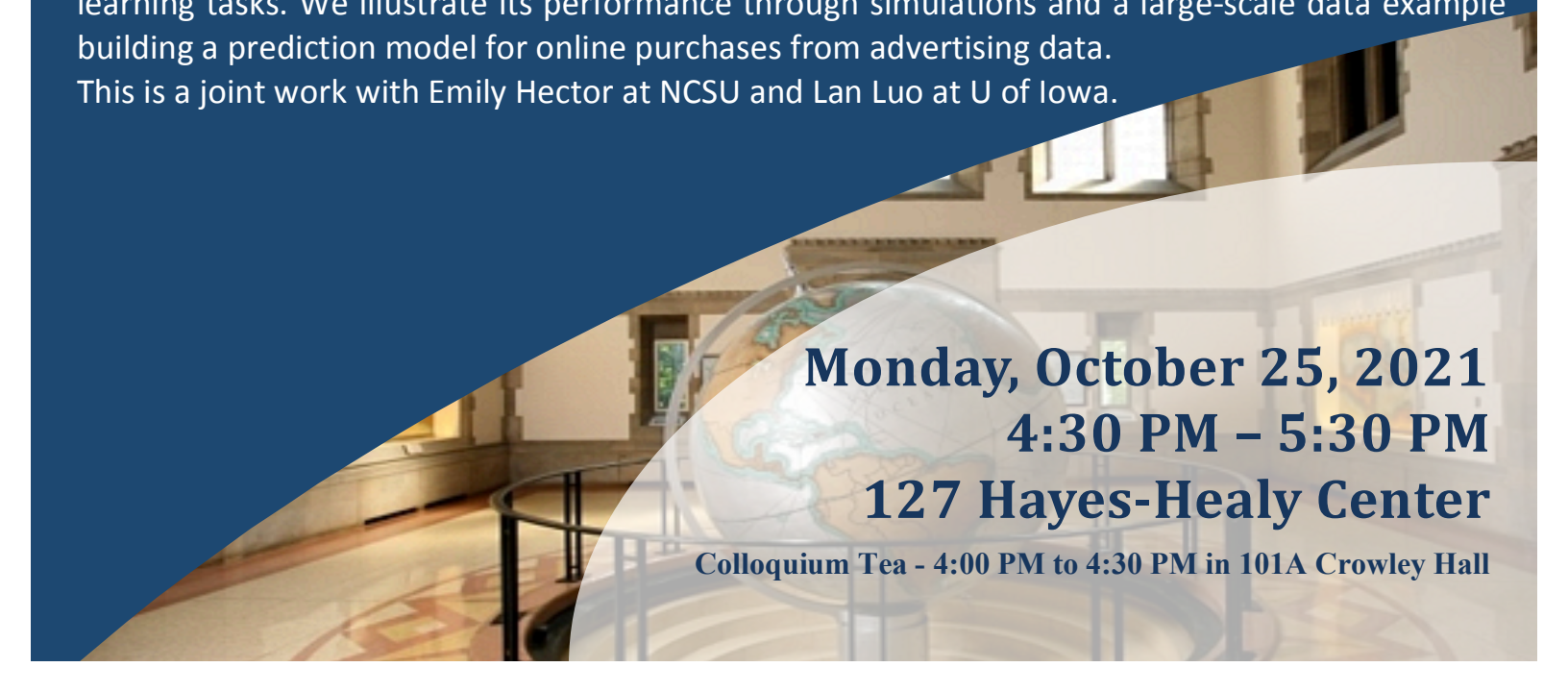
Department of Applied and Computational Mathematics and Statistics Colloquium

Peter Song

Department of Biostatistics
University of Michigan

Parallel-and-stream accelerator for computationally fast supervised learning with big data

Two dominant distributed computing strategies have emerged to overcome the computational bottleneck of supervised learning with big data: parallel data processing in the MapReduce paradigm and serial data processing in the online streaming paradigm. Although these two strategies are both common divide-and-combine approach, they differ in how they aggregate information, leading to different trade-offs between statistical and computational performances. We propose a new hybrid paradigm, termed a *Parallel-and-Stream Accelerator (PASA),* that uses the strengths of both distributed strategies for computationally fast and statistically efficient supervised learning. PASA's architecture nests online streaming processing into each distributed and parallelized data process in a MapReduce framework. PASA leverages the advantages and mitigates the disadvantages of both the MapReduce and online streaming approaches to deliver a more flexible paradigm satisfying practical computing needs. We study the analytic properties and computational complexity of PASA and detail its implementation for two key statistical learning tasks. We illustrate its performance through simulations and a large-scale data example building a prediction model for online purchases from advertising data.
This is a joint work with Emily Hector at NCSU and Lan Luo at U of Iowa.



Monday, October 25, 2021
4:30 PM – 5:30 PM
127 Hayes-Healy Center

Colloquium Tea - 4:00 PM to 4:30 PM in 101A Crowley Hall