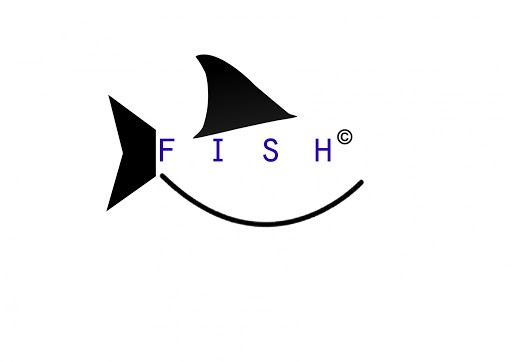
## horizontal line

Filming Industry Systematic Heuristic Classification (F.I.S.H)

10.16.2020

Zengxiaoran Kang

Yifei Liang

Joycelin Gu

Zichen Li

# Motivation:

Netflix is the dominant company in the streaming entertainment industry. It has over 190 subscribers, making it one of the world’s largest entertainment services. Its content varies from original shows, movies, and network TV. According to Investopedia's compelling analysis, as of 2019, an estimated 21.9 million American households will switch to streaming services such as Netflix, and this was expected to reach 34.9 million by 2023. The rating and genre of a movie/TV show usually indicates whether it is good or not and thus influences users’ preference. By understanding the composition of these two major factors, this project aims to uncover the popular trend of current entertainment streaming service.

# Target Audience:

The Netflix company itself and all of its users can benefit from our product. With the support of our models , the company can predict the rating when they try to acquire licensing for a TV show and movies. It also allows the users to make a prediction about a movie before it is released..

# Dataset

We are going to use the dataset “Netflix Movies and TV Shows” found on Kaggle. This dataset contains 6234 objects and 12 attributes as of January 2020. The attributes include show\_id, type (movie or TV show), title, director, cast, country, date\_added, release\_year, rating, duration, listed\_in, and description. In the first model, we will use rating as our label, and the rest of the attributes as the features. In the second model, type will be our label and others are the features. For our preliminary prediction, all attributes are useful. However, we might eliminate some of the attributes based on our analysis.

The dataset will be split into two parts. The first part will be used to train and develop the model, while the second part will be used to evaluate the precision/accuracy of the model about the movie ratings. After the set of predicted movie ratings are generated using the model, its accuracy can be determined by comparing the predicted ratings with actual ratings.

# Toolkit

We are planning on using the following libraries:

* Data processing
  + Numpy, Pandas - for data manipulation and analysis. In addition, we use it to filter out unqualified sample data.
  + Re - Extraction of patterned data
  + Ntlk - vectorize words in “Descriptions” attributes.
* Visualization
  + plotly, pygal - visually analyze the relationship between each feature and understand the correlation between the feature and label.
* Machine Model
  + Scikit-learn, keras - for classification, regression, and layering
  + Tensorflow - Neural networks construction and Deep learning operation.
* Evaluation
  + AUC ROC - performance measurement for classification problem at various thresholds.

# Preliminary Sketch

We will first perform Exploratory data analysis (EDA) to gain a better understanding of the data: identify popular categories, most frequently occurring words in description, the change of ratings of each category, and etc.. The team’s goal is to construct two comprehensive models. The first model is to predict the ratings of shows and perform detailed exploratory data analysis. The second model serves to classify the genre of the shows based on its summary.

We plan to use various methods to approach our goal, and compare their precision with AUC ROC. For example, linear regression would be a potential solution for the first model, where it will predict the rating of a show. Support vector machine would be an adequate model for the classification problem presented in the second model.

**Reference:**

Database: <https://www.kaggle.com/shivamb/netflix-shows>

Motivation: <https://www.investopedia.com/articles/investing/060815/how-netflix-changing-tv-industry.asp#:~:text=1%EF%BB%BF%20By%20creating%20compelling,the%20way%20they%20do%20business>.