

# Artificial Intelligence Research in Statistics

William A. Gale

Daryl Pregibon

AT&T Bell Laboratories, Murray Hill, NJ 07974

THE INITIAL RESULTS from a few AI research projects in statistics have been quite interesting to statisticians: Feasibility demonstration systems have been built at Stanford University, AT&T Bell Laboratories, and the University of Edinburgh. Several more design studies have been completed. A conference devoted to expert systems in statistics was sponsored by the Royal Statistical Society.

On the other hand, statistics as a domain may be of particular interest to AI researchers, for it offers both tasks well suited to current AI capabilities and tasks requiring development of new AI techniques.

Statisticians do a variety of tasks, some of which can now be assisted by expert system techniques. One common task is data analysis—application of statistical tools to a particular set of data to reach some conclusion. Another is experiment design—planning the collection of data so that it can be easily analyzed to reach a conclusion. These tasks are frequently done in a consulting environment. Expert system techniques have recently been applied to these two tasks.

The statistics discipline has been developing for the past 100 years, most dramatically since the widespread availability of modern computers. Some of the knowledge developed by research statisticians is regularly taught—for example the lore of the normal distribution, student's t-test, analysis of variance methods, and linear regression analysis. But some statistical knowledge is not yet so formalized—specifically how these and other methods are chosen and applied to analyze data in practice (which process we call *strategy*).

The formalized knowledge indicates depth in the domain, while the informal knowledge indicates an opportunity for AI techniques. Specifically, it appears that expert system techniques can provide a framework for formalizing strategies of data analysis, thereby opening the subject to research by statisticians. This prospect has excited many statisticians.

The existence of extensive statistical packages suggests that AI contributions to data analysis will be intelligent *interfaces*, an area of little development. These interfaces could provide guidance, interpretation, and instruction, which are needed by novice users and are not available in current packages.

Like AI, statistics is a study of tools for use in other domains. This attribute implies that knowledge from three domains impinges on intelligent statistical analysis systems—AI, statistics, and the “ground” domain. While current systems suggest it is possible to rely on a user to supply ground

domain knowledge, more intelligent treatment of the ground domain will require development of knowledge representation and acquisition techniques. This is a key challenge to AI techniques offered by the statistics domain.

The next three sections review specific projects applying AI methods in statistics. A section reviews the talks at the RSS meeting, and the last two sections focus on the challenges to AI methods posed by this domain.

## The Rx Project.

Robert L. Blum (1982, 1983) has been leading the RX project at the Stanford Heuristic Programming Project. RX aims to design and perform statistical analyses in medicine to establish causal relationships. The project requires:

- interfacing AI software to a large data base as well as to a statistics package
- representing knowledge in medicine and statistics
- developing techniques for:
  - automatic hypothesis generation
  - study design
  - data analysis.

A feasibility demonstration system has been constructed using a subset of a data base collected by the American Rheumatism Association and an interface to IDL (Kaplan, 1978). The study design module is the most elaborated, using stored knowledge of confounding relationships to derive a confirmatory test for a proposed causal hypothesis. The discovery module and data analysis module are relatively weak. A key feature of the system is its ability to incorporate newly confirmed causal relationships into the knowledge base.

## REX.

We (Gale & Pregibon, 1982) have built a Regression EXpert [REX] system at AT&T Bell Laboratories, continuing preliminary work (Chambers, 1981; Chambers, Pregibon, and Zayas, 1981). The goal of REX is to allow novices to perform regression analyses safely by using advanced tools to detect and correct violations of assumptions made by standard techniques. The project requires:

- interacting with a novice statistician
- interfacing to a statistics package
- representing knowledge in statistics
- and developing techniques in
  - statistical strategy
  - interpretation of results
  - tutoring

A feasibility demonstration system has been constructed using the S system (Becker and Chambers, 1984) as the underlying statistical package. Statistical knowledge is represented in a frame-based system modeled after Centaur (Aikins, 1980). A strategy has been implemented for regression analysis, an area of great interest to statisticians. The strategy implemented focuses on checking the assumptions implicit in an initial model, and suggesting changes to data, model, or fitting method as may be appropriate when an assumption fails. The interpretation of results and tutoring are relatively weak.

The areas covered by RX and REX are complementary at this point. REX does no design; RX runs its tests blindly.

### Other Projects.

Richard O'Keefe (1982) built the Automated Statistical Analyst [ASA] system at Edinburgh University as part of his thesis work. ASA is designed to help a client analyze an experiment he has already designed but has not yet performed. It assumes a particularly unhelpful (ignorant) user and attempts to guide him to an appropriate nonparametric analysis.

Neil C. Rowe (1982), Stanford Department of Computer Science, considered how one might abstract a statistical data base. His goal was to provide approximate answers to statistical queries of a large data base without actually accessing the data base. He developed means of forming and updating an abstract of a data base, and ways to reason with the abstract to provide answers with a storable accuracy.

George Lawton (1982), U.S. Army Research Institute, has designed an intelligent interface to SAS (Helwig, 1979) as an initial step in a projected Computer Assisted Research Design System [CARDS]. CARDS is a long-term project. The initial step is a tool to assist users in quantitative research integration—standardization and comparison of statistical results from different studies.

A system built by Jean Louis Roos (1983), University of Aix en Provence, shows heavy use of ground domain knowledge. The system guides the economist user in the construction of an econometric model with consistent economic assumptions for all equations. The model is then estimated by a standard statistical package without feedback to the expert system.

Petr Hajek and Tomas Havranek (1982), Czechoslovak Academy of Science, have described requirements for GUHA-80. Their aim is to develop interesting views of empirical data. While they expect to have a statistical package available to the program for hypothesis testing, the emphasis is on hypothesis formation. The intention is to try automatic hypothesis formation ideas, as in Lenat's (Davis and Lenat, 1982) AM program.

Gerald J. Hahn (1983), General Electric, prepared a review of statistical consultation systems for the 1983 meeting of the American Statistical Association. The review included an example strategy for product life data, including flow charts.

David J. Hand (1983), London Institute of Psychiatry,

has written on the requirements of an expert system in statistical consultation. He relates observations of actual statistical consultation processes, discusses differences between medical and statistical requirements, and gives a useful list of attributes for a statistical consultation system.

### RSS Meeting On Expert Systems In Statistics

The most overt sign of interest among statisticians has been a meeting convened by the Royal Statistical Society on Expert Systems in Statistics, held in London on Saturday, October 22, 1983.

Two of the talks dealt with using standard programming methods to make much more helpful interfaces to existing software packages. A. Baines, University of Leeds, has begun work on these lines, and G. B. Wetherill, Department Chairman, University of Kent, has made progress along these lines. Their work shows the perceived need for assistance in managing today's statistical packages and will provide a comparison point for the value of using AI techniques in rendering that assistance.

The remaining four talks were based on various experiences with AI techniques. Alan Bundy, University of Edinburgh, speaking first and representing AI research more than any other speaker, spent time on survey material, then described his work on Mecho and Richard O'Keefe's work on the Automated Statistical Analyst. Daryl Pregibon spoke on our experience using REX to develop a strategy for regression analysis. David Hand, London Institute of Psychiatry, presented his views on design criteria for expert systems in statistics based on a carefully reasoned appraisal of differences between statistics and other areas where expert systems have been used. J. Fox, Imperial Cancer Research Fund, addressed inference techniques used in existing expert systems (*i.e.*, Statistics in AI). He described knowledge-based techniques used in medicine and also the use of quasi-statistical techniques in decision making. His defense of non-Bayesian techniques was roundly attacked and stoutly maintained.

### Challenges For AI in Statistics

Analysis of data is done entirely with computer tools now that interactive graphical software is widely available. Many statistical packages exist, including large ones such as SAS and SPSS (Hull, 1981) and smaller ones such as Minitab (Ryan, 1981). Special mention should be made of IDL, since it is written in Interlisp and may therefore be particularly easy for some AI projects to use. IDL is a relatively small package, however. These packages are widely used and, in the eyes of statisticians, abused. The abuse is based on ignorance of statistics and provides an opportunity for the application of expert system techniques.

For the expert statistician, current packages are very helpful. They can compute a large variety of statistics for a data set, manage the storage of all data sets, provide nice graphical displays easily, and be extended easily as a new statistical methodology is developed. They require little *programming* expertise from the user.

But for the novice, current packages are lacking important features. They provide only numbers, not interpretations. They provide no guidance on to do next, or what should have been done before. Moreover, they provide no instruction. These areas—guidance, interpretation, and instruction—provide feasible targets for application of natural language techniques and instructional techniques as well as expert system techniques.

The existence of these large and powerful statistical packages provides an opportunity and a challenge to AI. The addition of an interface to existing software is always a high leverage opportunity. The AI challenge is to provide access to the software while understanding what is being done. The problem is that an expert human can easily do things that have a purpose, but whose purpose is obscure.

Another research direction in statistical packages will test the value of explicit knowledge representation. This direction uses standard techniques to make the packages easy to use. The UKCREG program being developed under direction of G. B. Wetherill (1982) is a good example. This system uses menus to show what can be done and to indicate why it might be good to do something. It responds with an interpretation of test results rather than a number. However, it does not keep the knowledge of the tests made for its own use, and it does not suggest actions to the user. This program can certainly be called “friendly,” if not “intelligent.”

Since statistics is a study of tools for use in other domains, statistical methods can be used to greater effect when informed by domain knowledge, just as AI has found power by restricting attention to particular domains. Therefore applications of AI in statistics may involve three domains of knowledge—AI, statistics, and the “ground” domain. Approaches to the ground domain have been varied. RX has taken medicine as a specific ground domain and has undertaken to represent knowledge from this domain. Since this approach requires expertise in three areas, and since it may limit the applicability of the system, most other projects have left the ground domain unspecified.

If an interactive system is provided, the user can be expected to provide the required ground domain knowledge. The apparent success of this approach by REX and ASA suggests that the statistics knowledge is fairly well closed, self-consistent, and separable from ground domain knowledge. This is helpful for applications of current expert system technology.

A limitation of not using ground domain vocabulary is the necessity of using Statistics vocabulary. This necessity in turn requires that the user be willing to learn statistics vocabulary (and concepts) and that the system be prepared to provide such instruction. While this task may be reasonable for technical workers, it will exclude most managers. Not using ground domain knowledge and procedures may be fatal in some cases. A possible means to including ground domain knowledge and procedures would assume a local statistician who could specialize a statistics-knowledgeable system

to local ground domain knowledge, procedures, and vocabulary. Providing a way to do this is a challenge to knowledge-acquisition techniques including machine learning.

No analysis is complete without a written report. The capability of generating a report from the trace of an analysis would be a useful area to apply natural language generation techniques.

Once it is possible to consult with a user on several kinds of data analysis, the problem will arise of deciding which analysis model the user should be using. In standard consultation practice, this crucial step is usually accomplished by a lot of give and take as the analyst and the consultant strive to understand each other. The analyst frequently does not understand the categories available, while the consultant may flounder in the ground domain vocabulary in which the problem is presented. Handling this process by computer will require substantial development of interactive discourse techniques and substantial tolerance for erroneous inputs.

Theoretical analysis of tests seems to be a longer range goal for AI techniques. Contributions here will require representing substantial amounts of knowledge in mathematical statistics. Some short-range progress might be made in automating studies done by Monte Carlo techniques.

### Future Directions

Many sub-domains of data analysis exist, *e.g.*, regression analysis and analysis of variance. Developing strategies in these areas will be one future activity for statisticians. They can be expected to require additional AI techniques as well.

Future consultation systems should include both an experiment design phase and a data analysis phase. It is necessary to accommodate experiments designed without the systems aid, however, and since the users’ understanding of the design may be poor, this task will be difficult.

No analysis is complete without a report. The reports now generated are crude and mechanical. Data structures exist to support far more polished reports if natural language generation capabilities are used.

Research at AT&T Bell Laboratories will include an examination of whether a statistician can develop a strategy without involving a “knowledge engineer.” Since data analysis is done entirely on a computer, it may be possible to devise a system that can learn by watching and questioning, compiling a strategy as an output. We are exploring the possibilities with a system named Student.

In summary, we look forward to increasing activity by both statisticians and Artificial Intelligence researchers applying AI methods in Statistics.

### References

- Aikins, J. S. (1980) Prototypes & Production Rules HPP-80-17, Stanford University.
- Becker, R. A. & Chambers, J. M. (1984), *S—An Interactive Environment for Data Analysis and Graphics*, Wadsworth International, Belmont, (available from Computer Information Service, AT&T Bell Laboratories: Murray Hill, NJ.)

- Blum, R. L. (1982) *Discovery and representation of causal relationships from a large time-oriented clinical database: The RX project*. New York: Springer-Verlag
- Blum, R. L. (1982) Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical database: The RX project. *Computers and Biomedical Research* 15(2), 164-187.
- Blum, R. L. (1983) Representation of empirically derived causal relationships. *IJCAI* 3.
- Chambers, J. M. (1981) Some Thoughts on Expert Software. In W. F. Eddy, (Ed.) *Computer Science and Statistics: Proc. of 13th Symposium on the Interface*. New York: Springer-Verlag, 36-40.
- Chambers, J. M., Pregibon, D. & Zayas, E. (1981) "Expert Software for Data Analysis: An Initial Experiment," Invited Paper, 43rd session of the International Statistical Institute, I. P. 18 2. Buenos Aires, Argentina.
- Davis, R. & Lenat, D. B. (1982) *Knowledge-Based Systems in Artificial Intelligence*. New York: McGraw-Hill.
- Gale, W. A. & Pregibon, D. (1982), An Expert System for Regression Analysis. In Heiner, K. W. (Ed.) *Computer Science and Statistics; Proc. of 14th Symposium on the Interface*, New York: Springer-Verlag.
- Hahn, G. J. (1983), Expert Systems for Statistical Consulting. Prepared for Annual Meeting American Statistical Association, Toronto, Canada
- Hajek, P. & Haveranek, T., (1982) GUHA-80 An Application of Artificial Intelligence to Data Analysis. *Pocitace a umela intelgenca* (1) 107-134
- Hand, D. J. (1983) *Statistical Expert System Design*. Institute of Psychiatry, DeCrespigny Park, London SE5 8AF.
- Helwig, J. T., & Council, K. A. (1979), *The SAS User's Guide, 1979 Edition*. Cary, NC: SAS Institute Inc.
- Hull, C. H., & Nie, N. H. (1981), *SPSS, Second Edition*. New York: McGraw Hill.
- Kaplan, R. M., Sheil, B. A. & Smith, E. R. (1978) *The Interactive Data-analysis Language Reference Manual*. Xerox Corporation, Palo Alto Research Center.
- Lawton, G. (1983) *Development of an Intelligent Computer Assisted Research Design System*. In 1983 Conference on Artificial Intelligence, Oakland University, Rochester Michigan (*proceedings in preparation*)
- O'Keefe, R. (1982), An Expert System for Statistics, presented at *Theory and Practice of Knowledge Based Systems* 14 September, 1982. Brunel University, Egham, Surrey, England
- Roos, J. L. (1983) A Knowledge Based System for Econometric Modeling, preprint.
- Rowe, N. C. (1982) Rule Based Statistical Calculations on a 'Database Abstract' In *Proceedings of the First LBL Workshop on Statistical Database Management*, Lawrence Berkeley Laboratory, University of California.
- Ryan, T. A., Joiner, B. L., & Ryan, B. F. (1981) *Minitab Reference Manual*. University Park: Minitab Project, Pennsylvania State University.
- Wetherill, G. B., UKC Regression Group (1982) A General Program for Multiple Regression and Response Surface Analysis entitled UKCREG. Mond Division, ICI Ltd., Runcorn, England



## COMPUTER SCIENCE AT SHELL

Shell is a technology-driven company. To meet Shell's computing needs we operate one of the largest technical/commercial data processing centers in the world. Computer Science research at Shell Development Company is performed by a small group with high impact and visibility, backed by the resources of a large and progressive corporation. To capitalize on the emerging computer and communications technologies, this group has challenging R&D opportunities in the following areas:

### Artificial Intelligence

- Expert systems development
- Tool and language evaluation

### Database

- Super-large spatial and numerical databases
- Distributed multi-database integration
- Design methodologies

### Operating Systems

- Optical storage and access methods for large data sets
- Functional processing (object-oriented systems)
- Application-specific languages

### Networking

- Local area networks
- Wide area networks
- High bandwidth channel extensions

### Graphics Workstations

- Advanced 2-D and 3-D displays
- Bitmap graphics
- Color plotting
- VLSI

### Large-Scale Scientific Computing

- Novel computer architecture designs
- Algorithms for highly parallel computing

Our central computing environment includes several large UNIVAC computers, each equipped with a

memory-attached Array Processor (120 MFLOPS) for geophysical processing, and many large IBM systems for commercial processing and time-sharing activities. In addition, there are numerous VAX 750's, VAX 780's, and IBM 4341's for special applications. The Computer Science group has a dedicated research facility consisting of a VAX 750, a VAX 780, an IBM 4341-II and many other smaller systems. A wide variety of systems and software is currently in use, including: UNIX\*, VMS, VM/CMS, C, FORTRAN, and LISP.

We are seeking Computer Science professionals with advanced degrees (PhD preferred, or experienced MS) to contribute in any of the areas listed above. Interested parties should contact: Shell Development Company, Research Recruitment, P.O. Box 1380, Houston, Texas 77001, or submit inquiries to uucp address... Shell!nielsen

## Shell Development Company

An Equal Opportunity Employer M/F

\*UNIX is a trademark of AT&T