The University of Chicago Computer Science Department
PRESENTS:

"Data Discovery: Unleashing the Value of Data"

Raul Castro Fernandez
Massachusetts Institute of Technology

Abstract:
Organizations use only a small portion of all data they own. Consequently, most of the potential value is untapped. This happens because their analysts suffer a data discovery problem: when solving a task that requires data, analysts spend more time finding the relevant data than solving the task at hand. The core problem is that there is not adequate infrastructure to support the many different discovery problems organizations face. Hence, finding data remains largely a manual and time-consuming process.

In this talk I'll present Aurum, a system that radically changes how users interact with their organizations' data. With Aurum users can solve discovery problems in minutes instead of weeks. To achieve this, Aurum has three novel features: 1) it makes data discovery programmable so users can solve many different discovery problems by writing different programs; 2) it solves data discovery queries fast, so users can solve their problems in minutes instead of weeks; 3) it scales to large amounts of data, so no relevant data is left behind. In addition, I'll explain how Aurum handles not only structured data such as tables in databases, data lakes, and spreadsheets, but also unstructured data such as PDF files, word documents, and even conversations from Slack channels.

I'll conclude with a vision for how to make data easier to work with and to program, a key ingredient needed to exploit all data available in organizations and enable new applications.

Bio:
In my research I build high-performance systems for discovering, preparing, and processing data. I often use techniques from data management, statistics, and machine learning. At MIT I work with professors Sam Madden and Mike Stonebraker. Before MIT, I completed my PhD at Imperial College London with Peter Pietzu.

Thursday, April 4, 2019
3:30 pm
JCL 390
Host: Aaron Elmore