Department of Statistics

STATISTICS COLLOQUIUM

GREGORY VALIANT

Department of Computer Science
Stanford University

When Your Big Data Seems Too Small

MONDAY, March 11, 2019 at 4:30 PM
Eckhart 133, 5734 S. University Avenue
Refreshments before the seminar at 4:00PM in Jones 111

ABSTRACT

We discuss several problems related to the challenge of making accurate inferences about a complex phenomenon, given relatively little data. We show that for several fundamental and practically relevant settings, including estimating the intrinsic dimensionality of a high-dimensional distribution, and learning a population of distributions given few data points from each distribution, it is possible to ``denoise" the empirical distribution significantly. We will also discuss the problem of estimating the ``learnability" of a data source: given too little data to train an accurate model, we show that it is often possible to estimate the extent to which a good model exists. Framed differently, even in the regime in which there is insufficient data to learn, it is possible to estimate the performance that could be achieved if you obtain a much larger amount of data (from the same source) and then train a model on that larger dataset. Our results, while theoretical, have a number of practical applications, and we also discuss some biological applications.

This talk is based on joint work with Weihao Kong.
