10

Statistical Themes and Lessons for Data Mining

Objectives:

- Data mining is a growing discipline, which originated outside statistics in the database management community, mainly for commercial concerns.
- We examine here data mining definitions, tools, and how data mining could be used in official statistics.
- Data mining is made on the interface of computer science and statistics, utilizing advances in both disciplines to make progress in extracting information from large database.
- This chapter highlights some statistical themes and lessons that are directly relevant to data mining.
- Attempts to identify opportunities where close cooperation between the statistical and computational communities might reasonably provide synergy for further progress in data analysis.
- This section briefly describes some of the central statistical ideas we think relevant to data mining.

Abstract. Data mining is a new field at the frontiers of statistics and information technologies (database management, artificial intelligence, machine learning, etc.), which aims at discovering structures and patterns in large datasets. We examine here its definitions, tools, and how data mining could be used in official statistics.

Data mining is on the interface of computer science and statistics, utilizing advances in both disciplines to make progress in extracting information from large database. It is an emerging field that has attracted much attention in a very short period of time. This section highlights some statistical themes and lessons that are directly relevant to data mining and attempts to identify opportunities where close cooperation between the statistical and computational communities might reasonably provide synergy for further progress in data analysis.

S. Sumathi and S.N. Sivanandam: *Statistical Themes and Lessons for Data Mining*, Studies in Computational Intelligence (SCI) **29**, 243–263 (2006)
www.springerlink.com

10.1 Data Mining and Official Statistics

Data mining is often presented as a revolution in information processing. Here are two definitions taken from the literature:

U.M. Fayyad: "Data mining is the nontrivial process identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

D.J. Hand: "Data mining (consists in) the discovery of interesting, unexpected, or valuable structures in large data sets."

The development of data mining is related with the availability of very large databases and the need of exploiting these bases in a new way. As J. Friedman in 1997 pointed out, the database management community has become recently interested in using DBMS for decision support. They realized that data, which have been collected for management organization such as recording transactions, might contain useful information for, e.g., improving the knowledge of, and the service to, customers.

The metaphor of data mining means that there are treasures (or nuggets) hidden under mountains of data, which may be discovered by specific tools. Of course this is not a new idea and there has been many developments in statistical methodology, which were oriented toward the discovery of patterns or models through techniques like EDA (exploratory data analysis) and multivariate exploratory analysis (dimension reduction methods: component, correspondence and cluster analysis among many others).

10.1.1 What is New in Data Mining is:

The availability of very large databases, which makes obsolete classical definitions of large samples: billions of records, terabytes of data are not unusual. More and more data are collected automatically, the massive use of new techniques coming from the computer science community like neural networks, decision trees, induction rules; commercial interests in valorizing existing information in order to propose individual solutions to targeted customers; new software packages, more user friendly, with attractive interfaces, directed as much toward decision makers as professionals analysts, but much more expensive!

It is interesting to remark that statisticians (those who are not reluctant to this new trend?) prefer the expression *data mining*, while computer scientists use knowledge discovery, but it is more a matter of expressing a belonging to a community than a real difference.

10.1.2 Goals and Tools of Data Mining

The purpose of data mining is to find structures in data. Following D. Hand in 2000, there are two kinds of structures that are sought in data mining activities: models and patterns. Building models is a major activity of many statisticians and econometricians, especially in NSI [National Statistical Institutes], and it will not be necessary to elaborate too long on this. Let us say that a model is a global summary of relationships between variables, which both helps to understand phenomenons and allows predictions. Linear models, simultaneous equations are widely used. But a model is generally chosen on an a priori basis, based upon a simplifying theory. Exploration of alternative models, possibly nonlinear or in a nonclosed mathematical form, is made feasible by DM algorithms. DM appears then as a collection of tools presented usually in one package, in such a way that several techniques may be compared on the same dataset.

DM does not reduce to decision trees, neural networks, or genetic algorithms, as some software vendors claim. DM algorithms offers extensive possibilities of finding models relating variables together: we have mentioned neural networks, decision trees, which finds nonlinear models, graphical models (or Bayesian belief networks), etc. give a valuable representation of relations between variables.

In contrast to the global description given by a model, a pattern is often defined as a characteristic structure exhibited by a few number of points: for instance a small subgroup of customers with a high commercial value, or conversely highly risked. Tools involved here are often cluster analysis, which is a well-known statistical technique, and also Kohonen self-organizing maps, based on neural networks.

Association rule discovery, or market basket analysis, is one of the favorite and perhaps new tools of data mining: its origin is in analyzing purchases in supermarket: one is interested in the percentage of customers whose purchase simultaneously two goods, more precisely in identifying couples of goods. A, B with high conditional probabilities of purchasing A, given that B is purchased. Of course the result is interesting only if P (A/not B) is low. The problem in pattern discovery is to be sure that patterns are real and useful: the probability of finding any given pattern increases with the size of the database. It is thus necessary to develop validation rules, and in this respect processing the whole database is not the best thing to do: sampling (or splitting the base into several sets) will be better than exhaustive processing. One should test if a model or a pattern remains valid in another part of the base than the one that has been explored. Usefulness should also be tested: associations are only correlations, and not causation and it is sure that promoting B will mechanically increase the purchase of A.

10.1.3 New Mines: Texts, Web, Symbolic Data?

Besides classical database where data is usually presented in the form of a rectangular array, new kinds of data are now present: Symbolic data, such as fuzzy data or intervals: the data is not known precisely but belongs to an interval, with or without a probability distribution.

Text mining: Most of the information that circulates is now digitized as word processor documents, and powerful techniques are currently available for a wide variety of applications. For example, one can categorize information from news agencies; analyze patent portfolios, customer complaints; classify incoming e-mail according to predefined topics, group-related documents based on their content, without requiring predefined classes; or assign documents to one or more user-defined categories.

Web mining is a new concept for statistical analysis of Website information (text mining) or frequentation, as well as of the behavior of the Web surfers.

10.1.4 Applications in Official Statistics

Of course in National Statistical Institutes there has always been some use of exploratory Data analysis, or of model choice algorithms, but as far as we know it seems that there are few, if not none, known applications of data mining techniques in the meaning of trying to discover new models or patterns in their databases by using the new tools described before.

It is not surprising because the main task of NSI is data production, and analysis is often done by different institutes. Furthermore it seems that the idea of exploring a database with the objective of finding unexpected patterns or models is not familiar to official statisticians who have to answer precise questions and make forecasts. One may consider that official statisticians are experts and do not need DM, which is an automatic process for end users who do not know what is hidden under their data. Statistical analyses are done generally if they can be repeated in a production framework. But as far as NSI manage large databases on population, trade, agriculture, companies, there is certainly great potentialities in exploiting their mines of data.

Let us point out a few fields where data mining tools could be useful:

Business statistics with special mention for innovation policy, financial health household equipment and savings health statistics, more precisely mortality and morbidity, in order to detect unexpected risk factors analysis of metadata information by means of text mining.

10.2 Statistical Themes and Lessons for Data Mining

Sta-tis-tics (noun). The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling (*American Heritage Dictionary*).

Statistics is enjoying a renaissance period. Modern computing hardware and software have freed the statistician from narrowly specified models and spawned a fresh approach to the subject, especially as it relates to data analysis. Today's statistical toolkit draws on a rich body of theoretical and methodological research (Table 10.1).

Table 10.1. Statisticians have developed a large infrastructure (theory) to support their methods and a language (probability calculus) to describe their approach to quantifying the uncertainty associated with drawing inferences from data. These methods enable one to describe relationships between variables for prediction, quantifying effects, or suggesting causal paths.

Area of Statistics	Description of Activities
Experimental design and sampling	How to select cases if one has the liberty to choose.
Exploratory data analysis	Hypothesis generation rather than hypothesis testing
Statistical graphics	Data visualization
Statistical modeling	Regression and classification techniques.
Statistical inference	Estimation and prediction techniques.

The field of data mining, like statistics, concerns itself with "learning from data" or "turning data into information." The context encompasses statistics, but with a somewhat different emphasis. In particular, data mining involves retrospective analyses of data: thus, topics such as experimental design are outside the scope of data mining and fall within statistics proper. Data miners are often more interested in understandability than accuracy or predictability per se. Thus, there is a focus on relatively simple interpretable models involving rules, trees, graphs, and so forth. Applications involving very large numbers of variables and vast numbers of measurements are also common in data mining. Thus, computational efficiency and scalability are critically important, and issues of statistical consistency may be a secondary consideration. Furthermore, the current practice of data mining is often pattern focused rather than model focused, i.e., rather than building a coherent global model that includes all variables of interest, data mining algorithms (such as any of the many rule induction systems on the market) will produce sets of statements about local dependencies among variables (in rule form).

In this overall context, current data mining practice is very much driven by practical computational concerns. However, in focusing almost exclusively on computational issues, it is easy to forget that statistics is in fact a *core* component. The term *data mining* has long had negative connotations in the statistics literature. Data mining without proper consideration of the fundamental statistical nature of the inference problem is indeed to be avoided. However, a goal of this section is to convince the reader that modern statistics can offer significant constructive advice to the data miner, although many problems remain unsolved. Throughout the section we highlight some major themes of statistics research, focusing in particular on the practical lessons pertinent of data mining.

10.2.1 An Overview of Statistical Science

This section briefly describes some of the central statistical ideas we think relevant to data mining.

Probability distributions. The statistical literature contains mathematical characterizations of a wealth of probability distributions, as well as properties of random variables – functions defined on the "events" to which a probability measure assigns values. Important relations among probability distributions include marginalization (summing over a subset of values) and conditionalization (forming a conditional probability measure from a measure on a sample space and some event of positive measure). Essential relations among random variables include independence, conditional independence, and various measures of dependence, of which the most famous is the correlation coefficient. The statistical literature also characterizes families of distributions by properties that are useful in identifying any particular member of the family from data, or by closure properties useful in model construction or inference, for example conjugate families, closed under conditionalization, and the multinormal family, closed under liner combination. Knowledge of the properties of distribution families can be invaluable in analyzing data and making appropriate inference.

Estimation, consistency, uncertainty, assumptions, robustness, and model averaging. An estimator is a function from sample data to some estimated, such as the value of a parameter. When the data comprise a sample from a larger actual or a potential collection governed by some probability distribution, the family of estimators corresponding to all possible samples from that collection also has a probability distribution. Classical statistics investigates such distributions of estimators in order to establish basic properties such as reliability and uncertainty. A variety of resampling and simulation techniques also exist for assessing estimator uncertainty.

Estimation almost always requires some set of assumptions. Such assumptions are typically false, but often useful. If a model (which we can think of as a set of assumptions) is incorrect, estimates based on it can be expected to be incorrect as well. One of the aims of statistical research is to find ways to weaken the assumptions necessary for good estimation. "Robust statistics" looks for estimators that work satisfactorily for larger families of distributions and have small errors when assumptions are violated.

Bayesian estimation emphasizes that alternative models and their competing assumptions are often plausible. Rather than making an estimate based on a single model several models can be considered and an estimate obtained as the weighted average of the estimates given by the individual models. In fact, such Bayesian model averaging is bound to improve predictive performance, on average. Since the models obtained in data mining are usually the results of some automated search procedure, accounting for the potential errors associated with the search itself is crucial. In practice, this often requires a Monte

Carlo analysis. The impression is that the error rates of search procedures proposed and used in the data mining and in the statistical literature are far too rarely estimated in this way.

Hypothesis testing. Since statistical tests are widely used, some of their important limitations should be noted. Viewed as a one-sided estimation method, hypothesis testing is inconsistent unless the alpha level of testing rule is decreased appropriately as the sample size increases. Generally, an α level test of one hypothesis and an α level test of another hypothesis do not jointly provide an α level test of the conjunction of the two hypotheses. In special cases, rules (sometimes called contrasts) exist for simultaneously testing several hypotheses. An important corollary for data mining is that the α level of a test has nothing directly to do with the probability of error in a search procedure that involves testing a series of hypothesis. If, for example, for each pair of a set of variables, hypotheses of independence are tested at $\alpha = 0.05$, then 0.05 is not the probability of erroneously finding some dependent set of variables when in fact all pairs are independent. Thus, in data mining procedures that use a sequence of hypothesis tests, the alpha level of the tests cannot generally be taken as an estimate of any error probability related to the outcome of the search.

Data miners should note that while error probabilities of tests have something to do with the truth of hypotheses, the connection is somewhat tenuous. Hypotheses that are excellent approximations may be rejected in large samples; tests of linear models, for example, typically reject them in very large samples no matter how closely they seem to fit the data.

Model scoring. The evidence provided by data should lead us to prefer some models or hypotheses to others, and to be indifferent between still other models. A score is any rule that maps models and data to numbers whose numerical ordering corresponds to a preference ordering over the space of models, given the data. For the reasons just considered, scoring rules are often an attractive alternative to tests. Typical rules assign models a value determined by the likelihood function associated with the model, the number of parameters, or dimension, of the model, and the data. Popular rules include the Akaike information criterion, Bayes information criterion, and minimum description length. Given a prior probability distribution over models, the posterior probability on the data is itself a scoring function, arguably a privileged one. The Bayes information criterion approximates posterior probabilities in large samples.

There is a notion of consistency appropriate to scoring rules; in the large sample limit, almost surely the true model should be among those receiving maximal scores. AIC scores are not, in general, consistent. There are also uncertainties associated with scores, since two different samples of the same size from the same distribution may yield not only different numerical values for the same model, but even different orderings of models.

For obvious combinatorial reasons, it is often impossible when searching a large model space to calculate scores for all models; it is, however, often feasible to describe and calculate scores for a few equivalence classes of models receiving the highest scores.

In some contexts, inferences made using Bayesian scores can differ a great deal from inferences made with hypothesis tests. Raftery in 1995 gives examples of models that account for almost all of the variance of an outcome of interest, and have very high Bayesian scores, but are overwhelmingly rejected by statistical tests.

Markov Chain Monte Carlo. Historically, insurmountable computational difficulties forced data analysts to eschew exact analysis of elaborate hierarchical Bayesian models and complex likelihood calculations. Recent dramatic advances in Monte Carlo methods have, however, liberated analysts from some of these constraints. One particular class of simulation methods, dubbed Markov chain Monte Carlo, originally developed in statistical mechanics, has revolutionized the practice of Bayesian statistics. Smith and Roberts in 1993 provided an accessible overview from the Bayesian perspective; Gilks et al. in 1996 provided a practical introduction addressing both Bayesian and non-Bayesian perspectives. Simulation methods may become unacceptably slow when faced with massive datasets. In such cases, recent advances in analytic approximations prove useful.

Generalized model classes. A major achievement of statistical methodological research has been the development of very general and flexible model classes. Generalized linear models, for instance, embrace many classical linear models, and unify estimation and testing theory for such models. Generalized additive models show similar potential. Graphical models represent probabilistic and statistical models with planer graphs, where the vertices represent (possibly latent) random variables and the edges represent stochastic dependencies. This provides a powerful language for describing models and the graphs themselves make modeling assumptions explicit. Graphical models provide important bridges between the vast statistical literature on multivariate analysis and such fields as artificial intelligence, casual analysis, and data mining.

Rational decision and planning. The theory of rational choice assumes the decision maker has available a definite set of alternative actions, knowledge of a definite set of possible alternative states of the world, knowledge of the payoffs or utilities of the outcomes of each possible action in each possible state of the world, and knowledge of the probabilities of various possible states of the world. Given all of this information, a decision rule specifies which of the alternative actions ought to be taken. A large literature in statistics and economics addresses alternative decision rules – maximizing expected utility, minimizing maximum loss, etc. Typically, rational decision making and planning are the goals of data mining, and rather than providing techniques or methods for data mining, the theory of rational choice poses norms for the use of information obtained from a database.

The very framework of rational decision making requires probabilities and knowledge of the effects alternative actions will have. To know the outcomes of actions is to know something of cause-and-effect relations, and extracting such causal information is often one of the principle goals of data mining and statistical inference more generally.

Inference to causes. Understanding causation is the hidden force behind the historical development of statistics. From the beginning of the subject, in the work of Bernoulli and Laplace, the absence of causal connection between two variables has been taken to imply their probabilistic independence, and the same idea is fundamental in the theory of experimental design. Early in this century, Wright in 1921 introduced directed graphs to represent causal hypotheses (with vertices as random variables and edges representing direct influences), and they have become common representations of causal hypotheses in the social sciences, biology, computer science and engineering.

Kiiveri and Speed in 1982 combined directed graphs with a generalized connection between independence and absence of casual connection in what they called the *Markov condition*: provided Y is not an effect of X, X and Y are conditionally independent given the direct causes of X. They showed that much of the linear modeling literature tacitly assumed the Markov condition; the same is true for casual models of categorical data, and virtually all causal models of systems without feedback. Under additional assumptions, conditional independence therefore provides information about causal dependence. The most common, and most thoroughly investigated, additional assumption is that all conditional independencies are due to the Markov condition applied to the directed graph describing the actual causal processes generating the data, a requirement that has been given many names, including "faithfulness." Directed graphs with associated probability distributions satisfying the Markov condition are called by different names in different literatures: Bayes nets, belief nets, structural equation models, path models, etc. Nonetheless, causal inferences from uncontrolled convenience samples are liable to many sources of error, and data miners should proceed with extreme caution.

Sources of error peculiar to casual inference from a database include the existence of unrecorded causes of associations (confounders) between recorded variables, associations among recorded variables produced by the influence such variables themselves have on whether a unit is included in the database or the sample drawn from it (sample selection bias), missing data (which can be produced either by confounding or by sample selection bias), and samples composed of units with different causal structures. Assuming that all independencies result from the Markov property of directed graphs representing causal hypotheses, for multinormal and multinomial distributions there are procedures based on hypothesis tests of constraints that, in the large sample limit, give correct causal information under most of these conditions. Scoring search algorithms have been developed for the case where there are no confounders, mixtures, and sample selection bias, but the best performance in search seems to be obtained using hypothesis-test initial searches followed by a Bayes score postsearch. Work is under way developing feasible scoring searches for latent variable models.

Prediction. Sometimes one is interested in using a sample, or a database, to predict properties of a new sample, where it is assumed that the two samples are obtained from the same probability distribution. As with estimation, in prediction we are interested both in reliability and uncertainty, often measured by the variance of the predictor.

Prediction methods for this sort of problem always assume some structure in the probability distribution. In data mining contexts, structure is typically either supplied by human experts, or inferred from the database automatically. Regression, for example, assumes a particular functional form relating variables. Structure can be also be specified in terms of constraints, such as independence, conditional independence, higher order conditions on correlations, etc. On average, a prediction method that guarantees satisfaction of the constraints realized in the probability distribution – and no others – will be more accurate and have smaller variance than one that does not. Finding the appropriate constraints to satisfy is the most difficult issue in this sort of prediction. As with estimation, prediction can be improved by model averaging, provided the prior probabilities of the alternative assumptions imposed by the model are available.

Often the results of data mining are used to attempt to predict the effects of a policy change or intervention of some kind. Such predictions are essentially causal inferences, and are not in general correctly given by estimating a probability distribution and predicting by conditioning on the variable to be manipulated.

10.2.2 Is Data Mining "Statistical Deja Vu" (All Over Again)?

In the mid 1960s, the statistics community referred to unfettered exploration of data as *fishing* or *data dredging.* The community, enamored by elegant (analytical) mathematical solutions to inferential problems, argued that since their theories were invalidated by "looking at the data," it was wrong to do so. The major proponent of the exploratory data analysis (EDA) school, J.W. Tukey, countered this argument with the obvious retort that statisticians were putting the cart before the horse. He argued that statistical theory should adapt to the scientific method rather than the other way around. Thirty years hence, the statistical community has largely adopted Tukey's perspective, and has made considerable progress in serving both masters, namely acknowledging that model search is a critical and unavoidable step in the modeling process, and devising formal methods to account for search in their inferential procedures.

Three themes of modern statistics that are of fundamental importance to data miners are: clarity about *goals*, appropriate *reliability* assessment, and adequate accounting for sources of *uncertainty.*

Clarity about goals. Sometimes data analysis aims at finding a convenient, easily computable representation of how the data are distributed in a particular database. In other cases, data analysis aims at predicting features of new cases, or new samples, drawn from outside the database used to develop a predictive model (this is particularly challenging in dynamic situations). In yet other cases, data analysis aims at providing a basis for policy. That is, the analysis is intended to yield insight into casual mechanisms that are used to form predictions about new samples that might be produced by interventions or actions that did not apply in the original database from which the model (or models) were developed. Each of these goals presents distinct inference problems, with distinct hazards. Confusing or equivocating over the aim invites the use of inappropriate methods and may result in unfortunate predictions and inferences.

As an example, consider the observational study reported by Chasnoff et al. in 1989 comparing babies born to cocaine-using mothers with babies born to noncocaine-using mothers. The authors concluded: "For women who become pregnant and are users of cocaine, intervention in early pregnancy with cessation of cocaine use will result in improved obstetric outcome." Fortunately, there exists independent evidence to support this causal claim. However, much of Chasnoff et al. focus on a statistical analysis (analysis of variance) that has little, if anything, to do with the causal question of interest.

Hand in 1994 provides a series of examples illustrating how easy it is to give the right answers to the wrong question. For example, he discusses the problem of analyzing clinical trial data where patients drop out due to adverse side effects of a particular treatment. In this case, the important issue is which population is one interested in modeling. The population at large versus the population who remain within the trail. This problem arises in more general settings than in clinical trails, e.g., nonrespondents (refusers) in survey data. In such situations it is important to be explicit about the questions one is trying to answer.

In this general context an important issue is that of formulating statistical strategy, i.e., how does one structure a data analysis problem so that the right question can be asked? Hand's conclusion is that this is largely an "art" because it is less well formalized than the mathematical and computational details of applying a particular technique. This "art" is gained through experience (at present at least) rather than taught. The implication for data mining is that human judgment is essential for many nontrivial inference problems. Thus, automation can at best only partially guide the data analysis process. Properly defining the goals of an analysis remains a human centered, and often difficult, process.

Use of methods that are reliable means to the goal, under assumptions the user (and consumer) understands and finds plausible in the context . Statistical theory applies several meanings to the word *reliability*, many of which also apply to model search. For example, under what conditions does a search procedure provide correct information, of the kind sought, with probability one as the sample size increases without bound. Answers to such questions are often elusive and can require sophisticated mathematical analysis. Where answers are available, the data analyst should pay careful attention to the reasonableness of underlying assumptions. Another key data mining question is this: what are the probabilities of various kinds of errors that result from using a method in finite samples? The answers to this question will typically vary with the kinds of errors considered, the sample size, and the frequency of occurrence of the various kinds of targets or signals whose description is the goal of inference. These questions are often best addressed by Monte Carlo methods, although in some cases analytic results may be available.

A sense of the uncertainties of models and predictions. Quite often background knowledge and even the best methods of search and statistical assessment should leave the investigator with a range of uncertainties about the correct model, or the correct prediction. The data analyst must quantify these uncertainties so that subsequent decisions can be appropriately hedged.

Another example involves a current debate in the atmospheric sciences. The question is whether or not specific recurrent pressure patterns can be clearly identified from daily geopotential height records, which have been compiled in the Northern Hemisphere since 1948. The existence of well-defined recurrent patterns (or "regimes") has significant implications for models of upper atmosphere low-frequency variability beyond the timescale of daily weather disturbances (and, thus, models of the earth's climate over large timescales). Several studies have used a variety of clustering algorithms to detect in homogeneities ("bumps") in low-dimensional projections of the gridded data. While this work has attempted to validate the cluster models via resampling techniques, it is difficult to infer from the multiple studies whether regimes truly exist, and, if they do, where precisely they are located. It seems likely that 48 winters' worth of data is not enough to identify regimes to any degree of certainly and that there is a fundamental uncertainty (given the current data) about the underlying mechanisms at work. All is not lost, however, since it is also clear that one could quantify model uncertainty in this context, and theorize accordingly. In what follows we elaborate on these points and offer a perspective one some of the hazards of data mining.

10.2.3 Characterizing Uncertainty

The statistical approach contends that reporting a single number for a parameter estimate or a prediction is almost always inadequate. Quantification of the *uncertainty* associated with a single number, while often challenging, is critical for subsequent decision making. As an example, Draper in 1995, considered the case of the 1980 Energy Modeling Forum (EMF) at Stanford University where a 43-person working group of economists and energy experts convened to forecast world oil prices from 1981 to 2020. The group generated predictions based on a number of econometric models and scenarios, embodying a variety of assumptions about supply, demand, and growth rates of relevant quantities. A plausible reference scenario and model was selected as representative, but the summary report cautioned against interpreting point predictions based on the reference scenario as "[the working group's] 'forecast' of the oil future, as there are too many unknowns to accept any projection as a forecast." The summary report did conclude, however, that most of the uncertainty about future oil prices "concerns not whether these prices will rise.... but how rapidly they will rise."

In 1980, the average spot price of crude oil was around $32 per barrel. Despite the warning about the potential uncertainty associated with the point estimates, governments and private companies around the world focused on the last sentence in the quotation above, and proceeded to invest an estimated $500 billion dollars, on the basis that the price would probably be close to $40 dollars per barrel in the mid-eighties. In fact, the actual 1986 world average spot price of oil was about $13 per barrel.

Using only the information available to the EMF in 1980, along with thoughtful but elementary statistical methods, Draper in 1995 shows that a 90% predictive interval for the 1986 price would have ranged from about $20 to over $90. Note that this interval does not actually contain the actual 1986 price-insightful statistical analysis does not provide clairvoyance. However, decision makers would (and should) have proceeded more cautiously in 1980, had they understood the full extent of their uncertainty.

Correctly accounting for the different sources of uncertainty presents significant challenges. Until recently, the statistical literature focused primarily on quantifying parametric and predictive uncertainty in the context of a particular model. Two distinct approaches are in common use. "Frequentist" statisticians focus on the randomness in sampled data and summarize the induced randomness in parameters and predictions by the so-called *sampling distributions.* "Bayesian" statisticians instead treat the data as fixed, and use *Bayes theorem* to turn prior opinion about quantities of interest (always expressed by a probability distribution), into a so-called *posterior distribution* that embraces all the available information. The fierce conflicts between previous generations of frequentists and Bayesians, have largely given way in recent years to a more pragmatic approach; most statisticians will base their choice of tool on scientific appropriateness and convenience.

In any event, recent research has led to increased awareness that *within-model* uncertainty (as discussed in the previous paragraph) may often, in practice, be dominated by *between-model* uncertainty. It is common practice nowadays for statisticians and data miners to use computationally intensive model selection algorithms to seek out a single optimal model from an enormous class of potential models. The problem is that *several* different models may be close to optimal, yet lead to different inferences. Intuitively, ambiguity over the model should dilute information about effect parameters and predictions, since "part of the evidence is spent to specify the model." Promising techniques for properly accounting for this source of uncertainty include

Bayesian model averaging and resampling methods. The main point here is that data miners need to think carefully about model assessment and look beyond commonly used goodness-of-fit measures such as mean square error.

10.2.4 What Can Go Wrong, Will Go Wrong

Data mining poses difficult and fundamental challenges to the theory and practice of statistics. While statistics does not have all the answers for the data miner, it does provide a useful and practical framework for which to search for solutions. In this section, we describe some lessons that statisticians have learned when theory meets data.

Data Can Lie

Data mining applications typically rely on observational (as opposed to experimental) data. Interpreting observed associations in such data is challenging; sensible inferences require careful analysis, and detailed consideration of the underlying factors. Here we offer a detailed example to support this position.

Wen et al. in 1995 (WHN) analyzed administrative records of all Ontario general hospital separations (discharges, transfers, or in-hospital deaths) from 1981 to 1990, focusing specifically on patients who had received a primary open cholecystectomy. Some of these patients had in addition received an incidental (i.e., discretionary) appendectomy during the cholecystectomy procedure. Table 10.2 displays the data on one outcome, namely in-hospital deaths. A chi-square test comparing this outcome for the two groups of patients shows a "statistically significant" difference. This "finding" is surprising since long-term prevention of appendicitis is the sole rationale for the incidental appendectomy procedure – no short-term improvement in outcomes is expected. This "finding" might lead a naïve hospital policy maker to conclude that all cholecystectomy patients should have an incidental appendectomy to improve their chances of a good outcome! Clearly something is amiss – how could incidental appendectomy improve outcomes?

WHN did separately consider a subgroup of low-risk patients. For these patients (using ten different definitions of "low-risk"), incidental appendectomy indeed resulted in poorer outcomes. Paradoxically, it could even be the case that appendectomy adversely affects outcomes for both high-risk patients and low-risk patients, but appears to positively affect outcomes when the low-risk

Table 10.2. In-hospital survival of patients undergoing primary open cholecystectomy with and without incidental appendectomy

	With appendectomy	Without appendectomy
In-hospital deaths, No. (%)	21 (0.27%)	1,394 (0.73%)
In-hospital survivors, No. (%)	7,825 (99.73%)	190,205 (99.27%)

Table 10.3. Fictitious data consistent with the Wen et al. (1995) data

	With appendectomy		Without appendectomy	
	Low risk	High risk	Low risk	High risk
Death	7	14	100	1294
Survival	7700	125	164009	26196

Table 10.4. Proportion of in-hospital deaths cross-classified by incidental appendectomy and patients risk grouping for the fictitious data of Table 10.3.

	With appendectomy	Without appendectomy
Low risk	0.0009	0.0006
High risk	0.10	0.05
Combined	0.003	0.007

and high-risk patients are combined. WHN did not provide enough data to check whether this so-called "Simpson's paradox" [Simpson, 1951] occurred in this example. However, Table 10.3 presents data that are plausible and consistent with WHN's data.

Table 10.4 displays the corresponding proportions of in-hospital death for these fictitious data. Clearly the risk and death categories are directly correlated. In addition, appendectomies are more likely to be carried out on low-risk patients than on high-risk ones. Thus, if we did not know the risk category (age) of a patient, knowing that they had an appendectomy allows us to infer that they are more likely to be lower risk (younger). However, this does not in any way imply that having an appendectomy will lower one's risk. Nonetheless, when risk is omitted from the table, exactly such a fallacious conclusion appears justified from the data.

Returning to the original data, WHN provide a more sophisticated regression analysis, adjusting for many possible confounding variables (e.g., age, sex, admission status). They conclude that "there is absolutely no basis for any short-term improvement in outcomes" due to incidental appendectomy. This careful analysis agrees with common sense in this case. In general, analyses of observational data demand such care, and come with no guarantees. Other characteristics of available data that connive to spoil casual inferences include:

- Associations in the database may be due in whole or part to unrecorded common causes (latent variables).
- The population under study may be a mixture of distinct causal systems, resulting in statistical associations that are due to the mixing rather than to any direct influence of variables on one another or any substantive common cause.
- Missing values of variables for some units may result in misleading associations among the recorded values.

- Membership in the database may be influenced by two or more factors under study, which will create a "spurious" statistical association between those variables.
- Many models with quite distinct causal implications may "fit" the data equally or almost equally well.
- The frequency distributions in samples may not be well approximated by the most familiar families of probability distributions.
- The recorded values of variables may be the result of "feedback" mechanisms, which are not well represented by simple "nonrecursive" statistical models.

There is research that addresses aspects of these problems, but there are few statistical procedures yet available that can be used "off the shelf" – the way randomization is used in experimental design – to reduce these risks. Standard techniques such as multiple regression, and logistic regression may work in many cases, such as in the appendectomy example, but they are not always adequate guards against these hazards. Indeed, controlling for possibly confounding variables with multiple regression can in some cases produce inferior estimates of effect sizes. Procedures recently developed in the artificial intelligence and statistics literature address some of the problems associated with latent variables and mixing, but so far only for two families of probability distributions, the normal and multinomial.

Sometimes It Is Not What Is In The Data That Matters

Classical statistical methods start with a random sample, yet in practice, data or the institutions that give rise to data, can be uncooperative. In such cases, inferences that ignore how the data were "selected" can lead to distorted conclusions.

Consider, for example, the Challenger Space Shuttle accident. The Rogers Commission concluded that an O-ring failure in the solid rocket booster led to the structural breakup and loss of the Challenger. In reconstructing the events leading up to the decision to launch, the Commission noted a mistake in the analysis of thermal-distress data whereby flights with no (i.e., zero) incidents of O-ring damage were excluded from critical plots of O-ring damage and ambient launch temperature since it was felt that they did not contribute any information about the temperature effect. This truncation of the data led to the conclusion that no relationship between O-ring damage and temperature existed, and ultimately, the decision to launch. Dalal et al. in 1989 throw statistical light on the matter by demonstrating the strong correlation between O-ring damage and temperature, and quantifying the risk (of catastrophic failure) at 31°F. Had the original analysis used all of the data, it would have indicated that the decision to launch was at best a risky proposition.

In the above case, the selection bias problem was one of "human error" and could easily have been avoided. In most problems, selection bias is an inherent characteristic of the available data and methods of analysis need to deal with it. It is our experience that every dataset has the potential for selection bias to invalidate standard inferences. The lessons to be learned here are

- that any technique used to analyze truncated data as if it was a random sample, can be fooled, regardless of how the truncation was induced;
- the data themselves are seldom capable to alert the analyst that a selection mechanism is operating – information external to the data at hand is critical in understanding the nature and extent of potential biases.

The Perversity of the Pervasive P-Value

P-values and associated significance (or hypothesis) tests play a central role in classical (frequentist) statistics. It seems natural, therefore, that data miners should make widespread use of P-values. However, indiscriminate use of P-values can lead data miners astray in most applications.

The standard significance test proceeds as follows. Consider two competing hypotheses about the world: the *Null Hypothesis*, commonly denoted by H_o, and the *Alternative Hypothesis*, commonly denoted by H_A. Typically H_o is "nested" within H_A; for example, H_o might state that a certain combination of parameters is equal to zero, while H_A might place no restriction on the combination. A *test statistic*, T is selected and calculated from the data at hand. The idea is that T (Data) should measure the evidence in the data against H_o. The analyst rejects H_o in favor of H_A if T (Data) is more extreme than would be expected if H_o were true. Specifically, the analyst computes the P-value, that is, the probability of T being greater than or equal to T (Data), given that H_o is true. The analyst rejects H_o if the P-value is less than a preset *significance level*, α.

There are three primary difficulties associated with this approach:

1. The standard advice that statistics educators provide, and scientific journals rigidly adhere to, is to choose α to be 0.05 or 0.01, *regardless of sample size.* These particular α-levels arose in Sir Ronald Fisher's study of relatively small agricultural experiments (on the order of 30–200 plots). Textbook advice has emphasized the need to take account of the power of the test against H_A when setting α, and somehow reduce α when the sample size is large. This crucial but vague advice has largely fallen on deaf ears.
2. Raftery in 1995 points out that the whole hypothesis testing framework rests on the basic assumption that only two hypotheses are ever entertained. In practice, data miners will consider very large numbers of possible models. As a consequence, indiscriminate use of P-values with "standard" fixed α-levels can lead to undesirable outcomes such as selecting a model with parameters that are highly significantly different from zero, even when the training data are pure noise. This point is of fundamental importance for data miners.

3. The P-value is the probability associated with the event that the test statistic was as extreme as the value observed, or more so. However, the event that actually happened was that a specific value of the test statistic was observed. Consequently, the relationship between the P-value and the veracity of H_o is subtle at best. Jeffreys in 1980 puts it this way:

   > I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trail value was improbable; that is, that it has not predicted something that has not happened.

Bayes Factors are the Bayesian analog of the frequentist P-values and admit to a more direct interpretation – the Bayesian analyst computes the posterior probability that a hypothesis is correct. With fixed α-levels, the frequentist and the Bayesian will arrive at very different conclusions. For example, Berger and Sellke in 1987 show that data that yield a P-value of 0.05 when testing a normal mean, result in a posterior probability for H_o that is at least 0.30 for any "objective" prior distribution. One way to reconcile the two positions is to view Bayes factors as a method for selecting appropriate α-levels.

Intervention and Prediction

A specific class of prediction problems involves interventions that alter the probability distribution of the problem, as in predicting the values (or probabilities) of variables under a change in manufacturing procedures, or changes in economic or medical treatment policies. Accurate predictions of this kind require some knowledge of the relevant causal structure, and are in general quite different from prediction without intervention, although the usual caveats about uncertainty and model averaging apply. For graphical representations of causal hypotheses according to the Markov condition, general algorithms for predicting the outcomes of interventions from complete or incomplete causal models were developed.

Consider the following example, Herbert Needleman's famous studies of the correlation of lead deposits in children's teeth with their IQs resulted, eventually, in removal of tertraethyl lead from gasoline in the United States. One dataset Needleman examined included more than 200 subjects and measured a large number of covariates. Needleman et al. in 1985 reanalyzed the data using backwards stepwise regression of verbal IQ on these variables and obtained six significant regressors, including lead. Klepper in 1988 reanalyzed the data assuming that all of the variables were measured with error. Their model assumes that each measured number is a linear combination of the true value and an error, and that the parameters of interest are not the regression coefficients but rather the coefficients relating the unmeasured "true value" variables to the unmeasured true value of verbal IQ. These coefficients are in fact indeterminate – in econometric terminology, “unidentifiable.” An interval estimate of the coefficients that is strictly positive or negative for each coefficient can be made, however, if the amount of measurement error can be bounded with prior knowledge by an amount that varies from case to case. Klepper found that the bound required to ensure the existence of a strictly negative interval estimate for the lead-IQ coefficient was much too strict to be credible; thus he concluded that the case against lead was not nearly as strong as Needleman’s analysis suggested.

Allowing the possibility of latent variables, Scheines in 1996 reanalyzed the correlations (using TETRAD methodology) and concluded that three of the six regressors could have no influence on IQ. The regression included the three extra variables only because the partial regression coefficient is estimated by conditioning on all other regressors, which is just the right thing to do for linear prediction, but the wrong thing to do for causal inference using the Markov condition. Using the Klepper model, but without the three irrelevant variables, and assigning to all of the parameters a normal prior probability with mean zero and a substantial variance, Scheines that used Markov chain Monte Carlo to compute a posterior probability distribution for the lead-IQ parameter. The probability is very high that lead exposure reduces verbal IQ.

10.2.5 Symbiosis in Statistics

Easy access to data in digital form and the availability of software tools for statistical analyses have made it possible for the man in street to set up shop and “do statistics.” Nowhere is this more true today than in data mining. Based on the arguments in this section, let us assume that statistics is a necessary but not sufficient component in the practice of data mining. How well will the statistics profession serve the data mining community? Hoerl et al. in 1993, for example, assert that:

We are our own best customers. Much of the work of the statistical profession is intended for other members of the statistical profession.

Despite this rather negative view of the relevance of statistical research, real-world applications do in fact drive much of what goes on it statistics, although often in a very indirect manner.

As an example, consider the field of signal processing and communications, an area where a specialized set of relatively sophisticated statistical methods and models have been honed for practical use. The field was driven by fundamental advances from Claude Shannon and others in the 1940s. Like most of the other contributors to the field, Shannon was not a statistician, but possessed a deep understanding of probability theory and its applications. Through the 1950s to the present, due to rapid advances in both theory and hardware, the field has exploded and relevant statistical methods such as estimation and detection have found their way into everyday use in radio and network communications systems. Modern statistical communications reflects the symbiosis of statistical theory and engineering practice. Engineering researchers in the field are in effect "adjunct" statisticians: educated in probability theory and basic statistics they have the tools to apply statistical methods to their problems of interest. Meanwhile statisticians continue to develop more general models and estimation techniques of potential applicability to new problems in communications.

This type of symbiosis can also be seen in other areas such as financial modeling, speech recognition (where for example hidden Markov models provide the state of the art in the field), and most notably, epidemiology. Indeed, if statistics can claim to have revolutionized any field, it is in the biological and health sciences where the statistical approach to data analysis gave birth to the field of biostatistics.

The relevance of this symbiosis for data mining is that data-miners need to understand statistical principles, and statisticians need to understand the nature of the important problems that the data mining community is attacking or being asked to attack. This has been a successful model in the past for fields where statistics has had considerable impact and has the potential to see ongoing success.

10.3 Summary

Data mining is a growing discipline, which originated outside statistics in the database management community, mainly for commercial concerns. The gap is now filled and data mining could be considered as the branch of exploratory statistics where one tries to find unexpected and useful models and patterns, through an extensive use of classic and new algorithms. The expression *unexpected* should not be misleading: one has greater chances to discover something interesting when one is previously acquainted with the data.

Caution is of course necessary to avoid drawing wrong conclusions and statisticians, because they are familiar with uncertainty and risk, and are the right people to derive validation tests. A completely automatic process of knowledge extraction is also a misleading idea: even with very efficient software, human expertise and interventions are necessary. Official statistics should be a field for data mining, giving a new life and profitability to its huge databases, but it may implies a redefinition of the missions of NSI.

The statistical literature has a wealth of technical procedures and results to offer data mining, but it also has a few simple methodological morals: prove that estimation and search procedures used in data mining are consistent under conditions reasonably thought to apply in applications; use and reveal uncertainty, do not hide it; calibrate the errors of search, both for honesty and to take advantages of model averaging; do not confuse conditioning with intervening; and finally, do not take the error probabilities of hypothesis tests to be the error probabilities of search procedures.

10.4 Review Questions

1. What are the goals and tools of data mining?
2. Give an overview of statistical science for application to data mining.
3. Justify – data mining is all over gain.
4. Define intervention, prediction, and P-value.