Recommender Systems

datasciencedojo
data science for everyone

Overview

- Introduction
    - Collaborative vs Content-based
- How do they work?
    - Ranking by similarity
    - Predicting
    - Evaluation
- Advantages/Disadvantages
- Example using Azure ML

INTRODUCTION

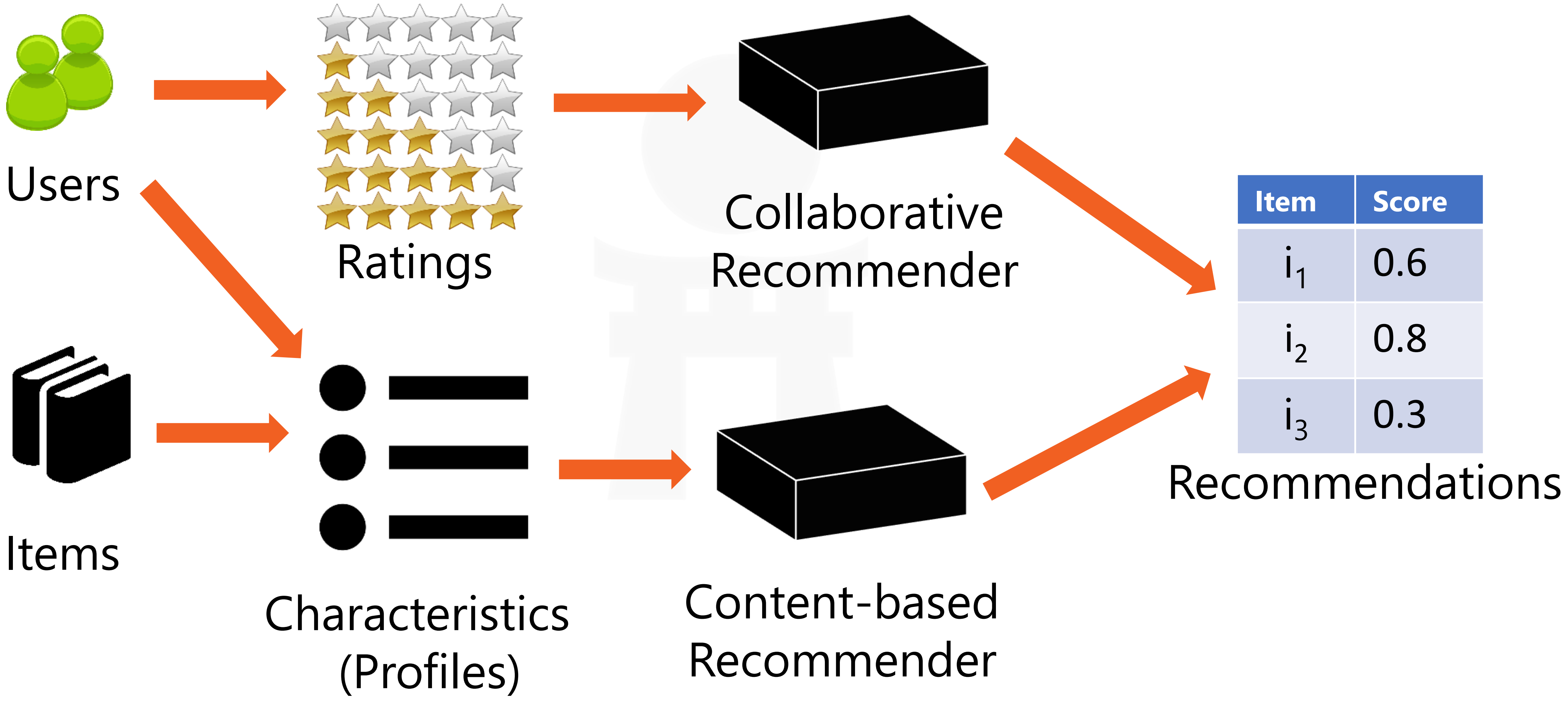
Recommendation Systems

- Automated systems to filter and recommend entities (products, ads, people) based on users' interest and taste.
- Designed to solve the information overload problem

Why recommendation systems?

- For Customers
  - Narrow down the set of choices
  - Discover new, interesting things
  - Save time
- For Business
  - Increase the number of items sold
  - Sell more diverse items
  - Better understand what the user wants

Collaborative vs. Content-based Recommenders



Item	Score
i_1	0.6
i_2	0.8
i_3	0.3

Collaborative vs. Content-based Recommenders

Collaborative

- 'Give me items that **people like me** enjoy'
- Users, Items, & Ratings

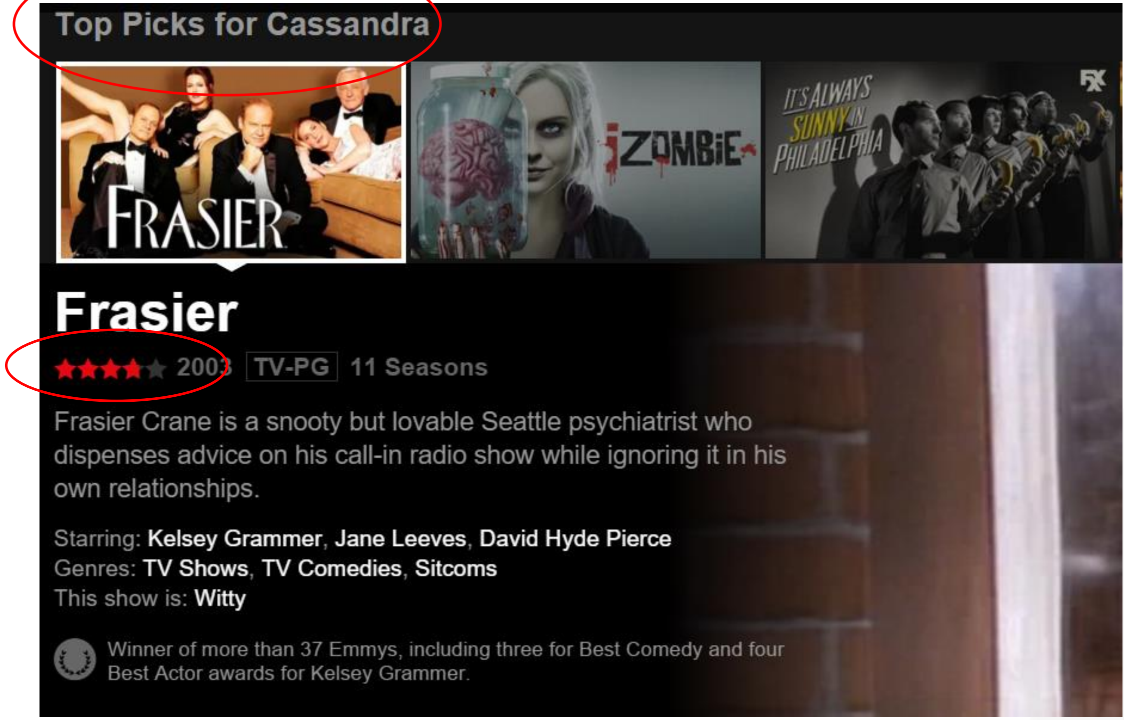
❿Use Ratings of similar Users to recommend unseen Items

Content-Based

- 'Give me items similar to **items I like**'
- User & Item profiles

❿Use overlap of User and Item characteristics to recommend unseen items

Example: Netflix



Top Picks for Cassandra

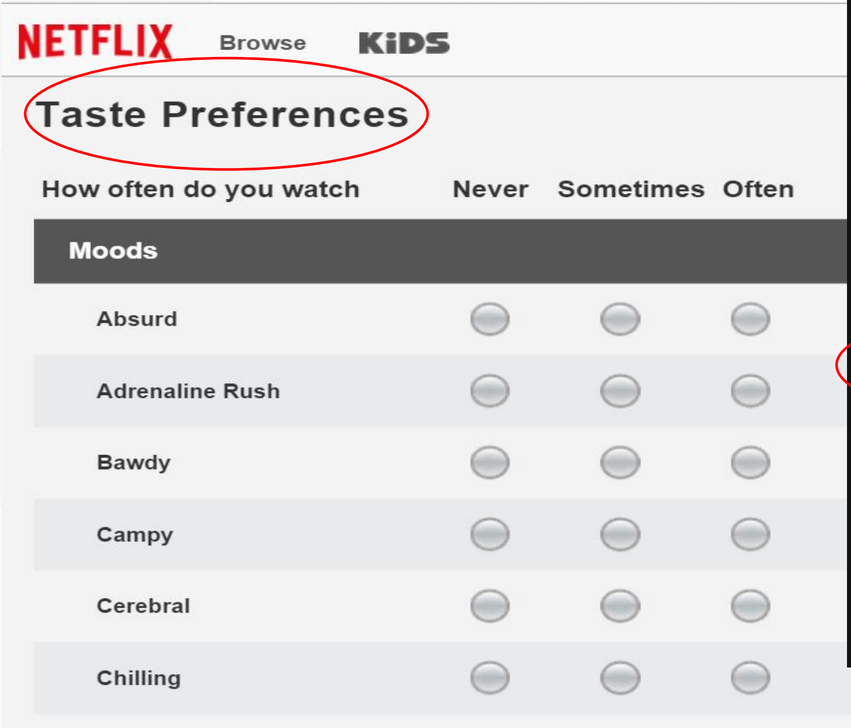
Frasier

★★★★★ 2003 TV-PG 11 Seasons

Frasier Crane is a snooty but lovable Seattle psychiatrist who dispenses advice on his call-in radio show while ignoring it in his own relationships.

Starring: Kelsey Grammer, Jane Leeves, David Hyde Pierce
Genres: TV Shows, TV Comedies, Sitcoms
This show is: Witty

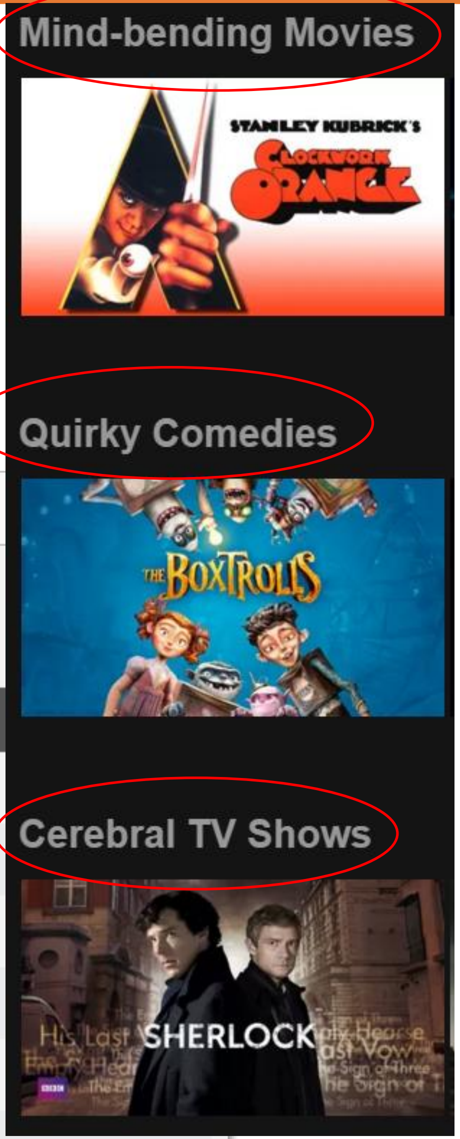
Winner of more than 37 Emmys, including three for Best Comedy and four Best Actor awards for Kelsey Grammer.



NETFLIX Browse KiDS

Taste Preferences

How often do you watch	Never	Sometimes	Often
Moods			
Absurd	○	○	○
Adrenaline Rush	○	○	○
Bawdy	○	○	○
Campy	○	○	○
Cerebral	○	○	○
Chilling	○	○	○

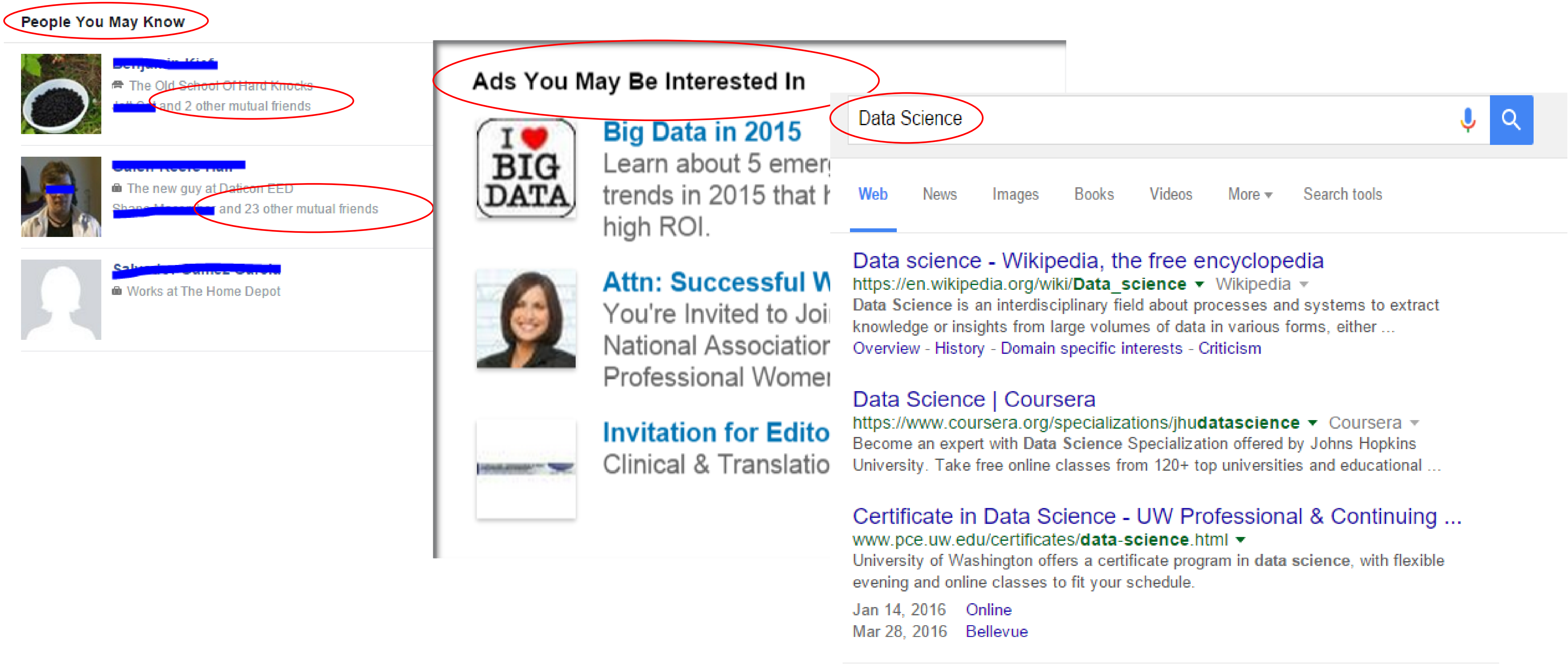


Mind-bending Movies

Quirky Comedies

Cerebral TV Shows

Example: Social Media & Search



People You May Know

[redacted]
The Old School Of Hard Knocks
[redacted] and 2 other mutual friends

[redacted]
The new guy at Daticon EED
[redacted] and 23 other mutual friends

[redacted]
Works at The Home Depot

Ads You May Be Interested In

Big Data in 2015
Learn about 5 emer
trends in 2015 that h
high ROI.

Attn: Successful W
You're Invited to Joi
National Associatior
Professional Womer

Invitation for Edito
Clinical & Translatio

Data Science

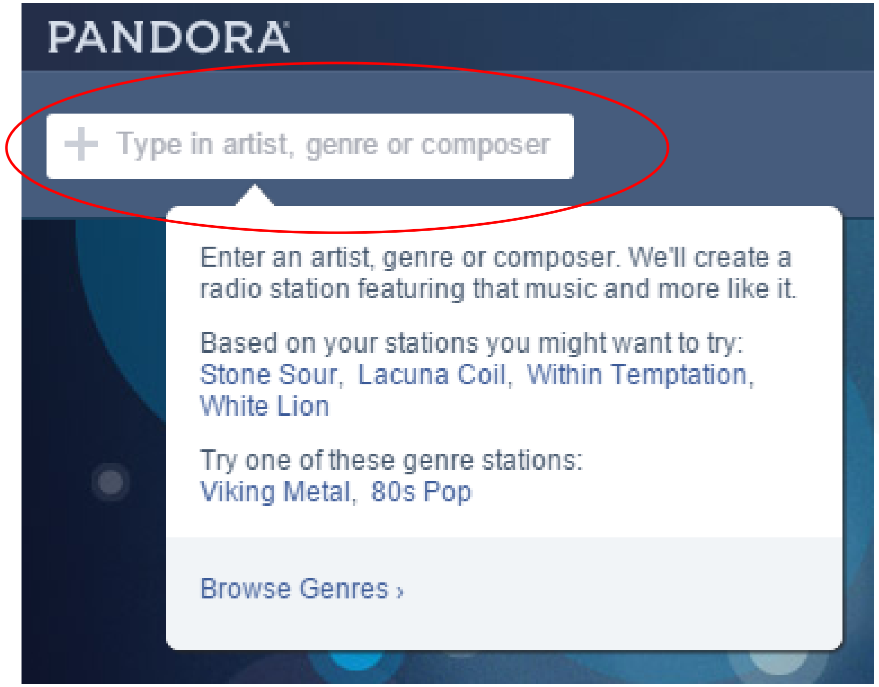
Web News Images Books Videos More ▾ Search tools

Data science - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Data_science** ▾ Wikipedia ▾
Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either ...
Overview - History - Domain specific interests - Criticism

Data Science | Coursera
https://www.coursera.org/specializations/jhu**datascience** ▾ Coursera ▾
Become an expert with **Data Science** Specialization offered by Johns Hopkins University. Take free online classes from 120+ top universities and educational ...

Certificate in Data Science - UW Professional & Continuing ...
www.pce.uw.edu/certificates/**data-science**.html ▾
University of Washington offers a certificate program in **data science**, with flexible evening and online classes to fit your schedule.
Jan 14, 2016 Online
Mar 28, 2016 Bellevue

Example: Pandora



PANDORA

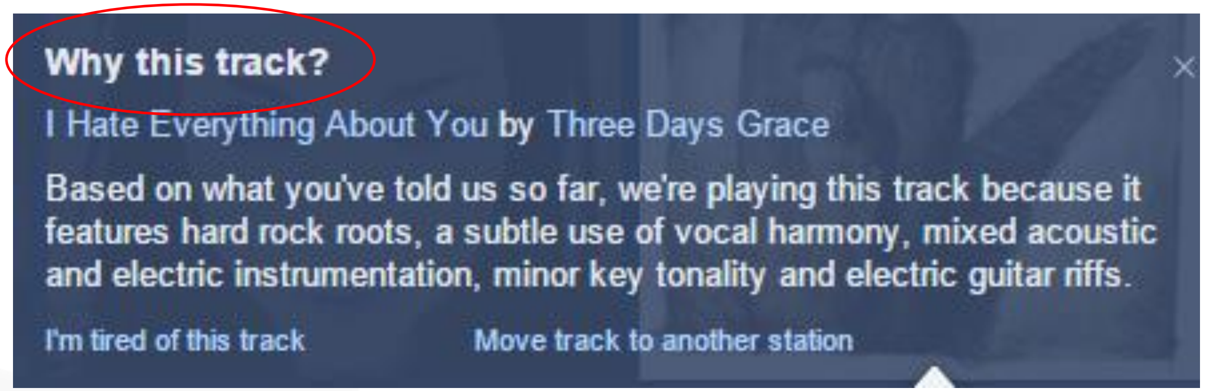
+ Type in artist, genre or composer

Enter an artist, genre or composer. We'll create a radio station featuring that music and more like it.

Based on your stations you might want to try:
Stone Sour, Lacuna Coil, Within Temptation, White Lion

Try one of these genre stations:
Viking Metal, 80s Pop

Browse Genres ›

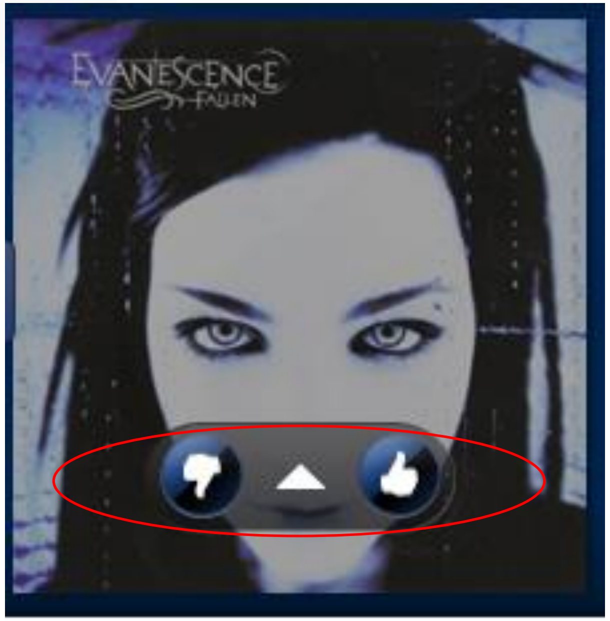


Why this track?

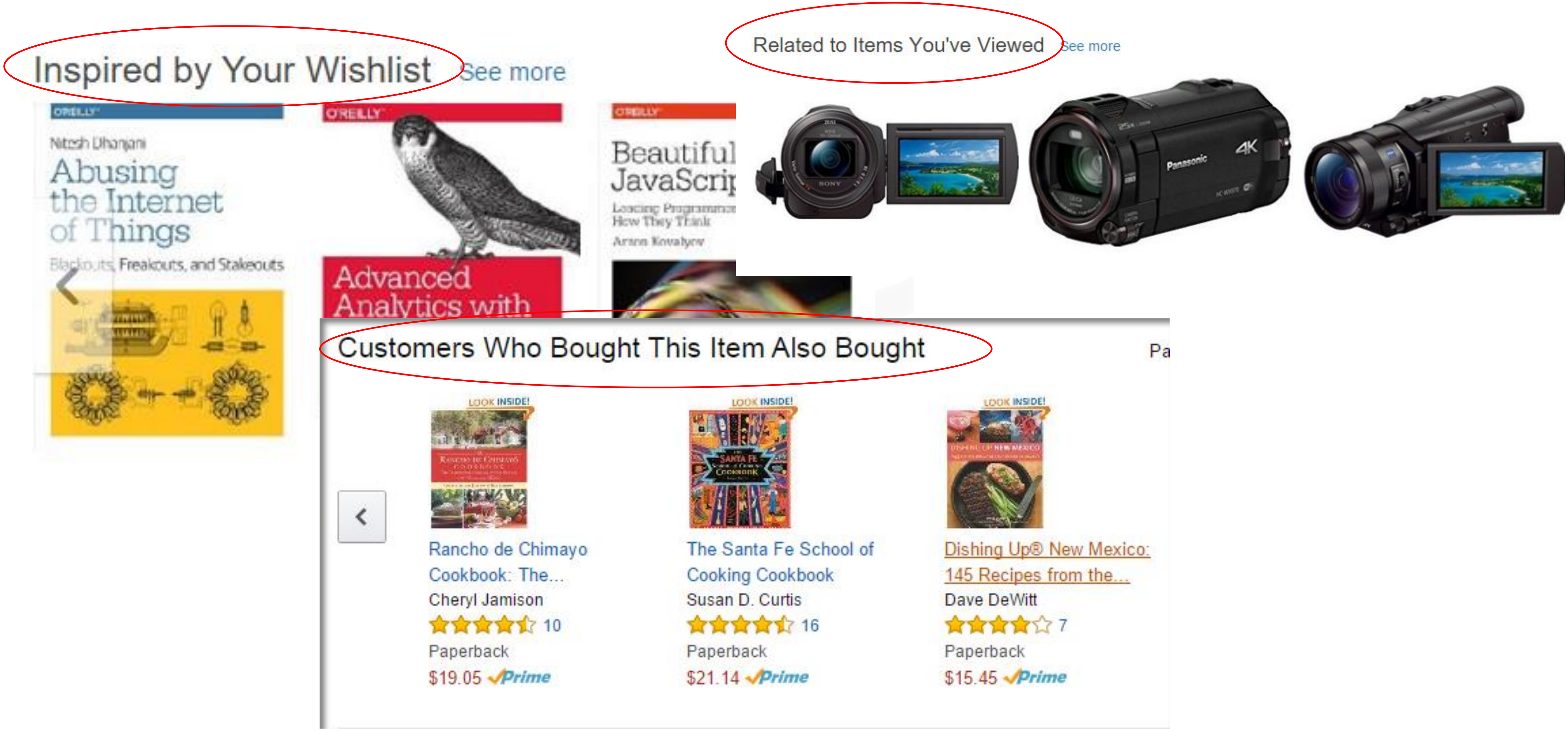
I Hate Everything About You by Three Days Grace

Based on what you've told us so far, we're playing this track because it features hard rock roots, a subtle use of vocal harmony, mixed acoustic and electric instrumentation, minor key tonality and electric guitar riffs.

I'm tired of this track   Move track to another station



Example: Amazon



Inspired by Your Wishlist See more

O'REILLY
Nitesh Dhanjani
Abusing the Internet of Things
Blackouts, Freakouts, and Stakeouts
O'REILLY
Advanced Analytics with
O'REILLY
Beautiful JavaScrip
Leading Programmers Explain How They Think
Anton Kovalyov

Related to Items You've Viewed See more

Customers Who Bought This Item Also Bought

Pa

Rancho de Chimayo Cookbook: The...
Cheryl Jamison
★★★★½ 10
Paperback
$19.05 ✓Prime

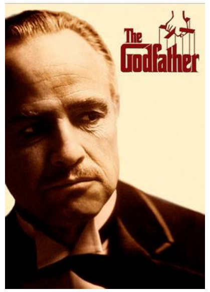
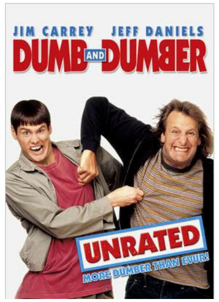
The Santa Fe School of Cooking Cookbook
Susan D. Curtis
★★★★½ 16
Paperback
$21.14 ✓Prime

Dishing Up® New Mexico: 145 Recipes from the...
Dave DeWitt
★★★★☆ 7
Paperback
$15.45 ✓Prime

Data Structure

- What kind of data?
  - Collaborative
    - Ratings of Items by Users
  - Content-based
    - Characteristic profiles of Users and Items

Data Structure – Collaborative

					
Alice	5	3	4	4	?
Bob	3	1	2	3	3
Chris	4	3	4	3	5
Donna	3	3	1	5	4
Evi	1	5	5	2	1

Data Structure – Content-based

Item/User	Drama?	Comedy?	Adventure?	Romance?
The Godfather	5	1	2	1
Titanic	4	3	2	5
Lord of the Rings	4	2	5	1
Dumb & Dumber	1	5	2	2
Spirited Away	5	3	5	2
Alice	5	4	1	4
Bob	3	1	1	1
Chris	4	2	5	2

Content-based: User Profiles

- **User Provided**
  - Ask for preferences
  - Needs accounts
  - Often low completion rates
- **Automated Generation**
  - Cookies follow behavior
  - No user persistence (often)

Content-based: Item Profiles

- **Expert Labeling**
  - Assign keywords based on content
  - May be provided by creators/distributors
  - Crowd sourcing?
- **Automated Indexing**
  - Used for text documents
  - Based on word content of document set
  - No expert knowledge involved

SIMILARITY

Similarity Measurements

- Given two vectors $\vec{x}$ and $\vec{y}$ with n components each
  - Ratings of User x and User y
  - Ratings for Item x and Item y
  - Profiles of User x and Item y
- How similar are the Users/Items?

Similarity Measurements

- Pearson's Correlation

$$sim(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Cosine Similarity

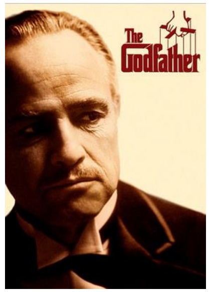
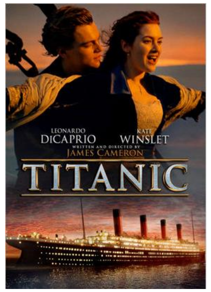
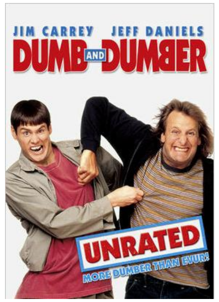
$$sim(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}| * |\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

User-Based Collaborative

- Goal: Predict User u's rating on a movie m they haven't seen
  - Find the N most similar Users to u who have seen movie m
  - Use their ratings to predict u's rating for movie m

User-Based Collaborative

Which metric should we use?

					
Alice	5	3	4	4	?
Bob	3	1	2	3	3
Chris	4	3	4	3	5
Donna	3	3	1	5	4
Evi	1	5	5	2	1

sim = ?

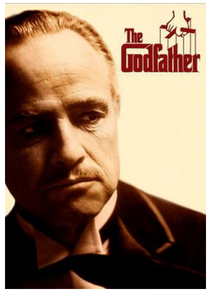
sim = ?

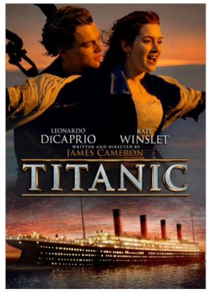
sim = ?

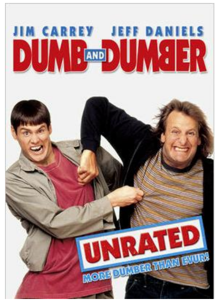
sim = ?

User-Based Collaborative

Pearson's correlation corrects for varied baselines





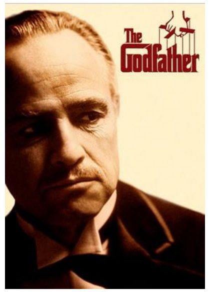


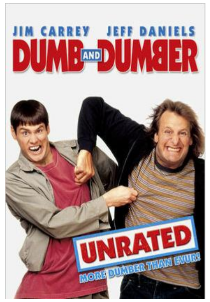
Alice	5	3	4	4	?	
Bob	3	1	2	3	3	sim=0.85
Chris	4	3	4	3	5	sim=0.90
Donna	3	3	1	5	4	sim=0.70
Evi	1	5	5	2	1	sim=0.79

Content-based: Similarity

- Goal: Return a recommendation list of items for each user
  - Find similarity of each User to each Item
  - Order Items by similarity

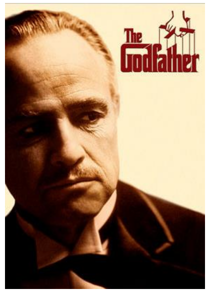
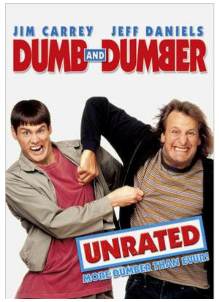
Content-based: Similarity





Item/User	Drama?	Comedy?	Adventure?	Romance?
The Godfather	5	1	2	1
Titanic	4	3	2	5
Lord of the Rings	4	2	5	1
Dumb & Dumber	1	5	2	2
Spirited Away	5	3	5	2
Alice	5	4	1	4
Bob	3	1	1	1
Chris	4	2	5	2

Content-based: Similarity

					
Alice	0.83	0.96	0.72	0.79	0.83
Bob	0.99	0.86	0.85	0.59	0.91
Chris	0.87	0.82	0.99	0.69	0.99

- Cosine similarity doesn't erase baselines
- Predicts order, not exact rating

PREDICTIONS

Collaborative: Predictions

- Use "Aggregation Function"
- Choose N nearest neighbors to User u
- Combine each neighbor j's rating on Item i $(r_{j,i})$
- Simple
  - $r_{u,i} = \frac{1}{N}\sum_{j=1}^{N} r_{j,i}$
- Weighted & Centered
  - $r_{u,i} = \overline{\boldsymbol{r}_u} + \alpha \sum_{j=1}^{N} sim(j,u)(r_{j,i} - \overline{r_j})$

Content-based: Predictions

- Simple
  - Rank in order of similarity
- Information retrieval techniques
  - Well studied, wide diversity of models
  - Classification algorithms

EVALUATION

Evaluating Recommendation

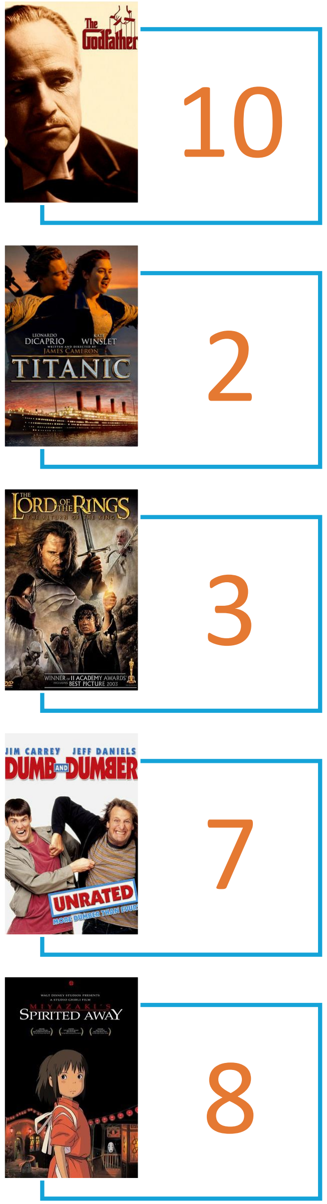
- **Mean Absolute Error (*MAE*)** computes the deviation between predicted ratings and actual ratings

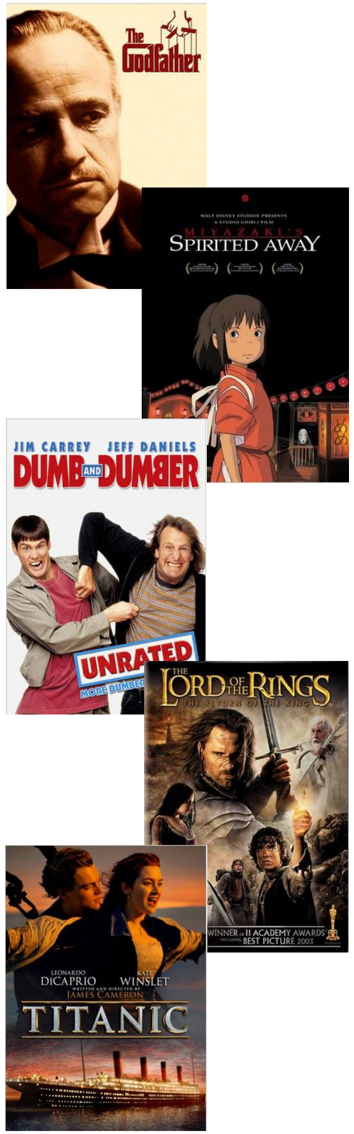
$$MAE = \frac{1}{n}\sum_{i=1}^{n} | p_i - r_i |$$

- **Root Mean Square Error (*RMSE*)** is similar to *MAE*, but places more emphasis on larger deviation

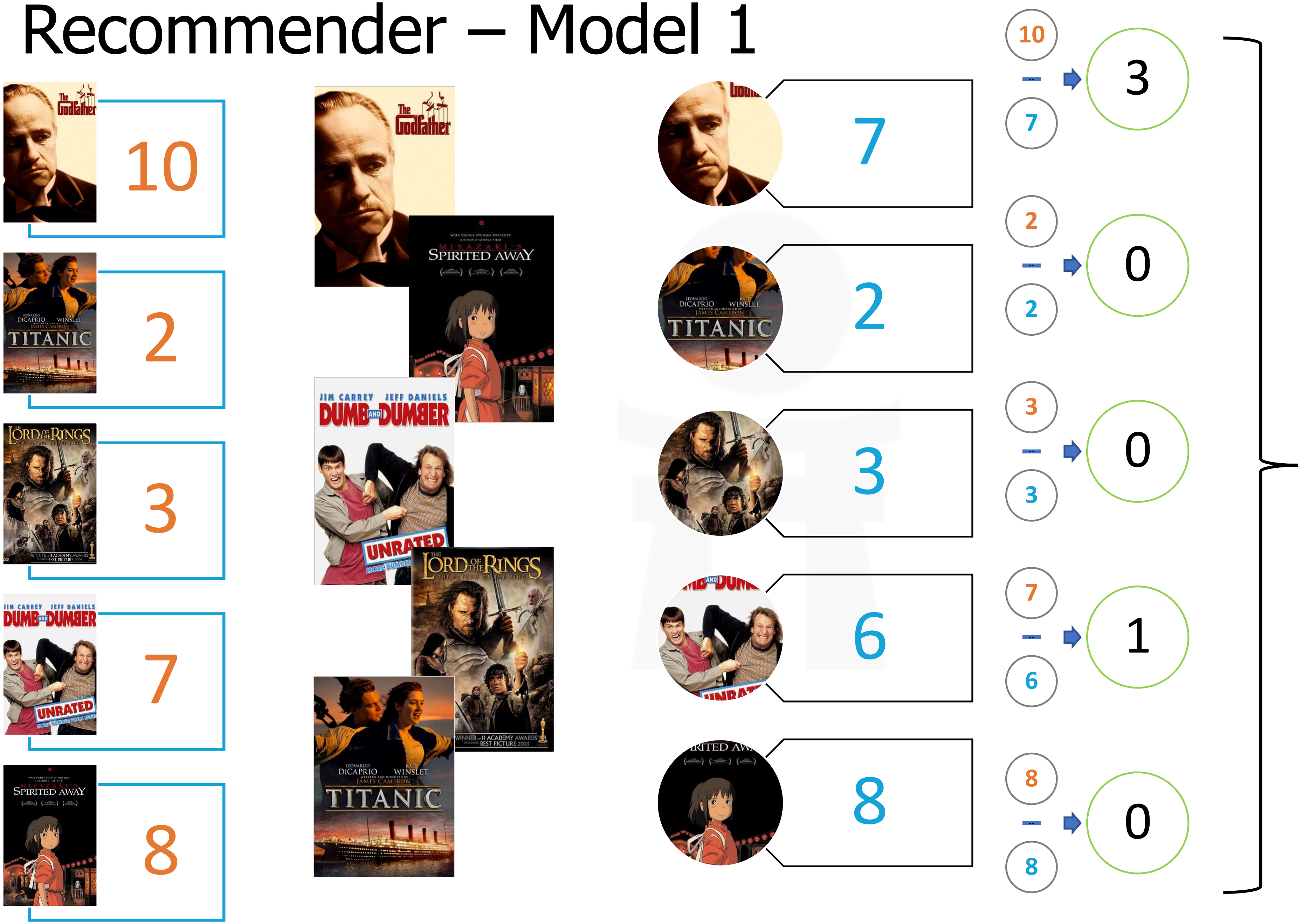
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2}$$

Evaluating a Ranker





Recommender – Model 1



10
2
3
7
8

7
2
3
6
8

10 − 7 → 3

2 − 2 → 0

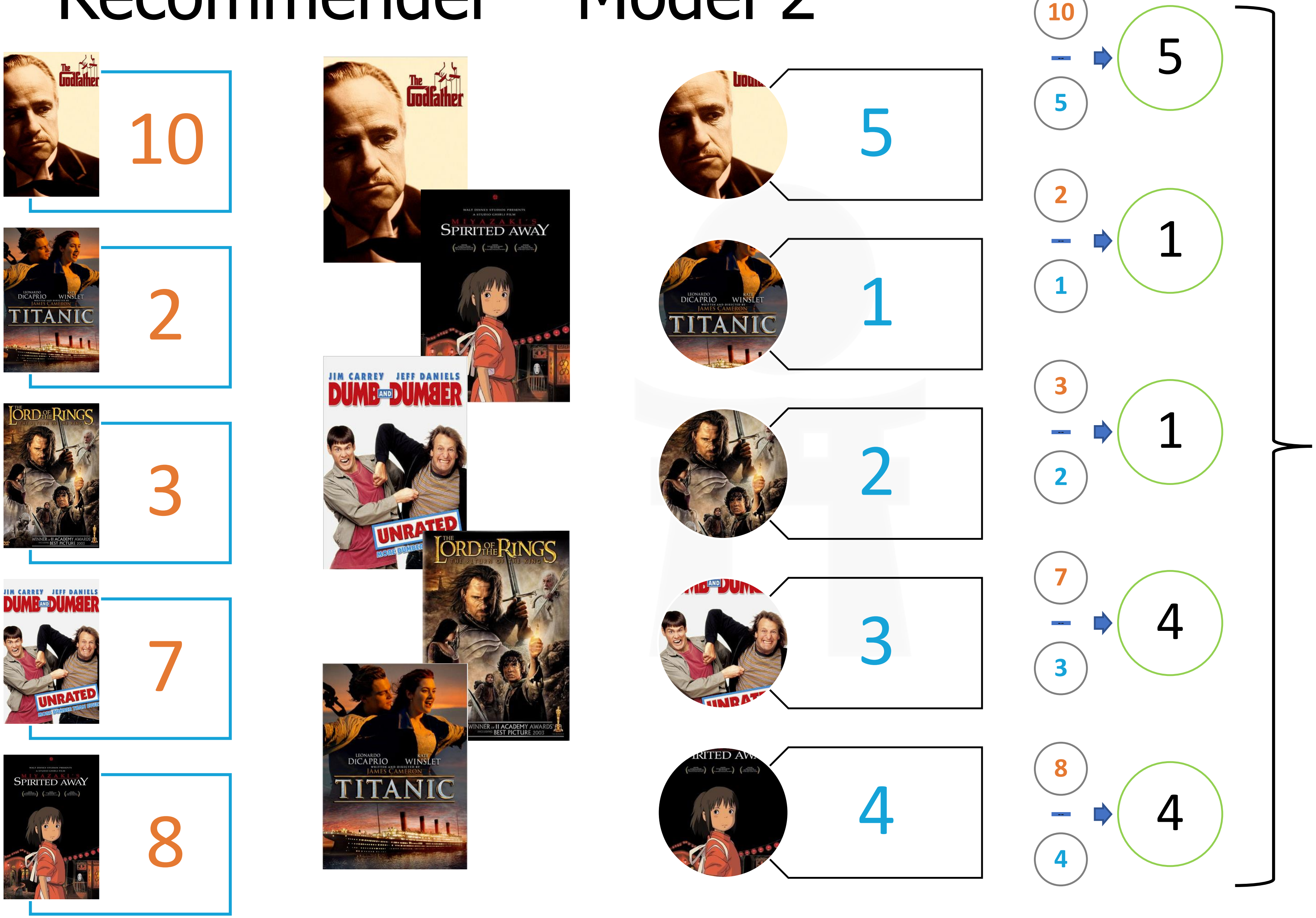
3 − 3 → 0

7 − 6 → 1

8 − 8 → 0

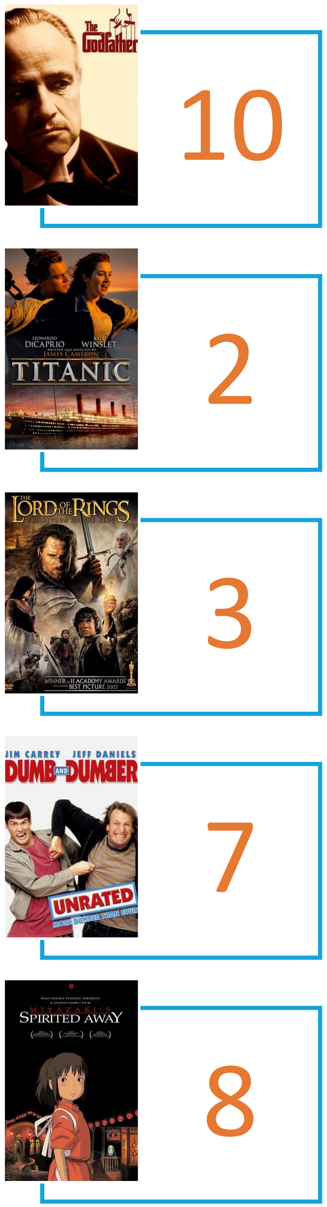
Total MAE =
4/5 = 0.8

Recommender – Model 2

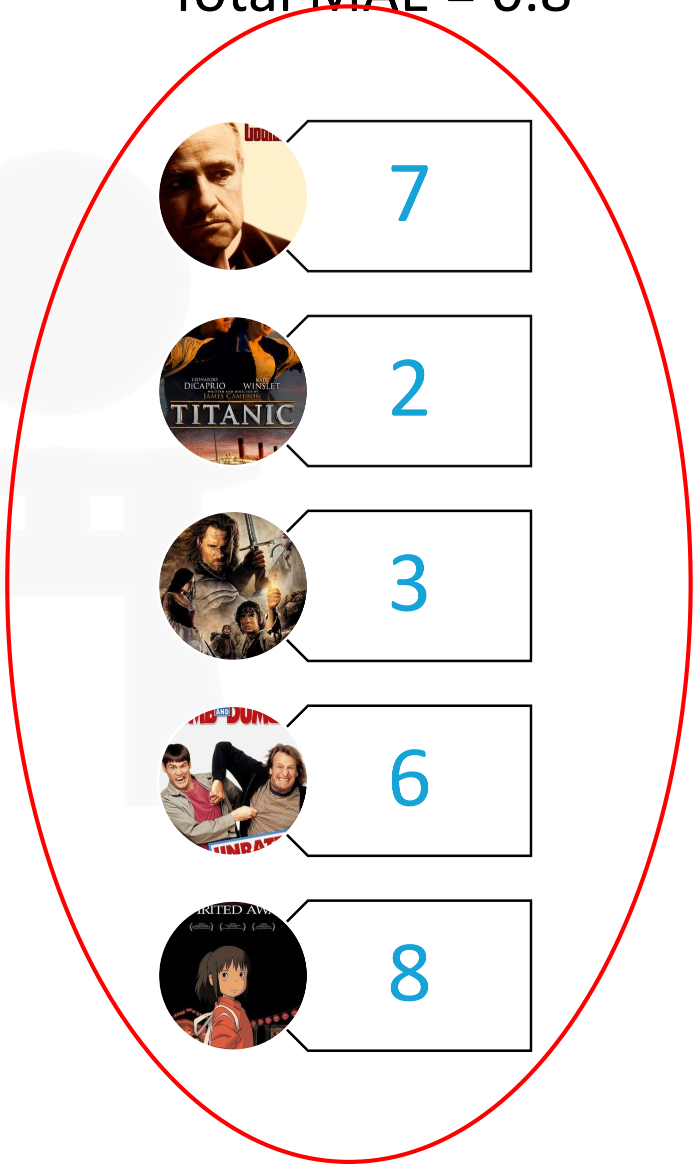


Total MAE =
15/5 = 3

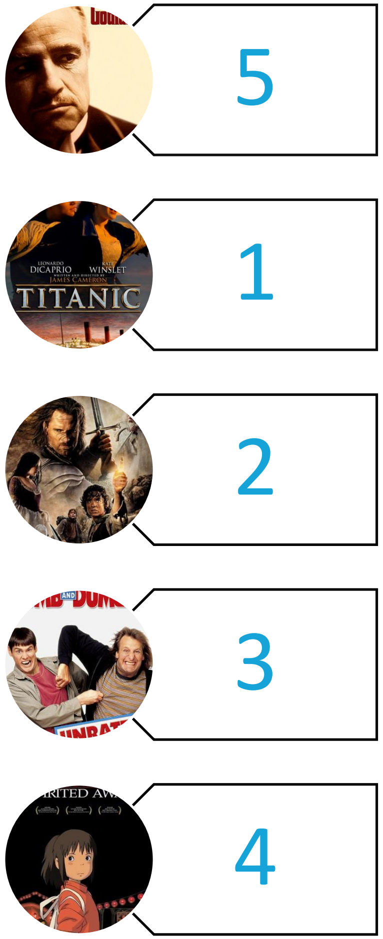
Which Recommender? – Model 1 or Model 2



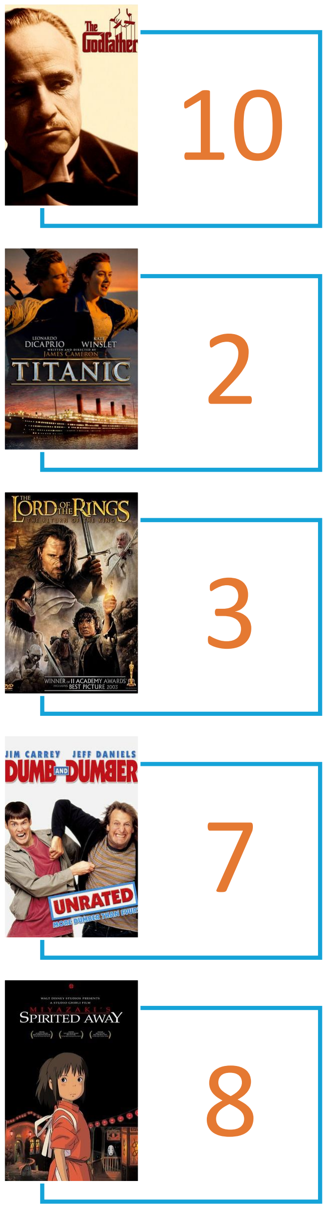
Total MAE = 0.8

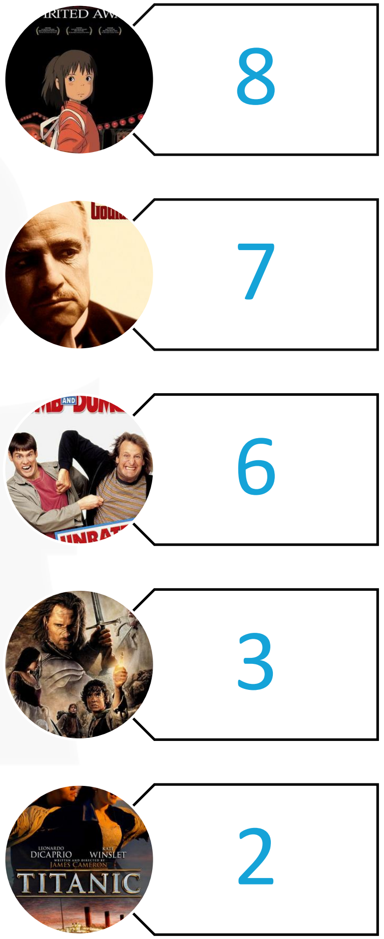


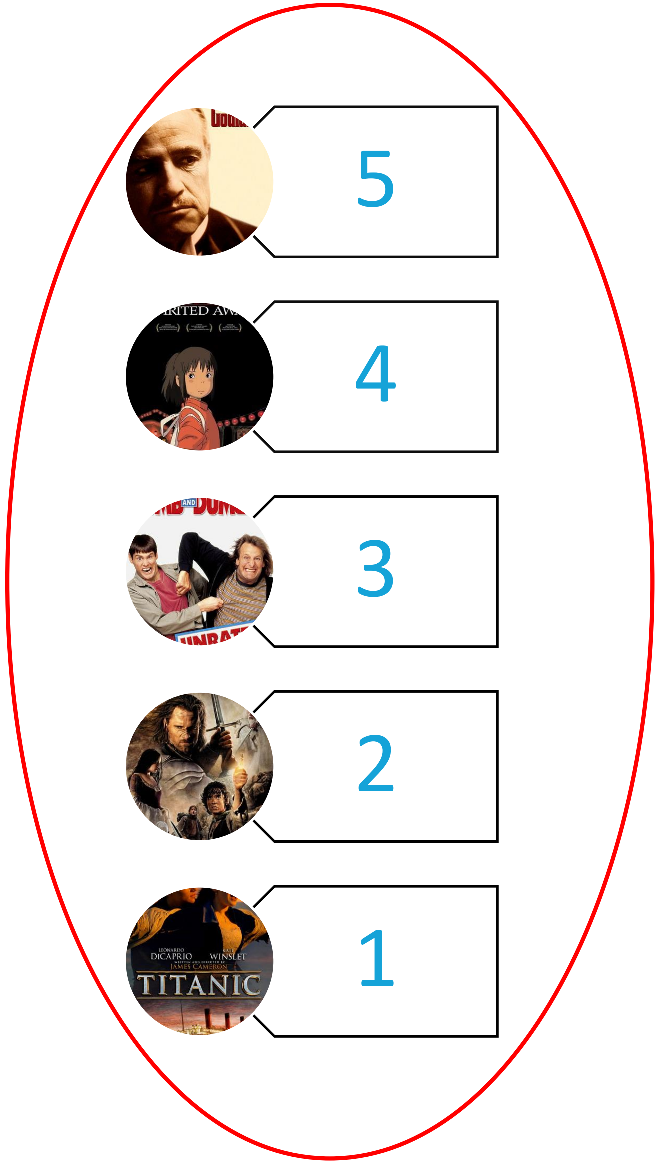
Total MAE = 3



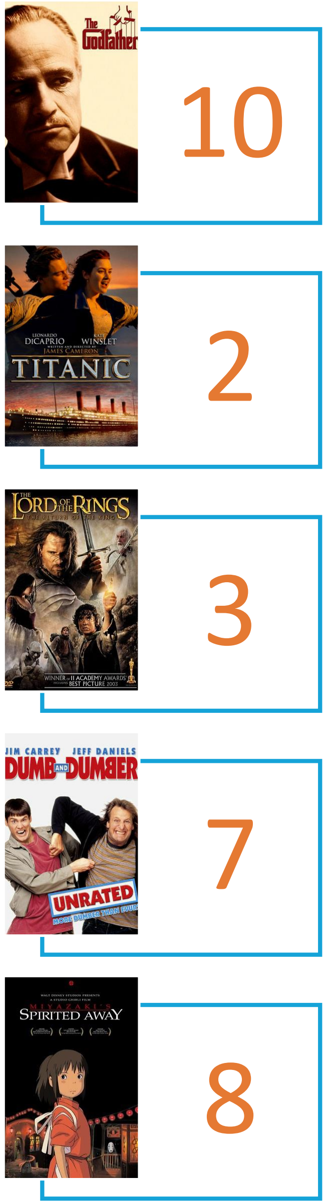
Which Recommender? – Model 1 or Model 2







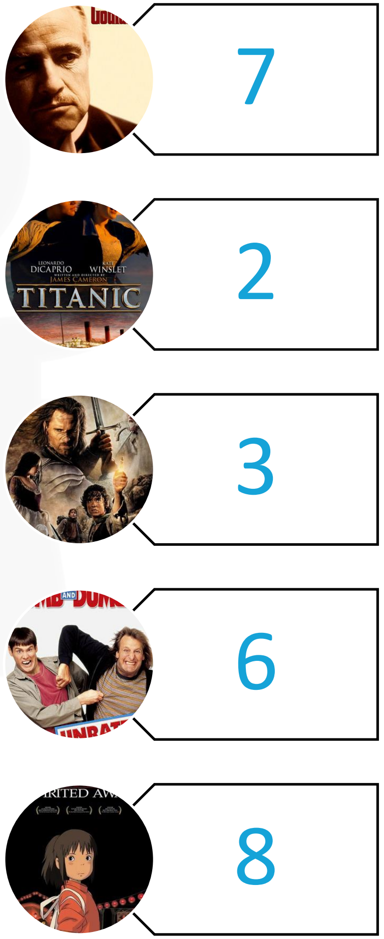
Recommender



Model 1 vs. Model 2

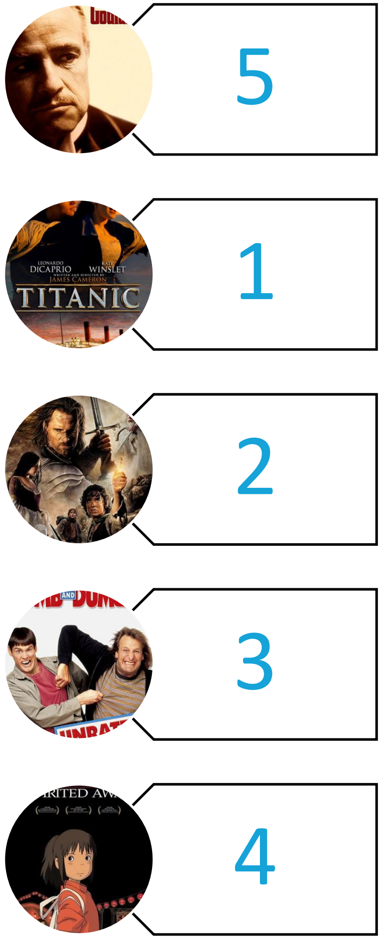
Predictor Model

Lower MAE value



Ranker Model

Follows same ranking as training



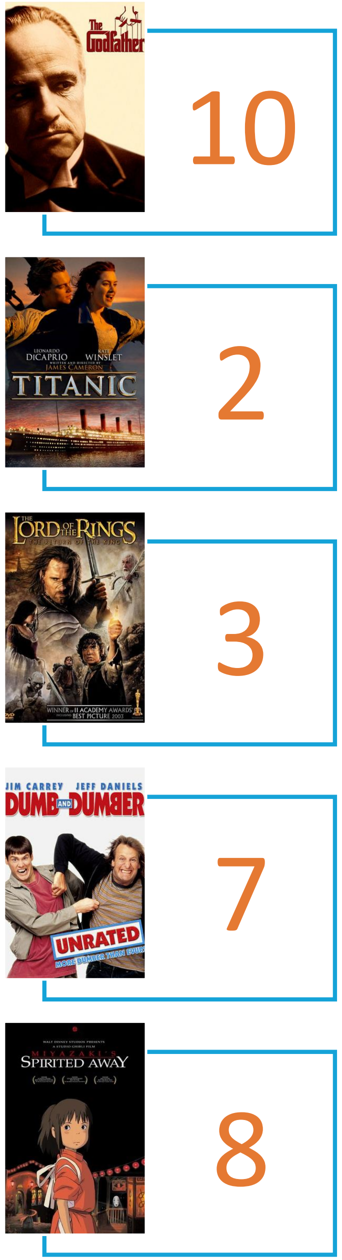
Metrics

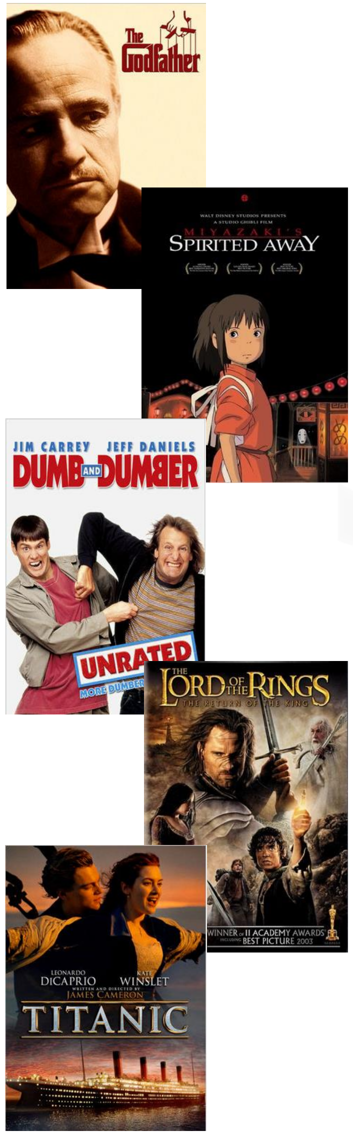
- Order matters, not exact rating value
- Graded Relevance
  - Have humans assign scores to possible results
  - Ideal results will be ordered by relevance, high to low
- Discounted cumulative gain (DCG)
  - Logarithmic reduction factor

$$DCG_N = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{\log_2 i}$$

Where:

- *N* is the length of the recommendation list
- rel_i returns the relevance of recommendation at position *i*





DCG Example

$$DCG_N = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{\log_2 i}$$

Following the formula above, the DCG for this set of movie ratings is:

$$10 + \frac{8}{log2(2)} + \frac{7}{log2(3)} + \frac{3}{log2(4)} + \frac{2}{log2(5)} \approx 24.78$$

Metrics

- **Ideal discounted cumulative gain (IDCG)**
  - DCG value when items are ordered perfectly

$$IDCG_N = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{\log_2 i}$$

- **Normalized discounted cumulative gain (nDCG)**

$$nDCG_N = \frac{DCG_N}{IDCG_N}$$

  - Normalized to the interval [0..1]

ADVANTAGES/DISADVANTAGES

Advantages

Collaborative

- Wide applicability
- Serendipity
- Simple

Content-based

- No community needed
- Transparency
- Good cold start

Disadvantages

Collaborative

- Poor cold start
- Grey Sheep
  - Shared accounts
- Shilling
- Poor scaling

Content-based

- Limited profiles
  - New users
  - Cost of expert labeling
- Over-specialization
  - Lack of diversity