1. Broad research topic: *Association Rule Mining*

2. Currently we have a focused topic: *Association Rule Mining and Decision Making for Closed-Ended Hierarchical Questionnaire Data*

3. We can turn this into a claim: *No one has used association rule mining to make decisions on closed-ended questionnaire data.*

4. Keywords: *("census" OR "application" OR "survey" OR "questionnaire" OR "poll" OR "canvass") AND ("closed") AND ("medical" OR "health") AND ("classification") AND ("anomaly" OR "anomalous" OR "imbalance" OR "rarity" OR "exception" OR "oddity" OR "inconsistency" OR "abnormality")*

5. Thesis statement: *Selecting a good job candidate, that matches a role, to progress to an actual medical is possible through association rule mining using only their closed health questionnaire responses.*

6. The main aim of the project is closely related to the most critical stakeholder and industry partner.

How can we apply machine learning techniques to a questionnaire to replace the role of high cost medical assessments used in selecting a candidate for a specific job role and yet still avoid the liability risk of an incorrect choice?

Question 1: Is it possible, in a timely manner, to reduce the need for a physical medical assessment for a job role by introducing a suitability predictor using only responses given in a medical questionnaire?

Question 2: Is it possible to improve upon the suitability predictor by allowing actual medical assessment results to be fed back into the live system?

Question 3: Would removing rare or anomalous candidates from the pool of candidates create a better suitability predictor?

Question 4: How to analyse and compare the results of repeat medical assessments from the same candidate for different job roles over time?

Question 5: How to verify and validate the above aims?

7. Our objectives are:

Objective 1. To classify a candidate into a small number of groups that give a sliding suitability score. Success will demonstrate that our industry partner is able to rely on the initial accuracy of the classification from this objective of at least 60%. Later objectives will refine this percentage.

Objective 2. To define a mechanism whereby results of physical medical assessments are fed back into the system for a better predictor. Success will show that our classification accuracy improves demonstrably by introducing dynamic membership functions.

Objective 3. To build an anomaly detection routine to predict a list of candidates of concern. Success will be measured by two measures. Firstly the ability to discover rare candidates from a dynamic feature set and also whether overall accuracy of classification improves with removal of such candidates.

Objective 4. To build a model whereby assessments maybe compared along a timeline so that assessments taken multiple times maybe analysed. Success will be measured by the reduction of candidate questionnaires for multiple roles. With this in mind the current system is being monitored for average questionnaire completion.

Objective 5. To evaluate the developed artefacts from the previous objectives. Success here involves comparing the developed artifacts using some very well defined methods such as confusion matrix, ROC graphs and F1 scores and so will be the most open to interpretation of success of all the objectives.