**Finding Combinatorial Effects in Big Data using Iterative Random Forest**Jonathon Romero, Ashley Cliff, Piet Jones, Ben Brown, Daniel Jacobson
Oak Ridge National Laboratory

romerojc@ornl.gov

**Abstract**

Discovering combinatorial effects of variables is an important problem to many scientific domains. To mathematically address all variables in a brute force manner requires a large amount of compute time that increases exponentially with the number of potentially interacting variables. For example, detecting epistatic associations between multiple genome variants and their effects on an organism (phenotypes) is a difficult problem in the field of genomics due to the exponential scale of possible combinatorial relationships between phenotypes and single nucleotide polymorphisms (SNPs). For sets of 8 SNPs, in a dataset that has ten million SNPs there are 2.48\*1051 possible combinations. Iterative random forests (iRFs) allow for the efficient exploration of the combinatorial space without having to exhaustively search the entire space. The iRF algorithm modifies the well-known random forest algorithm to handle variable importance and variable-interaction space in regression and classification by building multiple forests of decision trees, each with a random subset of the training data. These forests are then mined for recurring patterns. A generalized version, Tensor iterative Random Forest (TiRF), is able to effectively build forests that can be mined for interactions within a multidimensional X matrix, a multidimensional Y matrix, and interactions between multiple dimensions in X and Y, simultaneously. Scaling of this algorithm, to well utilize the resources offered via the Oak Ridge Leadership Computing Facility and allow for the faster creation of many trees, was done by building in data and message communication, via MPI, around an existing Random Forest implementation, Ranger. Applications range from gene interactions to materials physics to systems testing.