NISS

Workshop Report: Workshop on Statistics and Information Technology

Alan F. Karr, Jaeyong Lee, and Ashish P. Sanil

Technical Report Number 118
September, 2001

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

National Institute of Statistical Sciences

PO Box 14006 Research Triangle Park, NC 27709–4006 Tel 919.685.9300
FAX 919.685.9310

WORKSHOP REPORT

Workshop on Statistics and Information Technology
November 11–12, 1999

Alan F. Karr, Jaeyong Lee and Ashish Sanil
March 20, 2000

1 Introduction

This is the report of a "Workshop on Statistics and Information Technology" held on November 11-12, 1999 (Hurricane Floyd forced a postponement from the originally scheduled dates of September 16–17, 1999), at the NISS building in Research Triangle Park, NC. The purpose of the workshop was to expose statisticians to researchers and research problems in information technology (IT). The format was to present five important problem areas in IT from the *perspectives of both IT and statistics*, coupled with ample opportunity for questions and discussion.

The workshop responded, as does the Information Technology Research (ITR) program of the NSF, to the 1999 report of the President's Information Technology Advisory Council (PITAC). The PITAC report, available on-line at `www.ccic.gov/ac/report/`, pointedly notes that the future of our society depends heavily on IT. That the transformations to individual lives and society brought about by IT will be timely and effective is, however, not certain. Indeed, the report presents a compelling case that economic and social benefits of IT cannot be realized without massive, continuing investment in IT research and development.

Many problems that underlie the PITAC report are *inherently statistical* — they are driven by data, models and evaluation. However, the implications for the discipline of statistics are only beginning to be articulated, and the workshop was meant to accelerate that process.

Some general implications of IT for statistics seemed clear beforehand, and were borne out by the Workshop:

- In collaboration with the other disciplines, new statistical methods must be created that work in the staggeringly complex settings of the future.
- Scalable statistical techniques must be developed to cope with vast amounts of data of disparate types (for example, streaming audio and video) and varying quality.
- Methodology is necessary that merges, combines and assimilates data from laboratory experiments, observational studies and numerical (computer) models. Ultimately, such integration must happen substantially automatically.

- Risks and uncertainties must be quantified in increasingly complex and often non-standard settings.
- Comprehensible presentation and visualization of results to multiple audiences are crucial.

The goal of the Workshop was to examine these implications in specific settings, and to draw from them an agenda for future cross-disciplinary collaboration among researchers from IT and statistics.

The Workshop was organized by committee consisting of Melvyn Ciment (Potomac Institute for Policy Studies), Sallie Keller–McNulty (Los Alamos National Laboratory) and Jerome Sacks (NISS).

Approximately twenty-five people participated in the Workshop; their names, affiliations and E-mail addresses appear in the Appendix.

Financial support provided by the National Science Foundation is gratefully acknowledged.

2 Workshop Presentations

The Workshop consisted of two days of expository presentations and associated discussion, highlighting the five topics listed below. Each presentation was dual in nature. First, a researcher in information technology (computer science, electrical engineering, or software engineering) introduced the topic, its language and key research needs. Then, a researcher in statistics responded with a statistical interpretation of the IT issues and outline of the research implications for statistics. Each pair of presentations was interspersed with and followed by extended discussion.

The five focus areas, and the speakers and their backgrounds, were:

Human–Computer Interaction (§3.1 and 4.1), focusing on user interfaces to complex data and models, and on visualization of key outputs and results.

Information Science	Gary Marchionini, University of North Carolina at Chapel Hill
Visualization	Stephen G. Eick, Visual Insights, and Alan F. Karr, NISS

Digital Government (§3.2 and 4.2), focusing on methodology and technology for Web-based, disclosure-limited dissemination of statistical analyses based on confidential Federal data sets.

Computer Science/Security	Latanya Sweeney, Carnegie Mellon University
Statistics	Alan F. Karr, NISS

Data–Model Integration (§3.3 and 4.3), addressing systems that respond to queries by dynamically identifying, accessing and fusing relevant data, model components (computer simulations and statistical models) and computing resources.

Computational Linguistics	Vasileios Hatzivassiloglou, Columbia University
Statistics	Edward J. Wegman, George Mason University
Statistics	Sallie Keller–McNulty, Los Alamos National Laboratory

Internet Traffic Measurement and Analysis (§3.4 and 4.4), addressing needs for measurement and modeling of traffic in data networks, for example, in order to characterize quality of service for emerging applications such as streaming video.

Electrical Engineering	Daniel Stevenson, MCNC
Statistical Modeling	J. Stephen Marron, University of North Carolina at Chapel Hill

Software Development (§3.5 and 4.5), emphasizing assembly of software from components and prediction of the performance of complex software systems.

Software Engineering	Robert Horgan, Telcordia Technologies
Statistics	Siddhartha Dalal, Telcordia Technologies

PowerPoint versions of most of the presentations can be viewed on the NISS Web site: `www.niss.org/itworkshop/presentationindex.html`.

3 Summaries of Presentations

3.1 Human–Computer Interaction

Gary Marchionini: Human–Computer Interfaces for Statistical Data. The talk revolved around challenges in developing "human-centered" designs for human–computer interfaces, with a particular focus on the human–computer interactions involved in understanding tabular summaries of statistical data. Unlike the machine side of such systems (which is both relatively easy to understand and highly developed) human–centric designs are plagued with problems that stem from our limited ability to understand and classify human behavior.

Currently, researchers rely on crude guidelines and trial-and-iteration methods. Obstacles to be overcome in order to rectify this situation include:

- Imprecise articulation of the characteristics (such as beliefs, preferences, domain and system experience) and tasks (especially, their frequency and complexity) of users;
- Extreme diversity among users, who have a range of physical and mental capabilities, varying skill levels, and come from different corporate or computer–user cultures;
- Absence of good models, ranging from physical to psychological, for user behavior;
- Lack of abstractions and tools to measure system usability and effectiveness.

New technologies such as virtual reality, multimodal immersion and prothestic augmentations, offer opportunities, but can fail dramatically if not implemented effectively.

The "Citizen Access to Statistical Tables (CAST)" project was presented as an illustration. CAST is a NSF-funded Digital Government project to build systems that enable ordinary citizens to access tabulated Government data and the information contained in them. Some of the user

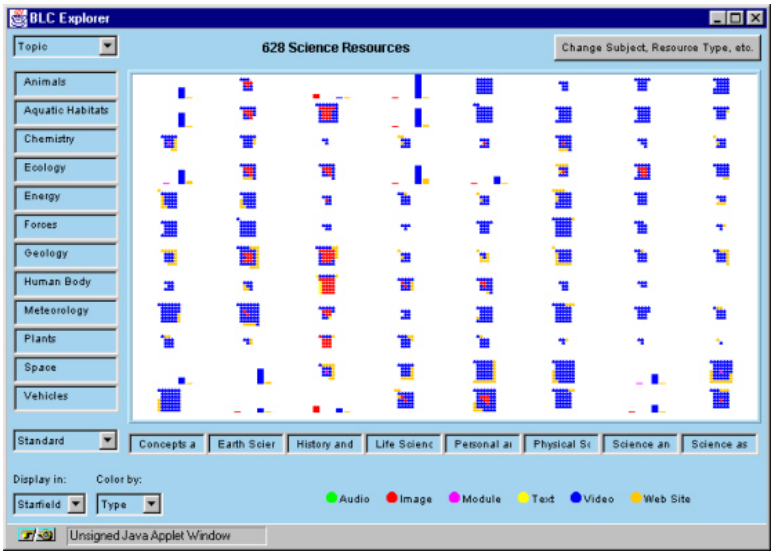

Figure 1: Potential view provided by the *Agileviews* interface to CAST.

behavior studies, interface prototypes and design strategies used to develop an effective interface were described. The latter employ *Agileviews*, an enhanced automobile metaphor for navigation, representing not only forward view, but also rear view, traffic reports, and shared views of what other drivers see. Figure 1 shows one potential view provided by the interface.

Steven Eick and Alan Karr: Visual Scalability. While much effort has focused on development, implementation and analysis of algorithms, and on their scalability, scalability analyses of visualizations (despite their importance) are have been almost entirely absent.

Visual scalability is the capability of visualization tools effectively to display large data sets, in terms of either the number or the dimension of individual data elements. The talk defined and structured the problem of visual scalability, abstractly in terms of responses that measure the business or scientific impact of visualizations and factors that affect the responses, and concretely in terms of measures of visual scalability and factors influencing them. Both current capabilities and future prospects were assessed along a number of dimensions. Approaches to increasing visual scalability include improved visual metaphors, interactivity and perspectives that link multiple views.

Issues and approaches were demonstrated (using Visual Insights' ADVIZOR™ software) for a variety of visualizations, including bar charts (see Figure 2), scatter plots and Multiscape views.

The strategy of linking multiple visual metaphors seems particularly promising: the ability interactively to examine the data at varying scales and in varying ways empowers the user to conduct explorations of the data in a that would be impossible in a predetermined algorithmic fashion. Figure 3 shows an example.

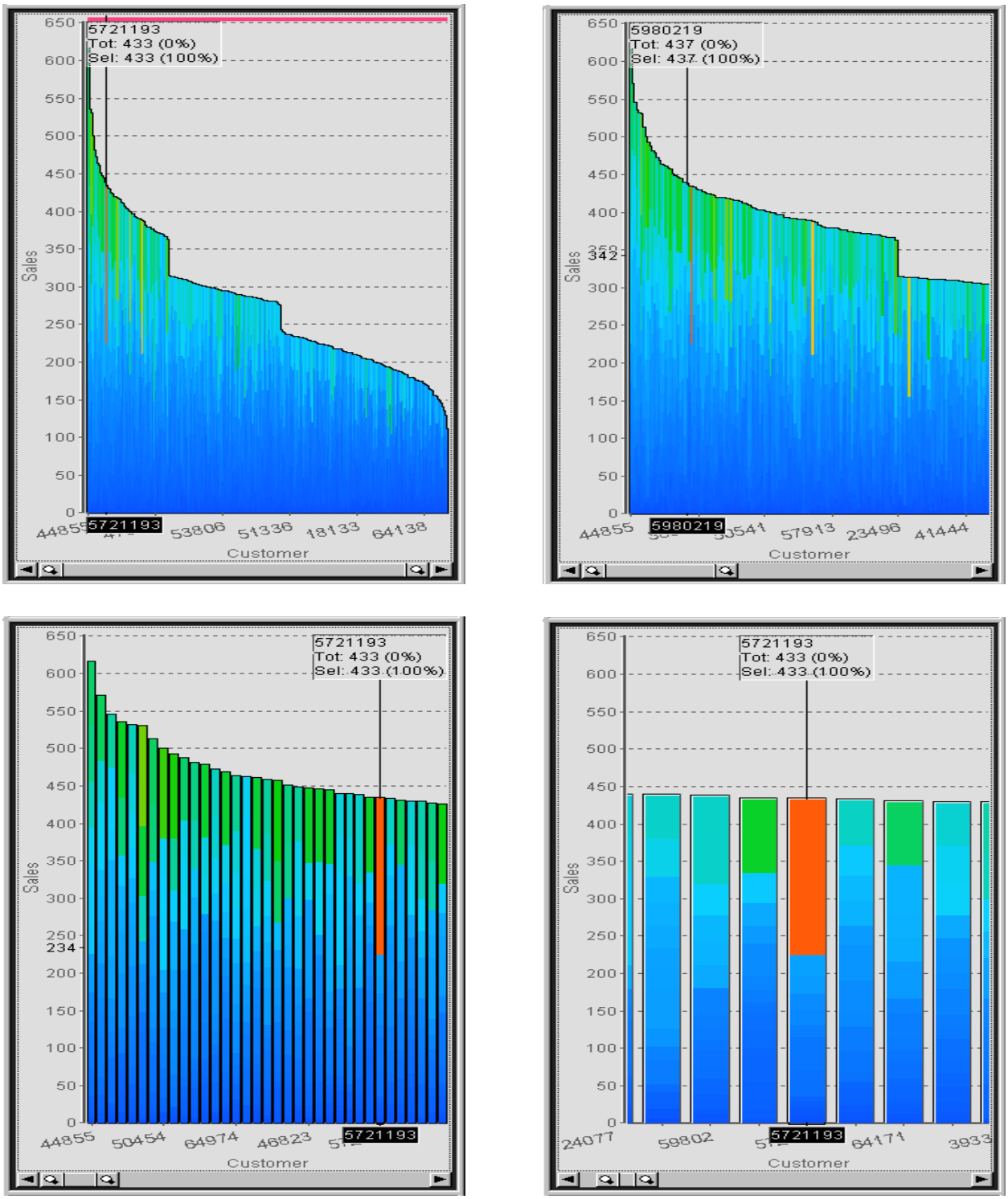

Figure 2: Bar chart scalability is increased by using levels of rendering detail and a red overplotting indicator (at the top of the view). Scalability in this case facilitates locating and then focusing attention on particular bars.

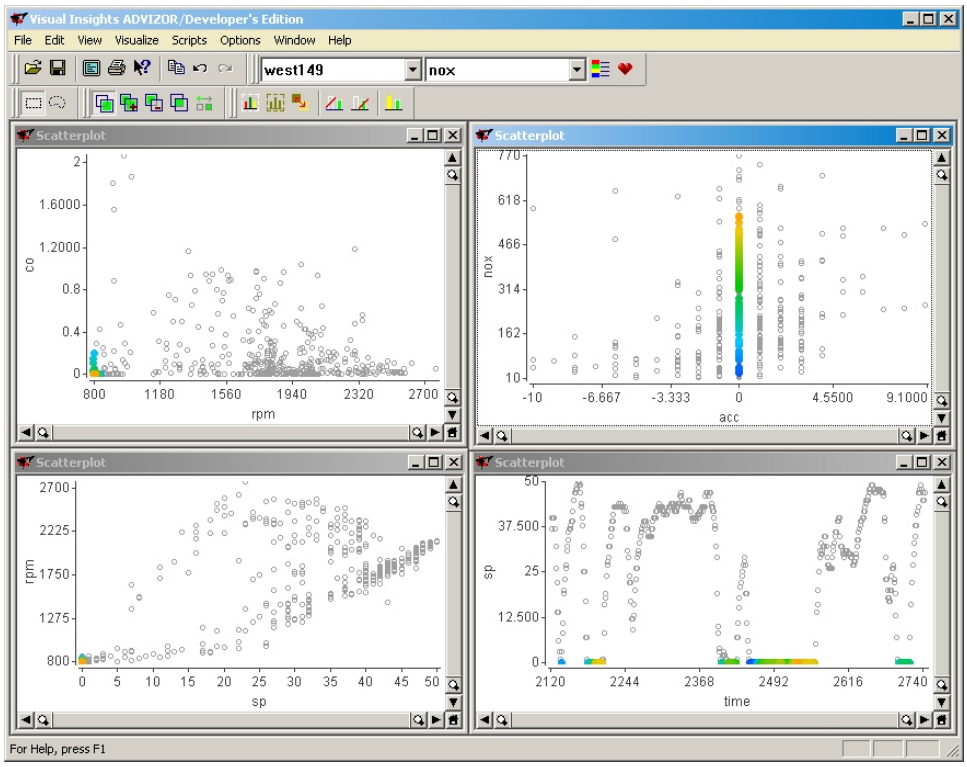

Figure 3: Perspective containing linked scatter plots of 7–dimensional data on automobile emissions during a single trip of 10 minutes. Different points correspond to measurements at 10-second intervals. The plots show CO emissions vs. engine RPM (upper left), NO_x emissions vs. acceleration (upper right), engine RPM vs. vehicle speed (lower left) and speed as a function of time (lower right). Points are colored according to the level of NO_x emissions. Iterated selection, first for low speeds (lower left view) and then for zero acceleration (upper right view), propagated to all four scatter plots, demonstrates that even with these restrictions, NO_x emissions vary dramatically, underscoring the difficulty of predicting emissions.

3.2 Digital Government

Latanya Sweeney: Threats to Data Confidentiality. This talk illustrated how publicly available resources can be exploited to uncover personal information on individuals and suggested some remedies.

The phenomenal growth of databases of information on individuals and the availability of electronic search tools creates alarming possibilities for recovering confidential information. Record linkage and profiling techniques could be applied to multiple databases (which were stripped of individual identifiers and supposedly considered safe) to effectively re-identify individuals' records. For example, supposedly anonymized records in a medical data set containing patient ID number, ZIP code, race, birth date, gender, visit date, diagnosis code, procedure code, physician ID

number, physician ZIP code and total charges were compared to a (public) voter registration list containing birth date, name, street address ZIP code. On the basis of birth data alone, 12% of the medical records were identified; birth date and gender identified 29%; birth date and 5-digit ZIP code identified 69%; and birth date and 9-digit ZIP code identified 97% of the records.

Even seemingly innocuous words in databases of text documents can lead to disclosure on individuals. Sweeney demonstrated how *Datafly* and *DataScrub* software can be used to "clean" potentially sensitive text.

Alan Karr: Web-Based Systems for Disclosure–Limited Analysis of Confidential Data. This talk described the NISS Digital Government project, which is building prototype systems to disseminate disclosure risk–limited statistical analyses of confidential data, rather than (more severely) disclosure–limited transformations of the data.

The project responds to tensions faced by Federal agencies that wish to or must make available (for example, to researchers) data collected at public expense, yet preserve the confidentiality of data elements, and hence the privacy of individual data subjects.

The systems being developed at NISS have three essential characteristics:

- They are query–based, responding to requests from users, declining to respond on the grounds that to do so creates too high a risk of disclosure, or invoking statistical risk reduction strategies before responding.
- They dispense statistical analyses, for example, cross–tabulations, log–linear models and linear regressions, instead of restricted data.
- They are *dynamic*: the risk associated with a query is assessed in light of responses (or not) to previous queries. Such systems allow the user community to have some influence over what information in the database is released.

A framework for such Web-based, history-dependent query systems was presented, and some design and implementation issues were discussed. Two prototype systems were described:

- A *table server* with sample Census data represented as a multidimensional contingency table of counts.
- A dissemination system for pesticide/herbicide usage survey data collected by the National Agricultural Statistics Service that uses geographical aggregation to achieve disclosability. For example, for the application rate (pounds per acre of a particular chemical applied to a particular crop) in a given county to be disclosable, at least three farms in that county must appear in the survey and no one of those may contain more than 60% of the total acreage (of surveyed farms). The system works by aggregating counties into disclosable "super–counties." Figures 4 and 5 illustrate, using synthetic data for the state of Ohio. The first shows whether counties are disclosable (the numbers are the sample counts), while the second shows the final aggregation (numbers label super–counties).

The talk concluded with some thoughts on disclosure risk models and risk reduction strategies, including visualization. Here, the "ordinary" rationale that visualization allows discovery of

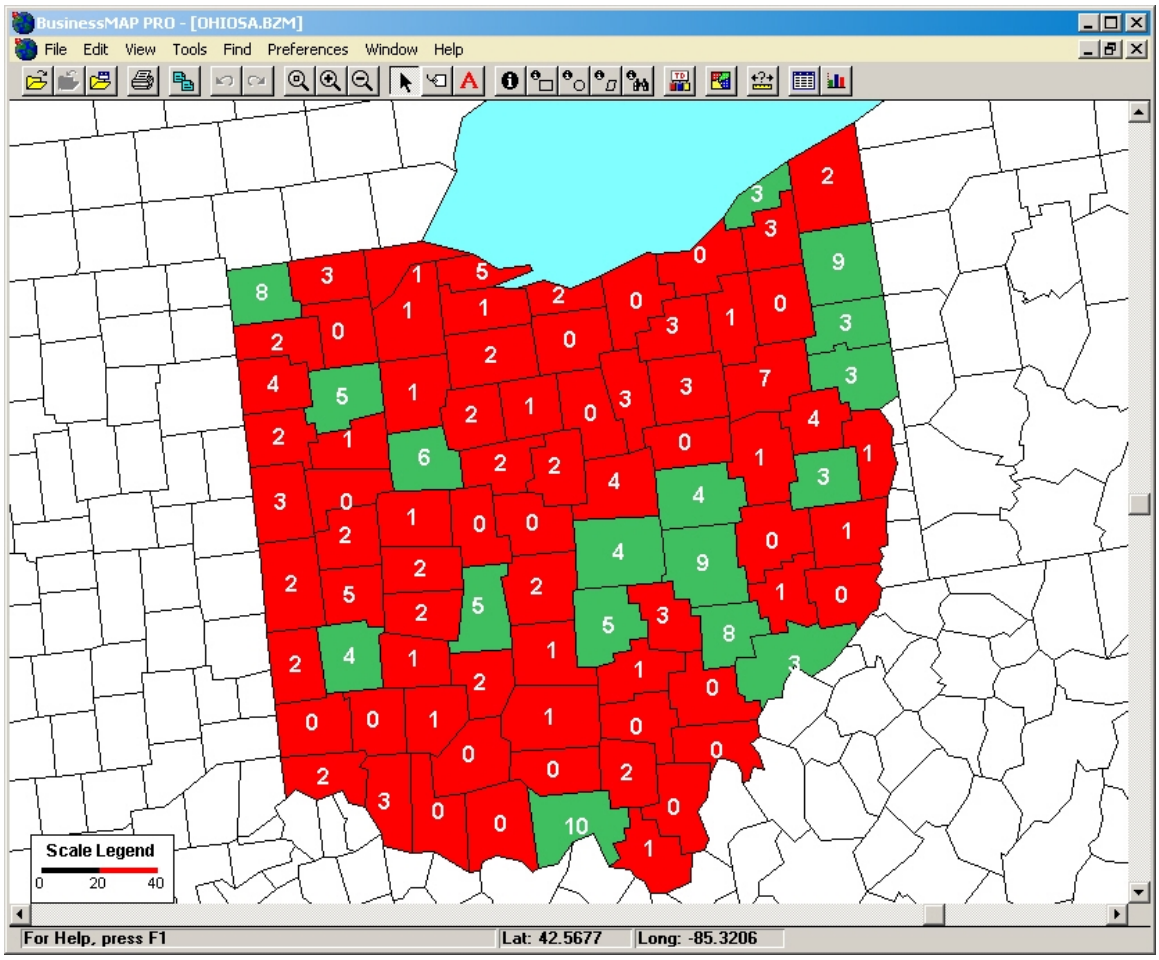

Figure 4: Farm survey data for Ohio. The map shows disclosable (green) and non-disclosable counties. Disclosability is based on both number of surveyed farms in each county (shown) and absence of dominance.

high-level structure in data without forcing attention to details is extended: for confidential data, visualization can disclose high-level structure without compromising details.

3.3 Data–Model Integration

Vasileios Hatzivassiloglou: Data and Model Integration — Examples and Challenges in Distributed Data Bases and Text Analysis. Information available on the Internet is enormous in amount and of great variety in type. To search (let alone digest) this information using bare hands and eyes is impossible. This talk presents the overview and challenges in retrieving information relevant to user queries.

Information retrieval can be divided into two steps. The first — search across multiple sources

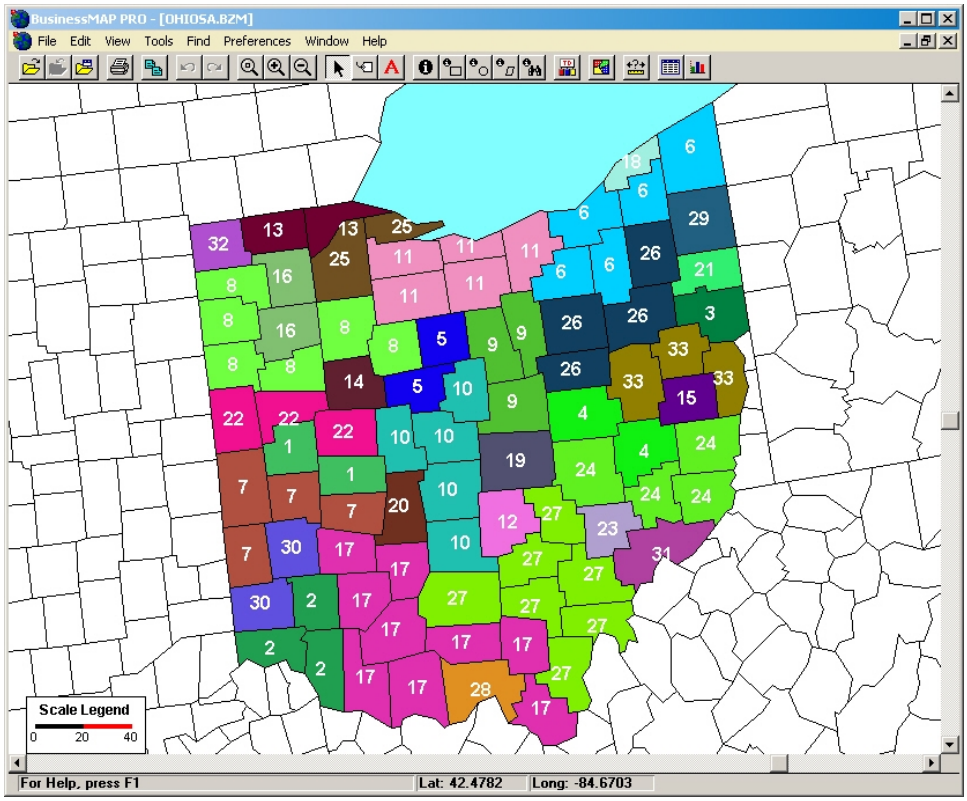

Figure 5: Geographical aggregation to achieve disclosability. Each of the "super–counties" shown satisfies disclosability requirements. (Numbers simply label the super–counties.) The aggregation is computed using simulated annealing to maximize the number of super–counties.

— is followed by search within a single source. Technologies are available for the latter: once a relevant source for information is identified, standard database methodologies such as text search and machine learning can be used, especially when the source is well structured. Searching across heterogeneous sources, on the other hand, poses many difficulties, especially in selecting relevant sources among relevant and irrelevant sources, poor– and high–quality sources, reliable and unreliable sources. Measurement and prediction of relevance, as well as extraction of terms to create domain ontologies, are particular challenges with statistical components.

Other obstacles include query translation, combining results from multiple sources and all aspects of metadata.

Applications at Columbia's Center for Advanced Research on Digital Government Information Systems (CARDGIS) were described.

Edward Wegman and Sallie Keller–McNulty: Data/Model Integration — Perspectives from Statistics. Wegman focused on coping with the *scale* of data: some systems generate one-half terabyte (or more) per day. At this scale, limits of human perception, network bandwidth and conventional algorithms (e.g., for clustering) all are exceeded.

One approach is to create data infrastructures to cope with scale, by emphasizing metadata that describe the scientific content of the data. Steps toward this goal include NASA's Distributed Active Archive Centers (DAAC) and Metadata Centers (MdC) to access the data. Central research issues are:

- Digital objects to index data;
- Automatic generation of metadata;
- Query refinement (using expert systems);
- Data quality.

Keller–McNulty addressed the problem from the perspective of "model mining:" what would it take to build a system to respond to user queries of the form, "What would be the effect of closing the 14th Street bridge in Washington, DC, during the evening rush hour?" A system that can do this must:

- Refine vague queries;
- Locate the relevant data among many sources available;
- Locate relevant (transportation/traffic) models;
- Combine the data and models;
- Present results together with associated uncertainties, which must first be quantified.

Figure 6 illustrates one possible result of the data–model assembly process.

Problems of human–computer interaction (HCI) are central in this context as well, especially models of user needs, knowledge and behavior.

3.4 Internet Traffic Measurement and Analysis

Daniel Stevenson: Network Traffic Modeling. Using the North Carolina Research and Educational Network (NCREN) — see Figure 7 — as a backdrop, Stevenson presented a number of case studies in which modeling of network traffic played a key role, and in which simple, aggregated measurements would not suffice. These case studies include:

- GTE's Mobile Subscriber Environment, where network traffic analysis led to a substantially lower cost estimate for building the system;
- MCNC's VISTAnet and GigaPoP networks, where engineers were forced to look at the data in novel ways in order to understand the network performance.

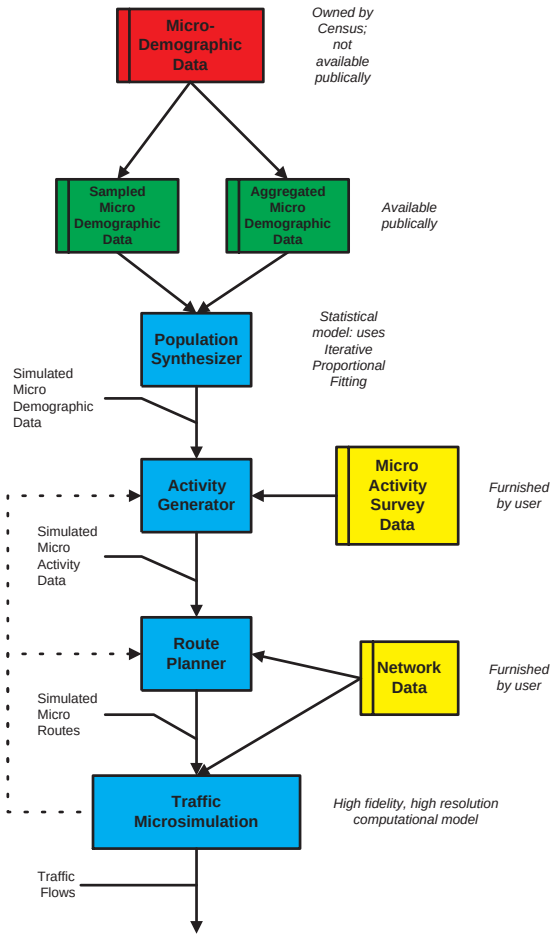

Figure 6: Data sets and models assembled to respond to a query about transportation, for example, regarding the effects of a change in the road network. The principal steps in the process: (1) Publicly available demographic data from the Census Bureau and a statistical model are used to simulate a population of individuals; (2) A survey of the activities that generate travel and a statistical model are used to synthesize activities and their locations for the synthetic population; (3) Routes are assigned on the basis of activities, road network data and an optimization model; (4) A traffic microsimulation model moves vehicles along their assigned routes, handling interactions among vehicles and traffic signals. Confidential data are shown in red, public data in green, user–provided data in yellow and model components in cyan.

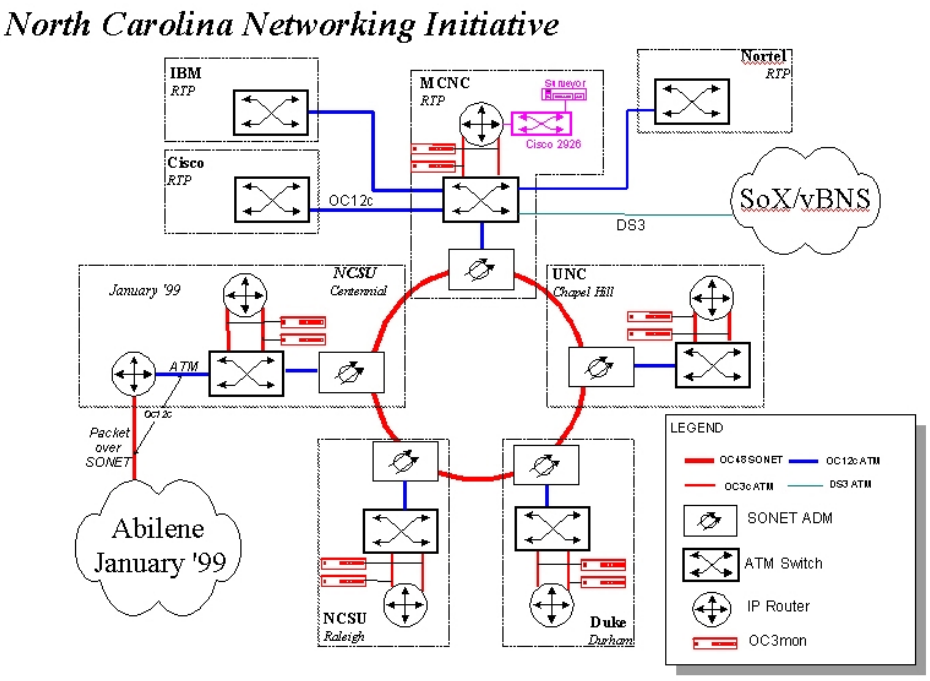

Figure 7: The North Carolina Research and Educational Network (NCREN). See www.ncni.org for additional information.

These and other networks exhibit complex and sometimes apparently contradictory behavior, such as an observed increase in loads following an increase in bandwidth connection, which cannot be characterized using existing models or on the basis of available measurements.

Some current trends in technology pose special challenges. For example, increasing amounts of voice and video traffic create new technical problems (such as the optimization echo perception in voice transmissions, and the measurement of the multimedia traffic flow), and raise questions of how the customers ought to be billed. Moreover, the dominant form of the data carried seems to be changing over time — from HTML to XML to MP3 files.

Standard analytical models based on Poisson or Erlang processes are inadequate, especially for forecasting, planning and prediction. More research on self-similar processes and multifractal models is needed.

J. Stephen Marron: Statistical Analysis of Internet Traffic Data — Smoothing and Multifractal Analysis. Marron described and demonstrated a of number models and analysis techniques for network data.

Many of these are based on the concept of "Scale Space," which maintains that there are different features of the data which become visible (and relevant) at varying scales of observation. Thus, data must be explored at several scales simultaneously, zooming in for a microscopic view and out for the "big picture," in order to obtain a sense of how apparent features change with the level of resolution.

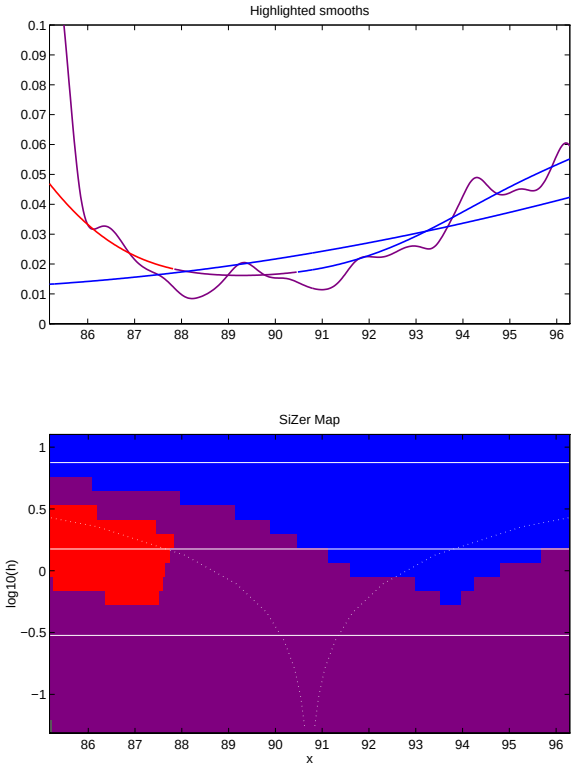

Figure 8: Multiscale analysis using *SiZer*. The data are the spans of changes to a particular software system, and were smoothed to estimate the span of changes as a function of time (the x-axis). (The span of a change is the numbers of files touched — an indicator of the difficulty or cost of making the change.) The degree of smoothing is on the y-axis. Color depicts (as a function of both time and scale $\equiv$ smoothing) whether the rate is positive (blue), negative (red) or statistically indistinguishable from zero (magenta). The upper plot shows estimated spans at several particular scales.

Two of Marron's programs, *SiZer* and *SiCon*, produce dynamic displays of data at various scales and were well-suited for multi-scale visualizations. Data on packet arrival times at a node in the NCREN were used as an illustration. Data sets simulated from a homogeneous Poisson process and a multifractal process were used for comparison. The *SiZer* and *SiCon* visualizations revealed that the network data did not look like the pure Poisson or multifractal processes, but appeared instead to share some features of each.

Directions for future research include better visualization tools, more realistic models (multifractal Cox-like processes) and strategies for model validation.

3.5 Software Development

Robert Horgan: Software Engineering Perspective. This talk presented software development and maintenance (40–70% of the lifecycle cost of software is for testing and maintenance) as a quality problem involving measurement, experimentation and prediction. Data are voluminous, and constitute a business asset, but tools to utilize the data are weak.

For example, prediction and control of defects in the software development process are currently not feasible. Existing models, such as the waterfall model (In which principal, sequential stages of software development are (1) Requirements analysis and definition; (2) System and software design; (3) Implementation and unit testing; (4) System testing; and (5) Operation and maintenance.), are inadequate to capture the complexity of the problem, and even the Capability Maturity Model (CMM) (see www.sei.cmu.edu/cmm/) has not been show to correlate with other measures of quality. Concepts from statistical quality control have not yet been applied effectively.

There is great need to link data quality with software quality. Following ideas of Wirth,

$$\text{Programs} = \text{Algorithms} + \text{Data}. \tag{1}$$

Even if not entirely successfully, quality of the algorithmic component of programs has been studied; that of data has substantially been ignored. For example, error models for terabytes of data have not been developed.

Siddhartha Dalal: Statistics Perspective. Dalal described the current state–of–the–art in application of statistical methods in software development, emphasizing the need to link such application to business objectives.

Particular opportunities include:

- Creation of taxonomies of defects (using, for example, techniques from exploratory data analysis or data mining);
- Use of experimental design techniques (for example, orthogonal arrays) for software testing;
- Model-based software testing;
- Strategies to model the performance of component-based (that is, virtually all) software.

4 Workshop Findings

In this section, we summarize the findings of the Workshop. These are stated as research issues to be pursued by appropriate cross-disciplinary teams.

Entry points for the research are essential. Many of the issues are very large, so that it is essential to begin the research intelligently, using specific problems, applications and testbeds.

4.1 Human–Computer Interaction

Human–centered design, to create improved, adaptable user interfaces that engage users more directly, and enable broad populations of users to access information contained in large, complex data sets. Strategies to be explored include explicitly asking users what they intend to do and then providing a customized interface for that purpose (whereas current interfaces are largely product- rather than task-determined); incorporating visual metaphors appropriate to the user's domain; and allowing highly configurable user preferences (perhaps by maintaining a database of preferences).

User interface usability metrics, such as time, accuracy, satisfaction, and both physiological and psychological effects. With these, quantitative evaluation of interfaces will become possible.

Visual Scalability, to extend visualization to very high-dimensional, massive data sets. Aspects relate to visual metaphors, algorithms (e.g., sampling from databases in order to apply computationally intensive techniques), displays, computational power and network-based data access. The consequence is that information contained in large data sets would be more readily discovered and acted upon.

Inference–Visualization Relationships, to merge the power of visual exploration and formal inference. As result, decisions based on visualizations will be sounder, and their effects more predictable.

Design–Visualization Relationships, to combine the parallel notions of data collection in experimental design (where to gather informative data) and data visualization (where in the data to look for information). Visualizations based on data collected for that purpose (rarely the case today) will be particularly powerful.

4.2 Digital Government

Systems that dispense statistical analyses as an alternative to restricted or transformed microdata, in order to preserve confidentiality of the data. Not only do such systems fit the emerging application service provider (ASP) paradigm, but also they may be the only workable solution in a world where record linkage cannot be prevented. Complementary strategies include dispensing visual displays and graphical summaries. Creation of user awareness regarding application of risk reduction strategies, so that analyses will be interpreted correctly, is a particular challenge.

Scalability of computational procedures to compute or reduce disclosure risk. Existing strategies, with few exceptions, either have been shown not to scale for use with very large databases or else it is not known whether they scale. Progress on this issue is necessary for extending statistical disclosure beyond its traditional "government data" context — for example, to transaction data collected in E-commerce.

Quantification of risk measures (which may also represent "cost," for example, of re-identification of sample elements) and accompanying probability models for risk. Similarly, ways to measure and model the "informativeness" of data released in response to user queries are necessary in order to balance risk and informativeness.

Synthetic data as a means of addressing disclosure issues. Using this strategy a synthetic data set is generated (which contains no microdata record associated with any real data subject) using a statistical model of the confidential data set that captures essential features and interrelationships among variables.

Legal issues, such as enforcement of disclosure protection. What is privacy in today's electronic world? Is it be possible to make "database snooping" illegal?

Dynamic databases containing longitudinal data, which present special confidentiality issues.

4.3 Data–Model Integration

Model-based search for information, especially across multiple, heterogeneous data sources. Models that predict the "usefulness" of different sources in answering a particular query would improve search significantly.

Quantification and measurement of relevance and quality of data sources.

User input to data–model integration processes.

Ontologies that support re-use of data and models. Statistical techniques are especially necessary to assist in extraction of terms from documents in order to build domain ontologies in the absence of explicit semantic models. Data mining (for example, discovery of patterns and associations) for metadata may be an essential capability.

Automatic generation of metadata, especially the scientific content of data sets, to construct digital objects that index the data. This capability would enable access to the information contained in increasing large and complex data sets.

Standards for description of data and models, such as XML (`www.xml.org`) and PMML (`www.oasis-open.org/cover/pmml.html`).

4.4 Internet Traffic Measurement and Analysis

Data exploration strategies to identify significant features (to be investigated further) in network data. Scale must be confronted: a single OC3Mon on a fully loaded link generates 1.5 terabytes of data per day.

Choice of appropriate time scale for modeling network traffic, in terms of both modeling implications and what action needs to be taken in light of the models.

Forecasting, planning and prediction, for which current network traffic models based on self-similarity are simply not effective. These models are descriptive rather than predictive, and cannot assess the effects of changes in the network itself or the behavior of users.

Design of data informative collection protocols for networks is effectively absent: today data are collected where it is easy to do so. Data collections designs that enhance both the low-level operation of networks (e.g., by leading to more efficient network protocols) or the high-level objectives of network operators (e.g., strategies to measure and enforce quality of service).

Global network behavior, which is not well understood, even by comparison with behavior of networks at single locations. Understanding the spatial structure of network traffic is an essential component.

Indirect measurements of network traffic, especially those that relate to performance seen by users (as opposed to that seen by network operators).

4.5 Software Development

Measurement of the quality of each step of software process and of software itself, using such metrics as reliability, fault density, usability, safety, security and maintainability.

Testing of software, which continues to lack a firm scientific basis. The role of experimental design in software testing (which is not well understood, despite initial efforts) requires illumination.

Software experimentation is desirable, but rarely possible at the industrial scale. "Real" software data, as a consequence, exhibit strong selection bias: they come from systems that function and survive.

SQC for software, in order to control software processes and predict defects, thereby helping managers set objectives.

Quality of component-based software, which has become the primary means by which new software is built. Strategies must be developed to control the quality of the final product, which is difficult even when behavior of the components is well understood. Models to simulate the behavior of component-based software are needed to address urgent needs identified in the PITAC report.

Appendix

A Participants

Name	Organization	E-Mail
Melvyn Ciment	Potomac Institute for Policy Studies	mciment@potomacinstitute.org
Lawrence Cox	U.S. EPA	cox.larry@epamail.epa.gov
Siddhartha Dalal	Telcordia Technologies	sid@research.telcordia.com
Stephen G. Eick	Visual Insights	eick@visualinsights.com
Sujit K. Ghosh	North Carolina State U	sghosh@stat.ncsu.edu
Vasileios Hatzivassiloglou	Columbia U	vh@cs.columbia.edu
Robert Horgan	Telcordia Technologies	jrh@research.telcordia.com
Wen-Hua Ju	Rutgers U	whju@stat.rutgers.edu
Alan F. Karr	NISS	karr@niss.org
Sallie Keller–McNulty	Los Alamos National Laboratory	sallie@lanl.gov
Jaeyong Lee	NISS	leej@server.niss.org
Deborah Leishman	IBM	leishman@us.ibm.com
Xiaodong Lin	Purdue U	linxd@stat.purdue.edu
Gary Marchionini	U of North Carolina — Chapel Hill	march@ils.unc.edu
J. Stephen Marron	U of North Carolina — Chapel Hill	marron@stat.unc.edu
George Michailidis	U of Michigan	gmichail@umich.edu
James Rosenberger	NSF	jrosenbe@nsf.gov
Alan Saalfeld	Ohio State U	saalfeld.1@osu.edu
Jerome Sacks	NISS	sacks@niss.org
Ashish Sanil	NISS	ashish@niss.org
Paul Speckman	U of Missouri–Columbia	speckman@stat.missouri.edu
Alex Stark	NISS	stark@niss.org
Daniel Stevenson	MCNC	stevenso@mcnc.org
Latanya Sweeney	Carnegie Mellon U	latanya@andrew.cmu.edu
Edward J. Wegman	George Mason U	ewegman@gmu.edu
Lara Wolfson	Brigham Young U	ljwolfson@byu.edu