

# Topic Models

HSS 510: NLP for HSS

Taegyeon Kim

May 8, 2024

# Agenda

## Things to be covered

- What is topic modeling
- Latent Dirichlet Allocation
- Extensions
  - Correlated Topic Model
  - Dynamic Topic Model
  - Structural Topic Model
- Model selection and interpretation
- Summary

# What is Topic Model

Multi-membership, unsupervised algorithms to discover topics

- Offers an automated method to discover topics in a corpus of documents
  - Used to understand and organize large collections of documents according to the discovered topics
  - Either for exploration or for measurement
- Documents can contain multiple topics ( $\iff$  clustering algorithms)
- Unsupervised learning (no manual labeling of topics)

# What is Topic Model

Multi-membership, unsupervised algorithms to discover topics

- Topic: what is being talked about/written about
  - Unclear what topic actually means in theoretical terms
  - Topic models assume an intuitive and abstract notion of a topic
- *“The meaning of a topic in an LDA topic model must be assessed empirically instead and defined against the background of substantive theoretical concepts, such as political issues or frames (Maier et al. 2018)*
  - See the labeling of topics in U.S. legislators’ tweets at the end of the lecture

# What is Topic Model

## LDA and its extensions

- Latent Dirichlet Allocation (LDA) is one fundamental approach ([Blei et al., 2003](#))
- Alternative approaches
  - Correlated Topic Model (CTM)
  - Structural Topic Model (STM)
  - Dynamic Topic Model (DTM)
  - Clustering of transformer-based embeddings (next lecture)

# Topic Model in Context

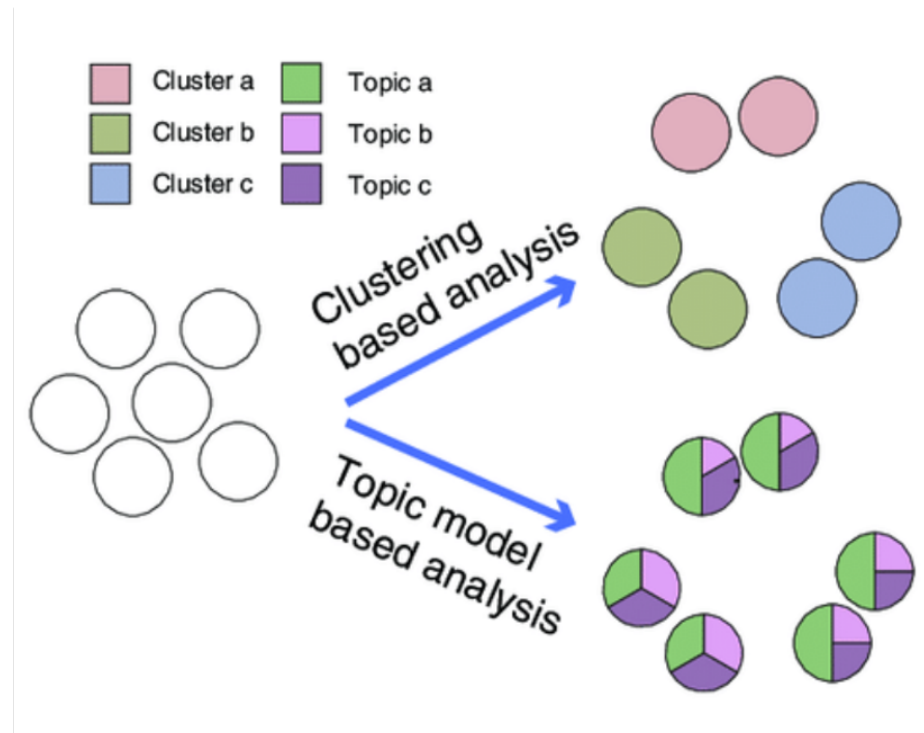
## Supervised vs. Unsupervised

	Supervised	Unsupervised
Objective	Trained on a labeled data to learn a mapping from input to output	Find patterns or structures within data without labels
Outcome	Pre-defined categories	Not quite pre-defined
Model evaluation	Explicit metrics such as accuracy, precision, recall, or MSE	Can involve qualitative assessment
Examples	Classification/regression for texts	<b>Topic models</b>

# Topic Model in Context

## Clustering algorithms vs. Topic models

- Clustering (e.g., K-means) assumes that each document belongs to one cluster
- Documents can have more than one idea in them (e.g., political speeches, newspaper articles, novels)
- [Click for figure source](#)



# Latent Dirichlet Allocation (LDA)

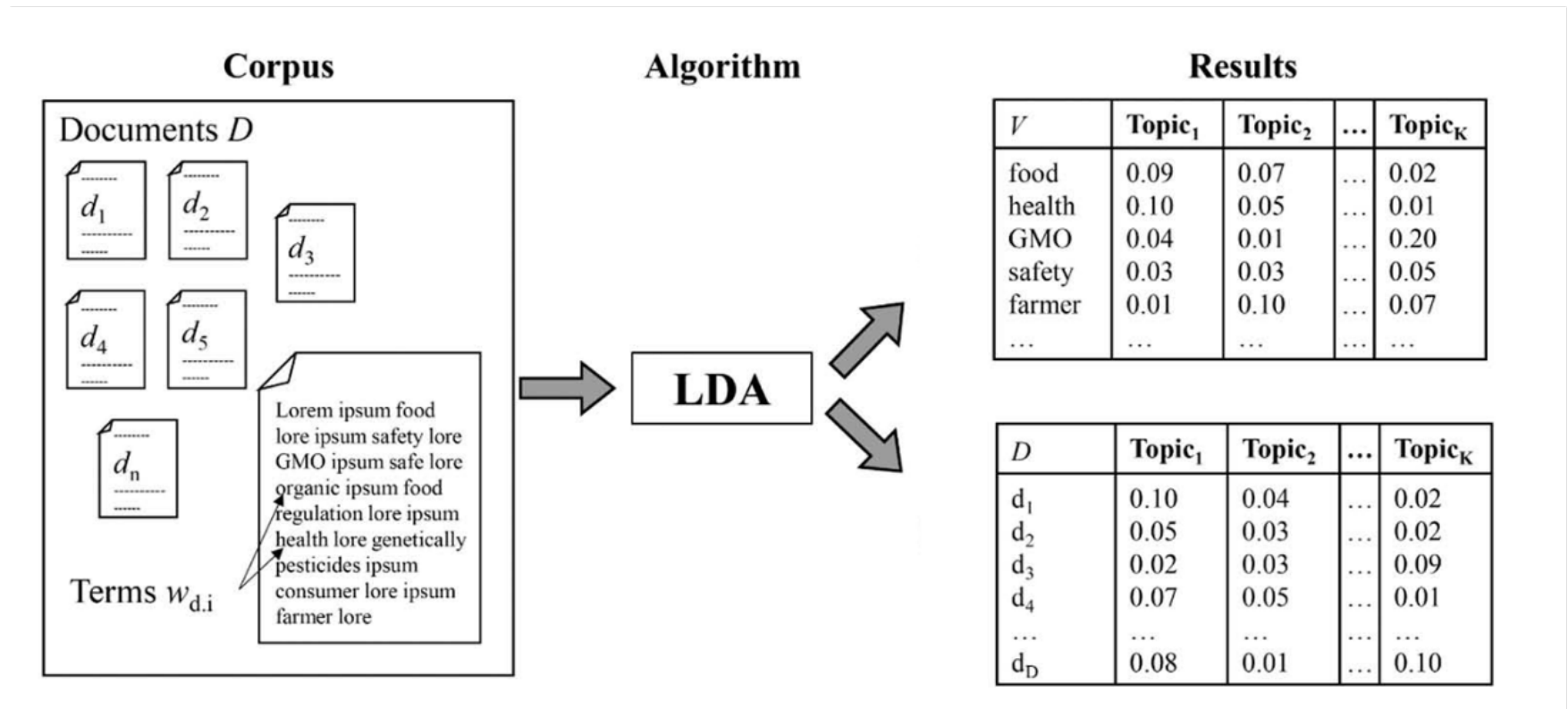
Used to identify the latent topic structure within a corpus

- One of many statistical algorithms for topic modeling
- Most seminar and widely used model
- Estimated via Bayesian frameworks
- “Documents” are seen as “distributions over topics”
- “Topics” are seen as “distributions over words”
- Make use of the BoW (Bag of Words) assumption



# Latent Dirichlet Allocation (LDA)

The framework of LDA (adaped from [Maier et al. 2018](#))



# LDA: Key Distributions

## Multinomial distribution

- Generalization of binomial distribution (e.g., flipping a coin)
- Probabilities of different outcomes (not just two)
- E.g., rolling a (6-sided) dice
  - Each side (from 1 to 6): discrete outcome
  - The probabilities sum up to 1 ( $\frac{1}{6} + \dots + \frac{1}{6} = 1$ )
- E.g., allocating a pie to one of 3 people
  - Each person: discrete outcome
  - The probabilities sum up to 1 ( $\frac{1}{4} + \frac{1}{4} + \frac{1}{2} = 1$ )

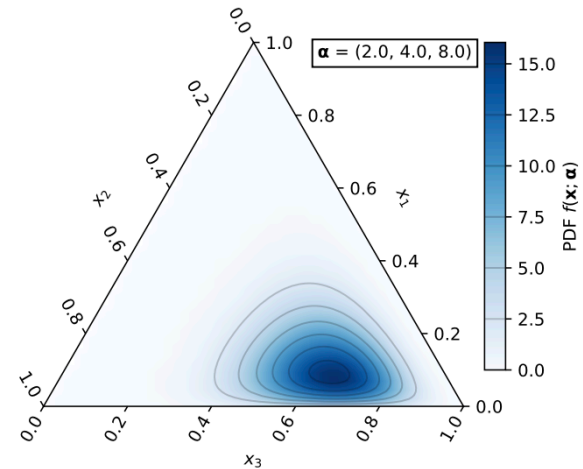
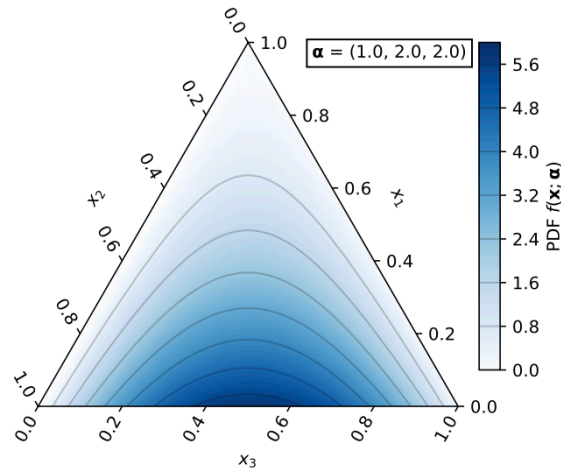
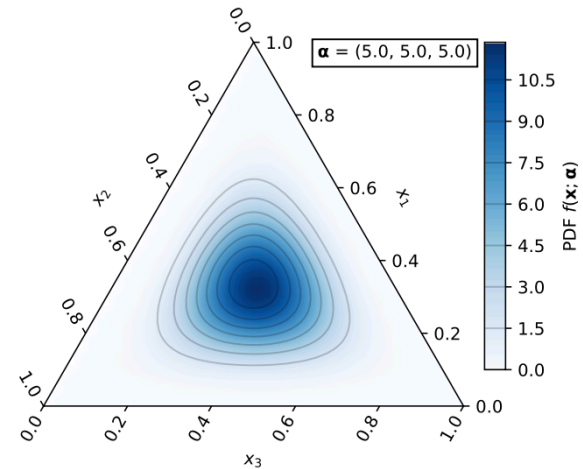
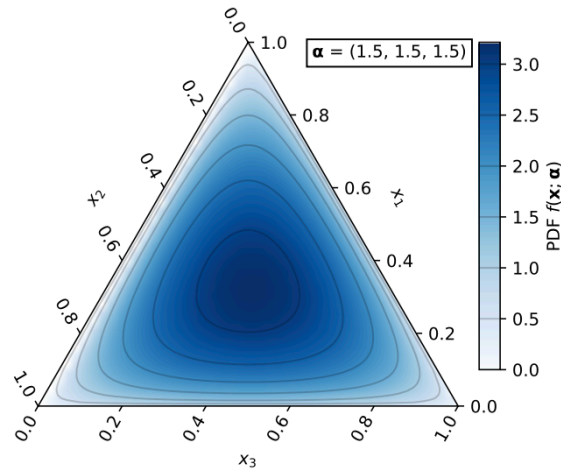
# LDA: Key Distributions

## Dirichlet distribution

- Provides a way to randomly generate *multinomial distributions*
- The pie split example
  - Imagine if you did not decide ahead of time exactly how to slice the pie
  - Instead, we specify a random process to decide how likely each person gets the pie
  - Equal chances among the three people, or perhaps it's more likely that one person gets it than the others
  - E.g.,  $[0, 0, 1]$ ,  $[1/3, 1/3, 1/3]$ ,  $[1/2, 1/4, 1/4]$ , etc.

# LDA: Key Distributions

$$\text{Dir}(x = 3; \alpha)$$



# LDA: Generative process

Consider a corpus of  $D$  documents, each with  $N_d$  words

- Assume a statistical model that generated our documents, then estimate the model and recover latent (unobserved) topics
- **Each document is seen as a multinomial distribution over topics**
- **Each topic is seen as a multinomial distribution over words**
- E.g., a stylized corpus with  $D = 5$ ,  $N = 8$  per document (and also 8 unique words in the corpus in total), and  $K = 3$

# LDA: Generative process

E.g.,  $D = 5$ ,  $K = 3$ ,  $N = 8$

- $\theta_d \sim \text{Dir}(\alpha)$ : for each  $d$ , its **topic distribution** is drawn from a Dirichlet distribution (e.g.,  $\theta_1 = [0.1, 0.7, 0.2]$ )
- $\beta_k \sim \text{Dir}(\eta)$ : for each  $k$ , its **word distribution** is drawn from another Dirichlet distribution (e.g.,  $\beta_2 = [0, 0, 0.2, 0.1, 0, 0, 0.3, 0.4]$ )
- $z_{d,n} \sim \text{Multinomial}(\theta_d)$ : for each document-word position, its topic is drawn from  $\theta_d$  (i.e., Topic 1 can be drawn)
- $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$ : for each document-word position, its word is drawn from the corresponding topic distribution (i.e., Word 8 can be drawn)

# LDA: Generative process

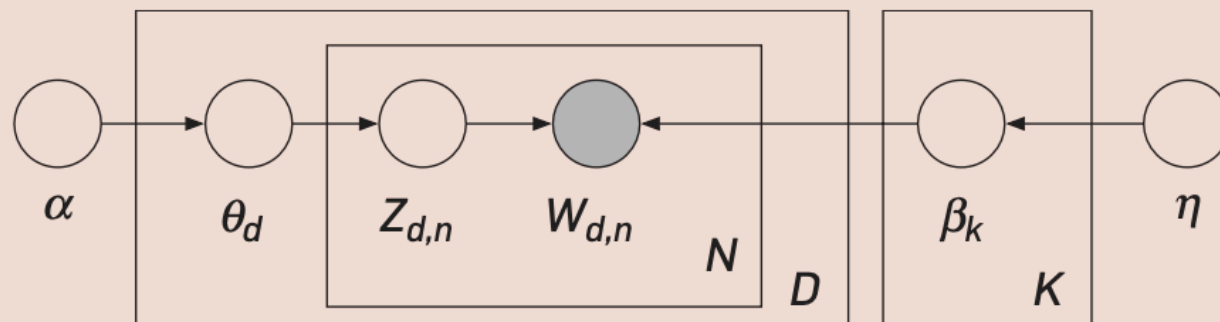
In summary, the joint probability is the following

- $\theta_d \sim \text{Dir}(\alpha)$  for  $d \in \{1, \dots, D\}$
- $\beta_k \sim \text{Dir}(\eta)$  for  $k \in \{1, \dots, K\}$
- $z_{d,n} \sim \text{Multinomial}(\theta_d)$
- $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

$$\begin{aligned} p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\ = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \\ \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \end{aligned}$$

# LDA: Generative process

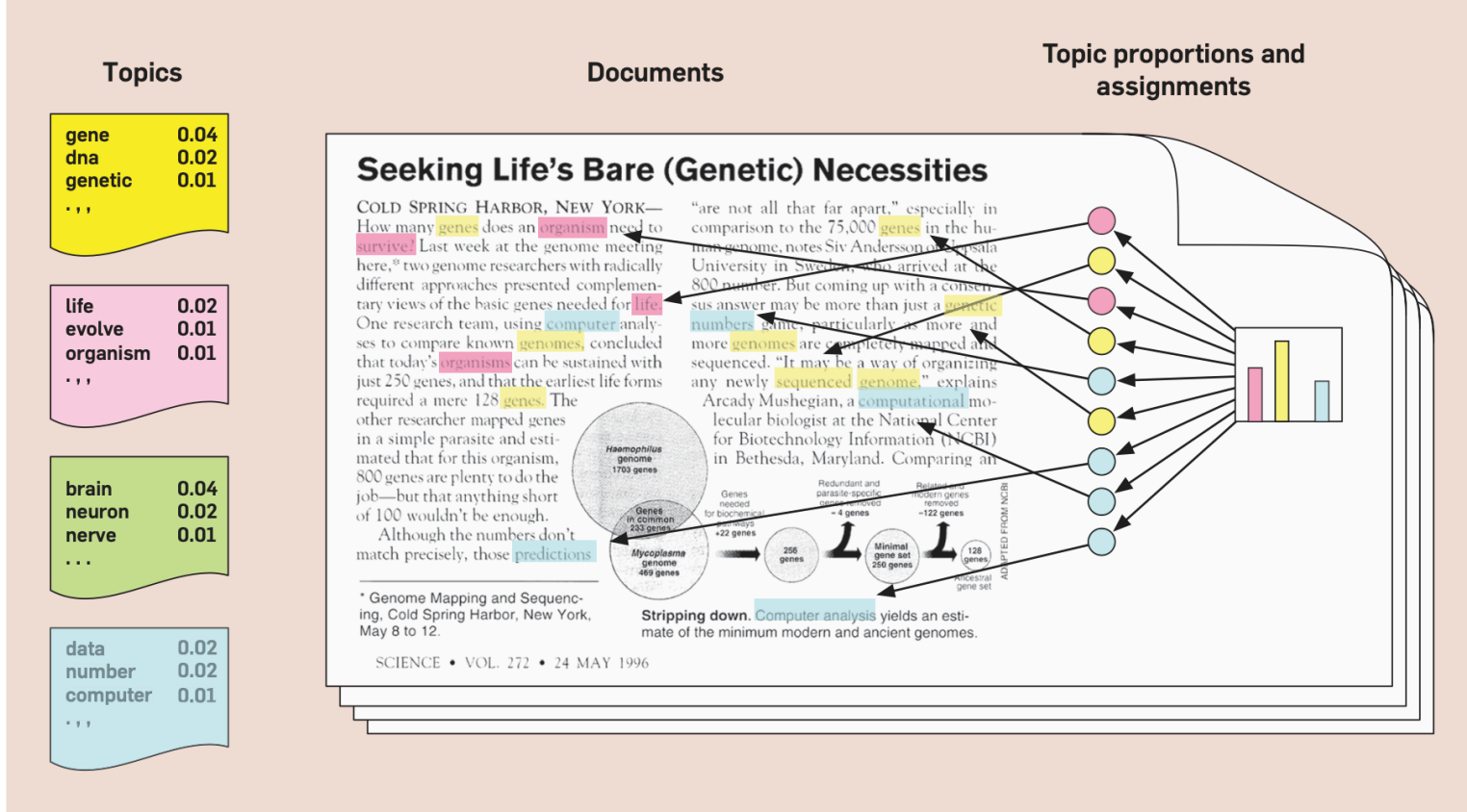
**Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are “plate” notation, which denotes replication. The  $N$  plate denotes the collection words within documents; the  $D$  plate denotes the collection of documents within the collection.**





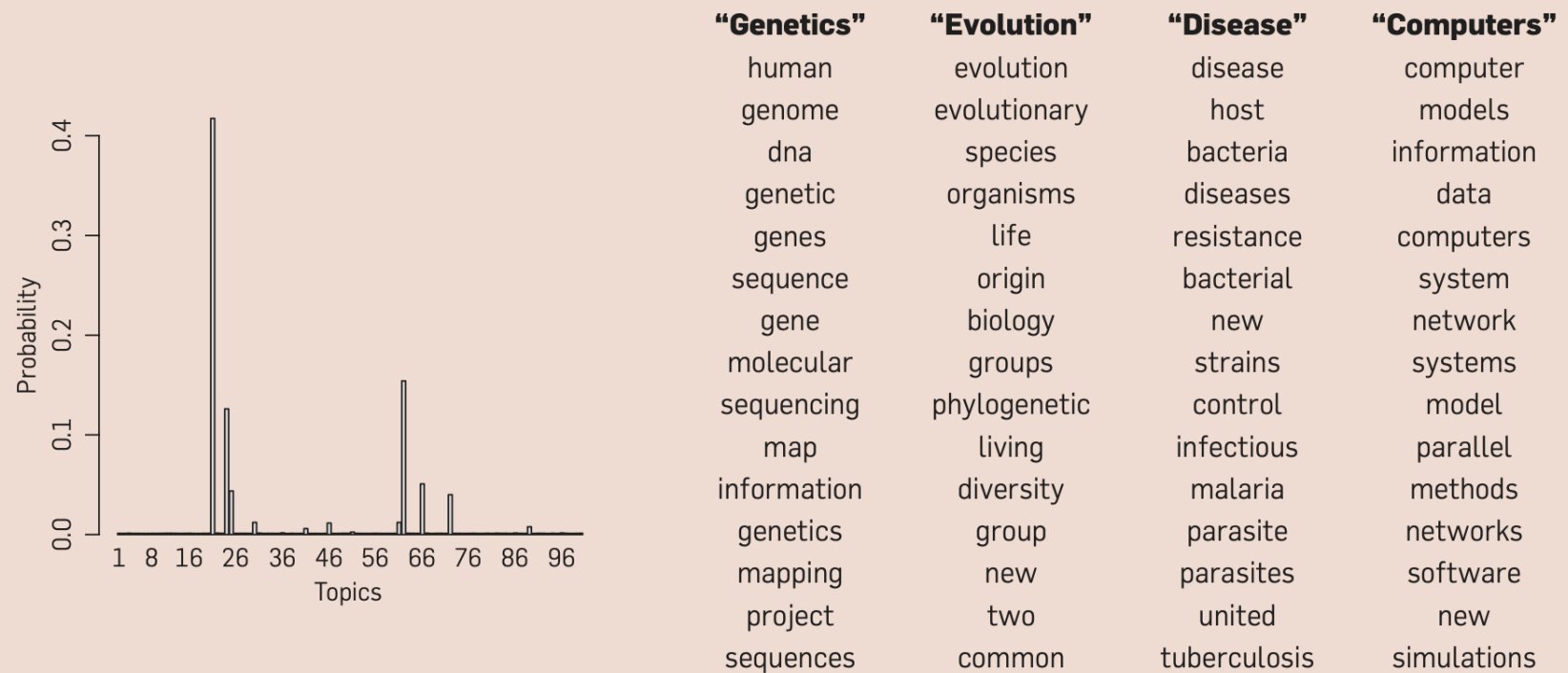
# LDA: Generative process

**Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of “topics,” which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.**



# LDA: Generative process

**Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.**



# LDA: Estimation

E.g., with  $D = 1000$ ,  $V = 10000$ , and  $K = 3$

- LDA estimates  $\theta$  (distributions over topics) and  $\beta$  (distributions over vocabulary)

$$\theta = \underbrace{\begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} \\ \dots & \dots & \dots \\ \theta_{1000,1} & \theta_{1000,2} & \theta_{1000,3} \end{pmatrix}}_{1000 \times 3} = \underbrace{\begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.6 \\ \dots & \dots & \dots \\ 0.1 & 0.8 & 0.1 \end{pmatrix}}_{1000 \times 3}$$

$$\beta = \underbrace{\begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,10000} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,10000} \\ \beta_{3,1} & \beta_{3,2} & \dots & \beta_{3,10000} \end{pmatrix}}_{3 \times 10000} = \underbrace{\begin{pmatrix} 0.04 & 0.01 & \dots & 0.0001 \\ 0.00002 & 0.001 & \dots & 0.05 \\ 0.00001 & 0.03 & \dots & 0.0001 \end{pmatrix}}_{3 \times 10000}$$

# LDA: Estimation

Estimation is done in a Bayesian framework

- $\text{Dir}(\alpha)$  and  $\text{Dir}(\eta)$  are the so called “prior distributions” of  $\theta$  and  $\beta$  ( $\text{Pr}(\mathbf{B})$ )
- With Bayes’ rule (likelihood:  $\text{Pr}(\mathbf{A}|\mathbf{B})$ ; evidence: ( $\text{Pr}(\mathbf{A})$ )), we update these prior distributions to obtain “posterior distributions for  $\theta$  and  $\beta$  ( $\text{Pr}(\mathbf{B}|\mathbf{A})$ )
- Estimation methods (see [Blei \(2012\)](#) for more discussions)
  - Gibbs sampling methods
  - Variational approximations

# Model Selection

Determining hyperparameters:  $K$  (and  $\alpha, \eta$ )

- A combination of quantitative metrics and human judgement
- The Dirichlet priors;  
 $\alpha$  (the topic distribution prior) and  
 $\eta$  (the word distribution prior)
  - Often set asymmetric (e.g., [0.1, 0.1, 0.1] for  $K = 3$ )
  - Typically both are set at 0.1 or 0.01 (the smaller, the fewer topics/words dominate) but this *can be tuned and affect* model performance
  - See ([Maier et al. 2018](#)) for more discussions

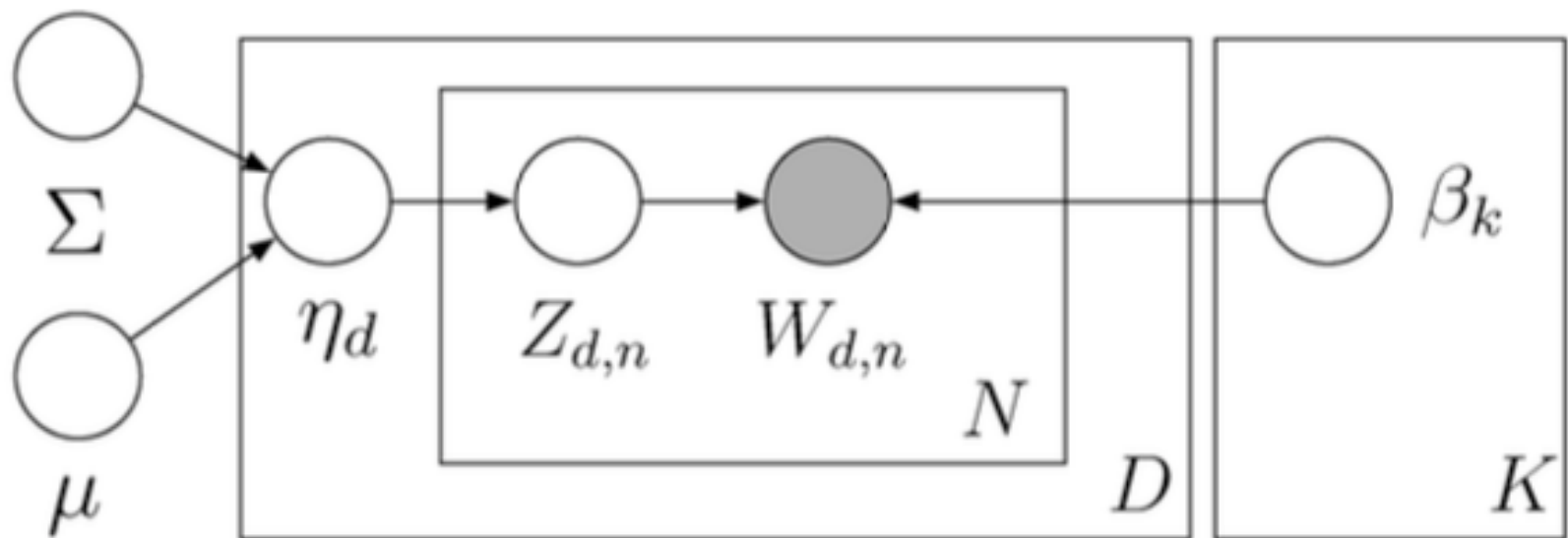
# Extensions

## Correlated Topic Model (CTM)

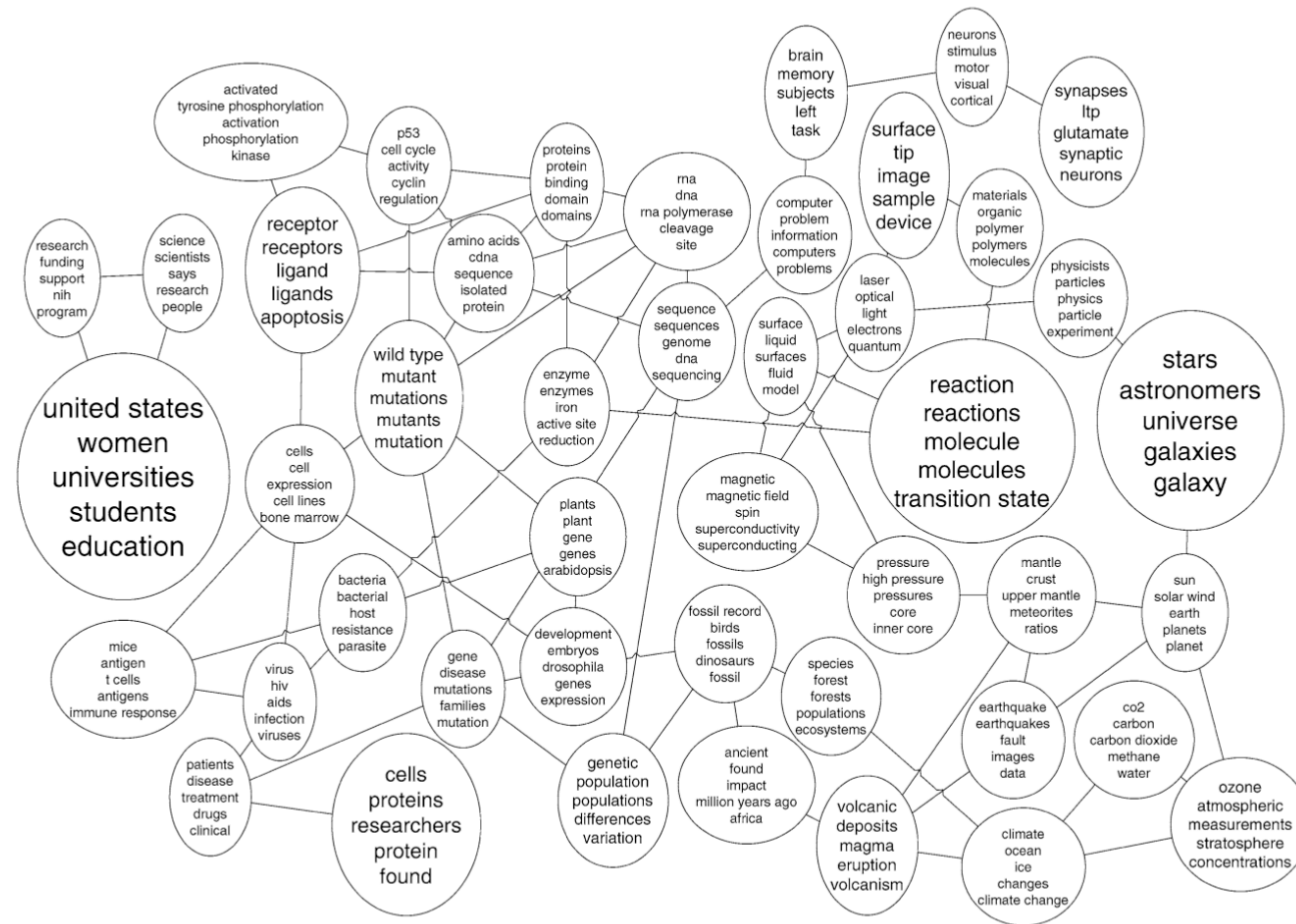
- [Blei and Lafferty \(2005\)](#) and [\(2007\)](#)
- A limitation of LDA is the inability to model topic correlation
- “A document about genetics is more likely to also be about disease than X-ray astronomy” ([Blei and Lafferty \(2007\)](#))
- CTM relies on logistic normal distribution, which allows for modeling correlations between topics through a covariance matrix
- The authors report that this can lead to a better model fit (what the model fit means will be discussed soon)

# Extensions

## Correlated Topic Model (CTM)



# Correlated Topic Model (CTM)





# Extensions

## Correlated Topic Model (CTM)

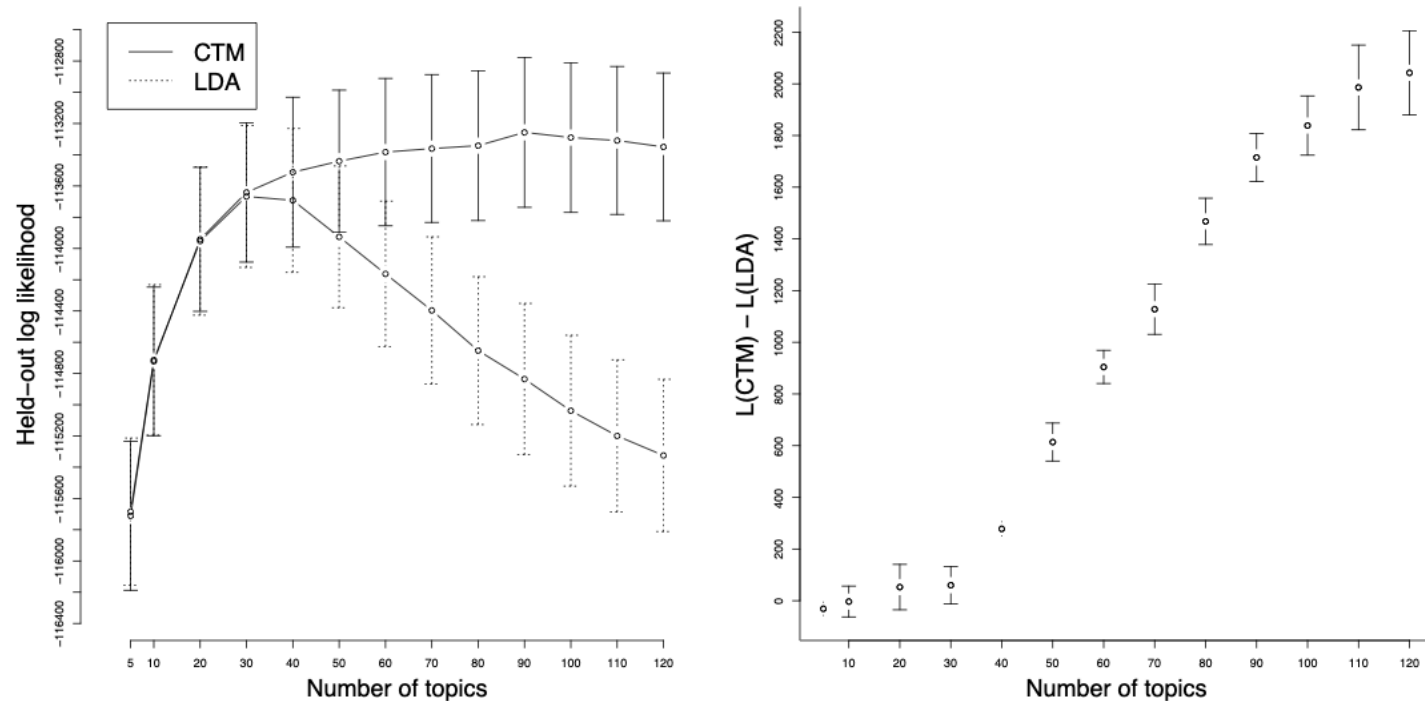


Figure 3: (L) The average held-out probability; CTM supports more topics than LDA. See figure at right for the standard error of the difference. (R) The log odds ratio of the held-out probability. Positive numbers indicate a better fit by the correlated topic model.

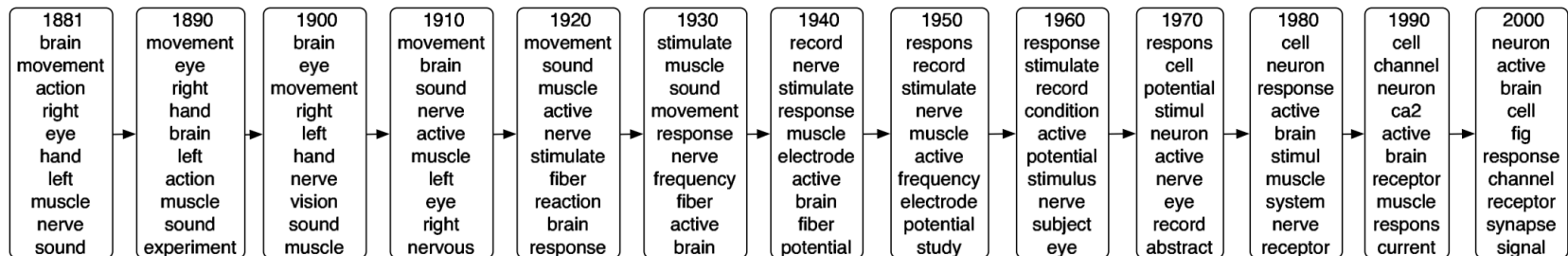
# Extensions

## Dynamic Topic Model (DTM)

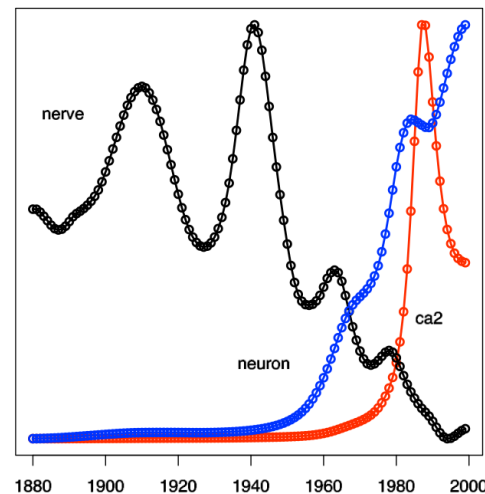
- [Blei and Lafferty \(2006\)](#) and [Blei \(2012\)](#)
- LDA assumes that the temporal order of documents does not matter
- This can be unrealistic when analyzing long-running collections
- Instead of a single distribution over words, a topic is now a sequence of distribution over words

# Extensions

## Dynamic Topic Model (DTM)



"Neuroscience"



- 1887 Mental Science
- 1900 Hemianopsia in Migraine
- 1912 A Defence of the ``New Phrenology''
- 1921 The Synchronal Flashing of Fireflies
- 1932 Myoesthesia and Imageless Thought
- 1943 Acetylcholine and the Physiology of the Nervous System
- 1952 Brain Waves and Unit Discharge in Cerebral Cortex
- 1963 Errorless Discrimination Learning in the Pigeon
- 1974 Temporal Summation of Light by a Vertebrate Visual Receptor
- 1983 Hysteresis in the Force-Calcium Relation in Muscle
- 1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

# Extensions

## Structural Topic Model (STM)

- [Roberts et al. \(2013\)](#) and [Roberts et al. \(2014\)](#)
- Key motivation: short documents do not provide the opportunity to observe the correlations between words—key information used to estimate topics
- “Structure”: how document-level covariates drives topics
  - E.g., group, time, etc.
- Covariates can be modeled to estimate **a) topic prevalence** and **b) topic content**
- Without any covariates, the model reduces to CTM

# Extensions

## Structural Topic Model (STM)

- Topic prevalence
  - Documents which have similar covariates will tend to talk about the same topics
  - Topic proportions within documents can vary through covariates
  - E.g., social media posts by Republican politicians might have different topic proportions than those posted by Democrats

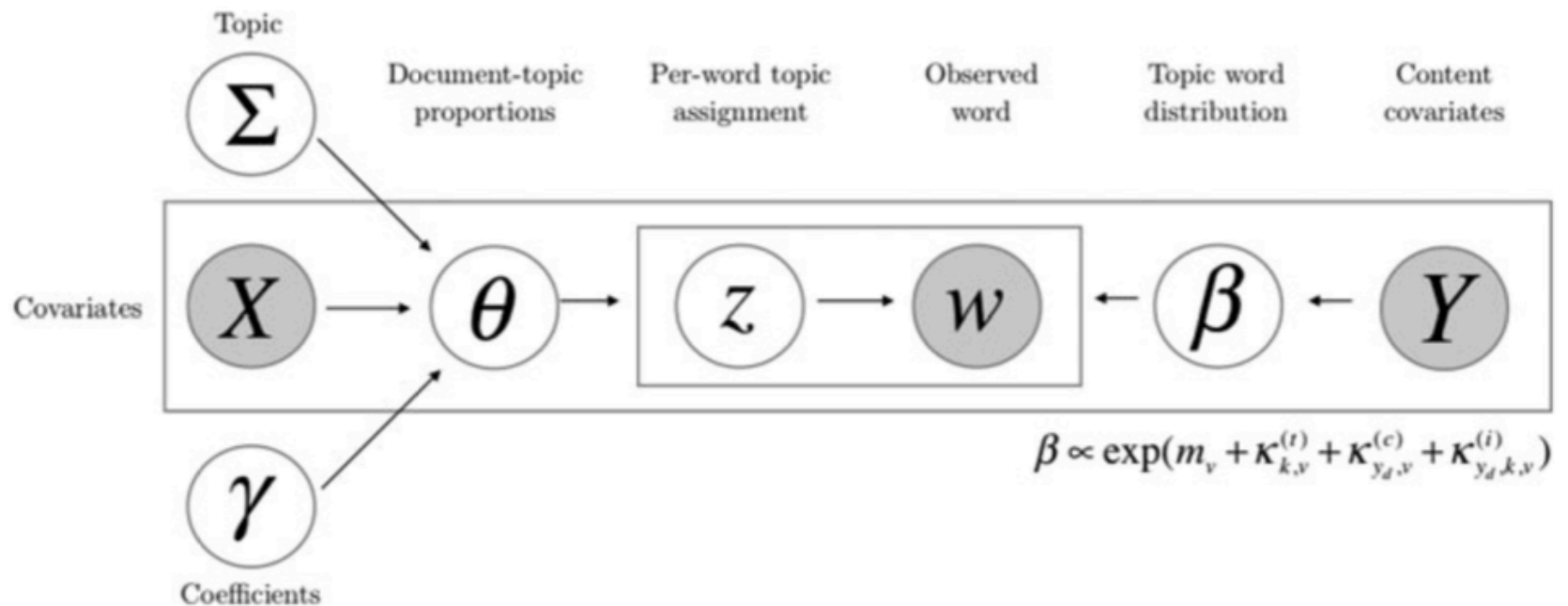
# Extensions

## Structural Topic Model (STM)

- Topic content
  - Word proportions within topics can vary through covariates
  - Documents which have similar covariates will tend to talk about topics in a similar way
  - E.g. when talking about a health care topic, Republican politicians might use different words than Democrats

# Extensions

## Structural Topic Model (STM)



# Model Selection and Interpretation

## Quantitative + qualitative approaches

- Quantitative
  - Perplexity (or held-out likelihood): for some held-out documents, how likely would the model have generated/predicted these documents?
    - Evaluation of predictive power  $\neq$  coherent topics
    - For how they differ, see [Chang et al.\(2009\)](#) and/or [this video](#)
  - Semantic coherence: how likely do the most common words for a topic co-occur in the same documents?
  - Exclusivity: do words with high probability in one topic have low probabilities in others?



# Model Selection and Interpretation

## Quantitative + qualitative approaches

- Qualitative
  - Careful reading of exemplar texts
    - Select a small subset of the documents with the highest proportion of the document assigned to the particular category under consideration
    - Read those documents to assess their common facets and interrogate whether a particular organization makes sense
  - Examine words that are indicative of a particular topic
    - The most straightforward method for obtaining these words is to select the highest probability words in each topic

# Model Selection and Interpretation

## Quantitative + qualitative approaches

- No quantitative metric can replace human judgement
- *“The most effective method for assessing model fit is to carefully read documents that are closely associated with particular topics to verify that the semantic concept covered by the topic is reflected in the text.”* ([Roberts et al. 2016](#))

# Summary

- Topic models help us explore or explicitly measure the thematic structure of a large collection of documents
- Select an algorithm that suits your goal
- Experiment with hyper-parameters, including the number of topics
- Quantitative measures for model selection is useful
- Manual reading plays a crucial role in model selection and labeling

# Guided Coding

- Guided Coding for LDA: [LDA with SOTU address data in R](#)
- Materials for STM
  - [Official documentation of stm package](#)
    - [Short demonstration](#)
  - [Tutorial with Facebook posts data](#)