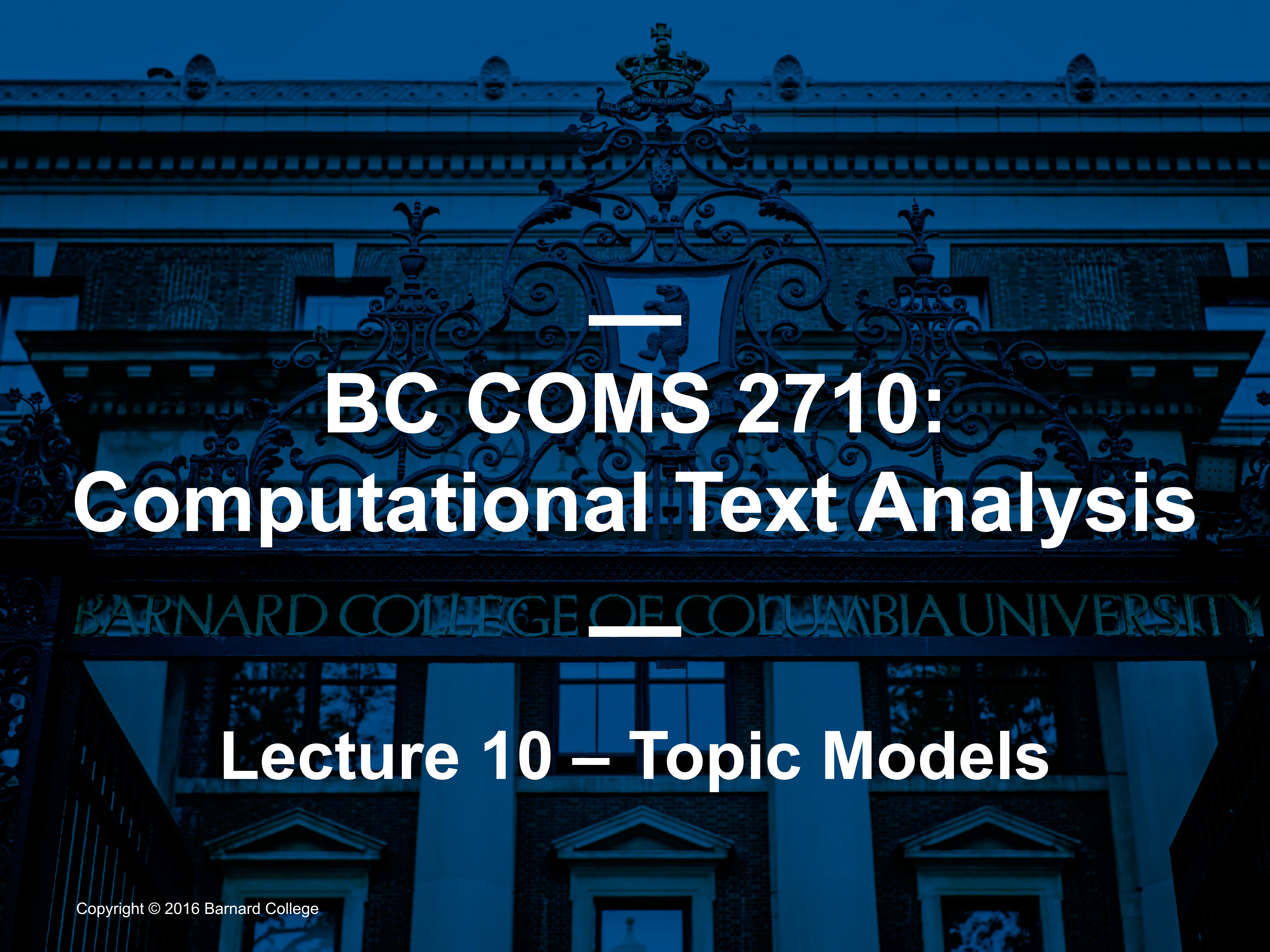


BC COMS 2710: Computational Text Analysis

Lecture 10 – Topic Models

Announcements – Assignments

- Tutorial 2.2
  - Due tomorrow night (Thursday, 04/20)
  - Long - Broken into lots of small steps

- Readings:
  - Reading 03 – link course site, due Sunday

- HW 02:
  - Released later today
  - Open ended assignment

- Office hours

Guest speakers: June 1st, 9th

- Guest Speakers:
  - Maria Antoniak:
    - PhD student @ Cornell – June 1st
  - Lucy Li
    - PhD student @ Berkeley – June 9th
    - Author of *Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas US History Textbooks*

- Attendance is required

Course so far

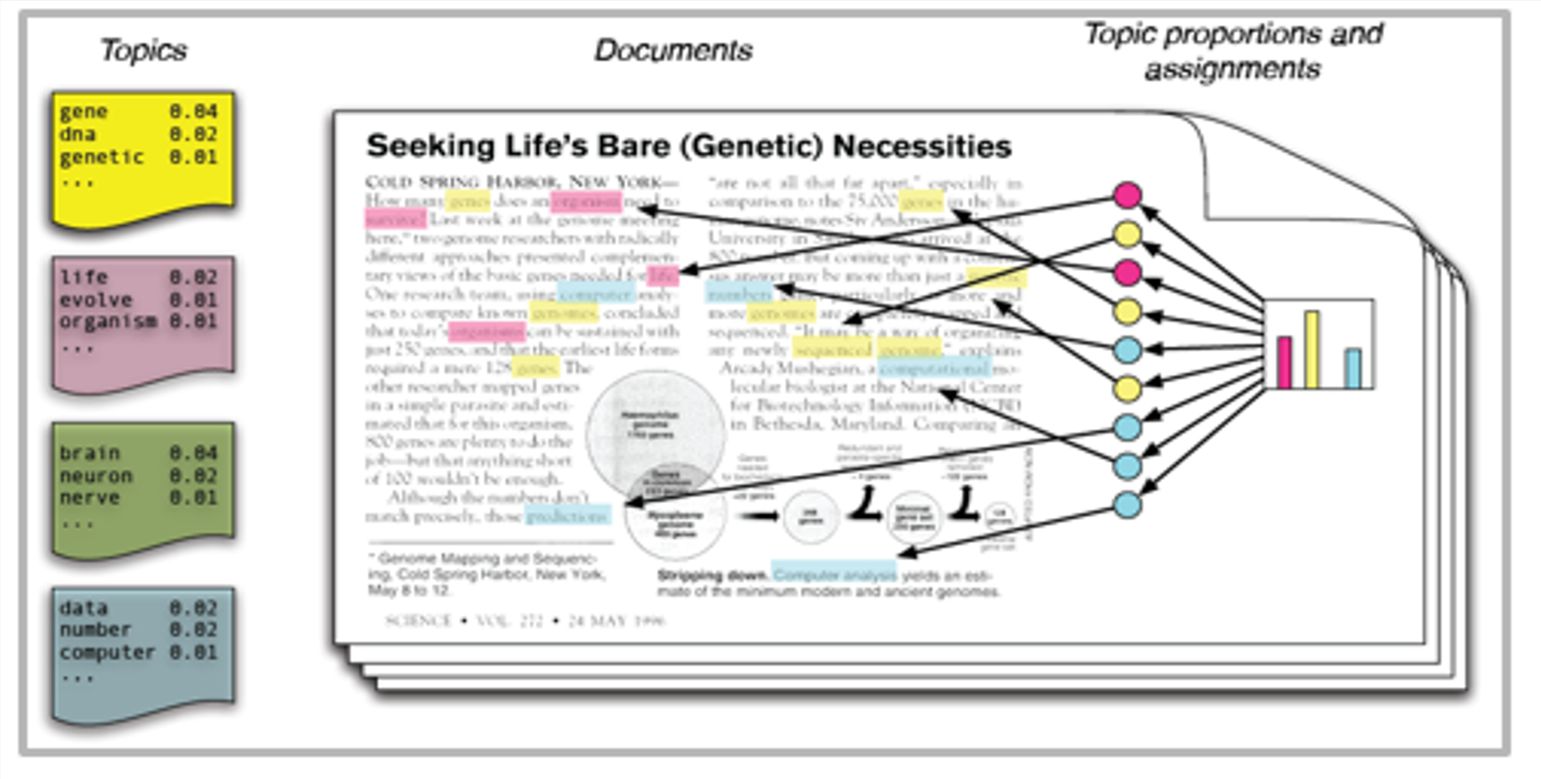
- Insights from individual words
  - TF-IDF
  - Dictionary based methods
- Insights from specific documents
  - Readability
- Today: Group individual words into larger constructs

Course Outline

▪ Python Overview	**Week 1**
▪ Lexical based analysis methods	**Week 2 - 3**
• Text Processing	
• Document Representation	
• Topic Modeling	
▪ Data Collection	**Week 4**
• Web Scraping	
• APIs	
▪ Machine Learning	**Week 5**
• Regression & Classification	
• Clustering	
▪ Advanced Topics & Final Projects	
• Dimensionality Reduction	
• Word Representations	**Week 6**

Topic Modeling

- Goal: Identify underlying topics across documents



Slide from Federico Nanni

What are topics?

Observation

Tokens that are likely to appear in the same context

Hidden structure that determines how **tokens** appear in a corpus

Want to uncover

Slide from Federico Nanni

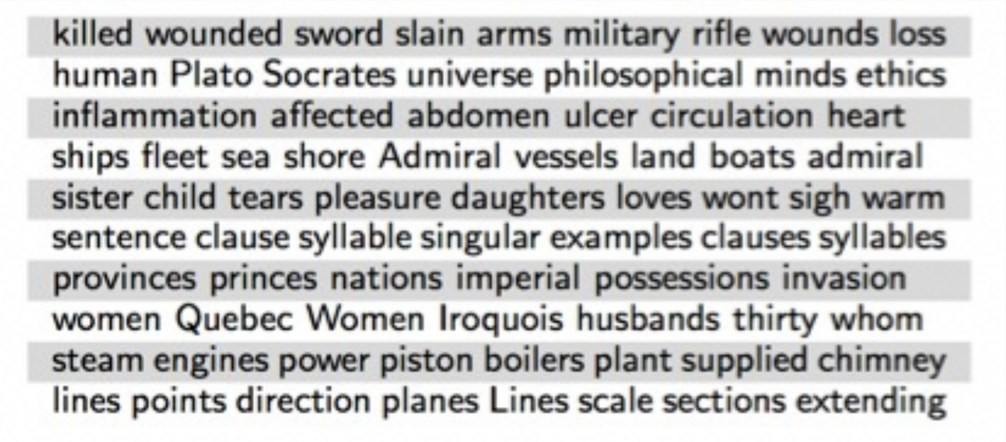
Topic Modeling: Corpora -> Topics

Input:

Millions of Books

Output: topics

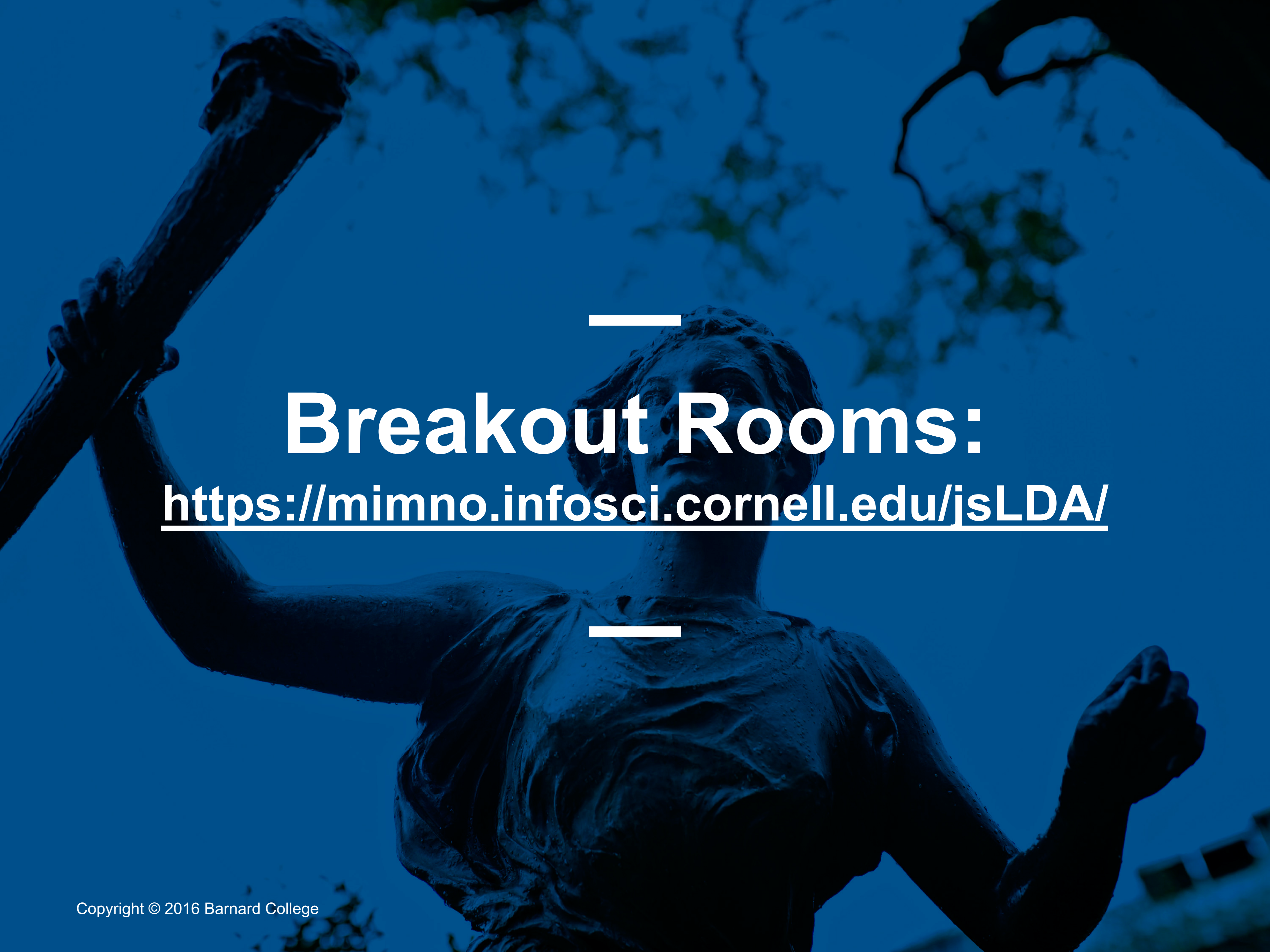
(distributions over words)



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

Each row is a topic

Slide from David Mimno



Breakout Rooms:

https://mimno.infosci.cornell.edu/jsLDA/

Discovering Topics

How do we discover topics?

- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Latent Dirichlet Allocation

How do we discover topics?

- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- **Latent Dirichlet Allocation**

LDA

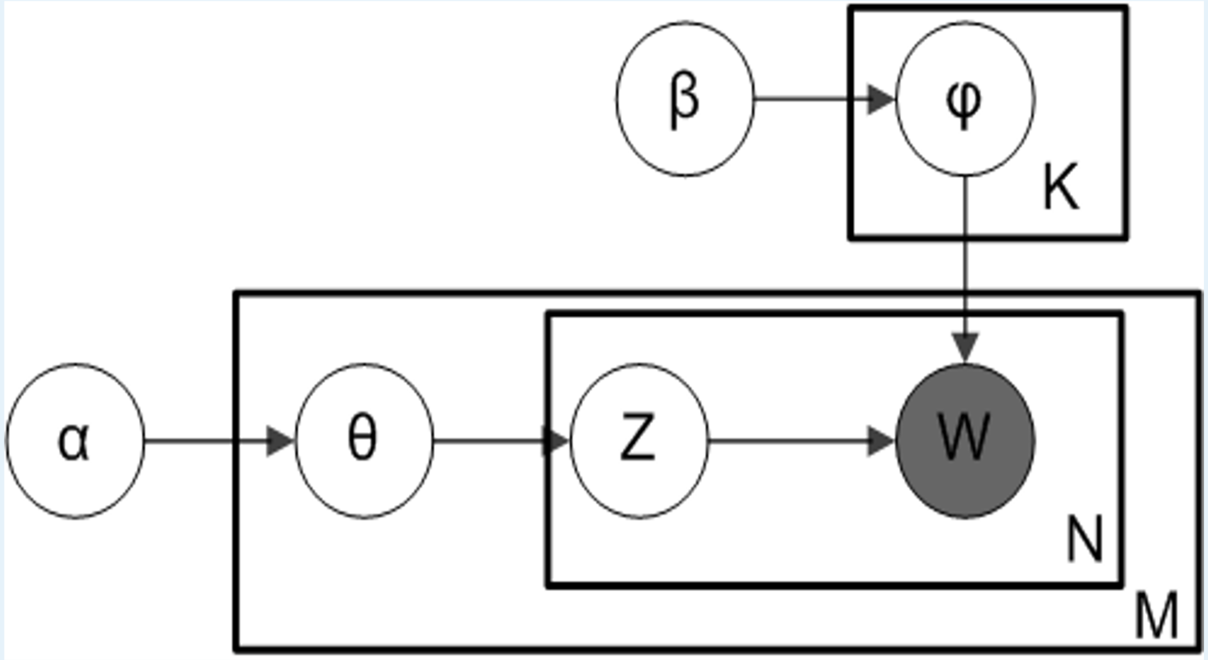
- Probabilistic model
- Generative model

LDA Generative Story

- Each word appears independent of each other

- Each word depends on the topic
  - Topics have a distribution of words
  - Topics have a distribution of documents

LDA Plate Notation

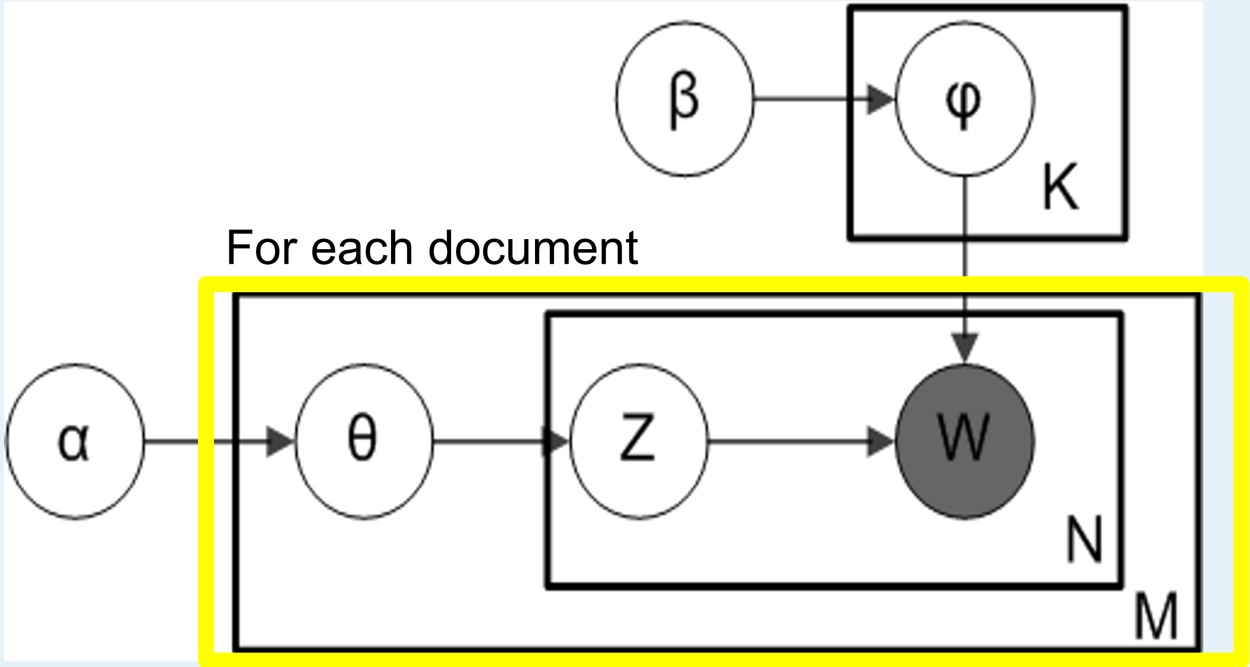


LDA Plate Notation

M = number of documents

N = number of words in a document

K = number of topics (we choose this)

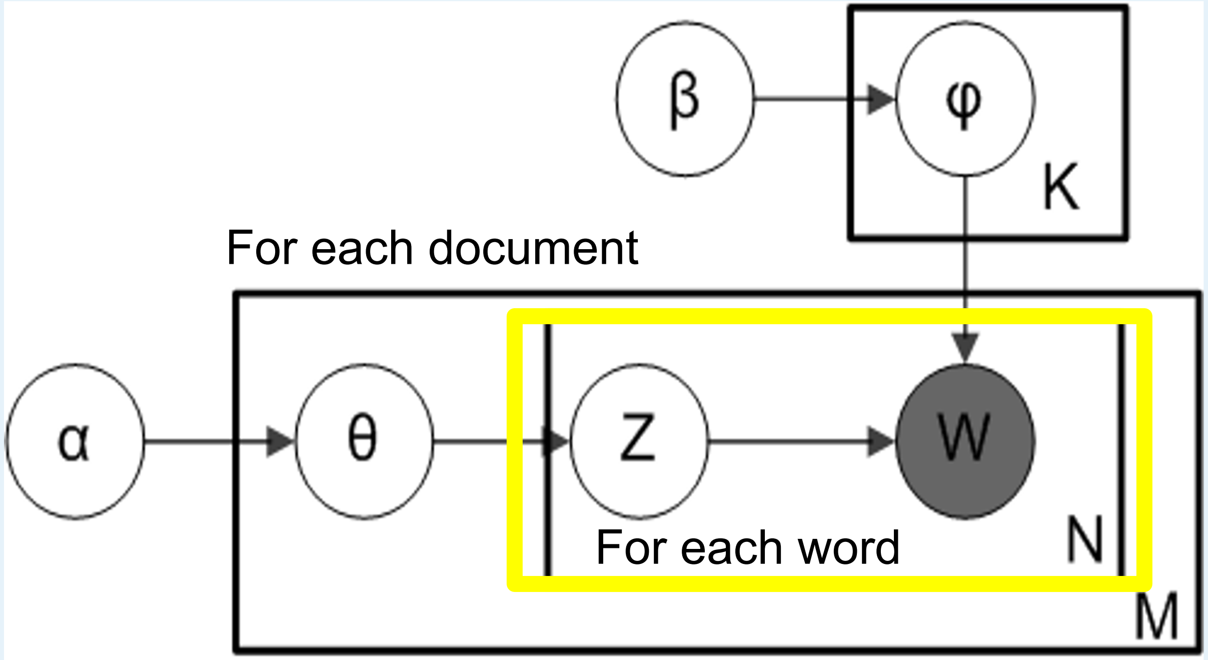


LDA Plate Notation

M = number of documents

N = number of words in a document

K = number of topics (we choose this)

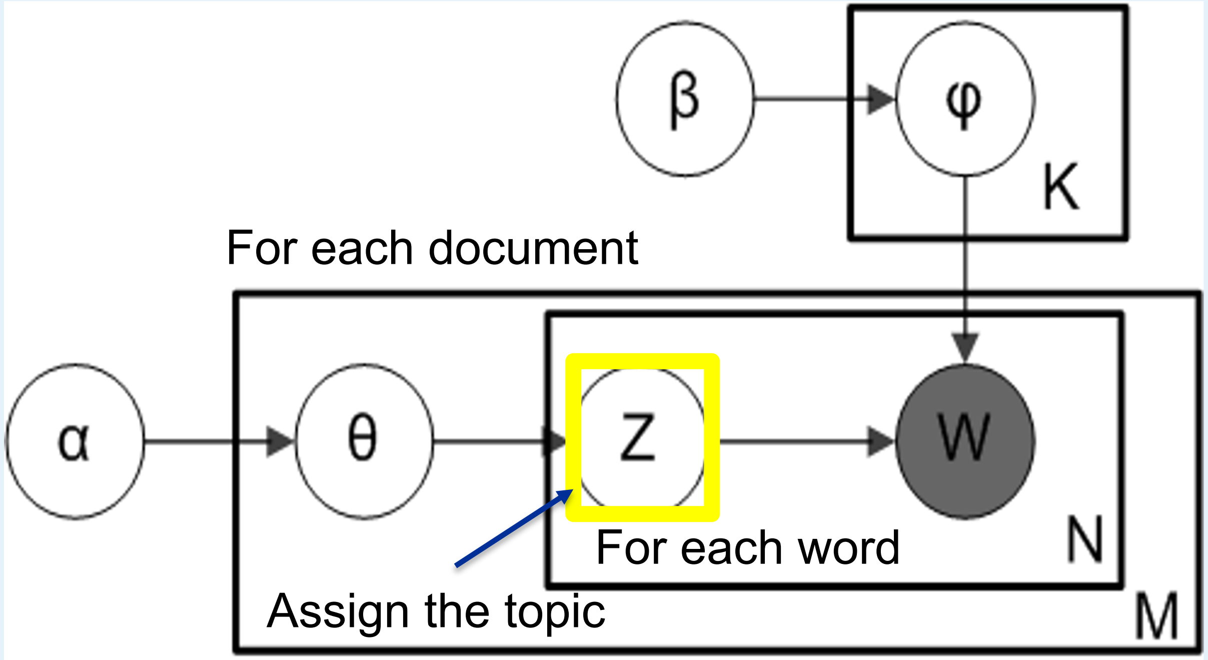


LDA Plate Notation

M = number of documents

N = number of words in a document

K = number of topics (we choose this)

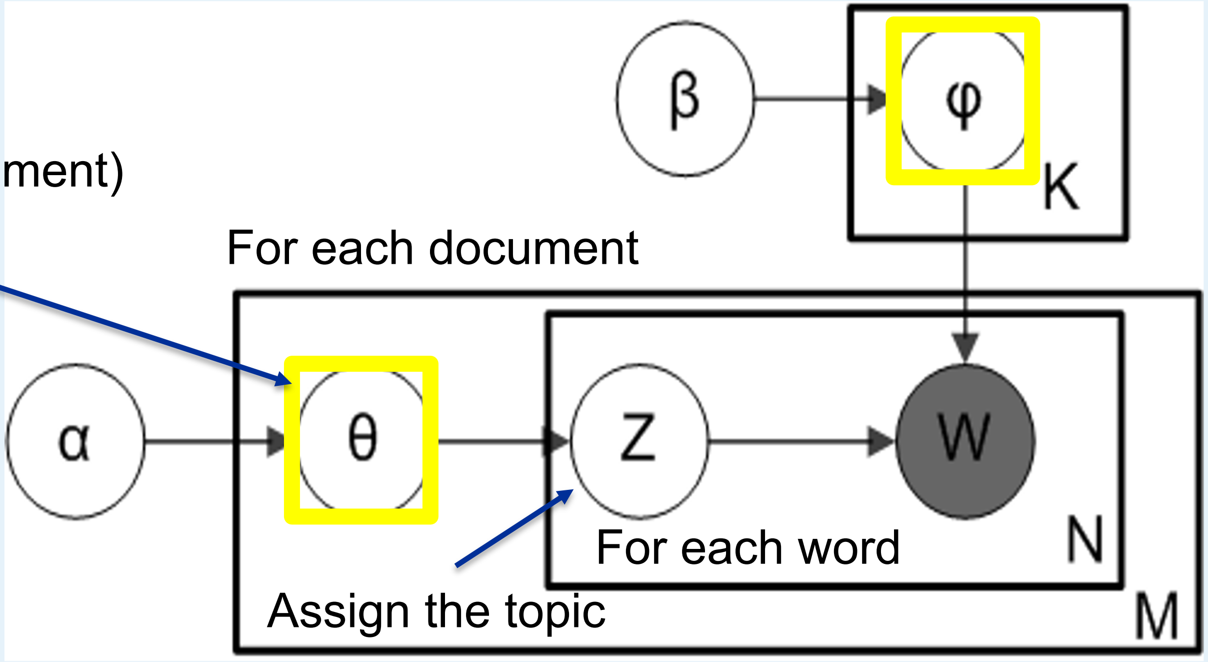


LDA Plate Notation

Given p(word | topic)

Given p(topic | document)

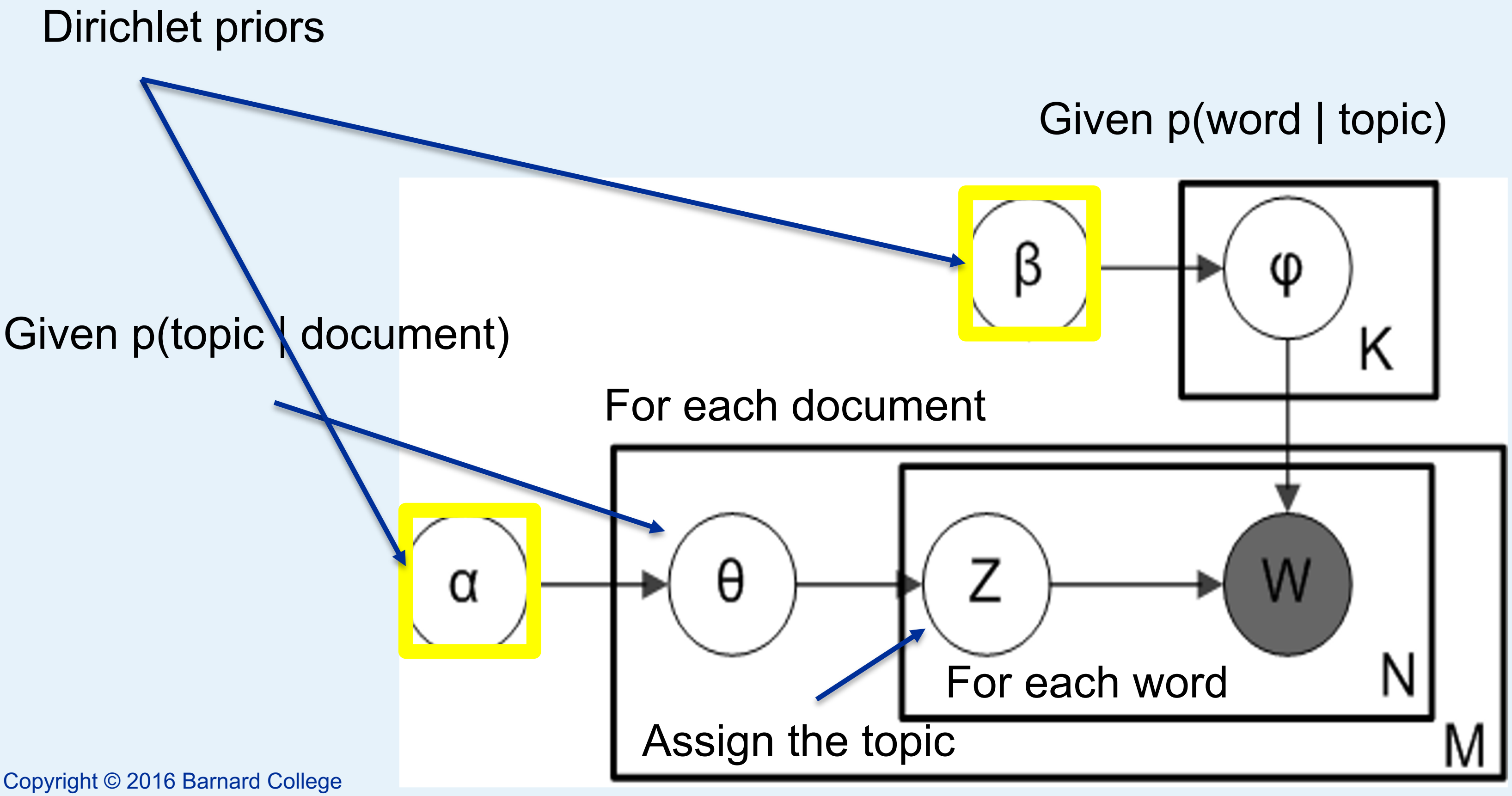
For each document



For each word

Assign the topic

LDA Plate Notation



Dirichlet priors

Given p(word | topic)

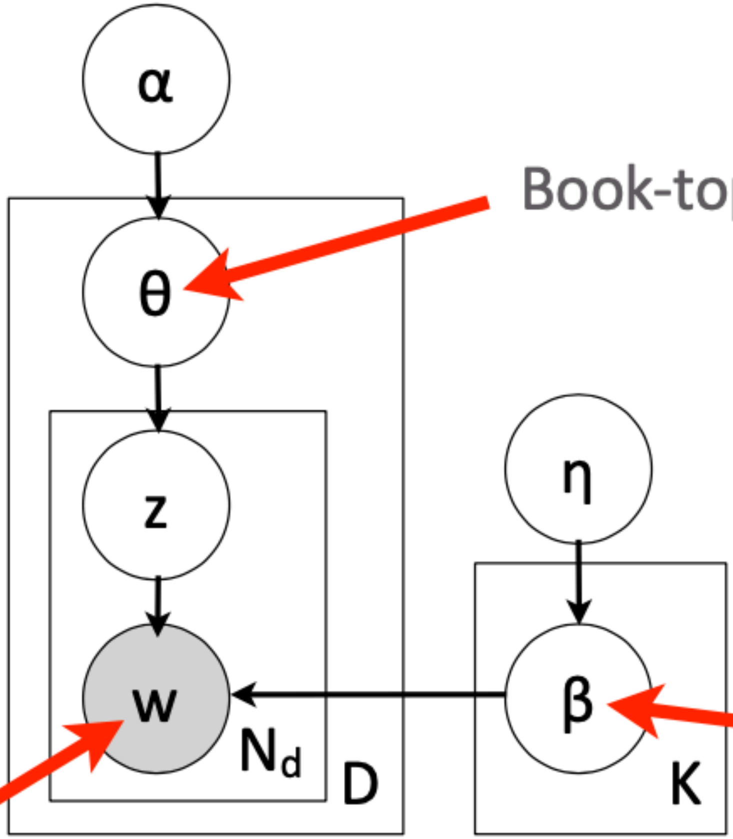
Given p(topic | document)

For each document

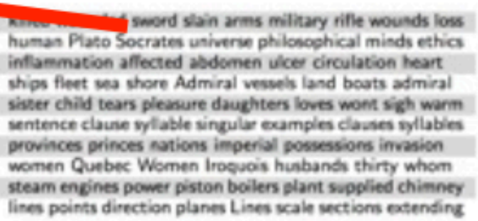
For each word

Assign the topic

LDA Plate Notation



Book-topic proportions



killed [illegible] 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

Slide from David Mimno

LDA Plate notation

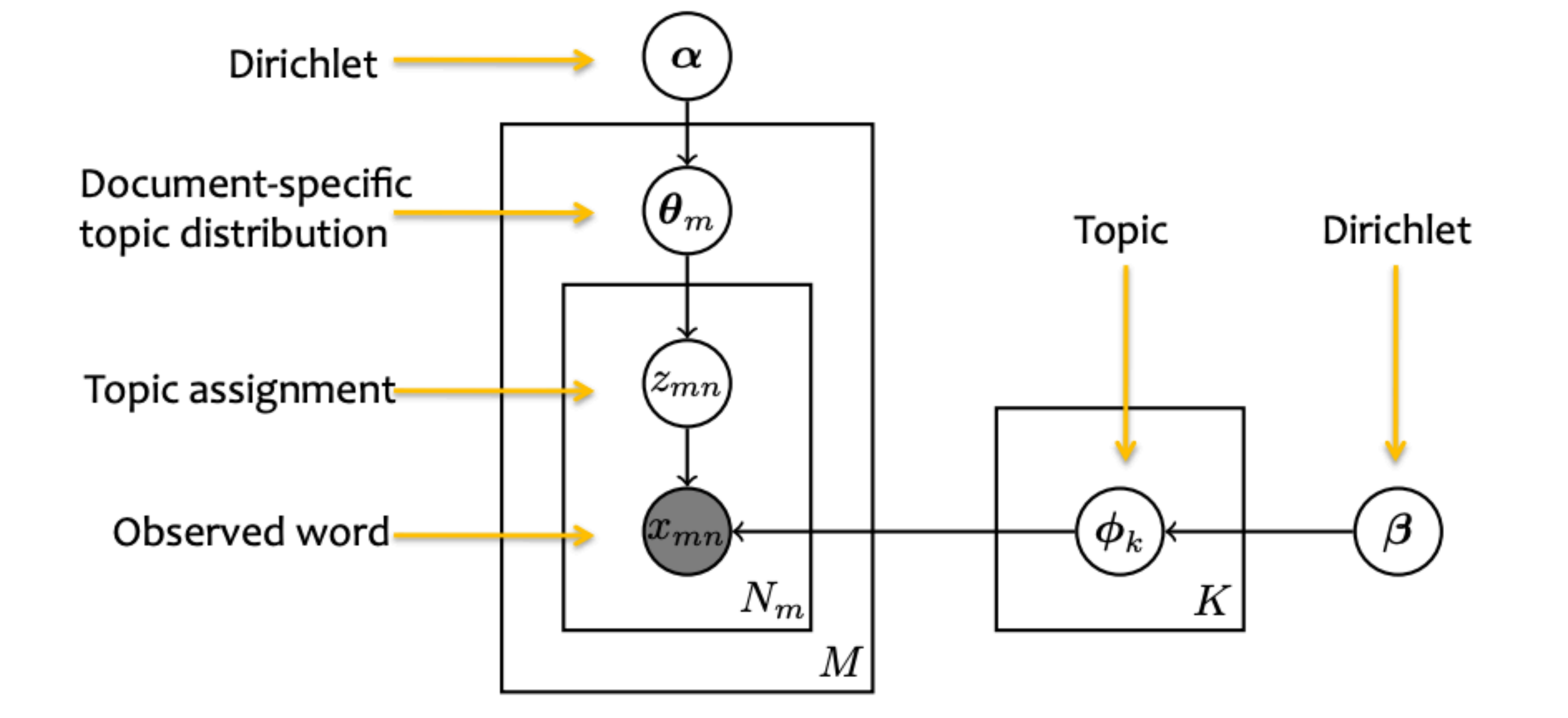
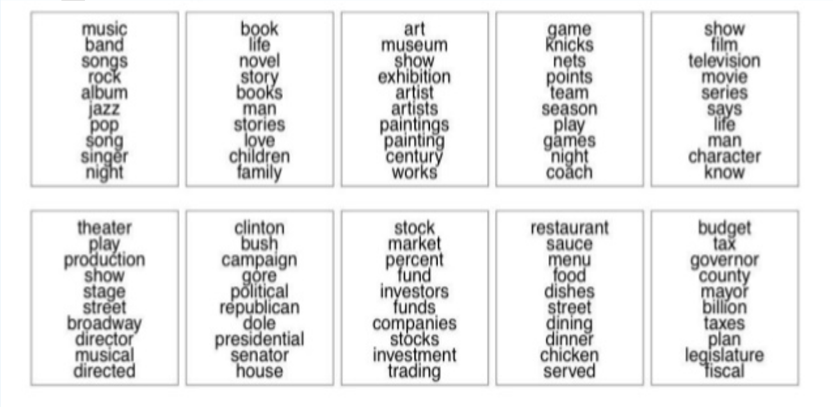


Figure from Matt Gormley

LDA Algorithm

Training LDA Model

1. Randomly assign words to topics
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens



music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

Training LDA Model

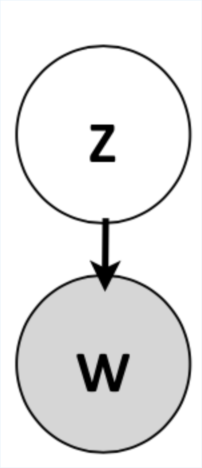
1. Randomly assign words to topics
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Randomly assign words to topics

Etruscan	trade	price	temple	market

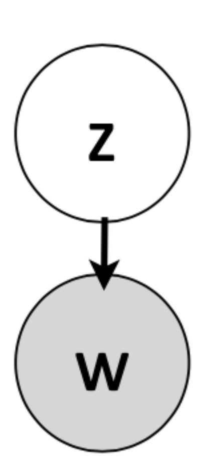
Example David Mimno

Randomly assign words to topics



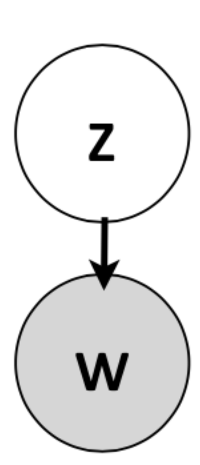
Etruscan	trade	price	temple	market

Randomly assign words to topics

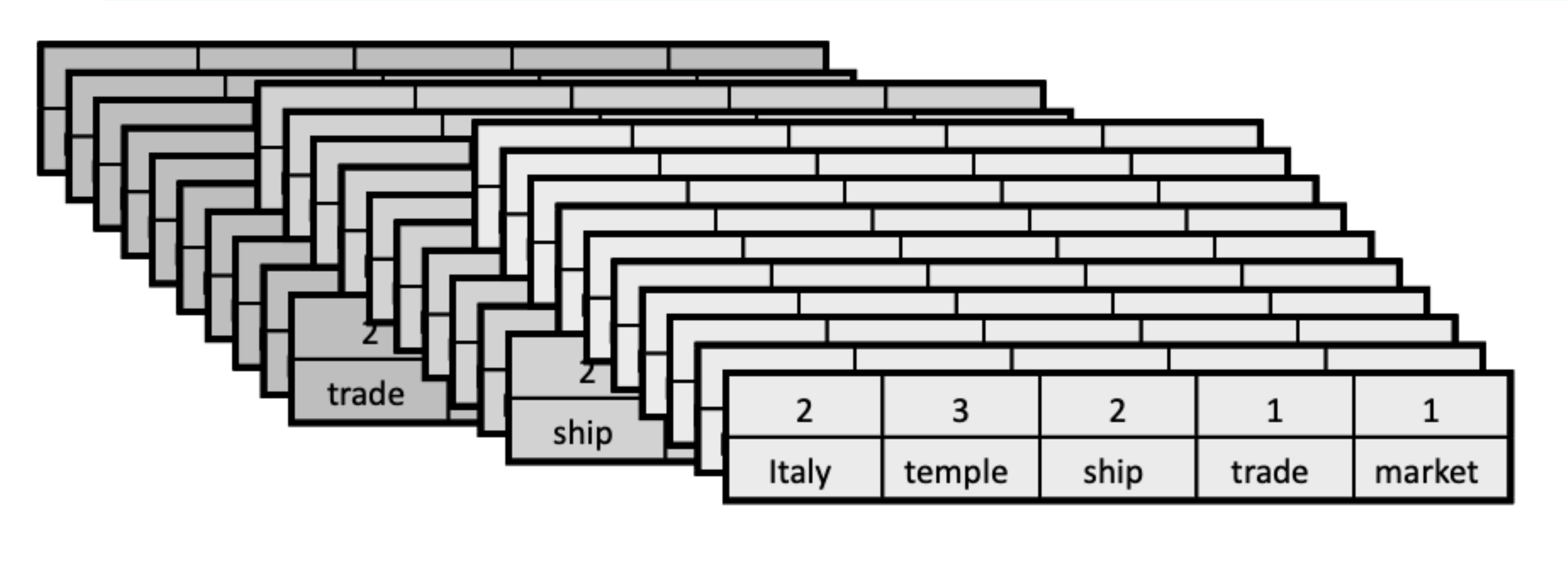


3	2	1	3	1
Etruscan	trade	price	temple	market

Randomly assign words to topics



3	2	1	3	1
Etruscan	trade	price	temple	market



2	3	2	1	1
Italy	temple	ship	trade	market

Global Statistics from Random Topic Assignments

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts across corpus

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

Training LDA Model

1. ~~Randomly assign words to topics~~
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Training LDA Model

1. ~~Randomly assign words to topics~~
2. Repeat many times:
    1. For each document:
        1. For each token, re-assign the topic based on:
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Reassign topic for “Trade”

3	2	1	3	1
Etruscan	trade	price	temple	market

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

Reassign topic for "Trade"

3	2	1	3	1
Etruscan	trade	price	temple	market

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

Reassign topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

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

Reassign topic for “Trade”

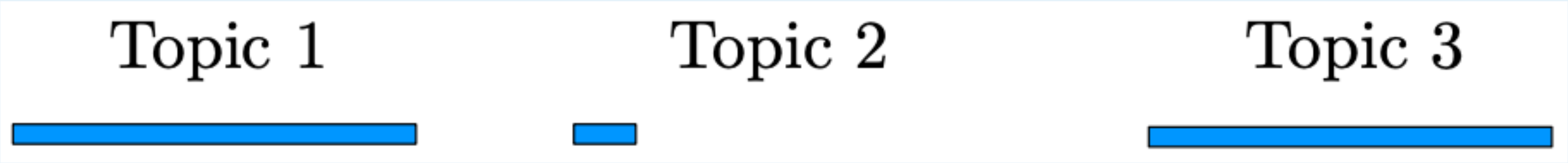
3	?	1	3	1
Etruscan	trade	price	temple	market

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

Pick a topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

Which topics occur in this document?



Pick a topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

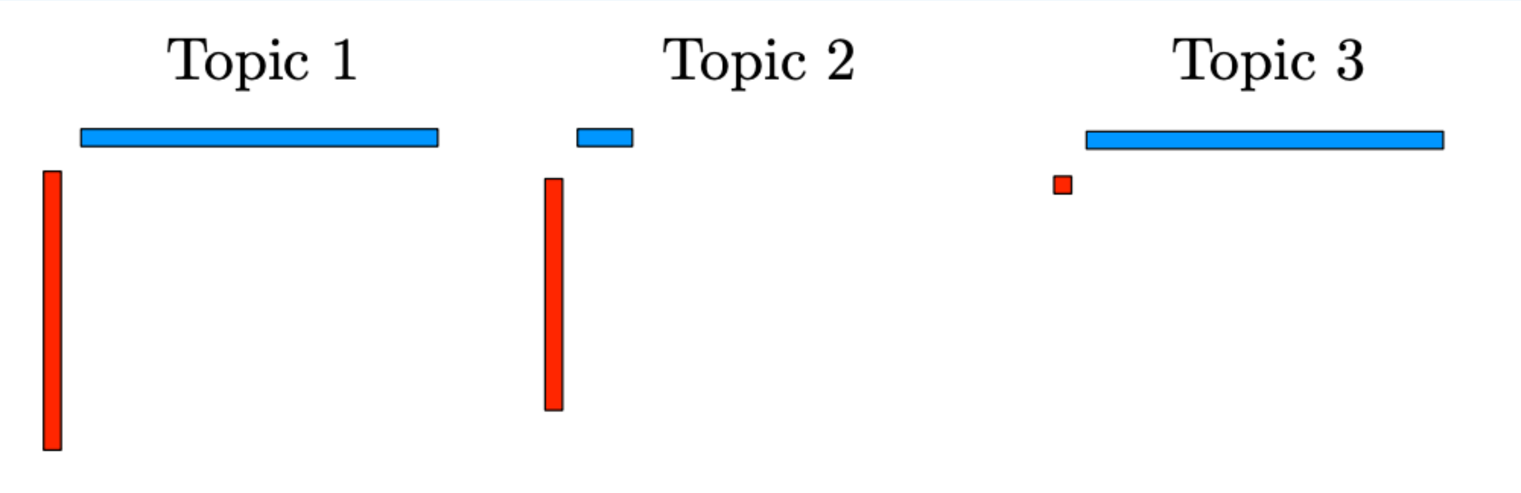
Which topics like the word-type “trade”?

	1	2	3
trade	10	7	1

Pick a topic for "Trade"

3	?	1	3	1
Etruscan	trade	price	temple	market

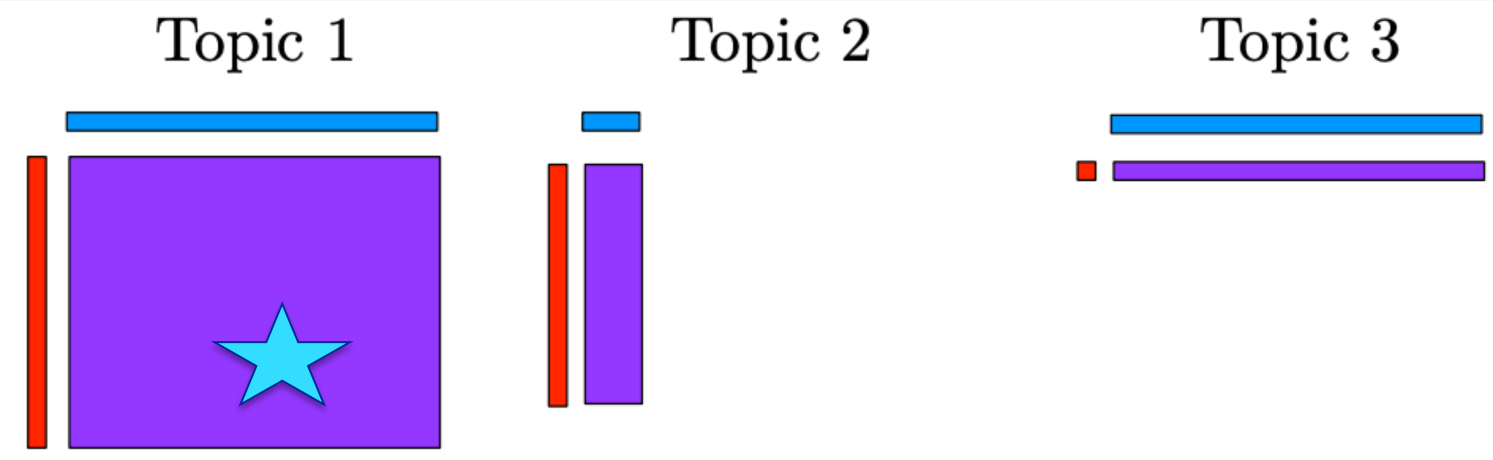
Which topics like the word "trade"?



Pick a topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

Pick a topic for “trade”?



Update topic for “Trade”

3	?	1	3	1
Etruscan	trade	price	temple	market

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

Update topic for “Trade”

3	1	1	3	1
Etruscan	trade	price	temple	market

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

Training LDA Model – Gibbs Sampling

1. **Randomly assign words to topics**
2. Repeat many times:
    1. For each document:
        1. **For each token, re-assign the topic based on:**
            1. Topic assignment for every other token in the document
            2. Topic assignment for every other instance of the type in the the corpus
3. Return: Topics assignments for all tokens

Modeling Decisions

Modeling decisions – hard choices

- Document definition

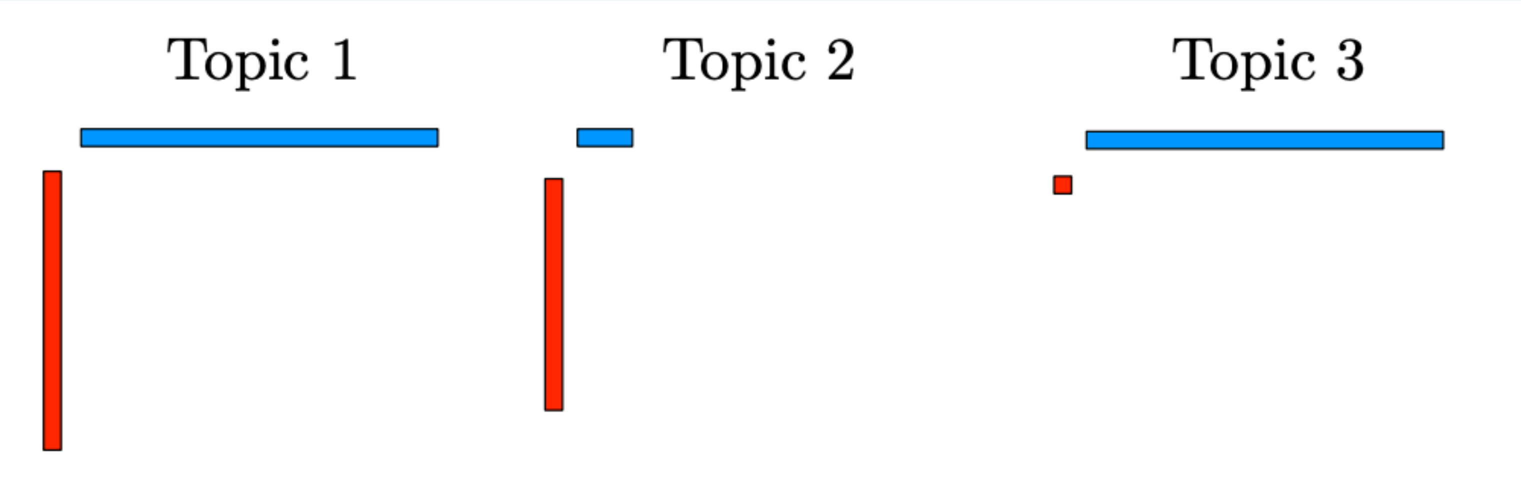
- Interesting words

- Knobs:
  - K - Number of topics
  - Hyper-parameters

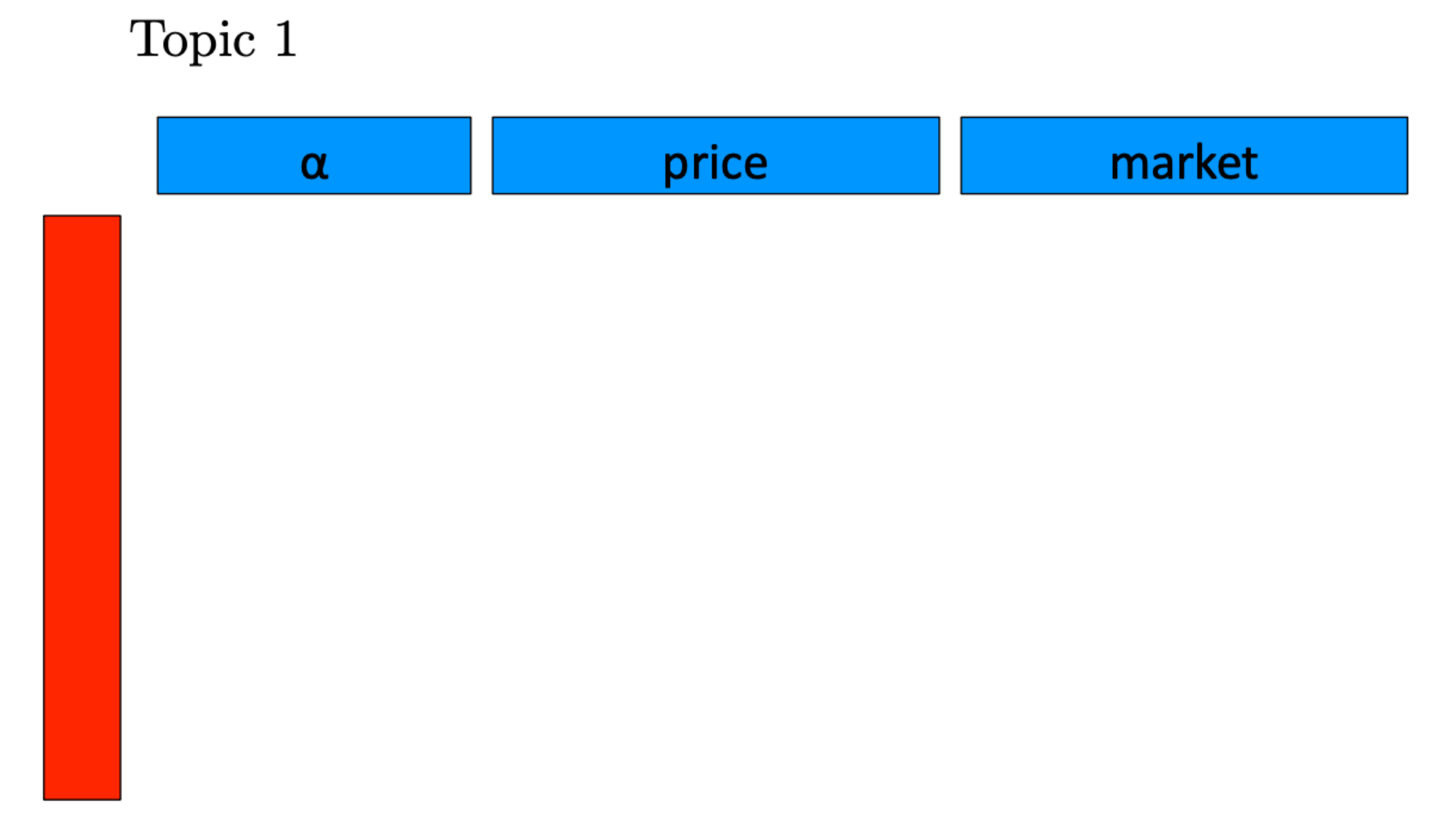
Hyperparemeters

3	?	1	3	1
Etruscan	trade	price	temple	market

Which topics like the word “trade”?



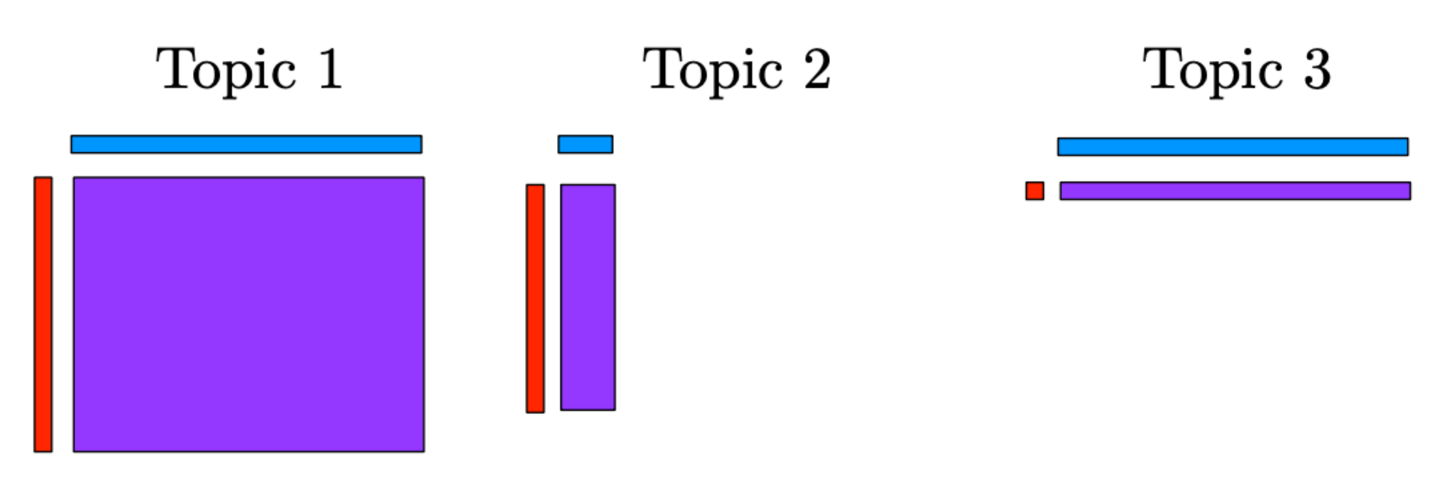
Hyperparameters - alpha



Hyperparemeters

3	?	1	3	1
Etruscan	trade	price	temple	market

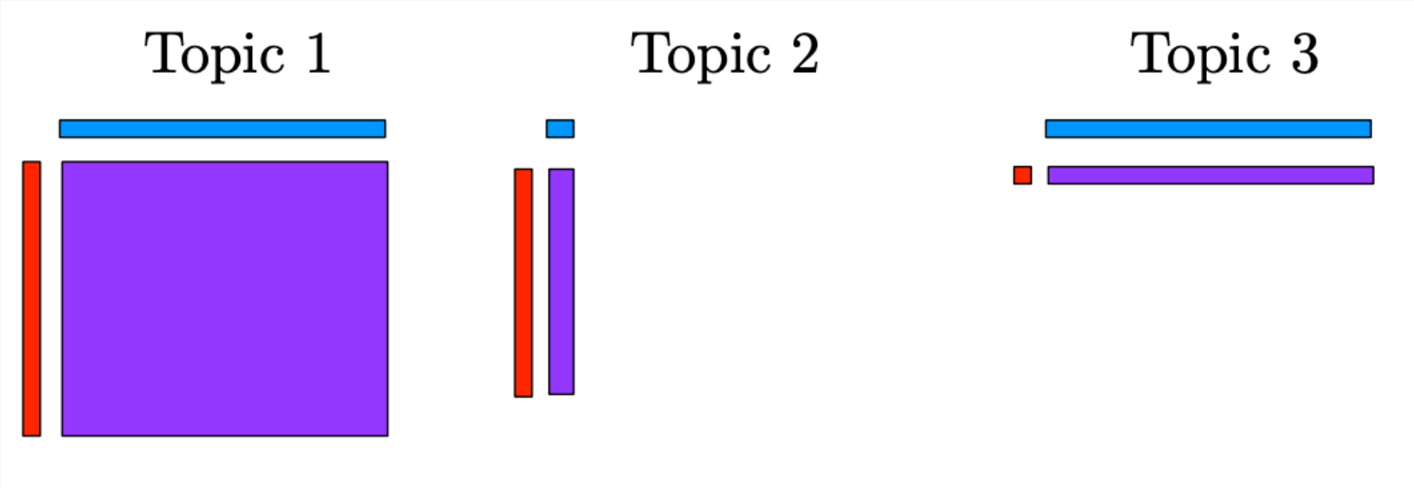
Which topics like the word “trade”?



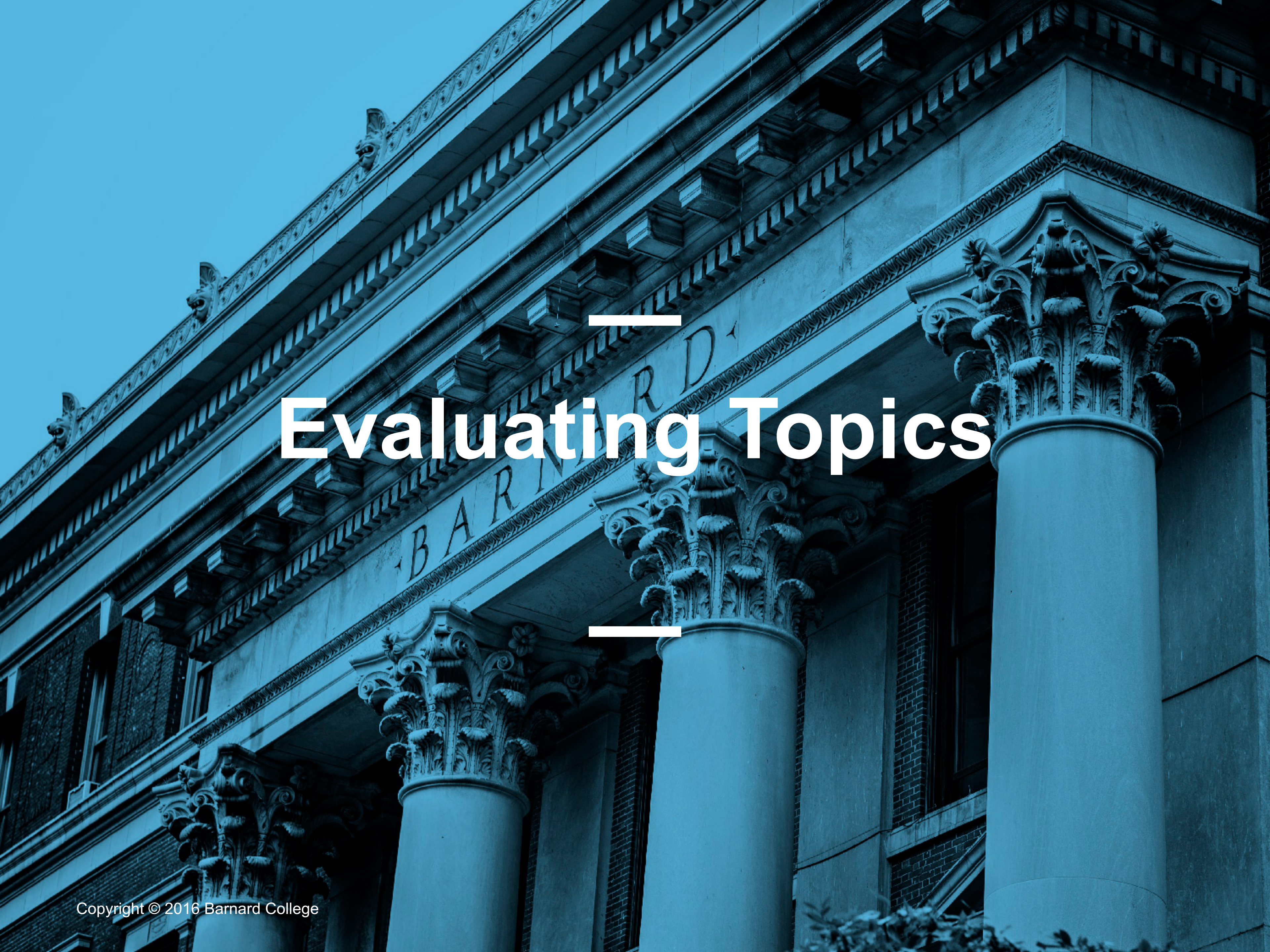
Hyperparemeters

3	?	1	3	1
Etruscan	trade	price	temple	market

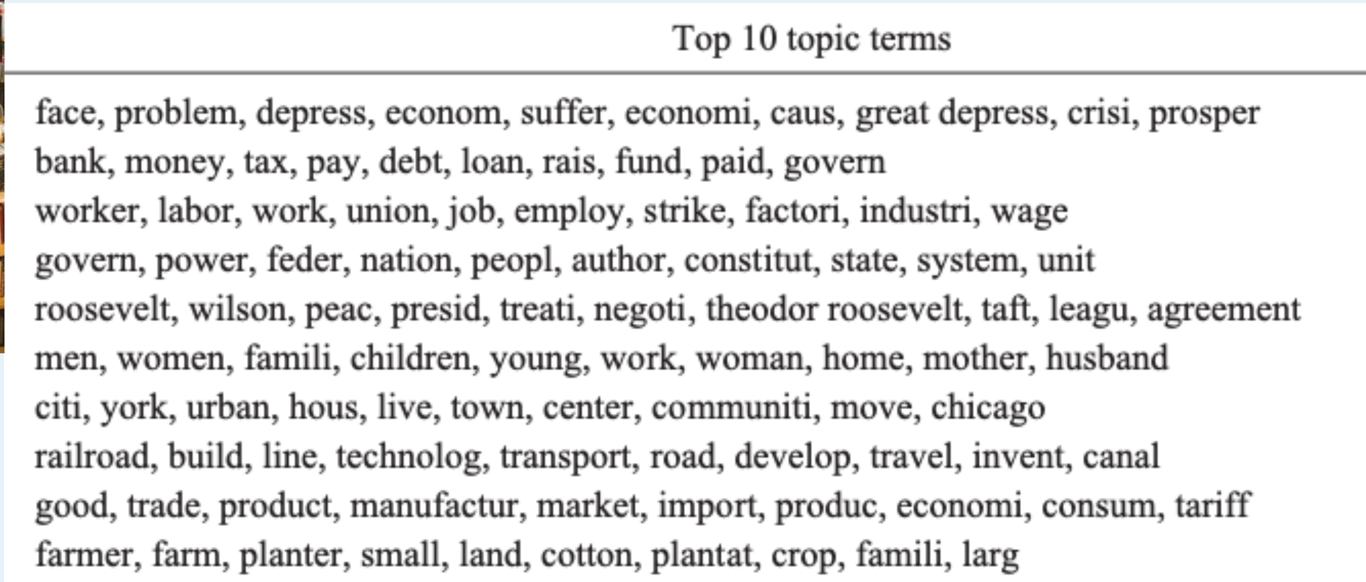
Which topics like the word “trade”?



Evaluating Topics



Output of topic models



Top 10 topic terms
face, problem, depress, econom, suffer, economi, caus, great depress, crisi, prosper
bank, money, tax, pay, debt, loan, rais, fund, paid, govern
worker, labor, work, union, job, employ, strike, factori, industri, wage
govern, power, feder, nation, peopl, author, constitut, state, system, unit
roosevelt, wilson, peac, presid, treati, negoti, theodor roosevelt, taft, leagu, agreement
men, women, famili, children, young, work, woman, home, mother, husband
citi, york, urban, hous, live, town, center, communiti, move, chicago
railroad, build, line, technolog, transport, road, develop, travel, invent, canal
good, trade, product, manufactur, market, import, produc, economi, consum, tariff
farmer, farm, planter, small, land, cotton, plantat, crop, famili, larg

What makes topics bad?

- **Random**, unrelated words
- ***Intruder*** words
- Boring, **overly general** words
- **Chimaeras:**
  - Multiple topics combined

Evaluation – Word Intrusion Task

- Take top k words in a topic
  - Usually 5 or 10
- Substitute 1 word with a top word from another topic
- Shuffle the works
- Ask someone to pick the intruder
  - If they can pick the intruder – it's a good topic