DATA 607
RECOMMENDER SYSTEMS

Week 1: Welcome!

GOALS FOR MEETUP

- Welcome!
- Course Overview: Syllabus and Blackboard, Slack?
- Overview types of recommender systems
- Introduce concepts for coming weeks

COURSE OVERVIEW

- This is an abbreviated 7-week course.
  - We will have assignments and discussion topics due each week
- Weekly rhythm:
  - Tuesday: Meetup @ 7 PM (EDT)
  - Thursday: Research Discussion item posted in BlackBoard
  - Tuesday: Assignment Due
  - Tuesday: Project Due
- Assignments need to be turned in by midnight on Tuesdays.
  - We will have a strict policy on no late submissions, but we will only grade the best 4/5 assignments and 4/5 research discussion items

ACADEMIC INTEGRITY

- In this course, you may collaborate and use base code from whatever sources you wish. But you must document what you started with, and what you added, so you are graded on your own contributed work! Please cite and provide a link to the original source.

- Reproducibility – students are providing all code and data so we can reproduce your work. All assignments – projects and research discussion assignments – need to be delivered in either a Jupyter notebook or an R Markdown format. This code should be posted into your GitHub repository, and you should provide the link in the project submission or discussion board, where applicable.

PROJECTS

- You will be responsible for presenting a topic during one of our course meetups. This is the Recommender Systems in Context presentation on the syllabus
  - You can sign up using the Doodle poll

- Final Project: Build a Recommender System on large dataset!
  - This will be a group project (2-3 students)
  - Planning Document Due 7/9
  - Final Code Due 7/16
  - Presentation 7/20

WHAT IS A RECOMMENDER SYSTEM?

1. Hand-curated
   a) "Staff Favorites" in a book store
   b) Movie Critics: Siskel and Ebert (past); Dana Stevens, Manohla Dargis, AO Scott (present)

2. Non—personalized aggregates
   a) Metacritic: Weighted Average of scores from small group of critics
      http://www.metacritic.com/about-metascores
   b) Rotten Tomatoes: Probabilistic scoring from larger group of critics
      https://www.rottentomatoes.com/about/
   c) IMDB Rating: Weighted Rating, "Bayesian Estimate"
      http://imdb.to/2ljPH90

3. Personalized
   a) Netflix, Spotify, Pandora, Amazon,
   b) The focus of this class!

POPULAR RECOMMENDERS

RECOMMENDER SYSTEMS IN THE WILD

Compare with similar items

	This item Cuisinart ICE-70 Electronic Ice Cream Maker, Brushed Chrome	Cuisinart ICE-30BC Pure Indulgence 2-Quart Automatic Frozen Yogurt, Sorbet, and Ice Cream Maker	Cuisinart ICE-21 1.5 Quart Frozen Yogurt-Ice Cream Maker (White) #1 Best Seller	Hamilton Beach 68330N 4-Quart Automatic Ice-Cream Maker,Cream
	Add to Cart	Add to Cart	Add to Cart	Add to Cart
Customer Rating	(432)	(3284)	(4312)	(1495)
Price	\$132.95	\$74.99	\$42.08	\$29.33
Shipping	prime	prime	prime	prime
Sold By	Amazon.com	Amazon.com	Amazon.com	Amazon.com
Color	Ice Cream Maker with Countdown Timer	Silver	White	Cream
Item Dimensions	8.62 x 9.74 x 13.22 in	9.11 x 8.31 x 11.26 in	10.8 x 10.8 x 15.8 in	12.5 x 15.3 x 11.1 in
Item Weight	13.5 lbs	12 lbs	10 lbs	6.5 lbs
Material Type	Chrome	Stainless Steel	man-made-material	Metal

RECOMMENDER SYSTEMS IN THE WILD

RECOMMENDER SYSTEMS IN THE WILD

(Amazon Prime has 100M users that bought $5B worth of items in 2017)

SYSTEM ENTITIES

"In a recommendation-system application there are two classes of entities, which we shall refer to as *users* and *items.* Users have preferences for certain items and these preferences must be teased out of the data. The data itself is represented as a *utility matrix*, giving for each user-item pair, a value that represents what is known about the degree of preference for that user for that item. Values come from an ordered set, e.g., integers 1-5 representing the number of stars that the user gave as a rating for that item. We assume that the matrix is sparse, meaning that most entries are "unknown". An unknown rating implies that we have no explicit information about the user's preference for the item."

Source: http://www.mmds.org

PERSONALIZED RECOMMENDER SYSTEMS

Types:

1. Content-Based

2. Collaborative Filtering
   - User-User
   - Item-Item

Methods:

1. Matrix Factorization

2. Alternations Least Squares

CONTENT-BASED RECOMMENDATIONS

	Signal and the Noise	Data Science for Business	Automated Data Collection With R	R for Data Science	Intro to Statistical Learning	ML w/ Scikit-Learn & Tensorflow	Weapons of Math Destruction	Programming Collective Intelligence	R in Action
Stats	1			1	1			1	
R			1	1					1
Python						1		1	
Big Data	1	1					1		
ML						1		1	

In Content-based recommenders, we examine properties of the items to be recommended.

MovieLens (a non-commercial recommender run by GroupLens a research lab out of the CS department at the University of Minnesota) allows community-applied tags! You can even make your own if you choose.

COLLABORATIVE FILTERING

	Signal and the Noise	Data Science for Business	Automated Data Collection With R	R for Data Science	Intro to Statistical Learning	ML w/ Scikit-Learn & Tensorflow	Weapons of Math Destruction	Programming Collective Intelligence	R in Action
Dave	4			5					4
Walt	4	5	4	4	5	NR	5	NR	5
Robert	NR	NR	NR	NR	NR	NR	NR	NR	NR
Tulasi	NR	NR	NR	5	5	NR	NR	NR	5
Logan	NR		4	5	5	NR	4	NR	4
Shyam	NR	4	5	NR	3	NR	NR	NR	NR
Yun	NR	5	4	4	5	NR	NR	NR	NR
Kumidini	NR	4	4	4	5	NR	NR	NR	4
Jason Joseph	NR	1	1	NR		NR	NR	NR	1

In Collaborative filtering, items are recommended based on similarity measures between items and/or users. In this case, would we be better off looking for item-item or user-user similarity?

Greg Linden, Brent Smith, and Jeremy York, Amazon.com: Item-to- Item Collaborative Filtering, http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf. 2003.

MATRIX FACTORIZATION

- A common and powerful tool to apply collaborative filtering methods to ratings matrices are matrix factorization tools. This helps identify latent factors that represent similarity between users or items.

MATRIX FACTORIZATION METHODS

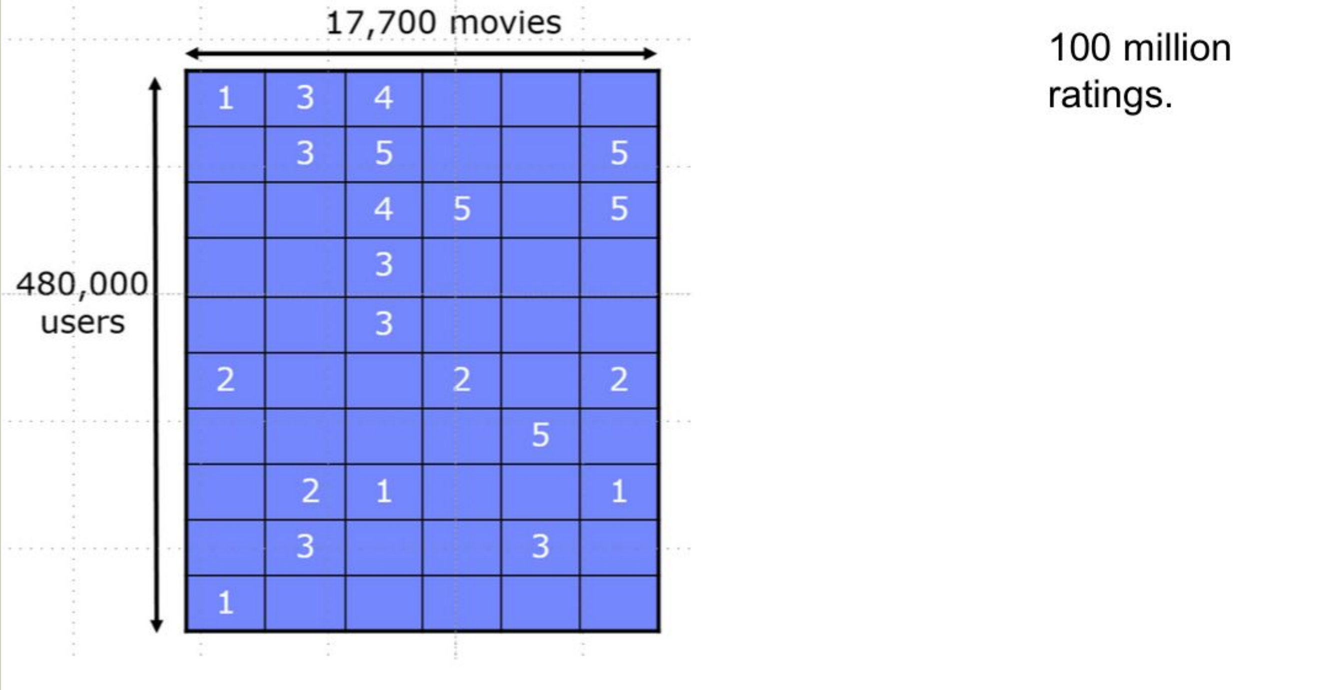
Some of the most successful realizations of latent factor models are based on *matrix factorization*. In its basic form, matrix factorization characterizes both items and users by vectors of factors inferred from item rating patterns. High correspondence between item and user factors leads to a

vector $q_i \in \mathbb{R}^f$, and each user u is associated with a vector $p_u \in \mathbb{R}^f$. For a given item i, the elements of q_i measure the extent to which the item possesses those factors, positive or negative. For a given user u, the elements of p_u measure the extent of interest the user has in items that are high on the corresponding factors, again, positive or negative. The resulting dot product, $q_i^T p_u$, captures the interaction between user u and item i—the user's overall interest in the item's characteristics. This approximates user u's rating of item i, which is denoted by r_{ui}, leading to the estimate

$$\hat{r}_{ui} = q_i^T p_u. \qquad (1)$$

- Source/Resource: https://datajobs.com/data-science-rep

NETFLIX PRIZE



- Data is over 99% sparse (there are no ratings for most of the user-item combinations
- Attributes: user, movie, ratings

Source: http://www.ics.uci.edu/~newman/courses/cs277/slides/netflix_overview.ppt.

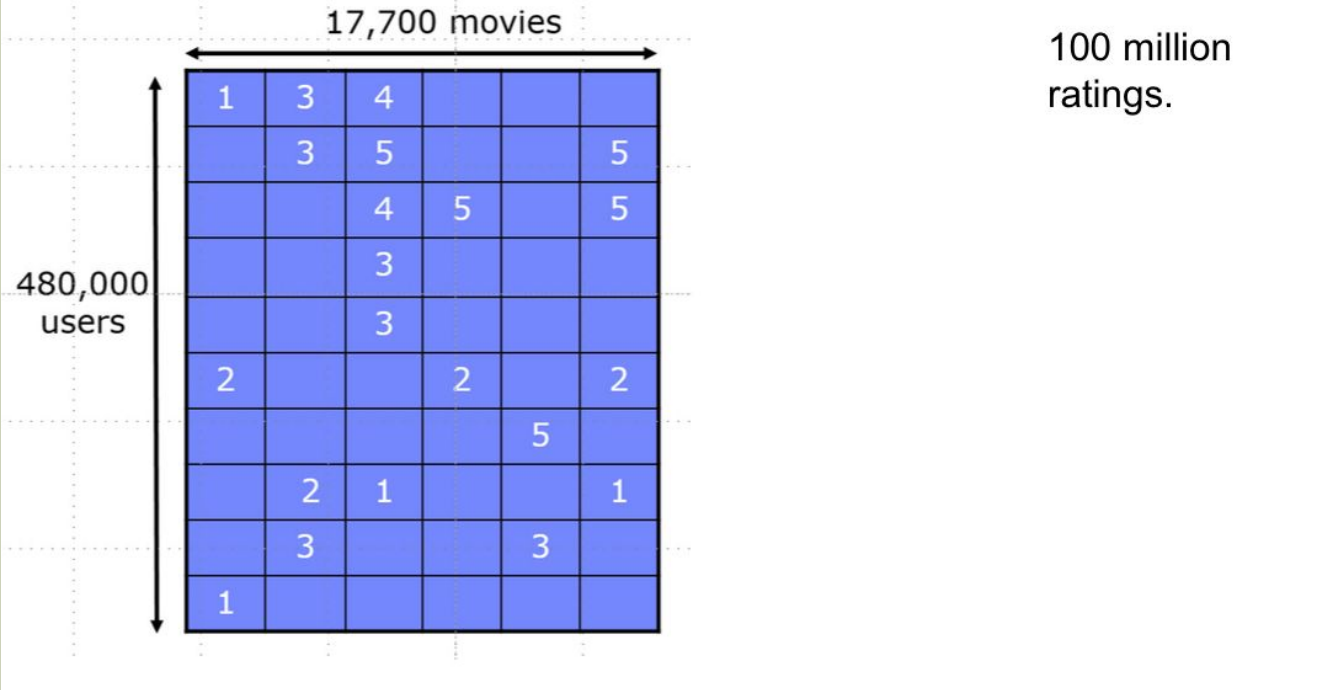
MATRIX FACTORIZATION

What kind of math do we need to make recommender systems?

Many recommendation systems rely on linear algebra. Techniques in matrix factorization and decomposition are essential to making recommender systems work at scale!

- Alternating Least Squares
- Singular Value Decomposition
- Stochastic Gradient Descent

NETFLIX PRIZE



- Data is over 99% sparse (there are no ratings for most of the user-item combinations
- Attributes: user, movie, ratings

Source: http://www.ics.uci.edu/~newman/courses/cs277/slides/netflix_overview.ppt.

SPARSITY

Sparsity = 1 – [Ratings / Users * Items)]

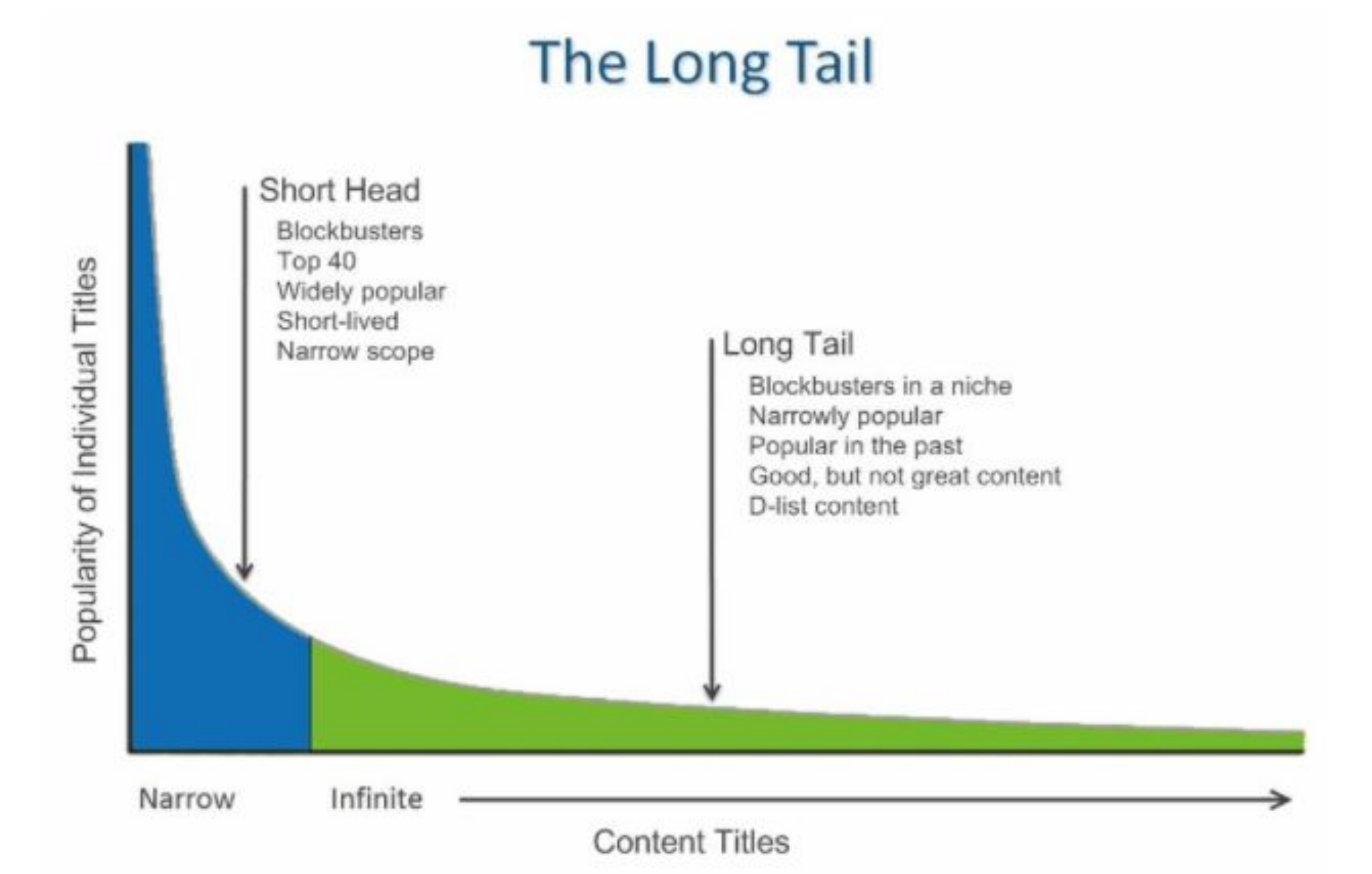
How do we address the issue of sparsity in RecSys design?

COLD START PROBLEM – a notorious issue with collaborative filtering.
(note how recommendations improve with user data)

This is an issue both with new users and new items.

Hybrid Systems? This is an umbrella term for multiple techniques that either mix strategies or using explicit features or preferences of the users or items.

THE LONG TAIL



Source: *Learning Solutions Magazine*, David Wilkins, August 17, 2009

ETHICS IN RECOMMENDER SYSTEMS

What role do they play in creating ‘filter bubbles”:

1. reinforce view points
2. increase polarization
3. lead users to extreme and conspiratorial content
4. perpetuate bias through algorithmic discrimination

“YouTube, the Great Radicalizer” by Zeynep Tufekci

https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Enroll in 643!

We will be offering Special Topics in Recommender Systems as a 7-week summer course this year. If you are interested and want to see a syllabus, you can email me at:

david.stern@sps.cuny.edu
david.c.stern@gmail.com