# MMDS 2010:

## Workshop on Algorithms for Modern Massive Data Sets

Cubberley Auditorium
Stanford University

**June 15–18, 2010**

The 2010 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2010) will address algorithmic and statistical challenges in modern large-scale data analysis. The goals of MMDS 2010 are to explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured scientific and internet data sets; and to bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote the cross-fertilization of ideas.

**Organizers:** *Michael Mahoney, Alex Shkolnik, Petros Drineas, Lek-Heng Lim, Gunnar Carlsson*

# Workshop Schedule

## Tuesday, June 15, 2010: Large-scale Data and Large-scale Computation

| Time | Event | Location/Page |
|---|---|---|
| **Registration & Opening** | | **Cubberley Auditorium** |
| 8:00–9:45am | *Breakfast and registration* | |
| 9:45–10:00am | Organizers<br>*Welcome and opening remarks* | |
| **First Session** | | **Cubberley Auditorium** |
| 10:00–11:00am | Peter Norvig, Google Research<br>*Internet-Scale Data Analysis (Tutorial)* | pp. 11 |
| 11:00–11:30am | Ashok Srivastava, NASA<br>*Virtual Sensors and Large-Scale Gaussian Processes* | pp. 12 |
| 11:30–12:00n | John Langford, Yahoo! Research<br>*A method for Parallel Online Learning* | pp. 9 |
| 12:00–2:00pm | *Lunch* (on your own) | |
| **Second Session** | | **Cubberley Auditorium** |
| 2:00–3:00pm | John Gilbert, University of California, Santa Barbara<br>*Combinatorial Scientific Computing: Experience and Challenges (Tutorial)* | pp. 8 |
| 3:00–3:30pm | Deepak Agarwal, Yahoo! Research<br>*Estimating Rates of Rare Events through Multiple Hierarchies* | pp. 6 |
| 3:30–4:00pm | James Demmel, University of California, Berkeley<br>*Minimizing Communication in Linear Algebra* | pp. 7 |
| 4:00–4:30pm | *Coffee break* | |
| **Third Session** | | **Cubberley Auditorium** |
| 4:30–5:00pm | Dmitri Kriukov, CAIDA<br>*Hyperbolic Mapping of Complex Networks* | pp. 9 |
| 5:00–5:30pm | Mehryar Mohri, New York University<br>*Matrix Approximation for Large-scale Learning* | pp. 11 |
| 5:30–6:00pm | David Bader, Georgia Tech College of Computing<br>*Massive Scale Analytics of Streaming Social Networks* | pp. 6 |
| 6:00–6:30pm | Ely Porat, Bar-Ilan University<br>*Fast Pseudo-Random Fingerprints* | pp. 11 |
| **Evening Reception** | | |
| 6:30–9:30pm | *Welcome Dinner Reception* | **New Guinea Garden** |

## Wednesday, June 16, 2010: Networked Data and Algorithmic Tools

| Time | Event | Location/Page |
|---|---|---|
| **First Session** | | **Cubberley Auditorium** |
| 9:00–10:00am | Peter Bickel, University of California, Berkeley<br>*Statistical Inference for Networks (Tutorial)* | pp. 6 |
| 10:00–10:30am | Jure Leskovec, Stanford University<br>*Inferring Networks of Diffusion and Influence* | pp. 9 |
| 10:30–11:00am | *Coffee break* | |
| **Second Session** | | **Cubberley Auditorium** |
| 11:00–11:30am | Michael W. Mahoney, Stanford University<br>*Geometric Network Analysis Tools* | pp. 10 |
| 11:30–12:00n | Edward Chang, Google Research<br>*AdHeat - A New Influence-based Social Ads Model and its Tera-Scale Algorithms* | pp. 7 |
| 12:00–12:30pm | Mauro Maggioni, Duke University<br>*Intrinsic Dimensionality Estimation and Multiscale Geometry of Data Sets* | pp. 10 |
| 12:30–2:30pm | *Lunch* (on your own) | |
| **Third Session** | | **Cubberley Auditorium** |
| 2:30–3:00pm | Guillermo Sapiro, University of Minnesota<br>*Structured Sparse Models* | pp. 11 |
| 3:00–3:30pm | Aleck Agarwal and Peter Bartlett, University of California, Berkeley<br>*Optimal regret in online learning* | pp. 6 |
| 3:30–4:00pm | John Duchi and Yoram Singer, UC Berkeley & Google Research<br>*Composite Objective Optimization and Learning for Massive Datasets* | pp. 12 |
| 4:00–4:30pm | *Coffee break* | |
| **Fourth Session** | | **Cubberley Auditorium** |
| 4:30–5:00pm | Steve Hillion, Greenplum<br>*MAD Analytics in Practice* | pp. 8 |
| 5:00–5:30pm | Matthew Harding, Stanford University<br>*Outlier Detection in Financial Trading Networks* | pp. 8 |
| 5:30–6:00pm | Neel Sundrahan, eBay Research<br>*Large Dataset Problems at the Long Tail* | pp. 12 |

## Thursday, June 17, 2010: Spectral Methods and Sparse Matrix Methods

| Time | Event | Location/Page |
|---|---|---|
| **First Session** | | **Cubberley Auditorium** |
| 9:00–10:00am | Sebastiano Vigna, Universit Degli Studi Di Milano<br>*Spectral Ranking (Tutorial)* | pp. 12 |
| 10:00–10:30am | Robert Stine, University of Pennsylvania<br>*Streaming Feature Selection* | pp. 12 |
| 10:30–11:00am | *Coffee break* | |
| **Second Session** | | **Cubberley Auditorium** |
| 11:00–11:30am | Konstantin Mischaikow, Rutgers University<br>*A Combinatorial Framework for Nonlinear Dynamics* | pp. 10 |
| 11:30–12:00n | Alfred Hero, University of Michigan, Ann Arbor<br>*Sparse Correlation Screening in High Dimension* | pp. 8 |
| 12:00–12:30pm | Susan Holmes, Stanford University<br>*Challenges in Statistical Analyses: Heterogeneous Data* | pp. 8 |
| 12:30–2:30pm | *Lunch* (on your own) | |
| **Third Session** | | **Cubberley Auditorium** |
| 2:30–3:30pm | Piotr Indyk, Massachusetts Institute of Technology<br>*Sparse Recovery Using Sparse Matrices (Tutorial)* | pp. 8 |
| 3:30–4:00pm | Sayan Mukherjee, Duke University<br>*Efficient Dimension Reduction on Massive Data* | pp. 11 |
| 4:00–4:30pm | *Coffee break* | |
| **Fourth Session** | | **Cubberley Auditorium** |
| 4:30–5:00pm | Padhraic Smyth, University of California, Irvine<br>*Statistical Modeling of Large-Scale Sensor Count Data* | pp. 12 |
| 5:00–5:30pm | Ping Li, Cornell University<br>*Adaptive Base Class Boost (ABC-Boost) for Multi-Class Classification* | pp. 9 |
| 5:30–6:00pm | Edo Liberty, Yahoo! Research<br>*Scaleable Correlation Clustering Algorithms* | pp. 9 |
| **Evening Reception** | | |
| 6:00–9:00pm | *Dinner Reception and Poster Session* | **Cubberley Courtyard** |

## Friday, June 18, 2010: Randomized Algorithms for Data

| Time | Event | Location/Page |
|---|---|---|
| **First Session** | | **Cubberley Auditorium** |
| 9:00–10:00am | Petros Drineas, Rensselaer Polytechnic Institute<br>*Randomized Algorithms in Linear Algebra and Large Data Applications (Tutorial)* | pp. 7 |
| 10:00–10:30am | Gunnar Martinsson, University of Colorado, Boulder<br>*Randomized methods for computing the SVD/PCA of very large matrices* | pp. 10 |
| 10:30–11:00am | *Coffee break* | |
| **Second Session** | | **Cubberley Auditorium** |
| 11:00–11:30pm | Ilse Ipsen, North Carolina State University<br>*Numerical Reliability of Randomized Algorithms* | pp. 9 |
| 11:30–12:00n | Patrick Wolfe, Harvard University<br>*Randomized Algorithms and Sampling Schemes for Large Matrices* | pp. 13 |
| 12:00–12:30pm | Alexandre d'Aspremont, Princeton University<br>*Subsampling, Spectral Methods & Semidefinite Programming* | pp. 6 |
| 12:30–2:30pm | *Lunch* (on your own) | |
| **Third Session** | | **Cubberley Auditorium** |
| 2:30–3:00pm | Gary Miller, Carnegie Mellon University<br>*Specialized System Solvers for Very Large Systems: Theory and Practice* | pp. 10 |
| 3:00–3:30pm | John Wright and Emmanuel Candes, Microsoft & Stanford<br>*Robust Principal Component Analysis?* | pp. 7 |
| 3:30–4:00pm | *Coffee break* | |
| **Fourth Session** | | **Cubberley Auditorium** |
| 4:00–4:30pm | Alon Orlitsky, University of California, San Diego<br>*Estimation, Prediction, and Classification over Large Alphabets* | pp. 11 |
| 4:30–5:00pm | Ken Clarkson, IBM Almaden Research<br>*Numerical Linear Algebra in the Streaming Model* | pp. 7 |
| 5:00–5:30pm | David Woodruff, IBM Almaden Research<br>*Fast $\ell_p$ Regression in Data Streams* | pp. 13 |

## Poster Presentations: Thursday, June 17, 2010

# Talk Abstracts

### Estimating Rates of Rare Events through Multiple Hierarchies

Deepak K Agarwal, Yahoo! Research

We consider the problem of estimating rates of rare events for high dimensional, multivariate categorical data where several dimensions are hierarchical. Such problems are routine in several data mining applications including computational advertising, our main focus in this talk. We propose a novel log-linear model that scales to massive data applications with billions of training records and several million potential predictors in a map-reduce framework. Our method exploits correlations in aggregates observed at multiple resolutions when working with multiple hierarchies; stable estimates at coarser resolution provide informative prior information to improve estimates at finer resolutions. Other than prediction accuracy and scalability, our method has an inbuilt variable screening procedure through a "spike and slab" prior that provides parsimony by pruning non-informative predictors without hurting predictive accuracy. We perform large scale experiments on data from Right Media Ad Exchange at Yahoo! and illustrate our approach on datasets with several billion records and hundreds of millions of predictors. Extensive comparisons with other benchmark methods show significant improvements in prediction accuracy.

### Subsampling, Spectral Methods & Semidefinite Programming

Alexandre d'Aspremont, Princeton University

We show how subsampling techniques can be used to approximate eigenvectors of large matrices and discuss applications in dimensionality reduction and ranking. We then use these results to reduce the per iteration complexity of first order algorithms for semidefinite programming. The subsampling ratio explicitly controls the algorithm's granularity, i.e. the tradeoff between cost per iteration and total number of iterations. In this setting, the method's total computational cost is also directly proportional to the complexity (i.e. rank) of the solution.

### Massive Scale Analytics of Streaming Social Networks

David Bader, Georgia Tech College of Computing

Emerging real-world graph problems include detecting community structure in large social networks, improving the resilience of the electric power grid, and detecting and preventing disease in human populations. Unlike traditional applications in computational science and engineering, solving these problems at scale often raises new challenges because of sparsity and the lack of locality in the data, the need for additional research on scalable algorithms and development of frameworks for solving these problems on high performance computers, and the need for improved models that also capture the noise and bias inherent in the torrential data streams. The explosion of real-world graph data poses a substantial challenge: How can we analyze constantly changing graphs with billions of vertices? Our approach leverages the Cray XMT's fine-grained parallelism and flat memory model to scale to massive graphs. On the Cray XMT, our static graph characterization package GraphCT summarizes such massive graphs, and our ongoing STINGER streaming work updates clustering coefficients on massive graphs at a rate of tens of thousands updates per second.

### Optimal regret in online learning

Aleck Agarwal & Peter Bartlett, University of California, Berkeley

We study the regret of optimal strategies for online learning. Using von Neumann's minimax theorem, we show that the optimal regret in an adversarial setting is closely related to the behavior of the empirical minimization algorithm in a stochastic process setting: it is equal to the maximum, over joint distributions of the adversary's action sequence, of the difference between a sum of minimal expected losses and the minimal empirical loss. We show that the optimal regret has a natural geometric interpretation, since it can be viewed as the gap in Jensen's inequality for a concave functional–the minimizer over the player's actions of expected loss–defined on a set of probability distributions. We use this result to obtain upper and lower bounds on the regret of an optimal strategy for a variety of online learning problems.

Joint work with Jacob Abernethy and Alexander Rakhlin.

### Statistical Inference for Networks

Peter Bickel, University of California, Berkeley

We'll discuss some different models for networks and approaches to community identification in the physics and statistics literature, as discussed in a recent book by M.E.J. Newman, and a potentially unifying nonparametric statistical model for unlabelled graphs. We'll discuss from the point of view of asymptotics fitting using modularities such as that proposed by Newman and Girvan, as well as by maximum likelihood and the "method of moments" and indicate further directions of research.

### Robust Principal Component Analysis?

John Wright & Emmanuel Candes, Microsoft Research Asia & Stanford University

This talk is about a curious phenomenon. Suppose we have a data matrix, which is the superposition of a low-rank component and a sparse component. Can we recover each component individually? We prove that under some suitable assumptions, it is possible to recover both the low-rank and the sparse components exactly by solving a very convenient convex program called Principal Component Pursuit; among all feasible decompositions, simply minimize a weighted combination of the nuclear norm and of the L1 norm. This suggests the possibility of a principled approach to robust principal component analysis since our methodology and results assert that one can recover the principal components of a data matrix even though a positive fraction of its entries are arbitrarily corrupted. This extends to the situation where a fraction of the entries are missing as well. We discuss an algorithm for solving this optimization problem emphasizing the suitability of this approach for large scale problems, and present applications in the area of video surveillance, where our methodology allows for the detection of objects in a cluttered background, and in the area of face recognition, where it offers a principled way of removing shadows and specularities in images of faces. Joint work with X. Li and Y. Ma.

### AdHeat - A New Influence-based Social Ads Model and its Tera-Scale Algorithms

Edward Chang, Google Research

In this talk, I present AdHeat, a social ad model considering user influence in addition to relevance for matching ads. Traditionally, ad placement employs the relevance model. Such a model matches ads with Web page content, user interests, or both. We have observed, however, on social networks that the relevance model suffers from two shortcomings. First, influential users (users who contribute opinions) seldom click ads that are highly relevant to their expertise. Second, because influential users' contents and activities are attractive to other users, hint words summarizing their expertise and activities may be widely preferred by non-influential users, whose information is too sparse to yield effective relevance analysis. Therefore, we propose AdHeat, which diffuses hint words of influential users to others and then matches ads for each user with aggregated hints. Our experimental results show AdHeat to enjoy a significant improved performance in CTR. Key tera-scale algorithms of AdHeat are presented to explain its effectiveness.

### Numerical Linear Algebra in the Streaming Model

Ken Clarkson, IBM Almaden Research

We give near-optimal space bounds in the streaming model for linear algebra problems that include estimation of matrix products, linear regression, low-rank approximation, and approximation of matrix rank.

In the streaming model, we take one pass over the matrix entries, in an arbitrary order (that is not up to us). We wish to store a sketch of each input matrix, that is, a compressed version of it, and then use the sketch for computation.

For matrices A and B, our sketches are simply S'A and S'B, where S is a random sign matrix, with each entry being +1 or -1 with equal probability. (Here A,B, and S all have the same number of rows, and S' denotes the transpose of S.) We sharpen prior quality guarantees for estimating A'B as (S'A)'S'B = A'SS'B, and prove novel lower bounds for the number of bits needed for sketches such that such guarantees can hold.

We also show related results for best rank-k approximation; for example, that the sketches S'A and AR, where R is another sign matrix, can be used to approximate A by a matrix $Q := AR(S'AR)^{-}S'A$ of rank at most k/epsilon, where $(S'AR)^{-}$ is the Moore-Penrose pseudo-inverse of S'AR, so that with probability at least $1 - \delta$,

$$||A - Q|| \leq (1 + \epsilon)||A - A_k||$$

where $A_k$ is the best approximation to A of rank k. Here $\|.\|$ denotes the Frobenius norm, and the number of columns of S and R depend on $\epsilon$, $\delta$, and k. This talk will focus on our upper bounds.

Joint with David Woodruff.

### Minimizing Communication in Linear Algebra

James Demmel, University of California, Berkeley

Algorithms have two kinds of costs: arithmetic and communication, by which we mean moving data either between levels of a memory hierarchy (in the sequential case) or between processors over a network (in the parallel case). Communication costs can already exceed arithmetic costs by orders of magnitude, and the gap is growing exponentially over time, so our goal is to design linear algebra algorithms that minimize communication. First, we show how to extend known communication lower bounds for $O(n^3)$ dense matrix multiplication to all direct linear algebra, i.e. for solving linear systems, least squares problems, eigenproblems and the SVD, for dense or sparse matrices, and for sequential or parallel machines. We also describe dense algorithms that attain these lower bounds; some implementations attain large speed ups over conventional algorithms. Second, we show how to minimize communication in Krylov-subspace methods for solving sparse linear system and eigenproblems, and again demonstrate new algorithms with significant speedups.

### Randomized Algorithms in Linear Algebra and Large Data Applications

Petros Drineas, Rensselaer Polytechnic Institute

The introduction of randomization into the design and analysis of algorithms for common matrix problems (such as matrix multiplication, least-squares regression, the Singular Value Decomposition (SVD), etc.) over the last decade has provided a novel paradigm and complementary perspective to traditional numerical linear algebra approaches to matrix computations. This paradigm was motivated by technological developments in many areas of scientific and Internet research that permit the automatic generation of large data

sets that are often modeled as matrices, and in the last few years this paradigm has borne fruit in numerical implementation and data applications. In this talk, we will provide an overview of this approach, including how it can be used to approximate problems ranging from matrix multiplication and the SVD of matrices to approximately solving least-squares problems and systems of linear equations. In addition, application of these algorithms to large-scale data analysis will also be discussed.

### Combinatorial Scientific Computing: Experience and Challenges

John R. Gilbert, University of California, Santa Barbara

High-peformance scientific computing is sometimes thought of as a domain of continuous models, differential equations, numerical approximations, and floating-point arithmetic. However, combinatorial algorithms have long played a crucial role in computational science, from sparse matrix algorithms to optimization to mesh generation and parallel computing. Modern computational biology, chemistry, and data analysis are now highly influenced by discrete methods.

This tutorial talk will review some of the history, applications, and techniques of combinatorial scientific computing. I will discuss a number of emerging challenges in the areas of tools and technologies, drawing examples from various applications; and I will highlight our group's work on algebraic tools for high-performance computation on large graphs and networks.

### Outlier Detection in Financial Trading Networks

Matthew Harding, Stanford University

Recent developments in financial markets have enabled high frequency algorithms to dominate trading on electronic exchanges. This paper analyzes the resulting pattern of trades between human and machine traders at very high frequencies. We show that the resulting trading networks can be interpreted as a series of exchangeable graphs. We introduce tensor decompositions as an efficient estimation procedure designed to detect outliers in the resulting trading patterns. The method is applied to recent events in financial markets and can be used to analyze the impact of algorithmic trading on market volatility and extreme events.

Joint work with Andrei Kirilenko.

### Sparse Correlation Screening in High Dimension

Alfred Hero, University of Michigan

We consider the problem of screening for a few high correlations among a large number p of variables based on a small number n of samples. This problem arises in many disciplines including gene expression analysis, finance, and security, where the number of variables can range from a few hundred to hundreds of thousands and the sample covariance matrix is low rank. We distinguish between three types of correlation screening: screening auto-correlations in a single treatment, screening cross-correlations in multiple treatments, and screening persistent auto-correlations in multiple treatments. In each of these applications screening is accomplished by thresholding the sample correlation matrix. We obtain asymptotic large p expressions for the phase transition thresholds and familywise error rates. These give accurate and useful approximations when the correlation matrices are sparse and p is finite. The methodology is illustrated on both simulations and real data.

Joint work with Bala Rajaratnam.

### MAD Analytics in Practice

Steven Hillion, Greenplum

The explosion of data within a broad array of sectors, including telecommunications, government, internet and retail, has been accompanied by increasing sophistication in the use of analytics, beyond simple reporting. This poses a technical challenge to the traditional data warehouse model. Rapid and frequent execution of statistical tests and models on very large data sets requires a fundamental change in the philosophy of managing an organizations data. One proposal, dubbed "MAD Skills," was put forward at the 2009 VLDB conference by a team of data scientists, researchers and engineers. The database company Greenplum, which builds massively-parallel systems for large-scale data processing and analytics, has developed an array of functional capabilities and customer solutions based on this philosophy. In this presentation, we describe the MAD Skills approach and provide specific examples that illustrate the application of Greenplums scalable computing power to common analytical problems.

### Challenges in Statistical Analyses: Heterogeneous Data

Susan Holmes, Stanford University

Combining and analysing heterogeneous data poses several challenges in statistical data integration. I will show how some data abide naturally by their structure. (Categorical, continuous, ordinal, treelike, spatial). However, the relevant summary obtained by combining data of different types poses a challenge we will discuss. Using examples from large biological data sets I will show how sometimes it is better to summarize continuous data by a category, binary data by a tree and a set of trees by a continuum.

### Title Sparse Recovery Using Sparse Matrices

Piotr Indyk, Massachusetts Institute of Technology

Over the recent years, a new *linear* method for compressing high-dimensional data has been discovered. For a high-dimensional vector x, its compressed version (a.k.a. "sketch") is equal to Ax, where A is an m x n matrix (possibly chosen at random). Although typically the sketch length m is much smaller than the number of dimensions n, the sketch contains enough information to recover a good

"sparse approximation" to x. At the same time, the linearity of the sketching method is very convenient for many areas, such as data stream computing and compressive sensing

In this talk we survey sparse recovery results that utilize *sparse* matrices A. Such matrices have several attractive properties: they support algorithms with low computational complexity, and make it easy to perform incremental updates to vectors x.

Joint work with Anna Gilbert.

### Numerical Reliability of Randomized Algorithms

Isle Ipsen, North Carolina State University

We discuss how randomized algorithms are affected by changes in the inputs, such as the effects of finite precision arithmetic or changes in the probability distributions. In particular, we are interested in subset selection algorithms, their sensitivity (i.e numerical conditioning), and their reliability (i.e. numerical stability).

### Hyperbolic Mapping of Complex Networks

Dmitri Krioukov, CAIDA/UCSD

We establish a connection between the heterogeneous topology of complex networks, and the hyperbolic geometry of hidden metric spaces underlying these networks. Given a hyperbolic space, heterogeneous networks with strong clustering naturally emerge on top of the space as topological reflections of its geometry. Conversely, for any heterogeneous network with strong clustering, there is an effective hyperbolic space underlying the network. This framework allows one to study transport and communication processes on networks using only local information. We show that the efficiency of such processes achieves its theoretical maximum on complex networks only if the underlying space is hyperbolic. We discuss methods to map real complex networks to their hyperbolic spaces using statistical inference techniques.

### A method for Parallel Online Learning

John Langford, Yahoo! Research

In many cases, the fastest method known for learning a function in some function class is via online gradient descent. Unfortunately, common architectures for parallelism over massive datasets often do not support these kinds of algorithms – and even the question of _how_ to parallelize these kinds of algorithms when there are large amount of data remains unclear. I will present a method for effectively parallelizing these algorithms along with experimental results showing that the method works well.

### Inferring Networks of Diffusion and Influence

Jure Lescovec, Stanford University

Tracking information diffusion or virus propagation over networks has been a topic of considerable interest. While directly observing nodes becoming infected is often possible, observing individual transmissions (i.e., who infects or influences whom) is typically very difficult. Furthermore, in many applications, the underlying network over which the diffusions and propagations spread is actually unobserved. We tackle these challenges by developing a method for tracing paths of diffusion and influence through networks and inferring the networks over which contagions propagate. Given the times when nodes adopt pieces of information or become infected, we wish to identify the optimal network that best explains the observed infection times. Since the optimization problem is NP-hard to solve exactly.

We demonstrate the effectiveness of our approach by tracing information cascades in a set of 170 million blogs and news articles over a one year period to infer how information flows through the online media space. The diffusion network of news tends to have a core-periphery structure with a small set of core blogs and news media sites that diffuse information to the rest of the Web. These sites tend to have stable circles of influence with more general news media sites acting as connectors between them.

### Adaptive Base Class Boost (ABC-Boost) for Multi-Class Classification

Ping Li, Cornell University

Classification is a fundamental task in statistics and machine learning. Many practical applications can be cast as multi-class classification problems. This talk will present an interesting recent development of boosting for multi-class classification. Boosting (Schapire 1990, Freund 1995) has been highly successful in machine learning theory and, in particular, industry practice. This talk will focus on the improvement over MART (multiple additive regression trees, Friedman 2001) and LogitBoost (Friedman, Hastie, and Tibshirani, 2000), as well as comparisons with SVM and Deep Learning.

### Scaleable Correlation Clustering Algorithms

Edo Liberty, Yahoo! Research

A Correlation Clustering (CC) algorithm receives as input a graph G and returns a clustering graph C (a disjoint set of cliques). The cost of the solution is the symmetric difference between C and G, i.e., the number of edges they disagree on. In the bipartite case (BCC) the input graph is bipartite and the output is a set of disjoint bi-cliques. For CC, randomized and deterministic algorithms are known that achieve a constant factor approximation. For BCC, a factor log(n) approximation can be achieved by a more general weighted correlation clustering algorithm, where n is the number of nodes in G. These however, require solving large Linear Programs which make them impractical for moderately sized problems. In this talk, I will present two randomized algorithms, one for CC and one for BCC. Both give constant

factor approximations to their respective problems and are highly scalable. Time permits, I will point to open problems. This is joint work with Nir Ailon.

### Intrinsic Dimensionality Estimation and Multiscale Geometry of Data Sets

Mauro Maggioni, Duke University

The analysis of large data sets, modeled as point clouds in high dimensional spaces, is needed in a wide variety of applications such as recommendation systems, search engines, molecular dynamics, machine learning, statistical modeling, just to name a few. Often times it is claimed or assumed that many data sets, while lying in high dimensional spaces, have indeed a low-dimensional structure. It may come perhaps as a surprise that only very few, and rather sample-inefficient, algorithms exist to estimate the intrinsic dimensionality of these point clouds. We present a recent multiscale algorithm for estimating the intrinsic dimensionality of data sets, under the assumption that they are sampled from a rather tame low-dimensional object, such as a manifold, and perturbed by high dimensional noise. Under natural assumptions, this algorithm can be proven to estimate the correct dimensionality with a number of points which is merely linear in the intrinsic dimension. Experiments on synthetic and real data will be discussed. Furthermore, this algorithm opens the way to novel algorithms for exploring, visualizing, compressing and manipulating certain classes of high-dimensional point clouds.

### Geometric Network Analysis Tools

Michael W. Mahoney, Stanford University

Large social and information networks are typically modeled as graphs, i.e., as data structures that have combinatorial but not-obviously geometric properties. Thus, many of the most popular tools in statistics and machine learning, e.g., SVD, PCA, RKHSs, manifolds, and their variants and generalizations, are far from immediately-applicable for such network data. Many of these popular tools have a geometric underpinning, which provides relatively nice algorithmic properties, robustness to noise, and a basis for inference and learning. On the other hand, it is known that the key to understanding the worse-case behavior of many graph algorithms lies in understanding the metric and geometric embedding properties of different classes of graphs.

Here, we describe the use of geometric properties underlying scalable worst-case graph approximation algorithms for analyzing the empirical properties of large and adversarially-structured informatics graphs. A key issue will be understanding the statistical properties implicit in worst-case approximation algorithms. These *geometric network analysis tools* can be used as very fine "experimental probes" of the structure of networks with millions or more of nodes, thus permitting one to test commonly-made data analysis hypotheses (such as that the data live in a nice low-dimensional place or have intuitive clustering structure). In addition, these tools can be used to characterize very finely the coupling between local (often roughly Euclidean) properties and global (typically expander-like or hyperbolic) properties in large networks arising in many areas of data analysis.

### Randomized Methods for Computing the SVD/PCA of Very Large Matrices

Gunnar Martinsson, University of Colorado, Boulder

The talk will describe a set of recently developed randomized techniques for standard linear algebraic computations such as computing a partial singular value decomposition of a matrix. The techniques are very well suited for implementation on multi- core or other parallel architectures and for processing data stored outside of RAM, or streamed. Remarkably, the randomized sampling that is used not only loosen communication constraints, but does so while maintaining, or even improving, upon the accuracy and robustness of existing deterministic techniques.

### Specialized System Solvers for Very Large Systems: Theory and Practice

Gary L. Miller, Carnegie Mellon University

We will discuss the latest linear system solvers for very large sparse Symmetric Diagonally Dominate system (SDD). This seemingly restrictive class of systems has received substantial interest and in the last 15 years both algorithm design theory and practical implementations have made substantial progress. Due to the nearly linear run times for these systems there is also a growing number of problems that can be efficiently solved using SDD solvers including: image segmentation, image denoising, finding solutions to elliptic equations, computing maximum flow in a graph, and other problems in graphics.

Theoretically, we have recently shown that if $A$ is an $n \times n$ SDD matrix with $m$ entries the linear system $Ax = b$ can be solve to constant precision in $O(m \log^2 n)$ time ignoring lower order factors. On the implementation side, the latest code runs in linear time experimentally for large sparse systems. In fact the problem sizes are now large enough that on modern multicore workstations the memory bandwidth is a limiting factor. We discuss compression techniques that substantially speedup the solver time.

### A Combinatorial Framework for Nonlinear Dynamics

Konstantin Mischaikow, Rutgers University

It is almost a tautology that any valid description of the global dynamics of a multiparameter nonlinear system arrived at through numerical simulation or the accumulation of experimental data requires that the structures being described can be represented via a finite amount of data and are robust with respect to perturbations. We shall outline a combinatorial approach to dynamics that has these properties. In particular, using a simple example multiple parameter problem arising from population dynamics we will build

a database that provides a coarse rigorous description of the global dynamics at every parameter value. We will describe the algorithms used to approximate the dynamics and how algebraic topological tools are used to provide rigorous interpretations of the underlying continuous system based on the approximate dynamics.

### Matrix Approximation for Large-scale Learning

Mehryar Mohri, New York University

A crucial technique for scaling many learning algorithms to very large data sets reaching or exceeding millions of instances is based on low-rank approximations of kernel matrices. We describe several Nystrom-based approximation algorithms, including an ensemble algorithm shown to generate more accurate low-rank approximations. We give learning bounds for these methods and further analyze the impact of such approximations on the learning performance of several commonly used algorithms, such as kernel ridge regression and support vector machines. We also report the results of extensive experiments with several data sets containing up to 1M points demonstrating significant improvements over the standard Nystrom approximation.

Joint work with Corinna Cortes, Sanjiv Kumar, and Ameet Talwalkar.

### Efficient Dimension Reduction on Massive Data

Sayan Mukherjee, Duke University

Motivated by a problem in population genetics, the inference of genetic population structure, we explore efficient dimension reduction on massive data sets. Our objective is to implement on a standard desktop various flavors of dimension reduction procedures based on spectral methods. The scale of the data is 100,000-1,000,000 variables and tens of thousands of observations. We discuss efficient methods for unsupervised dimension reduction, supervised dimension reduction, and non-linear embeddings. A key insight will be that certain aspects of numerical accuracy can be sacfriced with very little loss of accuracy with respect to inference. Illustrations will be shown using simulated data as well as real genetic (SNP) and molecluar (expression) data.

### Internet-Scale Data Analysis

Peter Norvig, Google Research

The Internet provides unprecedented amount of data: text, images, video, links, clicks, and other data types. This tutorial describes data analysis/learning algorithms that are scalable, widely-applicable, and useful for building applications.

### Estimation, Prediction, and Classification over Large Alphabets

Alon Orlitsky, University of California, San Diego

It is clear that when analyzing data, symbols can be replaced by their order of appearance. For example the words "Sam I am I am Sam" can be represented by "1 2 3 2 3 1." What is less clear, and perhaps counter intuitive, is that analyzing the latter "pattern" yields different, and typically better results. We will describe the pattern approach to probability estimation, sequential prediction, and classification, and provide several experimental demonstrations of its efficacy. Based on work with J. Acharya, H. Das, S. Pan, P. Santhanam, K. Viswanathan, and J. Zhang.

### Fast Pseudo-Random Fingerprints

Ely Porat, Bar-Ilan University

We propose a method to exponentially speed up computation of various fingerprints, such as the ones used to compute similarity and rarity in massive data sets. Rather then maintaining the full stream of $b$ items of a universe $[u]$, such methods only maintain a concise fingerprint of the stream, and perform computations using the fingerprints. The computations are done approximately, and the required fingerprint size $k$ depends on the desired accuracy $\epsilon$ and confidence $\delta$. Our technique maintains a single bit per hash function, rather than a single integer, thus requiring a fingerprint of length $k = O(\frac{\ln \frac{1}{\delta}}{\epsilon 2})$ bits, rather than $O(\log u \cdot \frac{\ln \frac{1}{\delta}}{\epsilon 2})$ bits required by previous approaches. The main advantage of the fingerprints we propose is that rather than computing the fingerprint of a stream of $b$ items in time of $O(b \cdot k)$, we can compute it in time $O(b \log k)$. Thus this allows an exponential speedup for the fingerprint construction, or alternatively allows achieving a much higher accuracy while preserving computation time. Our methods rely on a specific family of pseudo-random hashes.

### Structured Sparse Models

Guillermo Sapiro, University of Minnesota

In this talk I will describe some recent results related to the introduction of structure in sparse models. I will first describe joint work with G. Yu and S. Mallat on image enhancement based on learning a block-sparse dictionary, and show its relationship with EM, manifold learning, and collaborative filtering. Then we present a collaborative hierarchical sparse model where the signals can share structure at different hierarchies of the dictionary. This second part is in collaboration with P. Sprechmann, I. Ramirez, and Y. Eldar.

### Composite Objective Optimization and Learning for Massive Datasets

John Duchi & Yoram Singer, University of Califonia, Berkeley & Google Research

Composite objective optimization is concerned with the problem of minimizing a two-term objective function which consists of an empirical loss function and a regularization function. Application with massive datasets often employ a regularization term which is non-differentiable or structured, such as L1 or mixed-norm regularization. Such regularizers promote sparse solutions and special structure of the parameters of the problem, which is a desirable goal for datasets of extremely high-dimensions. In this talk, we discuss several recently developed methods for performing composite objective minimization in the online learning and stochastic optimization settings. We start with a description of extensions of the well-known forward-backward splitting method to stochastic objectives. We then generalize this paradigm to the family of mirror-descent algorithms. Our work builds on recent work which connects proximal minimization to online and stochastic optimization. We conclude the algorithmic part with a description of a new approach, called AdaGrad, in which the proximal function is adapted throughout the course of the algorithm in a data-dependent manner. This temporal adaptation metaphorically allows us to find needles in haystacks as the algorithm is able to single out very predictive yet rarely observed features. We conclude with several experiments on large-scale datasets that demonstrate the merits of composite objective optimization and underscore superior performance of various instantiations of AdaGrad.

### Statistical Modeling of Large-Scale Sensor Count Data

Padhraic Smyth, University of California, Irvine

Modern sensor technologies allow us to capture rich data sets related to human behavior. A common form of this data is aggregated time-series of counts, e.g., how many people enter and exit a building every few minutes, how many vehicles pass over a particular point on a road, how many users access a Web site, and so on. In this talk I will describe our recent work on learning normal patterns and detecting anomalous events in such data. Events are characterized as local bursts of activity that look anomalous relative to normal hourly and daily patterns of behavior. The difficulty with this approach (as is the case with any outlier detection problem) is how to identify what is normal and what is anomalous, given no labeled training data. I will describe a statistical learning framework to address this problem, where we model normal behavior by an inhomogeneous Poisson process, which is in turn modulated by a hidden Markov model for bursty events. Experimental results will be illustrated using large real-world data sets collected over several months, involving people entering and exiting a UC Irvine campus building and data from freeway traffic sensors in the Southern California area. The talk will conclude with a brief discussion of open problems and ongoing work in this area.

Joint work with Alex Ihler and Jon Hutchin.

### Virtual Sensors and Large-Scale Gaussian Processes

Ashok Srivastava, NASA

We discuss the research and development of algorithms to support large-scale Gaussian Processes and their application to high resolution spectral data in the Earth Science domain. These algorithms are used to implement Virtual Sensors, a technique of emulating one sensor measurement given a number of correlated spectral measurements.

### Large Dataset Problems at the Long Tail

Neel Sundaresan, eBay Research

A Large Scale Marketplace Network like eBay provides unique opportunities and challenges to mining and learning from the data to build useful applications. User behavior data comes from user session data, product listing data, and user transaction data. The long tail nature and the diverse nature of each of these along different dimensions demand different solutions to known problems like Search, Recommender Systems, Classification, Trust and Reputation Systems, and Network Analysis. We describe these problems and approaches to solving these problems at scale.

### Streaming Feature Selection

Robert Stine, University of Pennsylvania

Streaming feature selection provides a framework for rapidly searching large collections of explanatory features. Our approach combines multiple streams of features. Each stream collects related features that have common substantive motivation (e.g. specific parts of speech) or that share similar construction (e.g. basis vectors for an RKHS). Feature ordering permits an external expert to influence the selection process without overfilling. A simple martingale argument shows that the approach controls the expected number of false discoveries. Our implementation includes methods for robustly judging the significance of the proposed explanatory variables; these methods assure that the fitting process is not distorted by common violations of model assumptions (such as anomalous outlying observations or dependent observations). Our examples apply this methodology to text data and multiple time series.

### Spectral Ranking

Sebastiano Vigna, Universit Degli Studi Di Milano

I will present a sketch of the history of spectral ranking–a general umbrella name for techniques that apply the theory of linear maps (in particular, eigenvalues and eigenvectors) to matrices that do not represent geometric transformations, but rather some kind of relationship between entities. Albeit recently made famous by the ample press coverage of Google's PageRank algorithm, spectral ranking was devised more than fifty years ago, almost exactly in the same terms, and has been studied in psychology and social sciences. Even if it started as a way to identify popular children among small groups, spectral ranking is now being applied to massive graphs. I will discuss some variants of the basic ideas, including recent developments that have been applied to search engines.

**Randomized Algorithms and Sampling Schemes for Large Matrices**

Patrick J. Wolfe, Harvard University

Randomized algorithms provide an appealing means of dimensionality reduction, helping to overcome the computational limitations currently faced by practitioners with massive datasets. In this talk we describe how new insights from multilinear algebra and statistics can be brought to bear on the design and analysis of algorithms and sampling schemes for large matrices, giving a number of new results in this direction. We discuss in particular the roles of determinants and compound matrices as they arise in a variety of recent settings of interest to the theoretical computer science community, as well as connections to other approaches based on random projections.

**Fast $\ell_p$ Regression in Data Streams**

David P Woodruff, IBM Almaden Research Center

I will talk about the $\ell_p$-Regression Problem, namely, given an $n \times m$ matrix $A$ and an $n \times 1$ column vector $b$, together with a parameter $\epsilon > 0$, output a vector $x'$ in $R^m$ for which $|Ax' - b|_p \leq (1+\epsilon) \min_{x \in R^m} |Ax - b|_p$ with good probability. In practice, $\ell_p$-regression is more robust than $\ell_2$-regression. We focus on the heavily overconstrained version. We give one pass $poly(d \log n/\epsilon)$ space algorithms in the general turnstile streaming model for every $1 \leq p < 2$. For a large range of parameters, our time complexity is close to optimal, improving the time of all previous algorithms, which are not low-space streaming algorithms.

Joint work with Christian Sohler.

# Poster Abstracts

### Algorithms for Large, Sparse Network Alignment Problems

Mohsen Bayati, Stanford University

We propose a new distributed algorithm for sparse variants of the network alignment problem that occurs in a variety of data mining areas including systems biology, database matching, and computer vision. Our algorithm uses a belief propagation heuristic and provides near optimal solutions for an NP-hard combinatorial optimization problem. We show that our algorithm is faster and outperforms or nearly ties existing algorithms on synthetic problems, a problem in bioinformatics, and a problem in ontology matching. We also provide a unified framework for studying and comparing all network alignment solvers.

Joint work with Margot Gerritsen, David Gleich, Amin Saberi and Ying Wang.

### Composite Objective Optimization and Learning for Massive Datasets

John Duchi, University California, Berkeley

Composite objective optimization is concerned with the problem of minimizing a two-term objective function which consists of an empirical loss function and a regularization function. Application with massive datasets often employ a regularization term which is non-differentiable or structured, such as L1 or mixed-norm regularization. Such regularizers promote sparse solutions and special structure of the parameters of the problem, which is a desirable goal for datasets of extremely high-dimensions. In this talk, we discuss several recently developed methods for performing composite objective minimization in the online learning and stochastic optimization settings. We start with a description of extensions of the well-known forward-backward splitting method to stochastic objectives. We then generalize this paradigm to the family of mirror-descent algorithms. Our work builds on recent work which connects proximal minimization to online and stochastic optimization. We conclude the algorithmic part with a description of a new approach, called AdaGrad, in which the proximal function is adapted throughout the course of the algorithm in a data-dependent manner. This temporal adaptation metaphorically allows us to find needles in haystacks as the algorithm is able to single out very predictive yet rarely observed features. We conclude with several experiments on large-scale datasets that demonstrate the merits of composite objective optimization and underscore superior performance of various instantiations of AdaGrad.

### Topological stability of the hippocampal spatial map

Facundo Memoli, Stanford University

The crucial role of the hippocampus in creating a spatial representation of the environment and in forming spatial memories is well known. The fact that the rodent hippocampal neurons, commonly known as place cells, tend to fire in a restricted region of the animals environment suggests that the rodent hippocampus codes for space but the nature of that spatial representation remains unclear. Current theories suggest that the hippocampus explicitly represents geometric elements of space derived from a path integration process that takes into account distances and angles of self motion information. However this hypothesis has difficulty explaining the results of several experimental studies that indicate that the hippocampal spatial map is invariant with respect to a significant range of geometrical transformations of the environment. This invariance suggests an alternative framework where hippocampal neural activity is best understood as representing the topology of the animals environment. We therefore suggest that the actual role of the hippocampus is to encode topological memory maps, where the patterns of ongoing neural activity represent the connectivity of locations in the environment or the connectivity of elements of a memory trace.

From a computational perspective, this hypothesis suggests that the temporal ordering of spiking from hippocampal neurons is the key determinant of the spatial information communicated to downstream structures. If so, then the variation seen in hippocampal firing rates should be limited to a range that preserves the global topological information encoded in the ensemble spike trains. More generally, if the overall approach to spatial information analysis is correct, the experimentally observed parameters of firing activity must guarantee the topological stability the hippocampal map. We therefore investigate the robustness of the hippocampal topological map with respect to independent variations of various place cell activity parameters, such as the firing rate and the distribution of sizes of place fields. We used the Persistent Homology method, applied to simulated data, to probe the complete range for each parameter independently and hence to theoretically establish the range of spiking parameters that lead to topological stability.

Join work with Yu. DABAGHIAN, F. MEMOLI, G. SINGH, L. M. FRANK, G. CARLSSON

### Greatly improved allele-specific tumor copy numbers with DNA microarrays when a matched normal is available

Pierre Neuvial, University of California, Berlekey

High-throughput genotyping microarrays assess both total DNA copy number and allelic composition, which makes them a tool of choice for copy number studies in cancer, including total copy number and loss of heterozygosity (LOH) analyses. Even after state of the art preprocessing methods, allelic signal estimates from genotyping arrays still suffer from systematic effects that make them difficult to use effectively for such downstream analyses.

We propose a method, TumorBoost, for normalizing allelic estimates of one tumor sample based on estimates from a single matched normal. The method applies to any paired tumor-normal estimates from any microarray-based technology, combined with any preprocessing method. We demonstrate that it increases the signal-to-noise ratio of allelic signals, making it significantly easier to detect allelic imbalances. TumorBoost increases the power to detect somatic copy-number events (including copy-neutral LOH) in the tumor from allelic signals of Affymetrix or Illumina origin.

Importantly, high-precision allelic estimates can be obtained from a single pair of tumor-normal hybridizations, if TumorBoost is combined with single-array preprocessing methods such as (allele-specific) CRMAvII for Affymetrix or BeadStudio's (proprietary) XY-normalization method for Illumina. This makes the method suitable for both large and small studies, and practical for applied medical diagnostics, because each patient can be analyzed independently of others. Based on these results, we recommend the use of matched normal samples in cancer copy number studies.

Joint work with Henrik Bengtsson and Terry Speed.

### Motion correction for in vivo Two Photon Laser Scanning Microscopy

Mihaela Obreja, University of Pittsburgh

Two-photon laser-scanning microscopy (TPLSM) can be used for in vivo neuroimaging of small animals. Due to the very high resolution of the images, brain motion is a source of large artifacts; tissue may be displaced by 10 or more pixels from its rest position. Thus, because the scanning rate is relatively slow comparing with the cardiac and respiratory cycles, some tissue pixels are scanned several times while others are never scanned. Consequently, although the images superficially appear reasonable, they can lead to incorrect conclusions with respect to brain structure and function. As a line is scanned almost instantaneous ( 1ms), our problem is reduced to relocating each of the lines in a three-dimensional stack of images to its "correct" location. Addressing the motion effects, we describe a Hidden Markov Model to estimate the sequence of hidden states most likely to have generated the sequence of observations. Our algorithm assigns probabilities for the states based on concomitant physiological measurements and estimates the most likely path of observed lines from the areas which move the least. Because there is no gold standard for comparison we compare our result with an image collected after the animal is sacrificed.

### A Spectral Algorithm for Exploring Partitions near a Given Seed

Lorenzo Orecchia, University of California, Berkeley

We propose a local version of the spectral relaxation to the minimum conductance problem. Rather than computing an approximation to the best partition in the entire input graph, we are motivated by the problem of finding a sparse partition near an input seed set. Such a primitive seems quite useful in improving and refining clusters locally in many settings such as image segmentation and the analysis of social networks. From a theoretical perspective, we show that the solution to our non-convex optimization relaxation for this problem may be computed as the solution to a system of linear equations. Interestingly, this also provides an optimization characterization of a generalization of Personalized PageRank vectors. We also show that how to obtain Cheeger-like quality-of-approximation guarantees when we round the solution and empirically illustrate the application of our ideas to the analysis of data graphs.

### Multithreaded Asynchronous Graph Traversal for In-Memory and Semi-External Memory

Roger Pearce, Texas A&M and LLNL

Processing large graphs is becoming increasingly important for many computational domains. Unfortunately, many algorithms and implementations do not scale with the demand for increasing graph sizes. As a result, researchers have attempted to meet the growing data demands using parallel and external memory techniques. Our work, targeted to chip multi-processors, takes a highly parallel asynchronous approach to hide the high data latency due to both poor locality and delays in the underlying graph data storage.

We present a novel asynchronous approach to compute Breadth First Search (BFS), Single Source Shortest Path (SSSP), and Connected Components (CC) for large graphs in shared memory. We present an experimental study applying our technique to both In-Memory (IM) and Semi-External Memory (SEM) graphs utilizing multi-core processors and solid-state memory devices. Our experiments using both synthetic and real-world datasets show that our asynchronous approach is able to overcome data latencies and provide significant speedup over alternative approaches.

### Similarity Search and Locality Sensitive Hashing using TCAMs

Rajendra Shinde, Stanford University

Similarity search methods are widely used as kernels in various data mining and machine learning applications including those in computational biology, web search/clustering. Nearest neighbor search (NNS) algorithms are often used to retrieve similar entries, given a query. While there exist efficient techniques for exact query lookup using hashing, similarity search using exact nearest neighbors suffers from a "curse of dimensionality," i.e. for high dimensional spaces, best known solutions offer little improvement over brute force search and thus are unsuitable for large scale streaming applications. Fast solutions to the approximate NNS problem include Locality Sensitive Hashing (LSH) based techniques, which need storage polynomial in $n$ with exponent greater than 1, and query time sublinear, but still polynomial in $n$, where $n$ is the size of the database. In this work we present a new technique of solving the approximate NNS problem in Euclidean space using a Ternary Content Addressable Memory (TCAM), which needs near linear space

and has O(1) query time. In fact, this method also works around the best known lower bounds in the cell probe model for the query time using a data structure near linear in the size of the data base.

TCAMs are high performance associative memories widely used in networking applications such as address lookups and access control lists. A TCAM can query for a bit vector within a database of ternary vectors, where every bit position represents 0, 1 or $*$. The $*$ is a wild card representing either a 0 or a 1. We leverage TCAMs to design a variant of LSH, called Ternary Locality Sensitive Hashing (TLSH) wherein we hash database entries represented by vectors in the Euclidean space into $\{0, 1, *\}$. By using the added functionality of a TLSH scheme with respect to the $*$ character, we solve an instance of the approximate nearest neighbor problem with 1 TCAM access and storage nearly linear in the size of the database. We validate our claims with extensive simulations using both real world (Wikipedia) as well as synthetic (but illustrative) datasets. We observe that using a TCAM of width 288 bits, it is possible to solve the approximate NNS problem on a database of size 1 million points with high accuracy. Finally, we design an experiment with TCAMs within an enterprise ethernet switch (Cisco Catalyst 4500) to validate that TLSH can be used to perform 1.5 million queries per second per 1Gb/s port. We believe that this work can open new avenues in very high speed data mining.

Joint work with Ashish Goel, Pankaj Gupta and Debojyoti Dutta.

### Optimized Intrinsic Dimension Estimation for High Dimensional Data

Kumar Sricharan, University of Michigan, Ann Arbor

Intrinsic dimension estimation is important for analysis of massive datasets whose principal modes of variation lie on a lower dimensional subspace. In such cases dimensionality reduction can be accomplished to eliminate redundancy in the data without loss of information. An accurate estimator of intrinsic dimension is a prerequisite for specifying model order in factor analysis methods such as principal components analysis (PCA), ISOMAP, and Laplacian eigenmaps. The most common method for selecting an embedding dimension for these algorithms was to detect a knee in a residual error curve, e.g., scree plots of sorted eigenvalues. In this presentation we introduce a new dimensionality estimator that is based on fluctuations of the sizes of nearest neighbor balls centered at a subset of the data points. Unlike other dimension estimation approaches our new dimension estimator is derived directly from a mean squared error (M.S.E.) optimality condition for partitioned kNN estimators of multivariate density functionals. This enables us to predict the performance of the dimension estimator and optimize the M.S.E. convergence rate among estimators in its class. Empirical experiments are presented that show that this asymptotic optimality translates into improved performance in the finite sample regime. We will illustrate our dimension estimator by applying it to the analysis of large scale gene expression data, Internet traffic, and spam email traces.

Joint work with R. Raich and Alfred O. Hero III.

### Reproducibility in Massive Datasets

Victoria Stodden, Yale University

Reproducibility is a cornerstone of the scientific method. Issues of reproducibility in computational research have recently been coming to the fore, and the difficulty of this problem scales with dataset size. CERN anticipates 15 petabytes of data produced each year, and is reviewing how to make these data open to facilitate reproducibility. Other efforts to ensure access to massive datasets are underway, such as the DataStaR project at Cornell University. This poster will discuss the state of the art of reproducible computational research, including data hosting and access, and outline problems faced, such as permitting data query when the data is static on the web, as happens in this most challenging case of reproducibility. Issues of intellectual property and data ownership typically create a barrier to the open release of research data, and a consistent method for their resolution is suggested, the "Reproducible Research Standard," to align intellectual property rights in computational science with longstanding scientific norm

### Minimizing I/O contention at NERSC using data analysis

Daniela Ushizima, Lawrence Berkeley National Laboratory

High performance computers (HPC) are a key resource to support algorithmic, mathematical and statistical research in modern large-scale data analysis, so it is important to identify I/O contention, quantify its variability and minimize its impact on HPC systems. I/O volume transfers can be monitored from the server or client-side, however client-side statistics are memory intensive, therefore prohibitive at large scale. The National Energy Research Scientific Computing (NERSC) center deployed the Lustre Monitoring Tool (LMT) for collecting server-side I/O statistics. One of the challenges is to detect events from I/O time series from LMT data streams, collected the past two years. We propose a scheme that detects transients in the signal to mark significant activities, for later signal segmentation into events. The main steps of the algorithm are: (a) definition of a minimum amplitude level, estimated from a year of observations; (b) signal convolution with a smoothing kernel to guarantee differentiability and minimize background oscillations; (c) derivative calculation to determine the maximal and minimal turning points of the signal to identify event start and end points. This simple, yet fast automated Tevent recognizer T may be able to characterize a large fraction of the I/O as being from separate jobs. The goal is to correlate the I/O signal intervals (detected events) with records of running jobs for that time interval, with minimum interference on the workload of the HPC system. This is an important tool to manage the I/O system by means of characterizing the workload and measuring its impact on the I/O system utilization.

Joint work with Andrew Uselton, Katie Antypas and Jeffrey Sukharev.

### Clustering the data by Completely Positive Factorizations

Changqing Xu, University of the Pacific

An $n$ by $n$ real matrix $A$ is called completely positive ($CP$) if $A$ can be factorized as $A = BB^T$, where $B$ is an $n \times m$ entrywise nonnegative matrix. The smallest possible number of the columns of $B$, is called the $CP$ rank of $A$, and denoted by $cprank(A)$. $CP$ matrices are closely related to some unsupervised clustering methods such as Kernel method and spectral clustering. We show that a probabilistic clustering is also related to it. Furthermore, we use the result on the $cprank(A)$ of the data matrix $A$ to tackle the long pending question in the unsupervised clustering: How to determine the number of the clusters, and obtain some interesting result.

### Evolutionary Spectral Clustering with Adaptive Forgetting Factor

Kevin S. Xu, University of Michigan, Ann Arbor

Many practical applications of clustering involve data collected over time. Often, the data to be clustered are nonstationary, and the underlying cluster structure changes with time. In these applications, evolutionary clustering can be applied to the data to track changes in clusters. We present an evolutionary version of spectral clustering that applies a forgetting factor to past affinities between data points and aggregates them with current affinities. We propose to use an adaptive forgetting factor and provide a method to automatically choose this forgetting factor at each time step. The forgetting factor is chosen through a bias-variance trade-off to optimize mean squared error. We evaluate the performance of the proposed method through experiments on synthetic and real data and find that, with an adaptive forgetting factor, we are able to obtain improved clustering performance compared to a fixed forgetting factor.

Joint work with Mark Kliger and Alfred O. Hero III.

# Acknowledgements

## Sponsors

The Organizers of MMDS 2010 and the MMDS Foundation would like to thank the following institutional sponsors for their generous support:

- AFOSR: the Air Force Office of Scientific Research
- LBL: the Lawrence Berkeley National Laboratory
- NSF: the National Science Foundation
- ONR: the Office of Naval Research
- Department of Mathematics, Stanford University