Big Data Group Project – Phase 1

# Team Members and Contribution

The team members, their roles and responsibilities for the project are summarised below. With these defined roles the aim is that each individual team member contributes 25% to the total effort required to the completed project.

* **Alec David Vukovich**
  + Team lead – responsible for overall coordination of team meetings and ensuring the project is on track.
* **Darragh Brendan Ruddy**
  + Presentation lead – responsible for the coordination of phase 4 of the project.
* **Jemal Mohammed-Nour**
  + Machine Learning lead - responsible for the coordination of phase 3 of the project.
* **Matija Zivkovic**
  + ELT lead – responsible for the coordination of phase 2 of the project.

During each phase of the project the team lead for that phase will be responsible for delegating tasks to other team members as required to ensure the workload is evenly spread across the team.

# Introduction to the Dataset

The dataset selected for this project was collected from the popular crowdfunding platform Kickstarter and is publicly available on Kaggle via the URL below. Crowdfunding aims to raise small amounts of money from a large group of people to achieve funding goals. To reach a large audience, projects are usually listed on internet crowdfunding platforms, of which Kickstarter is one of the largest. **Dataset URL** [*https://www.kaggle.com/kemical/kickstarter-projects*](https://www.kaggle.com/kemical/kickstarter-projects)

The dataset consists of two .csv files with data extracted from Kickstarter over two periods of time. Combined these have over 700,000 records with 13 to 15 features. The features are summarised below:

1. ID – Unique identifier for the project.
2. name – name of the project.
3. category – specific category of the project.
4. main\_category – general category of project.
5. currency – Currency in which the funding is sourced.
6. deadline – deadline for the crowdfunding project to reach its funding goal.
7. goal – fundraising goal in the project currency.
8. launched – date and time the crowdfunding project was launched.
9. pledged – amount of money raised in project currency.
10. state – the state of the project e.g. cancelled, successful or failed.
11. backers – number of individuals pledging money.
12. country – country from which the project is listed.
13. usd pledged – the money raised normalized to USD. This conversion is done by Kickstarter.
14. usd\_pledge\_real\* - the money raised normalized to USD (done by Fixer.io API)
15. usd\_goal\_real\* - the goal normalized to USD (done by Fixer.io API)

\*Feature is only present in one .csv file in the dataset.

# Objectives and Motivation

Crowdfunding is a multibillion-dollar industry and being mainly internet based there are vast amounts of data being generated which can be leveraged to explore how to better tap into this market. This project aims to explore the variables which contribute to a successful project, either by achieving the funding goal or raising a large amount of money relative to similar projects. Ultimately this project will identify the “recipe” for a Kickstarter project which gives it the best chance of success.

# Proposed Methodology

A concept of the project methodology is summarised below.

The dataset will be extracted from Kaggle in the form of two .csv files with a structure as described in the Introduction to the dataset section. This data will be explored to detect inconsistent, incomplete, or inaccurate data and depending on the context will be repaired (e.g. imputation) or dropped. Features will also be transformed ready for the machine learning model, for example setting the “state” feature as binary whether the project is successful or not. The ID column will be dropped completely as this feature is not relevant to the goals of the project.

Once the data is cleaned, we will perform exploratory data analysis (EDA) on the clean data set. Additional features will also be engineered from the data set to feed into the ML model. Some examples include splitting the date-time features into individual features for year, month, day etc. The exchange rate at the time of the project can also be derived from the data to normalize the “goal” feature to USD. A feature can be generated by subtracting the launched date from the deadline date to find out how long a project is open for. More opportunities for additional features will be explored in stage 2 and 3.

For this project, a supervised classification model will be implemented to attempt to classify with some confidence whether a project will be successful or not based on select features, either original or engineered, from the dataset. We will also implement a regression model to see if we can predict how much money will be raised for a project, or how many backers will be attracted. The final pipeline will take a prospective project’s details and provide a prediction of how much money will be raised, how many backers will be attracted, and whether the project will be successful or not.