2015 NSF Workshop on Intelligent Systems for Geosciences
Participant Statement: Arindam Banerjee

1) Summarize 3-4 big ideas in your field in the last few years that you think may have an impact in geosciences research

1. High-dimensional structured estimation, learning dependencies: In several scientific problems, including problems in geosciences, the number of samples (examples) of a phenomenon of interest, e.g., Atlantic hurricanes, Indian summer monsoon rainfall, etc., is often far smaller than the number of features or covariates which could potentially be influencing the phenomenon. Considerable progress has been made in recent times on such high-dimensional problems using structured estimators such as the Lasso, the Dantzig selector, their variants and generalizations, which can automatically do feature selection while fitting a (parametric) predictive model using a small number of samples. Related techniques have also been developed for learning statistical dependencies in multivariate settings.

2. Semi-parametric and non-parametric approaches for predictive modeling: Progress is being made on effectively fitting certain types of nonlinear predictive models with relatively small number of samples. One class of semi-parametric models consider the response as a parametric linear combination of monotone nonlinear transformations of the given features, a form of nonlinearity commonly observed in scientific problems. Non-parametric models and their variants in the context of multi-task learning acknowledge that predictive models need to change based on the 'phase' of the input, where different phases may be represented on a manifold or a suitable nearest neighbor graph. In geosciences, the phase space may be determined by indices, e.g., ENSO, NAO, etc., or other factors.

3. Deep learning, robust representations: Deep learning constitutes a revival of earlier work on neural networks with a new set of ideas which focus on learning rich and robust representations for predictive problems. The robustness of the representation comes from training such models under suitable perturbations of the input, e.g., dropout, noisy auto-encoders, etc., so the representation and the corresponding predictive model is largely invariant to such perturbations. Neural networks have been successfully used in geosciences and hydrology before, and the new advances in deep learning can further these existing applications as well as lead to new applications.

4. Inference in latent variable models: Latent variable models, ranging from mixture of Gaussians (MoG) and hidden Markov models (HMMs) to random effect models and dynamic Bayes nets, provide a convenient and interpretable modeling framework for settings where some important variables are unavailable. Classically, inference in such models have been difficult because of non-convexity and related challenges, and one needed to use suitable alternating minimization schemes which can have poor convergence, say to a bad local minima, especially in high-dimensions. Recent work has shown that provable inference can be done in some such models using suitable spectral methods, which can lead to reliable usage and qualitatively better results from such models.

2) Highlight 2-3 important research trends in your area that can be relevant to the workshop goals

1. Probabilistic graphical models: Probabilistic graphical models provide a rich representation for multivariate problems, where statistical dependencies among the variables are represented as a graph, either directed, e.g., Bayes nets, or undirected, e.g., Markov random fields, and possibly having latent (unobserved) variables. Such models have been successful when dealing with multiple variables with dependencies which are direct, through other variables, or over space/time, which are all aspects commonly encountered in geoscience data.

2. Online learning: Online learning is well studied framework for sequential decision making, and includes multiplicative update algorithms, online Bayesian approaches, and online convex optimization as specific realizations. Online learning is particularly well suited for adaptive modeling, where a given model needs to be suitably updated as new data becomes available. In the context of geosciences, online learning may be appropriate for near term predictions, seasonal predictions, or multi-model ensembles.

3. Spatiotemporal data analysis: Much of data encountered in geosciences is spatiotemporal, and there is a growing body of literature on models and methods for spatiotemporal data analysis. The literature includes nonparametric approaches based on Gaussian processes and variants, multivariate spatial and temporal parametric models including vector auto-regressive models and variants, probabilistic graphical models for spatiotemporal data, and spatiotemporal pattern mining approaches, among others.