

# COMP3006 Computer Science Project

*Anomaly Detection in  
Complex Data Sets*

## Project Plan

Khoi-Nguyen Tran (u4315673)

## Background

Anomaly Detection is an important part of data mining with many real-world applications. Anomalies in a data set are irregular, inconsistent or discrepant data values relative to other values in the data set by some defined measure. Anomalies are usually the result of error in measurement or input of data values, contamination from other populations, or simply rare events.

The definition of anomalies is subjective and depends on the domain of the data set. For example, suppose a data set has values from a known normal distribution, then an anomaly could be defined as being 3 standard deviations from the mean; or suppose a data set has data that forms a cluster, then anomalies could be defined as data values that are far from the average center of the cluster by a certain unit of measurement.

Complex data sets are used in this project. Their definition is structured or unstructured data with different types of attributes. A few different types of attributes are numeric, categorical, logical, string, intervals, continuous, data streams, and many others. This project will focus on structured data with attributes such as continuous, categorical and numeric.

Detecting anomalies in complex data sets are difficult because many problems such as the ambiguity of defining measurement to classify anomalous data. For example, given data values with many numeric and categorical attributes, defining a measure of anomaly becomes difficult because of unclear numeric relationships or correlations between categorical attributes.

## Task Description

For my project, I will be looking at recent and past research on the project title and their associated implementations. My literature review is expected to cover the two well research areas of anomaly detection in numeric and categorical data sets, and the research of anomaly detection in mixed attribute and high-dimensional data sets.

Then my aim is to develop new ideas to the problem such as combining current methods, using additional algorithms such as classification algorithms to aid in finding anomalies in complex data sets, or to improve some aspect of developed implementations such as reducing memory consumption or reducing the execution time.

I will then implement those ideas and perform experiments to measure the effectiveness of those ideas. Refining and evaluating the ideas will follow,

which will form a cycle of implementation, refinement and evaluation of ideas until the result is satisfactory or the limited of time available ends.

If the above description of the task fails because of failure of implementations or the new ideas are insufficient for the problem, which should become clear weeks before the planned beginning of writing the report, then I will think about developing theory or define methods for possible improvements of existing theory.

## Plan

Week

- 1 Meeting supervisor and discuss potential topics
- 2 Finalise topic and begin literature review
  
- 3 Literature Review
- 4 Literature Review and form basic ideas for the problem
- 5 Refine and focus ideas, more literature review, begin evaluation
  
- 6 Evaluation of ideas and plan implementation
- 7 Implementation
- 8 Implementation and evaluation
- 9 Implementation and evaluation
  
- 10 Begin report and evaluate final results
- 11 Begin presentation and complete draft of report
- 12 Finalise presentation and report
  
- 13 Final checks, submit report and give presentation