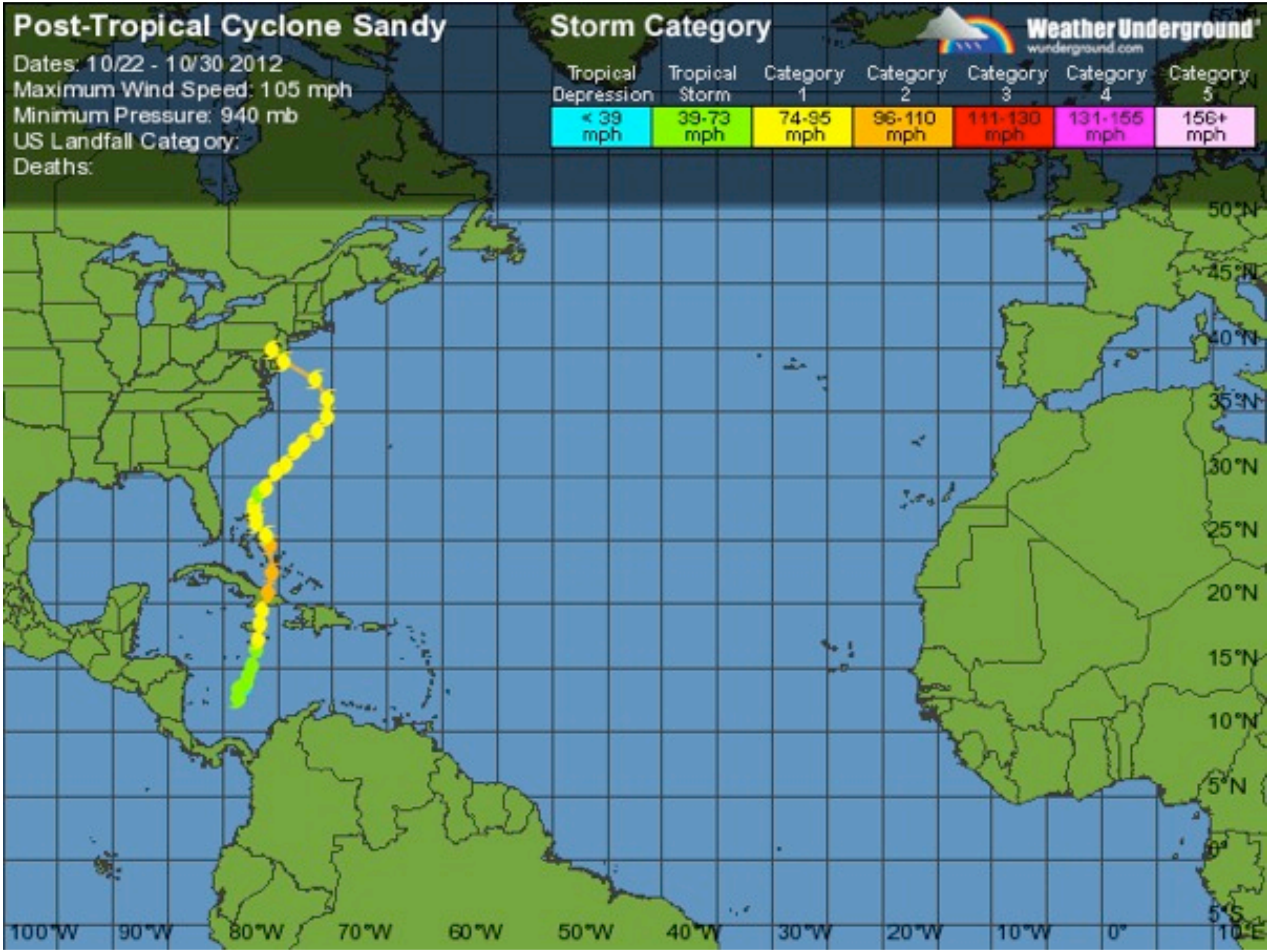
The details: training and validating big models on big data

PRINCETON UNIVERSITY

COMPUTING RESEARCH ASSOCIATION
CRA

MALLET
machine learning for language toolkit

David Mimno
Princeton, Computer Science


Post-Tropical Cyclone Sandy
Dates: 10/22 - 10/30 2012
Maximum Wind Speed: 105 mph
Minimum Pressure: 940 mb
US Landfall Category:
Deaths:
Storm Category
Weather Underground
wunderground.com
Tropical Depression
< 39 mph
Tropical Storm
39-73 mph
Category 1
74-95 mph
Category 2
96-110 mph
Category 3
111-130 mph
Category 4
131-155 mph
Category 5
156+ mph
50°N
45°N
40°N
35°N
30°N
25°N
20°N
15°N
10°N
5°N
0°
5°S
100°W
90°W
80°W
70°W
60°W
50°W
40°W
30°W
20°W
10°W
0°
10°E

George Dyson, Turing's Cathedral

- "The reaction of most meteorologists towards computer-assisted forecasting paralleled that of the Institute mathematicians towards computer-assisted mathematics: skepticism that a machine could improve upon what they were doing with brains alone."

George Dyson, Turing's Cathedral

- "The reaction of most meteorologists towards **computer-assisted** forecasting paralleled that of the Institute mathematicians towards computer-assisted mathematics: skepticism that a machine could improve upon what they were doing with brains alone."

Outline

- Training topic models
- Modeling choices
- Diagnostics

Outline

- **Training topic models**
- Modeling choices
- Diagnostics

Library-scale topic models

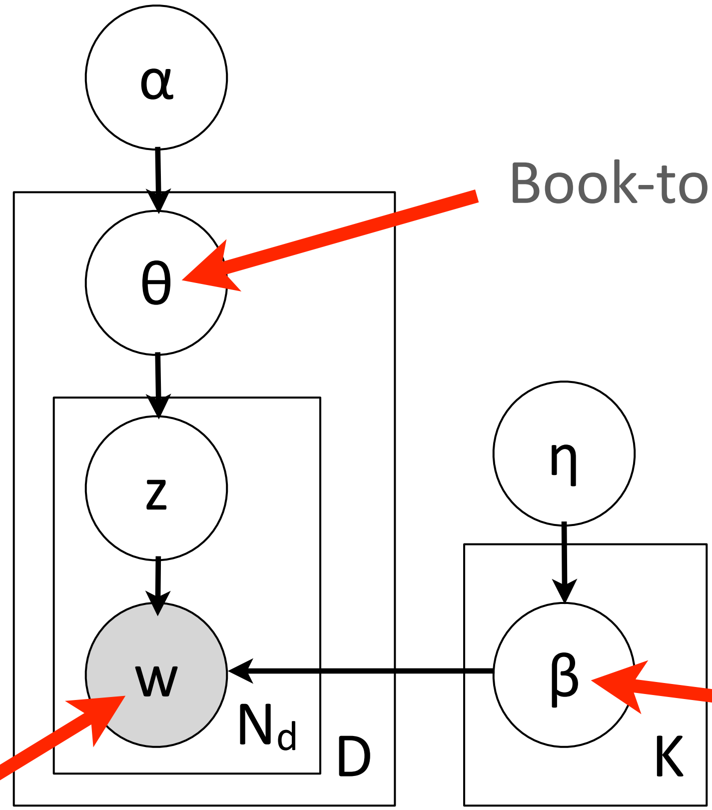
Input: 1.2M pre-1922 books
(33 billion non-stopwords)

Output: 2000 "topics"
(distributions over words)

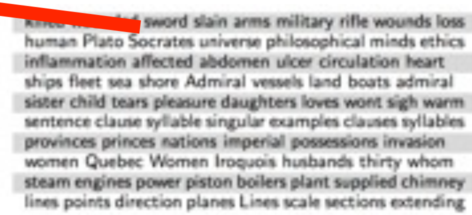
killed wounded sword slain arms military rifle wounds loss
human Plato Socrates universe philosophical minds ethics
inflammation affected abdomen ulcer circulation heart
ships fleet sea shore Admiral vessels land boats admiral
sister child tears pleasure daughters loves wont sigh warm
sentence clause syllable singular examples clauses syllables
provinces princes nations imperial possessions invasion
women Quebec Women Iroquois husbands thirty whom
steam engines power piston boilers plant supplied chimney
lines points direction planes Lines scale sections extending

Random examples, each row is a topic

Latent Dirichlet Allocation



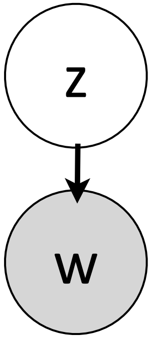
Book-topic proportions



An example document

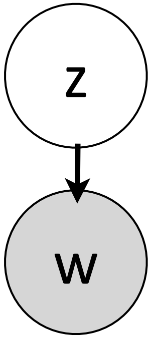
Etruscan	trade	price	temple	market

Assign topics

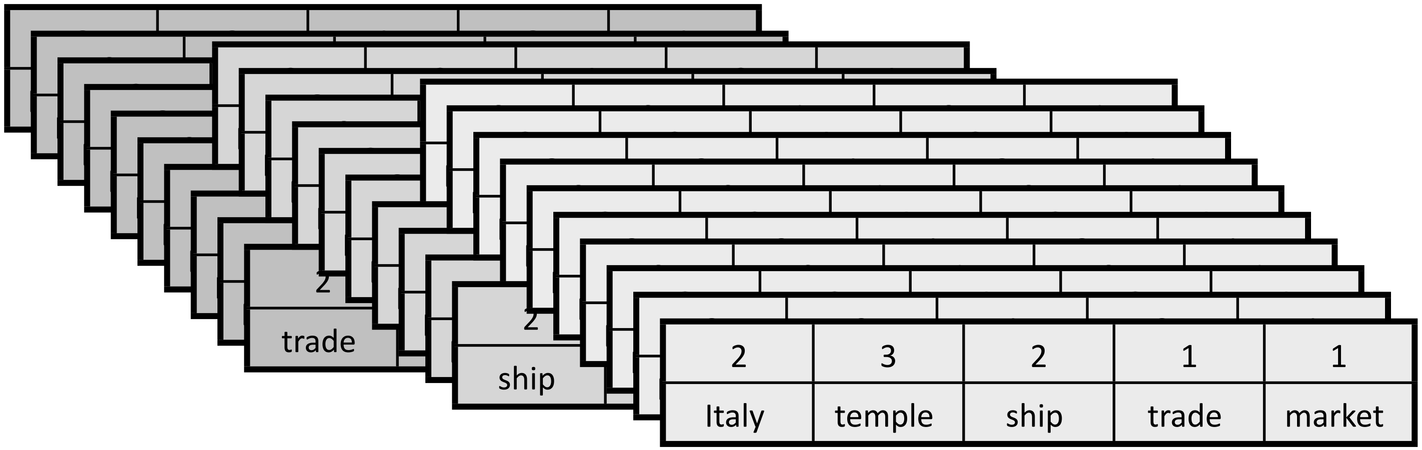


3	2	1	3	1
Etruscan	trade	price	temple	market

Assign topics



3	2	1	3	1
Etruscan	trade	price	temple	market



Global statistics

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from **all** docs

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

Algorithm

- Initialize topic assignments randomly
- For each iteration:
  - For each document:
    - For each word:
      - Resample topic for word, given all other words and their current topic assignments
- Produce reports

Algorithm

- Initialize topic assignments randomly
- For each iteration:
  - For each document:
    - For each word:
      - **Resample topic for word, given all other words and their current topic assignments**
- Produce reports

Sample topic for “trade”

3	**2**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
…			

Remove current assignment

3	**2**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**8**	1
...			

Remove current assignment

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**7**	1
...			

Pick a topic for “trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

Which topics occur in this doc?

3	?	1	3	1
Etruscan	trade	price	temple	market

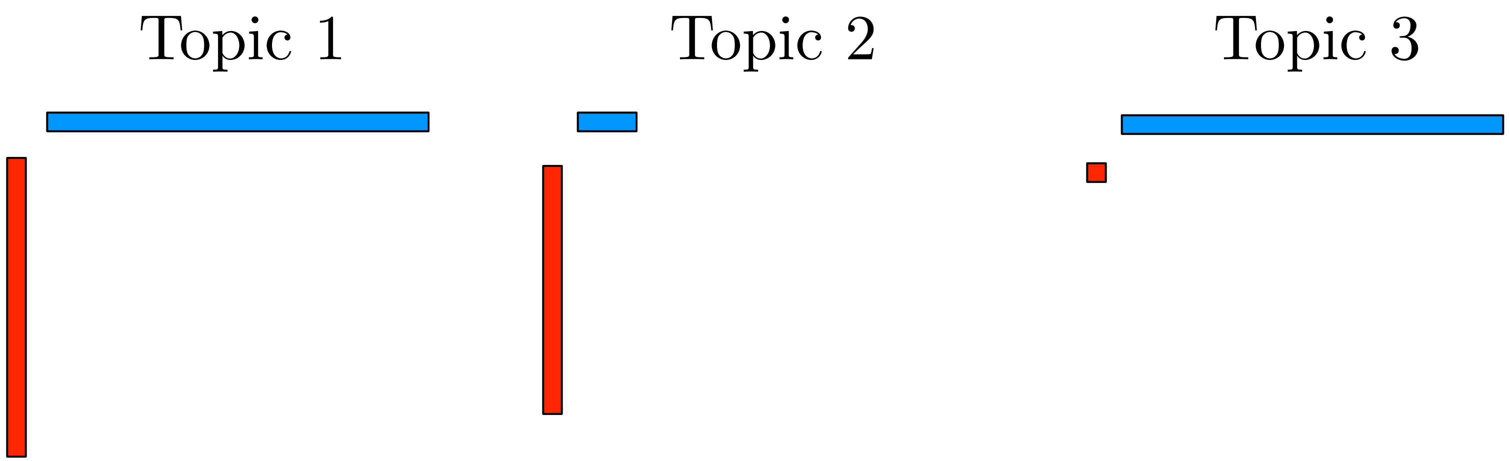
Which topics like the word “trade”?

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
trade	10	7	1

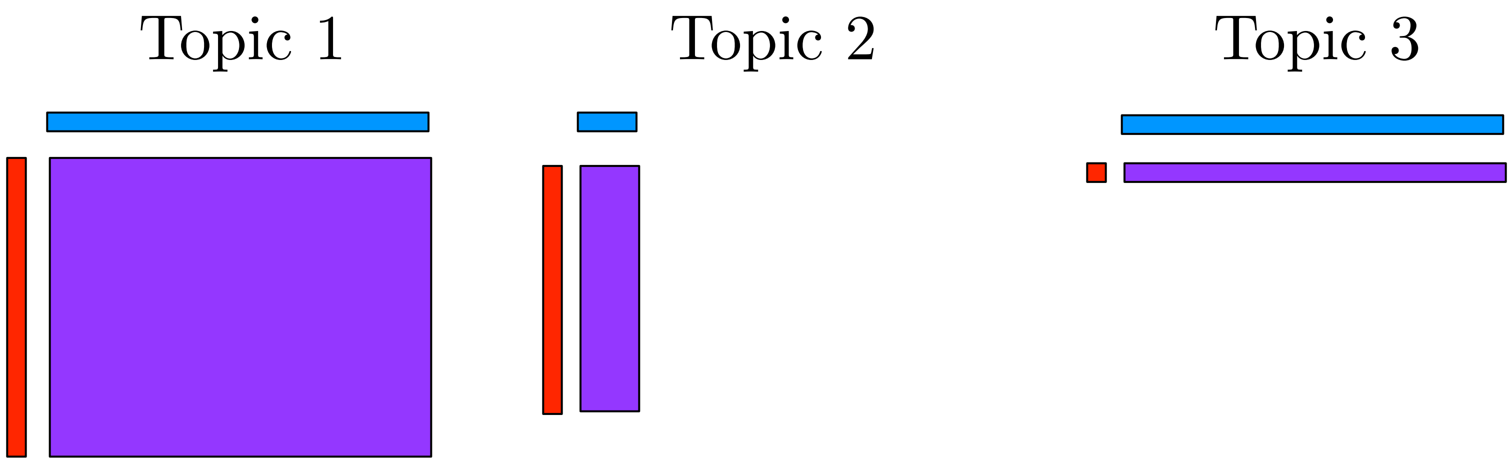
Which topics like the word "trade"?

3	?	1	3	1
Etruscan	trade	price	temple	market



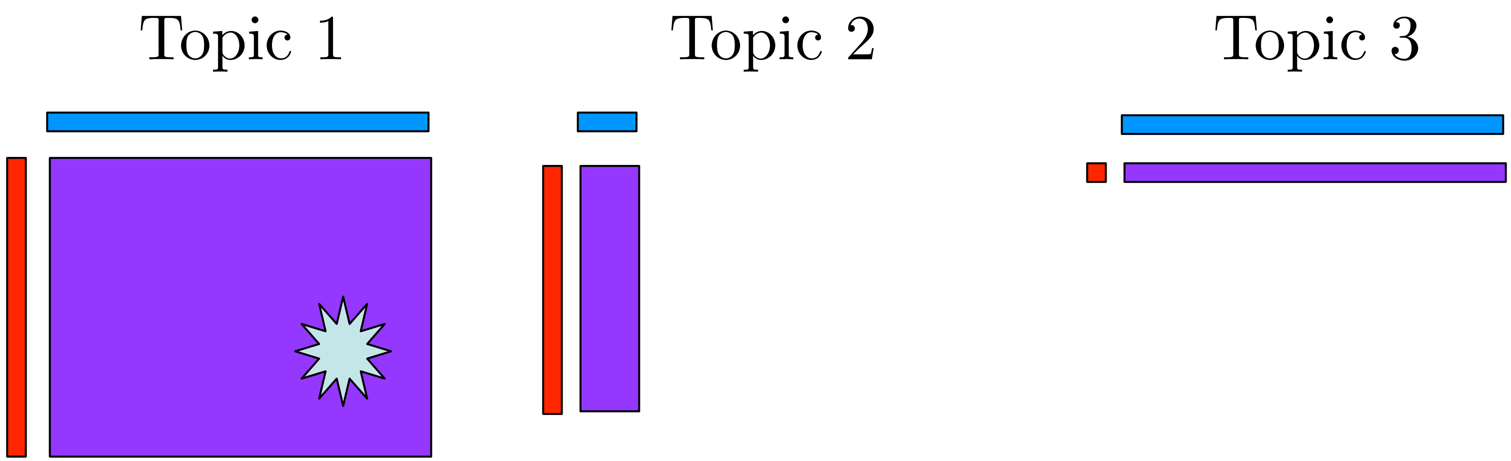
Pick a topic for “trade”

3	?	1	3	1
Etruscan	trade	price	temple	market



Pick a topic for “trade”

3	?	1	3	1
Etruscan	trade	price	temple	market



Pick a topic for "trade"

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**10**	7	1
...			

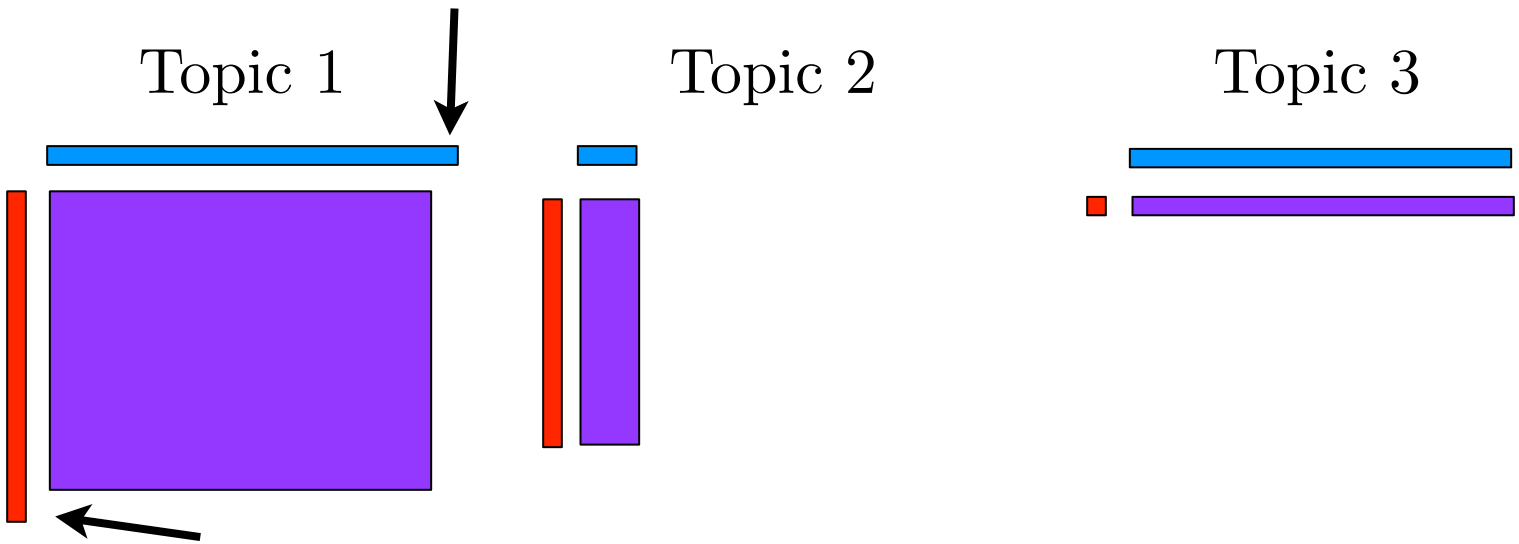
Pick a topic for "trade"

3	**1**	1	3	1
Etruscan	trade	price	temple	market

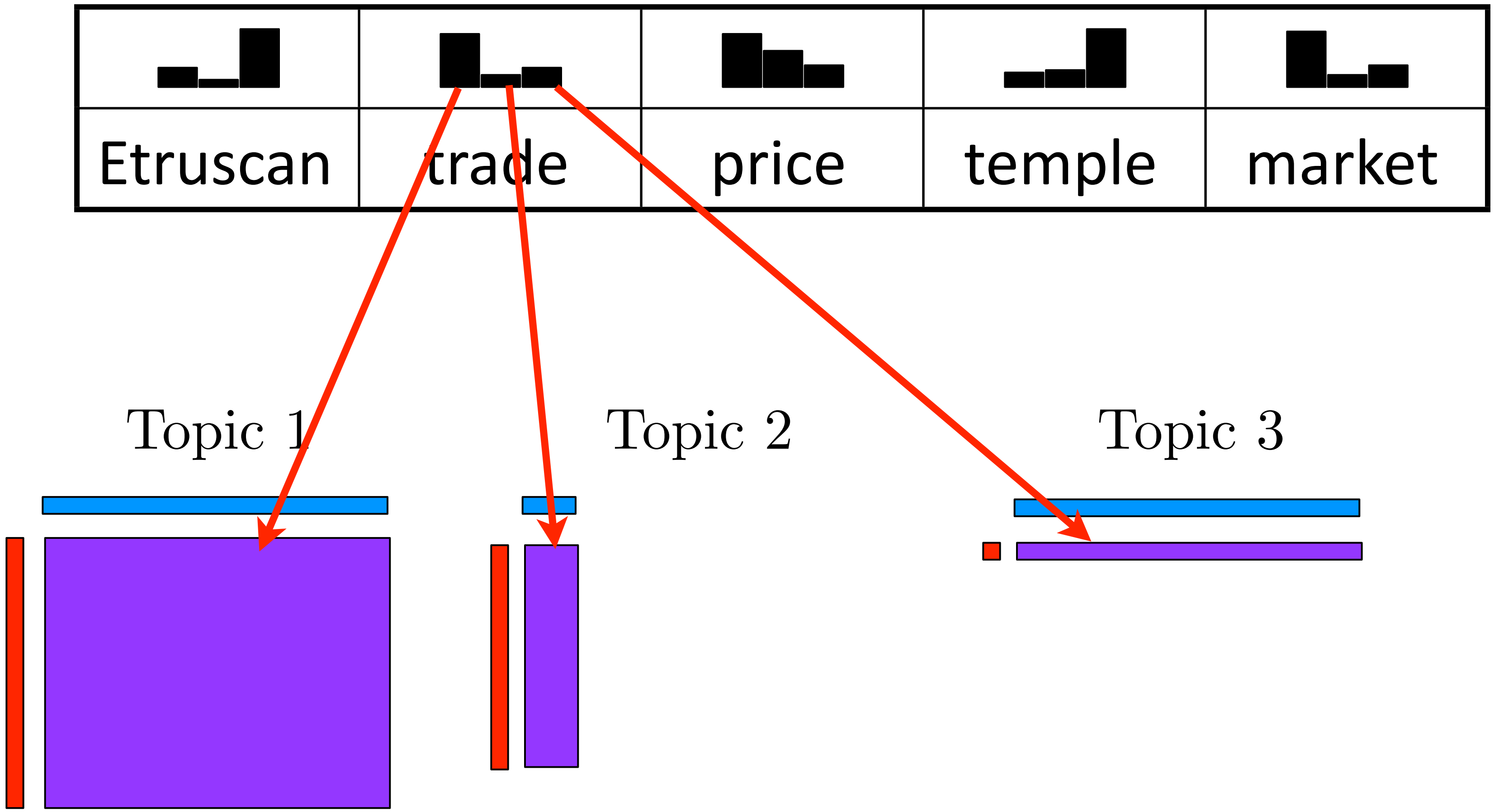
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**11**	7	1
…			

Increase counts for 1 and “trade”|1

3	1	1	3	1
Etruscan	trade	price	temple	market



Variational inference



Outline

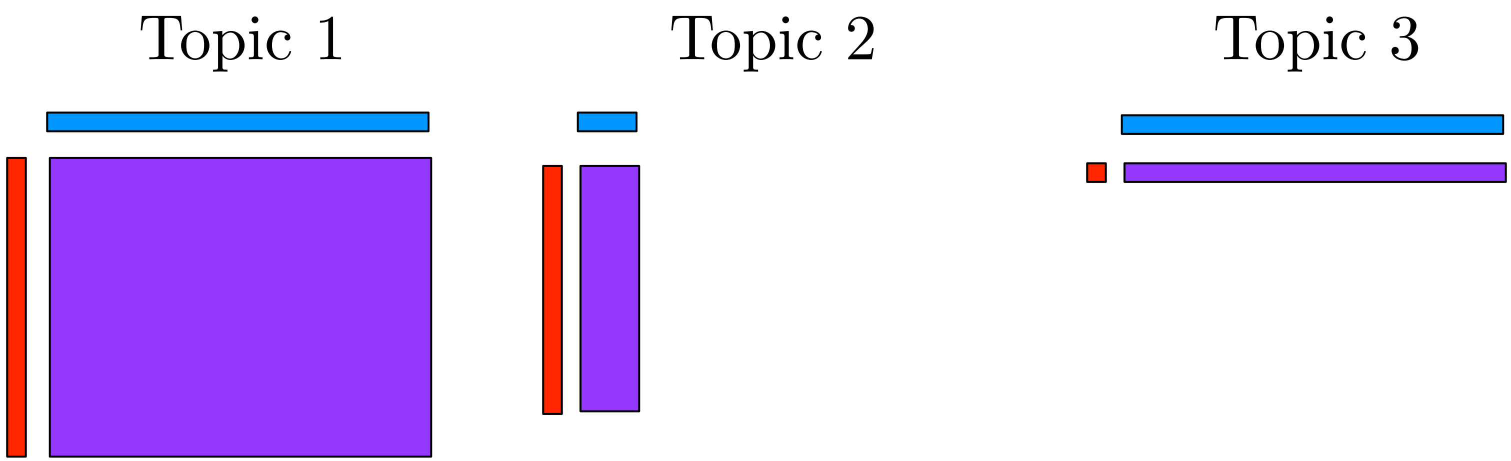
- Training topic models
- **Modeling choices**
- Diagnostics

Things people didn't know they had to think about

- What is a **document**?
- Which **words** are interesting?
- What is a word, anyway?
- Knobs:
  - Number of topics
  - Hyper-parameters

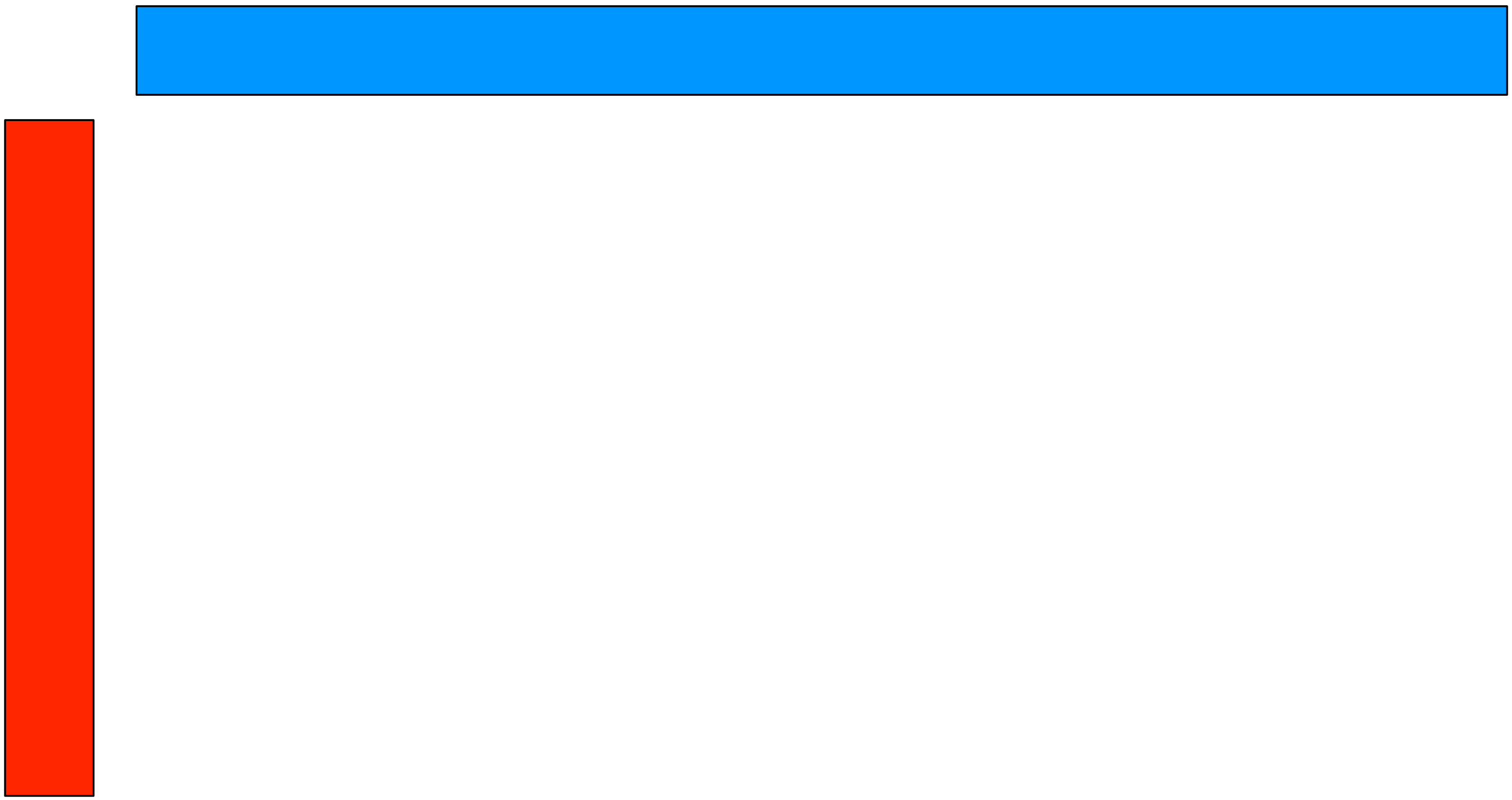
Pick a topic for "trade"

3	?	1	3	1
Etruscan	trade	price	temple	market

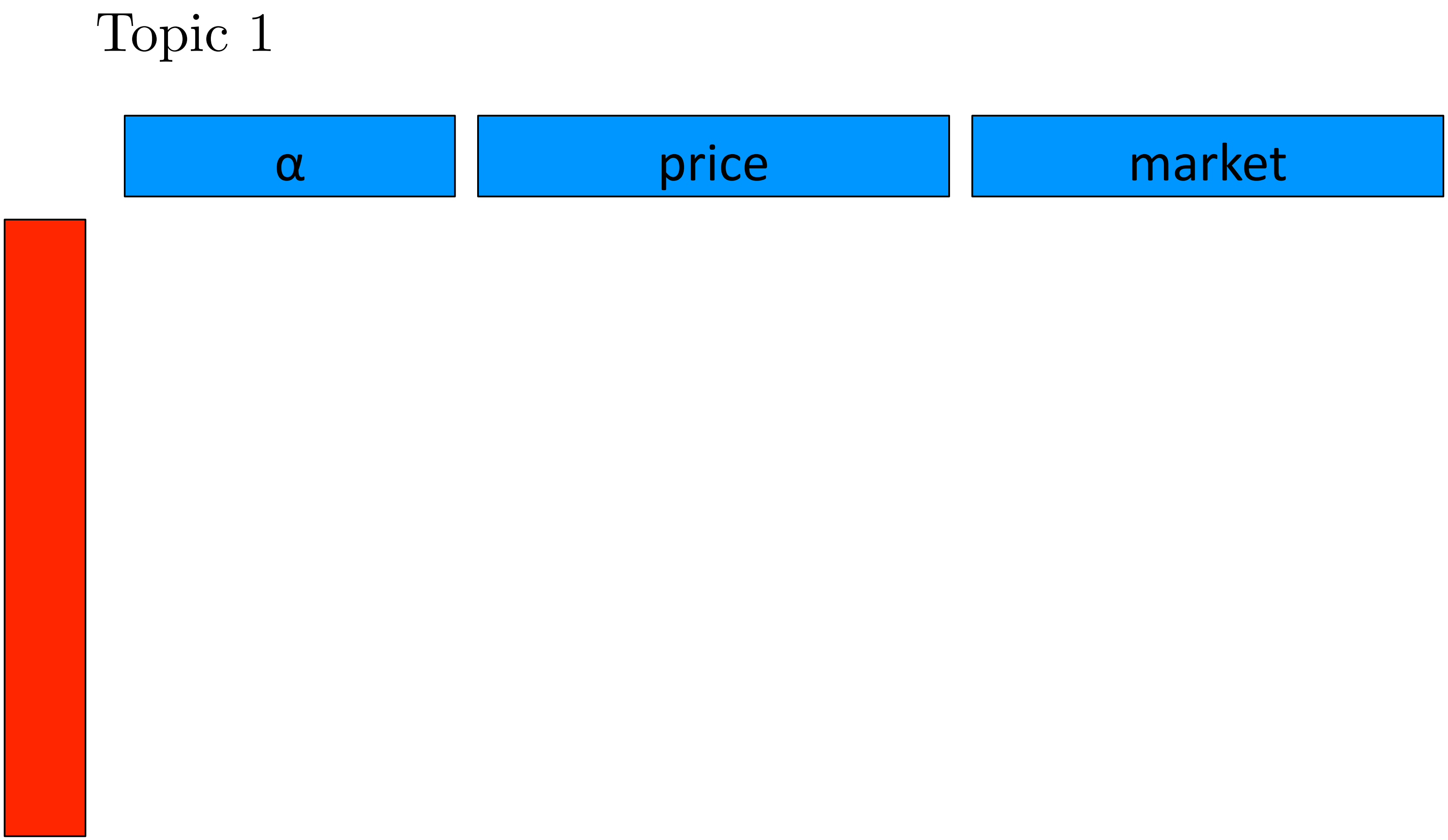


Which topics like the word “trade”?

Topic 1

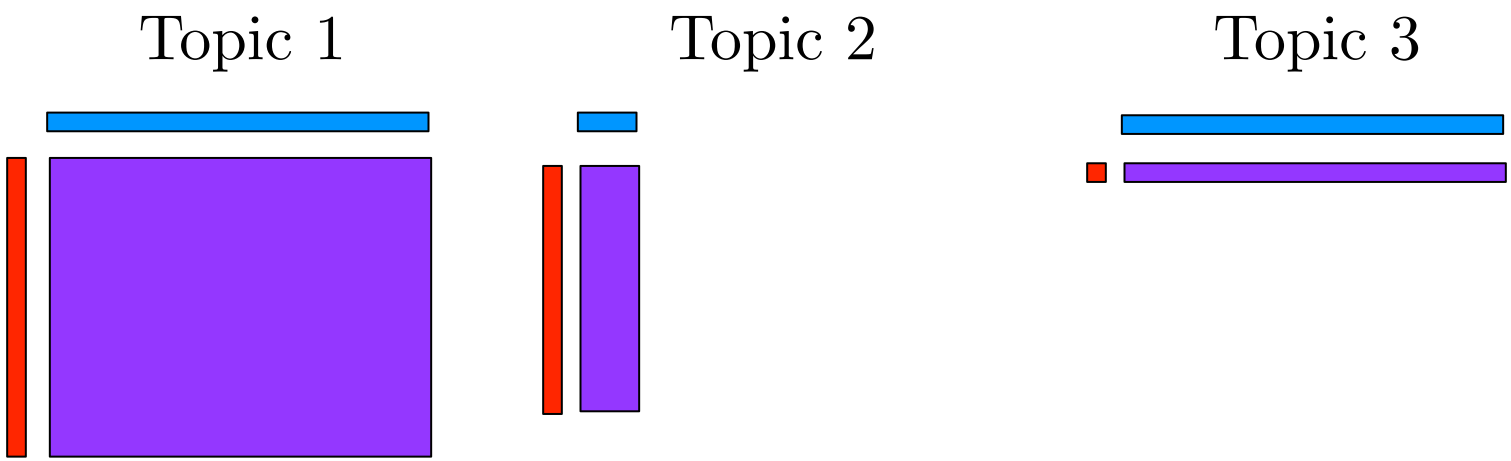


Which topics like the word “trade”?



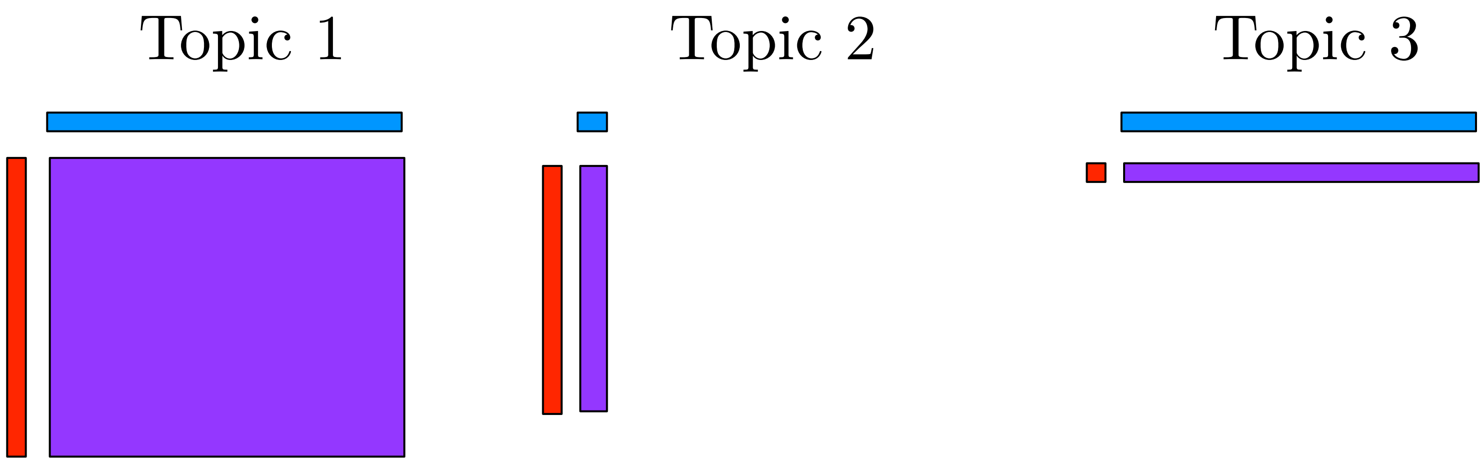
Pick a topic for “trade”

3	?	1	3	1
Etruscan	trade	price	temple	market



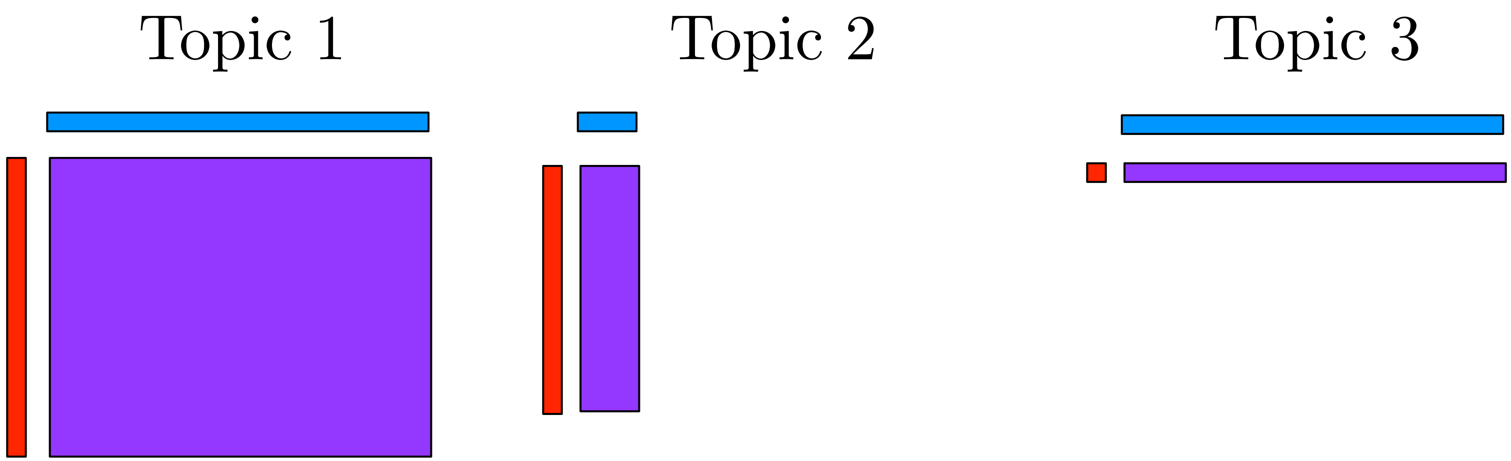
Pick a topic for "trade"

3	?	1	3	1
Etruscan	trade	price	temple	market



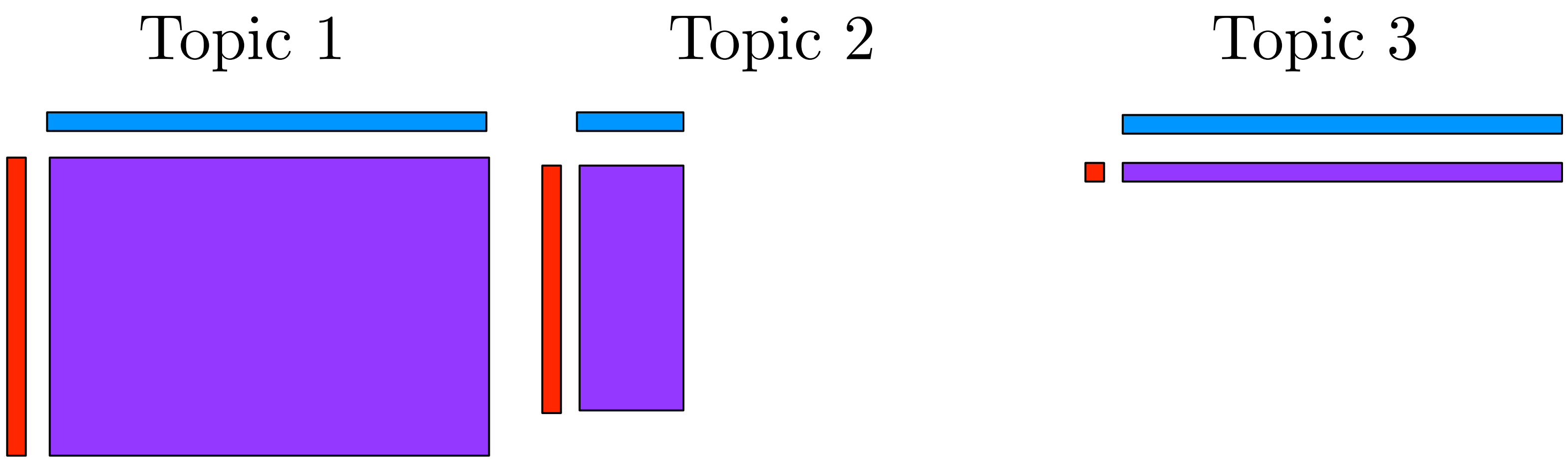
Pick a topic for “trade”

3	?	1	3	1
Etruscan	trade	price	temple	market



Pick a topic for "trade"

3	?	1	3	1
Etruscan	trade	price	temple	market



Hyper-parameters: learn or fix?

	Pros	Cons
Fixed	All topics similar size, quality	Duplicate topics, frequent words repeated
Learned	Some topics big, others small	Small topics may be low quality

Outline

- Training topic models
- Modeling choices
- **Diagnostics**

What makes topics bad?

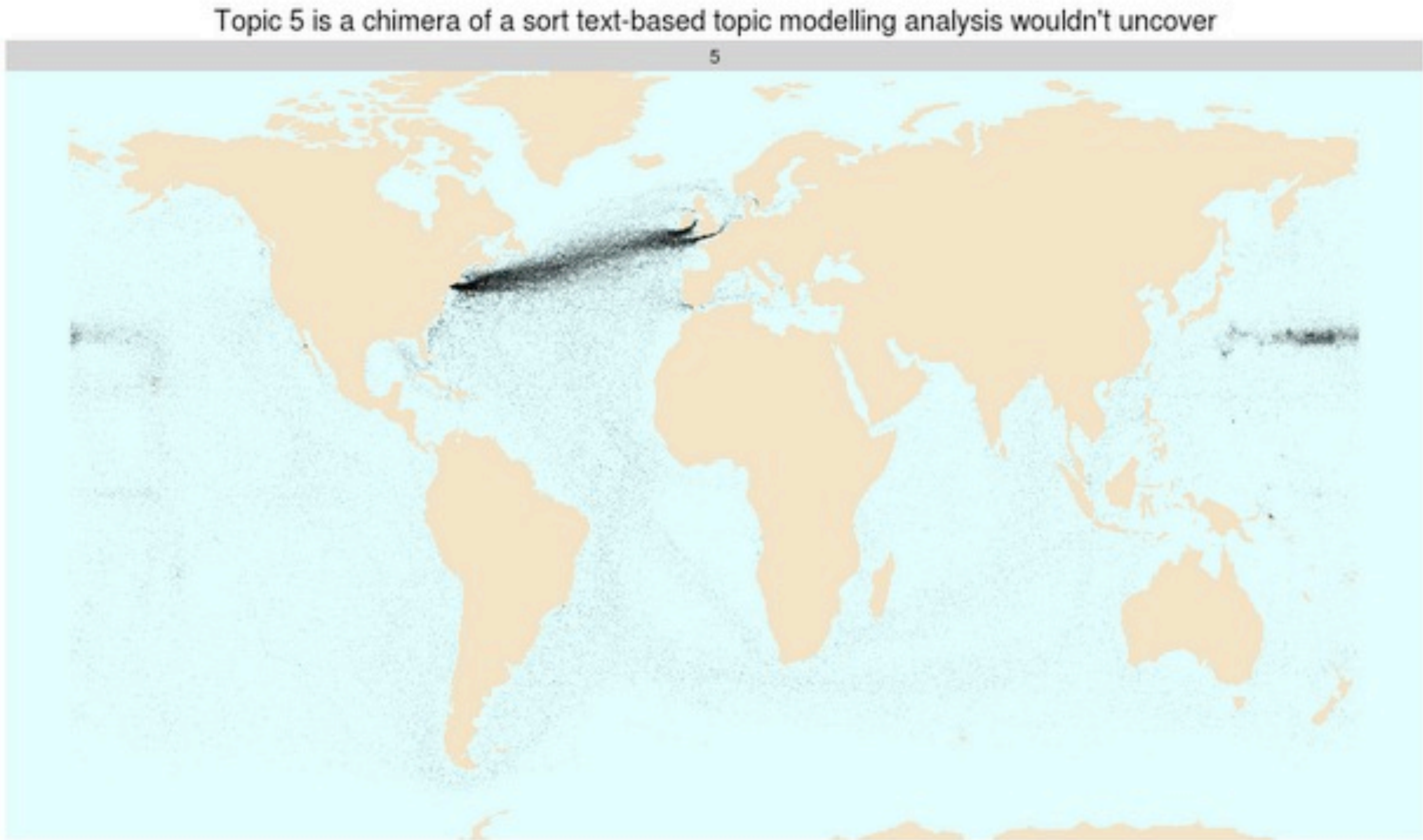
- **Random**, unrelated words
- One or two "**intruder**" words
- Boring, **overly general** words
- Two or more good topics combined, sometimes with a general word in common (**chimaeras**)

Example topic

- aging, lifespan, globin, age related, longevity, human, age, erythroid, sickle cell, beta globin, hb, senescence, adult, older, lcr

Example topic

- **aging, lifespan**, globin, **age related**, **longevity, human, age**, erythroid, sickle cell, beta globin, hb, **senescence**, **adult**, **older**, lcr


Topic 5 is a chimera of a sort text-based topic modelling analysis wouldn't uncover
5

@benschmidt

Evaluations of topic quality

1. Size (# of tokens assigned)
2. Within-doc rank
3. Similarity to corpus-wide distribution
4. Locally-frequent words
5. Co-doc Coherence

All of these are in Mallet 2.0.7!

Topic size

- How many words in the corpus are assigned to this topic?
- Fewer words, lower quality topics.

Within-doc rank

- For every doc, rank topics by frequency.
- In what proportion of documents is a topic the **most prominent** topic?
- General topics: **small** proportion of **many** documents.
- Focused topics: **large** proportion of **few** documents.

Similarity to corpus dist'n

- Rank **all** words in corpus in order.
- Measure similarity of each topic to this global ranking.
- Topics with **high similarity** to the whole corpus are usually uninteresting.

Locally frequent words

- If a rare word occurs in a document, it will occur often.
- In **long documents** unusual words can have high frequency.
- Compare "topics" generated by word token count to "topics" generated by **document** count.

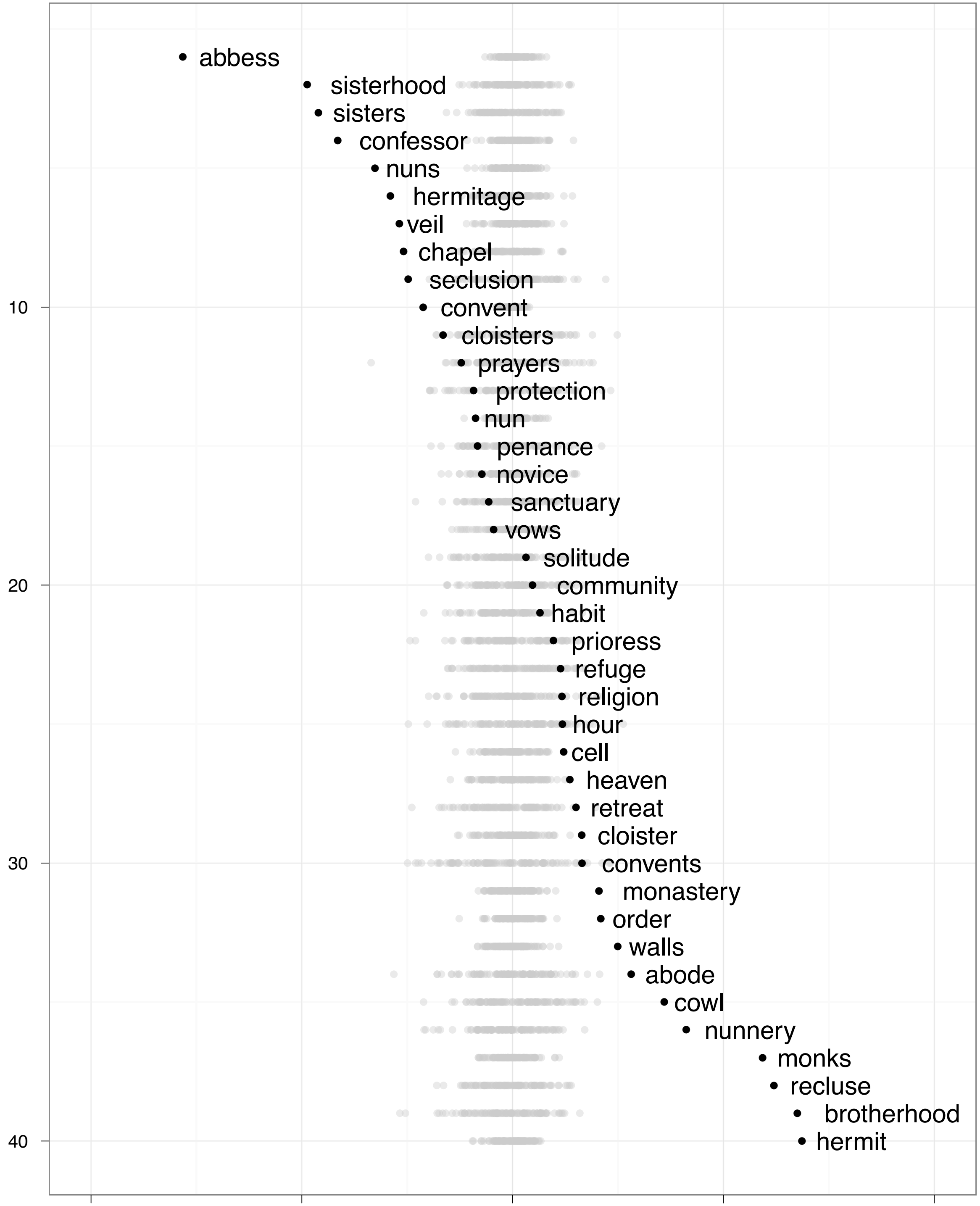
Co-doc "coherence"

- Use the training document set
- Create binarized co-document frequencies
- Compare conditional probability of each word to all *higher-ranked* words

log P('erythroid' | 'aging')

Co-document frequencies

	aging	lifespan	erythroid
aging	100	25	0
lifespan	25	50	0
erythroid	0	0	25



abbess
sisterhood
sisters
confessor
nuns
hermitage
veil
chapel
seclusion
convent
cloisters
prayers
protection
nun
penance
novice
sanctuary
vows
solitude
community
habit
prioress
refuge
religion
hour
cell
heaven
retreat
cloister
convents
monastery
order
walls
abode
cowl
nunnery
monks
recluse
brotherhood
hermit
10
20
30
40