Data Mining: Statistics and More?

David J. HAND

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. It is concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners. New problems arise, partly as a consequence of the sheer size of the data sets involved, and partly because of issues of pattern matching. However, since statistics provides the intellectual glue underlying the effort, it is important for statisticians to become involved. There are very real opportunities for statisticians to make significant contributions.

KEY WORDS: Databases; Exploratory data analysis; Knowledge discovery.

1. DEFINITION AND OBJECTIVES

The term *data mining* is not new to statisticians. It is a term synonymous with *data dredging* or *fishing* and has been used to describe the process of trawling through data in the hope of identifying patterns. It has a derogatory connotation because a sufficiently exhaustive search will certainly throw up patterns of some kind—by definition data that are not simply uniform have differences which can be interpreted as patterns. The trouble is that many of these "patterns" will simply be a product of random fluctuations, and will not represent any underlying structure. The object of data analysis is not to model the fleeting random patterns of the moment, but to model the underlying structures which give rise to consistent and replicable patterns. To statisticians, then, the term data mining conveys the sense of naive hope vainly struggling against the cold realities of chance.

To other researchers, however, the term is seen in a much more positive light. Stimulated by progress in computer technology and electronic data acquisition, recent decades have seen the growth of huge databases, in fields ranging from supermarket sales and banking, through astronomy, particle physics, chemistry, and medicine, to official and governmental statistics. These databases are viewed as a resource. It is certain that there is much valuable information in them, information that has not been tapped, and data mining is regarded as providing a set of tools by which that information may be extracted. Looked at in this positive light, it is hardly surprising that the commercial, industrial, and economic possibilities inherent in the notion of extracting information from these large masses of data have attracted considerable interest. The interest in the field is demonstrated by the fact that the Third International Conference on Knowledge Discovery and Data Mining, held in 1997, attracted around 700 participants.

Superficially, of course, what we are describing here is nothing but exploratory data analysis, an activity which has been carried out since data were first analyzed and which achieved greater respectability through the work of John Tukey. But there is a difference, and it is this difference that explains why statisticians have been slow to latch on to the opportunities. This difference is the sheer size of the data sets now available. Statisticians have typically not concerned themselves with data sets containing many millions or even billions of records. Moreover, special storage and manipulation techniques are required to handle data collections of this size—and the database technology which has grown up to handle them has been developed by entirely different intellectual communities from statisticians.

It is probably no exaggeration to say that most statisticians are concerned with *primary* data analysis. That is, the data are collected with a particular question or set of questions in mind. Indeed, entire subdisciplines, such as experimental design and survey design, have grown up to facilitate the efficient collection of data so as to answer the given questions. Data mining, on the other hand, is entirely concerned with *secondary* data analysis. In fact we might define data mining as *the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners.* We see from this that data mining is very much an *inductive* exercise, as opposed to the hypothetico-deductive approach often seen as the paradigm for how modern science progresses (Hand in press).

Statistics as a discipline has a poor record for timely recognition of important ideas. A common pattern is that a new idea will be launched by researchers in some other discipline, will attract considerable interest (with its promise often being subjected to excessive media hype—which can sometimes result in a backlash), and only then will statisticians become involved. By which time, of course, the intellectual proprietorship—not to mention large research grants—has gone elsewhere. Examples of this include work on pattern recognition, expert systems, genetic algorithms, neural networks, and machine learning. All of these might legitimately be regarded as subdisciplines of statistics, but they are not generally so regarded. Of course, statisticians have later made very significant advances in all of these fields, but the fact that the perceived natural home of these areas lies not in statistics but in other areas is demonstrated

David J. Hand is Professor of Statistics, Department of Statistics, The Open University, Milton Keynes, MK7 6AA, United Kingdom (E-mail: d.j.hand@open.ac.uk).

by the key journals for these areas—they are not statistical journals.

Data mining seems to be following this pattern. For the health of the discipline of statistics as a whole it is important, perhaps vital, that we learn from previous experience. Unless we do, there is a real danger that statistics—and statisticians—will be perceived as a minor irrelevance, and as not playing the fundamental role in scientific and wider life that they properly do. There is an urgency for statisticians to become involved with data mining exercises, to learn about the special problems of data mining, and to contribute in important ways to a discipline that is attracting increasing attention from a broad spectrum of concerns.

In Section 2 of this article we examine some of the major differences in emphasis between statistics and data mining. In Section 3 we look at some of the major tools, and Section 4 concludes.

2. WHAT'S NEW ABOUT DATA MINING?

Statistics, especially as taught in most statistics texts, might be described as being characterized by data sets which are small and clean, which permit straightforward answers via intensive analysis of single data sets, which are static, which were sampled in an iid manner, which were often collected to answer the particular problem being addressed, and which are solely numeric. None of these apply in the data mining context.

2.1 Size of Data Sets

For example, to a classically trained statistician a large data set might contain a few hundred points. Certainly a data set of a few thousand would be large. But modern databases often contain millions of records. Indeed, nowadays gigabyte or terabyte databases are by no means uncommon. Here are some examples. The American retailer Wal-Mart makes over 20 million transactions daily (Babcock 1994). According to Cortes and Pregibon (1997) AT&T has 100 million customers, and carries 200 million calls a day on its long-distance network. Harrison (1993) said that Mobil Oil aims to store over 100 terabytes of data concerned with oil exploration. Fayyad, Djorgovski, and Weir (1996) described the Digital Palomar Observatory Sky Survey as involving three terabytes of data, and Fayyad, Piatetsky-Shapiro, and Smyth (1996) said that the NASA Earth Observing System is projected to generate on the order of 50 gigabytes of data per hour around the turn of the century. A project of which most readers will have heard, the human genome project, has already collected gigabytes of data. Numbers like these clearly put into context the futility of standard statistical techniques. Something new is called for.

Data sets of these sorts of sizes lead to problems with which statisticians have not typically had to concern themselves in the past. An obvious one is that the data will not all fit into the main memory of the computer, despite the recent dramatic increases in capacity. This means that, if all of the data is to be processed during an analysis, adaptive or sequential techniques have to be developed. Adaptive and sequential estimation methods have been of more central concern to nonstatistical communities—especially to those working in pattern recognition and machine learning.

Data sets may be large because the number of records is large or because the number of variables is large. (Of course, what is a record in one situation may be a variable in another—it depends on the objectives of the analysis.) When the number of variables is large the curse of dimensionality really begins to bite—with 1,000 binary variables there are of the order of 10^{300} cells, a number which makes even a billion records pale into insignificance.

The problem of limited computer memory is just the beginning of the difficulties that follow from large data sets. Perhaps the data are stored not as the single flat file so beloved of statisticians, but as multiple interrelated flat files. Perhaps there is a hierarchical structure, which does not permit an easy scan through the entire data set. It is possible that very large data sets will not all be held in one place, but will be distributed. This makes accessing and sampling a complicated and time-consuming process. As a consequence of the structured way in which the data are necessarily stored, it might be the case that straightforward statistical methods cannot be applied, and stratified or clustered variants will be necessary.

There are also more subtle issues consequent on the sheer size of the data sets. In the past, in many situations where statisticians have classically worked, the problem has been one of lack of data rather than abundance. Thus, the strategy was developed of fixing the Type I error of a test at some "reasonable" value, such as 1%, 5%, or 10%, and collecting sufficient data to give adequate power for appropriate alternative hypotheses. However, when data exists in the superabundance described previously, this strategy becomes rather questionable. The results of such tests will lead to very strong evidence that even tiny effects exist, effects which are so minute as to be of doubtful practical value. All research questions involve a background level of uncertainty (of the precise question formulation, of the definitions of the variables, of the precision of the observations, of the way in which the data was drawn, of contamination, and so on) and if the effect sizes are substantially less than these other sources, then, no matter how confident one is in their reality, their value is doubtful. In place of statistical significance, we need to consider more carefully substantive significance: is the effect important or valuable or not?

2.2 Contaminated Data

Clean data is a necessary prerequisite for most statistical analyses. Entire books, not to mention careers, have been created around the issues of outlier detection and missing data. An ideal solution, when questionable data items arise, is to go back and check the source. In the data mining context, however, when the analysis is necessarily secondary, this is impossible. Moreover, when the data sets are large, it is practically certain that some of the data will be invalid in some way. This is especially true when the data describe human interactions of some kind, such as marketing data, financial transaction data, or human resource data. Contamination is also an important problem when large data sets, in which we are perhaps seeking weak relationships, are involved. Suppose, for example, that one in a thousand

records have been drawn from some distribution other than that we believe they have been drawn from. One-tenth of 1% of the data from another source would have little impact in conventional statistical problems, but in the context of a billion records this means that a million are drawn from this distribution. This is sufficient that they cannot be ignored in the analysis.

2.3 Nonstationarity, Selection Bias, and Dependent Observations

Standard statistical techniques are based on the assumption that the data items have been sampled independently and from the same distribution. Models, such as repeated measures methods, have been and are being developed for certain special situations when this is not the case. However, contravention of the idealized iid situation is probably the norm in data mining problems. Very large data sets are unlikely to arise in an iid manner; it is much more likely that some regions of the variable space will be sampled more heavily than others at different times (for example, differing time zones mean that supermarket transaction or telephone call data will not occur randomly over the whole of the United States). This may cast doubt on the validity of standard estimates, as well as posing special problems for sequential estimation and search algorithms.

Despite their inherent difficulties, the data acquisition aspects are perhaps one of the more straightforward to model. More difficult are issues of nonstationarity of the population being studied and selection bias. The first of these, also called population drift (Taylor, Nakhaeizadeh, and Kunisch 1997; Hand 1997), can arise because the underlying population is changing (for example, the population of applicants for bank loans may evolve as the economy heats and cools) or for other reasons (for example, gradual distortion creeping into measuring instruments). Unless the time of acquisition of the individual records is date-stamped, changing population structures may be undetectable. Moreover, the nature of the changes may be subtle and difficult to detect. Sometimes the situation can be even more complicated than the above may imply because often the data are dynamic. The Wal-Mart transactions or AT&T phone calls occur *every* day, not just one day, so that the database is a constantly evolving entity. This is very different from the conventional statistical situation. It might be necessary to process the data in real time. The results of an analysis obtained in September, for what happened one day in June may be of little value to the organization. The need for quick answers and the size of the data sets also lead to tough questions about statistical algorithms.

Selection bias—distortion of the selected sample away from a simple random sample—is an important and underrated problem. It is ubiquitous, and is not one which is specific to large data sets, though it is perhaps especially troublesome there. It arises, for example, in the choice of patients for clinical trials induced by the inclusion/exclusion criteria; can arise in surveys due to nonresponse; and in psychological research when the subjects are chosen from readily available people, namely young and intelligent students. In general, very large data sets are likely to have been subjected to selection bias of various kinds—they are likely to be *convenience* or *opportunity* samples rather than the statisticians' idealized random samples. Whether selection bias matters or not depends on the objectives of the data analysis. If one hopes to make inferences to the underlying population, then any sample distortion can invalidate the results. Selection bias can be an inherent part of the problem: it arises when developing scoring rules for deciding whether an applicant to be a mail order agent is acceptable. Typically in this situation comprehensive data is available only for those previous applicants who were graded "good risk" by some previous rule. Those graded "bad" would have been rejected and hence their true status never discovered. Likewise, of people offered a bank loan, comprehensive data is available only for those who take up the offer. If these are used to construct the models, to make inferences about the behavior of future applicants, then errors are likely to be introduced. On a small scale, Copas and Li (1997) describe a study of the rate of hospitalization of kidney patients given a new form of dialysis. A plot shows that the log-rate decreases over time. However, it also shows that the numbers assigned to the new treatment change over time. Patients not assigned to the new treatment were assigned to the standard one, and the selection was not random but was in the hands of the clinician, so that doubt is cast on the argument that log-rate for the new treatment is improving. What is needed to handle selection bias, as in the case of population drift, is a larger model that also takes account of the sample selection mechanism. For the large data sets that are the focus of data mining studies—which will generally also be complex data sets and for which sufficient details of how the data were collected may not be available—this will usually not be easy to construct.

2.4 Finding Interesting Patterns

The problems outlined previously show why the current statistical paradigm of intensive "hand" analysis of a single data set is inadequate for what faces those concerned with data mining. With a billion data points, even a scatterplot may be useless. There is no alternative to heavy reliance on computer programs set to discover patterns for themselves, with relatively little human intervention. A nice example was given by Fayyad, Djorgovski, and Weir (1996). Describing the crisis in astronomy arising from the huge quantities of data which are becoming available, they say: "We face a critical need for information processing technology and methodology with which to manage this data avalanche in order to produce interesting scientific results quickly and efficiently. Developments in the fields of Knowledge Discovery in Databases (KDD), machine learning, and related areas can provide at least some solutions. Much of the future of scientific information processing lies in the creative and efficient implementation and integration of these methods." Referring to the Second Palomar Observatory Sky Survey, the authors estimate that there will be at least 5×10^7 galaxies and 2×10^9 stellar objects detectable. Their aim is "to enable and maximize the extraction of meaningful information from such a large database in an efficient and timely manner" and they note that "reducing the images to

catalog entries is an overwhelming task which inherently requires an automated approach."

Of course, it is not possible simply to ask a computer to "search for interesting patterns" or to "see if there is any structure in the data." Before one can do this one needs to define what one means by patterns or structure. And before one can do that one needs to decide what one means by "interesting." Klösgen (1996, p. 252) characterized interestingness as multifaceted: "*Evidence* indicates the significance of a finding measured by a statistical criterion. *Redundancy* amounts to the similarity of a finding with respect to other findings and measures to what degree a finding follows from another one. *Usefulness* relates a finding to the goals of the user. *Novelty* includes the deviation from prior knowledge of the user or system. *Simplicity* refers to the syntactical complexity of the presentation of a finding, and *generality* is determined by the fraction of the population a finding refers to." In general, of course, what is of interest will depend very much on the application domain.

When searching for patterns or structure a compromise needs to be made between the specific and the general. The essence of data mining is that one does not know precisely what sort of structure one is seeking, so a fairly general definition will be appropriate. On the other hand, too general a definition will throw up too many candidate patterns. In *market basket analysis* one studies conditional probabilities of purchasing certain goods, given that others are purchased. One can define potentially interesting patterns as those which have high conditional probabilities (termed *confidence* in market basket analysis) as well as reasonably large marginal probabilities for the conditioning variables (termed *support* in market basket analysis). A computer program can identify all such patterns with values over given thresholds and present them for consideration by the client.

In the market basket analysis example the existing database was analyzed to identify potentially interesting patterns. However, the objective is not simply to characterize the existing database. What one really wants to do is, first, to make inferences to future likely co-occurrences of items in a basket, and, second and ideally, to make causal statements about the patterns of purchases: if someone can be persuaded to buy item A then they are also likely to buy item B. The simple marginal and conditional probabilities are insufficient to tell us about causal relationships—more sophisticated techniques are required.

Another illustration of the need to compromise between the specific and the general arises when seeking patterns in time series, such as arise in patient monitoring, telemetry, financial markets, traffic flow, and so on. Keogh and Smyth (1997) describe telemetry signals from the Space Shuttle: about 20,000 sensors are measured each second, with the signals from missions that may last several days accumulating. Such data are especially valuable for fault detection. One of the difficulties with time series pattern matching is potential nonlinear transformation of the time scale. By allowing such transformations in the pattern to be matched, one generalizes—but overdoing such generalization will make the exercise pointless. Familiarity with the problem domain and a willingness to try ad hoc approaches seems essential here.

2.5 Nonnumeric Data

Finally, classical statistics deals solely with numeric data. Increasingly nowadays, databases contain data of other kinds. Four obvious examples are image data, audio data, text data, and geographical data. The issues of data mining—of finding interesting patterns and structures in the data—apply just as much here as to simple numerical data. Mining the internet has become a distinct subarea of data mining in its own right.

2.6 Spurious Relationships and Automated Data Analysis

To statisticians, one thing will be immediately apparent from the previous examples. Because the pattern searches will throw up a large numbers of candidate patterns, there will be a high probability that spurious (chance) data configurations will be identified as patterns. How might this be dealt with? There are conventional multiple comparisons approaches in statistics, in which, for example, the overall experimentwise error is controlled, but these were not designed for the sheer numbers of candidate patterns generated by data mining. This is an area which would benefit from some careful thought. It is possible that a solution will only be found by stepping outside the conventional probabilistic statistical framework—possibly using scoring rules instead of probabilistic interpretations. The problem is similar to that of overfitting of statistical models, an issue which has attracted renewed interest with the development of extremely flexible models such as neural networks. Several distinct but related strategies have been developed for easing the problem, and it may be possible to develop analogous strategies for data mining. These strategies include restricting the family of models (c.f. limiting the size of the class of patterns examined), optimizing a penalized goodness-of-fit function (c.f. penalizing the patterns according to the size of the set of possible patterns satisfying the criteria), and shrinking an overfitted model (c.f. imposing tougher pattern selection criteria). Of course, the bottom line is that those patterns and structures identified as potentially interesting will be presented to a domain expert for consideration—to be accepted or rejected in the context of the substantive domain and objectives, and not merely on the basis of internal statistical structure.

It is probably legitimate to characterize some of the analysis undertaken during data mining as *automatic data analysis*, since much of it occurs outside the direct control of the researcher. To many statisticians this whole notion will be abhorrent. Data analysis is as much an art as a science. However, the imperatives of the sheer volume of data mean that we have no choice. In any case, the issue of where human data analysis stops and automatic data analysis begins is a moot point. After all, even standard statistical tools use extensive search as part of the model-fitting process—think of variable selection in regression and of the search involved in constructing classification trees.

In the 1980s a flurry of work on automatic data analysis occurred under the name of statistical expert systems re-

search (a review of such work was given by Gale, Hand, and Kelly 1993). These were computer programs that interacted with the user and the data to conduct valid and accurate statistical analyses. The work was motivated by a concern about misuse of increasingly powerful and yet easy to use statistical packages. In principle, a statistical expert system would embody a large base of intelligent understanding of the data analysis process, which it could apply automatically (to a relatively small set of data, at least in data mining terms). Compare this with a data mining system, which embodies a small base of intelligent understanding, but which applies it to a large data set. In both cases the application is automatic, though in both cases interaction with the researcher is fundamental. In a statistical expert system the program drives the analysis following a *statistical strategy* because the user has insufficient statistical expertise to do so. In a data mining application, the program drives the analysis because the user has insufficient resources to manually examine billions of records and hundreds of thousands of potential patterns. Given these similarities between the two enterprises, it is sensible to ask if there are lessons which the data mining community might learn from the statistical expert system experience. Relevant lessons include the importance of a well-defined potential user population. Much statistical expert systems research went on in the abstract ("let's see if we can build a system which will do analysis of variance"). Little wonder that such systems vanished without trace, when those who might need and make use of such a system had not been identified beforehand. A second lesson is the importance of sufficiently broad system expertise—a system may be expert at one-way analysis of variance (or identifying one type of pattern in data mining), but, given an inevitable learning curve, a certain frequency of use is necessary to make the system valuable. And, of course, from a scientific point of view, it is necessary to formulate beforehand a criterion by which success can be judged. It seems clear that to have an impact, research on data mining systems should be tied into real practical applications, with a clear problem and objective specification.

3. METHODS

In the previous sections I have spoken in fairly general terms about the objective of data mining as being to find patterns or structure in large data sets. However, it is sometimes useful to distinguish between two classes of data mining techniques, which seek, respectively, to find *patterns* and *models*. The position of the dividing line between these is rather arbitrary. However, to me a model is a global representation of a structure that summarizes the systematic component underlying the data or that describes how the data may have arisen. The word "global" here signifies that it is a comprehensive structure, referring to many cases. In contrast, a pattern is a local structure, perhaps relating to just a handful of variables and a few cases. The market basket associations mentioned previously illustrate such patterns: perhaps only a few hundred of the many baskets demonstrate a particular pattern. Likewise, in the time series example, if one is searching for patterns the objective is not to construct a global model, such as a Box–Jenkins model, but rather to locate structures that are of relatively short duration—the patterns sought in technical analysis of stock market behavior provide a good illustration.

With this distinction we can identify two types of data mining method, according to whether they seek to build models or to find patterns. The first type, concerned with building global models is, apart from the problems inherent from the sizes of the data sets, identical to conventional exploratory statistical methods. It was such "traditional" methods, used in a data mining context, which led to the rejection of the conventional wisdom that a portfolio of long-term mortgage customers is a good portfolio: in fact such customers may be the ones who have been unable to find a more attractive offer elsewhere—the less good customers. Models for both prediction and description occur in data mining contexts—for example, description is often the aim with scientific data while prediction is often the aim with commercial data. (Of course, again there is overlap. I am not intending to imply that only descriptive models are relevant with scientific data, but simply to illustrate application domains.) A distinction is also sometimes made (Box and Hunter 1965; Cox 1990; Hand 1995) between *empirical* and *mechanistic* models. The former (also sometimes called *operational*) seek to model relationships without basing them on any underlying theory. The latter (sometimes called *substantive*, *phenomenological*, or *iconic*) are based on some theory of mechanism for the underlying data generating process. Data mining, almost by definition, is chiefly concerned with the former.

We could add a third type of model here, which might be termed *prescriptive*. These are models which do not so much unearth structure in the data as impose structure on it. Such models are also relevant in a data mining context—though perhaps the interpretation is rather different from most data mining applications. The class of techniques which generally go under the name of cluster analysis provides an example. On the one hand we have methods which seek to discover naturally occurring structures in the data—to carve nature at the joints, as it has been put. And on the other hand we have methods which seek to partition the data in some convenient way. The former might be especially relevant in a scientific context, where one may be interested in characterizing different kinds of entities. The latter may be especially relevant in a commercial context, where one may simply want to group the objects into classes which have a relatively high measure of internal homogeneity—without any notion that the different clusters really represent qualitatively different kinds of entities. Partitioning the data in this latter sense yields a prescriptive model. Parenthetically at this point, we might also note that mixture decomposition, with slightly different aims yet again, but also a data mining tool, is also sometimes included under the term cluster analysis. It is perhaps unfortunate that the term "cluster analysis" is sometimes used for all three objectives.

Methods for building global models in data mining include cluster analysis, regression analysis, supervised classification methods in general, projection pursuit, and, in-

deed, any method for summarizing data. Bayesian networks (e.g., Heckerman 1997) are also receiving a great deal of attention in the data mining context. When there are many variables the overall probability distribution is very complex, and the curse of dimensionality means that estimating individual cell probabilities (with categorical variables) is out of the question. Bayesian networks seek to summarize the overall distribution in terms of the conditional dependencies between the variables. If there are only a relatively few nonzero conditional dependencies, the result is a dramatic reduction in the number of parameters which need to be estimated (and hence a concomitant increase in precision).

If there is a huge number of records, the question arises as to whether all are necessary for the model building process. As noted previously, there may well be difficulties in accessing all the records in a reasonable time, and adaptive methods may be necessary. If the basic estimation procedure does not permit explicit solutions, so that iterative methods are normally employed, one may need to think carefully about the algorithms which are used in a context when data set size or evolving nature means iteration is impossible. An approach which is sometimes used is to build one's model on a sample of the data, perhaps updating it using the remainder when one is sure one has the correct basic structure. This, of course, is predicated on the ability to draw a *random* or, at least, representative, sample from the entire database—an operation which may not be easy or, indeed, possible.

We should also bear in mind the complexity of the required model. We noted earlier that with huge data sets it is possible to model small idiosyncrasies which are of little practical import. So, for example, in the global modeling context, just how many clusters is it worth knowing about and just how small a conditional dependence should be retained in the model? Perhaps we should restrict ourselves to the few largest.

Moving on from global models, the second class of data mining method seeks to find patterns by sifting through the data seeking co-occurrences of particular values on particular variables. It is this class of strategies which has led to the notion of data mining as seeking nuggets of information among the mass of data. Pattern finding strategies are especially useful for anomaly detection. Examples include fault detection, either during manufacturing or in operation; fraud detection, for example in applications for loans, in credit card usage, or in tax returns; and in distinguishing between genuine and spurious alarm triggers. There is obviously a relationship between pattern detection methods and diagnostics in conventional statistics, but they are not identical. One big difference is that conventional diagnostic methods need a model with which to compare the data, while simple pattern detection does not. Another is the need, in the pattern-detection context, to search through very large collections of data and, indeed, of possible shapes of pattern. This makes automatic computer search vital in data mining: it is necessary to rely on the machine's ability to search through masses of data without getting bored, tired, and without making mistakes. These new aspects to the problem mean that the area is rich for potential research effort— new problems require new solutions.

Graphical methods, especially dynamic and interactive graphical methods, also have a key role to play here. Such tools allow one to take advantage of the particular power of the human eye and mind at digesting very complex information. The dynamic graphical display known as the *World Tour*—projecting multivariate data down into a two dimensional projection and letting the direction of projection vary—is an example of this. At present such methods have their limitations. Sitting watching such a display for any length of time can be a mind-numbing experience. However, here again the computer can come to our aid. We can define measures of interestingness for a scatterplot and let the machine apply these measures as it produces the projections. We are back at projection pursuit. This, of course, requires us to articulate and define beforehand what we consider "interesting." But we can go further. We can present the machine with a series of projections, telling it which ones we find interesting and which we do not, and (provided we have given it a basic alphabet of structures) we can let it learn appropriate internal representations of "interesting" for itself.

So far, in this discussion of methods, I have only referred to methods for the statistical aspects. But I noted previously that database technology was also a fundamental leg on which the data mining enterprise stood. Just as the size of the data sets and the range of the problems mean that standard statistical methods are inadequate for the challenge of data mining, so Imielinksi and Mannila (1996) argued that standard database technology is inadequate. Although the few primitives of structured query language (SQL) are sufficient for many business applications, they are insufficient for data mining applications and new tools are needed. The term OLAP, standing for online analytical processing, is often used to describe the sorts of query-driven analysis which must be undertaken. Expressions used in OLAP include such things as rolling up (producing marginals), drilling (going down levels of aggregation—the opposite of rolling up), slicing (conditioning on one variable), and dicing (conditioning on many variables). See, for example, Gray et al. (1997). Moving up a level, the term "data warehousing" is often used. A "data warehouse" is "a database that contains subject-oriented, integrated, and historical data that is primarily used in analysis and decision support environments. Data warehousing is the process of creating a data warehouse by collecting and cleaning transactional data and making it available for online retrieval to assist in analysis and decision making." (Uthurusamy 1996). A data warehouse thus has integrated data (rather than data contained in a number of separate databases), both raw data and summarized data, historical data if the data are accumulating over time, and metadata (descriptions of the meaning and context of the data).

4. CONCLUSION

Some authors (e.g., Fayyad 1997) see data mining as a "single step in a larger process that we call the KDD

process." "KDD" here stands for *Knowledge Discovery in Databases*. Other steps in this process include (Fayyad 1997): data warehousing; target data selection; cleaning; preprocessing; transformation and reduction; data mining; model selection (or combination); evaluation and interpretation; consolidation and use of the extracted knowledge. Apart from general issues arising from data set size and particular issues concerned with pattern search, most of these steps will be familiar to statisticians.

Given the commercial interest in data mining, it is hardly surprising that a number of software tools have appeared on the market. Some are general tools, similar to powerful statistical data exploration systems, while others essentially seek to put the capacity for extracting knowledge from data in the hands of the domain expert rather than a professional data analyst. These must thus use domain terminology. An example of general tool is Explora (Hoschka and Klösgen 1991; Klösgen 1996) and examples of more domain specific tools are the Interactive Data Exploration and Analysis system of AT&T (Selfridge, Srivastava, and Wilson 1996), which permits one to segment market data and analyze the effect of new promotions and advertisements, and Advanced Scout (Bhandari et al. 1997) which seeks interesting patterns in basketball games.

The promise and opportunities of data mining are obvious—and commercial organizations have not been slow to react to this. However, parallels with other fields, such as expert systems and artificial intelligence, suggest that some caution should be exercised. Data mining is not a universal panacea and it is not without its problems and difficulties. Care must be exercised to ensure that the claims are not overinflated or there will be a very real danger of a backlash. In particular, when searching for patterns it is necessary to consider how many of those discovered are real (rather than chance fluctuations in the database), how to make valid probability statements about them (given the probably nonrandom nature of the data), and how many of them are nontrivial, interesting, and valuable. Some cost-benefit analysis of data mining exercises seems appropriate.

Beyond all this, however, lie the very real contributions that statisticians can make. We must learn from the experience in other areas, and not hesitate to become involved in what is after all, essentially a statistical problem. There are real opportunities to make major advances in tackling genuinely important problems. It would be a great loss, for the reputation of statistics as a discipline as well as for individual statisticians, if these opportunities were not grasped.

Further general reading on data mining may be found in the new journal *Data Mining and Knowledge Discovery*, a special issue of the *Communications of the ACM* on data mining (November 1996, vol. 39, no. 11), and in Fayyad, Piatetsky-Shapiro, and Smyth (1996).

[Received September 1997. Revised December 1997.]

REFERENCES

Babcock, C. (1994), "Parallel Processing Mines Retail Data," *Computer World*, 6.

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., and Ramanujam, K. (1997), "Advanced Scout: Data Mining and Knowledge Discovery in NBA Data," *Data Mining and Knowledge Discovery*, 1, 121–125.

Box, G., and Hunter, W. (1965), "The Experimental Study of Physical Mechanisms," *Technometrics*, 7, 57–71.

Copas, J.B., and Li, H.G. (1997), "Inference for Nonrandom Samples" (with discussion), *Journal of the Royal Statistical Society*, Series B, 59, 55–95.

Cortes, C., and Pregibon, D. (1997), "Mega-monitoring," unpublished paper presented at the University of Washington/Microsoft Summer Research Institute on Data Mining, July 6–11, 1997.

Cox, D.R. (1990), "Role of Models in Statistical Analysis," *Statistical Science*, 5, 169–174.

Fayyad, U. (1997), "Editorial," *Data Mining and Knowledge Discovery*, 1, 5–10.

Fayyad, U.M., Djorgovski, S.G., and Weir, N. (1996), "Automating the Analysis and Cataloging of Sky Surveys," in *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA: AAAI Press, pp. 471–493.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996), "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA: AAAI Press, pp. 1–34.

Gale, W.A., Hand, D.J., and Kelly, A.E. (1993), "Artificial Intelligence and Statistics," in *Handbook of Statistics 9: Computational Statistics*, ed. C.R. Rao, Amsterdam: North-Holland, pp. 535–576.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997), "Data Cube: A Relational Aggregration Operator Generalizing Group-by, Cross-tab, and Sub-totals," *Data Mining and Knowledge Discovery*, 1, 29–53.

Hand, D.J. (1995), "Discussion contribution to Chatfield (1995)," *Journal of the Royal Statistical Society*, Series A, 158, 448.

——— (1997), *Construction and Assessment of Classification Rules*, Chichester: Wiley.

——— (in press), "Statistics and the Scientific Method," to appear in *Encyclopedia of Biostatistics*, Chichester: Wiley.

Harrison, D. (1993), "Backing Up," *Neural Computing*, 98–104.

Heckerman, D. (1997), "Bayesian Networks for Data Mining," *Data Mining and Knowledge Discovery*, 1, 79–119.

Hoschka, P., and Klösgen, W. (1991), "A Support System for Interpreting Statistical Data," in *Knowledge Discovery in Databases*, eds. G. Piatetsky-Shapiro and W. Frawley, Menlo Park: CA: AAAI Press, pp. 325–346.

Imielinski, T., and Mannila, H. (1996), "A Database Perspective on Knowledge Discovery," *Communications of the ACM*, 39, 58–64.

Keogh, E., and Smyth, P. (1997), "A Probabilistic Approach to Fast Pattern Matching in Time Series Databases," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, pp. 24–30.

Klösgen, W. (1996), "Explora: A Multipattern and Multistrategy Discovery Assistant," in *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA: AAAI Press, pp. 249–271.

Selfridge, P.G., Srivastava, D., and Wilson, L.O. (1996), "IDEA: Interactive Data Exploration and Analysis," in *Proceedings of SIGMOD 96*, New York: ACM Press, pp. 24–34.

Taylor, C.C., Nakhaeizadeh, G., and Kunisch, G. (1997), "Statistical Aspects of Classification in Drifting Populations," in *Preliminary Papers of the 6th International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, FL: pp. 521–528.

Uthurusamy, R. (1996), "From Data Mining to Knowledge Discovery: Current Challenges and Future Directions," in *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA: AAAI Press, pp. 560–569.