Statistics for Big Data

Today, data analysis methods in machine learning and statistics play a central role in industry and science. The growth of the Web and improvements in data collection technology in science have lead to a rapid increase in the magnitude and complexity of these analysis tasks. This growth is driving the need for scalable, parallel and online algorithms and models that can handle this "Big Data". This course will provide a broad foundation for this timely challenge.

In particular, we will focus on the challenges associated with datasets of massive size and dimensionality, including settings where the dimensionality of the data is growing faster than the number of data points. Framed by canonical examples of big data applications in science and industry, we will present a core set of techniques, both in terms of algorithms and models, to tackle these challenges. We will also explore the computational foundations associated with performing these analyses in the context of parallel and cloud architectures.

Large-scale modeling techniques covered will include linear models, graphical models, matrix and tensor factorizations, clustering, and latent factor models. Algorithmic topics include sketching, fast n-body problems, random projections and hashing, large-scale online learning, and parallel learning. The computational techniques covered in this course will provide a basic foundation in large-scale programming, ranging from the basic "parfor" to parallel abstractions, such as MapReduce (Hadoop) and GraphLab.

To be successful in this course, students should have prior exposure to basic statistical and machine learning concepts, such as those covered in one of these courses: CSE446, CSE546, CSE515, STAT535, EE511, or EE512. As needed, we will also provide background reading on certain topics throughout the quarter.