Second Annual Workshop

EMERGING DATA SCIENCE METHODS FOR COMPLEX BIOMEDICAL AND CYBER DATA
March 26-27, 2020

Invited Speakers, Talk titles, and Abstracts

Rachel Cummings, PhD
Department of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA

Talk title: Differential Privacy for Dynamic Databases

Abstract: Privacy concerns are becoming a major obstacle to using data in the ways we want. How can data scientists make use of potentially sensitive data, while providing rigorous privacy guarantees to the individuals who provided data? Over the last decade, differential privacy has emerged as the de facto gold standard of privacy preserving data analysis. Differential privacy ensures that an algorithm does not overfit to the individuals in the database by guaranteeing that if any single entry in the database were to be changed, then the algorithm would still have approximately the same distribution over outputs. In this talk, we will focus on recent advances in differential privacy for dynamic databases, where the content of the database evolves over time as new data are acquired. First, we will see how to extend differentially private algorithms for static databases to the dynamic setting, with relatively small loss in the privacy-accuracy tradeoff. Next, we see algorithms for privately detecting changes in data composition. We will conclude with a discussion of open problems in this space, including the use of differential privacy for other types of data dynamism. (based on joint works with Sara Krehbiel, Kevin Lai, Yuliia Lut, Yajun Mei, Uthaipon (Tao) Tantipongpipat, Rui Tuo, and Wanrong Zhang.)

George Cybenko, PhD
Dorothy and Walter Gramm Professor of Engineering
Thayer School of Engineering
Dartmouth College, Hanover NH

Talk title: Machine learning and cyber operations in an adaptive adversarial environment

Abstract: While machine learning has shown remarkable progress in a variety of domains, those successes have been made in environments that are stochastically stationary. That is, the statistics of the environment do not change which effectively assumes that an adversary is not adapting. This talk will review basic concepts and several recent relevant results, suggesting ways in which to both analyze, learn and operate in such environments.

Moazzam Khan, PhD
Software Engineer
IBM Security Systems

Talk title: Applying Data science to Detect Malicious User Behavior

Abstract: User behavior is a major indicator of security status of a network. Malicious user behavior may range from inadvertent, such as drive-by download from an infected website, to intentional misuse such as unauthorized access, stealing proprietary information etc. Every user action on a network leaves a trail behind in the form of device logs, we can apply data science on these device logs to extract useful analytics about a user's behavior. In this talk we will discuss IBM Qradar User Behavior Analytics as a use case to see how we can apply data science to the device log data and extract useful analytics about the user behavior.

Xiao-Li Meng, PhD
Whipple V.N. Jones Professor of Statistics,
Harvard University

Talk title: TBD

Abstract: **TBD**

Doug Miller, MD
Senior Associate Dean for Medical Education
Medical College of Georgia
Augusta University

Talk title: TBD

Abstract: **TBD**

Ashis SenGupta, PhD
Applied Statistics Unit
Indian Statistical Institute
Kolkata, India

Talk title: Statistical Machine Learning for Manifold Data - Applications to Gait Analysis

Abstract: TBD

Elizabeth Slate, PhD
Duncan McLean and Pearl Levine Fairweather Professor
Department of Statistics
Florida State University, FL

Talk title: A Joint Model for Biomarker Discovery in Heterogeneous Populations

Abstract: Identification of valid, clinically relevant biomarkers for disease has potential to provide less invasive diagnostic tools, to enhance understanding of initiation and progression at the cellular level, and to guide development of new therapeutic agents. When the biomarkers are binary, logic regression provides a means to discover Boolean combinations of the markers strongly associated with outcome. The interpretability of these Boolean marker combinations and, potentially, additional interactions with environmental and behavioral characteristics, is appealing and can provide insight. However, complex diseases such as cancer that arise from multiple pathways and present at varying stages of development and progression can lead to hidden population heterogeneity in the biomarker-disease association. We describe an extension of logic regression for jointly modeling binary and continuous outcomes that uses a latent class structure to accommodate subpopulation heterogeneity. Estimation and inference are compared for two Bayesian semiparametric formulations using a variety of computational approaches.

Heping Zhang, PhD
Susan Dwight Bliss Professor of Biostatistics, Professor of Statistics and Data Science, Yale University, New Haven, CT

Talk title: A Pursuit of Interaction

Abstract: TBD

Hongyu Zhao, PhD
Ira V. Hiscock Professor of Biostatistics, Professor of Genetics and Professor of Statistics and Data Science, Yale University, New Haven, CT

Talk title: Integrating multidimensional data for clustering analysis with applications to cancer genomics data

Abstract: Advances in high-throughput genomic technologies coupled with large-scale studies including The Cancer Genome Atlas (TCGA) project have generated rich resources of diverse types of omics data to better understand disease etiology and treatment responses. Clustering patients into subtypes with similar disease etiologies and/or treatment responses using multiple omics data type has the potential to improve the precision of clustering than using a single data type. However, in practice, patient clustering is still mostly based on a single type of omics data or ad hoc integration of clustering results from each data type, leading to potential loss of information. By treating each omics data type as a different informative representation from patients, we propose a novel multi-view spectral clustering framework to integrate different omics data types measured from the same subject. We learn the weight of each data type as well as a similarity measure between patients via a non-convex optimization framework. We solve the proposed non-convex problem iteratively using the ADMM algorithm and show the convergence of the algorithm. The accuracy and robustness of the proposed clustering method is studied both in theory and through various synthetic data. When our method is applied to the TCGA data, the patient clusters inferred by our method show more significant differences in survival times between clusters than those inferred from existing clustering methods. This is joint work with Seyoung Park.