



KDD-2000

The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

August 20-23, 2000
Boston, MA, USA

Final Program

Welcome to KDD-2000!

Letter from the Chairs

Welcome to the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. This premier conference is sponsored by the Association for Computing Machinery (ACM) and organized by its Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), with co-sponsorship from AAI, ACM SIGMOD, and ACM SIGART, and in cooperation with Interface and ASA SCS.

Acceptance of papers for the conference proceedings was extremely competitive. The program committee selected 26 papers for inclusion as full papers from 248 research paper submissions. Another 24 papers were selected as poster papers. The selected papers cover a wide range of topics and application areas, and reflect the vigor and excitement of this research area. KDD 2000 is also featuring an industrial track (IT) that received 36 submissions with 12 selected for inclusion in the proceedings as industrial track papers. Four other papers were selected for the IT track from the pool submitted to the research track, and 2 other papers were included by invitation.

Seven invited presentations by leading researchers and industry veterans cover topics ranging from data mining techniques to privacy and commercial implications of data mining. KDD 2000 also features three panels addressing personalization, privacy, and KDD process standards, and six tutorials cover hypertext, visualization, biology applications, customer relationship management, time series similarity, and high performance data mining. Five workshops have also been organized covering topics in data mining the web, distributed and parallel approaches, multimedia, text, and the role of post processing in Machine Learning and data mining in general. The popular KDD Cup competition for applications of data mining focuses on web clickstream data from an apparel retailer.

The quality of the SIGKDD 2000 Conference was ensured through the hard work of a large number of people. We would like to thank the 57 program committee members for reading, reviewing, and discussing a great many papers in order to identify the very best ones. The professionalism of the organizing committee is evident in the breadth and depth of the entire set of topics presented at the conference. We hope you find the conference insightful and helpful in continuing new and interesting lines of research in knowledge discovery and data mining.

Raghu Ramakrishnan and Sal Stolfo

Program Committee Co-chairs

Ismail Parsa

General Conference Chair

Program Highlights

- Keynote and Invited Talks by:
 - Bruce Buchanan
 - Jason Catlett
 - Matt Cutler
 - James Goodnight
 - Christos H. Papadimitriou
 - Michael Saylor
 - David Stodder
- Presentation of 26 research papers and 18 industry papers.
- Poster session featuring 24 research papers.
- Panels on:
 - Personalization
 - Privacy
 - KDD Process Standards
- Tutorials on:
 - Data Mining for Hypertext
 - Multidimensional Visualization
 - Knowledge Discovery in Biological Domains
 - Data Mining for Successful Customer Management
 - Time Series Similarity Measures
 - High Performance Data Mining
- Workshops on:
 - Web Mining for E-Commerce
 - Distributed & Parallel Knowledge Discovery
 - Multimedia Data Mining
 - Text Mining
 - Post Processing in Machine Learning and Data Mining
- KDD Cup Awards
- Exhibitors of commercial systems, research prototypes, and products for knowledge discovery and data mining.

Sunday, August 20th

9:00 am - 4:00 pm

Workshop on Distributed and Parallel Knowledge Discovery [Beacon Hill 2]

Chairs: Hillol Kargupta (Washington State University), Joydeep Ghosh (University of Texas at Austin), Vipin Kumar (University of Minnesota), Zoran Obradovic (Washington State University)

Workshop on Multimedia Data Mining [Cambridge Complex]

Chairs: Simeon J. Simoff (University of Sydney), Osmar R. Zaiane (University of Alberta)

Workshop on Post Processing in Machine Learning and Data Mining [Beacon Hill 1]

Chairs: A. (Fazel) Famili (National Research Council of Canada), Ivan Bruha (McMaster University)

Workshop on Text Mining [Waterfront 1]

Chairs: Marko Grobelnik (J.Stefan Institute), Dunja Mladenic (J.Stefan Institute and Carnegie Mellon University), Natasa Milic-Frayling (Microsoft Research Ltd)

Workshop on Web Mining for E-Commerce (WEBKDD'2000) [North End Complex]

Chairs: Ronny Kohavi (Blue Martini), Myra Spiliopoulou (Humboldt University), Jaideep Srivastava (Yodlee.com)

9:00 am - 12:00 pm

Tutorial Session M1: Data Mining for Hypertext [Waterfront 2]

Instructor: Soumen Chakrabarti (IIT Bombay)

Tutorial Session M2: Multidimensional Visualization for High Dimensional Datasets and Multivariate Relations [Amphitheater]

Instructor: Alfred Inselberg (Tel Aviv University)

Tutorial Session M3: Knowledge Discovery in Biological Domains [Waterfront 3]

Instructors: I. Jurisica (University of Toronto), I. Rigoutsos (IBM T. J. Watson), A. Floratos (IBM T. J. Watson)

10:00 am - 10:30 am

Catered Break [Commonwealth Hall]

12:00 pm - 1:00 pm

Lunch (on your own)

1:00 pm - 4:00 pm

Tutorial Session A1: Data Mining for Successful Customer Management (CRM) [Amphitheater]

Instructors: Gregory Piatetsky-Shapiro (Xchange), Steve Gallant (Xchange), Dorian Pyle (Xchange)

Tutorial Session A2: Time Series Similarity Measures [Waterfront 3]

Instructors: Gautam Das (Microsoft Research), Dimitrios Gunopulos (University of California, Riverside)

Tutorial Session A3: High Performance Data Mining [Waterfront 2]

Instructors: Vipin Kumar (University of Minnesota), Mohammed Zaki (Rensselaer Polytechnic Institute)

4:00 pm - 4:30 pm

Catered Break [Commonwealth Hall]

4:30 pm - 5:00 pm

Conference Opening and Awards [Commonwealth Hall]

5:00 pm - 6:00 pm

Keynote Talk I [Commonwealth Hall]

On Certain Rigorous Approaches to Data Mining, Christos H. Papadimitriou (UC-Berkeley)

6:00 pm - 7:00 pm

KDD-Cup Awards [Commonwealth Hall]

Monday, August 21st

7:00 am - 8:30 am

Continental Breakfast [Commonwealth Hall]

8:30 am - 10:00 am

Plenary Session [Commonwealth Hall]

Keynote Talk II: *Among Those Dark Electronic Mills: Privacy and Data Mining*, Jason Catlett (President and Founder, Junkbusters)

Best Research Paper Presentation: *Hancock: A Language for Extracting Signatures from Data Streams*, C. Cortes, K. Fisher, D. Pregibon, A. Rogers, (AT&T Labs—Research, Shannon Laboratory), F. Smith (Cornell University).

10:00 am - 10:30 am

Catered Break [Commonwealth Hall]

10:30 am - 12:00 pm

Research Track Session I: Constraints and Evaluation in the KDD Process [Commonwealth Hall]

An Empirical Analysis of Techniques for Constructing and Searching K-Dimensional Trees, D. Talbert and D. Fisher (Vanderbilt University).

Generating Non-Redundant Association Rules, M. Zaki (Rensselaer Polytechnic Institute).

Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation's Advanced Detection System (ADS). Runner-up best applications paper. T. Senator (NASD Regulation, Inc.).

Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns, B. Padmanabhan (University of Pennsylvania) and A. Tuzhilin (New York University).

Industrial Track Invited Talks [Amphitheater]

E-metrics: Tomorrow's Business Metrics Today, Matt Cutler (Co-Founder and Chief E-Business Intelligence Officer, NetGenesis)

After the Gold Rush: Data Mining in the New Economy, David Stodder (Editorial Director, Intelligent Enterprise).

12:00 pm - 1:30 pm

Catered Lunch [Commonwealth Hall]

1:30 pm - 3:00 pm

Research Track Session II: New KDD Algorithms [Commonwealth Hall]

Data Selection for Support Vector Machine Classifiers, G. Fung and O. L. Mangasarian (University of Wisconsin).

Mining High-Speed Data Streams, P. Domingos and G. Hulten (University of Washington).

Deformable Markov Model Templates for Time-Series Pattern Matching. Runner-up best research paper. X. Ge and P. Smyth (University of California, Irvine).

Active Learning using Adaptive Resampling, V. Iyengar, C. Apte, and T. Zhang (IBM Research).

Industrial Track Session I: Applications [Amphitheater]

Data Mining Solves Tough Semiconductor Manufacturing Problems, R. Gardner and J. Bieker (Motorola).

Genome Scale Prediction of Protein Functional Class from Sequence using Data Mining, R. D. King, A. Karwath, A. Clare (University of Wales, Aberystwyth), and L. Dehaspe (Katholieke Universiteit Leuven).

Data Mining to Detect Abnormal Behavior in Aerospace Data, J. M. Peña (DLSIIS, Facultad de Informática, UPM), F. Famili, and S. Létourneau (Institute for Information Technology, National Research Council, Ottawa, Canada).

Predictive Modeling in Automotive Direct Marketing: Tools, Experiences and Open Issues. R. Wirth, W. Gersten, and D. Arndt (DaimlerChrysler).

3:00 pm - 3:30 pm

Catered Break [Commonwealth Hall]

(continued on next page)

Monday, August 21st

3:30 pm - 5:00 pm

Research Track Session III: Efficiency and Scalability of KDD Algorithms [Commonwealth Hall]

Efficient Search for Association Rules, G. Webb (Deakin University).

Depth First Generation of Long Patterns, R. Agarwal, C. Aggarwal, and V. V. V. Prasad (IBM T. J. Watson Research Center).

Reversing the Dimensionality Curse for Similarity Indexing in High Dimensional Space, C. Aggarwal and P. Yu (IBM T. J. Watson Research Center).

The Generalized Bayesian Committee Machine, V. Tresp (Siemens).

Industrial Track Session II: Text Mining and Data Preparation [Amphitheater]

Agglomerative Clustering of a Search Engine Query Log, D. Beeferman (Lycos Inc.) and A. Berger (Carnegie Mellon University).

Textual Data Mining of Service Center Call Records, P.-N. Tan (University of Minnesota), H. Blau, S. Harp, R. Goldman (Honeywell Technology Center).

Automating Exploratory Data Analysis for Efficient Data Mining, J. Becher, P. Berkhin, and E. Freeman (Accrue Software, Inc.).

Exploration Mining in Diabetic Patients Databases: Findings and Conclusions. W. Hsu, M. Lee, B. Liu, and T. Ling (National University of Singapore).

5:00 pm - 6:00 pm

Poster Previews [Commonwealth Hall]

6:00 pm - 7:30 pm

Discovery Reception and Poster Presentations [Plaza Ballroom]

Tuesday, August 22nd

7:00 am - 8:30 am

Continental Breakfast [Commonwealth Hall]

8:30 am - 10:00 am

Plenary Session [Commonwealth Hall]

Industrial Track Keynote Talk I: *One to One Relationships via Web, Wireless and Voice*, Michael J. Saylor (Founder, President and CEO, MicroStrategy Incorporated).

Industrial Track Keynote Talk II: *Decision Support in the Booming E-World*, James H. Goodnight (President and CEO, SAS).

10:00 am - 10:30 am

Catered Break [Commonwealth Hall]

10:30 am - 12:00 pm

Research Track Session IV: Mining the Web

[Commonwealth Hall]

A General Probabilistic Framework for Clustering Individuals, I. Cadez, S. Gaffney, P. Smyth (University of California, Irvine).

Efficient Identification of Web Communities, G. Flake, S. Lawrence, and C. Lee Giles (NEC Research Institute).

Global Partial Orders from Sequential Data, H. Mannila (Nokia) and C. Meek (Microsoft Research).

Efficient Clustering of High-Dimensional Datasets with Application to Reference Matching, A. McCallum (Just Research), K. Nigam (Carnegie Mellon), and L. Ungar (University of Pennsylvania).

Industrial Track Session III: Targeting Prospects

[Amphitheater]

Cross-Sell: A Fast Promotion-Tunable Customer-Item Recommendation Method Based on Conditionally Independent Probabilities, B. Kitts, D. Freed, and J. Vrieze (Vignette Corp.).

Identifying Prospective Customers, P. Chou (IBM T. J. Watson Research), E. Grossman (IBM Global Intelligence Solutions), D. Gunopulos (University of California, Riverside), P. Kamesam (IBM Insurance Research Center).

(continued on next page)

Tuesday, August 22nd

Targeting the Right Students Using Data Mining, Y. Ma, B. Liu, C.-K. Wong (National University of Singapore), P. Yu (IBM T. J. Watson Research Center), and S. M. Lee (Gifted Education Branch, Ministry of Education, Singapore).

Evolutionary Algorithms in Data Mining: Multi-Objective Performance Modeling for Direct Marketing, S. Bhattacharyya (University of Illinois, Chicago).

12:00 pm - 1:30 pm

Catered Lunch [Commonwealth Hall]

12:45 pm - 1:30 pm

SIGKDD Meeting [Commonwealth Hall -- in conjunction with Lunch Reception]. Chair: Won Kim.

1:30 pm - 3:00 pm

Research Track Session V: Interactive Knowledge Exploration [Commonwealth Hall]

Towards an Effective Cooperation of the Computer and the User for Classification, M. Ankerst, M. Ester, and H.-P. Kreigel (University of Munich).

A Framework for Specifying Explicit Bias for Revision of Approximate Information Extraction Rules, R. Feldman, J. Scler, Y. Liberzon (Instinct Software).

Explicitly Representing Expected Cost: An Alternative to ROC Representation, C. Drummond and R. Holte (University of Ottawa).

Multi-Level Organization and Summarization of the Discovered Rules, B. Liu, M. Hu, and W. Hsu (National University of Singapore).

Industry Track Session IV: E-Commerce and Temporal Data [Amphitheater]

Hybrid Poisson Process, A. Farahat (HNC Software).

Data Mining Techniques for Optimizing Inventories for Electronic Commerce, A. Dhond, A. Gupta, and S. Vadhavkar (Massachusetts Institute of Technology).

Mining the Stock Market: Which Measure Is Best? M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani, (Stanford University).

Discovering Similar Patterns in Time Series, J. P. Caracalente and I. Lopez-Chavarrias (Universidad Politecnica Madrid).

3:00 pm - 3:30 pm

Catered Break [Commonwealth Hall]

3:30 pm - 4:45pm

Panel I: KDD Process Standards [Commonwealth Hall]
Chair: Ismail Parsa (Epsilon). Participants: B. Husick (CPEX), R. Wirth (CRISP-DM), R. Grossman (DMG), U. Fayyad (OLEDBM), E. Thomsen (TASF).

3:30 pm - 4:15 pm

Industrial Track Session V: Telephony/ISP Applications [Amphitheater]

Defection Detection: Using Online Activity Profiles to Predict ISP Customer Vulnerability, N. Raghavan, R. Bell (AT&T Labs Research), M. Schonlau (RAND), D. Pregibon (AT&T Labs Research), A. Karr (National Institute of Statistical Sciences).

Incremental Quantile Estimation for Massive Tracking, F. Chen, D. Lambert, J. Pinheiro (Bell Labs, Lucent Technologies).

4:15pm - 5:30 pm

Industrial Track Panel: Privacy [Amphitheater]

Chair: Jonathan Smith (University of Pennsylvania). Participants: D. Jaye (Engage Technologies), L. Ungar (University of Pennsylvania), Jane Swift (Lt. Governor of Massachusetts), Rakesh Agrawal (IBM Almaden).

4:45 pm - 6:00 pm

Panel II: Personalization and Data Mining [Commonwealth Hall]

Chair: Alex Tuzhilin (NYU). Participants: P. Hagen (Senior Analyst, Forrester Research), R. Kohavi (Director of Data Mining, Blue Martini Software), B. J. Lowell (President and CEO, YOUpowered, Inc.), J. Riedl (University of Minnesota & Co-Founder of Net Perceptions).

6:00 pm - 7:00 pm

KDD Transfer Meeting [Cambridge Complex]

7:00 pm - 7:30 pm

KDD-2000 Organizers Meeting

Wednesday, August 23rd

Keynotes

8:30 am - 10:00 am

Plenary Session [Commonwealth Hall]

Research Track Keynote III: *Informed Knowledge Discovery: Using Prior Knowledge in Discovery Programs*, Bruce Buchanan (University Professor of Computer Science and Professor of Philosophy, Medicine, and Intelligent Systems, University of Pittsburgh).

Best Application Paper Presentation: *Mining IC Test Data to Optimize VLSI Testing*, Tony Fountain (San Diego Supercomputer Center, UCSD), Thomas Dietterich (Oregon State University), Bill Sudyka (Hewlett Packard Company).

10:00 am - 10:30 am

Catered Break [Commonwealth Hall]

10:30 am - 12:00 pm

Research Track Session IV: Visualization

[Commonwealth Hall]

Visualization and the Process of Modeling: A Cognitive-Theoretic View, A. W. Crapo (GE Corporate Research & Development, Rensselaer Polytechnic Institute), L. B. Waisel (Carnegie Mellon University), W. A. Wallace (Rensselaer Polytechnic Institute), T. R. Willemain (Rensselaer Polytechnic Institute).

Visualizing Association Rules with Interactive Mosaic Plots, H. Hofmann (Augsburg University), A. Siebes (CWI), and A. Wilhelm (Augsburg University).

Interactive Exploration of Very Large Relational Data Sets Through 3D Dynamic Projections, L. Yang (Western Michigan University).

RuleViz: A Model for Visualizing Knowledge Discovery Process, J. Han and N. Cercone (University of Waterloo).

12:00 pm - 1:00 pm

Open Forum [Commonwealth Hall]

On Certain Rigorous Approaches to Data Mining

Christos H. Papadimitriou
C. Lester Hogan Professor and Associate Chair
Computer Science Division
EECS Department
University of California, Berkeley

Abstract: In a recent joint paper with Jon Kleinberg and Prabhakar Raghavan we proposed a novel formal approach to interestingness based on considerations from mathematical economics and optimization. Although this approach requires an understanding of the enterprise's business model and environment that is not realistically attainable, I shall argue that it can lead to interesting insights and novel styles of data mining. I will also discuss certain other foundational approaches to important current problems related to data mining, such as formalizing privacy, and sampling web documents uniformly at random.

Biography: Christos Papadimitriou has a Bachelors from Athens Polytechnic and a PhD from Princeton. He has taught at Harvard, MIT, Athens Polytechnic, Stanford, and UCSD. Since 1995 he has been teaching at the University of California Berkeley, where he is the C. Lester Hogan Professor of Electrical Engineering and Computer Science. He has written five books, and over 200 articles on algorithms, complexity, and their applications to various fields, including databases, optimization, artificial intelligence, the life sciences, and economics.

Time and location: Sunday, 5:00pm-6pm [Amphitheater]

Keynotes

Among Those Dark Electronic Mills: Privacy and Data Mining

Jason Catlett
President and Founder, Junkbusters Corporation

Abstract: The concept of privacy encompasses various claims by individuals over whether information about them is communicated to others. Data mining allows discovery of personal data that was previously unknown, possibly even to the individuals concerned. For example, one aim of collaborative filtering is to recommend to people books that they would enjoy but have not read or even heard of. How can the concept of privacy be extended to whether such information is even created? Is there a need to extend the principles of fair information practice, which form the framework for the data protection laws prevailing in most developed countries? What should practitioners consider when designing and implementing data mining projects? Will this technology lead to the promised land of personal empowerment and prosperity or an post-Orwellian dystopia of mass-customized cognitive ghettos? Or just more junk mail?

Biography: Jason Catlett is President and founder of Junkbusters Corp, a leading resource on the control of telemarketing calls, unwanted mail, spam, and commercial invasions of privacy. Described by the Associated Press as a nationally known privacy advocate, Catlett is frequently quoted in the media as a critic of intrusive marketing practices. Catlett's Ph.D. from the University of Sydney explored the induction of decision trees on very large data sets (or at least very large for 1989). He continued this research agenda at AT&T Bell Labs from 1992 to 1996. He is a member of the editorial board of the Machine Learning Journal.

Time and location: Monday, 8:30am-9:30am [Commonwealth Hall]

One to One Relationships via Web, Wireless and Voice

Michael J. Saylor
Founder, President and CEO, MicroStrategy Inc.

Abstract: The mixture of a powerful data warehouses and an advanced business intelligence solution opens up an incredibly effective tool for marketers. Corporations are recognizing that reaching millions of customers with generic broadcast messaging is not nearly as effective as one-to-one narrowcasting. Companies can now reach individuals with personalized information and relevant product offerings via web, wireless and voice, creating a strong value for the individual. Customers truly can be reached where they are with information and offers they want -- the bar for customer expectations will be raised and companies must be ready.

Three Key Bullet Points:

- Companies can better understand their customers
- Customers can be reached personally & inexpensively
- Wireless devices will make marketing ubiquitous

Biography: Michael J. Saylor co-founded MicroStrategy Incorporated ten years ago at the age of 24, eighteen months after graduating from MIT. He has helped grow MicroStrategy into a multi-billion dollar public company.

Saylor's vision is central to the current success of MicroStrategy. He is actively involved in all major strategic decisions, and is instrumental in shaping both the marketing and technology direction of the company.

Saylor is consistently recognized as one of the technology industry's leading visionaries as he pioneers the way for Intelligent E-Business. He has been featured by many of today's leading business and computer publications including Fortune, Forbes, The Washington Post, USA Today, Investor's Business Daily, PC Week and Information Week, in addition to being profiled on the NBC Nightly News, CNBC's Squawk Box, and CBS's 60 Minutes.

Time and location: Tuesday, 8:30am-9:15am [Commonwealth Hall]

Keynotes

Decision Support in the Booming E-World

Dr. James H. Goodnight
President and CEO, SAS

Abstract: According to Forrester Research, global e-commerce will be worth nearly \$7 trillion by the year 2004. That constitutes more than a 14,000% increase from the 1999 total of approximately \$1 billion. With this explosion, the need for business intelligence solutions will increase proportionally.

In his remarks, Goodnight will outline the strategy for helping customers succeed at e-business. He will focus on making effective use of the data generated by Internet activities, while moving more of the customers' business-to-business services to the Web.

Biography: Dr. James H. Goodnight is president, CEO and co-founder of SAS Institute, the largest privately held software company in the world and the leader in data warehousing and decision support. Goodnight has authored many of the procedures that comprise SAS® software, and he remains closely involved in the company's day-to-day development activities.

A native of Wilmington, N.C., Goodnight holds bachelor's and master's degrees as well as a doctorate in statistics from North Carolina State University. He served on the faculty of NCSU from 1972 to 1976, and continues to serve as an adjunct professor. Goodnight is a Fellow of the American Statistical Association, and has authored numerous papers on statistical computing.

Time and location: Tuesday, 9:15am-10am [Commonwealth Hall]

Informed Knowledge Discovery: Using Prior Knowledge in Discovery Programs

Bruce Buchanan
Professor of Computer Science and Professor of
Philosophy, Medicine, and Intelligent Systems
University of Pittsburgh

Abstract: Informed knowledge discovery uses background information about a domain to guide a discovery program toward finding interesting and novel relationships in a database. Background knowledge may be of several forms including relationships already found, semantic categories, causal preconditions, and taxonomic relationships. Recent work on discovery in science will illustrate these concepts but we will also argue for the domain-independence of the heuristics used.

Biography: Bruce Buchanan has worked in artificial intelligence since joining the Dendral project at Stanford in 1966. He was one of the principals in the development of the Dendral, Meta-Dendral, Mycin, and Protean programs at Stanford, and has continued working on symbolic learning and data mining since joining the faculty of the University of Pittsburgh in 1988, where he is now University Professor of Computer Science and Professor of Philosophy, Medicine, and Intelligent Systems. He is a member of the National Academy of Science Institute of Medicine and is currently President of the AAAI.

Time and location: Wednesday, 8:30am-9:30am [Commonwealth Hall]

Invited Talks

E-metrics: Tomorrow's Business Metrics Today

Matt Cutler

Co-founder & Chief E-Business Intelligence Officer,
NetGenesis

Abstract: Brick and mortar companies use well-understood, well-defined metrics to describe business performance and understand success. While e-business shares many principles with the offline world, Web sites require a fundamentally new approach and perspective to business analytics. Today, while many e-businesses are struggling with the concept of hits versus page views for measuring success, many forward thinking sites have begun to adopt new, Web-centric business metrics. These sites are now drilling deeper into customer interests and segments to track individual behaviors and clickstream patterns for more effective targeted marketing campaigns. With this increased customer knowledge, e-businesses are able to improve customer retention, build a more loyal customer base, and increase ROI. Attend this lively session to discuss topics like:

- The key differences between 'stickiness' and 'slipperiness'
- Specific action items based on your sites 'Personalization Index'
- How user retention can be improved through your 'Freshness Factor'
- And more!

Biography: Matt Cutler co-founded NetGenesis in January of 1994 and serves as Chief E-Business Intelligence Officer. He is responsible for leading NetGenesis' marketplace education and standards development efforts. A frequent contributing writer and speaker at major Internet industry tradeshow and conferences, Mr. Cutler's commentary has appeared in the Wall Street Journal, CNN, USA Today, Investors Business Daily, National Public Radio, and numerous other media outlets. From June 1995 to December 1997, he also served as Chairman of the Webmasters' Guild (now a part of the Association for Internet Professionals), the world's first professional association of Webmasters. Mr. Cutler holds a B.S. in Mechanical Engineering with honors from the Massachusetts Institute of Technology.

Time and location: Monday, 10:30am-11:15am [Amphitheater]

After the Gold Rush: Data Mining in the New Economy

David Stodder

Editorial Director, Intelligent Enterprise

Abstract: What is the market telling us about opportunities for data mining and knowledge discovery tools and products? Several years ago, data mining was sizzling: more recently, however, the market has forced consolidation and repackaging of tools. But is that the end of the story--or just the end of one generation and the beginning of another? Today, few large businesses would dispute that some form of data mining is essential to their efforts to understand customers and uncover knowledge to help them improve business processes. E-business models and strategies are pushing the demand for sophisticated data analysis and mining capabilities down into the midmarket. Service provider and hosting options are opening up new possibilities as well.

This talk will explore some of the market forces at work that are effecting the data mining and knowledge discovery industry. In particular, the talk will discuss customer relationship management and personalization--and how these key trends are going to affect far more applications than they do today. The talk will also offer industry perspectives on how some of the major industry players in the database, business intelligence, and corporate applications areas are integrating data mining and knowledge discovery into their solutions. Finally, the talk will consider trends in mining text and unstructured data: a huge opportunity given the kind of information resources that are important in the new economy.

Biography: David Stodder is editorial director of Intelligent Enterprise magazine (www.intelligententerprise.com), published by CMP Media. Stodder was the founding editor-in-chief of Intelligent Enterprise. Previously, he was chief editor of Database Programming & Design and editorial director of the Database Summit Series conferences. Stodder is also editorial director of Intelligent Enterprise's custom and ancillary publications and Web-based communities. He is a frequent contributor to industry conferences and events.

Time and location: Monday, 11:15am-12pm [Amphitheater]

Best Paper Talks

Hancock: A Language for Extracting Signatures from Data Streams

(Best Research Paper)

Corinna Cortes
Kathleen Fisher
Daryl Pregibon
Anne Rogers
(AT&T Labs-Research)
Frederick Smith
(Cornell University)

Abstract: Massive transaction streams present a number of opportunities for data mining techniques. Such transactions can represent calls on a telephone network, commercial credit card purchases, stock market trades, or HTTP requests to a web server. While historically such data have been collected for billing or security purposes, they are now being used to discover how customers or their intermediaries (called {transactors}) use the underlying services. For several years, we have been computing evolving profiles (called {signatures}) of the transactors in large data streams. The signature for each transactor captures the salient features of his transactions through time. Programs for processing signatures must be highly optimized because of the size of the data stream (several gigabytes per day) and the number of signatures to maintain (hundreds of millions). The original C programs to compute signatures often sacrificed readability for performance. Consequently, they were hard to verify and difficult to maintain. Hancock is a domain-specific language designed and implemented to express computationally efficient signature programs cleanly. In this paper, we describe the obstacles to computing signatures from massive transaction streams and explain how Hancock addresses these problems. For expository purposes, we present Hancock using a running example from the telecommunications industry; however, the language itself is general and applies equally well to other domains.

Time and location: Monday, 9:30am-10am [Amphitheater]

Mining IC Test Data to Optimize VLSI Testing

(Best Application Paper)

Tony Fountain
(San Diego SuperComputing Center, UCSD)
Thomas Dietterich
(Oregon State University)
Bill Sudyka
(Hewlett Packard Company)

Abstract: We describe an application of data mining and decision analysis to the problem of die-level functional test in integrated circuit manufacturing. Integrated circuits are fabricated on large wafers that can hold hundreds of individual chips (“die”). In current practice, large and expensive machines test each of these die to check that they are functioning properly (die-level functional test; DLFT), and then the wafers are cut up, and the good die are assembled into packages and connected to the package pins. Finally, the resulting packages are tested to ensure that the final product is functioning correctly. The purpose of die-level functional test is to avoid the expense of packaging bad die and to provide rapid feedback to the fabrication process by detecting die failures. The challenge for a decision-theoretic approach is to reduce the amount of DLFT (and the associated costs) while still providing process feedback. We describe a decision-theoretic approach to DLFT in which historical test data is mined to create a probabilistic model of patterns of die failure. This model is combined with greedy value-of-information computations to decide in real time which die to test next and when to stop testing. We report the results of several experiments that demonstrate the ability of this procedure to make good testing decisions, good stopping decisions, and to detect anomalous die. Based on experiments with historical test data from Hewlett Packard Company, the resulting system has the potential to improve profits on mature IC products.

Time and location: Tuesday, 9:30am-10am [Amphitheater]

Tutorials

Weaving the Semantic Web: Data Mining for Hypertext

Soumen Chakrabarti

Abstract: With over 800 million pages covering most areas of human endeavor, the World-wide Web is a fertile ground for data mining research to make a difference to the effectiveness of information search. Today, Web surfers access the Web through two dominant interfaces: clicking on hyperlinks and searching via keyword queries. This process is often tentative and unsatisfactory. Better support is needed for expressing one's information need and dealing with a search result in more structured ways than available now. Data mining and machine learning have significant roles to play towards this end.

In this tutorial we will survey recent advances in learning and mining problems related to hypertext in general and the Web in particular. We will review the continuum of supervised to semi-supervised to unsupervised learning problems, highlight the specific challenges which distinguish data mining in the hypertext domain from data mining in the context of data warehouses, and summarize the key areas of recent and ongoing research.

Biography: Soumen Chakrabarti received his B.Tech in Computer Science from the Indian Institute of Technology, Kharagpur, in 1991 and his M.S. and Ph.D. in Computer Science from the University of California, Berkeley in 1992 and 1996. At Berkeley he worked on compilers and runtime systems for running scalable parallel scientific software on message passing multiprocessors. He was a Research Staff Member at IBM Almaden Research Center between 1996 and 1999. He is currently an Assistant Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology, Bombay. His current research interests include hypertext information retrieval, web analysis and data mining. He designed the Focused Crawler and part of the Clever search engine, filing several patents in the process. His work on focused crawling got the Best Paper award at the 8th International World Wide Web Conference in 1999. He has served on the program committees for KDD 1998, KDD 1999, and WWW 2000.

Time and location: Monday, 9:00am-12pm [Waterfront 2]

Visualizing High Dimensional Datasets and Relations

Alfred Inselberg

Abstract: Intellectual curiosity and the abundance of important multivariate problems, motivate the quest for multidimensional visualization techniques to augment our 3-D perception and experience. Starting from the early successes of data visualization, like Dr. Snow's dot map in 1854 showing the connection of a cholera epidemic to a water pump, a short review of the development is given. It leads to guidelines for desirable and attainable properties in such methodologies.

With the emphasis being on the visualization of high dimensional data we focus on Parallel Coordinates; a leading multidimensional/multivariate visualization methodology for this field. The mathematical foundations for the display and discovery of multidimensional relations without loss of information are presented interlaced with a variety of applications. Several multivariate real datasets (i.e. financial, manufacturing, process control, trading etc.) will be displayed and explored interactively showing how some unsuspected complex relations were discovered from visual cues suggested by the picture. The derivation of algorithms can also be motivated from this visualization and is illustrated with examples from Computer Vision and Collision Avoidance for Air Traffic Control.

Then geometrical algorithms for Automatic Knowledge Discovery are derived and applied to real datasets. These algorithms have low computational complexity, provide explicit and comprehensible rules, do dimensionality selection by finding ONLY the parameters containing relevant information, and order these parameters according to some optimality criteria. Finally, the power to model and display complex nonlinear relations is illustrated by obtaining, from data, a model of a real country's economy and interactively discovering plausible economic policies, interrelationships, competition for the same resources, impact of constraints downstream, sensitivities as well as do trade-off analysis for Decision Support.

Biography: In ancient times AI received a Ph.D. in Applied Math and Physics from the Univ. of Illinois (Champaign-Urbana) and then held academic positions in the USA (Univ. of Ill., UCLA, USC) and abroad. In IBM, where he did research for several years, he became Senior Technical Corporate Staff Member (a sought after appellation of dubious value). Subsequently, in 1995 he was elected Senior Fellow in Visualization at the San Diego SuperComputing Center. He has his own company Multidimensional Graphs Ltd and now teaches at Tel Aviv University. AI invented (1959) and contributed to the development of Parallel Coordinates, has several patents, over 70 refereed technical papers, numerous professional and academic awards, and is now writing a book on Multidimensional Visualization... and Hi-Tech entertainment.

Time and location: Monday, 9:00am-12pm [Amphitheater]

Tutorials

Knowledge Discovery in Biological Domains

I. Jurisica

(Assistant Professor, Faculty of Information Studies,
University of Toronto & Visiting Scientist, IBM Toronto)

I. Rigoutsos

(Manager, Bioinformatics and Pattern Discovery Group,
Computational Biology Center, IBM T. J. Watson)

A. Floratos

(Research Staff Member, IBM T.J. Watson Research &
adjunct Professor of Computer Science, Courant Institute
of Mathematical Sciences, New York University)

Abstract: Biological research is generating data at an explosive rate. The Human Genome Project is expected to identify the codes for over 3 billion bases by the year 2003. This will provide code for about 100,000 proteins. Analyzing this volume of data and using it intelligently is a challenge because of its complexity, its multiple interdependent factors, the uncertainty of these dependencies, and the continuous evolution of our understanding of the data. In general, reasoning with biomedical information requires flexible knowledge representation structures and powerful knowledge-discovery tools.

This tutorial provides an introduction to the latest computational techniques for data mining and knowledge discovery in biological domains. We will explore the fit of the traditional data-mining techniques for alphanumeric, visual and relational data to biology. After characterizing biological problems, basic definitions and diverse algorithms will be presented. This will include scientific discovery, pattern identification, organization, summarization and description, clustering, classifying, associating and predicting, and information extraction. An overview of current state-of-the-art commercial and academic systems will be covered, with the emphasis on successful examples of data mining and knowledge discovery in biology. The examples will include amino acid sequence analysis, homology detection, elucidation of biological function, protein structure prediction and identification of related proteins, systematic generation of bio-dictionaries(TM) and their exploitation, analysis of biological effects, model generation and use, DNA microarrays analysis, data curation, hypothesis generation and testing. We will identify limitations of generic approaches, define problems and issues that must be addressed to successfully mine biological sequence and structure databases. We will close by discussing future directions of knowledge discovery in biology, and its relevance of knowledge visualization, knowledge evolution and management of scientific knowledge.

Time and location: Monday, 9:00am-12pm [Waterfront 3]

Data Mining for Successful Customer Management

Gregory Piatetsky-Shapiro

(Vice President and Chief Scientist, Analytics, Xchange)

Steve Gallant

(Director, Analytics Services, Xchange)

Dorian Pyle

(Director, Training & Methodology, Xchange)

Abstract: The promise of data mining in business environments is enormous. Until recently capitalizing on that promise in a real-world business environment has sometimes been very difficult. The promise is still as bright as ever and the recent past has taught practitioners of data mining for CRM (customer relationship management) much about delivering high-return, practical results.

Data mining is not a universal panacea for CRM success. Critical criteria include tools selection, business objective matching, data discovery, preparation & delivery. Successful data mining in a CRM environment is far more than the application of algorithms to data.

Customer Relationship Management is a broad approach to doing business. It is holistic in that it encompasses all aspects and functions of a company, focusing on managing the relationship between customer and company just as much between company and customer. CRM requires a two-way street – and exchange of information just as much as of goods and services. This tutorial selects five crucial CRM strategic business areas. Although part of a whole, each area is examined separately in the tutorial to enable a clear view of the problems to be met, the business problem to be solved, and the methods for delivering value. The areas chosen for scrutiny are: data preparation, customer segmentation, attrition, cross sell and e-commerce.

The tutorial assumes mastery all of the essential data mining concepts, practices and procedures. It focuses particularly on leveraging these basic skills to get the most out of them, and extend the miner's skill set within CRM.

Time and location: Monday, 1:00pm-4pm [Amphitheater]

Tutorials

Time Series Similarity Measures

Gautam Das

(Data Mining and Exploration, Microsoft Research)

Dimitrios Gunopulos

(Assistant Professor, Department of Computer Science and Engineering, University of California, Riverside)

Abstract: Time series data arise in a variety of domains, such as stock market analysis, environmental data, telecommunications data, medical and financial data. Typically each time series describes the evolution of an object as a function of time at a given data collection station. Examples are, the daily price fluctuations of a stock, or web data that count the number of clicks at different sites. Higher dimensional time series can be used to describe the evolution of more complex objects, for example digital image sequences.

Currently time series data account for a large fraction of the data stored in commercial databases. Recently there is increasing recognition of this fact, and support for time series as a new data type in commercial databases management systems is increasing. IBM DB2 for example implements support for time series using data-blades.

A fundamental problem of interest is to determine whether two given time series display similar behavior. The problem is interesting (and difficult) because the similarity measures should allow for imprecise matches. There are several applications of such measures. For example, they can be used to cluster the different time series into similar groups, or to classify a time series based on a set of known examples.

Another problem of interest is the indexing problem: given a set of time series Q , prepare an index offline such that given a query series q , the time series in Q that are most similar to q can be reported quickly. As an application, an investor may wish to know the stocks that behave similarly to a certain query stock.

In the database and data mining communities, various similarity measures and indexing techniques for time series have been proposed. In this tutorial we describe the state-of-art of this area by comparing and summarizing several of these techniques in detail.

Time and location: Monday, 1:00pm-4pm [Waterfront 3]

High Performance Data Mining

Vipin Kumar

Mohammed Zaki

Abstract: A fundamental problem in data mining is to develop algorithms and systems which scale with increase in the amount of data, and with increase in the data dimensions and complexity. Due to the huge size of data and amount of computation involved in mining algorithms, parallel and distributed processing is often considered an essential component for a successful data mining solution.

The goal of this tutorial is to provide researchers, practitioners, and advanced students with an introduction to high performance data mining. The focus will be on algorithms, software tools, and system architectures appropriate for mining massive data sets using techniques from scalable, parallel and distributed computing.

The tutorial will provide 1) an overview of fundamental parallel and distributed data mining algorithms covering common techniques like classification, associations, sequences, clustering, etc.; 2) an introduction to some of the basic architectural frameworks for high performance data mining systems; and 3) an understanding of some of the outstanding algorithmic and systems issues while mining large data sets. With this knowledge, the audience should be better prepared to mine larger data sets in practice or undertake research in this area.

Biographies: Vipin Kumar is a Professor of Computer Science at the University of Minnesota. His current research focuses on parallel computing and data mining. His past research has produced highly efficient algorithms and softwares such as Metis, hMetis, and PSPASES. He has authored over 100 research articles, and coedited or coauthored 5 books including the widely used text book "Introduction to Parallel Computing". Kumar serves on the editorial boards of several prominent journals in parallel computing. He is a Fellow of IEEE and the Minnesota Supercomputer Institute, and is a member of SIAM and ACM.

Mohammed J. Zaki is an Assistant Professor of Computer Science at Rensselaer Polytechnic Institute. His research interests include the design of efficient, scalable, and parallel algorithms and systems for various data mining tasks. He has published over 40 papers in this area, and he recently co-edited the book, "Large-scale Parallel Data Mining," LNAI State-of-the-Art-Survey, Vol. 1759, Springer-Verlag, 2000. He was co-chair for ACM SIGKDD workshop on Large-scale Parallel KDD Systems (1999), and is a co-chair for IEEE IPDPS Workshop on High Performance Data Mining (2000). He is a member of ACM and IEEE.

Time and location: Monday, 1:00pm-4pm [Waterfront 2]

Panels

Personalization and Data Mining: Exploring the Synergies

Chair:

Alex Tuzhilin (NYU)

Participants:

Paul Hagan

(Senior Analyst, Forrester Research)

Ron Kohavi

(Director of Data Mining, Blue Martini)

Bonnie J. Lowell

(President and CEO, YOUpowered)

John Riedl

(University of Minnesota and Co-Founder of
Net Perceptions)

Abstract: Personalization has recently become a “hot” area. Personalization companies focus on building customer profiles, providing recommendations to the customers, and delivering personalized Web content. However, since this is a relatively new area, it is still unclear to many people what personalization really means. This panel, consisting of experts in the area of personalization and data mining will provide insights into this question, explore possible connections between personalization and data mining, and examine how data mining can contribute to personalization and vice versa.

Time and location: Tuesday, 3:30pm-4:45pm [Commonwealth Hall]

Privacy

Chair:

Jonathan Smith (U. Penn.)

Participants:

Dan Jaye

(Engage Technologies)

Lyle Ungar

(University of Pennsylvania)

Jane Swift

(Lt. Governor of Massachusetts)

Rakesh Agrawal

(IBM Almaden)

Time and location: Tuesday, 4:15pm-5:30pm [Amphitheater]

KDD Process Standards Panel

Chair:

Ismail Parsa (Epsilon)

Abstract: Adoption of universally acceptable and sustainable “standards” -- in terms of terminology, evaluation and methodology -- is key to KDD's continuing success. Today, the process of knowledge discovery and data mining is far from automated (i.e., still requires human interaction) and, therefore, is difficult to deploy effectively. Many commercial tools lack functionality and scalability that customers require. There are institutions and on-going efforts working independently toward streamlining the whole or part of the KDD and/or related processes. Among them are CPEX, CRISP-DM, DMG, MDC, OLEDB DM and TASF.

This panel bring together representatives from these institutions and on-going efforts with the objective of informing the KDD community on the current state of their work and, possibly opening the doors for mutual collaboration. Processes (in alphabetical order of occurrence) and their participants:

- **CPEX** (Customer Profile Exchange) offers a vendor-neutral, XML-based open standard for facilitating the privacy-enabled interchange of customer information across disparate enterprise applications and systems. [www.cpex.org] **Participant:** Brad Husick.

- **CRISP-DM** is an industry consortium developing an industry-neutral and tool-neutral Cross-Industry Standard Process Model for Data Mining. [www.crisp-dm.org] **Participant:** Rudieger Wirth.

- **DMG**, the Data Mining Group, is a consortium of industry and academics formed to create standards, starting with PMML, (XML-based) for defining and sharing predictive models. [www.dmg.org] **Participant:** Robert Grossman.

- **OLEDB DM** (OLE DB for Data Mining), a Microsoft effort extending SQL databases through a new API to better support data mining operations. [www.microsoft.com/data/oledb/dm.htm] **Participant:** Usama Fayyad.

- **TASF** (The Analytic Solutions Forum), an industry consortium whose mission is to establish solution-oriented performance criteria and interoperability requirements within and between classes of decision support tools (including but not limited to OLAP or On-line Analytical Processing, data mining, data visualization, text processing, and decision analysis). [www.tasf.org] **Participant:** Erik Thomsen.

Time and location: Tuesday, 4:45pm-6pm [Commonwealth Hall]

Local Information

Boston, Massachusetts

KDD-2000 will take place in Boston at the World Trade Center. The World Trade Center is located on Boston's historic waterfront.

Cultural. Diverse. Distinctive. Dynamic. Aesthetic. Exciting. And of course, historic. This is Boston, where old charm meets contemporary style and sophistication. Where you can stroll along the cobblestone streets of Beacon Hill, and dance the night away at the city's hottest clubs. Where you can browse among the fashionable boutiques of Newbury Street and linger at an Italian bistro in the North End. Where you spend the day with the Boston Red Sox and the evening with the Boston Pops. Where you can immerse yourself in the academic world of Harvard and the out-of-this-world sights of Harvard Square. Where you'll find outstanding venues for music, art, theatre and more.

Boston is America's Walking City. It's premiere attraction, The Freedom Trail, is a walking tour through historic Boston, encompassing 16 of the most treasured sites in American history like Paul Revere's House, Old North Church, Old State House, Boston Massacre Site, Faneuil Hall and Bunker Hill Monument. Boston Common and the Public Garden are considered a national treasure.

Boston's many museums, concert halls, theatres and night-clubs are always abuzz with activity and excitement. From the internationally acclaimed Museum of Fine Arts, the Museum of Science and the John F Kennedy Library & Museum, to the famous Boston Symphony Orchestra and Boston Pops, to an abundant local and pre-Broadway theatre scene, Boston's cultural and entertainment options are bountiful.

Just a bridge away from Boston on the other side of the Charles River is Cambridge. Packed with a youthful vitality and international flair, Cambridge is often referred to as Boston's "Left Bank". It is the birthplace of higher education in America. Harvard College was founded in 1636 and, across town, MIT is renowned as the epicenter of the emerging cyber culture. North of Boston has the charm and the lure of the sea. West of Boston are the picturesque towns of Lexington and Concord. The famous shot that was heard 'round the world at the start of the American Revolution originated at the Old North Bridge in Concord.

Boston is the "Hub of New England" and major highways link Boston to points throughout the Northeast. In addition, Logan International Airport handles over 1,200 flights daily, with 52 carriers, serving the airport including 14 international airlines. Local transportation is provided by the subway system called the T.

The Greater Boston Area is home to these leading universities:

- Harvard University
- Massachusetts Institute of Technology
- Boston College
- Boston University
- Tufts University

Ground Transportation

The following information provided is the best available at press time. Please confirm fares when making reservations.

Taxis: Available at Logan Airport to Seaport Hotel. The approximate fare is \$15.

Subway: Shuttle buses provide free service between airline terminals and Airport 'T' Station on the MBTA Blue Line. Shuttle bus 22 serves terminals A and B. Shuttle bus 23 serves terminals C, D & E. Shuttle bus 11 is for transport between terminals, but does not stop at Airport 'T' Station. Blue Line fare is \$0.85.

City Transit System: Subways: Fare is \$0.85 regardless of distance traveled (subway token required). For general information, go to Massachusetts Bay Transport Authority.

Parking: Parking is available at the Seaport Hotel. The rate for valet parking is \$25.00 and \$21.00 for self parking.

Sponsorship

Sponsoring Organizations:

ACM SIGMOD

ACM SIGART

AAAI

In-Cooperation Partners: Interface, ASA SCS

Principle Corporate Contributors:

SAS

SGI

Corporate Contributors:

Blue Martini Software

Epsilon

MINEit

Salford Systems

SPSS/Clementine

Xchange

Contributors:

Microsoft Research

Sponsors:

DBMiner

digiMine

mohomine

SIG KDD would like to thank SAS for sponsoring the Discovery Reception, SGI for sponsoring Monday's luncheon, and Blue Martini, DBMiner Technology, digiMine, Epsilon, Magnify, Microsoft, MineIt, Mohomine, Salford Systems, SPSS, and Xchange for helping to make the conference a success through their corporate sponsorships.

Organization

SIGKDD Chair:

Won Kim, *Cyber Database Solutions*

Organizing Committee

General Chair:

Ismail Parsa, *Epsilon*

Program Co-Chairs:

Raghu Ramakrishnan, *U. of Wisconsin and QUIQ*

Sal Stolfo, *Columbia University*

Program Co-Chair (Advisory):

Daryl Pregibon, *AT&T Research*

Industrial Track Chairs:

Kenneth Church, *AT&T Research*

Mario Schkolnick, *SGI*

Awards Chair:

Heikki Mannila, *Nokia Research Center and Helsinki*

University of Technology

Student Travel Awards:

Wenke Lee, *North Carolina State University*

Conference Treasurer:

John Lin, *Epsilon*

Demos/Exhibits Chair:

Dorian Pyle, *Xchange*

Local Arrangements Chair:

Eleanor Tipa, *Epsilon*

Panels Chair:

Alexander Tuzhilin, *New York University*

Proceedings Chair:

Roberto Bayardo, *IBM Almaden Research Center*

Publicity Chair:

Paul Bradley, *Microsoft*

Registration Chair:

Amar Gupta, *MIT*

Sponsorship Chair:

Robert Grossman, *Magnify*

Tutorials Chair:

Raymond Ng, *University of British Columbia*

Workshops Chair:

Philip Chan, *Florida Institute of Technology*