LAB #7: RECOMMENDER SYSTEMS

CS 109A, STAT 121A, AC 209A: Data Science

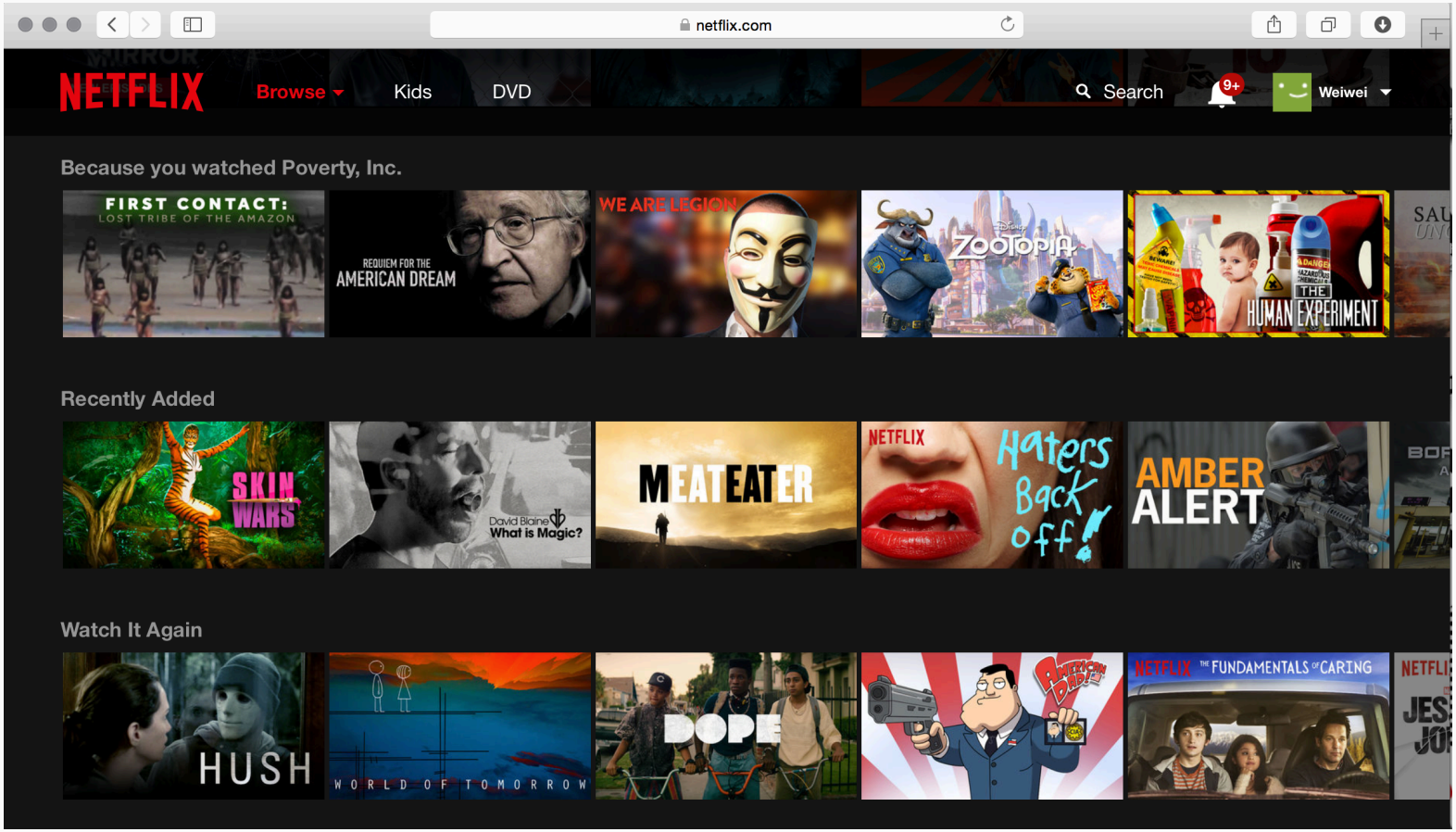
Fall 2016

Harvard University

NETFLIX RECOMMENDER SYSTEMS

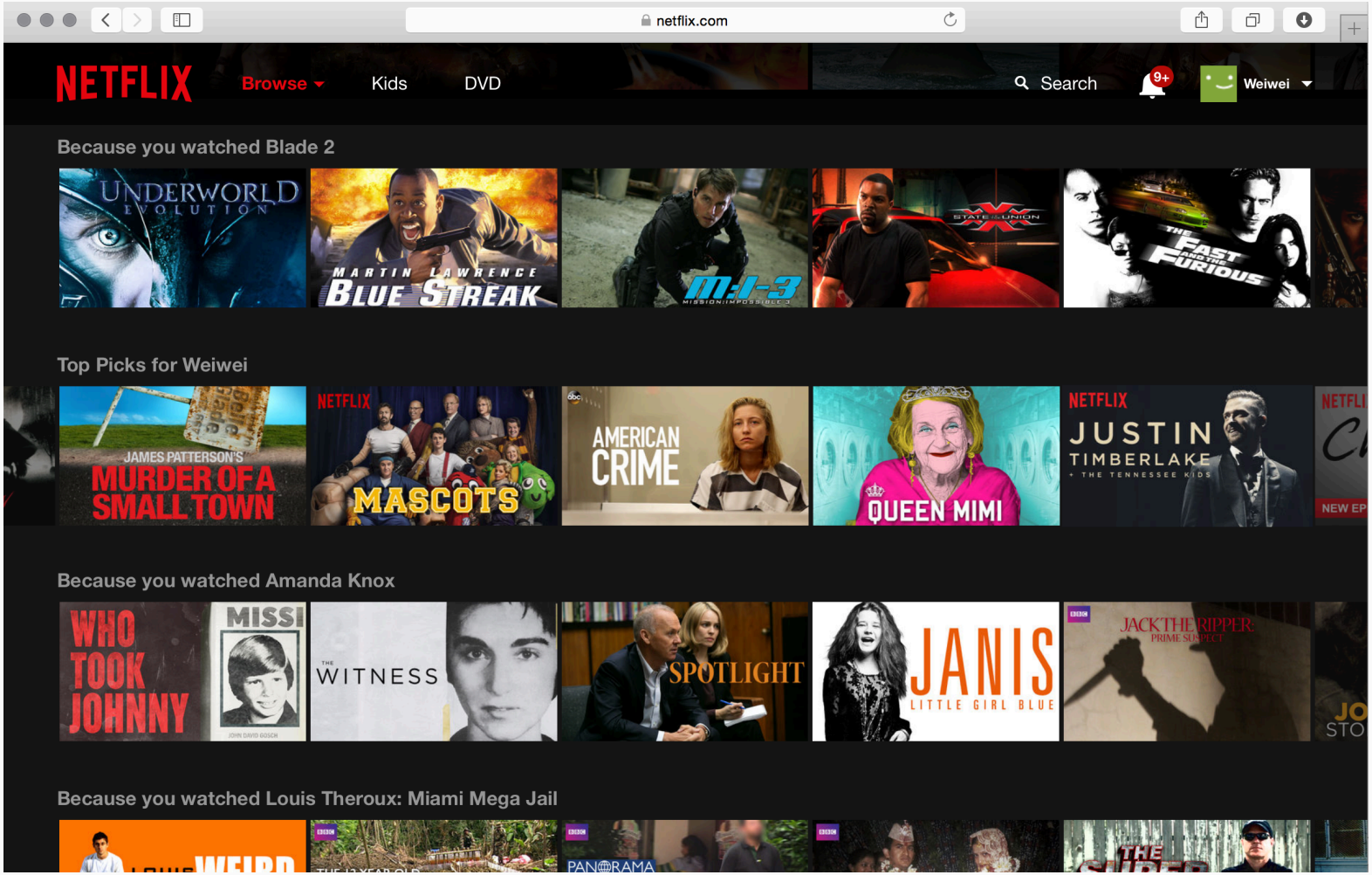
HOW DOES NETFLIX KNOW WHAT YOU LIKE?

What movies are on your Netflix landing page? Why are they there?



HOW DOES NETFLIX KNOW WHAT YOU LIKE?

What movies are on your Netflix landing page? Why are they there?



RECOMMENDING MOVIES

Neflix movie recommendations come in three flavors:

1. **Popularity based:** the movies everyone likes best
2. **Item based:** what's similar to this movie
3. **User based:** what you might like, given what you watch/like

We will explore popularity based movie recommendation in this lab.

POPULARITY BASED RANKING

THE "LIKABILITY" OF A MOVIE

We can phrase the movie recommendation problem based on popularity in terms of binary classification: for each movie, we classify it as "recommended" if its "likability" is in the top 25 and "not recommended" otherwise.

So how do we quantify "popularity" or "likability"?

THE "LIKABILITY" OF A MOVIE

We can phrase the movie recommendation problem based on popularity in terms of binary classification: for each movie, we classify it as "recommended" if its "likability" is in the top 25 and "not recommended" otherwise.

So how do we quantify "popularity" or "likability"?

1. Total number of likes
2. Percent of likes
3. Other suggestions?

THE "LIKABILITY" OF A MOVIE

For this lab, you have two datasets:

1. A dataset containing ratings from 100 users for 1000 movies. The first two columns contain the user and movie IDs. The last column contains a 1 if the user liked the movie, and 0 otherwise. Not every movie is rated by every user (i.e. some movies have more ratings than others).
2. A dataset with the names of the movies and corresponding IDs

Come up with a reasonable metric for "likability" and rank the movies! (**Steps 1-3**)

A PROBABILISTIC MODEL FOR "LIKABILITY"

Question: Why is `likes/ratings` a reasonable metric for "likability"?

Answer: We think `likes/ratings` for a movie approximates the probability of any user liking this particular movie.

Is this true?

A PROBABILISTIC MODEL FOR "LIKABILITY"

- Given a movie, the probability that any user will like it

$$p(\textbf{rating} = \text{like}) = \theta_{\text{movie}}$$

- How does this like-probability relate to ratings?
- What's the probability of getting 2 likes out of 3 ratings given θ_{movie}? Assuming that each user rates the movie independently using the same like-probability.

$$p(\textbf{likes} = 2 | \textbf{ratings} = 3, \theta_{\text{movie}}) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \theta_{\text{movie}}^2 (1 - \theta_{\text{movie}})^{3-2}$$

$\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ is the number of ways to arrange 3 ratings with 2 likes; θ_{movie}^2 is the probability of getting two likes; $(1 - \theta_{\text{movie}})^{3-2}$ is the probability to getting one unlike.

A PROBABILISTIC MODEL FOR "LIKABILITY"

- Given a movie, the probability that any user will like it

$$p(\textbf{rating} = \text{like}) = \theta_{\text{movie}}$$

- How does this like-probability relate to ratings?
- What's the probability of getting 2 likes out of 3 ratings given θ_{movie}? Assuming that each user rates the movie independently using the same like-probability.

$$\underbrace{p(\textbf{likes} = k | \textbf{ratings} = n, \theta_{\text{movie}})}_{\textit{Likelihood}} = \underbrace{Bin(k; n, \theta_{\text{movie}})}_{\text{binomial distribution}}$$

$$= \begin{pmatrix} n \\ k \end{pmatrix} \theta_{\text{movie}}^{k} (1 - \theta_{\text{movie}})^{n-k}$$

- We see the number of likes, from this, we want to estimate θ_{movie}. How?

MAXIMUM LIKELIHOOD ESTIMATE OF LIKE-PROBABILITY

We want to find the $\theta^{\text{MLE}}_{\text{movie}}$ that maximize the likelihood of the data:

- Work with the log Likelihood:

$$\log \underbrace{p(\textbf{likes} = k | \textbf{ratings} = n, \theta_{\text{movie}})}_{\textit{Likelihood}} = \log \left[\binom{n}{k} \theta^k_{\text{movie}} (1 - \theta_{\text{movie}})^{n-k} \right]$$

- Simplify:

$$\log \underbrace{p(\textbf{likes} = k | \textbf{ratings} = n, \theta_{\text{movie}})}_{\textit{Likelihood}} = \log \binom{n}{k} + k \log \theta_{\text{movie}} + (n - k) \log(1 - \theta_{\text{movie}})$$

- Take the derivative, set to zero and solve for θ_{movie}.

$$\begin{aligned}
0 &= \frac{k}{\theta_{\text{movie}}} - \frac{n-k}{1-\theta_{\text{movie}}} \\
\frac{n-k}{\theta_{\text{movie}}} &= \frac{k}{1-\theta_{\text{movie}}} \\
n\theta_{\text{movie}} - k\theta_{\text{movie}} &= k - k\theta_{\text{movie}} \\
\theta^{\text{MLE}}_{\text{movie}} &= \frac{k}{n} = \frac{\textbf{likes}}{\textbf{ratings}}
\end{aligned}$$

The like-percentage is just the MLE estimate of the like-probability!

A BAYSIAN PROBABILISTIC MODEL FOR "LIKABILITY"

Question: Is using like-percentage better than total likes?

Question: Is using like-percentage fool-proof (best ever)?

Question: Does our top 25 movies ranked using like-percentage look reasonable? What factors are we failing to account for in our model?

A BAYSIAN PROBABILISTIC MODEL FOR "LIKABILITY"

To prevent overfitting, we want to incorporate some prior beliefs regarding the like-probability.

- Pick a prior $p(\theta_{\text{movie}})$ to encode your beliefs
- Incorporate the prior by considering the posterior

$$\underbrace{p(\theta_{\text{movie}} \mid \textbf{likes} = k, \textbf{ratings} = n)}_{\text{Posterior}} \propto \underbrace{p(\textbf{likes} = k \mid \theta_{\text{movie}}, \textbf{ratings} = n)}_{\text{Likelihood}} \underbrace{p(\theta_{\text{movie}})}_{\text{Prior}}$$

- Use MAP, $\theta^{\text{MAP}}_{\text{movie}}$, the mode of the posterior to estimate like-probability.

Apply this to your dataset and recompute the top 25 movies using $\theta^{\text{MAP}}_{\text{movie}}$. (**Steps 4-5**)

EVALUATING OUR BAYESIAN MODEL

- What kind of beliefs about like-probability do the priors encode?
- Does adding a prior make a difference in our like-probability estimate?
- What is a good prior? What if we picked wrong???
- Is ranking by $\theta^{\text{MAP}}_{\text{movie}}$ a good idea?