Speaker: Prof. Ce Zhang, ETH Zurich

Date & Time: Nov. 11, 3:30 PM - 4:45 PM CT

Title: Towards Understanding End-to-end Learning in the Context of Data: Machine Learning Dancing over Semirings and Codd's Table

Abstract:
Recent advances in machine learning (ML) systems have made it incredibly easier to train ML models given a training set. However, our understanding of the behavior of the model training process has not been improving at the same pace. Consequently, a number of key questions remain: How can we systematically assign importance or value to training data with respect to the utility of the trained models, may it be accuracy, fairness, or robustness? How does noise in the training data, either injected by noisy data acquisition processes or adversarial parties, have an impact on the trained models? How can we find the right data that can be cleaned and labeled to improve the utility of the trained models? Just when we start to understand these important questions for ML models in isolation recently, we now have to face the reality that most real-world ML applications are way more complex than a single ML model.
In this talk, I will revisit these questions for an end-to-end ML pipeline, which consists of a noise model for data and a feature extraction pipeline, followed by the training of an ML model. In the first part of this talk, I will introduce some recent theoretical results in an abstract way: How to calculate the Shapley value of a training example for ML models trained over feature extractors, modeled as a polynomial in the provenance semiring? How to compute the entropy and expectation of ML models trained over data uncertainty, modeled as a Codd Table? As we will see, even these problems are #P-hard for general ML models, though, surprisingly, we can obtain PTIME algorithms for a simpler proxy model (namely a K-nearest neighbor classifier), for a large family of polynomials, input noise distributions, and utilities.
I will then put these theoretical results into practice. Given a set of heuristics and a proxy model to approximate a realworld end-to-end ML pipeline into these abstract problems, I will provide a principled framework for three applications: (1) certifiable defence of backdoor attacks, (2) targeted data cleaning for ML, and (3) data valuation and debugging for end-to-end ML pipelines. I will describe both our positive empirical results but also those cases that our current approach failed at.

Biography:
Prof. Ce Zhang is an Assistant Professor in Computer Science at ETH Zurich. He believes that by making data—along with the processing of data—easily accessible to non-expert users, we have the potential to make the world a better place. His current research focuses on understanding and building next-generation machine learning systems and platforms. Before

joining ETH, Ce was advised by Christopher Ré, finished his PhD roundtripping between the University of Wisconsin-Madison and Stanford University, and spent another year as a postdoctoral researcher at Stanford. He contributed to the research efforts that won the SIGMOD Best Paper Award and SIGMOD Research Highlight Award, and was featured in special issues, including in the Science magazine, the Communications of the ACM, “Best of VLDB”, and the Nature magazine. His work has also been reported by media such as the Atlantic, WIRED, and Quanta Magazine. More information about his group can be found at: ds3lab.org.