Semester: **Summer 2010**
MS Student: **Md. Shiblee Sadik**
Chair: **Dr. Le Gruenwald**
Thesis Title: **OUTLIER DETECTION FOR DATA STREAMS**

ABSTRACT:
Outlier detection is a well established area of study for statistical data. However, most of the existing outlier detection techniques are designed targeting the regular data model, where the entire dataset is available for random access. Typical outlier detection techniques construct a standard data distribution or model from the entire dataset and execute their detection algorithms over each data point. Evidently these techniques are not suitable for online data streams where the entire dataset, due to its unbounded volume, is not available for random access. Moreover, the data distribution in data streams change over time which challenges the existing outlier detection techniques that assume a constant standard data distribution for the entire dataset. In addition, data streams are characterized by uncertainty which imposes further complexity. In this work we propose two outlier detection techniques, called Distance Based Outline Detection for Data Streams (DB-ODDS) and Automatic Outlier Detection for Data Streams (A-ODDS). Both techniques are based on a novel continuously adaptive data distribution function that addresses all the issues of data streams; but DB-ODDS identifies outliers depending upon a user-defined minimum neighbor density, whereas A-ODDS identifies the most deviated data points as outliers. We also present efficient and online implementations of the two techniques and experiments evaluating their accuracy and execution time using three real-life datasets from meteorological and energy consumption monitoring applications. The performance results show that DB-ODDS and A-ODDS are superior compared with existing techniques, and while DB-ODDS requires domain knowledge to function, it offers better accuracy and execution time than A-ODDS.