Recommender Systems Designed for Yelp.com

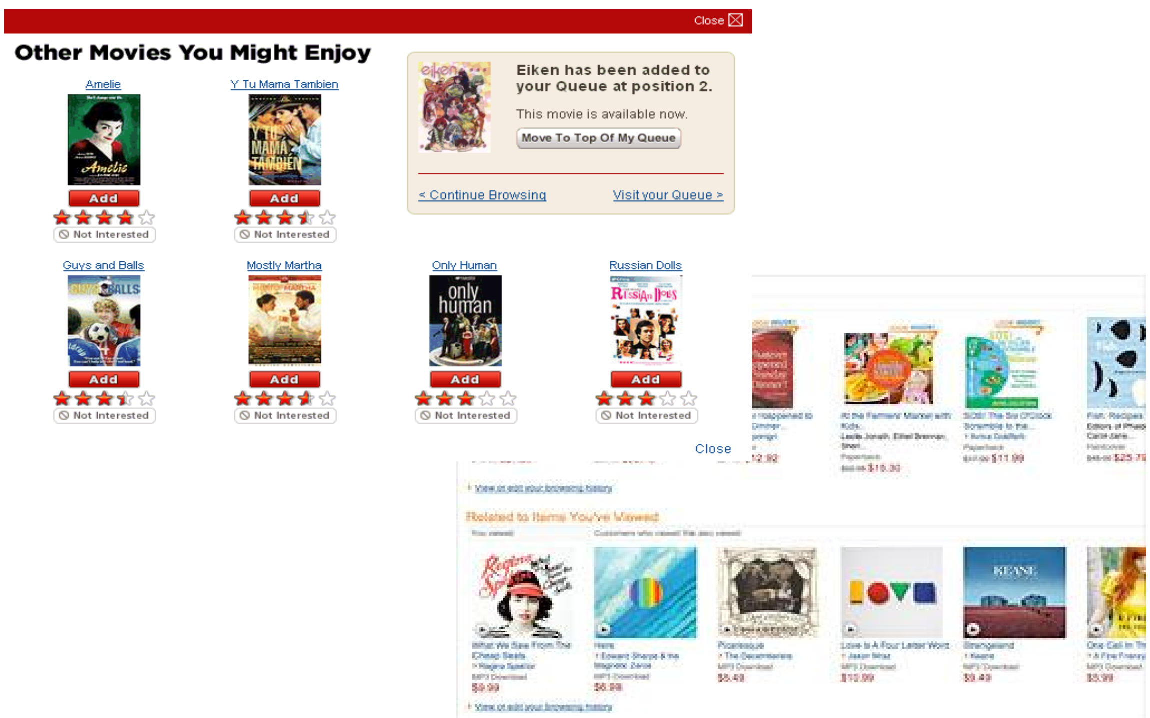
Naomi Carrillo
Idan Elmaleh
Rheanna Gallego
Zack Kloock
Irene Ng
Jocelyne Perez
Michael Schwinger
Ryan Shiroma

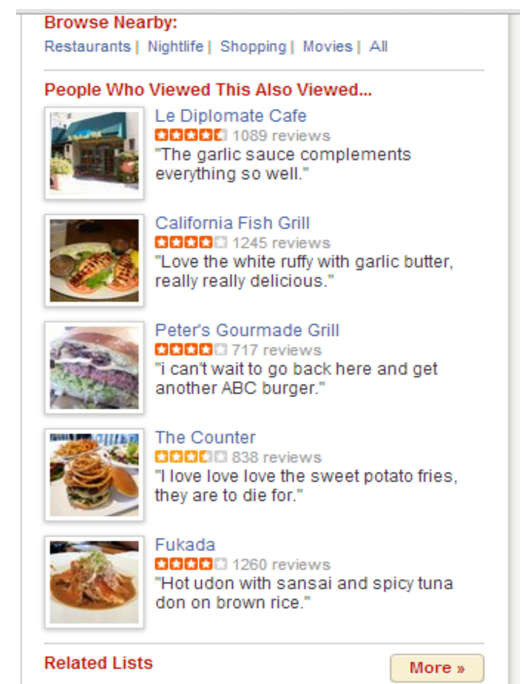
Outline

- Introduction
- Data
- Methods
- Other Findings
- Results

Intro

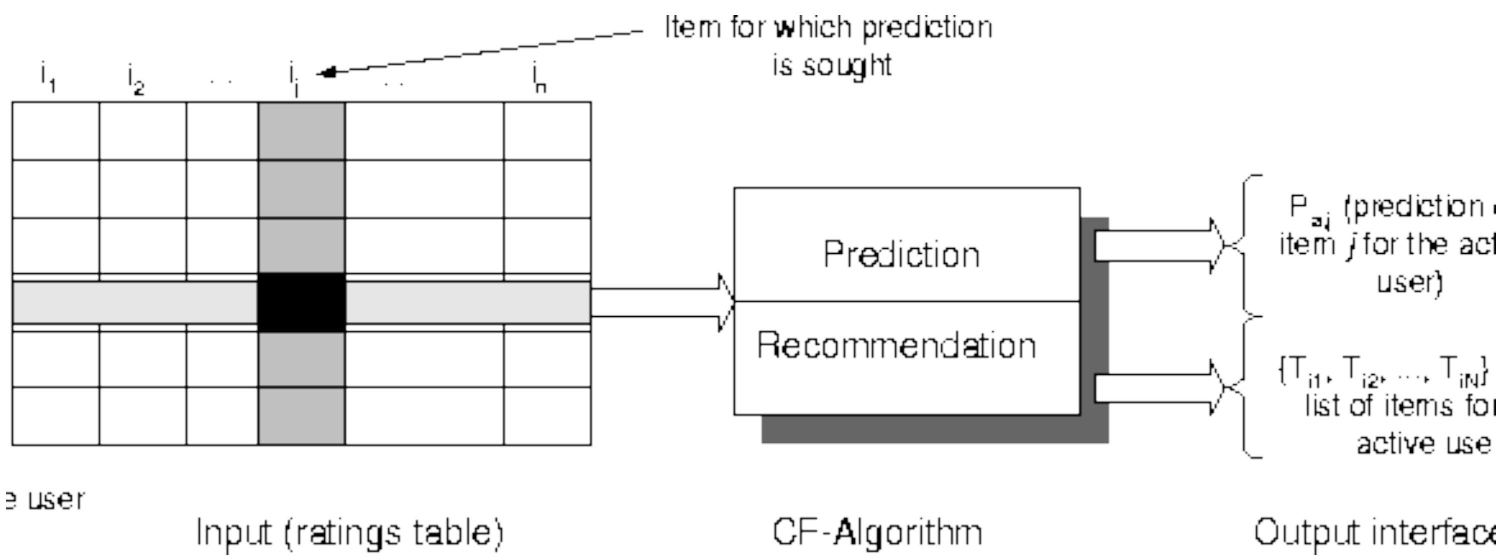
- Recommender systems: filtering system meant to ‘recommend’ items that may be of interest to the user
- Used often in electronic commerce





Intro

- Each entry a_{ij} represents ratings of i_{th} user for j_{th} rating
- Send information through prediction method
- Either predict user's rating for item j or list of recommended items for user j



Background

RecSys Challenge 2013: Yelp Business Rating Prediction

- Competition created by Yelp on Kaggle
- Asks competitors to create models and algorithms for predicting user ratings for businesses

- Graded on accuracy and RMSE

  N = # of review ratings to predict
  y_{pred} = predicted rating for review j
  y_{ref} = actual rating for review j

$$\mathrm{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

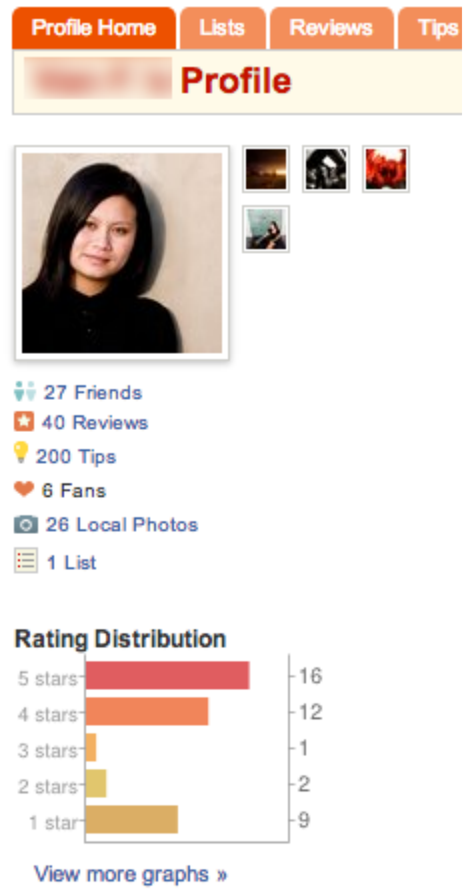
- $300 prize for 1st place

Problems specific to Yelp data

1. Sparsity
   99.9% empty

2. Cold Start
   large number of unknown users/businesses

3. ‘Grey Sheep’
   unpredictable ratings



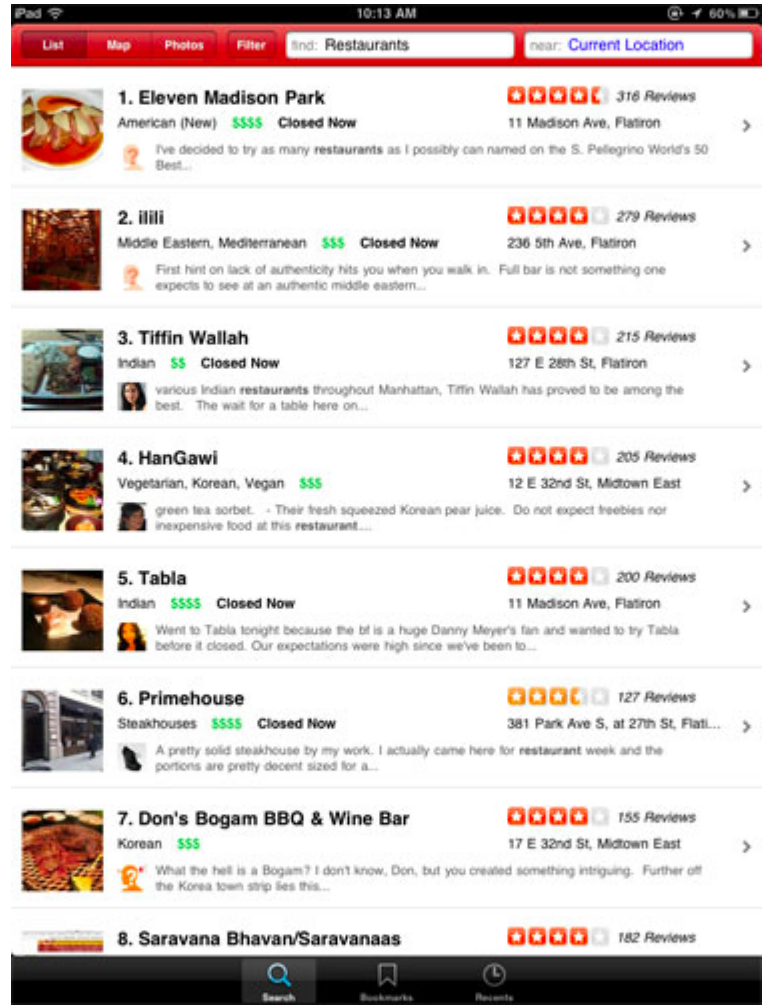
Available Data

In training set:

- 11,537 businesses
- 8,282 check-in sets
- 43,873 users
- 229,907 reviews

In test set:

- 1,205 businesses
- 734 check-in sets
- 5,105 users
- 22,956 reviews to predict



Information Known About Businesses

For 11,537 businesses we know:

Business ID

Categories

City

Full Address

Latitude & Longitude

Name

Neighborhood

Open

Review Count

Stars, State & Type

Information Known About Users

For 43,873 users, we know their:

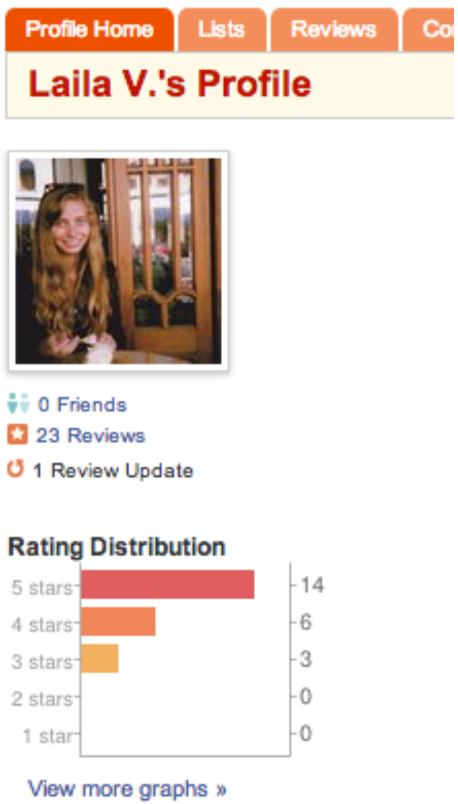
- Average Stars
- Name
- Review Count
- Type
- User ID
- Votes (useful, funny, cool)

For 4 users, we know all the above except their average stars.

For 2,104 users we have nothing for them.

Types of predictions

1. Business and User ratings are known
2. Business or User ratings are known
3. Both are unknown



The Big Picture

What are we looking at and how to make sense of it?

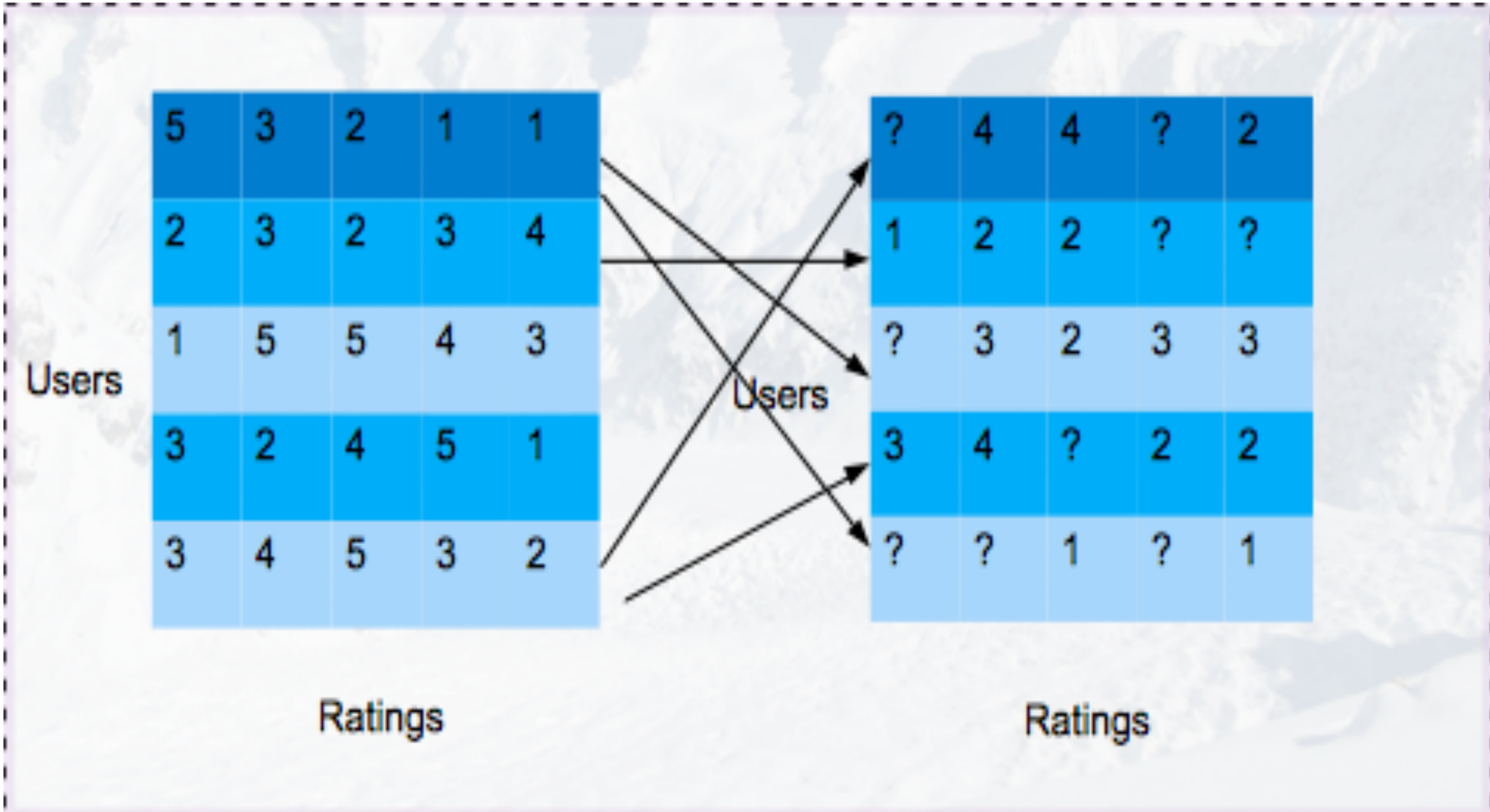
$$
\begin{array}{cc}
 & \begin{array}{cccccc} \text{col0} & \text{col1} & \text{col2} & \text{col3} & \text{col4} & \text{col5} \end{array} \\
\begin{array}{c} \text{row0} \\ \text{row1} \\ \text{row2} \\ \text{row3} \\ \text{row4} \\ \text{row5} \end{array} &
\begin{bmatrix}
15 & 0 & 0 & 22 & 0 & -15 \\
0 & 11 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & -6 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
91 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 28 & 0 & 0 & 0
\end{bmatrix}
\end{array}
$$

What to do with all this Information?

- Now that we know our data and all the information that is given to us what should we do? What method of predicting unknown ratings should we use?
- We want to get the most accurate ratings for each user on a business that they have never gone to before.
- What we used to help me do this is a Nearest Neighbor Method.

Row	Col	Value
1	1	11
1	2	15
1	3	5
4	3	20
3	6	7

Nearest Neighbor Method



Variables

- Gender- RMSE 1.1535 on validation set
- Average Stars – RMSE .9656 on validation set
- Review Count – RMSE 1.1689 on validation set

- How we used these variables was in combination with the Nearest Neighbor Method.
- Taking the mean of all 5 values would produce the best RMSE.

Problem with Euclidean Distance Alone

Comedy or Science Fiction?

Weighted Similarity-Jaccard Index

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

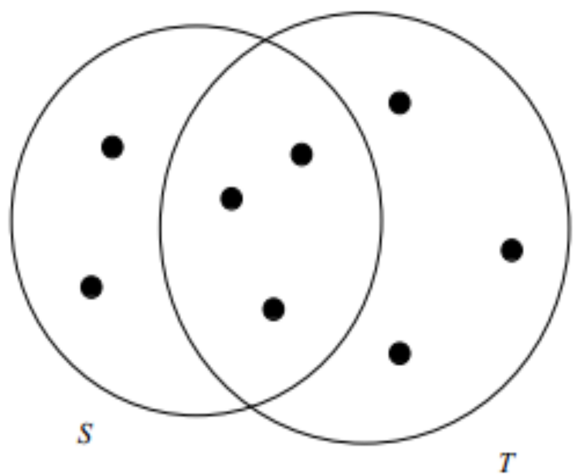


Figure 3.1: Two sets with Jaccard similarity 3/8

Problem with Jaccard Index Alone

Weighted Similarity-Jaccard Index

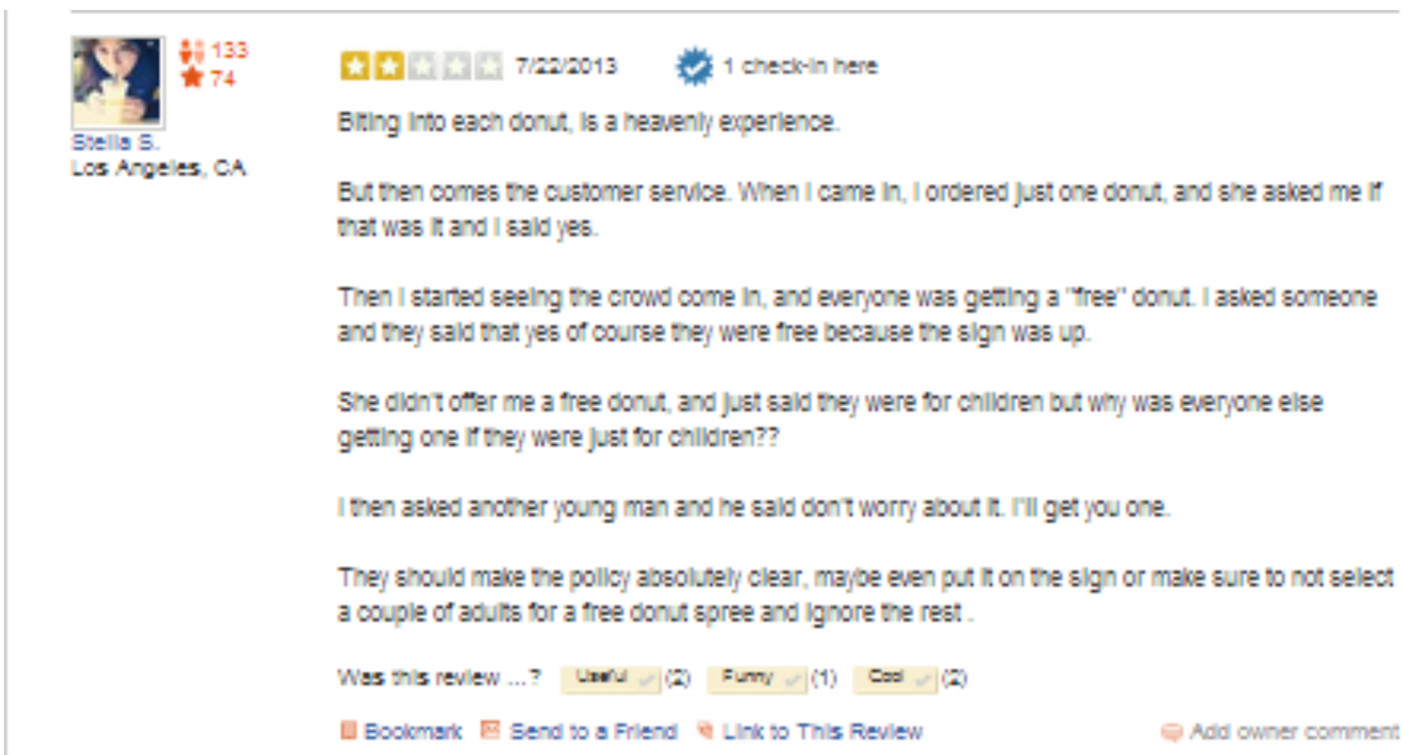
- Idea proposed by Laurent Candillier, Frank Meyer, Francoise Fessant
  - "Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems"
  - Tested on Netflix and MovieLens Ratings
- Product of Euclidean Distance and Jaccard Index
  - Combination of both
  - Gives a weight to the Euclidean Distance

Weighted Similarity:Results

- RMSE score of 1.32948 on Kaggle
- Higher Than Both User and Business Mean and Global Benchmark

Weighted Average-Funny, Useful, Cool Ratings

- Works for user reviews with Funny, Cool, Useful Ratings
  - User star rating is give a higher weight



Weighted Average:Results

- RMSE score of 1.28893 on Kaggle
- Lower Than Both User and Business Mean and Global Benchmark

Why are distance measures difficult with the Yelp Data set?

- Not having enough users who rate the same business to compare neighbors

	Business 1	Business 2	Business 3
User 1	2	4	**??**
User 2	--	4	5
User 3	5	--	3
User 4	--	--	5

Imputation

- Create psuedo-ratings from each user's personal average ratings

	Business 1	Business 2	Business 3
User 1	2	4	**<u>??</u>**
User 2	<u>3.8</u>	4	5
User 3	5	<u>2.5</u>	3
User 4	<u>4.1</u>	<u>4.1</u>	5

Compute Similarity

	Business 1	Business 2	Business 3
User 1	2	4	**<u>??</u>**
User 2	<u>3.8</u>	4	5
User 3	5	<u>2.5</u>	3
User 4	<u>4.1</u>	<u>4.1</u>	5

<u>User 2:</u> $d_{12} = (2 - 3.8)^2 + (4 - 4)^2 = 3.24$

User 4: $d_{14} = (2 - 4.1)^2 + (4 - 4.1)^2 = 4.42$

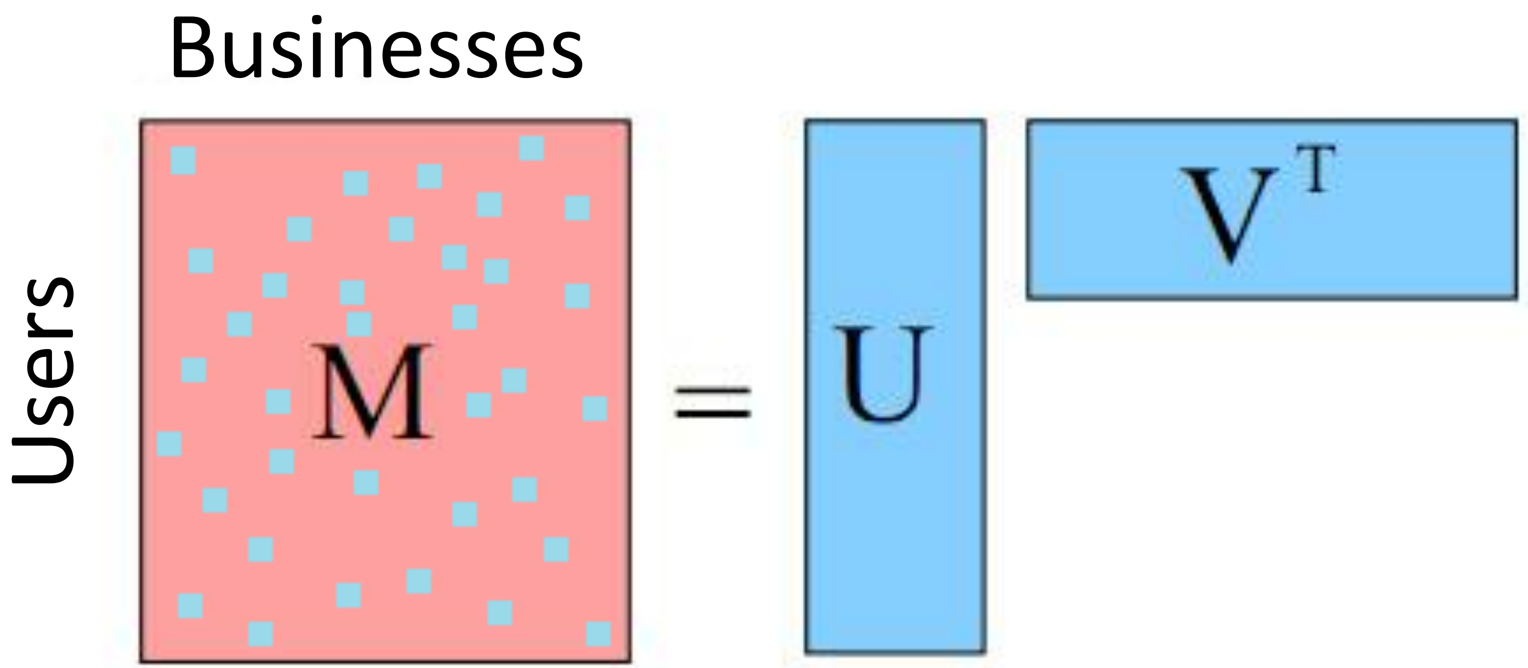
User 3: $d_{13} = (2 - 5)^2 + (4 - 2.5)^2 = 11.25$

Prediction

	Business 1	Business 2	Business 3
User 1	2	4	**<u>5</u>**
User 2	<u>3.8</u>	4	5
User 3	5	<u>2.5</u>	3
User 4	<u>4.1</u>	<u>4.1</u>	5

Result:
<u>RMSE ~1.249</u> on Kaggle

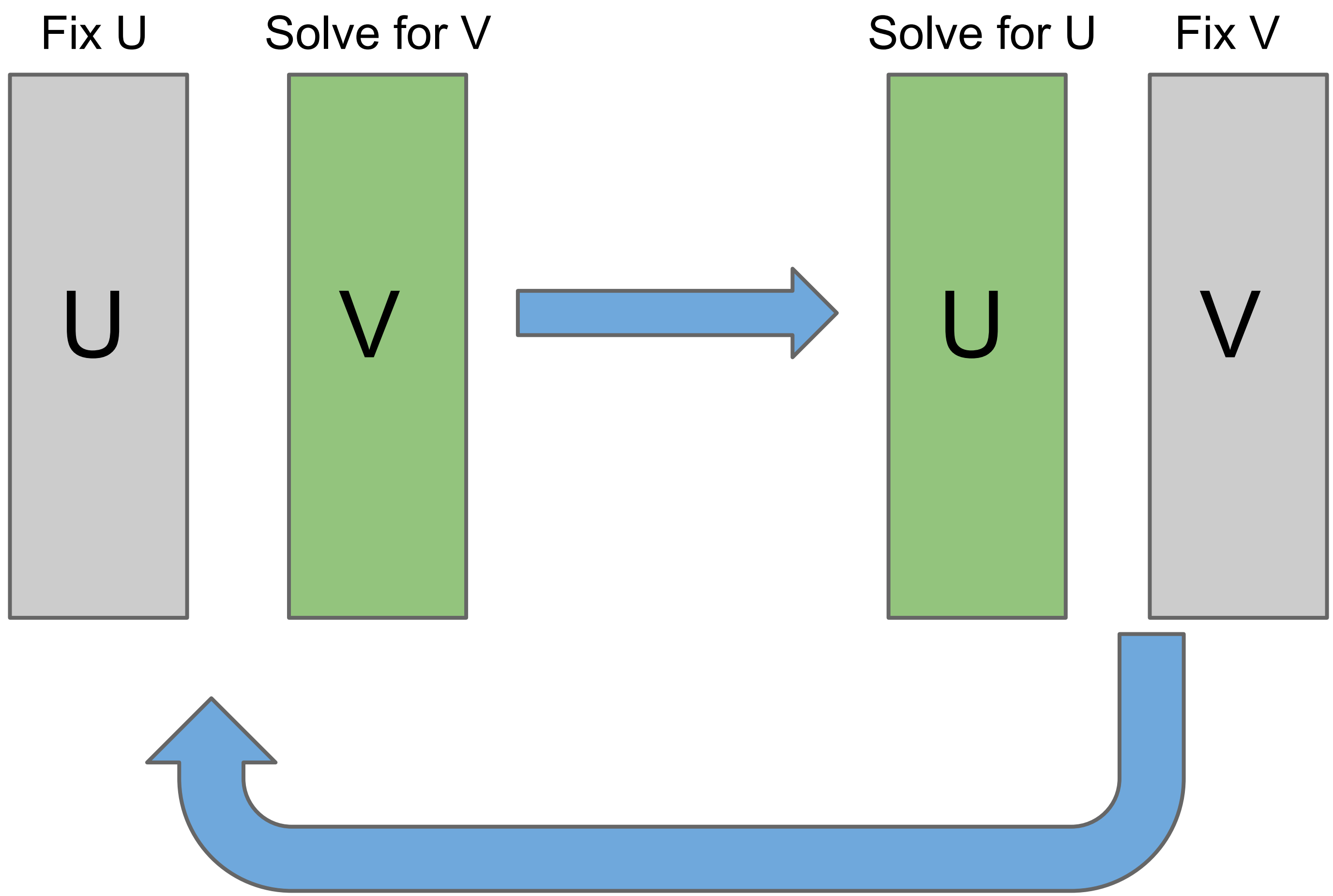
Matrix Decomposition



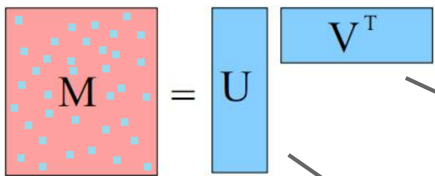
$$M_{m \times n} = U_{m \times k} V_{n \times k}^{\top}$$

Matrix Decomposition

Find matrices U and V by optimization with gradient descent or alternating least squares



Matrix Decomposition



Predicting values with U and V

user 1	4	-3	1
user 2	2	3	-1
user 3	2	-1	3
user 4	3	-3	0
user 5	0	2	-2

business 1	business 2	business 3	business 4	business 5
3	0	2	1	-2
2	-1	-1	0	2
-2	1	-2	1	3

$$Prediction = (2 \times 2) + (3 \times -1) + (-1 \times -2) = 4$$

Matrix Decomposition

Additions to the SVD algorithm

- Regularization (Tikhonov Regularization)
- Subtract global average rating (3.776 stars)
- Weighting business factors(V) by similarity/categories/time

Result:

Rank(factors) = 8

regularization constant = 0.55

<u>RMSE 1.256</u> on Kaggle

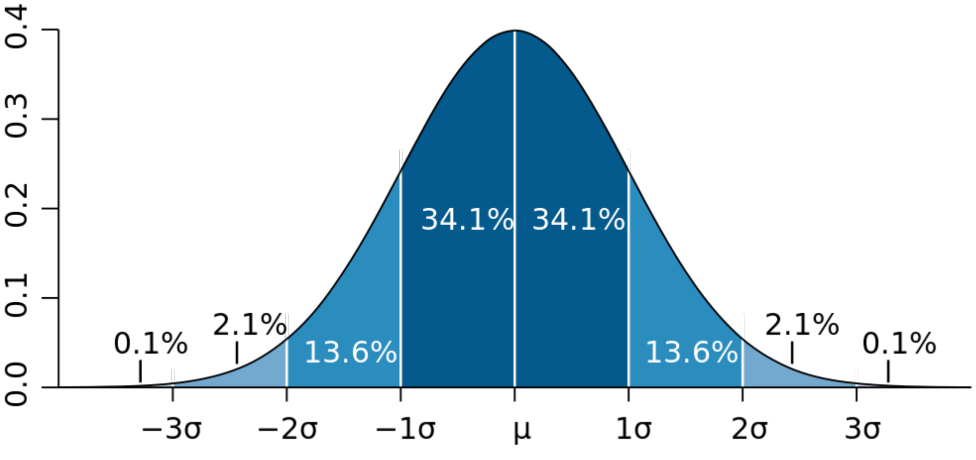
Mean Predictors

UxB	B1	B2	B3	B4	B5
U1	5				
U2	4				
U3	4				
U4	5				
U5	??		3	2	

Sandbag Ratings

"Sandbag" Rating: a 1-2 star rating for a business or user that generally receives ratings in the 4-5 range

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$



Approach

- After identifying a “sandbag” rating in either a row or column, the average of that vector is computed, but with the “sandbag” rating omitted to create a more reflective mean.

- The missing value is then replaced with that mean.

- When evaluated on Kaggle, this predictor received a RMSE of 1.29371

Combined Weighted Mean Predictor

- For certain values, a mean-item predictor works better and for others, a mean-user predictor does.
- To solve for this, we averaged these two predictors and weighted them based on how many reviews each had.

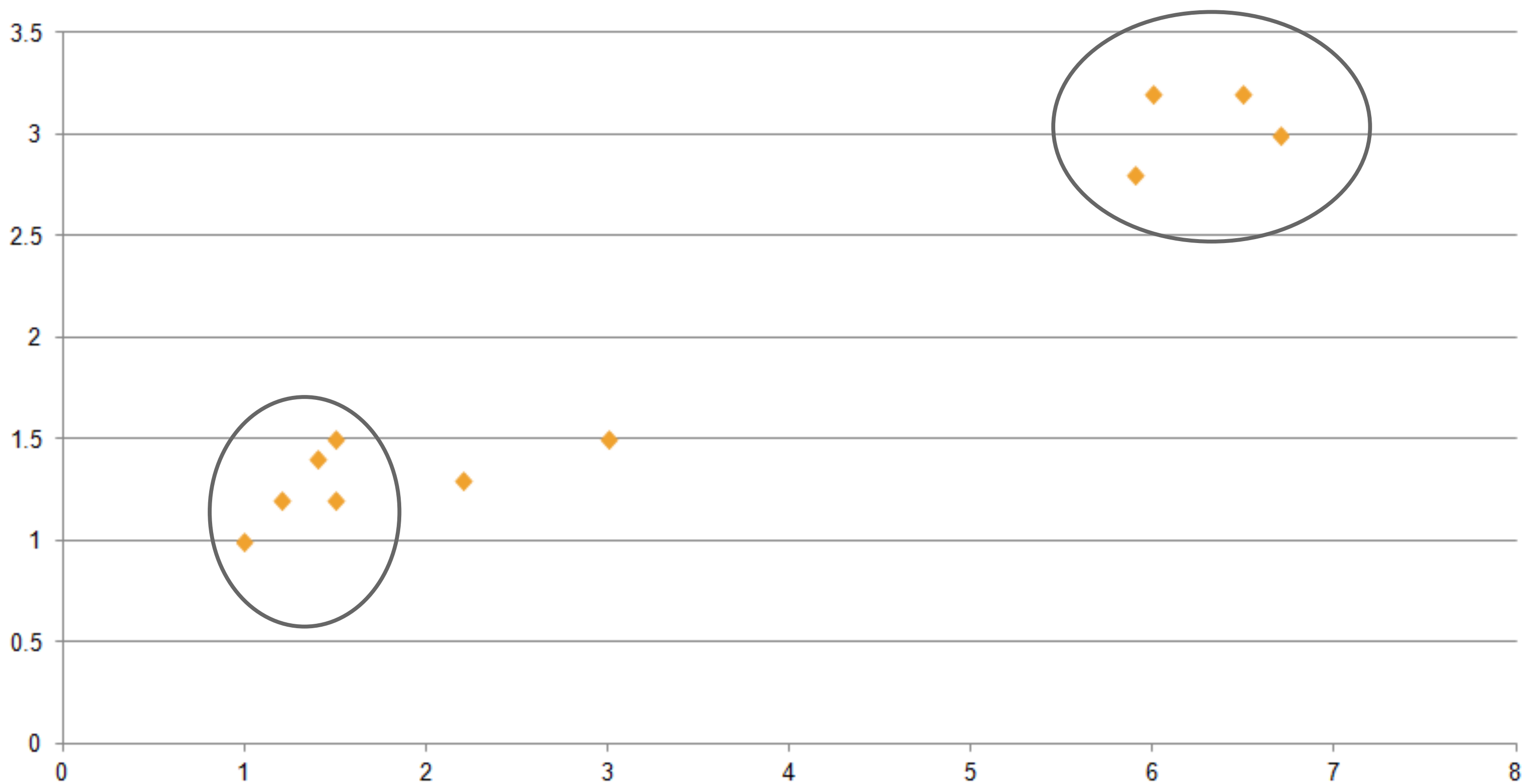
Example Matrix

UxB	B1	B2	B3	B4	B5
U1	5				
U2	4				
U3	4				
U4	5				
U5	-1		3	2	

- B1 columns averages 4.5
- U5 row averages 2.5
- 6 ratings available to predict B1xU5
- (4.5 * 4/6) + (2.5 * 2/6) = 3.833
- On Kaggle, this predictor received a RMSE of 1.252

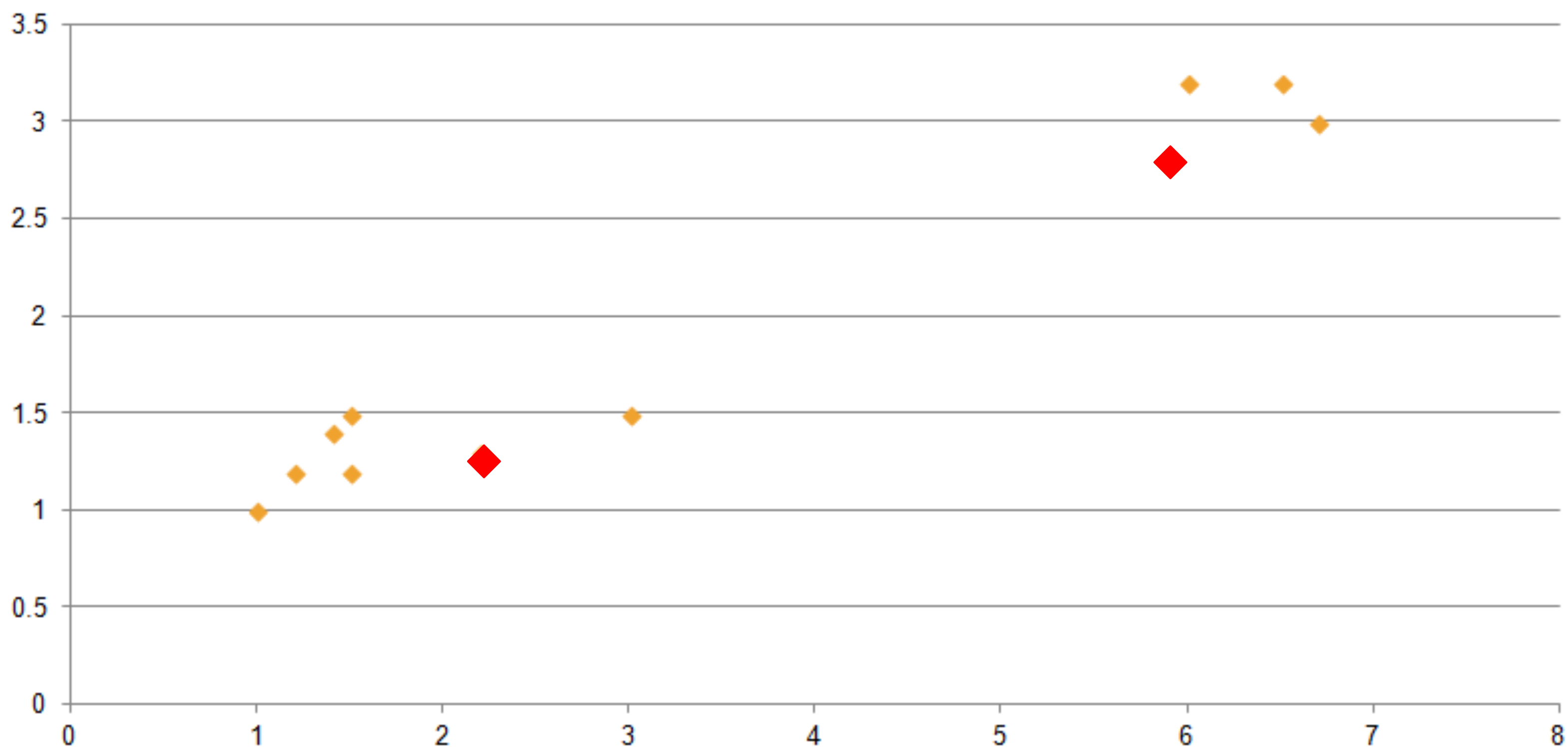
Clustering

- Clustering is a useful algorithm when the data can be separated into “types” or “groups” that are basically the same.



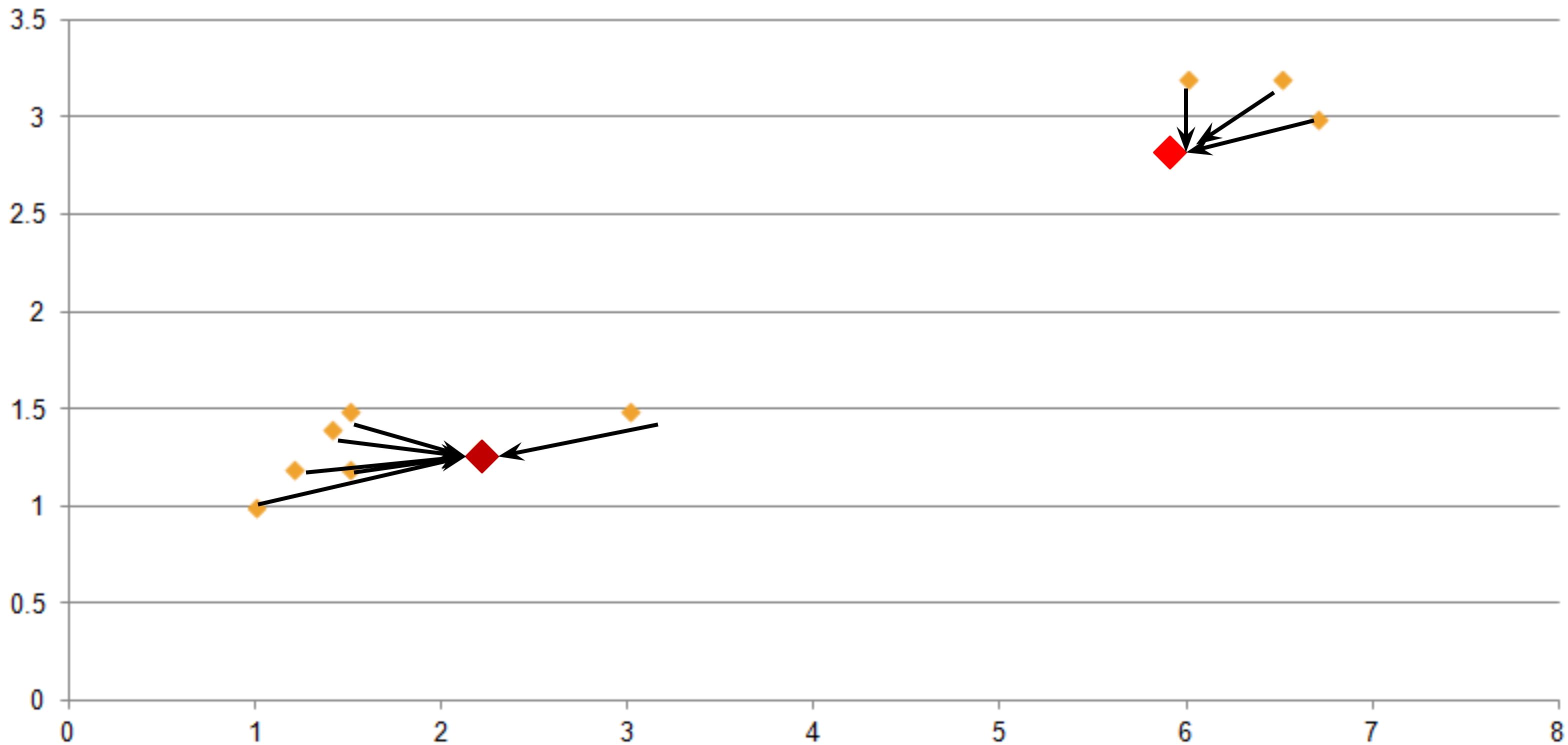
Clustering

- Clustering begins by choosing a set number of random data points (usually randomly selected from the given data points) to be "centers” (i.e., 2 centers)



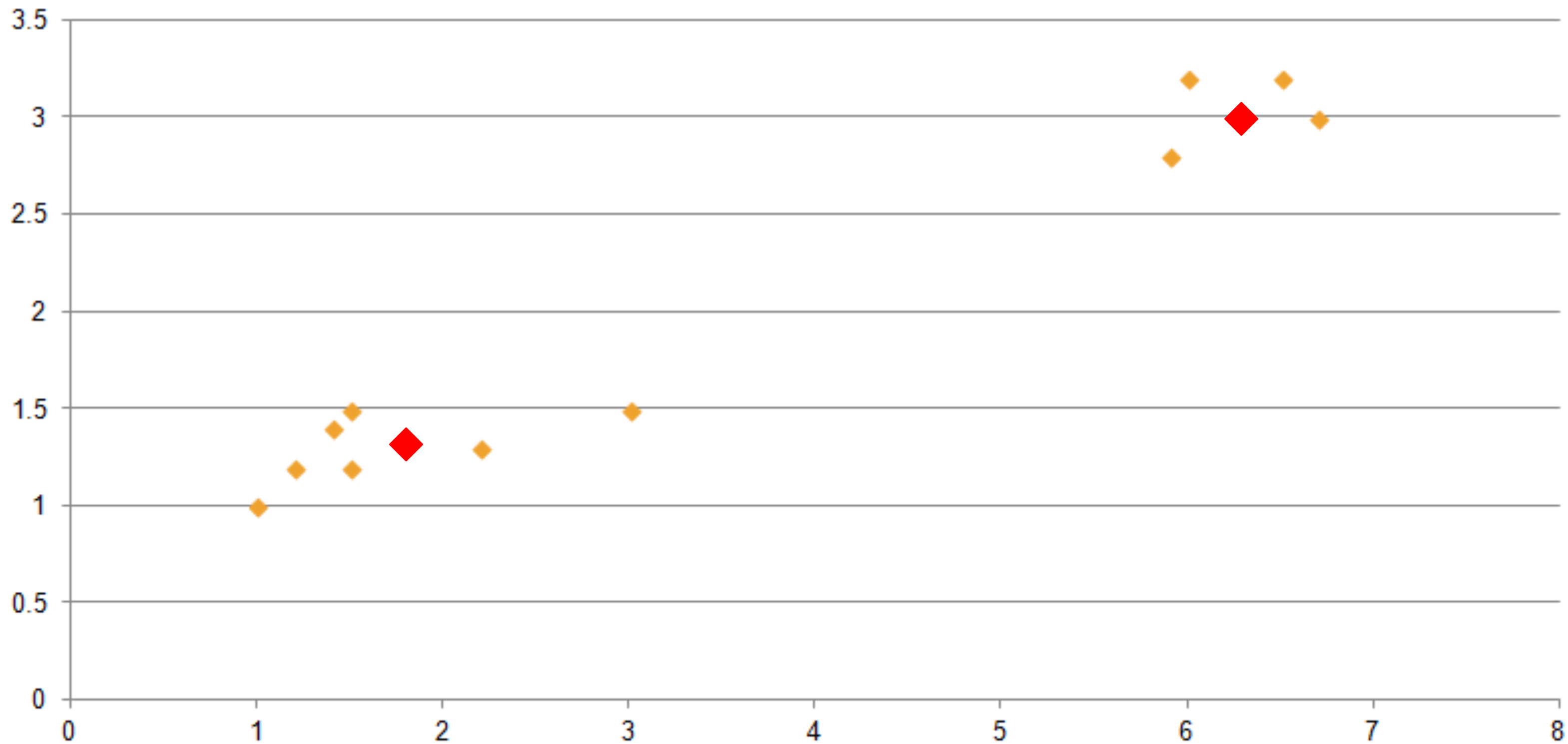
Clustering

- We then calculate the distance between every point and every center (which can take a long time) and select the closest center to each point.



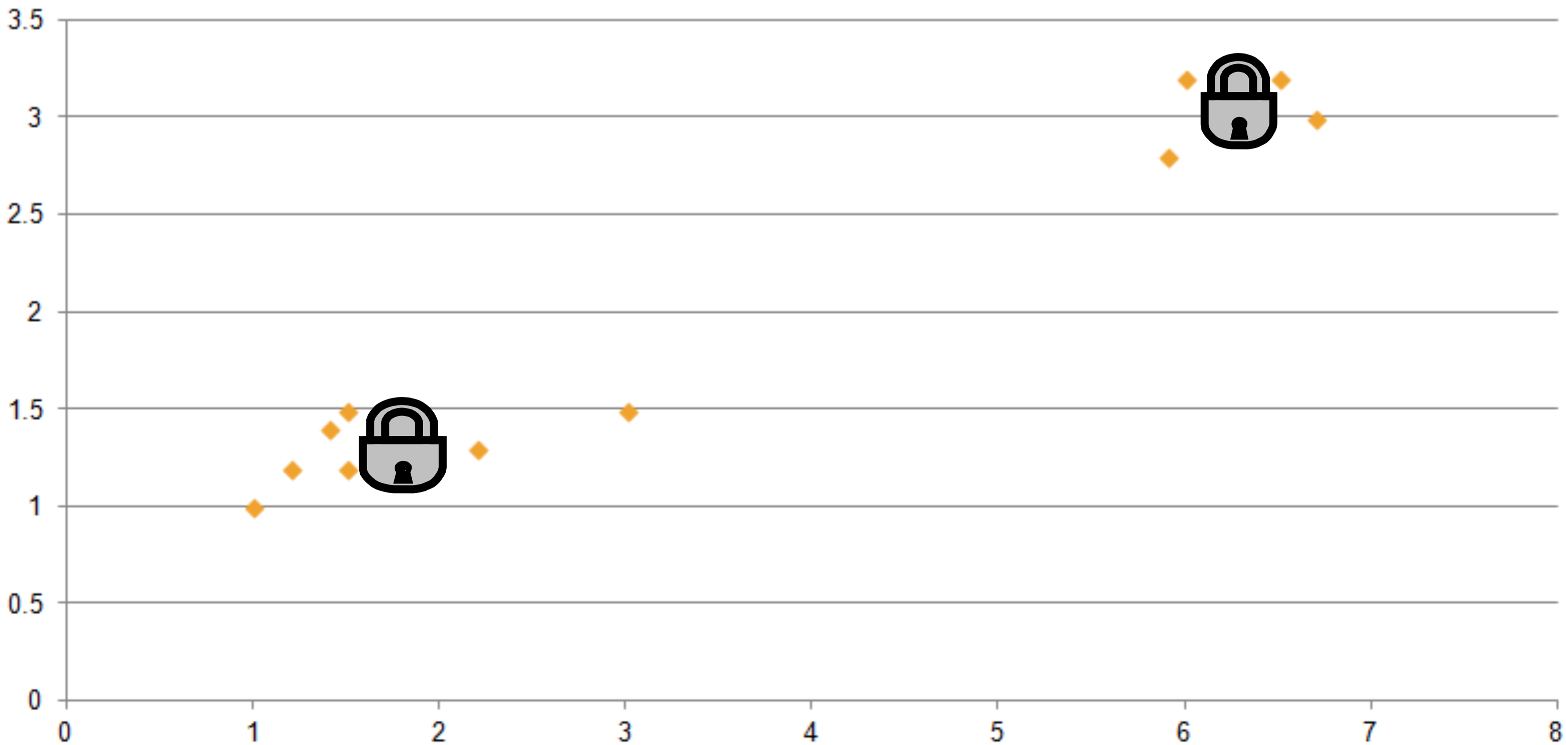
Clustering

- Then, we take each center and move it to the mean of the points assigned to it.



Clustering

- Recalculate the distance to each center, then move the centers, until the centers are fixed.



Clustering

- Now, we can predict each point as if it were the cluster center, which will fill in any missing information.
- Clustering relies somewhat on luck, if bad cluster centers are chosen at the beginning, you can get inaccurate groupings.

Clustering

- To increase accuracy, we clustered over several subgroups, chosen from the most popular business categories, such as restaurants or shopping. Each grouping had a different number of cluster centers.

Clustering

- For the Yelp! Data set, clustering was a relatively ineffective predictor, with an error of ~1.4 RMSE on Kaggle, compared to the user mean error of 1.28

Other Findings

Split Data by Gender

- Functions Used:
  - knnSparse
  - svdSparse
  - yelpMean

- Results: RMSE 1.0103 on the validation set

UxB	BusX	BusY	BusZ	BusA
Male1	4	-1	3	5
Male2	5	3	-1	-1
Male3	-1	3	-1	-1

UxB	BusX	BusY	BusZ	BusA
Female1	1	5	-1	5
Female2	2	3	3	-1
Female3	-1	3	-1	-1

UxB	BusX	BusY	BusZ	BusA
ng1	4	2	5	-1
ng2	5	4	-1	3
ng3	3	3	4	-1

Category

- Italian: average rating of 4.1
- Mexican: average rating of 4.2
- Bars: average rating 3.9

- Initial guess for Anthill Pub in category “Bar”

3.9 stars

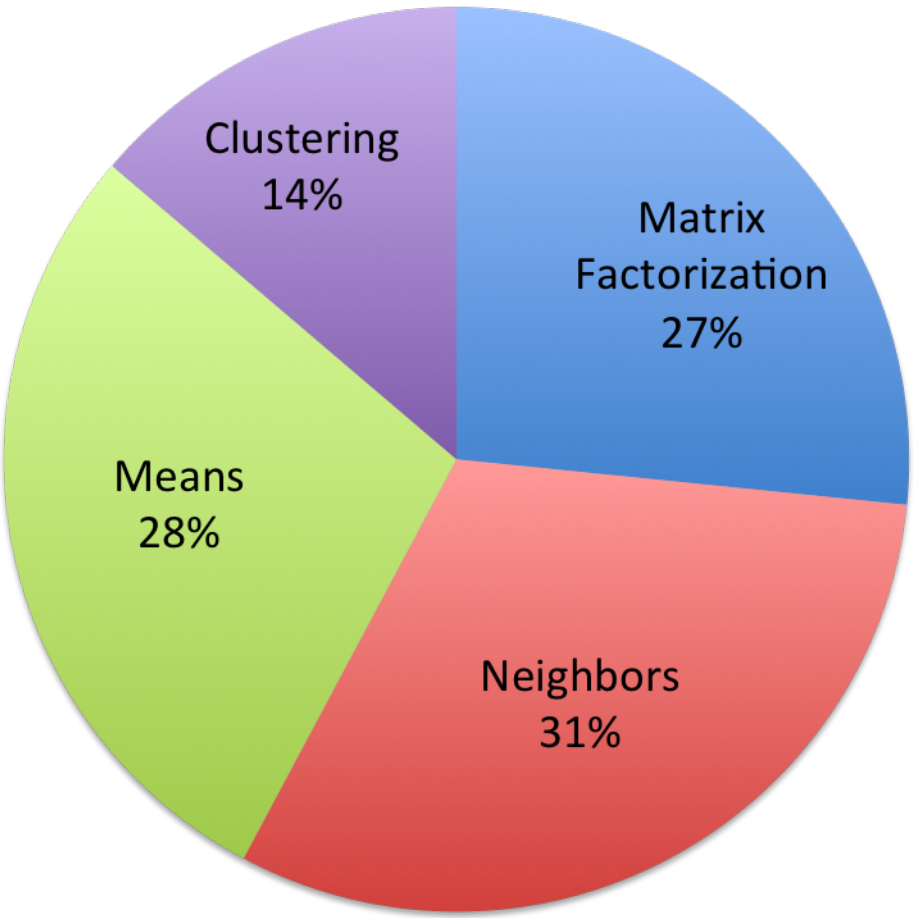
Types of predictions

		Businesses: Known	Businesses: Unknown
Users	Known	Neighborhood Methods Clustering Matrix Factorization 28%	User Means User-oriented Neighborhood 27%
	Unknown	Business Means Clustering by businesses Category means 33%	Global average Predicted user and business means Category means 12%

Results of Individual Models

Method	Ranking out of 405
Clustering	288
Matrix Factorization	132
User/Business means	139
Neighborhood model	122
Combined Weighted Mean	143

Blending



Method	Ranking out of 405
Clustering	288th
Matrix Factorization	132nd
User/Business means	130th
Neighborhood model	122nd
<u>Blended</u>	<u>51st</u>

Thanks

Advisors:

Dr. Alexander Ihler

Sholeh Forouzan