

# DATA MINING, STATISTICAL METHODS MINING AND HISTORY OF STATISTICS

by Emanuel Parzen  
Department of Statistics, Texas A&M University  
College Station, TX 77843-3143

## 0. Abstract.

The emerging field of “data mining” is (according to Pregibon (1996)) a blend of statistics, artificial intelligence, and database research. It is projected to be a multi-billion dollar industry by the year 2000. I propose the name “statistical methods mining” for the process of developing (and marketing) various *maps of the world of statistical knowledge* in order to apply to data mining the enormous virtual encyclopaedia of statistical knowledge that exists in the literature. I propose that planning modern core, applied, and computational statistical activities should be based on statistical history whose mottos are “the purpose of statistical computing is insight not numbers” and “the purpose of statistical history is insight (about future influences) not details (of past influences)”. The goal of statistical methods mining is to promote the optimum practice of analysis of diverse (possibly massive) data sets by enabling researchers in all fields to be made aware that statisticians have important roles to play to provide expertise about classical and modern statistical methods, theory, practice, applications, and computational techniques for data mining and statistical inference (extraction of information and identification of models from data). Contents of this paper are: 1. Future of statistics is bright, future of statisticians needs planning; 2. What went wrong in the recent history of statistical science; 3. Balancing core research and interdisciplinary research; 4. Statistical education and statistical modeling are analogous

iterative cycles; 5. Statistical history was never neglected in past thinking about the future of statistics.

## 1. Future of statistics is bright, future of statisticians needs planning

Why should the future of statistics, statisticians, statistical methods, and statistical science deserve to be discussed at an Interface Symposium whose themes are data mining and the analysis of massive data sets?

I believe that data mining requires a diversity of scientific, computing, statistical, and numerical analytical skills to provide visual (graphical) diagnostic presentations that can be used to make decisions and discover relations between variables. In order to accurately distill information, we need to apply and integrate the extensive range of classical and modern statistical methods. We propose that researchers in data mining and massive data set analysis can benefit from *statistical methods mining* which I define as providing a vision of the diversity of statistical methods which enables one to defer learning their details until they are used in applications. A brilliant example of such a paper is Elder and Pregibon (1996).

I believe that we need to plan the future of statistics and statisticians in order to (1) stimulate a new explosion of statistical science, (2) explore analogies between probability methods (to describe the population or “infinite” data sets) and statistical methods for massive data sets, and (3) stimulate educational opportunities for applied researchers who use some statistical techniques, and for statisticians who work with non-statisticians using statistical methods.

I believe we need to plan (1) the future of statistical methods, in a world over-

whelmed with data, and (2) the future of statistical education and practice, in a world where computers and statistical computing packages can provide engineers, scientists, managers, professionals, and the public with technical proficiency in statistical methods which the health and prosperity of society increasingly requires and rewards.

*The future of statisticians must be planned, not forecasted.* Statistical methods mining seeks to provide the vision and expertise that can prevent statistical computing from being used to unwisely apply statistical methods to reach faulty conclusions.

Statisticians must plan for their future to avoid looming troubles (outlined by Jon Kettenring, President of the American Statistical Association, in *Amstat News* May 1996, p. 3):

Other disciplines and professions are seizing our opportunities (to teach introductory statistics, to make quality job one, to mine data, to analyze massive data sets).

Our graduate statistics programs need improvements to attract and train bright statisticians for successful careers as “whole” statisticians (see Kettenring (1995)) who can balance applied statistics (jointly practiced with researchers in other disciplines) and core (theoretical and mathematical) statistics.

In order for statisticians to be important players in science and technology and to have good jobs and careers, statisticians should continuously improve their public relations. Current image is described by Kettenring (1996): “The public jokes about us. Employers do not see us as idea people and ‘money makers’. Key decision makers forget or do not know what we have to offer.”

## 2. What went wrong in the recent history of Statistical Science

In the 21st century every individual, every organization, every discipline is expected to prepare formal plans for the future which justify their economic (not just their scholarly) relevance in a world of downsizing and outsourcing. To plan for the future of statistics and statisticians we propose the use of “statistical methods mining” based on “statistical historical thinking”, defined as the history of statistical applications, methods, theory, computations, and science, studied as guides to planning future influences rather than study of past influences.

We can identify the outstanding fact about the history of American statistics (according to ads for the American Statistical Association 30 minute documentary video “Statistical Science: 150 Years of Progress”); it is the explosion of statistical science that occurred around World War II. We propose planning that the *next outstanding historical event will be the explosion of statistical methods mining* (around the year 2000).

The American Statistical Association was founded in 1839 in Boston by intellectuals who wanted to improve the world by studying social problems and wanted to use statistics as a “tool” to privately collect and interpret data relevant to the social sciences. Billard (1997) gives a brilliant account of the history of J.A.S.A. until 1939. The Institute of Mathematical Statistics was founded in 1935 (Hogg (1986) and Hunter (1996)) by statisticians in universities and industry; they wanted to be experts on statistical methods and on mathematical statistics (the mathematical study of the methods used to analyze data).

During the World War II years 1940-1945 the mathematical statisticians were

able to use their unique talents to make tremendous contributions to the war effort (by research on war problems and the education of workers in industry in methods of quality control). In the period 1945-1970 federal research grants and the entrepreneurial leaders of the 1950-1960 generation of mathematical statisticians built great university departments of mathematical statistics focused on discipline generated research and graduate education.

*It is my opinion that we need to worry about planning today because of leadership failures by the 1950-1960 generation of leading university statisticians.* I accuse them of failing to seize opportunities to develop “swinging” statistics programs interacting with and serving other disciplines. Many anecdotes can be told about negative behavior that happened among statisticians.

There is much positive behavior by statisticians. The Committee on Applied and Theoretical Statistics of the National Research Council, and the National Institute of Statistical Sciences at Research Triangle Park in North Carolina (inspired by Okin and Sacks (1990)), deserve credit and appreciation for organizing activities that demonstrate how research by mathematical statisticians could be oriented towards discipline generated problems and aimed at applicable methods for important applied interdisciplinary research problems.

How can academic statisticians plan for more improvements? We need leadership to develop ways of defining the frontiers of research which would mentor statistical faculty to integrate scholarship in core (discipline generated) research, applied research, and computer science. We need to award prizes to recognize outstanding achievements in integrating core and outreach research (such as the ASA Wilks Award and the Texas

A&M Parzen Prize).

Statistics departments in the 1960s were slow to seek interdisciplinary links with computer science departments to help develop computational statistics. To raise our historical consciousness about computational statistics and statistical computing we need to teach the history (Goodman (1994)) of the Interface Foundation and the annual symposiums on the Interface of Statistics and Computer Science.

As a historical note on failures of academic statistical leadership we should include in our curriculum history of the development of statistical software packages. The SPSS statistical computing system originated in the 1960s among social science graduate students at Stanford University without any Statistics Department faculty member being aware that it was happening.

What went wrong in the golden age of mathematical statistics departments is stated by George Box (1980) in his discussion of a report on “Preparing Statisticians for Careers in Industry”; he writes that statistical core research and graduate education went wrong by ignoring the history of the influence of important practical problems in the development of general statistical methods. We list in Table A the examples quoted by Box.

The problem of introductory statistical education is to change the attitude of students that statistics is their least relevant course, with no relation to their career goals. Working engineers report being ignorant of statistics, despite having had a course in statistics. The blame for this disaster is often assigned to academic core (mathematical) statisticians, and the solution is alleged to be “outreach” (teach statistics through relevant applications and deny any relation to mathematical thinking). There are many

developments in progress (for example, Newton and Harville (1997)) that demonstrate that statistical computing will provide curricula with more consumer satisfaction. As we attempt to modernize statistical curricula, I recommend an approach based on teaching statistical methods mining and advertising the need to involve professional statisticians in the broad range of applied statistical practice, while making clear that there is a discipline of statistics with core versions of methods that are elegant and applicable.

### **3. Balancing core research and interdisciplinary research**

I define core statistical research to be about mathematically synthesizing ideas drawn from many analogous applications. The goal is to create general statistical methods that provide technology transfer between statistical innovations arising in different disciplines that apply statistical methods. Education in core statistics is needed to teach methods that are elegant and applicable, and help statisticians promote the case for their expertise to be part of teams in the broad range of applied statistical practice.

Statistical historical thinking teaches us that every core statistical method began with real practical problems which stimulated theoretical problems (discipline generated problems) for theoretical study. Can statisticians plan their futures if the history of the stimulus of applied research to core research is not an explicit part of our research, education, service, public relations? We need to continuously mentor “whole statisticians” who achieve in their careers a balance between applications (applied research), theory (abstract core research), and computing. We need to award prizes to

continuously increase recognition by scientific and academic publics (and academies of science) of the roles and existence of statisticians as “experts about the discipline and practice of statistics”.

A bright future for statistics requires that individual statisticians practice the principle of continuous improvement, emphasized by Ed Deming, which recommends to decision makers that every action should be judged by how well it prepares you for subsequent actions, and how well it is based on knowledge (data) rather than opinions.

An important area of planning for the future is designing statistics graduate curricula. In many statistical programs when past graduates are polled about what courses they found most valuable for their careers, they reply that they found very useful the courses on classical applied parametric statistical methods and found no use at all for their courses on probability and mathematical statistics. What are the implications for the future?

Students should be informed that they need theory courses to prepare for multiple careers. They should learn classical statistics, computational statistics, modern statistical inference (modern nonparametric statistical methods) as part of a balance among core, applied, and computational statistics which helps educate “whole statisticians”.

To educate whole statisticians, I propose an approach which teaches probability and mathematical statistics not as theory (with emphasis on proofs of well known theorems) but as *operational* skills for “hands on theory ” needed to develop and evaluate new statistical methods; I believe a suitable name for this type of course is “mathematical statistics for non-mathematical statisticians”. We list in Table C some of the great theorems every statistician should know.

One can demonstrate a strong link between the history of statistics and motivating statisticians to have a consumer level operational knowledge of probability and mathematical statistics. In “Breakthroughs in Statistics”, an excellent collection of historically significant papers, edited by Johnson and Kotz (1991), Beran writes (p. 566) in his introduction to Efron’s paper on bootstrap: “By the late 1970’s theoretical workers in robust statistics and nonparametrics were familiar with estimates as statistical functionals—that is, as functions of the empirical cdf. This background prepared the way for subsequent interpretation and analysis of a bootstrap distribution as a statistical functional.”

In a review of some of the many books recently published on the history of statistics, Feinberg (1992) describes two approaches, externalism and internalism. The external approach looks to the social roots and larger intellectual issues causing developments. The internal approach looks to the technical outcomes of developments, details of formulas and proofs.

In my approach to the study of the history of statistics, which views it as a guide to statistical methods mining and planning modern core, applied, and computational statistical research activities, I propose that two kinds of skill need to be balanced:

1. technical competence (analogous to internal skills), hard work to accomplish goals and implement decisions;
2. vision to imagine alternative courses of action (analogous to external skills), inspiration to infer information about where to guide one’s technical power.

(I am inspired by the two kinds of skills which I once heard the great entrepreneur Steve Ross state that he looked for, and found hard to find, as he tried to recruit great

managers of enterprises).

As an example of teaching graduate students how to balance technical and vision skills, one should teach skills about what a theorem says, and vision to recognize when the theorem can play a significant role in solving problems.

#### **4. Statistical education and statistical modeling are analogous iterative cycles**

Practicing statistical methods mining requires understanding about strategies for solving statistical problems and learning statistical methods. We propose that both require a cycle of steps which one usually repeats (iterates) several times before reaching a satisfactory conclusion.

The cycle of statistical model building (whose motto is “no model is true, only useful”) consists of four stages:

Stage 1 (S): *Specify* very general class of models.

Stage 2 (I): *Identify* tentative parametric model.

Stage 3 (E): *Estimate* parameters of tentative models.

Stage 4 (T): *Test* goodness of fit, diagnose improved models.

The PDCA cycle of statistical problem solving (called in quality circles Shewhart’s or Deming’s wheel) consists of four stages:

Stage 1 (P): Plan; pose the question, form expectations.

Stage 2 (D): Do; collect the data, make observations.

Stage 3 (C): Check; analyze the data, compare observations and expectation.

Stage 4 (A): Act; interpret the results, find the best theory or decision that

fits the data.

We think of both cycles as EOCI (Expect, Observe, Compare, Interpret).

Reformers of mathematics education recommend that teachers should communicate the four aspects of learning which cognitive sciences recommend for success:

1. simple recall,
2. algorithmic learning,
3. conceptual learning, and
4. problem solving strategies.

In statistical teaching we can make these cognitive concepts more concrete by teaching that statistical concepts (such as the sample mean or sample variance) have three aspects:

1. how to define it (mean of sample distribution);
2. how to compute it (average the sample quantile function (values arranged in increasing order));
3. how to interpret it (estimate location parameter of sample).

The fourth aspect of statistical learning consists of ideas about combining concepts to conduct an iterative statistical investigation whose output is data models, which can be applied to *simulate* more data with the same distribution as the originally observed data.

## **5. Statistical history was never neglected in past thinking about the future of statistics**

This paper is about the future of statistics in the 21st century and the potential of

statistical methods mining to stimulate a new explosion of statistical science. I would like to emphasize that we should read proceedings of past conferences on the future of statistics. We give some examples to show that historical thinking was never neglected.

From the Proceedings of the *Conference on Directions for Mathematical Statistics*, organized by Ghurye (1975) at the University of Alberta in 1974, we quote Mark Kac (p. 6) and Herbert Robbins (p. 116, a provocative essay “Wither mathematical statistics?”). **Kac:** Johannsen took a large number of beans, weighed them and constructed a histogram; the smooth curve fitted to this histogram was what my teacher introduced to us as the Quetelet curve. That was my first encounter with the normal distribution and the name Quetelet.

At this time I would like to make a small digression which has its own point to make. Quetelet was an extraordinarily interesting man, who was a student of Laplace and was the first to introduce statistical methodology into social science (even though he was trained as an astronomer); he was also the author of some early books in the area (*Lettres sur la Théorie des Probabilités*, 1846; *Physique Sociale*, 1st ed. 1835, 2nd ed. 1869; *Anthropométrie*, 1870). It might interest some of you to know that Quetelet was private tutor to the two princes of Saxe-Coburg, one of whom Prince Albert, later became Queen Victoria’s Consort. He was the first major governmental figure to try to introduce some kind of rational thinking into the operations of the government, and thus may be considered as the forefather of Operations Research. Now, the point of this digression is that if you look back at this connection between Astronomy and Operations Research, via Laplace, Quetelet and Prince Albert, you will realize the significance of the

Danish proverb I quoted at the beginning (“it is difficult to predict, especially the future”).

Anyway, coming back to Johannsen, he argued that if all individual characteristics are inheritable, then if we take the small beans and plant them, take the large ones and plant them, and plot separately the two histograms for the progeny of the small and the large beans, then we should again obtain Quetelet curves, one centered around the mean weight of the small beans used as progenitors and the other around that of the large ones. Now, he did carry out such an experiment and did draw those histograms, and discovered that the two curves were almost identical with the original one. Actually, there was a slight shift to the left for the small ones, so that by repeated selection one could separate out the two populations. Of course, we now know that it is possible to distinguish between the genetic and the environmental factors, because the mean is controlled by the genetic factor while the variance is controlled by the environmental.

I have mentioned the above example because it illustrates a kind of unity of scientific thought: here was a basic problem in biology which was solved by a rather simple idea, which on closer analysis turns out to have an underlying mathematical foundation; and when one thinks some more about it, one is able to put a great deal of quantitative flesh on this extraordinarily interesting and impressive qualitative skeleton.

**Robbins:** Mathematical statistics is relatively new. Two sources that I have found most informative concerning its ‘early’ history are Karl Pearson’s three-

volume work on the life, letters and labours of Sir Francis Galton, and the British statistical journals during the period 1920–40 that contain the famous controversies involving R. A. Fisher, J. Neyman and E. S. Pearson, and their sometimes bewildered contemporaries of lesser rank. It would be a most useful thing, especially during times when nothing really new and important seems to be going on, for students and professors to acquaint themselves with at least this much of the historical background of their subject. An intense preoccupation with the latest technical minutiae, and indifference to the social and intellectual forces of tradition and revolutionary change, combine to produce the Mandarinism that some would now say already characterizes academic statistical theory and is more likely to describe its immediate future.

Planning the future of statistics has been discussed by Tukey (1962), Watts (1968), Healy (1978), and McPherson (1989).

Our research (Parzen (1991)) on unification of statistical methods can be interpreted to be about maps of statistical methods and statistical methods mining. In future papers we will describe our ideas for maps of statistical methods using three coordinates: data notation; whole statistician (application, theory, computation); data inference.

Parzen (1993) discusses the interaction between theory and computation (hammers) and applications (nails). Examples of the role of statisticians: apply (1) the “hammer” of wavelets to the “nails” of nonparametrically estimating for real data changepoints and regressions (see Ogden (1997)), and (2) the hammer of linear programming to estimate regression quantiles.

## Bibliography

- Billard, Lynne. (1996) “A Voyage of Discovery,” *Journal of the American Statistical Association*, Vol. 92, No. 437, Presidential Address. pp. 1–12.
- Box, G. E. P. (1980) Comment (Preparing statisticians for careers in industry: report of ASA Section on Statistical Education committee on training of statisticians for industry). *The American Statistician*, 34, 65–80.
- Elder, John and Daryl Pregibon. (1996) “A statistical perspective on knowledge discovery in databases” in Usama M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, MIT Press, 83–113.
- Fienberg, S. E. (1992) A brief history of statistics in three and one-half chapters: a review essay. *Statistical Science* 7, 208–225.
- Ghurye, S. G. (1976) Proceedings of the conference on directions for mathematical statistics. *Advances in Applied Probability Supplement*, 7, September 1975.
- Goodman, Arnold (1975) “American Statistical Association,” *Encyclopedia of Information Science and Technology, Volume 1* (Edited by Jack Belzer, Albert Holzman and Allen Kent), Marcel Dekker.
- Goodman, Arnold, (1994) “Interface Insights: From Birth into the Next Century,” *Proceedings of Interface '93: 25th Symposium on the Interface of Computing Science and Statistics*, Interface Foundation of North America.
- Healy, M. J. R. (1976) Is statistics a science? *Journal of the Royal Statistical Society*,

- Series A*, **141**, pp. 385–393.
- Hogg, R. V. (1986) On the origins of the Institute of Mathematical Statistics. *Statistical Science* **1**, 285–291.
- Hunter, P. W. (1996) Drawing the boundaries: mathematical statistics in 20th-century America. *Historia Mathematica*, **23**, 7–30.
- Kettenring, J. R. (1995) What industry needs (Symposium on Modern Interdisciplinary University Statistics Education). *The American Statistician*, **49**, 2–4.
- Kettenring, Jon. (1996) “Strategic Planning – The Future of Statistics,” *AMSTAT News*, American Statistical Association, May 1996, No. 231, p. 3.
- Kotz, S. and N. L. Johnson, (eds.). (1991) *Breakthroughs in Statistics*. Vol. I, II. Springer, New York, Berlin.
- McPherson, G. (1989) “The Scientists’ View of Statistics—A Neglected Area,” *Journal of the Royal Statistical Society, A*, **152**, 221–240.
- Newton, H .J. and Jane Harville. (1997) *StatConcepts: A Visual Tour of Statistical Ideas*. Duxbury.
- Ogden, R. Todd. (1997) *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhäuser, Boston.
- Olkin, I. and Sacks, J. (1990) “Cross-disciplinary research in the Statistical Sciences,” Institute of Mathematical Statistics Panel Report. *Statistical Science*, **5**, 121–146.
- Parzen, Emanuel. (1991) “Unification of Statistical Methods for Continuous and

- Discrete Data,” *Proceedings Computer Science–Statistics INTERFACE '90*, (ed. C. Page and R. LePage), Springer Verlag: New York, 235–242.
- Parzen, Emanuel. (1993) “History of Statistics in Real Time: Hammers and Nails”, *Proceedings Computer Science-Statistics, Interface Foundation*, Vol. 24, 602–608.
- Pregibon, D. (1996) Data mining. *Statistical Computing and Graphics Newsletter*, 7, p. 8.
- Tukey, J. W. (1962) The future of data analysis. *Ann. Math. Statist.* **33**, 1–67.
- Watts, D. G. (1968) *The Future of Statistics*, Academic Press: New York.

**Table A. Practical Problems Motivating General Concepts (George Box (1980))**

Practical Problem	Investigator	Derived General Concept
Analysis of Asteroid Data. How far is it from Berlin to Potsdam?	Gauss	Least squares
Are planetary orbits randomly distributed?	Daniel Bernouilli	Hypothesis testing
What is the population of France?	Laplace	Ratio estimators
How to handle small samples of brewery data.	Gosset	$t$ test
Improving agricultural practice by using field trials.	Fisher	Design of experiments
Do potato varieties and fertilizers interact?	Fisher	Analysis of variance
Accounting for strange cycles in U.K. wheat prices.	Yule	Parametric time series models
Economic inspection (of ammunition).	Wald Barnard	Sequential tests
Need to perform large numbers of statistical tests in pharmaceutical industry before computers were available.	Wilcoxon	Nonparametric tests

**Table B: Disciplines Active in  
Statistical Collaborative Research**

Health Sciences  
  Medicine  
  Public Health and Epidemiology  
  Biostatistics  
  Cancer Clinical Trials  
  AIDS Clinical Trials  
Life Sciences  
  Biology  
  Ecology  
  Fisheries and Wildlife  
  Environmental Sciences  
  Toxicology and Pharmacology  
  Genetics  
  Entomology  
  Forest Science  
  Physiology  
Agriculture  
  Animal Science  
  Soils and Crop Sciences  
  Agricultural Economics  
  Veterinary Medicine  
  Food Science  
Behavioral and Social Sciences  
  Psychology, Cognitive Sciences  
  Economics, Econometrics  
  Education  
  Sociology  
  Political Science  
  Sample Survey  
  Government Statistics  
Physical, Chemical, Earth and Atmospheric Sciences  
  Chemistry, Chemometrics  
  Geology, Geophysics  
  Physics, Astronomy, Chaos, Fractals  
  Meteorology, Climate Research  
  Oceanography  
Engineering and Mathematical Sciences  
  Engineering  
  Artificial Intelligence  
  Neural Nets  
  Massive Data Sets Analysis  
  Operations Research and Reliability  
  Mathematics  
  Signal Processing  
  Image Analysis and Pattern Recognition  
  Industrial Statistics  
  Defense Statistical Standards  
  Hydrology  
Business Administration  
  Finance  
  Forecasting  
  Insurance  
Law

### Table C: Some statistical great truths (theorems can be practical)

Introductory graduate statistics courses should try to communicate some of the great statistical truths that demonstrate the important role of theory in the practice of statistics and data analysis.

- A. Why are sample sizes of opinion polls always approximately 1100, no matter how large the population being polled
- B. Law of averages (rate of decrease of variances of sample averages)
- C. Bell shaped curve (normal distribution)
- D. Why does statistical inference work; limit theorems for order statistics, extreme values.
- E. Time series analysis; understand persistence (correlation) to diagnose spurious non-zero correlation, spurious non-zero mean, and spurious trend.
- F. Functional statistical inference: density estimation to diagnose spurious zero correlation, spurious lack of relationship.
- G. Accuracy and reliability of lab tests (Bayes reasoning).
- H. A statistical theorem of which many methods are extensions. To cut a path through the jungle of statistical methods, one needs analogies between analogies which make it easy to comprehend a diversity of methods. I claim that a unifying analogy is provided by the basic identities: for random variables  $X, Y$  and constant  $\theta$ ;

$$E [|X - \theta|^2] = \text{VAR}[X] + \{E[X] - \theta\}^2$$

$$\text{VAR}[Y] = E[\text{VAR}[Y|X]] + \text{VAR}[E[Y|X]].$$

- I. What is a Brownian Bridge, and what are their applications in statistical theory and practice.

## Table D

### Glossary of Modern Statistical Methods Mining

Robust and resistant  
Exploratory data analysis  
Generalized linear models  
EM algorithm  
Loglinear models  
Bootstrap  
Borrowing strength  
Regularization  
Shrinkage  
Simpson's paradox  
Automatic modeling criteria  
Density estimation  
Projection pursuit  
Decision trees, CART  
Adaptive splines, MARS  
Gibbs sampling  
Causal inference  
Wavelets  
Quantile domain  
Long tailed distributions  
Changepoint Analysis  
Spectral estimation  
Time Series Prediction  
Kalman filtering  
Long memory time series, fractals  
Clustering  
Neural networks  
Data mining