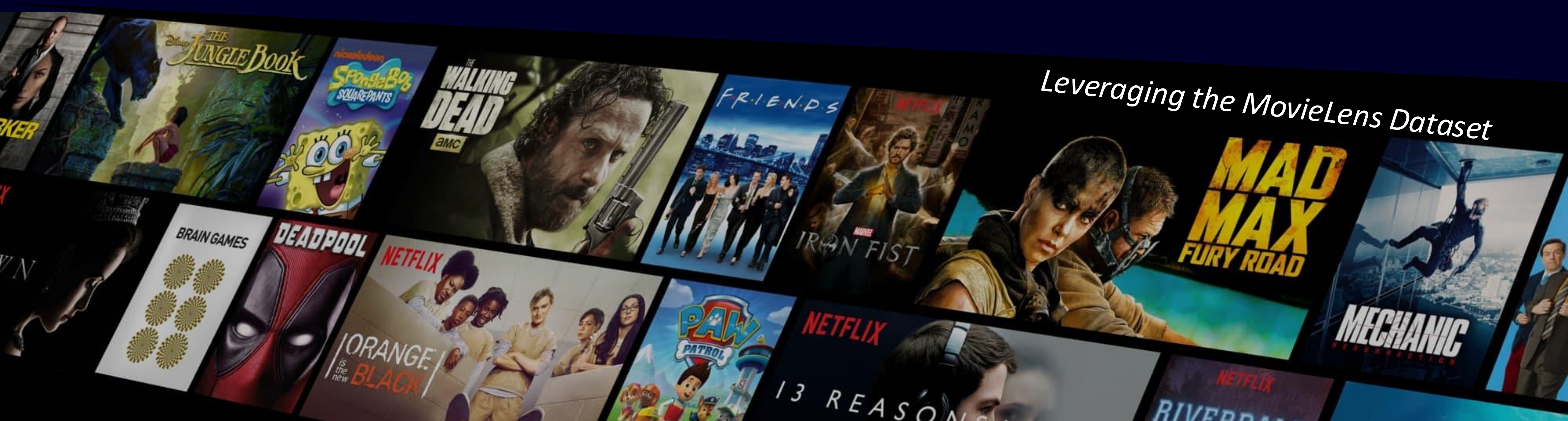
Project By

Tim Benjamin Hoffmann

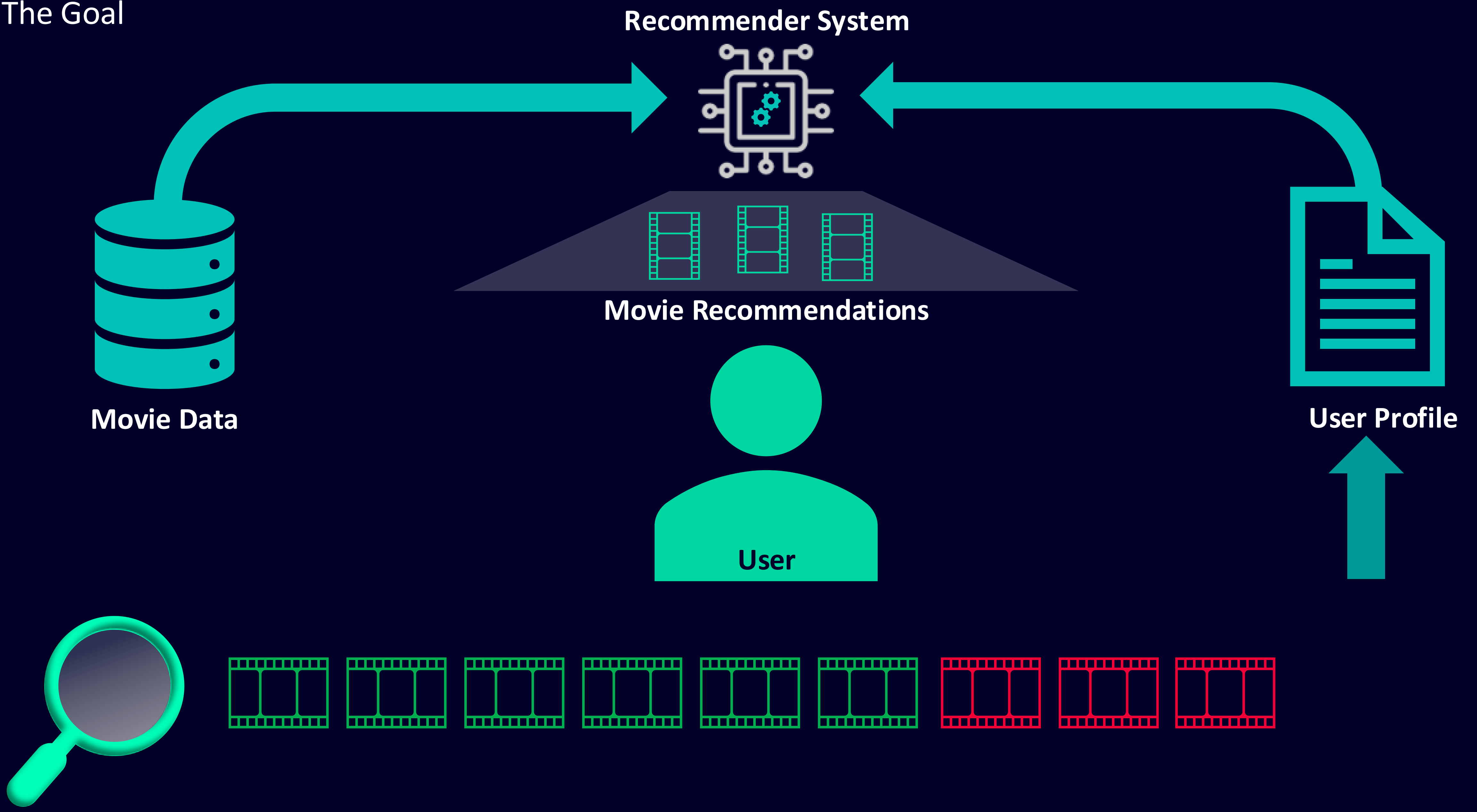
MŰEGYETEM 1782

Movie Recommender System



Leveraging the MovieLens Dataset

Motivation
The Goal



Data

The MovieLens Dataset: 3 essential files

ratings.csv

- Contains 100,836 ratings
- Attributes: userId, movieId, rating, timestamp
- Each row corresponds to a single rating
- Ratings from 0.5 to 5

tags.csv

- Contains 3,683 tags
  (Short descriptive phrase)
- Attributes: userId, movieId, tag, timestamp
- Each row corresponds to a single tag given by a user to a particular movie

movies.csv

- Contains 9,742 movies
- Attributes: movieId, title, genres
- Each row corresponds to a single movie within the dataset
- A movie can have multiple genres associated with it

Modeling
Two Models

User-based Collaborative Filtering

- Computes similarities between users to predict ratings
- Generally provides a diverse set of recommendations
- Suffers from data sparsity and scalability issues

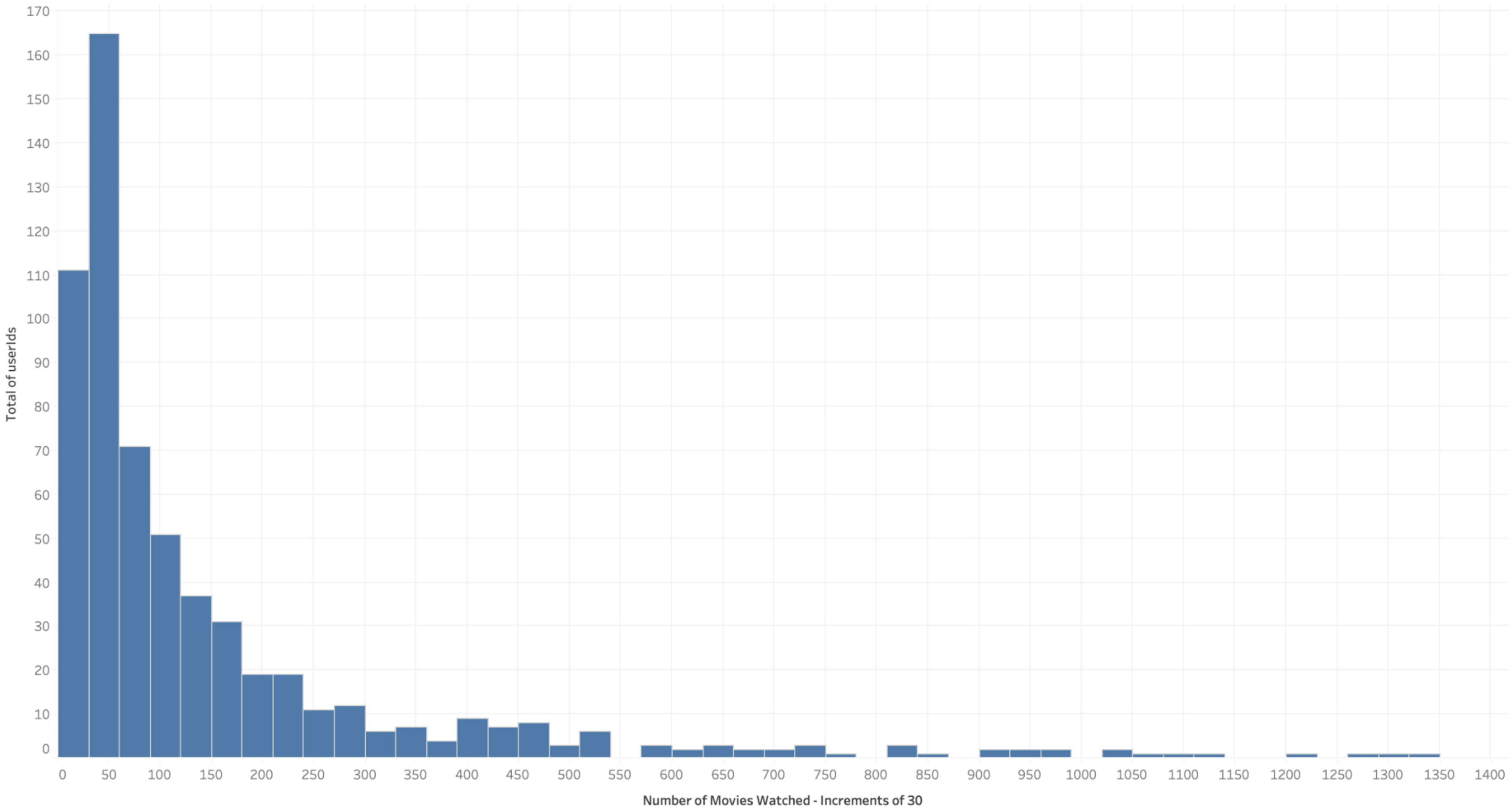
Content-based Filtering

- Recommends items based only on attributes a user likes
- Works well for new items
- Tends to recommend a narrow range of items

Evaluation mainly through Precision metric

Results
Exploratory Analysis

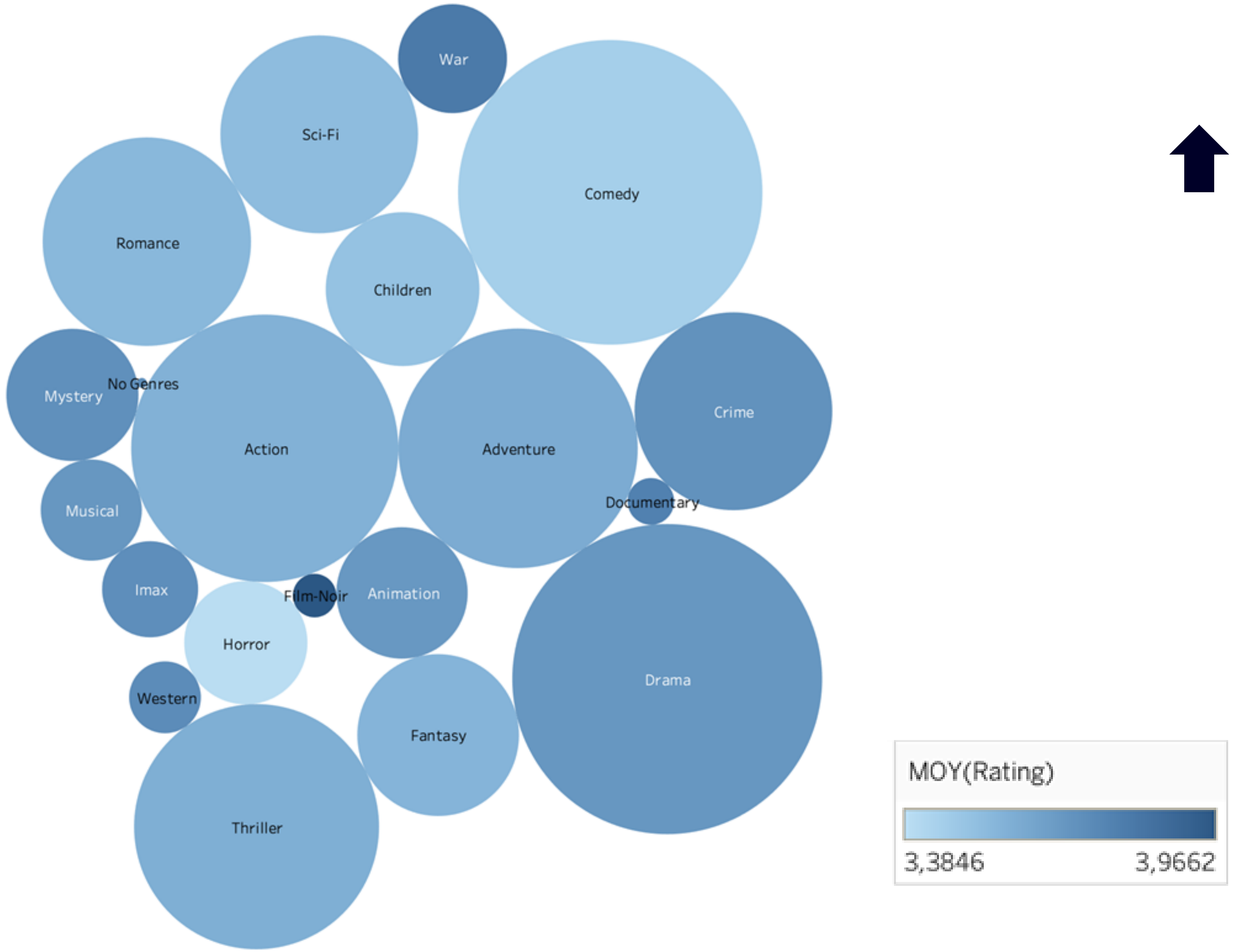
Number of users based on total number of ratings



70 Movies rated on average (Median)

Results
Exploratory Analysis

Number (size) and quality (color) of ratings per genre

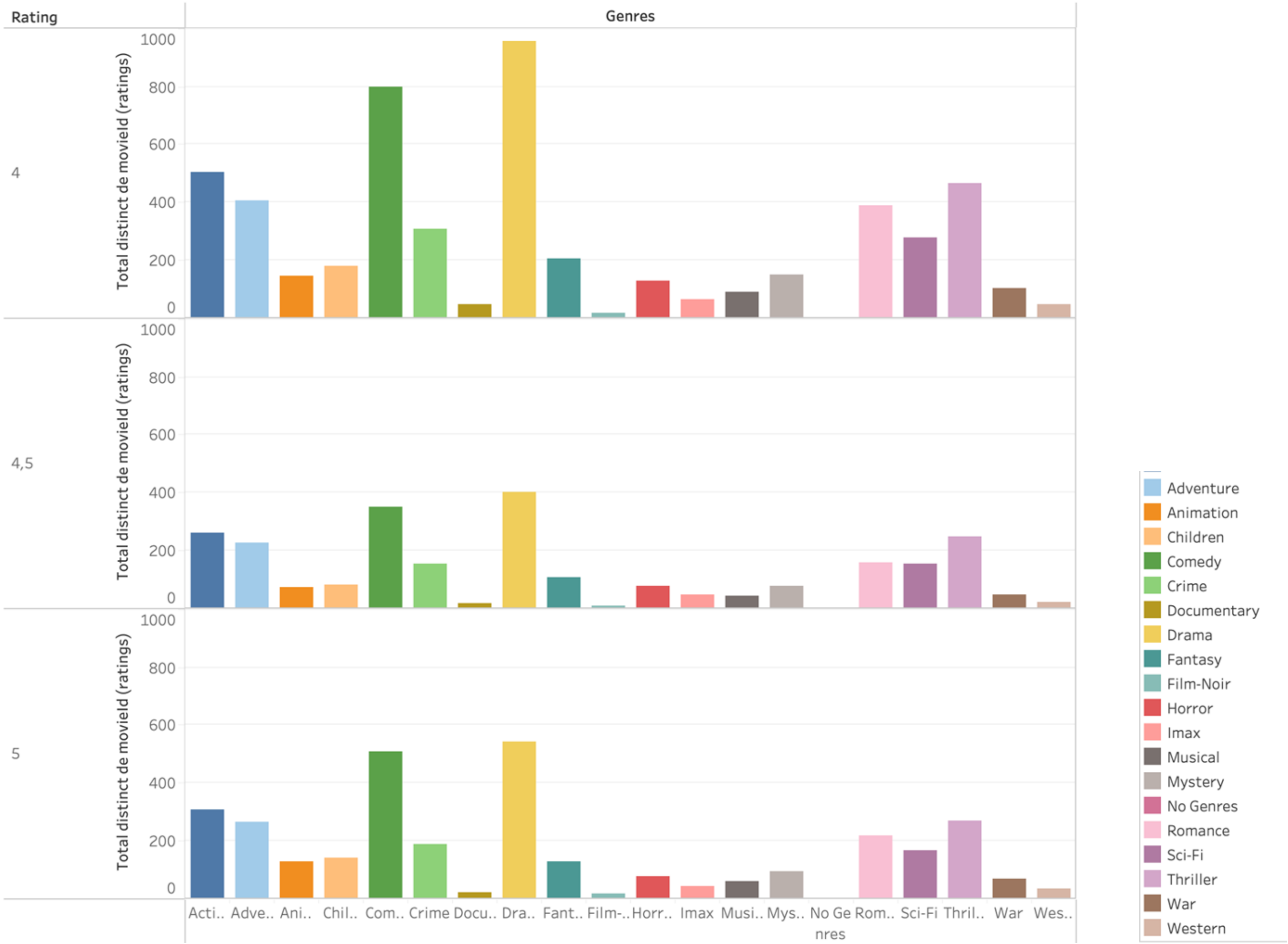


⬆ **Number of Ratings**

⬇ **Rating**

Results
Exploratory Analysis

Number of movies with best ratings for each genre

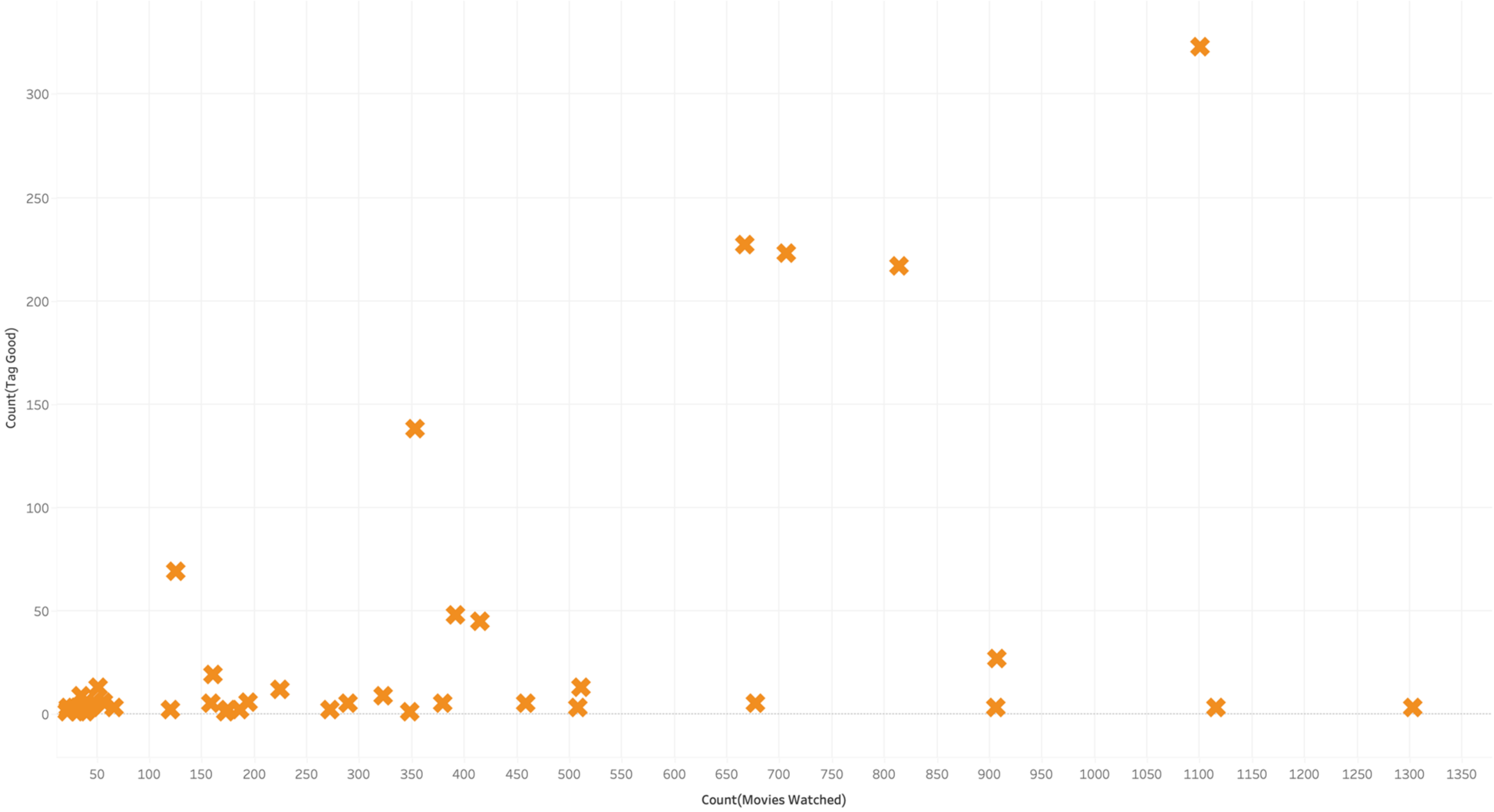


Relative Dominance of frequently rated genres

True quality could be dominated by prevalence

Results
Exploratory Analysis

Number good tags given per movies watched

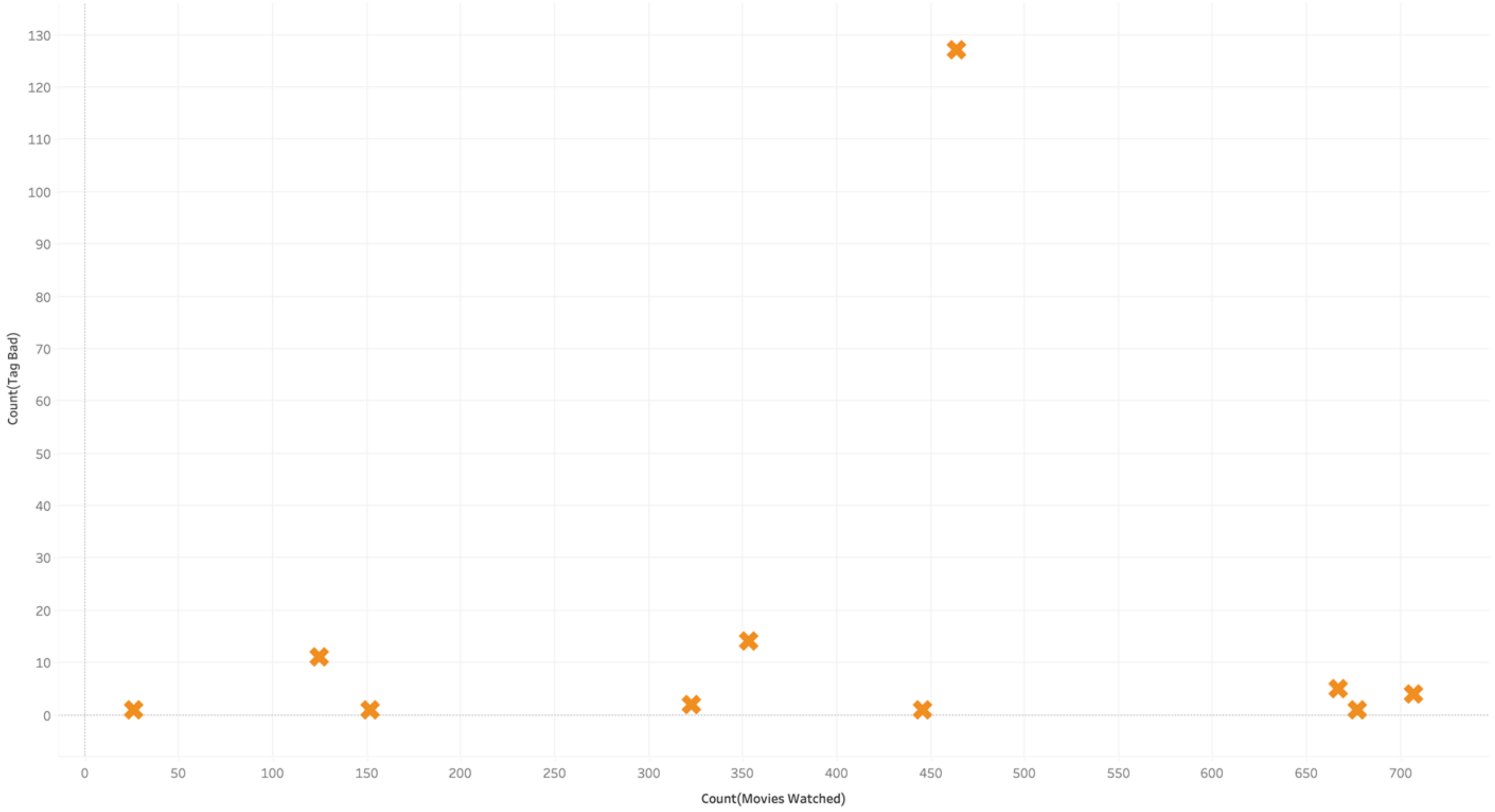


Tags do not correlate with movies watched

Few people provide most of the tags

Results
Exploratory Analysis

Number good tags given per movies watched

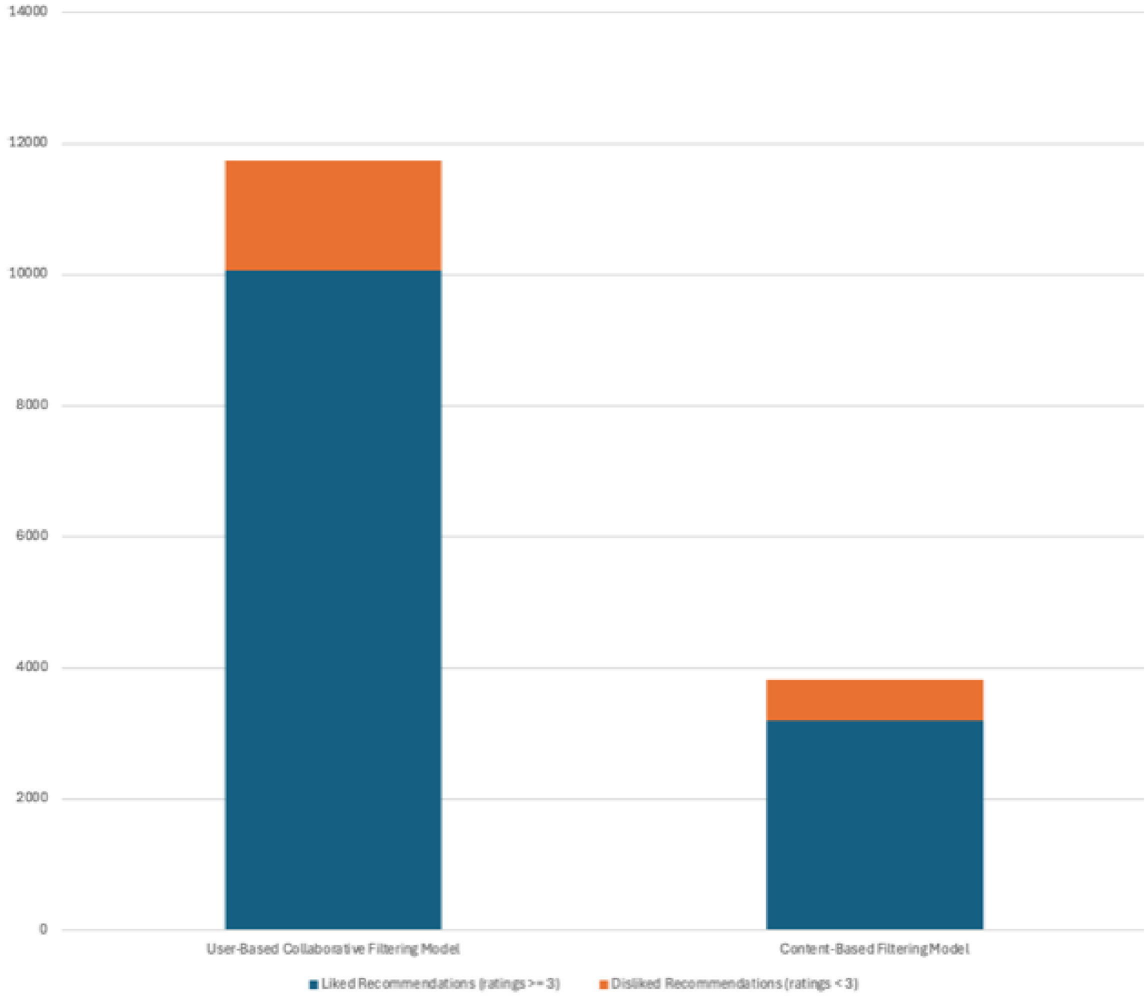


Very few tags associated with bad ratings

Bad tags overly represent the opinions of only ten people (And only <u>one</u> in particular)

Results
Model Evaluation

Number of positive and negative recommendations per model



Testing Strategy

- **80/20 Train/Test Split per User**
- **Stratified Sampling**
- **Evaluation solely on watched movies in test set**

User-based Filtering

Precision: 85.72 %

RMSE: 1.072

Content-based Filtering

Precision: 83.66 %

Results
Personal Evaluation

User-based Filtering

Precision: 83.3%

	Tim	Ruhan	Quentin	Total
Liked	8	7	10	25
Disliked	2	3	0	5

Table 1: Personal evaluation results of user-based collaborative filtering model

Content-based Filtering

Precision: 80.0 %

	Tim	Ruhan	Quentin	Total
Liked	7	8	9	24
Disliked	3	2	1	6

Table 2: Personal evaluation results of content-based filtering model

Results
Conclusion

- **Data Exploration provided valuable insights about:**
  - rating distribution
  - negative quality-quantity correlation
  - Relative dominance of frequently rated genres
  - Bias of tags due to limited user pool
- **Both models achieve a similar quality of recommendations:**
  - CF Precision: 85.72 %, CBF Precision: 83.66 %

Both models have advantages and disadvantages. The preference for one over the other depends on the use-case. Methods can be combined to produce more robust results.

Summary

M Ű E G Y E T E M 1 7 8 2

1	Exploratory Data Analysis of the MovieLens Dataset and Discovery of valuable insights	✓
2	Development of a User-Based Collaborative Filtering Model for Movie Recommendations	✓
3	Development of a Content-Based Filtering Model for Movie Recommendations	✓
4	Evaluation and Comparison of both Models using sampled Test Data	✓
5	Robust Recommendations with similar Precision Metrics between 80-85 %	✓

Contact

Tim Benjamin Hoffmann
Electrical Engineering Student
E-mail: tim.b.hoffmann@edu.bme.hu