Names: Tony Oh (do256), William Rhee (wr86)
STSCI 4100: Multivariate Analysis
Spring 2025
24 March 2025

## [1] Problem Background

- **IMDb documents** reviews of released movies and films in non-commercial datasets — these datasets contain vast amounts of metadata regarding viewer ratings, genres, release years, and more. In this project, we want to understand if these factors dictate any trends in film popularity and genre success.

## [2] Question(s) to be Studied

Some datasets focus on short films and episodes. For us, we will focus on **movies** (which will **later** require data filtering).

But for now, here are *possible* questions we can consider:

- How have **movie** ratings changed over time?
- Are there certain **movie** genres that consistently receive higher ratings?
- Do modern **movies** receive more polarizing ratings compared to past films?
- Does the **number** of votes have any effect on ratings?
- How does the **director** change the rating of a movie?

## [3] Data Source

Here is the link to IMDB's non-commercial datasets:
https://developer.imdb.com/non-commercial-datasets/

Among these datasets, we will look at the following .tsv files:

- title.basics.tsv
- title.ratings.tsv
- title.crew.tsv
- title.principals.tsv

## [4] Head of Dataset

There are four data frames.

```
> head(basics)
     tconst titleType           primaryTitle          originalTitle isAdult startYear endYear runtimeMinutes                  genres
1 tt0000001     short             Carmencita             Carmencita       0      1894      NA              1        Documentary,Short
2 tt0000002     short Le clown et ses chiens Le clown et ses chiens       0      1892      NA              5          Animation,Short
3 tt0000003     short           Poor Pierrot         Pauvre Pierrot       0      1892      NA              5 Animation,Comedy,Romance
4 tt0000004     short            Un bon bock            Un bon bock       0      1892      NA             12          Animation,Short
5 tt0000005     short       Blacksmith Scene       Blacksmith Scene       0      1893      NA              1                   Short
6 tt0000006     short      Chinese Opium Den      Chinese Opium Den       0      1894      NA              1                   Short
> head(crew)
     tconst directors   writers
1 tt0000001 nm0005690      <NA>
2 tt0000002 nm0721526      <NA>
3 tt0000003 nm0721526 nm0721526
4 tt0000004 nm0721526      <NA>
5 tt0000005 nm0005690      <NA>
6 tt0000006 nm0005690      <NA>
> head(principals)
     tconst ordering    nconst        category                    job characters
1 tt0000001        1 nm1588970            self                   <NA>     [Self]
2 tt0000001        2 nm0005690        director                   <NA>       <NA>
3 tt0000001        3 nm0005690        producer               producer       <NA>
4 tt0000001        4 nm0374658 cinematographer director of photography       <NA>
5 tt0000002        1 nm0721526        director                   <NA>       <NA>
6 tt0000002        2 nm1335271        composer                   <NA>       <NA>
> head(ratings)
     tconst averageRating numVotes
1 tt0000001           5.7     2142
2 tt0000002           5.5      290
3 tt0000003           6.4     2178
4 tt0000004           5.3      186
5 tt0000005           6.2     2912
6 tt0000006           5.0      209
```