UNIVERSITY OF CHICAGO
DEPARTMENT OF COMPUTER SCIENCE

PRESENTS:

"Data Integration with Unreliable Sources"

Theodoros (Theo) Rekatsinas
Stanford University

Abstract:

Data integration is an essential element of data-intensive science and modern analytics. Users often need to combine data from different sources to gain new scientific knowledge, obtain accurate insights, and create new services. However, today's upsurge in the number and heterogeneity——in terms of format and reliability——of data sources limits the ability of users to reason about the value of data. This raises the fundamental questions: what makes a data source useful to end users, how can we integrate unreliable data, and which sources we need to combine to maximize the user's utility?

In this talk, I discuss how to assess and leverage the quality and reliability of data to make data integration more efficient. Specifically, I demonstrate how statistical learning is the key to managing large volumes of heterogeneous sources effectively. Building upon this observation, I introduce new solutions to classical data integration problems, such as data conflict resolution and data cleaning, and show that these solutions outperform their traditional counterparts by large margins. I finish with an outlook on how recent advancements in machine learning have the potential to streamline the construction of end-to-end data curation systems and bring data closer to users.

Bio:

Theodoros (Theo) Rekatsinas is a Moore Data Postdoctoral Fellow at Stanford working with Christopher Ré; he earned his Ph.D. in Computer Science from the University of Maryland, where he was advised by Amol Deshpande and Lise Getoor. His research interests are in data management, with a focus on data integration, data cleaning, and uncertain data. Theo's work on using quality-aware data integration techniques to forecast the emergence and progression of disease outbreaks received the Best Paper Award at SDM 2015. Theo was awarded the Larry S. Davis Doctoral Dissertation award in 2015. Website: http://stanford.edu/~thodrek/

Monday, February 20, 2017
2:30 pm
Ryerson 251
Host: Aaron Elmore