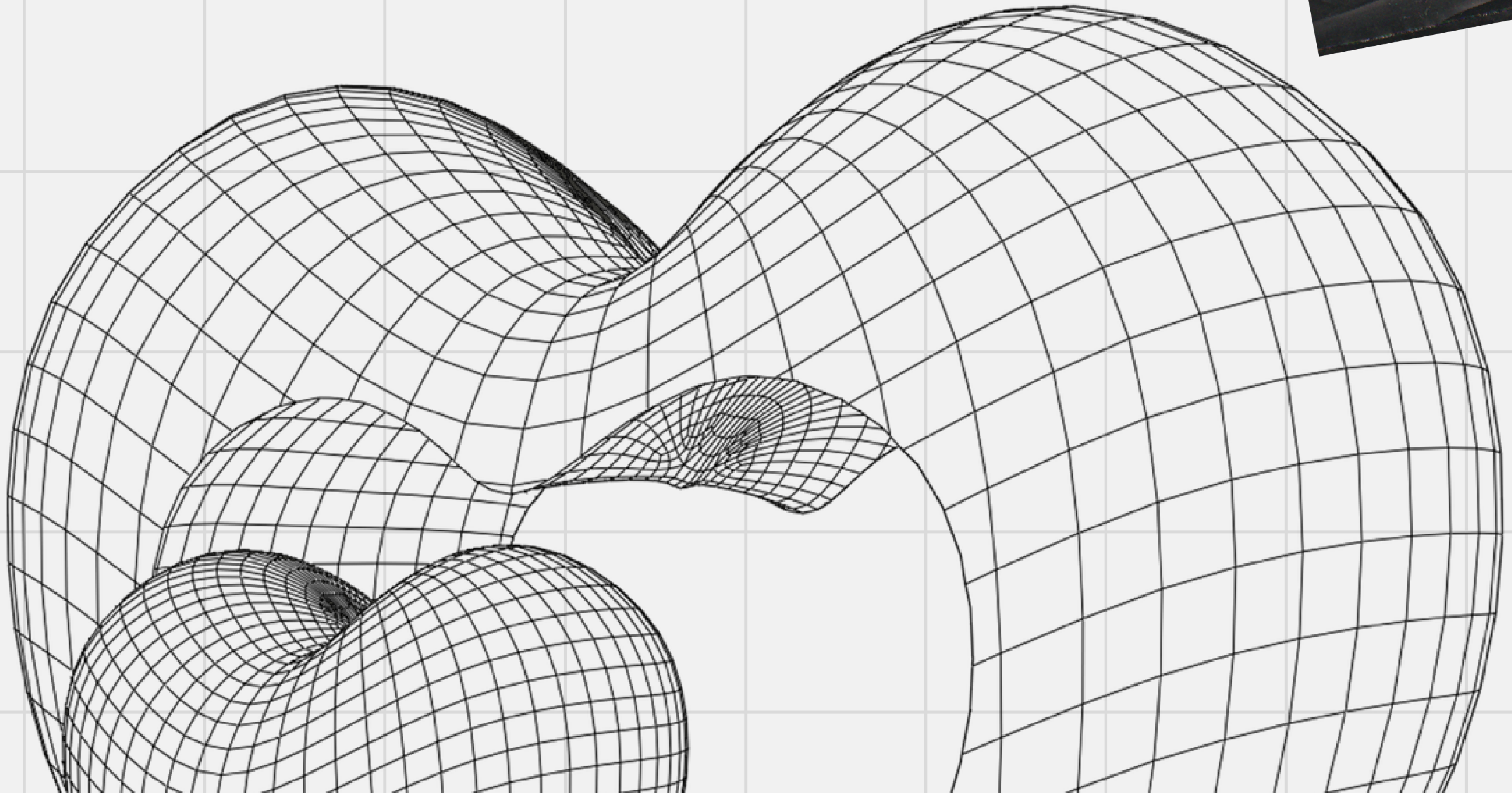
MOVIE RECOMMENDATION SYSTEM

ST312



Yiwei Zhou, Beining Pu

TABLE OF CONTENT

INTRODUCTION:

Information Overload

People make decisions when using the internet, like choosing what movies to watch, websites to visit, or which products to buy online. Vast amount of choices

Our project tackles this by using and comparing advanced **collaborative filtering** techniques on the MovieLens 100K dataset.

Improve the accuracy of movie rating prediction by users. Our project seeks to **personalise model**, ensuring that recommendations are both precise and meaningful to each viewer.

OVERVIEW OF MOVIELENS 100K DATASET

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy
...	...	...	...
9737	193581	Black Butler: Book of the Atlantic (2017)	Action|Animation|Comedy|Fantasy
9738	193583	No Game No Life: Zero (2017)	Animation|Comedy|Fantasy
9739	193585	Flint (2017)	Drama
9740	193587	Bungo Stray Dogs: Dead Apple (2018)	Action|Animation
9741	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931
...	...	...	...	...
100831	610	166534	4.0	1493848402
100832	610	168248	5.0	1493850091
100833	610	168250	5.0	1494273047
100834	610	168252	5.0	1493846352
100835	610	170875	3.0	1493846415

userId	movieId	tag	timestamp
2	60756	funny	1445714994
2	60756	Highly quotable	1445714996
2	60756	will ferrell	1445714992
2	89774	Boxing story	1445715207
2	89774	MMA	1445715200
...	...	...	...
606	7382	for katie	1171234019
606	7936	austere	1173392334
610	3265	gun fu	1493843984
610	3265	heroic bloodshed	1493843978
610	168248	Heroic Bloodshed	1493844270

The MovieLens 100K dataset comprises **100,836 ratings** from **610 users** on **9,742 movies**, with ratings ranging from 0.5 to 5 stars. The time span is from 1996 to 2018.

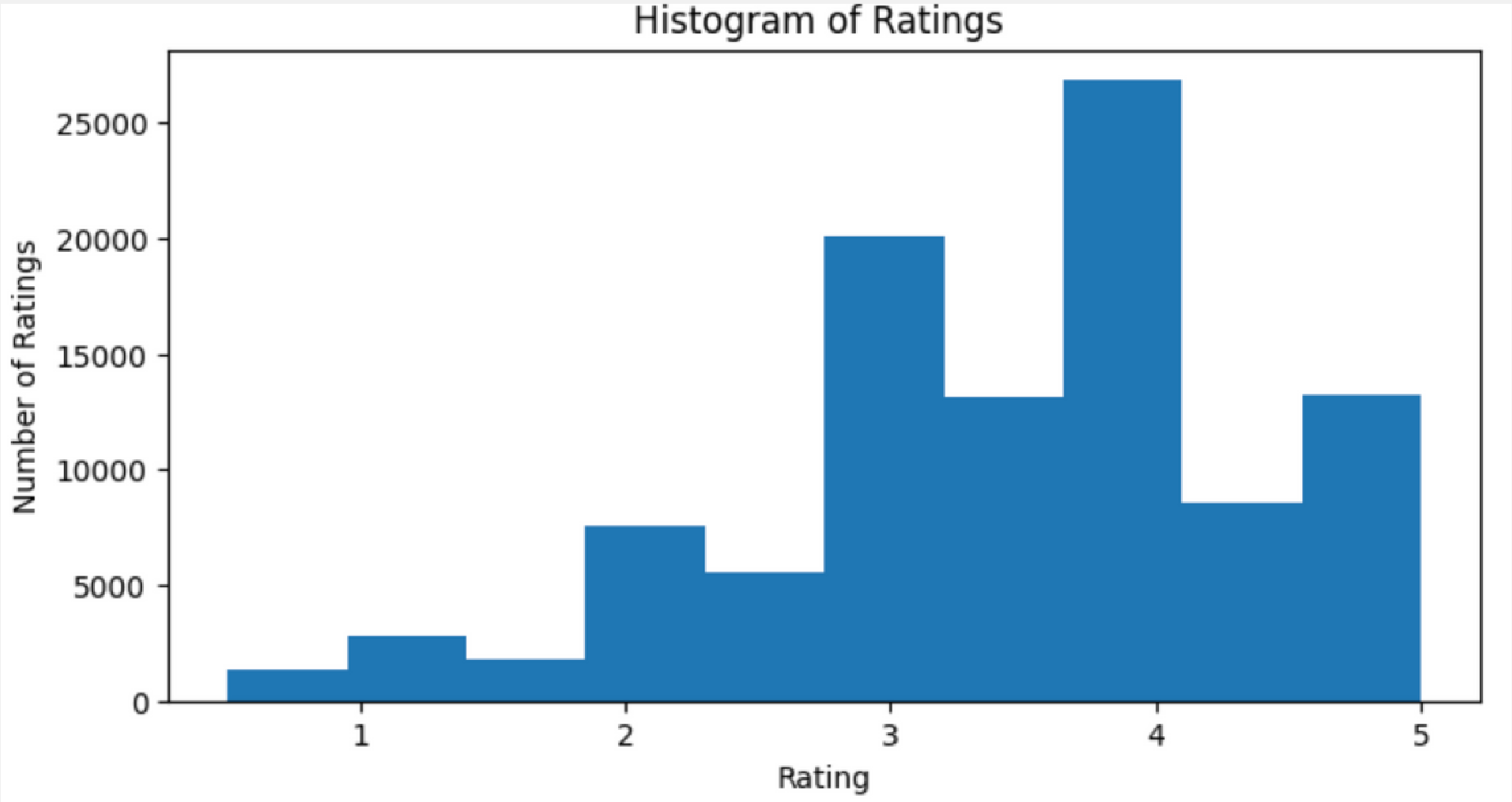
EXPLORATORY DATA ANALYSIS

Histograms and **line graphs**, **cloud map** to show the trends and patterns in data

- 1.DISTRIBUTION OF MOVIE **RATINGS**
- 2.NUMBER OF RATINGS PER MOVIE
- 3.MOVIE COUNTS BY GENRE
- 4.AVERAGE RATINGS BY GENRE
- 5.USER SEGMENTATION
- 6.RATINGS OVER TIME
- 7.TAG USAGE DISTRIBUTION
- 8.MOST COMMON TAGS

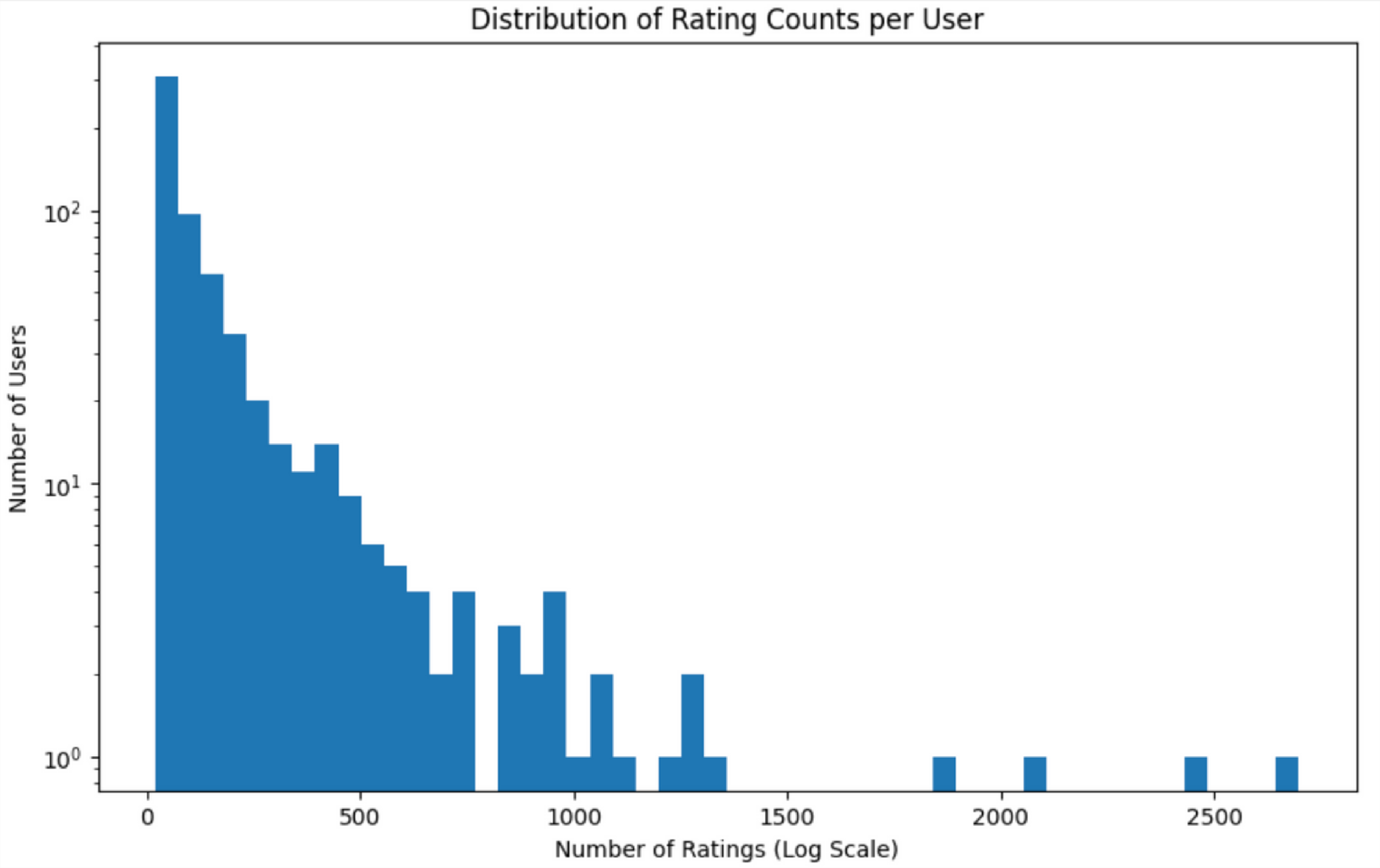
DISTRIBUTION OF MOVIE RATINGS

- It shows a leftward skew with a peak at 4-star ratings
- Users generally rate movies favorably. The trend suggests that users are less likely to give lower ratings, highlighting potential biases in user feedback



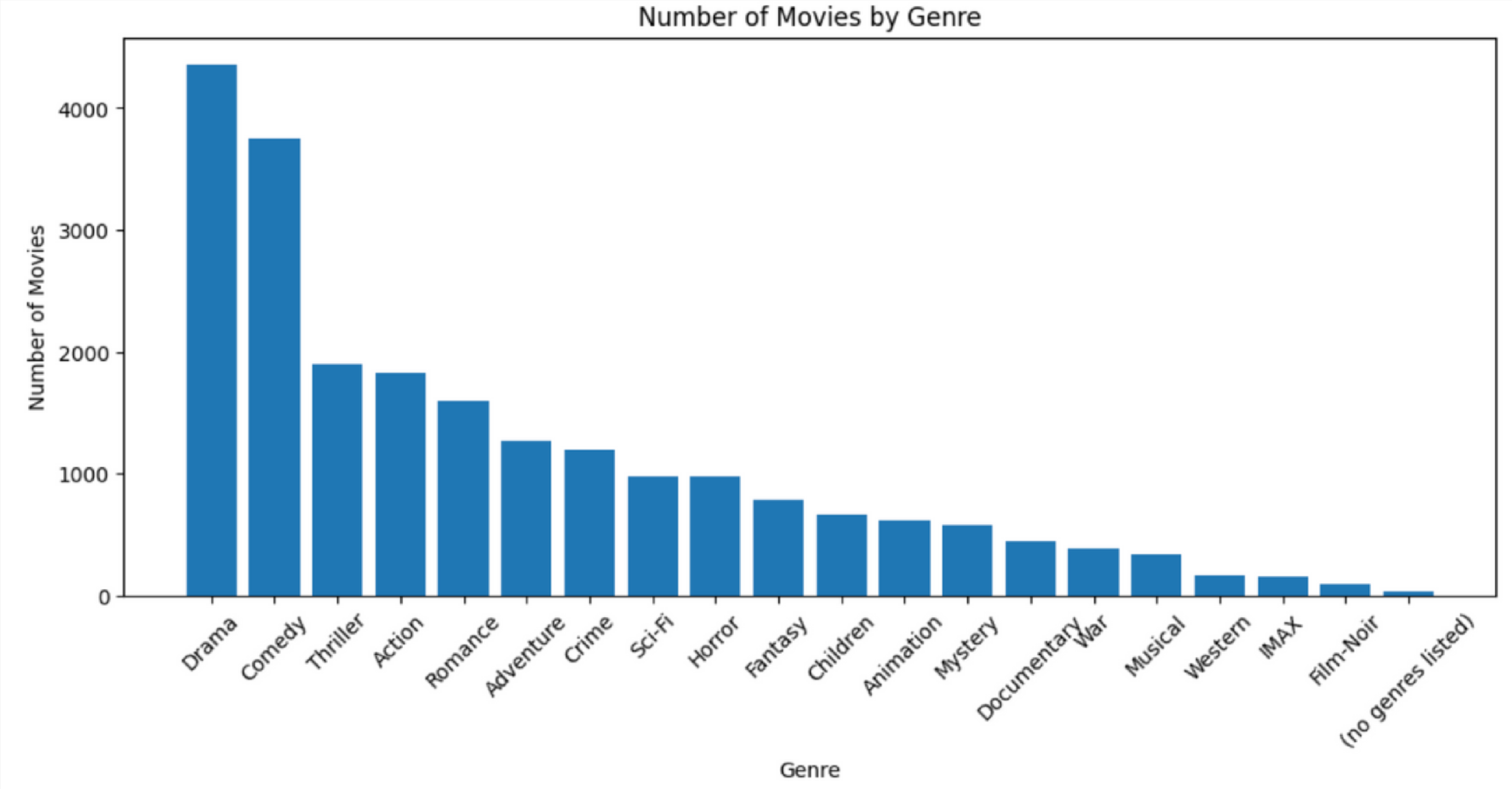
NUMBER OF RATINGS PER MOVIE

- Most movies have fewer than 100 ratings
- A rightward-skewed distribution.
- A few popular movies accumulate many ratings, the majority receive relatively few
- This impact the robustness of the recommendations



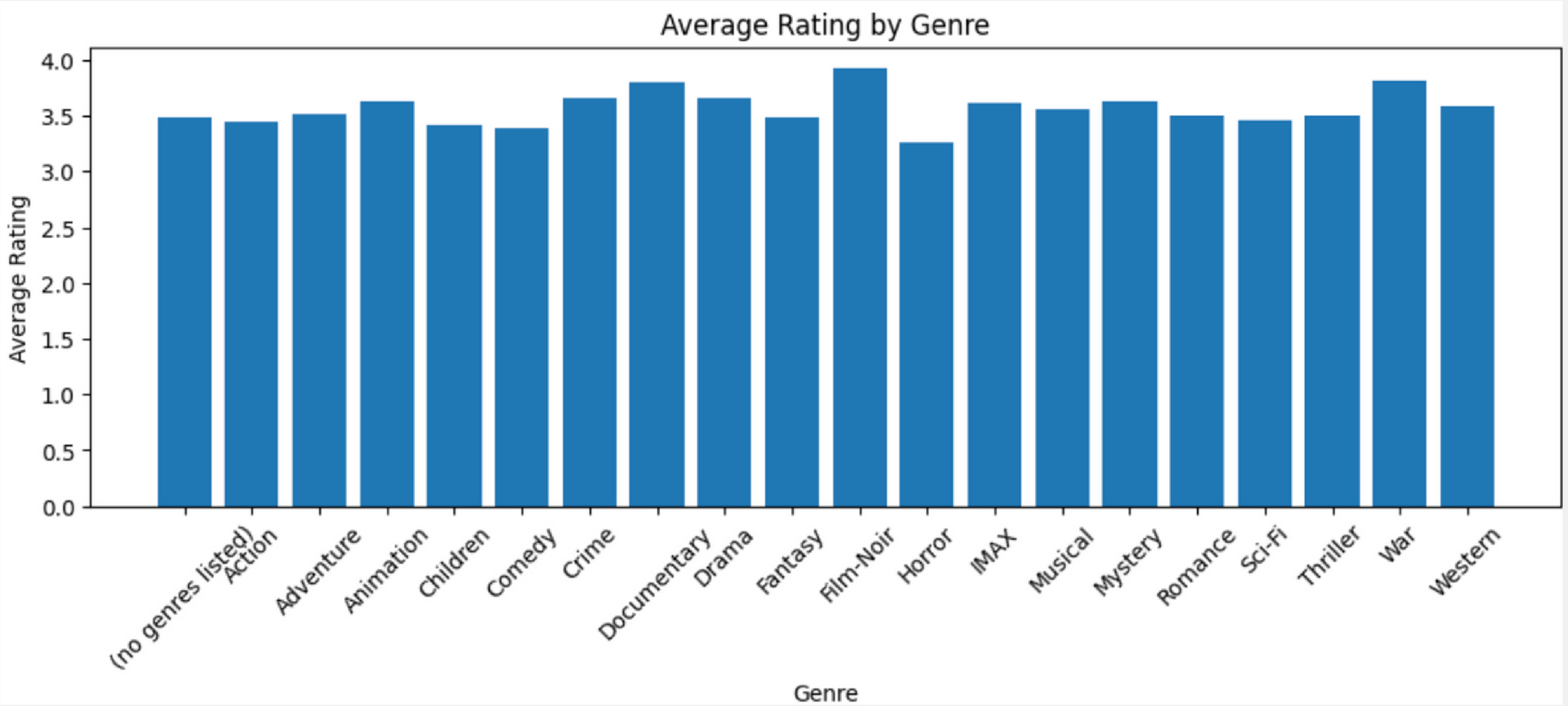
MOVIE COUNTS BY GENRE

- Drama, Comedy, and Thriller have the highest counts
- Genres like Documentary and War are less represented, which may affect the diversity of recommendations



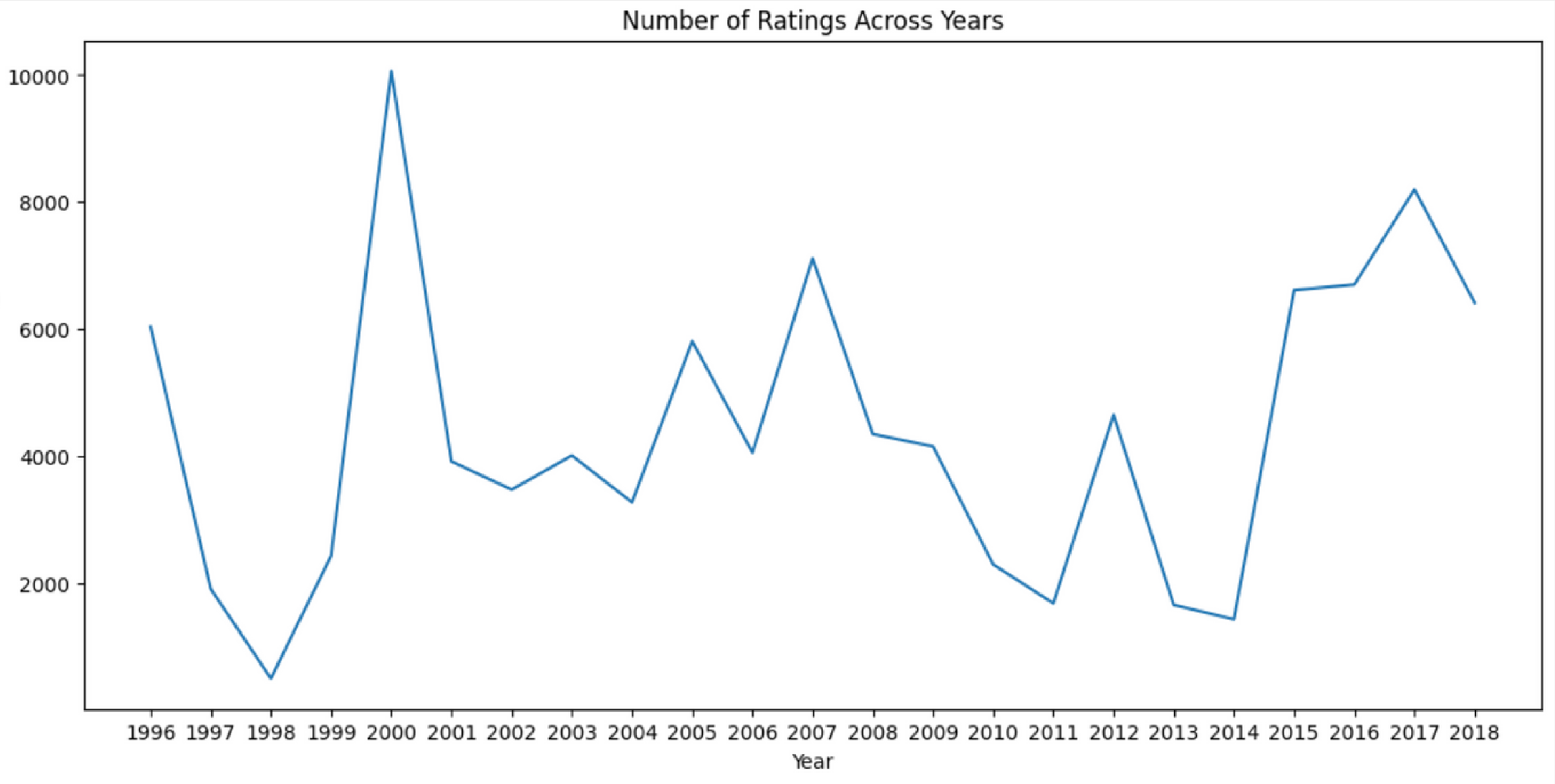
AVERAGE RATINGS BY GENRE

This bar chart compares the **average ratings across different genres**. Genres like Film-Noir and Documentary receive slightly higher average ratings, while popular genres such as Comedy and Children's movies receive lower scores. This variation in ratings by genre can guide targeted recommendations



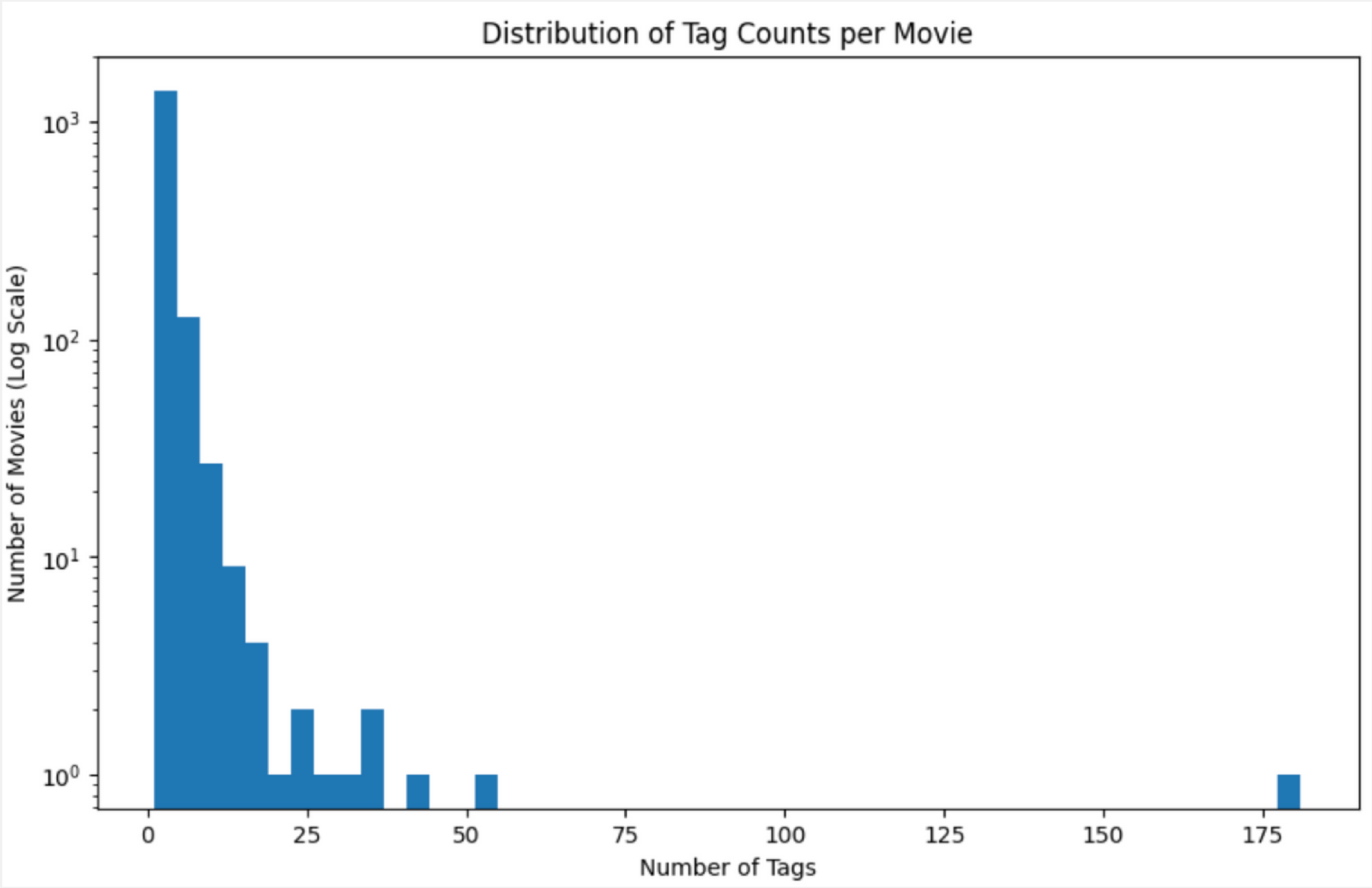
RATINGS OVER TIME

The line chart tracks the **number of ratings users provided over the years**. A notable peak in the year 2000 suggests increased activity, possibly influenced by popular movie releases or enhancements in the MovieLens platform.



TAG USAGE DISTRIBUTION

- Highly skewed. Few users utilize tagging extensively
- Most tags are concentrated among a small group.
- Its sparcity limits their effectiveness

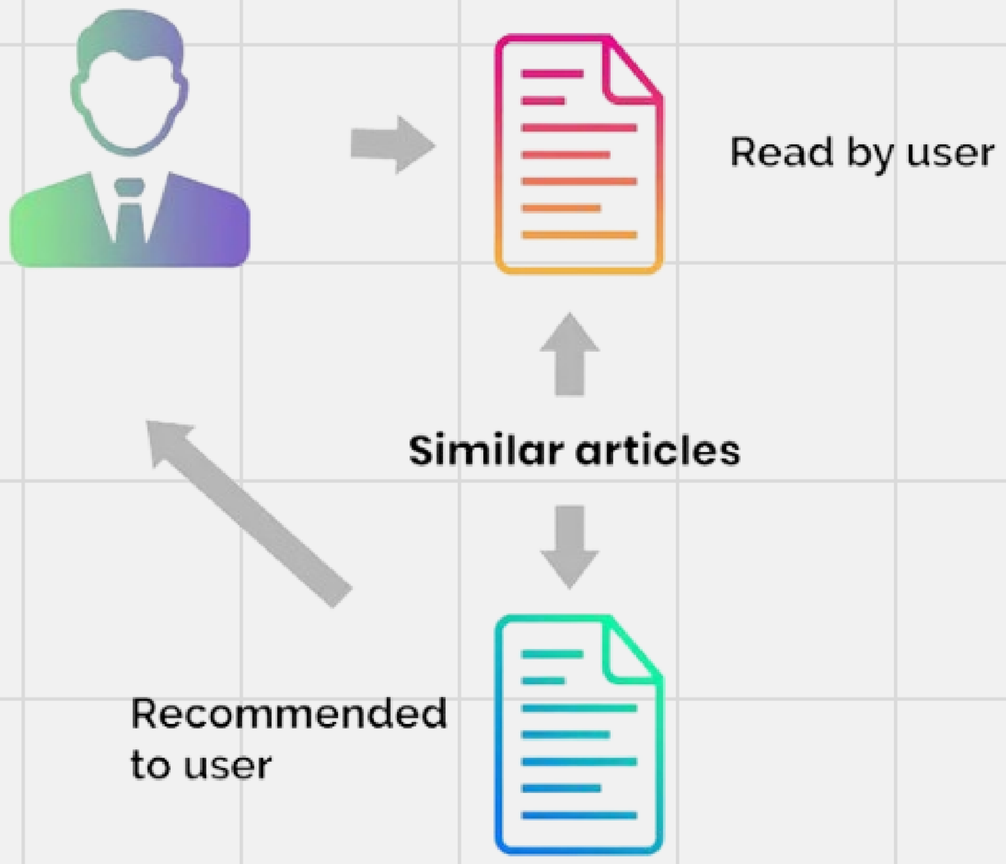


MOST COMMON TAGS

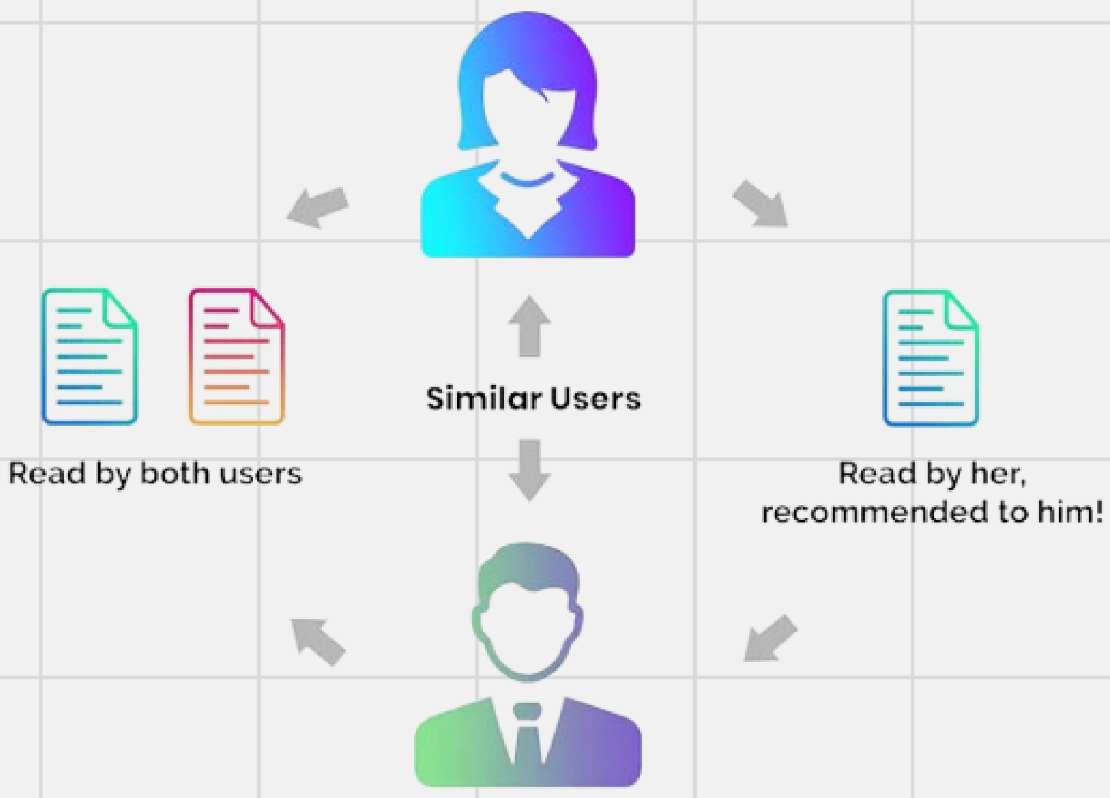
- Users mainly use movielens platforms as an alternative to Netflix's own recommendation system

CLASSIFICATION OF RECOMMENDATION SYSTEMS

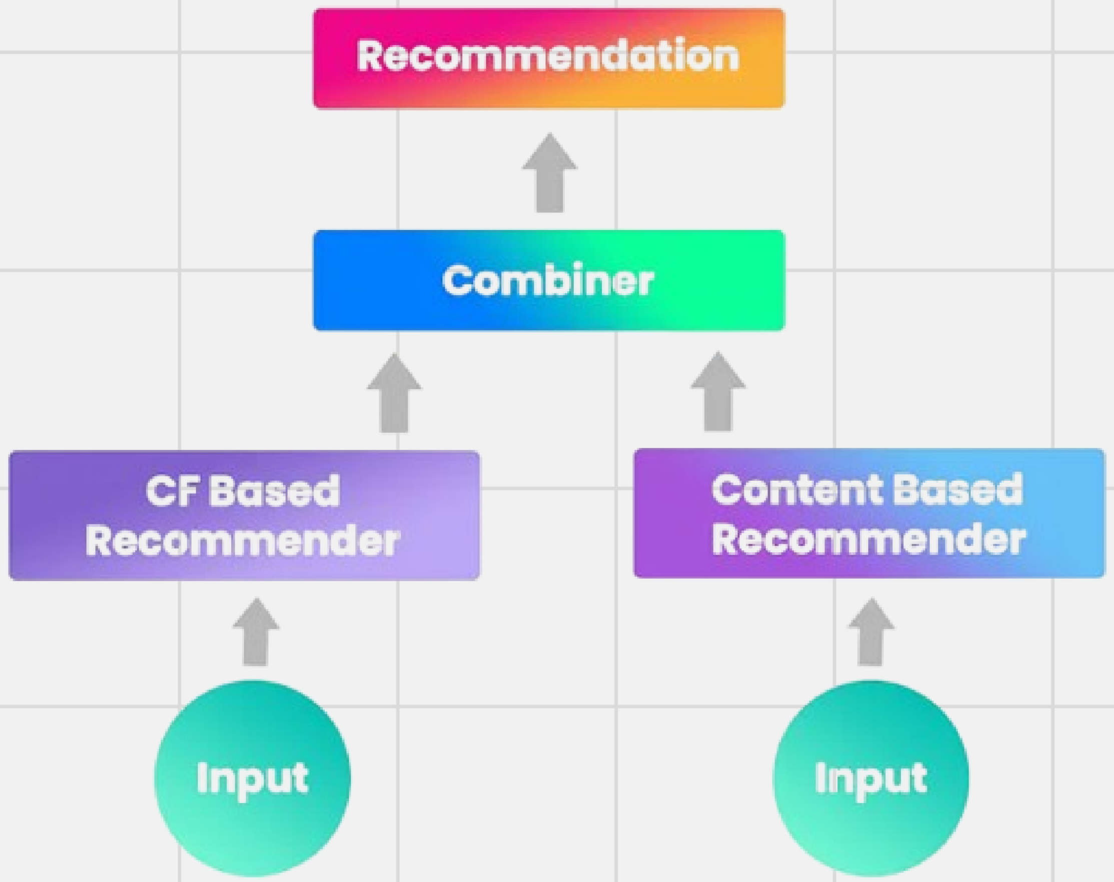
Content-based filtering



Collaborative filtering



Hybrid Recommendations



METHODS----- COLLABORATIVE FILTERING

Collaborative filtering is a technique that predicts a user's preferences based on the likes and dislikes of similar users. It operates under the **assumption that those who agreed in the past will agree again**. By analysing patterns in user behavior, collaborative filtering offers personalized recommendations.

Matrix Factorization

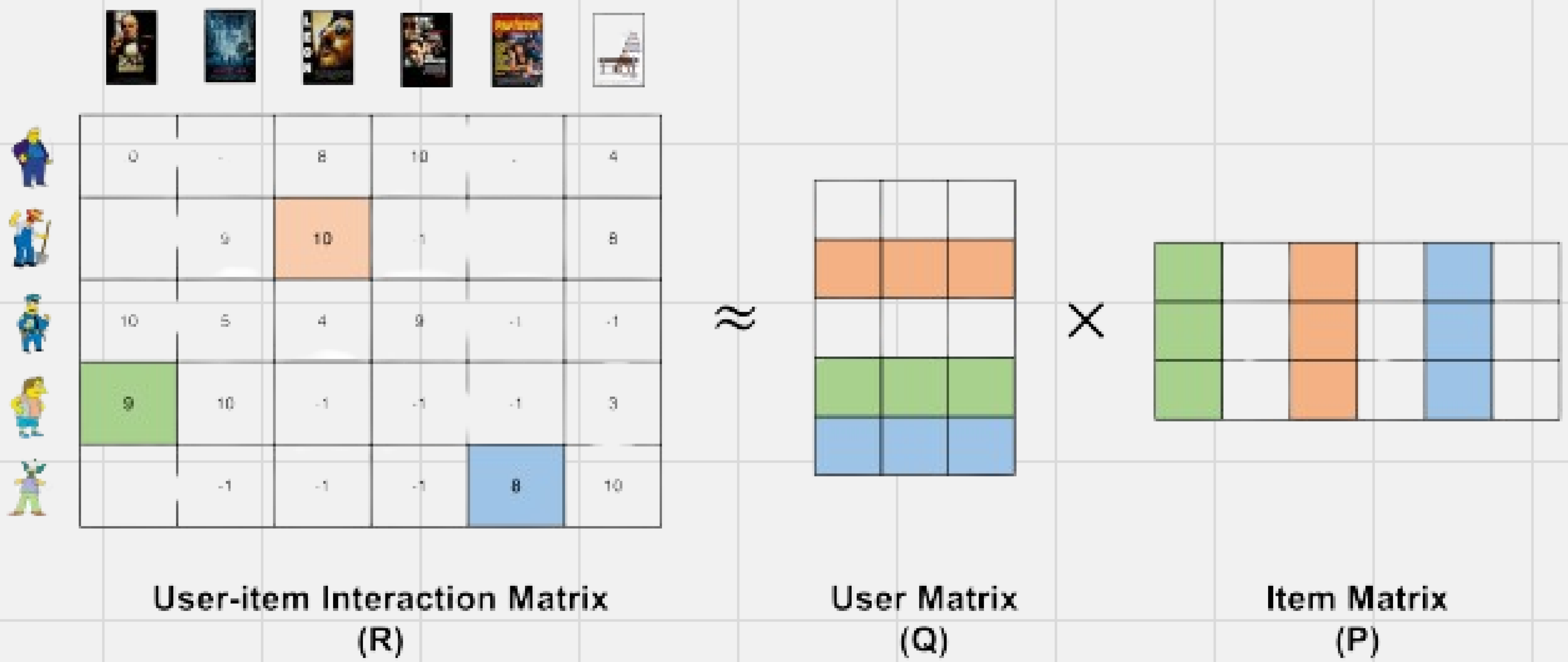
used to **decompose the user-item interaction matrix into lower-dimensional forms**, revealing latent features of users and items.

Transformer

utilise **self-attention mechanisms** to model all interactions between users and items within a **sequence**

MATRIX FACTORISATION

- The sparse user-item ratings matrix decomposed into lower dimensional user and item matrices.
- Traditional Singular Value Decomposition does not work for our dataset because the dataset has missing values.
- The predicted rating for the last 2 ratings by a user is calculated as the dot product



TRANSFORMERS

Self-attention mechanism

Learned Embedding

We can put **different embedding** into model for learning the pattern between them

+

Positional Encoding

Adds information about the **position** of items in the sequence to the input embeddings. Understanding **user behavior over time.**

STEP 1

Create suqences for input embeddings:

UserID embedding

MovieID embedding

Rating history embedding

Genre embedding

STEP 2

Model 1:

UserID embedding

MovieID embedding

+

Positional Encoding

STEP 3

Model 2:

UserID embedding

MovieID embedding

Rating embedding

+

Positional Encoding

STEP 4

Model 3:

UserID embedding

MovieID embedding

Rating embedding

genre embedding

+

Positional Encoding

TRANSFORMER MODELS

RESULTS COMPARISON

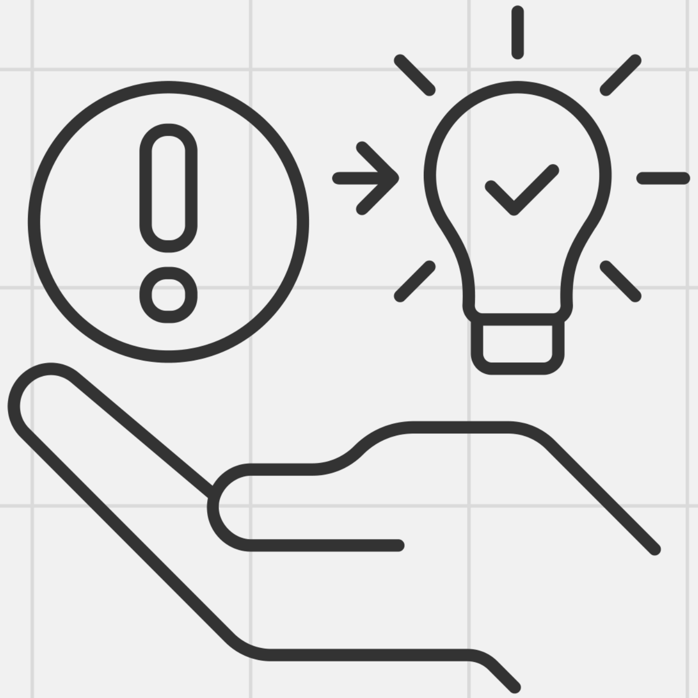
	MSE	MAE
Baseline Model	1.197	0.897

Baseline model set every predicted rating as **3.5** (median rating)

	MSE	MAE
with Bias	1.181	0.872
withou Bias	1.292	0.943

	MSE	MAE
Model 1	0.942	0.746
Model 2	0.893	0.729
Model 3	0.860	0.715

LIMITATIONS



Static Preferences: Models don't account for changing user tastes over time

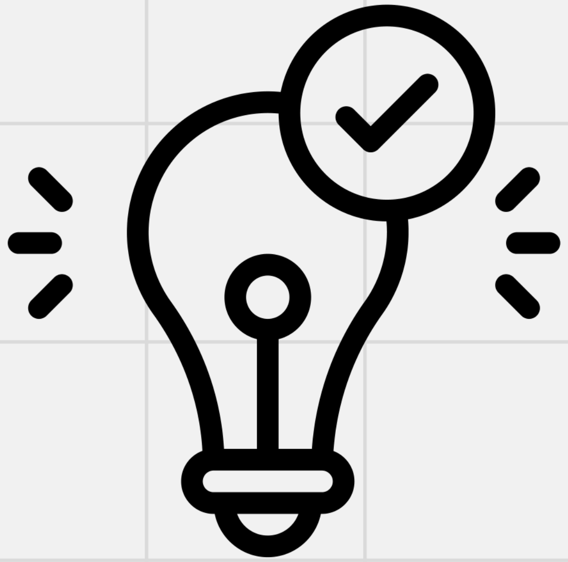
Data Sparsity: Most potential user-item interactions are not recorded.

Linearity Assumption: Matrix factorization assumes linear user preferences

Overfitting: Hard to capture connections across long user histories.

Computational Cost: Transformers require substantial computing resources

CONCLUSION AND FUTURE WORK



Key Findings: Transformers with more feasures improve accuracy over baselines. Transformers generally better than matrix factorisation.

Best Model: Transformer 3 with all avaible embeddings and positional encodings.

Future Directions: ‘*Hybrid Methods*’
Combine collaborative filtering with content-based filtering. Adding tags and comment into model

Temporal Models:
Account for changes in user preferences over time.

THANK YOU