**Tutorial on Parallel Programming in R**

Josh Hewitt & Henry Scharf

May 12, 2014

# 1 Description

As the size of data increases at a rapid pace, the number of individual computations for even standard analyes is growing at a staggering rate. The rate at which individual processors can perform these compuatations is no longer keeping pace with the volume of data, and so several useful techniques can become prohibitively slow to implement. Luckily, the cost of processors has continued to decline rapidly, and so multi-core systems are common place even in personal computers. Many of these slow techniques involve several independent tasks which may be spread over several cores, thereby significantly reducing the total computation time. In this tutorial, we will focus on identifying these situations, and using a few packages in R to 'parallelize' sequential code. We illustrate the process with several commonplace examples which will include some of the following:

(a) bootstrapping (CI for linear regression coef)

(b) cross validation ?

(c) simulation?

(d) sensitivity analysis in bayesian statistics?

(e) large datasets? (eg: genetic?)

(f) permutation tests?

R PACKAGES USED

(a) 'foreach'

(b) 'multicore'

(c) 'Rhadoop'

# 2 Outline and Objectives

1. Identify parallelizable computation tasks

2. Extend a working knowledge of R programming to take advantage of multiple cores

3.

4.

## 3 About the Instructor

Henry Scharf is a PhD student at Colorado State University with a strong background in both teaching and computational statistics. He received his Masters in Education from the University of Arizona, and is an instructor at CSU. He has worked in conjunction with the National Renewable Energy Labratory on questions surrounding prioritized compression of massive datasets sensistive to specific secondary analysis.

## 4 Relevance to Conference Goals

Theme 3: Model validation and comparison approaches.

Theme 4: Software and programming methods to obtain, clean, describe, or analyze data.