# **Overview**

In the data world, there are many different titles and a lot of them are overlapping in terms of the titles and the skills. It can be confusing for both potential employees and employers. Project 4 was designed to discover what data roles are companies most frequently hiring for? The questions guided to achieve the objective for this project are:

1. Determine factors that have the significant impact on the salary
2. Determine factors that distinguish job category.

Followed by data science workflow, there are four major steps in this project:

1. Data collection
2. Data exploration
3. Salary prediction modelling
4. Job title prediction modelling

## **Data Collection**

Data collection was done by web scraping seek.com.au to go through all the jobs under ‘Data Scientist’ in Australia, and store them in csv format for further analysis. There are 2795 rows and 7 columns in this dataset.

## **Data Exploration**

The information on seek.com are in different formats, for example, salary in daily/weekly/annually rate, and many of the job ads did not have salary information. Also, due to the limited ability of scrapy and the complexity of the html query, the salary column also includes job types in it. That’s why data cleaning is required to standardise all the data. Therefore, salary had to be updated for these job ads in order to build the prediction models. These are updated based on average salary. The title, location and area columns have been segregated into smaller category in order to remove the noise. For description column, snowballstemmer and lemmatizer have been implemented to remove any ‘ing’ and ‘y’ at the end of the each word and standardized all the verb and plural in the sentence.

## **Modelling**

Salary column have been grouped into three ranges – medium high and low. Therefore, this model should be treated as a classification problem rather than regression problem. The baseline accuracy for medium salary is around 89%. A few algorithms including logistic regression, random forest and decision tree with maximum depth of 5 were tested. Random forest ranked the highest accuracy score of 95%. The models feature importance information also informs that most data jobs are in located in Sydney CBD, followed by Melbourne. Date and job category don’t seem to be as important.

Similar to salary, job category has been divided into six groups – Senior, Entry Level, Data Engineer, Data Analyst, Data Scientist and Other. Random forest algorithm still outperforms logistic regression and decision tree. The most required data scientist in the job market are ‘senior’ and ‘engineer’ compared to other data-driven professions. There is no deny that machine learning is the state-of-the-art approaches in many business filed. It is great if someone knows how to use DecisionTreeClassifier.score(), however, what companies required in daily life is someone that can build and deploy models from the scratch.