

# Department of Applied and Computational Mathematics and Statistics Colloquium

**Bowei Xi**


Department of Statistics  
Purdue University

will give a lecture entitled:

*Large Complex Data: Divide and Recombine (D&R) with RHIPE*

## Abstract

D&R is a new statistical approach to the analysis of large complex data. The data are divided into subsets. Computationally, each subset is a small dataset. Analytic methods are applied to each of the subsets, and the outputs of each method are recombined to form a result for the entire data. Computations can be run in parallel with no communication among them, making them embarrassingly parallel, the simplest possible parallel processing. Using D&R, a data analyst can apply almost any statistical or visualization method to large complex data. Direct application of most analytic methods to the entire data is either infeasible, or impractical. D&R enables deep analysis: comprehensive analysis, including visualization of the detailed data, that minimizes the risk of losing important information. One of our D&R research thrusts uses statistics to develop “best” division and recombination procedures for analytic methods. Another is a D&R computational environment that has two widely used components, R and Hadoop, and our RHIPE merger of them. Hadoop is a distributed database and parallel compute engine that executes the embarrassingly parallel D&R computations across a cluster. RHIPE allows analysis wholly from within R, making programming with the data very efficient.



**Thursday, September 11, 2014**  
**4:15 p.m. to 5:15 p.m.**  
**127 Hayes-Healy Center**

Colloquium Tea

3:45 p.m. to 4:15 p.m. 154 Hurley Hall