Data Mining and Statistics: What's the Connection?

Jerome H. Friedman
Department of Statistics and
Stanford Linear Accelerator Center
Stanford University
Stanford, CA 94305
jhf@stat.stanford.edu

Abstract

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated exploratory data analysis of (usually) large complex data sets. In spite of (or perhaps because of) the somewhat exaggerated hype, this field is having a major impact in business, industry, and science. It also affords enormous research opportunities for new methodological developments. Despite the obvious connections between data mining and statistical data analysis, most of the methodologies used in Data Mining have so far originated in fields other than Statistics. This paper explores some of the reasons for this, and why statisticians should have an interest in Data Mining. It is argued that Statistics can potentially have a major influence on Data Mining, but in order to do so some of our basic paradigms and operating principles may have to be modified.

1 Introduction

General Disclaimer:

> The opinions expressed in this paper are those only of the author, and do not necessarily reflect the views of the editors, sponsors, Stanford University, or friends of the author.

The theme of The 29th Symposium on the Interface (May 1997, Houston, TX) is 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, 20 years later, to ask "How far have we come since 1977? Perhaps to Data Mining?" This paper addresses the following issues:

What is Data Mining?

What is Statistics?

What is the connection (if any)?

How can statisticians contribute (if at all)?

Should we want to?

2 What is Data Mining?

Data Mining (DM) is at best a vaguely defined field; its definition largely depends on the background and views of the definer. Here are some definitions taken from the DM literature:

Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. - Fayyad.

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. - Zekulin.

Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. - Ferruzza.

Data mining is the process of discovering advantageous patterns in data. - John

Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information. - Parsaye

Data Mining is ...

- Decision Trees
- Neural Networks
- Rule Induction
- Nearest Neighbors
- Genetic Algorithms
  - Mehta

Despite these somewhat lofty definitions, DM so far has been largely a commercial enterprise. As in most gold rushes of the past, the goal is to "mine the miners". The largest profits are made by selling tools to the miners, rather than in doing the actual mining. The concept of DM is used as a device to sell computer hardware and software.

Hardware manufacturers emphasize the high computational requirements associated with DM. Very large data bases must be stored and quickly accessed, and computationally intensive methodology applied to these data. This requires massive amounts of disk space and fast compute engines with large internal (RAM) memories. DM opens new markets for such hardware.

Software providers emphasize competitive edge. "Your competitor is doing it, so you had better keep up." Also emphasized is added value to existing legacy data bases. Many organizations have large transaction oriented data bases used for inventory, billing, accounting, etc. These data bases were very expensive to create and are costly to maintain. For a *relatively* small additional investment DM tools offer to discover highly profitable "nuggets" of information hidden in these data.

The goal of both hardware and software vendors has been to capitalize on the current publicity (hype) surrounding DM by quickly bringing to market DM products before the market becomes saturated. Once a company has invested $50K to $100K in a DM package, and perhaps much more in training, it is unlikely that they will buy a competing product anytime soon unless its superiority over the former can be overwhelmingly demonstrated. Examples of some current DM products (as of May 1997) are:

IBM: "Intelligent Miner"

Tandem: "Relational Data Miner"

Angoss Software: "KnowledgeSEEKER"

Thinking Machines Corporation: "DarwinTM"

NeoVista Software: "ASIC"

ISL Decision Systems, Inc.: "Clementine"

DataMind Corporation: "DataMind Data Cruncher"

Silicon Graphics: "MineSet"

California Scientific Software: "BrainMaker"

WizSoft Corporation: "WizWhy"

Lockheed Corporation: "Recon"

SAS Corporation: "SAS Enterprise Miner "

Besides these more or less "comprehensive" packages, there is a wide variety of specialized single purpose products. In addition, many consulting firms have been formed that specialize in DM. A difference between statisticians and computer scientists in this field seems to be that when a statistician has an idea he or she writes a paper; a computer scientist starts a company.

Current DM products are characterized by:

—*Attractive* GUI to:

- Data bases (query language)
- Suite of data analysis procedures.

—Windows style interface:

- Flexible convenient input
  - point and click icons and menus
  - input dialog boxes
  - diagrams to describe analyses
  - sophisticated graphical views of the output
  - a variety of data plots
  - slick graphical representations:
    trees, networks, flight simulation, etc.

—Convenient manipulation of the results.

These packages are usually directed as much towards decision makers as DM professionals.

The statistical analysis procedures provided by current 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
- Visualization

In addition, some include:

- Neural networks
- Bayesian belief networks ("graphical models")
- Genetic algorithms
- Self-organizing maps
- 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.

3 Why now? What's the rush?

The idea of learning from data has been around for a long time. So it is reasonable to ask why the interest in DM has suddenly become so intense. The principal reason is that the field of Data Base Management has recently 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.

Recently the Data Base Management community has become interested in using DBMS for "Decision Support". Such Decision Support systems (DSS) allow statistical queries from data collected for OLTP applications. For example "How many diapers did all of the stores in our chain sell last month?" A DSS requires the construction of a "Data Warehouse". Data Warehouses unify the data scattered throughout the many departments of an organization into a single centralized (usually very large $\sim$ 100 GB) data base with a common format. Sometimes smaller sub–data bases are also constructed for specialized analyses; these are called "Data Marts".

Decision Support systems are intended for "on–line analytic processing" (OLAP) and relational OLAP, called ROLAP. ROLAP is intended for "multidimensional analysis". ROLAP data bases are organized by dimension, that is logical grouping by attributes (variables). The conceptual framework is that of a "data–cube" which can be viewed as a large high–dimensional contingency table. ROLAP supports queries of the type:

- "Display total sales for sportswear departments during spring quarter, for stores in shopping malls, in large California cities"
- "Contrast this with stores in small towns."
- "Display all items for which profit margins are negative."

With ROLAP queries are issued manually by the user. The user formulates potentially relevant questions; the resulting answers may then suggest further questions, resulting in additional queries. The analysis proceeds in this manner until no more interesting questions are suggested, or until the analyst becomes tired or runs out of time. DM can be done with ROLAP but it requires a sophisticated (domain knowledge) user who (according to Parsaye) "does not sleep or age". The user must repeatedly formulate (guess) informative queries.

Data Mining may also be done by a (software) DM system that automatically searches for patterns by itself given only vague instructions from the user, and then displays important items, predictions, and/or anomalies.

- "What are the characteristics of items with negative profit margins?"
- "If we decide to market an item - predict (estimate) its profit margin."
- "Find the characteristics of all items for which one can accurately predict profit margin."

Not all very large data bases (VLDB) are commercial. Examples from science and engineering abound. These are usually associated with computer automated data collection:

- Astronomical (sky maps)
- Meteorological (weather, pollution monitoring stations)
- Satellite remote sensing
- High energy physics
- Industrial process control

These kinds of data also can profit (in principle) from Data Mining technology.

A combination of factors have recently come together to focus interest on DM. They include the emergence of very large data bases such as commercial data warehouses and computer automated data recording in science and engineering. Along with these have come advances in computer technology such as faster and bigger compute engines and parallel architectures. In combination they allow fast access to vast amounts of data, and the ability to apply computationally intensive statistical methodology to these data.

4 Is Data Mining and intellectual discipline?

The current interest in DM raises several issues to be addressed by the academic community. Although DM appears to be a viable commercial enterprise one can ask whether or not it qualifies as an intellectual discipline. Certainly there exists much important related research in Computer Science. This includes:

- Efficient computation of aggregates (ROLAP)
- Fast CUBE-BY ($X \times X$) queries
- Off-line precomputation of (selected) queries to speed–up on–line queries
- Parallel computation of on–line queries
- Direct interface of DBMS to DM algorithms
- Disk as opposed to RAM based implementations
- Parallel implementations of basic DM algorithms

From the perspective of statistical data analysis one can ask whether DM *methodology* is an intellectual discipline. So far, the answer is – not yet. DM packages implement well known procedures from the fields of machine learning, pattern recognition, neural networks, and data visualization. They emphasize "look and feel" (GUI) and the *existence* of functionality. There seems to be no real regard for performance (what's under the hood). The goal is to get to market quickly. Most academic research in this area so far has focused on incremental modifications to current machine learning methods, and the speed–up of existing algorithms.

However, in the *future* the answer is – almost surely, yes! Every time a technology increases in effectiveness by a factor of ten, one should completely rethink how to apply it. Consider the historical progression form walking to driving to flying. Each increases speed by roughly a factor of ten. However, each such purely quantitative increase has completely reoriented our thinking on the use of transportation in our society. A favorite quote of Chuck Dickens (former Director of Computing at SLAC) over the years has been " Every time computing power increases by a factor of ten we should totally rethink how and what we compute." A corollary to this might be " Every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it". Both computing power and data have increased by at least several orders of magnitude since nearly all currently used DM tools were invented. One can safely predict a big intellectual and academic (as well as commercial) future for new DM methodology.

5 Should Data Mining be part of Statistics?

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.

1. Pattern recognition* - CS / Engineering
2. Data base management - CS / Library Science
3. Neural networks* - Psychology / CS / Engineering
4. Machine Learning* - CS / AI
5. Graphical models (Bayes nets)* - CS / AI
6. Genetic programming - CS / Engineering
7. Chemometrics* - Chemistry
8. 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.

6 What is Statistics?

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."

> "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 [Tukey (1962)], 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.

7 Which way to go?

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 is the apparent perspective of our main professional organization, the American Statistical Association. In the five–year plan reported by its Strategic Planning Committee (Amstat News - Feb. 1997) there is a section on "Enhancing the reputation and health of our discipline". Three main approaches are suggested:

- Become involved pro-actively in policy issues.
- Build bridges to federal agencies.
- Promote Statistics education in K-12.

These are clearly important valuable suggestions, as far as they go, and should be pursued vigorously. However, the assumption inherit in them is that our present product is fine and that our only problem is in getting the word out. If we want Statistics to remain a relevant vital information science, one might also include suggestions on how to foster in our field a climate of innovation and change for meeting new data analytic challenges of the present and future.

Some statisticians argue that our field is in fact rapidly changing its perspective, perhaps too rapidly away from the procedures and principles that have served us so well in the past. This may be the case, but it is not obvious. As a counter example, the following is quoted from a recent (1997) editorial review of a paper submitted to the *Journal of the American Statistical Association*:

> "I am somewhat troubled by the absence of
> theoretical evidence for your procedure.
> Although JASA has, in the past, published
> papers with only simulation-based evidence,
> it is not a practice that I am comfortable with.
> For your procedure, an investigation of the
> asymptotic MSE or consistency and
> asymptotic normality should be doable"

This is not a criticism (or complement) to the JASA editors. Journal pages are a scarce resource and they have an obligation to see that only articles of the highest interest to the readers are published. It does suggest however that the primary importance of *mathematical* validation has not yet seriously diminished in our field.

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 considered 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.

If statisticians and data miners are to join together to address the data analysis challenges of the future, some DM paradigms may also require modification. The DM community may have to moderate its romance with "big". A prevailing attitude seems to be that unless an analysis involves gigabytes or terabytes of data, it cannot possibly be worthwhile. A dominant theme of many presentations, going as far back as the 1977 Dallas Conference, is "My data set is bigger than your data set." It seems to be a requirement that all of the data that has been collected must be used in every aspect of the analysis. Sophisticated procedures that cannot simultaneously "handle" data sets of such size are not considered relevant to DM.

Most DM applications routinely require data sets that are considerably larger than those that have been addressed by our traditional statistical procedures (kilobytes). However, it is often the case that the questions being asked of the data can be answered to sufficient accuracy with less than the entire (giga– or terabyte) data base. Sampling methodology, which has a long tradition in Statistics, can profitably be used to improve accuracy while mitigating computational requirements. Also, a powerful computationally intense procedure operating on a subsample of the data may in fact provide superior accuracy than a less sophisticated one using the entire data base.

8 Conclusion

Data Mining is an emerging discipline in a long list of other data related fields (Section 5) 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. Twenty years have passed since that conference. We again have the opportunity to reexamine our place among the information sciences. The DM community is meeting us half way. They scheduled their annual meeting KDD–97 at Newport Beach, CA just after the Joint Meetings in Anaheim, so as to encourage attendance by statisticians. They appreciate the importance of statistical thinking in data analysis. It is now up to us.

References

Tukey, J. W. (1962). The future of data analysis. *Ann. Statist.* **33**, 1-67.