# COMP6237 Data Mining Group Project Brief
# Mario Bros

## Introduction of The Project

The project is going to choose a dataset to have some experiments done by applying some technical data mining skills on these data to build a predictive model and find the solution of the problem for the dataset.

## The Dataset and Problem

This project using the dataset collected from IBM Watson Analytics [1]. The data is taken from human resources department and contains various information from employees such as age, education, department, income as well as whether they have left their jobs or not. This is a human resource employee attrition problem which is going to find out if a given employee is most likely going to leave his job or not by applying the data into a trained model of a serious of machine learning algorithms.

## Data Analysis and Evaluation

The dataset has to be cleaned before applying the algorithm to train a model. Some columns can be omitted as they will to make any reasonable effects on the problem such as employee number, standard hours as all data contain the same number of 80 and over18 as they are all over 18 years old. In addition to this, the data will also be simplified a bit due to some large amount of difference within a column. For example, the monthly income column contains different amount of numbers. These will be simplified as low, medium or high based on a range of incomes. In this stage, the dataset also needs analysing by finding the correlation of the data by comparing the attributes. For example, most of the employees who have left their job also contain very high value of 'NumCompaniesWorked'.

## Algorithm Approaches

The project may use some machine learning algorithms such as decision tree, k-means clustering, random forest and SVM. Due to the early stage of the project, they are still under consideration only but the main purpose of applying this is going to build a predictive model to be able to solve the problem. Decision tree may also be used to find the balance of the data to divide the data to have further analysis such as finding what take the most effect on the problem within a specific department.

## Details of Group Members

| Name | Email |
|---|---|
| Zilan Huang | zh4u17@soton.ac.uk |
| Bang Du | bd1m17@soton.ac.uk |
| Zihan Wang | zw1u17@soton.ac.uk |
| Yang Zhou | yz12a17@soton.ac.uk |
| Da An | da1g17@soton.ac.uk |
| Mardhiyyah Rafrin | mr1n17@soton.ac.uk |

## References

[1] M. Stacker IV, "IBM," 2 April 2015. [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/.