

# Movie Recommendation System

Heesuk Jeong  
(500350619)

# Recommender System

- Given an item, predicts preference of a user

The screenshot displays the IMDb homepage with a dark theme. At the top is a search bar with the text "Find Movies, TV shows, Celebrities and more..." and a dropdown menu set to "All". Navigation links include "IMDb", "Movies, TV & Showtimes", "Celebs, Events & Photos", "News & Community", "Watchlist", and a "Sign in with Facebook" button.

The main content area is divided into several sections:

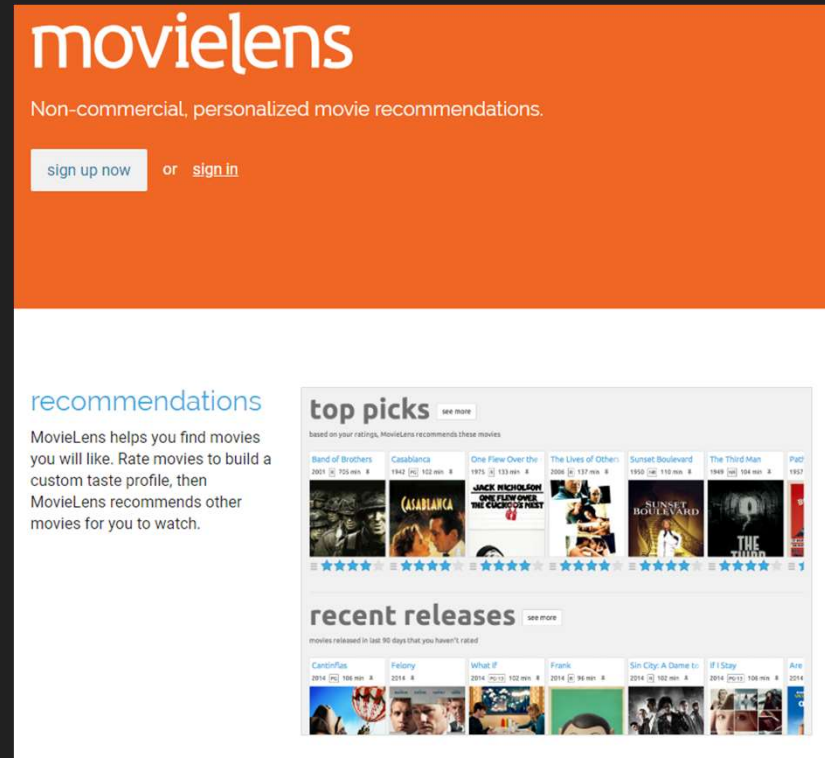
- top picks**: A section titled "based on your ratings, MovieLens recommends these movies". It features a grid of movie posters including "Band of Brothers", "Casablanca", "One Flew Over the Cuckoo's Nest", "The Lives of Others", "Sunset Boulevard", and "The Third Man". Each poster includes a star rating bar.
- recent releases**: A section titled "movies released in last 90 days that you haven't rated". It shows a grid of newer movie posters such as "Cantinflas", "Felony", "What If", "Frank", "Sin City: A Dame to Kill For", and "If I Stay".
- Trailers**: A section on the right side of the main content area featuring three large movie trailers with play buttons. The trailers are for "Goat" (King of the Goats), "Ralph" (Ralph Breaks the Internet), and "Avengers: Endgame".
- Opening This Week**: A vertical list on the far right showing a stack of movie posters with plus signs, including "Spider-Man: Into the Spider-Verse", "Deadpool", "The Mule", "Mortal Engine", "If Beale Street Could Talk", "The Wedding Guest", and "Capernaum". A link "See more opening movies" is at the bottom.

# Movielens – 1M Dataset

- 1 million ratings  
from 6,000 users on 4,000 movies.

[ml-1m.zip](#)

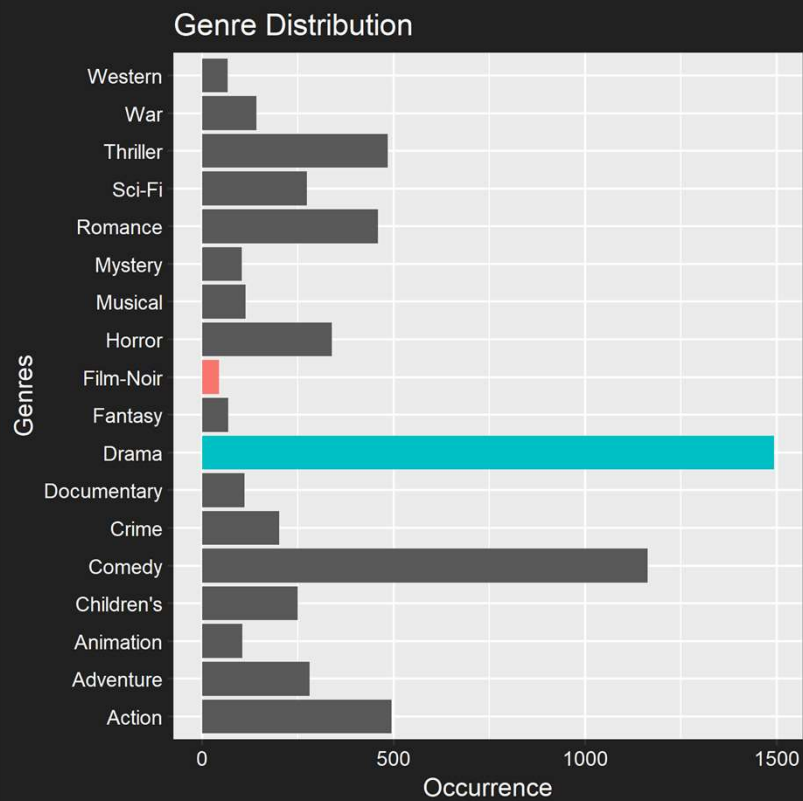
- movies.dat
- ratings.dat



# movies.dat

3,883 observations, 3 attributes

- movieId
- title
- genres

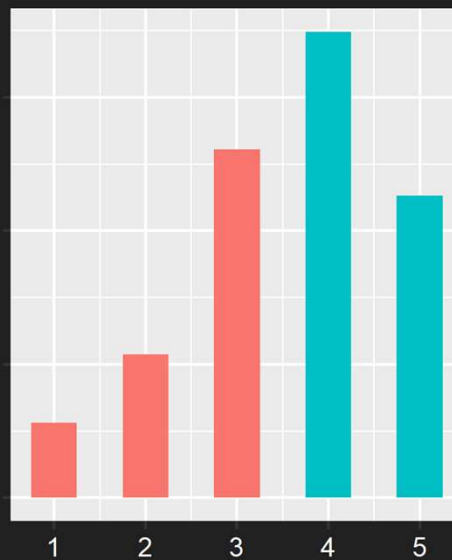


# ratings.dat

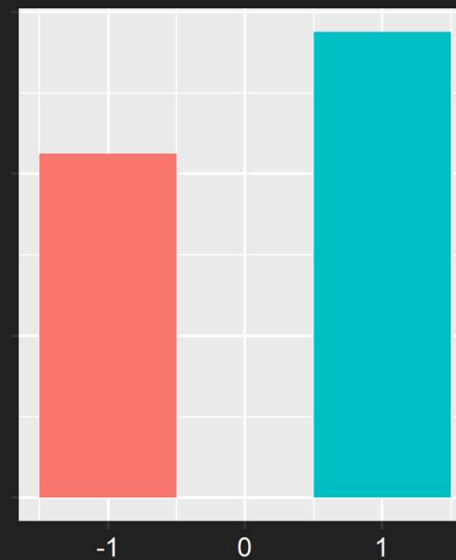
1,000,209 observations, 3 attributes

- userId
- movieId
- ratings

Distribution of Ratings

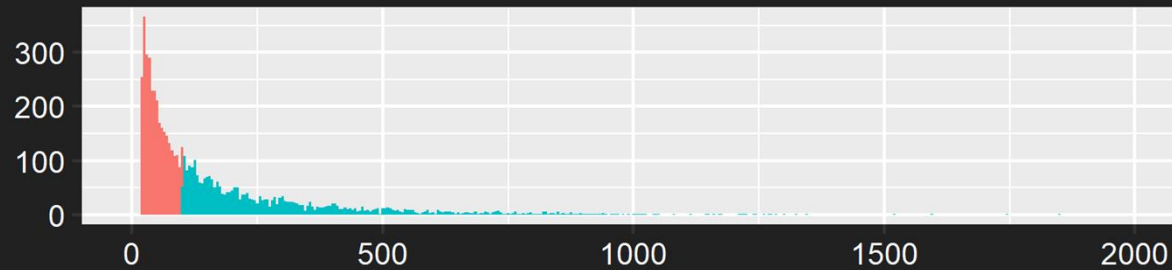


Binary Representation

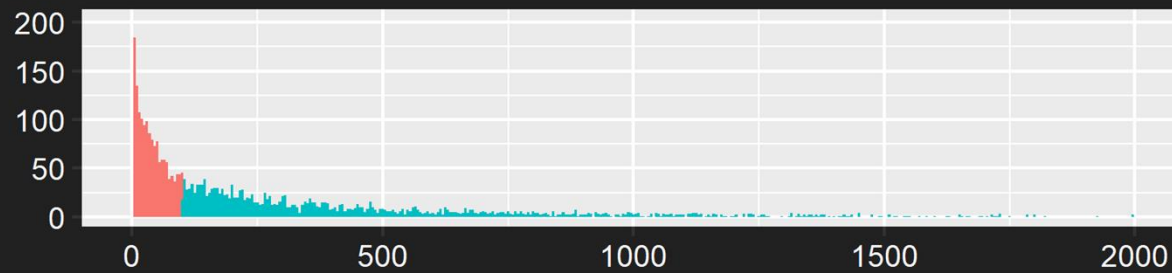


# Issue: Sparsity of Ratings

Number of Movies Rated by each User

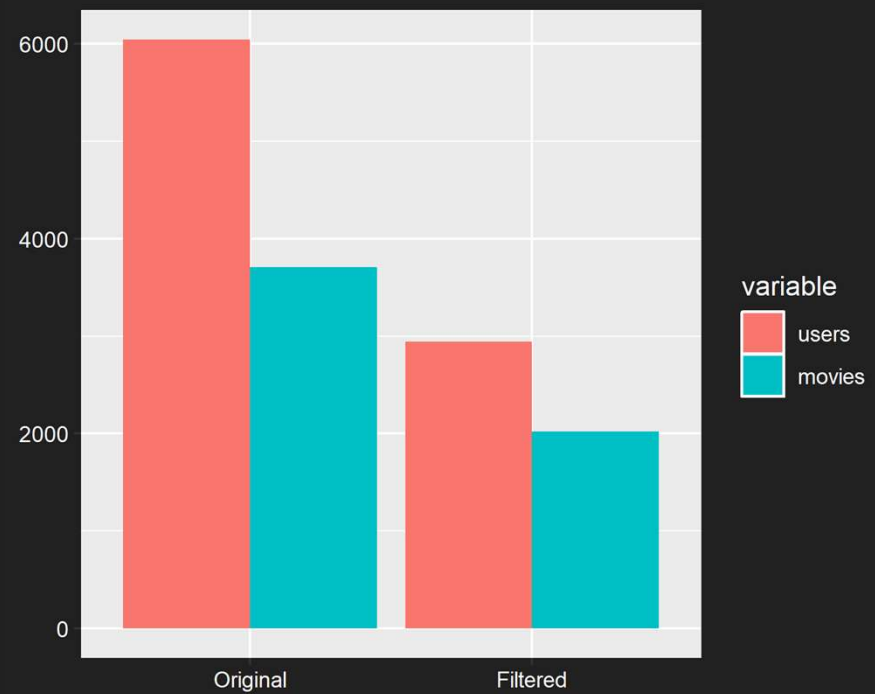


Number of Times each Movie is Rated



# Secondary Filtered Dataset

- Original Dataset
  - 1,000,209 Observations
  - 4.47% Dense
- Filtered Dataset ( $n \geq 100$ )
  - 795,382 Observations
  - 13.38% Dense



# Validation

- 60% Train Set
- 40% Test Set

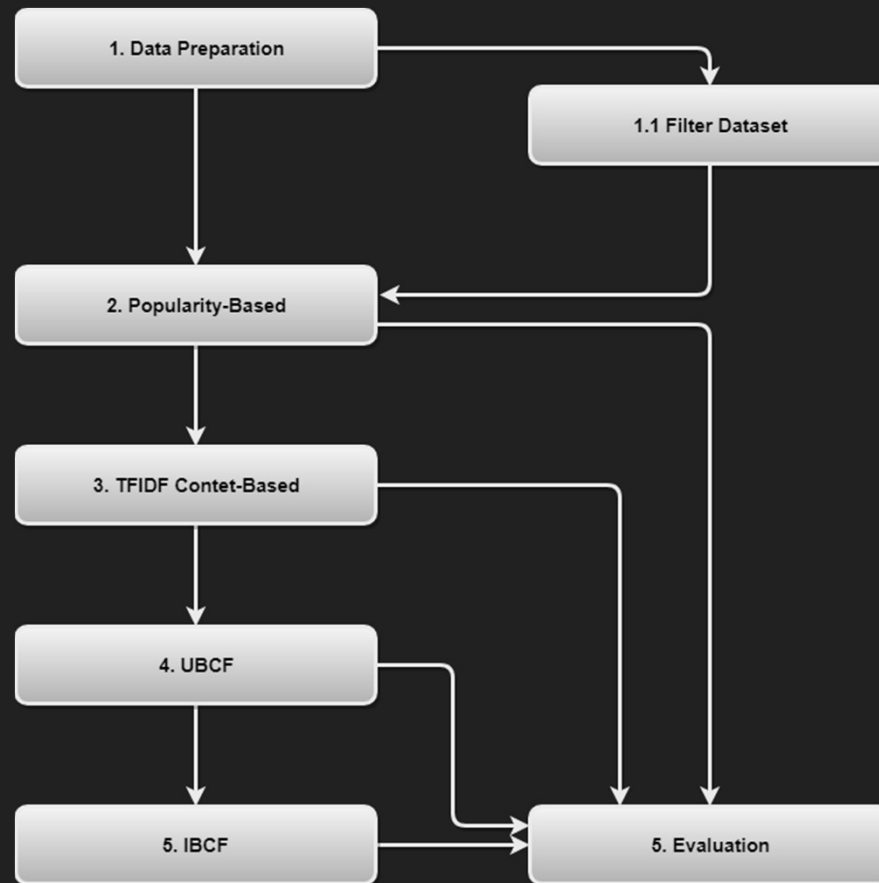
users	movies						m
	1	2	3	4	5	...	
1	1	-1	0	0	1		0
2	-1	0	-1	1	0	...	0
3	0	0	1	1	0		0
4	0	0	1	-1	0		1
5	0	1	0	0	1		0
	...						...
n	1	0	0	0	1	...	-1



users	movies						m
	1	2	3	4	5	...	
1	1	-1	0	0	1		0
2	?	0	-1	1	0	...	0
3	0	0	1	1	0		0
4	0	0	?	-1	0		1
5	0	?	0	0	1		0
	...						...
n	1	0	0	0	?	...	-1

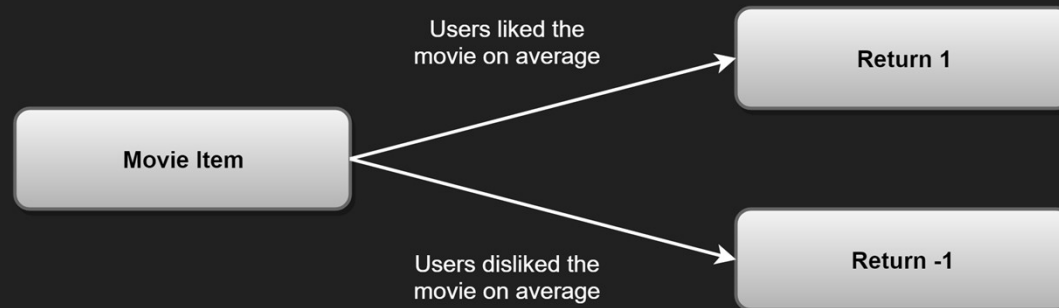


# Approach



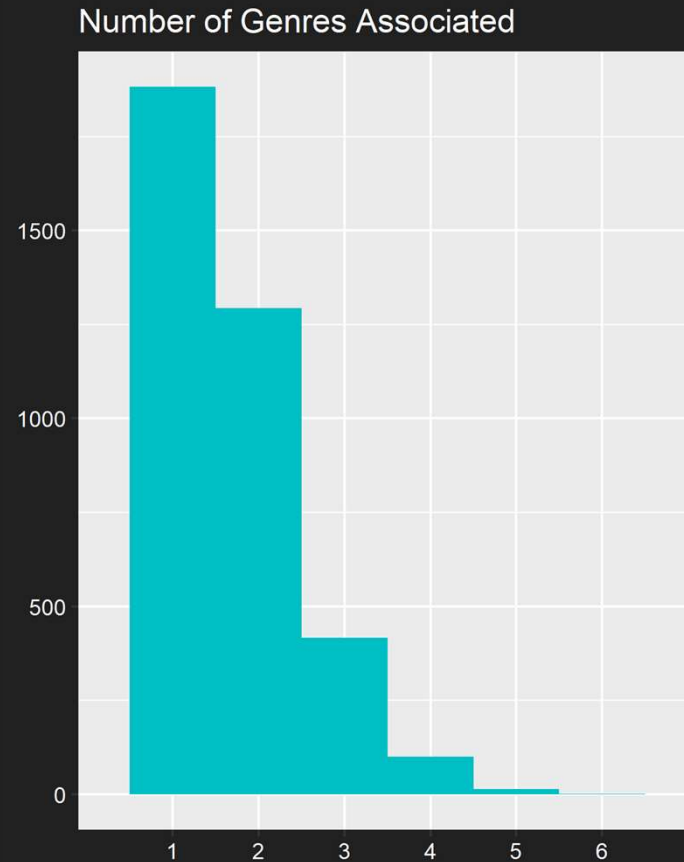
# Popularity-Based

- Average rating given a movie
- Used as a benchmark model



# Genre-Based Filtering

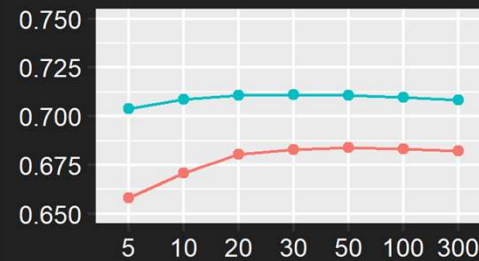
- Each movie has 1 to 6 genres
- Genres distributed unevenly
- TF-IDF values of genres used to create user-profile



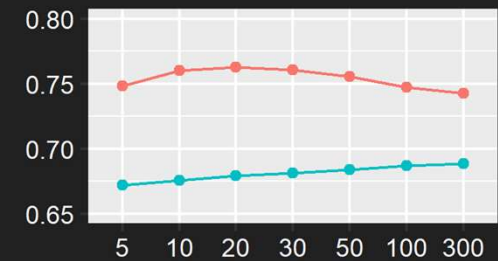
# Collaborative Filtering

- Cosine Similarity
- Number of Neighbors
  - $K = 30$

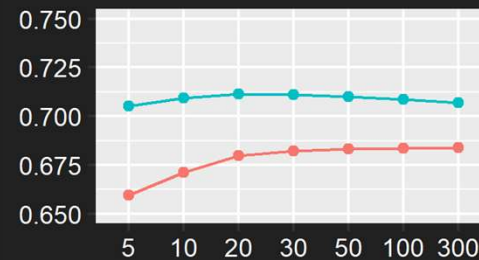
UBCF - Original



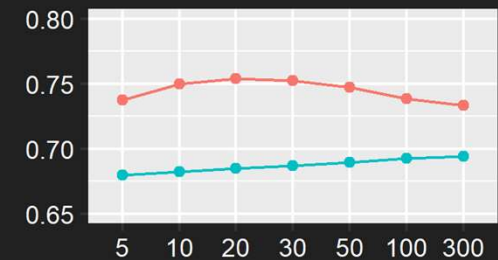
IBCF - Original



UBCF - Filtered

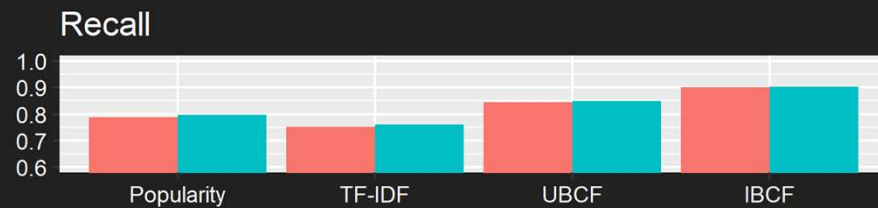
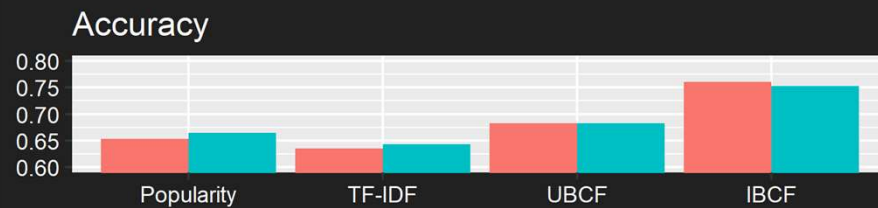
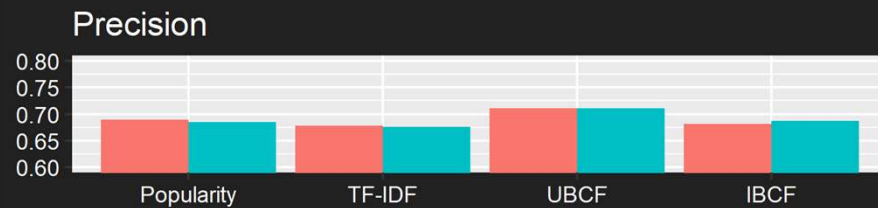


IBCF - Filtered



variable  accuracy  precision

# Effect of Filtering Dataset was Not Significant



Original Filtered

↑ Higher Density  
↓ Smaller Dataset

# Evaluation

- Primary Metric
  - Precision
- Secondary Metric
  - Accuracy
  - Recall

<i>Original Dataset</i>	Precision	Accuracy	Recall
Popularity-based	68.94%	65.39%	78.77%
TFIDF	67.78%	63.48%	75.12%
UBCF (k = 30)	71.11%	68.28%	84.44%
IBCF (k = 30)	68.10%	76.08%	90.11%

<i>Filtered Dataset</i>	Precision	Accuracy	Recall
Popularity-based	68.78%	66.03%	79.44%
TFIDF	67.79%	63.75%	75.62%
UBCF (k = 30)	71.28%	67.84%	84.46%
IBCF (k = 30)	68.90%	75.01%	90.26%

# Importance of Recommender Systems

- Amount of data available online is always increasing
- Broad scope of application

## Pandora helps users discover new podcasts on the platform

TECH NEWS

Tuesday, 11 Dec 2018  
11:00 AM MYT



Pandora announced that the platform now has hundreds of podcasts across a wide range of genres bringing users over 100,000 episodes to listen to. — AFP

After working on the Podcast Genome Project for the last year, Pandora is officially introducing podcasts and highly personalised podcast recommendations to the music streaming service.

Pandora announced that the platform now has hundreds of podcasts across a wide

November 2, 2018

## Movie Recommendations with Spark Collaborative Filtering

Rosaria Silipo



Collaborative filtering (CF)[1] based on the alternating least squares (ALS) technique[2] is another algorithm used to generate recommendations. It produces automatic predictions (filtering) about the interests of a user by collecting preferences from many other users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A

has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than a randomly chosen person. This algorithm gained a lot of traction in the data science community after it was used by the team winner of the [Netflix Prize](#).