BIOSTATISTICS SEMINAR

Chuanhai Liu

PhD, Professor of Statistics,
Purdue University

A Multithreaded and Distributed R for Big Data Analysis

The computer software R is one of the most popular computing tools for data analysis. In the past decade or so, tremendous efforts have been made to make R useful for big data analysis. These include Tessera, Revolution-R, and SparkR, to name a few. As we know, they are all making use of JAVA-based softwares such as Hadoop and Spark. In this talk, we introduce an entirely new alternative, a multithreaded and distributed R, called SupR. The prototype of SupR (http://www.stat.purdue.edu/~chuanhai/SupR/index.html) was made possible by modifying R (R-3.1.1) existing internal system implementation. The key features of the prototype include (1) a R-style front-end obtained by maintaining the existing R syntax and internal basic data structures, (2) a Java-like multithreading model, (3) a Spark-like cluster computing environment, and (4) a built-in simple distributed file system. With simple examples, including multithreaded Expectation-Maximization and distributed Linear Regression, we show how SupR can be potentially useful for big data analysis.

Thursday March 7, 2019 **3:30 pm - 4:30 pm**

Blue Cross and Blue Shield of North Carolina Foundation Auditorium

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH