Welcome to General Assembly

- **WiFi** GA Guest
- **Password** yellowpencil

GENERAL ASSEMBLY

DATA SCIENCE

Lesson 9 - Recommendation Engines

Course Plan

UNITS

UNIT 1: FOUNDATIONS OF DATA MODELING	Introduction to Data Science	Lesson 1
	Elements of Data Science	Lesson 2
	Data Visualisation	Lesson 3
	Linear Regression	Lesson 4
	Logistic Regression	Lesson 5
	Model Evaluation	Lesson 6
	Regularisation	Lesson 7
	Clustering	Lesson 8
UNIT 2: DATA SCIENCE IN THE REAL WORLD	Recommendations	Lesson 9
	SQL + Productivity	Lesson 10
	Decision Trees	Lesson 11
	Ensembles	Lesson 12
	Natural Language Programming	Lesson 13
	Cloud Computing	Lesson 14
	Time Series	Lesson 15
	Soft Skills	Lesson 16
	Network Analysis	Lesson 17
	Neural Networks	Lesson 18
	Final Projects Presentations	Lesson 19
	Final Projects Presentations	Lesson 20

Paul & James review final project ideas

PROJECTS

FINAL PROJECT

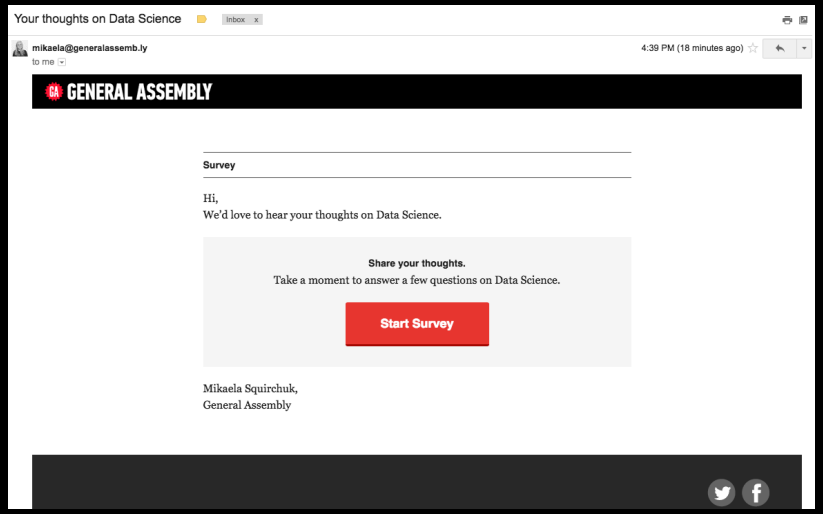
For the Data Science final project, you will address a data-related problem in your professional field or in a field you interested in. You will acquire a real-world data set, form a hypothesis about it, clean, parse, and apply modeling techniques and data analysis principles to ultimately create a predictive model. Students present their results and each write a report that includes the following:

- Clearly articulated problem statement
- Summary of data acquisition, cleaning, and parsing stage
- Clear presentation of your predictive model and the processes you took to create it
- Presentation style appropriate to the audience

Your instructional team will help you scope out your project so that you choose something that is feasible to accomplish given the skills you acquire in the course.

DATA SCIENCE PART TIME COURSE

Feedback Survey



DATA SCIENCE PART TIME COURSE

Drinks

Git & GitHub – 1 Pager Guide!

(Part B) EVERY CLASS:

At the START of the class, you'll need to sync the latest materials from the COURSE repo:

(1) **Make sure you are in the dat11syd directory:**

```
cd ~/workspace/dat11syd
```

(2) **Make sure to select the "master" branch of your repo:**

```
git checkout master
```

(3) **Fetch the latest changes from the UPSTREAM repo (i.e the course repo)**

```
git fetch upstream
```

(4) **Merge the changes from the upstream repo to your master branch:**

```
git merge upstream/master
```

DURING the class:

(5) **Before editing, either copy files to your "students/" folder, or rename them**

At the END of every class:

(6) **Make sure you are in the dat11syd directory:**

```
cd ~/workspace/dat11syd
```

(7) **Add any files that you've updated to your git registry:**

```
git add -A
```

(8) **Commit the changes with a sensible comment:**

```
git commit -m "my updates for lesson 7"
```

(9) **Push your changes to your PERSONAL repo:**

```
git push origin master
```

DONE!!!!!

AGENDA 8

1. What are Recommendations?
2. What is the motivation of recommendations?
3. What is Content-Based Filtering?
4. What is Collaborative Filtering?
5. Measuring Accuracy
6. Lab
7. Other Considerations
8. Discussion

DATA SCIENCE PART TIME COURSE

WHAT ARE RECOMMENDATION ENGINES?

- What are recommendations?
- Why are they important?
- Give one example people are likely to come across?

Work in two groups to answer the above questions and present back to the class.

5 mins

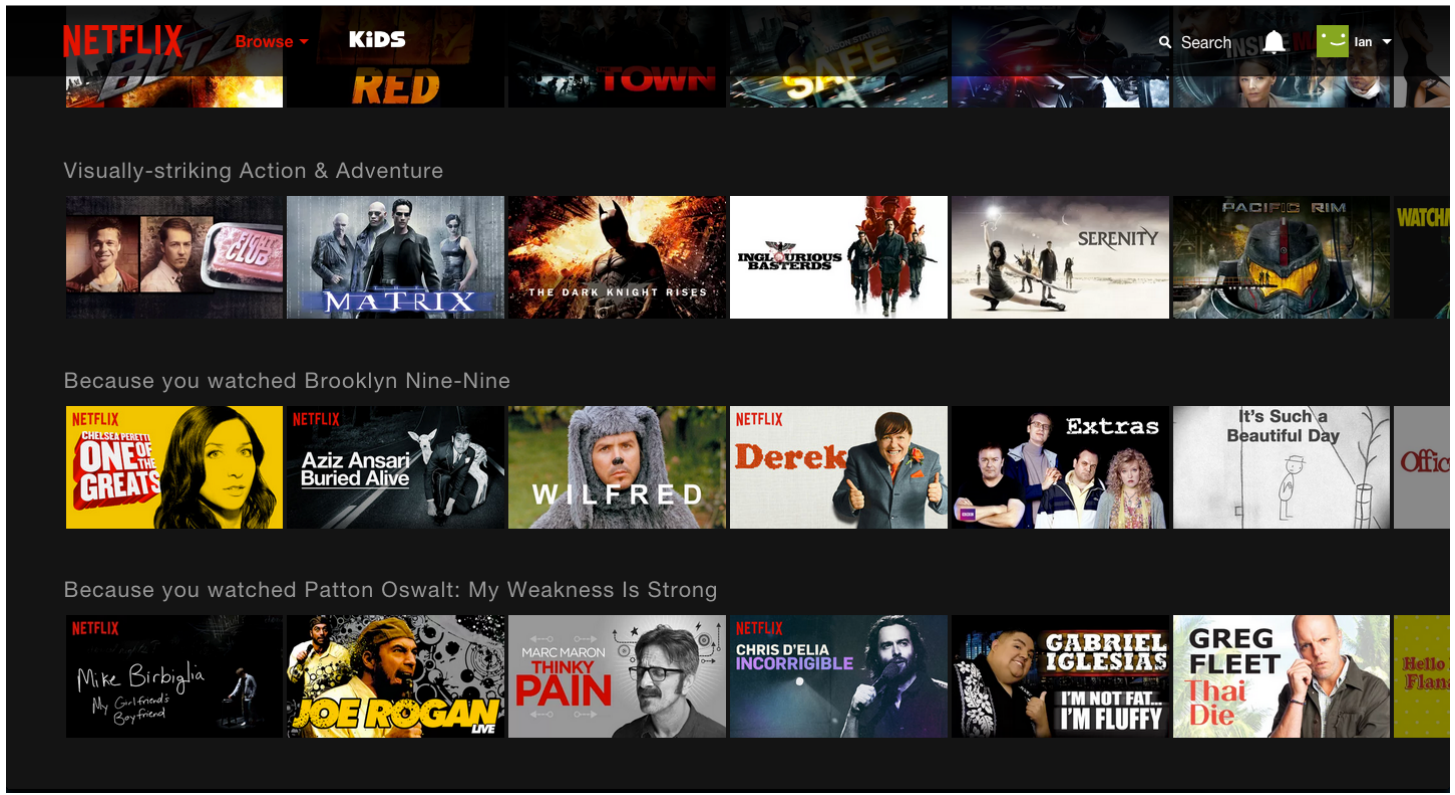
RECOMMENDATION ENGINES

Recommendation engines aims to match users to things (movies, songs, items, events, etc) they might enjoy but have not yet tried.

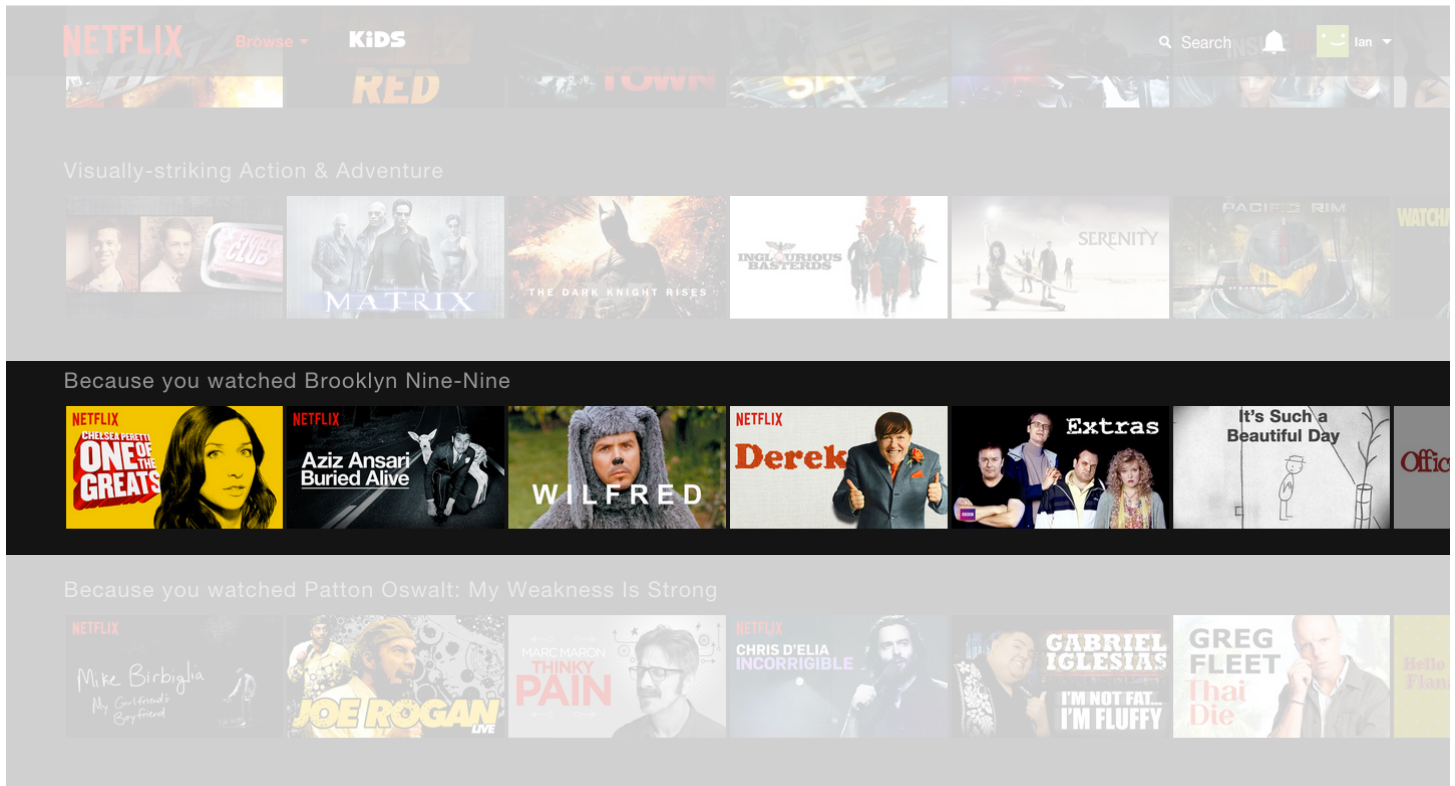
The rating is produced by analysing other user/item ratings (and sometimes item characteristics) to provide personalised recommendations to users.

DATA SCIENCE PART TIME COURSE

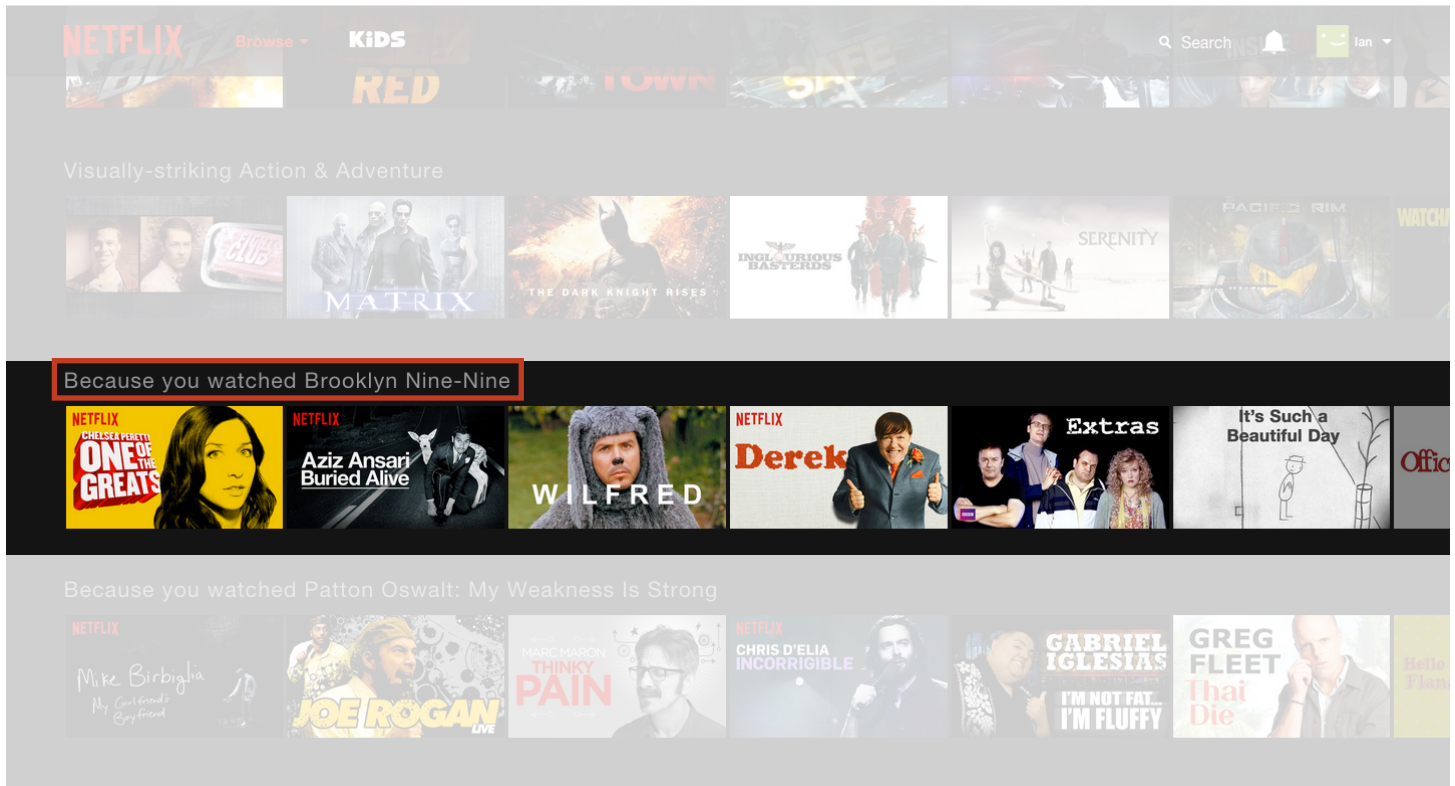
Netflix - Entertainment or Data Science?



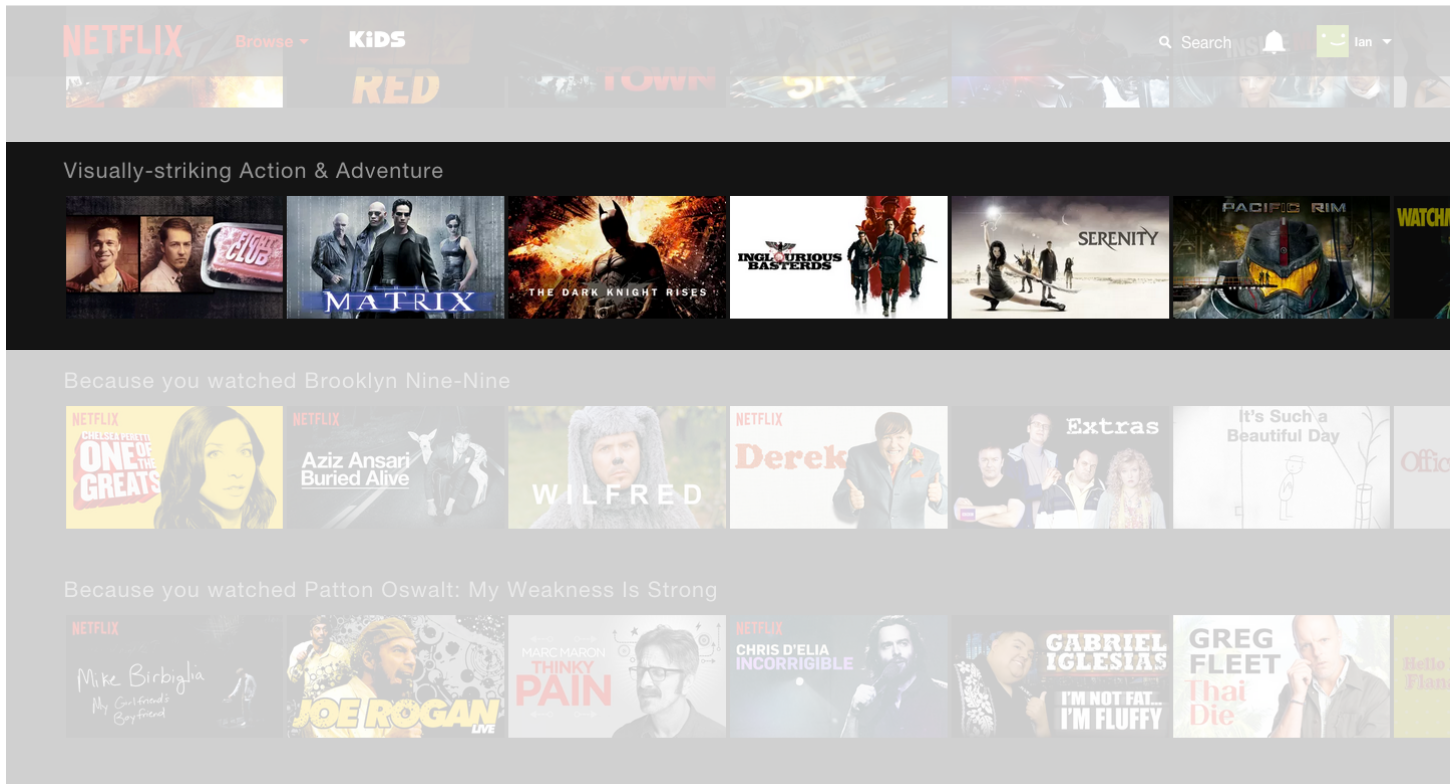
Netflix - Entertainment or Data Science?



Netflix - Entertainment or Data Science?

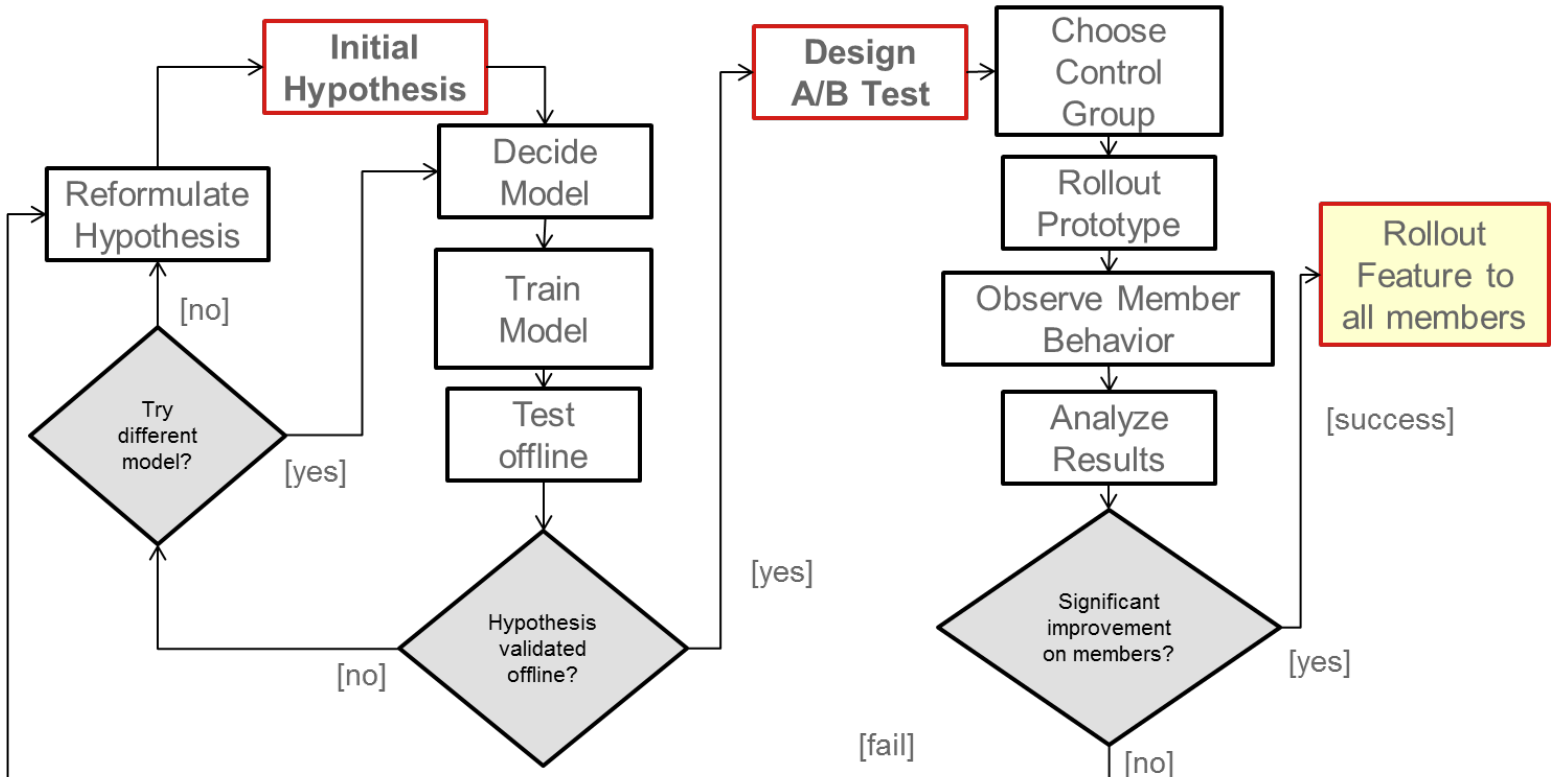


Netflix - Entertainment or Data Science?



Netflix - Entertainment or Data Science?

Netflix - Entertainment or Data Science? 19



There are two general approaches to the design:

Content-based filtering

items are mapped into a feature space, and recommendations depend on item characteristics (action / thriller / RomCom cast, reality vs fiction etc)

Collaborative filtering

Item ratings & recommendations based upon user preferences

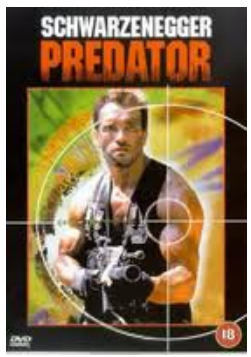
DATA SCIENCE PART TIME COURSE

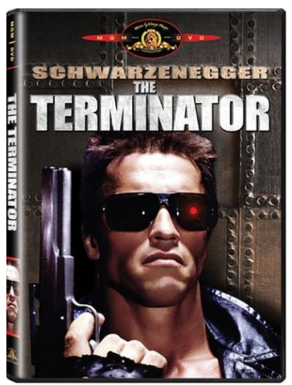
CONTENT-BASED FILTERING

CONTENT-BASED FILTERING

Looking at attributes of an item, you then make recommendations based on how similar those items are.

You liked Predator with Arnold Schwarzenegger you might also like The Terminator (because Arnie's in that too).



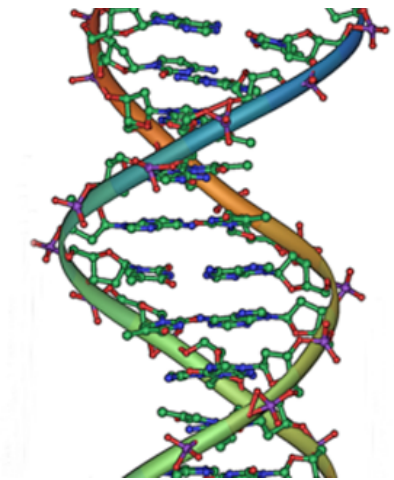


Content-based filtering begins by mapping each item into a feature space. Both users and items are represented by vectors in this space.

Item vectors measure the degree to which the item is described by each feature, and user vectors measure a user's preferences for each feature.

Ratings are generated by taking dot products of user & item vectors.

Pandora is an example of Content-Based filtering. A massive taxonomy of musical information. Trained musical analysts identify over 450 musical characteristics (lookup the Music Genome Project).



PANDORA®

Content-based filtering has some difficulties:

- **Must map items into a feature space (manual work)**
- **Recommendations are limited in scope (items must be similar to each other)**
- **Hard to create cross-content recommendations (eg books/music films...this would require comparing elements from different feature spaces)**

DATA SCIENCE PART TIME COURSE

COLLABORATIVE FILTERING

“Customers who purchased X also purchased Y”

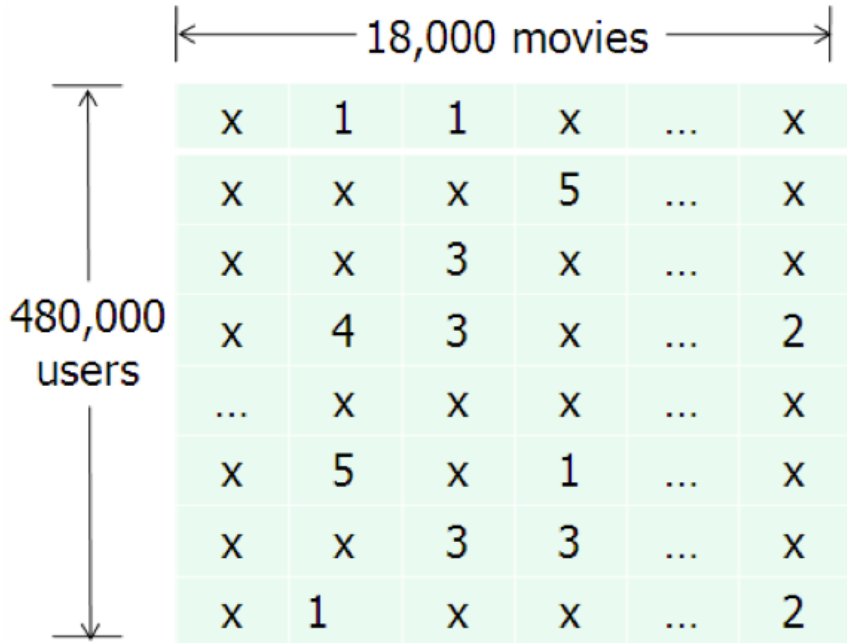
Someone with similar tastes to you will be able to recommend things you might like, e.g. people who watch ‘The Newsroom’ will probably enjoy ‘The Social Network’ because there is a large audience in common.

Collaborative filtering refers to a family of methods for predicting ratings where instead of thinking about users and items in terms of a feature space, we are only interested in the existing user-item ratings themselves.

In this case, our dataset is a ratings matrix whose columns correspond to items, and whose rows correspond to users.

This will be the general form of the data we analyse for collaborative filtering.

The method relies on previous user-item ratings (or feedback).



Collaborative filtering is susceptible to the Cold Start problem.

What happens if we don’t have any (or enough) reviews?

Collaborative filtering is susceptible to the Cold Start problem.

What happens if we don't have any (or enough) reviews?

Until users rate several items, we don't know anything about their preferences.

We can get around this by enhancing our recommendations using implicit feedback, which may include things like item browsing behaviour, search patterns, purchase history, etc. Or by using a hybrid model.

DATA SCIENCE PART TIME COURSE

SIMILARITY SCORES

Jaccard Similarity:
Defines similarity between two sets of objects

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

Jaccard Similarity:
Defines similarity between two sets of objects

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Number of similar elements

Total
Number of distinct elements

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

JS ({1, 2, 3}, {2,3,4}) = {2,3} / {1,2,3,4}
= 2/4
= 1/2

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

User one: {“LG-LCDTV 52” ”,“Blu Ray Player”,“HDMI Cable”}
User two: {“LG-LCDTV 52”,“Sony PS4”,“HDMI Cable”}

JS (User one, User two) =

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

User one: {“LG-LCDTV 52” ”,“Blu Ray Player”,“HDMI Cable”}
User two: {“LG-LCDTV 52”,“Sony PS4”,“HDMI Cable”}

JS (User one, User two) = 2/4
= 1/2

LAB

Lab 9 Recommendation Engines.ipynb

DATA SCIENCE PART TIME COURSE

OTHER CONSIDERATIONS

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Accuracy = Correct Recommendations / Total Possible Recommendations
= (a + d)/(a + b +c + d)

Precision = Correctly Recommended Items / Total Recommended Items
= d / (b + d)

Recall = Correctly Recommended Items / Total Useful Recommended Items
= d / (c + d)

Precision ~1 means the algorithm returned more relevant results than irrelevant.
Recall ~1 means that an algorithm returned most of the relevant results.

EXPLICIT VS IMPLICIT DATA

Explicit data is when you ask the user to rate something, e.g. 1-5 star rating for a movie

Implicit data is when you observe a users behaviour and record

- Alternating Least Squares (ALS)
- Stochastic Gradient Descent (SGD)
- Singular Value Decomposition (SVD)
- Factorization Machine (FM)
- Collaborative Less is More Filtering (CLiMF)

- Ranking
- Freshness
- Diversity
- Social Recommendations
- Context Aware Recommendations
- Hybrid Models (Combining Content based filtering and Collaborative filtering)
- Model Objectives (what are you trying to optimise)
- Sequences?

DATA SCIENCE

DISCUSSION TIME

Week 4 Review

Final Project 2

Pre-reading

Halfway Review

Project

WEEK 4 - Review

DISCUSSION TIME

- Regularisation
  - Why do we do this?
  - Name 1x type of regularised regression?
  - How does it work?
- Clustering
  - Why do we do this?
  - How does k-means work?

Final Project - Part 2

Goal: Create an outline of your research design approach, including hypothesis, assumptions, goals, and success metrics.

DELIVERABLES

Project Design Writeup

- **Requirements:**
  - Well-articulated problem statement with "specific aim" and hypothesis, based on your lightning talk
  - An outline of any potential methods and models
  - Detailed explanation of extant data available (ie: build a data dictionary or link to pre-built data dictionaries)
  - Describe any outstanding questions, assumptions, risks, caveats
  - Demonstrate domain knowledge, including specific features or relevant benchmarks from similar projects
  - Define your goals and criteria, in order to explain what success looks like
- **Bonus:**
  - Consider alternative hypotheses: if your project is a regression problem, is it possible to rewrite it as a classification problem?
  - "Convert" your goal metric from a statistical one (like Mean Squared Error) and tie it to something non-data people can understand, like a cost/benefit analysis, etc.
- **Submission:**
  - 19th Dec 2017

Lesson 9 - Review

PRE-READING

AirBnB article on AutoML: https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8

Lesson 9 - Review