**Mining the Big Data – Critical Features and Sample Selection**

The rapidly growing big data generated by connected sensors, devices, the web and social media, etc., have stimulated the advancement of data science, which holds tremendous potential for problem solving in various domains. How to properly utilize the big data to obtain useful analytics and, more excitingly, to build accurate models for knowledge discovery is a topic of great importance in data mining, and therein two issues arise: how to select a critical subset of features and how to select a critical subset of representative data points. This talk presents ongoing research that suggests: 1. the critical feature dimension problem is theoretically intractable, but simple heuristic methods may well be sufficient for practical purposes; 2. there are big data analytic problems where evidence suggest that the success of data mining depends more on the critical feature dimension than the specific features selected, thus a random selection of the features based on the dataset's critical feature dimension will prove sufficient; and 3. The problem of critical sampling has the same intractable complexity as critical feature dimension, but again simple heuristic methods may well be practicable in most applications; experimental results with several versions of the heuristic method are presented and discussed. Finally, a set of metrics for data quality is proposed based on the concepts of critical features and critical sampling.