

CHALLENGES FOR DATA MINERS, STATISTICIANS AND CLIENTS WHO DEPEND UPON AND FUND THEM

Arnold Goodman, Associate Director, UCI Center for Statistical Consulting, agoodman@uci.edu

Our world is increasingly overwhelmed by the massive amounts of complex data begging for an effective method to be transformed into interpretable knowledge. Discovering such knowledge from the data requires the informative patterns mined from data to be generalized into predictive models (sufficient for most business or practical purposes) that first suggest knowledge and then facilitate its acceptance in the world beyond this data (a goal for important or scientific purposes).

Uncertainties inherent in both the collection and processing of data demand to be accounted for in prediction confidence intervals, hypothesis test levels and serious knowledge evaluations. Data mining and statistics will inevitably grow toward each other in the next ten years because data mining will not become knowledge discovery without statistical thinking, statistics will not be able to succeed on massive and complex datasets without data mining approaches, and both will be driven by clients who own the data and compel them to work together in knowledge discovery.

A basic challenge is for the results to work almost all (not only some) of the time, account for uncertainty outside (not only inside) the data, and add valuable knowledge to a client's world.

Data miners tend to be ignorant of statistics and client's domain, statisticians tend to be ignorant of data mining and client's domain, and clients tend to be ignorant of data mining and statistics. Unfortunately, they also tend to be inhibited by myopic points of view: computer scientists focus upon database manipulations and processing algorithms, statisticians focus upon identifying and handling uncertainties, and clients focus upon integrating knowledge into the knowledge domain.

Knowledge discovery rests on the three balanced legs of computer science, statistics and client knowledge: it will not stand either on one leg or on two legs, or even on three unbalanced legs. Successful knowledge discovery needs a substantial commitment to collaboration from all three.

A maturity challenge is for data miners, statisticians and clients to recognize their dependence on each other and for all of them to widen their focus until true collaboration becomes reality.

When John Tukey said "we should sit loosely in the saddle of the data" to a Stanford Statistics Seminar around 1959, he showed amazing insight into the essence of exploratory data analysis. This admonition, initially for statisticians alone, applies equally as well today to data miners and clients. Although far more effort is spent on processing data beyond what the data really support, far less effort is spent on planning the data selection and knowledge evaluation before acceptance.

An investment challenge is to balance effort spent on analysis inside the database with effort spent on analysis outside the database in the knowledge domain, difficult though it might be.

Lack of critical knowledge is not excused by insufficient time in an outdated curriculum or any difficulties in acquiring it through personal efforts. Data miners, statisticians and scientists can overcome these difficulties by reading [Principles of Data Mining](#) and [The Elements of Statistical Learning](#), attending Interface Symposia, and attracting that critical knowledge to their meetings.

A leadership challenge is to create an atmosphere appreciating critical external knowledge, and to place those who possess it at the core of its conferences rather than in the periphery.

[The Brain Makers](#) and [Mind Matters](#) document artificial intelligence going from over-promising in the early 1960's to under-performing in the early 1970's and expert systems going from over-promising in the early 1980's to under-performing in the early 1990's, despite vestiges remaining when computer technology is finally able to keep the promises of computer science. Is machine learning under-promising and over-performing because it has learned its lesson from this history?

The critical challenge for us all is to view the challenges as opportunities for our joint success.

Refinement of comments graciously distributed to 11,000 data miners by [kdnuggets](#).