructure phase. Machine learning methods and Markov chain Monte Carlo (MCMC) algorithms are examples of approaches that intrinsically integrate computation.

3.1 The Loss Function Approach

Two machine learning methodologies stand out: boosting (Freund and Schapire 1997; Hastie, Tibshirani, and Freeman 2001) and support vector machines (SVM) (Scholkopf and Smola 2002). These both have impressive empirical performance on data sets that could have very high dimensions in terms of sample size and/or the number of predictors. Recently much theoretical understanding also has been obtained.

The current view of boosting is that it fits an additive model through gradient descent (or its variant) to minimize an objective or loss function. It is stopped early by monitoring the generalization or prediction error of the fitted model either estimated by cross-validation or assessed over a proper test set. That is, minimization of the loss function is a "pretense"; we are really interested in the solutions along the way to the minimum, not the minimum itself, and are prepared to stop early. In this way, the numerical convergence is not important at all, but the prediction performance is. SVM is based on a penalized optimization of a hinge loss function and computation is also the main focus through the "kernel trick." An implicit reproducing Hilbert space is induced by a kernel function, and a linear model is fitted in this space. However, all of the computation is done conveniently through the kernel function.

The machine learning approach based on minimizing a loss function can be viewed as a natural extension of the maximum likelihood approach where the loss function is the negated log-likelihood function. The penalized version is an extension of the maximum a posteriori (MAP) approach in Bayesian inference. What is new is the liberation from the negated log-likelihood function to a general loss function, in a way reminiscent of M-estimation. The motivation is not the same. In M-estimation, the goal is to obtain robust estimators in a parametric setting. In loss function machine learning, the goal is to have computationally feasible loss functions (often convex) to optimize over large data sets. Because robustness also can be desirable in the machine learning context, we now see some of the convex Huber functions being integrated into the machine learning literature.

This loss-function machine learning approach has been very successful in building up models for prediction. The measure of uncertainty has been based on perturbing the data in one way or another (permutation, resampling, and cross-validation). But a fundamental assumption to justify these perturbations to the original data is the iid assumption underlying most of the current machine learning methods.

3.2 Graphical Models

Graphical models represent another important development in statistics and machine learning. These models are widely used in the engineering and science communities. (See Lauritzen 1996 for a systematic treatment and Jordan 2004 for a recent review from the algorithmic stand point and graphical model applications.) Graphical models are effective in dealing with intricate dependencies and structures in the thousands or more variables present in today's large data sets. Examples include spatiotemporal modeling of temperature and precipitation in atmospheric science, image processing in the multiresolution framework of wavelets, gene network discovery based on gene expression and other modes of data, hidden Markov models in speech recognition, hierarchical models in information retrieval, and error-correction codes in communication. Obviously, models for dependent structures existed long before the formalism of graphical models that use the graph representation with variables as nodes so that general algorithms can be devised to compute marginal and conditional probabilities of interest.

One popular inference tool in graphical models is sampling algorithms, of which MCMC is the most prominent. If the Markov chain converges, then the MCMC method yields an estimate of the posterior distribution that provides an uncertainty measure. The design of an MCMC scheme to ensure a good mixing speed or convergence must be taken into account when laying out the distributions/conditional distributions for such a model. This is another example of third-generation computation. For many graphical models of high dimension, which are increasingly common, MCMC methods are more easily trapped in local modes of a posterior distribution, and convergence is difficult to guarantee. Adaptive sampling algorithms are more effective in these situations. (See Liu 2003 for MCMC and other sampling methods.)

On the other hand, the maximum likelihood principle turns a graphical model inference problem into an optimization problem. When the graphical model corresponds to a tree graph, we could use efficient message passing algorithms (or junction trees) for exact ML parameter estimation. This algorithm is very efficient when only local dependence exists in the graphical model and is very expensive otherwise (when a general graphical model gets embedded in a tree structure on an enlarged space). In general, the optimization function from the log-likelihood often is not convex, so that direct optimization is difficult. Searching for approximate loss functions that are more computationally feasible has been a focus of current research (cf. Yedidia, Freeman, and Weiss 2001; Wainwright and Jordan 2005). In particular, Wainwright (2006a) showed that for a specific graphical (mixture) model, computationally efficient algorithms also can have estimation advantages, as we have seen in the case of boosting and other methods discussed earlier.

Obtaining uncertainty measures through data perturbation is much harder than in the iid case, however. Parametric bootstrap seems to be a reasonable solution, but theoretical studies are needed to validate this approach, especially because optimizing an approximate loss function may lead to inconsistent estimates (cf. Wainwright 2006a).

3.3 Incorporating Auxiliary Information

For many massive data sets, the number p of predictors is much larger than the sample size n—the so-called "$p \gg n$" phenomenon. Furthermore, the cost of obtaining labeling or response information can be high (as in website classification). Therefore, using auxiliary information in the unlabeled data or the predictors is crucial for implicit regularization and increased well-posedness of our statistical inference problem. Semisupervised learning has emerged to incorporate classification information in the unlabeled data. Intuitively, if the predictor distribution is multimodal, then knowing the valleys of this distribution will help find the classification boundary if the classification boundary coincides with some of the valleys; otherwise, not. (For more information, see Chapelle, Scholkopf, and Zien 2006.)

In the loss function optimization approach of machine learning (Sec. 3.1.2), information on groupings of predictors has recently been built into the penalized loss function approach (see Yuan and Lin 2006; Kim, Kim, and Kim 2006; Zou and Hastie 2005). Zhao, Guilherme, and Yu (2006) proposed a general composite absolute penalty (CAP) framework to include the grouping structure and at the same time extend to the hierarchical structure among predictors. The CAP framework can facilitate group selection and enforce selection orders of predictors.

3.4 Sparsity and Interpretability

Interpretability of a statistical model is always desirable in any investigation, and it is indispensable for model building in science. One computationally efficient means of obtaining sparsity or interpretable models is through Lasso or the L_1-penalized least squares (Chen and Donoho 1994; Tibshirani 1996). Fast algorithms to produce the whole Lasso path are known (Osborne, Presnell, and Turlach 2000; Efron, Hastie, and Tibshirani 2004). Moreover, connections between Lasso and L2Boosting are observed and understood (Efron et al. 2004; Zhao and Yu 2004) and provide understanding into the sparsity property of the boosting estimates. Sparsity is also a well-known principle in low-level vision, as discussed by Wu, Li, Liu, and Zhu (2007). To capture sparsity of variables in a nonparametric setting, Rodeo (Laffterty and Wasserman 2005) combines boosting (or gradient descent) with kernel estimation to build sparse nonparametric models.

Because of its usefulness in practice, Lasso also has been the focus of much recent theoretical research in statistics (i.e., machine learning) and applied mathematics (Chen and Donoho 1994; Donoho 2004; Tropp 2004; Candes and Tao 2005; Meinshausen and Bühlmann 2006; Zhao and Yu 2006; Wainwright 2006b; Zou 2006; Greenshtein and Ritov 2004; Vander Geer 2006; Meinshausen and Yu 2006; Zhang and Huang 2006). Lasso is attractive because the L_1 penalty has a dual role: It simultaneously regularizes prediction and selects variables. What emerges from these studies is an incoherence or irrepresentable condition required for Lasso to select the correct variables if they exist and are sparse. This condition asks for the "irrelevant" variables to not be too correlated with the relevant or correct variables in the sparse model. These results also hold for the case of $p \gg n$, which has emerged as a valuable asymptotic setup for deriving analytical results relevant to high-dimensional data. Our recent work (Meinshausen and Yu 2006) indicates that when the irrepresentable condition is violated, Lasso still behaves sensibly in the sense that the Lasso estimates keep the order of the original coefficients with high probability, and the number of nonzero Lasso estimates cannot be too much larger than the nonzero "true" coefficients. This results hold in the $p \gg n$ case, and the sparsity assumption for the true model can be l_q for some $q \in [0, 1]$, suggesting the robustness of sparse model estimation through Lasso relative to the departure from the l_0 assumption that has been imposed conventionally in previous work.

All of the four aforementioned areas require some necessary eigenanalysis or convex optimization. It is of great interest to investigate parallel computation to mitigate the high demand on computation for large n or large p cases. However, as argued in Section 5.5, an "imprecise" parallel computation of eigenanalysis and convex optimization might be sufficient and actually may improve the statistical accuracy of the fitted models in terms of such tasks as parameter estimation and prediction.

4. NEW OPPORTUNITIES

Advances in cyberinfrastructure have helped energize some scientific disciplines as well as form new ones. In this section we describe two such fields: text processing (in detail) and sensor networks (briefly).

4.1 Text Processing

Texts are the data for information retrieval (e.g., web search), information extraction (e.g., title extraction from documents), natural language processing (e.g., machine translation), and question answering (e.g., "what is the distance between Berkeley and San Francisco?"). Research in text processing is being done mostly outside the traditional statistics community in such areas as signal processing, machine learning, and artificial intelligence. However, statisticians are getting involved, as illustrated by Genkin, Lewis, and Madigan (2007) in this special issue.

4.1.1 Information Retrieval. Information retrieval (IR) is the science and practice of indexing and searching data, especially in text or other unstructured forms. A typical IR task is searching for an image with horses in an image database, or searching for a document with a specific name on a computer. The volume of data is daunting, and the data structure is not traditional for statisticians.

4.1.2 Web Search. We all rely on web search for seeking information on almost anything and everything. Searches for articles on a topic, documents with specific titles, show times and locations of a particular movie, mortgage rates, e-mail addresses and telephone numbers of colleagues are just a few examples of web searches. Web search is the hottest topic in IR, but its scale is gigantic and requires a huge amount of computation. First, the target of web search is moving; the contents of websites change within 1 week for 30–40% of the web (Fetterly, Manasse, Najork, and Wiener 2004). A crawler is the tool that a search engine uses to collect websites into its database to answer queries. Because the web content is interlinked, very clumpy, and very diverse, random sampling to crawl cannot be easily carried out. Thus the crawling results, or search results for given queries, might be biased. Worse yet, the content of a website can include more than just text; images and videos, as well as interactive content, are common. Web data are highly unstructured, and its processing or data reduction/feature extraction is very challenging. (See Henzinger 2003 and Henzinger, Motwani, and Silverstein 2003 for more details on data collection and algorithm issues related to web search.)

Based on the websites collected in a database by a crawler, when a query is entered, the relevant websites are found and ranked. This fuels a very active research area in machine learning, ranking function estimation. A ranking function in web search usually depends on weighting the content of websites and links among the sites, as in the PageRank algorithm used by Google (Brin and Page 1998). When a weighting scheme is open to the public, however, opportunities arise for the so-called "search engine optimizers" (SEOs) to mislead the search engine to irrelevant websites of an SEO's customers. Therefore, search engines have to outwit SEOs in their search and ranking strategies, while also dealing with the fast-changing and growing world of websites.

4.1.3 Information Extraction. Information extraction (IE) attempts to do more than IR. Its goal is to extract useful facts for users in electronic documents (in most natural languages); that is, it aims to use text as a demonstration of understanding. IE had already existed early in natural language processing, but its potential is boosted dramatically by the IT revolution. For example, Lloyds of London, a shipping company, has been performing an IE task with human analysts for hundreds of years. The company wants to know all ship-sinking incidents around the world and put the information in a database. The IE question in the IT age is whether we can replace human analysts by computer algorithms automated to collect this information from data such as newspapers, web broadcasts, and government documents, possibly in different languages.

4.1.4 Question Answering. Question answering (QA) takes open-domain questions and searches over a collection of documents to find concise answers to the questions. It takes IR and IE further to deal with natural language sentence queries and return answers that need to be precise. Current QA systems can answer simple questions about facts like the one about the distance between San Francisco and Berkeley, but have difficulty answering complex questions requiring reasoning or analysis, such as determining the differences between a private university and a public university.

The data collection issue is present in information retrieval (web search), information extraction, and question answering, all of which rely on various databases. The most interesting is web search, in which web crawling is data collection. Web crawling is a difficult but exciting area for statisticians. Because of the dynamic nature of the web, new research might be called for in experimental design on how and when to select websites to crawl and what content and how much to take to store. Text data sets are often huge; for example, the sample size was 36 million and the dimension of features was 860,000 in the study of Gao, Suzuki, and Yu (2006). Transmitting and storing such data nontrivial. Often sending media such as DVDs through postal mail is the best route. Visualization also can be very useful, but cannot be readily done at this time.

Most of the time texts need to be represented by numeric forms before further actions. Programming skills are needed to process these text data, and statistical thinking in natural language processing is needed to keep the key information in the numeric form (or forming the feature vector) for downstream comparisons between data units. In addition, statistical modeling is often used to relate the feature vector to the goal of the task through a loss function formulation (Collins 2000).

Streaming data and data fusion are also issues of great interest. Collecting data online from a particular user for inputting Chinese or Japanese characters by typing in phonetic strings (Gao, Suzuki, and Wen 2002) would require streaming data algorithms. Fusing information across languages would be useful for Lloyds of London. For a web search, certainly different modes of data (e.g., text, image, audio) at a webpage should be integrated to form the feature vector for the webpage. Search engines like Google use these multimode data (e.g., *local.google.com*).

Some resources to help readers follow up on these topics beyond this article are as follows. For information retrieval, information extraction, and natural language processing, see Manning and Schütze (1999), Jurafsky and Martin (2000), and Manning, Raghavan, and Schütze (2007). For information retrieval and question answering, read TREC publications at *http://trec.nist.gov/pubs.html*. For current developments in these areas, see websites of the following conferences: StatNLP: Association for Computational Linguistics (ACL), North American ACL (NAACL), Empirical methods for NLP (EMNLP), European ACL (EACL), ICASSP, ICSLP, SIGIR, and WWW, and *http://trec.nist.gov/pubs.html*.

4.2 Sensor Networks

Sensor networks are self-networked small devices engineered to collaborate with one another and to collect information concerning the environment around them. Their flexibility greatly extends our ability to monitor and control the physical environment from remote locations. Their applications range from seismic, natural environmental monitoring to industrial quality control and to military uses. The primary functions of the sensors are data collection, communication (data transmission), and, to a much lesser extent, computation. But the sensors are constrained by the battery power. So far, sensor network research has been dominated by researchers from computer science and electrical engineering, but nonetheless it provides an ideal platform from which to integrate statistical analysis with computation, data compression, and transmission because the overriding power constraint forces us to consider all of the players in the same framework to maximize the utility of the limited battery energy. It would be interesting to try to devise a framework that encompasses components 2 (transmission and compression) and 4 (formal modeling) to answer optimality questions. Distributed algorithms are also very desirable because data transmission among the sensors is expensive. For up-to-date information on sensor networks, interested readers can follow research presented each year at the Information Processing in Sensor Networks (IPSN) series website: (*http://www.cse.wustl.edu/lu/ipsn07.html*).

5. LOOKING AHEAD

Atmospheric science, text processing, and sensor networks provide windows into the spectrum of opportunities offered by the IT revolution. In this section we examine statistical issues encountered in both new and old fields flooded with data and that have yet to be addressed adequately by the statistics community. We begin with a brief overview on the distributed trend of computing, and then offer some personal (possibly ignorant) views on how massive data and this trend might inspire new core research directions in statistics.

5.1 Parallelism or Distributed (Grid) Computing

As evident from the NSF report (Atkins et al. 2003), parallelism is now being used to increase the power of computation and storage capacity. Within a computer, multiple chips form parallel processing units, parallel or clustered high-speed computers are connected to further increase the computing power, and for storage, because disk prices are falling, many disks are used to host databases of a few terabytes. Wired and wireless networks are used for distributed computation. The costs of displays are dropping, and useful three-dimensional interaction on possibly many displays at the same time is becoming feasible. Whereas parallel computing is still useful in the midst these developments, the exciting new directions are virtual computing and grid computing. Relative to these modes, parallel computing is rigid because it specifies how iterative algorithms are to be cut up and requires a central supervisor. Grid computing attempts to break this bond and allow clusters to determine their own computation during free-cycle time. Virtual computing further abstracts this environment so that different programs and different operating systems can handle slices of data. In all of these modes, the greatest difficulty is that it takes more time to cut up the data than to run the actual computations. True distributed (peer-to-peer) algorithms can process data locally and merge their results pairwise. (These descriptions of computing modes are basically taken from an e-mail exchange with Dr. Leland Wilkinson commenting on an earlier version of this manuscript.)

Matlab is equipped with a distributed computing engine and toolbox, which “enable(s) you to develop distributed and parallel MATLAB applications and execute them on a cluster of computers without leaving your technical computing development environment” (from *http://www.mathworks.com/products/distribtb/*). Meanwhile, Apple has developed a grid computing toolkit installed on all iMacs, called XGrid, that “turns a group of Macs into a supercomputer, so they can work on problems greater than each individually could solve” (*http://www.apple.com/macosx/features/xgrid/*). Sun also provides grid solutions (*http://www.sun.com/software/grid/*). On a much larger scale, BOINC (Berkeley Open Infrastructure for Network Computing) provides “open-source software for volunteer computing and desktop grid computing” to solve problems in earth sciences, biology and medicine, mathematics, and physics (*http://boinc.berkeley.edu/*). However, as alluded to earlier, “carving up the data can take longer than actually processing it on these (grid) systems. That’s why Google uses a distributed model in which the data never get consolidated in one place. That way, computations are local, and the merging is done in a distributed fashion. Once data are merged into a single database, parallel architecture is not going to help much with speeding things up, because most of the time is spent accessing the data” (excerpt from an e-mail exchange with Dr. Wilkinson).

These distributed computing developments are examples of the parallelism used by the computer science community to mitigate the limitations of the current computer technology. They offer the users plug-ins to increase computing power. A first step is to use existing parallel environments (e.g., Matlab toolbox and XGrid) in our statistical investigation of data. More fundamentally, however, we need to devise statistical methods, exploratory or formal, that are well suited for distributed or grid computing environments.

With these computing trends in the background, we are ready to focus on statistical issues at the cutting edge of statistical research.

5.2 Data Collection and Management

The “Towards 2020 Science” report states that “the way scientists interact with data and with one another is undergoing a fundamental paradigm shift.” The paradigm is shifting from the traditional experiment → analysis → publication to the new experiment → data organization → analysis → publication. Data transmission by moving the data to standard packages (e.g., R or Matlab) does not deal with the new stage of data organization. In this section we broadly interpret this new stage of science research and discuss the implied statistical issues in data collection and management.

5.2.1 Online or Streaming Data. Online or streaming data analysis stems from the second characteristic of the IT revolution: high data rate. High-dimensionality has attracted much attention within the statistics community, as is evident from the numerous conferences and workshops with this phrase in the title. Nonetheless streaming data analysis remains mostly outside of the research spotlight in statistics, with the exception of the special 2003 issue of *Journal of Computational and Graphical Statistics* from which many references are cited in this article.

Due to the real-time requirement of streaming data analysis and the huge volumes of data coming in, the desired speed for extracting information from data is much higher than that in the batch or offline mode. However, designing a fast online algorithm may require batch data analysis to identify which features of the data to retain. On the other hand, before batch data can be collected, data reduction or feature selection must be carried out online to reduce the data volume for storage; for example, down-sampling or aggregation may be necessary for batch mode analysis.

As in the article of Chambers, Lambert, and Vander Weil (2006), many online or streaming data algorithms repeatedly update a low-dimensional feature distribution and identify outliers or anomalies relative to this distribution. Other streaming algorithms deal with records (e.g., of phone calls) and have a distinctly discrete flavor. (For more discussion of issues related to streaming data, see Gilbert and Strauss 2007 in this issue.)

5.2.2 Data Fusion. For complex problems (e.g., from genomics and atmospheric science), multiple data sources must be used, that is, data fusion is called for. In the arctic cloud detection project, we implemented a very simple form of data fusion from two sensors on the same satellite Terra; a consensus label was given only when our MISR-sensor based algorithm gave the same label as the MODIS-sensor operational algorithm. This is a fusion at the decision level. We have also fused the two sensor data at the feature level by applying quadratic discriminant analysis to the three MISR features and the five MODIS features for the final MISR–MODIS soft labeling. Dass and Jain (2007) address fusion of fingerprint with other biometric traits and mention three fusion levels: feature level, matching score level and decision level. It is clear that other levels of fusion might be considered depending on the problem. Data fusion is also necessary in many other fields, such as genomics research (fusion of gene expression and sequence data) and climate modeling (fusion of simulation and observation data). These tasks are often related to large research or government projects and involve huge amounts of data, online or streaming data, etc. It is therefore a research frontier for us to get involved in and make significant contributions to by, for instance, providing a framework to think about data fusion relative to communication and computation constraints.

5.2.3 Interacting With Database. Reducing the data size is an important prerequisite to dealing with massive data sets, but we run the obvious risk of losing important information. The goal in a digital sky survey is to find very sparse signals (e.g., quasars). Reducing the size by random sampling is likely to not include these targeted signals altogether. This dictates the need for statistical methods to interact with databases.

The recent story of Google's machine translation success (Norvig 2006) confirms this point. Google's translation system is based on analyzing an incredibly large database of documents and their human translations to carry out its own translation. Databases of such a size can only be housed by a few specialized companies, such as Google and Microsoft. It is commonly believed that results from a reduced-sized database housed by a research unit in a university would not come close to the Google results. It is also noteworthy that memory constraints prompted the Google group to use binning (e.g., regularization) on the character strings to achieve better translation accuracy. This is an example of how communication constraints and considerations led to better statistical accuracy.

The interaction with data-bases goes both ways. The easier direction is to understand database structures and design methods with fast implementation on the databases—the purpose of data mining. The recent advances of research in machine learning and statistics make it high time to integrate these areas into data mining software with the necessary modifications to suit the database. The harder direction is to influence database design so that data in a database can be accessed swiftly by a slew of statistical algorithms. To achieve this, statisticians or data analysts need to reach consensus on the basic operations that we need to conduct on a database for most if not all statistical analysis algorithms. The size of the database might prohibit the use of any algorithm that is worse than linearly scalable to the size of the data. This points to the same question of what is the most efficient statistical method subject to a computational constraint, which we discuss later.

To help access databases more easily in general, the data science group at Keio University (*http://www.stat.math.keio.ac.jp/index.html*) has created a front-end data management system, DandD, to acquire data directly from the web or a database. Its module DandDR interacts with R directly.

5.3 Expanding Exploratory Data Analysis for Massive Data

Exploratory data analysis (EDA) is an integral part of every statisticians' toolkit. Tukey's book on EDA in the 1970s (Tukey 1970), and especially his famous article in The *Annals of Statistics* in 1962 (Tukey 1962) did much to give the approach intellectual credibility. The basic tools of EDA are summary statistics and simple visual displays, such as histograms, boxplots, scatterplots, and time series plots. For high-dimensional data such as those in computer networks (Denby et al. 2007), two or three-dimensional visualization abilities become quite limited for understanding complex structures. Efforts have been made by the statistics community to accommodate higher dimensions through parallel coordinate plots (Inselberg 1999; Wegman 1990) and selective projections of data as in GGobi (*www.ggobi.org*) for continuous data and through mosaic plots (Hofmann 2000) for categorical data. Recent work has been done on an enhancement of parallel coordinate plot called textile plot (Kumasaka and Shibata 2007). An alternative and attractive approach is to use interactive graphics, linking low-dimensional plots to gain higher-dimensional insights. Paul Velleman's commercial package Data Desk (*www.datadesk.com*) already had consistent and powerful interactive graphics in the early 1990s. A number of research software packages have taken this approach further, particularly

Mondrian (*stats.math.uni-augsburg.de/Mondrian/*), which includes parallel coordinate plots and mosaic plots, and GGobi. (See Unwin, Theus, and Hofmann 2006 for more on visualizing large datasets up to one million.) (This paragraph is more or less taken from an e-mail exchange with Antony Unwin and improves on the original paragraph in an earlier version.)

For other communities (e.g., machine learning and signal processing) dealing with similar high-dimensional data, EDA is not yet part of their education, so few use it before formal algorithmic analysis or modeling. For such areas as information retrieval and information extraction, the original form of data is often text, which is not always well formulated or structured. It is of great value to visualize text data using traditional and recent EDA tools, and hopefully also with new tools that could incorporate fast modeling algorithms.

5.3.1 Seeing Helped by Enhanced Computing Power. Data volume and complexity are one side of the computer technology. The other side is the increased parallel computing power to visualize data (multimedia representation) and fit sophisticated models. The field of data visualization is advancing rapidly, mostly outside of statistics. Because visual processing units take more than one-third of the cortex in human's brain, it is necessary to use the superb information gathering ability of our vision. The efficient use of our vision is even more desirable for the complex data that we face today. It relies on an understanding of our vision from neuroscience and computer vision to render images from data, using spatial locations, perspectives, color, and ray tracing (e.g., shading). It goes without saying that we should use the existing EDA visualization tools as much as possible, while at the same time exploring new possibilities offered by the visual arts or multimedia community.

Distributed computing is being used for data visualization through clusters of graphical processing units and CPUs. For example, Levit (2006) brought out in real-time different aspects of the data coming in from a digital sky survey by parallelizing graphical processing units and hundreds of displays. The visualization group at Lawrence Berkeley Laboratory (*http://vis.lbl.gov*) also conducts research on parallel graphics and visualization. Moreover, a government agency headed by Dr. Jim Thomas, the National Visualization and Analytics Center (NVAC), chartered by the U.S. Department of Homeland Security, has set its objective in 2004 "to define a five-year research and development agenda for visual analytics to address the most pressing needs in R&D to facilitate advanced analytical insight" to help "counter future terrorist attacks in the U.S. and around the globe" (*http://nvac.pnl.gov/agenda.stm*).

It is natural to ask whether we can bring out more quantitative information in our data with these new visualization tools. Progress has been made in the field of scientific visualization in this direction (e.g., Ben Fry's website *http://acg.media.mit.edu/people/fry/*). (See Johnson 2004 for an overview of the most pressing problems in scientific visualization.) In particular, Wilkinson (2005) has developed the grammar of graphics to draw graphic displays for statistics. We can add more to this enterprise, I believe, if we collaborate with researchers in visualization to represent results from modern methods, such as machine learning and MCMC, for model revision and validation as we did with residual plots in simple linear regression.

5.3.2 Simplifying Data to See Through Modeling. Complementary to extending our abilities through computer graphics and visualization to see more in data, we can use computing power to simplify the data to visualize. Sophisticated modeling methods first can be used on data so that patterns can present themselves in output plots (after, e.g., some boosting or SVM fits) which are not possible in displays of the original data. Residual is an obvious output to investigate, as in classical statistics, but additional visualizations of other outputs (margin plots from SVM and fitting error plots along a boosting path reality come to mind) should be added to the routine diagnostics of a model. We also can search for meaningful low-dimensional structures in high-dimensional data. If we find these structures, then the high-dimensional data can be reduced to low dimensions for visualization. Subject knowledge often suggest such dimensionality reduction or models in low dimensions, as seen in the cloud project and in the reports by Faraway and Reed (2007) and Buvaneswari et al. (2007) in this special issue. When subject knowledge is not adequate, automatic dimensionality-reduction methods can be tried to suggest possible meaningful data reduction. These methods aid the search for these structures at a speed impossible before. Data visualization and model fitting should be conducted iteratively, however. Seeing suggests models to fit, and model fits give data to see. This is similar to what we do in residual analysis for regression models, but residual plots are replaced with multimedia data representation and regression models are replaced by more general methods. Admittedly, this is easier said than done, however.

Recent years have brought much activity in automatic data reduction, including Kernel PCA (Scholkopf, Smola, and Müller 1998), ISOMAP (Tenenbaum, de Silva, and Langford 2000), LLE (Roweis and Saul 2000) and its extension using the Hessian matrix (Donoho and Grimes 2003) and spectral clustering (Shi and Malik 2000; Ng, Jordan, and Weiss 2001; Belkin and Niyogi 2003; Zhou, Bousquet, Weston, Scholkopf, and Zien 2004). These methods generalize the traditional PCA and MDS because an eigenanalysis, either local or global, underlies all of them. (See Ham, Lee, Mika, and Scholkopf 2003 for a nice theoretical synthesis of different dimensionality-reduction techniques from a kernel standpoint.) Before these dimensionality-reduction methods become routine EDA analysis, much more experience in applying them to real data sets and theoretical analysis is needed to understand the pros and cons of each method both in absolute terms and relative to each other. This brings us naturally to the topic of the next section on expanding our analytical knowledge.

5.4 Use of Nontraditional Mathematical Tools

Like it or not, we are leaving the comfort of the classical paradigm founded mathematically on calculus and a large sample size relative to the dimension of the parameters. The pervasive existence of the $p \gg n$ in massive data sets suggests that even though we still need asymptotics to see regularity, the asymptotics should not be done with a fixed p. Hence the dimension of the parameter space is growing with increasing sample size. Most of our intuition is derived from the three-dimensional physical space that we live in. To paraphrase Aldous (1989), because we are at a point with not much intuition to go on, analytical derivations might help lead us toward the light. It is encouraging to see that theoretical results for the $p \gg n$ case are appearing in random matrices, linear modeling, Lasso under deterministic and stochastic assumptions, boosting, and covariance estimation. Much insight should be gained through such analysis. Becoming part of the new cyberinfrastructure demands that the analytical results take into account as much as possible the algorithmic implementation of the methods, instead of assuming that the entities to be analyzed are the exact maxima of objective functions which in practice cannot be obtained due to computational reasons. (For comprehensive tutorials and new research material, see SAMSI's recent workshop on high-dimensional inference and random matrices at *http://www.samsi.info/programs/2006ranmatprogram.shtml*.) It remains to be seen whether the random matrix results will become equivalent to the central limit theorem in classical statistics. In any event, much distilling is needed to simplify the methods used to derive these results before they enter the analysis toolkit of routine statistical investigations of massive data to provide insight.

5.5 Computation for Data With Uncertainty or Noise

Computation was not a concern of Fisher (1922) but is central to a statistical investigation today. There is something very novel about boosting (and fitting neural networks); the computation parameter, the number of iterations, also serves as a regularization parameter in statistical estimation. The BLasso algorithm by Zhao and Yu (2004) has a similar property. This is a componentwise gradient descent algorithm with a fixed step to minimize the generalized Lasso loss (convex loss and penalty functions) simultaneously for different values of λ's. It shares many similarities with boosting when a componentwise gradient descent algorithm, or the forward stagewise regression (FSR) (Efron et al. 2004), is used. That is, BLasso has a forward step just as in FSR, but with a backward step added to make sure the combined penalty is minimized, not just the loss function part, which is the aim of boosting. Moreover, BLasso solves a sequence of optimization problems corresponding to different λ's similar to the barrier method in optimization (Boyd and Vandenberghe 2004).

The coupling of computation and regularization in boosting and BLasso is reminiscent of the equivalence of computation and modeling in K-complexity theory. Relative to a universal turing machine, the K-complexity of a binary string is defined as the length of the shortest program that prints out the string and stops. Because of an equivalence of a (prefix) program and a probability distribution, there is an equivalence of computation (program) and modeling represented by the distribution. Despite the fact that K-complexity is not computable, this equivalence has an intriguing intellectual appeal.

Let us entertain ourselves further by looking into modeling and computation practiced today. We know that statistical model fitting uses scientific computing, but statistical computation is special. Even in the parametric case there is a well-known result that only one Newton or second-order step is needed to make a $\sqrt{n}$-consistent estimator efficient. That is, because our objective function is a random quantity, we do not need convergence of the minimization algorithm to get a statistically satisfying solution, as shown in boosting. In nonparametric methods such as boosting, neural nets, and BLasso, early stopping before convergence saves computation and regularizes the fitting procedure, and hence results in a better statistical model. Again, computation and model fitting seem to be working in the same direction: less computation and better statistical accuracy. These facts indicate the intimate relationship between computation and model fitting. They prompt us to ask the following question: Is there a minimal amount of computation needed for a certain statistical accuracy?

It is not clear whether or not this question can be answered, because fast algorithms in scientific computation often rely on closed-form equations or relationships derived through analytical means. Analytical calculations have infinite precision, whereas scientific computations are of finite precision. Nevertheless, we believe that this is a very interesting intellectual question, and the pursuit of the answer could lead to useful practical consequences for modeling IT data.

6. CONCLUSION

The main difference and advantage (and/or disadvantage) of our time from Fisher's time is the availability of computing technology and consequently the availability of massive amounts of data. We argue that for the healthy existence of our field, solving real data problems has to be our aim, and we need to find a way to join the ongoing cyberinfrastructure development.

Many exciting challenges must be met to achieve the goal of solving real problems. In particular, we need to take advantage of the distributed/grid computing trend, interact efficiently with databases and other data sources such as sensor networks, design EDA visualization tools, use new or nonconventional mathematical results, develop new statistical algorithms satisfying communication and computation constraints, and devise new statistical inference paradigms to encompass such endeavors.

At an organizational or cultural level, the statistics community also faces many challenges. As described herein, there are many large scientific projects in which huge amounts of data are collected and managed, for example, in atmospheric science (e.g., model simulation and remote sensing data), astronomy (e.g., digital sky surveys), and biology (e.g., genome or brain databases across different species). Individual statisticians can find and are finding collaborative roles in these big projects, but it is very difficult for individual statisticians to influence the fundamentals of these projects, for instance, to have a say in data collection and in the choice of algorithms to use in mining the huge databases. Collective thinking and leadership from our community are needed if our discipline is to have the necessary impact in the IT age.

In addition to statistical skills, social and interpersonal skills are needed to successfully collaborate with scientists and persuade them of the key role of statistics in scientific investigations. The importance of these nontechnical skills in interdisciplinary research suggests the need for a culture change in our community and for these nontraditional skills to be valued and recognized in, for example, tenure reviews, promotions, and awards. Last but not least, we need to educate our graduate and undergraduate students with the relevant technical and interpersonal skills. Insightful comments and concrete suggestions on the education front have been given by Madigan and Stuetzle (2004), a discussion on the NSF Future Statistics Workshop Report.

This is a time of data deluge; we can help build the ark and ride on it, if we so choose.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants DMS-03036508 and DMS-0605165, ARO grant W911NF-05-1-0104, NSFC 60628102, and a Guggenheim Fellowship. As incomplete as this work might seem, it would not have been possible without the support and encouragement of many colleagues. The author is very grateful for helpful discussions and comments from Leland Wilkinson, David Purdy, Tilmann Gneiting, Martin Wainwright, John Rice, Mike Jordan, Nicolai Meinshausen, Peter Bickel, Ritei Shibata, Antony Unwin, Dan Klein, Noureddine El Karoui, and Hang Li. She is also very grateful to two referees for their constructive criticism on the first version of the manuscript which led to the manuscript being essentially rewritten. Finally, she thanks the guest editor, Vijay Nair, for his encouragement, patience, on-target comments, and invaluable help with the presentation.

[Received February 2006. Revised January 2007.]

REFERENCES

Aldous, D. (1989), *Probability Approximations via the Poisson Clumping Heuristics*, New York: Springer-Verlag.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messina, P., Ostriker, J. P., and Wright, M. H. (2003), "Revolutionizing Science and Engineering Through Cyberinfrastructure," Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.

Ban, R. J., Andrew, J. T., Brown, B. G., and Changnon, D. (2006), "Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts," in *National Research Council Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts*, Washington, DC: National Academies Press, p. 112.

Belkin, M., and Niyogi, P. (2003), "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Nueral Computation*, 15, 1373–1396.

Berk, R. A., Bickel, P., Campbell, K., Fovelli, R., Keller-McNutty, S., Kelly, E., Linn, R., Park, B., Perelson, A., Rouphail, N., Sacks, J., and Scheonberg, F. (2002), "Workshop on Statistical Approaches for the Evaluation of Complex Computer Models," *Statistical Science*, 17, 173–192.

Boyd, S. P., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge, U.K.: Cambridge University Press.

Braverman, A., Dobinson, E., Graves, S. J., Burl, M. C., Bastano, B., Hinke, T., Lynnes, C. S., Minster, B., Ramachandran, R., Behnke, J., Carver, L., Garay, M., Granger, S., Hardin, D., and Wilson, B. (2006), "Final Report of the 2nd NASA Data Mining Workshop: Issues and Applications in Earth Science," available at *http://datamining.itsc.uah.edu/meeting06*.

Breiman, L. (2001), "Statistical Modeling: Two Cultures," *Statistical Science*, 16, 199–231.

Brin, S., and Page, L. (1998), "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proceedings of the 7th International Conference on World Wide Web/Computer Networks*, 30, 107–117.

Buvaneswari, A., Graybeal, J. M., James, D. A., Lambert, D., Liu, C., and MacDonald, W. M. (2007), "A Statistical View of a Wireless Call," *Technometrics*, 49, 305–317.

Candes, E. J., and Tao, T. (2005), "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n," available at *http://www.acm.caltech.edu/emmanuel/publications.html*.

Chambers, J. M., Lambert, D., and Vander Weil, S. (2006), "Monitoring Networked Applications With Incremental Quantile Estimation," *Statistical Science*, 21, 463–475.

Chapelle, O., Scholkopf, B., and Zien, A. (2006), *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*, Cambridge, MA: MIT Press.

Chen, S., and Donoho, D. (1994), "Basis Pursuit," technical report, Stanford University, Dept. of Statistics.

Collins, M. (2000), "Discriminative Reranking for Natural Language Parsing," in *ICML 2000*.

Dass, S. C., and Jain, A. K. (2007), "Fingerprint-Based Recognition," *Technometrics*, 49, 262–276.

Denby, L., Landwehr, J. M., Mallows, C. L., Meloche, J., Tuck, J., Xi, B., Michailidis, G., and Nair, V. N. (2007), "Statistical Aspects of the Analysis of Data Networks," *Technometrics*, 49, 318–334.

Donoho, D. (2004), "For Most Large Underdetermined Systems of Linear Equations, the Minimal l1-Norm Near-Solution Approximates the Sparsest Near-Solution," technical report, Stanford University, Dept. of Statistics.

Donoho, D., and Grimes, C. (2003), "Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data," *Proceedings of the National Academy of Sciences*, 100, 5591–5596.

Efron, B., Hastie, T., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Emmott, S., Rison, S., Abiteboul, S., Bishop, C., Blakeley, J., Brun, R., Brunak, S., Buneman, P., Cardelli, L., Cox, S., Emmerich, W., Ferguson, N., Finkelstein, A. M. F., Hannay, T., Herbert, A., Kuppermann, A., Landshoff, P., Moin, P., Muggleton, S., Parker, A., Peitsch, M., Radman, M., Sato, T., Searls, D., Shapiro, E., Shiers, J., Soberon, J., Syme, D., Szalay, A., Szyperski, C., Watkins, D., Young, M., Zauner, K., Beckman, B., Hagehulsmann, A., Harel, D., Lyutsarev, V., Martin, B., Phillips, A., and Wallace, R. (2005), "Towards 2020 Science," The Towards 2020 Science Committee, under the aegis of Microsoft Research Cambridge.

Faraway, J., and Reed, M. P. (2007), "Statistics for Digital Human Motion Modeling in Ergonomics," *Technometrics*, 49, 277–290.

Fetterly, D., Manasse, M., Najork, M., and Wiener, J. L. (2004), "A Large-Scale Study of the Evolution of Web Pages," *Software Practice and Experience*, 1, 1–27.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society London*, Ser. A, 222, 309–368.

Freund, Y., and Schapire, R. (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139.

Gao, J., Suzuki, H., and Wen, Y. (2002), "Exploiting Headword Dependency and Predictive Clustering for Language Modeling," in *EMNLP*.

Gao, J., Suzuki, H., and Yu, B. (2006), "Approximation Lasso Methods for Language Modeling," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 225–232.

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304.

Gilbert, A. C., and Strauss, M. J. (2007), "Group Testing in Statistical Signal Recovery," *Technometrics*, 49, 346–356.

Greenshtein, E., and Ritov, Y. (2004), "Persistence in High-Dimensional Predictor Selection and the Virtue of Overparametrization," *Bernoulli*, 10, 971–988.

Ham, J., Lee, D. D., Mika, S., and Scholkopf, B. (2003), "A Kernel View of the Dimensionality Reduction of the Manifolds," Technical Report TR-110, Max-Plank Institute for Biological Cybernetics.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer-Verlag.

Henzinger, M. R. (2003), "Algorithmic Challenges in Web Search Engines," *Internet Mathematics*, 1, 115–123.

Henzinger, M. R., Motwani, R., and Silverstein, C. (2003), "Challenges in Web Search Engines," in *The ACM 18th International Joint Conference on Artificial Intelligence*, pp. 1573–1579.

Hofmann, H. (2000), "Exploring Categorical Data: Interactive Mosaic Plots," *Metrika*, 51, 11–26.

Inselberg, A. (1999), "Don't Panic... Do It in Parallel," *Computional Statistics*, 14, 53–77.

Jacob, J. C., and Husman, L. E. (2001), "Large-Scale Visualization of Digital Sky Surveys," in *Virtual Observations of the Future: Astronomical Society of the Pacific Conference Series*, p. 225.

Johnson, C. (2004), "Visualization Viewpoints," *IEEE Transactions on Computer Graphics and Applications*, 24, 13–17.

Jordan, M. I. (2004), "Graphical Models," *Statistical Science*, 19, 140–155.

Jurafsky, D., and Martin, J. H. (2000), *Speech and Language Processing*, Englewood Cliffs, NJ: Prentice-Hall.

Kim, Y., Kim, J., and Kim, Y. (2006), "Blockwise Sparse Regression," *Statistica Sinica*, 16, 375–390.

Knuteson, B., and Padley, P. (2003), "Statistical Challenges With Massive Datasets in Partical Physics," *Journal of Computational and Graphical Statistics*, 12, 808–828.

Kumasaka, N., and Shibata, R. (2007), "High-Dimensional Visualization: The Textile Plot," manuscript, Keio University, Dept. of Math., *www.stat.math.keio.ac.jp/kumasaka/papers/csda.pdf.*

Laffterty, J., and Wasserman, L. (2005), "Rodeo: Sparse Nonparametric Regression in High Dimensions," in *Advances in Neural Information Processing Systems*, p. 18.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, U.K.: Oxford University Press.

Levit, C. (2006), "Using Graphics Processing Unit (GPU) Hardware for Interactive Exploration of Large Multivariate Data," presentation at the invited session on statistics and information technology, Interface 2006, Pasadena, CA.

Lindsay, B. G., Kettenring, J., and Siegmund, D. O. (2004), "A Raport on the Future of Statistics," *Statistical Science*, 19, 387–413.

Liu, J. (2003), *Monte Carlo Strategies in Scientific Computing*, New York: Springer.

Madigan, D., and Stuetzle, W. (2004), "Comment on 'A Report on the Future of Statistics,' " by B. G. Lindsay, J. Kettenring, and D. O. Siegmund, *Statistical Science*, 19, 408.

Mallows, C. (2006), "Tukey's Paper After 40 Years," *Technometrics*, 48, 319–336.

Manning, C., Raghavan, P., and Schütze, H. (2007), *Introduction to Information Retrieval*, available at *http://www-csli.stanford.edu/schuetze/information-retrieval-book.html.*

Manning, C., and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462.

Meinshausen, N., and Yu, B. (2006), "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," technical report, University of California Berkeley, Dept. of Statistics.

Ng, A., Jordan, M., and Weiss, Y. (2001), "On Spectral Clustering Analysis and an Algorithm," *NIPS2001*, 14, 849–856.

Norvig, P. (2006), "Theorizing From Data: Avoiding the Capital Mistake," presented at the CITRIS Distinguished Speaker Series, University of California Berkeley, September 25, 2006.

Osborne, M., Presnell, B., and Turlach, B. A. (2000), "A New Approach to Variable Selection in Least Squares Problems," *Journal of Numerical Analysis*, 20, 389–403.

Roweis, S., and Saul, L. (2000), "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, 290, 2323–2326.

Scholkopf, B., and Smola, A. J. (2002), *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: MIT Press.

Scholkopf, B., Smola, A., and Müller, K. R. (1998), "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, 10, 1299–1319.

Shi, J., and Malik, J. (2000), "Normalized Cuts and Image Segmentation," *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22, 888–905.

Shi, T., Clothiaux, E. E., Yu, B., Braverman, A. J., and Groff, G. N. (2006a), "Detection of Daytime Arctic Clouds Using MISR and MODIS Data," *Remote Sensing of Environment*, 107, 172–184.

Shi, T., Yu, B., Clothiaux, E., and Braverman, A. (2006b), "Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data With Case Studies," technical report, University of California Berkeley, Dept. of Statistics.

Speed, T. P. (2005), "Terence's Stuff: Interdisiciplinary Research," *IMS Bulletin*, 34, 16.

Tenenbaum, J. B., de Silva, V. F., and Langford, J. C. (2000), "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, 290, 2319–2323.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Tropp, J. (2004), "Just Relax: Convex Programming Methods for Subset Selection and Sparse Approximation," ICES Report 04-04, University of Texas Austin.

Tukey, J. (1962), "The Future of Data Analysis," *The Annals of Mathematical Statistics*, 33, 1–67.

——— (1970), *Exploratory Data Analysis*, Reading, MA: Addison Wesley.

Unwin, A., Theus, M., and Hofmann, H. (2006), *Graphics of Large Datasets: Visualizing a Million*, New York: Springer.

Vander Geer, S. (2006), "High-Dimensional Generalized Models and the Lasso," Technical Report 133, Seminar fur Statistik, ETH Zurich.

Wainwright, M. J. (2006a), "Estimating the "Wrong" Graphical Model: Benefits in the Computation-Limited Setting," *Journal of Machine Learning Research*, 7, 1829–1859.

——— (2006b), "Sharp Thresholds for High-Dimensional and Noisy Recovery," Technical Report 709, University of California Berkeley, Dept. of Statistics.

Wainwright, M. J., and Jordan, M. I. (2005), "A Variational Principle for Graphical Models," in *New Directions in Statistical Signal Processing: From Systems to Brain*, MIT Press, Chap. 11.

Wegman, E. J. (1990), "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of the American Statistical Association*, 85, 664–675.

Welling, J., and Dearthick, M. (2001), "Visualization of Large Multi-Dimensional Datasets," in *Virtual Observations of the Future: Astronomical Society of the Pacific Conference Series*, p. 225.

Wilkinson, L. (2005), *The Grammar of Graphics*, New York: Springer.

Wu, Y. N., Li, J., Liu, Z., and Zhu, S. C. (2007), "Statistical Principles in Low-Level Vision," *Technometrics*, 49, 249–261.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001), "Generalized Belief Propagation," *NIPS2000*, 13, 689–695.

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Ser. B, 68, 49–67.

Zhang, C.-H., and Huang, J. (2006), "Model-Selection Consistency of the Lasso in High-Dimensional Linear Regression," technical report, Rutgers University, Dept. of Statistics.

Zhao, P., Guilherme, R., and Yu, B. (2006), "Grouped and Hierarchical Model Selection Through Composite Absolute Penalties," technical report, University of California Berkeley, Dept. of Statistics.

Zhao, P., and Yu, B. (2004), "Boosted Lasso," technical report, University of California Berkeley, Dept. of Statistics.

——— (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zhou, D., Bousquet, T. N., Weston, J., Scholkopf, B., and Zien, A. (2004), "Learning With Local and Global Consistency," *Advances in Neural Information Processing Systems*, 16, 321–328.

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Ser. B, 67, 301–320.