With ever increasing internet and cellular speeds and dwindling bandwidth costs, watching movies on-demand on streaming platforms like Netflix, Hulu, and Amazon etc. has become common place. Despite providing movie viewers the luxury of streaming countless movies at the click of a button, movie viewers end up spending significant leisure time trying to search/browse and decide which movie to watch. Many times viewers compromise to watch less exciting movies because they are not able find a better movie which is available. Streaming platforms compete hard and spend huge money to improve their movie collection, but viewers often complain that there are no good enough movies available to stream. Viewers often learn about a great movie which was always available for streaming but watch it only after a friend suggests it. So, it does not need much convincing to anyone that the current recommendation systems have inefficiencies. An improved movie recommendation system benefits movie viewers and streaming providers.

Our team is proposing to leverage data science and address some of the inefficiencies in the movie recommendation systems. We intend to use publicly available datasets from Netflix (<https://www.kaggle.com/netflix-inc/netflix-prize-data>) and IMDB (<https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>) to build a recommendation system.

Netflix dataset contains information about how 480,189 Netflix consumers who have rated several of the 17,770 movies over a period of several years between 1999 and 2005. The data provides insights about how a particular customer has rated various movies at various time periods.

Example: Customer 1488844 rated the 1997 movie ‘Character, with 3 stars on 2005-09-06 and rated the 2004 movie ‘What the #$\*! Do We Know!?’ with 4 stars on 2005-05-12.

As you can see the data does not contain much information about the attributes of the movie other than the year and title. We intend to collate Netflix dataset with IMDB dataset which provides rich information about 23 attributes for each movie like director name, duration, actors, facebook likes etc.

It is our hypothesis that the collation of Netflix and IMDB datasets will improve the recommendations given to the customers. Upon doing data analysis using data science algorithms, we are hoping to discover newer way to make recommendations to movie viewers.

We intend to analyze the data using the following techniques

* Principal Component Analysis – Identify which features of the movie have the most influence on the ratings.
* Classification and Clustering – Do classification and cluster analysis to identify similarities among movies and customers to make recommendations.
* Time series analysis – Surface any seasonal trends in movies watched by customers.
* Bayes Net Analysis.

We run the risk of finding out that the datasets are insufficient to draw any meaningful conclusions. It could be because of lack of common movie titles between IMDB and Netflix datasets or other reasons. To mitigate this risk we can source additional data with webpage scraping programs using python programming, we can also limit the scope of the recommendation system to only the movie titles which are common to both datasets. Decision on the mitigation strategy will be taken keeping in mind the course timelines. We goal is to have a functional recommendation system ready before course completion.