# An Analysis Of Metadata To Predict Successful Movies

Group A22: Malinda Ham, Jeremy Johannsen, Uyen Tran

- **What features make us enjoy a movie?**
- **What characteristics can predict the success of an upcoming movie?**

**Motivation and Background:**
As movie fans, it can be difficult to understand what makes us like certain movies over another but we can usually identify aspects that we are drawn to, such as cast, director, genre, and plot. We are interested in finding out what the preferences of the general movie-going public are and using that information to predict which movies will be successful. In addition to satisfying our curiosity as movie fans, this information is useful to film studios and investors when deciding how to produce their movies and theaters to decide how to allocate screens and schedule showings.

## 1. How do the features among the most successful movies compare to the features in a larger set of movies?

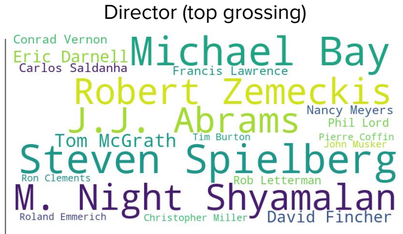
**Methodology:**
- Extract names of actors, directors, plot keywords, and genres from each movie.
- Create subsets of the most successful 1000 movies based on gross profit, popularity, and average user-voted rating
- Identify features that appear significantly more often in each subset successful movies than the entire movies dataset using a right-tailed hypothesis test with the assumption that appearance rates in the entire dataset are binomially distributed

**Results:**
- Significant features of each type were identified for each success metric

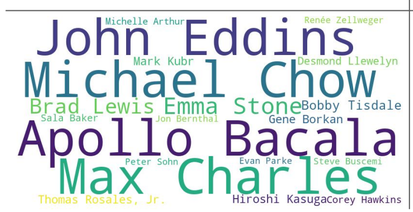
Genre (top grossing)



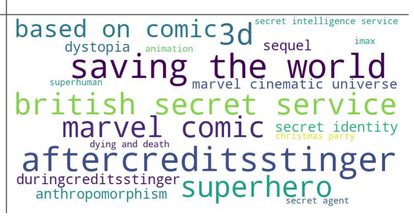

Director (top grossing)




Cast (top grossing)




Keywords (top grossing)

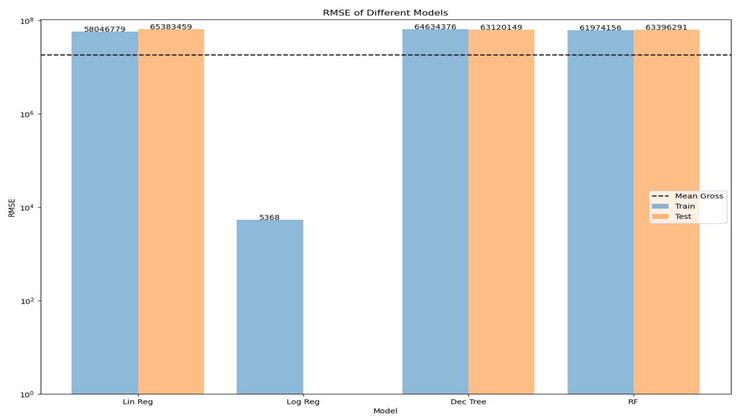



- Many more significant features than could be shown were found for each type and metric except genre, even with p-hacking to adjust for the large population size
- Significant cast and directors were different between success metrics, with more noticeable overlap for genres and keywords
- Many significant keywords are specifically related to Marvel movies
- Very uncommon values may be overrepresented in significant values if they happen to appear in a successful movie

## 2. Based on the keywords found, how can we predict whether the upcoming movies will be popular or not?

**Methodology:**
- Convert the keywords into a one hot encoded data frame
- Train 4 machine learning models, shown below, with the keywords as it features and gross as its target
  - Linear/Logistic Regression
  - Decision Tree and Random Forest Regressions
- Analyse the performance based on root mean squared error and coefficient of correlation
- Take one of the models and analyse how much it weighed each keyword, finding the most important keywords

**Results:**
- Most of the models had a large RMSE, with logistic regression giving a surprisingly, and worrying, low error (RMSE comparison shown in the graph below)

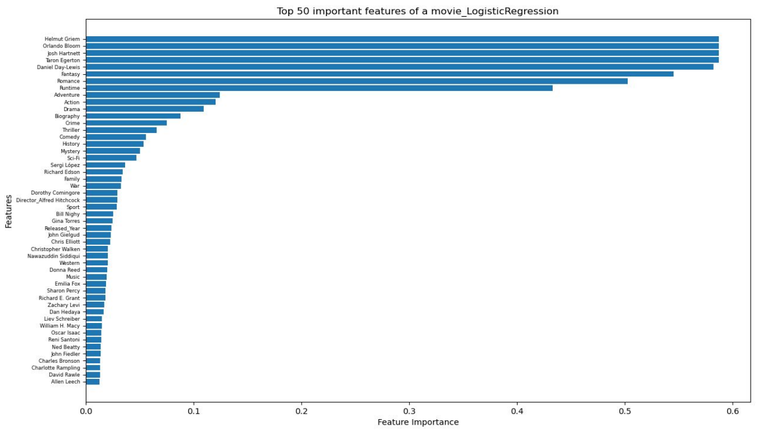


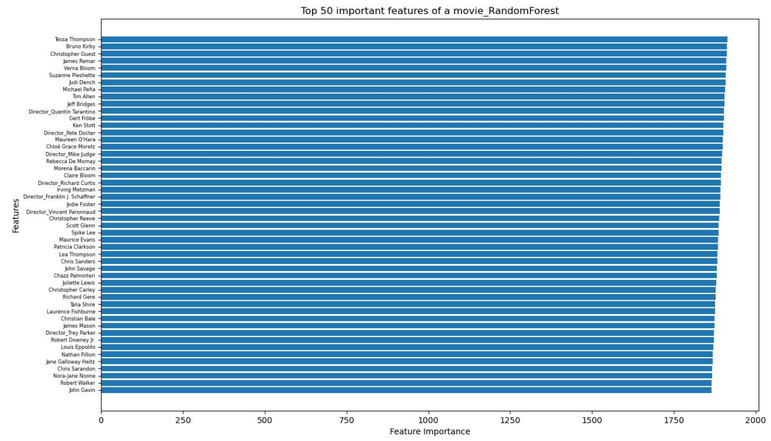

- Analysing the logistic regression's decision making showed it favored movies with many keywords and ones that containing 'soldier' as one of its keywords
  - The model also disfavored movies containing the 'woman director keyword', leading to questions about data set bias
- Overall, the models were very dependant on the specific samples they were feed, leading to different seeds having widely different error, within a similar distribution
  - Logistic regression occasionally jumped up to have a comparable error to the other models

## 3. What are the most important features that a movie should have?

**Methodology:**
- Combine 2 datasets: movies_metadata and imdb_top_1000
- Selected features: Genre, Director, Stars, Released_Year, etc.
- Target: Gross
- Logistic Regression
- Random Forest Regression

**Results:**
- **LR: more on genres**
  - Genres: Adventure, Action, Fantasy, Drama
  - Cast: Orlando Bloom, Zachary Levi
- **RF: more on cast and directors**
  - Cast: Robert Downey Jr., Morena Baccarin (Deadpool), Chloë Grace Moretz
  - Directors: Pete Docter (Pixar), Chris Sanders (Lilo & Stitch), Trey Parker (South Park, Despicable Me 3)




**Future work:**
- Choose a datasets better-suited to answering our questions
  - Consider bias factors (underrepresentation of certain groups in filmmaking)
  - Consider more recent movies to make better predictions about future movies and exclude deceased cast or crew members
  - Remove outliers that obscure general conclusions (e.g. Avatar)
  - Remove uncommon or overly-specific keywords
- Better training for machine learning models
  - Larger amount of data ( >> 10,,000 observations)
  - Test different hyperparameters (Max depth, C value, etc.)
  - Better encoding for keywords
  - Would require to model to train for longer
  - Potentially more broadly applicable results
    - Not as dependent on what movies get selected for testing