**Presenter:** Brian Bullins

**Title:** Beyond First-Order Methods for Large-Scale Optimization

**Abstract:** In recent years, stochastic gradient descent (SGD) has taken center stage for training large-scale models in machine learning. Although methods which go beyond first-order information may achieve better iteration complexity in theory, the per-iteration costs often render them unusable when faced with the current growth in both the available data and the size of the models, particularly when such models now have hundreds of billions of parameters.

In this talk, I will present results, both theoretical and practical, for several important settings which enable second-order optimization to be as scalable as first-order methods. To begin, I will describe a stochastic second-order algorithm for convex optimization which uses Hessian information to construct an unbiased Newton step in time linear in the problem dimension. In this case, bypassing the typical efficiency barriers for second-order methods relies on harnessing the ERM structure in standard machine learning tasks. Given the non-convexity of deep neural networks, it has further become important to develop a better understanding of non-convex guarantees. Thus, I will present a Hessian-based method which provably converges to first-order critical points faster than gradient descent, alongside guarantees for converging to second-order critical points. Optimization methods which may parallelize have also become increasingly critical when facing enormous deep learning models, and so I will show how we may leverage stochastic second-order information to attain faster methods in the distributed optimization setting.