# Movie Rating Prediction

**Question:** Given a movie, what is the possible score (in a scale of 5) is the movie?
**Data:** The Movies Dataset
Reference to: `https://www.kaggle.com/rounakbanik/the-movies-dataset`

In today's society, entertainment industries is growing in a rapid pace. Along with the growth of the industry, entertaining products are being made on each day. New movies, drama and TV series are published every day. Audience is to these products as consumers is to merchandise. Consumers choose merchandise according to its quality and price, which are quantifiable metrics. If a consumer is unsatisfied with the merchandise, he or she can choose to return and refund. But what about the audience of movies and other video products? No refunds are ever made for the time they spent watching them. Thus, predictions of the qualities those entertaining products is essential for people to better choose and allocate their time on movies. In this specific case, we focus on movie qualities. Even though every audience has his or her own tastes, a predicting score can serve as a crucial metric for them. For those who are hesitating because of quality concerns of a movie, this score might be a critical factor in making their decision; for those who still want to have a try, this score serves as a baseline expectation.

The dataset we have chosen allows us to answer such question mainly because of the following points:

- **Data Coverage** By coverage, we refer to the type of information that is given in the dataset. The more types of information there are, the more comprehensive we could be when analyzing the data. Namely, we are given the casts, keywords, genres, release dates and all kinds of other information (meta data). Such various aspects of information will help reduce variance and bias. The types of information included also guarantees the scope of our analysis.

- **Data Size** The data size is large enough, as there are over 45,000. The number of movies included in the dataset guarantees the depth of our analysis.

- **Feasible to Train a Model** In this dataset, we are already given the scores of these 45,000 movies in ratings.csv. With such a file, we can easily divide the data into training set, validation set, and testing set in the ratio of 8:1:1. Given such large amount of data, we can train a model directly without using cross validation.