Program
for the

Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

KDD-2004

Seattle, WA, USA
August 22-25, 2004

www.acm.org/sigs/sigkdd/kdd2004

Additional maps on page 27.

WESTIN SEATTLE—GRAND LEVEL

WESTIN SEATTLE—MEZZANINE LEVEL

Program Highlights

Invited Talks

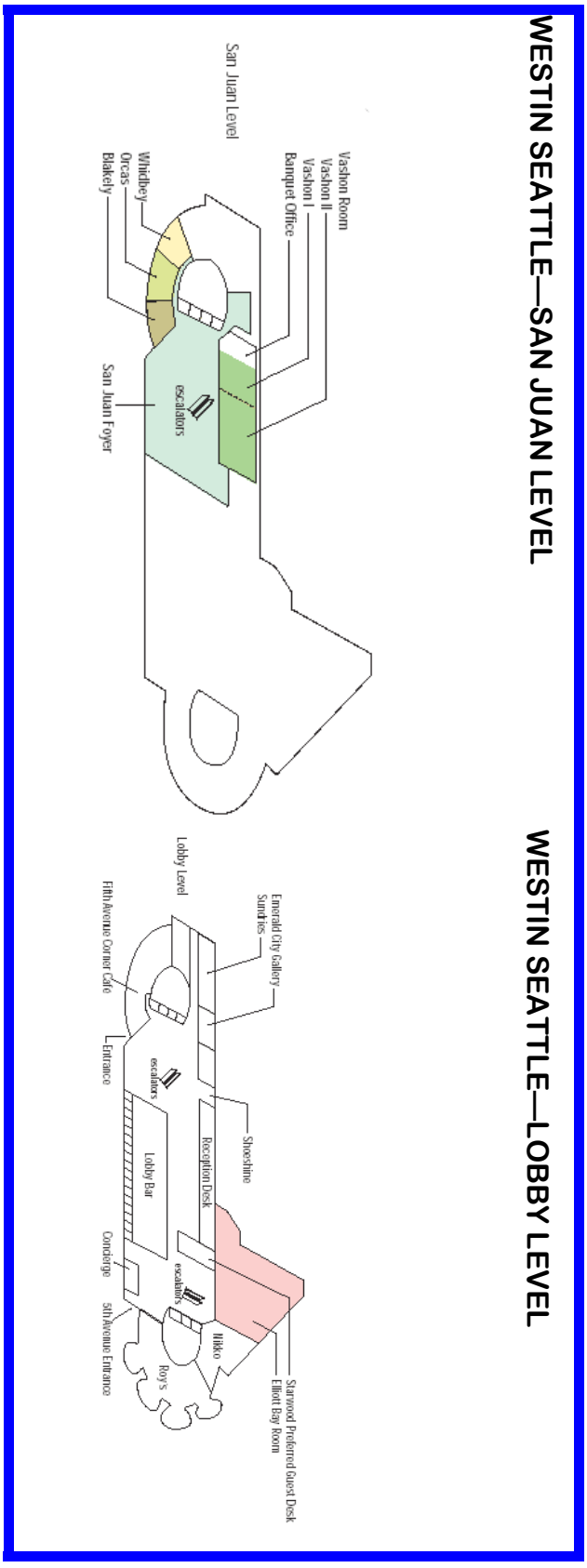
- User-Centered Design for KDD,
  Eric Haseltine, *National Security Agency*
- Graphical Models for Data Mining,
  David Heckerman, *Microsoft Research*

Research and Industrial/Government Tracks

- 40 research papers in 13 sessions
- 13 industrial/government papers in 4 sessions
- 44 research posters
- 13 industrial/government posters
- 2 panels
- 6 tutorials:
  - Online Mining Data Streams: Problems, Applications and Progress
  - Graph Structures in Data Mining
  - Data Quality and Data Cleaning: An Overview
  - Data Mining and Machine Learning in Time Series Databases
  - Junk E-mail Filtering
  - Mining Unstructured Data
- 8 workshops:
  - BIOKDD 2004: Data Mining in Bioinformatics
  - Mining Temporal and Sequential Data
  - MRDM 2004: Multi-Relational Data Mining
  - MDM/KDD 2004: Multimedia Data Mining
  - DM-SSP 2004: Data Mining Standards
  - LinkKDD 2004: "Link Discovery" Workshop
  - WebKDD 2004: Web Mining and Web Analysis
  - MSW 2004: Mining for and from the Semantic Web

Saturday, August 21

Ongoing:
Registration (Grand Foyer) **17:00-21:00**



Summarized Technical Program

Sunday

- 8 Workshops
- 6 Tutorials
- SIGKDD 2004 Opening
- Awards Ceremony
- Innovation Award Talk
- KDD Cup 2004

Monday

- Invited Talk by Eric Haseltine, *NSA*
- Research Track
  - Times Series (3 papers)
  - Multiple Objectives (3 papers)
  - Latent Models (2 papers)
  - Anomaly and Fraud Detection (2 papers)
  - Spatial Clustering (2 papers)
- Industrial/Government Track
  - Genes and Cancer (3 papers)
- Best Paper Award Talks
- Poster Highlights
- Poster Session

Tuesday

- Invited Talk by David Heckerman, *Microsoft Research*
- Research Track
  - Dimensionality Reduction (3 papers)
  - Supervised Learning (3 papers)
  - Constraints and Prior Knowledge (3 papers)
  - Analyzing Graphs (4 papers)
  - Data Streams (4 papers)
  - Frequent Sets and Association Rules (4 papers)
- Industrial/Government Track
  - Commerce (3 papers)
  - Detection (3 papers)
- Panel: Can Natural Language Processing Help Text Mining?

Wednesday

- Plenary Panel: Data Mining: Good, Bad, Or Just a Tool?
- Research Track
  - Correlation Analysis (3 papers)
  - Unsupervised Learning (3 papers)
- Industrial/Government Track
  - Visual and Image Mining (3 papers)

Sunday, August 22

Ongoing:
Registration (Grand Foyer) **7:30-20:00**

8:30-16:30
Full-Day Workshops

BIOKDD 2004: Data Mining in Bioinformatics
(Blakely—Starts at 8:45)

Mining Temporal and Sequential Data
(Vashon II—Starts at 9:00)

MRDM 2004: Multi-Relational Data Mining
(Whidbey—Starts at 9:00)

MDM/KDD 2004: Multimedia Data Mining
(Vashon—Starts at 8:40)

DM-SSP 2004: Data Mining Standards
(Baker—Starts at 9:00)

LinkKDD 2004: "Link Discovery" Workshop
(Orcas—Starts at 9:00)

WebKDD 2004: Web Mining and Web Analysis
(St. Helens—Starts at 8:30)

MSW 2004: Mining for and from the Semantic Web
(Stuart—Starts at 8:30)

9:00-12:00 (Cascade 1 A-B)
Tutorial: Online Mining Data Streams: Problems, Applications and Progress
Jian Pei, *SUNY Buffalo*
Haixun Wang, *IBM*
Philip S. Yu, *IBM*

9:00-12:00 (Cascade 1 C)
Tutorial: Data Quality and Data Cleaning: An Overview
Tamraparni Dasu, *AT&T Labs-Research*
Theodore Johnson, *AT&T Labs-Research*

9:00-12:00 (Cascade 2)
Tutorial: Graph Structures in Data Mining
Soumen Chakrabarti, *IIT Bombay*
Christos Faloutsos, *Carnegie Mellon University*

10:00-10:30 (Grand Foyer)
Coffee Break

Sunday, August 22

12:00-13:30
Lunch (on your own)

13:30-16:30 (Cascade 1 A-B)
Tutorial: Mining Unstructured Data
Ronen Feldman, *ClearForest*

13:30-16:30 (Cascade 1 C)
Tutorial: Junk E-mail Filtering
Joshua Goodman, *Microsoft Research*
Geoff Hulten, *Microsoft Research*

13:30-16:30 (Cascade 2)
Tutorial: Data Mining and Machine Learning in Time Series Databases
Eamonn Keogh, *University of California at Riverside*

15:00-15:30
Coffee Break

16:30-17:00 (Grand Foyer)
Ice Cream Break

17:00-17:20 (Grand Ballroom)
Opening Remarks and Award Presentations
Ronny Kohavi, General Chair
Johannes Gehrke, William DuMouchel, Program Chairs

17:20-18:15 (Grand Ballroom)
Innovation Award Talk by Jiawei Han

18:15-19:15 (Grand Ballroom)
KDD Cup Awards
Chairs: Rich Caruana, Thorsten Joachims

Monday, August 23

Ongoing:
Registration (Grand Foyer) **8:00-17:00**
Exhibits (Fifth Avenue Room)..... **10:00-17:00**

7:30-8:30 (Grand Foyer)
Continental Breakfast—sponsored by Fair Issac

Notes

Notes

Monday, August 23

8:30-10:00 (Grand Ballroom)
Invited Talk
Chair: John Elder

User-Centered Design for KDD
Eric Haseltine, *National Security Agency*

During initial development, KDD solutions often focus heavily on algorithms, architectures, software, hardware, and systems engineering challenges, without first thoroughly exploring how end-users will employ the new KDD technology. As a result of such "system-centered" design, many useless features are implemented that prolong development and significantly add to life cycle cost, while making the system hard to operate and use. This presentation will describe an alternate "user-centered" approach -- borrowed from the consumer products industry -- that can produce KDD solutions with shorter development cycles, lower costs, and much better usability.

10:00-10:30
Coffee Break

10:30-12:00 Research Track Session 1
(Grand Ballroom 1)
Time Series
Chair: Haixun Wang

Recovering Latent Time-Series from their Observed Sums: Networked Tomography with Particle Filters
Edoardo Airoldi, Christos Faloutsos

Mining, Indexing, and Querying Historical Spatiotemporal Data
Nikos Mamoulis, Huping Cao, George Kollios, Marios Hadjieleftheriou, Yufei Tao, David W. L. Cheung

Clustering Time Series from ARMA Models with Clipped Data
Anthony Bagnall, Gareth Janacek

10:30-12:00 Research Track Session 2
(Cascade Ballroom 1)
Multiple Objectives
Chair: Tina Eliassi-Rad

Regularized Multi-Task Learning
Theodoros Evgeniou, Massimiliano Pontil

Monday, August 23

Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions
Naren Ramakrishnan, Deept Kumar, Bud Mishra, Malcolm Potts, Richard Helm

Toward Parameter-Free Data Mining
Eamonn Keogh, Stefano Lonardi, Chotirat Ann Ratanamahatana

10:30-12:00 Industrial/Govt Track Session 1
(Cascade Ballroom 2)
Genes and Cancer
Chair: Karl Rexer

Mining Coherent Gene Clusters from Three-Dimensional Microarray Data *****
Daxin Jiang, Jian Pei, Murali Ramanathan, Chun Tang, Aidong Zhang

A Rank Sum Test Method for Informative Gene Discovery
Lin Deng, Jian Pei, Jinwen Ma, Dik Lun Lee

Predicting Prostate Cancer Recurrence via Maximizing the Concordance Index
Lian Yan, David Verbel, Olivier Saidi

12:00-13:30 (Grand Ballroom 2-3)
Lunch—sponsored by SAS

13:30-14:30 Research Track Session 3
(Grand Ballroom 1)
Latent Models
Chair: Dharmendra Modha

Web Usage Mining Based on Probabilistic Latent Semantic Analysis
Xin Jin, Yanzan Zhou, Bamshad Mobasher

Probabilistic Author-Topic Models for Information Discovery
Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, Thomas Griffiths

13:30-14:30 Research Track Session 4
(Cascade Ballroom 1)
Anomaly and Fraud Detection
Chair: Charles Elkan

Selection, Combination, and Evaluation of Effective Software Sensors for Detecting Abnormal Computer

* Runner-up, Best Application Paper

Program Committee

Wei Wang, *University of North Carolina, USA*
Geoff Webb, *Monash University, Australia*
Ed Wegman, *George Mason University, USA*
Walker White, *University of Dallas, USA*
Allan Wilks, *AT&T, USA*
Stefan Wrobel, *Fraunhofer AIS/University of Bonn, Germany*
Xindong Wu – *University of Vermont, USA*
Jun Yang, *Duke University, USA*
Philip Yu, *IBM, USA*
Osmar Zaïane, *University of Alberta, Canada*
Mohammed Zaki, *Rensselaer Polytechnic Institute, USA*

Industrial/Government Track

Dean Abbott, *Abbott Consulting, USA*
Scott Bennett, *SRA, USA*
Pavel Berkhin, *Yahoo!, USA*
Eric Bloedorn, *Mitre, USA*
Steve Donoho, *Mantas, USA*
Ashutosh Garg, *IBM Almaden Research Center, USA*
Paul Hess, *Hess Consulting, USA*
Cheryl Howard, *Elder Research, USA*
Chandrika Kamath, *Lawrence Livermore Natl. Labs, USA*
Bert Kappen, *University of Nijmegen / Promedas, Netherlands*
Brendan Kitts, *iProspect, USA*
Jacek Koronacki, *Polish Academy of Science / NuTech Solutions, Poland*
Kamal Nigam, *Intelliseek, USA*
Claudia Pearce, *NSA, USA*
Dorian Pyle, *Data Miners, USA*
Karl Rexer, *Rexer Analytics, USA*
Greg Ridgeway, *RAND, USA*
Sigal Sahar, *Intel, USA/Israel*
Joseph Sirosh, *Fair Isaac, USA*
Volker Tresp, *Siemens Research, Germany*
Jaffray Woodriff, *Biomind, USA*
Kenji Yamanishi, *NEC, Japan*

Best Paper Awards Committee

Jiawei Han, *Univ. of Illinois at Urbana-Champaign, USA*
Heikki Mannila, *University of Helsinki, Finland*
Rajeev Motwani, *Stanford University, USA*

ACM/SIGKDD Chair

Won Kim, *Cyber Database Solutions, USA*

Program Committee

George Karypis, *University of Minnesota, USA*
Daniel Keim, *University of Konstanz, Germany*
David Kempe, *University of Washington, USA*
Eamonn Keogh, *University of California at Riverside, USA*
Masaru Kitsuregawa, *University of Tokyo, Japan*
Jon Kleinberg, *Cornell University, USA*
Flip Korn, *AT&T Labs Research, USA*
Hans-Peter Kriegel, *University of Munich, Germany*
Vipin Kumar, *University of Minnesota, USA*
Laks V.S. Lakshmanan, *Univ. of British Columbia, Canada*
Diane Lambert, *Bell Labs, USA*
Jim Landwehr, *Avaya Corp, USA*
Terran Lane, *University of New Mexico, USA*
Wenke Lee, *Georgia Institute of Technology, USA*
Ying Li, *Microsoft, USA*
Bing Liu, *University of Illinois at Chicago, USA*
Jun Liu, *Harvard University, USA*
Wei-Yin Loh, *University of Wisconsin at Madison, USA*
Hongjun Lu, *Hong Kong University of Science and Technology, China*
David Madigan, *Rutgers University, USA*
Heikki Mannila, *University of Helsinki, Finland*
Brij Masand, *Data Miners Inc, USA*
Nina Mishra, *Hewlett Packard/Stanford University, USA*
Dunja Mladenic, *Jožef Stefan Institute, Slovenia*
Dharmendra Modha, *IBM, USA*
Raymond Mooney, *University of Texas at Austin, USA*
Alejandro Murua, *University of Washington, USA*
Dave Musicant, *Carleton College, USA*
Raymond Ng, *University of British Columbia, Canada*
Doug Nychka, *National Ctr. for Atmospheric Research, USA*
David Page, *University of Wisconsin at Madison, USA*
Srinivasan Parthasarathy, *Ohio State University, USA*
Jian Pei, *State University of New York at Buffalo, USA*
David Poole, *AT&T, USA*
Rajeev Rastogi, *Bell Labs, USA*
Greg Ridgeway, *RAND, USA*
Mirek Riedewald, *Cornell University, USA*
Joerg Sander, *University of Alberta, Canada*
Sunita Sarawagi, *IIT Bombay, India*
Matt Schonlau, *RAND, USA*
Thomas Seidl, *Aachen University, Germany*
Jude Shavlik, *University of Wisconsin at Madison, USA*
Kyuseok Shim, *Seoul National University, Korea*
David Skalak, *IBM, USA*
Padhraic Smyth, *University of California Irvine, USA*
Myra Spiliopoulou, *Otto-von-Guericke University Magdeburg, Germany*
Werner Stuetzle, *University of Washington, USA*
Hannu Toivonen, *University of Helsinki, Finland*
Agma Traina, *University of São Paolo, Brazil*
Alexandar Tuzhilin, *New York University, USA*
Marina Vannucci, *Texas A&M University, USA*
Chris Volinsky, *AT&T, USA*
Haixun Wang, *IBM, USA*

Monday, August 23

Usage
Jude Shavlik, Mark Shavlik

Adversarial Classification
Nilesh Dalvi, Pedro Domingos, Mausam Mausam, Sumit Sanghai, Deepak Verma

13:30-14:30 Research Track Session 5
(Cascade Ballroom 2)
Spatial Clustering
Chair: Martin Ester

Rapid Detection of Significant Spatial Clusters
Daniel Neill, Andrew Moore

Fast Mining of Spatial Collocations
Xin Zhang, Nikos Mamoulis, David W. L. Cheung, Yutao Shou

14:30-15:00 (Grand Foyer)
Coffee Break

15:00-16:00 Paper Award Talks (Grand Ballroom 3)
Best Paper Award Session
Chair: William DuMouchel

BEST RESEARCH PAPER AWARD
A Probabilistic Framework for Semi-Supervised Clustering
Sugato Basu, Mikhail Bilenko, Raymond Mooney

BEST INDUSTRIAL PAPER AWARD
Learning to Detect Malicious Executables in the Wild
Jeremy Kolter, Marcus A. Maloof

16:00-17:00 (Grand Ballroom 3)
Plenary Poster Presentations

17:00-17:20 (Grand Foyer)
Ice Cream Break

17:20-18:20 (Grand Ballroom 3)
Plenary Poster Presentations

18:30-20:30 (Grand Ballroom 1-2)
Poster Session and Reception

Monday, August 23

Poster Papers – Research Track

On Demand Classification of Data Streams
Charu Aggarwal, Jiawei Han, Jianyong Wang, Philip Yu

A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation
Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, Dharmendra Modha

An Objective Evaluation Criterion for Clustering
Arindam Banerjee, John Langford

Column-Generation Boosting Methods for Mixture of Kernels
Jinbo Bi, Tong Zhang, Kristin Bennett

IncSpan: Incremental Mining of Sequential Patterns in Large Database
Hong Cheng, Xifeng Yan, Jiawei Han

Belief State Approaches to Signaling Alarms in Surveillance Systems
Kaustav Das, Jeff Schneider

Locating Secret Messages in Images
Ian Davidson, Goutam Paul

A Microeconomic Data Mining Problem: Customer-Oriented Catalog Segmentation
Martin Ester, Rong Ge, Wen Jin, Zengjian Hu

A New Privacy Model and Association-Rule Mining Algorithm for Large-Scale Distributed Environments
Bobi Gilburd, Assaf Schuster, Ran Wolff

Discovering Additive Structure in Black Box Functions
Giles Hooker

Diagnosing Extrapolation: Tree-Based Density Estimation
Giles Hooker

SPIN: Mining Maximal Frequent Subgraphs from Graph Databases
Jun Huan, Wei Wang, Jan Prins, Jiong Yang

On Detecting Space-Time Clusters
Vijay Iyengar

Why Collective Inference Improves Rational Classification
David Jensen, Jennifer Neville, Brian Gallagher

A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts
Brian Kulis, Yuqiang Guan, Inderjit Dhillon

Clustering Moving Objects
Yifan Li, Jiawei Han, Jiong Yang

Program Committee

Research Track

Chid Apte, *IBM TJ Watson Research Center, USA*
Daniel Barbará, *George Mason University, USA*
Stephen Bay, *Stanford University, USA*
Roberto Bayardo, *IBM Almaden Research Center, USA*
Shai Ben-David, *Technion, Israel*
Michael Berthold, *University of Konstanz, Germany*
Richard Bolton, *KnowledgeBase Marketing, Inc., Canada*
Carla Brodley, *Purdue University, USA*
Andreas Buja, *University of Pennsylvania, USA*
Wray Buntine, *Ultimode Systems, USA*
Jamie Callan, *Carnegie Mellon University, USA*
Claire Cardie, *Cornell University, USA*
Soumen Chakrabarti, *IIT Bombay, India*
Chee-Yong Chan, *National Univ. of Singapore, Singapore*
Rada Chirkova, *North Carolina State University, USA*
Chris Clifton, *Purdue University, USA*
Graham Cormode, *Rutgers University, USA*
Mark Craven, *University of Wisconsin at Madison, USA*
Gautam Das, *Microsoft Research, USA*
Luc De Raedt, *Albert-Ludwigs University Freiburg, Germany*
Victor DeGruttola, *Harvard School of Public Health, USA*
Inderjit Dhillon, *University of Texas at Austin, USA*
Alin Dobra, *University of Florida, USA*
Carlotta Domeniconi, *George Mason University, USA*
Charles Elkan, *University of California at San Diego, USA*
Tina Elliassi-Rad, *Lawrence Livermore National Labs, USA*
Martin Ester, *Simon Fraser University, Canada*
Christos Faloutsos, *Carnegie Mellon University, USA*
Usama Fayyad, *DMX Group, USA*
Ronen Feldman, *Bar Ilan University, Israel*
Doug Fisher, *Vanderbilt University, USA*
Takeshi Fukuda , *IBM, Japan*
Thomas Gärtner, *Fraunhofer, Germany*
Venkatesh Ganti, *Microsoft Research, USA*
Minos Garofalakis, *Lucent Bell Labs, USA*
Ed George, *Wharton School, University of Pennsylvania, USA*
Lise Getoor, *University of Maryland, USA*
Joydeep Ghosh, *University of Texas at Austin, USA*
Aristides Gionis, *University of Helsinki, Finland*
Bart Goethals, *University of Helsinki, Finland*
Marko Grobelnik, *Jožef Stefan Institute, Slovenia*
Dimitrios Gunopulos, *University of California Riverside, USA*
Peter Haas, *IBM Almaden Research Center, USA*
Jiawei Han, *University of Illinois at Urbana-Champaign, USA*
David Heckerman, *Microsoft Research, USA*
Alexander Hinneburg, *Martin-Luther University Halle/ Wittenberg, Germany*
Howard Ho, *IBM Almaden Research Center, USA*
Piotr Indyk, *Massachusetts Institute of Technology, USA*
David Jensen, *University of Massachusetts, USA*
Chris Jermaine, *University of Florida, USA*
Thorsten Joachims, *Cornell University, USA*
Ted Johnson, *AT&T Labs Research, USA*
Hillol Kargupta, *Univ. of Maryland Baltimore County, USA*
Alan Karr, *NISS, USA*

Organizing Committee

General Chair: Ronny Kohavi, *Amazon.com, USA*

Program Chairs: Johannes Gehrke, *Cornell University, USA* and William DuMouchel, *AT&T Labs Research, USA*

Industrial/Government Track Chairs: John Elder, *Elder Research, USA* and Bharat Rao, *Siemens Medical, USA*

Best Paper Awards Chair: Surajit Chaudhuri, *Microsoft Research, USA*

Panels Chair: Raghu Ramakrishnan, *University of Wisconsin-Madison, USA*

Tutorials Chair: Mihael Ankerst, *Boeing, USA*

Workshops Chair: Myra Spiliopoulou, *Otto-von-Guericke University Magdeberg, Germany*

Student Awards Chair: David Madigan, *Rutgers University, USA*

Proceedings Chair: Joydeep Ghosh, *University of Texas at Austin, USA*

KDD-Cup Chairs:Rich Caruana, *Cornell University, USA* and Thorsten Joachims, *Cornell University, USA*

Publicity Chair: Gabor Melli, *PredictionWorks, USA*

Webmaster: Gabor Melli, *PredictionWorks, USA*

Local Publicity Chair: Zhaohui Tang, *Microsoft Research, USA*

Treasurer: Rajesh Parekh, *Blue Martini Software, USA*

Local Arrangements Chair: Ying Li, *Microsoft, USA*

Sponsorship Chairs: Kamal Ali, *Yahoo!, USA* and Tom Breur, *ING Card, USA*

Exhibits Chair: Llew Mason, *Blue Martini Software, USA*

Registration Chair: Marina Meila, *University of Washington, USA*

Monday, August 23

A Framework for Ontology-Driven Subspace Clustering
Jinze Liu, Wei Wang, Jiong Yang

The IOC algorithm: Efficient Many-Class Non-parametric Classification for High-Dimensional Data
Ting Liu, Ke Yang, Andrew Moore

When Do Data Mining Results Violate Privacy?
Murat Kantarcioglu, Jiashun Jin, Chris Clifton

Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization
Aleksander Kolcz, Abdur Chowdhury, Joshua Alspector

Learning Spatially Variant Dissimilarity (SVaD) Measures
Krishna Kummamuru, Raghu Krishnapuram, Rakesh Agrawal

Sleeved Co-clustering
Avraham Melkman, Eran Shaham

Semantic Representation, Search and Mining of Multimedia Content
Apostol Natsev, Milind Naphade, John R. Smith

A Quickstart in Frequent Structure Mining Can Make a Difference
Siegfried Nijssen, Joost N. Kok

Automatic Multimedia Cross-modal Correlation Discovery
Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, Pinar Duygulu

Estimating the Size of the Telephone Universe: A Bayesian Mark-Recapture Approach
David Poole

Cluster-Based Concept Invention for Statistical Relational Learning
Alexandrin Popescul, Lyle Ungar

Identifying Early Buyers from Purchase Data
Paat Rusmevichientong, Shenghuo Zhu, David Selinger

Privacy Preserving Regression Modelling via Distributed Computation
Ashish Sanil, Alan Karr, Xiaodong Lin, Jerome Reiter

Dense Itemsets
Jouni K. Seppänen, Heikki Mannila

Extending the Notion of Support
Michael Steinbach, Pang-Ning Tan, Hui Xiong, Vipin Kumar

Monday, August 23

Poster Papers – Research Track (cont.)

Ordering Patterns by Combining Opinions from Multiple Sources
Pang-Ning Tan, Rong Jin

A Generative Probabilistic Approach to Visualizing Sets of Symbolic Sequences
Peter Tino, Ata Kaban, Yi Sun

Rotation Invariant Measures for Trajectories
Michail Vlachos, Dimitrios Gunopulos, Gautam Das

Parallel Computation of High Dimensional Robust Correlation and Covariance Matrices
Alan Wagner, James Chilson, Raymond Ng, Ruben Zamar

Privacy-Preserving Bayesian Network Structure Computation on Distributed Heterogeneous Data
Rebecca Wright, Zhiqiang Yang

Mining Scale-Free Networks using Geodesic Clustering
Andrew Wu, Michael Garland, Jiawei Han

IMMC: Incremental Maximum Margin Criterion
Jun Yan, Benyu Zhang, Shuicheng Yan, Zheng Chen, Fan Weiguo, Qiang Yang, Ma Wei-Ying, Qiansheng Cheng

2PXMiner – Efficient Mining of Frequent XML Query Patterns with Repeated Siblings
Liang-Huai Yang, Mong Li Lee, Wynne Hsu

Redundancy Based Feature Selection for Microarray Data
Lei Yu, Huan Liu

A Cross-Collection Mixture Model for Comparative Text Mining
ChangXiang Zhai, Atulya Velivelli, Bei Yu

A Data Mining Approach to Modeling Relationships among Categories in Image Collection
Ruofei Zhang, Zhongfei (Mark) Zhang

A DEA Approach for Model Combination
Eric Zheng, Balaji Padamanabhan

Optimal Randomization for Privacy Preserving Data Mining
Yu Zhu, Lei Liu

Acknowledgements

GOLD Sponsors

amazon.com.

BOEING

FairIsaac.

Microsoft SQLServer

sas

SPSS

BRONZE Sponsors

MITRE

ORGANIZATIONAL Sponsors

acm SIGKDD
Association for Computing Machinery

ACM SIGMOD
ACM SIG on Management of Data

American Association of Artificial Intelligence

Wednesday, August 25

Exploiting a Support-Based Upper Bound of Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs
Hui Xiong, Shashi Shekhar, Pang-Ning Tan, Vipin Kumar

10:30-12:00 Research Track Session 13
(Cascade Ballroom 1)
Unsupervised Learning
Chair: Marko Grobelnik

Exploiting Dictionaries in Named Entity Extraction: Combining SemiMarkov Extraction Processes and Data Integration Methods
William Cohen, Sunita Sarawagi

Mining Reference Tables for Automatic Text Segmentation
Eugene Agichtein, Venkatesh Ganti

Mining and Summarizing Customer Reviews
Minqing Hu, Bing Liu

10:30-12:00 Industrial/Govt Track Session 4
(Cascade Ballroom 2)
Visual and Image Mining
Chair: Kenji Yamanishi

Visually Mining and Monitoring Massive Time Series
Jessica Lin, Jeff Lankford, Eamonn Keogh, Stefano Lonardi

V-Miner: Using Enhanced Parallel Coordinates to Mine Product Design and Test Data
Kaidi Zhao, Bing Liu, Thomas Tirpak, Andreas Schaller

Interactive Training of Advanced Classifiers for Mining Remote Sensing Image Archives
Selim Aksoy, Krzysztof Koperski, Giovanni Marchisio, Carsten Tusk

Monday, August 23

Poster Papers – Industrial/Government Track

ANN Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study
Nicolas de Abajo, Alberto B. Diez, Vanesa Lobato, Sergio R. Cuesta

Cross Channel Optimized Marketing by Reinforcement Learning
Naoki Abe, Naval Verma, Chidanand Apte, Robert Schroko

Exploring the Community Structure of Newsgroups
Christian Borgs, Jennifer Chayes, Mohammad Mahdian, Amin Saberi

Feature Selection in Scientific Applications
Erick Cantu-Paz, Shawn Newsam, Chandrika Kamath

A General Approach to Incorporate Data Quality Matrices into Data Mining Algorithms
Ian Davidson, Ashish Grover, Ashwin Satyanarayana, Giri Tayi

A System for Automated Mapping of Bill-of-Material Part Numbers
Jayant Kalagnanam, Moninder Singh, Sudhir Verma, Michael Patek, Yuk Wah Wong

Tracking Dynamics of Topic Trends Using a Finite Mixture Model
Satoshi Morinaga, Kenji Yamanishi

Mining Traffic Data from Probe-Car System for Travel Time Prediction
Takayuki Nakata, Jun-ichi Takeuchi

Document Preprocessing For Naive Bayes Classification and Clustering with Mixture of Multinomials
Dmitry Pavlov, Ramnath Balasubramanyan, Byron Dom, Shyam Kapur, Jignashu Parikh

Programming the K-means Clustering Algorithm in SQL
Carlos Ordonez

Learning a Complex Metabolomic Dataset Using Random Forests and Support Vector Machines
Young Truong, Chris Beecher, Adele Cutler, Leanna House, Xiaodong Lin, Stanley Young

1-Dimensional Splines as Building Blocks for Improving Accuracy of Risk Outcomes Models
David Vogel, Morgan Wang

Analytical View of Business Data
Adam Yeh, Jonathan Tang, Youxuan Jin

Tuesday, August 24

Ongoing:
Registration (Grand Foyer) **10:00-17:00**
Exhibits (Fifth Avenue Room)..... **10:00-17:00**

7:30-8:30 (Grand Foyer)
Continental Breakfast—sponsored by Boeing

8:30-10:00 (Grand Ballroom 2-3)
Invited Talk
Chair: Johannes Gehrke

Graphical Models for Data Mining
David Heckerman, *Microsoft Research*

I will discuss the use of graphical models for data mining. I will review key research areas including structure learning, variational methods and relational modeling, and describe applications ranging from web traffic analysis to AIDS vaccine design.

10:00-10:30 (Grand Foyer)
Coffee Break

10:30-12:00 Research Track Session 6
(Grand Ballroom 1)
Dimensionality Reduction
Chair: Chris Volinsky

GPCA: An Efficient Dimension Reduction Scheme for Image Compression and Retrieval
Jieping Ye, Ravi Janardan, Qi Li

IDR/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition
Jieping Ye, Qi Li, Hui Xiong, Park Haesun, Ravi Janardan, Vipin Kumar

Fast Galactic Morphology via Eigenimages
Brigham Anderson, Andrew Moore, Andrew Connolly, Bob Nichol

10:30-12:00 Research Track Session 7
(Cascade Ballroom 1)
Supervised Learning
Chair: Geoff Webb

A Bayesian Network Framework for Reject Inference
Andrew Smith, Charles Elkan

Wednesday, August 25

technology, while giving privacy advocates the opportunity to articulate their concerns. Three main issues will be discussed: (1) There is significant value to society in developing the science underpinning data mining, but also significant risk for misuse of the technology. The same techniques that could accurately identify malignant tumors could be used to classify individuals as potential terrorists, and the medical information that can be used to help doctors in emergency situations can also be used for invasive marketing. What should our response be? To disallow data mining altogether? To only apply it to "non-controversial" areas? To accept some risk if the need is acute or the benefits are compelling? (2) If our response is to develop data mining techniques and to apply them with care when appropriate or necessary, what checks and balances are required in order to safeguard individual rights? How can we constrain when and to what ends the technology is applied, and how the results are interpreted? What are the parallels to existing legal protections? What are the differences that make the problem of electronic privacy more challenging? (3) The Technology and Privacy Advisory Committee (TAPAC) recently issued its report. What are its main recommendations? How will, or should, it influence data mining research and practice?

Panelists:
Deirdre Mulligan, *University of California Berkeley*
David Jensen, *University of Massachusetts Amherst*
Michael J. Pazzani, *National Science Foundation*
Rakesh Agrawal, *IBM Almaden Research Center*

10:00-10:30 (Grand Foyer)
Coffee Break

10:30-12:00 Research Track Session 12
(Grand Ballroom 1)
Correlation Analysis
Chair: Aristides Gionis

Discovering Complex Matchings across Web Query Interfaces: A Correlation Mining Approach
Bin He, Kevin Chen-Chuan Chang, Jiawei Han

Fully Automatic Cross-Associations
Deepayan Chakrabarti, Spiros Papadimitriou, Dharmendra Modha, Christos Faloutsos

Tuesday, August 24

16:00-18:00 Research Track Session 11
(Cascade Ballroom 2)
Frequent Sets and Association Rules
Chair: Osmar Zaïane

The Complexity of Mining Maximal Frequent Itemsets and Maximal Frequent Patterns *
Guizhen Yang

Support Envelopes: A Technique for Exploring the Structure of Association Patterns
Michael Steinbach, Pang-Ning Tan, Vipin Kumar

On the Discovery of Significant Statistical Quantitative Rules
Hong Zhang, Balaji Padmanabhan, Alexander Tuzhilin

Approximating a Collection of Frequent Sets
Foto Afrati, Aristides Gionis, Heikki Mannila

18:00-18:45 (Grand Crescent)
KDD Transfer Meeting

19:15-22:30 Program Committee and Organizing Committee Dinner *(by invitation only)*
Buses depart from front of hotel at 19:15, or meet onsite at 19:30. Dinner (lower deck) starts at 20:15. See Ying Li for details.

* Runner-up, Best Research Paper

Wednesday, August 25

Ongoing:
Registration (Grand Foyer) **10:00-12:30**

7:30-8:30 (Grand Foyer)
Continental Breakfast—sponsored by SPSS

8:30-10:00 (Grand Ballroom 1)

Plenary Panel:	**Data Mining: Good, Bad, or Just a Tool?**
Chair:	Raghu Ramakrishnan, *University of Wisconsin, Madison*

This panel is intended to be a forum to argue for continued efforts in developing data mining as a

Tuesday, August 24

An Iterative Method for Multi-Class Cost-Sensitive Learning
Naoki Abe, Bianca Zadrozny

Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria
Richard Caruana, Alex Niculescu-Mizil

10:30-12:00 Industrial/Govt Track Session 2
(Cascade Ballroom 2)
Commerce
Chair: Cheryl Howard

Density-based Spam Detector
Kenichi Yoshida, Fuminori Adachi, Takashi Washio, Hiroshi Motoda, Teruaki Homma, Akihiro Nakashima, Hiromitsu Fujikawa, Katsuyuki Yamazaki

Predicting Customer Grocery Shopping Lists from POS Purchase Data
Chad Cumby, Andy Fano, Rayid Ghani, Marko Krema

TiVo: Making Show Recommendations Using a Distributed Collaborative Filtering Architecture
Kamal Ali, Wijnand Van Stam

12:00-14:00 (Grand Ballroom 2-3)
KDD Business Lunch—sponsored by amazon.com

14:00-15:30 (Grand Ballroom 1)

Panel:	**Can Natural Language Processing Help Text Mining?**
Chair:	Anne Kao, *Boeing Phantom Works*

Natural Language Processing (NLP) has been around for a number of decades. It has developed various techniques that are typically linguistically inspired, i.e. text is typically syntactically parsed using information from a formal grammar and a lexicon, the resulting information is then interpreted semantically and used to extract information about what was said. NLP may be deep or shallow, and even use statistical means to disambiguate word senses or multiple parses of the same sentence. It tends to focus on one document or piece of text at a time and be rather computationally expensive. It includes techniques like word stemming, multiword phrase grouping, synonym normalization, anaphora resolution, and role determination.

Text Mining is more recent, and uses techniques primarily developed in statistics and machine learning.

Its aim typically is not to understand all or even a large part of what a given speaker/writer has said, but rather to extract patterns across a large number of documents. It includes things like text classification according to some fixed set of categories, automatic text clustering, extraction of topics from texts or groups of text and the analysis of trends.

In this panel, we will discuss (1) Can traditional NLP methods help text mining? If so, can they help all areas of text mining? Or just some areas? Which NLP areas/ techniques are useful? (2) What is novel about text mining vs. NLP? In light of this, what would be some new future directions for NLP in light of requirements from text mining?

Panelists:

Jaime Carbonell, *Carnegie Mellon University*
Ken Church, *Microsoft Research*
Oren Etzioni, *University of Washington*
Nancy Lawler, *Department of Defense*
Marko Grobelnik, *Jozef Stefan Institute*
Dave Lewis, *David Lewis Consulting*
Giovanni Marchisio, *Insightful Corporation*

14:00-15:30 Industrial/Govt Track Session 3
(Cascade Ballroom 1)
Detection
Chair: Dustin Hux

Early Detection of Insider Trading in Option Markets
Steve Donoho

Eigenspace-based Anomaly Detection in Computer Systems
Tsuyoshi Ide, Hisashi Kashima

Effective Localized Regression for Damage Detection in Large Complex Mechanical Structures
Aleksandar Lazarevic, Ramdev Kanapady, Chandrika Kamath, Vipin Kumar, Kumar Tamma

14:00-15:30 Research Track Session 8
(Cascade Ballroom 2)
Constraints and Prior Knowledge
Chair: Xindong Wu

Interestingness of Frequent Itemsets Using Bayesian Networks as Background Knowledge
Szymon Jaroszewsicz, Dan Simovici

Efficient Closed Pattern Mining in the Presence of Tough Block Constraints
Krishna Gade, Jianyong Wang, George Karypis

Incorporating Prior Knowledge with Weighted Margin Support Vector Machines *
Xiaoyun Wu, Rohini Srihari

15:30-16:00 (Grand Foyer)
Coffee Break

16:00-18:00 Research Track Session 9
(Grand Ballroom 1)
Analyzing Graphs
Chair: Chris Clifton

Fast Discovery of 'Connection Subgraphs'
Christos Faloutsos, Kevin McCurley, Andrew Tomkins

Mining the Space of Graph Properties
Glen Jeh, Jennifer Widom

Scalable Mining Large Disk-Based Graph Databases
Chen Wang, Wei Wang, Jian Pei, Yongtai Zhu, Baile Shi

Cyclic Pattern Kernels for Predictive Graph Mining
Tamas Horvath, Thomas Gärtner, Stefan Wrobel

16:00-18:00 Research Track Session 10
(Cascade Ballroom 1)
Data Streams
Chair: Haixun Wang

Systematic Data Selection to Mine Concept-Drifting Data Streams
Wei Fan

Incremental Maintenance of Quotient Cube for Median
Cuiping Li, Gao Cong, Anthony K. H. Tung, Shan Wang

Machine Learning for Online Query Relaxation
Ion Muslea

A Graph-Theoretic Approach to Extract Storylines from Search Results
Ravi Kumar, Uma Mahadevan, D. Sivakumar

* Best Student Paper