LET'S GO TO THE MOVIES!

Machine Learning & IMDB 5000 Movie Dataset

Team Corn on da CAWB
Claire Broad, Andre Duarte, Will Young, Brigit Lawrence-Gomez
MSAN621
December 8, 2016

Overview

Show me the money!

Question:
Can we predict the **total gross revenue** that a movie will make?

Techniques:
Random Forest
Decision Tree
KNN Regression

Ask yourself one question: 'Do I feel lucky?'

Question:
Can we **predict the genre** of a movie using keywords or movie description?

Techniques:
Logistic Regression
LDA

...the beginning of a beautiful friendship.

Question:
Can we find movies that are **similar to each other,** beyond looking at genre?

Technique:
Principal Component Analysis

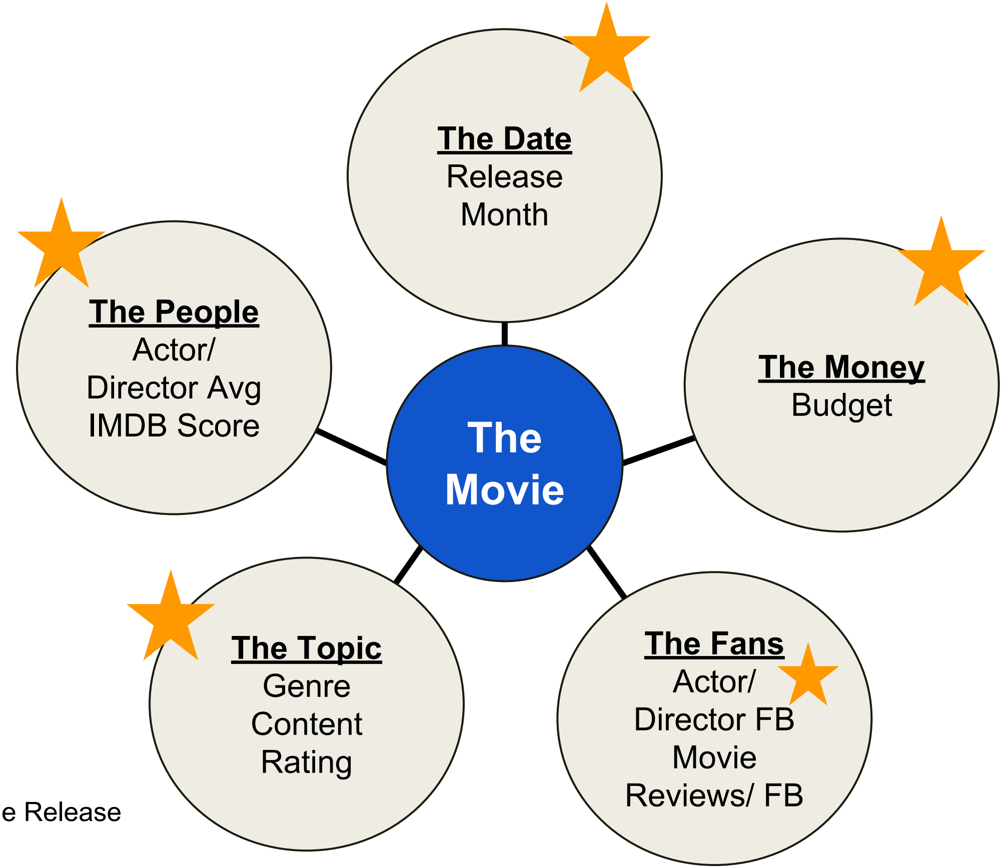
The Data

- **IMDB 5000 data (https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset)**
  - This is the top 5000 US movies of all time, ranked by production budget

- **We scraped IMBD API for additional features:**
  - Release date
  - Plot description
  - Awards
  - Type (Movie or TV Series)
  - Writer
  - Metascore (a weighted critic score)

- **We also generated a few of our own features to help predict revenue:**
  - Average IMDB Rating score for each actor and director

Revenue Prediction: Feature Extraction



The Movie

The Date
Release Month

The Money
Budget

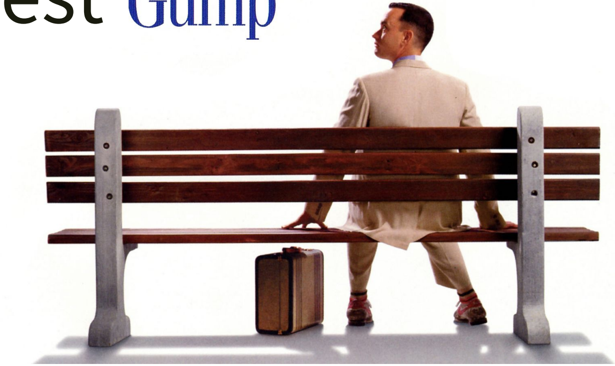
The Fans
Actor/ Director FB Movie Reviews/ FB

The Topic
Genre Content Rating

The People
Actor/ Director Avg IMDB Score

★ Actionable Before Movie Release

The Final Model: Random Forest Gump



	All Features	Pre-Release Features Only
Tree Depth	12	28
Test RMSE	44e6	46e6
R^2	0.66	0.56
Gross Revenue Std. Dev. (Test)	76e6	69e6

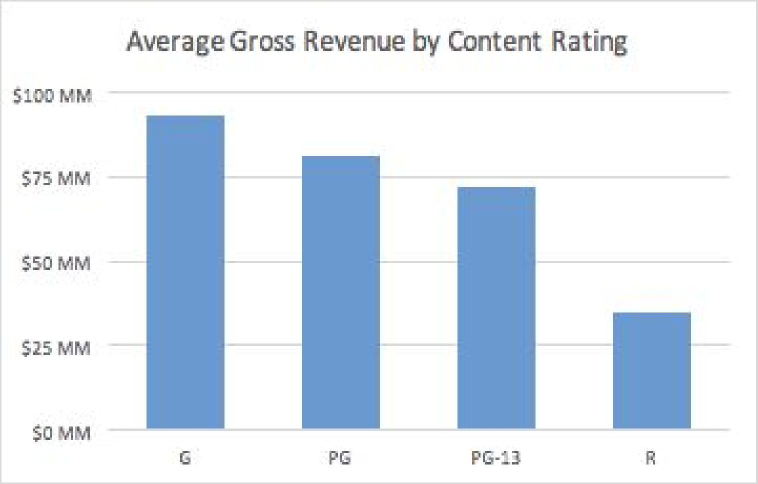
Rank	All Features Model Feature	Pre-Release Only Model Feature
1	# Users Voted (IMDB)	Budget
2	Budget	Avg Director IMDB Score
3	# Users Wrote Reviews	Avg Lead Actor IMDB Score
4	R Content Rating	Duration
5	Avg Lead Actor IMDB Score	Avg 3rd Actor IMDB Score
6	Total Cast Facebook Likes	Total Cast Facebook Likes
7	Duration	Year
8	Year	Director Facebook Likes
9	Avg 2nd Actor IMDB Score	Avg 2nd Actor IMDB Score
10	Avg 3rd Actor IMDB Score	Release Month

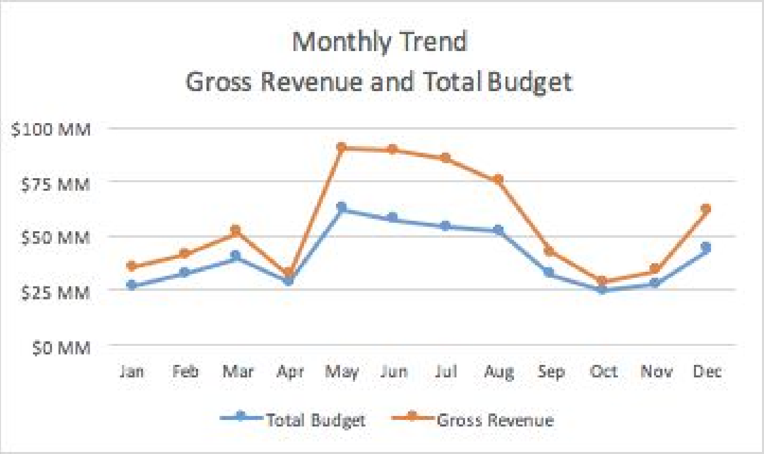
Additional Insights

- **Content Rating** correlated with Revenue
  - Incentive for studio executives to pad rating

- **Total Budget** and Gross Revenue highly correlated and reveal seasonal trend

- **Total Marketing Budget** could potentially improve prediction to identify movies “hyped” in advance





Genre Prediction

Given a 5 word description, can we guess the genre?

- Action
- Comedy
- Drama
- Romance
- Fantasy
- Horror
- Thriller
- True

Genre Prediction

Given a 5 word description, can we guess the genre?

VILLAIN

venom

spider man

Sandman

symbiote

Action
Comedy
Drama
Romance
Fantasy
Horror
Thriller
True

Genre Prediction

Given a 5 word description, can we guess the genre?

Correct: Action

Movie: Spiderman

Action

Comedy

Drama

Romance

Fantasy

Horror

Thriller

True

Genre Prediction

- Bucket genres into 8 categories
- Use top keywords as features
- Get accuracy score/ misclassification rate from several ML algorithms

Genre Prediction

- Bucket genres into 8 categories
- Use top keywords as features
- Get accuracy score/ misclassification rate from several ML algorithms

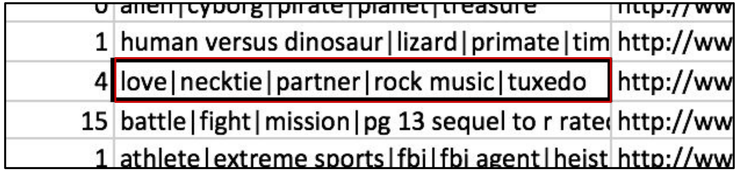
Problem: ‘Top 5’ keywords are not consistently good descriptors of the plot or tone

Genre Prediction

- Bucket genres into 8 categories

- Use top keywords as features

- Get accuracy score/ misclassification rate from several ML algorithms



0	alien|cyborg|pirate|planet|treasure	http://ww
1	human versus dinosaur|lizard|primate|tim	http://ww
4	love|necktie|partner|rock music|tuxedo	http://ww
15	battle|fight|mission|pg 13 sequel to r rate	http://ww
1	athlete|extreme sports|fbi|fbi agent|heist	http://ww

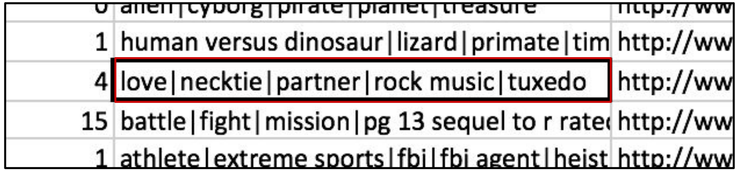
Problem: ‘Top 5’ keywords are not consistently good descriptors of the plot or tone

Genre Prediction

- Bucket genres into 8 categories

- Use top keywords as features

- Get accuracy score/ misclassification rate from several ML algorithms

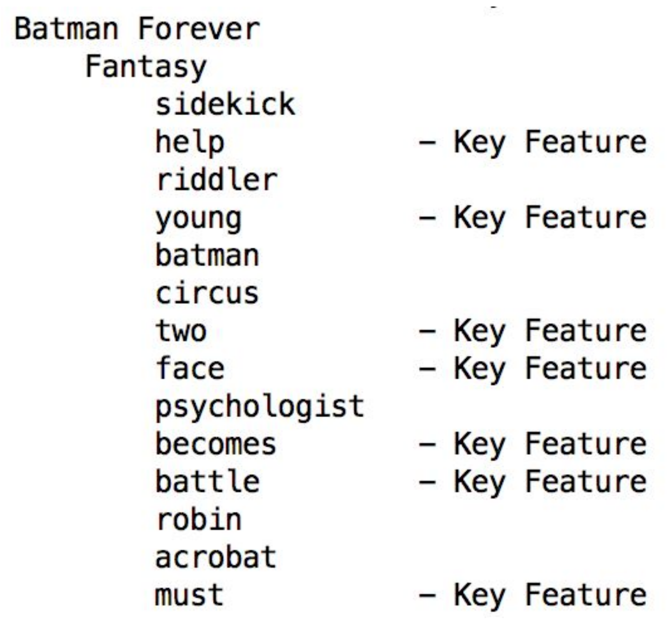


0	alien|cyborg|pirate|planet|treasure	http://ww
1	human versus dinosaur|lizard|primate|tim	http://ww
4	love|necktie|partner|rock music|tuxedo	http://ww
15	battle|fight|mission|pg 13 sequel to r rate	http://ww
1	athlete|extreme sports|fbi|fbi agent|heist	http://ww

Problem: ‘Top 5’ keywords are not consistently good descriptors of the plot or tone

Genre Prediction

- Possible solution: Tokenized Plot Descriptions
- Pull top n word - genre pairs (n = 1000)
- Eliminate words with >4 genres
- 383 keywords remaining → features



```
Batman Forever
    Fantasy
        sidekick
        help                - Key Feature
        riddler
        young               - Key Feature
        batman
        circus
        two                 - Key Feature
        face                - Key Feature
        psychologist
        becomes             - Key Feature
        battle              - Key Feature
        robin
        acrobat
        must                - Key Feature
```

Genre Prediction - Results

- Bucket accuracy
  - Keyword version: 40% (Logistic Regression)
  - Plot description version: 40% (LDA)

- Obscured accuracy
  - Keyword version: 20%
  - Plot description version: 17%

“Bad” Predictions

- Young Frankenstein

  Keywords: assistant, castle, experiment, frankenstein's monster, scientist
  Predicted: Horror
  Actual: Comedy

- As Good as It Gets

  Keywords: dog, friendship, neighbor, unlikely friendship, writer
  Predicted: Comedy
  Actual: Drama

- Batman Forever

  Keywords: love, necktie, partner, rock music, tuxedo
  Predicted: Drama
  Actual: Fantasy

Good Predictions

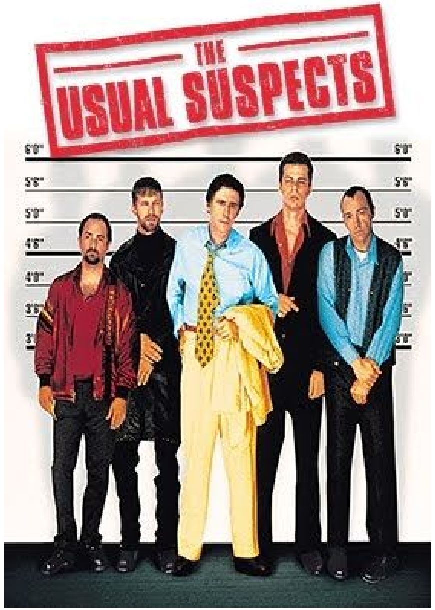
- Pokémon 3: The Movie
  Keywords: ash, father, mother, pokemon, professor
  Bucket: Fantasy

- Friday the 13th: A New Beginning
  Keywords: jason voorhees, murder, new jersey, nightmare, teenager
  Bucket: Horror

- Transporter 2
  Keywords: driver, french, kidnapping, police, sequel
  Bucket: Thriller

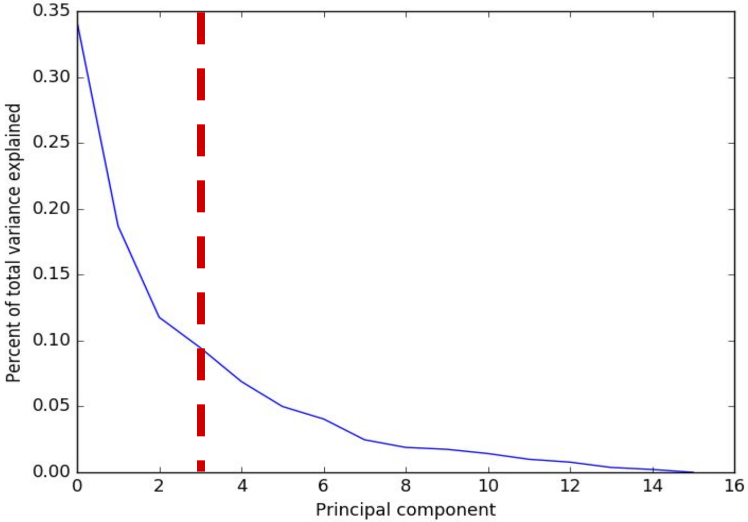
PCA Clustering

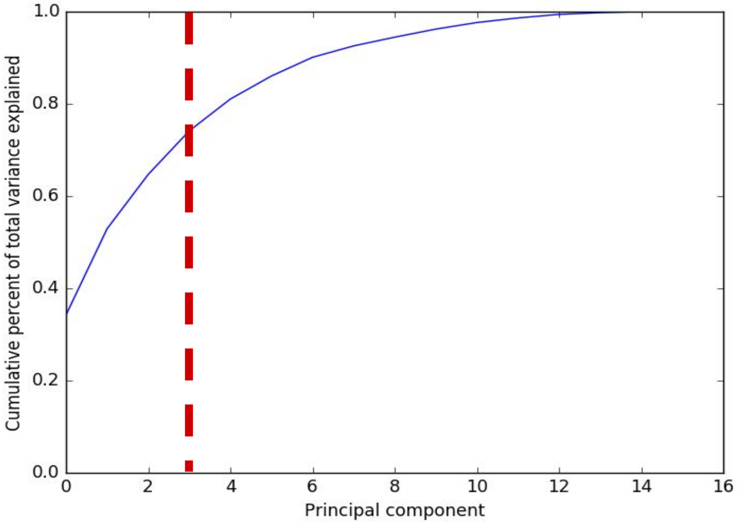
- What are the most defining characteristics of movies in the dataset?
- What do the principal components mean?
- *Round up the **Usual Suspects**...*



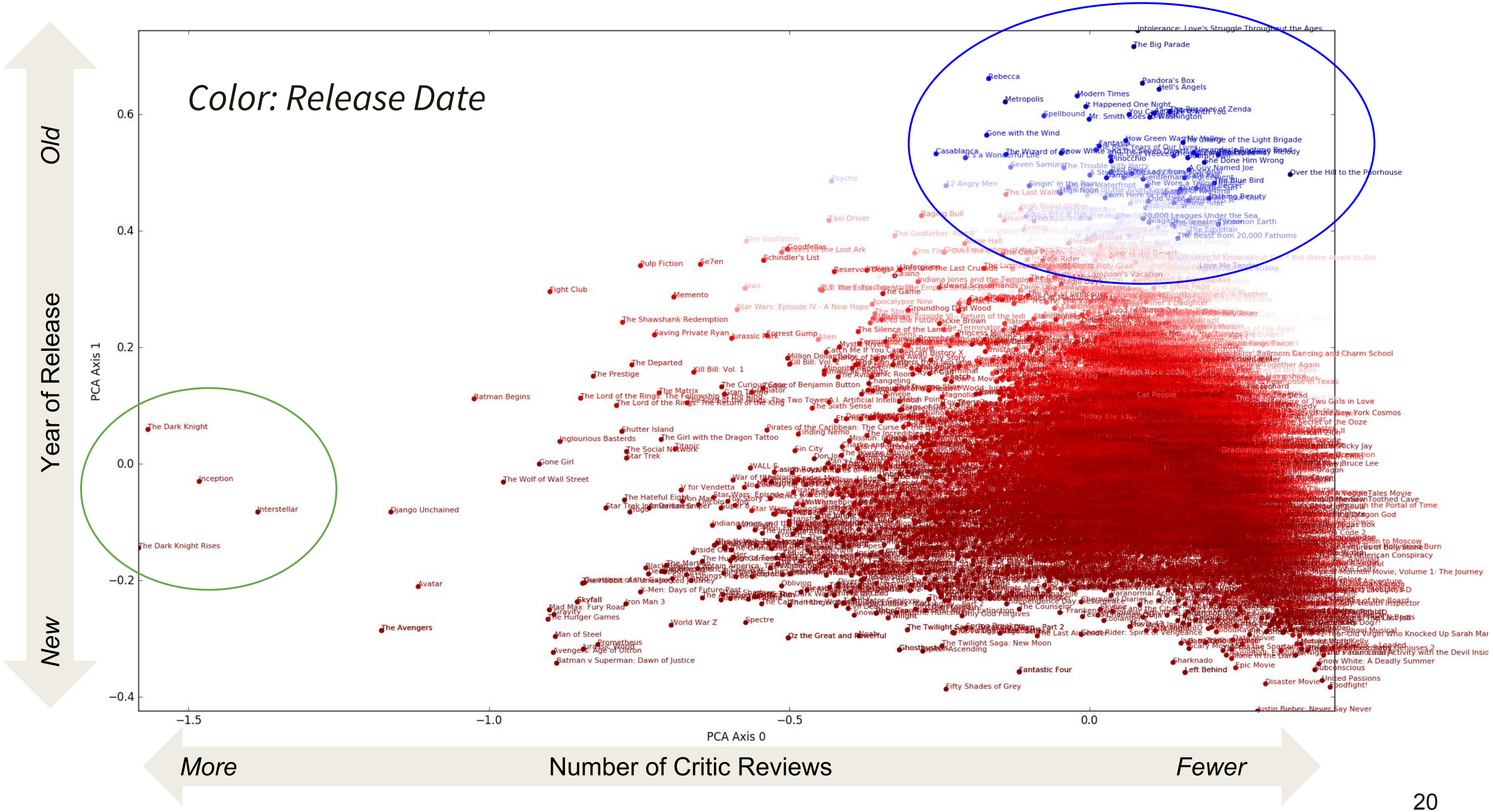
PCA Clustering

- **65%** of the Total Variance explained with **3** principal components


Percent of Total Variance
Explained by Principal Components


Cumulative Percent of Total Variance
Explained by Principal Components

PCA Clustering



Old

Year of Release

New

More

Number of Critic Reviews

Fewer

That's all Folks!

BLOOPERS

Revenue Prediction: Data Challenges

Data is like a box of chocolates. You never know what you're gonna get…

Data Challenges:

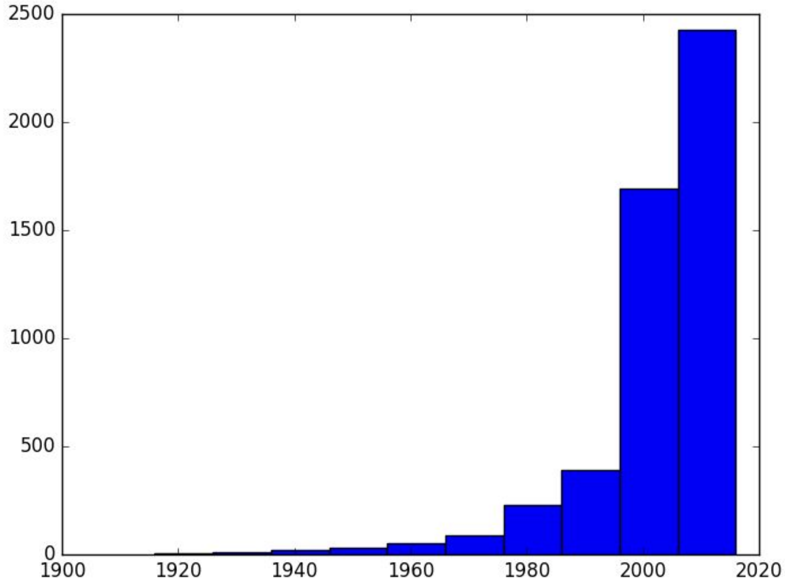
- Missing revenue data ~20% of the movies
- Movies combined with TV
- International movies had revenue in foreign currencies

Solution:

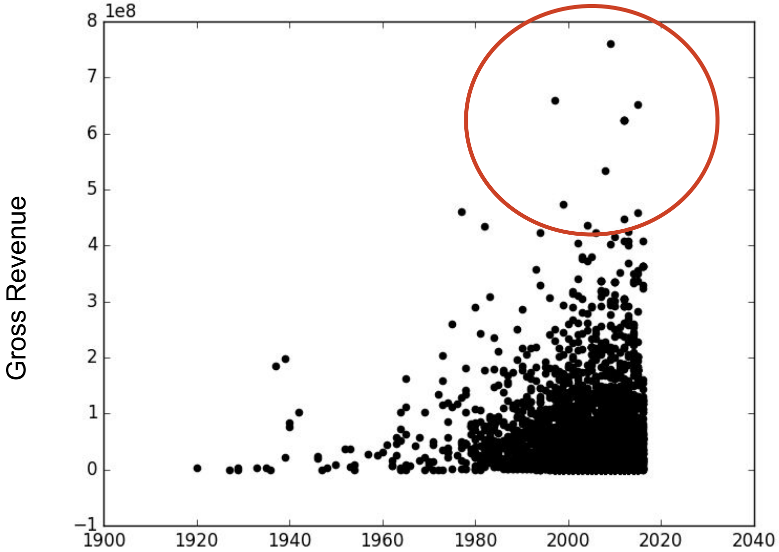
- Take them out

The Data

Movies in dataset skew more recent



“Blockbusters” make revenue hard to predict


1e8
Gross Revenue