Electrical & Computer ENGINEERING

Carnegie Mellon University

Graduate Seminar

Snorkel: Beyond hand-labeled data

Christopher Ré

Associate Professor
Department of Computer Science
Stanford University in the InfoLab

Thursday, September 28th

4:30 PM Doherty Hall A302
1:30 PM Room B23 118

Abstract:
This talk describes Snorkel, a software system whose goal is to make routine machine learning tasks dramatically easier. Snorkel focuses on a key bottleneck in the development of machine learning systems: the lack of large training datasets for a user's task. In Snorkel, a user implicitly defines large training sets by writing simple programs that create labeled data, instead of tediously hand-labeling individual data items. In turn, this allows users to incorporate many sources of training data, some of low quality, to build high-quality models. This talk will describe how Snorkel changes the way users program machine learning models. A key technical challenge in Snorkel is combining heuristic training data that may have uneven and unknown quality and an unknown correlation structure. This talk will explain the underlying theory, including methods to learn both the parameters and structure of generative models without labeled data. Additionally we'll describe our recent experiences with hackathons, which suggest the Snorkel approach may allow a broader set of users to train machine learning models and do so more easily than previous approaches.

Snorkel is being used by scientists in areas including genomics and drug repurposing, by a number of companies involved in various forms of search, and by law enforcement in the fight against human trafficking. Snorkel is open source on github. Technical blog posts and tutorials are available at Snorkel.Stanford.edu.

Bio:
Christopher (Chris) Ré is an associate professor in the Department of Computer Science at Stanford University in the InfoLab who is affiliated with the Statistical Machine Learning Group, Pervasive Parallelism Lab, and Stanford AI Lab.

His work's goal is to enable users and developers to build applications that more deeply understand and exploit data. His contributions span database theory, database systems, and machine learning, and his work has won best paper at a premier venue in each area, respectively, at PODS 2012, SIGMOD 2014, and ICML 2016. In addition, work from his group has been incorporated into major scientific and humanitarian efforts, including the IceCube neutrino detector, PaleoDeepDive and MEMEX in the fight against human trafficking, and into commercial products from major web and enterprise companies. He cofounded a company, based on his research, that was acquired by Apple in 2017.

He received a SIGMOD Dissertation Award in 2010, an NSF CAREER Award in 2011, an Alfred P. Sloan Fellowship in 2013, a Moore Data Driven Investigator Award in 2014, the VLDB early Career Award in 2015, the MacArthur Foundation Fellowship in 2015, and an Okawa Research Grant in 2016.

SEMINAR NOTES: (REFRESHMENTS SERVED AT 4:30 PM

ECE Seminar Committee

Aswin Sankaranarayanan
saswin@ece.cmu.edu

Maysam Chamanzar
mchamanz@andrew.cmu.edu

Swarun Kumar
swarun@cmu.edu