Third Annual Workshop

EMERGING DATA SCIENCE METHODS FOR COMPLEX BIOMEDICAL AND CYBER DATA

March 16-17, 2023

Invited Speakers, Talk titles, and Abstracts

Gagan Agrawal, PhD
Professor and Associate Dean of Research, School of Computer and Cyber Sciences, Augusta University
Augusta, GA

Talk title: Applications of Machine Learning to Privacy and Security

Abstract: This talk will describe multiple projects where machine learning has been applied to privacy and security problems. In the first project, we will describe how NLP-based approaches are able to help with digital forensics, especially determining the topic of a given document or a chat sequence. One important contribution here is an explainable method for detecting the topic in a chat exchange, such as those using social media. Next, we will describe how security of keyless systems can be improved using machine learning – more specifically, how some of the newer models can help detect the difference between an original signal and a signal amplified by an adversary. The third project will be at the intersection of adversarial machine learning and network intrusion detection.

Bio: Dr. Gagan Agrawal is a Professor and Associate Dean of research in the School of Computer and Cyber Sciences at Augusta University. He previously held faculty positions at University of Delaware and Ohio State University. His work in these areas has resulted in more than 275 peer-reviewed publications, significant funding from the National Science Foundation and the Department of Energy and 30 PhD graduates. He has served on the editorial board of four journals and served as program committee co-chair, area chair, or program committee major for many conferences. His notable research contributions include middleware systems for parallelizing data-analytics applications on clusters and other HPC architectures, techniques for managing scientific data, and parallel algorithms for data mining and machine learning.

Sandrine Dudoit, PhD
Associate Dean for the Faculty and Research, Division of Computing, Data Science, and Society,
Professor, Department of Statistics and Division of Biostatistics, School of Public Health, University of California, Berkeley, CA

Talk title: Learning from Data in Single-Cell Transcriptomics

Abstract: The ability to measure gene expression levels for individual cells (vs. pools of cells) is crucial to address many important biological questions, such as the study of stem cell differentiation, the detection of rare mutations in cancer, or the discovery of cellular subtypes in the brain. Single-cell transcriptome sequencing (RNA-Seq) allows the high-throughput measurement of gene expression levels for entire genomes at the resolution of single cells. RNA-Seq studies provide a great example of

the range of questions one encounters in a Data Science workflow, where the data are complex in a variety of ways, there are multiple analysis steps, and drawing on rigorous statistical principles and methods is essential to derive reliable and interpretable biological results. In this talk, I will provide a survey of statistical questions related to the analysis of single-cell RNA-Seq data to investigate the differentiation of stem cells in the brain, including, exploratory data analysis, dimensionality reduction, normalization, expression quantitation, cluster analysis, and the inference of cellular lineages.

Bio: Dr. Sandrine Dudoit is Professor of the Department of Statistics and Professor in the Division of Biostatistics, School of Public, at the University of California, Berkeley. Professor Dudoit's methodological research interests regard high-dimensional inference and include exploratory data analysis (EDA), visualization, loss-based estimation with cross-validation, and multiple hypothesis testing. Much of her methodological work is motivated by statistical inference questions arising in biological research and, in particular, the design and analysis of high-throughput microarray and sequencing gene expression experiments, e.g., single-cell transcriptome sequencing (RNA-Seq) for discovering novel cell types and for the study of stem cell differentiation. She is also interested in statistical computing and, in particular, reproducible research. She is a founding core developer of the Bioconductor Project, an open-source and open-development software project for the analysis of biomedical and genomic data. Professor Dudoit is a co-author of the book Multiple Testing Procedures with Applications to Genomics and a co-editor of the book Bioinformatics and Computational Biology Solutions Using R and Bioconductor. She is Associate Editor of three journals, including The Annals of Applied Statistics and IEEE/ACM Transactions on Computational Biology and Bioinformatics. Professor Dudoit was named Fellow of the American Statistical Association in 2010, Elected Member of the International Statistical Institute in 2014, and Fellow of the Institute of Mathematical Statistics in 2021.

James B D Joshi, PhD
Program Director, Secure and Trustworthy Cyberspace (SaTC), National Science Foundation (NSF) and Professor, University of Pittsburgh, PA, USA

Talk title: Privacy: challenges and road ahead

Abstract: Recent advances in computing and information technologies have enabled a hyper-connected cyberspace that has become an intricate part of our society. Enabled by such connectivity, myriads of ways to collect our personal data, and rapid advances in computing and analytics/AI, we have unprecedented opportunities to solve significant societal problems through data-driven, evidence based approaches. However, such technological advances also present equally difficult set of challenges with regards to the protection of our privacy. In this talk, I will first overview the Secure and Trustworthy Cyberspace (SaTC) program, and then I will discuss current state of privacy-preserving technologies, and discuss various challenges that we need to address to ensure our privacy while benefiting from innovations in computing.

Bio: Dr. James Joshi is a professor of School of Computing and Information at the University of Pittsburgh, and the director/founder of the Laboratory of Education and Research on Security Assured Information Systems (LERSAIS). He is currently serving as a Program Director in the Computer and Network System (CNS) division and its Secure and Trustworthy Cyberspace (SaTC) program at the US National Science Foundation. He also serves as the Co-Chair of the Privacy Interagency Working Group of the Networking and Information Technology R&D (NITRD); he also co-chairs the recently established

NITRD Fast Track Action Committees for: (1) Advancing Privacy Preserving Data Sharing and Analytics, and (2) Digital Assets R&D Agenda. He is an IEEE Fellow, an ACM Distinguished Member, and an IEEE CS Golden Core member. His research interests include access control models, security and privacy of distributed systems and AI/ML, and trust management. He is a recipient of the NSF CAREER award in 2006. He established and managed the NSF CyberCorp Scholarship for Service program at Pitt in 2006. He also established LERSAIS as a NSA designated Center of Academic Excellence in Cyber Defense (both CAE and CAE-R). He has served as a program co-chair and/or general co-chair of several international conferences/workshops. He currently serves as the steering committee chair of co-located conferences IEEE: CIC, TPS and CogMI, and had recently served as the EIC of IEEE Transactions on Services Computing. He has published over 140 articles as book chapters and papers in journals, conferences and workshops, and has served as a special issue editor of several journals including Elsevier Computer & Security, ACM TOPS, Springer MONET, IJCIS, and Information Systems Frontiers. His research has been supported by NSF, NSA/DoD, and Cisco.

Murat Kantarcioglu, Ph.D.
Ashbel Smith Professor of Computer Science, Director, UT Dallas Data Security and Privacy Lab, University of Texas at Dallas, Dallas, TX

Talk title: Securing Big Data in the Age of Artificial Intelligence

Abstract: In the age of big data and Artificial Intelligence (AI), protecting the security and privacy of stored data is paramount for maintaining public trust, accountability and getting the full value from the collected data. Therefore, we need to address security and privacy challenges ranging from allowing access to big data to building novel AI models using the privacy sensitive data. In this talk, I provide an overview of our end-to-end solution framework that addresses these security and privacy challenges arise in the age of AI. We first provide an overview of secure data processing techniques that leverages trusted execution environments for analyzing and protecting stored encrypted data. In addition, we discuss our federated learning framework that is designed to be robust against poisoning attacks and when humans can work with AI to improve decision outcomes and help preventing potential attacks.

Bio: Dr. Murat Kantarcioglu is an Ashbel Smith Professor in the Computer Science Department and Director of the Data Security and Privacy Lab at The University of Texas at Dallas (UTD). He received a PhD in Computer Science from Purdue University in 2005 where he received the Purdue CERIAS Diamond Award for Academic excellence. He is also a faculty associate at Harvard Data Privacy Lab and a visiting scholar at UC Berkeley RISE Labs. Dr. Kantarcioglu's research focuses on the integration of cyber security, data science and blockchains for creating technologies that can efficiently and securely process and share data. His research has been supported by grants including from NSF, AFOSR, ARO, ONR, NSA, and NIH. He has published over 170 peer reviewed papers in top tier venues such as ACM KDD, SIGMOD, ICDM, ICDE, PVLDB, NDSS, USENIX Security and several IEEE/ACM Transactions as well as served as program co-chair for conferences such as IEEE ICDE, ACM SACMAT, IEEE Cloud, ACM CODASPY. Some of his research work has been covered by the media outlets such as the Boston Globe, ABC News, PBS/KERA, DFW Television, and has received multiple best paper awards. He is the recipient of various awards including NSF CAREER award, the AMIA (American Medical Informatics Association) 2014 Homer R Warner Award and the IEEE ISI (Intelligence and Security Informatics) 2017 Technical Achievement

Award presented jointly by IEEE SMC and IEEE ITS societies for his research in data security and privacy. He is also a fellow of AAAS, and IEEE.

Xihong Lin, PhD
Professor of Biostatistics, Coordinating Director of the Program in Quantitative Genomics, T. H. Chan School of Public Health, Harvard University, Boston, MA

Talk title: Ensemble methods for testing a global null with applications to whole genome sequencing studies

Abstract: Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact of no uniformly most powerful test, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power, e.g., the class of alternatives in the scenario of the same effect sign or signal sparsity. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. The key component of the framework is to introduce a certain random procedure in the construction of base tests. We then apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations are conducted to demonstrate type I error control and power gain of the proposed ensemble tests. In an analysis of the WGS data from the Atherosclerosis Risk in Communities (ARIC) study, the ensemble tests demonstrate substantial and consistent power improvement compared to other existing tests. This is joint work with Yaowu Liu.

Bio: Dr. Lin is Professor and former Chair of the Department of Biostatistics, Coordinating Director of the Program in Quantitative Genomics at the Harvard T. H. Chan School of Public Health, and Professor of the Department of Statistics at the Faculty of Arts and Sciences of Harvard University, and Associate Member of the Broad Institute of Harvard and MIT. Dr. Lin is an elected member of the National Academy of Medicine. She received the 2002 Mortimer Spiegelman Award from the American Public Health Association, and the 2006 Committee of Presidents of Statistical Societies (COPSS) Presidents' Award and the 2017 COPSS FN David Award. She is an elected fellow of American Statistical Association (ASA), Institute of Mathematical Statistics, and International Statistical Institute. Dr. Lin's research interests lie in development and application of statistical and computational methods for analysis of massive data from genome, exposome and phenome, and scalable statistical inference and learning for big genomic, epidemiological and health data. Her theoretical and computational statistical research includes statistical methods for testing a large number of complex hypotheses, causal inference, statistical inference for large covariance matrices, prediction models using high-dimensional data, cloud-based statistical computing, and statistical methods for epidemiological studies. Dr. Lin's research has been well supported by various funding agencies, in particular, she was funded by the MERIT Award (R37) (2007-2015) and is an Outstanding Investigator Award (OIA) (R35) (2015-2022) from the National Cancer Institute (NCI).

Bani K. Mallick, PhD
University Distinguished Professor, Department of Statistics, Texas A&M University

Talk title: Modeling complex data using Bayesian Local Models

Talk abstract: All models are not wrong, in fact some of them could be correct, at least locally! Moreover, they are useful! Based on this principle, I will propose Bayesian local models using partitioning. The Bayesian partition model constructs arbitrarily complex models by splitting the covariate space into an unknown number of disjoint regions. Within each region the data are assumed to be generated by a simpler model. The partition can be created using Voronoi Tessellations or Trees. The main challenge is to determine the local regions (partitions) adaptively. I will discuss local models for density regression, survival analysis and spatial prediction. Some theoretical properties of the models will be discussed. I will show simulations and applications to real data analysis where the proposed method will successfully identify the partition structure as well as estimate the local model parameters.

Bio: Dr. Mallick is a Distinguished Professor and Susan M. Arseven `75 Chair in Data Science and Computational Statistics in the Department of Statistics at Texas A&M University in College Station. He is also the Director of the Center for Statistical Bioinformatics. He is a fellow of the American Association for the Advancement of Science, American Statistical Association, Institute of Mathematical Statistics, International Statistical Institute and the Royal Statistical Society. He received the Distinguished research award from Texas A&M University and the Young Researcher award from the International Indian Statistical Association. Dr. Mallick's many areas of research include semiparametric classification and regression, hierarchical spatial modeling, inverse problem, uncertainty quantification and Bioinformatics. He is equally renowned for his ability to do major collaborative research with scientists from myriad fields beyond his own. He has coauthored or co-edited six books and over 150 research publications. He has supervised and mentored 25 PhD students and 6 Post doctorate fellows. Dr. Mallick earned his undergraduate from the Presidency University in Kolkata, MS from the Calcutta University and PhD from the University of Connecticut.

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

Talk title: Privacy, Data Privacy, and Differential Privacy

Abstract: This talk invites curious minds to contemplate the notion of data privacy. It first traces the evasive concept of privacy to a legal right, derived from the frustration of the husband of a socialite attracting tabloids when yellow journalism and film photography became popular in 1890s. More than a century later, the rise of digital technologies and data science has made the issue of data privacy a central concern for essentially all enterprises, from medical research to business applications, and to census operations. Differential privacy (DP), a theoretically elegant and methodologically impactful framework developed in cryptography, is a major milestone in dealing with the thorny issue of properly balancing data privacy and data utility. However, the popularity of DP has brought both hype and scrutiny, revealing several misunderstandings and subtleties that have created confusions even among specialists. The technical part of this talk is therefore devoted to explicating such issues from a statistical

perspective, particularly via the prior-to-posterior semantics of DP. This semantics yields an intuitive statistical interpretation of DP, albeit it does not correspond in general to the commonly understood and desired data privacy protection. The determination of whether the traditional data swapping method is DP demonstrates the subtleties and their consequences when overlooked. Ultimately, the talk aims to highlight the challenges and research opportunities in quantifying data privacy, what DP does and does not protect, and the need to properly analyze DP data. (This talk is based on three articles with James Bailie and Ruobin Gong.)

Bio: Dr. Xiao-Li Meng is the Founding Editor-in-Chief of Harvard Data Science Review and is well known for his depth and breadth in research and his innovation and passion in pedagogy. His interests range from the theoretical foundations of statistical inferences (e.g., the interplay among Bayesian, Fiducial, and frequentist perspectives; frameworks for multi-source, multi-phase and multi- resolution inferences) to statistical methods and computation (e.g., posterior predictive p-value; EM algorithm; Markov chain Monte Carlo; bridge and path sampling) to applications in natural, social, and medical sciences and engineering (e.g., complex statistical modeling in astronomy and astrophysics, assessing disparity in mental health services, and quantifying statistical information in genetic studies). Dr. Li was the former Dean of the Graduate School of Arts and Sciences at Harvard. Dr. Li was named the best statistician under the age of 40 by COPSS (Committee of Presidents of Statistical Societies) in 2001, and he is the recipient of numerous awards and honors for his more than 150 publications in at least a dozen theoretical and methodological areas, as well as in areas of pedagogy and professional development. He has delivered more than 400 research presentations and public speeches on these topics.

Aditya Prakash, PhD
Associate Professor, College of Computing, Georgia Institute of Technology, Atlanta, Georgia

Talk title: New Methods for Robust Time-Series Mining

Abstract: Sequence modeling at scale is an important problem for many domains like traffic prediction, behavior modeling, network analytics and disease forecasting. Improvements in ML models for sequence prediction have led to improvements in tasks across multiple disciplines. However, most time series prediction models only focus on producing accurate so-called 'point' predictions and do not prioritize handling uncertainty and calibrating the (complex, possibly multimodal) predictive distribution itself. This is problematic as providing a trustworthy and reliable estimate is important for real world applications in forecasting, anomaly detection, and more generally dealing with uncertain cases during inference. This problem is exacerbated in spatio-temporal forecasting. In this talk, we discuss some of our recent work in advancing the field of multi-variate forecasting with uncertainty and calibration by bridging deep sequential models with Gaussian processes and also using neural models to handle data revisions in real-time. We will also talk about our recent work on using energy-based models for end-end decision-making using forecasts.

Bio: B. Aditya Prakash is an Associate Professor in the College of Computing at the Georgia Institute of Technology ("Georgia Tech"). He received a Ph.D. from the Computer Science Department at Carnegie Mellon University in 2012, and a B.Tech (in CS) from the Indian Institute of Technology (IIT) -- Bombay in 2007. He has published one book, more than 80 papers in major venues, holds two U.S. patents and has given several tutorials at leading conferences. His work has also received multiple best-of-conference, best paper and travel awards. His research interests include Data Science, Machine Learning and AI, with

emphasis on big-data problems in large real-world networks and time-series, with applications to computational epidemiology/public health, urban computing, security and the Web. Tools developed by his group have been in use in many places including ORNL, Walmart and Facebook. He has received several awards such as a Facebook Faculty Award, the NSF CAREER award and was named as one of 'AI Ten to Watch' by IEEE. His work has also won awards in multiple data science challenges (e.g the Facebook COVID19 Symptom Challenge) and been highlighted by several media outlets/popular press like FiveThirtyEight.com. He is also a member of the infectious diseases modeling MIDAS network and core-faculty at the Center for Machine Learning (ML@GT) and the Institute for Data Engineering and Science (IDEaS) at Georgia Tech.

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

Talk title: Bayesian Multiset Canonical Correlation Analysis through Latent Factors

Abstract: In biomedical studies, multiple sourced data with varied features are often observed on the same subjects. The complexity poses challenges for data analysis and interpretation. We refer to each source as providing a "set" of data and seek to learn the dependence structure of this multiset data. We propose a Bayesian framework that uses latent factors to decompose data into four conceptual parts: the joint structure induced by the latent factors shared across all the sets, partial association structures induced by the latent factors shared by the associated subsets of the multiset data, individual structure induced by latent factors unique to each set, and noise. Our method enables flexible selection of latent factors and discovery of association structure as well as variable selection in the joint structure (and, if desired, partial association structure). We assess variable importance using the posterior probability of selection and derive the pairwise correlations between the multiset data. We illustrate performance using simulation. The methods are applied to data from a study of the temporomandibular joint to explore association between skull and multiple muscle attachment measurements. This is a joint work with Yuxi Zhao, Xin Henry Zhang, Shuchun Sun, and Hai Yao.

Bio: Elizabeth is the Duncan McLean and Pearl Levine Fairweather Professor of Statistics and Distinguished Research Professor in the Department of Statistics at Florida State University. She received her PhD in Statistics from Carnegie Mellon University in Pittsburgh, PA and held positions on the faculty at Cornell University and the Medical University of South Carolina prior to joining FSU. She has held visiting positions at Stanford University, the Biometry Research Group of the National Cancer Institute in Bethesda, MD, the Statistics and Applied Mathematical Science Institute in Raleigh, NC and as the David C. Jordan Visiting Scholar at AbbVie, Inc. At FSU, she directs the program in Statistical Data Science and is involved in several clinical trials, including a SMART study, with the FSU Autism Institute. Elizabeth is a Fellow of the American Statistical Association.

Invited Panel leaders

Doug Miller, MD

Professor, Department of Medicine: Cardiology and Department of Population Health Sciences, Medical College of Georgia, Augusta University, Augusta, GA

Bio: A medical graduate of McGill University, Dr. Miller was trained in cardiovascular medicine at Emory University and Harvard University Having published over 200 original papers and book chapters in the field, Dr. Miller enjoys a global reputation as a cardiologist and academic medicine leader. He has served as the Dean of three research-intensive medical schools in the U.S. and Canada, and as a member of their national MD program accrediting bodies. He advises several global health policy and biomedical business organizations. An entrepreneur holding patents in the fields of new drug development, medical imaging and artificial intelligence (AI), his high impact journal publications and invited lectures make him a leading authority on AI applications to healthcare and medical education.

Jennifer Lewis Priestley, PhD
Professor Emerita of Statistics and Data Science, Kennesaw State University
Chief Data Officer, Flock Specialty Finance

Bio: Dr. Priestley is a Professor Emeritus of Statistics and Data Science, former Executive Director of Analytics and Data Science Institute, former Associate Dean of the Graduate College at Kennesaw State University, Kennesaw, GA and a Chief Data Officer, Flock Specialty Finance. She received a Ph.D. from Georgia State University, an MBA from The Pennsylvania State University, and a BS from Georgia Institute of Technology. She architected the first Ph.D. Program in Data Science, which was launched in February 2015. In 2012, the SAS Institute recognized Dr. Priestley as the 2012 Distinguished Statistics Professor of the Year. Datanami recognized Dr. Priestley as one of the top 12 "Data Scientists to Watch in 2016." Dr. Priestley has been a featured international speaker at the World Statistical Congress, the South African Statistical Association, the Nelson Mandela University, the Federal Reserve Bank, SAS Global Forum, Big Data Week, Technology Association of Georgia, Data Science ATL, the Atlanta CEO Council, Predictive Analytics World, INFORMS and dozens of academic and corporate conferences addressing issues related to the evolution of data science, women in data science and ethical data science. Prior to receiving a Ph.D. in Statistics, Dr. Priestley worked in the Financial Services industry for 11 years. Her positions included Vice President of Business Development for VISA EU in London, as well as Regional Vice President for MasterCard US and as a senior consultant with Accenture's strategic services group.