TEMPLE UNIVERSITY®

Fall 2021 Colloquium

Department of Computer and Information Sciences

Data Preparation: The Biggest Roadblock in Data Science

Dr. El Kindi Rezig

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

Monday, March 14th, 2022

Zoom Link: https://temple.zoom.us/j/96026357183

Abstract: When building Machine learning (ML) models, data scientists face a significant hurdle: data preparation. ML models are exactly as good as the data we train them on. Unfortunately, data preparation is tedious and laborious because it often requires human judgment on how to proceed. In fact, data scientists spend at least 80% of their time locating the datasets they want to analyze, integrating them together, and cleaning the result.

In this talk, I will present my key contributions in data preparation for data science, which address the following problems: (1) data discovery: how to discover data of interest from a large collection of heterogeneous tables (e.g., data lakes); (2) error detection: how to find errors in the input and intermediate data; and (3) data repairing: how to repair data errors with minimal human intervention. The developed systems are specifically designed to support data science development which poses particular requirements such as interactivity and modularity. The talk will feature demonstrations of data preparation systems as well as discussions of our developed algorithms and techniques that enable data preparation at scale.

Bio: El Kindi Rezig is a postdoctoral associate at the Computer Science and Artificial Intelligence Laboratory (CSAIL) of MIT where he works with Michael Stonebraker. He earned his Ph.D. in Computer Science from Purdue University under the supervision of Walid Aref and Mourad Ouzzani. His research interests revolve around data management in general and data preparation for data science in particular. He has developed systems in collaboration with several organizations including Intel, Massachusetts General Hospital, and the U.S. Air Force.