DATA SCIENCE TT FACULTY CANDIDATE 2023

Roee Shraga, Ph.D.
Post Doc, Data Lab
Khoury College of Computer Sciences
Northeastern University, Boston, MA

Tuesday, March 21, 2023 @ 11:00AM EST
Unity Hall 520

Title: Recovering Data Semantics

Abstract: In data science, it is increasingly the case that the main challenge is finding, curating, and understanding the data that is available to solve a problem at hand. To top it all, modern-day data is challenging in that it lacks many forms of semantics ("meaning of data"). Metadata may be incomplete or unreliable, data sources are unknown, and data documentation rarely exists. To address these challenges, the objective of my research is to recover data semantics throughout data discovery, versioning, integration, and quality.

In this talk, I will discuss current data science challenges and highlight two specific aspects of my research that assist with such challenges. In particular, I will present ALITE, the first scalable integration solution for tables that may have been discovered in data lakes (repositories of big data). ALITE relaxes previous assumptions that tables share common attribute names (which completely determine the join columns), are complete (without null values), and have acyclic join patterns. I will also introduce Explain-Da-V, a solution that explains dataset versions by generating data transformations that resolve data changes. Explain-Da-V represents a new research paradigm that explores the semantics of data versioning.

Bio: Roee Shraga is a Postdoctoral fellow at the Khoury College of Computer Science at Northeastern University in Boston. His research mainly revolves around data discovery and integration and combines techniques from data management, machine learning, information retrieval and human-in-the-loop. His research has been published in top-tier conferences such as SIGMOD, VLDB, SIGIR, WWW, and ICDE. He is a recipient of the Council for Higher Education [VATAT] scholarship for outstanding data science postdocs. He is also a recipient of several PhD fellowships including the Leonard and Diane Sherman Interdisciplinary Fellowship (2017), the Daniel Excellence Scholarship (2019), and the Miriam and Aaron Gutwirth Memorial Fellowship (2020).

Host: Prof. Elke Rundensteiner, Data Science