NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

CS SEMINAR

**Title:** **Data Cleaning, Linking, and Integration: From Raw Data to Actionable Information**

Speaker: Professor Wang-Chiew Tan
Computer Science Department
University of California

Date/Time: 9 July 2015, Thursday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

Chaired by: Dr Hsu, Wynne, Professor, School of Computing
(whsu@comp.nus.edu.sg)

Abstract:

Management and policy decisions are typically made based on information that is derived from datasets. To prepare datasets so that they are ready to be analyzed requires a number of tedious and, oftentimes, manual data curation activities, such as data cleaning, linking, and integration with other datasets.

In this talk, I will present some recent work on these three activities with particular emphasis on a recent work where we propose a query-oriented system for cleaning data with oracles. Unlike prior data cleaning techniques where the focus has largely been to correct all existing data upfront, our framework is driven by the correctness of query results, cleans data only as needed, and also permits one to augment the underlying dataset through the identification of missing tuples in the result of a query. Incorrect/missing tuples are removed/added to the result of a query through edits that are applied to the underlying dataset, where the edits are derived by interacting with domain experts which we model as oracle crowds.

We show that the problem of determining minimal interactions with oracle crowds to derive database edits for removing/adding incorrect/missing tuples to the result of a query is NP-hard in general and present heuristic algorithms that interact with oracle crowds to progressively clean the dataset, as needed. I will also present recent work on temporal record linkage and integration that allows one to identify which facts are temporally related and how they can be meaningfully combined together.

Biodata:

Wang-Chiew Tan is a Professor of Computer Science at University of California, Santa Cruz. She received her B.Sc. (First Class) in Computer Science from the National University of Singapore and her Ph.D. in Computer Science from the University of Pennsylvania. Her research interests are in general area of data management, with emphasis on topics such as (big) data integration and exchange, data provenance, crowdsourcing, and scientific databases. She is the recipient of an NSF CAREER award, a Google Faculty Award, and an IBM Faculty Award. She is the co-author of four best papers, a co-recipient of the 2014 ACM PODS Alberto O. Mendelzon Test-of-Time Award, and several of her publications have been invited to and appeared in special issues for selected papers. She has served on the program committees of top database conferences. She was the program committee chair of the International Conference on Database Theory (ICDT) 2013, she is currently on the VLDB Board of Trustees, and she is the 2016 ACM Principles of Database Systems (PODS) program committee chair.