UW IGERT/BDGN Retreat and NSF Graduate Data Science Workshop 2016

*Please use **#uwdatascienceworkshop2016** for your tweets!*

Participant List Index for Titles and Abstracts

Mahdi Ahmadi 3

Sultanah Alshammari 3

Sandipan Banerjee 4

David Caldwell 4

Tony Cannistra 5

Daniel Cook 5

Subrina Farah 6

David Fleming 6

Christopher Fu 7

Alexey Gilman 7

Kameron Harris 8

Erika Helgeson 8

Amnon Horowitz 9

Shiree Hughes 9

Ken Jean-Baptiste 9

Serena Liu 10

Ryan McGee *10*

Fred Morstatter *10*

Matt Murbach *11*

Cecilia Noecker *11*

Eliah Overbey *12*

Pearl Philip *12*

Nima Salehi *13*

Susmit Shannigrahi *13*

Lincoln Sheets *13*

Qi Song *14*

Alex Tank *14*

Jin Tao *15*

Fei Wang *15*

Caitlyn Wolf *16*

Wei Xie *16*

Jianbo Ye *17*

Lisheng Zhou *18*

PARTICIPANTS TITLES AND ABSTRACTS

Mahdi Ahmadi

University of North Texas

Mechanical and Energy Engineering

Toward higher resolution and accuracy in air pollution data

Accurate and spatially high resolution air pollution data is necessary for a correct assessment of health impacts of air pollutant. Air pollution measurement data is very sparse in space because it is expensive to build and operate measurement stations. On the other hand, computer simulations and forecasts of air pollutant concentration have very high spatial resolution but still are poor in accuracy. Several national and state agencies are producing and archiving both type of information (i.e. measurement and simulation) every day. Combining these two large sets of geospatial data and applying customized interpolation algorithms repetitively we can determine proper methods to accurately estimate air pollution concentration at any point between the measurement stations. In this poster I present my current research on the design, implementation, assessment, and challenges of such algorithms.

Sultanah Alshammari

University of North Texas

Computer Science

Big Data Opportunities in Detecting Infectious Diseases Outbreaks in Global Mass Gatherings

Spread of infectious diseases at global mass gatherings can pose health threats to both the hosting country and the countries where participants originate. The travel patterns at the end of these international events may result in epidemics that can grow to pandemic levels within a short period of time. The huge amount of data being generated in relation to these events will make important contributions to public health surveillance once being collected and processed. This open up the opportunity of applying big data approaches to detect infectious diseases outbreaks at global mass gatherings. In this review, we present the key data requirements to model infectious diseases epidemics at global mass gatherings and the possible big data opportunities in this emerging research area.

Sandipan Banerjee

Notre Dame

Computer Science and Engineering

Visual Recognition of Paper Analytical Device Images for Detection of Falsified Pharmaceuticals

Falsification of medicines is a big problem in many developing countries, where technological infrastructure is inadequate to detect these harmful products. We have developed a set of inexpensive paper cards, called Paper Analytical Devices (PADs), which can efficiently classify drugs based on their chemical composition, as a potential solution to the problem. These cards have different reagents embedded in them which produce a set of distinctive color descriptors upon reacting with the chemical compounds that constitute pharmaceutical dosage forms. If a falsified version of the medicine lacks the active ingredient or includes substitute fillers, the difference in color is perceivable by humans. However, reading the cards with accuracy takes training and practice, which may hamper their scaling and implementation in low resource settings. To deal with this, we have developed an automatic visual recognition system to read the results from the PAD images. At first, the optimal set of reagents was found by running singular value decomposition on the intensity values of the color tones in the card images. A dataset of cards embedded with these reagents is produced to ge4n4e4r4a4t4e4 4t4h4e4 4m4o4s4t4 4d4i4s4t4i4n4c4t4i4v4e4 4r4e4s4u4l4t4s4 4f4o4r4 4a4 4s4e4t4 4o4f4 42464 4d4i4f4f4e4r4e4n4t4 4a4c4t4i4v4e4 4p4h4a4r4m4a4c4e4u4t4i4c4a4l4 4i4n4g4r4e4d4i4e4n4t4s4 4(4A4P4l4s4)4 4a4n4d4 4e4x4c4i4p4i4e4n4t4s4.4 4T4h4e4n4,4 4w4e4 4t4r4a4i4n4 4t4w4o4 4p4o4p4u4l4a4r4 4c4o4n4v4o4l4u4t4i4o4n4a4l4 4n4e4u4r4a4l4 4n4e4t4w4o4r4k4 4(4C4N4N4)4 4m4o4d4e4l4s4,4 4w4i4t4h4 4t4h4e4 4c4a4r4d4 4i4m4a4g4e4s4.4 4W4e4 4a4l4s4o4 4e4x4t4r4a4c4t4 4s4o4m4e4 4h4a4n4d4-4c4r4a4f4t4e4d4 4f4e4a4t4u4r4e4s4 4f4r4o4m4 4t4h4e4 4

David Caldwell

UW - BDGN

Bioengineering

Neuromodulation by Direct Electrical Cortical Stimulation in Humans

Neuromodulation through direct electrical stimulation of human cortex may improve neurologic outcomes after injury such as stroke. Prior research demonstrated that neural activity dependent stimulation can enhance synaptic plasticity, and potentially rehabilitation outcomes. Corticocortical evoked potentials (CCEPs) offer a method to interrogate the effects of direct electrical stimulation (DES) on human cortex and its effect on synaptic plasticity. In order to better characterize the potential role of direct electrical stimulation for neurorehabilitation in humans, we studied the beta-oscillation (12-20 Hz) triggered stimulation through characterization of CCEPs in humans implanted with electrocorticographic (ECoG) grids for epilepsy monitoring. Enhancement of CCEPs, consistent with prior literature, served as a proxy for enhancement of short term synaptic plasticity. We demonstrated short term enhancement of CCEPs with beta-band neural activity dependent stimulation. This study lays the groundwork for future development of neural activity dependent electrical stimulation in humans. More generally, this study

demonstrates the implementation of simultaneous recording and stimulation in humans, with potential applications in neuroprosthetics for functional restoration. Furthermore, the neural responses to stimulation are being explored through dimensionality reduction techniques to better understand and characterize the neural response to stimulation.

Tony Cannistra

UW - IGERT

Biology

The Effects of Climate on Biological Organisms

Novel approaches to understand the long term effects of climate change on biological organisms are critical to future science and policy with regard to ecosystem and biodiversity conservation. With the Paris agreement from COP21 recommending a 1.5ºC cap on global average temperature increase, how will biological organisms respond to the corresponding environmental change? Already researchers have revealed species range shifts, shifts in seasonal timing, habitat fragmentation, and habitat loss resulting from changing climate. However, these studies have been largely manual, field-based, and reliant on low-resolution climate observations. I aim to bring this field into the data-centric age by intelligently combining environmental data and organismal data to predict species responses to climate change, to enhance the quality and resolution of these predictions, and to create an open-source, accessible, and adaptable data pipeline for ease of use and transparency.

Daniel Cook

Northwestern University

Molecular Biosciences

Developing data-science web applications using the cloud

Data scientists frequently develop new methods, tools, and/or data sets. Users who want access to these resources may be inhibited by the need to setup specific software environments, understand complex statistical methods, or download large files. However, these challenges can reduced or eliminated by providing interfaces for data access, or by running analytical pipelines using cloud-based infrastructure. As part of the Andersen lab at Northwestern University, I have spearheaded the development of the C. elegans Natural Diversity Resource (CeNDR) located at elegansvariation.org. CeNDR is a web-application that provides data and tools to facilitate the study of natural variation in the C. elegans species. For example, an application programming interface (API) for accessing trait or genetic variant data across C. elegans wild isolates is provided. Additionally, a genome-browser and several databases are utilized and allow for integrative analysis. Finally, a genome-wide association mapping portal allows users to submit association mapping jobs and browse results. CeNDR makes use of a number of cloud-based services including relational databases, virtual machines, and file storage. Using CeNDR as an example, I provide an

overview of how data-scientists can utilize cloud-based services to facilitate access to resources they have developed.

Subrina Farah

University of Rochester

Family Medicine Research

Determining Predictors of Utilization of Quality Health Information among Patients and Federally Qualified Health Centers (FQHC): A Big Data Approach

Little is known regarding how to facilitate and empower inner-city patients to self-manage their healthcare via online resources. This project focuses on determining best approaches to engage low-income, largely minority patients in the use of an online patient portal for their personal health records, and to access high quality patient health information.

This study is funded by National Library of Medicine (NLM). Interventions are designed to facilitate patients to access quality health information from trusted and established health resources. e.g., how to manage the Patient Portal, attending structured classes to be an active patient, participating group presentations, one on one session in waiting room to improve patients' skills for accessing quality health information through Medline Plus.

Data have been collected from 608 participants during pre and post interventions time points. According to survey results, gender, Medline Plus usage, interest on portal training sessions, smartphone usage, portal usage, interest in learning to use portal and interest in HEALOW have been found significant between pre-post sessions ($p < .05$). Then advance data mining techniques have been used for determining predictors of utilization of quality health information among patients and health centers.

David Fleming

UW - IGERT

Astronomy

Simulating the high-dimensional problem of planetary habitability using VPLANET

Habitability, defined for a terrestrial planet as having liquid surface water, is impacted by many diverse phenomena ranging from atmospheric escape to orbital dynamics. Using VPLANET, we simulate Proxima Centauri b (Anglada-Escudé 2016), the closest exoplanet within its star's Habitable Zone (HZ), to identify potential pathways towards habitability. Performing a robust habitability analysis necessitates coupling numerous physical processes to explore

this high-dimensional space for a large number of simulations. To tackle this problem, we present a new software tool, bigplanet, which allows for hierarchical storage and analysis of simulation outputs too large to fit in memory.

Christopher Fu

UW - IGERT

Chemical Engineering

Discovering and Characterizing Chemical Reactions Through Molecular Simulations

Determining and characterizing chemical reaction mechanisms would allow for industrial processes to be optimized and enhance our understanding of the fundamentals of chemistry. Molecular dynamics simulations offer a unique scope to learn about such systems at a level that is difficult to resolve entirely through experiment alone. However, these methods are often computationally expensive. This project focuses on methods used to discover and characterize transitions while effectively reducing the computational cost.

Alexey Gilman

University of Washington

Chemical Engineering

Comparative time-course transcriptomics for induced bacterial metabolic switch

Methanotrophic bacteria utilize methane as their sole source of carbon and energy. They play an important role in sequestering methane (potent greenhouse gas) from the atmosphere and are a major player in the global carbon cycle. Methane monooxygenase (MMO), is the only known enzyme that is able to break the extremely stable methane C-H bond during methane oxidation. Unlike industrial catalytic processes, methanotrophs carry out this reaction at ambient conditions. Methylomicrobium buryatense (5GB1) contains genes for redundant methane oxidizing machinery that can switch in response to environmental conditions. PMMO (particulate methane monooxygenase) is a membrane-bound Cu-containing enzyme and acts as primary MMO when Cu is available. sMMO (soluble methane monooxygenase) is an Fe-containing enzyme and is present within the cytoplasm. In this work, we have grown 5GB1 culture in a bench-scale bioreactor and induced a metabolic switch by the addition of Cu and collected time-course RNA samples for sequencing. Several notable differences in metabolism are observed during the transition including faster growth rate and increased membrane lipid synthesis. In the last chapter of my thesis, I will use data-science techniques to cluster genes that are co-regulated during the time-course switch and attempt to elucidate their function. The overall goal of this exploratory study is to better understand the metabolic network as the cell responds to changing conditions.

Kameron Harris

UW - BDGN

Applied Math

Construction of a voxel-based mesoscopic mouse connectome

Whole-brain neural connectivity data are now available from viral tracing experiments, which reveal the connections between a source injection site and elsewhere in the brain. These hold the promise of revealing spatial patterns of connectivity throughout the mammalian brain. To achieve this goal, we seek to fit a weighted, nonnegative adjacency matrix among 100 µm brain "voxels" using viral tracer data. Despite a multi-year experimental effort, the problem remains severely underdetermined: Injection sites provide incomplete coverage, and the number of voxels is orders of magnitude larger than the number of injections. Furthermore, projection data are missing within the injection site because local connections there are not separable from the injection signal.

We use a novel machine-learning algorithm to meet these challenges and develop a spatially explicit, voxel-scale connectivity map of the mouse visual system.

Erika Helgeson

University of North Carolina

Biostatistics

Non-Parametric Cluster Significance Testing

Cluster analysis is an unsupervised learning strategy that can be employed to identify subgroups of observations in data sets of unknown structure. This strategy is particularly useful for analyzing high-dimensional data such as microarray gene expression data. Many clustering methods are available, but it is challenging to determine if the identified clusters represent distinct subgroups. We propose a novel strategy to investigate the significance of identified clusters by comparing the within-cluster sum of squares from the original data to that produced by clustering an appropriate unimodal null distribution. The null distribution we present for this problem uses kernel density estimation and thus does not require that the data follow any particular distribution. We find that our method can accurately test for the presence of clustering even when the number of features is high.

Amnon Horowitz

UW - IGERT

Computer Science & Engineering

Tracking neurons in Hydra Vulgaris

In order to extract calcium traces in lengthy recordings of behaving Hydra Vulgaris, we need to track the positions of individual neurons in a Hydra that is moving and deforming with time.

Shiree Hughes

Florida Atlantic University

Sensing and Embedded Network Systems Engineering

Food Waste on College Campuses

America wastes 40% of the food it produces; a reduction of just 15% could feed 25 million people. The combination of behavioral suggestion, social norming, and sensors networks can provide us with the tools needed to convince people to make smarter decisions about their eating habits. The prototype has been deployed in a college cafeteria for evaluation. Data will be collected before and after feedback is given to visitors, to determine if these methods successfully alter behavior.

Ken Jean-Baptiste

UW - BDGN

Genome Sciences

Characterizing Chromosome Conformation during Heat Shock Response in Arabidopsis thaliana

In response to heat stress, Arabidopsis thaliana undergoes rapid genome-wide shifts in transcription. This shift in transcription increases the expression of genes that encode heat shock proteins. A possible mechanism that may facilitate this rapid change in transcription is a drastic change in chromatin conformation that brings heat shock genes into close spatial proximity. This spatial proximity allows for heat shock transcription factors to quickly

increase the expression of heat shock genes. In order to investigate this, Hi-C will be used to investigate changes in chromatin conformation post heat shock.

Serena Liu

UW - IGERT

Genome Sciences

Exploring the macrophage polarization landscape via single-cell RNA-seq

Macrophages are immune cells that perform a wide range of crucial functions, including fighting off infections and promoting tissue repair; these cells are also highly plastic and can shift phenotypic states in response to environmental changes. My research aims to characterize the molecular mechanisms that underlie macrophage plasticity by using single-cell technologies.

Ryan McGee

UW - IGERT

Biology

Degree-of-Interest Hierarchical Network Browser

Many interesting networks are too large to interpret without significant simplification. Inference of community structure facilitates understanding, but often abstracts away useful information. This work presents a novel framework that enables detailed exploration of individual elements while maintaining the context of the overall network structure. In particular, this framework is applied to the development of a browser for scientific literature that allows for intuitive search and discovery of literature using citation data, paper content, and metadata.

Fred Morstatter

Arizona State University

CIDSE

An Unbiased Sample or a Spammer's Paradise? The Implications of Sampling Strategies in Social Media

While social media mining continues to be an active area of research, obtaining data for research is a perennial problem. Even more, obtaining unbiased data is a challenge for researchers who wish to study human behavior, and not technical artifacts induced by the sampling algorithm of a social media site. In this work, we evaluate one social media data outlet that gives data to its users in the form of a stream: Twitter's Sample API. We show that in its current form, this API can be poisoned by bots or spammers who wish to promote their content, jeopardizing the credibility of the data collected through this API. We design a proof-of-concept algorithm that shows how malicious users could increase the probability of their content appearing in the Sample API, thus biasing the content towards spam and bot content and harming the representativity of this data outlet.

Matt Murbach

UW - IGERT

Chemical Engineering

Unlocking insight into battery dynamics with nonlinear electrochemical impedance spectroscopy (NLEIS)

Electrochemical impedance spectroscopy (EIS) is widely used in the study of lithium ion batteries due to the relative ease of using equivalent circuit analogs to extract kinetic/transport parameters and provide data for health prognostics. However, the use of a small perturbation restricts EIS to probing only the linear regime despite the inherently nonlinear nature of the electrochemical system. While this linearization leads to a relatively straightforward analysis, the degeneracy of linearized EIS signals can result in a loss of discriminating power and artificially limit the informational content of the technique. Here we present an extension of traditional small amplitude EIS measurements – called nonlinear EIS (NLEIS) – in which the nonlinearity of a lithium ion battery is probed by measuring the higher harmonic response to moderate amplitude perturbations. We discuss the insight gained through both physics-based modeling and experimental measurements and look at the increased informational content in nonlinear impedance measurements.

Cecilia Noecker

UW - IGERT

Statistics

MIMOSA: An integrative modeling framework for linking ecological and metabolic microbiome variation

Multi-omic technologies have enabled comprehensive profiling of the composition and activity of the human microbiome. In particular, taxonomic profiling combined with metabolomics is an increasingly common approach. Typical analyses of such datasets, however, aim to identify associations between specific taxa and metabolites, without utilizing extensive prior knowledge of the mechanisms that link them. We present MIMOSA (Model-based Integration of Metabolite Observations and Species Abundances), an analytical framework to systematically model,

interpret, and compare variation in community composition and metabolite concentrations. MIMOSA uses taxonomic composition, genomic data, and enzymatic information to construct a simple mathematical model of community metabolism that predicts biosynthetic and degradation potential for each metabolite in each sample. It then uses a permutation-based approach to identify metabolites whose concentration varies in accordance with predicted metabolic potential. MIMOSA further identifies key taxa and reactions underlying the metabolic potential estimates. We validated MIMOSA's capabilities using simulation data generated with a constraint-based dynamic model of community metabolism. We found that with complete genomic and metabolic reference information, MIMOSA significantly explains variation in 68% of the metabolome of a model mouse gut community in spite of several simplifications, and is robust to added noise in quantification. We further evaluated the impact of missing reference information on predictions, and the ability to detect key taxa and reactions from simulated perturbations. We have applied MIMOSA to several microbiome multi-omic datasets. In a large collection of vaginal microbiome samples, MIMOSA mechanistically explained variation in 37% of metabolites based on shifts in community composition, including 55% of those enriched in the bacterial vaginosis disease state. In a study of fecal samples from mice fed one of two commonly used chows, we found, as expected, that MIMOSA can explain variation in non-dietary (likely microbial) metabolites at a much higher rate than dietary compounds (48% versus 14%), and identified metabolite variation putatively attributable to diet-microbiome interactions. As multi-omic studies increase in prevalence, frameworks to integrate these datasets and gain an improved mechanistic understanding of the microbiome's dynamics are becoming essential. MIMOSA is available as an R package at http://www.github.com/borenstein-lab/MIMOSA.

Eliah Overbey

UW - BDGN

Genome Sciences

Finding Enhancer-Promoter Interactions in Neural Progenitor Cells: Hi-C on Neural Rosettes

Hi-C is DNA sequencing technique used to capture information about which pieces of DNA are in contact with one another within the nucleus. DNA contacts are most commonly between promoters, which lie directly next to the gene they regulate, and enhancers, which up-regulate gene transcription. I am currently working on an experiment to perform Hi-C in a type of neural progenitor cell, called neural rosettes, which are derived from human embryonic stem cells.

Pearl Philip

University of Washington

Chemical Engineering

Enzymatic Inhibition through Data-Driven Drug Discovery

We use the power of quantitative structure-activity relationships (QSAR) models for the prediction of a desired interaction between proteins and small-molecules. A large number of molecular descriptors are employed to characterize the structure of about 400,000 drug-like compounds from a high-throughput screening assay. The data - available on PubChem - pertains to the inhibition of the deubiquitinating USP1/UAF1 - an enzyme essential to DNA-repair in proliferating cancer cells. Using an ensemble of robust QSAR models with optimized feature selection, the algorithm seeks to design an ideal enzyme inhibitor, towards chemical synthesis in a lab for use as a drug compound. The work highlights the most suitable set of features and model parameters for this data set. The goal of this research is to build a cheminformatics software for chemical data analyses, improve existing approaches to the USP1-inhibition problem, and create open-source resources for the design of molecular structures.

Nima Salehi

University of Michigan

Industrial and Operations Engineering

A Random Forest Classifier for Multi-type Functional Neuroimaging Data

Recent advancement in functional neuroimaging has led to significant understanding of the human brain. Traditional meta-analysis techniques have been used on neuroimaging data widely. However, they are suffering from a very important constraint. Researchers often do not provide the statistic images or the complete datasets; they only provide the locations of the activated points for each emotion (foci). In this study, we designed a modified Random Forest (RF) classifier for multi-type functional neuroimaging data (foci) and a K-Centroids Cluster Analysis (KCCA) algorithm to pre-process the foci. Our results suggest that the activated points in the brain do not entirely belong to one specific emotion (mixed feelings). In order to improve the accuracy of the classifier, we need to take into account the scatter noise (foci that do not cluster) by modeling them as a homogeneous process. Another way to deal with this problem is to define mixed membership models.

Susmit Shannigrahi

Colorado State University

Computer Science

Scientific Data Management Using Named Data Networking

Lincoln Sheets

University of Missouri

MU Informatics Institute

Data Mining to Predict Healthcare Utilization

Background. Because 5% of patients incur 50% of healthcare expenses, population health managers need to be able to focus preventive and longitudinal care on those patients who are at highest risk of increased utilization. Predictive analytics can be used to identify these patients and to better manage their care. Data mining permits the development of models that surpass the size restrictions of traditional statistical methods and take advantage of the rich data available in the electronic health record (EHR), without limiting predictions to specific chronic conditions. Objective. The objective was to improve predictive analytics in managed healthcare by mining EHR data for clinical predictors of increased healthcare utilization. Methods. In a population of 9,568 Medicare and Medicaid beneficiaries, patients in the highest 5% of charges were compared to equal numbers of patients with the lowest charges. Contrast mining was used to discover the combinations of clinical attributes frequently associated with high utilization and infrequently associated with low utilization. The attributes found in these combinations were then tested by multiple logistic regression, and the accuracy of the model was evaluated by the area under the receiver operating characteristic curve (AUROC). Results. Of 19,012 potential EHR patient attributes, 61 were found in combinations frequently associated with high utilization, but not with low utilization (support>20%). Thirteen of these attributes were significantly correlated with high utilization ($p<0.05$). Prediction models composed of these thirteen attributes had accuracies ranging from 74% to 75%. Conclusions. EHR mining reduced an unusably high number of patient attributes to a manageable set of potential healthcare utilization predictors. Treating these results as hypotheses to be tested by conventional methods yielded a highly accurate predictive model. This novel, two-step methodology can assist population health managers to focus preventive and longitudinal care on those patients who are at highest risk for increased utilization.

Qi Song

Washington State University

Electrical Engineering and Computer Science

Knowledge Search Made Easy: Effective Knowledge Graph Summarization and Applications

Mining and searching heterogeneous and large-scale knowledge graphs is very challenging under real-world resource constraints (e.g., memory, response time). knowledge graph search is a challenging task among them. In this work, 1) We introduce a class of summaries characterized by graph patterns. 2) We formulate the computation of graph summarization as a bi-criteria pattern mining problem. Given a knowledge graph G, the problem is to discover k diversified summary patterns that maximizes the informativeness measure. Although this problem is NP-hard, we develop an anytime algorithm to solve this bi-criteria pattern mining problem. 3) We develop query evaluation algorithms by leveraging the graph summarization. These algorithms efficiently compute (approximate) answers with high accuracy by only accessing a small number of summary patterns and their materialized views. Using real-world knowledge graphs, we experimentally verify the effectiveness and efficiency of our parallel algorithms for computing summarizations; and query evaluation guided by summarization.

Alex Tank

UW - IGERT

Statistics

Bayesian Structure Learning for Stationary Time Series

While much work has explored probabilistic graphical models for independent data, less attention has been paid to time series. The goal in this setting is to determine conditional independence relations between entire time series, which for stationary series, are encoded by zeros in the inverse spectral density matrix. We take a Bayesian approach to structure learning, placing priors on (i) the graph structure and (ii) spectral matrices given the graph. We leverage a Whittle likelihood approximation and define a conjugate prior---the hyper complex inverse Wishart---on the complex-valued and graph-constrained spectral matrices. Due to conjugacy, we can analytically marginalize the spectral matrices and obtain a closed-form marginal likelihood of the time series given a graph. Importantly, our analytic marginal likelihood allows us to avoid inference of the complex spectral matrices themselves and places us back into the framework of standard (Bayesian) structure learning. In particular, combining this marginal likelihood with our graph prior leads to efficient inference of the time series graph itself, which we base on a stochastic search procedure, though any standard approach can be straightforwardly modified to our time series case. We demonstrate our methods on analyzing stock data and neuroimaging data of brain activity during various auditory tasks.

Jin Tao

Washington State University

EE / CS

Supervised Learning of Activity Recognition Policies with Dataset Aggregation

Because of the sequential nature of human activity, we can consider the previous activities as the context to achieve better activity recognition accuracy, making it a sequential supervised learning problem. Sequential learning problems, where future observations depend on previous predictions(actions), violate the common i.i.d. assumptions made in statistical learning. This leads to poor performance in theory and ofter in practice. In this work, we combine an iterative algorithm with the recurrent classifier based on the exact feature representation. We demonstrate that the presented approach with different parameter settings including the constant one, the decimal, the exponential decay and the linear decay based on the exact feature representation can achieve higher recurrent training accuracy and testing accuracy as the iteration goes on than those of the previous recurrent classifiers.

Fei Wang

Rutgers University

Waksman Institute

User Interactive Web Applications for Data Visualization and Science Communication

In the century of big data, it has become fashionable to make all information publicly available. Everyone seem to be in a race to make the most data available as soon as possible. However, with petabytes of data being published every day, What are we going to do with all the data? It is only the beginning to publish research paper and data online. We need to harness the data so that people can understand and interact with them. Furthermore, in an era of scientific complexity and with the deluge of data currently being produced by research, it becomes paramount to engage the general public so they do not become distanced from science and important societal issues.There are various possibilities in making data accessible. I will show a few examples of what others are doing to engage the public with scientific data.

Caitlyn Wolf

UW - IGERT

Chemical Engineering

Conjugated Polymer Molecular Dynamics: Optimization with Scattering

The development of more efficient organic electronics relies on improvement of conjugated polymers. Control over structure and dynamics of conjugated polymers would assist in the design of improved optoelectronic properties. Computational work would be an approach to this design, but currently available force fields are inadequate in capturing electronic conjugation behavior. This work focuses on improving force fields for modeling conjugated polymers by utilizing experimental wide-angle x-ray scattering and quasi-elastic neutron scattering methods. Here we present previously completed work of poly(3-hexylthiophene) (P3HT) that identifies key parameters, e.g. torsion potentials and partial charges, in the model. Furthermore, we begin to extend this work to more complex materials (P3DDT, PQT-12 and PBTTT-C12). In the future, an objective will be to apply machine learning for the acceleration of this optimization process, allowing for the eventual in-silico discovery of novel conjugated polymers with improved charge transport properties.

Wei Xie

Vanderbilt University

Electrical Engineering & Computer Science

PrivLogit: Fast Privacy-preserving Logistic Regression by Tailoring Numerical Optimization

Safeguarding privacy in machine learning is highly desirable, especially in collaborative studies across many organizations. Cryptography and distributed computing are increasingly popular for the task. However, existing cryptographic protocols incur excess computational overhead, partially due to naive adoption of mainstream model estimation algorithms (such as the Newton method) and failing to tailor for secure computing-specific characteristics. Here, we present a contrasting perspective on designing numerical optimization method for secure settings. We introduce a seemingly less-favorable optimization method that can in fact significantly accelerate privacy-preserving logistic regression. Leveraging this new method (called PrivLogit), we propose two new secure protocols for conducting logistic regression in a privacy-preserving and distributed manner. Extensive theoretical and empirical evaluations prove the competitive performance of our two secure proposals while without compromising accuracy or privacy: with speedup up to 2.3x and 8.1x, respectively, over state-of-the-art; and even faster as data scales up. Such drastic improvement makes privacy-preserving logistic regression more scalable and practical to large-scale studies which are common for modern science.

Jianbo Ye

Pennsylvania State University

Information Sciences and Technology

Probabilistic Multigraph Modeling for Improving the Quality of Crowdsourced Affective Data

We proposed a probabilistic approach to joint modeling of participants' reliability and humans' regularity in crowdsourced affective studies.Reliability measures how likely it will be that a subject will respond to a question thoughtfully;Regularity measures how often a human will agree with other thoughtfully-entered responses coming from a targeted population. Crowdsourcing-based studies or experiments, which rely on human self-reported affect, pose additional challenges as compared with typical crowdsourcing studies that attempt to acquire {\it concrete non-affective} labels of objects.

The reliability of participants has been massively pursued for typical non-affective crowdsourcing studies, whereas the regularity of humans in an affective experiment in its own right has not been thoroughly considered. It has been often observed that different individuals exhibit different feelings on the same test question, which does not have a single correct response in the first place.High reliability of responses from one individual thus cannot conclusively result in high consensus across individuals. Instead, global testing consensus of a population is of interest to investigators.Built upon the agreement multigraph among tasks and workers, our probabilistic model differentiates

subject regularity from population reliability. We demonstrate the method's effectiveness for in-depth robust analysis of large-scale crowdsourced affective data, including emotion and aesthetic assessments collected by presenting visual stimuli to human subjects.

Lisheng Zhou

Rutgers

Genetics

A Statistical Method for Phenotype-Genotype Association That is Robust to Sequencing Error

Regions of the human genome where genotype frequencies significantly differ among cases and control groups may be regions that harbor disease loci. The purpose of this study is to develop a statistical test of association between multi-locus genotype (MLG) frequencies and a disease phenotype in a case/control study utilizing Next Generation Sequencing (NGS) data. This statistic is a likelihood ratio test whose asymptotic null distribution is central chi-square (degrees of freedom dependent upon the number of MLGs). We designed the statistic to be robust to differential sequencing misclassification (sequencing errors) in NGS variant calls. The parameters utilized in the test are (for each individual): (i) the observed alternative read counts at the set of SNPs; (ii) corresponding sequencing coverage at the SNPs; and (iii) phenotypes. Maximum likelihood estimates of MLG frequencies, sequencing errors and log-likelihoods are determined by an expectation-maximization (EM) algorithm. We apply permutation and bootstrapping to assess the type I error and power. We also use the factorial design to determine the performance of this statistic under specific genetic-model parameter situations. We find that this method maintains the correct type I error rate in simulation data and the 1000 Genome data for permutation with significant levels of 1%, 5%, and 10%.