whoami

Consulting

- Human-in-the-loop machine learning + MLOps

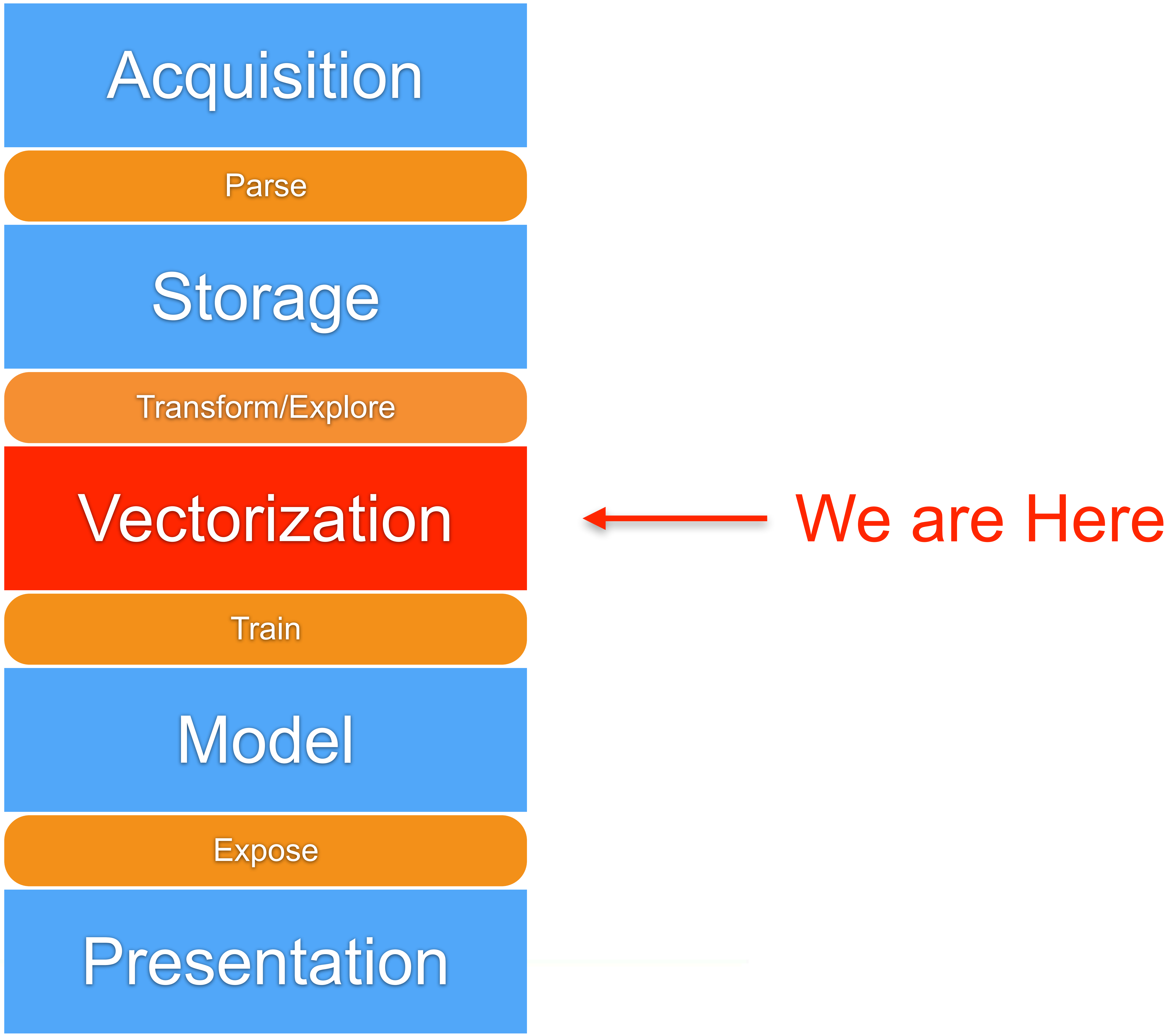
inquiries@jonathan.industries

Getting the Materials

https://github.com/jonathandinu/spark-livetraining

Working with Text: Introduction to NLP

Data Pipeline



Natural Language Processing

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

[1, 3, 1, 1, 2, 0, 1, 0]
[0, 1, 4, 0, 0, 1, 1, 1]
[3, 0, 1, 1, 2, 2, 3, 2]
[0, 1, 1, 1, 0, 3, 2, 3]
[1, 2, 1, 2, 2, 0, 0, 0]
[1, 0, 1, 1, 0, 1, 1, 1]
[0, 2, 0, 0, 2, 2, 0, 0]
[1, 1, 1, 1, 0, 1, 1, 1]

The Unreasonable Effectiveness of Data

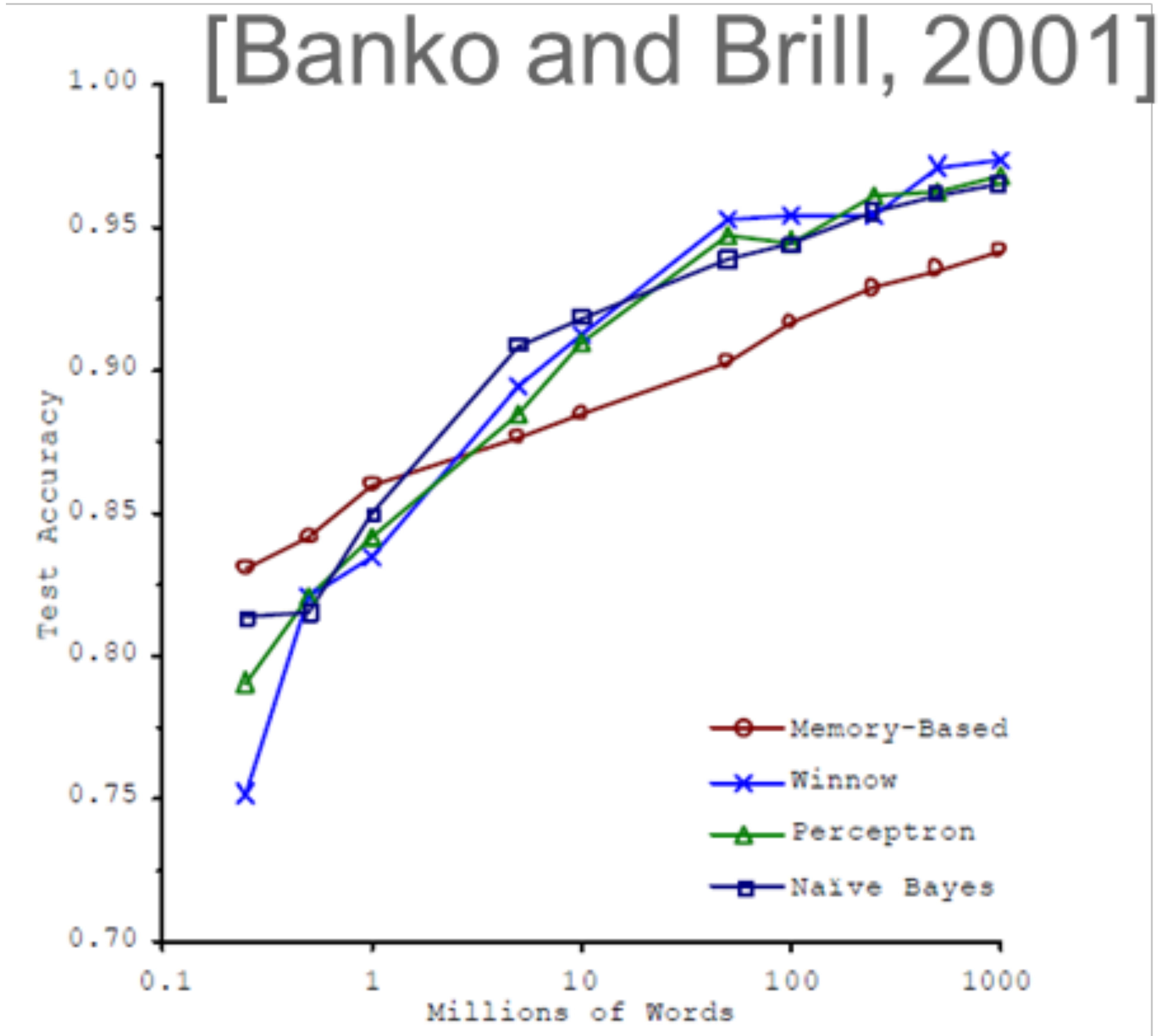


Figure 1. Learning Curves for Confusion Set Disambiguation

Source: https://www.microsoft.com/en-us/research/publication/scaling-to-very-very-large-corpora-for-natural-language-disambiguation/

Bag of Words

- **Document:** Single row of data/corpus
- **Corpus:** Entire set of all documents
- **Vocabulary:** Set of all words in corpus
- **Vector:** Mathematical representation of document (counts of word occurrences)

Bag of Words

The brown fox → { "the" : 1, "brown": 1, "fox" : 1 } → [0,0,1,0,1,0,...]

brown fox

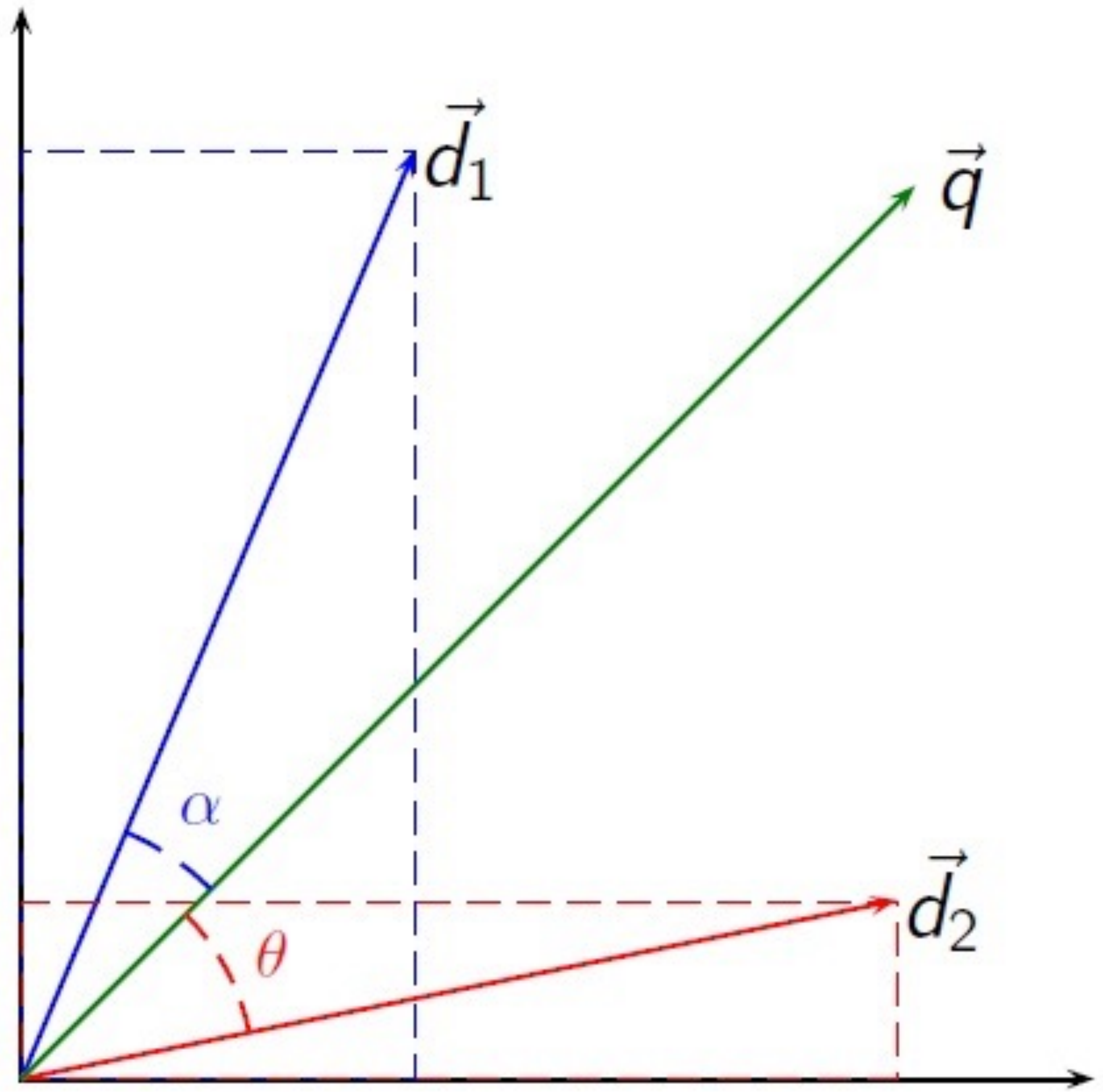
Tokenization

Vectorization

original document → dictionary of word counts → feature vector

Vector Space Model

Similarity is a measure of “distance”



TF-IDF

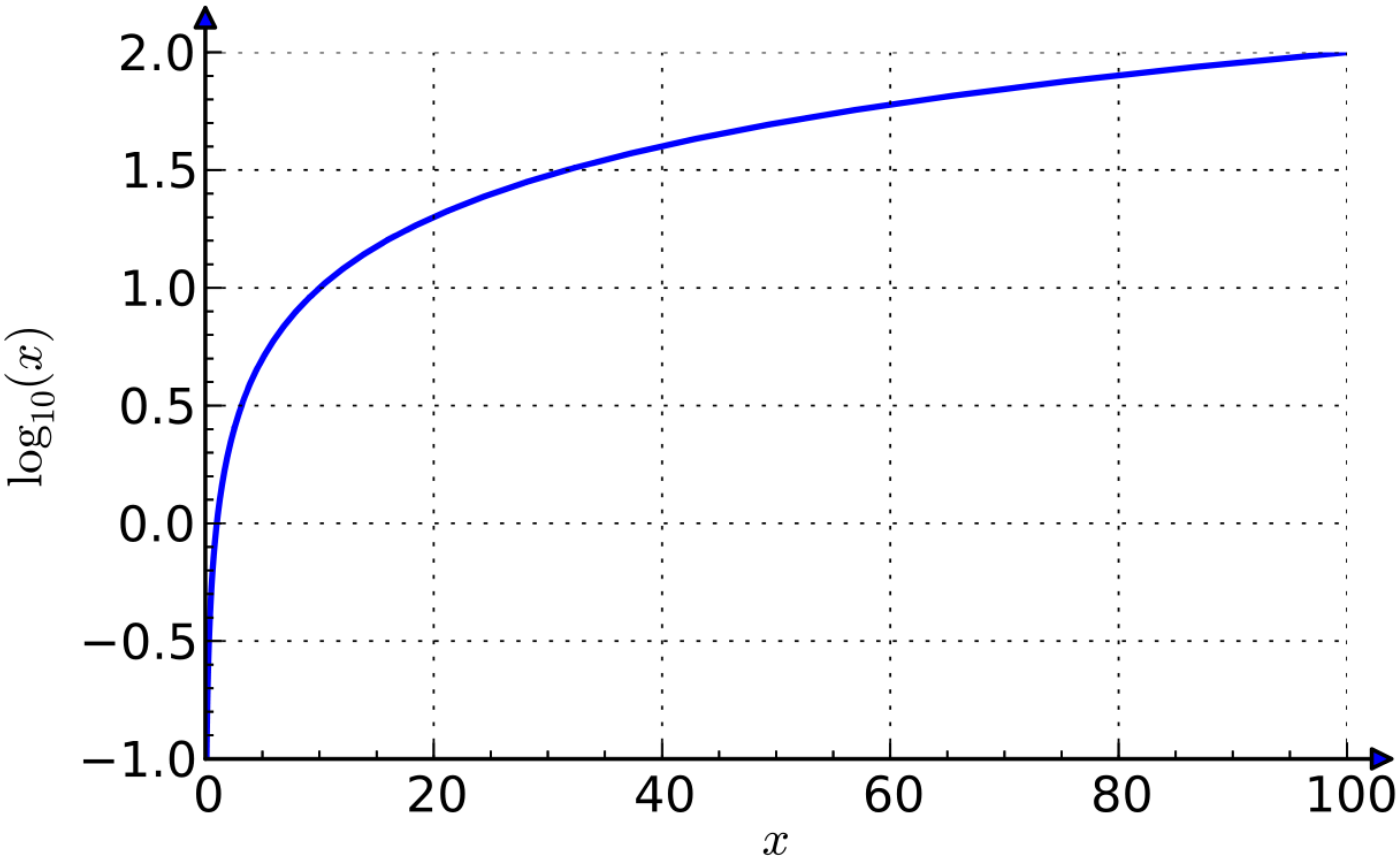
- Measure of discriminatory power of word (feature)
- Highest when term occurs many times in a small number of documents
- Lowest when term occurs few times in document or many times in corpus
- Useful for information retrieval (queries) and keyword extraction (among other things)

$$tf(t,d) = \frac{f_d(t)}{|d|} \qquad idf(t,D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

Tokenization and Vectorization with MLlib

Live Coding

TF-IDF



TF-IDF

Most Common

```
idf[:50]
```

```
[(u'students', 0.014067384597943282),
 (u'I', 0.15305316750494943),
 (u'school', 0.17010493952495984),
 (u'My', 0.3397655206814591),
 (u'The', 0.414913316782011 2),
 (u'help', 0.4188088461791251),
 (u'classroom', 0.5361023876769617),
 (u'learning', 0.5748186189046272),
 (u'need', 0.5820538952580256),
 (u'They', 0.5941434194555928),
 (u'learn', 0.6187002265438729),
 (u'able', 0.7452815794748304),
 (u'use', 0.7494117483916651),
 (u"''", 0.755060153205684),
 (u'We', 0.7552806889430156),
 (u'This', 0.7749201702459683),
 (u'class', 0.7913652190100225),
 (u'would', 0.8149828303863013),
 (u'make', 0.8239845109910496),
 (u'many', 0.8273389184929604),
```

Least Common

```
idf[:-50:-1]
```

```
[(u'beer', 10.378594025517652),
 (u'worsen', 10.378594025517652),
 (u'theorist', 10.378594025517652),
 (u'Beneath', 10.378594025517652),
 (u'.how', 10.378594025517652),
 (u'unchanged', 10.378594025517652),
 (u'lessons-', 10.378594025517652),
 (u'on-stage', 10.378594025517652),
 (u'interactiveness', 10.378594025517652),
 (u'GoogleEarth', 10.378594025517652),
 (u'peers\u2019', 10.378594025517652),
 (u'pre-schools', 10.378594025517652),
 (u'PER', 10.378594025517652),
 (u'Davies', 10.378594025517652),
 (u'Spalding', 10.378594025517652),
 (u'7:15am', 10.378594025517652),
 (u'geneticists', 10.378594025517652),
 (u'20-year-old', 10.378594025517652),
 (u'inservice', 10.378594025517652),
 (u'Conquering', 10.378594025517652),
```

```python
top_n = 10
summary = bag_of_words.map(lambda x: map(lambda idx: broadcast_idf.value[idx][0], np.argsort(x)[::-1][:top_n]))
```

```python
summary.take(15)
```

```
[[u'science',
  u'Outreach',
  u'17-21',
  u'one-year',
  u'resource',
  u'magazine',
  u'periodical',
  u'http',
  u'York',
  u'competency'],
 [u'Worlds',
  u'Hidden',
  u'microscopes',
  u'cell',
  u'stressing',
  u'6th',
  u'single',
  u'cluster',
  u'intense',
  u'organisms'],
 [u'corner',
  u'Harlem',
  u'calming',
  u'rug',
  u'soft',
  u'world.In',
  u'began',
  u'populate',
  u'putting',
  u'stain'],
 [u'Music',
  u'music',
  u'Appreciation',
```

Summarization

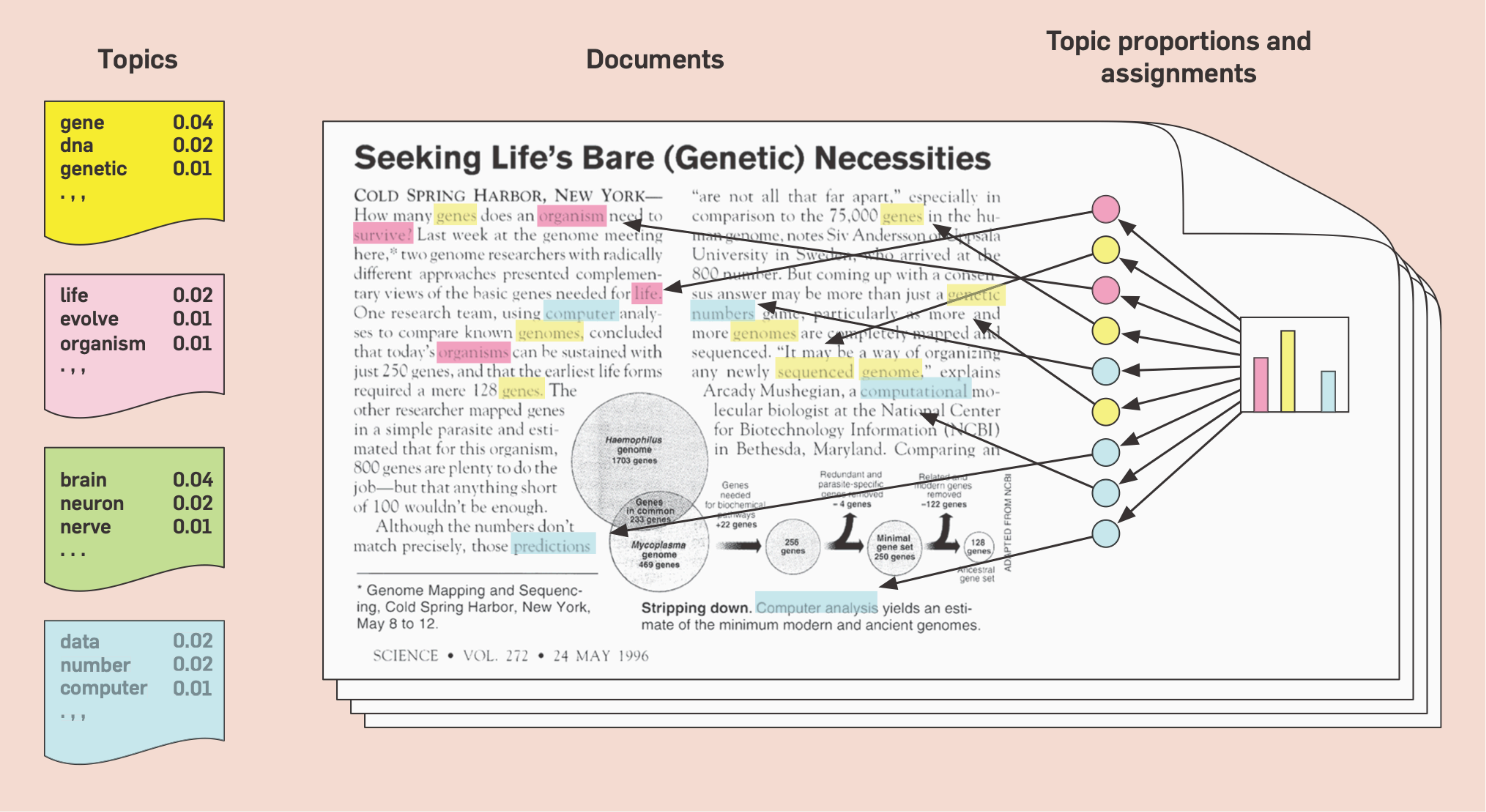
Understanding Corpuses with Topic Modelling

Latent Dirichlet Allocation*

- Generative probabilistic model for collections of discrete data
- “Killer application” has been topic modeling for text
- Unsupervised technique that explains sets of observed data as being generated from unobserved groups

** Equivalent to probabilistic latent semantic analysis (matrix factorization)*

Latent Dirichlet Allocation



Source: http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

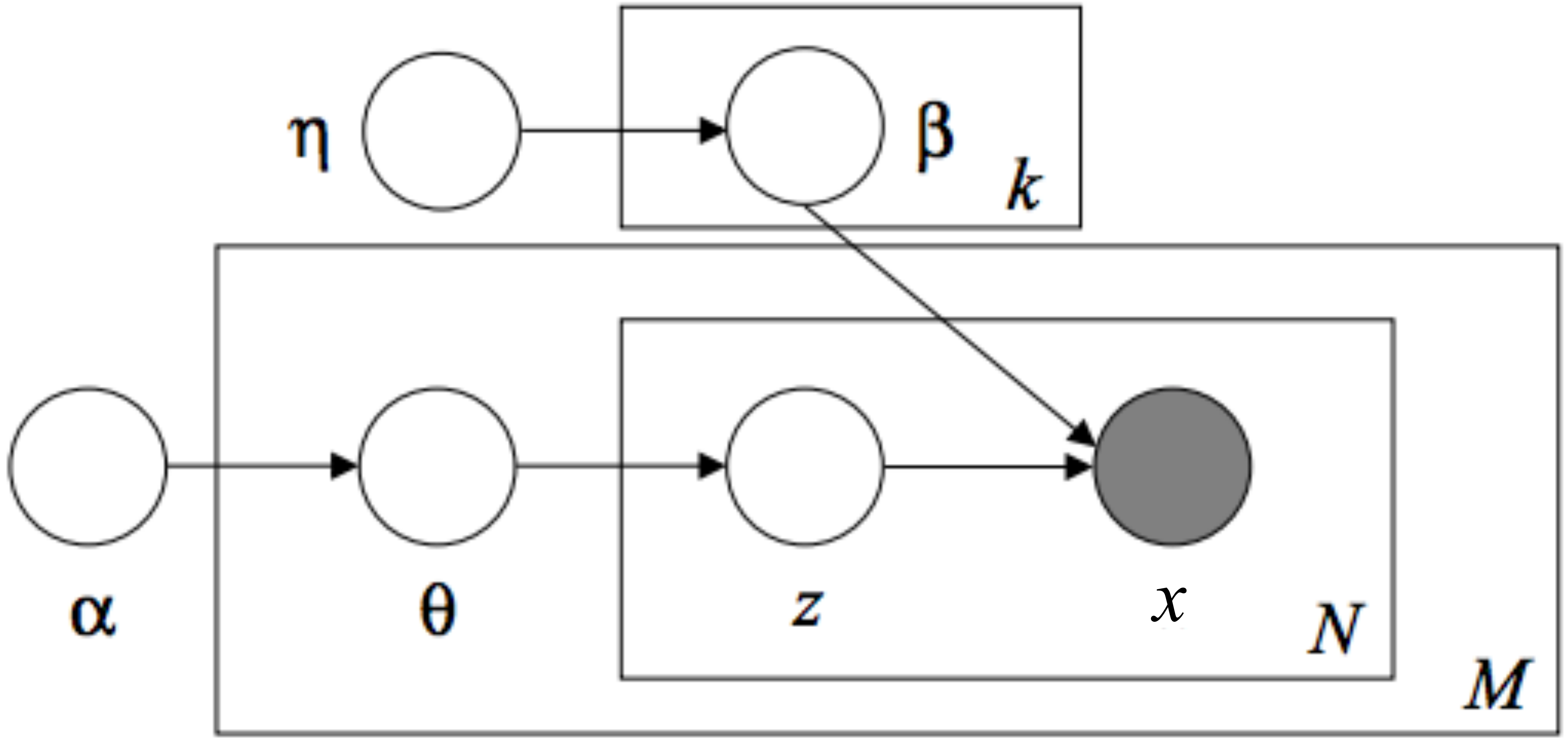
Generative Story

- Someone sits down to write a document.
- Assume that the (observed) words in each document are generated from a finite number of (unobserved) topics
- For each document:
  - Writer decides (mixture of) topics to write about: $z_{i,j} \sim Mulitnomial(\theta_i)$
  - Chooses words based on topic-word distribution: $x_{i,j} \sim Mulitnomial(\beta_{z_{i,j}})$

And we want to infer these

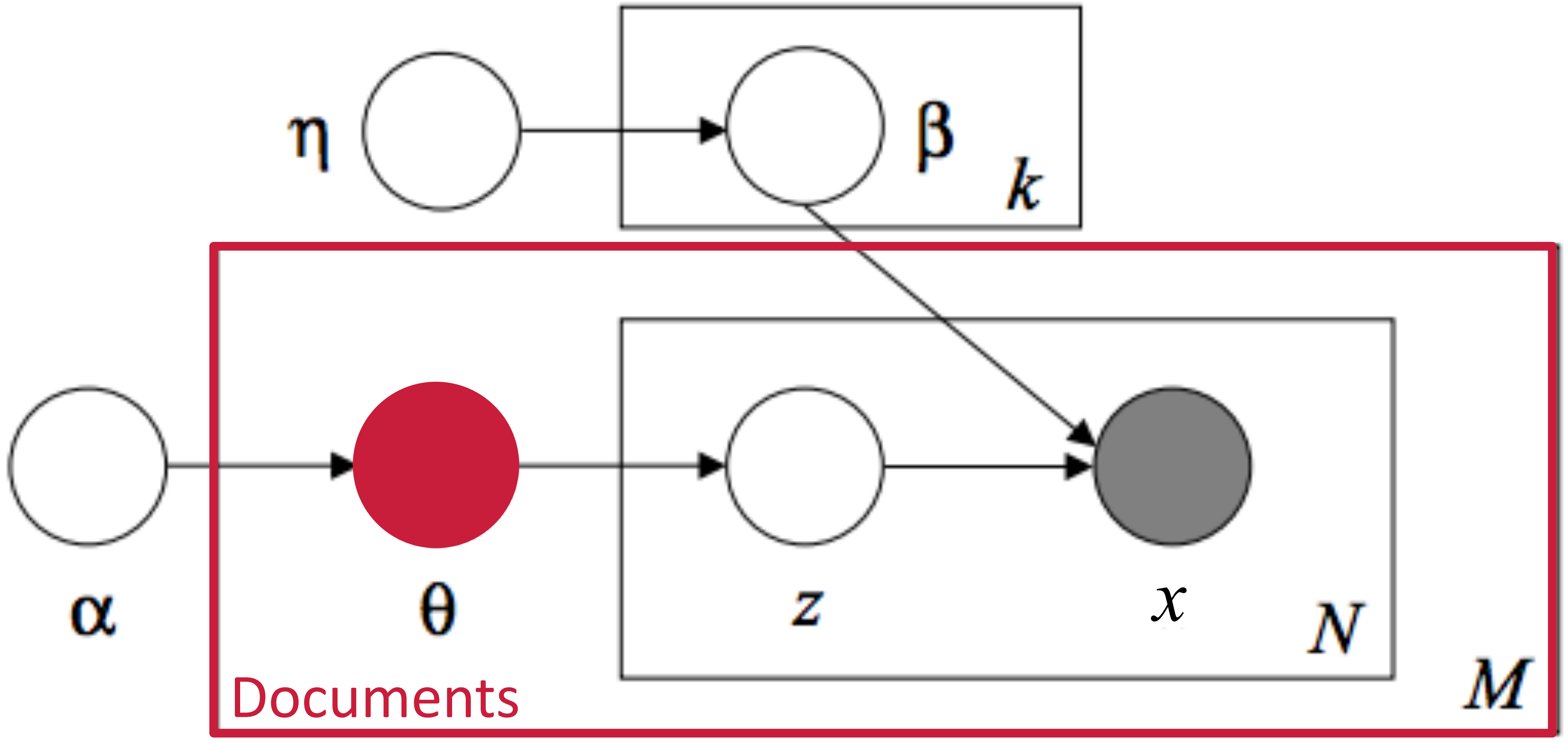
This is all we can see

Hierarchical Bayesian Model



Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent dirichlet allocation.

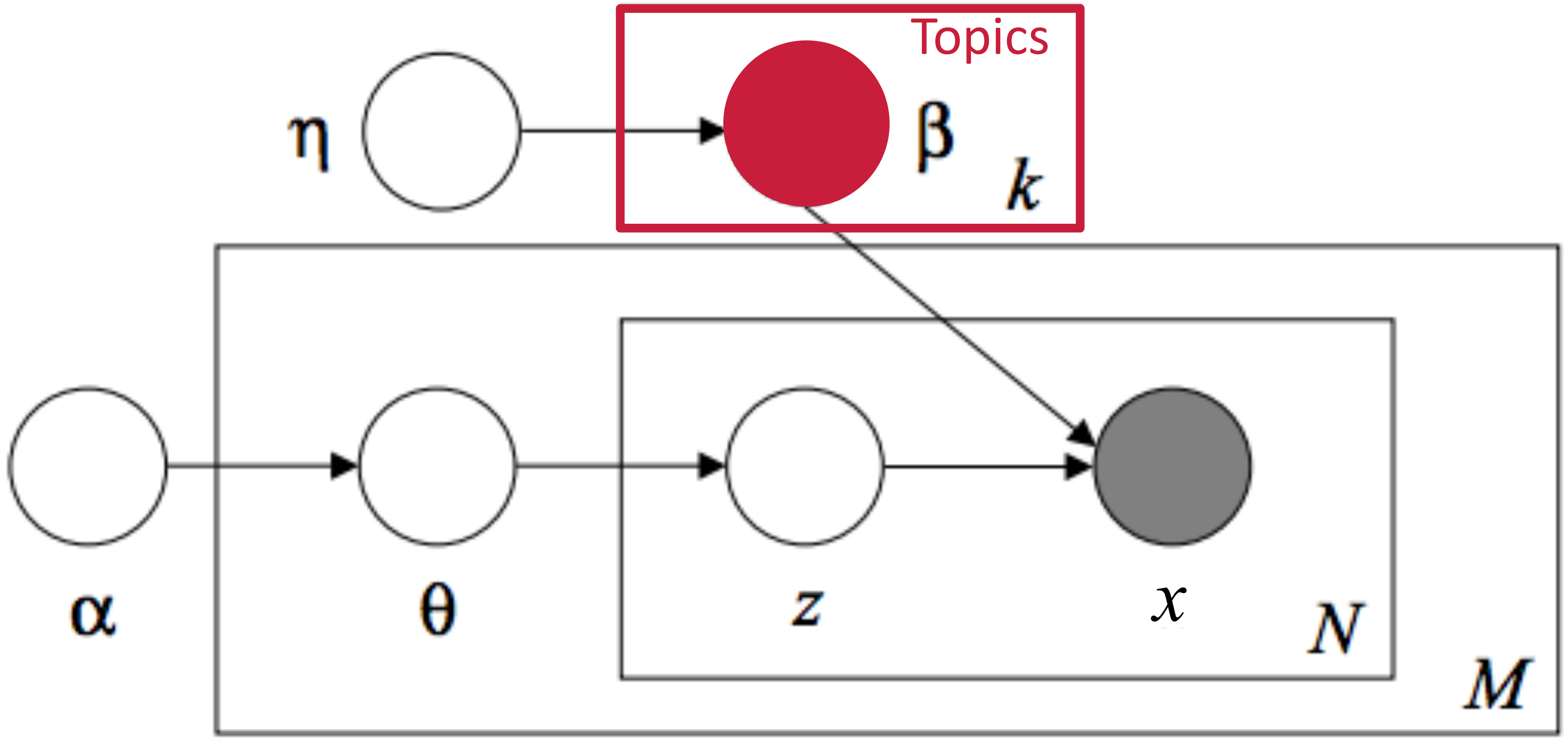
Hierarchical Bayesian Model



- Choose $\theta_i \sim Dir(\alpha)$ where $i \in \{1,\ldots,M\}$

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent dirichlet allocation.

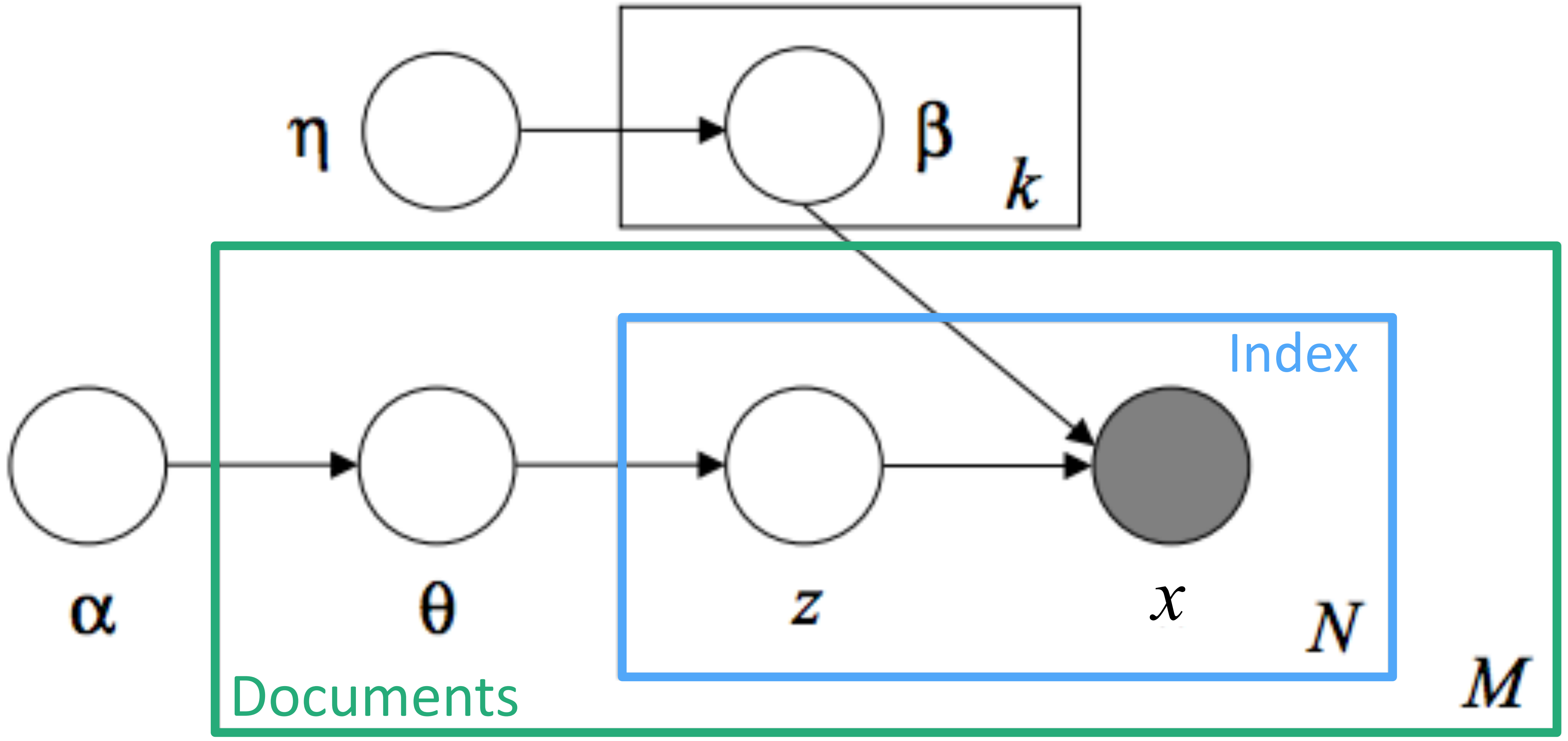
Hierarchical Bayesian Model



- Choose $\theta_i \sim Dir(\alpha)$ where $i \in \{1,\ldots,M\}$
- Choose $\beta_k \sim Dir(\eta)$ where $k \in \{1,\ldots,K\}$

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent dirichlet allocation.

Hierarchical Bayesian Model



Priors

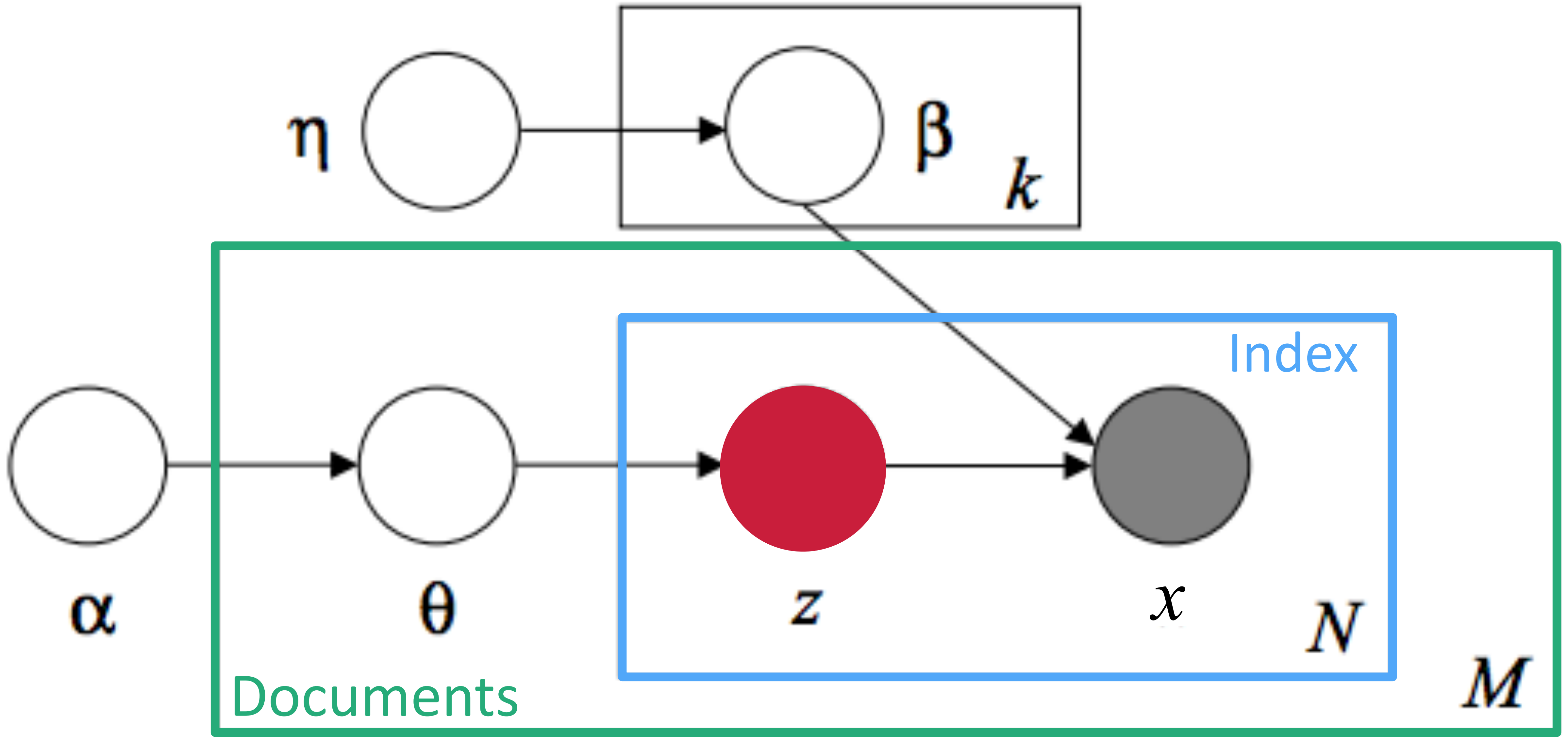
- Choose $\theta_i \sim Dir(\alpha)$ where $i \in \{1,\ldots,M\}$
- Choose $\beta_k \sim Dir(\eta)$ where $k \in \{1,\ldots,K\}$

Process

- For each position i in document j :

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent dirichlet allocation.

Hierarchical Bayesian Model



Priors

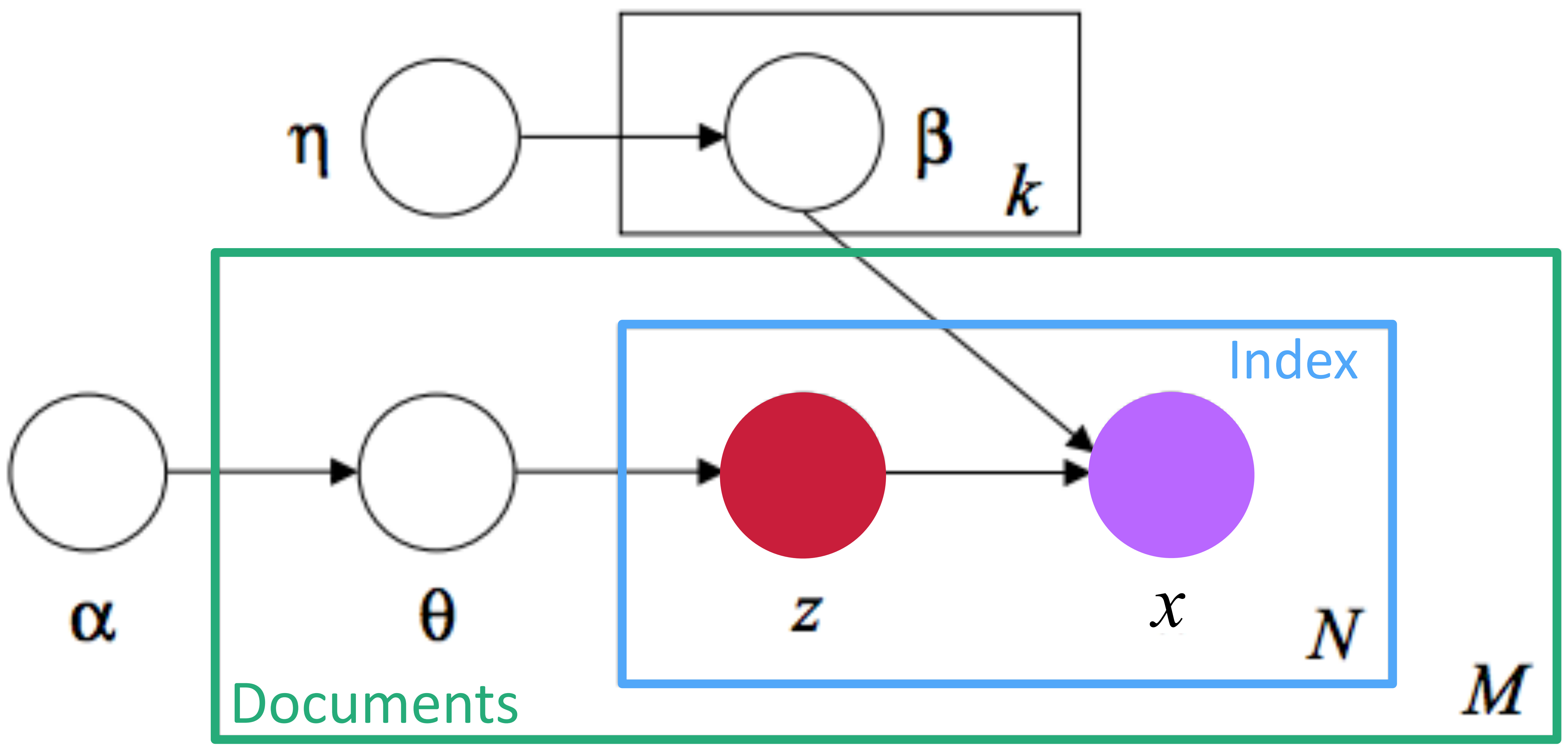
- Choose $\theta_i \sim Dir(\alpha)$ where $i \in \{1,\ldots,M\}$
- Choose $\beta_k \sim Dir(\eta)$ where $k \in \{1,\ldots,K\}$

Process

- For each position i in document j :
  - Choose a topic $z_{i,j} \sim Mulitnomial(\theta_i)$

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent dirichlet allocation.

Hierarchical Bayesian Model



Priors

- Choose $\theta_i \sim Dir(\alpha)$ where $i \in \{1,\ldots,M\}$
- Choose $\beta_k \sim Dir(\eta)$ where $k \in \{1,\ldots,K\}$

Process

- For each position i in document j :
  - Choose a topic $z_{i,j} \sim Mulitnomial(\theta_i)$
  - Choose a word $x_{i,j} \sim Mulitnomial(\beta_{z_{i,j}})$

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent dirichlet allocation.

Inference

$$p(z,\theta,\beta) = \frac{p(z,\theta,\beta \mid \alpha,\eta)}{p(x \mid \alpha,\eta)}$$

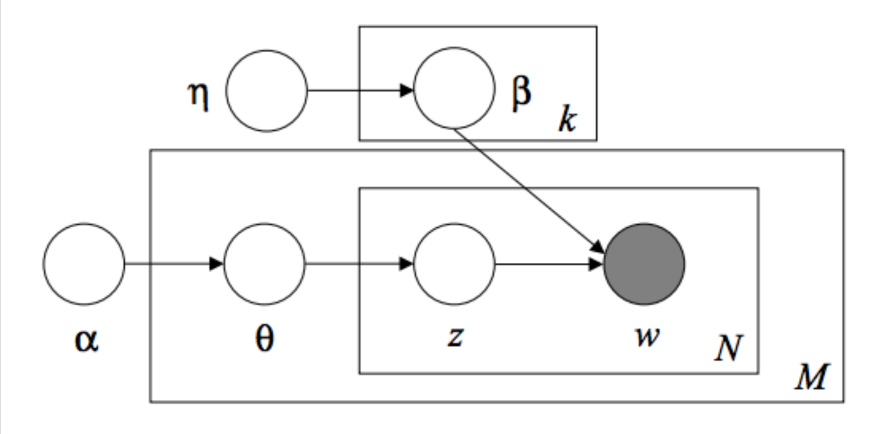
Posterior

- Estimate with online variational Bayes (with batch updates) maximizing the ELBO

Box’s Loop (the Blei method)

Domain Knowledge and Assumptions

Observed Data



Inference

Criticize and Evaluate

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models.

whoami

Consulting

- Human-in-the-loop machine learning + MLOps

inquiries@jonathan.industries

Appendix and References

inquiries@jonathan.industries

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
- Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, *1*, 203-232.
- Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (pp. 1385-1392).
- Johnson, M. J., & Willsky, A. S. (2013). Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, *14*(Feb), 673-701.
- Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.
- Fox, E., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2009). Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems* (pp. 457-464).

inquiries@jonathan.industries