Report from Interface'98:
Knowledge Discovery and the Interface of Computing and Statistics

Arnold Goodman
The UCI Center for Statistical Consulting
University of California, Irvine
Irvine, CA 92697-5105
agoodman@uci.edu

1. INTRODUCTION

KDD is that proactive new area of information technology driven by enormous data bases with significant knowledge buried deep inside them. It's objective is to make useful sense out of data. The data is not experimental but opportunistic; it just happens to be there --with alluring potential. It can be diverse, heterogeneous, overwhelming and maybe even non-stationary in space and time. Necessity demands it be analyzed all together.

KDD analysis is application-oriented and driven by computation. Probability and statistics are there to help but not drive the process of getting results. Although this situation is on the fringe of statistical tradition, it is not outside the boundary of statistical techniques and statistical thinking. KDD is actually reminiscent of the real beginnings of statistics: there was data to be mined before there was a theory to guide it.

KDD did not intend to become statistics; yet, it is not only uses, but also contributes to statistics. It is on the interface of computing and statistics, and there is much to be achieved -- by both data miners and statisticians -- in bridging the technique and thinking gap between these two fields.

2. Historical Highlights

It was a 1965 UCI seminar by Arthur Samuel, on teaching the computer to play checkers, that inspired me to conceive of the need for meetings at the interface of computer science and statistics. This area is composed of those employing computers for statistical problems, using statistics in computer problems, and utilizing both of them on those significant problems of other important knowledge areas.

The first formal meeting, Interface '67, had sessions on computational linguistics, artificial intelligence, applications within the Interface, and computer simulation. In 1972, Barry Merrill at State Farm Insurance became the first commercial customer of SAS, when he mined the first data warehouse of SMF records from IBM mainframe computers. From the perspective of the statistical world, decision trees were provided to KDD by Leo Breiman, Jerry Friedman, Richard Olshen and Charles Stone in their 1984 book on the extremely useful CART technique.

Both Interface '97 and Interface '98 Keynote Addresses dealt with KDD: Jerry Friedman said that statistics is no longer the only data game in town, and David Rocke said that the algorithm is the estimator. KDD poses great opportunities, as well as great challenges, for statistics as a field. KDD-97 appeared at Interface '98, Interface'98 is appearing at KDD- 98, and KDD-98 will itself appear at Interface '99. A relationship is being developed between the two meetings, to lead toward increasing their interaction and perhaps a co-located KDD-01 and Interface '01 meeting. Interaction will produce progress, and perhaps, synergy.

3. Some Interface '98 Papers Relevant for KDD

To give an idea of the topics discussed, I've identified a few papers (with brief content annotations) from the last Interface that might interest those following KDD.

Internet-measurement:

- Vern Paxson, "Statistical Challenges in Analyzing the Internet" -- pooling diverse and heterogeneous data to find islands of stability
- John Quarterman, "Visualization of Internet Data" -- using geographical maps and graphs to measure internet quality of service from data
- Walter Willinger, "Finding Order within Chaos" -- using wavelets to identify and detect scaling properties in non-stationary space/time

Tree-based methods:

- David Banks, "Maximum Entropy Models for Graph-Valued Random Variables" -- model selection to reflect application metrics.
- Hugh Chipman, Edward George & Robert McCulloch, "Making Sense of a Forest of Trees" -- metrics to see archetypes and clusters.
- William Shannon, "Averaging Classification Tree Models" -- using maximum likelihood and consensus estimates of "mean and variance"
- Andreas Buja & Yung-Seop Lee, "Criteria for Growing Classification and Regression Trees" -- better interpretation by splitting data based on the better performing of subsamples
- Steven Ellis, Christine Waternaux, Xinhua Liu & J. John Mann, "Comparison of Classification and Regression Trees in S-Plus and CART" -- CART has tree evaluation, error estimation and subsampling while S-Plus has better computing, graphics and interpreting
- Douglas Hawkins & Bret Musser, "One Tree or a Forest? Alternative Dendrographic Models" -- comparing trees plus their generating rules via membership and metrics
- Padraic Neville, "Growing Trees for Naive Bayes and Score Card Models"

- S. Stanley Young & Andrew Rusinko III, "Data Mining of Large High Throughput Screening Data Sets" -- extending recursive partitioning, for relating chemical structure to biological activity, for many-many variables

Software technology:

- R. Douglas Martin & Michael Sannella, "An Iconic Programming Interface for S-Plus and Mathcad" -- using component technology to integrate data and computing across packages
- David Wishart, "Exploiting the Graphical User Interface in Statistical Software: The Next Generation" -- with filtering, wizards, tutoring, visuals, model design, multiple windows, what-if's and internet linkages
- David Woodruff, "Heuristic Search Algorithms: Applications in and of Statistics" -- for combinatorial analysis in multivariate problem robustness and cluster analysis

Density distributions:

- David Marchette & Carey Priebe, "Alternating Kernel and Mixture Density Estimation" -- a semi-parametric approach
- Michael Minnotte, "Higher Order Histosplines: New Directions in Bin Smoothing" -- an extension to multivariate, nonparametric and boundary effect applications
- David Scott, "On Fitting and Adapting of Density Estimates" -- using moments of bins and polynomial patches for kernel estimates
- Sung Ahn & Edward Wegman, "A Penalty Function Method for Simplifying Adaptive Mixtures Density Estimates" -- reduces the complexity for mixtures of normal densities

Modeling:

- Julian Faraway, "Data Splitting Strategies for Assessing Model Selection Effects on Inference" -- performance is no better than using all the data for selection and inference
- Hakbae Lee, "Exploring Binary Response Regression Based on Dimension Reduction" -- sliced inverse regression, sliced average variance estimation and covariance differences
- Wei Pan, "Bias/Variance Tradeoff in Combining Subsample Estimates for a Very Large Data Set"
- Armin Roehrl, "Fast, Portable, Predictable and Scalable Bootstrapping" -- using bulk synchronous parallel computing to bootstrap
- Terry Therneau, "Penalized Cox Model in S-Plus" -- with smoothing splines and frailty
- Jimmy Ye, "On Measuring and Correcting the Effects of Data Mining and Model Selection" -- generalized degrees of freedom framework for comparing the "costs" of alternative modeling and mining techniques.

4. REFERENCES

[1] Goodman, A. and Elder, J. 1998. Knowledge Discovery and the Interface of Computing and Statistics. Proceedings: The Fourth International Conference on Knowledge Discovery and Data Mining.

[2] Weisberg, S. ed. 1999. Proceedings of Interface '98: 30th Symposium on the Interface of Computing and Statistics.

About the author:

Arnold Goodman is Associate Director of the UCI Center for Statistical Consulting.and Founder of the 33-year-old Symposia on the Interface of Computing and Statistics. After receiving his PhD in Statistics from Stanford, he spent 35 years in information technology at Rockwell, McDonnell Douglas, Atlantic Richfield and County of Los Angeles, before cofounding the Center at UCI.