Statistical Science
2004, Vol. 19, No. 3, 387–413
DOI 10.1214/088342304000000404

A Report on the Future of Statistics

Bruce G. Lindsay, Jon Kettenring and David O. Siegmund

Abstract. In May 2002 a workshop was held at the National Science Foundation to discuss the future challenges and opportunities for the statistics community. After the workshop the scientific committee produced an extensive report that described the general consensus of the community. This article is an abridgment of the full report.

Key words and phrases: Research funding, National Science Foundation, challenges, opportunities, statistical education.

1. INTRODUCTION

Evidence is all about us for the current unique opportunities for statistics. Consider, for example, the three pillars of the Mathematical Sciences Priority Area of the National Science Foundation: handling massive data, modeling complex systems and dealing with uncertainty. All three are primary interests of the discipline of statistics. Never before has statistical knowledge been more important—nor as widely useful—to the scientific enterprise.

Massive amounts of data are collected nowadays in many scientific fields, but unless there are proper data collection plans, this will almost inevitably lead to massive amounts of useless data. Without scientifically justified methods and efficient tools for the collection, exploration and analysis of data sets, regardless of their size, we will fail to learn more about the often complex and poorly or partly understood processes that yield the data.

To master this enormous opportunity, the statistics community must address the many challenges that are arising. Some of these are intellectual challenges; others are infrastructural, arising from the changing tides of external forces. This document records an attempt by the statistics community to identify and address these challenges and forces. It is based on a workshop on the future of statistics that was held at the National Science Foundation (NSF) in May of 2002.

1.1 The Workshop

The workshop was held at the request of the Foundation and was organized by a scientific committee of nine members. That same committee prepared a final report with the guidance and assistance of the workshop participants and many others. The full report (83 pages) is available online at *http://www.stat.psu.edu/~bgl/nsf_report.pdf*. This article is an abridged and updated version of the report.

There were about 50 participants in the workshop, chosen to represent the breadth of the statistical profession. There were a significant number of non-U.S. participants, but for the most part the workshop and the full report were focused on statistical sciences within the boundaries of the United States.

The scientific committee decided that, for maximum impact, the report should be directed to a wide range of audiences. Thus our target was not just statistics researchers, but also such important supporting players as collaborators, department heads, college deans and funding agencies.

The workshop was designed to focus on aspects of the statistics field that were particularly relevant to the NSF. As a consequence, biostatistics was not included. It is a large and thriving subdiscipline, with many departments of biostatistics associated with medical schools throughout the United States. Even though it has been excluded, we should point out that research in biostatistics constitutes a major component of the total research effort in statistics.

We should also note that the National Science Foundation (1998) report 98-95, widely known as the Odom

Bruce G. Lindsay is Distinguished Professor, Department of Statistics, Penn State University, University Park, PA 16802-2111, USA (e-mail: bgl@psu.edu). Jon Kettenring is Adjunct Professor, Drew University, Summit, NJ 07901, USA (e-mail: jon29@earthlink.net). David O. Siegmund is Professor, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA (e-mail: dos@stat.stanford.edu).

Report (and so cited here), was an important document for our report because of its role in generating Foundation policy. It was written to provide an external assessment of the needs of the mathematical sciences including statistics.

Since it is well written and important, and provides support for many of our main points, it is cited often in the full report. For example, the report identifies the three primary activities of mathematicians as follows:

(1) Generating concepts in fundamental mathematics;
(2) Interacting with areas that use mathematics, such as science, engineering, technology, finance, and national security; and
(3) Attracting and developing the next generation of mathematicians.

After substituting statistics for mathematics, this trichotomy serves well to describe the primary activities of statisticians as well. One fundamental distinction between mathematics and statistics lies in the balance between items (1) and (2), as will be discussed later.

1.2 What Is Statistics?

An important driving force behind the workshop and the resulting report was the perception that the role of the statistics profession is often only poorly understood by the rest of the scientific community. Much of the intellectual excitement of the core of the subject comes from the development and use of sophisticated mathematical and computational tools, and so falls beyond the ken of all but a few scientists. One of our goals was to improve this situation.

To fulfill this need, the first speaker at the workshop, the eminent D. R. Cox of Oxford University, was asked to start with the basics and identify "What is statistics?" This question was to be repeatedly addressed throughout the course of the workshop for the sake of its wider scientific audience. We summarize some of the key points here.

Statistics is the discipline concerned with the study of variability, with the study of uncertainty and with the study of decision-making in the face of uncertainty. Whereas these are issues that are crucial throughout the sciences and engineering, statistics is inherently an interdisciplinary science. Even though statistics does not have its own concrete scientific domain (like rocks, clouds, stars or DNA), it is united through a common body of knowledge and a common intellectual heritage.

A distinguishing feature of the statistics profession, and the methodology it develops, is the focus on a set of cautious principles for drawing scientific conclusions from data. This principled approach distinguishes statistics from a larger venue of data manipulation, organization and analysis. An overarching principle dictates that one should provide a measure of the uncertainty for scientific statements based on data. Such statistical tools as confidence coefficients, significance levels and credible regions were designed to provide easily interpreted measures of validity. When used appropriately, these tools help to curb drawing false conclusions from data.

Of course, statisticians do not own the tools of statistics any more than mathematicians own mathematics. Certainly most statistical applications and much statistical research is carried out by scientists in other subject matter areas. The essential role of statistical research is to develop new tools for use at the frontiers of science. In the later sections of this report we will demonstrate the very exciting statistical research possibilities that have arisen in recent years. In particular, the possibilities for data collection and storage have opened the need for whole new approaches to data analysis problems.

1.3 Our Scientific Domains

The scientific domains of statistical work are nearly as wide as all scientific endeavor. In the workshop we focused on two main areas: the central part of our subject, which we called the core, came first. The keynote speaker was Iain Johnstone of Stanford University.

Then came five domains of application: biological science (Warren Ewens, University of Pennsylvania), engineering and industrial statistics (Vijay Nair, University of Michigan), geological and environmental sciences (Richard Smith, University of North Carolina), information technology (Werner Steutzle, University of Washington), and social and economic science (Joel Horowitz, Northwestern University). These categories were chosen to correspond roughly to the different directorates of the National Science Foundation in which the research is supported. (In writing the report, a sixth domain—physical sciences—was added due to the increasing role of statistics in this area.)

In addition, there were talks by Chris Heyde (Australian National University and Columbia University) and James Berger (Duke University) on "Statistics in the International Scene" and "Institutes: The Role and Contribution to Statistics," respectively.

Note that the subject of statistics does not have an agreed-upon division of its heritage into distinct areas of research, such as "algebra, analysis, topology and geometry" in mathematics or "inorganic, organic, physical and biophysical" in chemistry. There are instead a central portion of the research, which we have chosen to call the core, and a variety of applications-oriented research that we have divided by scientific field.

In a later section of this article, each of these areas, save one, will be given an overview in terms of research activity. Unfortunately, social and economic science was excluded from the full report. The difficulty the editors faced is that this area is not only rather separated from the rest, but is also quite complex. Research workers in the field are most often housed not in statistics departments, but instead in such departments as economics, psychology or sociology. It includes several domains that have their own mature, specialized statistics literatures, such as psychometrics and econometrics. We felt that an adequate review of this rich and sophisticated area was beyond our time frame and resources.

As a supplement to this report, it is worth noting that the book *Statistics in the 21st Century* (Raftery, Tanner and Wells, 2002) contains 70 papers written by many of the leading scholars of today. It is recommended to statisticians as a valuable compendium of information, covering the current status and future directions of research in a wide variety of statistical topic areas.

1.4 The Statistical Community

By the nature of their work, statisticians work in a wide array of environments. In the United States there are many statisticians who work in departments of statistics. Such departments are found at most of the major research universities. There are now 86 Ph.D. programs in statistics, biostatistics and biometrics. They have tended to focus on graduate research, including collaboration with other disciplines, and education, as well as undergraduate service courses.

These departments largely arose by splitting off from mathematics departments in the second half of the twentieth century. Based on this long term relationship, statistics is often viewed as a branch of mathematics. This structural view is evidenced in the National Science Foundation itself, in which probability and statistics is one program of the Division of Mathematical Sciences, placed side by side with such "pure" branches as topology and algebra.

However, one of the key conclusions of the participants of the workshop was that statistics has become more and more distinct from the other mathematical areas. The scientific goals of statisticians and the directions of modern science point to a world where computer and information science tools are at least as important to statistics as those of probability theory.

A substantial fraction of the academic statistics community works in departments other than statistics. This can occur even in universities with statistics departments, where they can be found in business schools, social science and science departments across the spectrum. In schools without statistics departments, as for example in four-year colleges, there are often statisticians within the mathematics department, where they are needed for undergraduate education. Finally, there are also many statisticians who work in biostatistics departments and in medical schools.

Going beyond the academic community, but well connected to it, are many more statisticians employed in government and business, as well as many users of statistics. Regarding the field of statistics, the Odom Report stated:

> The interaction between the academic community and users in industry and government is highly developed, and hence there is a rapid dissemination of theoretical ideas and of challenging problems from applications, as well as a tradition of interdisciplinary work.

Statisticians are found in government agencies from the Census Bureau to the National Institute of Standards and Technology to the National Institutes of Health. They are employed across a wide range of industries, often for quality control work. In particular, the pharmaceutical industry has been a leading employer of statisticians, who carry out the design and analysis of experiments required for drug development.

2. CURRENT STATUS

In the full report, the second chapter provides an overview of the history of statistics. Since the main reason for its inclusion was to inform the general scientific audience about the development of statistics over the course of the twentieth century, it has been omitted from this article, and we proceed to the current status of the profession. For the material in this section, the editors are extremely grateful for the contributions of Iain Johnstone, who developed most of the data and content as part of his workshop address.

2.1 The Quality of the Profession

The Odom Report provided a strong endorsement of the quality of the U.S. effort in statistics, stating that "the statistical sciences are very healthy across all subareas in the United States, which is the clear world leader."

At the workshop Iain Johnstone presented an informal survey of four leading statistics journals (two of which are based in the United Kingdom) to provide evidence for this statement. Table 1 shows the affiliation of the U.S.-based authors in these journals. Approximately one-half of the authors had U.S. affiliations. Most of these authors are in academic institutions. Moreover, the vast majority come from statistics or biostatistics departments, with less than 1 in 10 coming from a department of mathematics or mathematical sciences. Table 2 shows the reported sources of funding for this published research.

Clearly the National Science Foundation and the National Institutes of Health are the major role players in funding research in statistics. However, as will come up later, this split in our funding is a key factor in diminishing our presence in either scientific agency.

2.2 The Size of the Profession

One way to gauge the size of the statistics profession is to compare it with the rest of mathematics. In Table 3 we give the approximate number of members in the leading statistics and mathematics societies. These numbers are somewhat difficult to compare due to overlapping membership lists and possible reporting biases, but they do suggest that the number of statistics professionals might be somewhere between one-fourth to one-half the number of mathematicians.

The American Mathematical Society annual survey for 2001 indicates that there are 86 doctoral programs in statistics, biostatistics and biometrics (Group IV). This can be compared with 196 programs in other areas of mathematics (Groups I, II, III and V). Again, the numbers are not easy to compare, but do provide some idea of the scale.

TABLE 1

Statistics	49%
Biostatistics	23%
Industry	6%
Math. Science	5%
Mathematics	4%
Other	13%

TABLE 2

NIH	40%
NSF	38%
NSA	9%
ARO/ONR/EPA	4%
Other	9%

TABLE 3

American Statistical Association (ASA)	16,000
Institute of Mathematical Statistics (IMS)	3,500
Biometric Society (ENAR/WNAR)	3,500
American Mathematical Society (AMS)	30,000
Mathematical Association of America (MAA)	33,000
Society for Industrial and Applied Mathematics (SIAM)	9,000

A better measure might be the annual number of statistics Ph.D.'s. However, these counts suffer from many of the usual data collection challenges: definition of population, quality of data and census nonresponse. Table 4 presents three rather different numbers for statistics, as well as two estimates for the rest of mathematics.

The AMS survey acknowledges problems with nonresponse from statistics programs. The NSF Survey of Earned Doctorates number is derived by aggregating "statistical subfields" from the nearly 300 fine categories by which fields of doctorate are categorized in this Foundation-wide survey.

If we consider the number of doctorates in math excluding statistics, there is greater coherence between the AMS and NSF surveys, again suggesting problems with the identification and collection of data for statistics in particular.

The NSF survey does provide data back in time that is useful for understanding how the relationship between statistics and the rest of mathematics has changed over the last 35 years. Figure 1 shows that the annual number of statistics Ph.D.'s (per NSF definition) started at 200, less than $1/3$ the number of mathematics degrees, but has grown more or less linearly ever since to 800, staying roughly equal with mathematics in the 1980s and falling behind mathematics slightly since.

The number of research doctorates is a noisy surrogate for the level of research activity. Regardless, one might find it surprising that there are three program directors in the Division of Mathematical Sciences (DMS) for Statistics and Probability as opposed to 19 for all other mathematical areas. This does not

TABLE 4

AMS Survey 2000 (excluding probability)	310
Amstat Online 2000 (self reports)	457
NSF Survey of Earned Doctorates 2000 (accumulated over statistical subfields)	822
For reference, math excluding statistics	
AMS Survey 2000	809
NSF Survey of Earned Doctorates	925

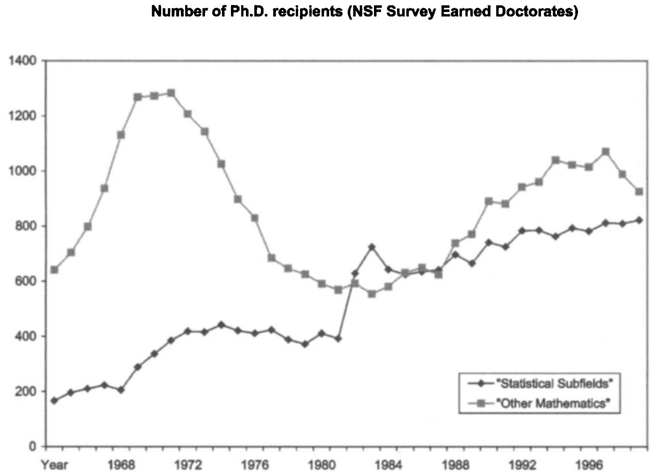

FIG. 1. *NSF survey on the number of doctorates by subject matter.*

seem proportionate to the size of the discipline or its potential importance in building connections between DMS and other sciences and engineering. This issue has been discussed with the DMS leadership. Their response is that the number of program officers in an area is strongly related to the number of research proposals received in that area, and at this time statistics is actually overrepresented relative to this number, the so-called proposal pressure.

Thus we are, in a sense, victims of our own success at being interdisciplinary and obtaining funding from other sources, particularly the National Institutes of Health (NIH). This plus the low funding rates in the DMS are the sources of reduced proposal pressure. One consequence is that we have a smaller role, and less influence, in this division of the NSF. Another consequence, to be developed later, is that there is not enough encouragement to do core research, as DMS is virtually the only source of funding.

2.3 The Odom Report: Issues in Mathematics and Statistics

The Odom Report provided some broad statements about the most important issues in mathematics as a whole. In this section we discuss them in the context of the current status of statistics. We will later revisit these themes in the designated subareas.

2.3.1 *Data collection.* A major theme of our report is that the statistics profession is experiencing a dramatic growth in its scientific value and its scientific workload due to changes in science and, in particular, data collection. The Odom Report stated that:

> With the advent of high-speed computers and sensors, some experimental sciences can now generate enormous volumes of data—the human genome is an example—and the new tools needed to organize this data and extract significant information from it will depend on the mathematical sciences.

Of all the mathematical sciences, the statistical sciences are uniquely focused on the collection and analysis of scientific data. Every senior statistician has felt the impact of this startling growth in the scale of data in recent years.

2.3.2 *Increased opportunities for scientific collaboration.* A second major theme of this report is that concurrent with the increased demand for statistical knowledge in the sciences, comes an increased pressure for statisticians to make major time commitments to gaining knowledge and providing expertise in a diverse set of scientific areas. As noted in the Odom Report:

> Both in applications and in multidisciplinary projects... there exist serious problems in the misuse of statistical models and in the quality of the education of scientists, engineers, social scientists, and other users of statistical methods. As observations generate more data, it will be essential to resolve this problem, perhaps by routinely including statisticians on research teams.

The Odom Report further noted the scientific problems of the future will be extremely complex and require collaborative efforts. It states that it will be virtually impossible for a single researcher to maintain sufficient expertise in both mathematics/computer science and a scientific discipline to model complex problems alone. We wholeheartedly agree with this finding.

2.3.3 *The next generation.* In several ways the future challenges to statistics differ from those of mathematics. For example, the Odom Report identifies three key issues:

> ...the mathematics community in the United States shares with other nations significant disciplinary challenges including a condition of *isolation* from other fields of science and engineering, a *decline* in the number of young people entering the field,

> and a *low level of interaction* with nonacademic fields, especially in the private sector. [Emphasis is ours.]

It is clear that the middle concern is of great importance to us. It is our observation that the number of U.S. residents entering the statistics field has shrunken over the years and the growth in Ph.D. degrees has come largely from foreign recruitment.

On the other hand, in the opinion of the scientific committee, the Odom Report's concern about isolation from other fields, scientific and nonscientific, applies less and less to the current statistics scene.

3. THE CORE OF STATISTICS

Outside the collaborative domain, the core activity of statisticians is the construction of the mathematical, conceptual and computational tools that can be used for information extraction. Much of the research has as its mathematical basis probability theory, but the end goal is always to provide results useful in empirical work. This distinguishes the theoretical research efforts of statisticians from most areas of mathematics in which abstract results are pursued purely for their intrinsic significance. As was stated in the Odom Report:

> Statistics has always been tied to applications, and the significance of results, even in theoretical statistics, is strongly dependent on the class of applications to which the results are relevant. In this aspect it *strongly differs* from all other disciplines of the mathematical sciences except computational mathematics. [Our emphasis.]

The terminology "core of statistics" is not routinely used by statisticians and so it is useful to describe more precisely its intended meaning. We define the core of statistics as that subset of statistical activity that is focused inward, on the subject itself, rather than toward the needs of statistics in particular scientific domains. As a synonym for "core," the word "inreach" might be offered. This would reflect the fact that this core activity is the opposite of outreach. As such, almost all statisticians are active in both inreach and outreach activities.

Research in the core area is focused on the development of statistical models, methods and related theory based on the general principles of the field. The objectives are to create unifying philosophies, concepts, statistical methods and computational tools. Although this is introspective activity, as noted above, a central philosophy of the core is that the importance of a problem is not dictated by its intrinsic beauty (as, say, in abstract mathematics). Rather, its importance is dictated by its potential for wide application or, alternatively, for its value to expand understanding of the scientific validity of our methods.

Through this combination of looking inward and looking outward, the core serves very much as an information hub. It is defined by its connectivity to, and simultaneous use in, virtually all other sciences. That core statistical concepts and methodology can be used simultaneously in a vast range of sciences and applications is a great source of efficiency in statistics and, as a consequence, provides high value to all of science.

Core research might be contrasted to "application-specific statistical research," which is more closely driven by the need to analyze data so as to answer questions in a particular scientific field. Of necessity, this research draws on core knowledge for tools as well as for an understanding of the limitations of the tools. It also provides raw material for future core research through its unmet needs.

3.1 Understanding Core Interactivity

As Johnstone noted in his workshop address, one way to demonstrate the amazing way that the core activities of statistics provide widespread value to the scientific community is to consider data on the citations of statistical literature. He did offer a strong caution that citation data should not be overinterpreted, because high citations for individual articles can reflect things other than quality or intrinsic importance. Just the same, it is offered here because it provides a simple and accessible measure of the widespread influence of statistical research on scientific fields outside of statistics.

The Institute of Scientific Information (ISI), which produces the *Science Citation Index* and its relatives, created several lists of the most cited scientists in the 1990s. Based on data provided by Jennifer Minnick, ISI (Oct. 11, 2000), 18 of the 25 most cited mathematical scientists of the period 1991–2001 were statisticians or biostatisticians. Citation counts per author are given in Figure 2. In addition, the *Journal of the American Statistical Association* was far and away the most cited mathematical science journal. (Editors' note: Since the full report the data on citations has shown an even more remarkable shift toward statisticians; see the website *http://in-cites.com/top/2003/index.html* for

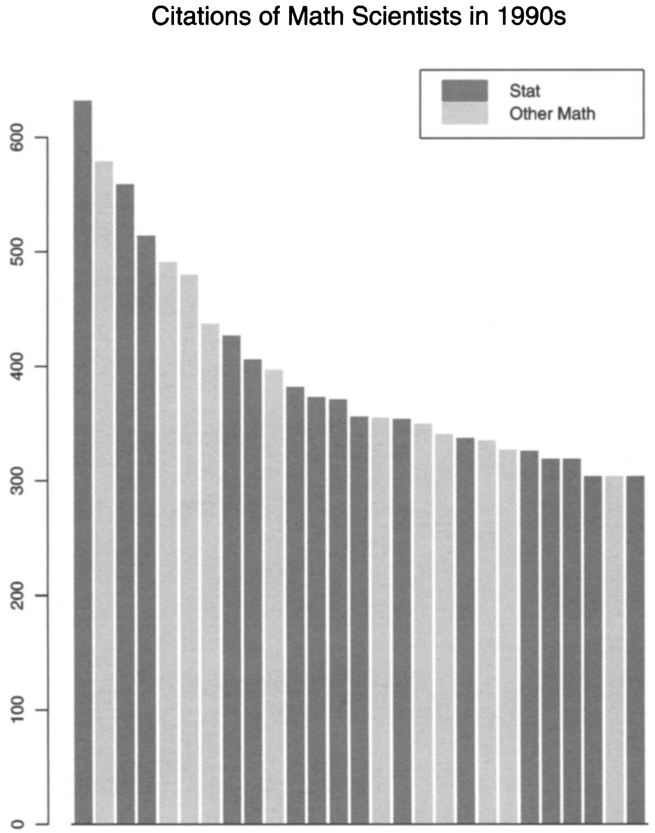

FIG. 2. *Citation counts of the most cited mathematical scientists.*

a list of the top 10 names; all were statisticians at the time of this writing.)

There is evidence that this high rate of citation of statistical articles, relative to mathematics as a whole, is related to its wide scientific influence. For example, the paper by Hall and Titterington (1987), which considers the thorny problem of choosing a smoothing parameter in nonparametric function estimation, has about 2/3 of its citations outside any definition of the core of statistics, including the IEEE journals, *Journal of Microscopy*, *Biomedical Engineering* and *Journal de Physique*. This is despite its appearance in a core research journal, and its theoretical cast.

One of the most important articles that leapt directly from core research into the mainstream of many scientific areas is the one by Efron (1979) that introduced bootstrap methods. An examination of 500 recent citations of this paper shows that only 152 of these citations appeared in the statistics literature. Figure 3 shows the wide dispersal of this innovation that was generated in the core of statistics.

Of course, the core also arrives at meaningful and useful methods for science because it reaches out to specific areas, finds important ideas, and creates the necessary generalizations that widen applicability. As an example, consider the development of methods that had their origins in age-specific death rates in actuarial work. In 1972 and 1975 the ideas of proportional hazards regression and partial likelihood analyses were introduced, which greatly enriched the tools available for the analysis of lifetime data when one has censored data along with covariate information. Since that time, these ideas and this methodology have grown and spread throughout the sciences to all settings where data that are censored or partially observed occur. This includes astronomy, for example, where a star visible with one measurement tool might be invisible due to inadequate signal with a second measurement tool.

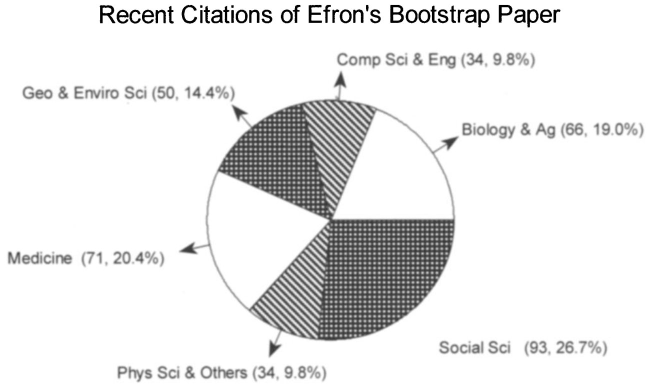

FIG. 3. *Of* 500 *recent citations of Efron's paper,* 152 *were in statistics. The distribution among sciences of the others is shown above.*

3.2 A Detailed Example of Interplay

The following recent example illustrates in more detail the theme that the core research in statistics feeds off and interacts with outreach efforts. Since at least some of the work is NSF funded, it indicates in part the kind of interactions that should be kept in mind when supporting core research.

In 2001 three astrophysicists published in *Science* (Miller, Nichol and Batuski, 2001) a confirmation of the Big Bang theory of the creation of the universe. They studied the imprint of so-called acoustic oscillations on the distribution of matter in the universe today and showed it was in concordance with the distribution of cosmic microwave background radiation from the early universe. It not only provided support for the Big Bang theory, it also provided an understanding of the physics of the early universe that enabled predictions of the distribution of matter from the microwave background radiation forward and backward in time.

The discovery was made using a new statistical method, the false discovery rate (known as the FDR), to detect the oscillations. At false discovery rate $1/4$, eight oscillations were flagged as possibly inconsistent with a smooth, featureless power spectrum. This and further analyses led the authors to conclude that the oscillations were statistically significant departures from a featureless matter-density power spectrum.

The method was developed through collaboration with two statisticians and published in *The Astronomical Journal* (Miller et al., 2001). Using this method, the authors were able to make their discovery and publish it in *Science* (Miller, Nichol and Batuski, 2001) while other competing groups were still plowing through the plethora of data.

It is interesting to trace the history of this success, because it illustrates quite well how the "information hub" operates. Figure 4 illustrates the migration route of the statistical idea.

When one tests many hypotheses on the same data set, one must adjust the significance levels of the tests to avoid spurious rejection of true null hypotheses. This "simultaneous inference" problem has perhaps received the most attention in medical statistics—at least, all of the references cited as motivation appeared in the medical literature. Indeed, the main statistical contribution here was not to propose the sequential P-value procedure that was used in this example per se, which actually went back to Simes in the 1980s (and maybe earlier) (Simes, 1986), but rather to establish a convincing theoretical justification. This theoretical justification, the FDR control, led other researchers to propose a version for estimation.

From Medicine to The Big Bang via FDR

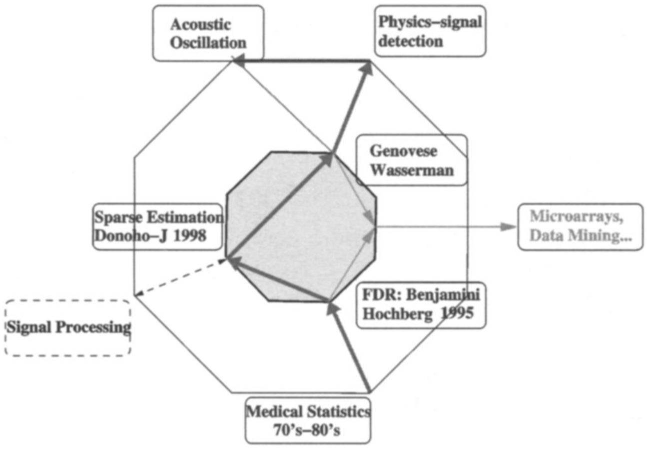
Acoustic Oscillation
Physics–signal detection
Genovese Wasserman
Sparse Estimation Donoho–J 1998
Microarrays, Data Mining...
Signal Processing
FDR: Benjamini Hochberg 1995
Medical Statistics 70's–80's

FIG. 4. *Illustraton of the migration of statistical ideas into the core from outside, their generalization in the core, followed by their export to new areas.*

The estimation proposal caught the attention of others because of its potential for threshold selection in wavelet shrinkage methods for statistical signal processing. Statisticians at Carnegie Mellon University (CMU) began work on FDR, both as a core statistics topic and also in collaboration with astrophysicists Miller and Nichol, two of the above-mentioned astrophysicists. Initially, they considered signal detection problems in huge pixel arrays. Later in their collaboration, the physicists recognized that this approach would apply to the acoustic oscillation signatures, which led to the *Science* article.

Miller and Nichol report that when they give talks to the physics community on this work there is great interest in the FDR approach. CMU physics professor Bob Nichol writes, in part, "I personally would like to emphasize the symbiotic relationship that has grown between the statisticians and astrophysicists here at CMU. It is now becoming clear that there are core common problems both sets of domain researchers find interesting e.g. application of FDR to astrophysical problems."

In fact, the astrophysicists appreciate the mathematical beauty of the statistics (and want to be involved), while the statisticians clearly relish their role in helping to understand the cosmos. In addition to these joint projects, this collaboration also is driving separate new research in the individual domains. In summary, this multiway collaboration has simulated both new joint research, as well as new separate research in the domain sciences. Therefore, it is a perfect marriage!

3.3 A Set of Research Challenges

What are the research challenges facing the core of statistics? Identifying such challenges in statistics is inherently different than in other sciences. Whereas in mathematics, for example, much focus has been given to famous lists of problems whose challenge is enduring, in statistics the problems always evolve relative to the development of new data structures and new computational tools. In addition, unlike the laboratory sciences, statistics does not have big expensive problems with multiple labs competing—or cooperating—on major frontiers. It is perhaps more true in statistical science than in other sciences that the most important advances will be unpredictable.

For this reason we need to maintain an underlying philosophy that is flexible enough to adapt to change. At the same time it is important that this future research not degenerate into a disparate collection of techniques.

One can identify some general themes driving modern core area research. The challenges are based on the development of conceptual frameworks and appropriate asymptotic approximation theories for dealing with (possibly) large numbers of observations with many parameters, many scales and complex dependencies. The following subsections identify these issues in more detail.

3.3.1 *Scales of data.* It has become commonplace to remark on the explosion in data being gathered. It is trite but true that the growth in data has been exponential, in data analysts quadratic and in statisticians linear. Huber's 1994 taxonomy of data sizes,

$$\text{Tiny } 10^2, \text{Small } 10^4, \text{Medium } 10^6,$$
$$\text{Large } 10^8, \text{Huge } 10^{10}$$

already looks quaint (Wegman, 1995). For example, a single data base for a single particle physics experiment using the BaBaR detector at the Stanford Linear Accelerator Center has 5×10^{15} bytes.

There will continue to be research issues at every scale—we have not solved all problems for data sets under 100. However, a new part of the challenge to statistics is that the mix of issues, such as generalizability, scalability, and robustness, as well as the depth of scientific understanding of the data, will change with scale and context. Moreover, it is clear that our research and graduate training has yet to fully recognize the computational and other issues associated with the larger scales.

3.3.2 *Data reduction and compression.* We need new "reduction principles." R. A. Fisher gave us many key ideas for data reduction, such as sufficiency, ancillarity, conditional arguments, transformations, pivotal methods and asymptotic optimality. Invariance came along later. However, there is a clear need for new ideas to guide us in areas such as model selection, prediction and classification.

One such idea is the use of "compression" as a guiding paradigm for data analysis. The basic idea is that good structural understanding of data is related to our ability to compactly store it without losing our ability to "decompress" it and recover nearly the original information. For example, in the domain of signal and image data, wavelets are actually not optimal for representing and compressing curved edges in images. This suggests the need for new representational systems for better compression.

3.3.3 *Data analysis outside statistics.* Many methods and computational strategies—such as machine learning and neural networks—have developed outside the field of statistics. For the most part these methods are not informed by the broader understanding they might gain from integration into mainstream statistics. Thus future research should involve coherently integrating the many methods of analysis for large and complex data sets being developed by the machine learning community and elsewhere into the core knowledge of statistics.

Following our tradition, this research could presumably be based on building models and structures that allow description of risk as well as its data-based assessment. It would then include developing principled tools for guided adaptation in the model building exercise. Another possibility was expressed by Breiman (2001), who stated that "If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."

3.3.4 *Multivariate analysis for large* p, *small* n. In many important statistical applications there are many more variables (p) than there are units being measured (n). Examples include analysis of curve data, spectra, images and DNA microarrays. A recent workshop titled "High Dimensional Data: $p \gg n$ in Mathematical Statistics and in Biomedical Applications," held in Leiden, Netherlands, highlighted the current research importance of this subject across many areas of statistics.

The following more specific example can be offered to illustrate how innovations in other fields might prove useful in this problem, thereby reinforcing the idea that the core continually looks outward for ideas. Random matrix theory describes a collection of models and methods that have developed over the last 40 years in mathematical physics, beginning with the study of energy levels in complex nuclei. In recent years these ideas have created much interest in probability and combinatorics.

The time now seems ripe to apply and develop these methods in high dimensional problems in statistics and data analysis. Scientists in many fields work with large data matrices [many observations (n) and many variables (p)] and there is little current statistical theory to support and understand heuristic methods used for dimensionality reduction in principal components analyses, canonical correlation analyses and so forth.

Early results suggest that "large n–large p" theory can in some cases yield more useful and insightful approximations than the classical "large n–fixed p" asymptotics. For example, the Tracy–Widom distribution for "Gaussian orthogonal ensembles" provides a single distribution, which with appropriate centering and scaling provides really quite remarkably accurate descriptions of the distributions of extreme principal components and canonical correlations in null hypothesis situations.

3.3.5 *Bayes and biased estimation.* The decade of the nineties brought the computational techniques and the power to make Bayesian methods fully implementable in a wide range of model types. A challenge for the coming decades is to fully develop and exploit the links between Bayesian methods and those of modern nonparametric and semiparametric statistics, including research on the possible combination of Bayesian and frequentist methodology.

One clear issue is that for models with huge data problems with large numbers of variables, the ideas of unbiasedness or "near" unbiasedness (as for the maximum likelihood estimator) become less useful, because the idea of data summarization implicit in statistical methodology becomes lost in the complexity and variability of any unbiased method. This points to the need for a more extensive "biased estimation theory" and new theories for huge data problems with large numbers of variables.

Given their ever-increasing use in all kinds of model-building exercises, it is also clear that there is a need for further analysis of Monte Carlo methods for inference.

3.3.6 *Middle ground between proof and computational experiment.* Yet another challenge for theoretical work in the coming decades is to develop an agreed-upon middle ground between the pace of proof (too slow), and the swamp of unfettered computational experimentation (too arbitrary and unconvincing). There are many problems in which rigorous mathematical verifications might be left behind in the development of methodology both because they are too hard and because they seem of secondary importance. For example, despite many years of work, there are important families of statistical models, such as mixture models, in which identifiability questions are largely ignored because of the difficult analysis that is involved and the ever-widening variety of model structures that must be investigated.

3.4 Building and Maintaining the Core Activities

Exploitation of the current manifold opportunities in science has led to an increased demand for greater subject matter knowledge and greater specialization in applications. This in turn has created a challenge for statistics by putting the core of the subject under stresses that could with time diminish its current effectiveness in the consolidation of statistical knowledge and its transfer back to the scientific frontiers. In essence, the multidisciplinary activities are becoming sufficiently large and diverse that they threaten professional cohesiveness.

If there is exponential growth in data collected and in the need for data analysis, why is core research relevant? Because unifying ideas can tame this growth, and the core area of statistics is the one place where these ideas can happen and be communicated throughout science. That is, promoting core statistics research is actually an important infrastructure goal for science from the point of view of efficient organization and communication of advances in data analysis.

A healthy core of statistics (through a lively connection with applications) is the best hope for efficient assimilation, development and portability between domains of the explosion of data analytic methods that is occurring. As such, it is a key infrastructure for science generally.

In Chapter 4 of the full report we identified and elaborated on the following opportunities and needs for the core:

- *Adapting to data analysis outside the core.* The growth in data needs provides a distinct challenge for statisticians to provide, in adequate time, intellectual structure for the many data analytic methods being developed in other arenas.
- *Fragmentation of core research.* Outreach activity is high and increasing for all sorts of good reasons. We think there has been an unintended consequence of this growth—a relative neglect of basic research and an attendant danger of our field fragmenting.
- *Manpower problem.* There is an ever-shrinking set of research workers in the United States who work in core area research. This manpower problem is destined to grow worse, partly from the general shortage of recruits into statistics and partly because outreach areas are pulling statisticians away from core research.
- *Increased professional demands.* The core research of statistics is multidisciplinary in its tools: It borrows from (at least) information theory, computer science and physics as well as from probability and traditional math areas. As statisticians have become more and more data-focused (in the sense of solving

real problems of modern size and scope), the skills needed in core areas have gone up. This need for ever-increasing technical skills provides a challenge to keeping the core vital as a place for integration of statistical ideas.

- *Research funding.* It seems clear that funding for core research has not kept pace with the growth of the subject. Investigators, rather than beating their heads against difficult funding walls, turn their efforts toward better funded outreach activities or consulting. The most basic needs remain the same: to nurture talent, to give senior people time and space to think, and to encourage junior people to buy into this line of research.
- *New funding paths.* New ideas for supporting research funding might help statistics meet its challenges. The full report contained an extended example of a funding strategy that might enable statisticians to enrich their basic statistical research with interdisciplinary activity without pulling them too far out of core research.

4. STATISTICS IN SCIENCE AND INDUSTRY

A distinguishing feature of statistics as a discipline is its interaction with the entire spectrum of natural and social sciences and with technology. This section is concerned with elucidation of the role of statistics in gathering knowledge across a wide spectrum of possibilities. Rather than give a broad, and necessarily incomplete, survey of areas in which statistics has had, and will continue to have, an impact, this section focuses on topics that illustrate important aspects of the interplay between statistics and other scientific disciplines.

4.1 Biological Sciences

Building on the foundations of agricultural and genetic statistics developed in the first half of the twentieth century, biostatistics, statistical epidemiology and randomized clinical trials have been cornerstones of the systematic attack on human disease that has dramatically increased life expectancy in advanced societies during the past half century.

Recent progress in molecular biology and genetics has opened entirely new areas of investigation, where for the foreseeable future there will be rapid advances in understanding fundamental life processes at the molecular level. The long term goals of this research are application of the knowledge of molecular processes to entire organisms and populations. These goals include improved tailoring of medical treatments to the individual (e.g., by devising treatment suited to the individual's genetic makeup), alleviation of malnutrition and starvation by improving agriculturally important plant species and domestic animals, improved public health, and better defense against bioterrorism.

At the risk of oversimplifying the many new developments in biological research, it is useful to consider four areas where statistical and computational methods have played and will continue to play an important role:

- *Biomolecular sequence analysis and functional genomics* refers to science based on analysis of DNA sequences (the building blocks of genes) and amino acid sequences (the building blocks of proteins), and global profiles of RNA and proteins in various cellular states, as used to discover the structure and evolution of genes and proteins, and their functions in normal and abnormal processes. An example is the identification of control regions imbedded in the genome that govern the amount of protein produced and the conditions under which it is produced.
- *Genetic epidemiology.* The goal of genetic epidemiology is to understand the relative importance of environment and genetics in human disease. For example, *gene mapping* involves the use of maps of molecular markers throughout the genome of a particular plant or animal to locate the genes that contribute to phenotypes of interest. It is frequently the first step toward better understanding and treatment of those diseases in plants and animals where inheritance plays an important role.
- *Evolution, population genetics and ecology* study the changes that occur at the population level in plants and animals in response to random mutational changes in the population's gene pool and changes in their environment. Although originally oriented toward the study of evolutionary relationships (e.g., the evidence supporting the hypothesis of a common African origin of modern humans), the ideas of population genetics are increasingly used to understand the evolution of bacteria and viruses (to provide appropriate vaccines and drugs) and the evolution of proteins in different species of plants and animals (to understand protein structure and function by identifying parts of related proteins in different species that have been conserved by evolution).
- *Computational neuroscience* uses modern methods of neuroimaging (PET, fMRI) to gain understanding of the functioning of nervous systems. This raises

questions both at the level of small numbers of interacting neurons and at the level of the entire brain: Which parts of the brain are activated under which conditions? How do the brains of normal and psychotic individuals differ in their structure and/or function? How can we use this knowledge for diagnosis and treatment?

4.1.1 *Statistical and computational methods.* As a consequence of this enormous diversity of scientific problems, an expansive set of statistical, probabilistic and computational methods has proved to be very useful. Some methods have proved themselves in a number of areas, while others have more specialized applications.

Stochastic processes, from finite Markov chains to point processes and Gaussian random fields, are useful across the entire spectrum of problems. Because of the large amount of data produced [e.g., expression levels on a microarray for tens of thousands of genes in a sample of individuals, or data from up to a thousand markers (in the future perhaps one hundred thousand) distributed across the genome of thousands of individuals], challenging issues of multiple comparisons fre quently arise in these areas.

Hidden Markov models and Markov chain Monte Carlo provide important computational algorithms for calculating and maximizing likelihood functions. Some of these statistical methods are classical (e.g., principal components, likelihood analysis), but even they may require adaptation (principal curves, likelihood analysis of stochastic processes) to deal with the large amounts of data produced by modern biological experiments. Other methods (hidden Markov models, Markov chain Monte Carlo) have developed relatively recently in parallel with the modern computing technology necessary to implement them.

A common feature of all the efforts described above is the amount, complexity and variability of data, with the result that computation (frequently including graphics) is an important aspect of the implementation of every idea. In view of the diverse mathematical and computational backgrounds of scientists engaged in biological research, it is important that computational algorithms be made as "user-friendly" as possible. This may require support for specialists to provide the "front end" and documentation necessary so that laboratory scientists can use tools developed by statisticians easily and correctly.

In summary, the large amounts of data produced by modern biological experiments and the variability in human response to medical intervention produce an increasing demand for statisticians who can communicate with biologists and devise new methods to guide experimental design and biological data analysis.

4.2 Engineering and Industry

Statistical concepts and methods have played a key role in industrial development over the last century. Applications in engineering and industry, in turn, have been major catalysts for research in statistical theory and methodology. The richness and variety of these problems have greatly influenced the development of statistics as a discipline.

Much of the early work was driven by the needs of the agricultural, manufacturing and defense industries. In recent years, the scope has expanded substantially into business and finance, software engineering, and service and health industries. Applications in these areas include credit scoring, customer profiling, design of intelligent highways and vehicles, e-commerce, fraud detection, network monitoring, and software quality and reliability.

Global competition and increasing customer expectations are transforming the environment in which companies operate. These changes have important implications for research directions in statistics. Following are brief descriptions of four general examples.

4.2.1 *Massive data sets with complex structure.* This topic cuts across all parts of business and industry (as well as other areas discussed in this report). Business and manufacturing processes are becoming increasingly complex. Consequently, engineers and managers are in greater need of relevant data to guide decision-making than ever before.

At the same time, advances in sensing and data capture technologies have made it possible to collect extensive amounts of data. These data often have complex structure in the form of time series, spatial processes, texts, images, very high dimensions with hierarchical structure and so on. Collection, modeling and analysis of these data present a wide range of difficult research challenges.

For example, monitoring, diagnosis and improvement of advanced manufacturing processes require new methods for data compression and feature extraction, development of intelligent diagnostics, and real-time process control. These problems also involve issues of a general nature such as selection biases, computing, scalability of algorithms and visualization.

4.2.2 *Large-scale computational models.* Computational models and simulation are being used more and more frequently in many areas of application. In manufacturing industries, competitive market forces and the concomitant pressure to reduce product development cycle times have led to less physical testing and greater use of computer-aided design and engineering methods. Finite-element analysis and other techniques are used extensively in the automobile industry for product design and optimization.

There are similar trends in semiconductor manufacturing, aircraft, defense and other industries. The computational models are very high dimensional, involving hundreds and even thousands of parameters and design variables. A single function evaluation can take several days on high-end computing platforms.

Experimentation, analysis, visualization and validation using large-scale computational models raise a variety of statistical challenges. These include (a) development of experimental designs for approximating and exploring response surfaces in very high dimensions; (b) incorporating randomness and uncertainty in the design parameters and material characteristics into the computational model; (c) modeling, screening, prediction and optimization.

4.2.3 *Reliability and safety.* The design, development and fabrication of highly reliable products that also meet safety and environmental goals represent another area of major challenge faced by industry. The traditional focus in reliability has been on the collection and analysis of "time-to-failure" data. This poses difficulties in high-reliability applications with few failures and high degrees of censoring.

Fortunately, advances in sensing technologies are making it possible to collect extensive amounts of data on degradation and performance-related measures associated with systems and components. While these data are a rich source of reliability information, there is a paucity of models and methods for analyzing degradation data and for combining them with physics-of-failure mechanisms for efficient reliability estimation, prediction and maintenance. Degradation analysis and device-level failure prediction are integral parts of predictive maintenance for expensive and high-reliability systems.

There are also vast amounts of field-performance data available from warranty and maintenance data bases. Mining these data for signals and process problems and using them for process improvement should be a major area of focus.

4.2.4 *Software engineering.* This is still a relatively new field compared with traditional branches of engineering. Its importance to the nation is underscored by the increasing reliance of the U.S. economy and national defense on high quality, mission critical software (National Research Council, 1996).

Statistics has a significant role to play in software engineering because data are central to managing the software development process, and statistical methods have proven to be valuable in dealing with several aspects of it. To mention a few examples, statistical considerations are essential for the construction and utilization of effective software metrics, and experimental design ideas are the backbone of technology for reducing the number of cases needed to test software efficiently (but not exhaustively). Furthermore, statistical quality control provides the basis for quantitative analysis of various parts of the software process and for continuous process improvement.

4.3 Geophysical and Environmental Sciences

The term "geophysical and environmental sciences" covers many specific fields of study, particularly if environmental sciences are taken to include the study of ecological phenomena and processes. This broad area of statistical activity does not have an easily summarized history nor a simple pattern of development. Indeed, the history of statistical work in the geophysical and environmental sciences is intertwined with fields as diverse as agriculture, biology, civil engineering, atmospheric chemistry and ecology, among others.

The collection and processing of large amounts of data are features in many of the major components of geophysical and environmental sciences such as meteorology, oceanography, seismology, the detection and attribution of climate change, and the dispersion of pollutants through the atmosphere.

Statisticians have been actively involved in all of these fields, but as statistical methodology has advanced to include, for example, complex models for spatiotemporal data and the associated methods of computation, the potential for direct interactions between statisticians and geophysical and environmental scientists has increased enormously. We offer a few examples of the types of challenges being addressed.

4.3.1 *Deterministic process models and stochastic models.* A substantial amount of emphasis is now being placed on the tandem use of deterministic process models and statistical models. Process models have typically taken fundamental scientific concepts such as mass balance in chemical constituents as a foundation and built up more elegant mathematical structures by overlaying equations that represent physical and chemical interactions, often in the form of sets of differential equations. Statistical models, on the other hand, typically rely on the description of observed data patterns as a fundamental motivation for model development. Increasingly, there is recognition that one's understanding of many geophysical and environmental processes can be advanced by combining ideas from these two modeling approaches.

One method that has been used to combine process and statistical models is to use the output of deterministic models as input information to a stochastic formulation. The full NSF report contains a detailed illustration of this process that arose in the analysis of bivariate time series representing northern and southern hemispheric temperature averages.

4.3.2 *Correlated data and environmental trends.* Many environmental problems involve the detection and estimation of changes over time. For example, an environmental monitoring agency such as the Environmental Protection Agency uses trend estimates to assess the success of pollution control programs and to identify areas where more stringent controls are needed. In climate modeling, a major preoccupation is to determine whether there is an overall trend in the data, not only for widely studied variables such as the global mean temperature, but also for numerous other variables where the conclusions are less clear-cut.

For statisticians, estimation of trend components with correlated errors is a problem with a long history, and much of this work involved substantial interaction between statisticians and geophysical and environmental scientists. For example, Sir Gilbert Walker, known to statisticians through his many contributions to time series analysis and, in particular, the Yule–Walker equations, was also a distinguished meteorologist who worked extensively on the El Niño—Southern Oscillation phenomenon, and these contributions were largely the result of the same research.

A long collaboration between statisticians and geophysicists has resulted in a series of papers on the detection of change in stratospheric ozone in which a large number of models with correlated errors are considered. This research, consisting largely of papers with a statistician as lead author but appearing in journals outside of the mainstream statistical outlets, is an excellent illustration of the outreach of statistics to other scientific fields.

4.3.3 *Statistical modeling and scientific conceptualization.* It is common for changes in environmental data records to be conceptualized within the statistical framework of *signal* plus *noise*. Indeed, this is the case for many of the models discussed above, in which various forms are given to the signal (or systematic) and noise (or error) components of the models to better represent the processes under study.

An example in which statistics has aided in the development of scientific thinking is the analysis of cycles in the populations of Canadian lynx and snowshoe hare, and a series of papers dealing with this has appeared in the *Proceedings of the National Academy of Sciences* (Stenseth et al., 1997, 1998, 2004a, b) and *Science*. Here, collaboration between statisticians and ecological scientists has resulted in a strengthening of scientific theory. A number of concepts have been developed through this work, including the relationship between autoregression order of a statistical model and the complexity of feedback systems between species (i.e., lynx and hare), and the idea that population cycles may exhibit spatial synchrony.

4.4 Information Technology

The rapid rise of computing and large-scale data storage has impacted many human endeavors, sometimes in profound ways. There has never been a more exciting time for statisticians working in areas related to information technology.

The development of the web and the exponentially increasing capabilities of computer systems have opened up possibilities previously undreamed for the exchange of information, the ability to collect and analyze extremely large data sets of diverse nature from diverse sources, and to communicate the results. The development of open source software magnifies the ability of researchers to leverage their talents and ideas. New challenges in statistical model-building and learning from data abound. The remainder of this section highlights a selected set of high-impact areas.

4.4.1 *Communications.* A wealth of communications records is generated every minute of every day. Each wireless and wireline call produces a record that reports who placed the call, who received the call, when and where it was placed, how long it lasted, and how it was paid for. Each user request to download a file from an Internet site is recorded in a log file. Each post to an on-line chat session in a public forum is recorded.

Such communications records are of interest to network engineers, who must design networks and develop new services, to sociologists, who are concerned with how people communicate and form social groups, to service providers, who need to ferret out fraud as quickly as possible, and to law enforcement and security agencies looking for criminal and terrorist activities.

There is a host of challenging statistical problems that need to be met before the wealth of data is converted into a wealth of information.

4.4.2 *Machine learning and data mining.* The line between research in machine learning and data mining, as carried out primarily in computer sciences departments, and research in nonparametric estimation, as carried out primarily in statistics departments, has increasingly become blurred. In fact the labels "machine learning" and "data mining" are now often used by statisticians. Primary areas of very active research within statistics departments include new methods for classification, clustering and predictive model-building. Statisticians have been developing classification tools for a long time, but the explosion in computational ability along with the fruits of recent research have led to some important new advances.

One such new advance in classification that takes advantage of these facts is *support vector machines*. The method is highly popular among computer science machine learning communities, but has greatly benefited from input by statisticians, who have contributed in important ways to understanding the properties of the method. However, there are important opportunities for further understanding of both the theoretical properties of this tool, and the most appropriate and efficient way to use this tool to recover information from data, in a broad variety of contexts.

4.4.3 *Networks.* The study of internet traffic can be roughly divided into traffic measurement and modeling, network topology, and network tomography. All of these areas present large-scale statistical challenges.

Further research in measurement and modeling is motivated by the need to jointly improve quality of service and efficiency. The current approach to quality of service is based on massive overprovisioning of resources, which is both wasteful and not completely effective, because of bursts in the traffic caused partly by improper protocol and routing procedures. Because many ideas for addressing these problems have been proposed, there is a major need for comparison, now done typically by simulation. This requires modeling to be done in a manner that addresses a serious statistical problem of goodness of fit. Classical statistical approaches and techniques are typically rendered impractical by the appearance at many points of heavy tailed distributions (often leaving even such standard tools as variance and correlation useless) and long-range dependence and nonstationarity (stepping beyond the most basic assumptions of classical time series).

Network topology presents different types of statistical problems. Here the goal is to understand the connectivity structure of the internet. Graph theoretic notions, combined with variation over time and also combined with sampling issues, are needed to make serious headway in this area.

Network tomography is about inferring structure of the Internet, based only on the behavior of signals sent through it. Proper understanding, analysis and modeling of the complex uncertainties involved in this process are important to make headway in this area.

4.4.4 *Data streams.* Statistical analyses of large data sets are often performed in what is essentially batch mode. Such data sets may require years to collect and prepare, and the corresponding statistical analyses may extend over a similar period of time. However, real-time data mining offers a rapidly growing niche for statisticians. Such situations arise, for example, in remote sensing, where limited bandwidth between an orbiting satellite and its ground station precludes transmission of all raw data. A second example is commercial web sites such as an airline reservation system, where detailed keystroke sequence data leading to actual or abortive reservations is not saved.

The challenge is to create statistical tools that run in almost linear time, that is, to design tools that can run in parallel with the real-time stream of data. For simple statistics such as sample moments, there are no difficulties. However, these tools must be able to adapt in real time. Furthermore, *data mining* makes use of virtually every modern statistical tool (e.g., clustering algorithms, trees, logistic regression). Transforming and recasting the statistical toolbox into this new and fundamentally important setting will require imagination, cleverness and collaboration with algorithmic experts in other areas of the mathematical sciences.

4.5 Physical Sciences

Historically, astronomy is one of the first and most important sources of inspiration for, and application of, statistical ideas. In the eighteenth century astronomers were using averages of a number of measurements of the same quantity made under identical conditions. This led, at the beginning of the nineteenth century, to the method of least squares.

Astronomy has expanded enormously, both in the size and complexity of its data sets, in recent years so as to estimate the Big Bang cosmological parameters from the anisotropic clustering of galaxies, the fluctuation spectrum of the cosmic microwave background radiation, and so forth. A host of other basic statistical problems arises from the Virtual Observatory, a federation of multiterabyte multiwavelength astronomical survey data bases.

Despite the common origins of statistics and astronomy, and our mutual interest in data analysis, only very recently have there been substantial collaborations between statisticians and astronomers. (One example of this type was presented earlier, in our discussion of the core.)

This long-standing gap between the fields of statistics and astronomy exemplifies a common pattern in the physical sciences. Statistics works by the efficient accrual of evidence from noisy individual information sources. In large part the historical spread of statistical methodology can be described as "noisy fields first": vital statistics, economics, agriculture, education, psychology, medical science, genetics and biology. The "hard sciences" earned their name from the almost perfect signal-to-noise ratios attainable in classical experimentation, so it is understandable that they have proved to be the most resistant to statistical methodology.

However, recent trends are softening the hard sciences, and so there is an increasing need for statistical principles and methods. Technology now enables bigger and more ambitious data-gathering projects such as those of the Sudbury neutrino observatory and the Wilkinson microwave anisotropy probe. These projects must extract crucial nuggets of information from mountains of noisy data. (The signal-to-noise ratio at Sudbury is less than one in a million.) Unsurprisingly, statistical methods play a big, sometimes crucial role in these projects.

To illustrate the promising future role of statistics in the physical sciences, we offer three brief statistics-intensive examples, from particle physics, chemical spectroscopy and astronomy.

4.5.1 *Confidence intervals in particle detection.* The following situation arises in the search for elusive particles: a detector runs for a long period of time, recording x interesting events; a similar run with the elusive particles shielded out yields a "background" count of y events. What is an upper confidence limit for the true rate of the particles of interest? Statistical issues become particularly sensitive if y exceeds x, so that the unbiased rate estimate is actually negative. The question then is whether the upper confidence limit is sufficiently positive to encourage further detection efforts.

Even in its simplest form—actual situations can involve much more elaborate background corrections—this problem has attracted widespread interest in the physics community. A much quoted reference is Feldman and Cousins (1998). Louis Lyons, professor of physics at Oxford, organized a September 2003 conference at the Stanford Linear Accelerator Center devoted to statistical problems in particle physics, astrophysics and cosmology (*www-conf.slac.stanford.edu/phystat2003/*).

4.5.2 *Comparative experiments in chemical spectroscopy.* Richard Zare of the Stanford chemistry faculty developed an advanced class of mass spectrometers able to simultaneously time the flights of large volumes of massive particles. This permits comparisons between collections of particles obtained under different conditions, for example, complex molecules grown in different chemical environments.

A typical spectrum consists of particle counts in binned units of time, perhaps 15,000 bins in a typical run. Comparing two such spectra, that is, looking for bins with significant count differences between the two conditions, is an exercise in simultaneous hypothesis testing. With 15,000 bins, the simultaneity is massive. Statistical methodology originally developed for microarray analysis can be brought to bear on spectroscopy comparisons, but the relationship between time bins is unlike that between genes, suggesting that new methodology will be needed.

4.5.3 *Survival analysis and astronomy.* In a spectacular example of parallel development, astronomy and biostatistics invented closely related theories to deal with missing data, the field called *survival analysis* in the statistics literature. The reasons for the missingness were different: astronomers are earth-bound so they cannot observe events too dim or too far away, leading to data "truncation." Data "censoring" occurs in medical trials when subjects fail to record a key event, such as relapse or death, before the end of the trial. Lynden–Bell's method and the Kaplan–Meier estimate, the astronomy and statistics solutions to the missing data problem, are essentially the same.

Mutual awareness did not occur until the 1980s. An influential series of joint astronomy–statistics conferences organized at Penn State by Babu and Feigelson led to collaborations and progress in the statistical analysis of astronomical data. For example, the extragalactic origin of gamma-ray bursts was demonstrated via survival analysis before the bursts could be identified with specific physical sources.

4.6 Enhancing Collaborative Activities

A distinguishing feature of the intellectual organization of statistics is the value placed on individual participation both in the development of statistical methodology and on multidisciplinary activities (e.g., in applications of statistics in biology, medicine, social science, astronomy, engineering, government policy and national security), which, in turn, becomes an important stimulus for new methodology. Although different people strike different balances between methodological research and subject matter applications, and the same people strike different balances at different times in their careers, essentially all statisticians participate in both activities.

Statistics, through these interactions, develops tools that are critical for enabling discoveries in other sciences and engineering. Statisticians are also instrumental in unearthing commonalities between seemingly unrelated problems in different disciplines, thereby contributing to or creating synergistic interactions between different scientific fields.

However, as noted by the Odom Report, our reach has not been wide or far enough. As they said:

> Both in applications and in multidisciplinary projects, however, there exist serious problems in the misuse of statistical models and in the quality of education of scientists, engineers, social scientists, and other users of statistical methods. As observations generate more data, it will be essential to resolve this problem, perhaps by routinely including statisticians on research teams.

One problem is that statisticians who attempt to participate widely in these activities face several steep challenges, including the need to stay current in all relevant fields and the need to provide the relevant software to carry out statistical analyses. In addition, in spite of a culture that encourages multidisciplinary activities, evaluating these activities has proven difficult and sometimes controversial.

In summary, we note that:

- The large amounts of data produced in almost all areas are producing an increasing demand for statisticians who can communicate with nonstatisticians, while devising new methods to guide experimental design and data analysis.
- As noted more extensively in the full report, there exists a software challenge that touches deeply in a number of areas. It corresponds to a wide need for statistical methods to be incorporated into open source software products, and a parallel lack of support for this infrastructure need.
- There exists a need for coordinated long term funding for interdisciplinary projects so that the statistician can afford to develop the scientific understanding vital to true collaboration.

5. STATISTICAL EDUCATION

We have identified a growing need for statistics and statisticians from wide areas of science and industry. As noted by the Odom Report, "There is ample professional opportunity for young people in statistics, both in academia, industry, and government." At the same time, the profession cannot meet demand with domestic candidates. Again, from the Odom Report, "A very high proportion of graduate students are foreign-born and many remain in the United States upon graduation."

At the same time that the demands on the profession have grown in the research arena, there has been a startling growth in demand in statistical education at lower levels:

- The statistics profession has started to feel the impact of the wide growth of statistical training in grades K–12, led by the implementation of an Advanced Placement (AP) course in statistics. This means many students are coming to college with substantial knowledge about statistics.
- Undergraduate enrollments in statistics were up sharply (45%) between 1990 and 2000.

These circumstances point to the need for the profession as a whole to consider how to handle this growth and how to build a statistics education infrastructure that can meet the changing and growing needs. The NSF workshop was not designed to give detailed consideration to the problems of statistical education. However, it was clear from the outset of the meeting that many of our current and future challenges are arising in the domain of education. As a result, a chapter of the full NSF report was dedicated to this topic. The goal of the chapter was to describe the current issues, as we saw them, and to invite further action by the community. Here is a summary of our suggestions.

5.1 Educational Reform

Clearly, the long-range solution to the shortage of statisticians must lie in improvements to the educational system that will attract, train, retain and reward a new generation of talented students. Improvements will be required across the spectrum from elementary school through continuing education of the workforce. The pool of K–16 teachers who are qualified to teach statistics needs to be increased.

We have identified the following key issues and needs in the area of educational improvements:

- Adequately trained teachers are needed for the statistics AP courses, as well as statistically literate instructors in other subjects in grades K–12. Considerable effort continues to be given to training K–12 teachers to take on AP statistics. Still, the need exceeds the supply. Dealing with the K–12 (and college level) shortage of trained teachers may well be the number one priority in statistics education.
- There is a need for an integrated K–16 curriculum that accounts for the improvement in high school statistics training. (We note that improving statistics literacy is just one piece of a larger challenge to provide quantitative preparation in this country.)
- Colleges and universities need expanded statistics minor and major options in both undergraduate and graduate programs. With the demand for undergraduate statistics courses on the rise, it is only natural to expect increased emphasis on providing a meaningful package of knowledge. Special pressure points can be seen in the demand from the pharmaceutical, biomedical and government statistical communities (including those involved in homeland security and national defense matters).
- We need to encourage and enable students, at all levels, to acquire deeper and broader subject matter knowledge in an area or areas of application. It has become increasingly difficult to be a successful statistician without involvement in a focused area of science and technology.
- At the graduate level, there is a large challenge to build training programs that can offer sufficient depth over the wide breadth of tools that the modern statistician is now using. This includes large chunks of mathematics, computer science and basic science. Unfortunately, the needs are expanding in all these areas.
- There is a growing need for postdoctoral training positions and other continuing education opportunities to help newly minted graduates develop their professional skills and help older statisticians stay up to date. Postdoctoral training should receive higher priority in the statistics community as a sensible way to launch research careers and broaden interests. The NSF's VIGRE plus NISS and SAMSI postdoctoral programs are signals of growth in this area.

In all aspects of the educational reform we have described, it is important that decisions are made using sound experimentation and good data, our own special forte.

5.2 Recruitment

The recruitment of the next generation defines a further set of challenges to our community. The number of people with training in statistics is not growing nearly fast enough for the exponential growth in the demands for statistical expertise. This trend must be changed dramatically to meet the high demands for statistical expertise for today's complex, large, interdisciplinary research problems in science and engineering.

There is no doubt that recruitment at the lowest levels has been helped by the AP course. At the same time, programs for enhanced recruitment that are focused on the mathematics profession as a whole, such as VIGRE, are very promising, but have many times lacked sensitivity to the special needs of statistics.

Much of what needs to be done in education and recruitment will require large investments. It is natural to look to the NSF for leadership and support. The time for action is now, as evidenced by such problems as the shortage of statisticians in key industries and many government agencies, heightened concerns about national security, and the increased reliance at the college level on adjuncts and instructors who are not members of the regular faculty.

6. A SUMMARY WITH RECOMMENDATIONS

The field of statistics has made profound contributions to society over the past century. Its impact is felt across virtually all branches of science, medicine, industry and government. The strong growth of the field—stimulated in large part by advances in computing technology—has also caused strains and frustrations as the opportunities have increased and the supporting infrastructure has struggled to keep pace. This report attempts to highlight many of the contributions that statistics has made and to identify priorities for continued progress.

Statistics is itself a science—the science of learning from data. It is grounded in a still growing core of knowledge that reflects its roots in probability and mathematics, and also the more recent influence of computer science. Statistics both draws from these roots and feeds back to them new mathematical and computational questions. Statistics is also an unusually interdisciplinary field. Indeed, applications are its lifeblood: they stimulate research on new theories and methods while providing valuable outlets for established techniques. Among the highest priorities for statistics today is adapting to meet the needs of data sets that are so large and complex that new ideas are required, not only to analyze the data, but also to design the experiments and interpret the experimental results. These problems are often the source of the widespread interdisciplinary collaborations—from astronomy to public policy to zoology—which statisticians engage in today.

A substantial proportion of the report describes contributions that statistics has made to other fields, including physical, biological and social sciences, as well as engineering and computer sciences. The motivation for this emphasis is to help clarify the often misunderstood role of statistical science and to illustrate its impact in a variety of ways. Special attention is also given to statistics education because there are substantial across-the-board opportunities for advancing the way the subject is taught and students are trained.

The sense of the workshop and the feeling of the leadership in the profession is that enormous opportunities lie ahead for statistics. The realization of this potential, however, will not come easily. Resources are too limited, the pipeline of students is too small and the infrastructure supporting the field is too constrained. To deal with these and other challenges, the following recommendations are put forth:

- *Promote understanding of statistical science.* Statistics is hard to pigeonhole. At the NSF, it falls largely under the mathematical sciences, yet most statisticians would agree that statistics is not a branch of mathematics. Modern statistics is also close to computer science, especially machine learning, yet most statisticians would agree that statistics is not a branch of computer science. Statistics is a science in itself, and attempts to group it here or there ultimately exacerbate misunderstandings about the field. Statisticians need to take responsibility for more effectively articulating the unique capabilities of their discipline. The NSF can help by assuring that, wherever it is housed, statistics can flourish without unproductive constraints.
- *Increase support for, and the autonomy of, the NSF statistics program.* To avoid stifling the momentum evident today in statistics (and partially documented in this report) and to reap the benefits of the multitude of opportunities presenting themselves, there are compelling reasons for providing a substantial boost in resources that support statistics at the NSF. (See below for some specific needs.) In addition, we suggest that the NSF provide the DMS statistics program with increased autonomy within its current organizational structure. This would be a logical step toward full division status that many feel is already overdue.
- *Develop more flexible funding models.* The creation of the new Statistics and Applied Mathematics Institute is an excellent example of creative new funding needed by the statistics profession. The needs, however, are not solely institutional. Increasingly, individual researchers are becoming involved in complex cross-disciplinary projects or in activities that are more akin to running a laboratory than doing individual research. One implication of this movement is the need for learning advanced programming techniques and development of sophisticated user-friendly software. We propose that the NSF develop novel funding arrangements that would encourage these new ventures while being careful not simply to extract these monies from the individual research grant pool.
- *Strengthen the core of statistics research.* The risk of fragmentation of the core of statistics has increased substantially as the field has diversified and expanded. More attention must be given to consolidation of knowledge and the development of new theories and methods with broad applicability. We urge the NSF to take responsibility for providing the level of support necessary to strengthen the statistics core.
- *Improve the support of multidisciplinary research activities.* Much of the excitement in science today stems from research that involves multiple disciplines. While statistics comes by this naturally, it often suffers from inclusion as an afterthought or exclusion as a minority player without a significant role. We encourage the NSF to experiment with new vehicles for funding this type of research and—when appropriate—to assure a role for statistics. For example, in many cases statisticians should be partners in projects with complex design and data analysis components. For such collaborations to succeed,

statisticians will need to have the time and support to understand the subject area.

- *Develop new models for statistics education.* The growth of AP statistics courses in high schools, the burgeoning enrollments in undergraduate statistics courses and major improvements in computing technology for data analysis underscore the need for reevaluation of the entire K–16 approach to statistics education. Graduate training is also due for reassessment: Keeping the right balance between training in the core parts of the science, preparing students for cross-disciplinary work and incorporating relevant parts of computer science into the curriculum are among the contributing factors to the awkward balancing act that departments face today. The role of postdoctoral training and continuing education more generally should also be part of the updated vision. To help the statistics community develop appropriate new models for education, and to do it both holistically and systematically, we suggest that the NSF sponsor or support a series of focused, coordinated workshops on statistics education with the aim of developing concrete plans for reform on these various fronts. It would be natural to carry out this undertaking in collaboration with the scientific and educational organizations that share responsibility for and concern about statistics education.
- *Accelerate the recruitment of the next generation.* Workshop participants pointed repeatedly to shortages in the pipeline of students and unmet demand from key industries and government laboratories and agencies. The long-range solution to this problem must lie in improvements to the education system, starting even in elementary school and continuing into high school and undergraduate programs. However, changes of this type will take much time and investment. Meanwhile, the shortage may prove quite damaging to the nation's infrastructure, especially in this period of heightened concerns about national defense and security—areas to which statistics has much to offer. Novel special programs designed to spur interest in undergraduate and graduate training in statistics should be considered. We encourage the NSF to join forces with leaders of the statistics profession to help solve the pipeline problem.

ACKNOWLEDGMENTS

The editors wish to thank the many people who contributed to the production of the NSF report. To start with, we are very indebted to the National Science Foundation for support of the workshop. Particular thanks go to Marianthi Markatou and John Stufken for their tremendous assistance at the time of the workshop and, more importantly, their strong enthusiasm and encouragement for its development.

The starting point for this report were the talks and written documents that the theme speakers at the workshop provided. After these presentations, the workshop participants divided into small groups according to theme and discussed the direction of statistics in the theme areas. The report was prepared based on a compilation of written material on the various topics that was organized by the editors. The editors also created material that would provide a wider view of the field as well as linkage between the topic areas.

Chapters 3 and 4 of the full report, which are Sections 2 and 3 of this article, dealt with the current status of statistics and the core of statistics. They were constructed based on speaking notes prepared by D. R. Cox and Iain Johnstone. Special thanks go to Iain Johnstone for the many figures in these chapters.

Members of the Scientific Committee who were designated as leaders for the various topic areas were in charge of preparing individual reports based on their committee's input. In particular, Section 5.1 of the full report was on biological sciences and was directed by David Siegmund; Section 5.2, engineering and industry, by Sam Hedayat; Section 5.3, geophysical and environmental sciences, by Larry Brown; and Section 5.4, information technology, by Grace Wahba. In addition, Bradley Efron and Eric Feigelson contributed material to Section 5.5 on the physical sciences. These sections are numbered 4.1–4.5 in this article.

Jon Kettenring prepared Chapter 6, found here as Section 5. Since education was not one of the themes of the workshop itself, material for this chapter was derived from secondary sources. Special thanks go to David Moore and Dick Schaeffer for their assistance in this area. Jon was also responsible for the material in the summary section of this article. (It was the executive summary in the full report.)

In addition, many helpful suggestions have been received since the material was first put into draft form, both from workshop participants and others. The report has been greatly improved as a result of their efforts.

REFERENCES

BREIMAN, L. (2001). Statistical modeling: The two cultures (with discussion). *Statist. Sci.* **16** 199–231.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

FELDMAN, G. J. and COUSINS, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57** 3873–3889.

HALL, P. and TITTERINGTON, D. M. (1987). Common structure of techniques for choosing smoothing parameters in regression problems. *J. Roy. Statist. Soc. Ser. B* **49** 184–198.

MILLER, C. J., NICHOL, R. C. and BATUSKI, D. J. (2001). Acoustic oscillations in the early universe and today. *Science* **292** 2302–2303.

MILLER, C. J. et al. (2001). Controlling the false-discovery rate in astrophysical data analysis. *Astronomical J.* **122** 3492–3505.

NATIONAL RESEARCH COUNCIL (1996). *Statistical Software Engineering*. National Academies Press, Washington.

NATIONAL SCIENCE FOUNDATION (1998). Report of the senior assessment panel of the international assessment of the U.S. mathematical science. Report 98-95, National Science Foundation, Arlington, VA.

RAFTERY, A. E., TANNER, M. A. and WELLS, M. T., eds. (2002). *Statistics in the 21st Century*. Chapman and Hall/CRC Press, London.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754.

STENSETH, N. C., FALCK, W., BJØRNSTAD, O. N. and KREBS, C. J. (1997). Population regulation in snowshoe hare and Canadian lynx: Asymmetric food web configurations between hare and lynx. *Proc. Natl. Acad. Sci. USA* **94** 5147–5152.

STENSETH, N. C. et al. (1998). From patterns to processes: Phase and density dependencies in the Canadian lynx cycle. *Proc. Natl. Acad. Sci. USA* **95** 15430–15435.

STENSETH, N. C. et al. (2004a). Snow conditions may create an invisible barrier for lynx. *Proc. Natl. Acad. Sci. USA* **101** 10632–10634.

STENSETH, N. C. et al. (2004b). The effect of climatic forcing on population synchrony and genetic structuring of the Canadian lynx. *Proc. Natl. Acad. Sci. USA* **101** 6056–6061.

WEGMAN, E. J. (1995). Huge data sets and the frontiers of computational feasibility. *J. Comput. Graph. Statist.* **4** 281–295.

Comment

Peter Bickel

As a member of the committee charged with initiating this report, but who, for various reasons, was unable to do much, I know how much effort went into the workshop that led to this report and its subsequent putting together. The profession owes a great debt to Jon, Bruce and David, as well as to Marianthi Markatou and John Stufken, who proposed this workshop. Having said this, I have to admit that I am uncomfortable with the abridged report as a presentation of the scope of statistics, as opposed to a well-justified appeal for a larger role for statistics at the NSF. The focus on the latter goal led to the elimination of an extensive discussion of biostatistics, which, as the authors admit, is perhaps the biggest field of employment and of much research for statisticians, as well as discussions about the very important fields connected with public policy such as sampling and econometric modeling. These fields are being just as overwhelmed by the data flood as the rest of society. That questions arising from these areas will not play a major role in our future seems doubtful to me. For instance, statistics of social networks—questions of sampling in graphs—are playing, and I expect will continue to play, a major role in security concerns as well as presenting major sources of novel problems. I do not know how it could have been done, given space limitations, but I would have liked to see more detailed connections drawn between the core and the various subject matter areas, and between the subject matter areas themselves such as the one above or between truncation and censoring in survival analysis, reliability theory, astronomy and other examples. I also believe the connections to the other mathematical sciences are understated. For instance, compression is an idea that, I believe, comes from and is highly developed in information theory. The tools needed to analyze graphs have been studied by graph theorists and specialists in discrete mathematics for some time; optimization, efficient algorithm construction, data base construction, not to mention both practical and theoretical aspects of machine learning, are all the province of computer scientists as well. I would agree that what has made statistics so exciting is the growing ubiquity of stochastic models in almost all fields of human endeavor. As inheritors of the traditions of formulating and analyzing such models, in fact of thinking in these ways, we have a lot to bring to the table and an important educational responsibility. However, to really participate, we need to

Peter Bickel is Professor, Department of Statistics, University of California, Berkeley, CA 94720-3860, USA (e-mail: bickel@stat.berkeley.edu).