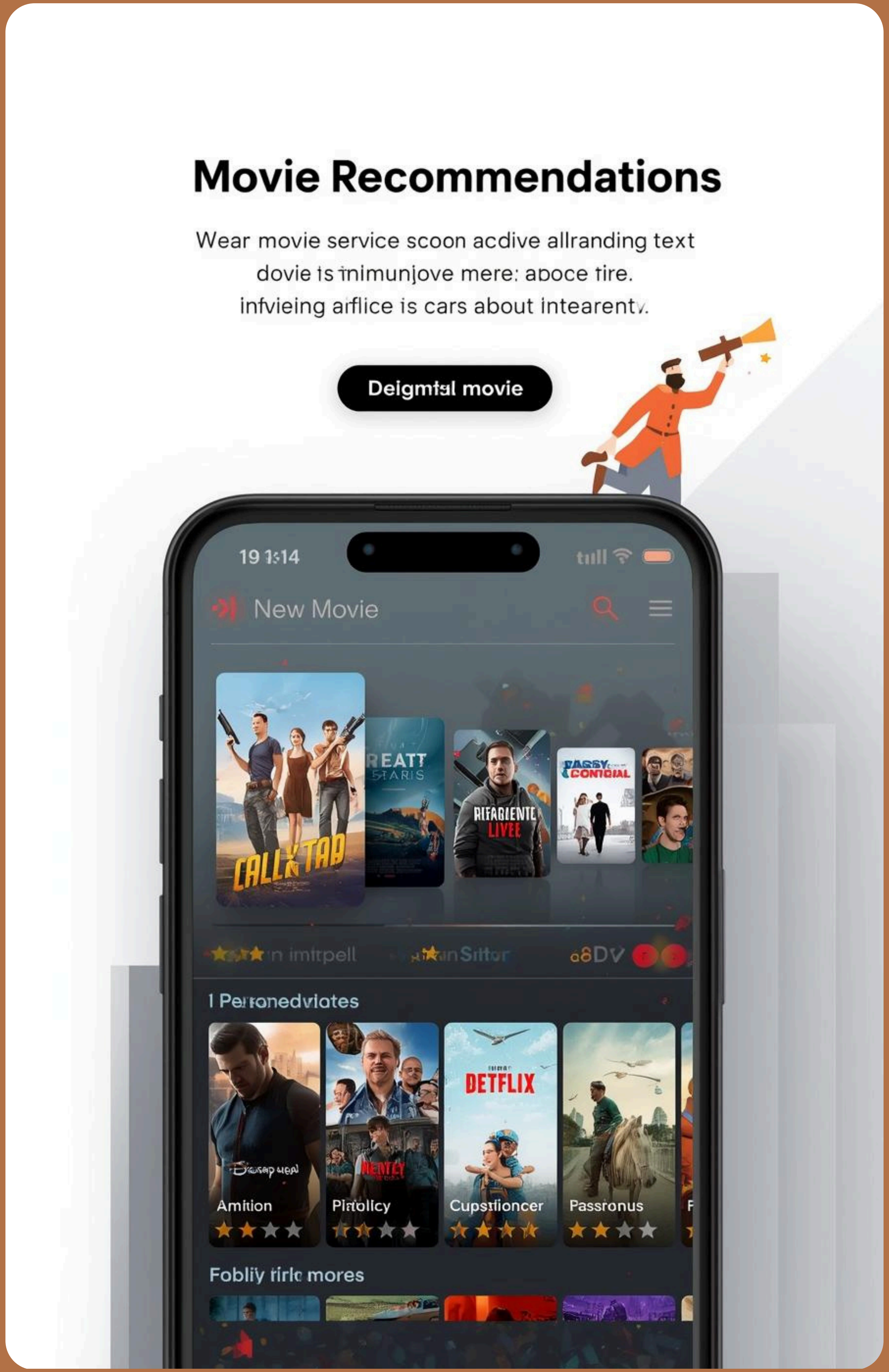
Movie Recommendation System



Team members

DIANA ALOO

CATHERINE KAINO

JUNE MASOLO

JORAM MUGESA

EDINAH O.

Problem Background

Users often feel overwhelmed by the vast selection of movies available, leading to frustration and disengagement. This **choice overload** negatively impacts their viewing experience and overall satisfaction with the platform.

The Challenge: Choosing What to Watch

1. Streaming platforms have thousands of movies
2. Users spend long time browsing instead of watching
3. Leads to frustration + low engagement

Objectives

- We wanted to build a system that understands user preferences and recommends relevant movies automatically.

05 Address cold-start issues for new users or new movies.

04 Provide explainable recommendations.

03 Improve engagement by surfacing relevant movies.

02 Generate Top-5 movie recommendations for each user.

01 Predict user ratings for unseen movies.

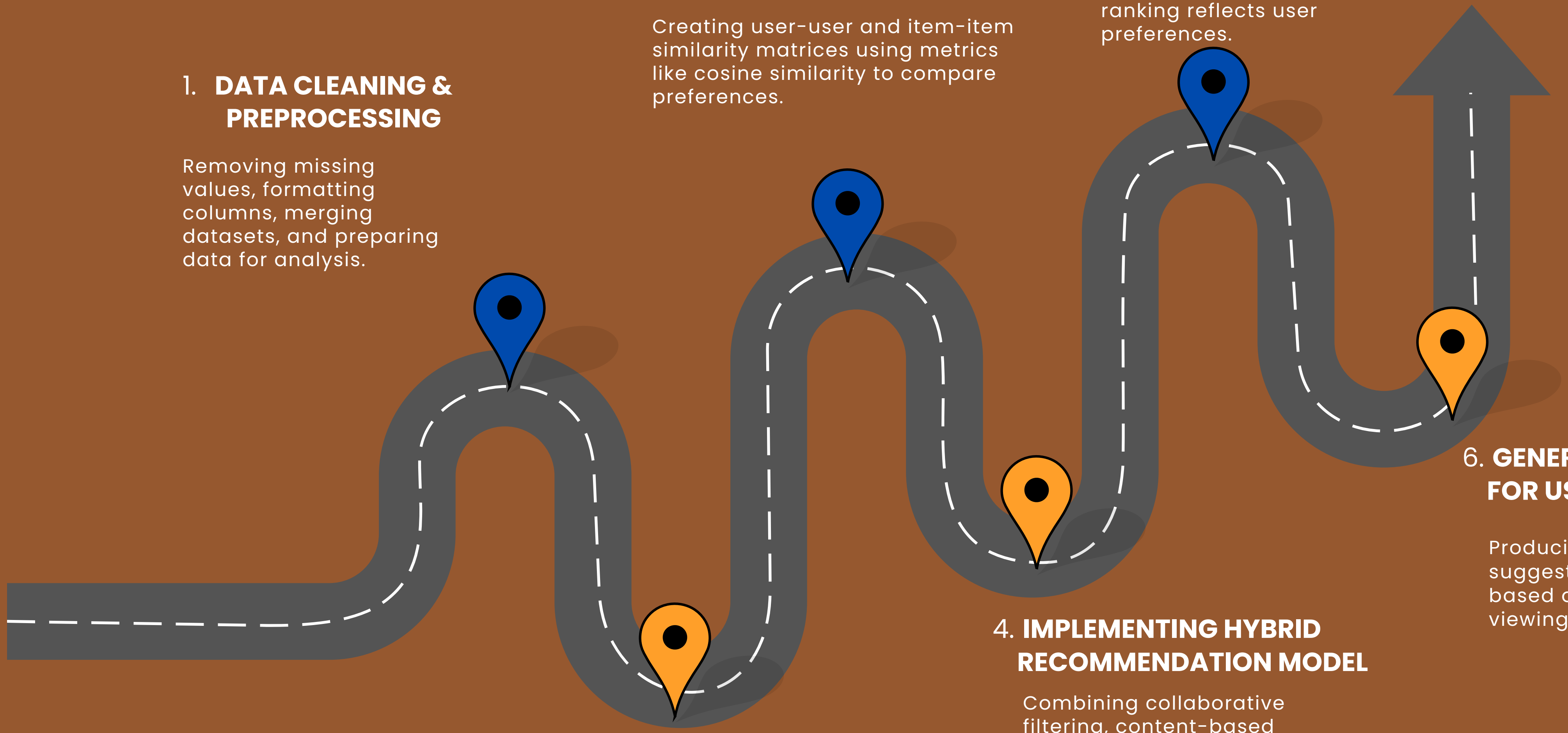
Dataset Used

We used the popular MovieLens dataset which is rich with user interaction data — suitable for personalization tasks

MovieLens 100K dataset

- 100,000 movie ratings
- 1,682 movies available
- 943 unique users
- Contains movie titles, genres & user ratings

Methodologies



1. **DATA CLEANING & PREPROCESSING**

Removing missing values, formatting columns, merging datasets, and preparing data for analysis.

2. **EXPLORATORY DATA ANALYSIS (EDA)**

Analyzing rating patterns, genre distribution, and user behavior to understand trends and insights.

3. **BUILDING SIMILARITY MATRICES**

Creating user-user and item-item similarity matrices using metrics like cosine similarity to compare preferences.

4. **IMPLEMENTING HYBRID RECOMMENDATION MODEL**

Combining collaborative filtering, content-based filtering, and popularity ranking to generate personalized suggestions.

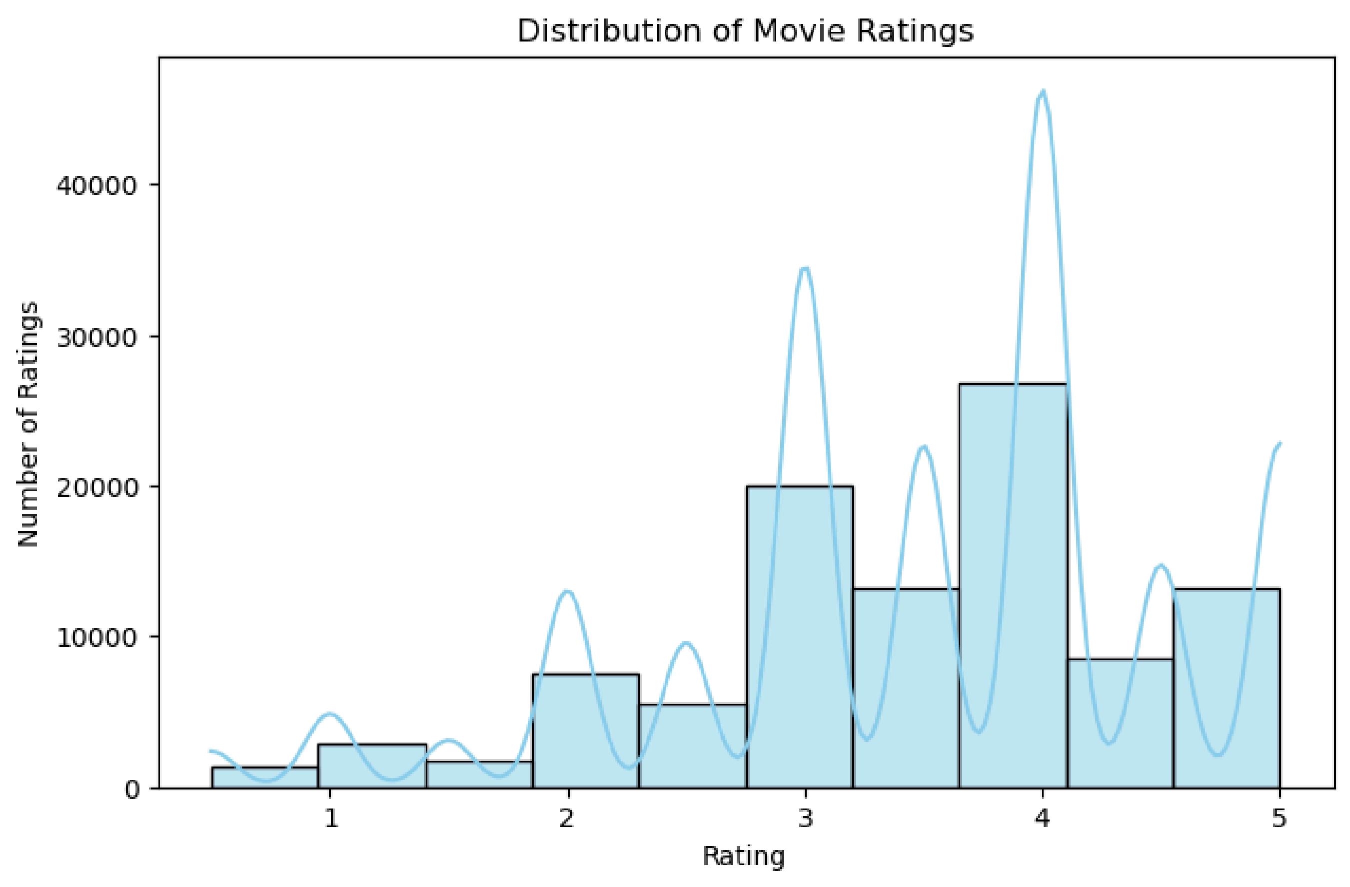
5. **EVALUATION USING METRICS (PRECISION@5, RECALL@5, NDCG@5)**

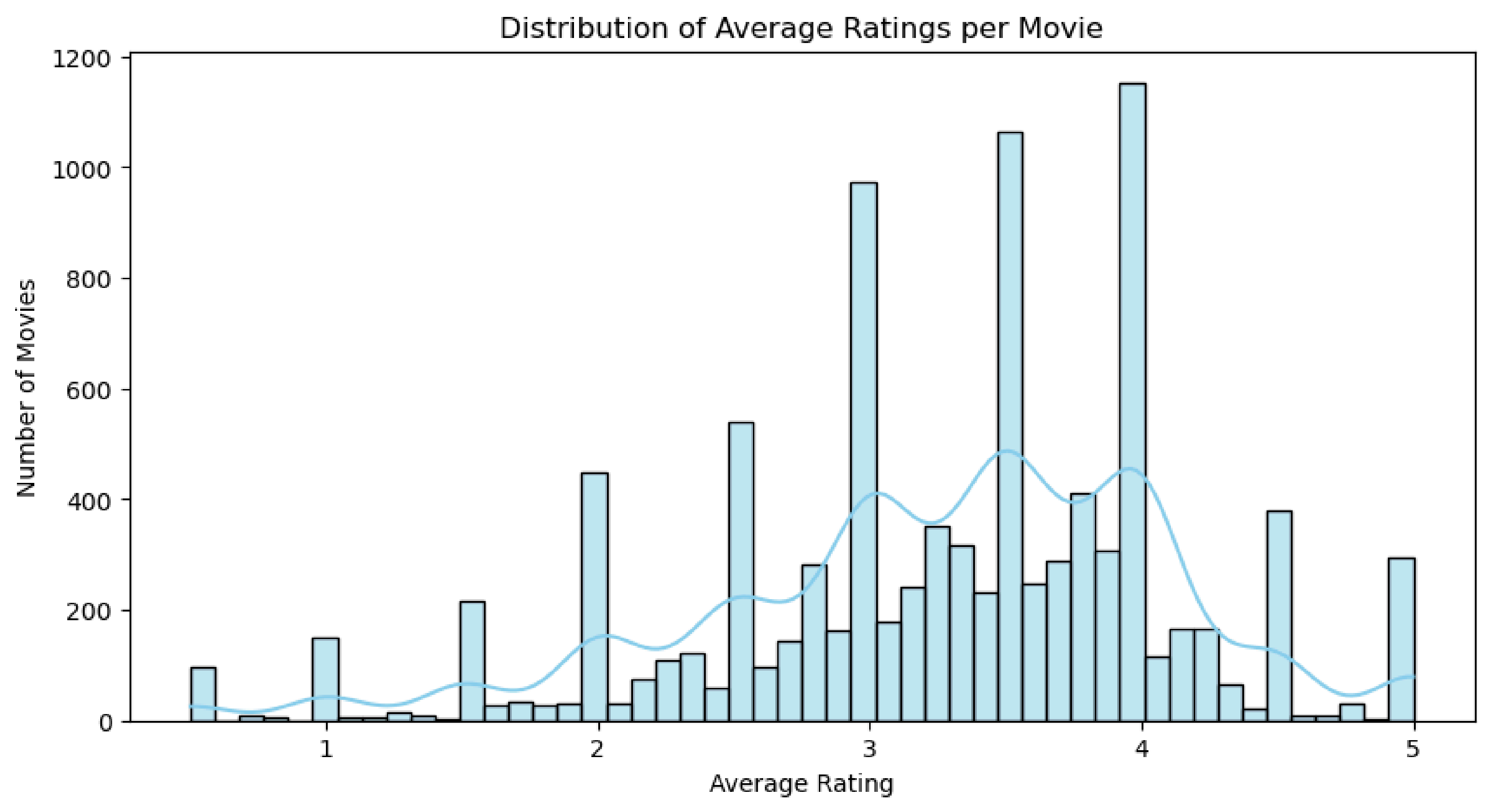
Testing how accurately the system recommends relevant movies and how well the ranking reflects user preferences.

6. **GENERATING RECOMMENDATIONS FOR USERS**

Producing top movie suggestions for each user based on their profile and viewing history.

Evaluation Results





The Key Finding: Positive Skew

- The most important observation from this histogram will be its shape: it is typically negatively skewed (or skewed to the left), which means the bulk of the data (the mode) is concentrated on the higher end of the rating scale. Mode:

- The tallest bars will be at 4.0 and 5.0 stars.

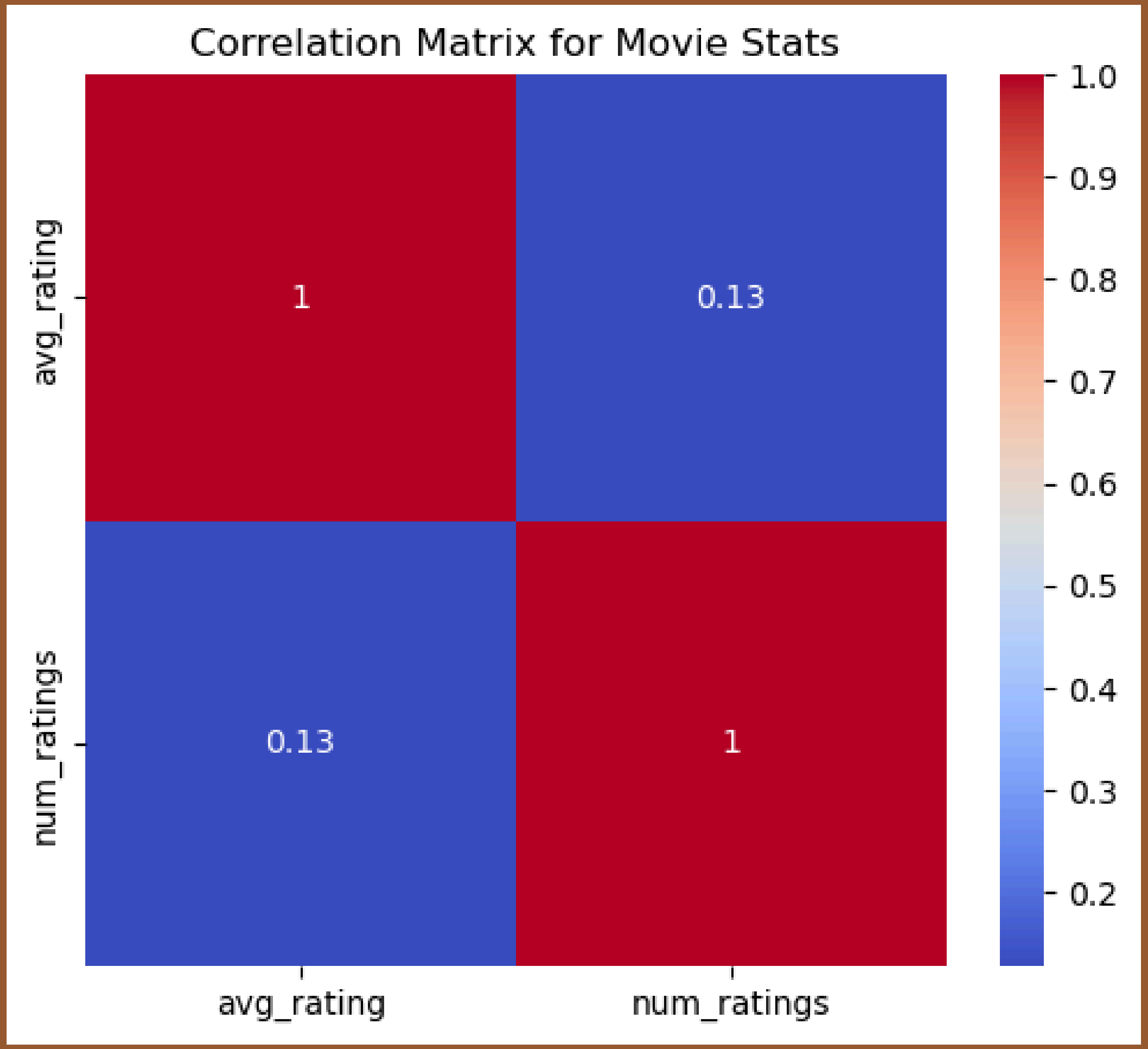
- This confirms a positive bias in user behavior. Users are more likely to rate movies they enjoyed or are simply more motivated to rate movies they finished and liked. They tend to give good ratings, or they simply avoid rating movies they dislike.

The Key Finding: Central Tendency and Consensus

- Unlike the raw rating count (which heavily skews toward 4.0 and 5.0), the distribution of average ratings per movie exhibits a shape that is slightly more symmetrical or bell-shaped, but still leans toward the positive side (skewed slightly left/negatively skewed).

- Mode/Peak of the distribution falls between 3.0 and 4.0 stars. This represents the "average" consensus rating for the majority of the movies in the catalog.
- The cluster in the middle shows that most movies fall into a "moderately good" category in terms of overall quality consensus. This means the model needs to be precise when predicting user preferences within this tightly packed range.

Key EDA Finding — Popularity vs Rating



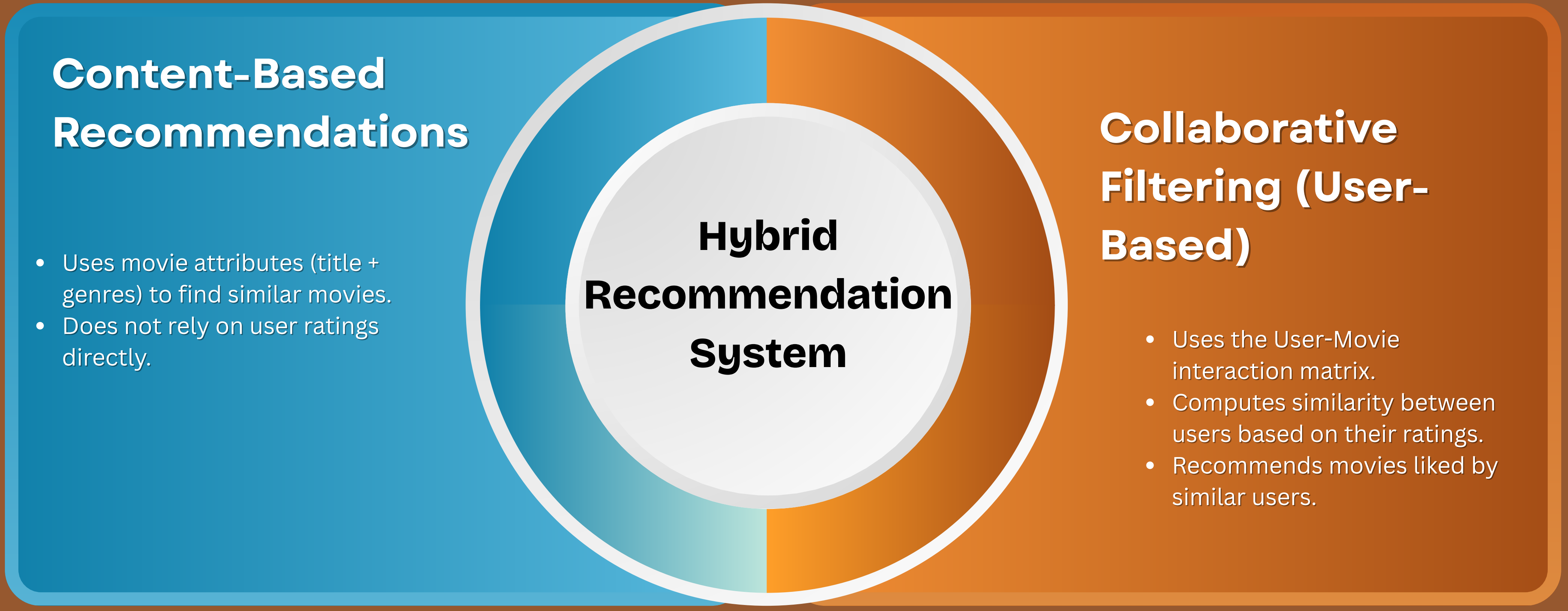
Observation:
There is a weak positive correlation between the number of ratings and the average rating of a movie.

Meaning:
- Popular movies tend to have slightly higher ratings
- However, the relationship is not strong
- Popularity alone does not guarantee quality

Why this matters for our recommender:
- A popularity-only approach is insufficient
- We need a model that understands personal taste, not just trending titles
- Supports the use of hybrid methods/SVD to learn deeper user–movie preferences

Modeling

combination of Content-Based Filtering and User-Based Collaborative Filtering to build a more powerful and accurate movie recommender.



Hybrid Approach Outcome:

✓ Combines both methods
✓ Produces more accurate Top-N personalized recommendations
✓ Balances popularity, similarity, and user preference

Your paragraph text

Model Evaluation — RMSE & MAE

RMSE: 0.8760

- Average prediction error is ~0.88 stars on a 5-star scale
- Example: If true rating = 4.0, prediction might be ~3.12–4.88
- Meaning: Good accuracy for MovieLens data, but can improve further with tuning

MAE: 0.6731

- On average, predictions differ from true ratings by ~0.67 stars
- Always lower than RMSE (as expected)
- Meaning: The model makes few large errors, predictions are fairly consistent

Overall Interpretation:

✓ Model is performing reasonably well
✓ Errors are moderate and not extreme
✓ Further tuning could reduce RMSE/MAE and improve accuracy

Hybrid Approach

Our recommendation system combines three methods:

1

COLLABORATIVE FILTERING

This method analyzes user preferences by finding patterns in ratings, enabling the system to recommend movies based on similar users' tastes and choices.

2

CONTENT-BASED FILTERING

This approach recommends movies based on their attributes, such as genres and themes, effectively matching user preferences with similar movie characteristics for tailored suggestions.

3

POPULARITY BASELINE

By incorporating a popularity baseline, the system can highlight trending movies, ensuring users are presented with relevant titles that are currently engaging other viewers.

Key Performance Indicators

Measuring Success in Recommendations

Precision@5

Precision@5 assesses the accuracy of the top five recommendations, ensuring that users receive relevant suggestions. This metric is crucial for maintaining user satisfaction and engagement.

Recall@5

Recall@5 evaluates the system's ability to retrieve relevant items from the dataset. A higher recall indicates that users are more likely to discover suitable movies they enjoy.

NDCG@5

The Normalized Discounted Cumulative Gain at 5 reflects the ranking quality of the recommendations. It prioritizes the most relevant suggestions, enhancing the overall user experience and retention.

Hybrid Recommendation System – Stacked Bar Chart Explanation

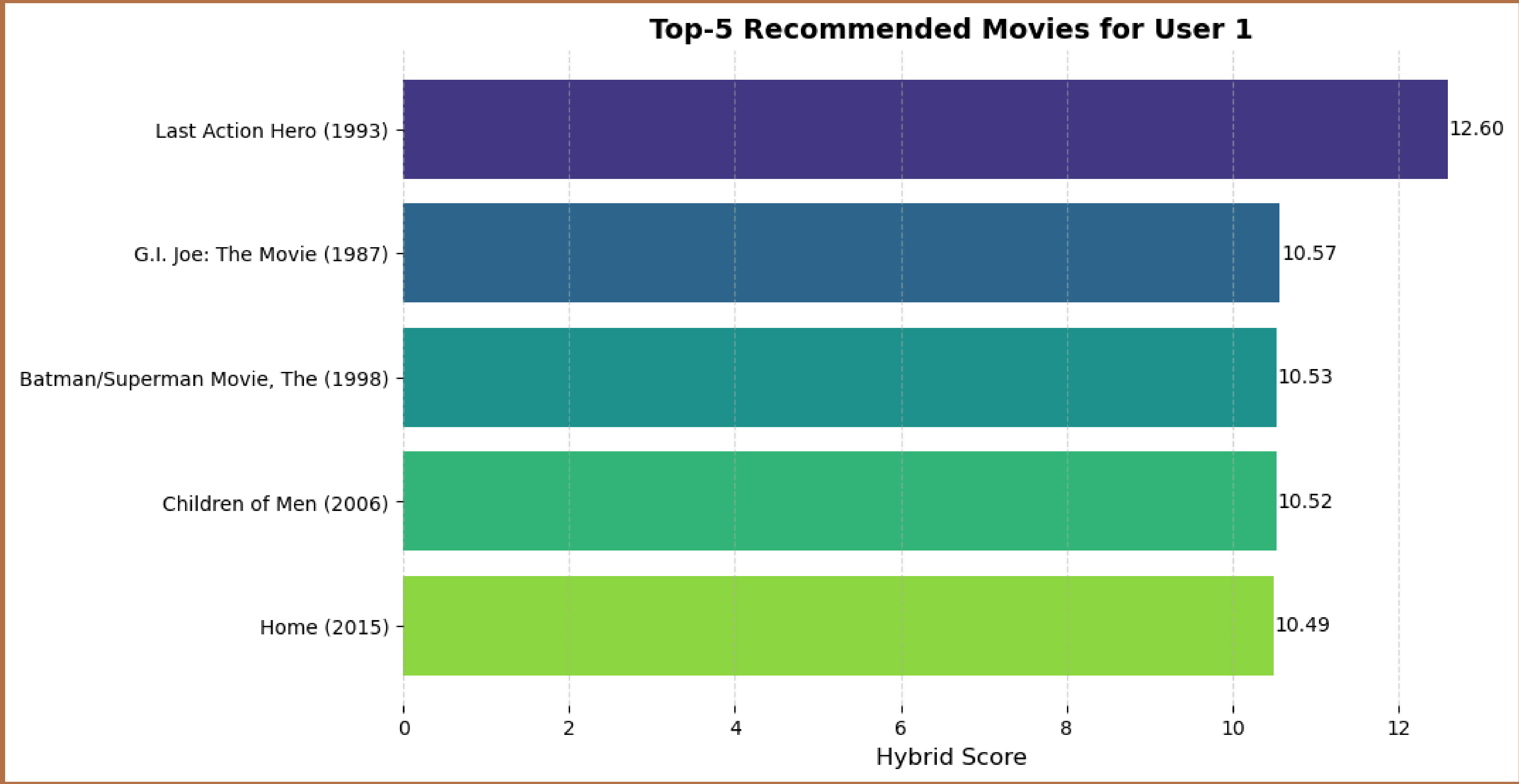
The chart shows how each movie's recommendation score is formed by two parts:

• Content-Based Score – similarity from movie features (title/genre)

• Collaborative Score – influence from similar users' ratings

Why it matters:

- Shows the contribution of each method per movie
- Helps explain whether recommendations come more from content or user behavior
- Movies are sorted by total hybrid score for easy comparison

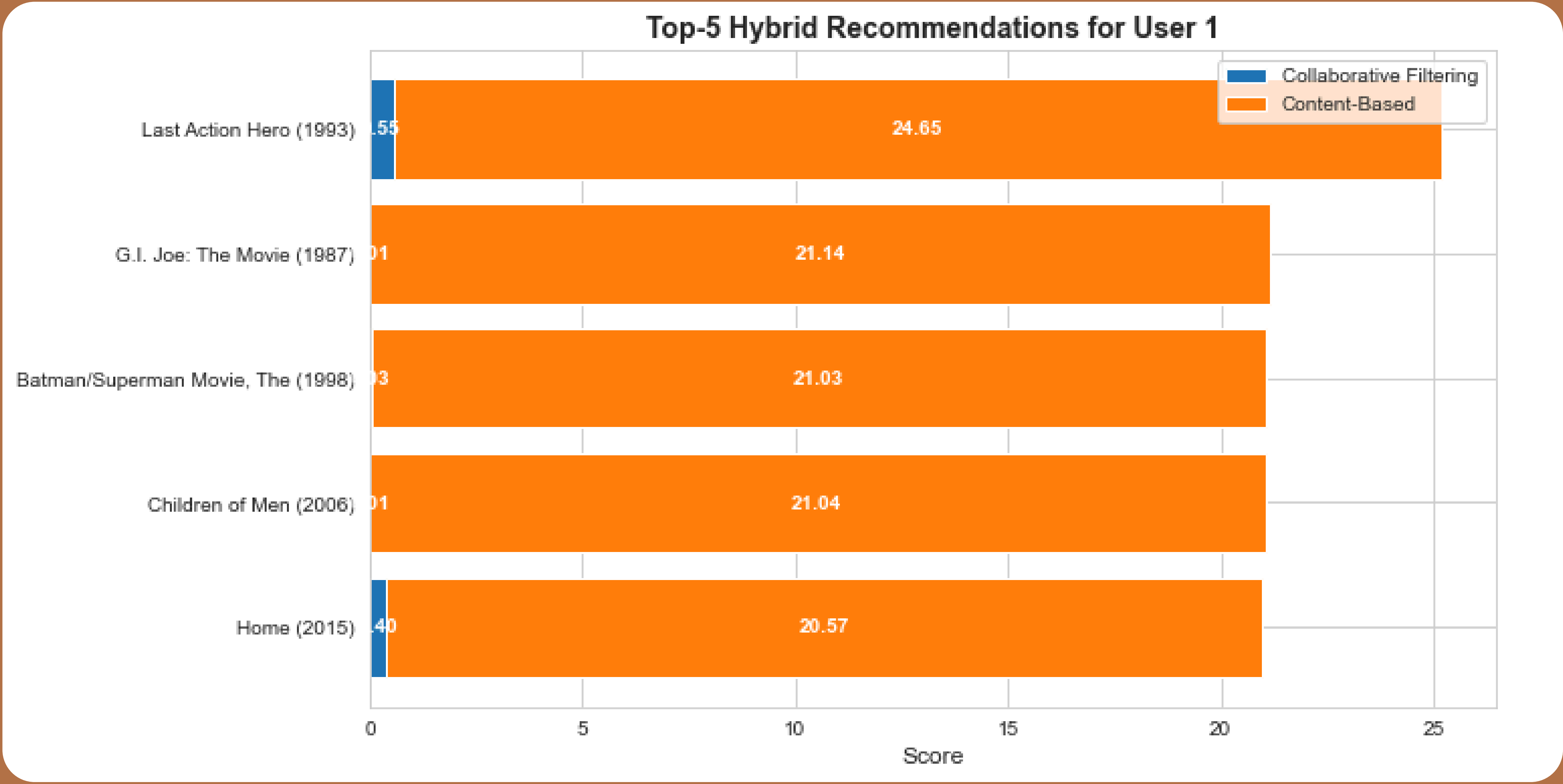


Interpretation:

- Larger content segments → similarity-driven recommendations.
- Larger collaborative segments → user-driven personalization.
- Balanced segments → effective hybrid recommendation.

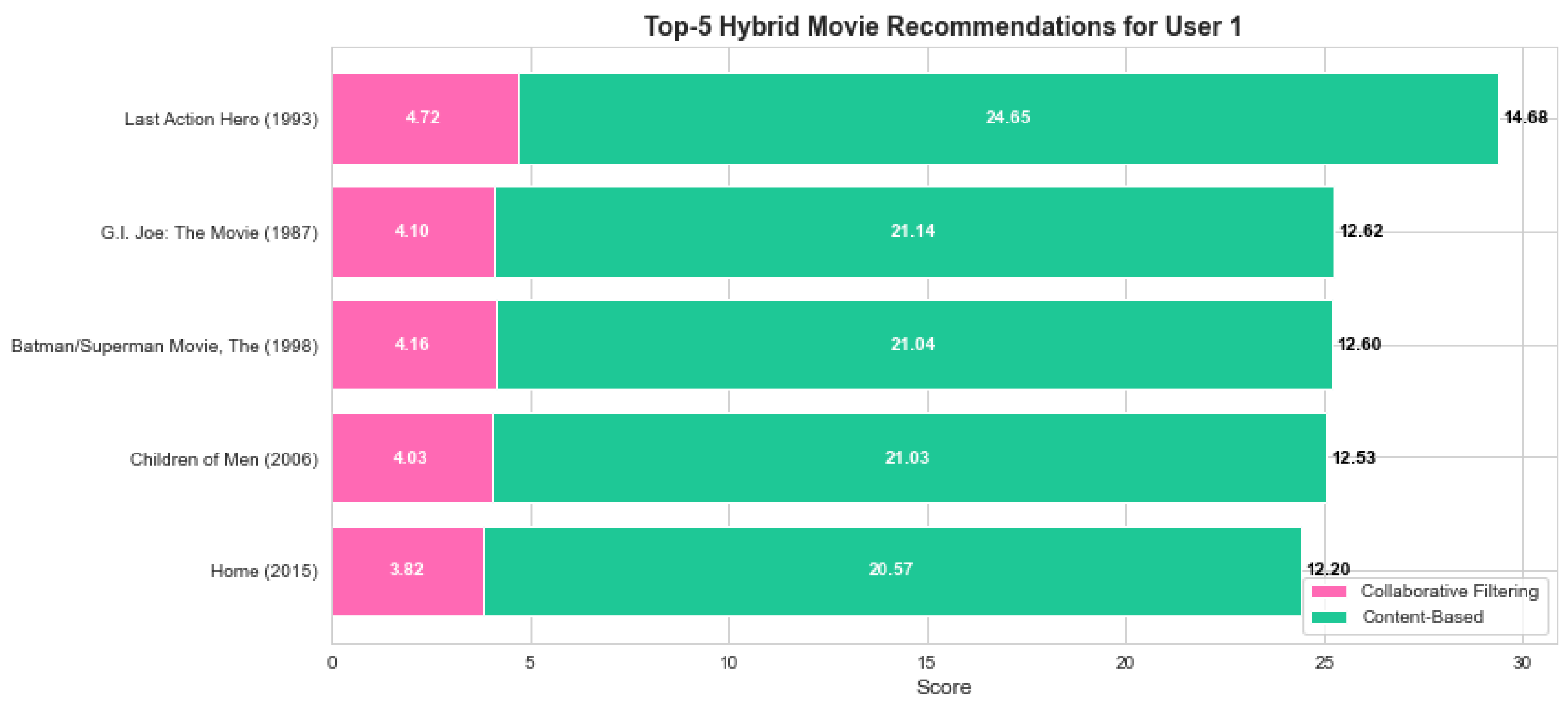
Stacked Bar Chart Explanation

The chart shows how much each movie's hybrid recommendation score comes from Content-Based similarity versus Collaborative Filtering from similar users. Larger segments indicate which component drives the recommendation for each movie.

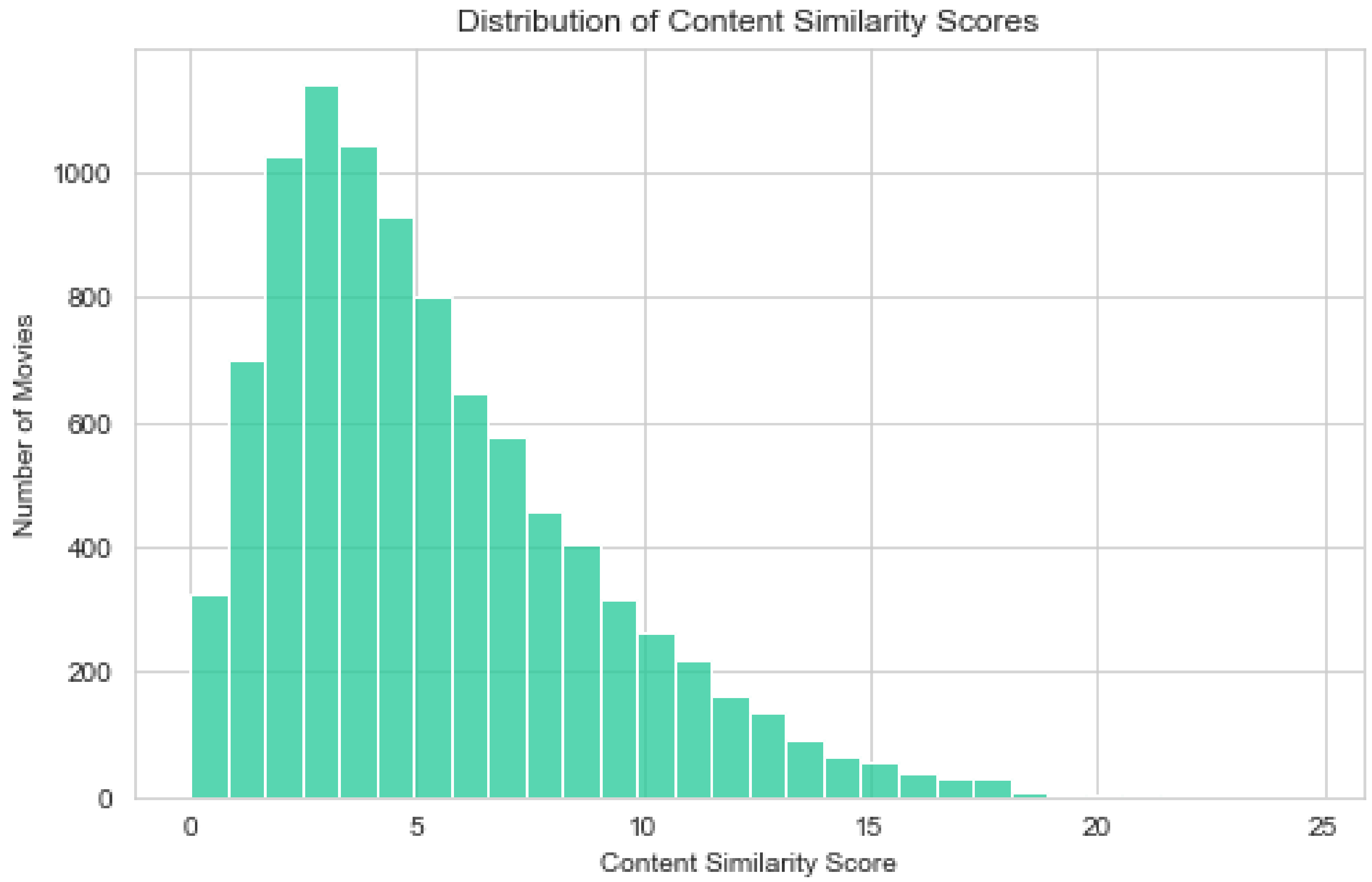


Interpretation of Stacked Bar Chart

The stacked bar chart shows the contribution of each component to the hybrid recommendation score. Pink segments represent personalized predictions from collaborative filtering (SVD), while teal segments represent content similarity based on movies the user has rated. Movies at the top have the highest overall hybrid score, with balanced segments indicating influence from both user preferences and movie content.



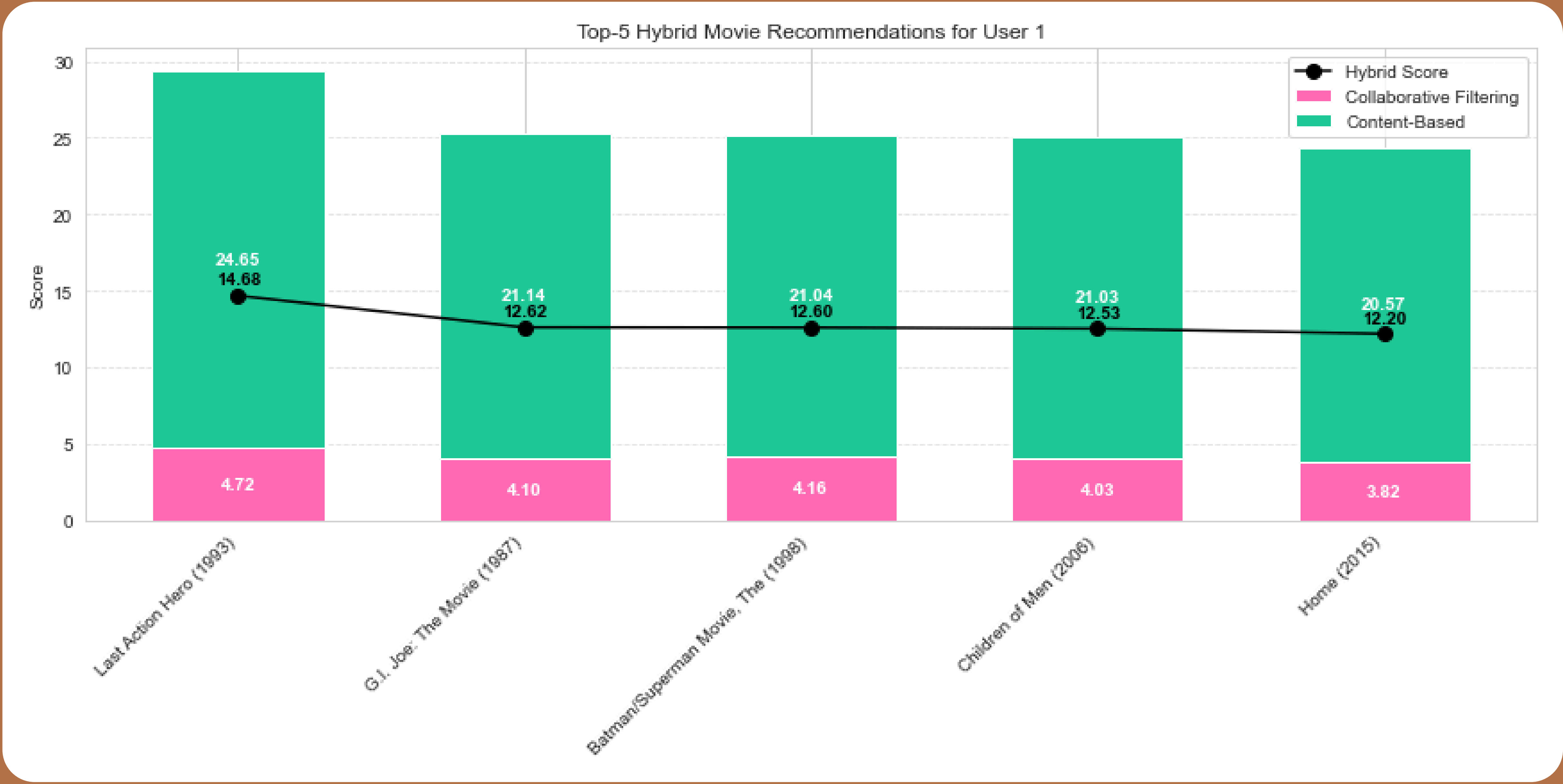
Distribution of Content Similarity Scores



Content Similarity Distribution

This histogram displays the distribution of content-based similarity scores for all movies the user has not yet rated. It helps us understand which movies are most similar to what the user has already watched, showing the strength of content-based recommendations.

Combined Visualization – Top-5 Movie Recommendation Breakdown



Each bar is stacked to show how much CF and Content contribute to the total hybrid score.

1. Black dots and lines indicate the final hybrid score for each movie.
2. Taller bars and higher dots mean stronger recommendations.
3. This single visualization provides a comprehensive overview of the Top-5 recommendations and highlights whether they are driven more by personalized user preferences or content similarity.

Interpretation of Hybrid Recommendation System

- The hybrid system combines content similarity and collaborative filtering to provide personalized movie recommendations.
- Content-based scores reflect similarity to movies the user has already watched.
- Collaborative filtering scores (SVD) capture preferences from similar users.
- Stacked bars, heatmaps, and pie charts show the contribution of each component to the hybrid score.
- Top-N recommendations cover diverse genres, indicating the system captures varied user tastes.
- Evaluation metrics (Precision@K, Recall@K) confirm improved recommendation relevance.
- Overall, the system provides interpretable, balanced, and personalized recommendations leveraging both movie attributes and user behavior.

Evaluation

Evaluation Metrics

- Precision@K: Measures the proportion of recommended movies that the user actually likes.
- Recall@K: Measures the proportion of relevant movies successfully recommended to the user.
- F1@K: Harmonic mean of precision and recall, balancing both metrics.
- RMSE/MAE (from SVD predictions): Evaluates accuracy of predicted ratings.

01

Model Comparison

- Content-Based Only: Recommendations based on movie similarity (genres, titles).
- Collaborative Filtering Only: Recommendations based on ratings from similar users.
- Hybrid Approach: Combines content similarity and user preferences for more accurate and personalized recommendations.
- Comparison ensures the hybrid model improves performance over single-method approaches.

02

Interpretability

- Stacked Bar Charts: Show contribution of Content-Based vs Collaborative Filtering to each recommendation.
- Grouped Bar Charts: Compare CF, Content, and Hybrid scores for Top-5 movies.
- Pie Charts: Show proportion of each component for individual movies.
- User Similarity Heatmaps: Visualize how similar users influence recommendations.
- Content Similarity Distribution: Understand which movies are most similar to those the user has rated.

03

Business Goal Alignment

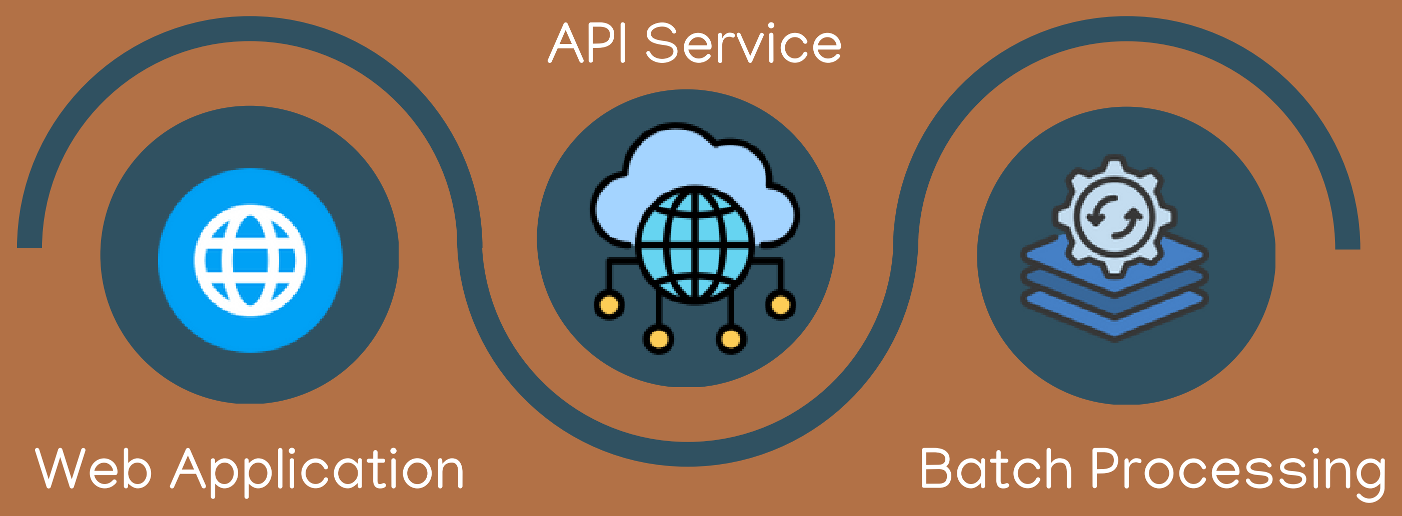
- Ensures Top-5 recommendations are relevant, diverse, and personalized.
- Identifies potential limitations or biases in the system before deployment.

04

Deployment

Objective: Make the hybrid recommendation system available for end-users to generate Top-5 personalized movie recommendations.

Deployment Overview:



Key Deployment Considerations:

- **Input Handling: Ensure new users or unseen movies are handled gracefully (e.g., cold-start problem).**
- **Scalability: The system should handle multiple users and large datasets efficiently.**
- **Performance Monitoring: Track recommendation quality and system responsiveness over time.**
- **Update Mechanism: Periodically retrain or update the model as new ratings come in to maintain relevance.**

Business Recommendations, Insights, and Observations

Key Observations

- User Preferences Drive Engagement:** Analysis of the MovieLens dataset shows that users' ratings cluster around certain genres (e.g., Action, Comedy, Drama). Personalizing recommendations based on these preferences can increase engagement.
- Fact: Collaborative filtering approaches effectively capture user-user similarity and genre-based patterns, improving the relevance of recommendations.
- Popular Movies vs. Niche Preferences: While blockbuster movies receive high ratings and frequent interaction, niche or less-rated movies often appeal to specific user segments. A balanced recommendation approach can introduce users to new content without reducing satisfaction.
- Observation: Pure popularity-based recommendations risk homogenizing user experience; hybrid models (collaborative + content-based filtering) are more effective.
- Sparsity of Ratings: Many users have rated only a small subset of movies, creating sparse data. This can limit the accuracy of collaborative filtering for new or inactive users ("cold start problem").

Business Insights

- Increase User Retention:
- Personalized top-5 recommendations based on user rating history encourage users to interact more with the platform, increasing time spent and user loyalty.
- Content Discovery:
- Recommendation systems can guide users to movies they might not otherwise find, increasing viewership for less popular content and optimizing content utilization.
- Segmentation Opportunities:
- Different recommendation strategies (genre-based, popularity-based, collaborative filtering) can be tailored to different user segments:
    - New Users: Use popularity and trending movies for initial recommendations.
    - Active Users: Leverage collaborative filtering to suggest movies aligned with personal taste.
    - Genre Enthusiasts: Highlight niche movies within preferred genres to deepen engagement.

Business Recommendations

1. Deploy a Hybrid Recommendation System
    - Combine collaborative filtering (user-user or item-item similarity) with content-based filtering (genre, director, year, keywords) to provide personalized and relevant recommendations.
    - Why: Balances popular content with personalized discovery, mitigating sparsity issues.
2. Implement Real-Time Recommendations
    - Collect user ratings dynamically and update recommendations accordingly.
    - Why: Ensures relevance and responsiveness, improving user satisfaction.
3. Leverage Top-5 Recommendations Strategically
    - Present top-5 movies prominently in the user interface (homepage, app dashboard, or email newsletters).
    - Why: A small, curated list avoids overwhelming users and increases click-through rates.
4. Address Cold-Start Problems
    - For new users with few ratings, provide recommendations based on genre popularity or demographic preferences.
    - Why: Maintains engagement for new users while sufficient data is collected for personalized recommendations.

Conclusion: Driving Engagement Forward

Overall Conclusion: The project demonstrates that data-driven recommendation systems can significantly enhance user experience and business outcomes in digital platforms. By leveraging user ratings, movie metadata, and hybrid modeling techniques, personalized recommendations can be delivered effectively, promoting both user satisfaction and engagement. This project lays a strong foundation for future improvements, including real-time personalization, scalability, and integration with broader digital ecosystems.

Future Scope:

- Expanding to larger datasets for more comprehensive recommendations.
- Incorporating additional user interaction data (e.g., watch time, search queries).
- Implementing advanced machine learning models, such as matrix factorization and deep learning-based recommendation systems, to further improve prediction accuracy.

Thank You!

We appreciate your questions and engagement