Text Analysis for Social Sciences in Python

Week 10: Text as Data – Topic Models

WINTER SEMESTER 21/22

HUYEN NGUYEN

HTTP://HUYENTTNGUYEN.COM

Last week…

Word Embeddings

…any questions with the materials and exercises?

Thank you to everyone who sent me the codes for the Bonus Exercise 1!

Final presentation date

Monday 31.01: 14.00 – 18.00 (20-25' presentation/group + 10' Q&A)

(i.e. code & slides submission by Thursday 27.01 23.59 to Open Olat folder)

Slot allocation (Google sheet link on your Slack channel #group_matching):

https://docs.google.com/spreadsheets/d/1E-IFBNDNk7-5GxALoYRfK-P8GO5aYJ8v9j0uEjUsZ5M/edit#gid=548845839

Today's agenda

Word Embedding (cont)

- Cosine Similarity
- Collocation

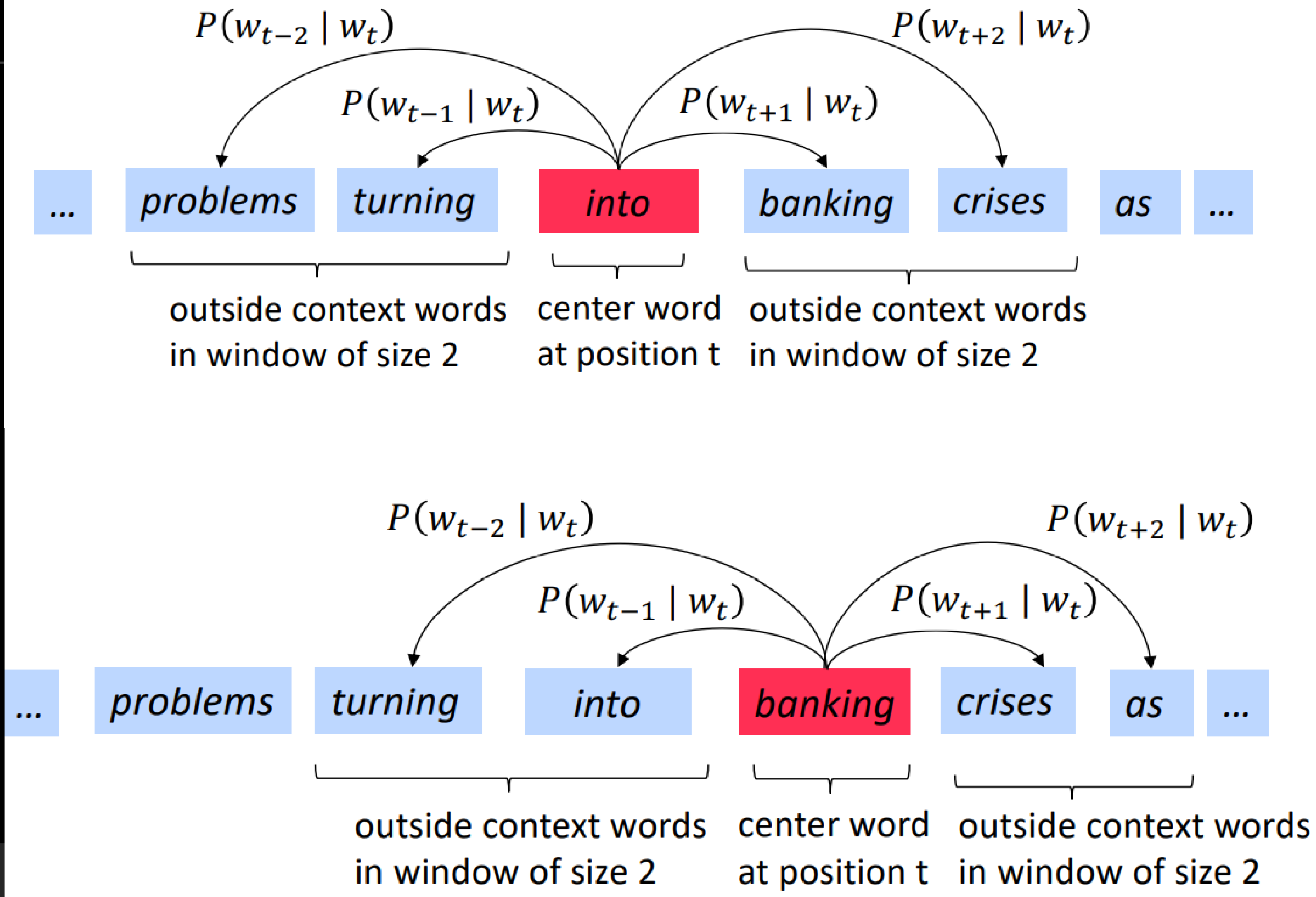
Topic models

- Word Embeddings vs. Topic Models?
- Topic models in Social Sciences
- Latent Dirichlet Allocation (LDA)
- Co-occurrence
- Structural Topic Model (STM)

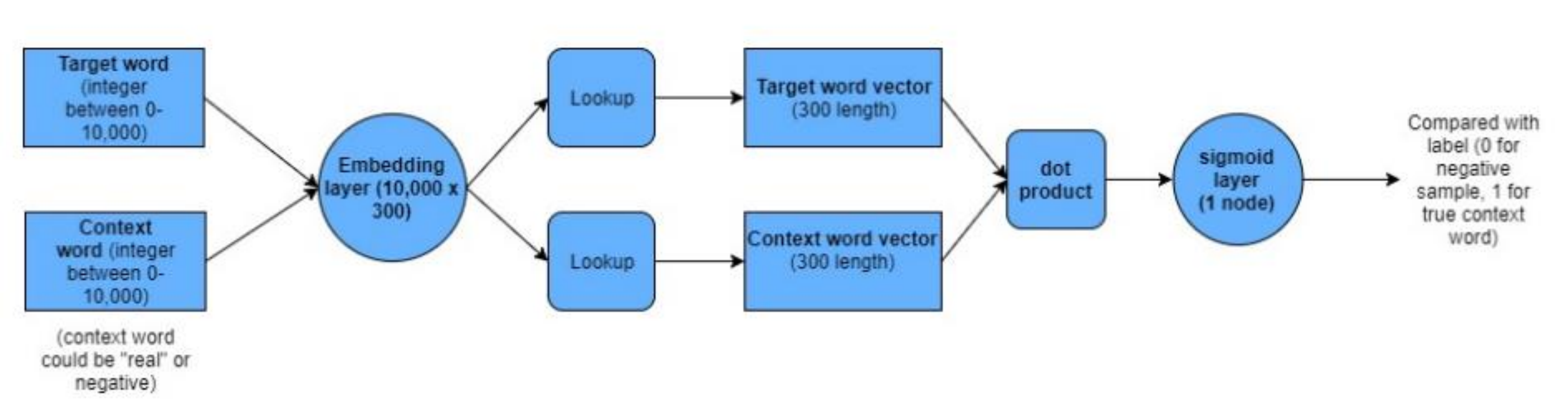
Word meaning as a neural word vector – visualization

$$\textit{expect} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$$

Example windows and process for computing $P(w_{t+j} \mid w_t)$



Word2vec schema



CBOW or Skip-gram Word2Vec?

CBOW is several times faster than skip gram and provides a better frequency for frequent words

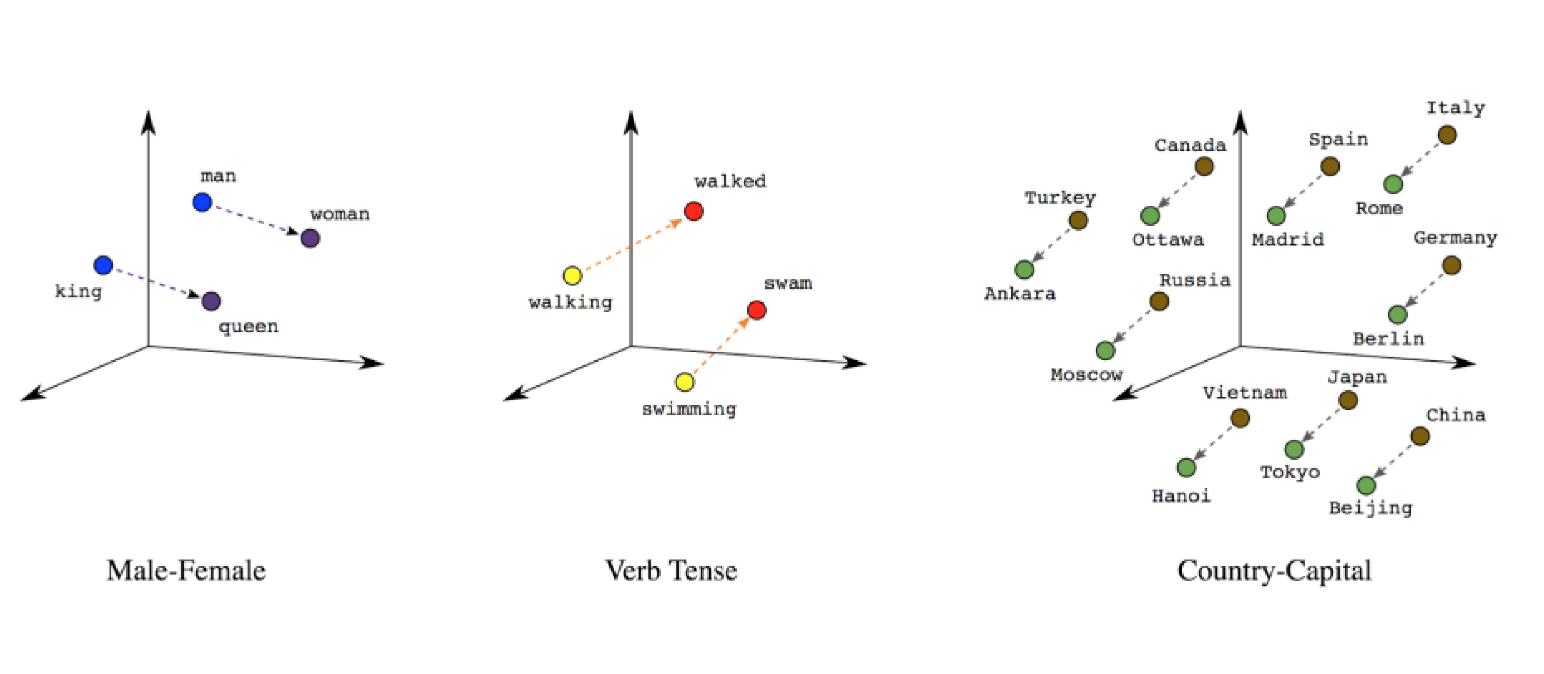
BUT skip gram needs a small amount of training data and represents even rare words or phrases.

→ Choose the model that works best for your data set.

How to measure cosine similarity?

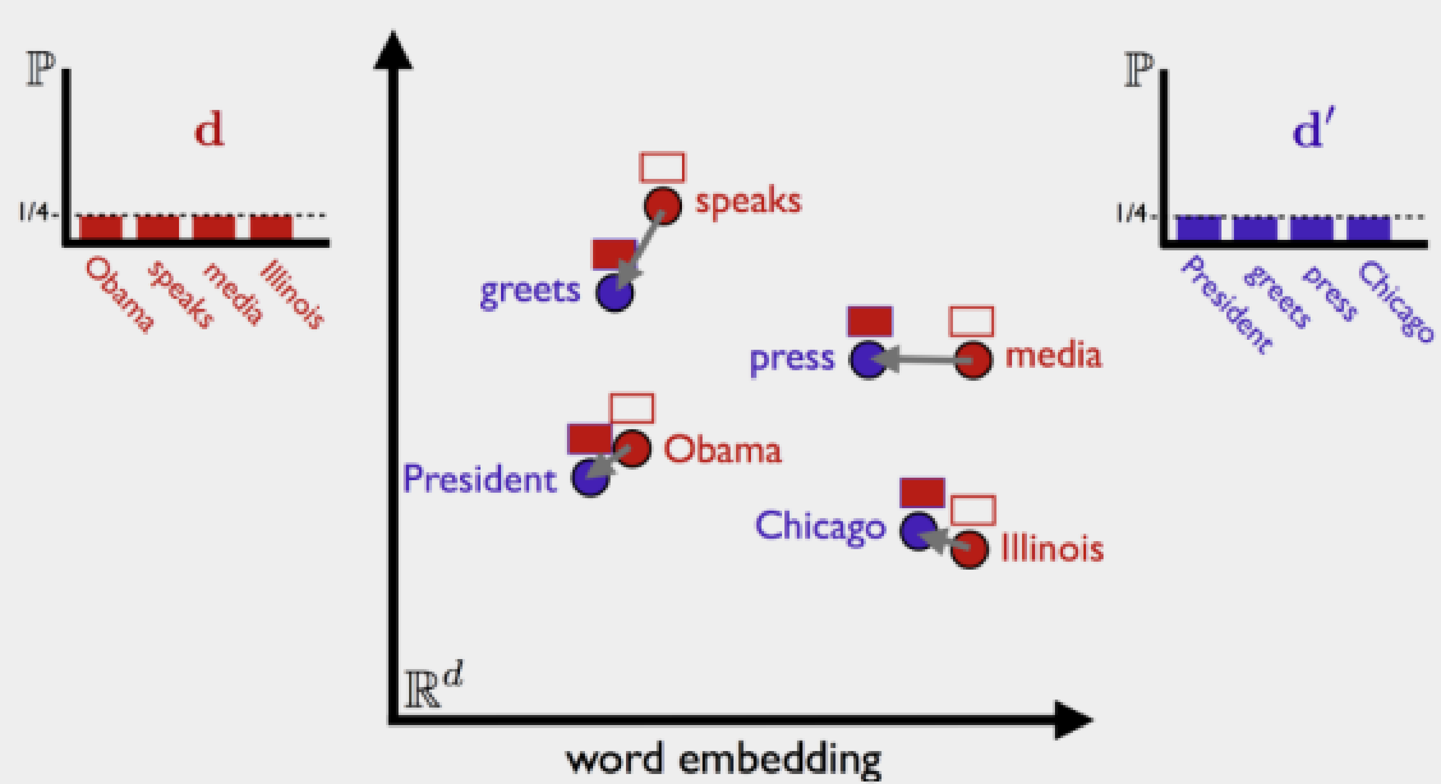
$$cos\theta = \frac{v_1 . v_2}{\|v_1\| \|v_2\|}$$

where v1 and v2 are vectors representing words (Word Embeddings/ documents (topic models).



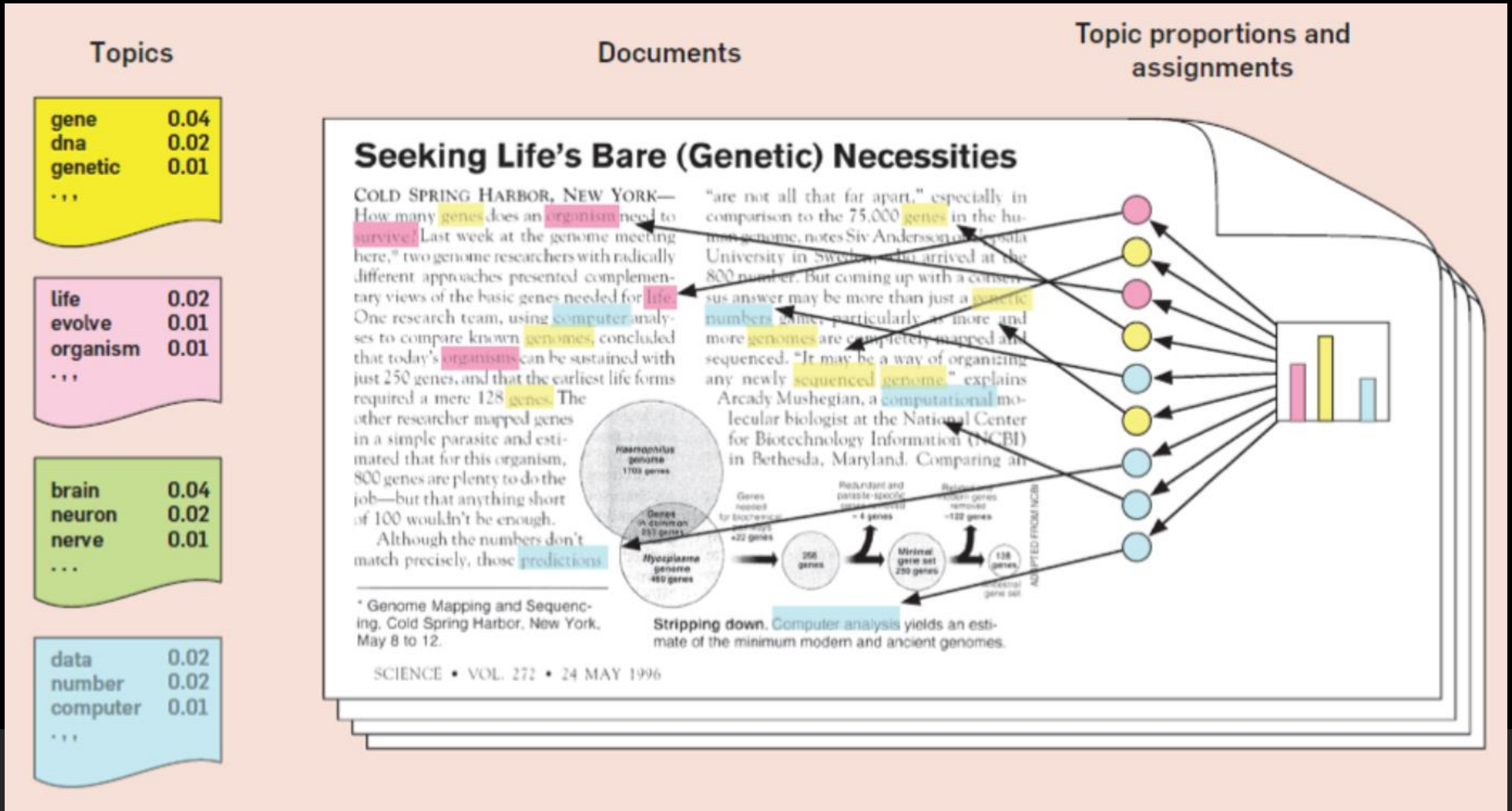
Word embeddings vs. topic models?

Word embedding models ignore information about individual documents → better understand the relationships between words.



Word embeddings vs. topic models?

Topic models reduce words to core meanings → understand documents more clearly.



Topic Models in Social Sciences

Core methods for topic models were developed in computer science and statistics to…

- summarize unstructured text
- use words within document to infer its subject
- reduce dimensionality

How are they used in social sciences?

- use topics as a form of measurement
- check how observed covariates drive trends in language
- tell a story not just about what, but how and why I topic models are more interpretable than other methods, e.g. PCA

Some example research papers

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Do presidential candidates move to the center/left/right in their political speeches? (Sim et al 2013)
- How do central bankers respond to an increase in transparency over their discussions? (Hansen, McMahon, and Prat 2018)

Document-term Matrix X

	W1	W2	W3	Wm
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

- A corpus of N documents D1, D2, D3,....Dn
- Vocabulary of M words W1, W2, W3,..... Wn
- The value of i, j cell gives the frequency count of word Wj in Document Di.

Matrix Factoring

LDA converts the document-term matrix into two lower-dimensional matrices, M1 and M2:

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

M1 is a N ×K document-topic matrix.

M2 is a K ×M topic-term matrix.

Latent Dirichlet Allocation (LDA)

Idea: documents exhibit each topic in some proportion.

- Each document is a distribution over topics.
- Each topic is a distribution over words.

Latent Dirichlet Allocation (LDA)

Idea: documents exhibit each topic in some proportion.

- Each document is a distribution over topics.
- Each topic is a distribution over words.

Latent Dirichlet Allocation estimates:

- The distribution over words for each topic.
- The proportion of a document in each topic, for each document.

Latent Dirichlet Allocation (LDA)

Idea: documents exhibit each topic in some proportion.

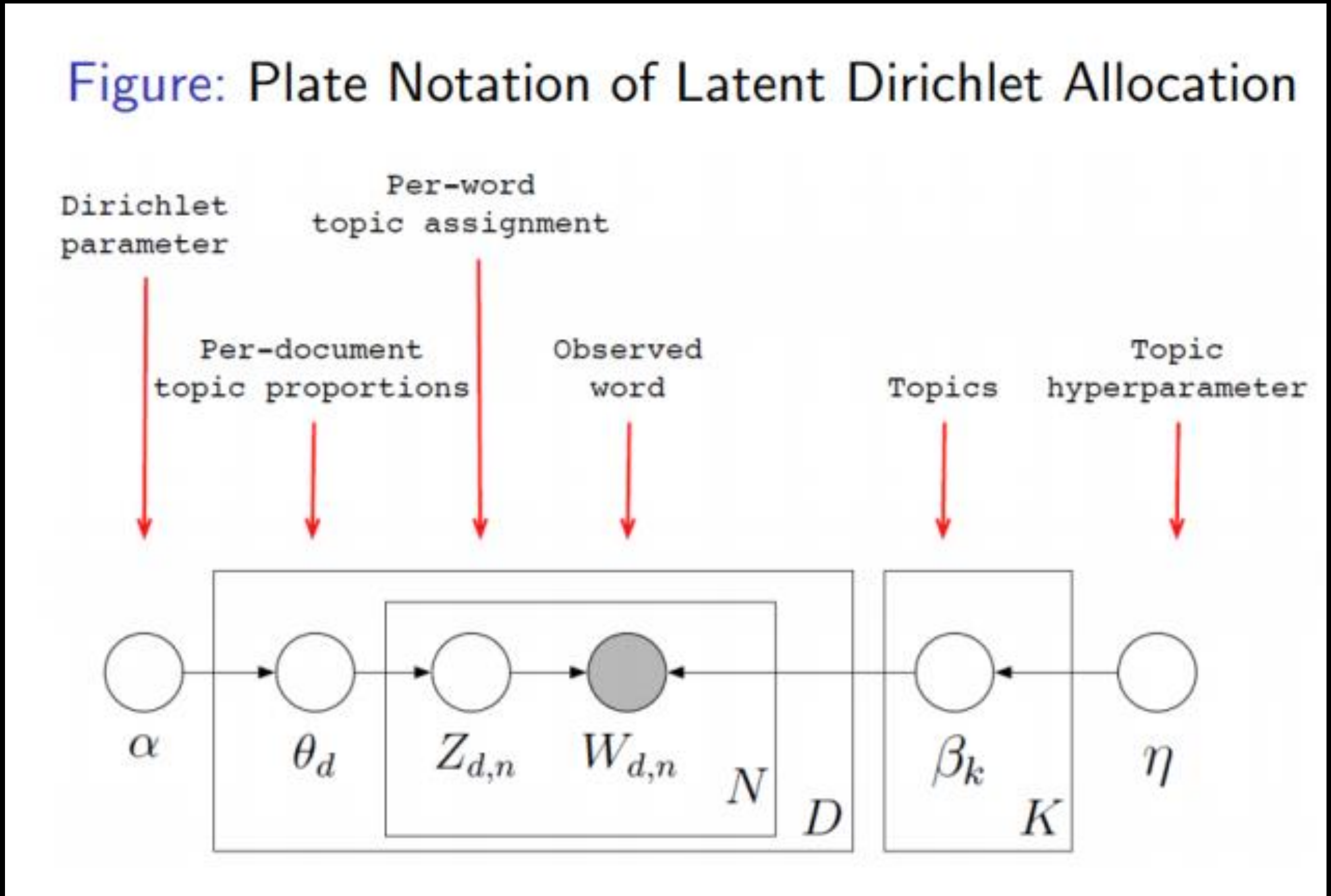
- Each document is a distribution over topics.
- Each topic is a distribution over words.

Latent Dirichlet Allocation estimates:

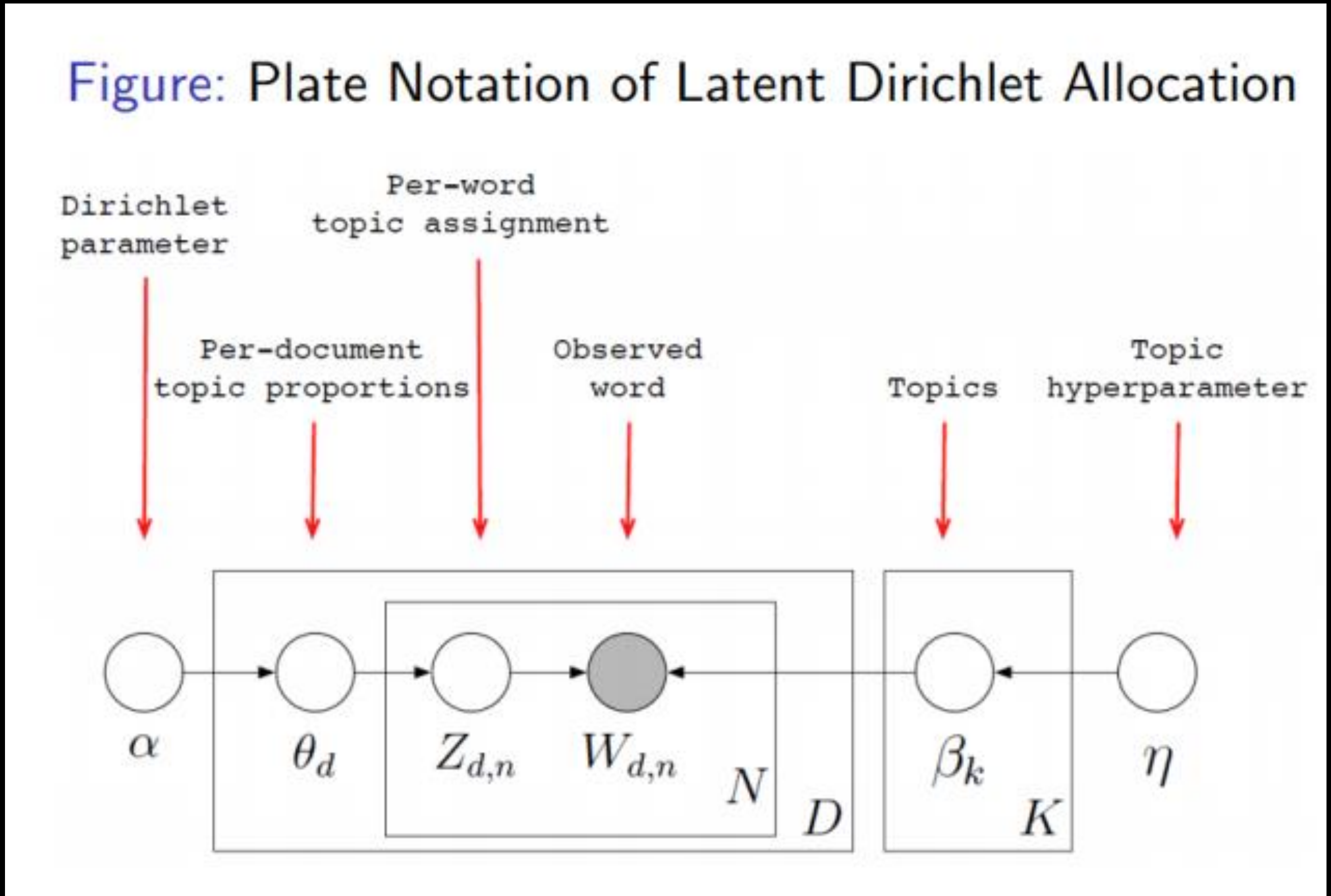
- The distribution over words for each topic.
- The proportion of a document in each topic, for each document.

Key assumptions: Bag of words/phrases, and fix number of topics ex ante.

A Bayesian model



A Bayesian model



α : document-topic density
↑↓ α = documents have ↑↓ topics

β: topic-word density
↑↓ β = topics have ↑↓ words

Number of topics:

- this is specified in advance, or can be chosen to optimize model fit.
- the “statistically optimal” topic count is usually too high for the topics to be interpretable/useful.

Why does it work? Co-occurrence

- Where is the information for each word's topic?
- We are learning the pattern of what words <u>occur together</u>.
- The model wants a topic to contain as few words as possible, but a document to contain as few topics as possible.
- This tension is what makes the model work.

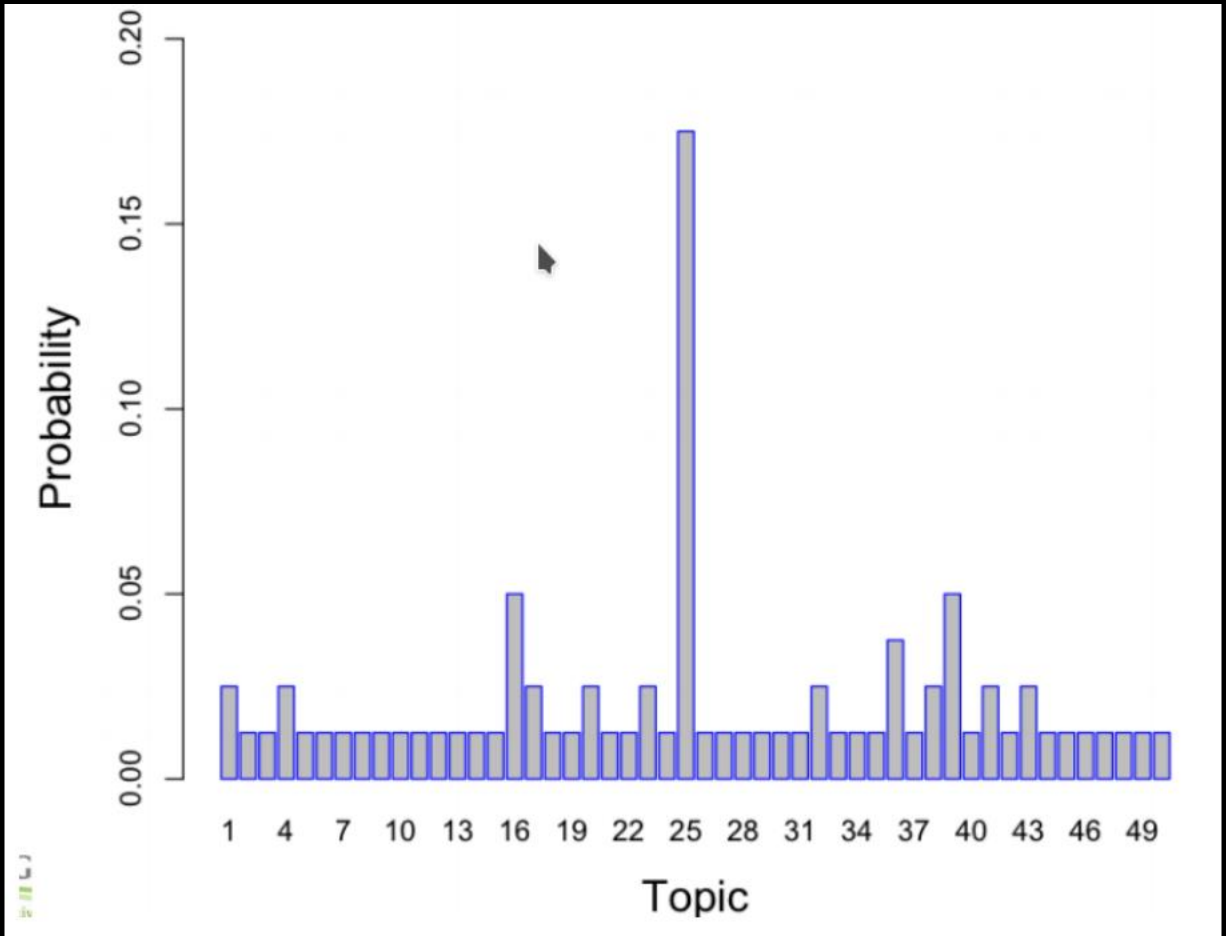
LDA in text application: An example (FOMC)

Hansen, McMahon, and Prat 2018 use LDA to analyze speech at the FOMC (Federal Open Market Committee).

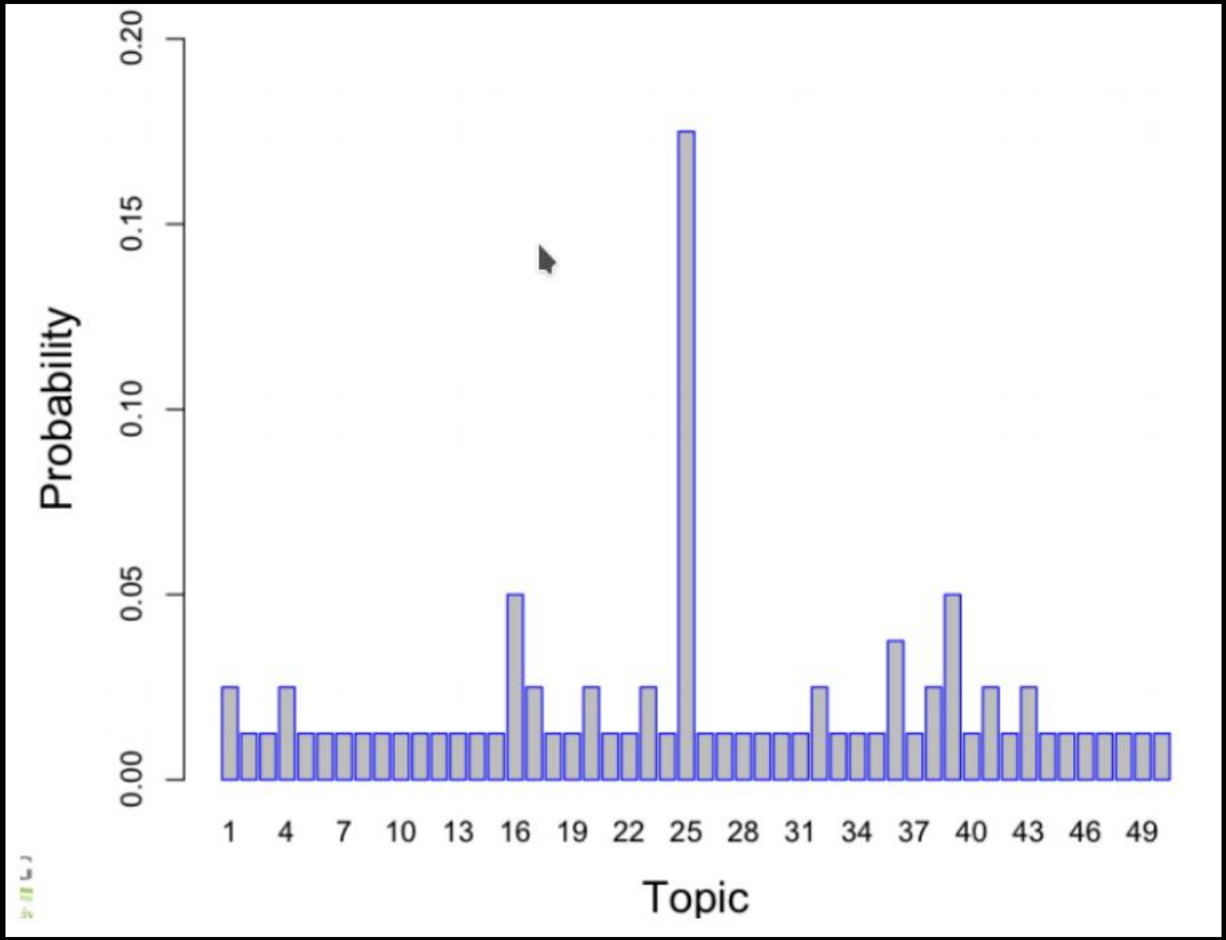
- 150 meetings, 20 years, 46000 speeches

They take the standard pre-processing steps and train LDA.

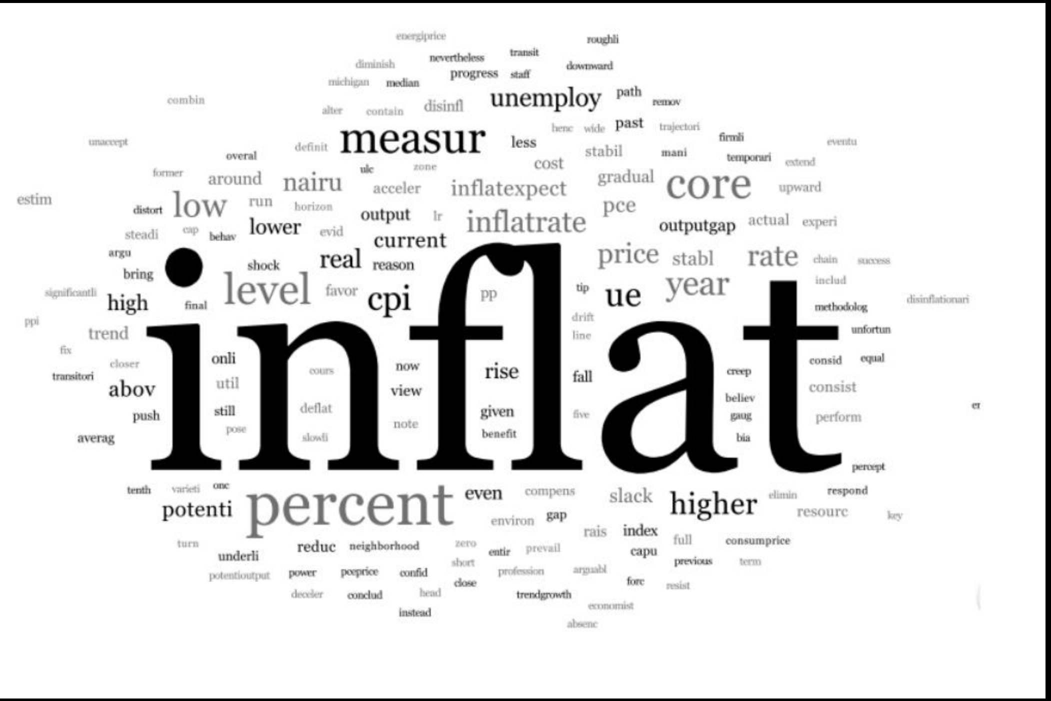
FOMC distribution of attention



FOMC distribution of attention



Topic 25 is…



What is their conclusion?

- Increasing transparency results in:
  - higher discipline / technocratic language (which is beneficial)
  - higher conformity (which is costly)

- Highlights tradeoffs from transparency in bureaucratic organizations

LDA on non-text applications?

LDA can be used on any type of co-occurence data. E.g:

- [Bandiera, Hansen, Prat, and Sadun (2017)]
  - use LDA to do dimension reduction on CEO tasks using time allocation data.
  - model endogenously identifies a “leader” CEO topic, and a “micro-manager” CEO topic
- [Draca and Schwarz (2018)]
  - use LDA to reduce dimensions of the features of political attitudes
  - model endogenously identifies “conservative” and “liberal” attitudes clusters.
- [Draca and Schwarz (2021)]
  - identify the underlying ideologies of citizens using LDA in political survey data.
  - evidence of a `disappearing centre' in a sub-group of countries: citizens shift away from centrist ideologies into anti-establishment `anarchist' ideologies over time, especially in the US.

Structural Topic Model = LDA + metadata

STM provides two ways to include contextual information:

- Topic prevalence can vary by metadata
- e.g. Republicans talk about military issues more then Democrats

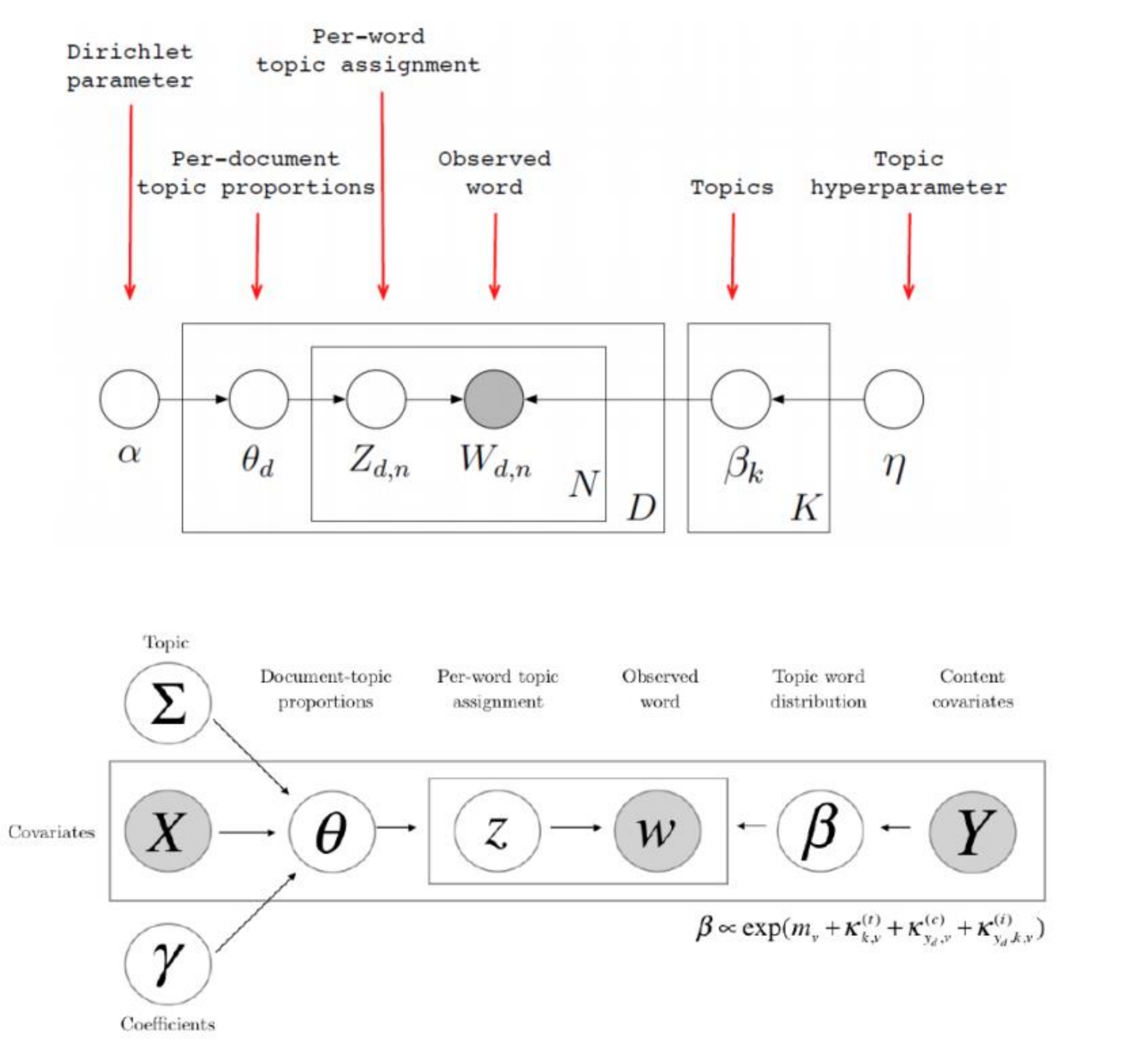
Topic content can vary by metadata

- e.g. Republicans talk about military issues differently from Democrats

Caveat: Structural topic model is not a prediction model.

- It will tell you which topics or features correlate with an outcome, but it will not provide an in-sample or out-of-sample prediction for an outcome.

LDA vs. STM illustration



Practice time ☺

Open your Google Colab/ Jupyter Notebook

Practice time ☺

Open your Google Colab/ Jupyter Notebook

https://github.com/httn21uhh/Text-Analysis-for-Social-Sciences-in-Python

→ W10_topic_model.ipynb + utils.py + txt_utils.py + death-penalty-cases.csv + X.pkl + X_tfidf.pkl

- Download and put them in the same directory/environment where you store other packages
- IMPORTANT: For txt_utils.py, open it first with Jupyter Notebook/Spyder/Pycharm and change the line of code WORKING_DIR to your own current working directory
- If none of these methods work, copy and run the relevant function code lines in these utils.py and txt_util.py into your own .ipynb. Then run them as usual.

This week…

- Topic models in action
- Follow closely the illustrated examples and replicate yourself as the session proceeds.
- Unmute yourself to ask questions at any point, including when you think things go too fast/slow for you.