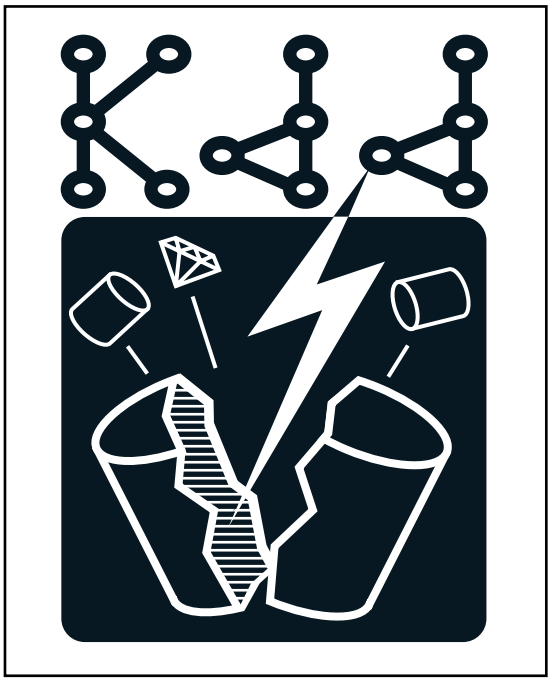


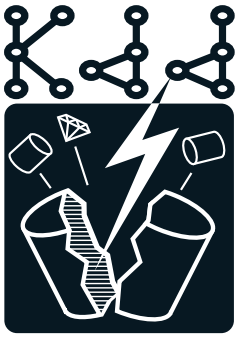
# The Third International Conference on Knowledge Discovery and Data Mining

August 14–17 1997, Newport Beach, California

Sponsored by the American Association for Artificial Intelligence

Cosponsored by
General Motors Research,
The National Institute for Statistical Sciences, and
NCR Corporation

In cooperation with
The American Statistical Association



# Welcome to KDD-97!

## KDD-97 Organization

### General Conference Chair

Ramasamy Uthurusamy, *General Motors Corporation*

### Program Cochairs

David Heckerman, *Microsoft Research*
Heikki Mannila, *University of Helsinki*, Finland
Daryl Pregibon, *AT&T Laboratories*

### Publicity Chair

Paul Stolorz, *Jet Propulsion Laboratory*

### Tutorial Chair

Padhraic Smyth, *University of California, Irvine*

### Demo and Poster Sessions Chair

Tej Anand, *NCR Corporation*

### Awards Chair

Gregory Piatetsky-Shapiro, *GTE Laboratories*

### Program Committee Members

Helena Ahonen, *University of Helsinki*
Tej Anand, *NCR*
Ron Brachman, *AT&T Laboratories*
Carla Brodley, *Purdue University*
Dan Carr, *George Mason University*
Peter Cheeseman, *NASA AMES Research Center*
David Cheung, *University of Hong Kong*
Wesley Chu, *University of California at Los Angeles*
Gregory Cooper, *University of Pittsburgh*
Robert Cowell, *City University*, UK
Bruce Croft, *University of Massachusetts at Amherst*
William Eddy, *Carnegie Mellon University*
Charles Elkan, *University of California at San Diego*
Usama Fayyad, *Microsoft Research*
Ronen Feldman, *Bar-Ilan University*, Israel
Jerry Friedman, *Stanford University*
Dan Geiger, *Technion*, Israel
Clark Glymour, *Carnegie Mellon University*
Moises Goldszmidt, *SRI International*
Georges Grinstein, *University of Massachusetts, Lowell*
Jiawei Han, *Simon Fraser University*, Canada
David Hand, *Open University*, UK
Trevor Hastie, *Stanford University*
David Heckerman, *Microsoft Corporation*
Haym Hirsh, *Rutgers University*
Jim Hodges, *University of Minnesota*
Se June Hong, *IBM T .J. Watson Research Center*
Tomasz Imielinski, *Rutgers University*
Yannis Ioannidis, *University of Wisconsin*
Lawrence D. Jackel, *AT&T Laboratories*
David Jensen, *University of Massachusetts*
Michael Jordan, *Massachusetts Institute of Technology*
Daniel Keim, *University of Munich*
Willi Kloesgen, *GMD*, Germany
Ronny Kohavi, *Silicon Graphics*
David D. Lewis, *AT&T Laboratories – Research*
David Madigan, *University of Washington*
Heikki Mannila, *University of Helsinki*
Brij Masand, *GTE Laboratories*
Gary McDonald, *General Motors Research*
Eric Mjolsness, *University of California at San Diego*
Sally Morton, *Rand Corporation*
Richard Muntz, *University of California at Los Angeles*
Raymond Ng, *University of British Columbia*
Steve Omohundro, *NEC Research*
Gregory Piatetsky-Shapiro, *GTE Laboratories*
Daryl Pregibon, *Bell Laboratories*
Raghu Ramakrishnan, *University of Wisconsin, Madison*
Patricia Riddle, *Boeing Computer Services*
Ted Senator, *NASD Regulation Inc.*
Jude Shavlik, *University of Wisconsin at Madison*
Wei-Min Shen, *University of Southern California*
Arno Siebes, *CWI*, Netherlands
Avi Silberschatz, *Bell Laboratories*
Evangelos Simoudis, *IBM Almaden Research Center*
Andrzej Skowron, *University of Warsaw*
Padhraic Smyth, *University of California at Irvine*
Ramakrishnan Srikant, *IBM Almaden Research Center*
John Stasko, *Georgia Institute of Technology*
Salvatore J. Stolfo, *Columbia University*
Paul Stolorz, *Jet Propulsion Laboratory*
Hanna Toivonen, *University of Helsinki*
Alexander Tuzhilin, *New York University, Stern School*
Ramasamy Uthurusamy, *General Motors R&D Center*
Graham Wills, *Bell Laboratories*
David Wolpert, *IBM Almaden Research Center*
Jan Zytkow, *Wichita State University*

# Thursday, August 14th

8:00AM–6:00 PM

## Tutorials

8:00–10:00 AM (SINGLE SESSION)

*Tutorial 1*

Data Mining and KDD: An Overview

*Usama Fayyad, Microsoft Research and Evangelos Simoudis, IBM*

10:00–10:30 AM

## Coffee Break

10:30 AM–12:30 PM

*Tutorial 2*

Modeling Data and Discovering Knowledge

*David Hand, Open University, UK*

10:30 AM–12:30 PM

*Tutorial 3*

Text Mining - Theory and Practice

*Ronen Feldman, Bar-Ilan University, Israel*

12:30–1:30 PM

## Lunch

1:30–3:30 PM

*Tutorial 4*

Exploratory Data Analysis Using Interactive Dynamic Graphics

*Deborah Swayne, Bell Communications Research and Diane Cook, Iowa State University*

1:30–3:30 PM

*Tutorial 5*

OLAP and Data Warehousing

*Surajit Chaudhuri, Microsoft Research and Umesh Dayal, Hewlett Packard Laboratories*

3:30–4:00 PM

## Coffee Break

4:00–6:00 PM

*Tutorial 6*

Visual Techniques for Exploring Databases

*Daniel Keim, University of Munich, Germany*

4:00–6:00 PM

*Tutorial 7*

Statistical Models for Categorical Response Data

*William DuMouchel, AT&T Research*

6:30–7:30 PM

## Opening Reception

*California Ballroom, Newport Beach Marriott Hotel and Tennis Club*

## KDD-97 Tutorial Program

All tutorials will be presented on Thursday, August 14, 1997 in sections of the Pacific Ballroom. Admission to the tutorials is included in your conference registration fee. Registrants can attend up to four consecutive tutorials, including four tutorial syllabi.

| TIME | SESSION 1 | SESSION 2 |
|---|---|---|
| 8:00 to 10:00 AM | T1 – Fayyad and Simoudis *(single session)* | |
| 10:30 AM to 12:30 PM | T2 – Hand | T3 – Feldman |
| 1:30 to 3:30 PM | T4 – Swayne and Cook | T5 – Chaudhuri and Dayal |
| 4:00 to 6:00 PM | T6 – Keim | T7 – DuMouchel |

# Friday, August 15th

*All sessions are plenary sessions and will be held in the Pacific Ballroom, Newport Beach Marriott Hotel and Tennis Club*

8:30–10:00 AM

## Keynote Address and Invited Talk Session

8:30 –8:45 AM

### Welcome and Introduction

*KDD-97 General Conference Chair: Ramasamy Uthurusamy and KDD-97 Program Cochairs: David Heckerman, Heikki Mannila, and Daryl Pregibon*

8:45–9:30 AM

### Keynote Address: From Large to Huge: A Statistician's Reactions to KDD & DM

*Peter J. Huber, University of Bayreuth, Germany*

The statistics and AI communities are confronted by the same challenge—the onslaught of ever larger data collections—but the two communities have reacted independently and differently. What could they learn from each other if they looked over the fence? What is amiss on either side?

9:30–10:00 AM

### Invited Talk: Machine Learning and KDD: New Developments

*Thomas Dietterich, University of Oregon*

Machine learning research is pursuing many directions relevant to KDD. This talk will review two exciting lines of research: learning with ensembles (committees, bagging, boosting, etc.) and learning with stochastic models. I will also briefly mention current research in reinforcement learning and methods for scaling up machine learning algorithms.

10:00–10:30 AM

## Coffee Break

10:30–12:15 PM

## Keynote Address and Poster Previews Session

10:30–11:15 AM

### Keynote Address: Searching for Causes and Predicting Interventions

*Clark Glymour, University of California San Diego and Carnegie Mellon University*

The problem in many data mining tasks is to predict features of a new, unobserved sample from a recorded sample, assuming the sampling distribution does not change. In causal data mining the task is to predict the features of new unobserved samples from a recorded sample, assuming any unobserved sample results from an intervention which directly alters the underlying probability distribution of some variables in a known way, and alters others indirectly through the influences of the directly manipulated variables. Data mining in the sciences and for business, military or public policy is often concerned with discovering causal structure, although that focus is sometimes obscured by the belief that causation is especially mysterious. Causation is not probability, but it is no more (and no less) mysterious. This lecture will illustrate why data mining for causes using techniques designed for recognition or classification is unwise, and will illustrate principles of causal data mining with several interesting cases that use both Bayesian and constraint based methods.

11:15 AM–12:15 PM

*Poster Previews Session*

## (Authors A-J)

12:15–2:00 PM

## Lunch

(Program Committee Luncheon, Newport North Room, lobby level, Newport Beach Marriott Hotel and Tennis Club)

2:00–3:00 PM

*Poster Previews Session*

## (Authors K-Z)

3:00–7:00 PM

## Exhibits, Demos, Poster Sessions

California Ballroom, Newport Beach Marriott Hotel and Tennis Club

3:00 –4:20 PM

*Poster Session 1*

## Authors A-G

4:20 –5:40 PM

*Poster Session 2*

## Authors H-O

5:40 –7:00 PM

*Poster Session 3*

## Authors P-Z

(Coffee and soft drinks will be available in the foyer from 3:00–7:00 PM.)

# Saturday, August 16th

8:30–9:45 AM

*Paper Session*

## Database Methodology

8:30–8:55 AM

### Computing Optimized Rectilinear Regions for Association Rules

*Kunikazu Yoda, Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, and Takeshi Tokuyama, IBM Tokyo Research Laboratory, Japan*

8:55–9:20 AM

### Density-Connected Sets and their Application for Trend Detection in Spatial Databases

*Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, University of Munich, Germany*

9:20–9:45 AM

### Mining Association Rules with Item Constraints

*Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal, IBM Almaden Research Center*

9:45–10:15 AM

## Coffee Break

10:15 AM–12:30 PM

## Keynote Address/Invited Talks/Panel Session

10:15–11:00 AM

### Keynote Address: Are Spatial Data Special: A Data Mining Perspective?

*Raymond Ng, University of British Columbia, Canada*

From a data mining standpoint, Ng will discuss how spatial data are similar to and different from conventional alphanumeric data. He will use characteristic and discriminant rules as examples, and discuss how the characteristics of spatial data affect the ways these rules can be found. Finally, Ng will comment on the generalization from spatial data to discuss whether temporal data or image data are special.

11:00–11:30 AM

### Invited Talk: Highlights from the Joint Statistical Meetings

*Jim Hodges, University of Minnesota*

Statisticians and data-miners have a lot in common but their paths tend not to cross. Thus, for example, many statisticians are missing out on computing and other ideas familiar to data-miners, and data-miners sometimes reinvent things that are well-known to statisticians. This talk is intended to toss a small bridge across the gap, by summarizing some talks and trends from the 1997 Joint Statistical Meetings, held in Anaheim, California just before KDD-97.

11:30–12:30 PM

### Panel Discussion: Validity of Data Mining Results

*Moderator: David Hand, The Open University, United Kingdom*

*Panelists: Jim Hodges, University of Minnesota; Reza Nakhaeizadeh, Daimler-Benz AG, Germany; Gregory Piatetsky-Shapiro, GTE Laboratories; Arno Siebes, CWI, The Netherlands; and Padhraic Smyth, University of California, Irvine*

Data mining methods typically rely on large-scale search for patterns or models. How can statistically valid results be ensured in this search-intensive approach? What kind of statistical tests are appropriate? How can one reduce overfitting? What techniques exist for efficient and data-adaptive sampling strategies for specific classes of algorithms? How can one combine statistical and non-statistical measures of the utility of discovered patterns? These and other questions from the audience will be discussed by the panelists.

12:30–2:00 PM

## Lunch

2:00–3:30 PM

*Paper Session*

## Applications

2:00–2:30 PM

### Detecting Atmospheric Regimes Using Cross-Validated Clustering

*Padhraic Smyth, University of California, Irvine; Michael Ghil and Kayo Ide, University of California, Los Angeles; Joe Roden, Jet Propulsion Laboratory, California Institute of Technology; Andrew Fraser, Portland State University*

2:30–3:00 PM

### Automated Discovery of Active Motifs in Three Dimensional Molecules

*Xiong Wang and Jason T.L. Wang, New Jersey Institute of Technology; Dennis Shasha, New York University; Bruce Shapiro, National Cancer Institute; Sitaram Dikshitulu, New Jersey Institute of Technology; Isidore Rigoutsos, IBM T. J. Watson Research Center; Kaizhong Zhang, University of Western Ontario, Canada*

3:00–3:30 PM

### JAM: Java Agents for Meta-Learning over Distributed Databases

*Salvatore Stolfo, Andreas L. Prodromidis, Shelley Tselepis, Wenke Lee, and Dave W. Fan, Columbia University; Philip K. Chan, Florida Institute of Technology*

3:30–4:00 PM

## Coffee Break

4:00–6:00PM

## Invited Talk/Paper Presentation/Awards Presentation

4:00–4:30 PM

### Invited Talk: Highlights from the Graphical Modeling Workshop

*David Madigan, University of Washington*

This talk will provide a very brief tutorial introduction to basic concepts of graphical models. A survey of current research topics will follow, highlighting in particular the work presented at the June 1997 AMS-IMS-SIAM workshop on graphical models.

4:30–5:00 PM

### Knowledge = Concepts: A Harmful Equation

*Jan M. Zytkow, Wichita State University and Polish Academy of Sciences, Poland*

5:00–6:00 PM

## Awards Presentation

# Sunday, August 17th

8:30–9:45 AM

*Paper Session*

## Methodology I

8:30–8:55 AM

### A Probabilistic Approach to Fast Pattern Matching in Time Series Databases

*Eamonn Keogh and Padhraic Smyth, University of California, Irvine*

8:55–9:20 AM

### Discriminative Versus Informative Learning

*Y. Dan Rubinstein and Trevor Hastie, Stanford University*

9:20–9:45 AM

### Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions

*Foster Provost and Tom Fawcett, NYNEX Science and Technology*

9:45–10:15 AM

## Coffee Break

10:15 AM–12:30 PM

*Paper Session*

## Methodology II/Invited Panel

10:15–10:40 AM

### Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms

*Gholamreza Nakhaeizadeh, Daimler-Benz AG, Germany and Alexander Schnabl, Technical University Vienna, Austria*

10:40–11:05 AM

### A Visual Interactive Framework for Attribute Discretization

*Ramesh Subramonian, Ramana Venkata, and Joyce Chen, Intel Corporation*

11:05–11:30 AM

### Using General Impressions to Analyze Discovered Classification Rules

*Bing Liu, Wynne Hsu, and Shu Chen, National University of Singapore, Singapore*

11:30 AM–12:30 PM

### Panel Discussion: Data Mining and the Web

*Moderator: Usama Fayyad, Microsoft Research*

*Panelists: Sue Dumais, Bellcore; Ronen Feldman, Bar-Ilan University, Israel; Marti Hearst, Xerox PARC; Haym Hirsh, Rutgers University; Willi Kloesgen, GMD German National Research Center; Kamran Parsaye, Information Discovery Inc.; and Michael Pazzani, University of California, Irvine*

The Web is becoming the premium source of information for a growing number of people, and Web-based systems are increasingly applied for various applications, whether for commercial transactions or for a group organizing and performing its collaboration. Two challenges are predominant for data mining on the Web. The first goal is to help users in finding useful information on the Web and in discovering knowledge about a domain that is represented by a collection of Web-documents. The second goal is to analyze the transactions run in a Web-based system, be it to optimize the system or to find information about the clients using the system. This panel will discuss the approaches, chances, and problems of applying data mining techniques for the Web.

12:30–2:00 PM

## Lunch

2:00–3:15 PM

*Paper Session*

## Tools/Environments for DM

2:00–2:25 PM

### An Interactive Visualization Environment for Data Exploration

*Mark Derthick, John Kolojejchick, and Steven F. Roth, Carnegie Mellon University*

2:25–2:50 PM

### Visualization Techniques to Explore Data Mining Results for Document Collections

*Ronen Feldman, Bar-Ilan University, Israel; Willi Klösgen, German National Research Center for Information Technology, Germany; Amir Zilberstein, Bar-Ilan University, Israel*

2:50–3:15 PM

### Anytime Exploratory Data Analysis for Massive Data Sets

*Padhraic Smyth, University of California, Irvine and David Wolpert, IBM Almaden Research Center*

3:15–4:00 PM

## Open Forum: Comments on KDD-97

8:30AM–5:00 PM

*Newport North Room, Lobby Level*

## Workshop on Issues in the Integration of Data Mining and Data Visualization

*(by invitation only)*

*Organizers: Georges Grinstein, Andreas Wierse, and Usama Fayyad*

# Poster Sessions

3:00–4:20 PM

*Poster Session I*

## Authors A-G

### Discovery of Actionable Patterns in Databases: The Action Hierarchy Approach
*Gediminas Adomavicius, New York University and Alexander Tuzhilin, Stern School of Business, New York University*

### Partial Classification Using Association Rules
*Kamal Ali and Stefanos Manganaris, IBM Global Business Intelligence Solutions; Ramakrishnan Srikant, IBM Almaden Research Center*

### Increasing the Efficiency of Data Mining Algorithms with Breadth-First Marker Propagation
*John M. Aronis, University of Pittsburgh and Foster J. Provost, NYNEX Science and Technology*

### Brute-Force Mining of High-Confidence Classification Rules
*Roberto J. Bayardo, Jr., The University of Texas at Austin*

### Applying Data Mining and Machine Learning Techniques to Submarine Intelligence Analysis
*Ulla Bergsten, Johan Schubert, and Per Svensson, Defence Research Establishment, Sweden*

### Process-Based Database Support for the Early Indicator Method
*Christoph Breitner and Jörg Schlösser, University of Karlsruhe, Germany; Rüdiger Wirth, Daimler-Benz AG, Germany*

### MineSet: An Integrated System for Data Mining
*Cliff Brunk, James Kelly, and Ron Kohavi, Silicon Graphics, Inc.*

### Proposal and Empirical Comparison of a Parallelizable Distance-Based Discretization Method
*Jesús Cerquides and Ramon López de Màntaras, Spanish Council for Scientific Research, Spain*

### Large Scale Data Mining: Challenges and Responses
*Jaturon Chattratichat, John Darlington, Moustafa Ghanem, Yike Guo, Harald Hüning, Martin Köhler, Janjao Sutiwaraphun, Hing Wing To, and Dan Yang, Imperial College of London, United Kingdom*

### Using Artificial Intelligence Planning to Automate Science Data Analysis for Large Image Databases
*Steve Chien, Forest Fisher, and Helen Mortensen, Jet Propulsion Laboratory, California Institute of Technology; Edisanter Lo and Ronald Greeley, Arizona State University*

### Mining Multivariate Time-Series Sensor Data to Discover Behavior Envelopes
*Dennis DeCoste, Jet Propulsion Laboratory, California Institute of Technology*

### Why Does Bagging Work? A Bayesian Account and its Implications
*Pedro Domingos, University of California, Irvine*

### Fast Committee Machines for Regression and Classification
*Harris Drucker, Monmouth University and ATT*

### A Guided Tour through the Data Mining Jungle
*Robert Engels, University of Karlsruhe, Germany; Guido Lindner, Daimler Benz AG, Germany; and Rudi Studer, University of Karlsruhe, Germany*

### Maximal Association Rules: A New Tool for Mining for Keyword Co-occurrences in Document Collections
*Ronen Feldman, Yonatan Aumann, Amihood Amir, and Amir Zilberstein, Bar-Ilan University, Israel; Willi Kloesgen, German National Research Center for Information Technology, Germany*

### Improving Scalability in a Scientific Discovery System by Exploiting Parallelism
*Gehad Galal, Diane J. Cook, and Lawrence B. Holder, University of Texas at Arlington*

4:20-5:40 PM

*Poster Session II*

## Authors H-O

### Deep Knowledge Discovery from Natural Language Texts
*Udo Hahn and Klemens Schnattinger, Freiburg University, Germany*

### Integrating and Mining Distributed Customer Databases
*Ira J. Haimowitz, Özden Gür-Ali, and Henry Schwarz, General Electric Corporate Research and Development*

### GA-Based Rule Enhancement in Concept Learning
*Jukka Hekanaho, Turku Center for Computer Science and Åbo Akademi University, Finland*

### Target-Independent Mining for Scientific Data: Capturing Transients and Trends for Phenomena Mining
*Thomas H. Hinke, John Rushing, Heggere Ranganath, and Sara J. Graves, University of Alabama in Huntsville*

### Zeta: A Global Method for Discretization of Continuous Variables
*K. M. Ho and P. D. Scott, University of Essex, United Kingdom*

### Adjusting for Multiple Comparisons in Decision Tree Pruning
*David Jensen and Matt Schmill, University of Massachusetts, Amherst*

### SIPping from the Data Firehose
*George H. John, IBM Almaden Research Center and Brian Lent, Stanford University*

### Mining Generalized Term Associations: Count Propagation Algorithm
*Jonghyun Kahng, Wen-Hsiang Kevin Liao, and Dennis McLeod, University of Southern California*

### Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes
*Micheline Kamber, Jiawei Han, and Jenny Y. Chiang, Simon Fraser University, Canada*

# Poster Sessions

Scalable, Distributed Data Mining–An Agent Architecture
*Hillol Kargupta, Ilker Hamzaoglu, and Brian Stafford, Los Alamos National Laboratory*

Clustering Sequences of Complex Objects
*A. Ketterlin, LSIIT, France*

A Unified Notion of Outliers: Properties and Computation
*Edwin M. Knorr and Raymond T. Ng, University of British Columbia, Canada*

Mining for Causes of Cancer: Machine Learning Experiments at Various Levels of Detail
*Stefan Kramer, Austrian Research Institute for Artificial Intelligence, Austria; Bernhard Pfahringer, University of Waikato, New Zealand; and Christoph Helma, University of Vienna, Austria*

Trellis Graphics Displays: A Multi-Dimensional Data Visualization Tool for Data Mining
*R. Douglas Martin, MathSoft, Inc.*

Fast Robust Visual Data Mining
*Ted Mihalisin, Temple University and John Timlin, Mihalisin Associates, Inc.*

5:40–7:00 PM

*Poster Session III*

## Authors P-Z

Beyond Concise and Colorful: Learning Intelligible Rules
*Michael J. Pazzani, Subramani Mani, and W. Rodman Shankle, The University of California, Irvine*

Scaling Up Inductive Algorithms: An Overview
*Foster Provost, NYNEX Science and Technology and Venkateswarlu Kolluri, University of Pittsburgh*

Visualizing Bagged Decision Trees
*J. Sunil Rao, Cleveland Clinic and William J.E. Potts, SAS Institute Inc.*

KESO: Minimizing Database Interaction
*Arno Siebes and Martin L. Kersten, CWI, The Netherlands*

Learning to Extract Text-Based Information from the World Wide Web
*Stephen Soderland, University of Washington*

Image Feature Reduction through Spoiling: Its Application to Multiple Matched Filters for Focus of Attention
*Timothy M. Stough and Carla E. Brodley, Purdue University*

Autonomous Discovery of Reliable Exception Rules
*Einoshin Suzuki, Yokohama National University, Japan*

An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases
*Shiby Thomas, Sreenath Bodagala, Khaled Alsabti, and Sanjay Ranka, University of Florida*

Bayesian Inference for Identifying Solar Active Regions
*Michael Turmon and Saleem Mukhtar, Jet Propulsion Laboratory, California Institute of Technology; Judit Pap, University of California, Los Angeles*

Schema Discovery for Semistructured Data
*Ke Wang and Huiqing Liu, National University of Singapore, Singapore*

Selecting Features by Vertical Compactness of Data
*Ke Wang and Suman Sundaresh, National University of Singapore, Singapore*

Knowledge Discovery in Integrated Call Centers: A Framework for Effective Customer-Driven Marketing
*Paul Xia, EIS International Inc.*

New Algorithms for Fast Discovery of Association Rules
*M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, University of Rochester*

Fast and Intuitive Clustering of Web Documents
*Oren Zamir, Oren Etzioni, Omid Madani, and Richard M. Karp, University of Washington*

KDD Process Planning
*Ning Zhong, Yamaguchi University, Japan; Chunnian Liu, Beijing Polytechnic University, China; Yoshitsugu Kakemoto, The University of Tokyo, Japan; Setsuo Ohsuga, Waseda University, Japan*

Optimal Multiple Intervals Discretization of Continuous Attributes for Supervised Learning
*D. A. Zighed, R. Rakotomalala, and F. Feschet, University of Lyon 2, France*

A Dataset Decomposition Approach to Data Mining and Machine Discovery
*Blaz Zupan and Marko Bohanec, Institute Jozef Stefan, Slovenia; Ivan Bratko, Institute Jozef Stefan and University of Ljubljana, Slovenia; Bojan Cestnik, Temida and Institute Jozef Stefan, Slovenia*

# Exhibitors

## AcknoSoft

Contact: Michel Manago
58 rue du dessouse des berges
75013, Paris France
Tel: 331 44 24 88 00
Fax: 331 44 24 88 66
Email: manago@ibpc.fr
Web: www.acknosoft.com

The KATE suite includes KATE datamining to generate automatically decision trees from cases, KATE-CBR to retrieve cases that are similar to a query and reason by analogy, KATE-editor to build a database that is object-oriented or link to most existing market databases. As an option, users may add KATE for R/3 Service Management to link to SAP and KATE WebServer to access the decision support system over an intranet or the internet. Written in C, KATE is also available as dynamic link libraries in Windows or as shared libraries on Unix (Sun Solaris). The KATE suite is tailor made for analyzing complex data in technical domains. Applications of KATE include maintenance of Boeing 737 engines, of marine diesel engines, of Naples' metro, of the French high speed train TGV, help desk for robots in the plastic industry, telecom networks, CAD/CAM workstations, reliability analysis of offshore platforms, nuclear power plants, industrial gas meters, the Ariane space center, quality management in manufacturing, evaluation of costs for manufacturing plastic parts, sales support for electronic devices.

## IBM Corporation

Contact: Carolina Salcedo
Route 100, Maildrop 3128
Sommers, NY 10589
Tel: (914) 766-3929
Fax: (914) 766-8328
Email: csalcedo@vnet.ibm.com16

Researchers at many IBM labs around the world are continuously developing powerful algorithms for analyzing large data sets stored in databases or flat files. These algorithms cover a full spectrum of data mining technologies and enable analyses ranging from classification and predictive modeling to association discovery and database segmentation. They can run on small workstations but are highly scaleable since they have parallel implementations optimized to handle large and parallel super computers and databases. Along with a comprehensive set of data analysts and application developers through IBM's flagship data mining technology product, the Intelligent Miner. Leveraging the Intelligent Miner technology, IBM has developed a collection of applications, Business Discovery Solutions (BDS), to make data mining more accessible to business users. Through an easy to use Java GUI, BDS addresses business problems such as customer retention, risk analysis, store layout optimization, and cross selling. IBM's top-caliber data mining analysts have extensive industry expertise and have helped more than 60 companies exploit the new developments in data mining. IBM, on whose machines 70% of the world's data reside, supports all the components required to guarantee a customer's success in data mining.

## Information Discovery, Inc.

Contact: Pamela Lerwick
703B Pier Avenue, Suite 169
Hermosa Beach, CA 90254
Tel: (310) 937-3600
Fax: (310) 937-0967
Email: datamine@ix.netcom.com
Web: www.datamining.com1

The Data Mining Suite™ is a comprehensive and integrated set of data mining products that provide complete solutions for knowledge discovery, predictive modeling and internet/intranet-based applications within a unified frame-work. Each product is novel and useful in its own right, but the joint application of the techniques used within the suite delivers unprecedented benefits to corporate users. The Data Mining Suite™ is highly scaleable and accesses large SQL databases directly without sampling or extracts.

## Kluwer Academic Publishers

Contact: Adam Chesler
PO Box 358, Accord Station
Hingham, MA 02018-0358
Tel: (617) 871-6600
Fax: (617) 871-6528
Email: achesler@wkap.com19
Web: www.wkap.nl20

Kluwer Publishers will have the journal *Data Mining and Knowledge Discovery* (Editors-in-chief: Usama Fayyad, Heikki Mannila, and Gregory Shapiro-Piatetsky) on display! Many other fine journals are also available for review, as well as over 25 new books, discounted 20% for KDD '97 attendees!

## MathSoft, Inc.

Contact: Tina Styer
Data Analysis Products Division
1700 Westlake Ave. N. #500
Seattle, WA 98109
Tel: 800-569-0123
Email: mktg@statsci.com
Web: www.mathsoft.com

MathSoft, Inc., is a leading provider of knowledge discovery software, with more than a million users worldwide. Mathsoft has a flexible family of products, offering solutions for data analysis, statistical data mining and decision support. S-Plus is the premier solution for powerful data analysis, visualization and statistical data mining. S-Plus offers the richest data analysis environment available, with over 2,000 built-in functions including both classical and robust techniques, all within a customizable, intuitive user interface. In addition, exclusive TRELLIS graphics allow you to reveal hidden meanings in complex, multidimensional data. MathSoft Stat Server is a powerful new approach that delivers a competitive edge in decision support. The first enterprise-wide solution for distributing sophisticated analyses and

# Exhibitors

graphics, StatServer leverages your company's existing client/server and Internet/intranet technology to put information in the hands of decision makers. Stop by to see a demonstration of the power of S-Plus and StatServer.

## Morgan Kaufmann Publishers

Contact: Patricia Kim
340 Pine Street, 6th Floor
San Francisco, CA 94104-3205
Tel: (650) 392-2665
Fax: (650) 982-2665
Email: pkim@mkp.com23
Web: www.mkp.com24

Since 1984, Morgan Kaufmann has published the finest technical information resources for computer and engineering professionals. Our audience includes the research and development communities, information technology (IS/IT) managers, and students in professional degree programs. We publish in book and digital form in such areas as databases, computer networking, computer systems, human computer interaction, computer graphics, multimedia information and systems, artificial intelligence, and software engineering. Many of our books are considered to be the definitive works in their fields. Please stop by our display and receive a 15% Knowledge Discovery and Data Mining 1997 conference discount.

## *PC AI* Magazine

Contact: Robin Okun
P.O. Box 30130
Phoenix, AZ 85046
Tel: (602) 971-1869
Fax: (602) 971-2321
Email: robin@pcai.com25
Web: www.pcai.com/pcai/26

*PC AI Magazine* provides the information necessary to help managers, programmers, executives, and other professionals understand the quickly unfolding realm of artificial intelligence (AI) and intelligent applications (IA). PC AI addresses the entire range of personal computers including the Mac, IBM PC, neXT, Apollo, and more. *PC AI* features developments in expert systems, neural networks, object oriented development, and all other areas of artificial intelligence. Feature articles, product reviews, real-world application stories, and a Buyer's Guide present a wide range of topics in each issue.

## Salford Systems

Contact: Kerry Martin
8880 Rio San Diego Dr.
Suite 1045
San Diego, CA 92108
Tel: (619) 543-8880

Salford Systems is exhibiting CART 3.0 (Classification and Regression Trees), a tree-structured data-mining tool, co-developed with the original authors of CART at UC Berkeley and Stanford (Breiman, Freidman, Olshen, and Stone). The premier decision tree tool-complete with built-in n-fold cross-validation, user-definable variable misclassification costs, linear combination splits, efficient handling of high-dimensional categorical predictors, and the new feature of combining multiple trees-is now available in one affordable package. CART 3.0 has a completely revised Windows graphical user interface and the ability to interact with the tree after database analysis. Useful diagnostics include dynamic pruning, gains charts for all sub-trees, and simultaneous viewing of training and test data accuracy scores. Experienced users can control CART 3.0 via command scripts, while newcomers can use point-and-click menu selections. Throughout an interactive session using the menu, CART 3.0 records command equivalents, providing an audit trail for the session. An in-depth, comprehensive manual explains every feature and nuance of CART within the context of over 30 examples. Salford Systems has been developing advanced tools for data analysis for PC and Unix platforms since 1983, and also provides consulting services to the telecommunications, financial services, health care and direct mail industries.

## Silicon Graphics

Contact: Ron Kohavi
Mailstop 80-876
2011 N. Shoreline Blvd.
Mountain View, CA 94043
Tel: (650) 933-3126
Fax: (650) 932-2874
Web: www.sgi.com27

MineSet™ version 2.0 is the fourth release of SGI's product for exploratory data analysis. Combining powerful integrated, interactive tools for data access and transformation, data mining, and visual data mining, MineSet provides you with a revolutionary paradigm for getting maximum value from your vast data resources. MineSet enables you to gain a deeper, intuitive understanding of your data, by helping you to discover hidden patterns, important trends and new knowledge. It is this deep understanding which can be used for developing powerful business strategies leading to greater competitive advantage.

## SRA International

Contact: Jim Hayden
4300 Fair Lakes Court
Fairfax, VA 22033
Tel: (703) 803-1689
Fax: (703) 803-1793
Email: Jim_Hayden@SRA.com28
Web: www.SRA.com29

SRA International has been creating innovative solutions to practical problems faced by businesses and government agencies for over eighteen years. We specialize in the fields of intelligent information retrieval; machine learning; knowledge-based systems; database engineering; and natural language processing. SRA empowers organizations with the ability to discover and detect patterns critical to their success through the use of a complete line of scaleable

# Exhibitors

data mining tools and professional services. SRA's KDD Toolset includes multi-strategy algorithms for discovering associations, classifications, sequences, and clusters, as well as high-speed rule and sequence-based pattern matching algorithms. These algorithms employ direct database access for mining data. Additionally, they are parallelized to take advantage of multiprocessor platforms for rapid analysis of extremely large data sets. Finally, we employ a comprehensive set of JDBC-compliant Java-based user interfaces for configuration and execution of algorithms as well as visualization of results for analysis and interpretation. SRA's knowledge discovery specialists understand how best to apply these advanced capabilities to enable you to utilize your most strategic asset: electronic information. Together, SRA's KDD Toolset and professional services provide solutions that are ideal for tackling such large-scale problems as fraud detection and prevention, competitive intelligence, and market behavior.

## Torrent Systems Inc.

Contact: Sondra Barrison
5 Cambridge Center
Cambridge, MA 02138
Tel: (617) 354-8484
Fax: (617) 354-6767
Email: sondra@torrent.com30

Torrent Systems, Inc. develops and markets tools, component software and applications to support system integrators and applications developers in building advanced data warehousing and data mining applications that run on parallel processing systems. ORCHESTRATE, Torrentis parallel development environment, hides the complexity of parallel programming and facilitates the creation of fully parallel, high-performance data-processing solutions for your MPP or SMP systems. ORCHESTRATE is fully compatible with all major scalable servers including the IBM RS/6000, SP server using IBM DB2, Parallel Sysplex, Sun, HP, NCR, Digital and Intel and supports Oracle Parallel Server, Informix XPS, and IBM DB2/PE. Torrent Systems is headquartered in Cambridge, MA.

## Toshiba Corporation

Contact: Hiroshi Tsukimoto
70 Yanagi-cho, Saiwai-ku
Kawasaki
Japan
Tel: 81-44-548-5469
Fax: 81-44-520-5856
Email: tukimoto@ssel.toshiba.co.jp31

NNE is a neural network based data mining tool. NEX (Neural network EXplainer) is the explanation module of NNE. NEX provides the explanation for trained neural networks by extracting rules from the networks. Since trained neural networks are black boxes and are difficult for humans to understand, the neural network is incomplete as a technique of data mining, which aims to discover understandable knowledge from databases. So NEX makes the neural network a complete data mining technique. NEX has the following features. (1) NEX can be applied to any neural network including recurrent neural networks. (2) NEX can be applied to any training method. (3) NEX can be applied not only to discrete values but also to continuous values. (4) NEX extracts accurate and simple rules in a short time. Especially, when classes are continuous, there is no other systematic method which can discover understandable knowledge. NEX discovers understandable knowledge, that is, rules, from trained neural networks. The accuracies of the rules are based on the trained networks. NEX can be connected to any neural network. So any neural network user can obtain an explanation for trained neural networks just by connecting NEX to their neural networks.

## WizSoft Ltd.

Contact: Abraham Meidan
3 Beit Hillel St.
Tel Aviv
Israel, 67017
Tel: 972-2-5631948
Fax: 972-3-5611945
Email: abraham@wizsoft.com32
Web: www.wizsoft.com33

Wiz Why for WINDOWS 95/WINDOWS NT is a data mining application for issuing prediction and classification. WizWhy analyzes the data, reveals all the if-then rules and mathematical formula rules, and calculates the significance level of each rule. WizWhy then predicts future cases based on the discovered rules. In empirical tests WizWhy was found to be faster and more accurate than neural networks, decision trees and genetic algorithms.

WizRule for WINDOWS 95/WINDOWS NT is a data cleansing and auditing tool. WizRule reveals all if-then rules and formula rules in the database. It then points at the deviations from the set of all the discovered rules as suspected errors, and calculates the level of unlikelihood of each deviation. WizRule avoids false alarms; almost every deviation with a high level of unlikelihood is indeed an error.

# Demos

## An Interactive Visualization Environment for Data Exploration

*Paper Title:* An Interactive Visualization Environment for Data Exploration

*Development Team:* Mark Derthick, John Kolojejchick, and Steven F. Roth, Carnegie Mellon University

*Telephone:* 412-268-8812

We will demonstrate an information-centric interface architecture for unifying the subtasks of knowledge discovery, allowing the analyst to focus on the process rather than the tools. These subtasks include data cleaning, creating a dataset, data reduction and projection, and exploratory visualization. Architectural integration is achieved with a shared object-oriented database, which is accessible to the user via a visual query language. Tight integration between queries and visualizations make exploration more interactive and less of a feed forward process than in previous systems.

*What Is Unique about the System?* Tight integration of querying and visualization. Interactive visualizations involving attributes of multiple objects.

## Document Explorer

*Paper Titles:* (1) Visualization Techniques to Explore Data Mining Results for Document Collections, and (2) Maximal Association Rules: A New Tool for Mining for Keyword Co-Occurrences in Document Collections

*Development Team:* Ronen Feldman and Amir Zilberstein, Bar-Ilan University; Willi Kloesgen, GMD

*Telephone:* 972-3-5318629 or 972-3-9326702

Document Explorer is a data mining system for document collections. Such a collection represents an application domain, and the primary goal of the system is to derive patterns that provide knowledge about this domain. Additionally, the derived patterns can be used to browse the collection. Document Explorer searches for patterns that capture relations between concepts of the domain. The patterns which have been verified as interesting are structured and presented in a visual user interface allowing the user to operate on the results to refine and redirect mining queries or to access the associated documents. The system offers preprocessing tools to construct or refine a knowledge base of domain concepts and to create an intermediate representation of the document collection that will be used by all subsequent data mining operations. The main pattern types the system can search for are frequent sets, associations, concept distributions, and keyword graphs. To provide some explicit bias, the system provides a dedicated query language for searching the vast implicit spaces of pattern instances that exist in the collection. This query language offers syntactical, background, quality and redundancy constraints. The query language is embedded in a GUI which makes it easy even for novice users to explore the document collections.

*What Is Unique about the System?* (1) Dealing with unstructured information; (2) unique visualization tools; (3) special query language designed for text mining; and (4) unique browsers that enable interactive exploration.

## GeoMiner: A Geo-Spatial Data Mining Engine

*Paper Title:* Described in Proceedings of SIGMOD'97

*Development Team:* Jiawei Han, Krzysztof Koperski Nebojsa Stefanovic, and Qing Chen, Simon Fraser University Intelligent Database Systems Research Lab

*Telephone:* (604) 291-4411

Spatial data mining is to mine high-level spatial information and knowledge from large spatial databases. A spatial data mining system prototype, GeoMiner, has been designed and developed based on our years of experience in the research and development of relational data mining system, DBMiner, and our research into spatial data mining. The data mining power of GeoMiner includes five spatial data mining modules: characterization, comparison, association, clustering, and classification. The SAND (Spatial And Nonspatial Data) architecture is applied in the modeling of spatial databases, whereas GeoMiner includes the spatial data cube construction module, spatial on-line analytical processing (OLAP) module, and spatial data mining modules. A spatial data mining language, GMQL (Geo-Mining Query Language), is designed and implemented as an extension to Spatial SQL, for spatial data mining. Moreover, an interactive, user-friendly data mining interface is constructed and tools are implemented for visualization of discovered spatial knowledge.

*What Is Unique about the System?* A spatial data mining system performing knowledge discovery based on both spatial and non-spatial properties of objects. It also includes spatial OLAP modules and tools for visualization of discovered spatial knowledge.

# Demos

## Id-Vis

*Paper Title:* A Visual Interactive Framework for Attribute Discretization

*Development Team:* Ramesh Subramonian, Ramana Venkata, and Joyce Chen, Microcomputer Research Laboratory, Intel Corporation

*Telephone:* (408) 653 - 6794 (Ramana Venkata)

We will demonstrate the Discretizer module of Id-Vis, our interactive platform for visual data mining on a client-server architecture. This module features multiple available algorithms, a drag-and-drop cut-point editor, multiple levels of data visualization with drill-down capability etc. A number of the variables are exposed to user experimentation. The system provides visual cues to the "optimal" number and locations of the cut-points. It also provides feedback to the user about the extra "badness," over the system-derived optima, incurred during the experimentation.

The central philosophy is that the system should place the user within the appropriate context of system-derived values, and provide the user with the opportunity to intelligently modify (e.g. display the impact on accuracy of these modifications) these optima and propagate them downstream. The server originally ships a compacted version of the raw data, or a equi-probabilistically distilled density function to the client. If, during the drill-down process, the user's information request and the associated accuracy constraints cannot be satisfied by the locally available data, the client obtains the appropriate data from the server and displays it.

*What Is Unique about the System?* Visual interactivity; opportunity for the user to encode his or her intuition/domain knowledge into and during the mining process; a feedback-loop paradigm of data mining as a learning process.

## Interactive Knowledge Exploration Using DBMiner

*Paper Title*: Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes

*Development Team:* Jiawei Han, Jenny Y. Chiang, Sonny Chee, Shan Cheng, Wan Gong, Micheline Kamber, Kris Koperski, Yijun Lu, Nebojsa Stefanovic, Lara Winstone, Betty Xia, Osmar R. Zaiane, and Hua Zhum, Simon Fraser University

*Telephone:* (604) 291-4411

With years of research and development efforts, the DBminer system developed in CS/SFU, Canada has incorporated many advanced research results into our system. This includes multiple-level knowledge in large relational databases and data warehouses, a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction, and clustering, and a data visualization package. The major technologies adopted are integration with data warehouse and OLAP technology, attribute-oriented induction, statistical analysis, progressive deepening for mining multiple-level knowledge, and meta-rule guided mining. The system provides a user-friendly, interactive data mining environment with good performance.

*What Is Unique about the System?* Integration with OLAP, multiple-level mining and multiple data mining modules.

## Kensington - High-Performance Distributed and Parallel Data Mining

*Paper Title:* Large Scale Data Mining: Challenges and Responses

*Development Team:* J. Chattratichat, J. Darlington, M. Ghanem, Y. Guo, M. Kohler, A. Saleem, J. Sutiwaraphun, and D. Yang, Department of Computing, Imperial College

*Telephone:* +44-171-5948360

Kensington is a prototype of an open, distributed, web-based, high-performance data mining system for use on parallel servers. A web-based client-tool, written in Java, gives access to a distributed collection of data bases and data mining modules which are executed on a high-performance parallel machine. The application consists of components which are integrated using Java/Corba middleware. The main components are: a database server, a data mining server, a visualisation and Web-server, and the client control tool. The data mining servers are embedded in Corba objects and distributed across a LAN or WAN. The database server is accessed via JDBC. The client-tool can access and control the data mining actions from anywhere. The high-performance mining modules are portable parallel implementations in C and MPI, and cover commonly needed functions such as classification (C4.5), prediction (neural networks), association discovery and self-organizing maps. The light-weight client tool is enriched by visualization applets for data mining results, whenever they become available. The overall goal is to (a) integrate distributed servers, e.g. data and computation servers, and (b) to make the integrated system universally accessible over the Web.

*What Is Unique about the System?* The Kensington architecture combines a flexible integration approach, based on distributed object technology, with high-performance datamining components. The component technology is

# Demos

platform-independent, and allows straightforward extension of the system. Accessibility over the Web is an inherent part of all components. In summary, the key features of Kensington are: accessibility, extensibility, distributed object architecture, platform-independence, high performance components.

## Mining For Many Kinds of Knowledge

*Paper Title:* Knowledge = Concepts: A Harmful Equation

*Development Team:* Arun Sanjeev and Jan Zytkow, Wichita State University

*Telephone:* 316-978-3015 (Arun Sanjeev) or 316-978-3925 (Jan Zytkow)

We will demonstrate Forty-Niner (49er), an automated discovery system that discovers knowledge in databases. We will apply the system to several databases to demonstrate how different forms of knowledge can be automatically discovered. 49er searches for many types of knowledge. It starts from contingency tables (CTs) and then recognizes special types of CTs which lead to other, more specialized forms of knowledge, such as equations, equivalence relations, or subset relations. When many relations of the same type have been discovered, 49er combines them into forms such as taxonomies and subset graphs. We will contrast 49er with specialized systems that are focused on a single form of knowledge, even if other forms of knowledge are much more appropriated for a given dataset. Further, we will contrast 49er's focus on knowledge that contains as much of empirical contents as possible with the focus on concept definitions typical to machine learning. We will show how CTs can be used to generate decision trees and how 49er's search for additional "redundant" knowledge makes decision trees more flexible and statistically significant. We will also contrast 49er's approach to taxonomy formation (combine many approximate equivalence relations) with statistical and conceptual clustering.

*What Is Unique about the System?* 49er is an autonomous knowledge discoverer. It automatically tunes itself to the forms of knowledge that are appropriate for a given dataset: equations, contingency tables, taxonomies, decision trees, and the like. 49er explores huge hypotheses spaces, evaluating the strength (to ensure predictive power) and significance of results (to prevent overfit).

## SONAR (System for Optimized Numeric Association Rules)

*Paper Title:* Computing Optimized Rectilinear Regions for Association Rules

*Development Team:* Takeshi Fukuda, Yasuhiko Morimoto, Hirofumi Matsuzawa, Shinichi Morishita, Takeshi Tokuyama, and Kunikazu Yoda, IBM Tokyo Research Laboratory

*Telephone:* +81-462-73-4946 (Kunikazu Yoda)

Recent progress in technologies for data input have made it easier for finance and retail organizations to collect massive amounts of data and to store them on disk at a low cost. Such organizations are interested in extracting from these huge databases previously unnoticed information that inspires new marketing strategies. In this demonstration, we introduce a system for mining optimized association rules and for generating decision/regression trees from databases with numeric data as well as categorical data.

*What Is Unique about the System?* Our system uses novel algorithms for efficiently creating ranges and regions with respect to various optimization criteria such as maximization of confidence or support, and minimization of entropy and mean squared error.

## S-PLUS DataBlade for Informix Universal Server

*Paper Title:* Data Mining with Trellis Graphics

*Development Team:* Kevin Brown and Jun Luo, Informix; Vikram Chalana, Scott Blachowicz, Marianna Clark and Doug Martin, Mathsoft

*Telephone:* (206) 283-8802 x229 (Doug Martin)

Informix Universal Server (IUS) is an object-relational data base. S-PLUS is an object-oriented language and system for data analysis, statistical modeling, visualization and programming with data. An IUS DataBlade is a collection of types (classes) of objects and access methods, that closely integrates applications software with the IUS database, typically on the server. The S-PLUS datablade for IUS provides new data types in IUS corresponding to intrinsic S-PLUS datatypes and functions to apply any S-PLUS expression on these datatypes. It also provides functions to convert IUS native data to these S-PLUS datatypes and vice-versa. The demonstration will include several data mining applications examples, including: integrated query and statistical data mining, visualizing and modeling the relationship between equity returns, firm size and book-to-market; robust beta mining (finding firms listed on the AMEX, NASDAQ and NYSE or which the beta calculation is influenced by outliers, and visualizing the data for such firms); hexagonal binding visualization of scatterplots for largish data sets; application of trellis graphics to a Lucent customer value analysis (CVA) study.

*What Is Unique about the System?* The object-oriented aspects of IUS and S-PLUS make the S-PLUS DataBlade a natural marriage of the two technologies: arbitrary object types can have mirror images in the IUS data base. Furthermore, the use of the IUS SQL93 is smoothly integrated with the use of S-PLUS functions.