Statistical Foundations of Latent Structure and Representation Learning

Columbia University • November 4–5, 2026

Each talk is 35 minutes including Q&A.

School of Social Work (SSW) Building
1255 Amsterdam Avenue, New York, NY 10027

Wednesday, November 4

Morning talks: **SSW C06** | Afternoon talks: **SSW 903**

8:30 am	Breakfast
8:55–9:00 am	Opening remarks: Yuqi Gu
9:00–9:35 am	**Bin Yu** | UC Berkeley Not All Interpretations Are Created Equal: Principled Evaluation of AI Interpretability
9:35–10:10 am	**Carey Priebe** | Johns Hopkins University Performance guarantees for LLMs
10:10–10:45 am	**Yuxin Chen** | University of Pennsylvania Transformers Meet In-Context Learning: A Universal Approximation Theory
10:45–11:10 am	Coffee break
11:10–11:45 am	**Jianqing Fan** | Princeton University Factor Informed Double Deep Learning For Average Treatment Effect Estimation
11:45 am–12:20 pm	**Florentina Bunea** | Cornell University Softmax-Structured Gaussian Mixtures for Distribution Ensembles
12:20–1:40 pm	Lunch
1:40–2:15 pm	**Lorenzo Mauri** | Duke University Scalable Bayesian Learning of Latent Structure: Integrating Data and Biological Knowledge
2:15–2:50 pm	**Hongtu Zhu** | UNC Chapel Hill FAST-Brain: A Flow-Aligned Spatio-Temporal Surrogate Brain Model
2:50–3:15 pm	Coffee break
3:15–3:50 pm	**Kun Zhang** | Carnegie Mellon University Causal representation learning and causal generative AI
3:50–4:25 pm	**Yihong Gu** | Harvard University Unveiling Invariant and Transferable Latent Factors Across Heterogeneous Environments via ATLAS
4:25–4:45 pm	Transition to poster session
4:45–6:15 pm	Early-career researcher poster session • SSW 214

Thursday, November 5

Morning talks: **SSW C06** | Afternoon talks: **SSW 903**

8:30 am	Breakfast
9:00–9:35 am	**Annie Qu** | UC Santa Barbara Meta Fusion: A Unified Framework For Multimodal Fusion with Adaptive Mutual Learning
9:35–10:10 am	**Tracy Ke** | Harvard University Topic-aware Ranking with Text-estimated Context
10:10–10:40 am	Coffee break
10:40–11:15 am	**Jiashun Jin** | Carnegie Mellon University Estimating weak spikes with Newton and Prony
11:15–11:50 am	**George Michailidis** | UCLA Learning Latent Representations and Structured Dynamics in High-Dimensional Time Series
11:50 am–1:40 pm	Lunch
1:40–2:15 pm	**Hui Zou** | Johns Hopkins University Foe or Ally: The Role of Dimensionality in Prediction with Measurement Errors
2:15–2:50 pm	**Joseph Feldman** | Washington University in St. Louis Scalable Independent Factor Analyzers using Spectral Initialization for Evaluating Large Language Models
2:50–3:25 pm	**Gongjun Xu** | University of Michigan Beyond Vintage Rotation: Bias-Free Sparse Representation Learning with Oracle Inference
3:25 pm	Closing remarks, followed by coffee

Talk abstracts

Wednesday, November 4, 2026

Bin Yu | UC Berkeley

9:00–9:35 am | SSW C06

Not All Interpretations Are Created Equal: Principled Evaluation of AI Interpretability

As AI transitions to agentic sociotechnical deployments, uninterpretable models pose serious threats to reliability, public trust, and contestability. However, not all interpretation methods are created equal: for interpretability to enhance deployment safety and effectiveness, interpretation methods must be evaluated as a whole, rather than validated through isolated interpretation instances. In this paper, we advocate elevating the evaluation of interpretation methods as a core assessment metric for AI system deployment, with clearly stated principles. To this end, we propose TRUST, a principled framework that builds upon Veridical Data Science (VDS) and the Predictive, Descriptive and Relevance (PDR) foundations. To differentiate interpretation methods effectively, TRUST systematically grades interpretation methods across five progressive tiers: Task-Aligned PDR Foundations, Rigorous Epistemic Grounding, Unifying Semantic Understanding, Scope Transferability, and Targeted Prospective Validation. We present a start-to-finish case study of the iterative Random Forest (iRF) family under TRUST and contrast with Generative Causal Testing (GCT) and other interpretation methods that are not able to clear different tiers of TRUST. Ultimately, TRUST principles foster contestability and epistemic grounding across both human-machine and fully autonomous agentic workflows.

Carey Priebe | Johns Hopkins University

9:35–10:10 am | SSW C06

Performance guarantees for LLMs

Performance guarantees for LLMs are essential to myriad critical applications, enabling for any application for which quantifiable confidence in LLM performance is nonnegotiable. We develop theory and methods for providing performance guarantees for LLMs in the context of a map between vector embeddings. The Data Kernel Perspective Space (DKPS) provides a framework to generate empirically validated theoretical predictions allowing guaranteed performance quantification for these transformative models. I will present probability, statistics and linear algebra for DKPS, and illustrative examples.

Yuxin Chen | University of Pennsylvania

10:10–10:45 am | SSW C06

Transformers Meet In-Context Learning: A Universal Approximation Theory

Modern large language models are capable of in-context learning, the ability to perform new tasks at inference time using only a handful of input-output examples in the prompt, without any fine-tuning or parameter updates. We develop a universal approximation theory to better understand how transformers enable in-context learning. For any class of functions (each representing a distinct task), we demonstrate how to construct a transformer that, without any further weight updates, can perform reliable prediction given only a few in-context examples. In contrast to much of the recent literature that frames transformers as algorithm approximators — i.e., constructing transformers to emulate the iterations of optimization algorithms as a means to approximate solutions of learning problems — our work adopts a fundamentally different approach rooted in universal function approximation. This alternative approach offers approximation guarantees that are not constrained by the effectiveness of the optimization algorithms being approximated, thereby extending far beyond convex problems and linear function classes. Our construction sheds light on how transformers can simultaneously learn general-purpose representations and adapt dynamically to in-context examples.

Talk abstracts

Wednesday, November 4, 2026

Jianqing Fan | Princeton University

11:10–11:45 am | SSW C06

Factor Informed Double Deep Learning For Average Treatment Effect Estimation

We investigate the problem of estimating the average treatment effect (ATE) under a very general setup where the covariates can be high-dimensional, highly correlated, and can have sparse nonlinear effects on the propensity and outcome models. We present the use of a Double Deep Learning strategy for estimation, which involves combining recently developed factor-augmented deep learning-based estimators, FAST-NN, for both the response functions and propensity scores to achieve our goal. By using FAST-NN, our method can select variables that contribute to propensity and outcome models in a completely nonparametric and algorithmic manner and adaptively learn low-dimensional function structures through neural networks. Our proposed novel estimator, FIDDLE (Factor Informed Double Deep Learning Estimator), estimates ATE based on the framework of augmented inverse propensity weighting AIPW with the FAST-NN-based response and propensity estimates. FIDDLE consistently estimates ATE even under model misspecification and is flexible to also allow for low-dimensional covariates. Our method achieves semiparametric efficiency under a very flexible family of propensity and outcome models. We present extensive numerical studies on synthetic and real datasets to support our theoretical guarantees and establish the advantages of our methods over other traditional choices, especially when the data dimension is large. (Joint work with Soham Jana, Sanjeev Kulkarni, and Qishuo Yin)

Florentina Bunea | Cornell University

11:45 am–12:20 pm | SSW C06

Softmax-Structured Gaussian Mixtures for Distribution Ensembles

We consider the problem of uncovering interpretable latent structure shared across an ensemble of complex probability distributions, while allowing for differences among its members. We introduce Softmax-Structured Gaussian Mixtures (SSGM), a model that combines the flexibility of a common Gaussian dictionary with a softmax-mixture parametrization of its weights. The resulting model admits two equivalent representations: as mixtures of many Gaussian components, or as mixtures of a much smaller number of shared, potentially complex latent densities. These representations provide complementary interpretations of the ensemble. We then investigate whether useful latent information can be recovered without fitting the entire softmax mixture. This talk will present a new procedure for screening softmax mixtures for near-purity, based on directly estimating a functional of the latent mixing measure, without estimating its atoms or weights, or knowing the number of components. We establish finite-sample guarantees for the screening criterion, without requiring minimum separation between mixture atoms. For near-pure mixtures, we further derive rates for dominant-atom estimation that adapt to the geometry of the mixture, and extend these methods to ensembles of softmax mixtures. We discuss the implications for learning shared latent representations, with contextual embeddings of text providing one motivating application.

Talk abstracts

Wednesday, November 4, 2026

Lorenzo Mauri | Duke University

1:40–2:15 pm | SSW 903

Scalable Bayesian Learning of Latent Structure: Integrating Data and Biological Knowledge

Latent factor models provide a powerful framework for learning structure in high-dimensional data, but scientific applications demand more than dimension reduction. How can we integrate heterogeneous data sources, incorporate existing scientific knowledge, and quantify uncertainty while keeping computation tractable? I will discuss two complementary approaches motivated by these challenges in 'omics studies. The first addresses inference on covariance structure within and across multiple data views. Spectral estimation and alignment of latent factors accommodate differences in dimensionality and signal strength across views, while subsequent Bayesian inference yields computationally efficient covariance estimates and calibrated uncertainty. The second incorporates biological pathway annotations through structured priors on factor loadings, encouraging interpretable representations while allowing the discovery of variation unexplained by existing annotations. Both approaches combine spectral estimation of latent factors with tractable Bayesian inference, avoiding Markov chain Monte Carlo and using data-adaptive regularization. I will discuss the statistical justification for this strategy, including uncertainty calibration after factor estimation, and illustrate its practical implications through applications to cancer multi-omics and host-response transcriptomics. Together, these developments show how latent variable methods can combine scalable computation, scientific structure, and principled inference. Based on joint work with David B. Dunson, Federica Stolf, Amy H. Herring, and Cameron Miller.

Hongtu Zhu | UNC Chapel Hill

2:15–2:50 pm | SSW 903

FAST-Brain: A Flow-Aligned Spatio-Temporal Surrogate Brain Model

Modeling resting-state functional magnetic resonance imaging (rs-fMRI) data is crucial for understanding brain-wide neural activity. However, traditional methods struggle to capture complex temporal dynamics over long horizons, to account for the brain's anatomical spatial structure, and to model high-dimensional ambient signals that lie on a low-dimensional intrinsic subspace. We propose FAST-Brain, a unified flow-aligned spatio-temporal surrogate brain model that addresses all three challenges. At its core is a flow-aligned generative framework that directly predicts the clean blood-oxygen-level-dependent (BOLD) signal, paired with a graph convolutional network that captures spatial structural constraints and a Transformer that models long-range temporal dependencies. Theoretically, we show that under a low-dimensional subspace assumption, the approximation error of our model scales with the intrinsic dimension rather than the ambient dimension, which justifies our direct modeling of the BOLD signal. Extensive experiments on synthetic and Human Connectome Project datasets demonstrate that FAST-Brain achieves state-of-the-art performance in recovering functional connectivity, effective connectivity, and the implicit low-dimensional signal subspace.

Talk abstracts

Wednesday, November 4, 2026

Kun Zhang | Carnegie Mellon University

3:15–3:50 pm | SSW 903

Causal representation learning and causal generative AI

As a core pillar of science and engineering, causality is transforming our approach to modern machine learning and artificial intelligence. Uncovering the causal process underlying observed data naturally helps answer 'why' and 'what-if' questions, informs optimal decision-making, and enables adaptive prediction. In many scenarios, observed variables, such as image pixels and questionnaire responses, are often reflections of the underlying hidden causal variables rather than being causal variables themselves. Causal representation learning aims to reveal the underlying hidden causal variables and their relations. In this talk, we show how the modularity property of causal systems makes it possible to recover the underlying causal representations from observational data with identifiability guarantees. We further demonstrate how identifiable causal representation learning can directly benefit generative AI, using image generation / editing and extrapolative data generation as illustrative examples.

Yihong Gu | Harvard University

3:50–4:25 pm | SSW 903

Unveiling Invariant and Transferable Latent Factors Across Heterogeneous Environments via ATLAS

We consider a multi-environment factor model in which high-dimensional covariates are collected from heterogeneous environments, with auxiliary labels available in a subset of these environments. The joint distribution of the covariates may vary across environments, whereas the latent structure is decomposed into invariant factors with shared loadings and heterogeneous factors with environment-specific loadings. Such a model is motivated by transfer learning and latent factor regression, where one seeks stable low-dimensional representations for both interpretation and robust out-of-sample prediction of the response Y. Leveraging the invariance principle, we show that the invariant and heterogeneous factors are disentangled under a minimal structural condition. Based on this, we propose ATLAS, an Auxiliary-label and invariance-guided Transfer via Latent Alignment across heterogeneous environmentS. ATLAS is a unified procedure that leverages the invariance principle to separate aligned invariant and unaligned heterogeneous factors, and further exploits supervision from auxiliary labels to extract prediction-invariant and transferable factors from those unaligned heterogeneous factors. ATLAS yields near-oracle performance for downstream latent factor regression, enables transferable prediction in new environments through the full latent signal when auxiliary labels are available, and reduces to robust invariant-factor-only prediction otherwise. We establish sharp non-asymptotic error bounds for recovering invariant and heterogeneous factors, identifying all the response-invariant factors, and estimating the invariant signal in Y.

Talk abstracts

Thursday, November 5, 2026

Annie Qu | UC Santa Barbara

9:00–9:35 am | SSW C06

Meta Fusion: A Unified Framework For Multimodal Fusion with Adaptive Mutual Learning

Developing effective multimodal data fusion strategies has become increasingly essential across a wide range of applications, from autonomous driving to medical diagnosis. Traditional fusion methods, including early, intermediate, and late fusion, integrate data at different stages, each offering distinct advantages and limitations. In this paper, we introduce Meta Fusion, a flexible and principled framework that unifies these strategies as special cases. Motivated by deep mutual learning and ensemble learning, Meta Fusion constructs a cohort of models based on various latent representation combinations, and further boosts predictive performance through soft information sharing within the cohort. Our approach is model-agnostic in learning the latent representations, allowing it to flexibly adapt to the unique characteristics of each modality. Theoretically, our soft information sharing mechanism reduces the generalization error. Furthermore, we develop an uncertainty quantification framework that provides rigorous coverage guarantees with minimal assumptions. Empirically, Meta Fusion consistently outperforms conventional strategies across extensive numerical studies, including Alzheimer's disease detection and neural decoding.

Tracy Ke | Harvard University

9:35–10:10 am | SSW C06

Topic-aware Ranking with Text-estimated Context

Pairwise comparisons often depend on latent contexts that can be inferred from auxiliary text. We develop a unified framework for topic-specific ranking using Topic-Aware Bradley(-Terry-Luce) (TAB) models, in which each entity has a score for each latent topic and comparison probabilities depend on document-specific topic weights. The framework accommodates two comparison designs motivated by journal ranking from citations and large language model evaluation from user preferences: entity-specific document contexts and a shared prompt context. The central statistical challenge is that estimated topic weights enter the comparison likelihood nonlinearly, inducing plug-in bias that can persist as the number of comparisons grows. We address this challenge by coordinating topic-weight estimation with an explicit third-order correction for the likelihood gradient. First, a tailored representation of the text yields weight estimators that are approximately linear in multinomial counts. Second, multinomial moment identities enable a two-step correction that accounts for both the original plug-in bias and the additional bias introduced by estimating the correction itself. We establish uniform error bounds for the corrected gradient, local geometric convergence of projected gradient ascent, and statistical error bounds for the resulting score estimators. The convergence rates reflect the distinct roles of comparison noise, text-estimation error, residual bias, and repeated document use. Applications to MADStat journal citations and Chatbot Arena preferences reveal variation in journal influence across research areas and in LLM performance across skills.

Talk abstracts

Thursday, November 5, 2026

Jiashun Jin | Carnegie Mellon University

10:40–11:15 am | SSW C06

Estimating weak spikes with Newton and Prony

How to estimateweak spikes is a fundamental problem in high-dimensional data analysis, where the empirical eigenvalue approach is known to be significantly biased. To correct the bias, we may use random matrix theory to derive the precise asymptotical behavior of the empirical eigenvalues, but such an approach is quite hard and so far researchers are only able to solve the idealized GOE case.

We developed a new approach where we first use the cycle count statistics to estimate the order-m power sums of the spikes for $m = 1, 2, \ldots$ and then use these power sums to estimate the spikes with Newton-Prony identities. In a setting much broader than the GOE case, we show that our approaches are consistent with algebraic rates. To implement our approaches, we must address a long-lasting open problem: how to efficiently compute the cycle count statistics. We solve it with a humAI approach where we combine complicated graph theory with the intelligence of AI.

George Michailidis | UCLA

11:15–11:50 am | SSW C06

Learning Latent Representations and Structured Dynamics in High-Dimensional Time Series

High-dimensional time series exhibit complex dependence structures that can be substantially simplified through appropriate low-dimensional representations. A central statistical challenge is to learn such representations while preserving the temporal relationships needed for prediction and inference. A motivating example arises in macroeconomic panels, where multiple economic indicators are observed across countries. A small set of primary indicators, such as grade domestic product, unemployment rate and inflation rate, may influence a much larger collection of secondary indicators, while economic linkages across countries generate both shared and country-specific temporal dynamics. In this talk, we present a two-stage framework for modeling high-dimensional matrix-valued time series that combines latent factor extraction with structured dynamic estimation. In the first stage, latent factors are extracted from the secondary variables after accounting for the temporal influence of the primary indicators. In the second stage, the joint dynamics of the primary variables and estimated factors are modeled through a panel vector autoregression whose transition matrices admit a low-rank-plus-sparse decomposition, capturing shared temporal patterns and entity-specific heterogeneity. We discuss identifiability, estimation under high-dimensional scaling, and the propagation of factor estimation error into downstream dynamic estimation. We establish statistical guarantees for the proposed estimators under high-dimensional scaling, explicitly accounting for the propagation of factor estimation error into the estimation of the dynamic model. Numerical experiments demonstrate the benefits of structured dimension reduction, while an application to macroeconomic data across selected Eurozone countries illustrates the framework's ability to capture common economic dynamics alongside country-specific temporal dependencies.

Talk abstracts

Thursday, November 5, 2026

Hui Zou | Johns Hopkins University

1:40–2:15 pm | SSW 903

Foe or Ally: The Role of Dimensionality in Prediction with Measurement Errors

The curse of dimensionality has long motivated sparsity-driven learning algorithms. This talk examines how dimensionality interacts with measurement error in regression and reveals a striking dichotomy rooted in where sparsity lives. For models sparse in the original predictor space, higher dimension only exacerbates the damage from measurement error. For models sparse in the principal component space, ℓ_1-penalized principal component regression automatically achieves the minimax-optimal prediction rate of clean data, requiring neither calibration nor any knowledge of the error covariance matrix. More remarkably, the impact of measurement error on prediction diminishes as the number of covariates grows, so that high dimensionality becomes a blessing.

Talk abstracts

Thursday, November 5, 2026

Joseph Feldman | Washington University in St. Louis

2:15–2:50 pm | SSW 903

Scalable Independent Factor Analyzers using Spectral Initialization for Evaluating Large Language Models

Large language model (LLM) benchmarks are typically summarized using aggregate accuracies or leaderboard rankings, obscuring multidimensional latent patterns of strengths, weaknesses, and heterogeneity across models. We introduce *probit independent factor analysis* (P-IFA), a latent variable model for binary benchmark responses that jointly estimates continuous LLM abilities and factor-specific latent subtypes. P-IFA extends classical probit factor models by assigning each latent factor an independent finite Gaussian-mixture distribution. This non-Gaussian structure both captures heterogeneous populations of LLMs and identifies the latent factor orientation up to sign and permutation under suitable conditions. To enable estimation in the high-dimensional regime where the number of benchmark prompts greatly exceeds the number of evaluated models, we develop a scalable initialization strategy motivated by the (latent) spectral geometry underlying P-IFA. The procedure first estimates a low-rank factor subspace using spectral decomposition, and then recovers the P-IFA orientation by optimizing a profiled Gaussian-mixture criterion. These estimates then initialize a parallelizable *maximum a posteriori* refinement of the full model. We study factor recovery, subtype estimation, and computational performance through simulations, and apply P-IFA to the IFEval instruction-following benchmark to characterize latent abilities and heterogeneous LLM performance profiles.

Gongjun Xu | University of Michigan

2:50–3:25 pm | SSW 903

Beyond Vintage Rotation: Bias-Free Sparse Representation Learning with Oracle Inference

Learning low-dimensional latent representations is a central topic in statistics and machine learning, and rotation methods have long been used to obtain sparse and interpretable representations. Despite nearly a century of widespread use across many fields, rigorous guarantees for valid inference for the learned representation remain lacking. In this paper, we identify a surprisingly prevalent phenomenon that suggests a reason for this gap: for a broad class of vintage rotations, the resulting estimators exhibit a non-estimable bias. Because this bias is independent of the data, it fundamentally precludes the development of valid inferential procedures, including the construction of confidence intervals and hypothesis testing. To address this challenge, we propose a novel bias-free rotation method within a general representation learning framework based on latent variables. We establish an oracle inference property for the learned sparse representations: the estimators achieve the same asymptotic variance as in the ideal setting where the latent variables are observed. To bridge the gap between theory and computation, we develop an efficient computational framework and prove that its output estimators retain the same oracle property. Our results provide a rigorous inference procedure for the rotated estimators, yielding statistically valid and interpretable representation learning.