Statistics Winter Workshop: 2026

Frontiers in Learning Under Data Heterogeneity

Friday, January 16, 2026

9:00am-4:00pm

J. Wayne Reitz Union Chamber Room (Ground Level)

Saturday, January 17, 2026

9:00am to 11:30am

J. Wayne Reitz Union Chamber Room (Ground Level)

Venue

J. Wayne Reitz Union
655 Reitz Union Drive
University of Florida
Campus Gainesville, FL
32611

Meeting Rooms

Presentations:	Chamber
Refreshments:	Room G320
Posters: Room:	Room G310

Parking

The Visitor Welcome Center and Bookstore parking garage is located at the Reitz Union, at the corner of Museum Road and Reitz Union Drive. The garage is an unrestricted pay facility available to all members of the university community. This garage can accommodate 300 vehicles. There are 45 short-term parking spaces located in the garage.

The garage hours of operation are Monday through Friday, 7:30 am to 4:30 pm. Short-term and daily fees apply during this time. The garage may be used during non-operating hours for short-term parking, free of charge.

Program — Day One
Friday, January 16, 2026

8:30 AM Breakfast (Room G320)

9:00 AM Welcome-- Mike Daniels
Professor and Chair, Department of Statistics, University of Florida

9:10 AM Annie Qu
Professor, University of California-Irvine
Representation Retrieval Learning for Heterogeneous Data Integration

9:55 AM Arya Mazumder
Chair and Professor, University of California-San Diego
Continual Learning with Gradient Descent for Neural Networks

10:40 AM Coffee Break (Room G320)

10:55 AM Xinwei Shen
Assistant Professor, University of Washington-Seattle
Generalization through the Lens of Distributional Learning

11:40 AM Lunch (Room G320)

1:10 PM Yang Feng
Professor, New York University
Regularized Fine-Tuning for Representation Multi-Task Learning: Adaptivity, Minimaxity, and Robustness

1:55 PM Coffee Break (Room G320)

2:10 PM Hamsa Bastani
Associate Professor, University of Pennsylvania
Beating the Winner's Curse via Inference-Aware Policy Optimization

3:00 PM Poster Session (Room G310)

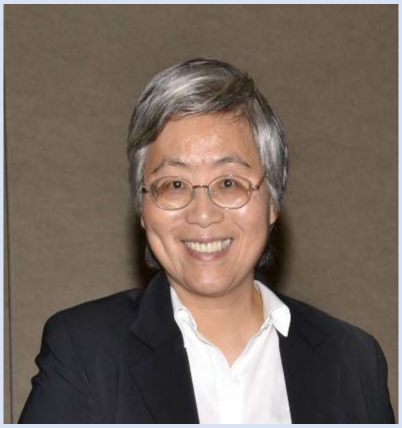
Program — Day Two
Saturday, January 17, 2026

8:30 AM	Breakfast (Room G310)
9:00 AM	Rui Duan Assistant Professor, Harvard University ***Collaborative Learning Methods for Multi-Site Data Integration***
9:45 AM	Lingzhou Xue Professor, Pennsylvania State University ***Statistical Accuracy-Communication Trade-off in Personalized Federated Learning with Minimax Guarantees***
10:30 AM	Coffee Break (Room G310)
10:45 AM	Yuekai Sun Associate Professor, University of Michigan ***How to train your LLM***

Guest Speakers

Annie Qu

Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara



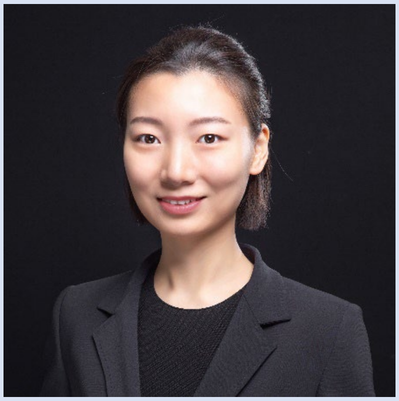
Title: Representation Retrieval Learning for Heterogeneous Data Integration

Abstract: In the era of big data, large-scale, multi-modal datasets are increasingly ubiquitous, offering unprecedented opportunities for predictive modeling and scientific discovery. However, these datasets often exhibit complex heterogeneity, such as covariate shift, posterior drift, and missing modalities which can hinder the accuracy of existing prediction algorithms. To address these challenges, we propose a novel Representation Retrieval (R2) framework, which integrates a representation learning module (the representer) with a sparsity-induced machine learning model (the learner). Moreover, we introduce the notion of "integrativeness" for representers, characterized by the effective data sources used in learning representers, and propose a Selective Integration Penalty (SIP) to explicitly improve the property. Theoretically, we demonstrate that the R2 framework relaxes the conventional full-sharing assumption in multi-task learning, allowing for partially shared structures, and that SIP can improve the convergence rate of the excess risk bound. Extensive simulation studies validate the empirical performance of our framework, and applications to two real-world datasets further confirm its superiority over existing approaches.

Biography: Annie Qu is Professor at Department of Statistics and Applied Probability, University of California, Santa Barbara starting July 2025. She received her Ph.D. in Statistics from the Pennsylvania State University in 1998. Qu's research focuses on solving fundamental issues regarding structured and unstructured large-scale data and developing cutting-edge statistical methods and theory in machine learning and algorithms for personalized medicine, text mining, recommender systems, medical imaging data, and network data analyses for complex heterogeneous data. Dr. Qu was a Data Science Founder Professor of Statistics and the Director of the Illinois Statistics Office at the University of Illinois at Urbana-Champaign during her tenure in 2008-2019, and Chancellor's Professor at UC Irvine in 2020-2025. She was a recipient of the NSF Career award from 2004 to 2009. She is a Fellow of the Institute of Mathematical Statistics (IMS), the American Statistical Association, and the American Association for the Advancement of Science. She is also a recipient of IMS Medallion Award and Lecturer in 2024. She serves as Journal of the American Statistical Association Theory and Methods Co-Editor from 2023 to 2025, IMS Program Secretary from 2021 to 2027 and ASA Council of Sections of Governing Board Chair in 2025. She is the recipient of the 2025 Carver Medal of IMS.

Rui Duan

Assistant Professor, Department of Biostatistics and Department of Epidemiology, Harvard University



Title: Collaborative Learning Methods for Multi-Site Data Integration

Abstract: In this talk, we introduce methods for leveraging models trained across multiple institutions to improve performance in new environments. We consider two settings: one where limited local training data are available and another where no local training data exist. For each case, we develop tailored strategies, including transfer learning approaches and model aggregation techniques. We further provide theoretical analysis of their asymptotic performance in high-dimensional regimes. The methods are evaluated through simulation studies and applied to real-world datasets, combining information from multiple biobank and clinical systems.

Biography: Rui Duan, PhD is an Assistant Professor in the Department of Biostatistics and the Department of Epidemiology, at the Harvard T.H. Chan School of Public Health, with additional affiliations at the Harvard Data Science Initiative, and the Center for Precision Psychiatry at Massachusetts General Hospital. Her research develops statistical and machine learning methods for the integration and analysis of large-scale biomedical data across multiple institutions. She has led methodological advances in federated learning, transfer learning, and data fusion, with a focus on applications to electronic health records, biobank-scale genetic data, and mental health outcomes.

Xinwei Shen

Assistant Professor, Mathematical Statistics, University of Washington

Title: Generalization through the Lens of Distributional Learning

Abstract: Generative AI has achieved remarkable success across various domains, but its potential for addressing statistical challenges is less explored. This talk focuses on generalization beyond the observed data distribution, including problems such as extrapolation, distribution shifts, and causal inference. These tasks require generalizing beyond what has been directly observed. We propose tackling such problems through a distributional perspective: instead of fitting only point summaries like conditional means, we estimate the entire distribution of the observed data. While natural from an identifiability standpoint, this approach has been underutilized in estimation. We introduce engression, a distributional learning method that balances the flexibility of generative models with conceptual simplicity. Under various structural settings, we show how engression can be adapted to out-of-support covariate shifts, conditional distribution shifts, and causal effect estimation. We hope to demonstrate how estimating the distribution yields stronger identification and hence better generalization than approaches targeting only low-dimensional summaries.

Biography: Xinwei Shen is an assistant professor in the Department of Statistics at the University of Washington. She was a postdoctoral researcher at ETH Zürich working with Peter Bühlmann and Nicolai Meinshausen and an incoming assistant professor at University of Washington. She obtained her PhD at HKUST advised by Tong Zhang and a Bachelor of Science degree at Fudan University. Her research interests include distributional learning, causality, robustness, and applications in climate science.

Yang Feng

Professor, Department of Biostatistics, New York University

Title: Regularized Fine-Tuning for Representation Multi-Task Learning: Adaptivity, Minimaxity, and Robustness

Abstract: We study multi-task linear regression for a collection of tasks that share a latent, low-dimensional structure. Each task's regression vector belongs to a subspace whose dimension, denoted intrinsic dimension, is much smaller than the ambient dimension. Unlike classical analyses that assume an identical subspace for every task, we allow each task's subspace to drift from a single reference subspace by a controllable similarity radius, and we permit an unknown fraction of tasks to be outliers that violate the shared-structure assumption altogether. Our contributions are threefold. First, adaptivity: we design a penalized empirical-risk algorithm and a spectral method. Both algorithms automatically adjust to the unknown similarity radius and to the proportion of outliers. Second, minimaxity: we prove information-theoretic lower bounds on the best achievable prediction risk over this problem class and show that both algorithms attain these bounds up to constant factors; when no outliers are present, the spectral method is exactly minimax-optimal. Third, robustness: for every choice of similarity radius and outlier proportion, the proposed estimators never incur larger expected prediction error than independent single-task regression, while delivering strict improvements whenever tasks are even moderately similar and outliers are sparse. Additionally, we introduce a thresholding algorithm to adapt to an unknown intrinsic dimension. We conduct extensive numerical experiments to validate our theoretical findings.

Biography: Yang Feng is a Professor of Biostatistics in the School of Global Public Health at New York University, where he is also affiliated with the Center for Data Science. He earned his Ph.D. in Operations Research from Princeton University in 2010. His research centers on the theoretical and methodological foundations of machine learning, high-dimensional statistics, network models, and nonparametric statistics, with applications in Alzheimer's disease prognosis, cancer subtype classification, genomics, electronic health records, and biomedical imaging, enabling more accurate models for risk assessment and clinical decision-making. His work has been supported by grants from the National Institutes of Health and the National Science Foundation (NSF), including the NSF CAREER Award. He currently serves as Associate Editor for several leading journals, including the Journal of the American Statistical Association (JASA), the Journal of Business & Economic Statistics, the Journal of Computational & Graphical Statistics, and the Annals of Applied Statistics. In addition, he will serve as Review Editor for JASA and The American Statistician from 2026 to 2028. His professional recognitions include being named as a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, as well as an elected member of the International Statistical Institute.

Yuekai Sun

Yuekai Sun, Associate Professor, Department of Statistics, University of Michigan

Title: How to train your LLM

Abstract: K2-V2 is a fully/totally open (open-weights, open training pipeline, open training data) 70B LLM that is SOTA for fully/totally open LLMs and LLMs in its size bracket. In this talk, I will discuss some statistical challenges and problems we encountered during its development, emphasizing problems that statisticians are well-equipped to tackle.

Biography: Yuekai Sun is an associate professor of statistics at the University of Michigan and a researcher at the Institute of Foundation Models (IFM), where he helps train fully/totally open LLMs. His research leverages statistical science to make AI smarter and more reliable. Some topics of recent interest include: AI evaluation, algorithmic fairness, transfer learning. Before coming to Michigan, Yuekai obtained his PhD and BA in computational math from Stanford University, advised by Michael Saunders and Jonathan Taylor, and Rice University respectively. Yuekai was born in Hangzhou, but he spent his formative years in Singapore and the SF Bay Area.

Arya Mazumdar

Arya Mazumdar Chair and Professor in AI, University of California San Diego

Title: Continual Learning with Gradient Descent for Neural Networks

Abstract: Continual learning, the ability of a model to adapt to an ongoing sequence of tasks without forgetting the earlier ones, is a central goal of artificial intelligence. To shed light on its underlying mechanisms, we analyze the limitations of continual learning in a tractable yet representative setting. In particular, we study one-hidden-layer quadratic neural networks trained by gradient descent on an XOR cluster dataset with Gaussian noise, where different tasks correspond to different clusters with orthogonal means. Our results obtain bounds on the rate of forgetting during train and test-time in terms of the number of iterations, the sample size, the number of tasks, and the hidden-layer size. Our results reveal interesting phenomena on the role of different problem parameters in the rate of forgetting. Numerical experiments across diverse setups confirm our results, demonstrating their validity beyond the analyzed settings. Time permitting, we will also discuss some results on Transfer learning.

Joint work with Hossein Taheri.

Biography: Arya Mazumdar is a Halıcıoğlu Data Science Institute Endowed Chair Professor in AI at University of California San Diego, and the Deputy Director and Associate Director of Research of the NSF-funded AI Institute for Learning-Enabled Optimization at Scale (TILOS) Institute. He is also a Co-PI and UCSD Site Lead of NSF-TRIPODS Phase II institute, EnCORE. His research areas cover algorithmic and statistical aspects of machine learning, error correcting codes, optimization, and signal processing. Currently, his main focus is machine learning, specifically computational aspects of statistical estimation/inference and distributed learning. Previously, his work in information and coding theory led to major advancements in the problem of local data recovery and sparse signal recovery. Awards received by Prof. Mazumdar include the 2011 ECE Distinguished Dissertation Award and 2025 ECE Distinguished Alumni Award (both from Univ. of Maryland), EURASIP Best Paper Award, IEEE ISIT Jack Keil Wolf paper award, and a 2015 NSF CAREER Award.

Hamsa Bastani

Hamsa Bastani, Associate Professor, University of Pennsylvania

Title: Beating the Winner's Curse via Inference-Aware Policy Optimization

Abstract:

There has been a surge of recent interest in automatically learning policies to target treatment decisions based on rich individual covariates. A common approach is to train a machine learning model to predict counterfactual outcomes, and then select the policy that optimizes the predicted objective value. In addition, practitioners also want confidence that the learned policy has better performance than the incumbent policy according to downstream policy evaluation. However, due to the winner's curse—an issue where the policy optimization procedure exploits prediction errors rather than finding actual improvements—predicted performance improvements are often not substantiated by downstream policy optimization. To address this challenge, we propose a novel strategy called inference-aware policy optimization, which modifies policy optimization to account for how the policy will be evaluated downstream. Specifically, it optimizes not only for the estimated objective value, but also for the chances that the policy will be statistically significantly better than the observational policy used to collect data. We mathematically characterize the Pareto frontier of policies according to the tradeoff of these two goals. Based on our characterization, we design a policy optimization algorithm that uses machine learning to predict counterfactual outcomes and then plugs in these predictions to estimate the Pareto frontier; then, the decision-maker can select the policy that optimizes their desired tradeoff, after which policy evaluation can be performed on the test set as usual. Finally, we perform simulations to illustrate the effectiveness of our methodology. Joint work with Osbert Bastani and Bryce McLaughlin.

Biography:

Hamsa Bastani is an Associate Professor of Operations, Information, and Decisions at the Wharton School, University of Pennsylvania. Her research focuses on developing novel machine learning algorithms for data-driven decision-making, with applications to healthcare operations, social good, and revenue management. Her work has received several recognitions, including the Wagner Prize for Excellence in Practice (2021), the Pierskalla Award for the best paper in healthcare (2016, 2019, 2021), the Behavioral OM Best Paper Award (2021), as well as first place in the George Nicholson and MSOM student paper competitions (2016). She previously completed her PhD at Stanford University and spent a year as a Herman Goldstine postdoctoral fellow at IBM Research.

Lingzhou Xue

Professor of Statistics, Pennsylvania State University

Title: Statistical Accuracy-Communication Trade-off in Personalized Federated Learning with Minimax Guarantees

Abstract:

Personalized federated learning (PFL) offers a flexible framework for aggregating information across distributed clients with heterogeneous data. This work considers a personalized federated learning setting that simultaneously learns global and local models. While purely local training has no communication cost, collaborative learning among the clients can leverage shared knowledge to improve statistical accuracy, presenting an accuracy-communication trade-off in personalized federated learning. However, the theoretical analysis of how personalization quantitatively influences sample and algorithmic efficiency and their inherent trade-off is largely unexplored. This paper contributes to filling this gap by providing a quantitative characterization of the personalization degree on the tradeoff. The results further offer theoretical insights for choosing the personalization degree. As a side contribution, we establish the minimax optimality in terms of statistical accuracy for a widely studied PFL formulation. The theoretical result is validated on both synthetic and real-world datasets, and its generalizability is verified in a non-convex setting.

Biography: Lingzhou Xue is a Professor of Statistics at Penn State. He received his B.Sc. in Statistics from Peking University in 2008 and his Ph.D. in Statistics from the University of Minnesota in 2012. He was a postdoctoral research associate at Princeton University from 2012-2013. His research interests include high-dimensional statistics, nonparametric statistics, statistical and machine learning, large-scale optimization, and statistical modeling in biomedical, environmental, and social sciences. His recent research focuses on causal inference, federated learning, graphical models, high-dimensional inference, optimal transport, random objects, and reinforcement learning. He is a dedicated mentor to Ph.D. students and postdoctoral researchers, and five of his former advisees have become tenure-track faculty members in statistics. He became an Elected Fellow of the Institute of Mathematical Statistics (IMS) in 2024, an Elected Fellow of the American Statistical Association (ASA) in 2023, and an Elected Member of the International Statistical Institute (ISI) in 2016. He received the inaugural Committee of Presidents of Statistical Societies (COPSS) Emerging Leader Award in 2021, the inaugural Bernoulli Society New Researcher Award in 2019, and the International Consortium of Chinese Mathematicians Best Paper Award in 2019.Currently, he serves as an Associate Editor for the *Journal of the American Statistical Association, Annals of Applied Statistics, Stat, ACM Transactions on Probabilistic Machine Learning,* and *Data Science in Science*.

Poster Winners

Chen Cheng, University of Chicago

Title: Quantifying population-level robustness by distribution shifts

Abstract: We revisit the stability of optimizers in statistical estimation and stochastic optimization problems, but instead of providing guarantees on the stability of the minimizers themselves, we investigate what shifts to the underlying data-generating process perturb solutions the most. To do so, we develop new mathematical tools for stability analyses, with guarantees beyond typical differentiable problems. We also make connections with statistical hypothesis testing and discovery, showing how these new results provide certificates of validity---or potential invalidity---of statistical estimate. We apply this framework to both simulated data and real-world datasets to evaluate the impact of distribution shifts with uncertainty quantification.

Jae Ho Chang, Ohio State University

Title: TBD

Abstract: TBD

Bhaskar Ray, North Carolina State University

Title: PACE: Privacy Aware Collaborative Estimation for Heterogeneous GLMs

Abstract: With sensitive data collected across various sites, restrictions on data sharing can hinder statistical estimation and inference. The seminal paper on Federated Learning proposed Federated Averaging (FedAvg) to perform Maximum Likelihood estimation. However, FedAvg and other algorithms for parameter estimation can lead to erroneous estimation or fail to converge under model heterogeneity across sites. We propose a novel method of parameter estimation for a broad class of Generalized Linear Models, where sites are partitioned into unobserved clusters based on the parameter value of the data distribution. It accounts for uncertainties in both the local MLE and the optimization iterates. We provide a theoretical analysis of the method encompassing both sources of uncertainty to provide non-asymptotic risk bounds. We conduct a hypothesis test-type classification based on one-shot estimation and utilize the inference to conduct a decentralized collaborative estimation, improving upon local estimation with high probability. We also prove asymptotic accuracy of the clustering algorithm and the consistency of the estimates. We validate our results with simulation studies.

Yizhe Ding, Pennsylvania State University

Title: TBD

Abstract: TBD

Sijian Fan, University of South Carolina

Title: Binary Inductive Matrix Completion with Spike-and-slab Group Lasso.

Abstract: Inductive matrix completion (IMC) leverages side information on rows and columns to improve prediction from partially observed matrices, but effective feature selection remains challenging in high-dimensional settings. This is particularly important in applications such as drug repositioning, where large-scale genomic and clinical features can aid prediction but vary widely in relevance. Existing feature-selective IMC methods, such as sparse-group penalty IMC, impose uniform shrinkage and assume global feature effects, which can be restrictive for binary outcomes. We propose binary inductive matrix completion with spike-and-slab group lasso (BiSSGL), a Bayesian framework that performs adaptive, group-structured feature selection for binary matrix completion. By inducing sparsity in latent factors while allowing heterogeneous feature effects, BiSSGL improves predictive performance and identifies informative side features. Simulation studies demonstrate its advantages, and we discuss applications to drug–disease association prediction and extensions to more complex settings.

Shubhangi Ghosh, Columbia University

Title: TBD

Abstract: TBD

Seunghyun Lee, Columbia University

Title: CLT in high-dimensional Bayesian linear regression

Abstract: This poster discusses posterior inference for linear statistics in high-dimensional Bayesian linear regression with product priors. In contrast to the vast literature under a contracting posterior, we consider a regime where neither the likelihood nor the prior dominates the other. This non-contracting regime is motivated by modern high-dimensional datasets with a low signal-to-noise ratio. We take a first step towards understanding limit distributions by providing posterior CLTs for the linear statistic and its posterior mean. Analogous to the Bernstein-von Mises theorem for contractive settings, the resulting limiting distributions are Gaussian, but heavily depend on the prior and center around the Mean-Field approximation of the posterior. As an application, we derive asymptotic coverage of posterior credible intervals when the prior is mis-specified**.**

Sponsors

McClave + Associates

Infotech

UF Department of Statistics

UF College of Liberal Arts and Sciences

UF Artificial Intelligence and Informatics Research Institute (AIIRI)

National Science Foundation