I aim to develop novel **machine learning** frameworks to learn from and make predictions on **large-scale spatiotemporal data**, that optimize both statistical accuracy and computational efficiency.

Spatiotemporal data is ubiquitous in our daily life ranging from climate science, via transportation, to social media. Today, data is being collected at unprecedented scale. Yesterday's concepts and tools are insufficient to serve tomorrow's data-driven decision makers. A key challenge is that spatiotemporal data often demonstrate complex dependence structures, come in multiple resolutions, and are of high dimensionality. This requires algorithms that can handle non-iid data, reason at different granularity, and generate structured predictions. In order to fully harness the power of large-scale spatiotemporal data, **my goal is to 1) develop new statistical models and computational tools for discovering spatial and temporal patterns, 2) understanding system dynamics and 3) making long-range predictions.** My primary research lies in the fundamental areas of structure learning, spatial statistics and optimization, and with a strong emphasis on applications in computational sustainability and social science. My previous work has yielded new methodologies for climate modeling [1], traffic forecasting [2], mobile intelligence [3], and social network analysis [4].

Past Work

My dissertation research is concerned with developing new machine learning tools, specifically tensor methods, for large-scale spatiotemporal analysis. Statistical analysis of data that take spatial, temporal, and spatio-temporal information into account has been intensively studied. Traditional methods often use Gaussian Processes. Gaussian processes (GPs) describe the dependency structure by constructing covariance functions, i.e, kernels, from observations. However, GP methods scale cubically with the data size in the worst case. Furthermore, designing GP kernels requires domain-specific knowledge, which is time-consuming and often introduces significant economic cost. To address the scalability issue and reduce human effort, my thesis work developed a low-rank tensor learning framework. It trades uncertainty for scalability, and learns the correlation structure in a nonparametric fashion. I completed the framework from three stages of learning: offline learning [NIPS 2014], online learning [ICML 2015] and memory-efficient learning [ICML 2016].

*Fast greedy **offline learning** for spatiotemporal analysis*

Spatiotemporal analysis has two central tasks: 1) cokriging (missing-value imputation) interpolates the data of one variable for unknown locations by taking advantage of the observations of variables from known locations; 2) forecasting estimates the future value of multivariate time series given historical observations. Prior works have identified two key principles for efficient modeling of spatiotemporal dependencies: 1) global consistency: the data on the same structure are likely to be similar, and 2) local consistency: the data in close locations are likely to be similar. To encode such dependency structures of the data in a concise way, I first represented the data as a three-dimensional tensor over (location, time, features). To achieve global consistency, we constrained the model tensor to be low-rank. The low-rank assumption is based on the belief that there exists low-dimensional representations for features, locations and time. For local consistency, we constructed a regularizer via a spatial Laplacian matrix, which approximates the covariance matrix commonly used in the GP literature, but that avoids the expensive matrix inversion operation. We incorporate these principles and show that both co-kriging and forecasting tasks reduce to a unified low-rank tensor learning framework. To optimize for such non-convex objectives, we proposed a greedy batch learning algorithm with theoretical convergence guarantees. This work was accepted as a **spotlight** presentation in NIPS 2014 (acceptance rate 5%).

*Accelerated **online learning** for tensor streams*

Spatiotemporal analysis frequently uses large-scale data streams. Here, batch learning algorithms would suffer from computational bottlenecks, especially when short response times are required. Therefore, effective and fast online learning algorithms are crucial to enable real-time analysis. Online learning of low-rank tensors aims to dynamically update a tensor model while preserving the low-rank structure. A core challenge is that performing low-rank tensor decomposition at every iteration is expensive for high dimensional tensors. To address this issue, I designed ALTO, an accelerated online low-rank tensor learning algorithm that keeps track of the low-rank components. For each batch of data, ALTO first projects the tensor model into a low-dimensional space, in order to perform efficient tensor decomposition. It then updates those components to obtain the low-rank tensor approximation. A similar heuristic called Streaming Tensor Analysis (STA) has achieved wide practical success. However, STA suffers from local optima as the projection step restricts the tensor to a fixed subspace. ALTO resolves this issue via randomization. I introduced random noise to perturb the model in order to jump out of the local optima. Theoretical analysis shows that our technique can significantly reduce the variance at a cost of very minor biases. This work gained high recognition by Prof. Christos Faloutsos from CMU who was the original author of the STA paper.

*Simple and **memory-efficient** algorithm for tensor regression*

Spatiotemporal data is one example of multi-way data. Tensors provide a natural representation for such data to extract high order correlations. For exploratory multi-way data analysis, tensor decomposition has been a popular technique. In contrast, tensor regression, a supervised method for learning with multi-directional relatedness, has not been fully examined. Part of the reason is that the general tensor regression problem is NP-hard. Statistical methods for tensor regression often use MCMC sampling, thus are hardly scalable. Another key challenge of tensor regression is that most state-of-art algorithms require unfolding tensors into matrices. This unfolding operation quickly leads to a memory bottleneck when dealing with large high-dimensional tensors. However, our investigation on tensor regression showed that efficient solutions are possible under feasible assumptions on the tensor model. To address the memory bottleneck, I designed the subsampled Tensor Projected Gradient (TPG), whose memory requirement is only *linear* in data size. With randomized sketching and fast tensor power iterations, the algorithm converges in fixed number of iterations. We provided a theoretical analysis of our algorithm, which is guaranteed to find the correct solution assuming the objective is ``nearly-convex''. In fact, the algorithm only needs a fixed number of iterations, depending solely on the logarithm of signal to noise ratio (which include dimensionality and sample size).

Research Agenda

I am excited to develop theoretically grounded algorithms for large-scale spatiotemporal analysis, translate technologies into an accessible software platform, and contribute to the emerging field of computational sustainability and social science. In the presence of noisy, non-iid spatiotemporal data, and on a scale where human analysis is no longer feasible, we need new methods that are not only **efficient** but also **robust**. To push towards practicality, my research agenda involves a blend of methodological and applied components.

Methodologies

Scalability: Exponential growth of sensory and satellite data requires scalable analytical tools. I am interested in addressing this issue from an algorithmic perspective. Spatiotemporal analysis requires modeling spatial and temporal correlations, which renders many state-of-the-art machine learning techniques invalid. For example, **sketching** is a known method to randomly down-sample data for speed-up. But how can we perform sketching while still preserving the underlying spatiotemporal distributions? Similarly, distributed infrastructures are easily configured for algorithms that are ``embarrassingly parallel''. But how can we **parallelize inference** for spatiotemporal models given their complex correlations? In all these settings, it is critical to find the optimal trade-off between statistical accuracy and computational efficiency. My thesis work achieves such a trade-off. The low-rank tensor learning frameworks performs dimension reduction while preserving the inherent structures of spatiotemporal data. I am also investigating other instances of this trade-off in several ongoing projects. One of them is to compress **deep neural networks** using tensor factorizations that can account for the multi-linear nature of network weights and reduce memory overhead.

Reliability: For risk-sensitive spatiotemporal applications such as self-driving cars and medical diagnosis, it is important to associate a reliable measure of confidence with predictions. Ensuring reliability of machine learning is a significantly magnified challenge in spatiotemporal data analysis. For example, the common ``missing at random'' assumption is generally invalid for spatiotemporal data. How can we learn with **non-random missing** observations? Existing methods can only handle input corruptions that are independent among features. How can we design robust methods that can model **correlated corruptions**? It is common practice to use cross-validation, resampling or regularization to improve model generalization. But how can we develop algorithms that **generalize** well in the presence of long-range spatiotemporal dependencies, hidden causal relationships and non-stationary, nonlinear dynamics? To address these challenges, we need learning algorithms that are reliable in extreme settings. Therefore, I am excited to explore new statistical techniques to handle irregular spatiotemporal signals, characterize correlated noise and to improve model generalization in a broader context. In this context, I am collaborating with Dr. Peter Kuhn, a USC professor in Oncology to develop robust Bayesian spatiotemporal models to study the progression patterns of prostate cancer.

Applications

Sustainability: Many important environmental problems, e.g. climate change, energy consumption and transportation, are intrinsically spatiotemporal in nature. Accurate modeling and efficient inference of the underlying spatiotemporal dynamics is the key to progress in these fields. For instance, El Nino is an oscillation in the ocean-atmospheric system that causes significant weather changes. Understanding the spatiotemporal impact of El Nino produces more accurate predictions of flooding and landslides and reduces socio-economic cost. As an active member of the Climate Informatics (CI) community, I have presented my climate model research at the National Center for Atmospheric Research (NCAR), which has triggered wide interest in my findings. Given the success and potential future impact, I aim to continue such collaborations to tackle societal challenges.

Social science: Complex social processes exhibit spatiotemporal patterns in both structure and content. For example, news article topics often exhibit geographic variation, while social network friendships and senator voting patterns are known to change over time. Current research has focused mainly on single snapshots of structure and content. But studying social media from both a spatial and temporal perspective can reveal nuanced information that is crucial to understanding and reasoning about complex social processes. Previously, I took a temporal approach to social media group anomaly detection [5] and presented a principled method to detect social network anomalies using Bayesian hierarchical models. This has opened up a new line of social media research that I am excited to expand into a full spatiotemporal framework.

Albert Einstein once said: "Time and space are modes by which we think and not conditions in which we live." I hope to fundamentally change the way spatiotemporal data is being analyzed and understood, especially in the domain of computational sustainability and social science. My research on novel machine learning frameworks will help to address the technical challenges and lay the foundation for the next generation of large-scale data analytics.

References

[1] Rose Yu*, Mohammad Taha Bahadori*, Yan Liu. "Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning." In Proceeding of Advances in Neural Information Processing Systems (NIPS), 2014 Spotlight

[2] Rose Yu, Dehua Cheng, Yan Liu. "Accelerated Online Low Rank Tensor Learning for Multivariate Spatiotemporal Streams." In Proceedings of the 32th International Conference on Machine Learning (ICML), 2015

[3] Rose Yu, Yan Liu. "Learning from Multiway Data: Simple and Efficient Tensor Regression." In Proceedings of the 33th International Conference on Machine Learning (ICML), 2016

[4] Rose Yu, Andrew Gelfand, Suju Rajan, Cyrus Shahabi, Yan Liu. "Geographic Segmentation via Latent Poisson Factor Model." in ACM International Conference on Web Search and Data Mining (WSDM), 2016

[5] Rose Yu, Xinran He, Yan Liu. "GLAD: Group Anomaly Detection in Social Media Analysis." In Proceeding of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD),2014