

Recommender Systems

Data Science Dojo

Overview

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

INTRODUCTION

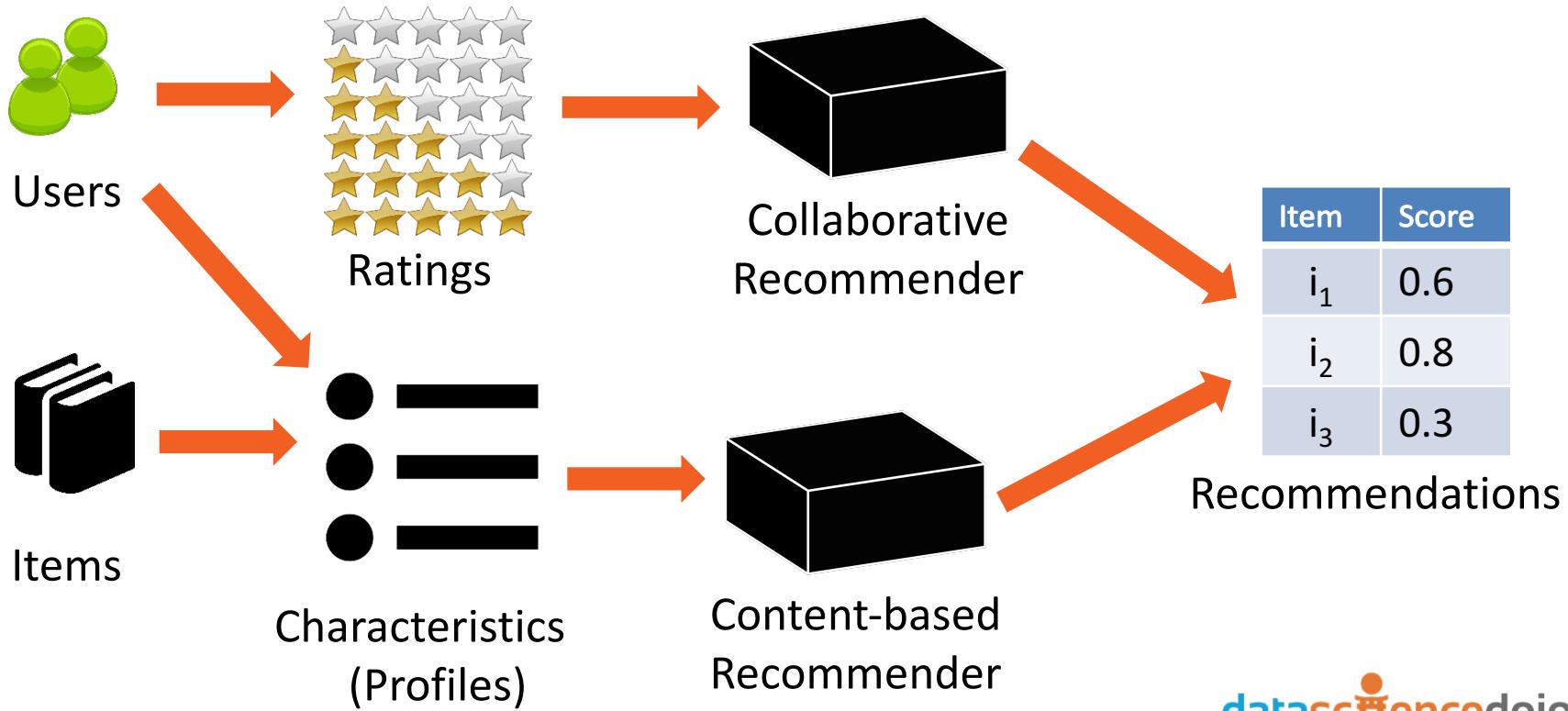
Recommendation Systems

- 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



Collaborative vs. Content-based Recommenders

Collaborative

- 'Give me items that people like me enjoy'

Content-Based

- 'Give me items similar to items I like'

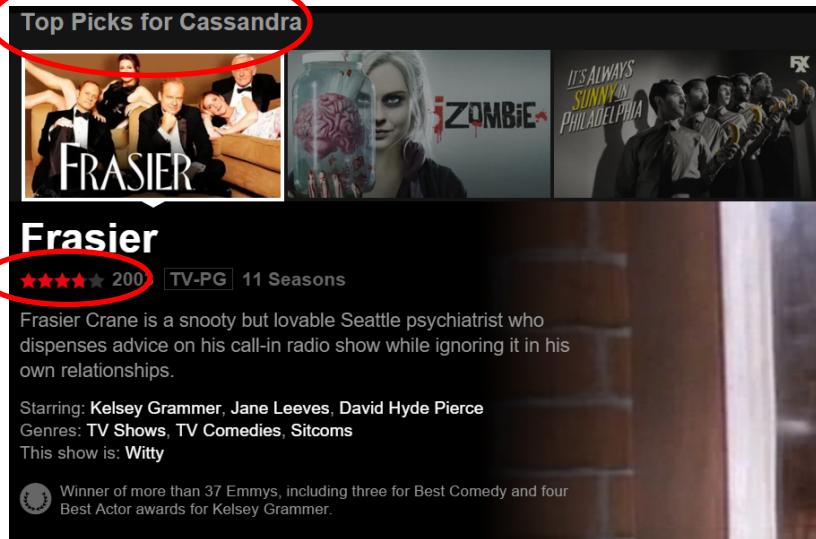
Collaborative Recommenders

- Users, Items, & Ratings
- Use Ratings of similar Users to recommend unseen Items

Content-based Recommenders

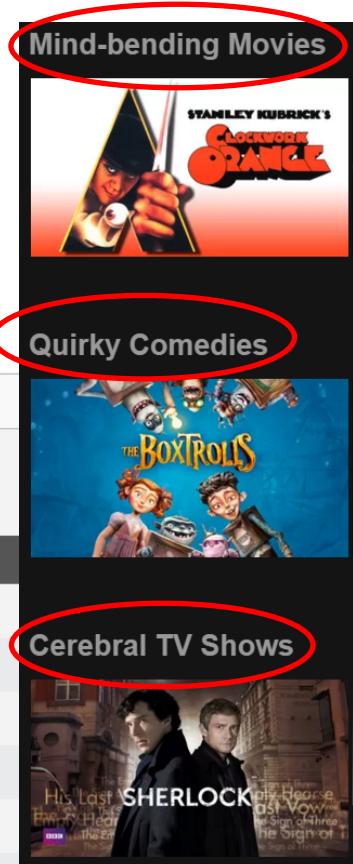
- User & Item profiles
- Use overlap of User and Item characteristics to recommend unseen items

Example: Netflix



The screenshot shows the "Taste Preferences" section of the Netflix website. At the top, there are tabs for "NETFLIX", "Browse", and "KIDS". Below that is a heading "Taste Preferences" with a red circle around it. There are two rows of filters: "How often do you watch" (Never, Sometimes, Often) and "Moods". The moods listed are Absurd, Adrenaline Rush, Bawdy, Campy, Cerebral, and Chilling, each with three radio buttons next to them.

How often do you watch	Never	Sometimes	Often
Absurd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adrenaline Rush	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bawdy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Campy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cerebral	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Chilling	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Example: Social Media & Search

People You May Know

 **Sergio in Kite**
The Old School Of Hard Knock
[redacted] and 2 other mutual friends

 **John Puerto Ram**
The new guy at Dalton LED
[redacted] and 23 other mutual friends

 **S-Jones James Clegg**
Works at The Home Depot

Ads You May Be Interested In

Big Data in 2015
Learn about 5 emerging trends in 2015 that have high ROI.

Attn: Successful Women
You're Invited to Join National Association of Professional Women

Invitation for Editors
Clinical & Translation

Data Science

Web News Images Books Videos More ▾ Search tools

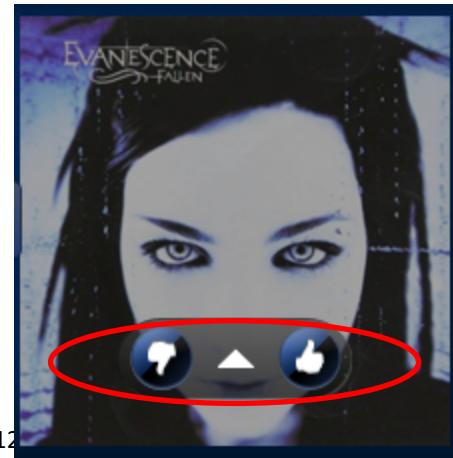
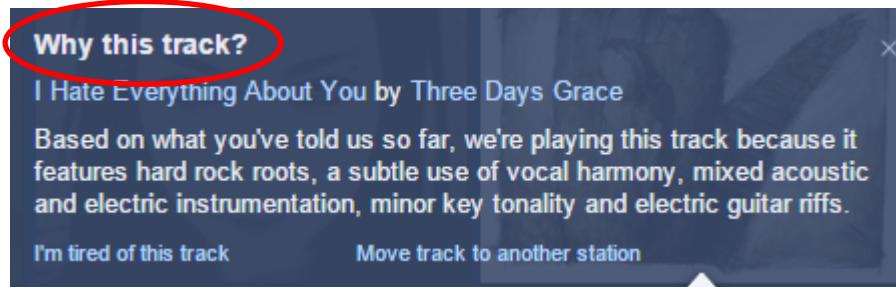
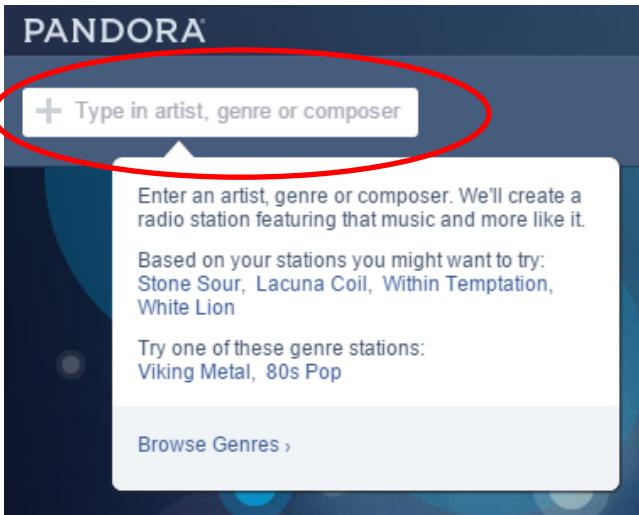
[Data science - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Data_science)
https://en.wikipedia.org/wiki/Data_science ▾ Wikipedia ▾
Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either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

[Data Science | Coursera](https://www.coursera.org/specializations/jhudatascience)
<https://www.coursera.org/specializations/jhudatascience> ▾ 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 Education](https://www.pce.uw.edu/certificates/data-science.html)
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



Example: Amazon

Inspired by Your Wishlist See more

Related to Items You've Viewed See more

Customers Who Bought This Item Also Bought

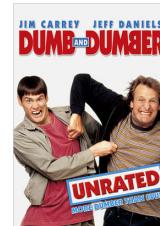
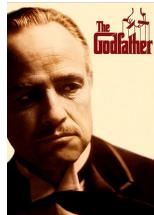
15

DATA STRUCTURE

Data Structure

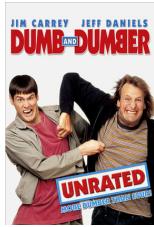
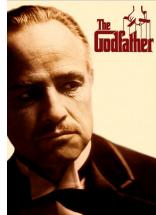
- What kind of data?
 - Collaborative
 - Ratings of Items by Users
 - Content-based
 - Characteristic profiles of Users and Items

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?
<i>The Godfather</i>	5	1	2	1
<i>Titanic</i>	4	3	2	5
<i>Lord of the Rings</i>	4	2	5	1
<i>Dumb & Dumber</i>	1	5	2	2
<i>Spirited Away</i>	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

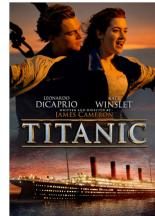
$$sim(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| * |\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

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

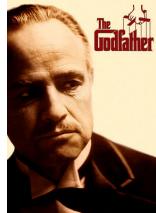
Which metric should we use?



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

Item-based Collaborative

- Alternate approach:
 - Use the similarity between items (and not users) to make predictions
 - Look for movies that are similar to movie m
 - Take Alice's ratings for these items to predict the rating for movie m

Collaborative: Item-based

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 = ?

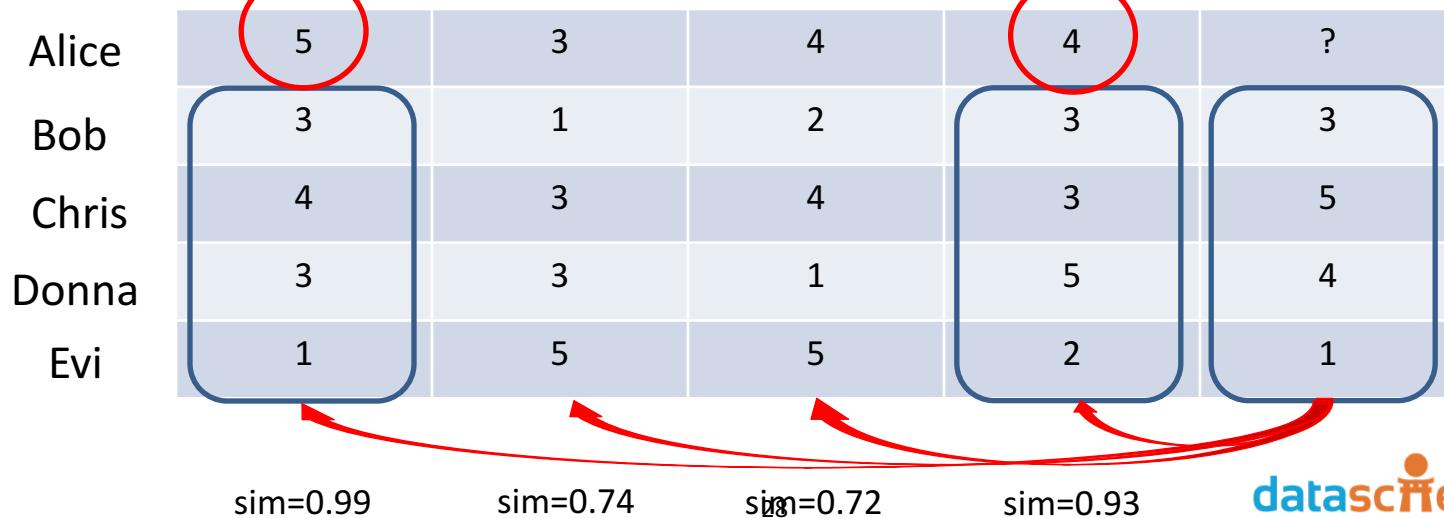
sim = ?

sim = ?

sim = ?

Collaborative: Item-based

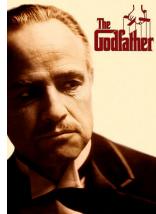
Cosine similarity
allows for
objective
good/bad



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?
<i>The Godfather</i>	5	1	2	1
<i>Titanic</i>	4	3	2	5
<i>Lord of the Rings</i>	4	2	5	1
<i>Dumb & Dumber</i>	1	5	2	2
<i>Spirited Away</i>	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
- Predict 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

- $r_{u,i} = \bar{r}_u + \alpha \sum_{j=1}^N sim(j, u)(r_{j,i} - \bar{r}_j)$

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

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

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

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

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)

$$nDCG_{pos} = \frac{DCG_{pos}}{IDCG_{pos}}$$

- 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

Back to 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.

Mind-bending Movies

Quirky Comedies

Cerebral TV Shows

The Netflix interface showing curated movie and TV show recommendations. The sections are highlighted with red ovals: "Mind-bending Movies" (top), "Quirky Comedies" (middle), and "Cerebral TV Shows" (bottom).

QUESTIONS

EXAMPLE USING AZURE ML