SFU Data Science Research Group presents

Seminar Series on Trustworthy Data Science and AI

SFU SIMON FRASER UNIVERSITY

Ihab Ilyas

Professor, University of Waterloo

HoloClean and Kamino: Structured Learning for Data Cleaning and Private Data Generation

Wednesday, June 30, 2021
04:00pm Pacific time

Scan or visit URL to register:
https://data.cs.sfu.ca/pi0f

Abstract

Data scientists spend a big chunk of their time preparing, cleaning, and transforming raw data before getting the chance to feed this data to their well-crafted models. This massive labor-intensive exercises to clean data remain the main impediment to automatic end-to-end AI pipeline for data science.

Real world scenarios are further complicated by data privacy and the need to protect sensitive user information. Data privatization efforts often compromise the utility of the data; for example, existing differentially private data synthesis methods usually fail in preserving one of the most fundamental data properties: the underlying correlations and dependencies among tuples and attributes (i.e., the structure of the data), compromising the utility of the output to downstream tasks.

In this talk I describe our work in Kamino and HoloClean: systems for data cleaning and private data synthesis.

HoloClean builds two main probabilistic models: a data generation model (describing how data was intended to look like); and a realization model (describing how errors might be introduced to the intended clean data). The framework uses few-shot learning, data augmentation, and self supervision to learn the parameters of these models, and use them to predict both error and their possible repairs.

Kamino is a new system that learns the HoloClean models privately and use them to synthesize "useful" private data instances. Kamino shows that structured-preserving private data generation can be a powerful tool to help data scientists work with samples of sensitive data sets.

Biography

Ihab Ilyas is a professor in the Cheriton School of Computer Science and the NSERC-Thomson Reuters Research Chair on data quality at the University of Waterloo. His main research focuses on the areas of big data and database systems, with special interest in data quality and integration, managing uncertain data, machine learning for data curation, and information extraction. Ihab is a co-founder of Tamr, a startup focusing on large-scale data integration, and he is also the co-founder of inductiv (acquired by Apple), a Waterloo-based startup on using AI for structured data cleaning. He is a recipient of the Ontario Early Researcher Award, a Cheriton Faculty Fellowship, an NSERC Discovery Accelerator Award, and a Google Faculty Award, and he is an ACM Fellow.

All are welcome.
Enquires: Jian Pei (jpei@sfu.ca)

SFU Data Science Research Group
https://data.cs.sfu.ca