NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

CS SEMINAR

Title: **Helping Scientists Connect Their Datasets**

Speaker: Professor David Maier
Maseeh Professor of Emerging Technologies
Portland State University

Date/Time: 31 July 2015, Friday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

Chaired by: Dr Chan Chee Yong, Associate Professor, School of Computing
(chancy@comp.nus.edu.sg)

Abstract:

Scientific datasets associated with a research project can proliferate over time as a result of activities such as sharing datasets among collaborators, extending existing ones with new measurements, and extracting subsets of data for analysis. As such datasets begin to accumulate, it becomes increasingly difficult for a scientist to keep track of their derivation history, which complicates data sharing, provenance tracking, and scientific reproducibility. Understanding what relationships exist between datasets can help scientists recall their original derivation history. For instance, if dataset A is contained in dataset B, then the connection between A and B could be that A was extended to create B.

In our initial work, we developed a set of relevant relationships, proposed the relationship-identification methodology for testing relationships between pairs of datasets, developed a set of algorithms for efficient discovery of these relationships, and organized these algorithms into a new system called ReConnect to assist scientists in relationship discovery. We evaluated existing alternative approaches that rely on flagging differences between two spreadsheets and found that they were impractical for many relationship-discovery tasks. Additionally, a user study showed that ReConnect can improve scientists' ability to detect useful relationships between datasets.

While ReConnect helps with identifying relationships between two datasets, it is infeasible for scientists to use it for determining relationships between all possible pairs in a large collection. In this talk, we introduce an end-to-end prototype system, ReDiscover, that identifies, from a collection of datasets, the pairs that are most likely related. Our preliminarily evaluation shows that ReDiscover can predict selected relationships with high precisions and within reasonable computational cost.

Biodata:

David Maier is Maseeh Professor of Emerging Technologies at Portland State University. Prior to his current position, he was on the faculty at SUNY Stony Brook and Oregon Graduate Institute. He has spent extended visits with INRIA, University of Wisconsin at Madison, Microsoft Research and National University of Singapore. He is the author of books on relational databases, logic programming and object-oriented databases, as well as many papers in database theory, object-oriented technology, scientific databases and data-stream processing. He received the Presidential Young Investigator Award from the National Science Foundation in 1984 and was awarded the 1997 SIGMOD Innovations Award for his contributions in objects and databases. He is also an ACM Fellow and IEEE Senior Member and serves on the Board on Mathematical Sciences and Their Applications at The National Academies. He holds a dual B.A. in Mathematics and in Computer Science from the University of Oregon (Honors College, 1974) and a Ph.D. in Electrical Engineering and Computer Science from Princeton University (1978).