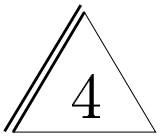

A Statistical Perspective on Knowledge Discovery in Databases

John F. Elder IV
Computational and Applied Mathematics Department
& Center for Research on Parallel Computation,
Rice University

Daryl Pregibon
Statistics and Data Analysis Research,
AT&T Bell Laboratories

Abstract

The quest to find models usefully characterizing data is a process central to the scientific method, and has been carried out on many fronts. Researchers from an expanding number of fields have designed algorithms to discover rules or equations that capture key relationships between variables in a database. The task of this chapter is to provide a perspective on statistical techniques applicable to KDD; accordingly, we review below some major advances in statistics in the last few decades. We next highlight some distinctives of what may be called a "statistical viewpoint." Finally we overview some influential classical and modern statistical methods for practical model induction.

4.1 Recent Statistical Contributions

It would be unfortunate if the KDD community dismissed statistical methods on the basis of courses that they took on statistics several to many years ago. The following provides a rough chronology of "recent" significant contributions in statistics that are relevant to the KDD community. The noteworthy fact is that this time period coincides with the significant increases in computing horsepower and memory, powerful and expressive programming languages, and general accessibility to computing that has propelled us into

the Information Age. In effect, this started a slow but deliberate shift in the statistical community, whereby important influences and enablers were to come from computing rather than mathematics.

4.1.1 The 1960s

This was the era of *robust* and *resistant* statistical methods. Following ideas of G. E. P. Box and J. W. Tukey,

Huber (1964) and Hampel (1974) formalized the notion that the usual estimators of location and regression coefficients were very sensitive to "outliers", "leverage values", and otherwise unreasonably small amounts of contamination. Key concepts are the

- *influence* function of Hampel (essentially the derivative of an estimator with respect to the data)
- *M*-estimators of Huber, so-called because they generalize maximum likelihood estimators (which require a probability distribution) to a closely related class of estimating equations
- *diagnostics*, where implicit downweighting of observations afforded by robust estimators is replaced by empirical derivatives that quantify the effects of small changes in the data on important aspects of regression-like models (see for example, Belsley, Kuh, and Welsch, 1980)

The theory supporting these ideas is elegant and important as it unifies many seemingly unrelated concepts (*e.g.* trimmed means and medians) and more so because it reflects the realism that data does not usually obey assumptions as required by (mathematical) theorems. Thus the robustness era freed statisticians of the shackles of narrow models depending on unrealistic assumptions (*e.g.* normality).

The only downside of the era was that too much effort was placed on deriving new estimators that deviated only slightly from each other both qualitatively and quantitatively.[1] What was needed instead, was the leadership and direction in *using* these methods in practice and dealing with the plethora of alternatives available. Partly because of this misguided effort, many of the techniques of the era never made it into commercial software and therefore never made it into the mainstream of methods used by nonstatisticians.

4.1.2 The Early 1970s

The term Exploratory Data Analysis (EDA) characterizes the notion that statistical insights and modeling are driven by data. John Tukey (1977; Mosteller and Tukey,

[1] Basically reflecting R. A. Fisher's insight (*Statistical Methods for Researchers*, 1924) that there is nothing easier than inventing a new statistical estimator.

1977) reinforced these notions in the early 70's using a battery of ultra-simple methods, *e.g.* what could be done with pencil and paper. But the deeper message was to dispel the traditional dogma stating that one was not allowed to "look at the data prior to modeling". On the one side the argument was that hypotheses and the like must not be biased by choosing them on the basis of what the data seem to be indicating. On the other side was the belief that pictures and numerical summaries of data are necessary in order to understand how rich a model the data can support.

A key notion in this era characterized statistical modeling as decomposing the data into structure and noise,

$$data = fit + residual \tag{4.1.1}$$

and then examining residuals to identify and move additional structure into the fit. The fitting process would then be repeated and followed by subsequent residual analyses.

The iterative process described above has its roots in the general statistical paradigm of partitioning variability into distinct parts (*e.g.,* explained and unexplained variation; or, in classification, within-group and between-group variation). The EDA notion simply uses the observed scale of the response rather than the somewhat unnatural squared units of "variability". While this might seem like a trivial distinction, the difference is critical since it is only on the observed scale that diagnosis *and* treatment is possible. For example, a component of variance can indicate that nonlinearity is present but cannot prescribe how to accommodate it.

Graphical methods (not to be confused with graphical *models* in Bayes nets) enjoyed a renaissance during this period as statisticians (re-)discovered that nothing outperforms human visual capabilities in pattern recognition. Specifically, statistical tests and models focus on *expected* values, and in many cases, it is the *un*expected that upsets or invalidates a model (*e.g.,* outliers). Tukey argued that (good) graphical methods should allow *un*expected values to present themselves — once highlighted, models can be expanded or changed to account for them.

Another important contribution was to make data *description* respectable once more. Statistics has its roots in earlier times when descriptive statistics reigned and mathematical statistics was only a gleam in the eye. Data description is concerned with simplicity and accuracy, while not being overly formal about quantifying these terms (though an important area of research tries to do just that; *e.g.*, Mallows (1973), Akaike (1973), and Rissanen (1978)). A key notion popularized in this era was that there is seldom a single *right* answer – in nearly all situations there are many answers. Effective data description highlights those that are simple, concise, and reasonably accurate. Simple transformations of a dataset are used to effect such descriptions, the two most common

ones being *data reexpression, e.g.* using $\log(age)$ instead of age, and *data splitting, e.g.* setting aside outliers to simplify the description of the bulk of the data.

4.1.3 The Late 1970s

To an outsider much of the statistical literature would seem fragmented and disjoint. But the fact of the matter is that much is closely related, but that specific details of individual contributions hide the real similarities. In the late 70's, two review papers and one book elegantly captured the essence of numerous prior publications. The first of these, *Generalized Linear Models* (Nelder and Wedderburn, 1974; McCullagh and Nelder, 1989) extended the usual normal theory linear model to a much wider class of models that included probability models other than the normal distribution, and structural models that were nonlinear. The theory accomplished this by decomposing the variation in a response variable into systematic and random components, and allowed the former to capture covariate effects through a strictly monotone *link* function, $g(\mu) = \sum x_j \beta_j$, and allowing the latter to be a member of the exponential family of distributions, $\mathcal{E}(\mu, \sigma)$. In so doing, these models provided a unifying theory for regression-like models for binary and count data, as well as continuous data from asymmetric distributions. The second major review paper is well known outside of statistics as the EM algorithm (Dempster, Laird, and Rubin, 1977). This paper neatly pulled together numerous ways of solving estimation problems with incomplete data. But the beauty of their general treatment was to instill the concept that even if data are complete, it is often useful to treat it as a missing value problem for computational purposes. Finally, the analysis of nominal or discrete data, specifically counts, had several disconnected streams in the literature and inconsistent ways to describe relationships. Bishop, Fienberg, and Holland (1975) pulled this material together into the class of *loglinear* models. The associated theory allowed researchers to draw analogies to models for continuous data (for example, analysis of variance ideas) and further provided computational strategies for estimation and hypothesis testing. It is also noteworthy that this work anticipated current work in so-called *graphical models*, a subset of the class of loglinear models for nominal data.

4.1.4 The Early 1980s

Resampling methods had been around since the late 1950s under the moniker *jackknife*, so-named by Tukey because it was a "trusty general purpose tool" for eliminating low-order bias from an estimator (Schreuder, 1986). The essence of the procedure is to replace the original n observations by n or more (possibly) correlated estimates of the quantity of interest (called *psuedovalues*). These are obtained by systematically leaving out one or more observations and recomputing the estimator. More precisely, if θ is the parameter

of interest, the ith psuedovalue is defined by

$$p_i = n\hat{\theta}_{all} - (n-k)\hat{\theta}_{-i} \tag{4.1.2}$$

where the last quantity is the estimator $\hat{\theta}$ based on leaving out the ith subset (of size k). The jackknife estimate of θ is the arithmetic mean of the psuedovalues, $\bar{p} = \sum p_i/n$.

While the jackknife was originally proposed as a bias reduction tool, it was quickly recognized that the ordinary standard deviation of the psuedovalues provides an honest estimate of the error in the estimate. Thus an empirical means of deriving a measure of uncertainty for virtually *any* statistical estimator was available. One interpretation of the procedure is that the construction of psuedovalues is based on repeatedly and systematically sampling *without* replacement from the data at hand. This led Efron (1979) to generalize the concept to repeated sampling *with* replacement, the so-called *bootstrap* (since it allowed one to "pick oneself up by the bootstraps" in constructing a confidence interval or standard error). This seemingly trivial insight opened the veritable flood gates for comprehensive analytic study and understanding of resampling methods. The focus on estimating precision of estimators rather than bias removal coupled with the advance of computing resources, allowed standard errors of highly nonlinear estimators to be routinely considered.

Unfortunately, as with robustness, the bulk of the research effort was directed at theoretical study of resampling ideas in what KDD researchers would regard as uninteresting situations. The most nonlinear procedures, such as those resulting from combining model identification and model estimation (see Section 4.1.6), received only cursory effort (*e.g.* Efron and Gong, 1983; Faraway, 1991).

4.1.5 The Late 1980s

One might characterize classical statistical methods as being "globally" linear whereby the explanatory/prediction/classification variables affect the distribution of the response variable via linear combinations. Thus the effect of x_j on y is summarized by a single regression coefficient β_j. Nonlinear relationships could only be modeled by specifically including the appropriate nonlinear terms in the model, *e.g.* x_j^2 or $\log x_j$. Cleveland (1979) helped seed the notion that globally linear procedures could be replaced with locally linear ones by employing scatterplot smoothers in interesting ways. A scatterplot smoother $s(x)$ is a data-dependent curve defined pointwise over the range of x. For example, the *moving average* smoother is defined at each unique x, as the mean $\bar{y}(x) = \sum y_i/k$ of the k (symmetric) nearest neighbors of x. The ordered sequence of these pointwise estimates traces out a "smooth" curve through the scatter of (x, y) points. Originally smoothers were used simply to enhance scatterplots where clutter or changing density of plotted points hindered visual interpretation of trends and nonlinear features.

But by interpreting a scatterplot smoother as an estimate of the conditional mean $E(y|x)$, one obtains an adaptive, nonlinear estimate of the effect of x on the distribution of y. Moreover, this nonlinearity could be tamed while simultaneously reducing bias caused by end-effects, by enforcing "local" linearity in the smoothing procedure (as opposed to local constants as provided by moving averages or medians). Thus by moving a *window* across the data and fitting linear regressions within the window, a globally nonlinear fit is obtained, *i.e.* the sequence of predictions at each point x_i, $s_i(x) = a_i + b_i x$, where the coefficients a_i and b_i are determined by the least squares regression of y on x for all points in the window centered on x_i.

This notion has been applied now in many contexts (*e.g.* regression, classification, discrimination) and across many "error" distributions (*e.g.* the generalized additive model of Hastie and Tibshirani, 1985). While this work reduced the emphasis on strict linearity of the explanatory variables in such models, it did not ameliorate the need for having previously identified the relevant variables to begin with.

4.1.6 The Early 1990s

Within the statistics community, Friedman and Tukey (1974) pioneered the notion of allowing a model to adapt even more nonlinearly by letting the data determine the interesting structure present with "projection pursuit" methods (Section 4.4.3). These are less restrictive than related nonlinear methods such as neural networks (Section 4.4.4), supposing a model of the form

$$\mu(y|x) = \sum_{k=1}^{K} g_k(\sum_{j=1}^{J} x_j \beta_{jk}) \tag{4.1.3}$$

where both the regression coefficients β_{jk} and the *squashing* functions $g_k()$ are unknown.

Important algorithmic developments and theory resulted from these models even though they failed to achieve widespread use within the statistics community. Part of the reason was that these models were regarded as *too* flexible in the sense that arbitrarily complex functions could be provably recovered (with big enough K). The community instead retreated back to additive models that had limited flexibility but afforded much greater interpretability. Indeed, interpretability was the focus of much of the work in this area as alternative formulations of the locally linear model were derived, *e.g.*, penalized likelihood and Bayesian formulations (O'Sullivan *et al.* 1986).

Still, these ambitious methods helped to nudge the community from focusing on model estimation to model selection; for modern methods (see Sections 4.4.3 to 4.4.7) the modeling search is over structure space as well as parameter space. It is not uncommon now for many thousands of candidate structures to be considered in a modeling run

– which forces one to be even more conservative when judging whether improvements are significant, since any measure of model quality optimized by a search is likely to be over-optimistic (see *e.g.*, Miller, 1990 in the context of regression subset selection). When considering a plethora of candidates it usually becomes clear that a wide variety of models, with different structures and even inputs, score nearly as well as the single "best". Therefore, following the ancient statistical adage that "in many counselors there is safety"[2] some researchers are now explicitly *blending* the outputs from several viable models to obtain estimates with reduced variance and (almost always) better accuracy on new data (*e.g.*, Wolpert, 1992; Breiman, 1994b). Such techniques are especially promising when the models being merged are from completely different families (for example, trees, polynomials, kernels, and splines), and if the local influence of each is a function of its estimated accuracy in that region of design space.

4.2 Distinctives of a Statistical Viewpoint

4.2.1 Interpretability

Researchers from different fields seem to emphasize different qualities in the models they seek. For example, Breiman (1994a) noted that the "neural network community" appears not to be wedded to variations on that approach, but may experiment with a wide variety of techniques under the overriding goal of developing a model minimally misclassifying new data. Statisticians, on the other hand, are usually interested in interpreting their models, and may sacrifice some performance to be able to extract meaning from the model structure. If the accuracy is acceptable they reason that a model which can be decomposed into revealing parts is often more useful than a "black box" system, especially during early stages of the investigation and design cycle.

4.2.2 Characterizing Uncertainty

The randomness in sampled data is inherited by estimated model parameters since these are functions of the data. Statisticians summarize the induced randomness by so-called *sampling distributions* of estimators. By judicious assumptions, exact sampling distributions are analytically tractable; more typically asymptotic arguments are invoked. The net result is often the same, the estimated parameters are approximately normally distributed. This distribution characterizes the uncertainty in the estimated parameters, and owing to normality, the uncertainty is succinctly captured in the standard deviation of the sampling distribution, termed the *standard error* of the estimate. Standard

[2] Proverbs 24:6b

statistical practice requires stating the standard errors of estimated model parameters. Parameters associated with estimates that are small in comparison to their standard errors, (*e.g.*, $t = \hat{\beta}/s.e.(\hat{\beta}) < 2$) are not likely to be part of the "true" underlying process generating the data, and it is often prudent to drop such parameters from the model. A term by term analysis such as this breaks down in the presence of collinear variables and is weakened also by nonlinear models that stretch the applicability of the asymptotic normal sampling theory. Yet, the basic insight is very useful: estimates should be accompanied by uncertainty measures (*e.g.*, error bars) to be useful.

The Bayesian paradigm provides a different though related perspective. Here one treats the parameter itself as a random variable and merges prior beliefs about the parameter together with observed data. The resulting *posterior* distribution, $p(\theta|data)$, can often itself be approximated by a normal distribution, and thereby a single number summary of parameter uncertainty is available. Of course, recent computational advances and ingenious algorithms (*e.g.* Markov chain monte carlo) obviate the need for analytically derived normal approximations. But lacking a picture of the posterior distribution, the second moment is often used to summarize the spread of the induced posterior distribution.

Other disciplines also deal with unavoidable variation. For instance, electrical engineers design circuits to filter noisy signals using components with inexact values themselves (*e.g.*, resistors with 15% tolerances). Similarly, financial analysts know that potential investments need to be evaluated not only on their expected return, but on their risk – usually, the standard deviation of those returns. Investments with higher historical or implied deviations make sense only if they are accompanied by an appropriate "risk premium" (higher expected return). On the other hand, logicians and computer scientists have been slow to appreciate the importance of explicitly handling uncertainty. Arguably, Statistics has had a head start on this problem and seems to have the natural language, the *probability* calculus, to propagate and characterize uncertainty in models. The "certainty factors" in early expert systems and the "fuzzy logic" of later ones, are weak attempts to do what probability has done for centuries.

In some modeling contexts, emphasis is on prediction rather than estimation (of model parameters). This change in emphasis does not reduce the need to characterize and report uncertainty. Properly formulated models provide not only the prediction at each point in the design space, $E(y|x)$), but also the associated variance, $var(y|x)$. Monitoring local variance is useful for more than confidence estimates. For example, Cox and John (1993) and Elder (1993a) employ conditional variances of a response surface to guide global search algorithms very efficiently in low dimensions. Unfortunately, only a few nonlinear inductive modeling techniques (see Section 4.4) explicitly incorporate conditional variance into their estimates – a clear area for potential improvement.

4.2.3 Borrowing Strength

It is often the case in statistical problems that inferences are desired but data is sparse. Consider an example from retail marketing. An SKU (stock keeping unit) is a unique label assigned to a retail product, for example, men's size 12 blue socks. Predictions of SKUs are required at a store level in a large chain of department stores to build up sufficient inventory for promotions and seasonal demand or other "predictable" events. The problem is that detailed historical data on individual SKU sales at each and every store in the chain is not available; for example, it may be that no men's size 12 blue socks sold in the Florida store since last November. The concept of *borrowing strength* allows us to build forecasts at the site-SKU level by exploiting hierarchies in the problem, possibly in more ways than one. By aggregating across stores, sufficient information is available to build a site-independent prediction for each SKU. This prediction can be used to add stability to predictions of SKUs in each of several regions, which can in turn be used to add stability at the site level. Similar types of decompositions could allow us to borrow strength by looking at sales of, say, all blue socks independent of size, then all socks, then men's undergarments, then menswear overall. Such "hierarchical models" have their origins in *empirical Bayes* models, so-called because inferences are not truly Bayesian, as maximum likelihood estimates are used in place of "hyperparameters" (the parameters in prior distributions) at the highest levels of the hierarchy where data is most numerous. This typically results in estimates of the form $\hat{y}_i = \alpha \bar{y}_i + (1-\alpha)\bar{y}$ where $\bar{y}_i$ is the estimate specific to the ith level of the hierarchy and $\bar{y}$ to that of its parent (where data is more abundant). The mixing parameter, α, captures the similarity of the individual estimate to its parent relative to the tightness of the distribution of the $\bar{y}_i$'s.

4.2.4 Explicit Assumptions

Statisticians are typically very aware of the explicit and implicit assumptions associated with their models. Though some of the appeal of non-traditional models and methods undoubtedly stems from their apparent ability to bypass statistical analysis stages many see as cumbersome, it is clear that matching the assumptions of a method with the characteristics of a problem is beneficial to its solution. Statistical analysts usually take the useful step of checking those assumptions; chiefly, by examining:

1. residuals (model errors)
2. diagnostics (model sensitivity to perturbations)
3. parameter covariances (redundancy)

Not all violations of assumptions are equally bad. For example, assumptions about stochastic behavior of the data are typically less important than the structural behavior;

the former might lead to inefficient estimates or inaccurate standard errors, but the later could result in biased estimates. Within these two broad classes, normality and independence assumptions are typically less important than constancy (homogeneity) of variance (*e.g.,* $var(y|x) = constant$ for all x). A single outlier from the structural model can bias the fit everywhere else. Likewise, leverage values are those observations that have undue influence on the fit, for example if deleting the ith observation resulted in a large change in the estimate of a key parameter. An important distinction is that leverage values need not correspond to large residuals – indeed by virtue of their "leverage", they bias the fit toward themselves resulting in small or negligible residuals. Colinearity among the predictor variables confuses the interpretation of the associated parameters, but can also be harmful to prediction; the new data must strictly abide by the interrelationships reflected in the training data or the model will be extrapolating beyond the confines of the training space, rather than interpolating within it.

4.2.5 Regularization

The aim of statistical inference and inductive modeling is to infer general laws from specific cases – to summarize observations about a phenomenon into a coherent model of the "underlying data-generating mechanism" which can be tested for explanatory power on new cases. To perform well on data not seen during "training", models need appropriate structure and complexity; they must be powerful enough to approximate the known data, but constrained enough to generalize successfully. "Ockham's razor" is often invoked as a guiding principle in model selection, which suggests one select for use, from competing hypotheses with similar explanatory power, the simplest one. In many cases the simpler, less accurate model will generalize better to new data arising from the process that generated the training set.

In statistical terms, the tradeoff is between model "underfit" (bias) and "overfit" (variance), and the imposition of modeling restraint is called "regularization". If data are plentiful, model overfit can be avoided by reserving representative subsets of the data for testing as the model is constructed. When performance on the test set systematically worsens, model growth is curtailed. In the more common scenario in which the design space is less densely populated with data, all the cases can be employed for training, and model complexity (*e.g.,* number of parameters) or roughness (*e.g.,* integrated squared slope of its response surface) is used to constrain the fit. The criterion to be minimized is then a weighted sum of the training error and the measure of model complexity or roughness. Note that nonlinear and adaptively-selected parameters can have more influence on a model than is typical for linear terms, so their inclusion must be accompanied by a correspondingly greater increase in training accuracy to "pay their way".

Two other regularization methods are employed, the first in statistics and the other in

nonstatistical communities. The first method, parameter *shrinkage*, uses all the variables but constrains their overall influence to make the models more robust. For example, with collinear variables, there can be infinite solutions to a linear estimation problem. Even with nearly collinear variables, the estimated "optimal" parameters may have huge variances – a clear type of overfit. By shrinking the parameters, *e.g.,* through a singular value decomposition of the ill-conditioned design matrix, a relatively robust weight solution is selected more near the origin (where all parameters are zeroed out). Likewise, ridge regression pushes unstable solutions in the direction of smaller values, effectively reducing the complexity of the model. Shrinkage can also be performed on trees (Pregibon and Hastie, 1990) and neural networks (known as "optimal weight decay"). Most shrinkage procedures have a Bayesian interpretation whereby the user-defined prior guides the direction and degree of regularization.

The second method, which applies to iterative procedures (and is considered a relatively crude approach by many statisticians), is to halt the adjustment procedure some time before convergence. This has been the primary means by which artificial neural network training (weight modification) is halted but has also been reported in other contexts such as taming the EM algorithm in positron emission tomography.

To summarize, it is a hallmark of the statistical approach to *regularize* models; *i.e.,* to employ simplifying constraints alongside accuracy measures during model formulation in order to best generalize to new cases – the true goal of most empirical modeling activities.

4.3 Reservations to Automatic Modeling in Statistics

The experienced statistician, perhaps the most capable of guiding the development of automated tools for data analysis, may also be the most acutely aware of all the difficulties that can arise when dealing with real data. This hesitation has bred skepticism of what automated procedures can offer and has contributed to the strong focus by the statistical community on model *estimation* to the neglect of the logical predecessor to this step, namely model *identification*. Another culprit underlying this benign neglect is the close historical connection between mathematics and statistics whereby statisticians tend to work on problems where theorems and other analytical solutions are attainable (*e.g.* sampling distributions and asymptotics). Such solutions are necessarily conditional on the underlying model being specified up to a small number of identifiable parameters that summarize the relationship of the predictor variables to the response variable through the first few moments of the conditional distribution, $f(y|x)$. For example the common regression model takes the form:

$$\mu(y|x) = \sum_{j=1}^{J} x_j \beta_j \tag{4.3.4}$$

$$\sigma(y|x) = constant \tag{4.3.5}$$

The implicit parameter J is not part of the explicit formulation nor is the precise specification of which x_j's define the model for the mean parameter μ. Traditional statistics provides very useful information on the sampling distribution of the estimates $\hat{\beta}_j$ for a fixed set of x_j's but no formalism for saying which x's are needed.

The relatively small effort by the statistical community in model identification has focused on marrying computing horsepower with human judgment (as opposed to fully automated procedures). The general problem is deciding how large or complex a model the available data will support. By directly and explicitly focusing on mean squared prediction error, statisticians have long understood the basic tradeoff between bias (too small a model) and variance (too large a model) in model selection. Algorithms (Furnival and Wilson, 1978) and methods (Mallows, 1973) have been used extensively in identifying candidate models summarized by model accuracy and model size. The primary reason that human judgment is crucial in this process is that algorithmic optimality does not and cannot include qualitative distinctions between competing models of similar size – for example, if the accuracy/availability/cost of the variables differ. So it is largely human expertise that is used to select (or validate) a model, or a few models, from the potentially large pools of candidate models.

The statistician's tendency to avoid complete automation out of respect for the challenges of the data, and the historical emphasis on models with interpretable structure, has led that community to focus on problems with a more manageable number of variables (a dozen, say) and cases (several hundred typically) than may be encountered in KDD problems, which can be orders of magnitude larger at the outset.[3] With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention, are required. The existence of such tools can free one up to, for instance, posit a wider range of candidate data features and basis functions (building blocks) than one would wish to deal with, if one were specifying a model structure "by hand".

This obvious need is gaining sympathy but precious little has resulted. The subsections below highlight some of the areas that further underlie the hesitation of automating model identification by the statistical community.

[3]Final models are often of similar complexity; it's the magnitude of the initial candidate set of variables and cases that is usually larger in KDD.

4.3.1 Statistical significance versus practical significance

A common approach to addressing the complexity and size of *model space* is to limit model growth in the model fitting/learning stage. This is almost always accomplished using a statistical test of significance at each step in the incremental model building stage. Thus for example, one could use a standard χ^2 test of independence between two nominal variables as a means to limit growth of a model that searches for "significant" association. The main problem with this approach is that significance levels depend critically on n, the sample size, such that as n increases, even trivial differences attain statistical significance. Statisticians ameliorate this problem by introducing context to better qualify findings as "significant."

4.3.2 Simpson's paradox

A related problem with automated search procedures is that they can often be completely fooled by anomalous association patterns, even for small datasets. An accessible and easily understood example (Freedman, Pisani, and Purves, 1978) concerns admission to graduate school at UC Berkeley in 1973. Across major departments, 30% of 1835 female applicants were admitted while 44% of 2691 male applicants were admitted. Do these disparate fractions indicate sex bias? On the face yes, but if the applicants and admissions are broken down by department, then the fractions of the two sexes admitted shows a very different story, where one might even argue that "reverse" sex bias is present! The "paradox" here is that the choice of major is *confounded* with sex – namely that females tend to apply to majors that are harder to get into while males apply to "easy" majors.

The implication of this paradox is that KDD tools which attack large databases looking for "interesting" associations between pairs of variables must also contain methods to search for potential confounders. Computationally, this changes the problem from an n^2 to an n^3 operation (or higher if one considers more confounders). The computational burden can only be avoided by providing knowledge about potential confounders to the discovery algorithm. While this is in principle possible, it is unlikely to be sufficient since common sense knowledge often suggests what confounders might be operating. Statisticians have long brought these common sense insights to the problem rather than delegate them to automata.

4.3.3 Selection bias

Automated knowledge discovery systems are applied to databases with the expectation of translating *data* into *information*. The bad news is that often the available data is not representative of the population of interest and the worse news is that the data itself contains no hint that there is a potential bias present. Namely, it's more an issue of what

is *not* in a data set rather than what information it contains. For example [4], suppose that the White House Press Secretary is using a KDD (*e.g.* information retrieval) tool to browse through email messages to PRESIDENT@WHITEHOUSE.GOV for those that concern health care reform. Suppose that she finds a 10:1 ratio of pro-reform to anti-reform messages, leading her to assert that "Americans favor reform by a 10:1 ratio" followed by the worrisome rejoinder "and Government can fix it." But it may well be that people dissatisfied with the health care system are *more likely* to "sound off" about their views than those who are satisfied. Thus even if the true distribution of views on health care reform has mean "score" of zero, self-selected samples that are heavily biased towards one of the tails of this distribution will give a very misleading estimate of the true situation. It is not realistic to expect automated tools to identify such instances. It is probably even less realistic to expect users (*e.g.* lawyers) of such systems to critically question such interesting "facts."

4.3.4 Quantifying Visual Capabilities

Today's data analyst is very dependent on interactive analysis where numerical and graphical summaries are computed or displayed "on the fly". Successful instances of data mining by statisticians are often sprinkled with cries of "aha" whereby some subject matter (context) information, or unexpected behavior in a plot, is discovered in the course of the interaction with the data. This discovery can change the intended course of subsequent analysis steps in quite unpredictable ways. Assuming that it is a very hard problem to include common sense and context information in automated modeling systems, this leaves automated interpretation of plots as a promising area to explore. There are two problems that have served as a barrier to statisticians in this regard:

1. it is hard to quantify a procedure to capture the *unexpected* in plots.
2. even if this could be accomplished, one would need to describe how this maps into the next analysis step in the automated procedure.

What is sorely needed in the statisticians armory is a way to represent meta-knowledge about the problem at hand and about the procedures commonly used. This suggests an opportunity where the KDD and statistical communities can complement their skills and work together to provide an acceptable and powerful solution.

4.4 Statistical Methods

[4] A less modern but more realistic situation occurred in US politics when three major polls overwhelmingly projected Dewey over Truman in the 1948 presidential election — too bad for Dewey (the Republican) that there was a discrepancy between the voting public and those with phone service.

In this section we review some classical and more recent methods that are used in knowledge discovery problems where interest centers on a single response variable, y, and a collection of predictors, $\mathbf{x} = (x_1, x_2, ..., x_J)$. All the models assume the availability of training data, and the goal is to find a model to predict y from $\mathbf{x}$ that performs well on new data. This problem had a well-defined solution (least squares) for many decades until computing advances made it possible to relax classical assumptions. Statisticians have since been on a feeding frenzy devising new estimation methods (*e.g.,* M-estimates) and models (*e.g.,* additive models) to exploit the less restricted formulation.

Others have been caught up in the race to develop increasingly flexible models, perhaps encouraged by the famous result of Kolmogorov (1957) that *all* multi-dimensional functions can be represented by a composition of one-dimensional functions. But statisticians are not comforted by this result as any such class of models has far too much flexibility to be useful in practice where finite and noisy data prevail. We need models that scale up to real data, which due to its size and complexity (*e.g.,* missing values) beguiles all but the simplest of analyses.

Following discussion of classical linear methods and nonparametric techniques, we briefly describe five modeling algorithms selected to span "statistical method space": projection pursuit, neural networks, polynomial networks, decision trees, and adaptive splines.[5] While each can be treated as a "black box" (with a few "knobs" to set) that performs variable selection and feature extraction from a set of candidate inputs, we do not recommend such reckless abandon. Rather, careful modeling and familiarity with the subject matter domain can lead to greatly improved performance.

Several recent references are recommended for further information on this subject. Friedman (1995) provides an excellent overview of the major issues involved in building models from data, applicable to all induction techniques. Weiss and Kulikowski (1991) describe, in a very accessible manner, the basics of major inductive or "machine learning" classification techniques, including linear discriminant analysis, decision trees, neural networks, and expert systems. Useful (and more advanced) recent surveys focusing on neural networks and their statistical properties include those by Ripley (1993) and Cheng and Titterington (1994). Barron and Barron (1988) provide a unifying view of many methods as "statistical learning networks". A comparison of approximately twenty public-domain classification algorithms is summarized on a number of diverse applications in the European StatLog project (Michie, Spiegelhalter, and Taylor, 1994). Lastly, modern developments in statistical density estimation and data visualization are

[5] Clearly, tree methods dominate work in the KDD, machine learning, and expert system communities – and not without reason. Trees can be mapped into rules, can more easily handle categorical data and missing values, and are usually far more interpretable. However, the "smooth" methods deserve consideration where applicable, as their basis functions can often be more appropriate for the data, and thus lead to improved performance.

effectively presented by Scott (1992).

4.4.1 Linear Models

The classical models for prediction and classification are linear regression and linear discriminant analysis, respectively. The term "linear" in these models pertains primarily to the fact that the regression or classification surface is a plane — a linear combination of the available predictors (equations 4.3.4-5) (which may be nonlinear functions of the original data). The flexibility and straightforward computation involved in linear regression leads to its wide use within other techniques. For example, *radial basis function* networks are merely the linear regression of a set of *kernel* features – nonlinear functions of the separation of each case from several (potentially adaptively-selected) data centroids (next Section). Lowe and Webb (1991) employ a neural network architecture (Section 4.4.4) to compute nonlinear data features which feed into a final linear regression stage, and polynomial networks (Section 4.4.5) use linear regression in every node to combine previous (nonlinear) polynomial data transformations. Even linear discriminant analysis, with appropriate pre- and post- processing, can be formulated as a problem with a linear regression stage (Hastie, Tibshirani, and Buja, 1994). This allows one to replace the linear regression module with an advanced nonlinear/nonparametric estimation method, greatly increasing the types of patterns that can be handled by such classification techniques.

The models are also linear in a second important respect; namely, that the estimated parameters in the model are linear in the response variable, y. For example, in the usual linear regression model,

$$\hat{\beta}_j = \sum_{i=1}^{n} c_{ij} y_i. \tag{4.4.6}$$

This type of linearity enables an exact sampling theory for estimated model parameters (Section 4.2.2), unless the x's were selected during the course of the analysis (in which case a hard to untangle nonlinearity is involved).

4.4.2 Nonparametric Methods

Linear models are parametric methods; they replace sample data with a model representing a global "consensus" of the pattern the data represents (to the degree to which the patterns can be captured by the particular building blocks used; typically, lines or quadratic curves). Nonparametric, "model-free", or "code book" methods instead keep the data around and refer to it when estimating the response or the class of a new point. The simplest such method is *nearest neighbors*, which returns the response of

the closest known point (as measured in the input-variable, or design, space according to some distance metric, *e.g.*, Euclidean). The resulting estimation surface or discrimination boundaries are thus extremely responsive to local variations. To smooth these somewhat, the data set can be pared of unusual points, or the responses of the nearest K neighbors can be averaged. This simple method is often quite competitive in performance and asymptotically, as the data density increases, results in no worse than twice the Bayes optimal error (Cover and Hart, 1967).

Kernel estimation (*e.g.*, Parzen, 1962) provides a smoothed, generalized weighting of near-neighbors. A density function (*e.g.*, uniform, triangular, or normal) is centered over each point to be predicted.[6] The prediction is the kernel-weighted average of all the data. A single parameter representing the spread of the kernel can be adjusted to govern the roughness (local responsivity) of the prediction.

While such methods appear to be model-free, a type of model is implicit in the choice of distance function. Even if one considers Mahalanobis distance between two points i and j,

$$d_{ij} = (\mathbf{x_i} - \mathbf{x_j})^{\mathbf{T}} \mathbf{\Sigma^{-1}} (\mathbf{x_i} - \mathbf{x_j}) \tag{4.4.7}$$

there is considerable latitude in deciding which x's should enter into the distance calculation, and in what form (*e.g.,* x or $\log x$). Scaling issues, as reflected in Σ, can make or break the resulting prediction. As with all modeling methods, such issues need to be carefully considered and experimented with on the training data.

More so than parametric methods, nonparametric techniques are essentially constrained to operate in low dimensions; they depend heavily on *local* structure, and high-dimensional data is so sparse that "local neighborhoods are empty and non-empty neighborhoods are not local" (Scott, 1992). For example, for data uniformly distributed throughout a 10-dimensional unit cube, U^{10}[0,1], only 0.1% of the data is in a histogram bin of width 0.5 – not a very local neighborhood! Also, as dimension grows, nearly every point both views itself as an outlier with respect to the rest of the training data, and becomes closer to an outer boundary of the space than to its next nearest neighbor (Friedman, 1995). Thus, methodological intuition gained from experience in low dimensions is thoroughly out of place in high-D spaces – a phenomenon known as the "curse of dimensionality".

Still, when intelligently selecting variables to reduce the dimensionality to where samples can reasonably densely populate predictor space, these simple methods can work very well, often outperforming parametric methods. Accordingly, even when automated induction methods (described below) are employed, it is useful to examine the perfor-

[6] A rectangular kernel leads to a type of histogram with flexible bin edges.

mance of simple- or kernel-weighted nearest neighbors on the subset of variables selected for use by the adaptive algorithms.

4.4.3 Projection Pursuit

In low dimensions, the human ability to recognize patterns is unlikely to be matched by automata. Straightforward visual examination of data using histograms, scatterplots, and rotating 3-D plots, can often reveal structure which is missed by automated induction algorithms (Elder 1993b). Under the "grand tour' strategy (Asimov 1985) the data is rotated smoothly through all (or most) 2-D views, allowing one to discover interesting perspectives. However, the number of different views explodes exponentially with dimension, limiting such visual coverage methods to problems with a moderate number of candidate predictor variables. Accordingly, statisticians have sought to quantify measures of "interestingness" which can be optimized by the computer to identify revealing views in high-D.

In a procedure known as Exploratory Projection Pursuit (Friedman and Tukey, 1974), one searches for 1-D projections that maximally deviate from normality, robustly smoothes the data along that projection, and subtracts the smooth from the response. This process is repeated, projection by projection, until the error reduction cannot justify the added complexity. The anti-normal projection index is a reasonable one to employ since, regardless of the true density, most random projections of high-D data are normal (Diaconis and Freedman, 1984). Other exploratory projection indices are designed to seek holes or clusters. Projection Pursuit Regression (Friedman and Stuetzle, 1981) utilizes a maximal correlation index while maximal class separation is used for building a classifier.

However, it is very difficult to capture "interestingness" in a single criterion; structure which would be obvious to an analyst can be missed (see, *e.g.*, Elder 1994). Even if a particular quality could be well-quantified in an index, an analyst employing visualization has the advantage of "multiple end-points"; that is, of recognizing any of a wide variety of patterns encountered without explicitly choosing them as a goal beforehand. This weakness is shared by all automated modelling techniques to varying degrees, so to maximize performance, the automated search for structure in high-D space must be complemented by visualization of the lower-D manifolds discovered. (In practice therefore, techniques producing models with interpretable components have the additional advantage of speeding up the "design cycle" or entire iterative process of model development.)

4.4.4 Neural Networks

Artificial neural networks (ANNs) are a useful class of models consisting of layers of nodes, each implementing a linearly-weighted sum of its inputs with an adjustable sigmoidal (S-

shaped) output transformation as the bounded squashing function. The outputs of every node on a layer feed into each node on subsequent layers as their inputs. With the back-propagation weight adjustment procedure (*e.g.*, Werbos, 1974), cases are fed through one at a time, and errors are used to adjust the weights of the final output node to a degree proportional to their contribution (magnitude). Then, weights for nodes which feed into it are similarly adjusted, and so forth, back to the first layer. Initial weights are typically set randomly.

Statisticians have been suspicious of ANNs due to their overzealous promotion, but also because they appear so over-parameterized, and the weight adjustment procedure is a local gradient method (missing global optima), sequential (allowing early cases to have too much influence), and interminable (leading to a crude type of regularization wherein moderating the runtime becomes the principle way to avoid overfit). However, these weaknesses cancel somewhat, as the slow, local search doesn't allow the excess of parameters to be overfit easily. Note also that the true degrees of freedom employed by an ANN are usually fewer than at first glance. The danger of overfit can depend on the training duration, since many random node weights lead to essentially linear functions (nodes operating in either the middle or an extreme of the sigmoid) and such linear functions are absorbed by subsequent layers. Only as nodes get pushed into the curved part of the sigmoids during training do many parameters become active. (This may explain the common observation that the performance of an ANN on a problem is often surprisingly robust with respect to changes in its network structure.)

4.4.5 Polynomial Networks

Regression terms are often adaptively selected from a candidate pool in a forward stepwise (greedy) manner: choose the single best term, add the term which best combines with it, add the third term which works best with the pair, and so forth (occasionally deleting a term which is not useful enough) until the accuracy improvement is too small to justify the increment in complexity. The first polynomial network algorithm, the Group Method of Data Handling (GMDH) (Ivakhnenko, 1968; see also Farlow, 1984), expanded this idea by considering "chunks" of terms simultaneously. The GMDH uses linear regression to fit quadratic polynomial nodes to an output variable using all input variable pairs in turn. The best M nodes are retained as the first layer, and their outputs are the candidate inputs for the next layer, and so on, until complexity impairs performance on a checking set of data (whence the name). The best node on the final layer, and all nodes feeding into it, become the model, thereby forming a hierarchical composition of functions (a feed-forward network).

Considerable improvements to the GMDH approach were introduced in the 1970s and 1980's with versions of the Polynomial Network Training (PNETTR) algorithm (Barron

et al., 1984) and the Algorithm for Synthesis of Polynomial Networks (ASPN, Elder 1985). Some details of the history and methodology of these algorithms are presented in (Elder and Brown, 1995). Like ANNs, polynomial network estimation surfaces are smooth and global, but with nonlinearities entering through higher-order polynomial terms and cross-products, rather than sigmoids. The structure is adaptive rather than fixed, and the parameters are adjusted in sets, using all the data, rather than globally using one case at a time. Polynomial networks can take orders of magnitude less time to train than back-propagation ANNs (*e.g.,* Shewhart, 1992) and typically achieve better results (*e.g.,* Tenorio and Lee, 1989). However, also like ANNs, polynomial networks are rather opaque; they are difficult to dissect, unlike trees, which can be interpreted straightforwardly.

4.4.6 Decision Trees

The neural and polynomial network methods are global estimators, and hence will poorly estimate a function everywhere if it is sufficiently badly behaved anywhere (*e.g.*, deBoor, 1978). Decision trees, which recursively divide the space into different regions, instead have sharp breaks in their estimation surfaces, allowing great local responsivity. Also, the variables selected for splits may be different in each adaptively-partitioned region of the space. The flexibility of the method seems often, in practice, to make up for the crude basis function (a constant). Note that in a classification problem in which a variable (fortuitously) has a different constant value for each class, a decision tree could capture the rule perfectly, whereas classical linear discriminant analysis would actually encounter numerical instabilities, due to the negligible within-class pooled variance of the variable.

Decision trees were legitimized in the statistical community by the pioneering work on Classification and Regression Trees (CART) of Breiman *et al.* (1984). These authors neatly describe the problem and provide theory and methods to grow a tree and validate it. They depart from most previous work in that they propose expanding nodes until they reach a prescribed minimal size or are themselves pure. A cost-complexity parameter is introduced that characterizes a nested sequence of subtrees and cross-validation is used to decide how far back to prune the overly large initial tree. As with other statistical models, the usual precautions and careful pre- and post-fitting analysis are required. An advantage held by trees in this regard is that the tree metaphor can be exploited for graphical analysis.

Trees are natural for classification, but are also useful in difficult estimation problems where their simple piecewise-constant response surface and lack of smoothness constraints make them highly robust to outliers in either the predictors or the response variable. They automatically select variables, and construct models quite rapidly for an adaptive method. Importantly, trees are also probably the easiest model form to interpret (so

long as they are shallow) which, in our experience, greatly improves the model's chances of actually being used. The main problem with trees is that they devour data at a rate exponential with depth; so, to uncover complex structure, extensive data is required.

4.4.7 Adaptive Splines

The extreme local responsiveness of trees can sometimes be a disadvantage. Friedman's (1991) Multiple Adaptive Regression Splines (MARS) model employs recursive partitioning to locate product spline basis functions of adjustable degree, rather than constants. This results in smooth adaptive function approximation as opposed to the crude steps or plateaus provided by regression trees. The method also considers splines involving interactions between previously selected variables, so it can orient its basis functions on other than the original data axes. To aid interpretation, model terms are collected according to their inputs and their influence is reported in an ANOVA manner, namely, the effects of individual variables and pairs of variables are collected together and graphically presented as function plots. Like CART, MARS employs cross-validation, prunes terms after over-growing, and can handle categorical variables. As a new (and somewhat complex) method, there is less accumulated experience with its use, though it has been favorably compared with ANNs (*e.g.*, DeVeaux *et al.*, 1993) One would expect that enforcing continuity of the response surface (and perhaps that of its slope) will be very useful for applications which have a design space densely enough populated to support (and require) the local responsivity of the spline-like basis functions.

4.5 Statistical Computing

Arguably, no matter how brilliant the model or method to describe and summarize data, software is essential if a methodology is actually to be used. KDD and machine learning communities implicitly provide software as their methods are largely described algorithmically. Statisticians on the other hand, are perfectly capable of generating scores of models and methods with well defined operating characteristics (at least asymptotically) without ever writing a line of code. But these days are largely over and it is rare that statistical methods are described or promoted without application to data.

Early general purpose statistical packages included BMDP and SPSS, from the biomedical and social sciences, respectively. Of the two most important newer systems – SAS and S, SAS is most similar in style, containing many special "procedures" for standard statistical models. S was designed as a *language* to express statistical computations, rather than as a complete package. For example, two sample t-tests were built-in to SAS, while S simply provided a high-level language to express the relevant computations

(which could be assigned to a function for repeated usage). Both SAS and S are widely used for exploratory data analysis, modeling, and graphics. Each provides some degree of data management to remove that burden from users and can be extended to tailor methods for specific applications. Emerging useful packages include Lisp-Stat (Tierney, 1990), an S-inspired, object-oriented (Common Lisp) system that is being used in research circles, as well as a host of PC-based systems that demonstrate remarkable breadth and accessible interfaces.

It is still the case that specialized methods (such as discussed in the previous Section) appear first in isolation, rather than as part of a bigger system. (A useful repository for many such statistical research algorithms is the "StatLib" archive.[7]) For example, CART fits trees and trees alone.[8] Features of these stand-alone programs usually eventually make their way into more general systems, losing some efficiencies, but gaining the capabilities of an *integrated* data analysis environment essential to analysis quality and analyst productivity. Thus the CART-inspired implementation of tree-based models in the S language (Clark and Pregibon, 1990) not only allows users to manipulate their data in a variety of ways prior to fitting, but also provides an interactive graphical interface to the model and the opportunity to painlessly explore alternatives (*e.g.,* additive models).

The story is somewhat similar on the graphical front. Most stand-alone statistical graphics systems provide real-time dynamic motion that many find essential for exploring complex high-dimensional data sets. General purpose systems adequately handle most plots but lack the degree of specialization allowing friendly user interfaces or state-of-the-art graphics. (Lisp-Stat is an exception in that certain advanced features such as case-linking multiple plots are provided.) The XGobi system (Swayne, Cook, and Buja, 1992) provides a comprehensive projection and real-time motion tool set that includes "grand tours" and other "guided tours"; the graphics system can be used stand-alone or within a cooperative statistics system (*e.g.,* S).

Finally, we want to emphasize that computing is more to statistics than a vehicle for data analysis. It has revolutionized the field through the computational methodologies that statisticians now take for granted (*e.g.,* resampling methods, cross-validation, and Markov Chain Monte Carlo). We expect the influence of Computer Science on Statistics to increase in the future.

[7] Send the one line message "send index" to "statlib@lib.stat.cmu.edu" for contents and retrieval instructions.

[8] Though it's recently been offered as an optional module for a popular PC package.

4.6 Conclusions

The tendency of the statistical community to propagate uncertainty in their models through sampling distributions, their familiarity with the need to regularize models (trade off accuracy and complexity), and their dogged perseverance in checking model assumptions and stability (through residual and graphical analyses) are strengths. Still, alternative heuristic modeling techniques have gained in popularity partly as a way to "avoid statistics" yet still address challenging induction tasks. Statisticians should learn from this the need to do a better job of communicating the value of such considerations, as well as clarifying and streamlining ways of injecting extra-data information into the modeling process.

A great deal of work goes into identifying, gathering, cleaning, and labeling the data, into specifying the question(s) to be asked of it, and into finding the right way to view it (literally and figuratively) to discover useful patterns. Despite the central importance of actually modeling the data (the focus of this chapter) that stage can take up only a small proportion of the project effort. It is hard to conceive that the entire process will ever be automated. Increased automation has not absolved researchers of the need to think in statistical terms, including matching model assumptions to the problem, seeking interpretability, quantifying variance, regulating complexity to improve generalization, and keeping a lookout for the unexpected. However, modern statistical modeling tools do make it possible for an analyst to think about the problem at a higher level (by handling some routine or massive tasks), to try numerous approaches, to estimate the uncertainty of conclusions arising out of even complex processes, and to iterate through several stages of a solution design before settling on a representation scheme (or even a blend of them). When one is comparing KDD techniques, or attempting to extract the most out of a database, it makes sense to try some of these accessible modern statistical algorithms.

Bibliography

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In Proceedings of the Second International Symposium on Information Theory, eds. Petrov and Csaki, 267–281, Budapest: Kiado Academy.

Asimov, D. 1985. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal on Scientific and Statistical Computing* 6: 128–143.

Barron, A. R.; and Barron, R. L. 1988. Statistical Learning Networks: A Unifying View. In Proceedings of the Twentieth Symposium on the Interface: Computing

Science and Statistics, Reston, Virginia.

Barron, R. L.; Mucciardi, A. N.; Cook, F. J.; Craig, J. N.; and Barron, A. R. 1984. Adaptive Learning Networks: Development and Application in the United States of Algorithms Related to GMDH, Ch. 2 in *Self-Organizing Methods in Modeling: GMDH Type Algorithms,* ed. S. J. Farlow, 25–65. New York: Marcel Dekker.

Belsley, D. A.; Kuh, E.; and Welsch, R. E. 1980. *Regression Diagnostics: Identifying influential data and sources of collinearity.* New York: John Wiley & Sons.

Bishop, Y. M. M.; Fienberg, S. E.; and Holland, P. W. 1975. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: MIT Press.

Breiman, L. 1994a. Comment on "Neural Networks" by Cheng and Titterington, *Statistical Science* 9(1): 38–42.

Breiman, L. 1994b. Stacked Regressions, Technical Report 367, Dept. Statistics, UC Berkeley.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.* Monterey, California: Wadsworth & Brooks.

Cheng, B.; and Titterington, D. M. 1994. Neural Networks: A Review from a Statistical Perspective (with discussion). *Statistical Science* 9(1): 2–54.

Clark, L. A.; and Pregibon, D. 1992. Tree-based Models. Ch. 8 in *Statistical Models in S*, eds. J. M. Chambers and T. Hastie. Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books and Software.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical. Association* 74: 829–836.

Cover, T. M.; and Hart, P. E. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13: 21–27.

Cox, D. D., and John, S. 1993. A Statistical Method for Global Optimization. In Proceedings of the IEEE Systems, Man, and Cybernetics Society, Chicago, Oct.

deBoor, C. 1978. *A Practical Guide to Splines.* New York: Springer-Verlag.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39: 1–38.

DeVeaux, R. D.; Psichogios, D. C.; and Ungar, L. H. 1993. A Comparison of Two nonparametric Estimation Schemes: MARS and Neural Networks. *Computers in Chemical Engineering* 17(8): 819–837.

Diaconis, P.; and Freedman, D. 1984. Asymptotics of Graphical Projection Pursuit. *Annals of Statistics* 12: 793–815.

Efron, B.; and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* 37: 36–48.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7: 1–26.

Elder, J. F. IV 1994. Comment on "Prosection Views" by Furnas and Buja. *Journal of Computational and Graphical Statistics* 3(4): 355–362.

Elder, J. F. IV 1993a. Global R^d Optimization when Probes are Expensive: the GROPE Algorithm. Ph.D. diss., Dept. of Systems Engineering, University of Virginia, May.

Elder, J. F. IV 1993b. Assisting Inductive Modeling Through Visualization. In Proceedings of the Joint Statistical Meeting, San Francisco, California, Aug. 7-11.

Elder, J. F. IV 1985. User's Manual: ASPN: Algorithm for Synthesis of Polynomial Networks, Barron Associates, Inc., Stanardsville, Virginia. (4th Edition, 1989.)

Elder, J. F. IV; and Brown, D. E. 1995. Induction and Polynomial Networks. Ch. 3 in *Advances in Control Networks and Large Scale Parallel Distributed Processing Models (Vol. 2),* ed. M. D. Fraser. Norwood, New Jersey: Ablex. Forthcoming. (Available as Technical Report IPC-TR-92-9, University of Virginia.)

Faraway, J. J. 1991. On the cost of Data Analysis, Technical Report 199, Dept. Statistics, Univ. Michigan, Ann Arbor.

Farlow, S. J., ed. 1984. *Self-Organizing Methods in Modeling: GMDH Type Algorithms.* New York: Marcel Dekker.

Freedman, D.; Pisani, R.; and Purves, R. 1978. *Statistics.* New York: WW Norton & Co.

Friedman, J. H. 1995. An Overview of Predictive Learning and Function Approximation. Ch. 1 in *From Statistics to Neural Networks: Theory and Pattern Recognition Applications,* eds. V. Cherkassky, J. H. Friedman, and H. Wechsler, Springer.

Friedman, J. H. 1991. Multiple Adaptive Regression Splines (with discussion). *Annals of Statistics* 19: 1–141.

Friedman, J. H.; and Stuetzle, W. 1981. Projection Pursuit Regression. *Journal of the American Statistical Association* 76(376) 817–823.

Friedman, J. H.; and Tukey, J. W. 1974. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers* 23: 881–889.

Furnival, G. M.; and Wilson, R. W. 1974. Regression by leaps and bounds. *Technometrics* 16: 499–511.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 62: 1179–1186.

Hastie, T. and Pregibon, D. 1990. Shrinking Trees, Technical Report, AT&T Bell Laboratories.

Hastie, T.; and Tibshirani, R. 1990. *Generalized Additive Models.* London: Chapman & Hall.

Hastie, T.; Tibshirani, R.; and Buja, A. 1994. Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association* 89(428): 1255–1270.

Huber, P. J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35: 73–101.

Ivakhnenko, A. G. 1968. The Group Method of Data Handling – A Rival of the Method of Stochastic Approximation. *Soviet Automatic Control* 3: 43–71.

Kolmogorov, A. N. 1957. On the Representation of Continuous Functions of Several Variables by Superpositions of Continuous Functions of One Variable and Addition. *Dokladi* 114: 679–681.

Lowe, D.; and Webb, A. R. 1991. Optimized Feature Extraction and the Bayes Decision in Feed-Forward Classifier Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13: 355–364.

Mallows, C. L. 1973. Some Comments on Cp. *Technometrics* 15: 661–675.

McCullagh, P.; and Nelder, J. A. 1989. *Generalized Linear Models (2nd Ed.)* London: Chapman & Hall.

Michie, D.; Spiegelhalter, D. J.; and Taylor, C. C., eds. 1994. *Machine Learning, Neural and Statistical Classification.* New York: Ellis Horwood.

Miller, A. J. 1990. *Subset Selection in Regression.* New York: Chapman & Hall.

Mosteller, F.; and Tukey, J. W. 1977. *Data Analysis and Regression.* Reading Massachusetts: Addison-Wesley.

Nelder, J. A.; and Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society A* 135: 370–384.

O'Sullivan, F.; Yandell, B. S.; and Raynor, W. J. Jr. 1986. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* 81: 96–103.

Parzen, E. 1962. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics* 33: 1065–1076.

Ripley, B. 1993. Statistical Aspects of Neural Networks. In *Chaos and Networks – Statistical and Probabilistic Aspects,* eds. O. Barndorff-Nielsen, D. Cox, J. Jensen, and W. Kendall, London: Chapman & Hall.

Rissanen, J. 1978. Modeling by Shortest Data Description. *Automatica* 14: 465–471.

Schreuder, H. T. 1986. Quenouille's estimator. *Encyclopedia of Statistical Science* 7: 473–476. New York: John Wiley & Sons.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: Wiley.

Shewhart, M. 1992. A Neural-Network-Based Tool. *IEEE Spectrum* February: 6.

Swayne, D. F.; Cook, D.; and Buja, A. 1992. XGobi: Interactive Dynamic Graphics in the X Window System with a Link to S. In Proceedings of the 1991 American Statistical Association Meetings.

Tenorio, M. F., and Lee, W. T. 1989. Self-Organizing Neural Networks for the Identification Problem. In *Advances in Neural Information Processing Systems,* ed. D. S. Touretzky, 57–64. San Mateo, California: Morgan Kauffman.

Tierney, L. 1990. *LISP-STAT* New York: John Wiley & Sons.

Tukey, J. W. 1977. *Exploratory Data Analysis.* Reading, Massachusetts: Addison-Wesley.

Weiss, S. M.; and Kulikowski, C. A. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems.* San Mateo, California: Morgan Kaufmann.

Werbos, P. 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. diss., Harvard, August.

Wolpert, D. 1992. Stacked Generalization. *Neural Networks* 5: 241–259.