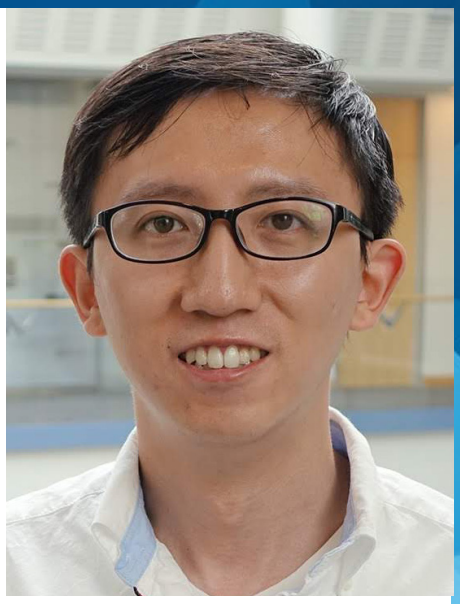
COLLEGE OF ENGINEERING
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Dissertation Defense



Zhongjun Jin

Democratizing Self-Service Data Preparation Through Program Synthesis for Non-expert Users

Monday, March 9, 2020
11:30 am – 1:30 pm 3316 EECS Bldg.

ABSTRACT: The majority of real-world data we can access today have one thing in common: they are not immediately usable in their original state. Trapped in a swamp of hairy data usability issues like non-standard data formats and heterogeneous data sources, most data analysts and machine learning practitioners have to burden themselves with “data janitor” work, which frequently entails writing ad-hoc Python or SQL scripts which is tedious and inefficient. In this dissertation, we will demonstrate with three systems that, by harnessing knowledge such as examples and other useful hints from the end user, program synthesis techniques guided by heuristics and machine learning can effectively make data preparation less painful and more efficient to perform by data users.

Data transformation, or data wrangling, is an important data preparation task that converts data from one format to a different format (usually more structured). Our system Foofah can discover meaningful data transformation programs using input-output transformation examples offered by end users, which significantly reduces the overall user effort. Our second system, CLX, demonstrates that, to fix the non-standard data format issue in a set of data instances with varied formats, users may not even need to provide complete examples, but only label desirable ones in the original dataset. The system is still capable of suggesting reasonable and explainable transformation operations. Prism, our third system, targets the task of data integration, i.e., combining multiple relations to formulate a desired schema. Prism allows the user to describe the target schema using not only high-resolution (precise) constraints of complete example data records in the target schema, but also (imprecise) constraints of varied resolutions, such as example data records with missing values, value ranges, or multiple possible values in cells, so as to require less familiarity of the database contents from the end user.

Co-Chairs: Profs. Michael Cafarella and HV Jagadish