

MMDS 2016:

Workshop on Algorithms for Modern Massive Data Sets

Stanley Hall
University of California at Berkeley

June 21–24, 2016

The 2016 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2016) will address algorithmic and statistical challenges in modern large-scale data analysis. The goals of MMDS 2016 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, Alexander Shkolnik, and Petros Drineas*

Workshop Schedule

Tuesday, June 21, 2016: Data Analysis and Statistical Data Analysis

Time	Event	Location/Page
Registration & Opening		Lobby outside Stanley Hall
8:00–9:45am	<i>Breakfast and registration</i>	
9:45–10:00am	Organizers <i>Welcome and opening remarks</i>	
First Session		Stanley Hall
10:00–11:00am	Peter Wang, Continuum Analytics <i>Meaningful Visual Exploration of Massive Data</i>	pp. 13
11:00–11:30am	Lise Getoor, UC Santa Cruz <i>Scalable Collective Inference from Richly Structured Data</i>	pp. 8
11:30–12:00pm	Julian Shun, UC Berkeley <i>A Framework for Processing Large Graphs in Shared Memory</i>	pp. 12
12:00–2:00pm	<i>Lunch (on your own)</i>	
Second Session		Stanley Hall
2:00–2:30pm	Aarti Singh, Carnegie Mellon University <i>Minimax optimal subsampling for large sample linear regression</i>	pp. 12
2:30–3:00pm	Cameron Musco, Massachusetts Institute of Technology <i>Randomized Low-Rank Approximation and PCA: Beyond Sketching</i>	pp. 10
3:00–3:30pm	Alex Dimakis, UT Austin <i>Restricted Strong Convexity Implies Weak Submodularity</i>	pp. 7
3:30–4:00pm	<i>Coffee break</i>	
Third Session		Stanley Hall
4:00–4:30pm	Bin Yu, UC Berkeley <i>The Stability Principle for Information Extraction from Data</i>	pp. 14
4:30–5:00pm	Constantine Caramanis, UT Austin <i>New Results in Non-Convex Optimization for Large Scale Machine Learning</i>	pp. 7
5:00–5:30pm	Kristofer Bouchard, Lawrence Berkeley National Laboratory <i>The Union of Intersections Method</i>	pp. 7
5:30–6:00pm	Alex Smola, Carnegie Mellon University <i>Head, Torso and Tail - Performance for modeling real data</i>	pp. 12
Evening Reception		Lobby outside Stanley Hall
6:00–8:00pm	<i>Dinner Reception</i>	

Wednesday, June 22, 2016: Industrial and Scientific Applications

Time	Event	Location/Page
First Session		Stanley Hall
9:00–10:00am	Jasjeet Sekhon, UC Berkeley <i>New Methods for Designing and Analyzing Large Scale Randomized Experiment</i>	pp. 11
10:00–10:30am	Jonathan Berry, Sandia National Laboratories <i>Cooperative Computing for Autonomous Data Centers Storing Social Network Data</i>	pp. 7
10:30–11:00am	<i>Coffee break</i>	
Second Session		Stanley Hall
11:00–11:30am	Marina Meila, University of Washington <i>Is manifold learning for toy data only?</i>	pp. 10
11:30–12:00pm	Jake VanderPlas, University of Washington <i>Exploring Galaxy Evolution through Manifold Learning</i>	pp. 13
12:00–2:00pm	<i>Lunch (on your own)</i>	
Third Session		Stanley Hall
2:00–2:30pm	Daniela Witten, University of Washington <i>Fast, flexible, and interpretable regression modeling</i>	pp. 14
2:30–3:00pm	Vahab Mirrokni, Google Research <i>Randomized Composable Core-sets for Distributed Computation</i>	pp. 10
3:00–3:30pm	Kimon Fountoulakis, University of California Berkeley <i>Local graph clustering algorithms: an optimization perspective</i>	pp. 8
3:30–4:00pm	<i>Coffee break</i>	
Fourth Session		Stanley Hall
4:00–4:30pm	Dacheng Xiu, Chicago Booth <i>Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data</i>	pp. 14
4:30–5:00pm	Lisa Goldberg, UC Berkeley <i>Identifying Broad and Narrow Financial Risk Factors with Convex Optimization: Part 1</i>	pp. 9
5:00–5:30pm	Alex Shkolnik, UC Berkeley <i>Identifying Broad and Narrow Financial Risk Factors with Convex Optimization: Part 2</i>	pp. 12
5:30–6:00pm	Serena Ng, Columbia University <i>Learning about business cycle conditions from four terabytes of data</i>	pp. 10

Thursday, June 23, 2016: Novel Algorithmic Approaches

Time	Event	Location/Page
First Session		
9:00–10:00am	Prabhat, Lawrence Berkeley National Laboratory <i>Top 10 Data Analytics Problems in Science</i>	Stanley Hall pp. 11
10:00–10:30am	Alex Gittens, ICSI and UC Berkeley <i>Low-rank matrix factorizations at scale: Spark for scientific data analytics</i>	pp. 8
10:30–11:00am	<i>Coffee break</i>	
Second Session		
11:00–11:30am	Abbas Ourmazd, Univ. of Wisconsin Milwaukee <i>Structure and Dynamics from Random Observations</i>	Stanley Hall pp. 19
11:30–12:00pm	Dimitris Achlioptas, UC Santa Cruz <i>Stochastic Integration via Error-Correcting Codes</i>	pp. 7
12:00–2:00pm	<i>Lunch (on your own)</i>	
Third Session		
2:00–2:30pm	Charles Martin, Calculation Consulting <i>Why Deep Learning Works: Perspectives from Theoretical Chemistry</i>	Stanley Hall pp. 9
2:30–3:00pm	Surya Ganguli, Stanford <i>A theory of multineuronal dimensionality, dynamics and measurement</i>	pp. 8
3:00–3:30pm	Fred Roosta, ICSI and UC Berkeley <i>Sub-sampled Newton Methods: Uniform and Non-Uniform Sampling</i>	pp. 11
3:30–4:00pm	<i>Coffee break</i>	
Fourth Session		
4:00–4:30pm	Sebastiano Vigna, Universita degli Studi di Milano <i>In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond</i>	Stanley Hall pp. 13
4:30–5:00pm	David Gleich, Purdue University <i>Higher-order clustering of networks</i>	pp. 9
5:00–5:30pm	Charalampos Tsourakakis, Harvard University <i>Mining Tools for Large-Scale Networks</i>	pp. 13
5:30–6:00pm	Jimeng Sun, Georgia tech <i>Building Scalable Predictive Modeling Platform for Healthcare Applications</i>	pp. 12
Evening Reception		
6:00–8:00pm	<i>Dinner Reception and Poster Session</i>	Lobby outside Stanley Hall

Friday, June 24, 2016: Novel Matrix and Graph Methods

Time	Event	Location/Page
First Session		Stanley Hall
9:00–10:00am	Christopher White, Microsoft <i>Scalable interaction with data: where artificial intelligence meets visualization</i>	pp. 14
10:00–10:30am	Christopher Re, Stanford <i>Ameliorating the Annotation Bottleneck</i>	pp. 11
10:30–11:00am	<i>Coffee break</i>	
Second Session		Stanley Hall
11:00–11:30am	Bryan Graham, UC Berkeley <i>Homophily and transitivity in dynamic network formation</i>	pp. 9
11:30–12:00pm	Mark Flood, Office of Financial Research <i>Systemwide Commonalities in Market Liquidity</i>	pp. 8
12:00–2:00pm	<i>Lunch (on your own)</i>	
Third Session		Stanley Hall
2:00–2:30pm	Moritz Hardt, Google Research <i>Train faster, generalize better: Stability of stochastic gradient descent</i>	pp. 9
2:30–3:00pm	Rachel Ward, University of Texas at Austin <i>Extracting governing equations from highly corrupted data</i>	pp. 13
3:00–3:30pm	Cosma Shalizi, Carnegie Mellon University <i>Nonparametric Network Smoothing</i>	pp. 11
3:30–4:00pm	<i>Coffee break</i>	
Fourth Session		Stanley Hall
4:00–4:30pm	Joakim Anden and Amit Singer, Princeton University <i>PCA from noisy linearly reduced measurements</i>	pp. 12
4:30–5:00pm	Robert Anderson, UC Berkeley <i>PCA with Model Misspecification</i>	pp. 7
5:00–5:30pm	Nesreen Ahmed and Ted Willke, Intel Labs <i>Fast Graphlet Decomposition</i>	pp. 14

Poster Presentations: Thursday, June 23, 2016

Event	Location/Page
Poster Session	Lobby outside Stanley Hall
John Arabadjis, State Street - GX Labs <i>A Data-Driven Approach to Multi-Asset Class Portfolio Simulations with Latent-Factor-Based Dimensionality Reduction</i>	pp. 15
Daniel Ahfock, University of Cambridge <i>A statistical perspective on sketched regression</i>	pp. 15
Ian Bolliger, UC Berkeley <i>Capturing spatiotemporal variability in the influence of topography and vegetation on snow depth in the Tuolumne River Basin</i>	pp. 15
Grigor Aslanyan, UC Berkeley <i>Cosmo 4D: Towards the beginning of the Universe</i>	pp. 15
Guangliang Chen, San Jose State University <i>Novel Machine Learning Techniques for Fast, Accurate Parameter Selection in Gaussian-kernel SVM</i>	pp. 15
Sebastien Dery, McGill University <i>Web-Scale Distributed Community Detection using GraphX</i>	pp. 16
Derek Driggs, University of Colorado Boulder <i>Parallelization of Stable Principal Component Pursuit</i>	pp. 16
N. Benjamin Erichson, University of Washington <i>Compressed Dynamic Mode Decomposition</i>	pp. 16
Jovile Grebliauskaite, Sqrrl Data, Inc. <i>Latent Behavior Analysis of Large Amounts of Network Security Data</i>	pp. 16
Aditya Grover, Stanford University <i>node2vec: Scalable Feature Learning for Networks</i>	pp. 16
Deborah Hanus, Harvard University <i>Inferring missing data and accounting for patient variation to predict effective HIV treatments</i>	pp. 17
Amin Jalali, University of Washington <i>Variational Gram Functions: Convex Analysis and Optimization</i>	pp. 17
Qingkai Kong, Berkeley Seismological Lab <i>MyShake - Smartphone crowdsourcing for earthquakes</i>	pp. 17
Vivek Kulkarni, Stony Brook University <i>Freshman or Fresher? Quantifying the Geographic Variation of Internet Language</i>	pp. 17
Aubrey Laskowski, University of Illinois at Urbana-Champaign <i>Algorithms for Computing Elements in a Free Distributive Lattice</i>	pp. 18
Doris Jung Lin Lee, UC Berkeley, and Robert J. Brunner, UIUC <i>Pattern Discovery and Large-Scale Data mining on cosmological datasets</i>	pp. 18
Zhen Li, Tsinghua University, Beijing, PRC. <i>Point Integral Method for PDEs on Point Clouds</i>	pp. 18
Shuyang Ling, UC Davis <i>Rapid, Robust, and Reliable Blind Deconvolution via Nonconvex Optimization</i>	pp. 18

Poster Presentations, continued.

Event	Location/Page
Poster Session	Lobby outside Stanley Hall
Miles Lopes, UC Davis <i>Compressed Sensing without Sparsity Assumptions</i>	pp. 18
Emaad Ahmed Manzoor, Stony Brook University <i>Streaming Pairwise Document Similarity by Shingling, Sketching and Hashing</i>	pp. 19
Chirag Modi, University of California, Berkeley <i>Analytic Derivatives of High Dimensional Forward Models in Cosmology</i>	pp. 19
Abbas Ourmazd, Univ. of Wisconsin Milwaukee <i>Structure and Dynamics from Random Observations</i>	pp. 19
Vu Pham, UC Berkeley <i>Robust sketching for multiple square-root LASSO problems</i>	pp. 19
Mert Pilanci, UC Berkeley <i>Fast Randomized Algorithms for Convex Optimization</i>	pp. 19
Pooja Rajkumar, UC Davis <i>Rectools: A recommendation engine package</i>	pp. 20
Sebastian Rodriguez, University of California, Merced <i>Using Play-by-Play Data to Model, Simulate, and Predict NBA Games</i>	pp. 20
Ramyar Saeedi, Washington State University <i>A Transfer Learning Approach for Autonomous Reconfiguration of Wearable Systems</i>	pp. 20
Divya Sardana, University of Cincinnati <i>Core periphery structures to analyse a spatio-temporal dataset of crimes in San Francisco</i>	pp. 20
A. Erdem Sariyuce, Sandia National Labs <i>Fast Hierarchy Construction for Dense Subgraphs</i>	pp. 21
Xiaofeng Shao, University of Illinois at Urbana-Champaign <i>A Subsampled Double Bootstrap for Massive Data</i>	pp. 21
Shaden Smith, University of Minnesota <i>SPLATT: Enabling Large-Scale Sparse Tensor Analysis</i>	pp. 21
Veronika Strnadova-Neeley, UC Santa Barbara, Lawrence Berkeley National Lab <i>A New Similarity Score for Large-Scale, Sparse, and Discrete-Valued Data</i>	pp. 21
Javier Turek, Parallel Computing Lab - Intel <i>Enabling Brain Functional Alignment for a Thousand Subjects</i>	pp. 22
Peng Xu and Jiyan Yang, Stanford University <i>Sub-sampled Newton Methods with Non-uniform Sampling</i>	pp. 22

Talk Abstracts

Stochastic Integration via Error-Correcting Codes

Dimitris Achlioptas, UC Santa Cruz

Several problems in applied mathematics and statistics require integrating a function f over a high-dimensional domain. For example, estimating the partition function of a graphical model for a fixed set of parameters requires integrating (summing) its unnormalized probability function f over all possible configurations (value assignments). The most common methods for performing such integration stochastically involve Markov Chains (MCMC).

We present an alternative to MCMC based on ideas from the theory of modern error-correcting codes. Specifically, we will see that stochastically integrating a non-negative function f over its entire domain can be achieved at the cost of maximizing f over suitable random subsets of its domain. The key lies in choosing these random subsets so that besides conferring good statistical properties, they also do not dramatically increase the difficulty of maximization. Using real-life satisfiability formulas as benchmarks, we see that selecting as subsets the codewords of Low Density Parity Check codes yields dramatic speedup and levels of accuracy that were previously unattainable.

Cooperative Computing for Autonomous Data Centers Storing Social Network Data

Jonathan Berry, Sandia National Laboratories

We consider graph datasets that are distributed among several data centers with constrained sharing arrangements. We propose a formal model in which to design and analyze algorithms for this context along with two representative algorithms: s-t connectivity and planted clique discovery. The latter forced us to rethink recent conventional wisdom from social science regarding the clustering coefficient distributions of social networks. I will describe our process of cleaning social networks of human-implausible network structure in order to test our algorithms.

PCA with Model Misspecification

Robert Anderson, UC Berkeley

The theoretical justifications for Principal Component Analysis (PCA) typically assume that the data is IID over the estimation window. In practice, this assumption is routinely violated in financial data. We examine the extent to which PCA-like procedures can be justified in the presence of two specific kinds of misspecification present in financial data: time-varying volatility, and the presence of regimes in factor loadings.

The Union of Intersections Method

Kristofer Bouchard, Lawrence Berkeley National Laboratory

The increasing size and complexity of biomedical data could dramatically enhance basic discovery and prediction for applications. Realizing this potential requires analytics that are simultaneously selective, accurate, predictive, stable, and scalable. However, current methods do not generally achieve this. Here, we introduce the Union of Intersections method, a novel, modular paradigm for recovering accurate, interpretable and predictive parameters in regression and classification. We briefly summarize new theoretical results proving superior performance of our method under less restrictive conditions than previous methods, conclusions supported by extensive simulations. We further demonstrate: increased prediction accuracy on a variety of benchmark biomedical data; extraction of meaningful functional networks from human electrophysiology recordings; and dramatically more parsimonious prediction of behavioral and physiological phenotypes from genetic data. As our methods are broadly applicable across domains, we provide several open-source implementations to improve data prediction and discovery across science and medical fields.

New Results in Non-Convex Optimization for Large Scale Machine Learning

Constantine Caramanis, UT Austin

The last few years has seen a flurry of activity in non-convex approaches to enable solution of large scale optimization problems that come up in machine learning. The common thread in many of these results is that low-rank matrix optimization or recovery can be accomplished while forcing the low-rank factorization and then solving the resulting factored (non-convex) optimization problem. We consider two important settings, and present new results in each: dealing with projections – and important and generic requirement for convex optimization – and dealing with robustness (corrupted points) – a topic in robust high dimensional statistics that has received much attention in theory and applications.

Restricted Strong Convexity Implies Weak Submodularity

Alex Dimakis, UT Austin

We connect high-dimensional subset selection and submodular maximization. Our results extend the work of Das and Kempe (2011) from the setting of linear regression to arbitrary objective functions. This connection allows us to obtain strong multiplicative performance bounds on several greedy feature selection methods without statistical modeling assumptions. This is in contrast to prior work that requires data generating models to obtain theoretical guarantees. Our work shows that greedy algorithms perform within a constant factor from the best possible subset-selection solution for a broad class of general objective functions. Our methods allow a direct control over the number of obtained

features as opposed to regularization parameters that only implicitly control sparsity.

Systemwide Commonalities in Market Liquidity

Mark Flood, Office of Financial Research

We explore statistical commonalities among granular measures of market liquidity with the goal of illuminating systemwide patterns in aggregate liquidity. We calculate daily invariant price impacts described by Kyle and Obizhaeva (2014) to assemble a granular panel of liquidity measures for equity, corporate bond, and futures markets. We estimate Bayesian models of hidden Markov chains and use Markov chain Monte Carlo analysis to measure the latent structure governing liquidity at the systemwide level. Three latent liquidity regimes — high, medium, and low price-impact — are adequate to describe each of the markets. Focusing on the equities subpanel, we test whether a collection of systemwide market summary time series can recover the estimated liquidity dynamics. This allows an economically meaningful attribution of the latent liquidity states and yields meaningful predictions of liquidity disruptions as far as 15 trading days in advance of the 2008 financial crisis.

Local graph clustering algorithms: an optimization perspective

Kimon Fountoulakis, University of California Berkeley

Locally-biased graph algorithms are algorithms that attempt to find local or small-scale structure in a typically large data graph. In some cases, this can be accomplished by adding some sort of locality constraint and calling a traditional graph algorithm; but more interesting are locally-biased graph algorithms that compute answers by running a procedure that does not even look at most of the graph. This corresponds more closely with what practitioners from various data science domains do, but it does not correspond well with the way that algorithmic and statistical theory is typically formulated. Recent work from several research communities has focused on developing locally-biased graph algorithms that come with strong complementary algorithmic and statistical theory and that are useful in practice in downstream data science applications. We provide a review and overview of this work, highlighting commonalities between seemingly-different approaches, and highlighting promising directions for future work.

A theory of multineuronal dimensionality, dynamics and measurement

Surya Ganguli, Stanford

While technological revolutions in neuroscience now enable us to record from ever increasing numbers of neurons, the number we will be able to record in the foreseeable future remains an infinitesimal fraction of the total number of neurons in mammalian circuits controlling complex behaviors.

Nevertheless, despite operating within this extreme under-sampling limit, a wide array of statistical procedures for dimensionality reduction of multineuronal recordings uncover remarkably insightful, low dimensional neural state space dynamics whose geometry reveals how behavior and cognition emerge from neural circuits. What theoretical principles explain this remarkable success; in essence, how is it that we can understand anything about the brain while recording an infinitesimal fraction of its degrees of freedom?

We develop an experimentally testable theoretical framework to answer this question. By making a novel conceptual connection between neural measurement and the theory of random projections, we derive scaling laws relating how many neurons we must record to accurately recover state space dynamics, given the complexity of the behavioral or cognitive task, and the smoothness of neural dynamics. Moreover we verify these scaling laws in the motor cortical dynamics of monkeys performing a reaching task.

Along the way, we derive new upper bounds on the number of random projections required to preserve the geometry of smooth random manifolds to a given level of accuracy. Our methods combine probability theory with Riemannian geometry to improve upon previously described upper bounds by two orders of magnitude.

Scalable Collective Inference from Richly Structured Data

Lise Getoor, UC Santa Cruz

In this talk, I will introduce hinge-loss Markov random fields (HL-MRFs), a new kind of probabilistic graphical model that supports scalable collective inference from richly structured data. HL-MRFs unify three different approaches to convex inference: LP approximations for randomized algorithms, local relaxations for probabilistic graphical models, and inference in soft logic. I will show that all three lead to the same inference objective. HL-MRFs typically have richly connected yet sparse dependency structures, and I will describe an inference algorithm that exploits the fine-grained dependency structures and is much more scalable than general-purpose convex optimization approaches. Along the way, I will describe probabilistic soft logic, a declarative language for defining HL-MRFs.

Low-rank matrix factorizations at scale: Spark for scientific data analytics

Alex Gittens, ICSI and UC Berkeley

We explore the trade-offs of performing linear algebra in Apache Spark versus the traditional C and MPI approach by examining three widely-used matrix factorizations: NMF (for physical plausibility), PCA (for its ubiquity), and CX (for model interpretability). We apply these methods to TB-scale problems in particle physics, climate modeling, and bioimaging using algorithms that map nicely onto Spark's data-parallelism model. We perform scaling experiments on

up to 1600 Cray XC40 nodes, describe the sources of slow-downs, and provide tuning guidance to obtain high performance.

Higher-order clustering of networks

David Gleich, Purdue University

Spectral clustering is a well-known way to partition a graph or network into clusters or communities with provable guarantees on the quality of the clusters. This guarantee is known as the Cheeger inequality and it holds for undirected graphs. We discuss a new generalization of the Cheeger inequality to higher-order structures in networks including network motifs. This is easy to implement and seamlessly generalizes spectral clustering to directed, signed, and many other types of complex networks. In particular, our generalization allows us to re-use the large history of existing ideas in spectral clustering including local methods, overlapping methods, and relationships with kernel k-means. We illustrate the types of clusters or communities found by our new method in biological, neuroscience, ecological, transportation, and social networks. This is joint work with Austin Benson and Jure Leskovec at Stanford.

Identifying Broad and Narrow Financial Risk Factors with Convex Optimization: Part 1

Lisa Goldberg, UC Berkeley

While the use of statistical methods to identify financial risk factors is a long-standing practice, the use of convex optimization for this purpose is a recent innovation. Specifically, a combination of convex programs developed by Chandrasekaran, Parillo and Witsky (2012) and Saunderson et al. (2012) can be used to extract financial risk factors from a sample return covariance matrix. I will outline an application of this extraction to financial risk forecasting and develop finance-oriented metrics to assess accuracy. Finally, I provide an example that highlights the difference between this approach and principal component analysis, which is the academic standard for risk factor identification. In a companion talk, Alex Shkolnik will discuss the convergence properties of the convex programs, and he will analyze the accuracy of the algorithm on simulated data.

Homophily and transitivity in dynamic network formation

Bryan Graham, UC Berkeley

In social and economic networks linked agents often share additional links in common. There are two competing explanations for this phenomenon. First, agents may have a structural taste for transitive links – the returns to linking may be higher if two agents share links in common. Second, agents may assortatively match on unobserved attributes, a process called homophily. I study parameter identifiability in a simple model of dynamic network formation with both effects. Agents form, maintain, and sever links over time

in order to maximize utility. The return to linking may be higher if agents share friends in common. A pair-specific utility component allows for arbitrary homophily on time-invariant agent attributes. I derive conditions under which it is possible to detect the presence of a taste for transitivity in the presence of assortative matching on unobservables. I leave the joint distribution of the initial network and the pair-specific utility component, a very high dimensional object, unrestricted. The analysis is of the ‘fixed effects’ type. The identification result is constructive, suggesting an analog estimator, whose single large network properties I characterize.

Train faster, generalize better: Stability of stochastic gradient descent

Moritz Hardt, Google Research

We show that any model trained by a stochastic gradient method with few iterations has vanishing generalization error. We prove this by showing the method is algorithmically stable in the sense of Bousquet and Elisseeff. Our analysis only employs elementary tools from convex and continuous optimization. Our results apply to both convex and non-convex optimization under standard Lipschitz and smoothness assumptions. Applying our results to the convex case, we provide new explanations for why multiple epochs of stochastic gradient descent generalize well in practice. In the nonconvex case, we provide a new interpretation of common practices in neural networks, and provide a formal rationale for stability-promoting mechanisms in training large, deep models. Conceptually, our findings underscore the importance of reducing training time beyond its obvious benefit.

Why Deep Learning Works: Perspectives from Theoretical Chemistry

Charles Martin, Calculation Consulting

We present new ideas which attempt to explain why Deep Learning works, taking lessons from Theoretical Chemistry, and integrating ideas from Protein Folding, Renormalization Group, and Quantum Chemistry.

We address the idea that spin glasses make good models for Deep Learning, and discuss both the p-spherical spin glass models used by LeCun, and the spin-glass-of-minimal frustration, proposed by Wolynes for protein folding some 20 years ago.

We argue that Deep Learning energy models resemble the energy models developed for protein folding, and, in contrast to the p-spin spherical models, suggest the energy landscape of a deep learning model should be ruggedly convex. We compare and contrast this to hypothesis to current suggestions as to why Deep Learning works.

We show the relationship between RBMs and Variational Renormalization Group, and explain the importance in modeling neuro-dynamics. We then discuss how the RG transform can be used as a path to construct an Effective Hamiltonian for Deep Learning that would help illuminate why these models work so well.

Is manifold learning for toy data only?

Marina Meila, University of Washington

Manifold learning algorithms aim to recover the underlying low-dimensional parametrization of the data using either local or global features. It is however widely recognized that the low dimensional parametrizations will typically distort the geometric properties of the original data, like distances and angles. These unpredictable and algorithm dependent distortions make it unsafe to pipeline the output of a manifold learning algorithm into other data analysis algorithms, limiting the use of these techniques in engineering and the sciences.

Moreover, accurate manifold learning typically requires very large sample sizes, yet existing implementations are not scalable, which has led to the commonly held belief that manifold learning algorithms aren't practical for real data.

This talk will show how both limitations can be overcome. I will present a statistically founded methodology to estimate and then cancel out the distortions introduced by any embedding algorithm, thus effectively preserving the distances in the original data. And I will demonstrate that with careful use of sparse data structures manifold learning can scale to data sets in the millions. Both points will be exemplified in the exploration of data from a large astronomical survey.

Joint work with Dominique Perrault-Joncas, James McQueen, Jacob VanderPlas, Zhongyue Zhang, Grace Telford

Randomized Composable Core-sets for Distributed Computation

Vahab Mirrokni, Google Research

An effective technique for solving optimization problems over massive data sets is to partition the data into smaller pieces, solve the problem on each piece and compute a representative solution from it, and finally obtain a solution inside the union of the representative solutions for all pieces. Such an algorithm can be implemented easily in 2 rounds of MapReduces or be applied in an streaming model. This technique can be captured via the concept of *composable core-sets*, and has been recently applied to solve diversity maximization problems as well as several clustering problems. However, for coverage and submodular maximization problems, impossibility bounds are known for this technique. In this talk, after a initial discussion about this technique and applications in diversity maximization and clustering problems, we focus on the submodular maximization problem, and show how to apply a randomized variant of composable core-set problem, and achieve 1/3-approximation for monotone and non-monotone submodular maximization problems. We prove this result by applying a simple greedy algorithm and show that a large class of algorithms can be deployed in this framework. Time-permitting, we show a more complicated algorithm that achieves 54% of the optimum in two rounds of MapReduces.

The main part of the talk is to appear in STOC 2015 and is a joint work with Morteza ZadiMoghaddam. The initial

parts are from two recent papers that appeared in PODS 2014 and NIPS 2014. Time-permitting, I will talk about recent results about applying sketching techniques to improve the state-of-the-art for coverage problems.

Randomized Low-Rank Approximation and PCA: Beyond Sketching

Cameron Musco, Massachusetts Institute of Technology

I will discuss recent work on randomized algorithms for low-rank approximation and principal component analysis (PCA). The talk will focus on efforts that move beyond the extremely fast, but relatively crude approximations offered by random sketching algorithms.

In particular, we will see how advances in Johnson-Lindenstrauss projection methods have provided tools for improving the analysis of classic iterative SVD algorithms, including the block power method and block Krylov methods. The key insight is to view the iterative algorithms as denoising procedures for coarse sketching methods.

I will discuss how this view can be used to analyze a simple block Krylov method, showing that the algorithm gives $(1+\epsilon)$ near optimal PCA and low-rank approximation in just $O(1/\sqrt{\epsilon})$ iterations. Despite their long history, this analysis is the first of a Krylov subspace method that does not depend on the matrix's spectral gaps.

I will also survey open questions in the analysis of iterative methods, promising work on approximate PCA via stochastic optimization, fast sampling methods for low-rank kernel matrix approximation, and faster techniques for singular value decomposition targeted at specific downstream tasks, such as principal component regression.

Learning about business cycle conditions from four terabytes of data

Serena Ng, Columbia University

This paper sets out to extract any business cycle information that might exist in four terabytes of weekly scanner data. This dataset has two unique features: the Great Recession of 2008 is included in the sample, and observations for both price and quantity.

The main challenge is to handle the volume, variety, and characteristics of the data within the constraints of our computing environment. While the available subsampling algorithms are computationally efficient, the sampling scheme may not be desirable from the viewpoint of economic analysis. This data also require the researcher to perform seasonal filtering, a task that is usually undertaken by government agents. The problem requires a particular type of low rank decomposition. In short, economic data have specific needs and require specific tools. Cross-disciplinary work is needed to develop improved alternatives.

Top 10 Data Analytics Problems in Science

Prabhat, Lawrence Berkeley National Laboratory

Lawrence Berkeley National Lab and NERSC are at the frontier of scientific research. Historically, NERSC has provided leadership computing for the computational science community, but we now find ourselves tackling Big Data problems from an array of observational and experimental sources. In this talk, I will review the landscape of Scientific Big Data problems at all scales, spanning astronomy, cosmology, climate, neuroscience, bioimaging, genomics, material science and subatomic physics. I will present a list of Top 10 Data Analytics problems from these domains, and highlight NERSC's current Data Analytics strategy and hardware/software resources. I will highlight opportunities for engaging with NERSC, Berkeley Lab and the scientific enterprise.

Ameliorating the Annotation Bottleneck

Christopher Re, Stanford

The rise of automatic feature-generation techniques, including deep learning, has the potential to greatly enlarge the pool of machine-learning users. Such methods require large labeled training sets to obtain high-quality results. This raises two related questions: First, how does one scale deep-learning systems? And second, how can one make it easier for users to build training sets? We describe some very recent work on these questions. Our contribution for the first question is a recent result that characterizes asynchronous learning as equivalent to changing the momentum term. Importantly, this result does not depend on convexity and, so, applies to deep learning. For the second question, we describe a new paradigm called Data Programming that enables users to programmatically cheaply generate large but noisy training sets. Nonconvex analysis techniques then allow us to model and denoise these noisy training data sets. We also report on how nonexperts are able to obtain high-quality end-to-end performance using our prototype information extraction framework, DDlite, that implements these ideas.

All projects available from <https://github.com/HazyResearch>

Sub-sampled Newton Methods: Uniform and Non-Uniform Sampling

Fred Roosta, ICSI and UC Berkeley

Many data analysis applications require the solution of optimization problems involving a sum of large number of functions. We consider the problem of minimizing a sum of n functions over a convex constraint set. Algorithms that carefully sub-sample to reduce n can improve the computational efficiency, while maintaining the original convergence properties. For second order methods, we first consider a general class of problems and give quantitative convergence results for variants of Newton's methods where the Hessian or the gradient is uniformly sub-sampled. We then show that, given certain assumptions, we can extend our analysis and apply non-uniform sampling which results in modified algorithms

exhibiting more robustness and better dependence on problem specific quantities, such as the condition number.

New Methods for Designing and Analyzing Large Scale Randomized Experiment

Jasjeet Sekhon, UC Berkeley

The rise of massive datasets that provide fine-grained information about human beings and their behavior provides unprecedented opportunities for evaluating the effectiveness of social, behavioral, and medical treatments. We have also become more interested in fine-grained inferences. Researchers and policy makers are increasingly unsatisfied with estimates of average treatment effects based on experimental samples that are unrepresentative of populations of interest. Instead, they seek to target treatments to particular populations and subgroups. These fine-grained inferences lead to small data problems: subgroups where the dimensionality of data is high but the number of observations is small. To make the best use of these new methods, randomized trials should be constructed differently, with an eye towards how they will be combined with observational data down the road. I discuss new methods for designing and analyzing randomized experiments that make the most of these opportunities. For example, inferences from randomized experiments can be improved by blocking: assigning treatment in fixed proportions within groups of similar units. However, the use of the method is limited by the difficulty in deriving these groups. Current blocking methods are restricted to special cases or run in exponential time; are not sensitive to clustering of data points; and are often heuristic, providing an unsatisfactory solution in many common instances. We present an algorithm that implements a new, widely applicable class of blocking—threshold blocking—that solves these problems. Given a minimum required group size and a distance metric, we study the blocking problem of minimizing the maximum distance between any two units within the same group. We prove this is a NP-hard problem and derive an approximation algorithm that yields a blocking where the maximum distance is guaranteed to be at most four times the optimal value. This algorithm runs in $O(n \log n)$ time with $O(n)$ space complexity. This makes it the first blocking method with an ensured level of performance that works in massive experiments. While many commonly used algorithms form pairs of units, our algorithm constructs the groups flexibly for any chosen minimum size. This facilitates complex experiments with several treatment arms and clustered data. I also discuss extensions of this method for observational data (e.g., matching) and for exploratory data analysis (e.g., clustering).

Nonparametric Network Smoothing

Cosma Shalizi, Carnegie Mellon University

Modern data analysis makes great use of non-parametric smoothing for estimation and of resampling to assess uncertainty. In this talk, I will describe two methods for non-parametric estimation of stochastic network models, one

based on density estimation in negatively curved spaces, the other based on treating adjacency matrices as spatial data and smoothing them. Both techniques will work as model-based bootstraps, and both can be used to assess the statistical significance of differences between networks.

Identifying Broad and Narrow Financial Risk Factors with Convex Optimization: Part 2

Alex Shkolnik, UC Berkeley

A combination of convex programs developed by Chandrasekaran, Parillo and Witsky (2012) and Saunderson et al. (2012) can be used to extract financial risk factors from a sample return covariance matrix. I will examine the convergence properties of the convex programs and look at their performance on simulated and empirical data. Using the finance-oriented metrics developed in Lisa Goldbergs companion talk, I analyze the accuracy of the algorithm on simulated data. The results point to modifications that may lead to improved performance.

A Framework for Processing Large Graphs in Shared Memory

Julian Shun, UC Berkeley

In this talk, I will discuss Ligra, a shared-memory graph processing framework that has two very simple routines, one for mapping over edges and one for mapping over vertices. The routines can be applied to any subset of the vertices and automatically adapt to their density, which makes the framework useful for many graph traversal algorithms that operate on subsets of the vertices. Ligra is able to express a broad class of graph algorithms including breadth-first search, betweenness centrality, eccentricity estimation, connectivity, PageRank, single-source shortest paths, and local clustering algorithms. I will describe implementations of parallel algorithms in Ligra and present performance results. I will also discuss Ligra+, an extension of Ligra that uses graph compression to reduce space usage and improve parallel performance.

PCA from noisy linearly reduced measurements

Joakim Anden and Amit Singer, Princeton University

We consider the problem of estimating the covariance of X from measurements of the form $y_i = A_i x_i + \varepsilon_i$ (for $i = 1, \dots, n$) where x_i are i.i.d unobserved samples of X , A_i are given linear operators, and ε_i represent noise. Our estimator is constructed efficiently via a simple linear inversion using conjugate gradient performed after eigenvalue shrinkage motivated by the spike model in high dimensional PCA. Applications to 2D image denoising and 3D structure classification in single particle cryo-EM will be discussed.

Minimax optimal subsampling for large sample linear regression

Aarti Singh, Carnegie Mellon University

We investigate statistical aspects of subsampling for large-scale linear regression under label budget constraints. In many applications, we have access to large datasets (such as healthcare records, database of building profiles, and visual stimuli), but the corresponding labels (such as customer satisfaction, energy usage, and brain response, respectively) are hard to obtain. We derive computationally feasible and near minimax optimal subsampling strategies for both with and without replacement settings for prediction as well as estimation of the regression coefficients. Experiments on both synthetic and real-world data confirm the effectiveness of our subsampling algorithm for small label budgets, in comparison to popular competitors such as uniform sampling, leverage score sampling and greedy methods.

Head, Torso and Tail - Performance for modeling real data

Alex Smola, Carnegie Mellon University

Real data is high dimensional and multi variate. A naive application of optimization techniques leads to rather poor performance, due to large memory footprint, latency and network cost. In this talk I will address how to overcome these constraints by design that takes advantage of distributional properties of real data for recommendation, classification and modeling.

Building Scalable Predictive Modeling Platform for Healthcare Applications

Jimeng Sun, Georgia tech

As the adoption of electronic health records (EHRs) has grown, EHRs are now composed of a diverse array of data, including structured information and unstructured clinical progress notes. Two unique challenges need to be addressed in order to utilize EHR data in clinical research and practice: 1) Computational phenotyping: How to turn complex and messy EHR data into meaningful clinical concepts or phenotypes? 2) Predictive modeling: How to develop accurate predictive models using longitudinal EHR data? To address these challenges, I will present our approaches using a case study on early detection for heart failure. For computational phenotyping, we present EHR data as data as inter-connected high-order relations i.e. tensors (e.g. tuples of patient-medication-diagnosis, patient-lab, and patient-symptoms), and then develop expert-guided sparse nonnegative tensor factorization for extracting multiple phenotype candidates from EHR data. Most of the phenotype candidates are considered clinically meaningful and with great predictive power. For predictive modeling, I will present how using deep learning to model temporal relations among events in EHR improved model performance in predicting heart failure (HF) diagnosis compared to conventional methods that ignore temporality.

Mining Tools for Large-Scale Networks

Charalampos Tsourakakis, Harvard University

Finding large near-cliques in massive networks is a notoriously hard problem of great importance to many applications, including anomaly detection in security, community detection in social networks, and mining the Web graph. How can we exploit idiosyncrasies of real-world networks in order to solve this NP-hard problem efficiently? Can we find dense subgraphs in graph streams with a single pass over the stream? Can we design near real time algorithms for time-evolving networks? In this talk I will answer these questions in the affirmative. I will also present state-of-the-art exact and approximation algorithms for extraction of large near-cliques from large-scale networks, the k-clique densest subgraph problem, which run in a few seconds on a typical laptop. I will present graph mining applications, including anomaly detection in citation networks, and planning a successful cocktail party. I will conclude my talk with some interesting research directions.

Exploring Galaxy Evolution through Manifold Learning

Jake VanderPlas, University of Washington

The decade-long Sloan Digital Sky Survey produced detailed spectral observations of over a million distant galaxies: a dataset that is still yielding new insights years after its release. The data are very high-dimensional: observations of each galaxy include fluxes measured at nearly 4000 distinct wavelengths. Astronomers have long employed dimensionality reduction algorithms such as PCA to understand the structure of this dataset, but the nonlinear nature of the structure means such linear transformations miss important features, and nonlinear manifold learning approaches, though promising, have historically been too slow to be applicable on the entire dataset. Here I will report on some initial exploratory work with the new megaman package for scalable manifold learning (introduced in another talk by my colleague Marina Meila), particularly the approaches we have used to apply such algorithms to noisy and incomplete observations.

In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond

Sebastiano Vigna, Universit degli Studi di Milano,
Dipartimento di Informatica

We approach the problem of computing geometric centralities, such as closeness and harmonic centrality, on very large graphs; traditionally this task requires an all-pairs shortest-path computation in the exact case, or a number of breadth-first traversals for approximated computations, but these techniques yield very weak statistical guarantees on highly disconnected graphs. We rather assume that the graph is accessed in a semi-streaming fashion, that is, that adjacency lists are scanned almost sequentially, and that a very small amount of memory (in the order of a dozen bytes) per node is available in core memory. We leverage the newly discovered algorithms based on HyperLogLog counters, making it possible to approximate a number of geometric centralities

at a very high speed and with high accuracy. While the application of similar algorithms for the approximation of closeness was attempted in the MapReduce framework, our exploitation of HyperLogLog counters reduces exponentially the memory footprint, paving the way for in-core processing of networks with a hundred billion nodes using just 2TiB of RAM. Moreover, the computations we describe are inherently parallelizable, and scale linearly with the number of available cores.

Meaningful Visual Exploration of Massive Data

Peter Wang, Continuum Analytics

The size and complexity of modern datasets has far outstripped the capabilities of common existing methods for visualization. Current tools work well for small numbers of data points, where each point is represented by many pixels and is thus individually perceivable by the human visual system. However, when there are many more data points than pixels, it is crucial to accurately convey the *distribution* of points, i.e., the overall structure and pattern of the data, which typically emerges only indirectly via patterns of overplotting and summing of pixel values. Achieving a faithful visualization with scatterplots or heatmaps usually requires knowledge of the underlying distribution of datapoints, either from a priori expertise or through exploration, which presents serious conceptual and practical problems for understanding new, unknown large datasets.

In this talk, I will describe a new approach, called Datashading, that is a novel adaptation of the visualization pipeline which provides a principled way to visualize large and complex datasets. Datashading is based on the simple idea of using binning techniques to retain the original data values as far into the visualization pipeline as possible, even to the pixel level. As implemented in the new Python datashader library (<https://github.com/bokeh/datashader>), this approach allows algorithmic computations to replace trial-and-error approaches at each stage of processing.

I will demonstrate how these techniques can be used for flexible, interactive, and practical visualization of even extremely large datasets with billions of points.

Extracting governing equations from highly corrupted data

Rachel Ward, University of Texas at Austin

Learning the nonlinear governing equations $F(x(t)) = dx(t)/dt$ from a finite number of snapshots of the system $x(1), x(2), x(3), \dots, x(N)$ is of great interest across different scientific fields. When such data is moreover corrupted by outliers, for example, due to the recording mechanism failing over unknown intervals of time, the task of recovering $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ becomes even more challenging. In this talk, we consider the general setting where the governing equations are multivariate polynomials. Combining recent results on quantitative central limit theorems for maps of chaotic flows with results from compressive sensing, we show that

an ℓ_1 -minimization algorithm can exactly recover the governing equations from a finite number of snapshots, even if a near-constant fraction of the data is arbitrarily corrupted.

This is joint work with Giang Tran, UT Austin.

Scalable interaction with data: where artificial intelligence meets visualization

Christopher White, Microsoft

The number of professional jobs that require working across collections of documents and data continues to grow, from social media marketing through cyber security to health and wellness. Many of these new jobs contain underspecified problems and unknown metrics, unlike the metrics found in industries such as stock trading, online advertising, and sales forecasting. Additionally, the volume and characteristics of relevant data, including noise and lack of structure, create problems for query planning, assessing true quantities and distributions, understanding structure at several scales, and detecting key elements of a data-generating process. This presentation will cover several approaches to these problems, including using automated methods to infer representations, enabling scalable navigation through new visualizations, and building modular and disposable systems for a wide range of users at low financial cost. The talk will include motivation from previous experience at DARPA, detailed real-world examples, and software demonstrations. It will conclude by covering next directions for special projects at Microsoft.

Fast Graphlet Decomposition

Nesreen Ahmed and Ted Willke, Intel Labs

From social science to biology, numerous applications often rely on graphlets for intuitive and meaningful characterization of networks at both the global macro-level as well as the local micro-level. While graphlets have witnessed a tremendous success and impact in a variety of domains, there has yet to be a fast and efficient approach for computing the frequencies of these subgraph patterns. We proposed a fast, efficient, and parallel algorithm for counting k -graphlets. On a large collection of 300+ networks from a variety of domains, our graphlet counting strategies are on average 460x faster than the state-of-the-art methods. This talk will describe some of the largest graphlet computations to date on billion node graphs, as well as the largest systematic investigation on over 300+ networks from a variety of domains, such as social, biological, and technological graphs.

Fast, flexible, and interpretable regression modeling

Daniela Witten, University of Washington

In modern applications, we are interested in fitting regression models that are fast (i.e. can be fit to high-dimensional data), flexible (e.g. do not assume a linear conditional mean relationship between the response and the predictors), and interpretable (i.e. a person can make sense of the fitted model). I'll present two recent developments towards this end – the fused lasso additive model (FLAM), and convex regression via interpretable sharp partitions (CRISP).

Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data

Dacheng Xiu, Chicago Booth

This paper constructs an estimator for the number of common factors in a setting where both the sampling frequency and the number of variables increase. Empirically, we document that the covariance matrix of a large portfolio of US equities is well represented by a low rank common structure with sparse residual matrix. When employed for out-of-sample portfolio allocation, the proposed estimator largely outperforms the sample covariance estimator.

The Stability Principle for Information Extraction from Data

Bin Yu, UC Berkeley

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to “reasonable” perturbations to data and to the model used. Jackknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models. Moreover, the stability principle goes beyond data and model perturbations. In this talk, a case is made for the importance of the stability principle in modern statistics and data science. Stability is not only necessary for interpretable statistical models, but also at the heart of CLT (a cornerstone of classical statistics) and modern limiting results. We put the principle into use to derive two new methods: ESCV (estimation stability with cross-validation) for regularization parameter selection in regression and staNMF (stability-driven non-negative matrix factorization) for number of component selection. ESCV and staNMF's successes are then demonstrated with movie reconstruction and *Drosophila* embryonic image data, respectively.

Poster Abstracts

A Data-Driven Approach to Multi-Asset Class Portfolio Simulations with Latent-Factor-Based Dimensionality Reduction

John Arabadjis, State Street - GX Labs

Sudden portfolio losses as experienced by many financial institutions in recent years create downside losses materially impacting portfolio and business strategies. This results in highly undesirable consequences of large portfolio losses. Downside losses typically arise from excessive levels of concentration risk in a portfolio that appear too infrequently to be closely managed, but still pose risk. In addition, most risk-assessment approaches focus primarily on a portfolio's volatility, often ignoring the longer-term risks. In this poster we present a framework that utilizes a latent-factor model of market risk, a security-level simulation engine, and new open source technologies running on a massively parallel high performance computing environment, to quantify forward-looking portfolio risk across multiple risk metrics.

A statistical perspective on sketched regression

Daniel Ahfock, University of Cambridge

Random projections are a useful technique for reducing the computational complexity of routine data analysis computations. We focus on their application to ordinary least squares regression, a benchmark predictive technique that can be computationally expensive on massive datasets. Sketched regression algorithms use random projections to compress the original dataset to a manageable size. Sketched regression algorithms have a number of highly desirable algorithmic properties, particularly in terms of worst case performance. We make a new connection between sketched regression and statistical linear modelling. The linear modelling framework reveals new statistical properties of sketched regression algorithms. The statistical view complements the existing theoretical framework for sketched regression and leads to analytical results about the accuracy of the procedure.

Cosmo 4D: Towards the beginning of the Universe

Grigor Aslanyan, UC Berkeley

Our understanding of Cosmology, the theory of the Universe, has improved dramatically in the past two decades, thanks to the advances in Observational techniques. Modern galaxy surveys contain millions of galaxies, and the amount of data is going to continue increasing exponentially with next generations surveys. Traditional analysis methods reduce the data to its statistical properties, such as the two-point function, which is then used to constrain cosmological models and parameters. We propose a new method which aims at reconstructing the full initial density distribution of the Universe from the observational data. This is a computationally very difficult problem, since the dimensionality of the

parameter space is very large (millions to billions of degrees of freedom). In addition, the problem is highly nonlinear, involving gravitational collapse to form the structure in the Universe (stars, galaxies, clusters of galaxies). We present a new efficient method for reconstructing the initial conditions of the Universe from the observations, which combines the L-BFGS optimization algorithm and our fast code for non-linear structure formation. Our methodology also allows to estimate the full initial power spectrum and the Fisher matrix (which can be used to estimate error bars, for example).

Capturing spatiotemporal variability in the influence of topography and vegetation on snow depth in the Tuolumne River Basin

Ian Bolliger, UC Berkeley

Understanding spatial nonstationarity in the influence of topography and vegetation on patterns of snow water equivalent (SWE) can improve distributed SWE models, improve river runoff forecasts, and guide investigations of physical processes that generate this variability. In this study, we seek to understand the nature of this nonstationarity and improve upon a basin-scale statistical snow depth model by allowing for spatial patterns within observed relationships.

Novel Machine Learning Techniques for Fast, Accurate Parameter Selection in Gaussian-kernel SVM

Guangliang Chen, San Jose State University

Despite the popularity and superior performance of Gaussian-kernel support vector machine (SVM), the two hyperparameters σ (scale) and C (tradeoff) remain hard to be tuned. Many techniques have been developed to address this challenge such as grid search (full/random) and Bayesian methods. However, they all ignore the intrinsic geometry of training data as well as the interpretation of the two parameters and tackle the problem solely from an optimization point of view, thus being computationally expensive. A few methods such as the Jaakkola and Caputo heuristics try to learn the σ parameter directly from training data. Unfortunately, both of them require computing distances between points from different training classes and also sorting them which can be a great computational burden for large data sets.

In this poster we present two novel techniques for efficiently selecting σ and C , respectively: (1) For the σ parameter we point out that it corresponds to the local scale of the training classes and propose a fast nearest-neighbor approach for directly setting its value. (2) For the C parameter, we have observed in many cases that the validation accuracy peaks at some C or stabilizes after some C (as C is set to increase). This motivated us to propose a procedure that starts from the lower end of the C grid and detects

an elbow point on the validation accuracy curve (which is gradually computed). Such a choice of C leads to robust margins and avoids testing the bigger C values which tend to be computationally more expensive.

We have tested our combined algorithm against the existing approaches on a number of data sets and obtained very favorable results. In most cases the predictive accuracy of our method is at least comparable to the best of those methods, though our method uses only a single value of σ . In addition, our method is much faster than the other methods. Finally, our method is insensitive to the number k of the nearest neighbors used, and in general any k between 6 and 10 gives competitive results.

Web-Scale Distributed Community Detection using GraphX

Sebastien Dery, McGill University

From social networks to targeted advertising, graph analysis are increasingly considered a powerful approach for structuring and analyzing data. Indeed, connected entities are often seen to spontaneously form clusters of internally dense link, hereby termed community, yielding interesting avenues for machine learning such as collaborative filtering. Growth in data promise to better capture the key properties of these communities. Unfortunately, directly applying existing data-parallel tools to graph computation tasks can be cumbersome and inefficient. The need for intuitive, comprehensive and scalable tools for graph computation has lead to the recent development of GraphX. By combining the advantages of both data-parallel and graph-parallel systems within the Spark data-parallel framework, users have now accessed to powerful graph computation in already established infrastructure. In this work we present the implementation of a modularity optimization inside this framework. We further evaluate its usefulness for community detection by clustering researchers embedded within the coauthorship network of biomedical literature.

Parallelization of Stable Principal Component Pursuit

Derek Driggs, University of Colorado Boulder

Stable principal component pursuit (SPCP) seeks to decompose noisy signals into their underlying low-rank and sparse components. In this study, we introduce two new methods to parallelize SPCP in order to take advantage of GPU, multiple CPU, and hybridized architectures. Our first development is a parallelizable algorithm for the randomized singular value decomposition (rSVD) that ameliorates the problem of tall-skinny matrix structures that arise often in applications. Our second development is a reformulation of SPCP using Burer-Monteiro splitting, which eliminates the need for an SVD and is particularly well suited for the GPU. We demonstrate, using synthetic data, surveillance video, and data from fMRI brain scans, that both algorithms offer significant speedup over traditional SPCP solvers.

Compressed Dynamic Mode Decomposition

N. Benjamin Erichson, University of Washington

The dynamic mode decomposition (DMD) is a data-driven matrix decomposition for spatio-temporal data grids. Originally introduced in the fluid mechanics community, DMD has been applied to several new domains, including neuroscience, epidemiology and robotics. Specifically, DMD is a regression technique which integrates two of the leading data analysis methods in use today: Fourier transforms and principal components. Leveraging on matrix sketching, we present a compressed dynamic mode decomposition algorithm, which enables the decomposition of massive spatio-temporal data grids or streams. The method is illustrated on high-resolution video data for robustly performing foreground/background separation. Moreover, we present a GPU accelerated implementation, which significantly decreases the compute time.

Latent Behavior Analysis of Large Amounts of Network Security Data

Jovile Grebliauskaite, Sqrrl Data, Inc.

Detecting malicious behavior in network and endpoint logs is an extremely challenging task: large and complex data sets, highly dynamic innocuous behavior, and intelligent adversaries contribute to the difficulty. Even analyzing the subtle structures in the normal behavior of network entities is a challenging task.

We present a novel technique for learning latent behavioral structures in large amounts network cyber-security event data. Our technique is inspired by the DeepWalk [Perozzi, Al-Rfou, and Skiena 2014] approach and consists of three main steps. First we generate a property graph from network log information, and run random walks on the resulting graph. We then use random walks to create sentences in a synthetic language and use this language to create Word2Vec model. Our language contains a synthetic grammar of network entity nouns corresponding to graph nodes, behavioral verbs corresponding to typed graph edges, and log property adjectives. We then use k-means clustering on the generated Word2Vec model to find and analyse latent behavioral structures such as behavioral-based clusters of network users and identification of outlying behaviors. We have implemented our technique in Apache Spark and applied it to the public Los Alamos National Laboratory network cyber dataset containing login, network flow, and endpoint process data.

Joint work with Christopher McCubbin (Sqrrl Data, Inc.).

node2vec: Scalable Feature Learning for Networks

Aditya Grover, Stanford University

Prediction tasks over nodes and edges in networks require careful effort in engineering features for learning algorithms. Recent research in the broader field of representation learning has led to significant progress in automating prediction

by learning the features themselves. However, present approaches are largely insensitive to local patterns unique to networks. Here we propose node2vec, an algorithmic framework for learning feature representations for nodes in networks. In node2vec, we learn a mapping of nodes to a low-dimensional space of features that maximizes the likelihood of preserving distances between network neighborhoods of nodes. We define a flexible notion of nodes network neighborhood and design a biased random walk procedure, which efficiently explores diverse neighborhoods and leads to rich feature representations. Our algorithm generalizes prior work which is based on rigid notions of network neighborhoods and we demonstrate that the added flexibility in exploring neighborhoods is the key to learning richer representations. We demonstrate the efficacy of node2vec over existing state-of-the-art techniques on multi-label classification and link prediction in several real-world networks from diverse domains. Taken together, our work represents a new way for efficiently learning state-of-the-art task-independent node representations in complex networks.

Inferring missing data and accounting for patient variation to predict effective HIV treatments

Deborah Hanus, Harvard University

Human Immunodeficiency Virus (HIV) is a potential precursor to Acquired Immune Deficiency Syndrome (AIDS), which commonly treated with a mixture of reverse transcriptase inhibitors (RTI) and protease inhibitors (PI). Adams, et al (2004) used a simulator to dynamically model the patients health as a state space of patient health indicators that can be modified by actions, providing different types of treatment. Ernst, et al, (2006) found, by applying reinforcement learning to this model of states and actions, they could predict potentially effective HIV treatment policies. However, Ernst 2006 assumed that patients reliably present themselves for treatment every five days and that all patients respond to prescribed treatments similarly. These assumptions are unrealistic.

We investigate whether we can produce similarly effective treatment policies even when faced with sporadically sampled data, and we account for the physiological variation of individual patients. By imputing the patients missing personal health indicators, we find that we can often still compute effective treatment policies. We also show that parameter variation results in different optimal policies, a first step to motivate the need for transfer learning in this domain.

Variational Gram Functions: Convex Analysis and Optimization

Amin Jalali, University of Washington

We propose a new class of convex penalty functions, called variational Gram functions (VGFs), that can promote pairwise relations, such as orthogonality, among a set of vectors in a vector space. These functions can serve as regularizers in convex optimization problems arising from hierarchical classification, multitask learning, estimating vectors

with disjoint supports, and other applications. We study necessary and sufficient conditions under which a VGF is convex, and give a characterization of its subdifferential. In addition, we show how to compute its proximal operator, and discuss efficient optimization algorithms for some structured loss-minimization problems using VGFs. Numerical experiments are presented to demonstrate the effectiveness of VGFs and the associated optimization algorithms.

MyShake - Smartphone crowdsourcing for earthquakes

Qingkai Kong, Berkeley Seismological Lab

We are building a seismic network that harnesses the accelerometers in personal smartphones to record earthquake shaking data for research, hazard information and warnings. We developed an android application MyShake, that running on the phone that has the function to distinguish earthquake shakings from daily human activities based on the different patterns behind the movements. It has a built-in artificial neural network detection algorithm to distinguish earthquake signal from human activities. When MyShake is triggered by the earthquake-like movements, it sends the trigger information back to our server which contains time and location of the trigger, at the same time, it stores the waveform data on local phone first, and upload to our server later. After release the application in Feb, we currently have 60 Gb waveform data uploaded to us every day that need to be processed. We will show some initial development of the infrastructure and algorithms to process the data.

Freshman or Fresher? Quantifying the Geographic Variation of Internet Language

Vivek Kulkarni, Stony Brook University

We present a new computational technique to detect and analyze statistically significant geographic variation in language. Our meta-analysis approach captures statistical properties of word usage across geographical regions and uses statistical methods to identify significant changes specific to regions. While previous approaches have primarily focused on lexical variation between regions, our method identifies words that demonstrate semantic and syntactic variation as well. We extend recently developed techniques for neural language models to learn word representations which capture differing semantics across geographical regions. In order to quantify this variation and ensure robust detection of true regional differences, we formulate a null model to determine whether observed changes are statistically significant. Our method is the first such approach to explicitly account for random variation due to chance while detecting regional variation in word meaning. To validate our model, we study and analyze two different massive online data sets: millions of tweets from Twitter spanning not only four different countries but also fifty states, as well as millions of phrases contained in the Google Book Ngrams. Our analysis reveals interesting facets of language change at multiple scales of geographic resolution – from neighboring states to distant continents

Algorithms for Computing Elements in a Free Distributive Lattice

Aubrey Laskowski, University of Illinois at Urbana-Champaign

A free distributive lattice categorizes logically associated data and removes redundant connections, leaving only the significant nodes and relations. The number of generators of a free distributive lattice is the same as the number of independent factors, with the elements being the irredundant relationships between the generators. Dedekind numbers describe the number of elements in a free distributive lattice with n generators. These numbers grow rapidly; the largest known Dedekind number uses 8 variables and is on the order of magnitude of 10^{23} . In research originating in the NSF-funded Euler Project, Dr. Bertram Ludscher has compiled algorithms surrounding the Dedekind numbers to facilitate further research. Together we are employing the parallel computing capabilities of the National Center for Supercomputing Applications in pursuit of the Dedekind number with 9 variables. Using analogs between free distributive lattices and monotone Boolean functions, we will be using mathematical results which greatly reduce the runtimes to calculate Dedekind numbers. Through analysis of existing algorithms, we hope to find new directions and methods which would put the Dedekind number with 9 variables within reach.

Pattern Discovery and Large-Scale Data mining on cosmological datasets

Doris Jung Lin Lee, UC Berkeley, and Robert J. Brunner, UIUC

Pattern Discovery and Large-Scale Data mining on cosmological datasets

With next-generation telescopes capturing tens of TB/night of observational data, the role of scalable and efficient data analysis methods have become central to the knowledge discovery process in Astronomy. The diversity and scale of astronomical datasets also presents challenging research problems to the data mining and machine learning community. This poster describes three projects highlighting our recent work in these areas: 1) Current state-of-the-art ML algorithms (SVM, LDA, DNNs) are capable of classifying galaxy morphology at above 90% accuracy, but their results reflect the inherent errors due to human classifiers. We propose a scalable, hybrid technique that integrates active learning in crowdsourcing citizen science platforms for improving the data quality of the training labels. 2) We developed a recursive, source-finding algorithm that automatically corrects for positional inaccuracies in outdated astronomical catalogs. By applying this technique to imaging data from two different sky survey, we recovered all 23,011 sources in a widely used astronomical catalog. 3) Traditional friends-of-friends algorithms and density-estimation methods designed for halo-finding are not only computationally intensive, but especially problematic for detecting substructures within haloes. We explore non-parametric, unsupervised methods for finding haloes in the Dark Sky Simulation,

a 34TB N-body simulation containing trillions of particles.

Point Integral Method for PDEs on Point Clouds

Zhen Li, Tsinghua University, Beijing, PRC.

In many cases, massive data can be represented as unstructured point clouds in high dimensional spaces. Partial differential equation (PDE) provides a powerful tool to deal with such huge amount of data. We proposed a numerical method, the Point Integral Method (PIM), to solve PDEs over point clouds.

In order to use PIM, we do not need to generate meshes from point clouds. With this advantage, it can solve Laplace equation and its generalizations directly on point clouds in high dimensional spaces. PIM has been successfully applied to solve problems in computer vision and image processing. It is also a potential tool to attack certain problems in data science.

Rapid, Robust, and Reliable Blind Deconvolution via Nonconvex Optimization

Shuyang Ling, UC Davis

Suppose we are given the convolution of two signals, $y = f * g$. When, under which conditions, and how can we reconstruct f and g from the knowledge of y alone if both f and g are unknown? This challenging problem, known as *blind deconvolution*, pervades many areas of science and technology, including astronomy, medical imaging, optics, and wireless communications. A key challenge of this difficult non-convex optimization problem is that it exhibits many local minima. We present an efficient numerical algorithm that is guaranteed to recover the exact solution, when the number of measurements is (up to log-factors) slightly larger than the information-theoretical minimum, and under reasonable conditions on g and f . The proposed regularized gradient descent algorithm converges at a geometric rate and is provably robust in the presence of noise. To the best of our knowledge, our algorithm is the first blind deconvolution algorithm that is numerically efficient, robust against noise, and comes with rigorous recovery guarantees under certain subspace conditions. Moreover, numerical experiments do not only provide empirical verification of our theory, but they also demonstrate that our method yields excellent performance even in situations beyond our theoretical framework.

Compressed Sensing without Sparsity Assumptions

Miles Lopes, UC Davis

The theory of Compressed Sensing (CS) asserts that an unknown p -dimensional signal x can be accurately recovered from an underdetermined set of n linear measurements with $n \ll p$, provided that x is sufficiently sparse. However, in applications, the degree of sparsity $\|x\|_0$ is typically unknown, and the problem of directly estimating $\|x\|_0$ has been a longstanding gap between theory and practice. A closely related issue is that $\|x\|_0$ is a highly idealized measure of

sparsity, and for real signals with entries not equal to 0, the value $\|x\|_0 = p$ is not a useful description of compressibility. In previous work, we considered an alternative measure of “soft” sparsity, and designed a procedure to estimate this measure without relying on sparsity assumptions.

The present work offers a new deconvolution-based method for estimating unknown sparsity, which has wider applicability and sharper theoretical guarantees. In particular, we introduce an entropy-based sparsity measure that generalizes $\|x\|_0$. Also, we propose estimator for this measure whose relative error converges at the dimension-free rate of $1/\sqrt{n}$, even when p/n diverges. Our main results also describe the limiting distribution of the estimator, as well as some connections to Basis Pursuit Denoising, the Lasso, deterministic measurement matrices, and inference problems in CS.

Streaming Pairwise Document Similarity by Shingling, Sketching and Hashing

Emaad Ahmed Manzoor, Stony Brook University

We present a technique to maintain the pairwise cosine similarities of documents arriving as a stream of individual words. Words from different source documents interleave in the stream to either initiate new documents or grow existing ones, and may arrive out-of-order relative to their sequence in their respective source document. The technique conforms to a single-pass, constant memory data stream model and needs no prior information about the size of the incoming vocabulary. When applied to detect anomalous graphs arriving as a stream of edges derived from system call traces of executing processes, the technique was demonstrably performant (ca. 10,000 edges/second) and accurate (ca. 0.95 average precision) while consuming bounded memory (ca. 256MB) that is also user-configurable. Project website: <http://bit.ly/streamspot>

Analytic Derivatives of High Dimensional Forward Models in Cosmology

Chirag Modi, University of California, Berkeley

With the modern galaxy surveys containing millions of galaxies, we are in the era of Precision Cosmology and accurate theoretical predictions are necessary for efficient use of data. However, the structures observed today are highly nonlinear due to gravitational and fluid forces moving particles over billions of years. Hence an alternative approach is to study these structures in initial conditions where linear theory is valid. This boils down to the problem of constructing accurate estimates of initial conditions corresponding to present day observations. Due to huge parameter space of tens of millions at the very least, this optimization problem is computationally very hard. However, since we have accurate forward models for this non-linear evolution, we can in principle use optimization algorithms such as L-BFGS/conjugate gradients which are assisted with knowledge of analytic derivatives. In this study, we begin with the observation that perturbation theory at second order agrees well with full N-body evolution and thus use a hybrid model

wherein we use the exact N-body evolution as forward model and perturbation theory to calculate derivatives. We show that this is incompatible in general. We then present results using perturbative models at different orders as the forward models and show that this on the other hand gives accurate estimates of initial conditions for large scales.

Structure and Dynamics from Random Observations

Abbas Ourmazd, Univ. of Wisconsin Milwaukee

I will describe on-going efforts to extract structure and dynamics from noisy snapshots recorded at uncertain time points. Examples will include YouTube videos, the structure and conformations of molecular machines such as the ribosome, and the femtosecond dynamics of bond-breaking in small molecules like nitrogen.

Robust sketching for multiple square-root LASSO problems

Vu Pham, UC Berkeley

Many learning tasks, such as cross-validation, parameter search, or leave-one-out analysis, involve multiple instances of similar problems, each instance sharing a large part of learning data with the others. We introduce a robust framework for solving multiple square-root LASSO problems, based on a sketch of the learning data that uses lowrank approximations. Our approach allows a dramatic reduction in computational effort, in effect reducing the number of observations from m (the number of observations to start with) to k (the number of singular values retained in the low-rank model), while not sacrificing sometimes even improving the statistical performance. Theoretical analysis, as well as numerical experiments on both synthetic and real data, illustrate the efficiency of the method in large scale applications.

Fast Randomized Algorithms for Convex Optimization

Mert Pilanci, UC Berkeley

With the advent of massive data sets, statistical learning and information processing techniques are expected to enable unprecedented possibilities for better decision making. However, existing algorithms for mathematical optimization, which are the core component in these techniques, often prove ineffective for scaling to the extent of all available data. We study random projection methods in the context of general convex optimization problems to address this challenge. The proposed method, called the Newton Sketch, is a faster randomized version of the well-known Newton’s Method with linear computational complexity in the input data. We show that Newtons sketch enables solving large scale optimization and statistical inference problems orders-of-magnitude faster than existing methods. Moreover, due to the special structure of certain random projections, it is

possible to speed up computation even further using dedicated hardware implementations such as graphical processing units (GPUs).

Rectools: A recommendation engine package

Pooja Rajkumar, UC Davis

Recommendation engines have a number of different applications. From books to movies, they enable the analysis and prediction of consumer preferences. The prevalence of recommender systems in both the business and computational world has led to clear advances in prediction models over the past years.

Current R packages include recosystem and recommenderlab. However, our new package, rectools, currently under development, extends its capabilities in several directions. One of the most important differences is that rectools allows users to incorporate covariates, such as age and gender, to improve predictive ability and better understand consumer behavior.

Our software incorporates a number of different methods, such as non-negative matrix factorization, naive latent factor and maximum likelihood random effects models, and nearest neighbor methods. In addition to our incorporation of covariate capabilities, rectools also integrates several kinds of parallel computation.

Code is available on GitHub, at <https://github.com/Pooja-Rajkumar/rectools>

Using Play-by-Play Data to Model, Simulate, and Predict NBA Games

Sebastian Rodriguez, University of California, Merced

In an attempt to improve our ability to predict the outcome of NBA games, we seek to model both end-game score differential and playing time for players and team lineups. To do this, we calculate scoring rates for each lineup and create stochastic lineup substitution models for all 30 teams. We examine different linear and non-linear methods to calculate the lineup scoring rates. The substitution model involves a continuous-time Markov Chain for each team in which the transition rates were inferred from the data. Training on 2014-15 and 2015-16 NBA regular season play-by-play data, we compare results of our proposed method for simulating and predicting playoff games to that of our previous study based on lineup and player play-time, point-spread accuracy, and predicted winners.

A Transfer Learning Approach for Autonomous Reconfiguration of Wearable Systems

Ramyar Saeedi, Washington State University

Recent advances in sensors, cloud computing, and related technologies are making wearables a powerful component of Cyber-Physical Systems (CPSs). These advances promise to provide wearables with the ability to observe patients or

users remotely and take actions or give feedback regardless of their locations. Wearables rely on computation and communication deeply embedded in and interacting with the human body and the environment. Because of the need for integration into the everyday life of users, wearables need to meet higher reliability, usability and predictability standards than general-purpose computing systems. But, utilization of wearables is currently limited to controlled environments, laboratory settings, and predefined protocols. This limitation creates major obstacles in scaling these systems up and advancing their utility in real-world environments. Therefore, there is a growing need for designing autonomous wearables that withstand uncontrolled and unexpected conditions and deliver an acceptable level of accuracy to users. Computational methods, including machine learning and signal processing algorithms, are used as the core intelligence of wearables for real-time extraction of clinically important information from sensor-collected data. Current approaches employing such methods, however, suffer from several deficiencies: 1) the accuracy of the computational algorithms decreases as the configuration of the system changes (e.g., due to sensor misplacement). 2) Computational algorithms for such tasks as classification and template matching typically need training data for building processing models for each configuration. In order to re-train the computational algorithms across configurations, there is a need to collect sufficient amount of labeled training data, a process that is known to be time-consuming and expensive. 3) Data collected for a specific type or a sensor brand may not be accurate enough in a new setting. In this ongoing work, our goal is to devise effective transfer learning methods (at the signal level) for adapting a computational model of wearables developed in one domain to a different but related domain. The target domain for the system could arise as a result of user-specific uncertainty, sensor variation, environmental changes, etc. Our main focus is on motion sensors (e.g., accelerometer) We developed network analysis methods to map signal patterns from the source domain to the target domain. In particular, we build a suitable graph model for the sensor data and use community detection methods on the graph to cluster the sensor data based on signal similarity. Our results show that by using the community detection methods, data can be better clustered into a similar set of signals for different domains (e.g., different smartphones). In a subsequent step of our method, the similar communities are used to transfer knowledge from the source to the target domain. This is a joint work with Assefaw Gebremedhin, Seyed Ali Rokni and Hassan Ghasemzadeh, all at Washington State University.

Core periphery structures to analyse a spatio-temporal dataset of crimes in San Francisco

Divya Sardana, University of Cincinnati

Core periphery structure is a meso-scale property of complex networks. A widely accepted definition of core periphery structures defines core as a dense cohesive cluster, surrounded by a sparsely connected periphery. Traditionally, core periphery structures have been successfully

used in diverse fields such as world systems, economics, social networks, and organizational studies. In this work, we demonstrate the effectiveness of core-periphery structures to analyse patterns in a large scale spatio-temporal dataset of crimes in the San Francisco city. Such an analysis can prove to be very useful to study the modus operandi of different crime types over time. We compare and contrast the results obtained for two core-periphery structure finding algorithms that we have developed, namely, GMKNN-CP and Clusterone-CP.

Fast Hierarchy Construction for Dense Subgraphs

A. Erdem Sariyuce, Sandia National Labs

Discovering dense subgraphs and understanding their relations is important. Peeling algorithms (k-core, k-truss, and nucleus decomposition) have been shown to be effective to locate many dense subgraphs. However, constructing the hierarchy and even correctly computing the connected k-cores and k-trusses are mostly overlooked in the literature. Finding k-cores, k-trusses and nuclei, and constructing the hierarchy requires an additional traversal operation which is as expensive as the peeling process. In this work, we first fix the mistake in the literature by a thorough review of history, and then propose efficient and generic algorithms to construct the hierarchy of dense subgraphs for k-core, k-truss or any nucleus decomposition. Our algorithms leverage the disjoint-set forest data structure to efficiently construct the hierarchy during traversal. Furthermore, we introduce a new idea to get rid of the traversal. We construct the subgraphs while visiting neighborhoods in the peeling process, and build the relations to previously constructed subgraphs. We also bring out an existing idea for k-core hierarchy, and adapted to our objective efficiently. Experiments on different types of large scale real-world networks show significant speedups over naive algorithms and existing alternatives.

A Subsampled Double Bootstrap for Massive Data

Xiaofeng Shao, University of Illinois at Urbana-Champaign

The bootstrap is a popular and powerful method for assessing precision of estimators and inferential methods. However, for massive datasets which are increasingly prevalent, the bootstrap becomes prohibitively costly in computation and its feasibility is questionable even with modern parallel computing platforms. Recently Kleiner, Talwalkar, Sarkar, and Jordan (2014) proposed a method called BLB (Bag of Little Bootstraps) for massive data which is more computationally scalable with little sacrifice of statistical accuracy. Building on BLB and the idea of fast double bootstrap, we propose a new resampling method, the subsampled double bootstrap, for both independent data and time series data. We establish consistency of the subsampled double bootstrap under mild conditions for both independent and dependent cases. Methodologically, the subsampled double bootstrap is superior to BLB in terms of running time, more sample coverage and automatic implementation with less tuning parameters for a given time budget. Its advantage relative

to BLB and bootstrap is also demonstrated in numerical simulations and a data illustration.

SPLATT: Enabling Large-Scale Sparse Tensor Analysis

Shaden Smith, University of Minnesota

Modeling multi-way data can be accomplished using tensors, which are data structures indexed along three or more dimensions. Tensor factorization is increasingly used to analyze extremely large and sparse multi-way datasets in life sciences, engineering, and business.

Tensor factorization is a computationally challenging task. The time and memory required to factor sparse tensors limits the size and dimensionality of the tensors that can be solved on a typical workstation, often making distributed solution approaches the only viable option. Existing tools are either totally unable to factor large tensors or can require days or even weeks to complete.

To that end, we present SPLATT, a software toolkit for large-scale sparse tensor factorization. SPLATT is a hybrid MPI+OpenMP code designed with performance from the start. SPLATT uses a compressed data structure that reduces memory requirements while reducing the operation count and improving cache locality. In result, SPLATT can factor tensors with billions of non-zeros in only a few minutes time on a supercomputer.

A New Similarity Score for Large-Scale, Sparse, and Discrete-Valued Data

Veronika Strnadova-Neeley, UC Santa Barbara, Lawrence Berkeley National Lab

I present a new similarity score, based on a statistical model, that is useful for clustering problems with high missing data rates and discrete data values. In settings that range from genomics to recommender systems, I demonstrate how this score can be used to develop fast algorithms for large-scale clustering. Together with collaborators at Lawrence Berkeley National Lab, UC Berkeley, and UC Santa Barbara, I developed the new similarity score by comparing the likelihood of observed data under an assumed clustering model, to the probability of observing the same data by chance. The advantage of our score over traditional similarity scores is its ability to leverage more data to make more accurate similarity comparisons, as long as a certain underlying clustering structure exists in the data. We applied our score to the recommender systems domain, where the challenges of high missing data rates and high dimensionality abound. We have shown that this new score is more effective at identifying similar users than traditional similarity scores, such as the Pearson correlation coefficient, in user-based collaborative filtering. We argue that our approach has significant potential to improve both accuracy and scalability in collaborative filtering. In ongoing work, we are building on the success of this similarity score in the sparse recommender systems setting to design new clustering algorithms for general discrete-valued data with high missing rates.

Enabling Brain Functional Alignment for a Thousand Subjects

Javier Turek, Parallel Computing Lab - Intel

Over the past few decades, neuroscience has seen major advances in understanding the mind thanks to functional magnetic resonance imaging (fMRI), which produces three-dimensional snapshots of brain activity. Recently, large multi-subject datasets have become widely available, leading to the development of new methods that can leverage this data. A challenge of working with multi-subject datasets is that subjects have distinct anatomical and functional structure, making direct aggregation of subject data unfeasible. Standard anatomical alignment methods provide a partial solution to this problem as functional brain topographies do not necessarily align. Lately, functional alignment techniques have been developed to overcome these limitations. The Shared Response Model (SRM) is the state-of-the-art method that maps functional topographies from every subject data to a low-dimensional shared response. This unsupervised machine learning method learns a mapping for each subject and a low-dimensional shared response, which can be applied to map one subject to another, reduce noise, discriminate between groups, and more. Despite its great predictive performance, SRM is hard to compute over a few tenths of subjects. We perform algorithmic optimizations that reduce the memory and runtime complexity and allow us to estimate the model for hundreds of subjects. We further develop a code-optimized distributed version of SRM. This distributed algorithm exhibits promising results of strong scaling results of up to 5x with 20 nodes on real datasets, and successful weak scaling by aligning a synthetic dataset of 1024 subjects in 512 nodes in 51 seconds.

Sub-sampled Newton Methods with Non-uniform Sampling

Peng Xu and Jiyan Yang, Stanford University

We consider the problem of minimizing a function $F(w)$ whose Hessian can be written as $\nabla^2 F(w) = \sum_{i=1}^n H_i(w) + Q(w)$ where n is much larger than the dimension of w , denoted by d . A high value of n and d makes it prohibitive to use a second-order method in which it can take $O(nd^2)$ time to form the Hessian and $O(d^3)$ time to compute an update. In this work, we propose a randomized Newton-type algorithm that exploits non-uniform sampling the terms in $\sum_{i=1}^n H_i(w)$ and inexact update. Two sampling distributions, namely, *row norm squares sampling* and *leverage scores sampling*, are considered. In addition, when computing the update, to further accelerate the solver, we consider using an iterative solver such as Conjugate Gradient (CG) to compute an inexact solution. With an analysis that improves the recent related work, we show that at each iteration non-uniformly sampling at most $O(d \log d)$ terms from $\sum_{i=1}^n H_i(w)$ is sufficient to achieve a linear-quadratic convergence rate when a suitable initial point is provided. Furthermore, we show that our sub-sampled Newton methods with inexact update have a better complexity than other stochastic second-order methods in terms of the dependence on the condition number when a constant linear convergence rate is desired. Finally, we show that this algorithm is applicable to many problems including Generalized Linear Models (GLM) and Semi-definite Programming (SDP). For the former, we propose a scheme to fast compute leverage scores based on information of previous iterations. Advantages of our algorithm are numerically verified.

Acknowledgements

Sponsors

The Organizers of MMDS 2016 and the MMDS Foundation would like to thank the following institutional sponsors for their generous support:

- **BIDS**, the Berkeley Institute for Data Science
- **Interana**, Behavioral Analytics for Event Data at Scale

