Cloud-Scale Data Analytics

SEMINAR ANNOUNCEMENT

Speaker: Kalyan Veeramachaneni
Date: Friday, April 22
Time: 10:45 a.m.
Location: Rice 242
Host: Laura Barnes

Title: Teaching a Computer to Be a Data Scientist

Abstract: In recent years, scientists have made great strides in scaling and automating big data collection, storage, and processing. However, deriving real, useful insights from relational and semantic data still requires time-consuming guesswork and human intuition. Designing novel approaches across diverse domains, including education, medicine, and energy, has helped me identify the foundational issues that are holding such analysis back. To address these roadblocks, we developed the “Data Science Machine,” an end-to-end automated system that generates predictive models from raw data.

While scaling our own projects, two fundamental questions arose: Can we develop general-purpose tools and utilities that are useful across a broad range of applications? And how can we make the data science process more structured and systematic while still considering the maximum amount of complexity? In this talk, I will introduce three technologies we developed to answer these questions: deep feature synthesis, prediction problem synthesis, and deep mining. These novel methods have allowed us to tackle a number of previously unsolvable problems, and to streamline the most inefficient steps in the data science pipeline: feature engineering, prediction engineering, and model tuning.

Combined, these new technologies make up what we call "The Data Science Machine": an automated system that turns raw data into predictive models with minimal human input. The Data Science Machine is as talented as its human counterparts—over the course of three competitions held at premiere machine learning conferences, it out competed 615 out of 906 human teams.

Our goal is to use the DSM and other technologies to change and broaden human-data interactions. With their help, experts can tighten the reins on their information - becoming feature selectors instead of feature creators, and using complexity to their advantage instead of getting lost in noise. Ultimately, we hope to supply different interfaces to economists, analysts, educators, and an array of other professionals in order to bring them closer to their goals - and to make it easier, more effective, and more enjoyable for everyone, across domains, to work closely with data.

About the speaker: Kalyan is currently a Research Scientist at the Computer Science and Artificial Intelligence laboratory at MIT. He co-leads a group called Any Scale Learning for All at CSAIL, MIT. He received his Ph.D in Electrical and Computer Engineering from Syracuse University in 2009. His research currently focuses on developing automation technologies for data science - the field that focuses on methods and technologies that enable deriving insights from data. To enable a computer algorithm to perform the same tasks as a data scientist, without supervision, Kalyan and his team observed the tasks humans do as data scientists for years and ultimately built algorithms that emulate human intelligence involved in a data science endeavor. He has founded “The human data interaction” project which aims to enable humans and machines to interact seamlessly and thus efficiently derive insights from the data. Additionally, during the past three years, Kalyan has built predictive models from data for a number of important societal problems. These models have enabled educators to enhance online learning experience for students, and helped doctors personalize treatment and predict outcomes for patients. More at http://www.kalyanv.org/