UNIVERSITY OF CHICAGO
DEPARTMENT OF COMPUTER SCIENCE

PRESENTS:

"Data Analysis over Large Scale Data Commons - Managing the Tail at Scale for Scientific Workflows "

Robert Grossman
Center for Data Intensive Science
University of Chicago

Abstract:

Data commons co-locate data, storage and computing infrastructure with commonly used services, tools and applications for analyzing and sharing data to create an interoperable resource for the research community.   Data commons, such as the NCI Genomic Data Commons, reanalyze all of the data as new workflows and pipelines are developed.   This needs to be done as all the other services continue to run on a data commons, including browsing and exploring the data, and submitting new datasets.

For some applications, such as analyzing genomic data, there can be quite a long tail to the distribution describing the time required to complete workflows. One of the challenges with designing and operating data commons is managing the workflows that contribute to this tail. The importance of managing the tail of latency distribution for building responsive large-scale web services is well known.

In this talk, we describe the design and architecture of a large scale data commons. We also discuss some of the challenges posed by the long tail of scientific workflows when analyzing data in a data commons and some of the techniques for managing the tail.

Bio:

Robert Grossman is the Director of the Center for Data Intensive Science (CDIS); a Senior Fellow and Core Faculty in the Institute for Genomics and Systems Biology and the Computation Institute; and a Professor and Co-Chief of the Section of Computational Biomedicine and Biomedical Data Science in the Department of Medicine at the University of Chicago. His research group focuses on data science, data intensive computing, biomedical informatics and related areas.   He is also the Director of the not-for-profit Open Commons Consortium that develops and operates data commons and data clouds to support research in science, medicine, health care, and the environment.

Wednesday, January 18, 2017
3:00 pm
Ryerson 251
Host: Michael Franklin