Making data analysis really about analysis

Dr. Jiannan Wang

March 10, 2015 @ 11:45 a.m., Room: E404

Abstract: With the increasing amount of available data, turning raw data into actionable information is a requirement in every field. One bottleneck that impedes the process is data cleaning. Data scientists can spend over half of their time cleaning data that is dirty — inconsistent, inaccurate, incomplete, and so on — before they even begin to do any real analysis. How can we make data analysis really about analysis?

In this talk, I will present CrowdER and SampleClean, two systems that I built to reduce cleaning cost while providing good answer quality. CrowdER is a hybrid human-machine data cleaning system. I will describe how CrowdER combines humans with machines and achieves both good efficiency and high accuracy compared to machine-only or human-only alternatives. As data volumes continue to grow, even with hybrid human-machine approaches, data cleaning still becomes increasingly time consuming. To further reduce cleaning cost, I built SampleClean, a fast and accurate query processing system for dirty data. SampleClean aims to obtain accurate query results from dirty data, by only cleaning a small sample of data. I will describe how SampleClean achieves this goal and provides a flexible trade-off between cleaning cost and answer quality.

Jiannan Wang is a postdoc in the AMPLab at UC Berkeley, where he works with Prof. Michael Franklin and leads the SampleClean project. His research is focused on developing algorithms and systems for extracting value from "dirty" data. He obtained his PhD from the Computer Science Department at Tsinghua University. During his PhD, he has been a visiting scholar at Chinese University of Hong Kong and UC Berkeley, and an intern at Qatar Computing Research Institute. His PhD research work was supported from a Google PhD Fellowship, a Boeing Scholarship, and a "New PhD Researcher Award" by Chinese Ministry of Education. His PhD dissertation won the China Computer Federation (CCF) Distinguished Dissertation Award. His similarity-join algorithm won first place of EDBT String Similarity Search/Join Competition.