RISE Talk

Who? Javier Cabrera, Ph.D., Department of Statistics & Cardiovascular Institute, Rutgers University

What? "Data nuggets" tools for clustering big data

When? 12:00-1:00 on Tuesday, October 8

Where? Hall of Sciences, Room 326

Big data has created new challenges for data analysis due to the large size of the datasets, with millions of observations and/or thousands of variables which are typical of many medical and business applications. An issue with many standard clustering algorithms is that since they require pairwise distance calculation, they are limited by the number of observations to around 100K in a basic computer. A work-around is to conduct the analysis with a random sample of the dataset and a recent proposal is to replace the random sample with a set of "support vectors". The pitfall of these solutions is that the structure of the dataset, particularly at the tails or edges of the dataset, is not guaranteed to be captured very well. I will present is a new solution for analyzing large datasets through the concept of "data nuggets". These data nuggets reduce a very large dataset into a small collection of nuggets of data, each containing a center, weight, and a scale parameter. Once the data is re-expressed as data nuggets, we may apply algorithms that compute standard unsupervised and supervised statistical methods, such clustering.

"Data nuggets" are also applicable to principal components analysis (PCA), linear models, and other datamining methods.

The methodology will be illustrated with an example of single cell flow cytometry involving millions of individual cells.