# Research Statement

Michael Schweinberger

My research is concerned with practical, computational, and theoretical aspects of learning from

- dependent and high-dimensional data without independent replications: e.g., network, spatial, and temporal data;
- structured data, that is, data with additional structure, either observed or unobserved: e.g., block, multilevel, spatial, and temporal structure;
- social science data: e.g., educational data, epidemiological data, and social network data.

My research was made possible by NWO award Rubicon-44606029 (sole PI), NSF award DMS-1513644 (sole PI), NSF award DMS-1812119 (sole PI), and ARO award W911NF-21-1-0335 (PI).

## Overview

My research is motivated by dependent and high-dimensional data without independent replications, such as network, spatial, and temporal data. My ideas of how to learn from dependent data without independent replications are elaborated in one of the simplest possible settings: statistical exponential families (Wainwright and Jordan, Foundations and Trends in Machine Learning, 2008). Exponential families are widely used throughout data science, as stand-alone models or building blocks of more complex models. The fundamental role of exponential families in data science is exemplified by the prominent role of multivariate Gaussians, but there are numerous other examples, including generalized linear models, undirected graphical models, Markov random fields in machine learning, and Boltzmann machines in artificial intelligence. In fact, my research (Schweinberger, JASA 2011) has contributed to the understanding of generative deep learning models in artificial intelligence (Kaplan et al., *On the S-instability and degeneracy of discrete deep learning models,* Information and Inference, 2020).

## Selected highlight

Consider network data, which are dependent data without independent replications. Since the 1950s, social scientists have pointed out that connections depend on other connections: e.g., the frequent observation that "a friend of a friend is a friend" suggests that friendships are dependent. In applications, population probability models are learned from a single observation of a population network or subnetworks sampled from a population network. That raises an important question:

***What can we learn about an interconnected and interdependent world where connections depend on other connections, without having the benefit of independent observations from the same source?***

In a decade-long sequence of first- and single-authored papers (e.g., Annals of Statistics; Bernoulli; Journal of the Royal Statistical Society, Series B; Journal of the American Statistical Association; Statistical Science) and additional papers in progress (e.g., Stewart and Schweinberger, 2021), I have

1. studied the properties of ill-posed models of dependent random variables, with applications to models of dependent network data and generative deep learning models in artificial intelligence. My work (Schweinberger, JASA, 2011) preceeded Chatterjee and Diaconis (AOS, 2013).
2. shown how well-posed models of dependent network data can be constructed based on statistical exponential families, with desirable properties.
3. demonstrated that statistical learning of an unbounded number of parameters based on a single observation of dependent random variables is possible, with theoretical guarantees.

4. developed scalable methods for statistical learning of an unbounded number of parameters based on a single observation of dependent random variables, with theoretical guarantees.

There is a common thread that connects these advances: **the importance of additional structure.** Models that lack mathematical structure to control the dependence among random variables can be ill-posed, but endowing models with additional structure can help control dependence and result in well-posed models with desirable properties. In addition, weak dependence facilitates concentration-of-measure results, which in turn facilitate consistency results. In other words, endowing models with additional structure has at least two advantages:

1. It facilitates the construction of well-posed models with desirable properties.
2. It facilitates statistical learning with theoretical guarantees.

Examples of additional structure are block, multilevel, spatial, and temporal structure. Moreover, additional structure helps answer fundamental questions about the statistical analysis of dependent network data raised by leading probabilists (e.g., Chatterjee and Diaconis, AOS, 2013) and statisticians (e.g., Fienberg, JCGS, 2012), as discussed in Schweinberger et al. (Statistical Science, 2020).

## Selected directions of future research

**Online educational assessment data:** In collaboration with Minjeong Jeon (Graduate School of Education & Information Studies, University of California, Los Angeles), I am working on educational assessment data. Among other things, we are developing statistical interaction and learning progression maps, with a view to providing educators with visual tools for monitoring student progress and detecting (underrepresented) groups of students who need more, and different support than other students.

**Stochastic processes involving networks, space, and time:** Many real-world processes involve networks, space, and time: e.g., infectious diseases spread through contacts, contacts depend on geographical space, and contacts change over time. While there are existing stochastic processes indexed by networks, space, time or combinations of them, many of them make either simplifying assumptions or have unknown probabilistic and statistical properties. One of my directions of future research is to design stochastic processes indexed by networks, space, and time that do justice to the complexity of network-mediated phenomena, and develop scalable statistical and computational methods and theoretical guarantees for learning them from data.

**Scalable selection of models of dependent data without independent replications and intractable likelihood functions:** Developing scalable model selection procedures with theoretical guarantees is non-trivial when the likelihood function is intractable, the number of parameters is large, and the data consists of a single observation of dependent random variables. Such scenarios arise in the statistical analysis of discrete and dependent data, such as discrete network, spatial, and temporal data. For example, there are many models of dependent network data, but no scalable model selection procedures with theoretical guarantees are known. I am working on a scalable approach to model selection in dependent data problems with intractable likelihood functions based on pseudo- and composite-likelihood-based regularization procedures.

**Quantifying uncertainty of statistical learning based on dependent data without independent replications:** In applications of data science, it is important to provide a disclaimer, stating how uncertain we are about statistical conclusions based on data. In scenarios when the number of parameters is unbounded and a single observation of dependent random variables is available, it is not obvious how to quantify uncertainty, because the distributions of many statistical quantities are unknown. A natural approach to capturing uncertainty is a Bayesian approach. I intend to elaborate on scalable Bayesian approaches to uncertainty quantification for discrete and dependent data without independent replications based on pseudo- and composite-likelihood functions.