

SECOND ANNOUNCEMENT AND REGISTRATION FORM

**38th Symposium on the Interface:
Computing Science and Statistics**

Theme:

Massive Data Sets and Streams

May 24-27, 2006

Westin Hotel, Pasadena, California

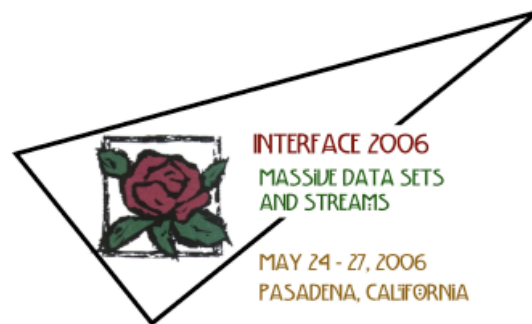
<http://www.interfacesymposia.org>

<http://galaxy.gmu.edu/Interface2006/i2006webpage.html>

**Keynote Speaker: Usama Fayyad
Yahoo, Inc.**

Title

**Grand Challenges in Data Mining:
the Technical, the Pragmatic, and the Ugly**



SPONSORED BY

The Interface Foundation of North America

HOSTED BY

The Jet Propulsion Laboratory
George Mason University

Financial Sponsors

ASA Section on Statistical Computing

ASA Section on Statistical Graphics

The Jet Propulsion Laboratory

SAS, Inc.

The National Institute of Statistical Science

U. S. National Security Agency

COOPERATING ORGANIZATIONS

ASA, CSNA, ENAR, IASC, IMS, INFORMS, SIAM and WNAR

This symposium is a long-standing forum focusing on the
Interface between computing science and statistics.

General Information

AN INVITATION

The Interface Foundation of North America cordially invites you to participate in the 38th Interface Symposium, the premier annual conference on the interface of computing and statistics. The Foundation is a non-profit educational corporation founded in 1987 to sponsor the symposium and publish the proceedings. IFNA also co-publishes the *Journal of Computational and Graphical Statistics*. For further information about IFNA visit our web site at <http://www.interfacesymposia.org>.

The theme this year is Massive Data Sets and Streams. This has been a major effort not only in the statistics community but also in the computing science community. The focus on data streams has been a major focus of the National Academy of Science's Committee on Applied and Theoretical Statistics. We are especially lucky this year to have Usama Fayyad as our keynote speaker. Usama has been one of the great innovators in the area of data mining. A number of other distinguished researchers in the area are also participating.

CONTACT INFORMATION

Conference Chairs:

Dr. Amy Braverman, Jet Propulsion Laboratory, Amy.Braverman@jpl.nasa.gov
Dr. Yasmin Said, Johns Hopkins University, ysaid99@hotmail.com, yhs@jhu.edu

Mail: Interface Foundation of North America, Inc.
P.O. Box 7460
Fairfax Station, VA 22039-7460

Phone: (703) 993-1680 Fax: (703) 993-1700

CALL FOR PARTICIPATION

In 2006 we are focusing on modern problems related to data and information overload and what we as professional data analysts can contribute to their solution. Broadly speaking, the program is organized around two complementary notions: (1) analysis of new data types brought to us by new technological capabilities, and (2) new methods and approaches enabled by modern technology. While these are both interesting in their own right, perhaps the most interesting problem of all is how to bridge the gap between them. In other words, new data types demand new techniques, and new technology enables new techniques, but are we really using the latter in service of the former? Papers and presentations from all areas of statistics, computer science and application areas relevant to modern data analysis have been invited. In this age of information overload, data analysis almost always requires analysts to cope with large data volumes. However, submissions have not been limited to massive applications, but include small-scale applications and case studies with an eye toward larger problems. Discipline scientists are often at the frontier of work with new data types and we have encouraged them share their problems and solutions with us whether or not those solutions appeal to traditional statistical or data mining methods. Sessions will include invited, special focus, refereed, and contributed paper sessions.

Invited sessions are organized by members of the Program Committee and designed to cover both the depth and breadth of current research related to the theme. See the section below on the Invited Program.

Special focus sessions are new this year: any participant can organize a special focus session on the topic of their choice by submitting a set of four or five abstracts related to the topic. The Program Chairs have reviewed the abstracts and have endeavored to keep the papers together in a devoted session.

Refereed paper sessions offer participants the opportunity to develop their research records with a relatively rapid review process. Refereed contributions are to be full, conference-style papers (not to exceed 12 pages). They have been reviewed by members of the Program Committee or others appointed by the Program Chairs. Refereed papers which are accepted as such will be collected into special sessions and speakers will be given a full 30 minutes to present their results - the same as invited speakers. Papers submitted as refereed, but not accepted as such will still be part of the contributed program.

Contributed sessions have been organized by the Program Chairs from contributed abstracts.

REGISTRATION AND FINANCIAL SUPPORT

The early registration deadline has passed. The regular member of Interface registration is \$265. For individuals who are not members of Interface, registration is \$350. Members of cooperating societies receive discounted registration. For members of co-operating societies, registration is \$300. Regular participants in the Interface Symposium Series are strongly encouraged to join Interface. There also is a student registration fee of \$150, and a one-day registration fee \$150. **In order to enroll in the short courses, participants must be registered for the conference.** The one-day registration does not include the banquet. All other registered participants attend the banquet at no additional charge. Guest tickets for the banquet may be purchased for an additional \$50 per ticket. Limited funds may still be available to support travel and per diem expenses: preference will be given to young researchers and graduate students, with a priority to those presenting papers.

HOTEL ACCOMODATIONS

The Interface 2006 conference hotel is the Westin Pasadena, located at 191 North Los Robles Avenue, telephone 626-792-2727. The hotel is newly remodeled and features wireless Internet service in the meeting rooms and public areas. Historic Old Town Pasadena is just a few blocks away, with many new restaurants, theaters and upscale shopping. The Metro Gold Line can take you to downtown Los Angeles in about 20 minutes, and bus service in the area is good. Other nearby attractions include Caltech, the Jet Propulsion Laboratory, the Huntington Library, the Norton Simon Museum, the Pasadena Museum of California Art, the Rose Bowl, and the neighborhood known as Bungalow Heaven. We have reserved a block of rooms for Interface attendees at the special rate of \$135 per night (plus tax). We expect these rooms to go quickly, so book early! The Westin has set up a special site for online reservations. <http://www.starwoodmeeting.com/StarGroupsWeb/booking/reservation?id=0509281286&key=EA041>.

Important notice for those who wish to make reservations for the nights of May 21 and 22 (NASA Data Mining Workshop participants): The Westin's online reservation system is experiencing a problem making reservations at the special Interface 2006 rate of \$135.00 per night for these two nights. If you need reservations for May 21 and/or 22, 2006, please call Janice Glosup at the Westin at 626-304-1419. Please continue to use the online reservation system for reservations that do not involve the problem nights.

SHORT COURSES

Short Course I: Random Graphs for Statistical Pattern Recognition

David Marchette (Naval Surface Warfare Center and Johns Hopkins University)

Time: 8:00 am to 12:00 noon. Location: San Marino.

Abstract: We will discuss various types of random graphs that have been suggested for pattern recognition. These include the relative neighborhood graphs of Toussaint and variations on these proposed by others; sphere graphs and digraphs; Voronoi tessellations and Delaunay triangularizations; minimum spanning trees; and nearest neighbor graphs. The use of graphs to analyze the complexity of the classification problem, provide visualization, dimensionality reduction, clustering and classification algorithms will be discussed. The various graphs and techniques will be illustrated on a wide range of interesting pattern recognition problems, as well as on simulated data sets. The material will be taken primarily from the book, *Random Graphs for Statistical Pattern Recognition* (by David Marchette, John Wiley & Sons, 2004).

Short Course II: Statistics and Information Theory

Bin Yu (UC Berkeley) and Mark Hansen (UCLA)

Time: 1:00 pm to 5:00 pm. Location: San Marino.

Abstract: Information Theory deals with a basic challenge in communication: How do we transmit information efficiently? In addressing that issue, Information Theorists have created a rich mathematical framework to describe communication processes with tools to characterize so-called fundamental limits of data compression and transmission. What might Statisticians learn from Information Theory? Basic concepts like entropy and Kullback-Leibler divergence have certainly played a role in statistics. But so too have estimation frameworks like the Maximum Entropy principle; novel decompositions like ICA; and even model selection methodologies like AIC and the Principle of Minimum Description Length. In this course we will illustrate how the basic questions and tools of Information Theory relate to statistical practice and theory.

THE INVITED PROGRAM

Computer Models in National Defense and Homeland Security

David Banks, organizer

David Banks, *Statistical Issues in Model Validation*

Kirstie Bellman, *The Challenges of Modeling Complex Embedded Systems Correctly and Usefully*

Leslie M. Moore, *Computer Experiment Designs to Achieve Multiple Objectives*

Defense and Security

Susan Paddock, organizer

C. Shane Reese, *A Hierarchical Model for the Reliability of an Anti-aircraft Missile System*

Thomas J. Sullivan, *Quantitative Tools for the Counter-IED Campaign*

Alyson Wilson, *Systems Reliability and Experiment Planning*

Geospatial and Other Massive Data Sets in Alcohol Studies

Yasmin Said, organizer

Paul Gruenewald, *Alcohol Problems: System Complexity as a Challenge to Data Systems*

Yasmin H. Said & Edward J. Wegman, *Time-of-Day, Day-of-Week, and Month-of-Year Effects for Alcohol Use*

William F. Wiecek, *Developing an Analytical Approach to Massive Data Sets on Alcohol Use*

Genomics

Chiara Sabatti, organizer

Sunduz Keles, *Mixture Modeling for Genome-wide Localization of Transcription Factors*

Jon McAuliffe, *An Infinite-state Generalized Hidden Markov Phylogeny for Multi-species Regulatory Module Detection*

Susan Service, *Dense SNP Genotyping on Chromosome 22 in 200 Persons from Each of 12 Populations*

Medicine

Emery Brown, organizer

Loren Frank, *Adaptive Analysis of Neural Plasticity in the Hippocampal System*

Matti Hamalainen, *Dynamic Imaging of Brain Function with MEG and EEG*

Nicho Hatsopoulos, *Current Developments in a Cortically Controlled Brain Machine Interface*

Uncertainty in Nature's Protein and Cellular Dance of Life

Arnold Goodman, organizer

Arnold F. Goodman, *Bioinformatics of DNA-RNA-Protein Processes: Uncertainty, a Flow Model and Collaborations*

Astronomy and Astrophysics

David van Dyk, organizer

George Djorgovski, *Virtual Astronomy, Information Technology, and the Revolution in Scientific Methodology*

Christopher Genovese, *Quasi-adaptive Confidence Bands and the Dark Energy Equation of State*

Robin Morris, *Modern Statistical Methods for GLAST Event Analysis*

Climate and Weather

Doug Nychka, organizer

Jeffrey L. Anderson, *Data Assimilation for Weather and Climate Models with Ensemble Filters*

Steve Sain, *Assessments of Climate Change using Regional Climate Models*

Doug Nychka, *Discussant*

Space and Solar Physics

Jay Johnson, organizer

Jay Johnson, *An Information-Dynamical Approach to Identify Nonlinearity in Magnetospheric Activity*

Brian Wilson, Lukas Mandrake, George Hajj, Chunming Wang, *JPL/USC GAIM: A Real-Time Global Ionospheric Data Assimilation Model*

Simon Wing, *Cross-Fertilization between Machine Learning and Space Physics*

Solid Earth Geophysics

Kristy Tiampo, organizer

Tim Ahern, *Management of Massive Seismic Data Sets: Activities within the IRIS Data Management System*

John Rundle and Andrea Donnellan, *Process, Pattern, Prediction: Using Space Data to Understand and Predict Complexity in Driven Dynamical Earth Systems*

Carola Tiede, *Some Ideas about the Use of Global Sensitivity Analyses in Geoscience Applications*

Statistical Aspects of Ranking Search Results

Daryl Pregibon, organizer

Junghoo Cho, *Controlled Randomization of Search Results: Its Implication and Benefit*

Guy Lebanon, *Conditional Models on the Ranking Poset- A Unifying Theme for Classification and Ranking*

D. Sivakumar, *Comparing and Aggregating Partial Rankings*

Network Traffic

George Michailidis , organizer

Andre Broido, *Diversity and Disparity in IP Traffic*
George Michailidis, *Flexicast Delay Network Tomography*
David Rolls, *Semi-experiment Investigations of Network Traffic*

Telecom Data Streams

Diane Lambert, organizer

David A. James, *Stream-based Data Analysis and Visualization of High-speed Wide-area Wireless Networks*
Pat Tendick, *Measurement and Analysis of Large-Scale Stochastic Service Systems*
Allan Wilks, *Waterworks for Stream Processing*

Curve Fitting for Massive, Complex Data Sets

Michael G. Schimek, organizer

Simon Wood, *Smooth Modelling of Large Datasets*
Marlene Muller, Michael G. Schimek, *Classification of High-dimensional Data by Semiparametric Generalized Regression Models*
Lijian Yang, Li Wang, *Efficient and Fast Spline-backfitted Kernel Smoothing of Additive Regression Models*

Spatio-temporal Data Mining

Zoran Obradovic, organizer

Richard A. Berk, *Using Ensemble Statistical Procedures for Imputing Homeless Counts for Los Angeles County*
Zoran Obradovic, *Integration of Deterministic and Statistical Algorithms for Retrieval and Analysis of Geophysical Parameters*
Shashi Shekhar, *What is special about mining spatial datasets?*

Streaming Data I

David Scott, organizer

Suhrid Balakrishnan, *A Streaming/Sequential Algorithm for Learning Sparse Classifiers*
Mark Hansen, *Viewing Machines: Embedded Coupled Human-observational Systems*
David Marchette, *Analysis of Streaming Text*

Streaming Data II

Bill Szewczyk, organizer

Pedro Domingos, *A General Framework for Mining Massive Data Streams*
William F. Szewczyk, *Time-evolving Adaptive Regression*
Olivier Verscheure, *Quantization for Adapted GMM-Based Speaker Verification*

Text Mining

Edward Wegman, organizer

John Rigsby, *Multi-Mode Co-clustering of Different Text Data Attributes*
Padhraic Smyth, *Text Mining Using Statistical Topic Models*
Jeffrey L. Solka, *Literature-based Discovery for the Identification of New Methods of Water Purification*

Statistics and Information Technology

Bin Yu, organizer

Akshay Adhikari, Lorraine Denby, Jim Landwehr, Jean Meloche, *Monitoring the Converged Network*
John Lafferty, *The Evolution of Science: Time Series Models of Scientific Journals and Other Large Text Databases*
Creon Levit, *Using Graphics Processing Unit (GPU) Hardware for Interactive Exploration of Large Multivariate Data*

Best of SIAM Data Mining 2006

Vipin Kumar, organizer

Indrajit Bhattacharya, Lisa Getoor, *A Latent Dirichlet Model for Unsupervised Entity Resolution*
Jerry Scripps, Pang-Ning Tan, *Clustering in the Presence of Bridge-Nodes*
Hongyan Liu, Jiawei Han, Dong Xin, Zheng Shao, *Mining Interesting Patterns from Very High Dimensional Data: A Top-Down Row Enumeration Approach*

Challenges in Modern Data Analysis

Amy Braverman, organizer

Anthony Freeman, *Earth Science Data: A Look Ahead*
William Szewczyk, *What is a Datum?*
Tom Torda, *Problems in Meta-Analyses: Studies Are Many But Cases Are Few*

Forensic Statistics

Karen Kafadar, organizer

Benjamin Bachrach, *Statistical Techniques Applied to the Surface Topography Analysis of Firearms-related Forensic Evidence*
James J. Filliben, *Statistical Methods for Assessing the Feasibility of a National Ballistics Imaging Database*
Hal Stern, *Discussant*

Data Fusion

Amy Braverman, organizer

Brian J. Smith, Mary Kathryn Cowles, *Fusing Point-Referenced Radon Data with Areal Uranium Data Arising from a Common Spatial Process*
Linda J. Young, Carol A. Gotway, *The Effects of Change of Support on Data Fusion*
Thomas Bengtsson, *Sample Based Data Fusion in Remote Sensing*

NASA Massive Data Mining Workshop, May 23-24, 2006

Data from Earth-orbiting satellites have been accumulating at a very high rate for several years now. In combination with in-situ observations and physical model output, this enormous, distributed repository holds the answers to important questions about our planet's past, present and future. However, the information is only accessible if effective analysis capabilities can be brought to bear. Data mining has the potential to provide these capabilities, and, if employed in close coordination with Earth science research, could increase the science return from NASA's vast Earth science data collection. The objectives of this second NASA Data Mining Workshop are to bring together Earth scientists and data miners to match the needs of the scientific community to existing capabilities provided by computer scientists and statisticians, and suggest future research directions they may pursue to help advance Earth science research. In particular, we seek to facilitate formation of collaborative relationships between Earth and data scientists, and identify specific problems those collaborations can address.

Date	Events
Tues. May 23	<ul style="list-style-type: none"> •Registration begins for NASA Workshop
Wed. May 24	<ul style="list-style-type: none"> •Registration begins for Interface 2006 •Short Course I: <i>Random Graphs for Statistical Pattern Recognition</i> (8:00 a.m. - noon) David Marchette, Naval Surface Warfare Center and Johns Hopkins University •Short Course II: <i>Statistics and Information Theory</i> (1:30 p.m. - 5:30 p.m.) Bin Yu, UC Berkeley and Mark Hansen, UCLA •Evening Mixer (8:00 p.m. - 10:00 p.m.)
Thurs. May 25	<ul style="list-style-type: none"> •Keynote Address (9:00 a.m. - 10:00 a.m.) •Technical sessions begin (10:30 a.m. - 5:45 p.m.) •Conference Banquet (7:00 p.m. - 10:30 p.m.)
Fri. May 26	<ul style="list-style-type: none"> •Technical sessions continue (8:00 a.m. - 5:45 p.m.)
Sat. May 27	<ul style="list-style-type: none"> •Technical sessions continue (8:00 a.m. - noon)

REGISTRATION NOTES

•For inquiries about registration, please contact Ms. Liz Quigley, (703) 993-1680, or via email, liz@galaxy.gmu.edu.

- Registration for Interface includes all Interface sessions, the banquet, and the conference CD (mailed after the conference).
- Full-time students: Please attach proof of your full-time student status - a letter from your department chair or major professor is required.
- Participants in short courses must be registered for the meeting (at least for a single day).
- On-site registration will be available.
- If you select and pay for IFNA membership then you can register at IFNA rates for Interface and the courses (i.e., membership is effective immediately).
- Cooperating Societies are: ASA, CSNA, ENAR, IASC, IMS, INFORMS, SIAM, and WNAR.

TRANSPORTATION FROM THE AIRPORT

There are five Los Angeles area airports: Los Angeles International (LAX), Ontario International, John Wayne-Orange County, Long Beach, and Burbank. LAX, Ontario and Long Beach are each approximately 35 miles from Pasadena. John Wayne is further and the closest airport is Burbank (about 15 miles). All airports are served by Super Shuttle and other shared ride van services. A taxi from anywhere but Burbank will cost between 60 and 100 dollars. Madre, a picturesque foothill community and Arcadia, where the Los Angeles County Arboretum and the Santa Anita racetrack are located.

ABOUT PASADENA

Pasadena is a city of approximately 140,000 people located 10 miles (16 kilometers) northeast of downtown Los Angeles. The city is bordered by the San Gabriel Mountains to the north, which have an extensive network of hiking and biking trails as well as mountain villages, lakes and skiing in the winter. On the other side of the mountains (approximately a one hour drive) is the Mojave Desert.

MEETING REGISTRATION FOR INTERFACE 2006

38th Meeting of the Interface, INTERFACE 2006, May 24-27, 2006

Name _____ (As will appear on your nametag)

Job Affiliation _____ (As will appear on your nametag)

(Indicate whether this is a work address or home address.)

Address _____

City _____ State or Province _____

ZIP or Postal Code _____ Country _____

Telephone _____ E-mail _____

Join Interface now (and attend at the Member rates of \$265). Pay \$35 membership plus \$265 registration.

Conference fees (with 1 Banquet ticket included, except single-day registrants)

- Interface member, \$265
- Member of Cooperating Societies, \$300
- Not a member of either Interface or Cooperating Societies, \$350 (I do not wish to join Interface.)
- Student Registration (no proceedings), \$150
- Single Day Registration (no banquet, no proceedings), \$150

Interface Short Course Registrations (At least one-day Conference Registration Required)

- Attendees who are members of Interface or Cooperating Societies, one course, \$165
- Attendees who are members of Interface or Cooperating Societies, both courses, \$225
- Attendees who are not members of Interface or Cooperating Societies, one course, \$225
- Attendees who are not members of Interface or Cooperating Societies, both courses, \$350
- Students, one course, \$75
- Students, both courses, \$100

If only one course is taken, circle appropriate one: **I: Random Graphs** **II: Statistics and Information**

Guest Banquet Tickets @ \$50 each Total Amount Enclosed _____

Make checks payable to Interface. Mail to: Interface 2004, P.O. Box 7460, Fairfax Station, VA 22039-7460, USA, or fax to: (703) 993-1700 (attention: Ms. Elizabeth Quigley). Visa, MasterCard, or Discover Credit Cards are accepted. Please supply also the following information.

Type of Card _____ Expiration Date _____

Credit Card Number _____ / _____

All card users, please note: Please include the 3-digit number that follows credit card number. This is found on the back of credit card in the signature line.

Cardholder's Name _____ Signature _____

Credit Card Billing Address & Email address: (If different from mailing address or attendee information. If Cardholder is not the attendee, please furnish email address for confirming receipt.

(Please Print.)



George Mason University, MS 4A7
Center for Computational Statistics
158 Science-Technology II Building
4400 University Drive
Fairfax, VA 22030-4444 USA

Non-Profit Organization
U.S. Postage
PAID
Fairfax, VA
Permit 1532

Interface 2006