The Role of Statistics in the Data Revolution?

Jerome H. Friedman
Stanford University
(jhf@stat.stanford.edu)

The theme of The 29th Symposium on the Interface (May 1997, Houston, TX) was "Data Mining and the analysis of large data sets". It is perhaps a coincidence that almost exactly 20 years before a "Conference on the Analysis of Large Complex Data Sets" was held in neighboring Dallas, organized by Leo Breiman, and sponsored by the ASA and IMS(!). It seems appropriate now, over 20 years later, to ask "How far have we come since 1977? In some respects things have changed a lot, in others not very much.

There has been considerable change in the nature of data. In 1977 large and complex data sets were fairly rare, and little need was seen to attempt to analyze those few that did exist. This of course has changed. The computer revolution over the last 20 years has completely altered the economics of data collection. Twenty years ago most data was still collected manually. The cost of collecting it was proportional to the amount collected. This made the cost of collecting large amounts prohibitively expensive. The goal was to carefully design experiments so that maximal information could be obtained with the fewest possible measurements.

Now much (if not most) data is automatically recorded with computers. There is a very high initial cost associated with purchasing the hardware, and especially developing the software, that is incurred before any data at all is taken. After the system is set up and working, the incremental expense of taking the data is only proportional to the cost of the magnetic medium on which it is recorded. This cost has been exponentially decreasing with time. Thus one tries to record as much data as possible to amortize the high initial set–up cost. The result has been the proliferation of very large data bases in terms of both the number of observations, and the number of measurements recorded for each one.

The idea of learning from data has been around for a long time. So it is reasonable to ask why the interest in analyzing these large and complex data sets has recently become so intense. The principal reason is that the field of Data Base Management has become involved. Data, especially large amounts of it, reside in data base management systems (DBMS). Conventional DBMS are focused on on–line transaction processing (OLTP); that is, the storage and fast retrieval of individual records for purposes of data organization. They are used to keep track of inventory, pay–roll records, billing records, invoices, etc. The process of analyzing such data for purposes other than that for which it was collected has become know as Data Mining (DM). To many in the information sciences, this term has replaced "Statistics" as being synonymous with data analysis.

Hardware and software vendors have been attempting to capitalize on the current publicity (hype) surrounding DM by quickly bringing to market products for analysis of large and complex data sets. The statistical analysis procedures provided by these DM packages nearly always include: decision tree induction (C4.5, CART, CHAID), rule induction (AQ, CN2, Recon, etc.), nearest neighbors ("case based reasoning"), clustering methods ("data segmentation"), association rules ("market basket analysis"), feature extraction, and data visualization. In addition, some include: neural networks, Bayesian belief networks ("graphical models"), genetic algorithms, self-organizing maps, and neuro-fuzzy systems.

Almost *none* of these DM packages offer: hypothesis testing, experimental design, response surface modeling, ANOVA, MANOVA, etc., linear regression, discriminant analysis, logistic regression, GLM, canonical correlation, principal components, factor analysis. These latter procedures are of course the main–stay of our standard statistical packages. Thus, nearly all of the methodology currently being marketed (and used) for DM has been developed and promoted in fields other than Statistics. Our core methodology has largely been ignored.

Even if one were to grant the intellectual viability of DM methodological development, the issue remains as to whether Statistics as a discipline should be concerned with it. Should we consider it part of our field? What does that mean? At a minimum it means that we should: publish articles about it in our journals, teach its practice in our undergraduate programs, teach relevant research topics in our graduate programs, provide recognition (jobs, tenure, awards) for those who do it well.

The answer is not obvious. One can catalog a long history of Statistics (as a field) ignoring useful methodology developed in other data related fields. Here are some of them along with their associated fields. The "*" labels those that had seminal beginnings in Statistics but for the most part were subsequently ignored in our field: pattern recognition* (CS / Engineering), data base management (CS / Library Science), neural networks* (Psychology / CS / Engineering), machine learning* (CS / AI), graphical models* (CS / AI), genetic programming (CS / Engineering), chemometrics* (Chemistry), and data visualization* (CS / Scientific Computing). To be sure, individual *statisticians* have contributed to many of these areas, but it is fair to say that they have not been embraced (at least with enthusiasm) by our field.

Since all of the above topics involve learning from data it is natural to ask why our field has remained so aloof from them. One reason often given is "That's not *statistics*". If being data related is not a sufficient reason for a topic to be considered part of our discipline, then what other qualifications are required? The answer so far seems to be that Statistics is being defined in terms of a set of *tools*, namely those currently being taught in our graduate programs. A few examples are: probability theory, real analysis, measure theory, asymptotics, decision theory, markov chains, martingales, ergotic theory, etc...

The field of Statistics seems to be defined as the set of problems that can be successfully addressed with these and related tools. It is clear that these tools have served (and continue to serve) us very well. As Brad Efron reminds us: "Statistics has been the most successful information science." and "Those who ignore Statistics are condemned to reinvent it."

One view recognizes that while the amount of data (and related applications) continue to grow exponentially, the number of statisticians is not growing that fast. Therefore our

field should concentrate that small part of information science that we do best, namely probabilistic inference based on mathematics. This is a highly defensible point of view that may well turn out to be the best strategy for our field. However, if adopted, we should become resigned to the fact that the roll of Statistics as a player in the "information revolution" will steadily diminish over time. This strategy has the strong advantage that it requires relatively little change to our current practice and academic programs.

Another point of view, advocated as early as 1962 by John Tukey (*Ann. Statist.* **33**, 1-67), holds that Statistics ought to be concerned with data analysis. The field should be defined in terms of a set of *problems* (as are most fields) rather than a set of tools, namely those problems that pertain to data. Should this point of view ever become the dominant one, a big change would be required in our practice and academic programs.

First (and foremost) we would have to make peace with computing. It is here to stay; that's where the data is. Computing has been one of the most glaring omissions in the set of tools that have so far defined Statistics. Had we incorporated computing methodology from its inception as a fundamental statistical tool (as opposed to simply a convenient way to apply our existing tools) many of the other data related fields would not have needed to exist. They would have been part of our field.

Coming to grips with computing means more than simply becoming conversant with statistical packages, although that is quite important. If computing is to become one of our fundamental research tools we will have to teach, or be sure that our students learn, the relevant Computer Science topics. These include numerical linear algebra, numerical and combinatorial optimization, data structures, algorithm design, machine architecture, programming methodology, data base management, parallel architectures and programming, etc. We will also have to expand our curriculum to include current computer oriented data analysis methodology, much of which has been developed outside our field.

If we are to compete with other data related fields in the academic (and commercial) marketplace, some of our basic paradigms will have to be modified. We may have to moderate our romance with mathematics. Mathematics (like computing) is a tool, a very powerful one to be sure, but not the only one that can be used to validate statistical methodology. Mathematics is not equivalent to theory, nor vice versa. Theories are intended to create understanding and mathematics, although quite valuable, is not the only way to do this. For example, the germ theory of disease (in and of itself) has little mathematical content, but it leads to considerable understanding of much medical phenomena. We will have to recognize that empirical validation, although necessarily limited (as is mathematics), does constitute a form of validation.

We may also have to modify our culture. Any statistician who has worked in other data related fields is struck by their "culture gap" with statistics. In these other fields the "currency" tends to be *ideas* rather than mathematical technique. Heuristically motivated ideas are initially evaluated on the merits of their heuristic arguments. Final value judgements are postponed until more thorough validation (theoretical or empirical) becomes available. The paradigm is "innocent until proven guilty" as opposed to the opposite one applied in our field. In the past we have tended to denigrate, or at best refused to accept, new methodology until it was completely validated using (preferably challenging) mathematics. This may have made sense years ago when all data sets were small and noise to signal

was high. This is a less viable strategy in many present day data analytic contexts. In particular, we may have to moderate our tendency to disregard developments (especially in other fields) that appear to work well, simply because the reasons for their success are not yet well understood by us.

Perhaps more than at anytime in the past Statistics is at a crossroads; we can decide to accommodate or resist change. As noted above, there are highly persuasive arguments for both points of view. Although opinions abound, no one knows for sure which strategy will best insure the health and viability of our field. Most statisticians seem to agree that Statistics is becoming relatively less influential among the information sciences. There tends to be less agreement as to what (if anything) should be done about it. The dominant perspective seems to be that we have a marketing problem; our customers and colleagues in other fields simply don't understand our value and importance. This may be the case. However, another explanation is that they do understand and have decided to go elsewhere.

In deciding whether or not to compete with other information sciences in new areas such as DM, several considerations should be taken into account. To quote Brian Joiner "Statistics has no God given right to exist". One cannot imagine a university without, for example, departments in mathematics, physics, chemistry and biology, etc. However, statistics departments are not always deemed essential. We prosper to the extent that we produce useful methodology. If data analytic techniques originating in other fields become dominant, our field will correspondingly suffer.

We are no longer the only game in town. Until recently, if one were interested in data analysis, Statistics was one of the very few (even remotely) appropriate fields in which to work. This is no longer the case. There are now many other exciting data oriented sciences that are competing with us for customers, students, jobs, and our own statisticians. If there exists a market for a new methodology it will be filled, with or without our blessing. Ignoring it will not make it go away. These fields now compete with us for the brightest students in terms of offering relevant curricula, exciting research projects, and the best jobs after graduation. Some of our prominent statisticians are becoming more interested in researching problems embraced by these other fields and prefer to publish in their journals. This "brain drain" of students and researchers away form Statistics may represent the most serious threat to the future health of our discipline.

Data Mining is an emerging discipline in a long list of other data related fields that have had their origins outside Statistics. In this case it is the Data Base Management community. In many ways this field (DM) represents the closest match to Statistics in terms of the types of problems it addresses. It is open to debate whether Statistics as a field should embrace Data Mining as a subdiscipline or leave it to the Computer Scientists. The intent of this paper is to stimulate that debate.

Over the years this discussion has been driven mainly by two leading visionaries of our field, John Tukey in his 1962 *Annals of Statistics* paper, and Leo Breiman at the 1977 Dallas conference. Over twenty years have passed since that conference. We again have the opportunity to reexamine our place among the information sciences.