Recommendation Systems
Part I

Devavrat Shah

Massachusetts Institute of Technology

Outline — Part I

Introduction, simple methods

- Module 1: background
  - Recommendation systems: why and what?
  - Example datasets
- Module 2: problem statement
  - Recommendation systems: a prediction problem
  - Model: from caricature to extremely complex
- Module 3: simple solutions
  - Solution I: averaging
  - Solution II: content-based

Module 1: background

Recommendations

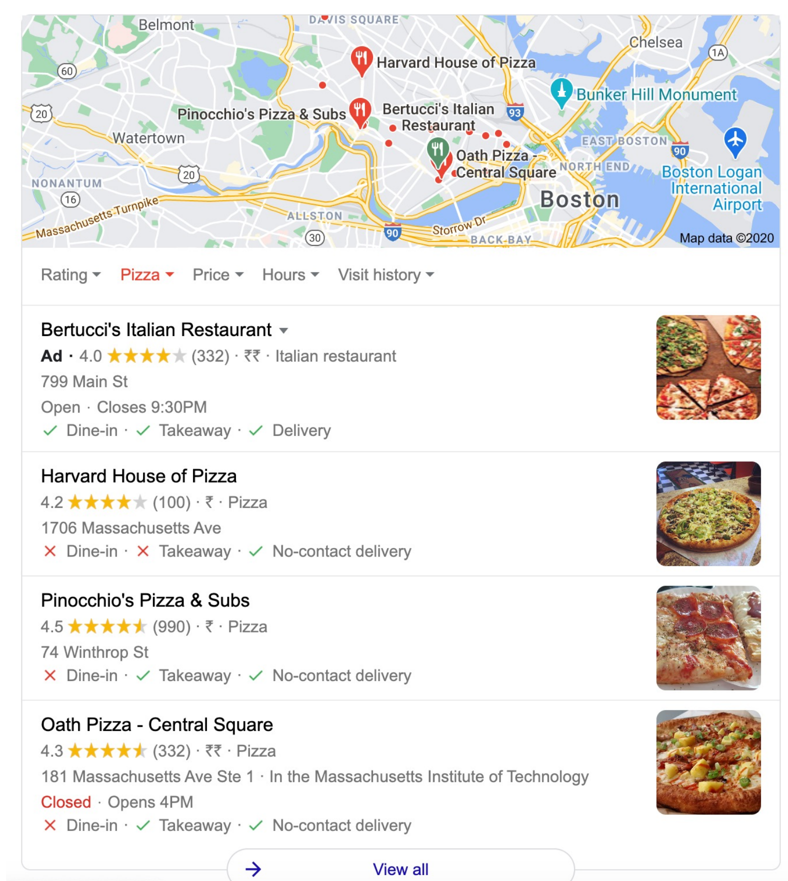
What is it, really?

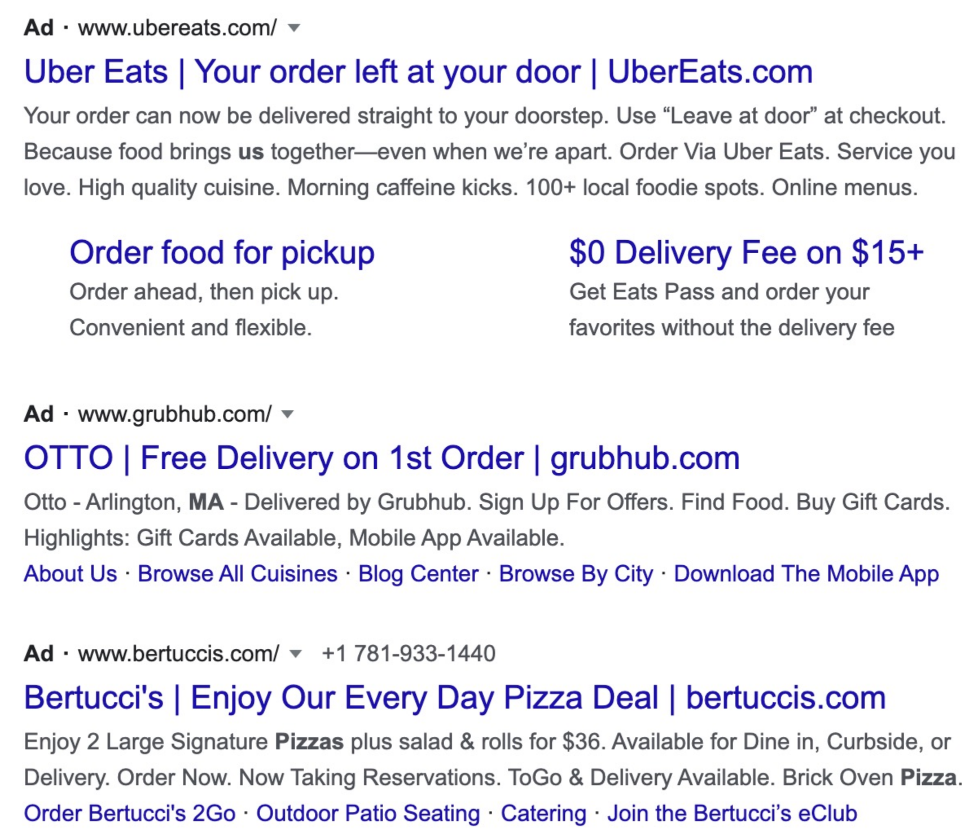
What food should you eat today?

Recommendations

Why not just search?

Want to eat Pizza, search for Pizza places near you





But

Which of these *many* options should you be *recommended?*

Which of these *many* be *advertised?*

Recommendations

What is it, really?

What food should you eat today?

Which activities should you plan for the upcoming weekend?

Where should you plan your holidays?

Whom should you date (and marry)?

What are the professional connections of your interest?

Which advertisements should you be subjected to?

What content on YouTube or music on Spotify will fancy you?

o o o o

Recommendations

May be *data* can help. What data?

Recommendations

May be *data* can help. What data?

Example: Yelp data

- Businesses: attributes (locations, category), hours
- Users: attributes, friends
- Reviews: rating, description, time
- Check-ins: time
- Tip

Exercise: go to link below, explore data and reproduce statistics reported

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

Businesses: 209393 (~200k) in total

	address	attributes	business_id	categories	city	hours	is_open	latitude	longitude	name	postal_code	review_count	stars	state
0	10913 Bailey Rd	{u'BusinessParking': u'{'garage': False, 'stre...	f9NumwFMBDn751xgFiRbNA	Active Life, Gun/Rifle Ranges, Guns & Ammo, Sh...	Cornelius	{u'Monday': u'10:0-18:0', u'Tuesday': u'11:0-2...	1	35.462724	-80.852612	The Range At Lake Norman	28031	36	3.5	NC
1	8880 E Via Linda, Ste 107	{u'GoodForKids': u'True', u'ByAppointmentOnly'...	Yzvjg0SayhoZgCljUJRF9Q	Health & Medical, Fitness & Instruction, Yoga,...	Scottsdale	None	1	33.569404	-111.890264	Carlos Santo, NMD	85258	4	5.0	AZ
2	3554 Rue Notre-Dame O	None	XNoUzKckATkOD1hP6vghZg	Pets, Pet Services, Pet Groomers	Montreal	None	1	45.479984	-73.580070	Felinus	H4C 1P4	5	5.0	QC
3	1015 Sharp Cir	{u'DogsAllowed': u'True', u'BusinessParking': ...	6OAZjbxqM5ol29BuHsil3w	Hardware Stores, Home Services, Building Suppl...	North Las Vegas	{u'Friday': u'7:0-16:0', u'Tuesday': u'7:0-16:...	0	36.219728	-115.127725	Nevada House of Hose	89030	3	2.5	NV
4	4827 E Downing Cir	{u'BusinessAcceptsCreditCards': u'True', u'ByA...	51M2Kk903DFYl6gnB5I6SQ	Home Services, Plumbing, Electricians, Handyma...	Mesa	{u'Friday': u'9:0-16:0', u'Tuesday': u'9:0-16:...	1	33.428065	-111.726648	USE MY GUY SERVICES LLC	85205	26	4.5	AZ

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

Users: 1968703 (~2M) in total

	average_stars	compliment_cool	compliment_cute	compliment_funny	compliment_hot	compliment_list	compliment_more	compliment_note
0	3.57	22	0	22	3	1	2	11
1	3.84	63	2	63	36	1	4	33
2	3.44	17	1	17	9	0	6	3
3	3.08	7	0	7	2	0	1	7
4	4.37	31	1	31	8	1	9	22

elite	fans	friends	funny	name	review_count	useful	user_id	yelping_since
	14	oeMvJh94PiGQnx_6GlndPQ, wm1z1PaJKvHgSDRKfwhfDg...	225	Rafael	553	628	ntlvfPzc8eglqvk92iDIAw	2007-07-06 03:27:11
2008,2009,2010,2011,2012,2013	27	ly7EnE8leJmyqyePVYFlug, pRlR63iDytsnnniPb3AOug...	316	Michelle	564	790	FOBRPlBHa3WPHFB5qYDlVg	2008-04-28 01:29:25
2010	5	Uwlk0txjQBPw_JhHsQnyeg, Ybxr1tSCkv3lYA0l1qmnPQ...	125	Martin	60	151	zZUnPeh2hEp0WydbAZEOOg	2008-08-28 23:40:05
2009	6	iog3Nyg1i4jeumiTVG_BSA, M92xWY2Vr9w0xoH8bPplfQ...	160	John	206	233	QaELAmRcDc5TfJEylaaP8g	2008-09-20 00:08:14
2009,2010,2011,2012,2014,2015,2016,2017,2018	78	3W3ZMSthojCUirKEqAwGNw, eTlbuu23j9tOgmla9POyLQ...	400	Anne	485	1265	xvu8G900tezTzbbfqmTKvA	2008-08-09 00:30:27

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

Reviews: 8021122 (~8M) in total

	business_id	cool	date	funny	review_id	stars	text	useful	user_id
0	-MhfebM0QIsKt87iDN-FNw	0	2015-04-15 05:21:16	0	xQY8N_XvtGbearJ5X4QryQ	2.0	As someone who has worked with many museums, I...	5	OwjRMXRC0KyPrIlcjaXeFQ
1	lbrU8StCq3yDfr-QMnGrmQ	0	2013-12-07 03:16:52	1	UmFMZ8PyXZTY2QcwzsfQYA	1.0	I am actually horrified this place is still in...	1	nIJD_7ZXHq-FX8byPMOkMQ
2	HQl28KMwrEKHqhFrrDqVNQ	0	2015-12-05 03:18:11	0	LG2ZaYiOgpr2DK_90pYjNw	5.0	I love Deagan's. I do. I really do. The atmosp...	1	V34qejxNsCbcgD8C0HVk-Q
3	5JxlZaqCnk1MnbgRirs40Q	0	2011-05-27 05:30:52	0	i6g_oA9Yf9Y31qt0wibXpw	1.0	Dismal, lukewarm, defrosted-tasting "TexMex" g...	0	ofKDkJKXSKZXu5xJNGiiBQ
4	IS4cv902ykd8wj1TR0N3-A	0	2017-01-14 21:56:57	0	6TdNDKywdbjoTkizeMce8A	4.0	Oh happy day, finally have a Canes near my cas...	0	UgMW8bLE0QMJDCkQ1Ax5Mg

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

Check-ins: 175187 (~175k) in total

	business_id	date
0	--1UhMGODdWsrMastO9DZw	2016-04-26 19:49:16, 2016-08-30 18:36:57, 2016...
1	--6MefnULPED_I942VcFNA	2011-06-04 18:22:23, 2011-07-23 23:51:33, 2012...
2	--7zmmkVg-IMGaXbuVd0SQ	2014-12-29 19:25:50, 2015-01-17 01:49:14, 2015...
3	--8LPVSo5i0Oo61X01sV9A	2016-07-08 16:43:30
4	--9QQLMTbFzLJ_oT-ON3Xw	2010-06-26 17:39:07, 2010-08-01 20:06:21, 2010...

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

Tip: 1320761 (~1.3M) total

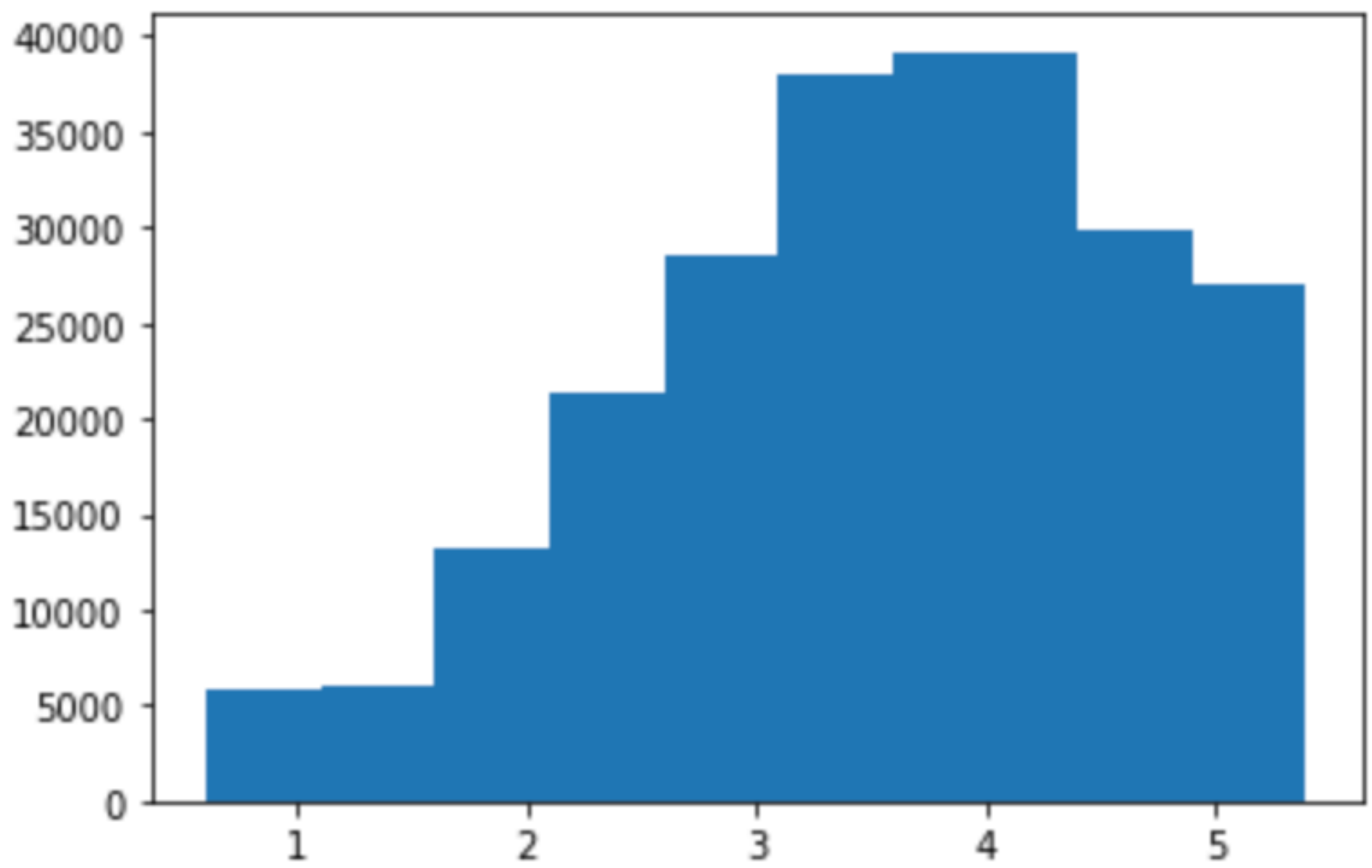
	business_id	compliment_count	date	text	user_id
0	UYX5zL_Xj9WEc_Wp-FrqHw	0	2013-11-26 18:20:08	Here for a quick mtg	hf27xTME3EiCp6NL6VtWZQ
1	Ch3HkwQYv1YKw_FO06vBWA	0	2014-06-15 22:26:45	Cucumber strawberry refresher	uEvusDwoSymbJJ0auR3muQ
2	rDoT-MgxGRiYqCmi0bG10g	0	2016-07-18 22:03:42	Very nice good service good food	AY-laIws3S7YXNl_f_D6rQ
3	OHXnDV01gLokiX1ELaQufA	0	2014-06-06 01:10:34	It's a small place. The staff is friendly.	Ue_7yUlkEbX4AhnYdUfL7g
4	GMrwDXRlAZU2zj5nH6l4vQ	0	2011-04-08 18:12:01	8 sandwiches, $24 total...what a bargain!!! An...	LltbT_fUMqZ-ZJP-vJ84lQ

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

(Aggregate) Star Rating Distribution



Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: Yelp Data

What fraction of reviews are known?

Users = ~2M

Businesses = ~200k

Total possible reviews = ~2M x ~200k = ~0.4T

Known reviews = ~8M

Fraction known = ~8M / 0.4 T = 2 x 10^{-5}

i.e. 2 in every 100k reviews is known, rest are *unknown*

Finding these *unknown* reviews is the primary goal of Rec Sys

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Explore: MovieLens Data

MovieLens Data

- Movies: attributes including title, release date, genre, actors, director
- Users: demographics including age, gender, occupation, zip code
- Reviews: ratings, timestamp

Exercise: go to link below, explore data and reproduce statistics reported

https://grouplens.org/datasets/movielens/100k

Explore: MovieLens Data

Reviews: 100000 (100k) in total

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205

Data URL:

https://grouplens.org/datasets/movielens/100k

Explore: MovieLens Data

Users: 1682 (~1.7k) in total

```
user id | age | gender | occupation | zip code
```

```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
```

Data URL:

https://grouplens.org/datasets/movielens/100k

Explore: MovieLens Data

Movies: 943 (~1k) in total

```
movie id | movie title | release date | video release date |
IMDb URL | unknown | Action | Adventure | Animation |
Children's | Comedy | Crime | Documentary | Drama | Fantasy |
Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi |
Thriller | War | Western |
```

```
1|Toy Story (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)|0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0
2|GoldenEye (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?GoldenEye%20(1995)|0|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
3|Four Rooms (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995)|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
4|Get Shorty (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Get%20Shorty%20(1995)|0|1|0|0|0|1|0|0|1|0|0|0|0|0|0|0|0|0|0
5|Copycat (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Copycat%20(1995)|0|0|0|0|0|0|1|0|1|0|0|0|0|0|0|0|1|0|0
6|Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)|01-Jan-1995||http://us.imdb.com/Title?Yao+a+yao+yao+dao+waipo+qiao+(1995)|0|0|0|0|0|0|0|0|1|0|
7|Twelve Monkeys (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Twelve%20Monkeys%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|1|0|0|0
8|Babe (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Babe%20(1995)|0|0|0|0|1|1|0|0|1|0|0|0|0|0|0|0|0|0|0
9|Dead Man Walking (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Dead%20Man%20Walking%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0
10|Richard III (1995)|22-Jan-1996||http://us.imdb.com/M/title-exact?Richard%20III%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|1|0
11|Seven (Se7en) (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Se7en%20(1995)|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|1|0|0
12|Usual Suspects, The (1995)|14-Aug-1995||http://us.imdb.com/M/title-exact?Usual%20Suspects,%20The%20(1995)|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|1|0|0
13|Mighty Aphrodite (1995)|30-Oct-1995||http://us.imdb.com/M/title-exact?Mighty%20Aphrodite%20(1995)|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0|0|0|0
14|Postino, Il (1994)|01-Jan-1994||http://us.imdb.com/M/title-exact?Postino,%20Il%20(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|1|0|0|0|0
15|Mr. Holland's Opus (1995)|29-Jan-1996||http://us.imdb.com/M/title-exact?Mr.%20Holland's%20Opus%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0
16|French Twist (Gazon maudit) (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Gazon%20maudit%20(1995)|0|0|0|0|0|1|0|0|0|0|0|0|0|0|1|0|0|0|0
17|From Dusk Till Dawn (1996)|05-Feb-1996||http://us.imdb.com/M/title-exact?From%20Dusk%20Till%20Dawn%20(1996)|0|1|0|0|0|1|1|0|0|0|0|1|0|0|0|0|1|0|0
18|White Balloon, The (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Badkonake%20Sefid%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0
19|Antonia's Line (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Antonia%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0
20|Angels and Insects (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Angels%20and%20Insects%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|1|0|0|0|0
```

Data URL:

https://grouplens.org/datasets/movielens/100k

Explore: MovieLens Data

Top rated movies

```
rating['movieId'].value_counts()
```

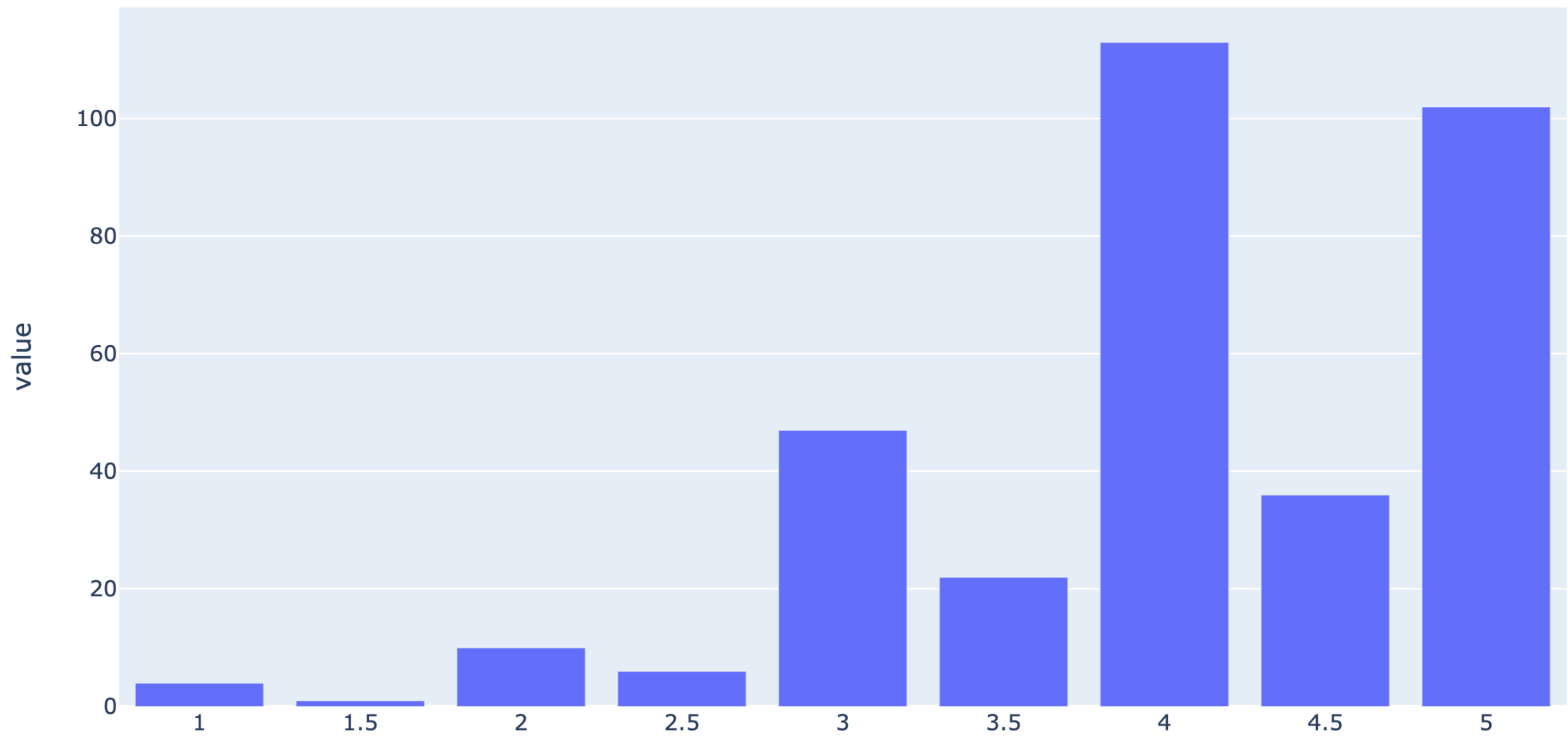
```
356     341
296     324
318     311
593     304
260     291
480     274
2571    259
1       247
527     244
589     237
1196    234
```

Data URL:

https://grouplens.org/datasets/movielens/100k

Explore: MovieLens Data

Distribution of top-rated movie (356)



Data URL:

https://grouplens.org/datasets/movielens/100k

Explore: MovieLens Data

What fraction of reviews are known?

Users = ~1.7k

Movies = ~1k

Total possible reviews = ~1.7k x ~1k = ~1.7M

Known reviews = ~100k

Fraction known = ~100k / 1.7M = ~0.058 or ~6%

i.e. 6 in every 100 reviews is known, rest are *unknown*

Finding these *unknown* reviews is the primary goal of Rec Sys

Data URL:

https://www.kaggle.com/jagangupta/what-s-in-a-review-yelp-ratings-eda

Module 2: problem statement

Recommendation: Problem statement

What does Yelp want to do?

Provide list of businesses

that satisfy your search criteria

ordered as per

user's interest or preference list at that moment

advertisement revenue considerations

so that ultimately *matching* happens within *few* clicks

that provides instant gratification to both user and Yelp

and continues bringing user back to Yelp

Recommendation: Problem statement

What about?

Linked In: connect people professionally

Facebook: filter friends' feed

Poshmark, Etsy: organize content of display

Amazon, Retail: display products and sellers

Tinder, Match: find suitable partner

Netflix, YouTube, Spotify: entertainment of interest

o o o o

So that ultimately *matching* happens within *few* clicks

providing instant gratification to *user, provider* and *platform*

and *users* and *providers* continue engaging with the *platform*

Recommendation: Problem statement

Prediction problem

what is the likelihood of *matching*

user

provider

at a given time

in a given *context*

Accurate solution to this prediction problem

provides essential ingredient for connecting users, providers on the platform

while respecting interests of users, providers and the platform

Recommendation: Problem statement

Prediction problem

N users

M providers or items (or other users)

Given user *i and* item *j*

find the likelihood of *i* matching with *j*

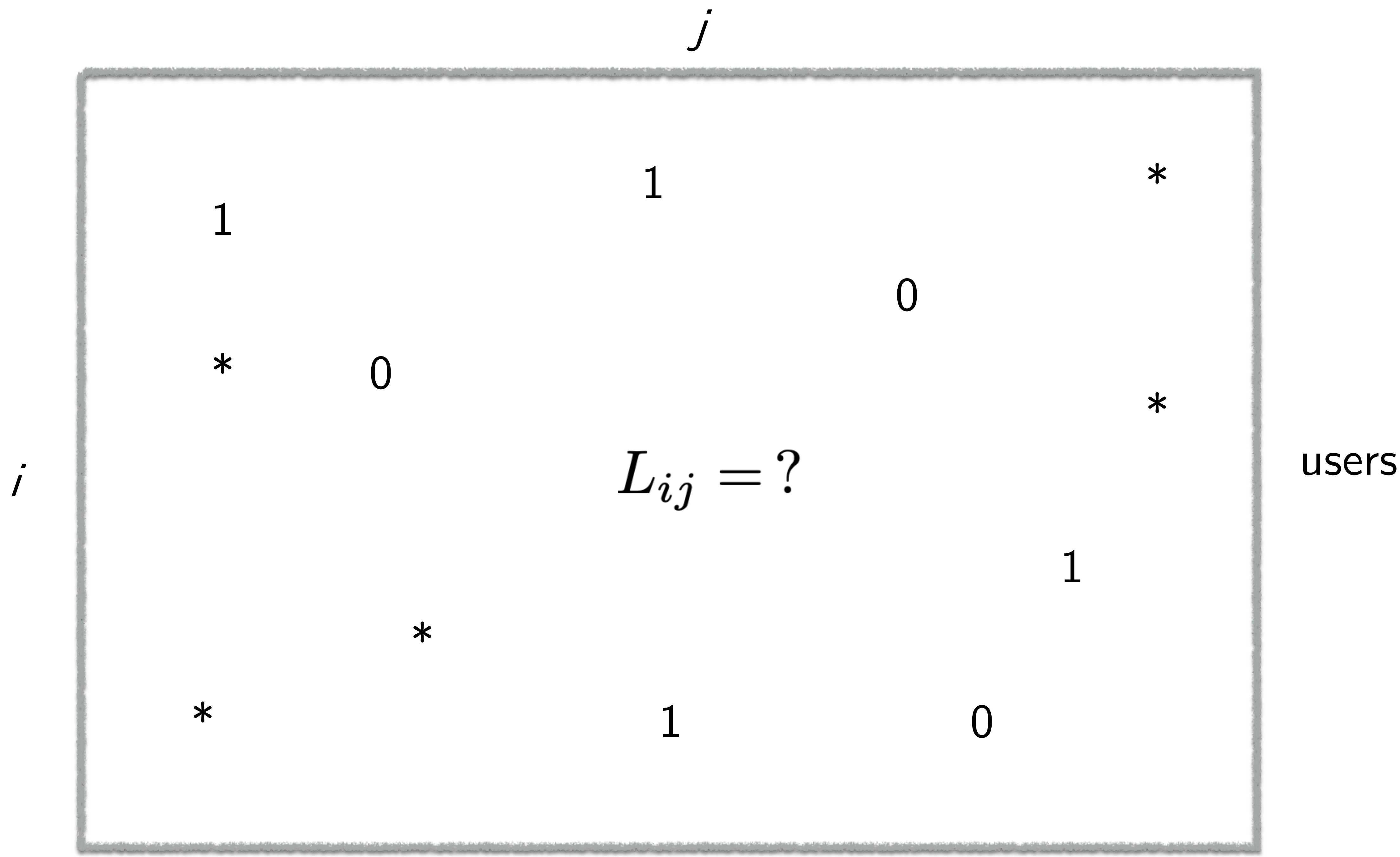
$$L_{ij} = ?$$

Using data

Historical matchings or preferences

Recommendation: Problem statement

Prediction problem: complete the matrix



items / providers

Recommendation: Problem statement

Complete the matrix

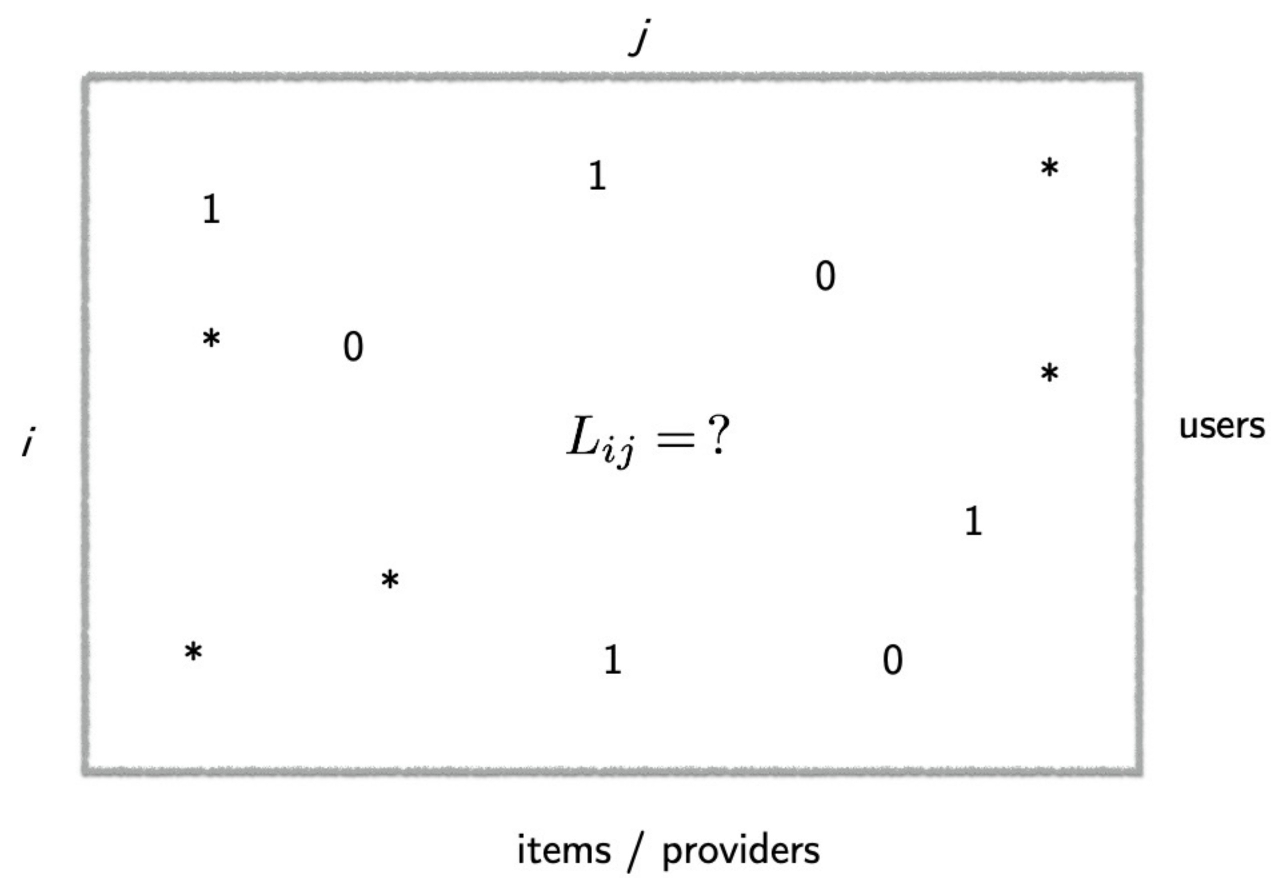
Observations: Y_{ij} over i in users, j in items

If (i, j) is observed

$$E[Y_{ij}] = L_{ij}$$

If (i, j) is *not* observed

$$Y_{ij} = \star \text{ or } ?$$

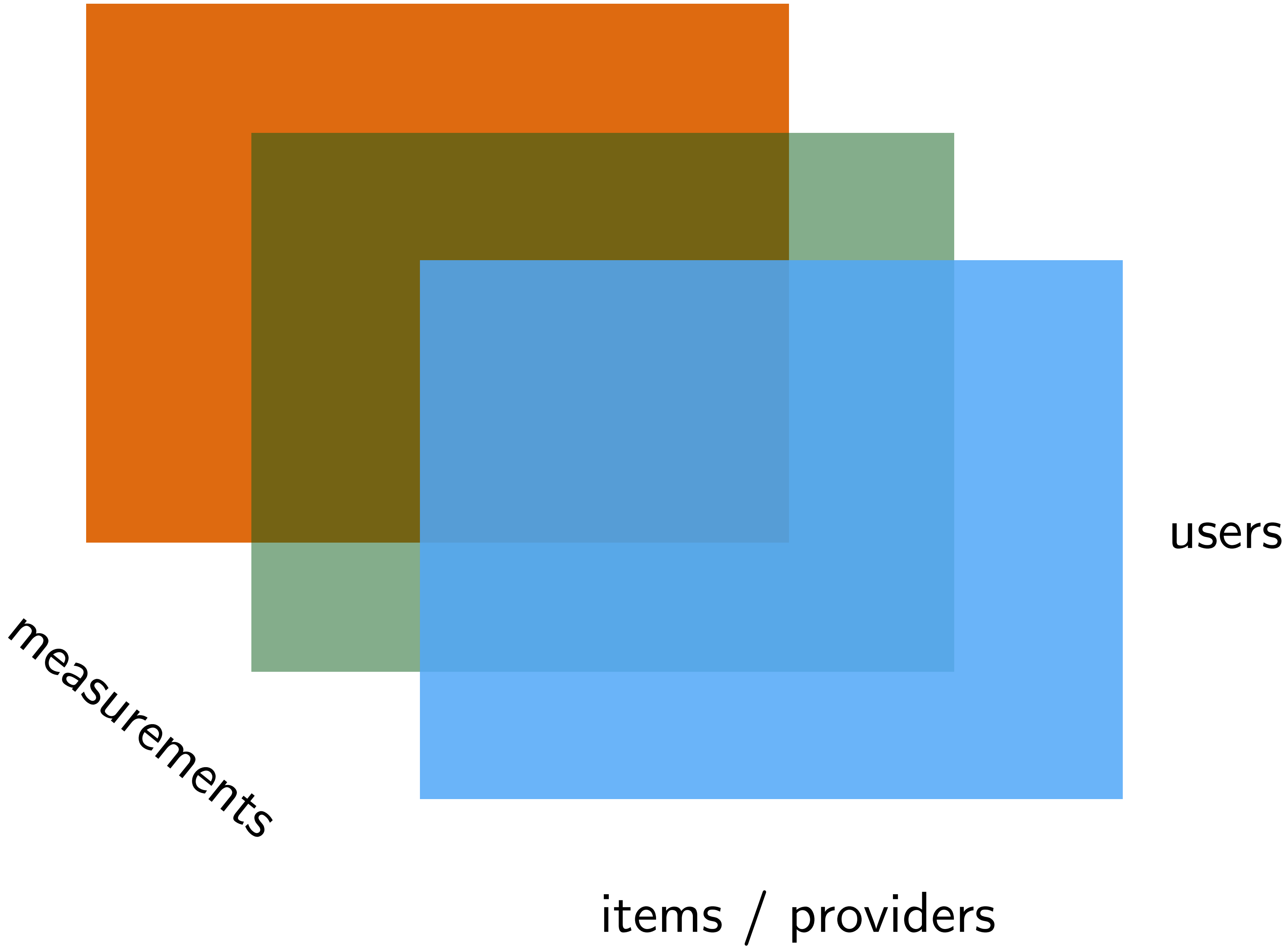


Goal: produce estimation $\hat{L}_{ij}$ for all i, j

so that $\hat{L}_{ij} \approx L_{ij}$ for all i, j

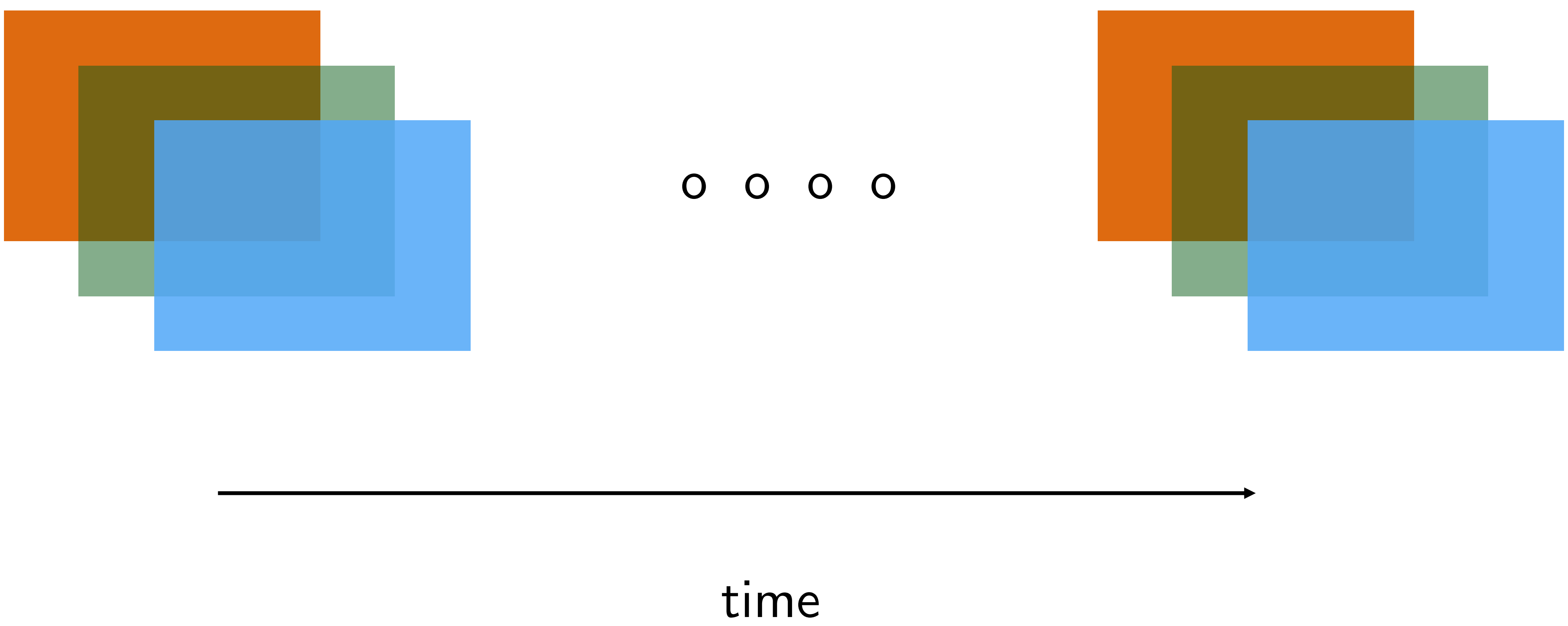
Recommendation: Problem statement

Prediction problem: complete the tensor



Recommendation: Problem statement

Prediction problem: complete the time varying tensor



Recommendation: Problem statement

Challenges that we will not discuss

- user provided data
  - not generated at random
  - can be strategic
  - can be driven by innate preference

- provider / item
  - can be systematically manipulated
    - convenient location
    - strategically modified content

Recommendation: Problem statement

Challenges that we will not discuss

- Side effects of recommendation systems
  - Information bubble
  - Feedback loops
  - Too powerful platforms
- Regulation of platforms
  - Interplay with recommendation systems

Module 3: simple solutions

Recommendation: Problem statement

We will start with simple problem statement: complete the matrix

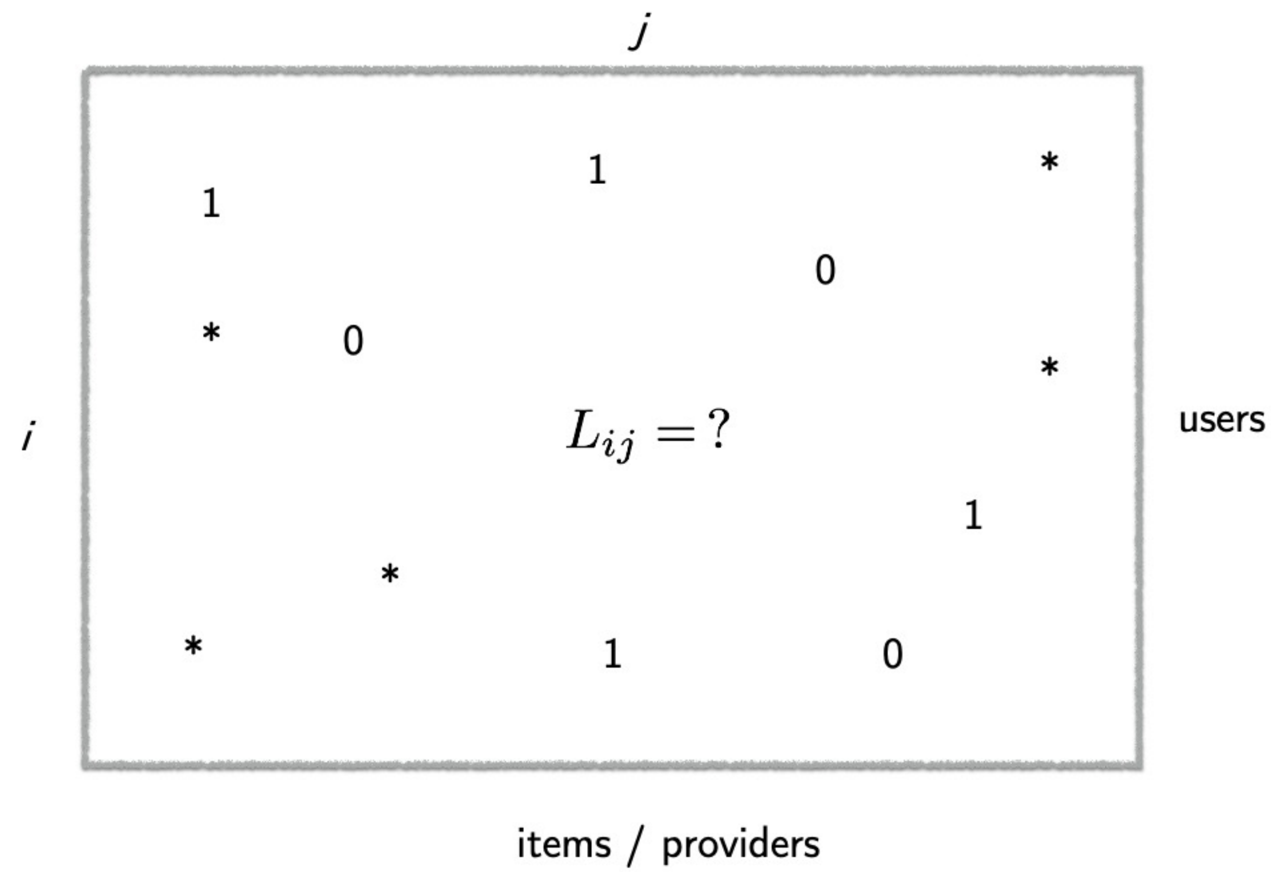
Observations: Y_{ij} over i in users, j in items

If (i, j) is observed

$$E[Y_{ij}] = L_{ij}$$

If (i, j) is *not* observed

$$Y_{ij} = \star \text{ or } ?$$



Goal: produce estimation $\hat{L}_{ij}$ for all i, j

so that $\hat{L}_{ij} \approx L_{ij}$ for all i, j

Solution 1: Averaging

A simple assumption to get started

All users are identical

That is, all rows of the matrix are the *same*

We wish to predict L_{ij}

All rows in the column *j* are the same

Each observation in column *j* is outcome of a coin toss with bias L_{ij}

Solution

Estimate L_{ij} as the average of all observations in column *j*

Solution 1: Averaging

How accurate?

By law of large numbers

as number of observations in column j grow

the estimate converges to the true likelihood

But how large should it be?

one (or few) good rating does not mean the place is excellent

By central limit theorem

the estimation error scale as $\frac{1}{\sqrt{n}}$ with n observations

An improved estimate: average $+\frac{1}{\sqrt{n}}$

Exercise: why correction is $+\frac{1}{\sqrt{n}}$ and not -$\frac{1}{\sqrt{n}}$

Solution 1: Averaging

What if, instead we assume

All items or providers are identical

Then estimate: row average (+ correction for number of observations)

How to put these two simple estimators together?

$$2L_{ij} \;=\; L_{i\cdot} + \frac{1}{\sqrt{n_{i\cdot}}} + L_{\cdot j} + \frac{1}{\sqrt{n_{\cdot j}}}$$

where $L_{i\cdot}$ is average of observed entries in row i

$n_{i\cdot}$ is number of observed entries in row i

$L_{\cdot j}$ is average of observed entries in column j

$n_{\cdot j}$ is number of observed entries in column j

Solution 2: Content Based

A little more involved assumption

Users and items have features

that are observed and can predict the likelihood

Let features of user i be x_i

MovieLens: demographics (age, gender, occupation, zip) of users

Let features of item j be y_j

MovieLens: attributes (genre, director, actors, year, title) of movies

Then, goal is to learn f where $L_{ij} = f(x_i, y_j)$

Exercise: how to utilize features of MovieLens operationally?

Solution 2: Content Based

This is *supervised learning* problem we have already seen

Labeled data:

each observed entry in matrix (i, j) corresponds to labeled data

$$((x_i, y_j); L_{ij})$$

Learning problem:

learn the model / function that maps features to label

For likelihood setting with observations being 0 or 1, it is *classification*

Exercise:
What method would you use for classification?
What if observations were not 0/1 but continuous numbers?

Appendix: Converting Content to Features

Challenge: Content is *not* structured

e.g. recall user information from MovieLens data

```
user id | age | gender | occupation | zip code

1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
```

How do we convert these "attributes" or "content" to features

Appendix: Converting Content to Features

Challenge: Content is *not* structured

How do we convert these "attributes" or "content" to features

Age: It's a number. That's easy.

Gender: Two classes or binary. Convert into 0 / 1.

Occupation:

Treat as a class. Use one-hot encoding.

```
user id | age | gender | occupation | zip code
```

```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
```

Appendix: Converting Content to Features

Challenge: Content is *not* structured

What about "Tip" data. It has free-form text.

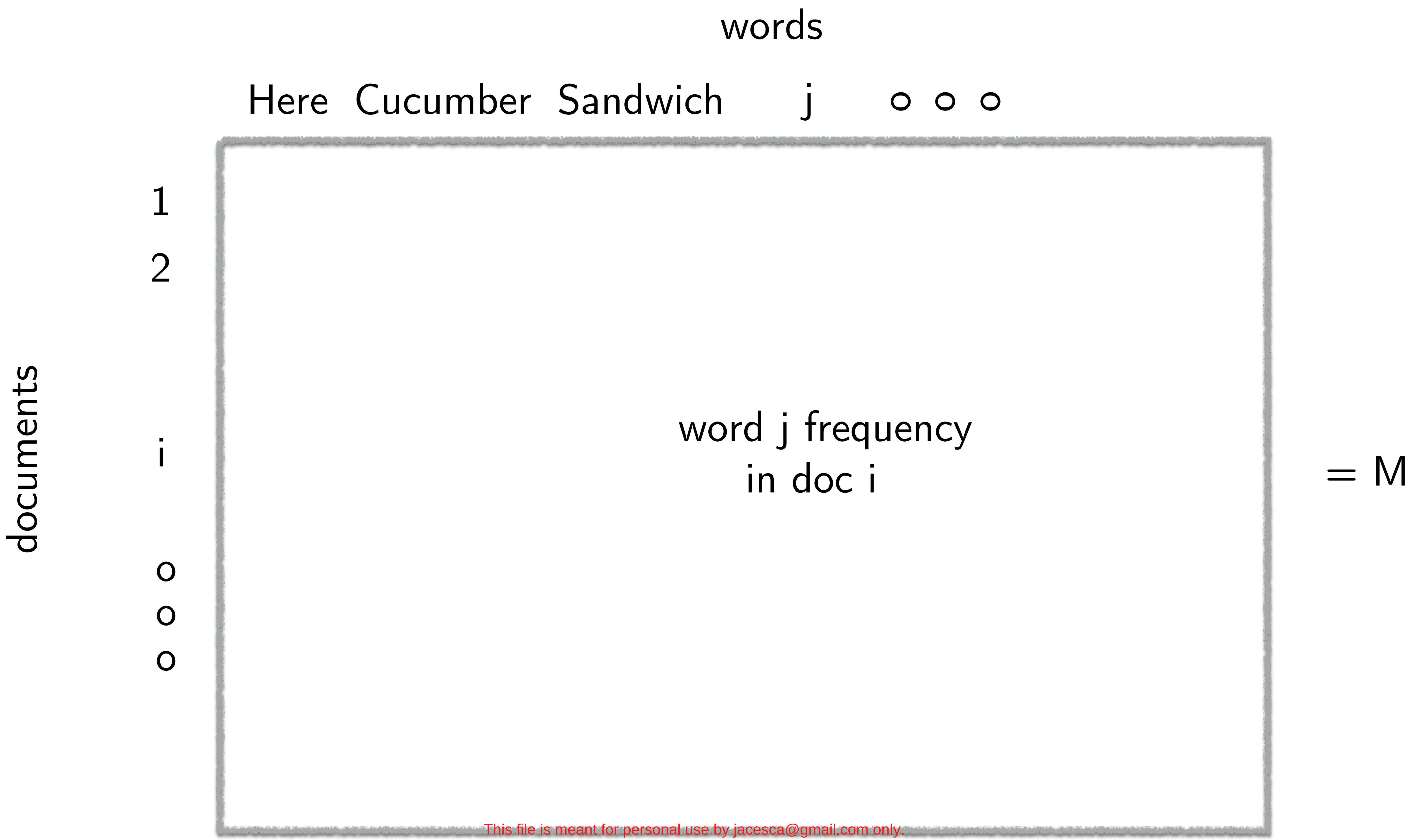
	business_id	compliment_count	date	text	user_id
0	UYX5zL_Xj9WEc_Wp-FrqHw	0	2013-11-26 18:20:08	Here for a quick mtg	hf27xTME3EiCp6NL6VtWZQ
1	Ch3HkwQYv1YKw_FO06vBWA	0	2014-06-15 22:26:45	Cucumber strawberry refresher	uEvusDwoSymbJJ0auR3muQ
2	rDoT-MgxGRiYqCmi0bG10g	0	2016-07-18 22:03:42	Very nice good service good food	AY-laIws3S7YXNl_f_D6rQ
3	OHXnDV01gLokiX1ELaQufA	0	2014-06-06 01:10:34	It's a small place. The staff is friendly.	Ue_7yUlkEbX4AhnYdUfL7g
4	GMrwDXRlAZU2zj5nH6l4vQ	0	2011-04-08 18:12:01	8 sandwiches, $24 total...what a bargain!!! An...	LltbT_fUMqZ-ZJP-vJ84IQ

Need an approach to convert text into number or vector of numbers

Appendix: Converting Content to Features

Text to vector of number:

Create word-frequency in documents matrix M



Appendix: Converting Content to Features

Text to vector of number:

- Create word-frequency in documents matrix M
- Perform Principal Component Analysis of M
  - Each document receives k co-ordinates
    - via k principal components
- This is the vector representing the text features (restricted to data)

Another (more classical) option:

- TF-IDF vector
- But it can be very large