DATA SCIENCE
11 WEEK PART TIME COURSE

Week 5 – Recommendation Engines
Monday 18th January 2016

1. What are Recommendations? (You tell me)
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 - Talk on Wednesday, Week 4

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 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.
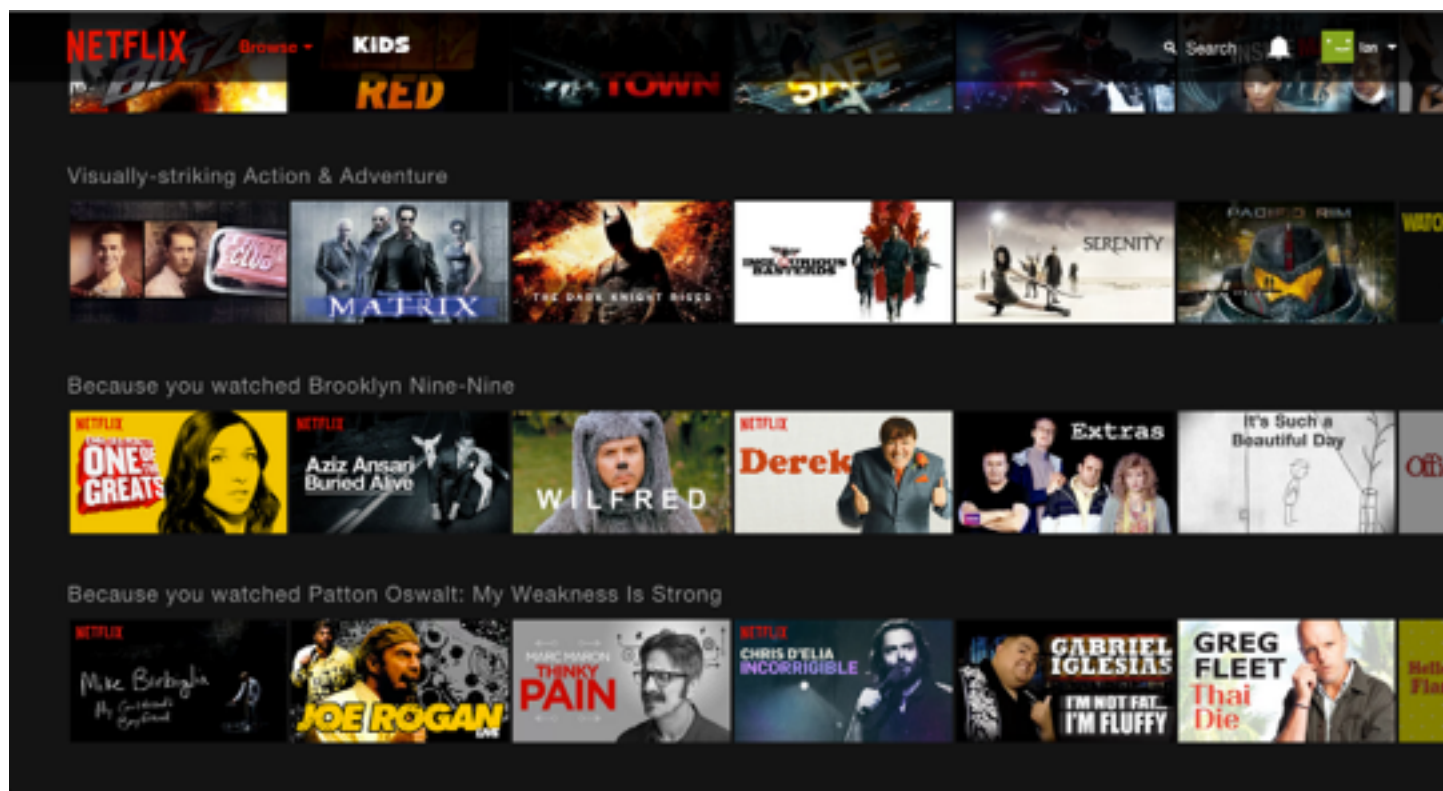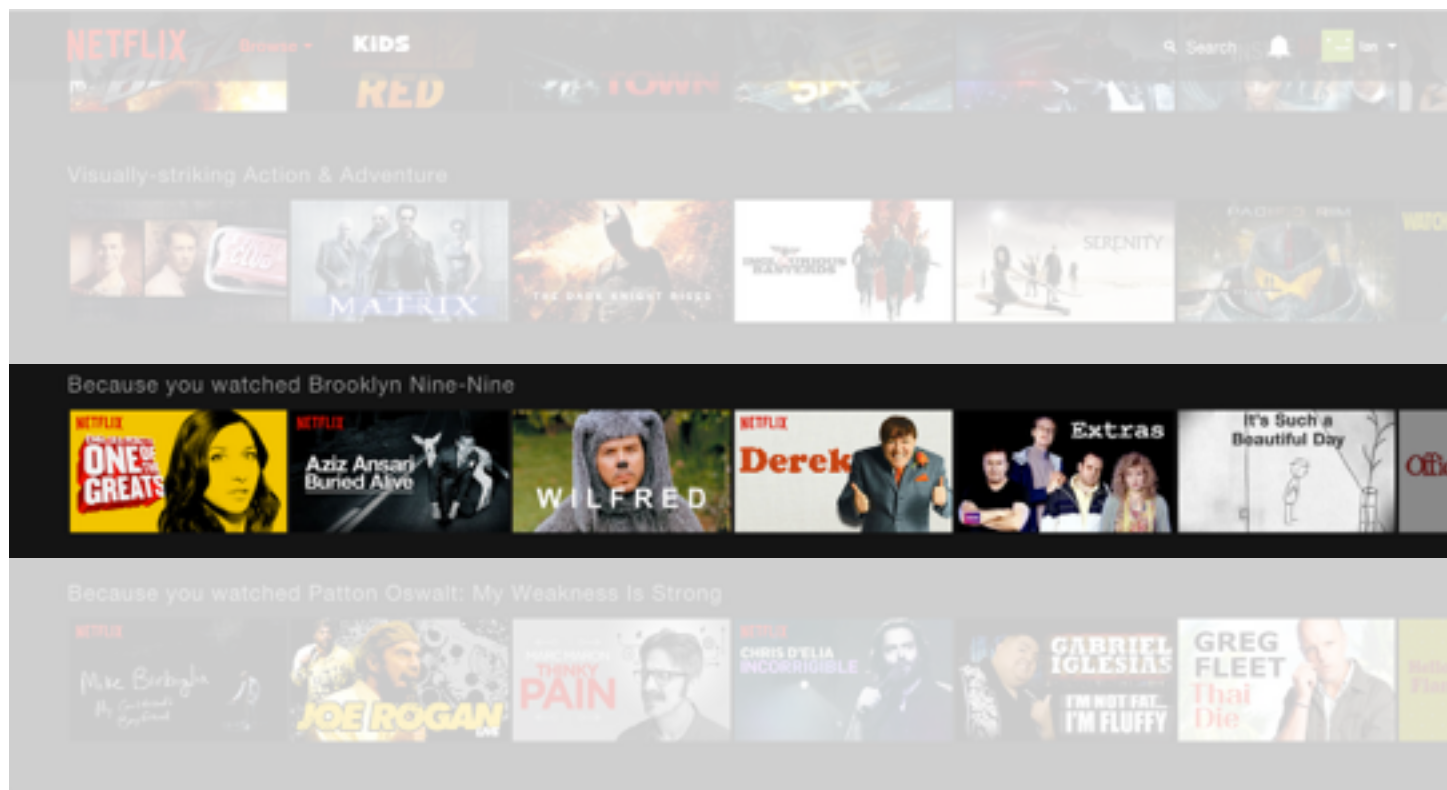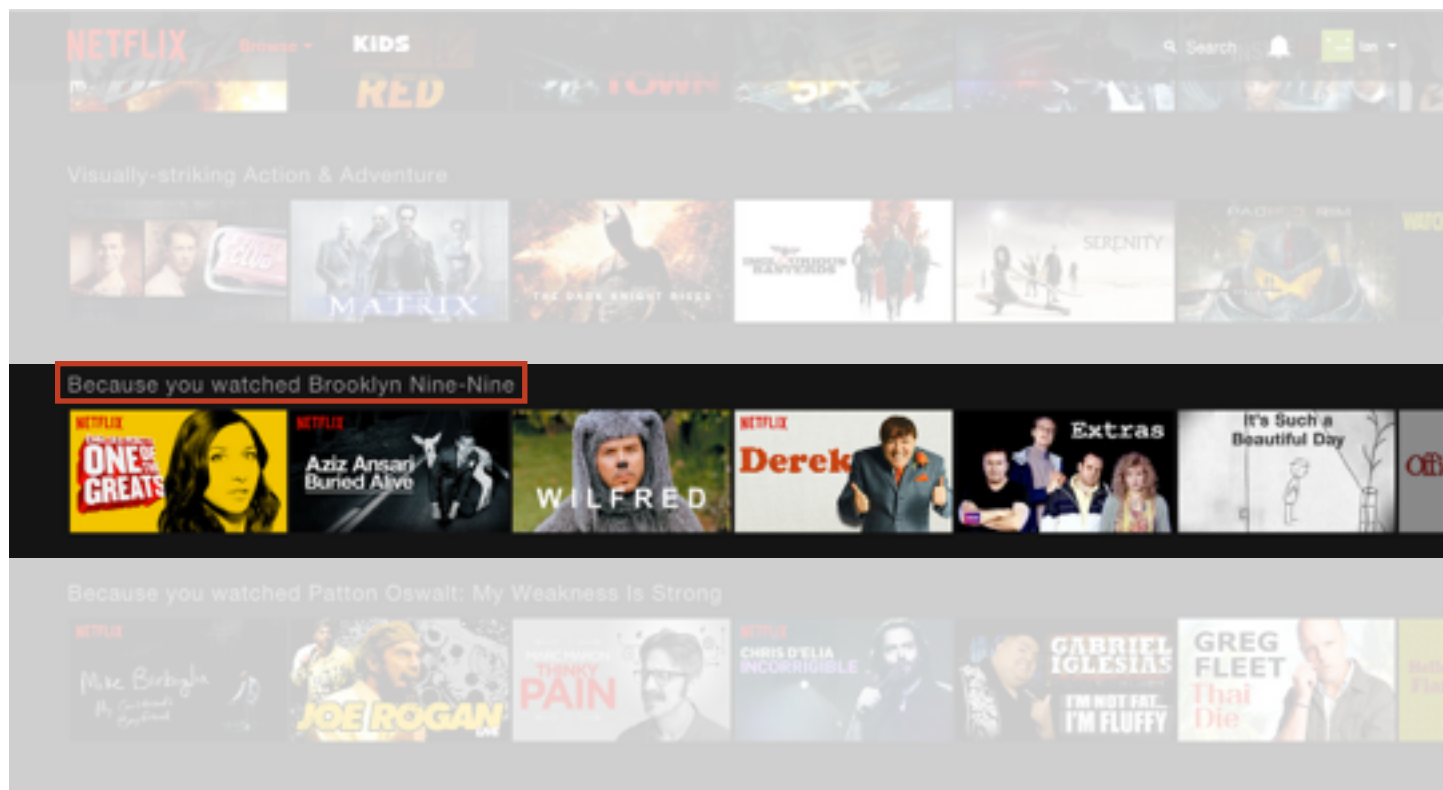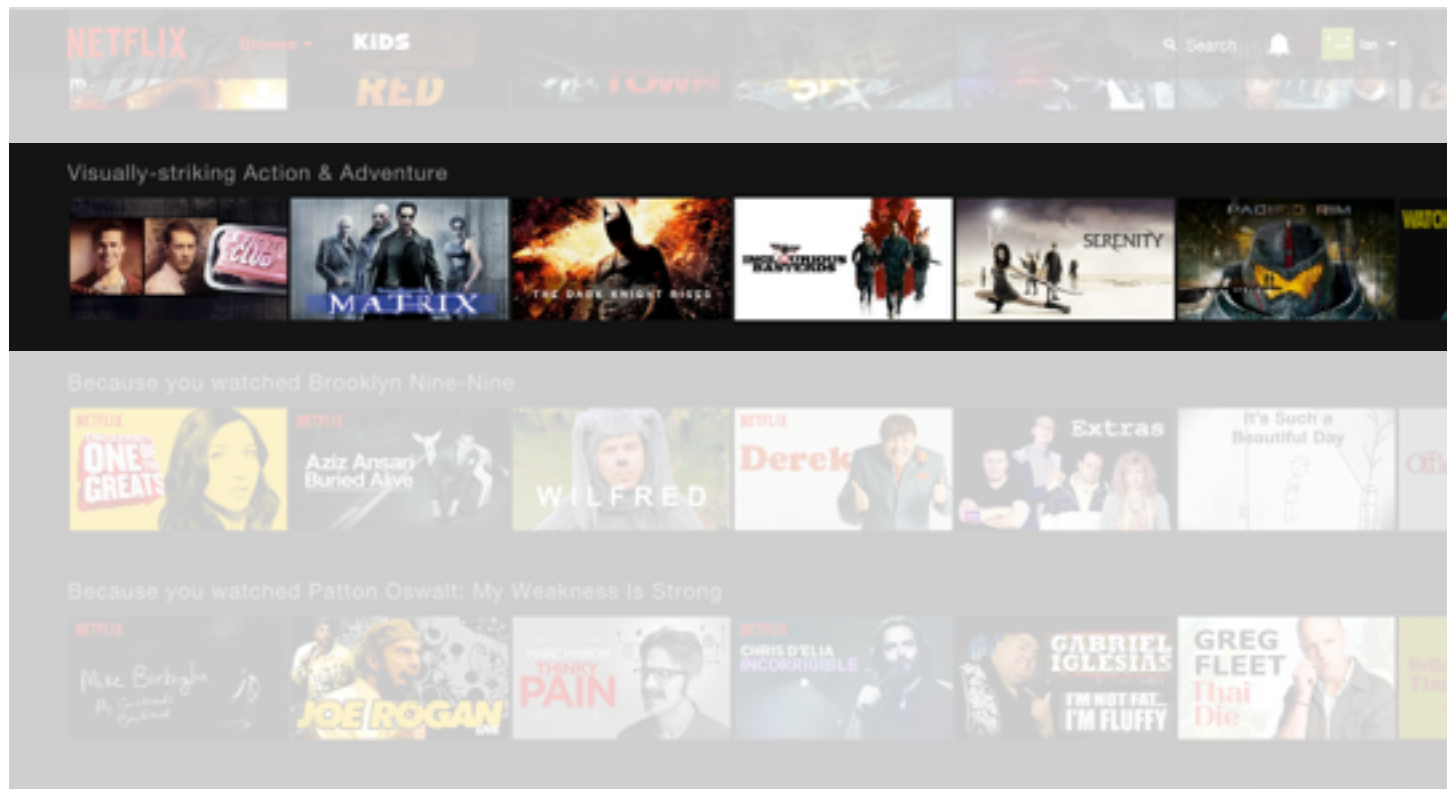
There are two general approaches to the design:

In **content-based** filtering, items are mapped into a feature space, and recommendations depend on item characteristics.

In contrast, the only data under consideration in **collaborative filtering** are user-item ratings, and recommendations depend on user preferences.

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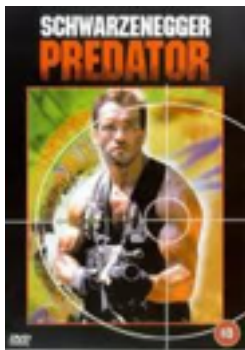
Content-based filtering begins by mapping each item into a feature space. Both users and items are represented by vectors in this space.

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).
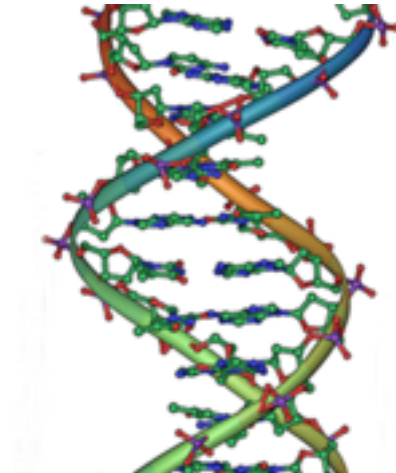
Content-based filtering has some difficulties:

‣ 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)

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).

		18,000 movies				
x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

480,000 users

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.

SIMILARITY SCORES

Jaccard Similarity:

Defines similarity between two sets of objects

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

Jaccard Similarity:
Defines similarity between two sets of objects

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

Number of similar elements

Number of distinct elements

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

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

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

User one: {"Target","Banana Republic","Old Navy"}
User two: {"Banana Republic","Gap","Kohl's"}

JS (User one, User two) =

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

User one: {"Target","Banana Republic","Old Navy"}
User two: {"Banana Republic","Gap","Kohl's"}

JS (User one, User two) = 1/5

LAB

git remote -v

git remote add upstream https://github.com/ihansel/SYD_DAT_3.git

git remote -v

git fetch upstream

git checkout master

git merge upstream/master

OR git reset –hard upstream/master

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 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 (SDG)
- 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?

DISCUSSION TIME

- ‣ Week 4
- ‣ Presentation next Wednesday – 5 slides on a data science topic of your choice
- ‣ Project
- ‣ Kaggle

Monday 11th January - Regularization

- [x] Select Variables for a Regression Model
- [x] Learn 3 Methods to automatically select variables
- [x] Explain difference between Ridge Regression & Lasso
- ? [x] Tie together concepts of X-Validation, Sub-Set

Wednesday 13th January

Clustering

- [x] Revise Unsupervised Learning
- [x] Know _1_ method for Clustering
- [x] Assess the "fit" of a Cluster
- [x] Summarise a Cluster

Presentations

‣ Goal of this exercise is to practice communication skills

‣ Majority of your communication has been written so far.

‣ Everyone should have a topic, if not just let me know and I'll suggest a few

How do you communicate to someone a complex idea in a way they feel they can understand the objective, method and results?