



JOHNS HOPKINS

WHITING SCHOOL  
*of* ENGINEERING

# Topic Modeling (LDA)

1/29/24

# Recap

---

- Last class:
  - Metrics to measure differences in word usage across subsets of corpora
    - Log Odds with Dirichlet Prior (Fightin' Words)
    - PMI Scores
- Today
  - Topic modeling (LDA)
  - Practical considerations and evaluation
  - Example application: Structured topic model and media manipulation



JOHNS HOPKINS

WHITING SCHOOL  
*of* ENGINEERING

**LDA**



# Odds ratio in Congressional data

Top Republican Words	Score	Top Democrat Words	Score
spending	-66.26	republican	56.63
obamacare	-59.90	wealthiest	40.78
government	-47.92	rhode	39.43
going	-45.33	women	38.16
that	-44.58	pollution	33.66
trillion	-43.43	republicans	32.86
taxes	-42.39	gun	32.45
you	-40.85	investments	32.22
administration	-39.07	families	31.93
debt	-38.92	violence	30.88

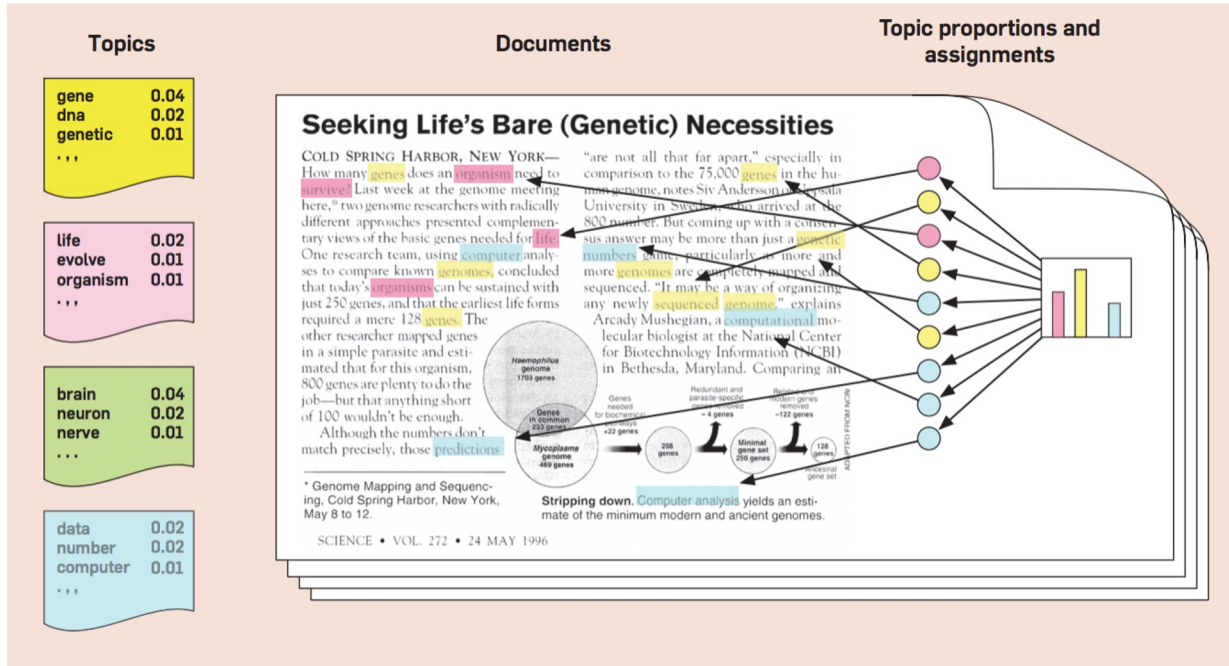
Probably all about budget and government spending

“Gun violence” is probably one topic

# Topic Modeling: Motivation

- Sometimes we care about specific words (more on this later)
- Often we want to group words into broader *topics*

# Latent Dirichlet Allocation



- Assume each document contains a mixture of "topics"
- Each topic uses mixtures of vocabulary words
- **Goal: recover topic and vocabulary distributions**

# Definitions

	Topic 1	Topic 2	...	Topic 30
administration	0.01	<b>0.12</b>	...	0.02
advertising	0.02	0.001	...	<b>0.25</b>
debt	0.1	0.001	...	0.01
...	...	...	...	...
government	0.01	<b>0.15</b>	...	0.01
...	...	...	...	...
spending	<b>0.12</b>	0.01	...	0.03
taxes	<b>0.15</b>	0.02	...	<b>0.35</b>
trillion	<b>0.19</b>	0.003	...	0.02

Each “topic” is defined by  $\phi$ , a multinomial distribution over the entire vocabulary

	Doc 1	Doc 2	...	Doc N
Topic 1	0.10	<b>0.60</b>	...	
Topic 3	0.02	0.05	...	
Topic 4	<b>0.30</b>	0.1	...	
...	...	...	...	...
Topic 15	<b>0.20</b>	0.01	...	<b>0.40</b>
...	...	...	...	...
Topic 28	0.01	0.03	...	<b>0.20</b>
Topic 29	<b>0.25</b>	0.15	...	
Topic 30	0.03	0.01	...	

Each document has associated  $\theta$ , a multinomial distribution over topics

# LDA Generative Story

Basic idea:

- Assume a story for generating our data (sampling from distributions)
- Estimate the parameters of the distribution
- [There are other approaches to topic modeling, this is specifically LDA]

Generative story for log-odds with a Dirichlet Prior:

1. Draw  $\boldsymbol{\pi}^{(i)} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. For  $n^{(i)}$  steps:
  1. Draw  $w \sim \text{Multinomial}(\boldsymbol{\pi}^{(i)})$



# LDA Generative Story

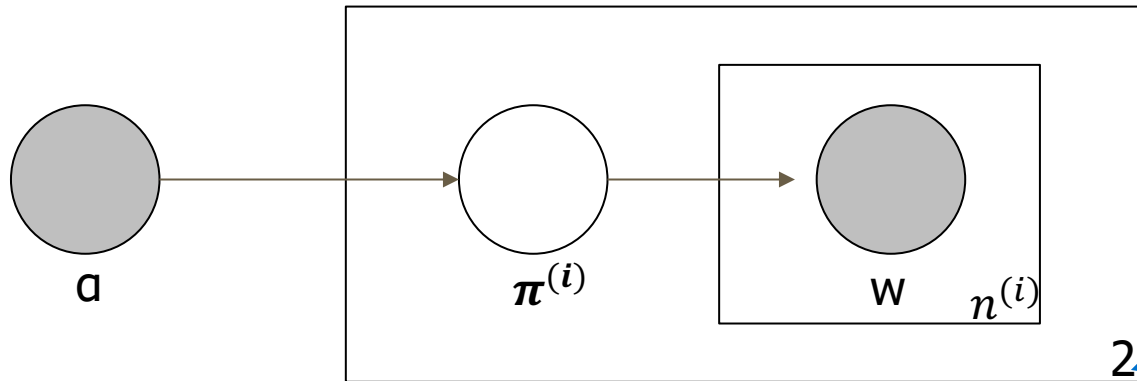
- For each topic  $k$ :
  - Draw  $\phi_k \sim \text{Dir}(\beta)$
- For each document  $d$ :
  - Draw  $\theta_d \sim \text{Dir}(\alpha)$
  - For each word in  $d$ :
    - Draw topic assignment  $z \sim \text{Multinomial}(\theta_d)$
    - Draw  $w \sim \text{Multinomial}(\phi_z)$

We use the data to estimate these two sets of parameters:

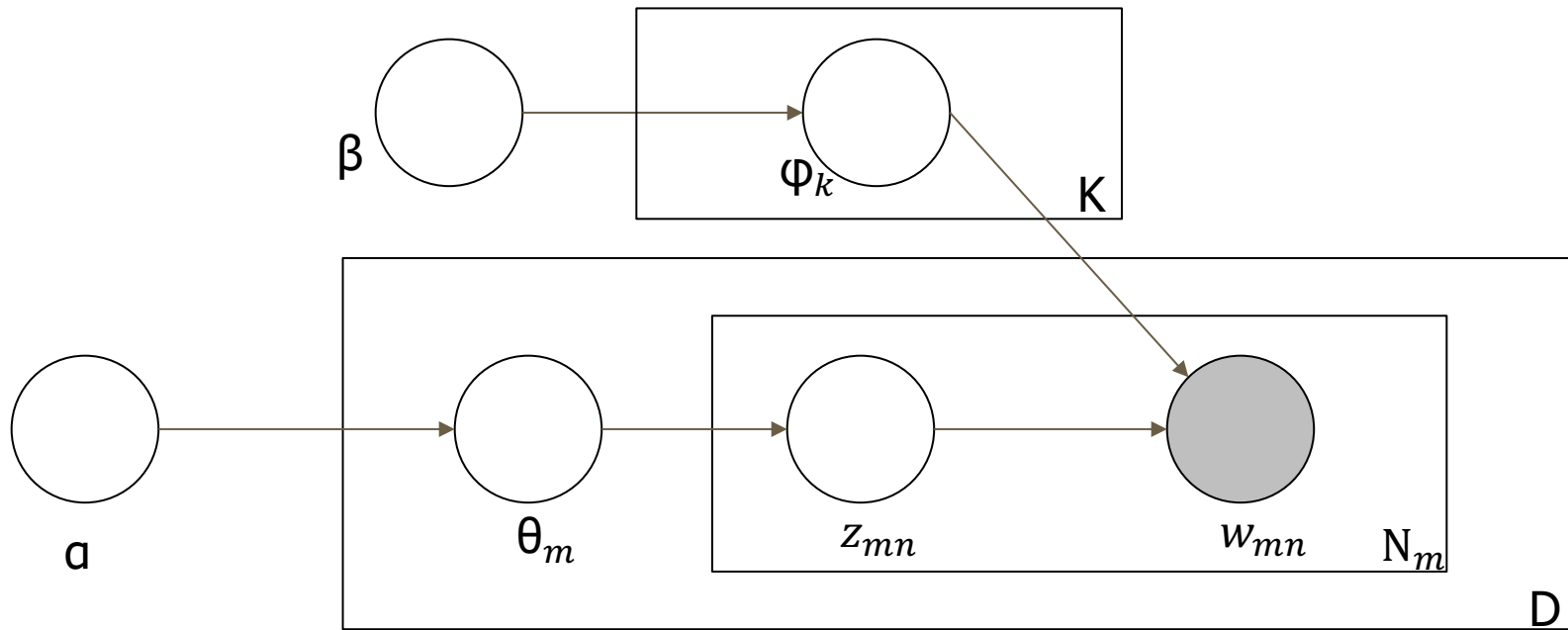
- $\phi$ , a distribution over vocabulary (1 for each topic)
- $\theta$ , a distribution over topics (1 for each document)

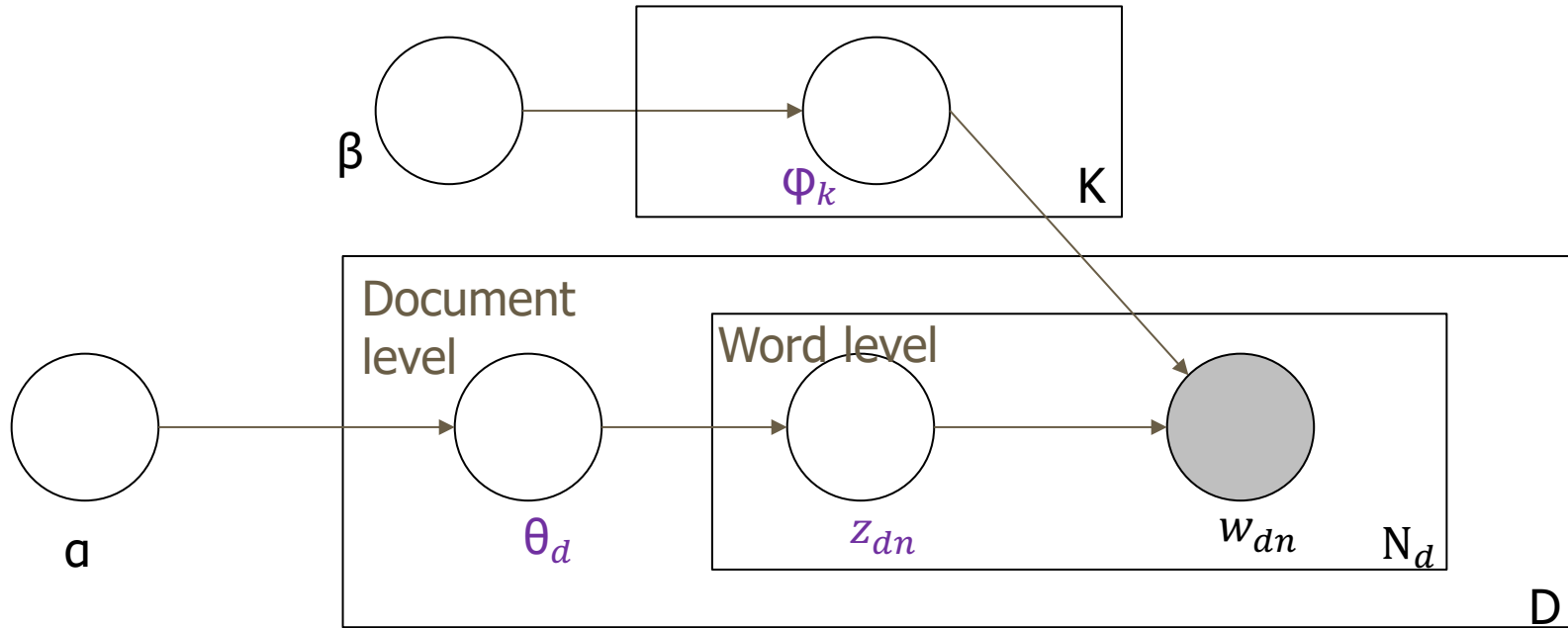
# Plate Notation: Log-odds with Dirichlet prior

- Shaded circle: value we observe
- Rectangles: values that are repeated (with number in corner reflecting # of repetitions)



We drew  $\pi$  for Democrats and  $\pi$  for Republicans





Variables we observe:  $D$  = number of documents;  $N$  = number of words per document,  $w$  words in document

Variables we want to estimate:  $\theta$ ,  $\phi$ ,  $z$  are latent variables

Variables we choose:  $\alpha$ ,  $\beta$  are hyperparameters.  $K$  = number of topics

# General Estimators [Heinrich, 2005]

Goal: estimate  $\theta, \phi$

$$p(\theta, \phi, z|w) = \frac{p(w|\theta, \phi, z)p(\theta, \phi, z)}{p(w)}$$

- MLE approach
  - Maximize likelihood:  $p(w | \theta, \phi, z)$
- MAP approach
  - Maximize posterior:  $p(\theta, \phi, z | w)$  OR  $p(w | \theta, \phi, z) p(\theta, \phi, z)$
- Bayesian approach
  - Approximate posterior:  $p(\theta, \phi, z | w)$
  - Take expectation of posterior to get point estimates

# LDA: Bayesian Inference

- Goal: estimate  $\theta, \phi$
- Bayesian approach: we estimate full posterior distribution

$$p(\theta, \phi, z|w) = \frac{p(\theta, \phi, z, w)}{p(w)}$$

$p(w)$  is the probability of your data set occurring under *any* parameters -- this is intractable!



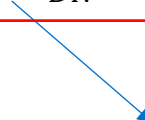
Solutions: Gibbs Sampling, Variational Inference

# Gibbs Sampling

Vastly available digitized text data has created new opportunities for understanding social phenomena. Relatedly, social issues like toxicity, discrimination, and propaganda frequently manifest in text, making text analyses critical for understanding and mitigating them. In this

- Assume we know topic assignments for all words in the corpus
- One word at a time, re-sample the topic assignment

# Gibbs Sampling

- Initialize  $z$  (e.g. randomly)
- for  $t = 1$  to  $T$  do:  For each iteration
  - for  $d = 1$  to  $D$ ; for  $n = 1$  to  $d$  do:  For each word in the corpus
    - $z_{dn}^{(t+1)} \sim P(Z_{dn} \mid z_{11}^{(t+1)}, \dots, z_{dn-1}^{(t+1)}, z_{dn+1}^{(t)}, \dots, z_{DN}^{(t)})$  Sample a new topic assignment
  - end for
- end for



# Gibbs Sampling

$$P(z_{dn} = k | z_{d,-n}, w, \alpha, \beta, \phi, \theta)$$

- We integrate out  $\phi, \theta$  (we can do this because of conjugacy)

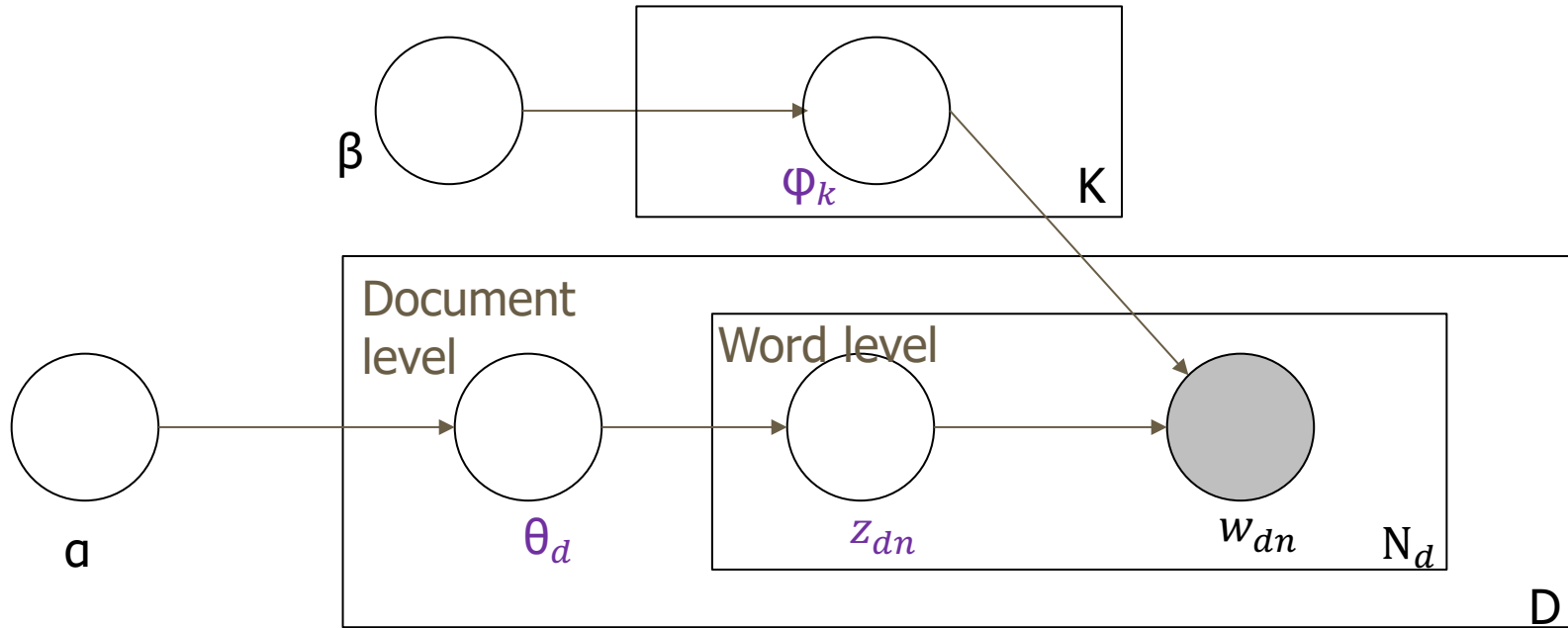
Number of times document  $d$  uses topic  $k$

From prior

Number of times topic  $k$  uses word  $w_{dn}$

From prior

$$P(z_{dn} = k | z_{d,-n}, w, \alpha, \beta) = \frac{c_{dk} + \alpha_k}{\sum_i^K c_{di} + \alpha_i} \frac{v_{kw_{dn}} + \beta_{w_{dn}}}{\sum_i^K v_{ki} + \beta_i}$$



Variables we observe:  $D$  = number of documents;  $N$  = number of words per document,  $w$  words in document

Variables we want to estimate:  $\theta$ ,  $\phi$ ,  $z$  are latent variables

Variables we choose:  $\alpha$ ,  $\beta$  are hyperparameters.  $K$  = number of topics

# Gibbs Sampling

$$P(z_{dn} = k | z_{d,-n}, w, \alpha, \beta, \varphi, \theta)$$

- We integrate out  $\varphi, \theta$  (we can do this because of conjugacy)

$$P(z_{dn} = k | z_{d,-n}, w, \alpha, \beta) = \frac{c_{dk} + \alpha_k}{\sum_i^K c_{di} + \alpha_i} \frac{v_{kw_{dn}} + \beta_{w_{dn}}}{\sum_i^K v_{ki} + \beta_i}$$

Prevalence of topic in document      Prevalence of word in topic

# Variational Inference

- Compared to Gibbs Sampling:
  - Deterministic, easy to determine convergence, requires fewer iterations
  - Doesn't require conjugacy
  - Math is more difficult
- Key ideas:
  - Pick a family of distributions ( $q$ ) over the latent variables with its own *variational parameters*
  - Find the setting of the parameters that makes  $q$  close to the posterior of interest
  - Use  $q$  with the fitted parameters as a proxy for the posterior

Xanda Schofield and Jordan Boyd-Graber

<https://www.youtube.com/watch?v=-tKmyHoVZ-g>

<https://www.youtube.com/watch?v=smfWKhdcaoA>

David Blei Lecture Notes

<https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf>

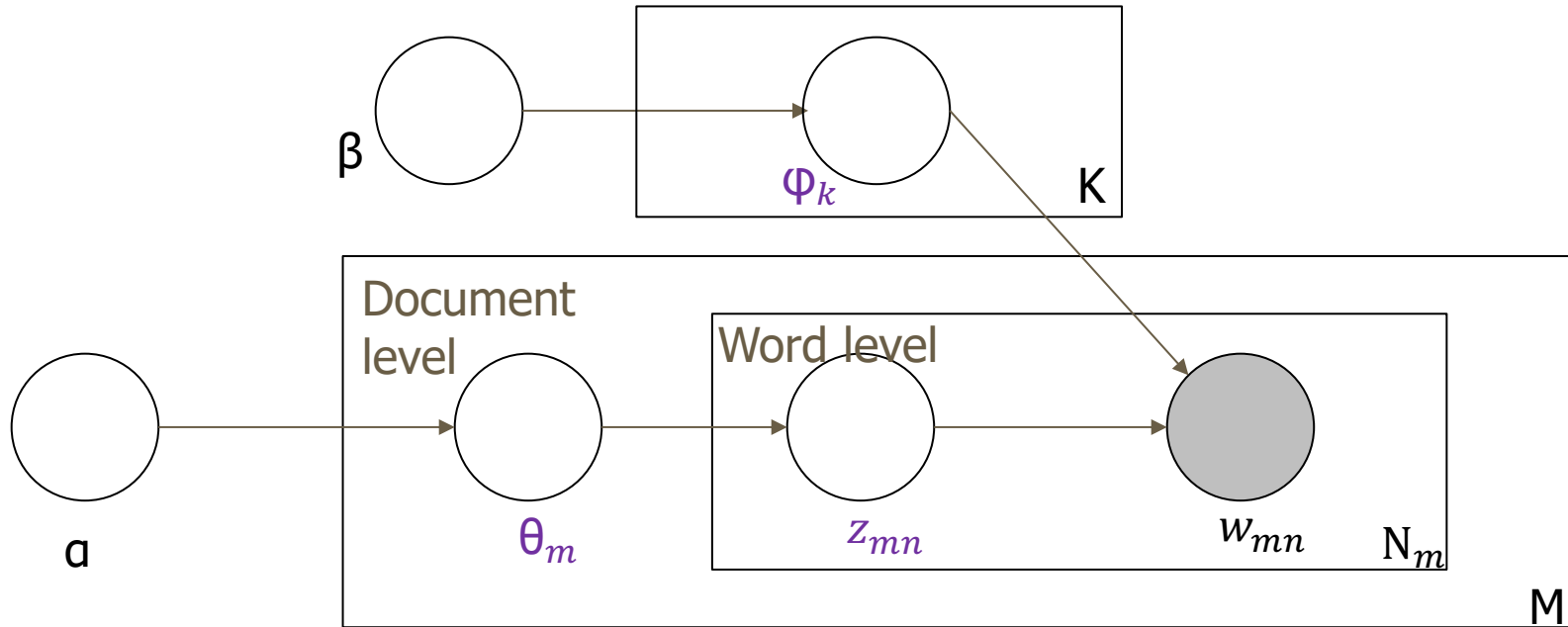


JOHNS HOPKINS

WHITING SCHOOL  
*of* ENGINEERING

# Practical considerations and evaluation





Variables we observe:  $M$  = number of documents;  $N$  = number of words per document,  $w$  words in document

Variables we want to estimate:  $\theta$ ,  $\phi$ ,  $z$  are latent variables

Variables we choose:  $\alpha$ ,  $\beta$  are hyperparameters.  $K$  = number of topics

# Choosing $\alpha$ , $\beta$ and $K$

- In practice, typically choose *symmetric* Dirichlet priors, e.g.  $\alpha, \beta = [1, 1, 1, 1, \dots], [0.1, 0.1, 0.1, 0.1, \dots]$  but some research has explored alternatives
- In practice, try a few  $K$  values and judge if topics look reasonable, but there are approaches that estimate the best value

# How do we describe a topic?

---

- Most probable words for each topic
- Words common in this topic *relative* to other topics
- Examining documents that contain high proportion of topic



# Sample Topics from NYT Corpus

#5	#6	#7	#8	#9	#10
10	0	he	court	had	sunday
30	tax	his	law	quarter	saturday
11	year	mr	case	points	friday
12	reports	said	federal	first	van
15	million	him	judge	second	weekend
13	credit	who	mr	year	gallery
14	taxes	had	lawyer	were	iowa
20	income	has	commission	last	duke
sept	included	when	legal	third	fair
16	500	not	lawyers	won	show

# LDA: Evaluation

---

- Held out likelihood
  - Hold out some subset of your corpus
  - Says NOTHING about coherence of topics
- Intruder Detection Tasks [Chang et al. 2009]
  - Give annotators 5 words that are probable under topic A and 1 word that is probable under topic B
  - If topics are coherent, annotators should easily be able to identify the intruder

# LDA: Advantages and Drawbacks

---

- When to use it
  - Initial investigation into unknown corpus
  - Concise description of corpus (dimensionality reduction)
  - [Features in downstream task]
- Limitations
  - Can't apply to specific questions (completely unsupervised)
  - Simplified word representations
    - BOW model
    - Can't take advantage of similar word
  - Strict assumptions
    - E.g. Independence assumptions

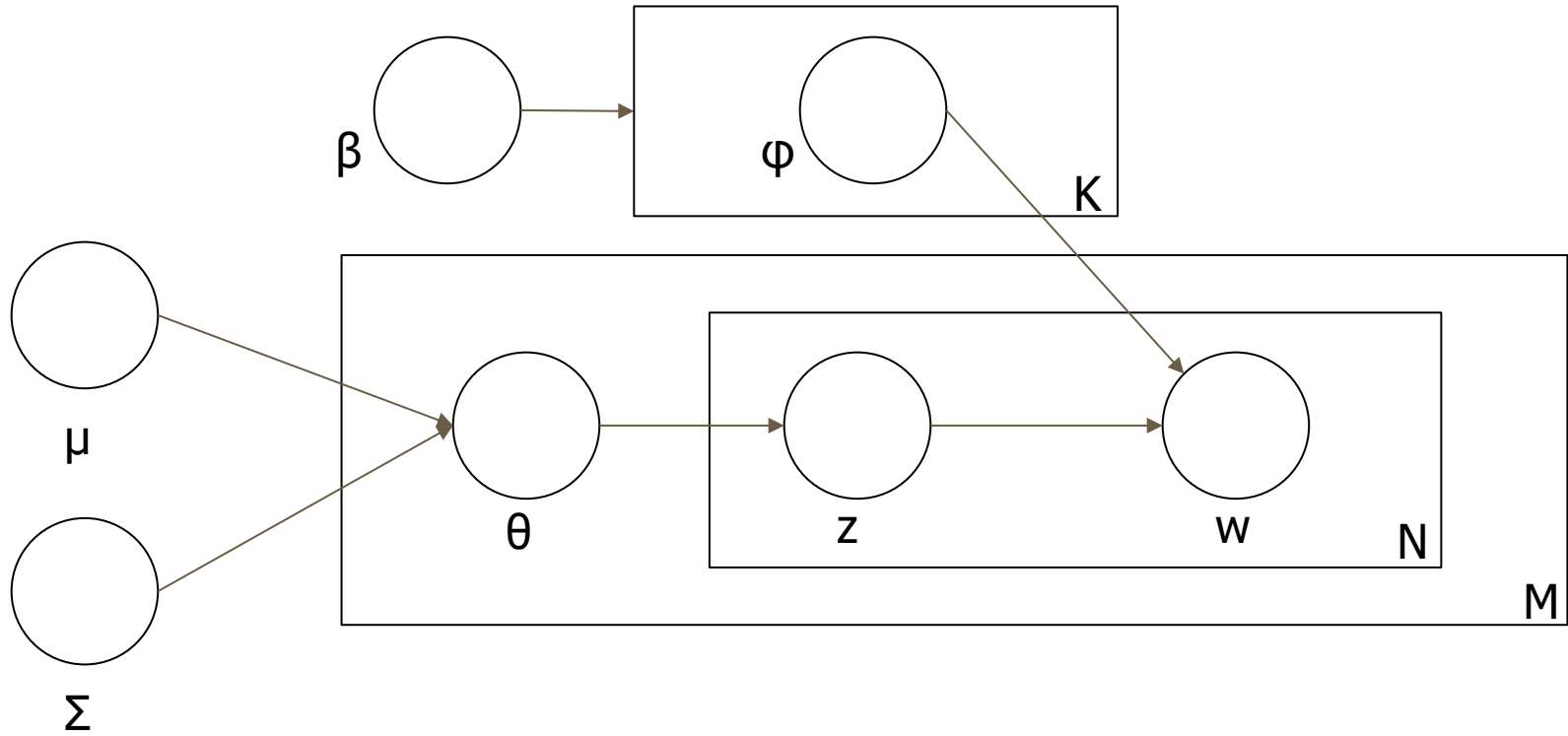
# Problem 1: Topic Correlations

---

- LDA
  - In a vector drawn from a Dirichlet distribution ( $\theta$ ), elements are nearly independent
- Reality
  - A document about biology is more likely to also be about chemistry than skateboarding

# Solution to Problem 1: Correlated Topic Model [Blei and Lafferty, 2006]

- For each topic  $k$ :
  - Draw  $\phi_k \sim \text{Dir}(\beta)$
- For each document  $D$ :
  - Draw  ~~$\theta_D \sim \text{Dir}(\alpha)$~~  Draw  $\eta_D \sim N(\mu, \Sigma); \theta_D = f(\eta_D)$   $\Sigma = \text{Topic covariance matrix}$
  - For each word in  $D$ :
    - Draw topic assignment  $z \sim \text{Multinomial}(\theta_D)$
    - Draw  $w \sim \text{Multinomial}(\phi_z)$
- $\phi$  is a distribution over your vocabulary (1 for each topic)
- $\theta$  is a distribution over topics (1 for each document)



# Short Break

---





JOHNS HOPKINS

WHITING SCHOOL  
*of* ENGINEERING

# Example application: Structured topic model and media manipulation



# Motivating application: Communications theory of media manipulation

- Communications scholarship on media influence:
  - “the media may not be successful much of the time in telling people what to think, but is stunningly successful in telling its readers what to think about” [Cohen, 1963]
  - Given a corpus of newspaper articles, we can determine how it may be influencing public opinion by analyzing changes in topic coverage
    - We don’t know exactly what topics are in advance: we need to be able to discover them from the corpus

# Motivating application: Communications theory of media manipulation

- Agenda setting
  - **What** topics are covered
- Framing
  - **How** topics are covered
- Priming
  - What effect reporting has on public opinion
  - “Framing works to shape and alter audience members’ interpretations and preferences through priming”

Entman’s thesis: we can use this framework to understand bias in the media

**“agenda setting, framing and priming fit together as *tools of power*”**

# Motivating application: Communications theory of media manipulation

---

- “process of culling a few elements of perceived reality and assembling a narrative that highlights connections among them to promote a particular interpretation” [Entman, 2007]
- Topic Level
  - Abortion is a moral issue
  - Abortion is health issue
  - [This should remind you agenda setting]
- Word Level
  - “Estate tax” vs. “Death tax”

# Framing: Additional Background

- Equivalence: different presentations of logically-identical information (Scheufele and Iyengar, 2012)
- *Emphasis*: “qualitatively different yet potentially relevant considerations” (Chong and Druckman, 2007, p.114)

	Issue-Specific	Issue-generic
Equivalence	90% unemployment; 10% employment	90%; 10%
Emphasis	Immigration: hero/worker vs. threat/job security	Morality, Economy, Security and Defense

Mendelsohn, Julia, Ceren Budak, and David Jurgens. "Modeling Framing in Immigration Discourse on Social Media." NAACL. 2021.

Card, Dallas, et al. "The media frames corpus: Annotations of frames across issues." ACL. 2015.

# Framing: Additional Background

Topic Model?

General taxonomy;  
classification models

	Issue-Specific	Issue-generic
Equivalence	90% unemployment vs. 10% employment	90% vs. 10%
Emphasis	Immigration: hero/worker vs. threat/job security	Morality, Economy, Security and Defense

Media Frames Corpus – issue-generic policy frames

Mendelsohn, Julia, Ceren Budak, and David Jurgens. "Modeling Framing in Immigration Discourse on Social Media." NAACL. 2021.

Card, Dallas, et al. "The media frames corpus: Annotations of frames across issues." ACL. 2015.

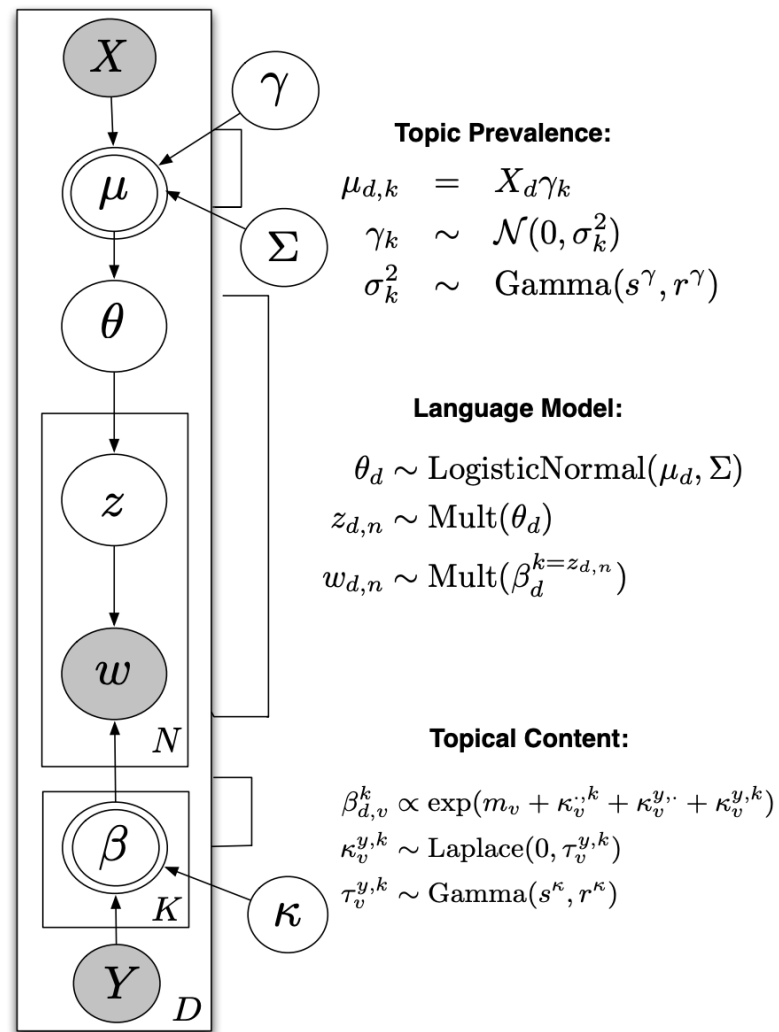
# Problem: LDA assumptions conflict with analysis goals

- LDA
  - The topic distributions ( $\theta$ ) are drawn from the same distribution  $\text{Dir}(\alpha)$  for all documents
- Reality
  - We often use LDA to look at how topics vary across documents
  - Example
    - We run LDA on a corpus of Democratic/Republican speeches.
    - Look at topic prevalence in Republican speeches and Democratic speeches
    - Conclude Republicans talk about taxes more than Democrats
  - But we've assumed that all speeches are drawing topics the same way
  - We need something other than LDA

# Solution: Structured Topic Model

- Topical prevalence: the proportion of document devoted to a given topic
  - $X$  - matrix of covariate information
  - Useful for *agenda setting*
- Topical content: the rate of word use within a given topic
  - $Y$  - matrix of covariate information
  - Useful for *framing*

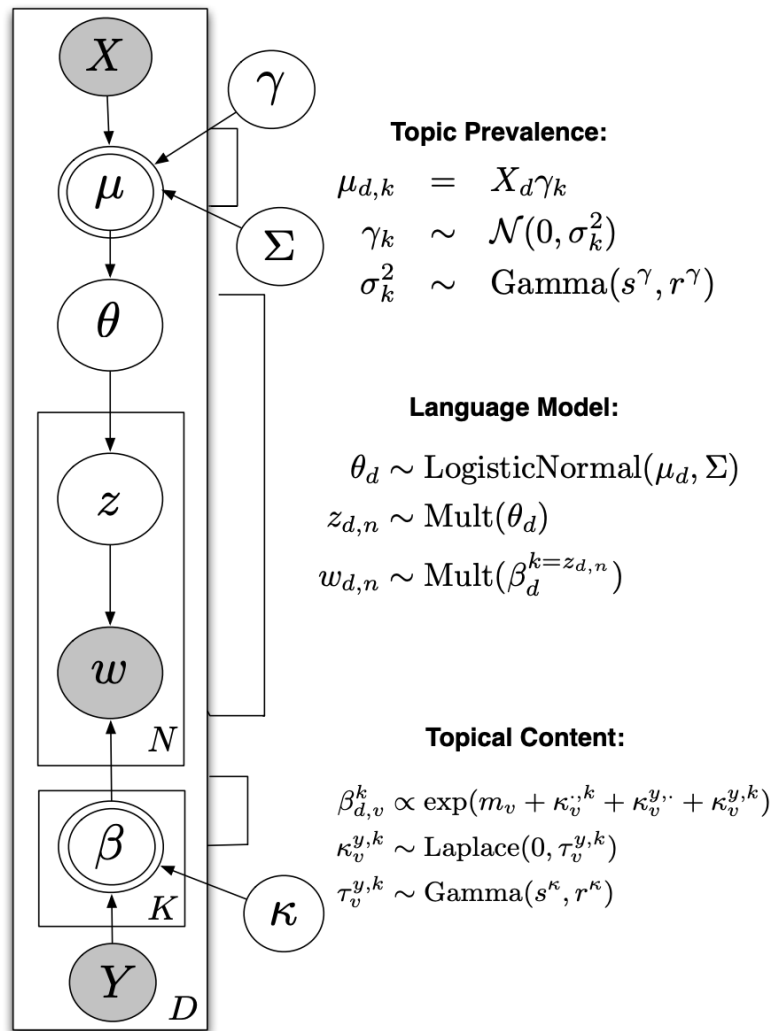
Roberts, Margaret E., et al. "The structural topic model and applied social science." *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. Vol. 4. No. 1. 2013.



# Solution: Structured Topic Model

- X could be Democrat/Republican as well as date of speech
  - Captures that Republicans talk more about *taxes* but rate varies by year
- Y could be Democrat/Republican
  - Captures that Democrats focus on social benefits and Republicans focus on government imposition

Roberts, Margaret E., et al. "The structural topic model and applied social science." *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. Vol. 4. No. 1. 2013.





# STM Example

21-year corpus on media coverage of grey wolf recovery in France

Nice-Matin = local newspaper

Le Monde = national newspaper

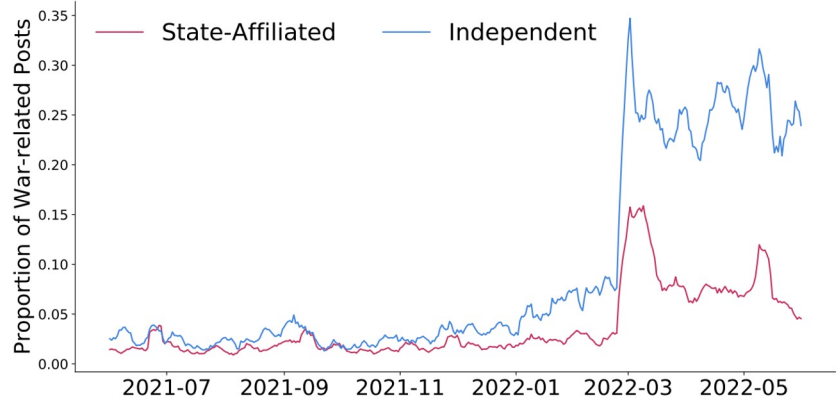
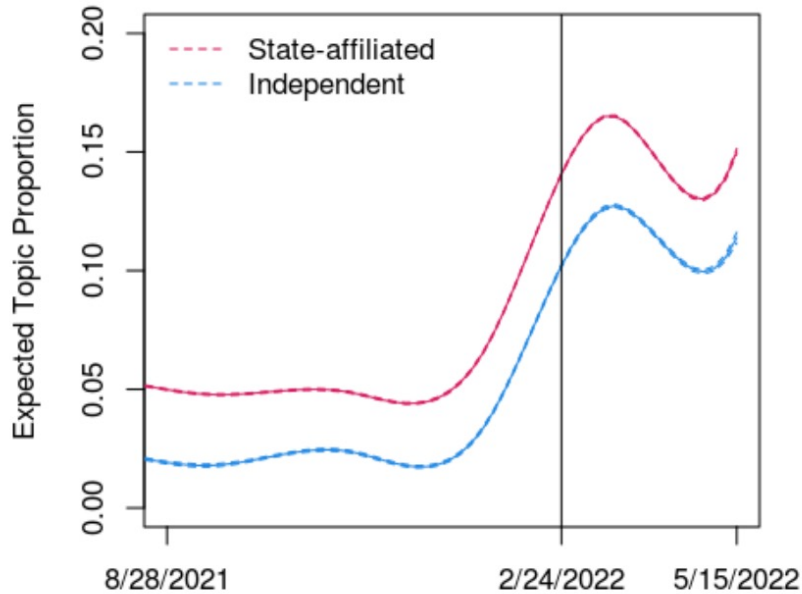
Topic 6: “Lethal Regulation”



<https://www.structuraltopicmodel.com/>

[Chandelier et al. 2018]

# STM topic with the highest probability of Ukraine and military related



# stm: R Package for Structural Topic Models

Margaret E. Roberts    Brandon M. Stewart    Dustin Tingley  
UCSD                                  Princeton                                  Harvard

- Extremely popular go-to tool for computational social science (Cited 1000+ times)
- Flexible inclusion of covariates
- Tools for visualizing topic outputs
  - E.g. expected proportions, selecting example documents for each topic, representing topics with top words
- [Implemented in R package]

# Today's takeaways

---

- Motivation behind topic modeling
- High-level understanding of LDA formulation and inference
- Key strengths and limitations of LDA
  - When later variants of LDA might be more useful
- Agenda setting and framing

Next class:

- Word Embeddings

# References

---

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
2. Roberts, Margaret E., et al. "The structural topic model and applied social science." *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. Vol. 4. No. 1. 2013.

Optional sources for more depth:

- Gibbs Sampling:
  - Jordan Boyd-Graber's Introduction:  
<https://www.youtube.com/watch?v=u7l5hhmdc0M>
  - <https://api.drum.lib.umd.edu/server/api/core/bitstreams/a36ce44d-0732-427d-8a81-a18c9b0b4dfa/content>
- [See previous slide for variational inference]