# CAPSTONE 2 - ABSTRACT

### Recommendation Systems on Movie Lens

**Title:** Build a recommendation system to find the most similar movies:

**Dataset:** https://www.kaggle.com/bakostamas/movie-recommendation-algorithm/data

There are different versions of the dataset (100K observations, 1M, 10M, 20M observations). Given the computational resources I have access to, choosing a dataset with the optimal sample size that my machine can handle is crucial. For this project, I am working on the dataset with 100K rows.

This dataset contains multiple files. Here is the list of files and corresponding features.

u.data -- This is a tab separated file contains 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies.

user id | item id | rating | timestamp.

Note: The time stamps are unix seconds since 1/1/1970 UTC

u.info -- The number of users, items, and ratings in the u data set.

u.item -- Information about the movies (tab separated) There are 19 genres (dummy variables). 1 indicates the movie is of that genre, a 0 indicates it is not. Note: movies can be in several genres at once.

movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation | Children's | Comedy | Crime | Documentary | Drama | Fantasy | Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi |Thriller | War | Western |

u.genre -- A list of the genres.

u.user -- Tab separated file with Demographic information of the users.

user id | age | gender | occupation | zip code

Note: The user ids are the ones used in the u.data data set.

u.occupation -- A list of the occupations.

**Goal:** The goal of this project is to apply content-based methods and collaborative filtering and build a recommendation system.