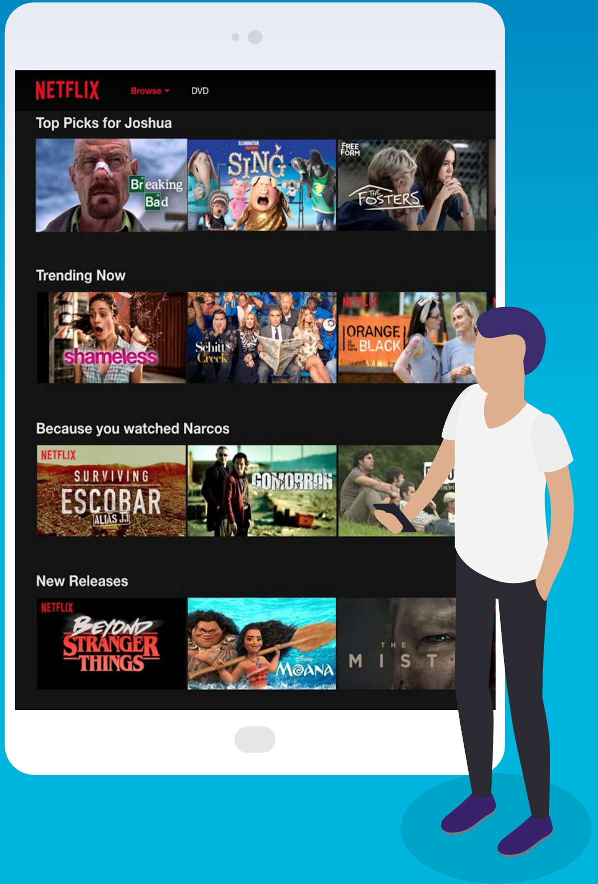
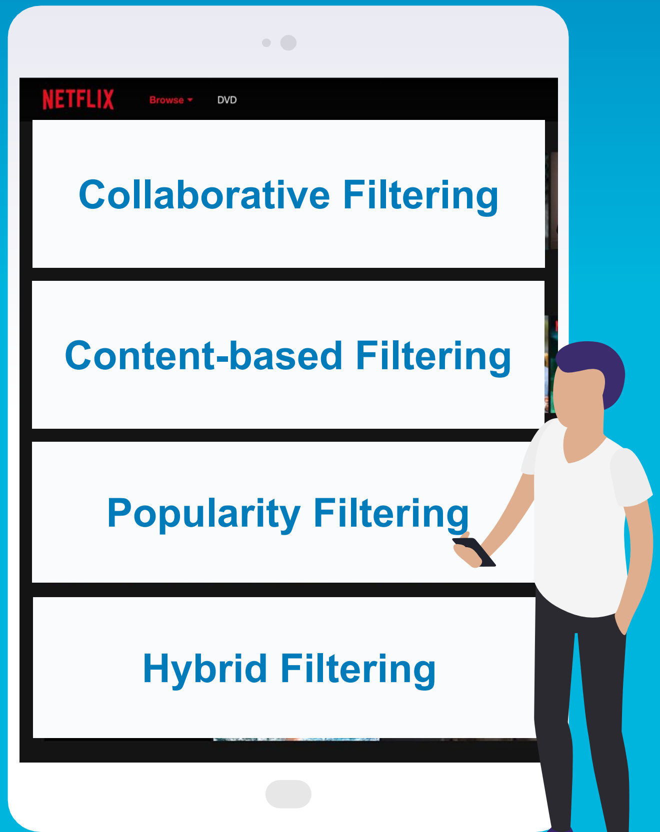
MovieLens Recommendation Systems

How do companies know what you like?



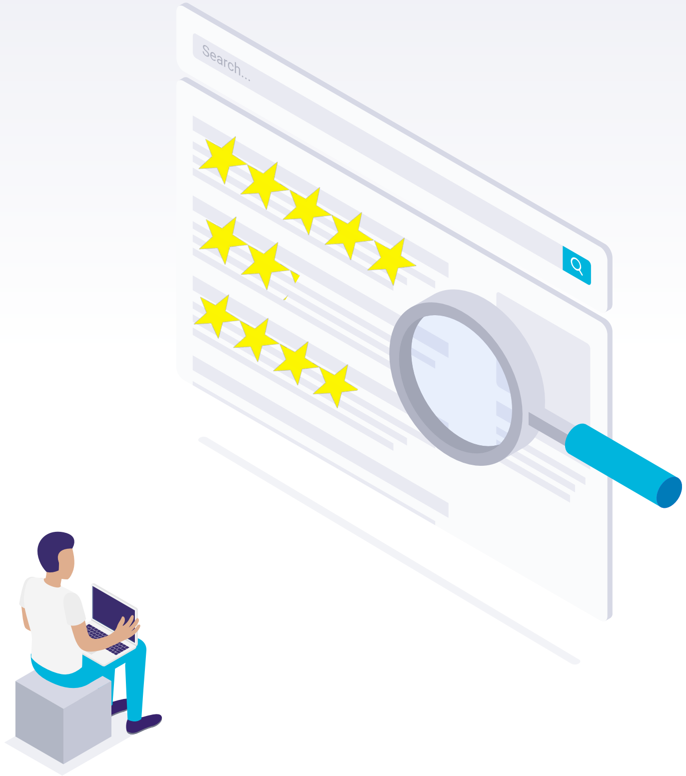
Recommendation Systems:



MovieLens Data

This dataset describes 5-star rating and free-text tagging activity from a movie recommendation service called MovieLens.

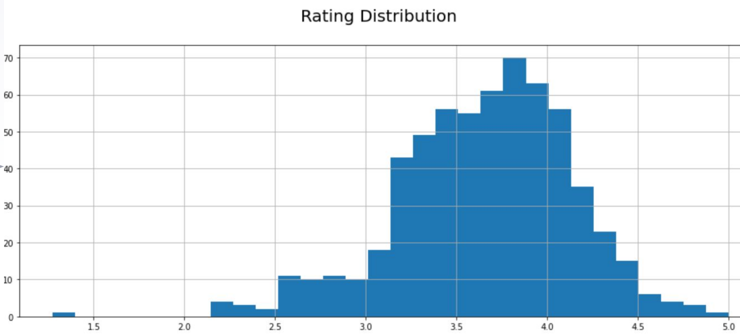
It contains **100836 ratings** across **58098 movies** based on **610 users**



F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS)
https://doi.org/10.1145/2827872

Files:

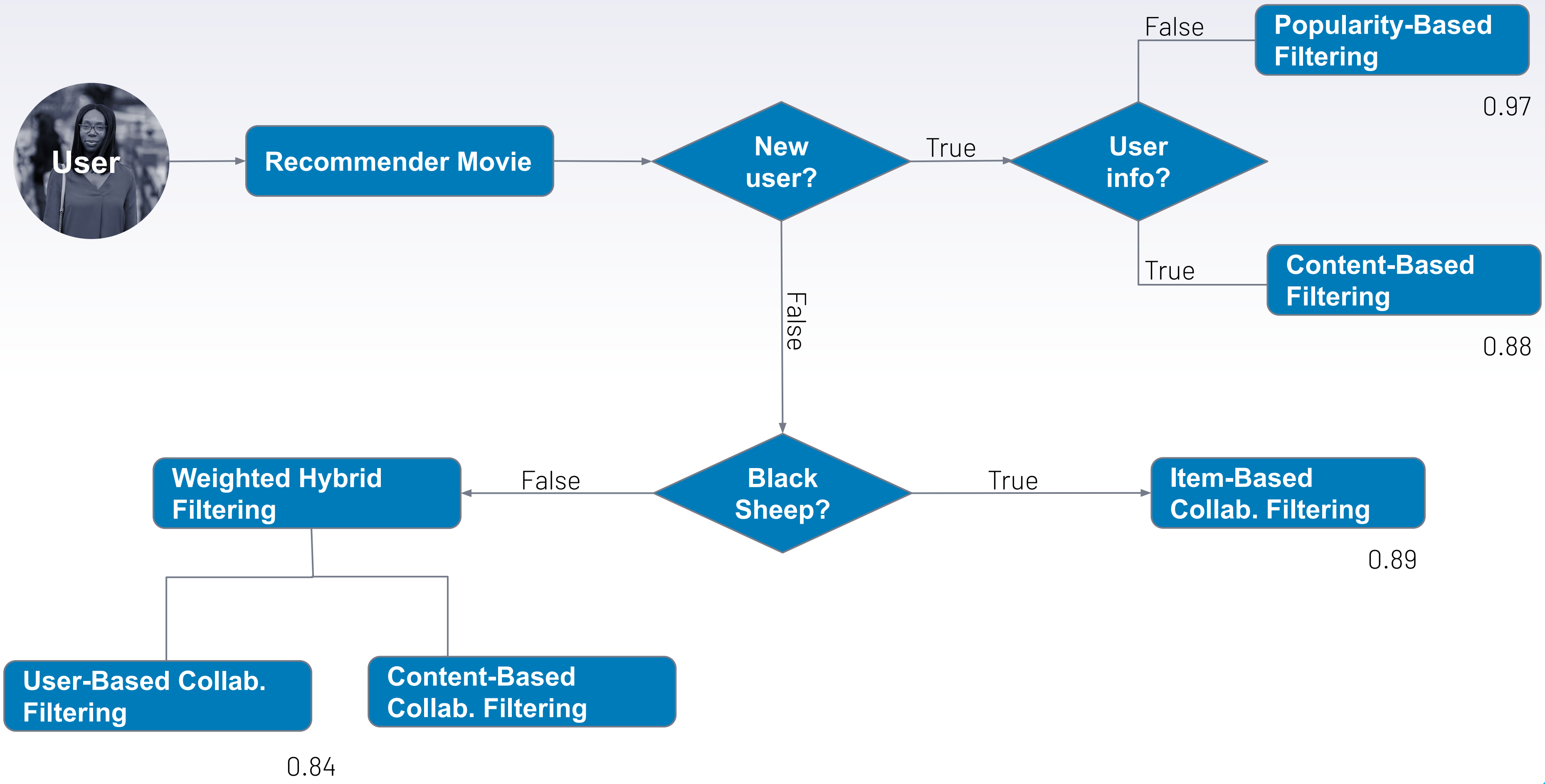
Ratings

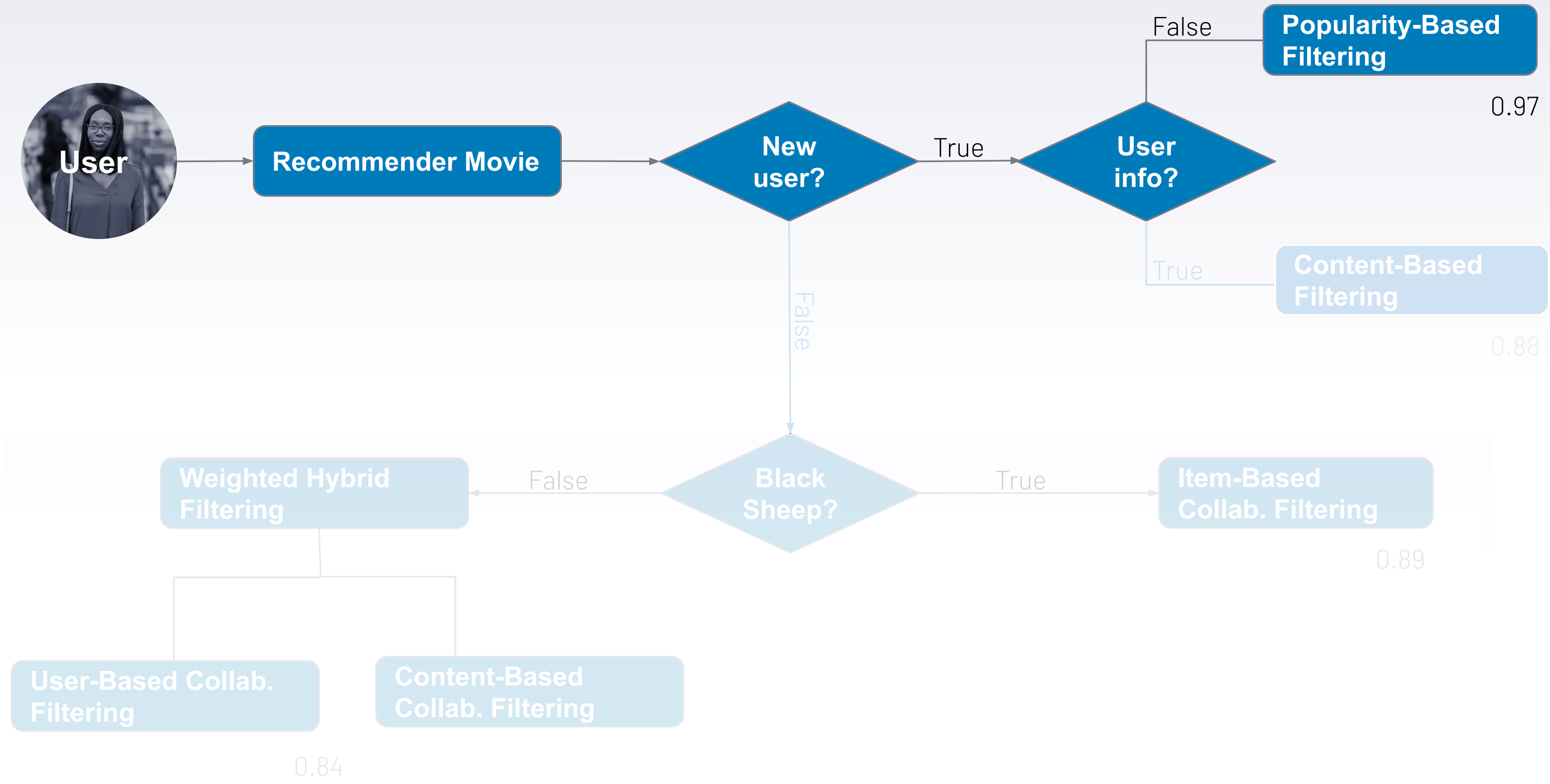


Movies

Users

Our Model


User
Recommender Movie
New user?
True
User info?
False
Popularity-Based Filtering
0.97
True
Content-Based Filtering
0.88
False
Black Sheep?
True
Item-Based Collab. Filtering
0.89
False
Weighted Hybrid Filtering
User-Based Collab. Filtering
Content-Based Collab. Filtering
0.84

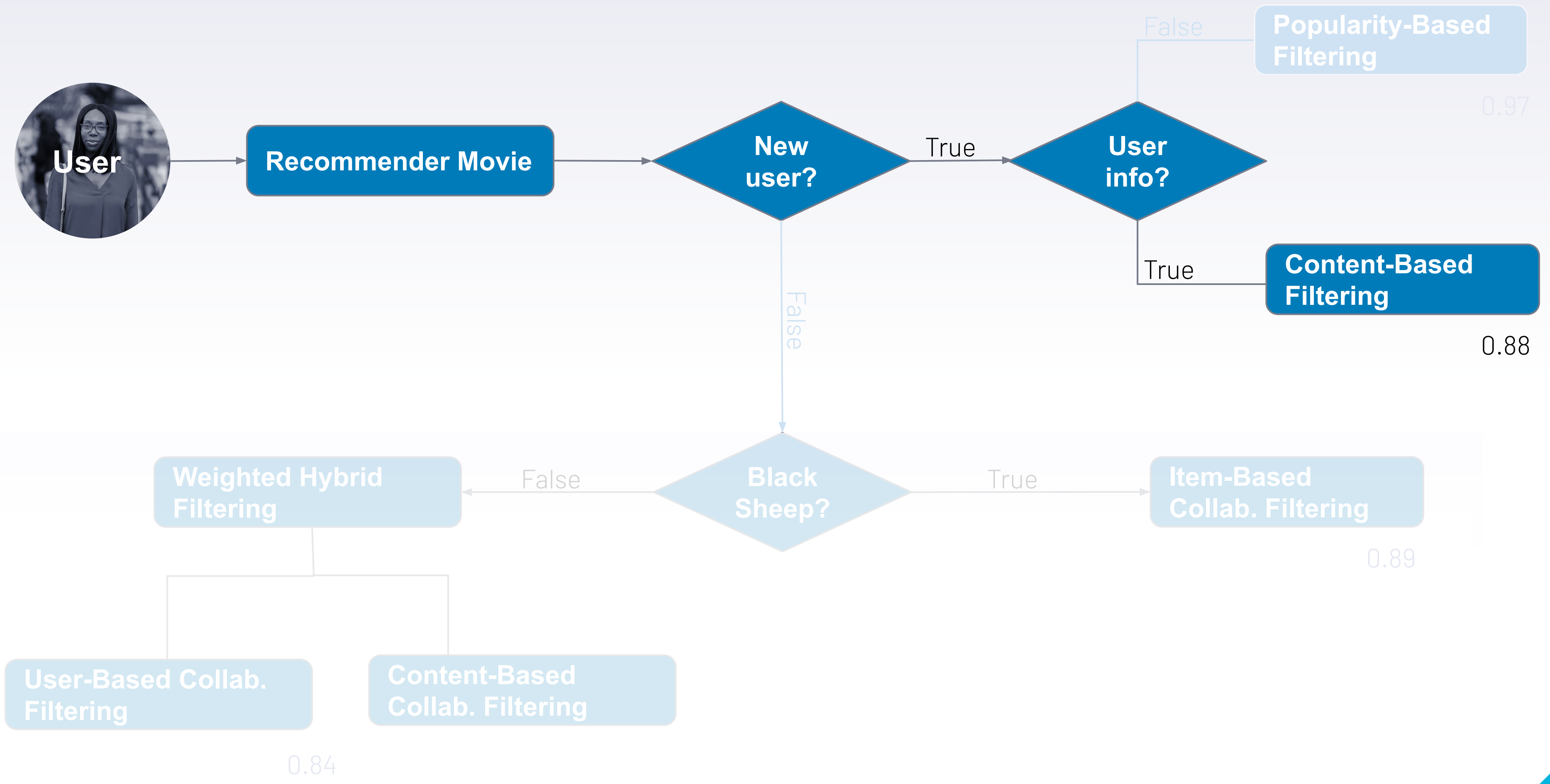

User
Recommender Movie
New user?
True
User info?
False
Popularity-Based Filtering
0.97
True
Content-Based Filtering
0.88
False
Black Sheep?
False
Weighted Hybrid Filtering
True
Item-Based Collab. Filtering
0.89
User-Based Collab. Filtering
Content-Based Collab. Filtering
0.84

Popularity Based Recommendation

- Non-personalized recommendations for when we have no data!
- We set **Item Popularity as a baseline** as it avoids creating niches. Users might want to be suggested movies that are popular even though they do not match their taste 100%

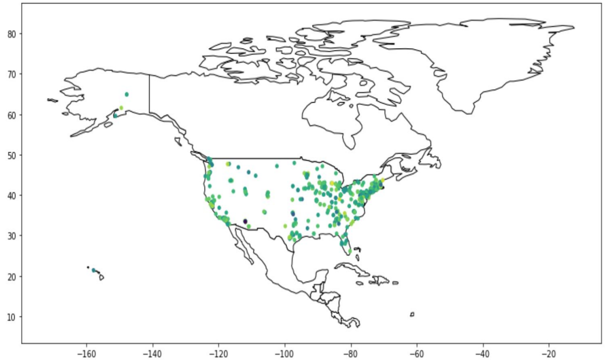
The winners of the Netflix Prize Competition used item popularity as a ranking function. As it enables to find a personalized ranking function rather than just item popularity, to satisfy taste of members with varying tastes.

https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-2-d9b96aa399f5


User
Recommender Movie
New user?
True
User info?
False
Popularity-Based Filtering
0.97
True
Content-Based Filtering
0.88
False
Black Sheep?
False
Weighted Hybrid Filtering
True
Item-Based Collab. Filtering
0.89
User-Based Collab. Filtering
Content-Based Collab. Filtering
0.84

Content-Based Preprocessing

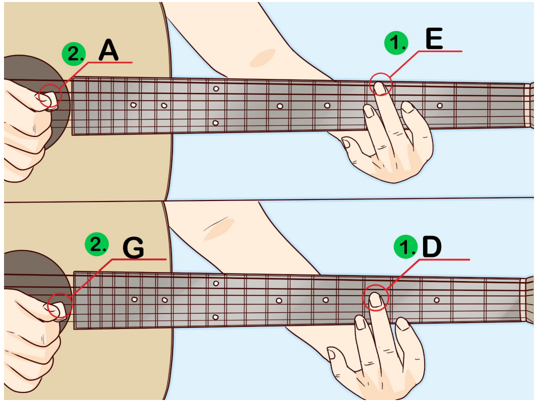
- **Ratings** - Flagged data issues to try out various models
  - Sampling
  - Guy in a bad mood
  - Transform Ratings
- **Movies** - Parse title, years, dummified genres,
- **User** - Demographics, mapped zip code to coordinates
- **User/Movies** - Low usage: 7.38% (Ohsoso troll 13%)



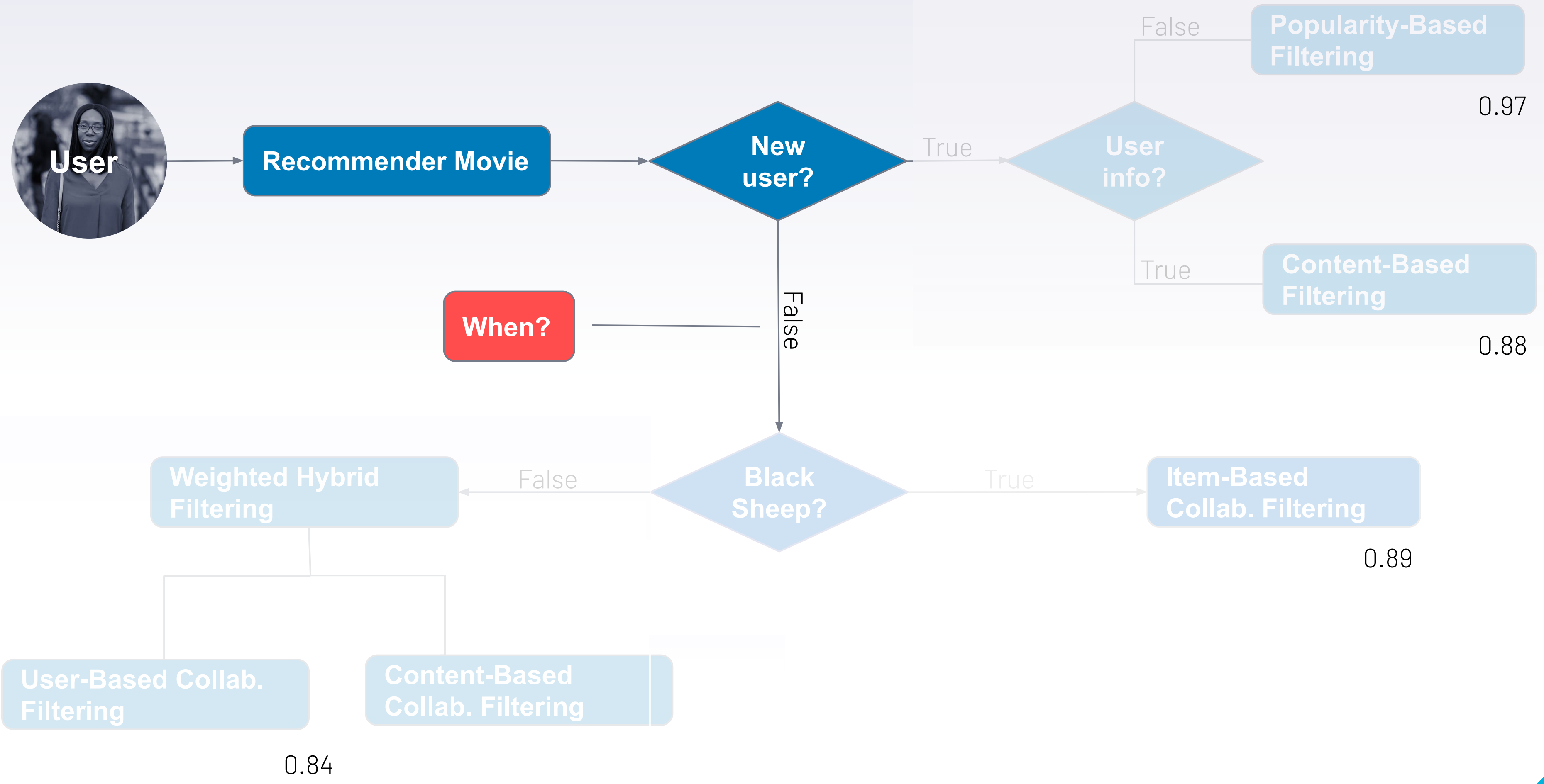
Next Steps: Web crawling to get more movie information

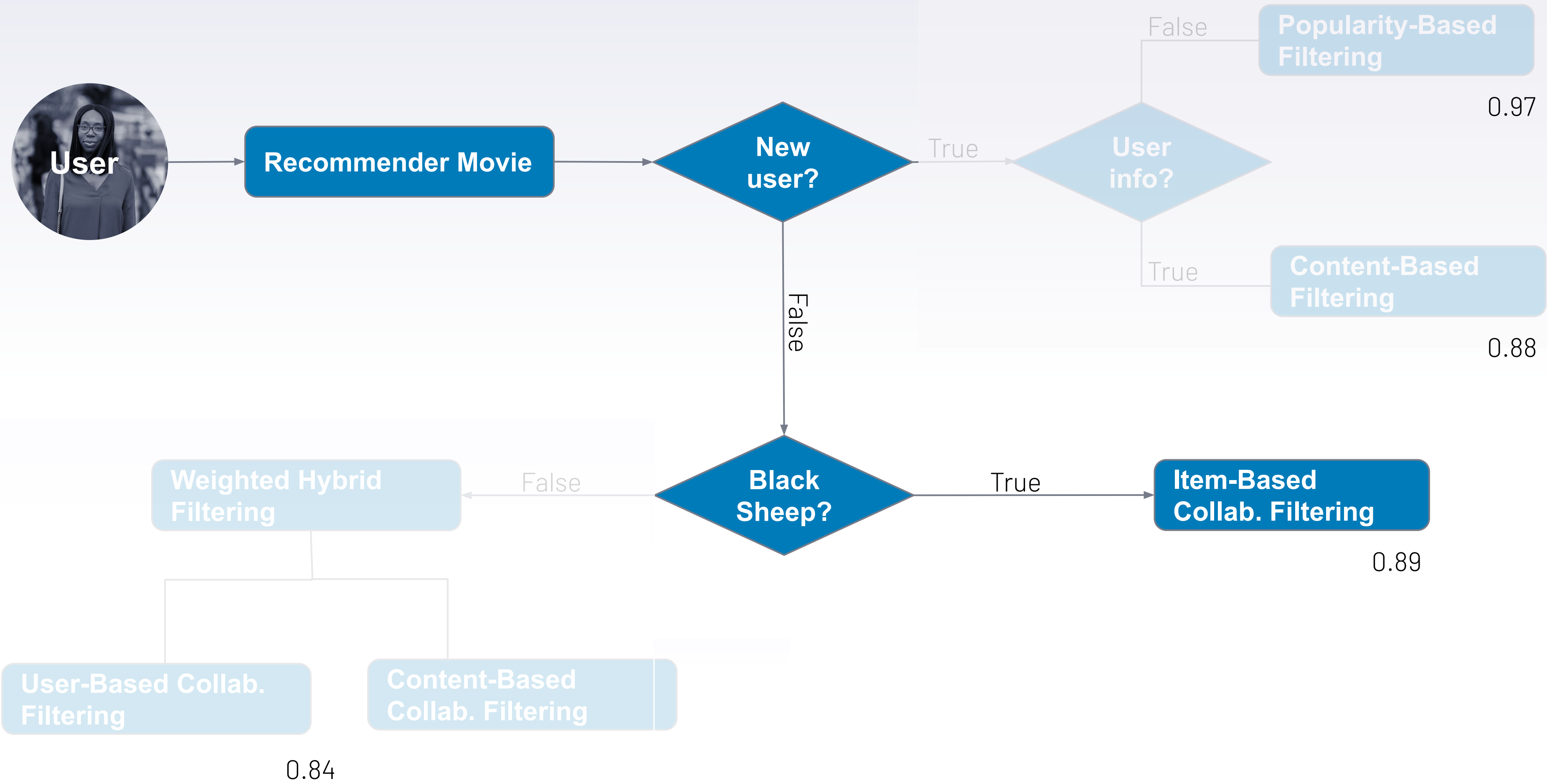
Content-Based Filtering

- **Lazy Predict:** Getting intuition of what works best
- **Preprocessing:** Scaling & Dummifying variables
- **Pipeline & Tuning:** Optimizing hyperparameters with 5-fold CV



Random Forest

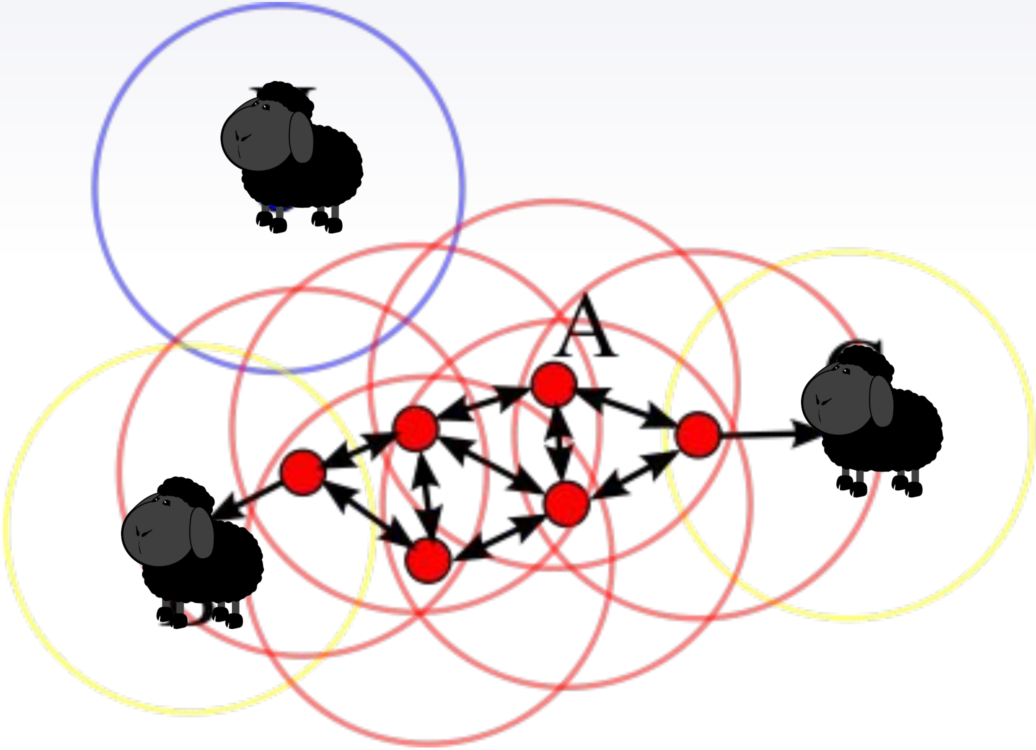

User
Recommender Movie
New user?
True
User info?
False
Popularity-Based Filtering
0.97
True
Content-Based Filtering
0.88
When?
False
Black Sheep?
False
Weighted Hybrid Filtering
True
Item-Based Collab. Filtering
0.89
User-Based Collab. Filtering
Content-Based Collab. Filtering
0.84

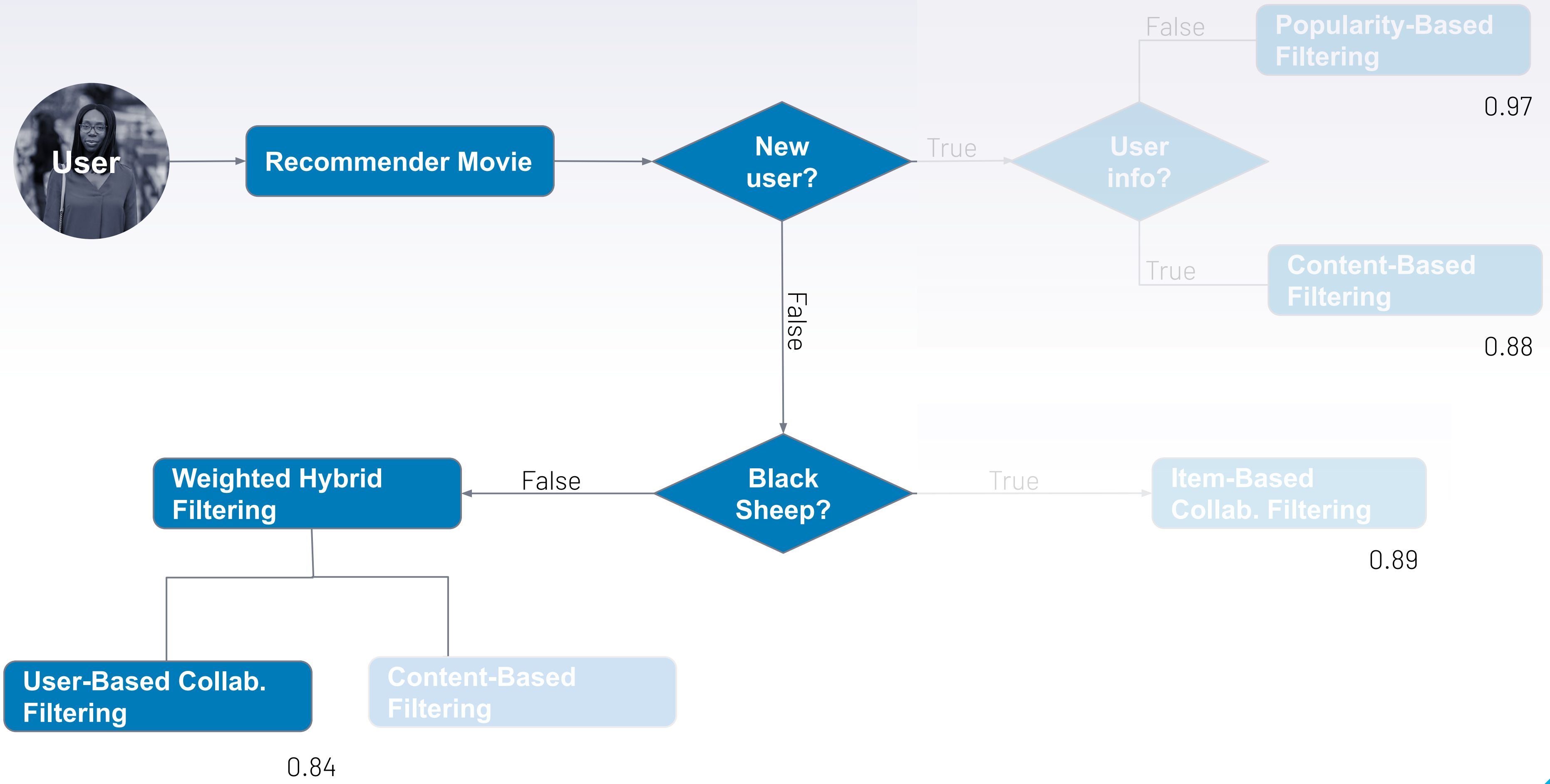

User
Recommender Movie
New user?
True
User info?
False
Popularity-Based Filtering
0.97
True
Content-Based Filtering
0.88
False
Black Sheep?
True
Item-Based Collab. Filtering
0.89
False
Weighted Hybrid Filtering
User-Based Collab. Filtering
Content-Based Collab. Filtering
0.84

Black Sheeps

- **DBSCAN with:**
  - User Demographics
  - Preferred Genre
- **Item-Based KNN:**

Recommending 'weird' people similar things to what they like



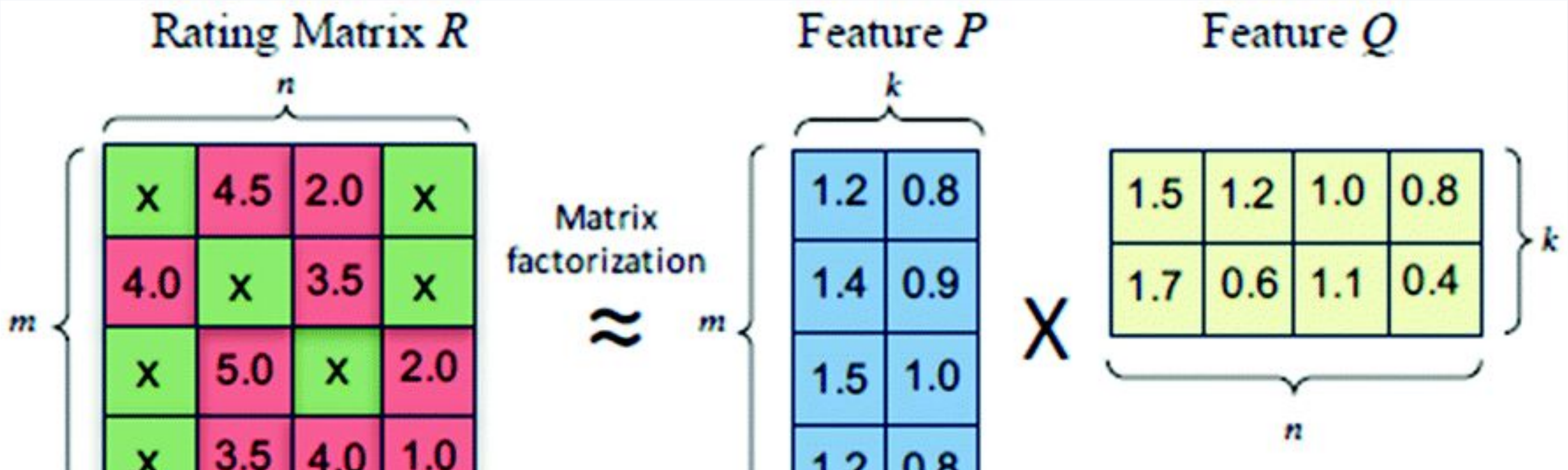

User
Recommender Movie
New user?
True
User info?
False
Popularity-Based Filtering
0.97
True
Content-Based Filtering
0.88
False
Black Sheep?
True
Item-Based Collab. Filtering
0.89
False
Weighted Hybrid Filtering
User-Based Collab. Filtering
0.84
Content-Based Filtering

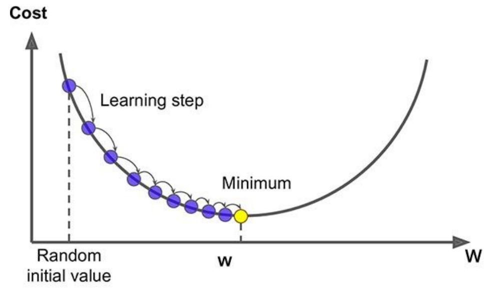
Collaborative Filtering

We implemented **Singular Value Decomposition (SVD)** model since it was giving us the lowest RMSE after hyperparameter tuning compared to other collaborative filtering models like KNN inspired ones.

	RMSE
NormalPredictor	1.42
BaselineOnly	0.87
SlopeOne	0.89
CoClustering	0.94
KNNBasic	0.93
KNNWithMeans	0.87
KNNWithZScore	0.87
KNNBaseline	0.86
SVD	0.85

https://data-flair.training/blogs/data-science-at-netflix/


Rating Matrix R
n
m
x 4.5 2.0 x
4.0 x 3.5 x
x 5.0 x 2.0
x 3.5 4.0 1.0
Matrix factorization
≈
Feature P
k
m
1.2 0.8
1.4 0.9
1.5 1.0
1.2 0.8
X
Feature Q
1.5 1.2 1.0 0.8
1.7 0.6 1.1 0.4
k
n
x Unknown samples
Observed samples


Cost
Learning step
Minimum
Random initial value
w
W

Hybrid Recommendation

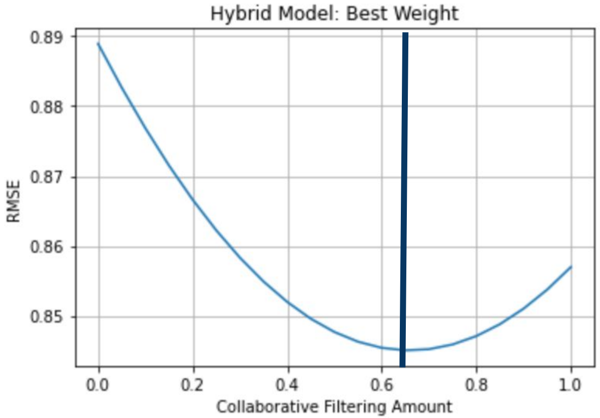
Collaborative Filtering (SVD)
65%

+

Content-Based (Random Forest)
35%

→

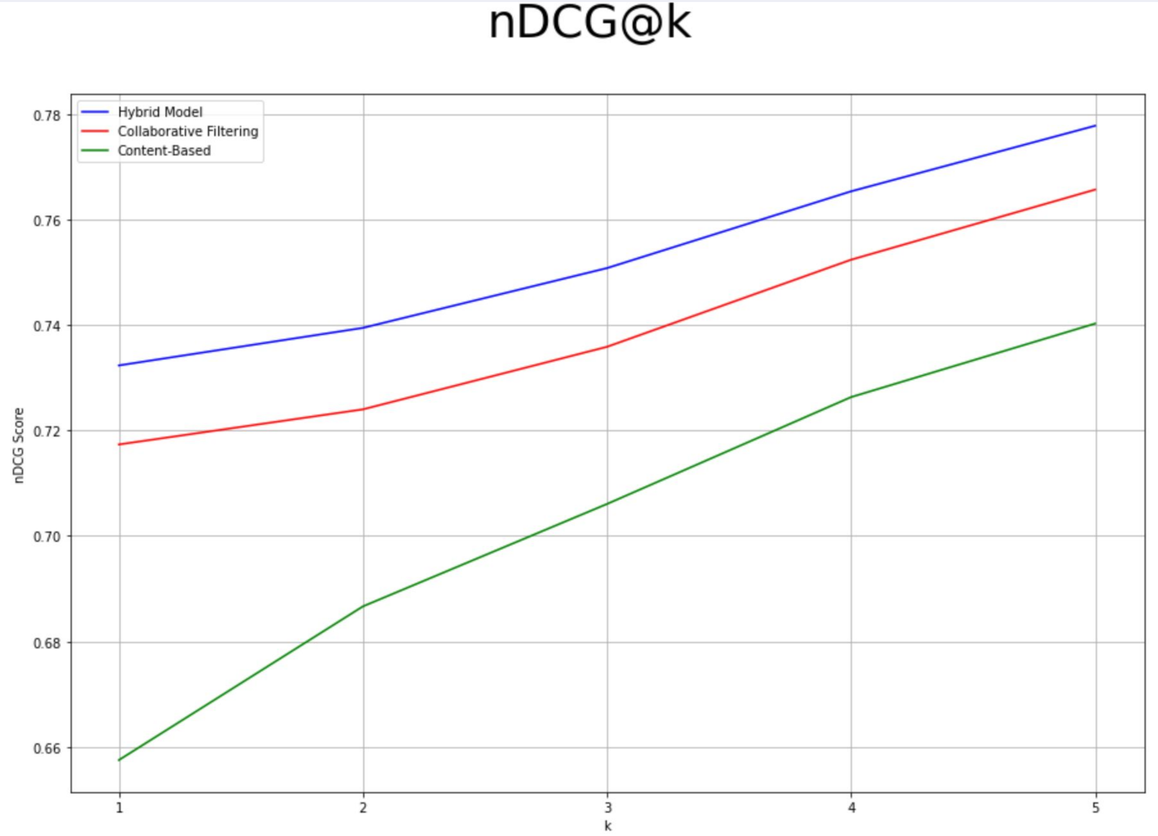
Weighted Hybrid Recommendation
RMSE: 0.84



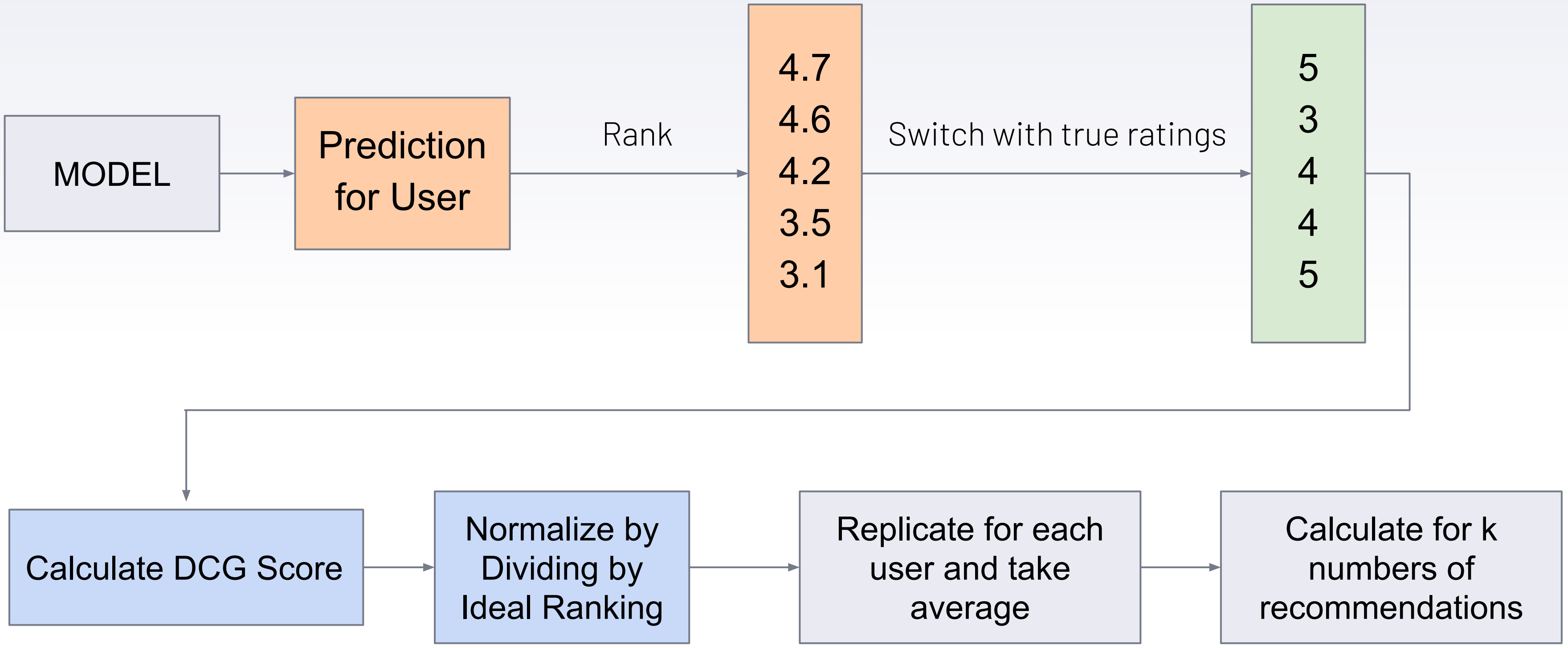
	RMSE
Collaborative Filtering	0.856962
Content-Based	0.888943
Hybrid Approach	0.845025

nDCG @ k

- Range: 0-1



nDCG @ k



$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Let's try it together!

Future Improvements

- Cold Start: resolve "new items" issue
- Find optimal minimal number of watched movies so that a user stops being new
- Let our model decide which users are black sheep!
- Combining Popularity Based and Content Based for diversity
- Linearly blend SVD and RBM to reduce error
- Display percentage of how much the user might enjoy a recommended movie

THANKS!

Any questions?