**Capstone Project Proposal**

1. **What is the problem you want to solve?**

I want to build a anime recommendation system.

1. **Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

My client anime websites. A good recommendation system will bring more traffic to their site because it makes enables users to find new anime to watch with minimal effort. With more traffic, the anime websites earns more from ads and membership subscription.

1. **What data are you going to use for this? How will you acquire this data?**

MyAnimeList.com is like the IMDB of japanese content, providing anime info and user ratings for anime. The score of a certain anime is actually the global mean of all the ratings given by the users

The MyAnimeList crawled dataset is available on Kaggle; it contains information about anime and people who watch them. The dataset contains 3 files:

* AnimeList.csv contains list of anime, with title, title synonyms, genre, studio, licensor, producer, duration, rating, score, airing date, episodes, source (manga, light novel etc.) and many other important data about individual anime providing sufficient information about trends in time about important aspects of anime. Rank is in float format in csv, but it contains only integer value. This is due to NaN values and their representation in pandas.
* UserList.csv contains information about users who watch anime, namely username, registration date (join\_date), last online date, birth date, gender, location, and lots of aggregated values from their anime lists.
* UserAnimeList.csv contains anime lists of all users. Per each record, here is username, anime ID, score, status and timestamp when was this record last updated.

1. **In brief, outline your approach to solving this problem (knowing that this might change later).**

Using the cleaned data, we will filter users who have rated at least 100 anime. Using this data, we will use 75 rows to train the recommendation system and 25 rows to test. To test the performance of the recommender, we will compare the predicted user score for the anime in the test data with the actual user score. We will train and compare a number of different machine learning (KNN,SVD,…) and select which one gives the best result.

1. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

Code in iPython notebook, and a report.