Announcements

Assignments

- HW8 (written + programming)
  - Due Thu 4/9, 11:59 pm

Introduction to Machine Learning

Recommender Systems

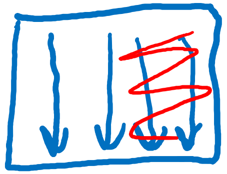
Instructor: Pat Virtue

Background: Low Dimensional Embeddings

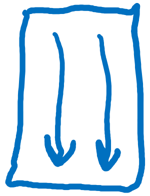
PCA: What did we do?

$$X \in \mathbb{R}^{N\times M}$$

$$\mathrm{eig}(X^T X) \rightarrow V \in \mathbb{R}^{M\times M}$$



$$\rightarrow V_k \in \mathbb{R}^{M\times K}$$



$\vec{x} \in \mathbb{R}^M$ $\qquad \vec{z} = V_k^T \vec{x} \qquad \vec{z} \in \mathbb{R}^K \qquad K < M$

$$\vec{x}' = V_k \vec{z}$$

$$\vec{x}' = V_k V_k^T \vec{x} \qquad \min_V \|\vec{x} - \vec{x}'\|$$

Background: Low Dimensional Embeddings

Why might low dimensional embeddings be useful?

- Example: MNIST digit classification with nearest neighbor

$X_{train} \in \mathbb{R}^{N \times 784}$

$\vec{x} \in \mathbb{R}^{784}$

$\|\vec{x} - \vec{x}^{(i)}\|_2$

$\hat{y} = y^{(i)}$

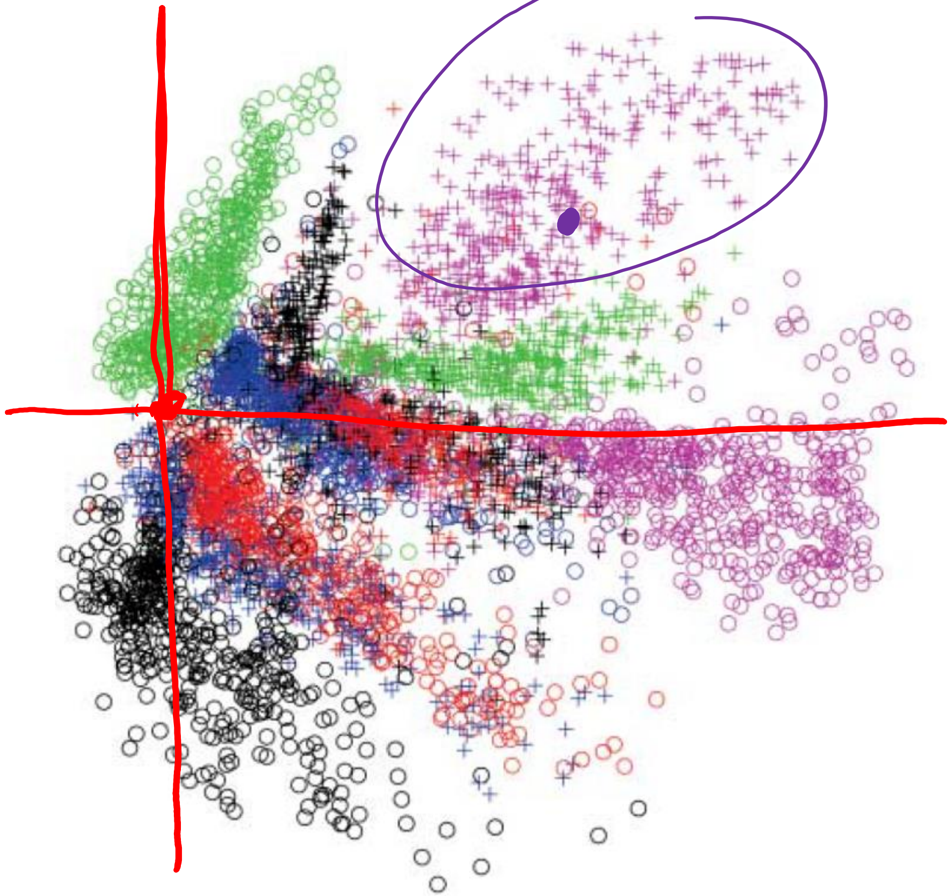
Background: Low Dimensional Embeddings

Why might low dimensional embeddings be useful?

- Example: MNIST digit classification with nearest neighbor

x

$$z = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

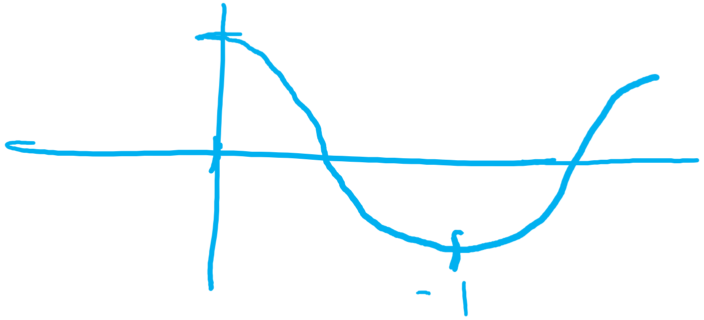


$$\|z - z^{(i)}\|_2$$

$$\hat{y} = y^{(i)}$$

Image: Hinton & Salakhutdinov. *Science* 313.5786 (2006): 504-507.

Background: Measure of Similarity



We've been using Euclidean distance

- $d(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_2$

Cosine similarity

- To vectors are similar if the angle between them is small
- $d(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^T\boldsymbol{z}$ $= x \cdot z$

$= \|x\| \, \|z\| \cos \theta$

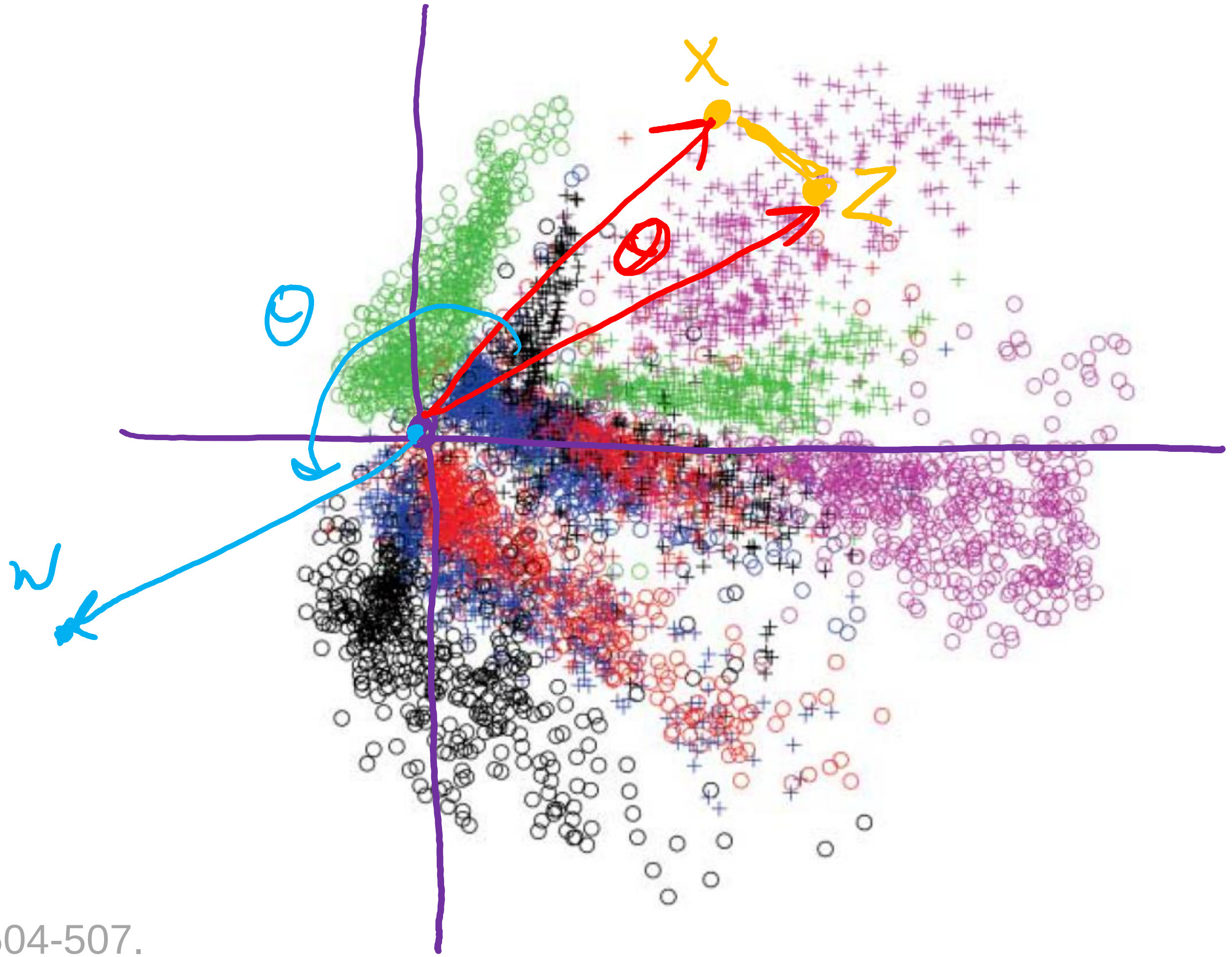


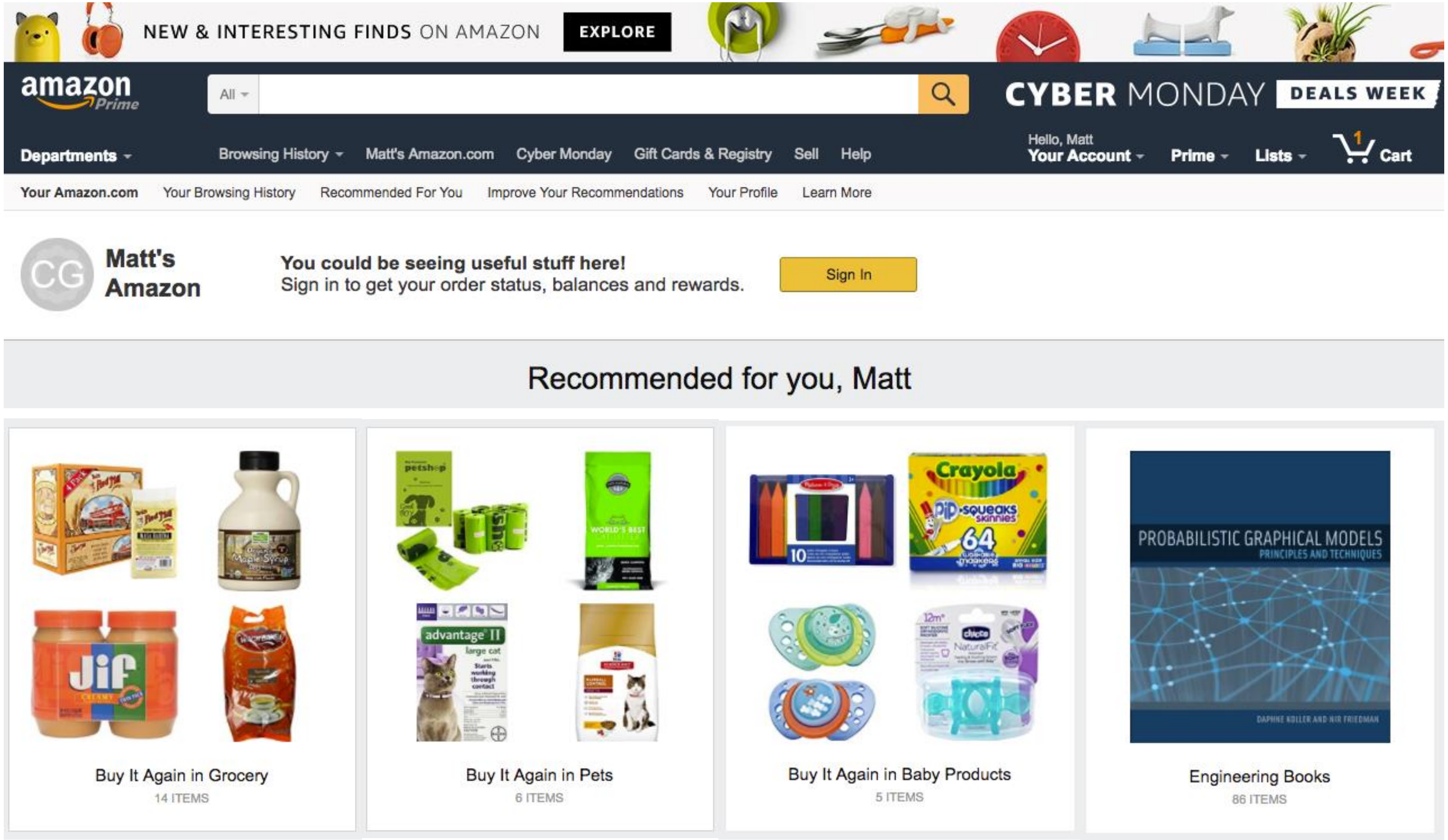
Image: Hinton & Salakhutdinov. *Science* 313.5786 (2006): 504-507.

Recommender Systems

A Common Challenge:

- Assume you're a company selling **items** of some sort: movies, songs, products, etc.
- Company collects millions of **ratings** from **users** of their **items**
- To maximize profit / user happiness, you want to **recommend** items that users are likely to want

Recommender Systems



Slide credit: CMU MLD Matt Gormley

Recommender Systems

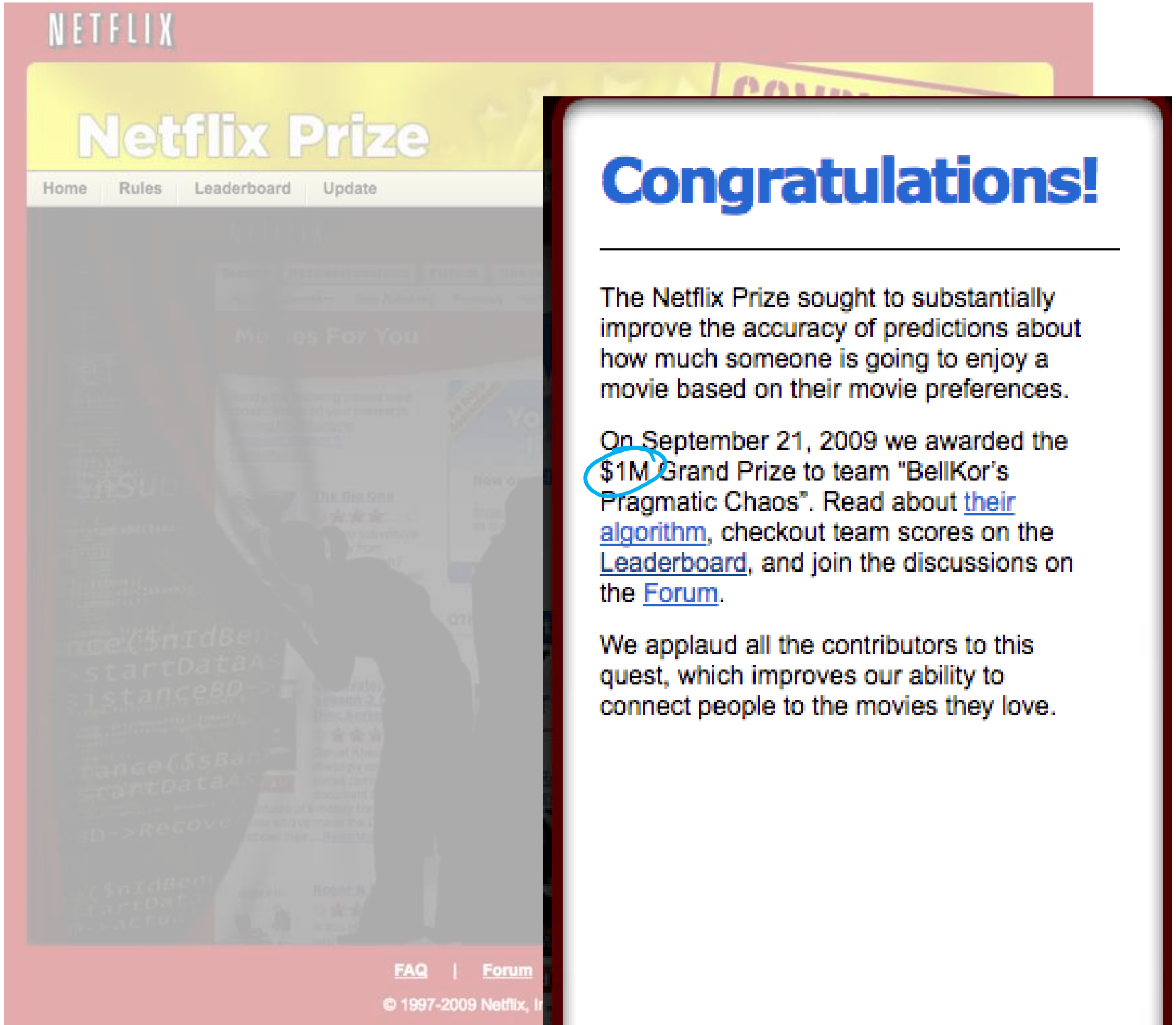
$u^{(i)}$ $v^{(j)}$

$\downarrow$

R_{ij}

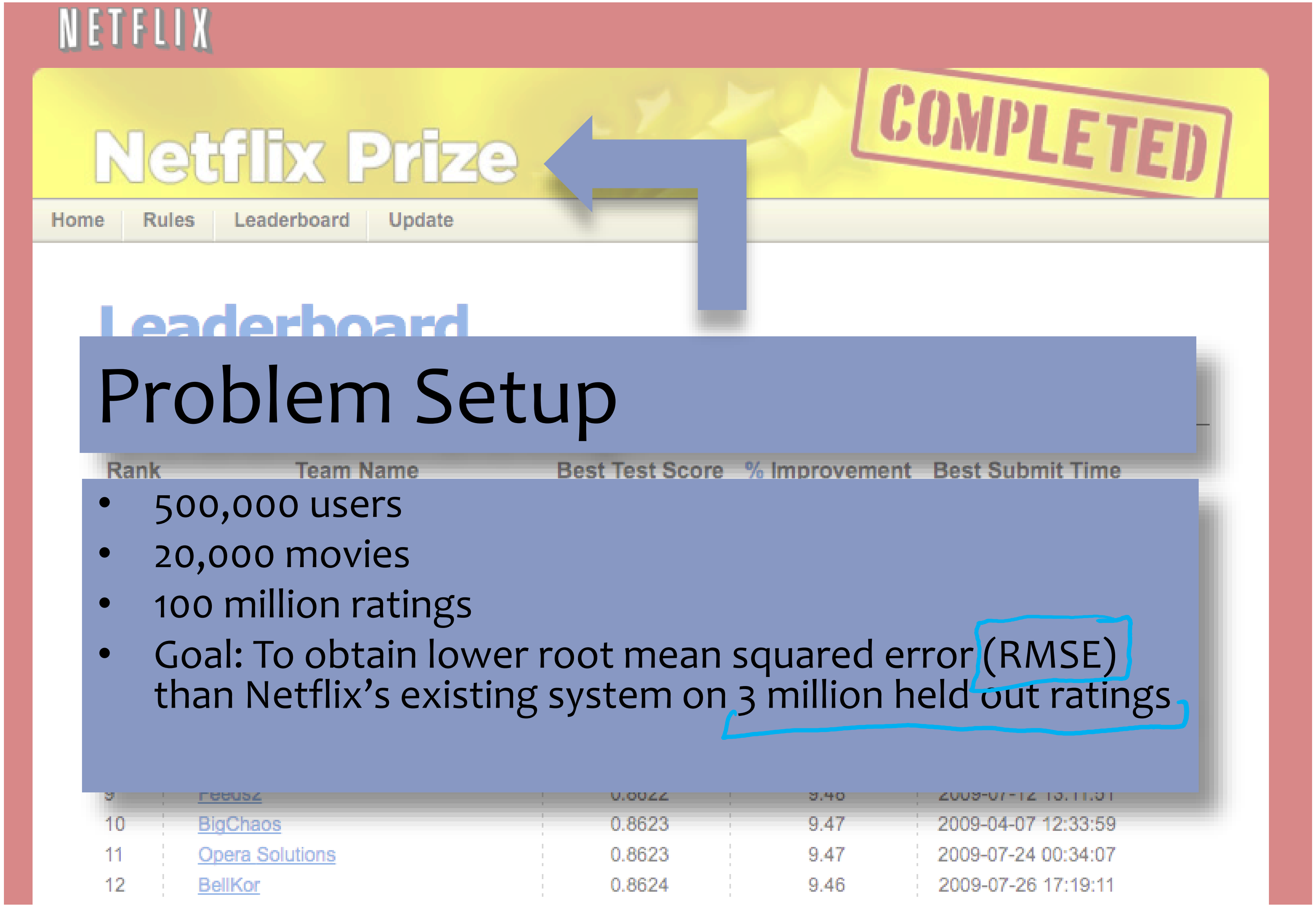
Slide credit: CMU MLD Matt Gormley

Recommender Systems



Slide credit: CMU MLD Matt Gormley

Recommender Systems



Problem Setup

- 500,000 users
- 20,000 movies
- 100 million ratings
- Goal: To obtain lower root mean squared error (RMSE) than Netflix's existing system on 3 million held out ratings

Slide credit: CMU MLD Matt Gormley

Recommender Systems

- **Setup:**
  - Items:
    movies, songs, products, etc.
    (often many thousands)
  - Users:
    watchers, listeners, purchasers, etc.
    (often many millions)
  - Feedback:
    5-star ratings, not-clicking 'next',
    purchases, etc.
- **Key Assumptions:**
  - Can represent ratings numerically as a user/item matrix
  - Users only rate a small number of items (the matrix is sparse)

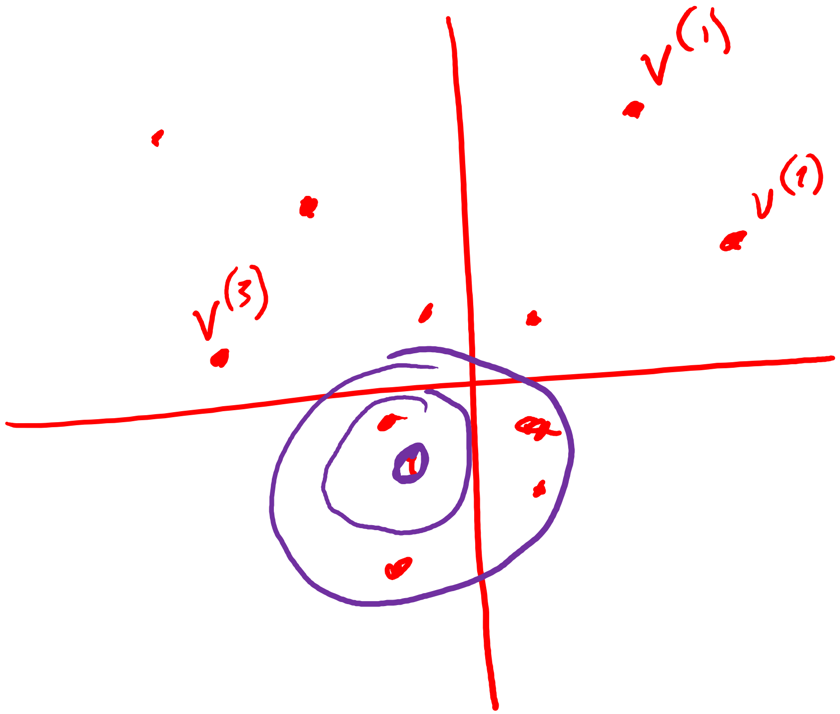
	Doctor Strange	Star Trek: Beyond	Zootopia
Alita	1	?	5
BB-8	3	4	
C-3Po	3	5	2

Different Approaches

Item-based (*Content filtering*)

$v^{(i)} \in R^M \Longrightarrow R^k$

- Features about each item
- Given an item, other "close" items have similar values
- e.g. Pandora.com, music genome project



Different Approaches

Item-based (*Content filtering*)

- Features about each item
- Given an item, other “close” items have similar values
- e.g. Pandora.com, music genome project

User-based

- Features about each user
- Given a user, other “close” users have similar preferences
- *Market segmentation*

Learning user-item relationship

- Can be done without features on either user or item
- Collaborative filtering techniques

COLLABORATIVE FILTERING

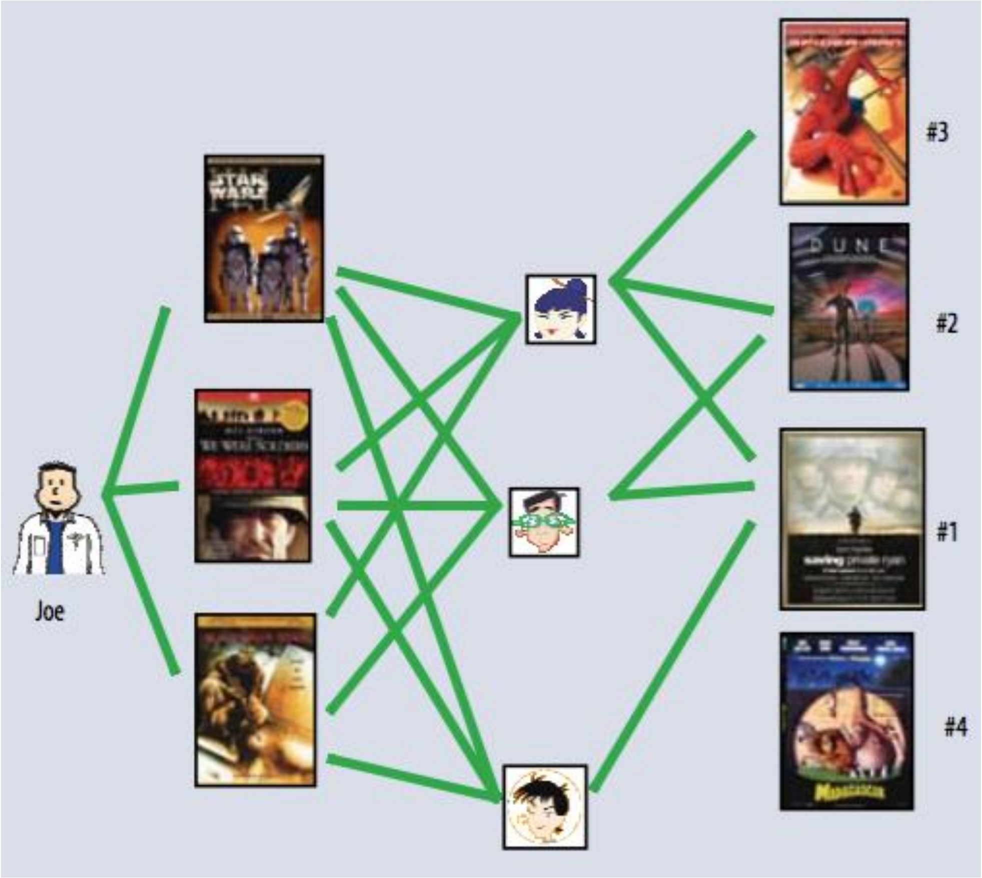
Collaborative Filtering

- **Everyday Examples of Collaborative Filtering...**
  - Bestseller lists
  - Top 40 music lists
  - The "recent returns" shelf at the library
  - Unmarked but well-used paths thru the woods
  - The printer room at work
  - "Read any good books lately?"
  - …
- **Common insight:** personal tastes are correlated
  - If Alita and BB-8 both like X and Alita likes Y then BB-8 is more likely to like Y
  - especially (perhaps) if BB-8 knows Alita

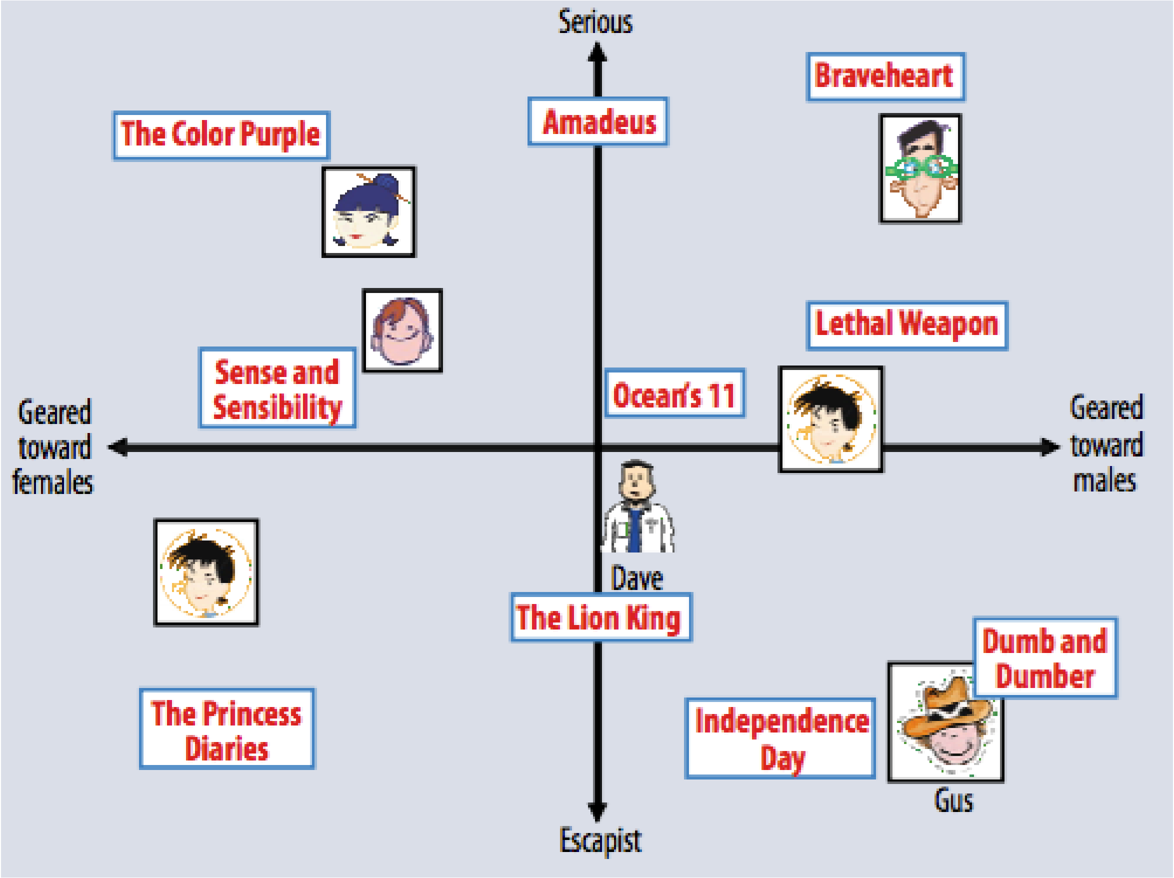
Slide from William Cohen

Two Types of Collaborative Filtering

1. Neighborhood Methods



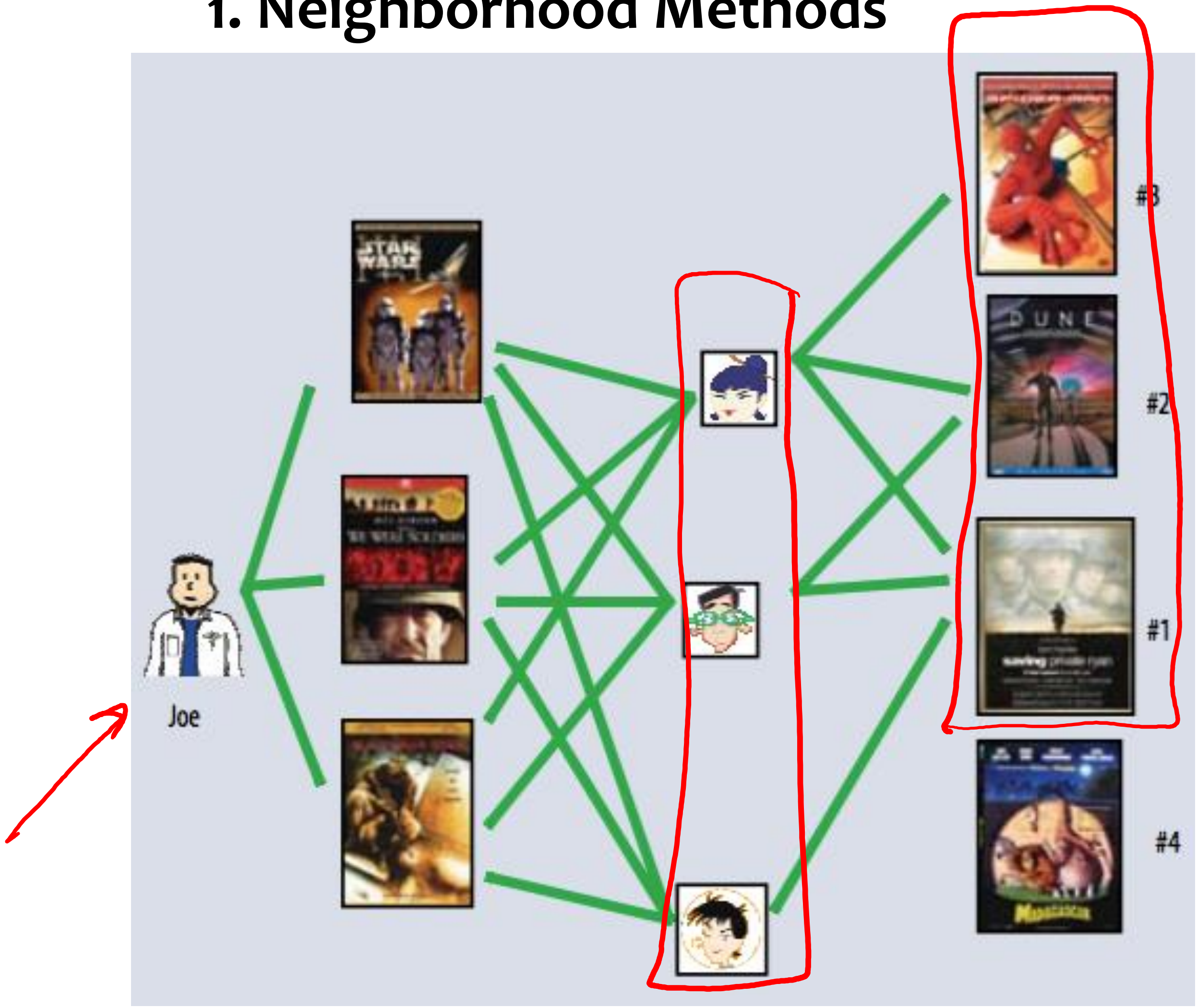
2. Latent Factor Methods



Figures from Koren et al. (2009)

Two Types of Collaborative Filtering

1. Neighborhood Methods



In the figure, assume that a green line indicates the movie was **watched**

Algorithm:

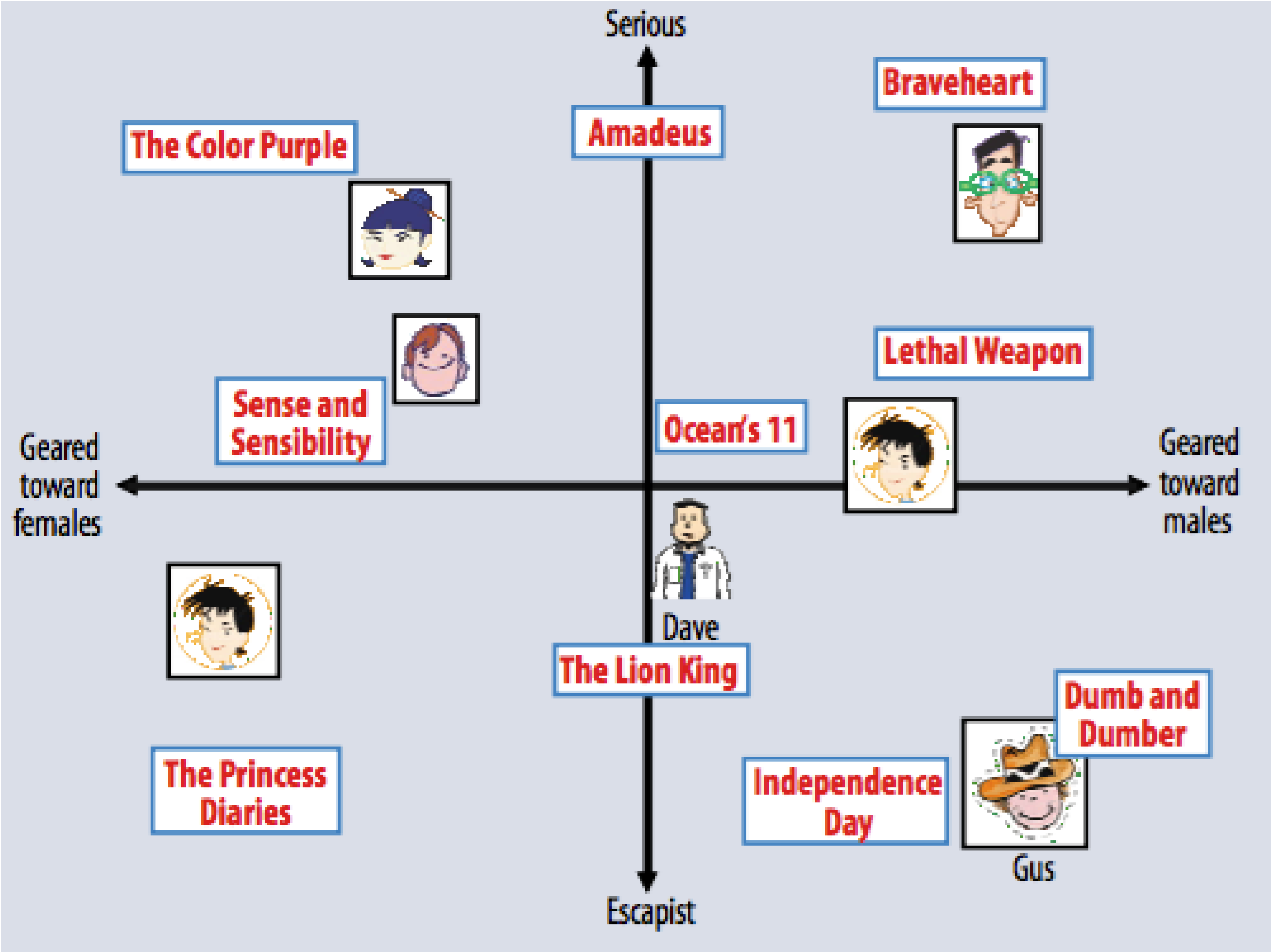
1. **Find neighbors** based on similarity of movie preferences
2. **Recommend** movies that those neighbors watched

Figures from Koren et al. (2009)

Two Types of Collaborative Filtering

2. Latent Factor Methods

- Assume that both movies and users live in some **low-dimensional space** describing their properties
- **Recommend** a movie based on its **proximity** to the user in the latent space
- **Example Algorithm:** Matrix Factorization



Figures from Koren et al. (2009)

Recommender System: Matrix Factorization

Learning to map items and users to the same lower dimensional space

$u^{(i)} \in \mathbb{R}^K$

$v^{(j)} \in \mathbb{R}^K$

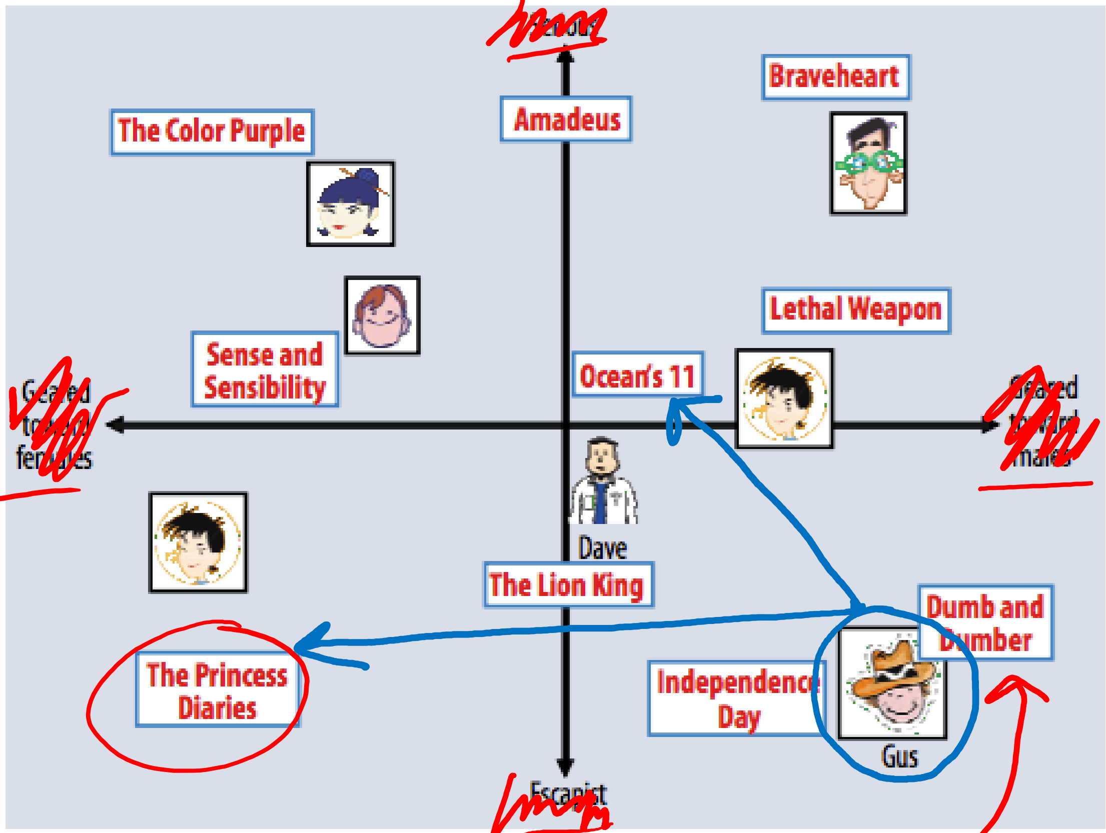
Similarity

- Euclidean $\|\vec{u}^{(i)} - \vec{v}^{(j)}\|_2$
- Cosine $\vec{u}^{(i)T} \vec{v}^{(j)}$

Label

R_{ij} rating user i on item j

$K=2$



Figures from Koren et al. (2009)

Recommender System: Matrix Factorization

Optimization

$$\hat{r}_{ij} = \vec{u}^{(i)T} v^{(j)}$$

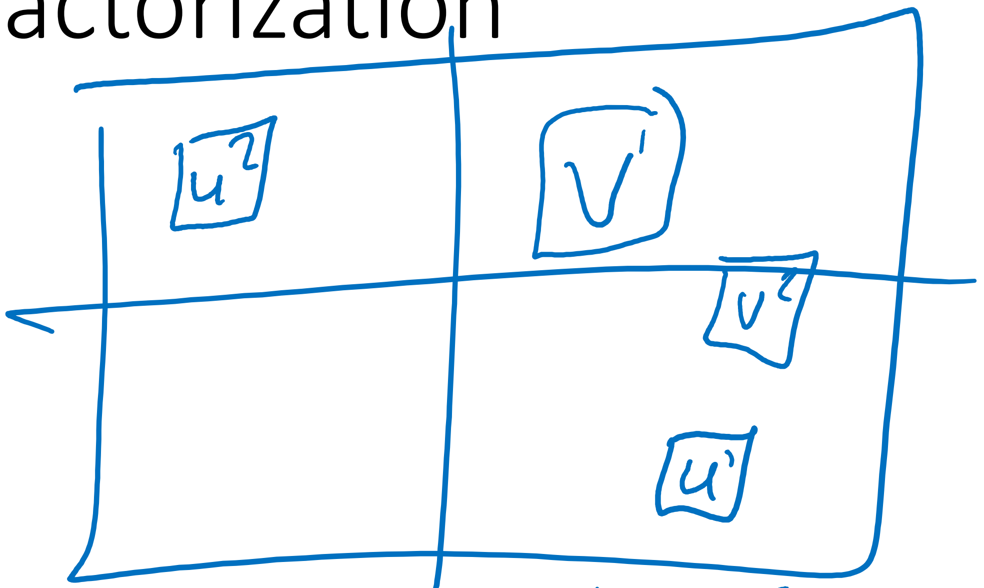
$$\min_{U,V} \sum_{ij} \left(R_{ij} - u^{(i)T} v^{(j)}\right)^2$$

$$\min_{UV} \left\| R - \underbrace{UV^T} \right\|_F^2$$

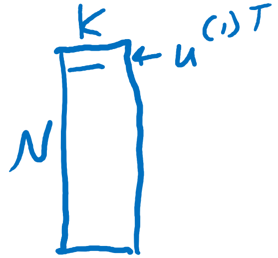
$$X^T X \approx V_k \Lambda V_k^T$$

$$X = USV^T$$

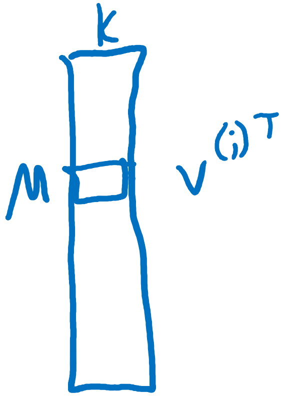
$$\tilde{X} = U_k \underline{S_k} V_k^T$$



$U \in \mathbb{R}^{N \times K}$



$V \in \mathbb{R}^{M \times K}$



$R \in \mathbb{R}^{N \times M}$

Recommender System: Matrix Factorization

Sparse labels ☹

R

M

N

	Doctor Strange	Star Trek: Beyond	Zootopia
Alita	1	?	5
BB-8	3	4	?
C-3PO	3	5	2

Example: MF for Netflix Problem

R

		NERO	JULIUS CAESAR	CLEOPATRA	SLEEPLESS IN SEATTLE	PRETTY WOMAN	CASABLANCA
HISTORY	1	1	1	1	0	0	0
	2	1	1	1	0	0	0
	3	1	1	1	0	0	0
BOTH	4	1	1	1	1	1	1
ROMANCE	5	-1	-1	-1	1	1	1
	6	-1	-1	1	1	1	1
	7	-1	-1	-1	1	1	1

≈

U

	HISTORY	ROMANCE
1	1	0
2	1	0
3	1	0
4	1	1
5	-1	1
6	-1	1
7	-1	1

X

V^T

	NERO	JULIUS CAESAR	CLEOPATRA	SLEEPLESS IN SEATTLE	PRETTY WOMAN	CASABLANCA
HISTORY	1	1	1	0	0	0
ROMANCE	0	0	1	1	1	1

(a) Example of rank-2 matrix factorization

$\underline{U_k}$ $\underline{V_k^T}$

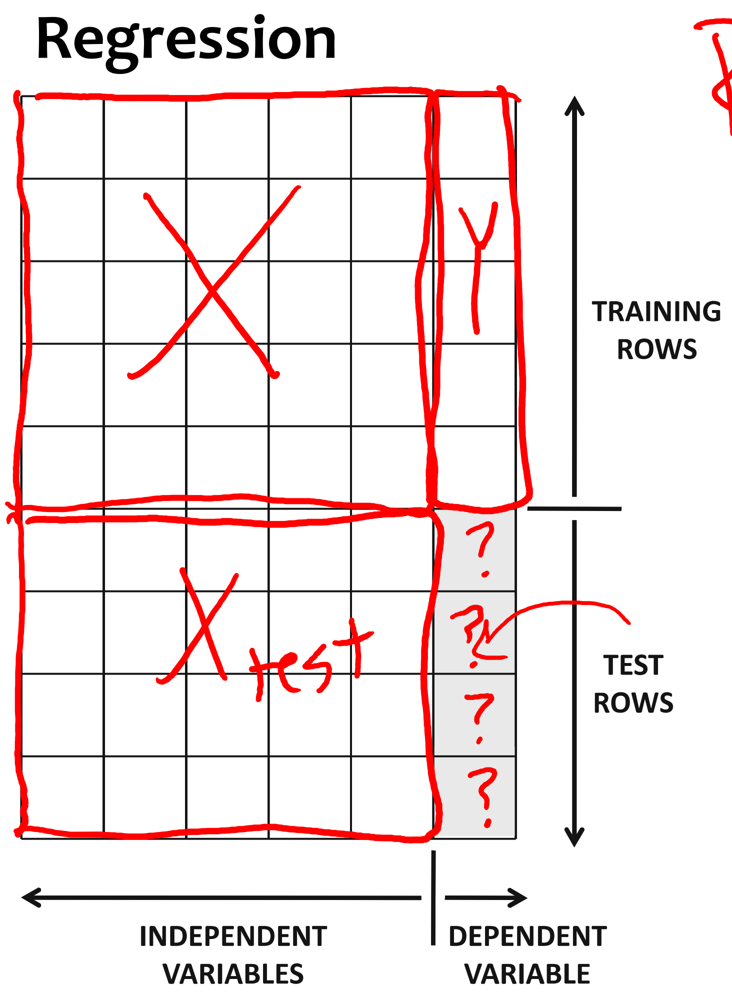
R

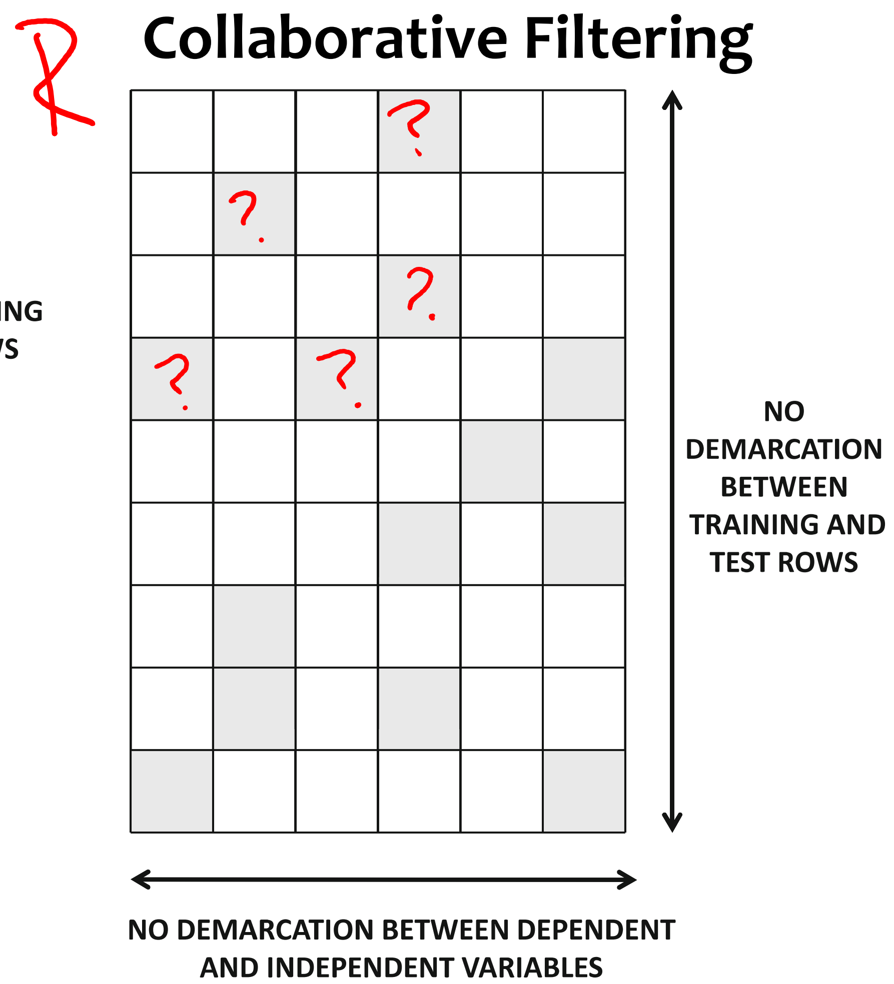
		NERO	JULIUS CAESAR	CLEOPATRA	SLEEPLESS IN SEATTLE	PRETTY WOMAN	CASABLANCA
HISTORY	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	0	0	0	0	0
BOTH	4	0	0	-1	0	0	0
ROMANCE	5	0	0	-1	0	0	0
	6	0	0	1	0	0	0
	7	0	0	-1	0	0	0

(b) Residual matrix

Figures from Aggarwal (2016)

Regression vs. Collaborative Filtering





Figures from Aggarwal (2016)

Matrix Factorization: SVD

We can use SVD, but as you'll see it has issues

$SVD(R) \rightarrow U, S, V$

$\rightarrow \hat{U}_k = U_k S_k, \quad \tilde{V}_k = V_k$

$\tilde{R} = \hat{U}_k \tilde{V}_k^T$

$\min_{U V} \|R - \tilde{R}\|$

U, V orth

	Doctor Strange	Star Trek: Beyond	Zootopia
Alita	1	?	5
BB-8	3	4	?
C-3Po	3	5	2

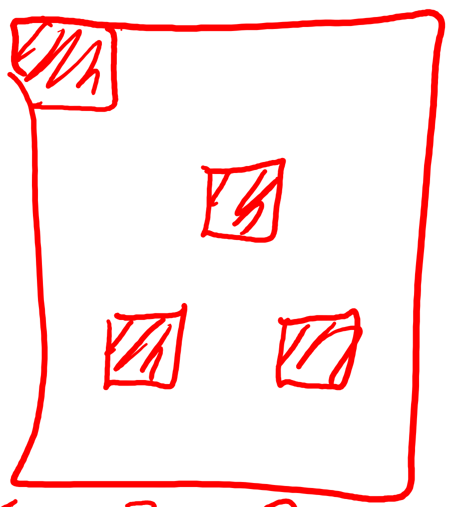
Alita	-1	0	1
BB-8	0	0,5	0
C-3Po	0	1	-0,5

Matrix Factorization: SGD

Objective function using only the labels we have

$$\min_{U,V} \|R - UV^T\|_F^2$$

$$\min_{U V} \sum_{i,j \in S} \left(R_{ij} - u^{(i)T} v^{(j)}\right)^2$$



$S = \{i,j\}$ R_{ij} is known

Piazza Poll 1

Is the following optimization a quadratic optimization?

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{i,j \in \mathcal{S}} \left(R_{ij} - \boldsymbol{u}^{(i)T} \boldsymbol{v}^{(j)} \right)^2$$

A. Yes

B. Calamity

C. No

Matrix Factorization: SGD

Method of *alternating minimization*

$$\min_{\boldsymbol{U},\boldsymbol{V}} J(\boldsymbol{U},\boldsymbol{V}), \quad J(\boldsymbol{U},\boldsymbol{V}) = \sum_{i,j\in\mathcal{S}}\left(R_{ij} - \boldsymbol{u}^{(i)^T}\boldsymbol{v}^{(j)}\right)^2$$

① $\min_{U} J(U, \underline{\underline{V}})$ $\quad U^{t+1} \leftarrow U^{t} - \alpha \nabla_U J$

② $\min_{V} J(\underline{\underline{U}}, V)$ $\quad V \leftarrow V - \alpha \nabla_V J$

Matrix Factorization: SGD

Method of *alternating minimization*

$$\min_{\boldsymbol{U},\boldsymbol{V}} J(\boldsymbol{U},\boldsymbol{V}) \qquad J(\boldsymbol{U},\boldsymbol{V}) = \sum_{i,j\in\mathcal{S}} \underbrace{\left(R_{ij} - \boldsymbol{u}^{(i)^T}\boldsymbol{v}^{(j)}\right)}^2$$

$$J_{ij}(U,V) = \underbrace{\left(R_{ij} - u^{i^T} v^j\right)}^2$$

$$\nabla_{ij\,u} J = -\left(R_{ij} - u^{i^T} v^j\right) v^j$$

① $\min_{u} J_{ij}(U,V)$

② $\min_{V} J_{ij}$

Matrix Factorization: SGD

Add regularization to avoid overfitting

$$\min_{\boldsymbol{U},\boldsymbol{V}} J(\boldsymbol{U},\boldsymbol{V}) \qquad J(\boldsymbol{U},\boldsymbol{V}) = \sum_{i,j\in\mathcal{S}} \left[\left(R_{ij} - {\boldsymbol{u}^{(i)}}^T \boldsymbol{v}^{(j)}\right)^2 + \lambda \|v^i\|_2^2 + \lambda \|u^i\|_2^2 \right]$$

Summary

Recommender systems solve many **real-world** (*large-scale) **problems**

Collaborative filtering by Matrix Factorization (MF) is an **efficient** and **effective** approach

(SVD for MF is a bit broken)

MF is just another example of a **common recipe**:

1. define a model
2. define an objective function
3. optimize with SGD

Optimization

- Need alternating minimization
- Add regularization to avoid overfitting