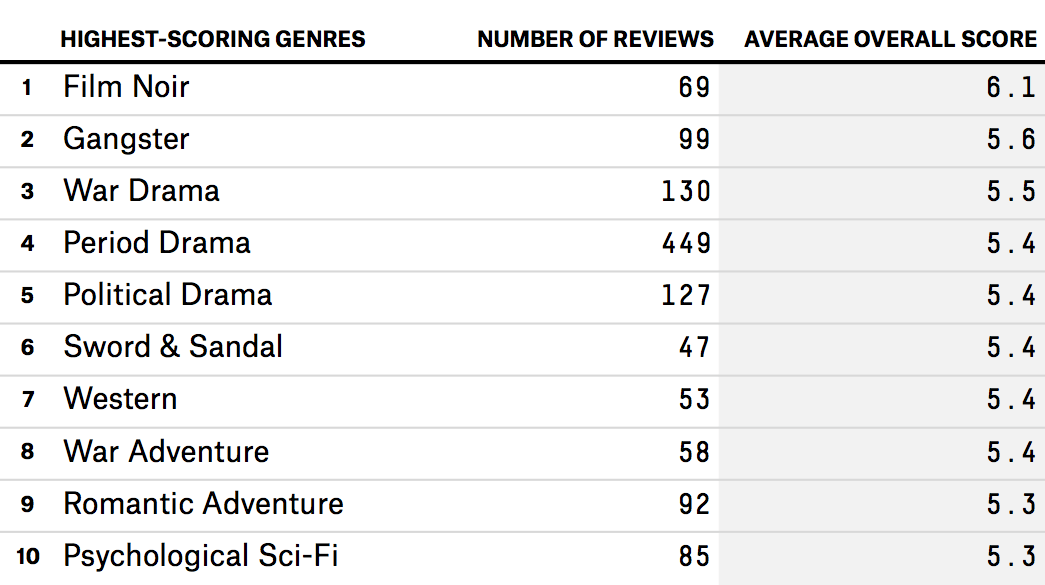
W201-5: Selecting Movie Scripts the

Data Science Way

Natarajan Shankar , Varadarajan Srinivasan, Leslie Teo, and Qian Yu



**W201-5 Final Case Study**

**Summer 2016**

# Introduction

Film-making is a highly collaborative endeavor and what makes a successful movie is a complex mix of story, cast, and technical movie-making under overall orchestration of a director. Nevertheless, there is a strong case to support the idea that movie script(or the screenplay) is the most vital component of a film’s success.[[1]](#footnote-2) At the very least, even if a good script is not sufficient to guarantee a successful movie, it is necessary.

At the same time, movie studios face a difficult challenge when choosing movie scripts. First, there are a lot of submissions i.e. between 50,000 to 250,000 per year.[[2]](#footnote-3) Second, only a few are chosen - perhaps 50 per year for larger production. Third, while only a few are chosen, history is littered with many examples of hit movies whose scripts were originally rejected.[[3]](#footnote-4)

At the very broad level, the movie industry is beginning to use data science to produce more successful movies. Specific applications include methods to analyze consumer preferences, including analyzing feedback via social media platforms and real-time sensors tracking individual emotional response during movie screening; construction of tools to predict success (win an Oscar, attendance); or tools to design better marketing strategies.

Though, some technology has been used to help write and analyze scripts, these approaches have, however, so far been rudimentary.

The Business Challenge (Core Question)

We are a new data science team in MovieBox, a major Hollywood studio. We have yet to prove ourselves but we have been given a free hand to use “data” to help improve the company's success at the box office.

We propose to design **a tool that will analyze movie scripts and rank them for potential success, called MovRank.** We will operationalize and measure success as: 1) Recommended script’s attractiveness (score, 0-5) when compared against publicly available ratings, 2) Recommended script’s potential for commercial success and 3) over time, percentage of recommendations which are produced.

We think this makes sense for three reasons.

* First, as noted above, while not the only factor, a good script is a key to a film’s success. Our tool will improve the efficiency of script evaluation and filtering - this should increase the number of “**good**” scripts and eventually hit movies.
* Second, our tools should not only increase the number of “**good**” scripts but also the **quality (potential for success)** of the chosen especially when we expand our tool to incorporate broader data sets. We note that our company, like others in the industry, are increasingly reliant on sequels and remakes for profits. Our tool should enable us to find more original gems.
* Third, our tools will reduce the likelihood of rejecting high potential scripts.

To be accepted as an alternative, our tool must do better than current approaches. One possible benchmark is the “The Blacklist” ([www.blckist.com](https://blcklst.com/)) site which lists human-picked “most liked” but yet to be produced scripts. Our tool would be a success if it leads to an improvement in speed to market (by 50% or more) and a rise in the conversion ratio of movies i.e. number made (which is currently around 33%).

Research Design

## Research Question & Models

The broad research question is fairly simple: ***Can MovRank analyze movie screenplays/scripts and predict how successful (audience likability and box office success) a movie based on the script will be?***

MovRank sets out to rank a script’s potential (score, 0-5) based on characteristics in the script, combined with external data on reviews, consumer preferences, box office, and casting.

1. First, our algorithm learns the characteristics of movies scripts themselves. We use linear classification method to analyze scripts and to categorize them into specific attributes/categories (genre, location, themes etc.) Further, the un-produced scripts are converted into a bag of words/word clusters.
2. Using word clusters, we apply clustering and document retrieval algorithm such as k-nearest neighbor or k-means to select scripts that are most similar to the past produced movies. This becomes an important mechanism to assess un-produced scripts.
3. These document clusters of similar past produced movies and corresponding data on user ratings and box office financial performance are fed into an aggregator module which computes the predicted rating and predicted box-office revenue for an un-produced script. These measures are used to develop the ranked list of script recommendations
4. Last we mine public/internet data for key themes on user preferences and adjust ranking to accommodate shift in user preferences based on recent trends.

We use **out of sample and cross-validation samples** to further refine our model.

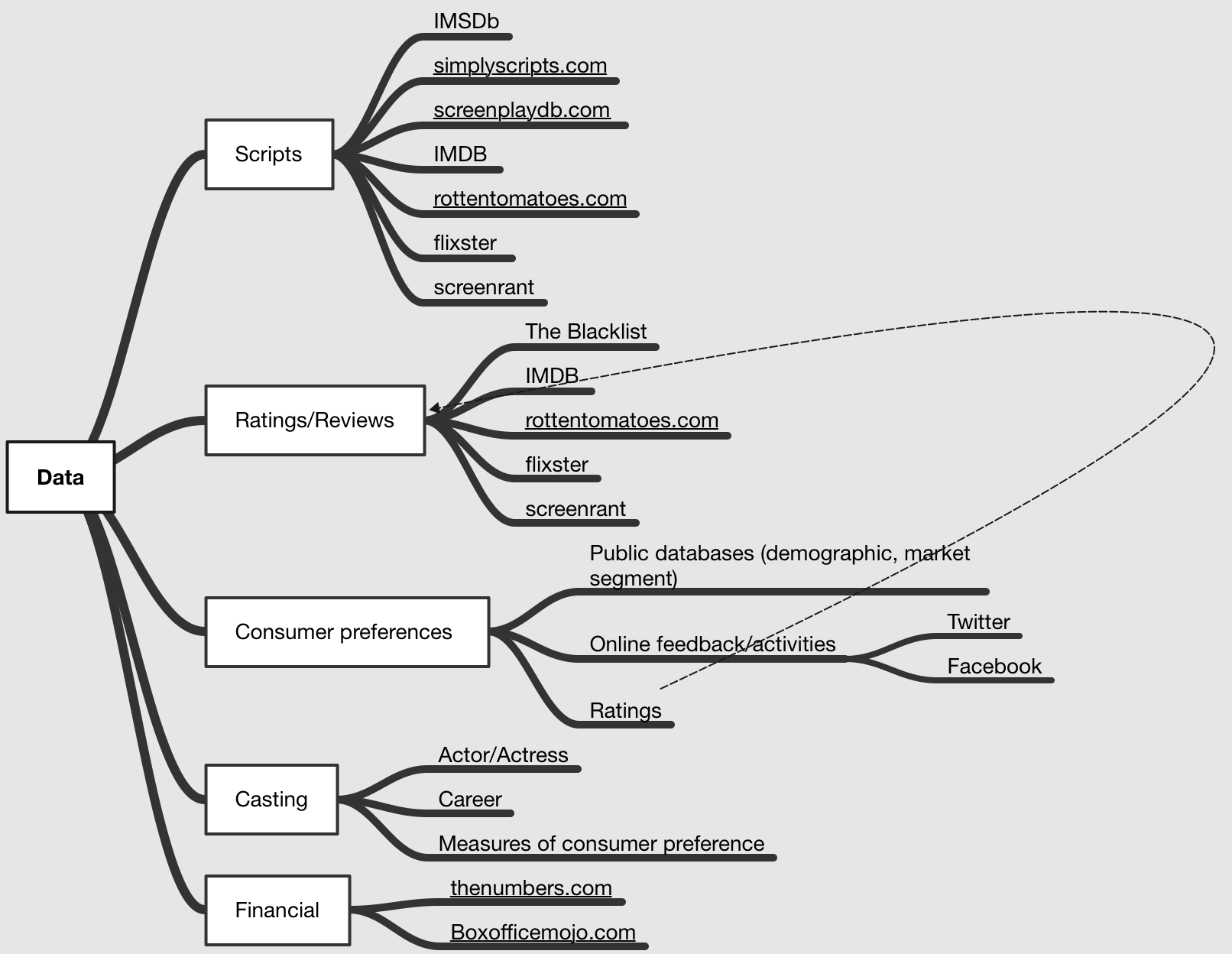
We emphasize **iteration and learning** in this exploration, as we add additional factors. We don’t have a good theoretical model~~,~~ and preferences change. Another reason is that individual behavior can change. In this respect, building machine learning algorithms that can re-validate results and re-learn are useful.

## Data Collection and Issues

### *Building the Algorithm - Data Needed for Training Sample*

To train MovRank, we source produced and un-produced scripts from various internet databases such as flixster, imdb.com, IMSDB, rottentomatoes, screenplaydb.com, and simplyscripts.com.[[4]](#footnote-5) (Figure 1) Scripts that have been produced provide some reference to estimate or validate our scores. We are cognizant of two challenges. When we observe crowd sourced rankings, these are based on the movies not scripts per se. Second, we only observe movies that have been produced. A lesser challenge is that given each website uses a different scale, these ratings will have to be transformed to a standardized scale of measurement.

Figure 1. Some Required Data



Boxoficemojo.com and thenumbers.com offers data on boxoffice performance and production budget. IMDB, Rottentomatoes, flixter, screerant etc. offers user and critic reviews which are beneficial to understanding what aspect of the movie resonated with the audience. Further, we will also mine Twitter, Facebook and movie discussion forum data to further delve into audience psyche.

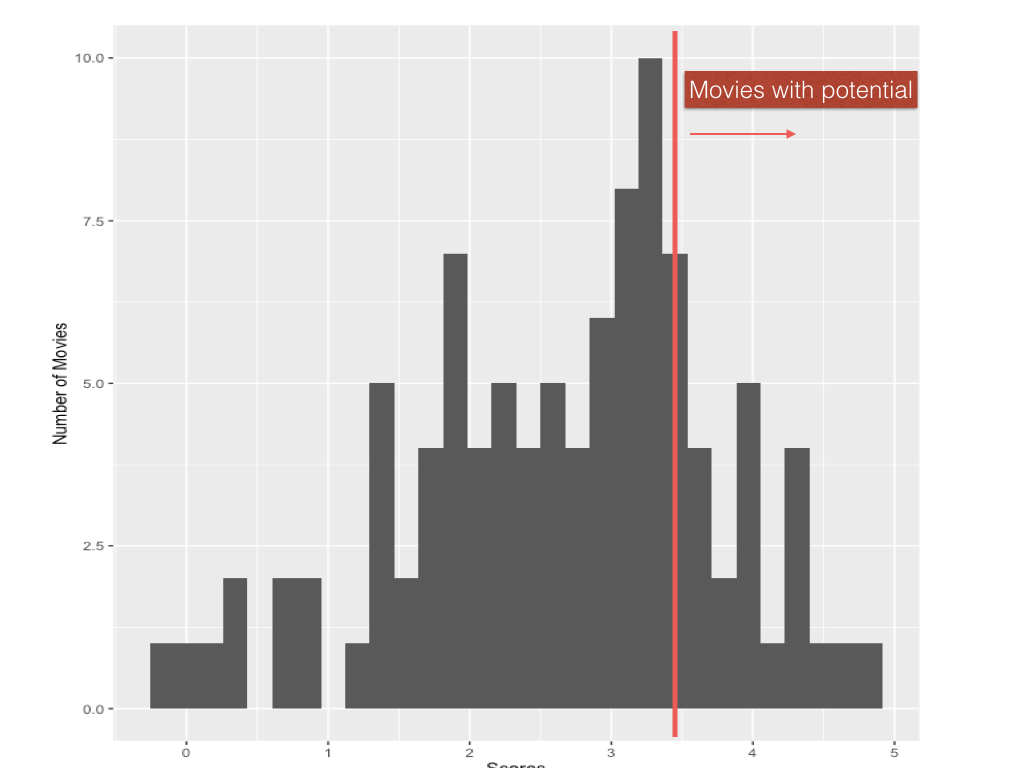
### *Validation data*

We will validate MovRank in two ways. First, we will compare our ranking to ranking from sites like The SSN or other evaluation services where scripts are scored by humans. Second, we will track how well our tool does over time by comparing predicted ratings against actual user ratings post production and release.

# Decision-Making And Organization Challenges

We have run a prototype of the simple model using our own database (Figure 2). Our model seems to be able to replicate “human ranking” with a 87% accuracy. We have also explored combining this with not-in-script data to see if we can rank how successful a movie will be in commercial terms. Our predicted values are better than random noise.

Figure 2. Proto-Type Results



We believe that MovRank is now ready for more systematic and larger breadth of data. Fortunately, we have a supportive CEO who sees the benefits of our tool. However, we face several challenges:

* There are groups that are **philosophically** opposed to this approach because it seems to fly in the face of the creative process. Our response is to emphasize that our tool is to help the creative process rather than replacing it.
* The current human evaluators who believe t**hey can do a better job and are concerned about their jobs**. Our response is to explain that the tool can free their energy from repetitive aspects of scripting selection and encourage them to focus on artistic details of screenplay review. We also want to explain that given the current technology trend, the automation of script selection is inevitable in the longer term.
* Senior management will face pressure from well-known producers and directors in adopting movie scripts. We explain that the benefit of the tool is to equip them with concrete data and evidence in those difficult decision making scenarios.

More broadly we do not underestimate how many in the industry, audience, and academia would feel threatened or concerned with this approach. This is broadly because of concern enumerated below.

# Ethical, Fallacy, and Bias

We see relatively few legal issues. However, our approach faces fallacy and bias concerns as well as ethical dilemmas.

First, MovRank is as **subjective as any human endeavor** but might give the impression of “objectivity”. It is based upon frameworks/models/data defined, designed, and built by a composite team who bring in vastly divergent expertise along with a strong subset dose of subjective interpretation in each domain. Past successful movies also ~~form the majority of baseline data used to create~~ influence the baseline data the most. ~~For example,~~ Also, script analysis assumes that large amounts of quantified information can reveal valuable insights. However, extractions are subjectively interpretable and could lead to **interpretation** bias. Quantitative interpretation of art also progressively gets **content** biased as many independent subjective attributes are quantified thus taking away the free form of what art is intended to depict. This **cognitive** bias, along with the quantification of successful movies, reaffirms confirmation bias because the eventual algorithm implicitly tends to favor information or evidence that confirms the theme and construct of just the successful movies.

Second, we could also be faced with difficult dilemmas if MovRank propagates current or historical biases (such as against ethnic groups) or, without biases, chooses to rank highly scripts that humans would find objectionable. As an example, ISIS is a heavily discussed topic in TV, online and social media and MovRank might favor a pro-ISIS theme. Nevertheless, we note that ours is just a tool not a binding decision rule so ultimately humans are still making the decision.

Third, MovRank could give **“false precision”**. A major fallacy of a quantitative algorithm is that it can be predictive to a measurable level of certainty. Our data might be biased or missing. Our tool must be augmented and refined through Bayesian study to evaluate not just the analytics data but also the conditions under which the movie was produced. Unless the baseline database is quite diverse and large, MovRank will not be statistically or practically significant.

Fourth, scriptwriters tend to express individuality and do not want to be compared against others. The ethics of **quantifying scriptwriters’** work where the quantification may not be an exact representation of the original work, and then subsequently ranked needs to be properly explained and justified.

Fifth, **fairness** is an issue to contend with as well. Since the original medium is no longer the basis of comparison, the new playing ground leads to new rules that must be ethically applied.

Finally, by simple virtue of economics, **commercial** considerations might dominate. For the studio, business might trump art. MovRank could also push creative talent towards commercial bias.

# Future

While MovRank is bootstrapped by analyzing a few chosen scripts and ranking them, the dictionaries and emotional indicators that are built with the training set can then be adaptively enhanced with the additional training using a wide spectrum of less known scripts. We will then add more data to see if scripts can be ranked by potential for success (in commercial or viewing terms).

As we gather more information and improve MovRank, we see three possible areas for further exploration. One, fine tuning our approach to a wide spectrum of less prominent scripts so we can target segments and subsegments of major themes. Two, using our tool upstream in the script writing process to “write” better scripts and put together better movies. This approach particularly goes beyond just word or cluster analysis of scripts but is extended to also measure plot direction/arch, dialogue quality and other more complex attributes. Third, we will expand our tool to combine script scores with other aspects of movie-making (casting, choice of director) to rank the likelihood of success more robust releases.

1. <http://www.inquiriesjournal.com/articles/172/the-significance-of-the-screenplay> [↑](#footnote-ref-2)
2. <http://www.screenwriterunknown.com/observations/odds-of-selling-a-spec-screenplay> [↑](#footnote-ref-3)
3. <http://mentalfloss.com/article/79197/8-hit-movies-were-originally-rejected-studios> [↑](#footnote-ref-4)
4. <http://www.filmmakerspot.com/download-movie-scripts/> [↑](#footnote-ref-5)