# The Formula for a Blockbuster

## Can movie producers predict their next hit?

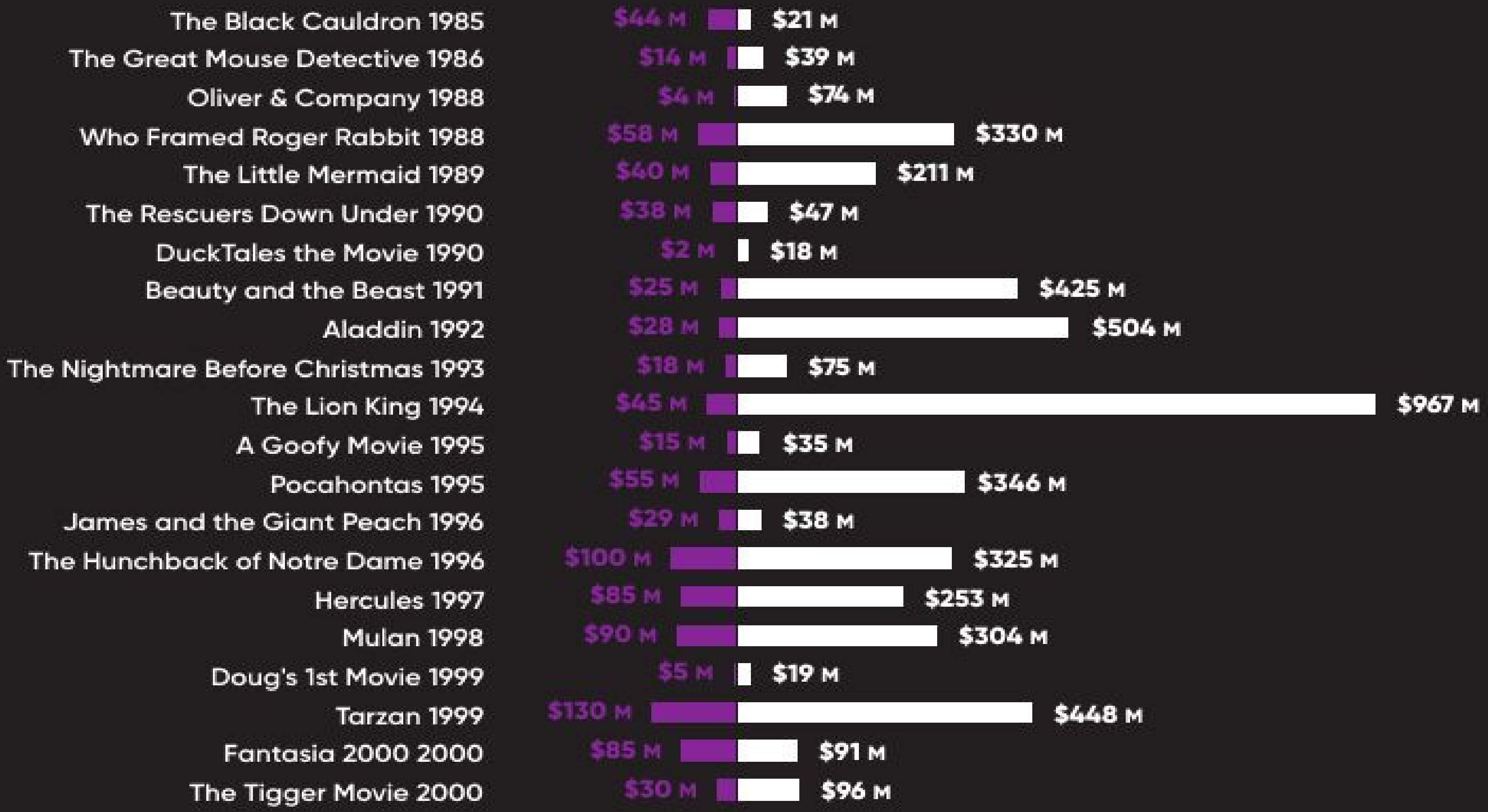
Alex Chung
Nathaniel Schub
Shan He

School of Information, UC Berkeley

### Context

## Movie producers have mixed success when predicting the outcome of their investments.

Producers typically rely on focus groups.


The Black Cauldron 1985 $44 M $21 M
The Great Mouse Detective 1986 $14 M $39 M
Oliver & Company 1988 $4 M $74 M
Who Framed Roger Rabbit 1988 $58 M $330 M
The Little Mermaid 1989 $40 M $211 M
The Rescuers Down Under 1990 $38 M $47 M
DuckTales the Movie 1990 $2 M $18 M
Beauty and the Beast 1991 $25 M $425 M
Aladdin 1992 $28 M $504 M
The Nightmare Before Christmas 1993 $18 M $75 M
The Lion King 1994 $45 M $967 M
A Goofy Movie 1995 $15 M $35 M
Pocahontas 1995 $55 M $346 M
James and the Giant Peach 1996 $29 M $38 M
The Hunchback of Notre Dame 1996 $100 M $325 M
Hercules 1997 $85 M $253 M
Mulan 1998 $90 M $304 M
Doug's 1st Movie 1999 $5 M $19 M
Tarzan 1999 $130 M $448 M
Fantasia 2000 2000 $85 M $91 M
The Tigger Movie 2000 $30 M $96 M

### The emergence of big data

## Netflix analyzed viewing habits of their 33 million subscribers to predict House of Cards would be a hit.

# Goal

- **Mine public data on movies to see if certain variables predict box-office success:**

**Does it matter which actors appear in your film?**

**What influence does release date have?**

**During opening week, does more conversation on Twitter improve sales?**

**Are bigger budgets associated with bigger sales?**

**....**

# Challenges

- **Data Quality**
  Majority of the movie metadata we use will come from a community built database, in which the data consistency and accuracy can be an issue

- **Data Accessibility**
  One of the planned analyses involves web-scraping for movie-relevant tweets. Accessibility of such data has legal and financial limitations

- **Number of Relevant Factors**
  There is certainly not a single factor that drives the box-office success for a movie. But our analysis will aim to use the breadth of data and depth of analyses to uncover the significant factors

# Data Acquisition

**TMDb API**

TMDb is a community built movie and TV database, which contains a wide variety of metadata on movies dating back to 2008.

https://www.themoviedb.org/documentation/api

**MovieLens | GroupLens**

Rating datasets from the MovieLens web site, which feature a MovieLens 20M dataset (20M ratings on 27,000 movies by 138,000 users) and a MovieLens Full dataset (26M ratings on 45,000 movies by 270,000 users)

https://grouplens.org/datasets/movielens/

**Rotten Tomatoes API**

Rotten tomatoes is a review aggregation website. Their “Tomatometer rating” - based on the published opinions of hundreds of film and television critics - is a trusted measurement of movie quality for millions of moviegoers.

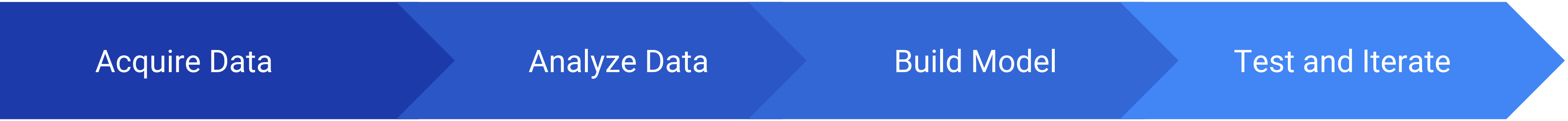
https://developer.fandango.com/Rotten_Tomatoes

**Twitter Search API**

The Twitter Search API is part of Twitter’s REST API. It allows queries against the indices of recent or popular Tweets

https://developer.twitter.com/en/docs/tweets/search/overview/basic-search

# Project Plan



## Acquire Data

- Download data from sources
- Twitter search API
- Consolidate data
- Build ER Diagram to direct analysis

## Analyze Data

- High level exploratory data analysis and cleansing to ensure data quality (particularly for data acquired through web-scraping)
- Design and impose schema as necessary

## Build Model

- Build hypothetical model incorporating all datasets to identify indicators of box office performance

## Test and Iterate

- Refine model as needed
- Consider adding or removing data
- Review takeaways, pressure test hypotheses

# Thank you