



# Clustering

Data Science: Jordan Boyd-Graber  
University of Maryland

SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

## Clustering

Questions:

- how do we fit clusters?
- how many clusters should we use?
- how should we evaluate model fit?

## K-Means

How do we fit the clusters?

- simplest method: K-means
- requires: real-valued data
- idea:
  - pick  $K$  initial cluster means
  - associate all points closest to mean  $k$  with cluster  $k$
  - use points in cluster  $k$  to update mean for that cluster
  - re-associate points closest to new mean for  $k$  with cluster  $k$
  - use new points in cluster  $k$  to update mean for that cluster
  - ...
  - stop when no change between updates

## K-Means

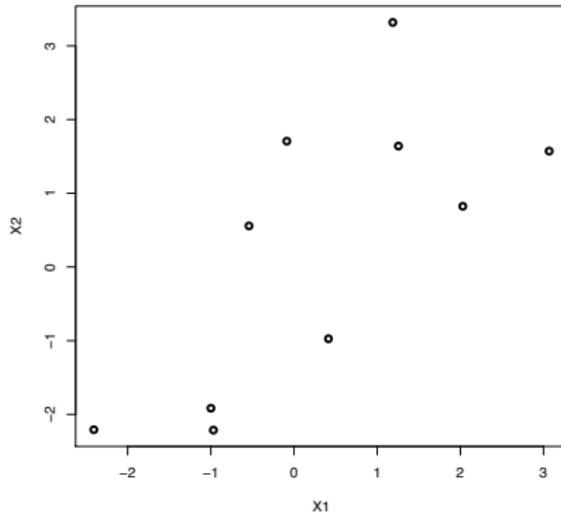
Animation at:

<http://shabal.in/visuals/kmeans/1.html>

## K-Means: Example

Data:

$x_1$	$x_2$
0.4	-1.0
-1.0	-2.2
-2.4	-2.2
-1.0	-1.9
-0.5	0.6
-0.1	1.7
1.2	3.3
3.1	1.6
1.3	1.6
2.0	0.8



## Why topic models?



- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes

## Why topic models?



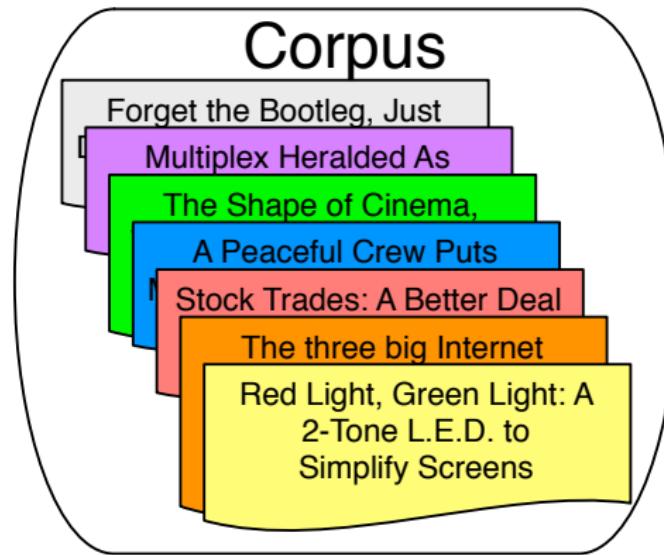
- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

## Roadmap

- What are topic models
- How to know if you have good topic model
- How to go from raw data to topics

## Conceptual Approach

From an **input corpus** and number of topics  $K \rightarrow$  words to topics



## Conceptual Approach

From an input corpus and number of topics  $K \rightarrow$  **words to topics**

TOPIC 1

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

TOPIC 2

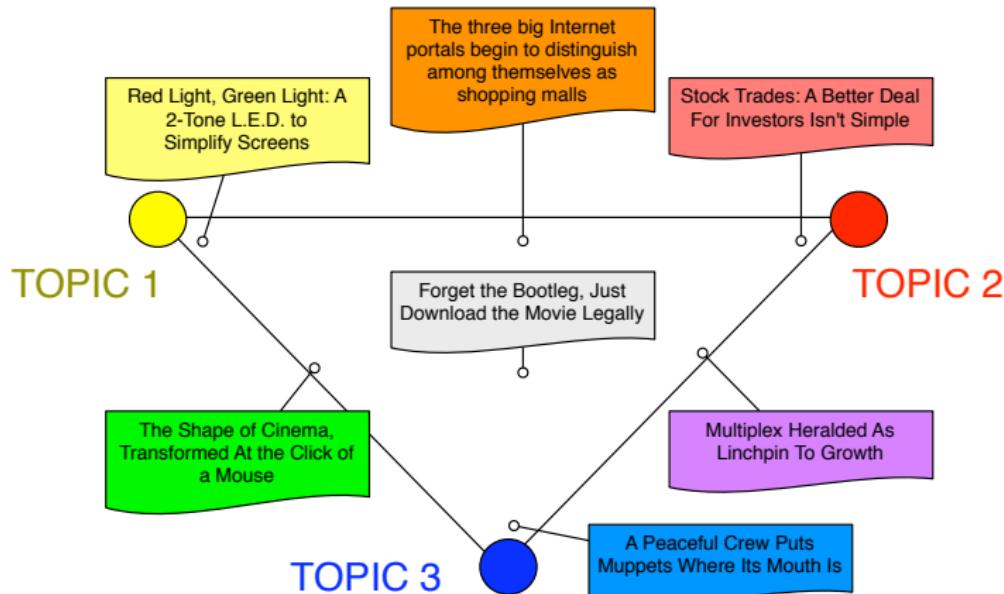
sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

TOPIC 3

play, film,  
movie, theater,  
production,  
star, director,  
stage

## Conceptual Approach

- For each document, what topics are expressed by that document?



## Topics from Science

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

## Why should you care?

- Neat way to explore / understand corpus collections
  - E-discovery
  - Social media
  - Scientific data
- NLP Applications
  - Word Sense Disambiguation
  - Discourse Segmentation
  - Machine Translation
- Psychology: word meaning, polysemy
- Inference is (relatively) simple

## Matrix Factorization Approach

$$\begin{bmatrix} M \times K \\ \text{Topic Assignment} \end{bmatrix} \times \begin{bmatrix} K \times V \\ \text{Topics} \end{bmatrix} \approx \begin{bmatrix} M \times V \\ \text{Dataset} \end{bmatrix}$$

**K** Number of topics

**M** Number of documents

**V** Size of vocabulary

## Matrix Factorization Approach

$$\begin{bmatrix} M \times K \\ \text{Topic Assignment} \end{bmatrix} \times \begin{bmatrix} K \times V \\ \text{Topics} \end{bmatrix} \approx \begin{bmatrix} M \times V \\ \text{Dataset} \end{bmatrix}$$

**K** Number of topics

**M** Number of documents

**V** Size of vocabulary

- If you use singular value decomposition (SVD), this technique is called latent semantic analysis.
- Popular in information retrieval.

## Alternative: Generative Model

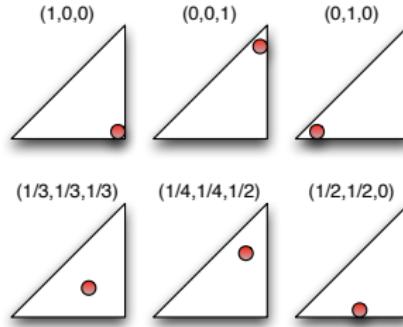
- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference

## Alternative: Generative Model

- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference
- Blei, Ng, Jordan. Latent **Dirichlet** Allocation. JMLR, 2003.

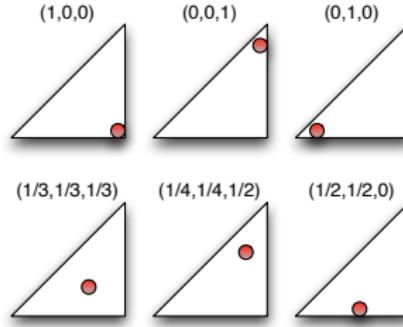
## Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



## Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



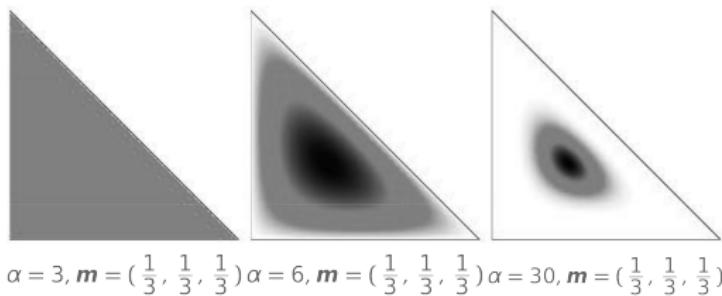
- Come from a Dirichlet distribution

## Dirichlet Distribution

$$P(\mathbf{p} \mid \alpha \mathbf{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

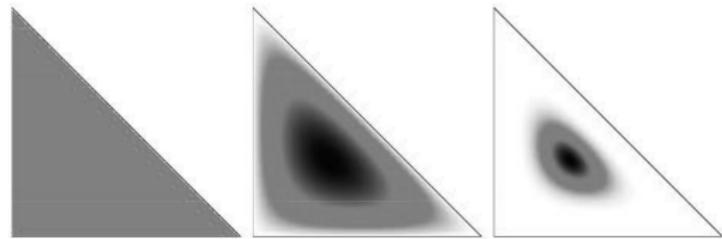
## Dirichlet Distribution

$$P(\mathbf{p} | \alpha \mathbf{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

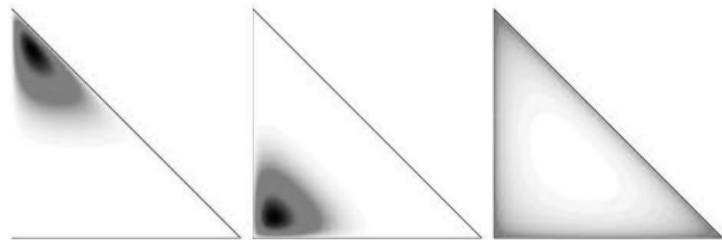


## Dirichlet Distribution

$$P(\mathbf{p} | \alpha \mathbf{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

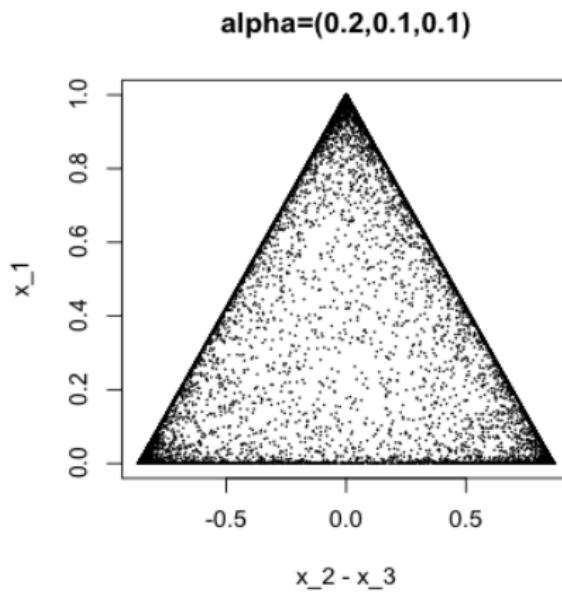


$$\alpha = 3, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \quad \alpha = 6, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \quad \alpha = 30, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$



$$\alpha = 14, \mathbf{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7}) \quad \alpha = 14, \mathbf{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7}) \quad \alpha = 2.7, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$

## Dirichlet Distribution



## Dirichlet Distribution

- If  $\vec{\phi} \sim \text{Dir}((\alpha))$ ,  $\vec{w} \sim \text{Mult}((\phi))$ , and  $n_k = |\{w_i : w_i = k\}|$  then

$$p(\phi | \alpha, \vec{w}) \propto p(\vec{w} | \phi) p(\phi | \alpha) \quad (1)$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \quad (2)$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \quad (3)$$

- Conjugacy: this **posterior** has the same form as the **prior**

## Dirichlet Distribution

- If  $\vec{\phi} \sim \text{Dir}((\alpha))$ ,  $\vec{w} \sim \text{Mult}((\phi))$ , and  $n_k = |\{w_i : w_i = k\}|$  then

$$p(\phi | \alpha, \vec{w}) \propto p(\vec{w} | \phi) p(\phi | \alpha) \quad (1)$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \quad (2)$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \quad (3)$$

- Conjugacy: this **posterior** has the same form as the **prior**

## Generative Model

TOPIC 1

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

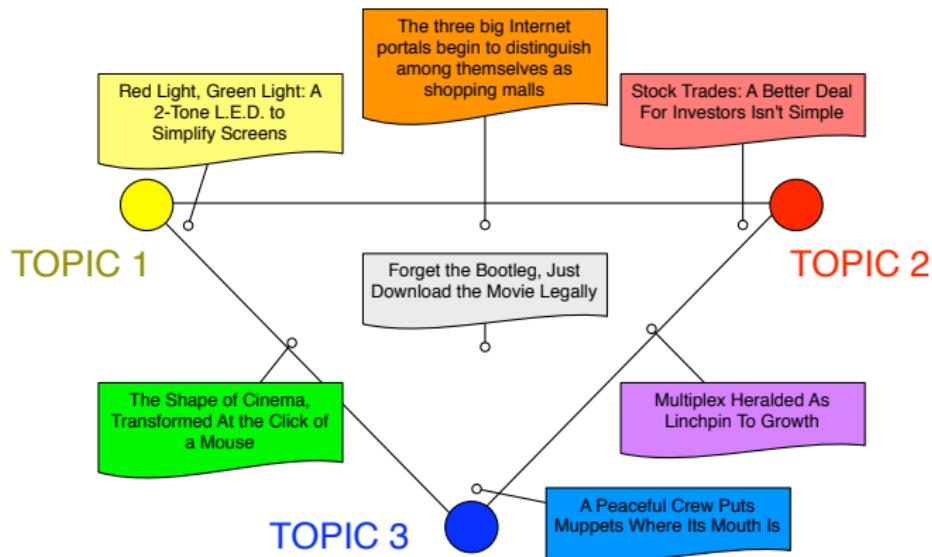
TOPIC 2

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

TOPIC 3

play, film,  
movie, theater,  
production,  
star, director,  
stage

## Generative Model



## Generative Model

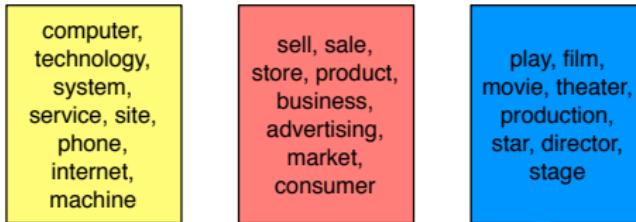
computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

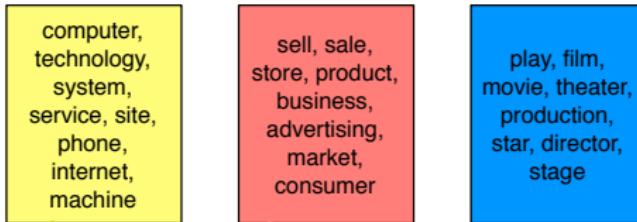
## Generative Model



Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...



## Generative Model



Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Generative Model

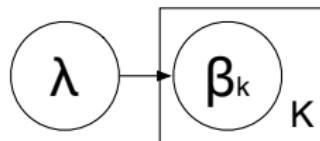
computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

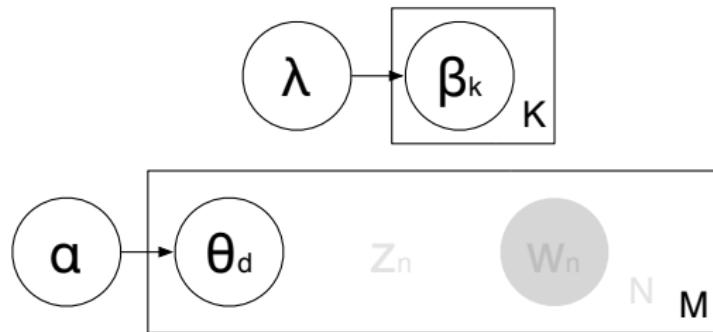
## Generative Model Approach



$\alpha$        $\theta_d$        $z_n$        $w_n$        $N$        $M$

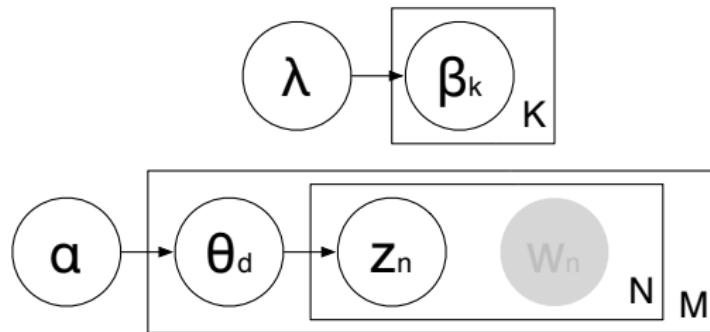
- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$

## Generative Model Approach



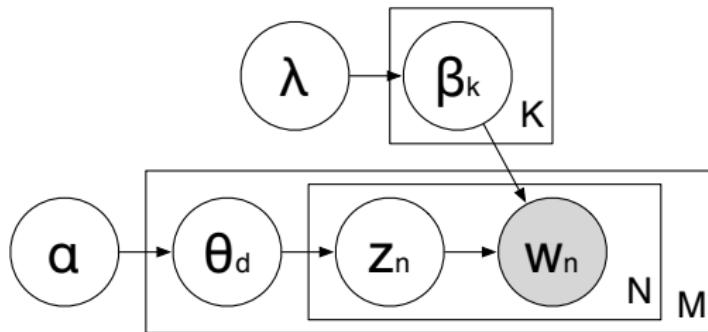
- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$

## Generative Model Approach



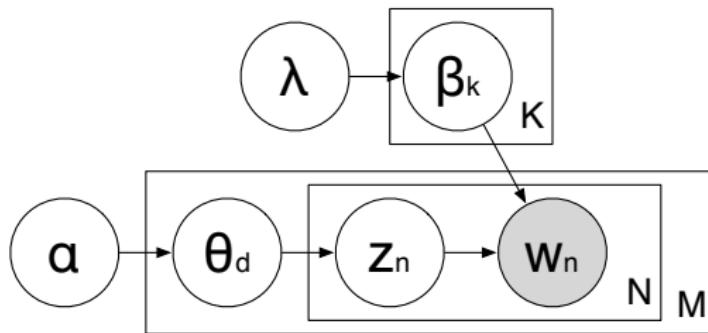
- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$
- For each word position  $n \in \{1, \dots, N\}$ , select a hidden topic  $z_n$  from the multinomial distribution parameterized by  $\theta$ .

## Generative Model Approach



- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$
- For each word position  $n \in \{1, \dots, N\}$ , select a hidden topic  $z_n$  from the multinomial distribution parameterized by  $\theta$ .
- Choose the observed word  $w_n$  from the distribution  $\beta_{z_n}$ .

## Generative Model Approach



- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$
- For each word position  $n \in \{1, \dots, N\}$ , select a hidden topic  $z_n$  from the multinomial distribution parameterized by  $\theta$ .
- Choose the observed word  $w_n$  from the distribution  $\beta_{z_n}$ .

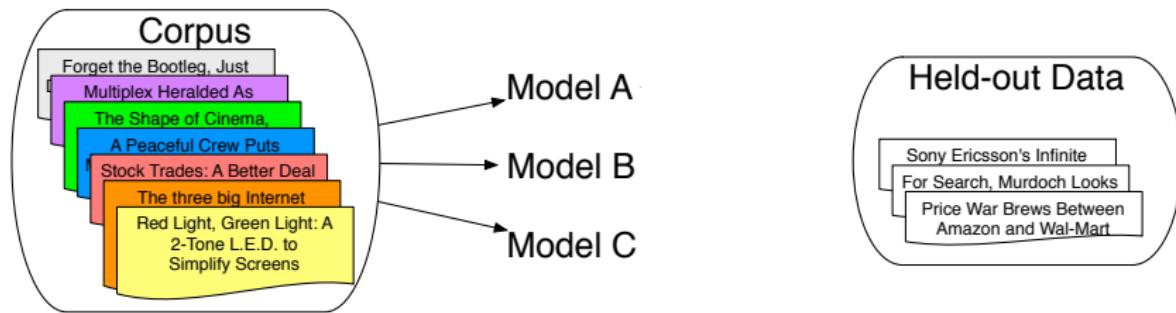
## Topic Models: What's Important

- Topic models
  - Topics to word types
  - Documents to topics
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA), a fully Bayesian version of pLSI, probabilistic version of LSA

## Topic Models: What's Important

- Topic models (latent variables)
  - Topics to word types
  - Documents to topics
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA), a fully Bayesian version of pLSI, probabilistic version of LSA

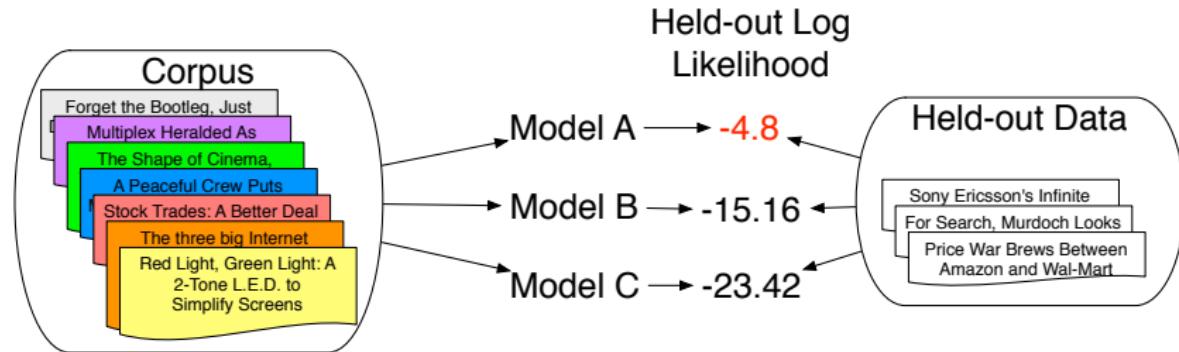
## Evaluation



$$P(\mathbf{w} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u}) = \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u})$$

How you compute it is important too (Wallach et al. 2009)

## Evaluation



Measures predictive power, not what the topics are

$$P(\mathbf{w} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u}) = \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u})$$

How you compute it is important too (Wallach et al. 2009)

## Word Intrusion

### TOPIC 1

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

### TOPIC 2

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

### TOPIC 3

play, film,  
movie, theater,  
production,  
star, director,  
stage

## Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

## Word Intrusion

1. Take the highest probability words from a topic

### Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

### Topic with Intruder

dog, cat, **apple**, horse, pig, cow

## Word Intrusion

1. Take the highest probability words from a topic

### Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

### Topic with Intruder

dog, cat, apple, horse, pig, cow

3. We ask users to find the word that doesn't belong

### Hypothesis

If the topics are interpretable, users will consistently choose true intruder

## Word Intrusion

1 / 10

crash      accident      board      agency      tibetan      safety

2 / 10

commercial      network      television      advertising      viewer      layoff

3 / 10

arrest      crime      inmate      pitcher      prison      death

4 / 10

hospital      doctor      health      care      medical      tradition

## Word Intrusion

**1 / 10****Reveal additional response**

crash

accident

board

agency

tibetan

safety

**2 / 10**

commercial

network

television

advertising

viewer

layoff

**3 / 10**

arrest

crime

inmate

pitcher

prison

death

**4 / 10**

hospital

doctor

health

care

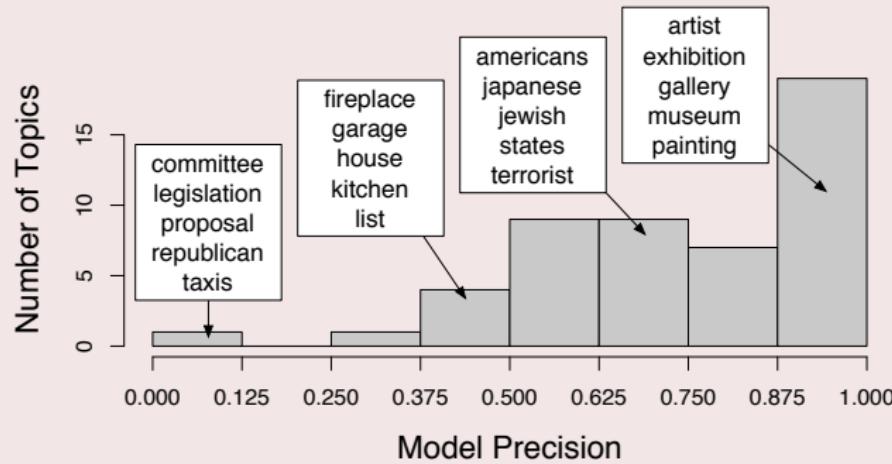
medical

tradition

- Order of words was shuffled
- Which intruder was selected varied
- Model precision: percentage of users who clicked on intruder

## Word Intrusion: Which Topics are Interpretable?

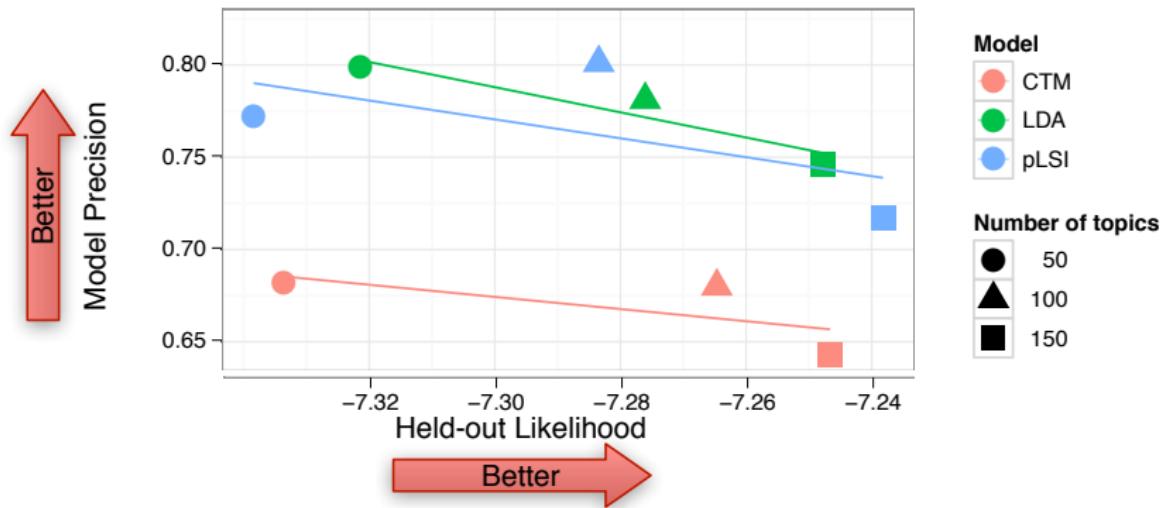
New York Times, 50 LDA Topics



Model Precision: percentage of correct intruders found

## Interpretability and Likelihood

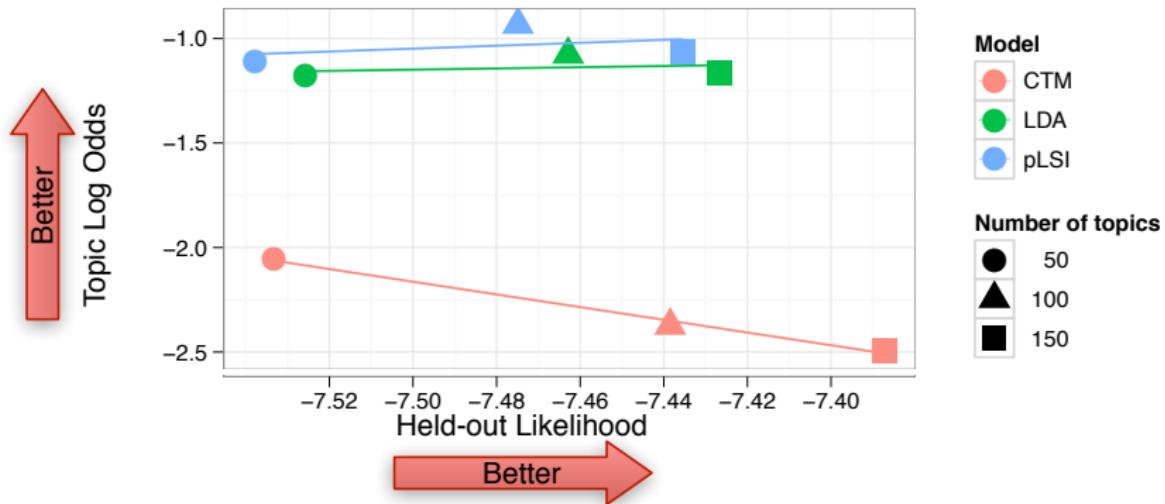
Model Precision on New York Times



within a model, higher likelihood  $\neq$  higher interpretability

## Interpretability and Likelihood

Topic Log Odds on Wikipedia



across models, higher likelihood  $\neq$  higher interpretability

## Evaluation Takeaway

- Measure what you care about
- If you care about prediction, likelihood is good
- If you care about a particular task, measure that

## Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \quad (4)$$

## Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \quad (4)$$

- Here, latent variables are topic assignments  $z$  and topics  $\theta$ .  $X$  is the words (divided into documents), and  $\Theta$  are hyperparameters to Dirichlet distributions:  $\alpha$  for topic proportion,  $\lambda$  for topics.

$$p(\vec{z}, \vec{\beta}, \vec{\theta} | \vec{w}, \alpha, \lambda) \quad (5)$$

## Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \quad (4)$$

- Here, latent variables are topic assignments  $z$  and topics  $\theta$ .  $X$  is the words (divided into documents), and  $\Theta$  are hyperparameters to Dirichlet distributions:  $\alpha$  for topic proportion,  $\lambda$  for topics.

$$p(\vec{z}, \vec{\beta}, \vec{\theta} | \vec{w}, \alpha, \lambda) \quad (5)$$

$$p(\vec{w}, \vec{z}, \vec{\theta}, \vec{\beta} | \alpha, \lambda) = \\ \prod_k p(\beta_k | \lambda) \prod_d p(\theta_d | \alpha) \prod_n p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{z_{d,n}})$$

## Gibbs Sampling

- A form of Markov Chain Monte Carlo
- Chain is a sequence of random variable states
- Given a state  $\{z_1, \dots, z_N\}$  given certain technical conditions, drawing  $z_k \sim p(z_1, \dots, z_{k-1}, z_{k+1}, \dots, z_N | X, \Theta)$  for all  $k$  (repeatedly) results in a Markov Chain whose stationary distribution is the posterior.
- For notational convenience, call  $\vec{z}$  with  $z_{d,n}$  removed  $\vec{z}_{-d,n}$

## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

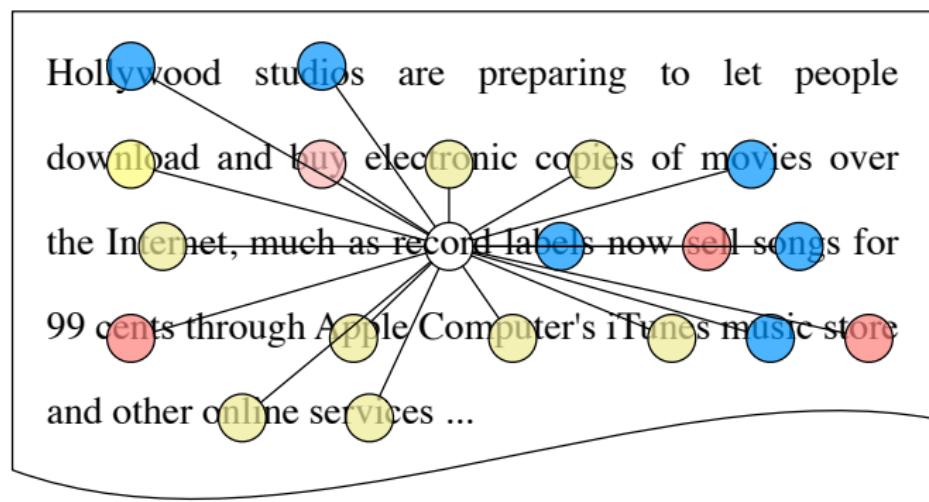
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage



## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

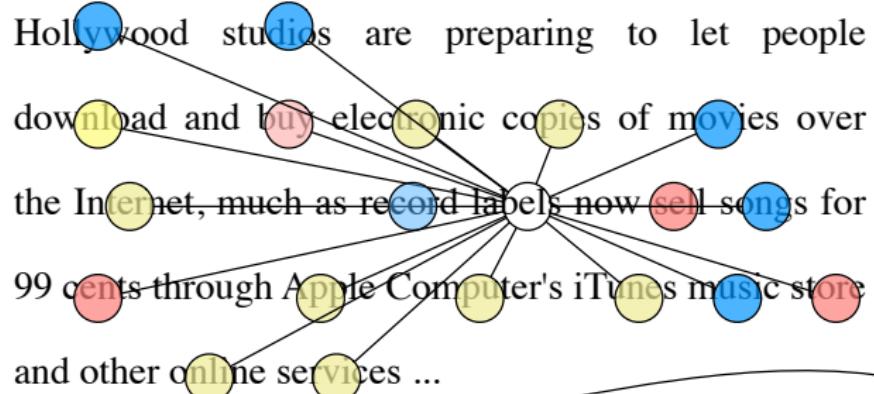
## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...



## Inference

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Gibbs Sampling

- For LDA, we will sample the topic assignments
- Thus, we want:

$$p(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda) = \frac{p(z_{d,n} = k, \vec{z}_{-d,n} | \vec{w}, \alpha, \lambda)}{p(\vec{z}_{-d,n} | \vec{w}, \alpha, \lambda)}$$

## Gibbs Sampling

- For LDA, we will sample the topic assignments
- Thus, we want:

$$p(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda) = \frac{p(z_{d,n} = k, \vec{z}_{-d,n} | \vec{w}, \alpha, \lambda)}{p(\vec{z}_{-d,n} | \vec{w}, \alpha, \lambda)}$$

- The topics and per-document topic proportions are integrated out / marginalized
- Let  $n_{d,i}$  be the number of words taking topic  $i$  in document  $d$ . Let  $v_{k,w}$  be the number of times word  $w$  is used in topic  $k$ .

$$= \frac{\int_{\theta_d} \left( \prod_{i \neq k} \theta_d^{\alpha_i + n_{d,i}-1} \right) \theta_d^{\alpha_k + n_{d,i}} d\theta_d \int_{\beta_k} \left( \prod_{i \neq w_{d,n}} \beta_{k,i}^{\lambda_i + v_{k,i}-1} \right) \beta_{k,w_{d,n}}^{\lambda_i + v_{k,i}} d\beta_k}{\int_{\theta_d} \left( \prod_i \theta_d^{\alpha_i + n_{d,i}-1} \right) d\theta_d \int_{\beta_k} \left( \prod_i \beta_{k,i}^{\lambda_i + v_{k,i}-1} \right) d\beta_k}$$

## Gibbs Sampling

- For LDA, we will sample the topic assignments
- The topics and per-document topic proportions are integrated out / marginalized / Rao-Blackwellized
- Thus, we want:

$$p(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda) = \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{\nu_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i \nu_{k,i} + \lambda_i}$$

## Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left( \prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) d\beta_k = \frac{\prod_i^V \Gamma(\beta_i + v_{k,i})}{\Gamma(\sum_i^V \beta_i + v_{k,i})}$$

## Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left( \prod_i \beta_{k,i}^{\lambda_i + v_{k,i}-1} \right) d\beta_k = \frac{\prod_i^V \Gamma(\beta_i + v_{k,i})}{\Gamma(\sum_i^V \beta_i + v_{k,i})}$$

- So we can simplify

$$\frac{\int_{\theta_d} \left( \prod_{i \neq k} \theta_d^{\alpha_i + n_{d,i}-1} \right) \theta_d^{\alpha_k + n_{d,k}} d\theta_d \int_{\beta_k} \left( \prod_{i \neq w_{d,n}} \beta_{k,i}^{\lambda_i + v_{k,i}-1} \right) \beta_{k,w_{d,n}}^{\lambda_k + v_{k,i}} d\beta_k}{\int_{\theta_d} \left( \prod_i \theta_d^{\alpha_i + n_{d,i}-1} \right) d\theta_d \int_{\beta_k} \left( \prod_i \beta_{k,i}^{\lambda_i + v_{k,i}-1} \right) d\beta_k} = \\ \frac{\frac{\Gamma(\alpha_k + n_{d,k} + 1)}{\Gamma(\sum_i^K \alpha_i + n_{d,i} + 1)} \prod_{i \neq k}^K \Gamma(\alpha_i + n_{d,i})}{\frac{\prod_i^K \Gamma(\alpha_i + n_{d,i})}{\Gamma(\sum_i^K \alpha_i + n_{d,i})}} \frac{\frac{\Gamma(\lambda_{w_{d,n}} + v_{k,w_{d,n}} + 1)}{\Gamma(\sum_i^V \lambda_i + v_{k,i} + 1)} \prod_{i \neq w_{d,n}}^V \Gamma(\lambda_i + v_{k,i})}{\frac{\prod_i^V \Gamma(\lambda_i + v_{k,i})}{\Gamma(\sum_i^V \lambda_i + v_{k,i})}}$$

## Gamma Function Identity

$$z = \frac{\Gamma(z+1)}{\Gamma(z)} \quad (6)$$

$$\begin{aligned} & \frac{\Gamma(\alpha_k + n_{d,k} + 1)}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i} + 1\right)} \prod_{i \neq k}^K \Gamma(\alpha_k + n_{d,k}) \frac{\Gamma(\lambda_{w_{d,n}} + v_{k,w_{d,n}} + 1)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i} + 1\right)} \prod_{i \neq w_{d,n}}^V \Gamma(\lambda_k + v_{k,w_{d,n}}) \\ & \frac{\prod_i^K \Gamma(\alpha_i + n_{d,i})}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i}\right)} \frac{\prod_i^V \Gamma(\lambda_i + v_{k,i})}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i}\right)} \\ &= \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \end{aligned}$$

## Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \quad (7)$$

- Number of times document  $d$  uses topic  $k$
- Number of times topic  $k$  uses word type  $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic  $k$
- How much this topic likes word  $w_{d,n}$

## Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \quad (7)$$

- Number of times document  $d$  uses topic  $k$
- Number of times topic  $k$  uses word type  $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic  $k$
- How much this topic likes word  $w_{d,n}$

## Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \quad (7)$$

- Number of times document  $d$  uses topic  $k$
- Number of times topic  $k$  uses word type  $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic  $k$
- How much this topic likes word  $w_{d,n}$

## Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \quad (7)$$

- Number of times document  $d$  uses topic  $k$
- Number of times topic  $k$  uses word type  $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- **Dirichlet parameter for topic to word distribution**
- How much this document likes topic  $k$
- How much this topic likes word  $w_{d,n}$

## Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \quad (7)$$

- Number of times document  $d$  uses topic  $k$
- Number of times topic  $k$  uses word type  $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- **How much this document likes topic  $k$**
- How much this topic likes word  $w_{d,n}$

## Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \quad (7)$$

- Number of times document  $d$  uses topic  $k$
- Number of times topic  $k$  uses word type  $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic  $k$
- **How much this topic likes word  $w_{d,n}$**

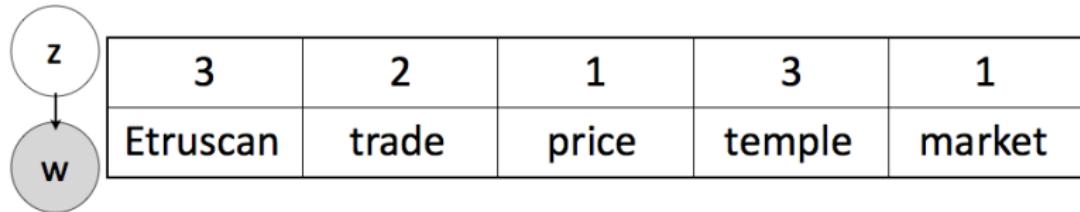
## Sample Document

Etruscan	trade	price	temple	market

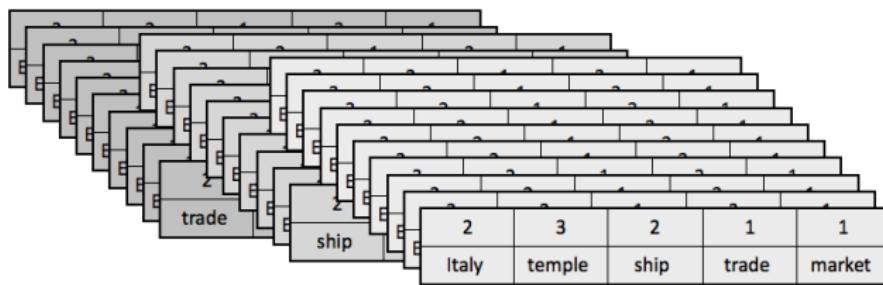
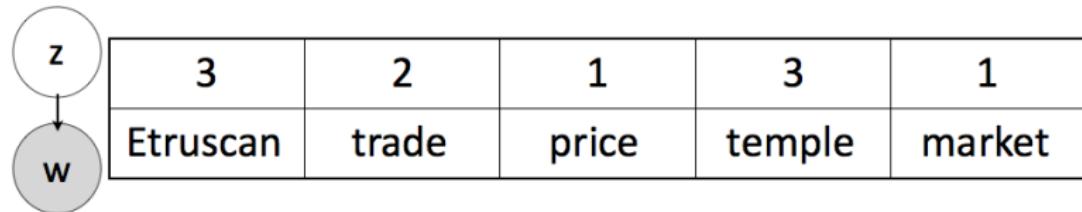
## Sample Document

Etruscan	trade	price	temple	market

## Randomly Assign Topics



## Randomly Assign Topics



## Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total  
counts  
from all  
docs



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

## Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total

	1	2	3
Etruscan	1	0	35
market	0	0	1

## Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

## Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total

	1	2	3
Etruscan	1	0	35
market	0	0	1

## Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

We want to sample this word ...

3	2	1	3	1
Etruscan	trade	price	temple	market



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

We want to sample this word ...

3	2	1	3	1
Etruscan	trade	price	temple	market

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

## Decrement its count

3	?	1	3	1
Etruscan	trade	price	temple	market

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



What is the conditional distribution for this topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

## Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

## Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1



Topic 2



Topic 3



## Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1

Sampling Equation

Topic 2

Topic 3

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

## Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1

Sampling Equation

Topic 2

Topic 3

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

## Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1



Topic 2



Topic 3



	1	2	3
trade	10	7	1

## Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1

Sampling Equation

Topic 2

Topic 3

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

## Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1

Sampling Equation

Topic 2

Topic 3

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

## Geometric interpretation

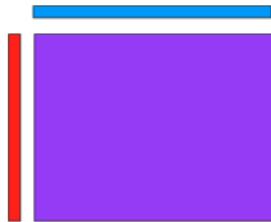
3	?	1	3	1
Etruscan	trade	price	temple	market



## Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1



Topic 2



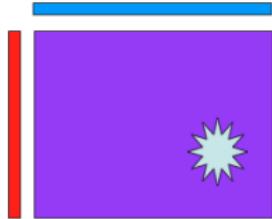
Topic 3



## Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1



Topic 2



Topic 3



## Update counts

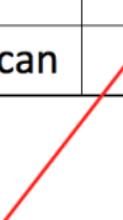
3	?	1	3	1
Etruscan	trade	price	temple	market

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



## Update counts

3	1	1	3	1
Etruscan	trade	price	temple	market



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

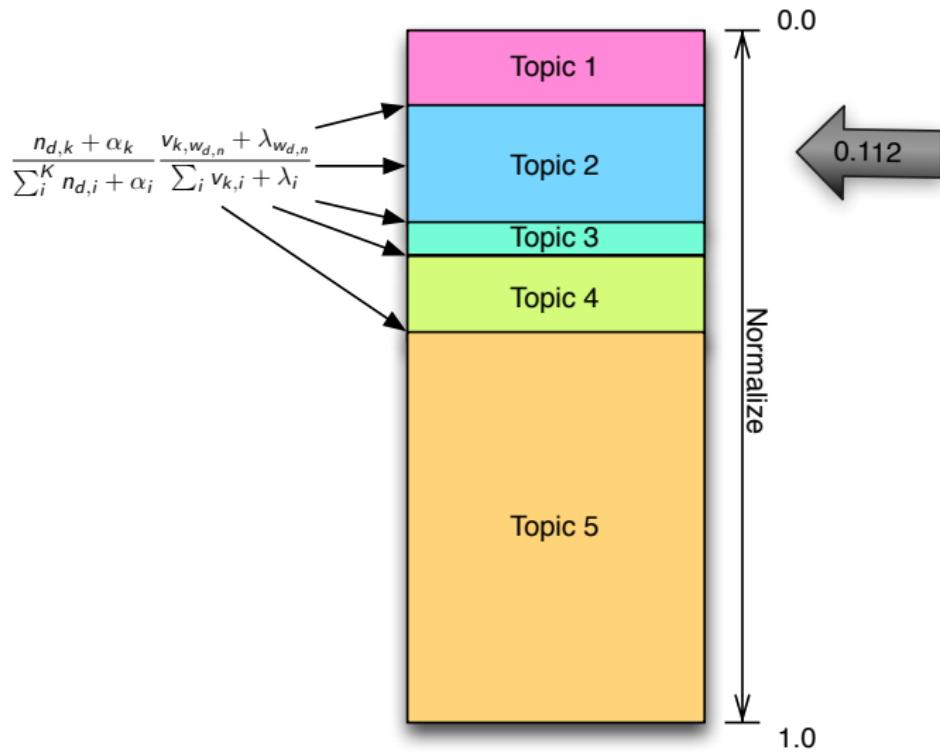


## Update counts

3	1	1	3	1
Etruscan	trade	price	temple	market



## Details: how to sample from a distribution



## Algorithm

1. For each iteration  $i$ :

1.1 For each document  $d$  and word  $n$  currently assigned to  $z_{old}$ :

1.1.1 Decrement  $n_{d,z_{old}}$  and  $v_{z_{old},w_{d,n}}$

1.1.2 Sample  $z_{new} = k$  with probability proportional to  $\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$

1.1.3 Increment  $n_{d,z_{new}}$  and  $v_{z_{new},w_{d,n}}$

## Implementation

### Algorithm

1. For each iteration  $i$ :

    1.1 For each document  $d$  and word  $n$  currently assigned to  $z_{old}$ :

        1.1.1 Decrement  $n_{d,z_{old}}$  and  $v_{z_{old},w_{d,n}}$

        1.1.2 Sample  $z_{new} = k$  with probability proportional to  $\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$

        1.1.3 Increment  $n_{d,z_{new}}$  and  $v_{z_{new},w_{d,n}}$

## Desiderata

- Hyperparameters: Sample them too (slice sampling)
- Initialization: Random
- Sampling: Until likelihood converges
- Lag / burn-in: Difference of opinion on this
- Number of chains: Should do more than one

## Available implementations

- Mallet (<http://mallet.cs.umass.edu>)
- LDAC (<http://www.cs.princeton.edu/~blei/lda-c>)
- Topicmod (<http://code.google.com/p/topicmod>)

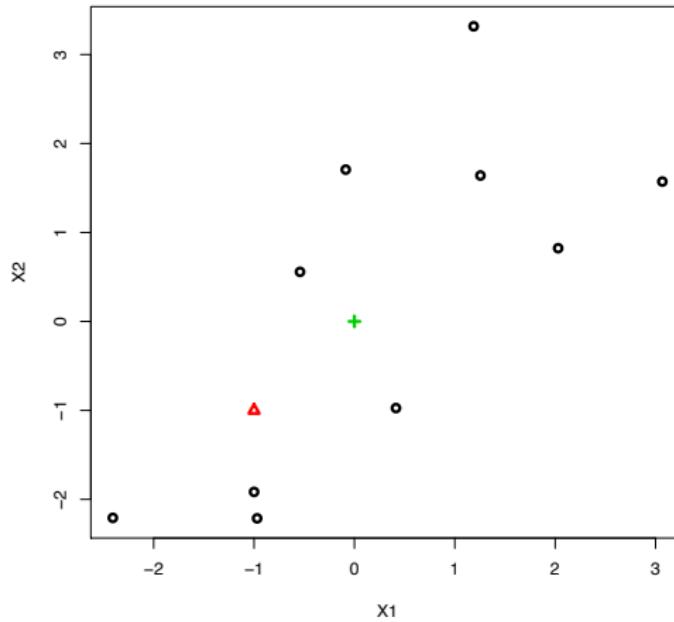
## Wrapup

- Topic Models: Tools to uncover themes in large document collections
- Another example of Gibbs Sampling
- In class: Gibbs sampling example

## K-Means: Example

Pick  $K$  centers (randomly):

$$(-1, -1) \text{ and } (0, 0)$$



## K-Means: Example

Calculate distance between points and those centers:

$x_1$	$x_2$	(-1, -1)	(0, 0)
0.4	-1.0	1.4	1.1
-1.0	-2.2	1.2	2.4
-2.4	-2.2	1.9	3.3
-1.0	-1.9	0.9	2.2
-0.5	0.6	1.6	0.8
-0.1	1.7	2.9	1.7
1.2	3.3	4.8	3.5
3.1	1.6	4.8	3.4
1.3	1.6	3.5	2.1
2.0	0.8	3.5	2.2

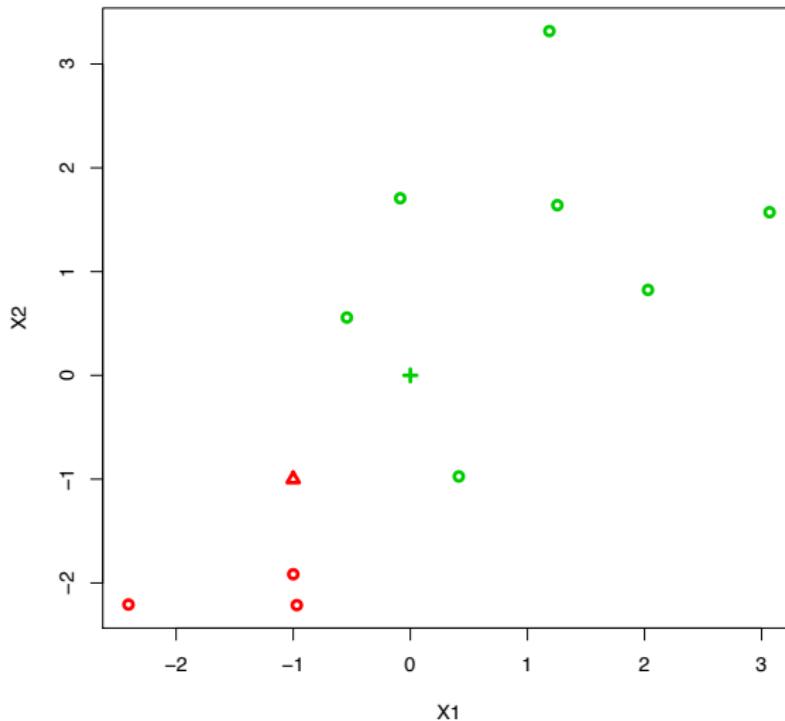
## K-Means: Example

Choose mean with smaller distance:

$x_1$	$x_2$	(-1, -1)	(0, 0)
0.4	-1.0	1.4	<b>1.1</b>
-1.0	-2.2	<b>1.2</b>	2.4
-2.4	-2.2	<b>1.9</b>	3.3
-1.0	-1.9	<b>0.9</b>	2.2
-0.5	0.6	1.6	<b>0.8</b>
-0.1	1.7	2.9	<b>1.7</b>
1.2	3.3	4.8	<b>3.5</b>
3.1	1.6	4.8	<b>3.4</b>
1.3	1.6	3.5	<b>2.1</b>
2.0	0.8	3.5	<b>2.2</b>

## K-Means: Example

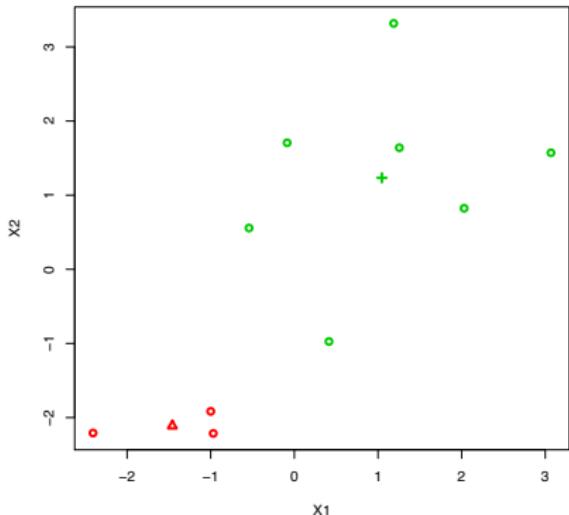
New clusters:



## K-Means: Example

Refit means for each cluster:

- cluster 1:  $(-1.0, -2.2)$ ,  
 $(-2.4, -2.2)$ ,  $(-1.0, -1.9)$
- new mean:  $(-1.5, -2.1)$
- cluster 2:  $(0.4, -1.0)$ ,  $(-0.5, 0.6)$ ,  
 $(-0.1, 1.7)$ ,  $(1.2, 3.3)$ ,  $(3.1, 1.6)$ ,  
 $(1.3, 1.6)$ ,  $(2.0, 0.8)$
- new mean:  $(1.0, 1.2)$



## K-Means: Example

Recalculate distances for each cluster:

$x_1$	$x_2$	(-1.5, -2.1)	(1.0, 1.2)
0.4	-1.0	2.2	2.3
-1.0	-2.2	0.5	4.0
-2.4	-2.2	1.0	4.9
-1.0	-1.9	0.5	3.8
-0.5	0.6	2.8	1.7
-0.1	1.7	4.1	1.2
1.2	3.3	6.0	2.1
3.1	1.6	5.8	2.0
1.3	1.6	4.6	0.5
2.0	0.8	4.6	1.1

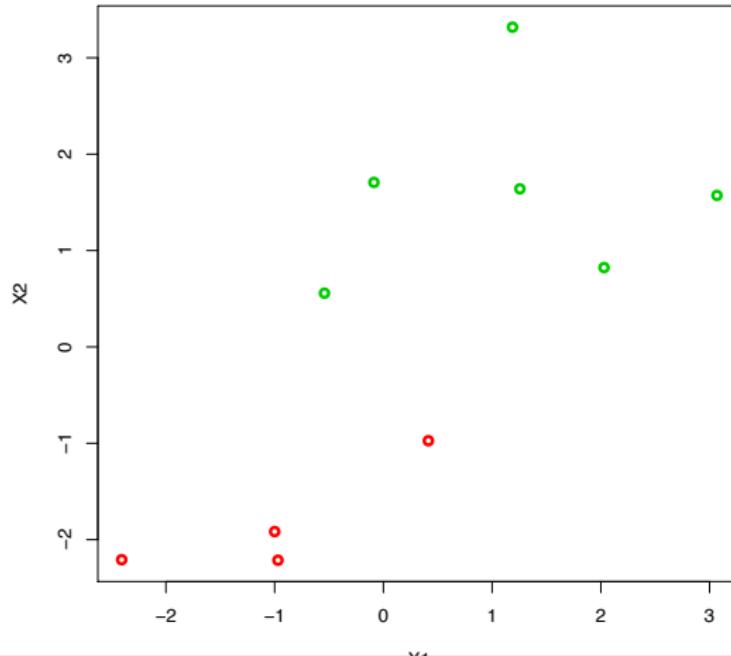
## K-Means: Example

Choose mean with smaller distance:

$x_1$	$x_2$	(-1.5, -2.1)	(1.0, 1.2)
0.4	-1.0	<b>2.2</b>	2.3
-1.0	-2.2	<b>0.5</b>	4.0
-2.4	-2.2	<b>1.0</b>	4.9
-1.0	-1.9	<b>0.5</b>	3.8
-0.5	0.6	2.8	<b>1.7</b>
-0.1	1.7	4.1	<b>1.2</b>
1.2	3.3	6.0	<b>2.1</b>
3.1	1.6	5.8	<b>2.0</b>
1.3	1.6	4.6	<b>0.5</b>
2.0	0.8	4.6	<b>1.1</b>

## K-Means: Example

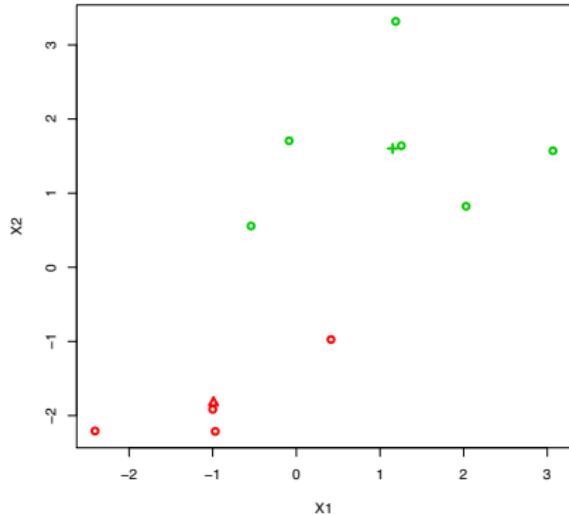
New clusters:



## K-Means: Example

Refit means for each cluster:

- cluster 1:  $(0.4, -1.0)$ ,  
 $(-1.0, -2.2)$ ,  $(-2.4, -2.2)$ ,  
 $(-1.0, -1.9)$
- new mean:  $(-1.0, -1.8)$
- cluster 2:  $(-0.5, 0.6)$ ,  $(-0.1, 1.7)$ ,  
 $(1.2, 3.3)$ ,  $(3.1, 1.6)$ ,  $(1.3, 1.6)$ ,  
 $(2.0, 0.8)$
- new mean:  $(1.2, 1.6)$



## K-Means: Example

Recalculate distances for each cluster:

$x_1$	$x_2$	(-1.0, -1.8)	(1.2, 1.6)
0.4	-1.0	1.6	2.7
-1.0	-2.2	0.4	4.4
-2.4	-2.2	1.5	5.2
-1.0	-1.9	0.1	4.1
-0.5	0.6	2.4	2.0
-0.1	1.7	3.6	1.2
1.2	3.3	5.6	1.7
3.1	1.6	5.3	1.9
1.3	1.6	4.1	0.1
2.0	0.8	4.0	1.2

## K-Means: Example

Select smallest distance and compare these clusters with previous:

Table: New Clusters

$x_1$	$x_2$	(-1.0, -1.8)	(1.2, 1.6)
0.4	-1.0	<b>1.6</b>	2.7
-1.0	-2.2	<b>0.4</b>	4.4
-2.4	-2.2	<b>1.5</b>	5.2
-1.0	-1.9	<b>0.1</b>	4.1
-0.5	0.6	2.4	<b>2.0</b>
-0.1	1.7	3.6	<b>1.2</b>
1.2	3.3	5.6	<b>1.7</b>
3.1	1.6	5.3	<b>1.9</b>
1.3	1.6	4.1	<b>0.1</b>
2.0	0.8	4.0	<b>1.2</b>

Table: Old Clusters

(-1.5, -2.1)	(1.0, 1.2)
<b>2.2</b>	2.3
<b>0.5</b>	4.0
<b>1.0</b>	4.9
<b>0.5</b>	3.8
2.8	<b>1.7</b>
4.1	<b>1.2</b>
6.0	<b>2.1</b>
5.8	<b>2.0</b>
4.6	<b>0.5</b>
4.6	<b>1.1</b>

## K-Means in Practice

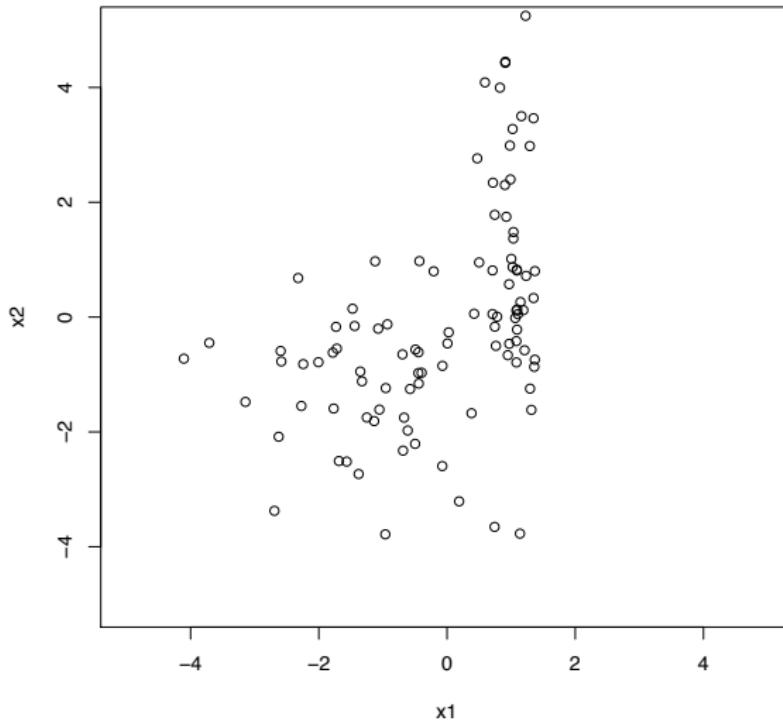
K-means can be used for *image segmentation*

- partition image into multiple segments
- find boundaries of objects
- make art



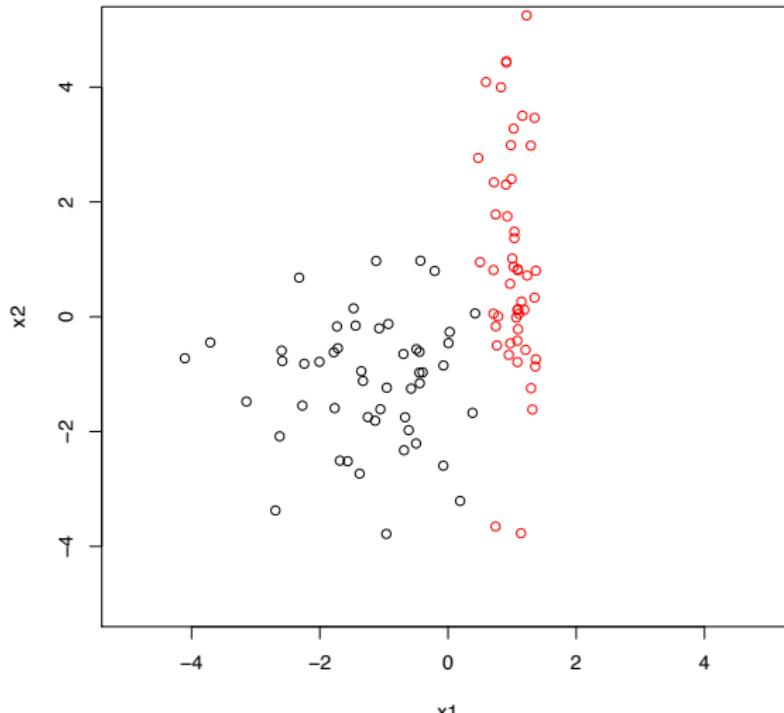
## K-Means Clustering

What if our data look like this?



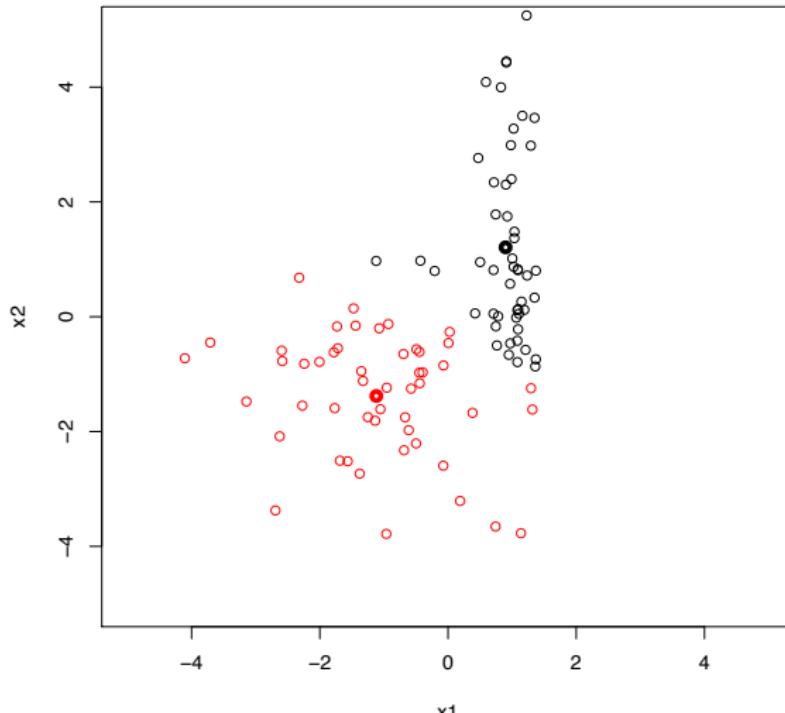
## K-Means Clustering

True clustering:



## K-Means Clustering

K-means clustering ( $K = 2$ ):



## Why topic models?



- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes

## Why topic models?



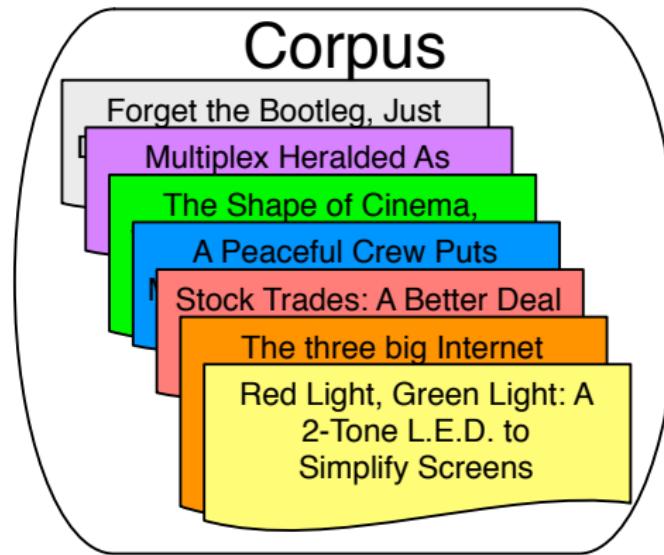
- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

## Roadmap

- What are topic models
- How to know if you have good topic model
- How to go from raw data to topics

## Conceptual Approach

From an **input corpus** and number of topics  $K \rightarrow$  words to topics



## Conceptual Approach

From an input corpus and number of topics  $K \rightarrow$  **words to topics**

TOPIC 1

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

TOPIC 2

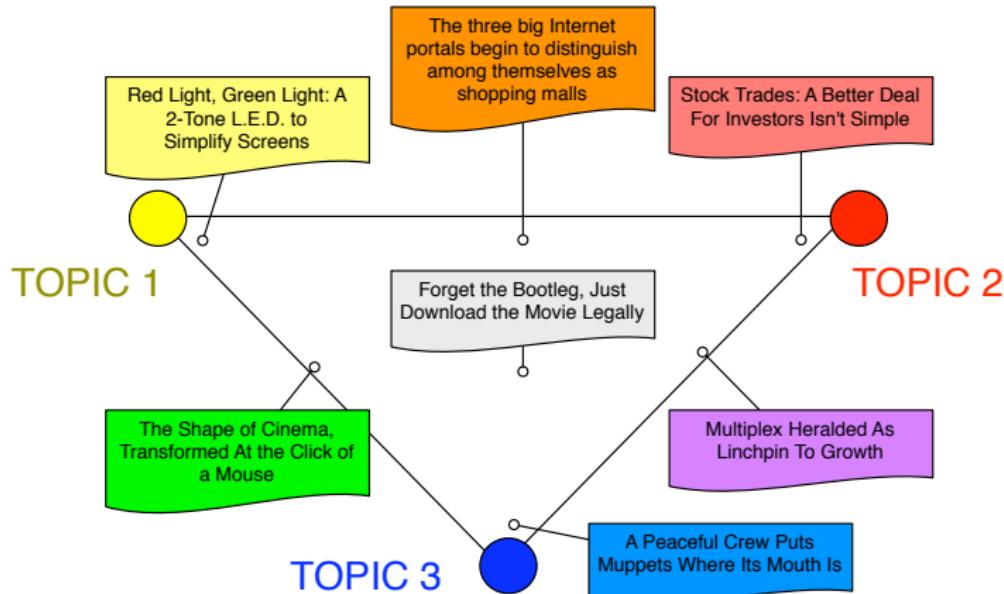
sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

TOPIC 3

play, film,  
movie, theater,  
production,  
star, director,  
stage

## Conceptual Approach

- For each document, what topics are expressed by that document?



## Topics from Science

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

## Why should you care?

- Neat way to explore / understand corpus collections
  - E-discovery
  - Social media
  - Scientific data
- NLP Applications
  - Word Sense Disambiguation
  - Discourse Segmentation
  - Machine Translation
- Psychology: word meaning, polysemy
- Inference is (relatively) simple

## Matrix Factorization Approach

$$\begin{bmatrix} M \times K \\ \text{Topic Assignment} \end{bmatrix} \times \begin{bmatrix} K \times V \\ \text{Topics} \end{bmatrix} \approx \begin{bmatrix} M \times V \\ \text{Dataset} \end{bmatrix}$$

**K** Number of topics

**M** Number of documents

**V** Size of vocabulary

## Matrix Factorization Approach

$$\begin{bmatrix} M \times K \\ \text{Topic Assignment} \end{bmatrix} \times \begin{bmatrix} K \times V \\ \text{Topics} \end{bmatrix} \approx \begin{bmatrix} M \times V \\ \text{Dataset} \end{bmatrix}$$

**K** Number of topics

**M** Number of documents

**V** Size of vocabulary

- If you use singular value decomposition (SVD), this technique is called latent semantic analysis.
- Popular in information retrieval.

## Alternative: Generative Model

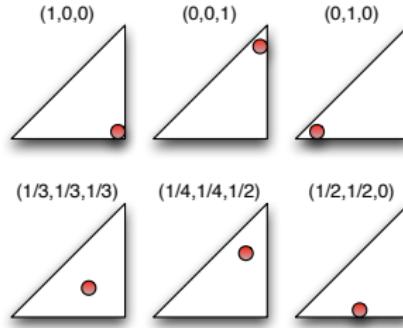
- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference

## Alternative: Generative Model

- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference
- Blei, Ng, Jordan. Latent **Dirichlet** Allocation. JMLR, 2003.

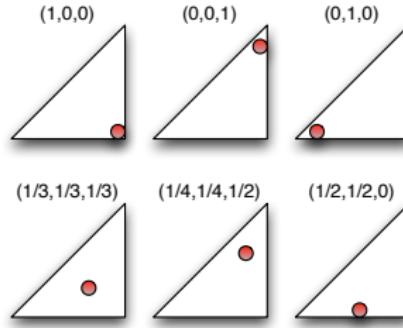
## Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



## Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



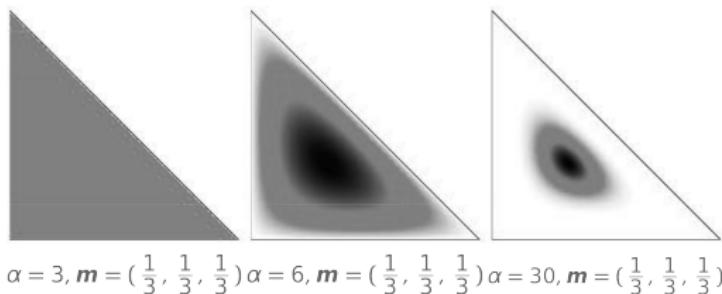
- Come from a Dirichlet distribution

## Dirichlet Distribution

$$P(\mathbf{p} \mid \alpha \mathbf{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

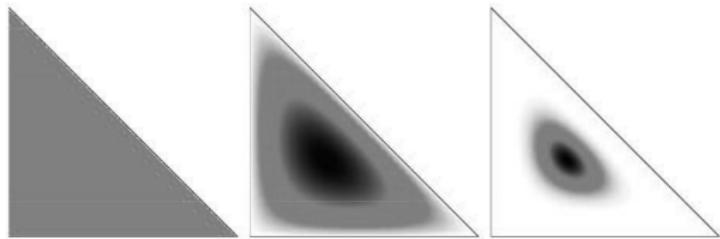
## Dirichlet Distribution

$$P(\mathbf{p} | \alpha \mathbf{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

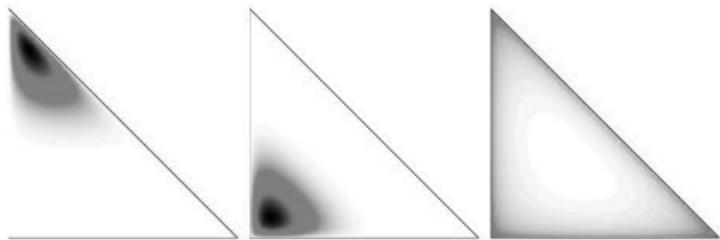


## Dirichlet Distribution

$$P(\mathbf{p} | \alpha \mathbf{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

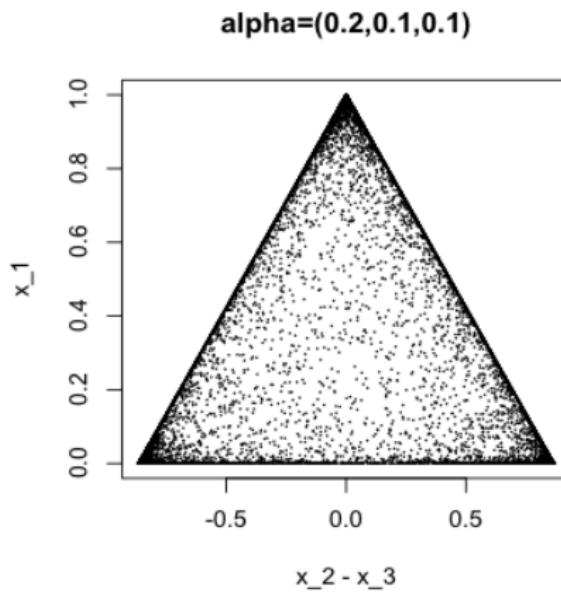


$$\alpha = 3, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \quad \alpha = 6, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \quad \alpha = 30, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$



$$\alpha = 14, \mathbf{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7}) \quad \alpha = 14, \mathbf{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7}) \quad \alpha = 2.7, \mathbf{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$

## Dirichlet Distribution



## Dirichlet Distribution

- If  $\vec{\phi} \sim \text{Dir}((\alpha))$ ,  $\vec{w} \sim \text{Mult}((\phi))$ , and  $n_k = |\{w_i : w_i = k\}|$  then

$$p(\phi | \alpha, \vec{w}) \propto p(\vec{w} | \phi) p(\phi | \alpha) \quad (1)$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \quad (2)$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \quad (3)$$

- Conjugacy: this **posterior** has the same form as the **prior**

## Dirichlet Distribution

- If  $\vec{\phi} \sim \text{Dir}((\alpha))$ ,  $\vec{w} \sim \text{Mult}((\phi))$ , and  $n_k = |\{w_i : w_i = k\}|$  then

$$p(\phi | \alpha, \vec{w}) \propto p(\vec{w} | \phi) p(\phi | \alpha) \quad (1)$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \quad (2)$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \quad (3)$$

- Conjugacy: this **posterior** has the same form as the **prior**

## Generative Model

TOPIC 1

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

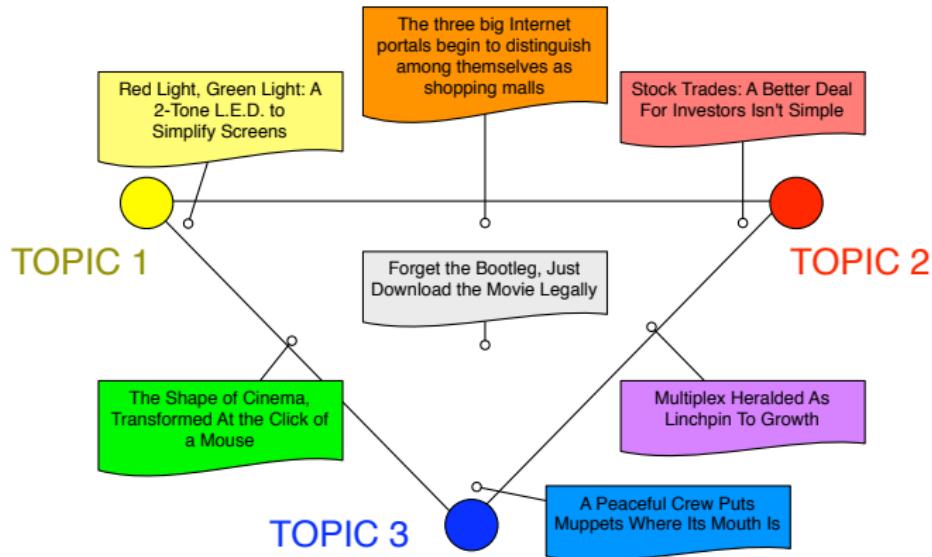
TOPIC 2

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

TOPIC 3

play, film,  
movie, theater,  
production,  
star, director,  
stage

## Generative Model



## Generative Model

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Generative Model



Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...



## Generative Model



Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Generative Model

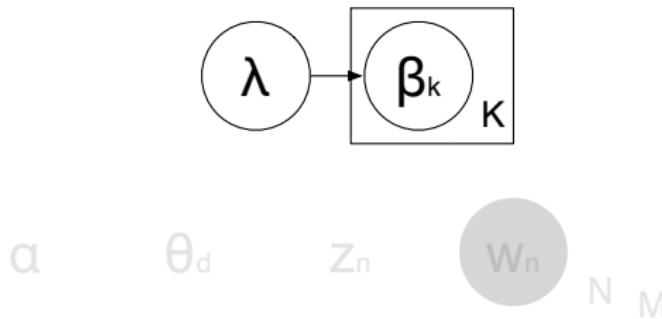
computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

play, film,  
movie, theater,  
production,  
star, director,  
stage

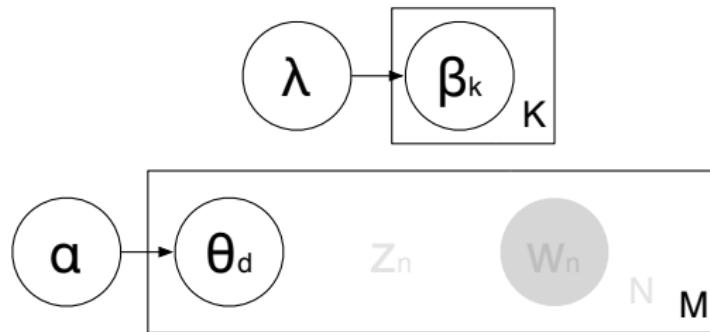
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Generative Model Approach



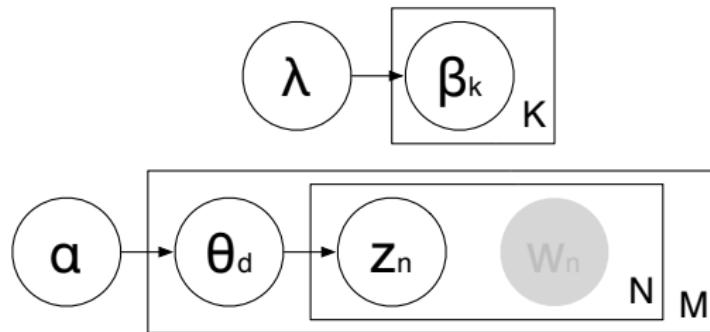
- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$

## Generative Model Approach



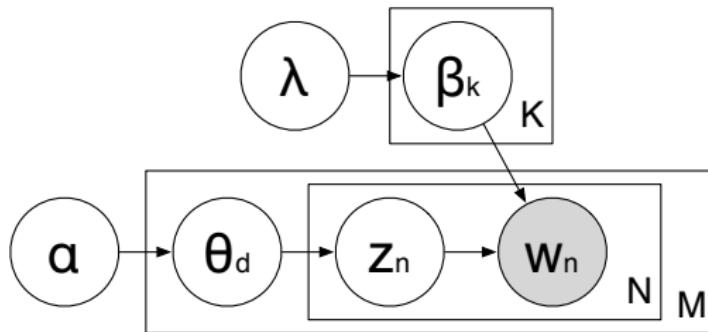
- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$

## Generative Model Approach



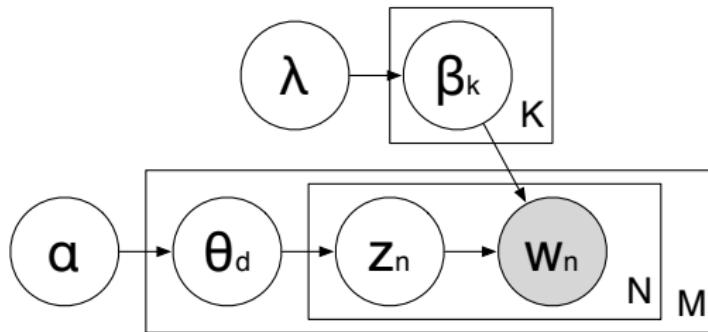
- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$
- For each word position  $n \in \{1, \dots, N\}$ , select a hidden topic  $z_n$  from the multinomial distribution parameterized by  $\theta$ .

## Generative Model Approach



- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$
- For each word position  $n \in \{1, \dots, N\}$ , select a hidden topic  $z_n$  from the multinomial distribution parameterized by  $\theta$ .
- Choose the observed word  $w_n$  from the distribution  $\beta_{z_n}$ .

## Generative Model Approach



- For each topic  $k \in \{1, \dots, K\}$ , draw a multinomial distribution  $\beta_k$  from a Dirichlet distribution with parameter  $\lambda$
- For each document  $d \in \{1, \dots, M\}$ , draw a multinomial distribution  $\theta_d$  from a Dirichlet distribution with parameter  $\alpha$
- For each word position  $n \in \{1, \dots, N\}$ , select a hidden topic  $z_n$  from the multinomial distribution parameterized by  $\theta$ .
- Choose the observed word  $w_n$  from the distribution  $\beta_{z_n}$ .

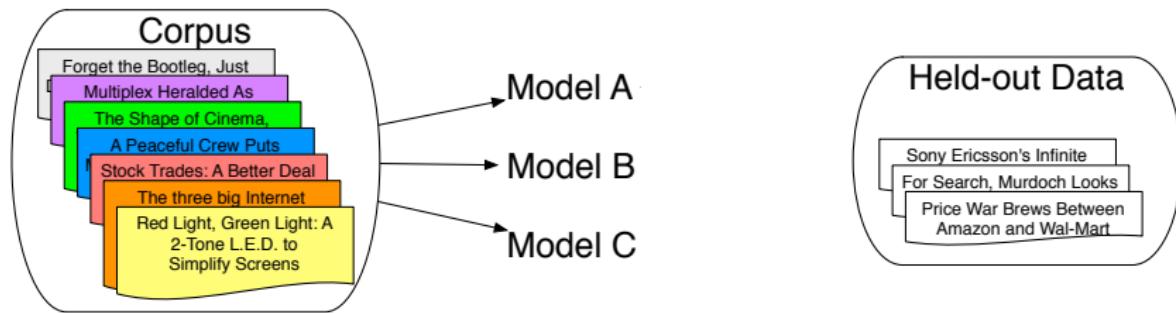
## Topic Models: What's Important

- Topic models
  - Topics to word types
  - Documents to topics
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA), a fully Bayesian version of pLSI, probabilistic version of LSA

## Topic Models: What's Important

- Topic models (latent variables)
  - Topics to word types
  - Documents to topics
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA), a fully Bayesian version of pLSI, probabilistic version of LSA

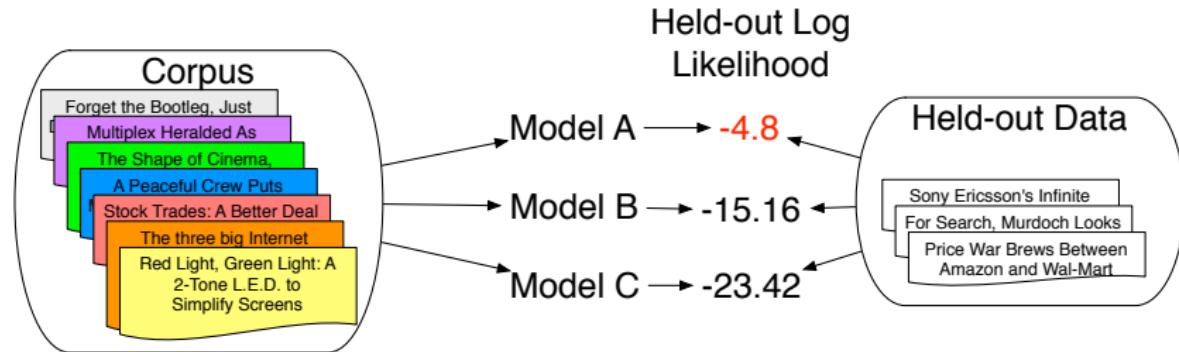
## Evaluation



$$P(\mathbf{w} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u}) = \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u})$$

How you compute it is important too (Wallach et al. 2009)

## Evaluation



Measures predictive power, not what the topics are

$$P(\mathbf{w} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u}) = \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z} | \mathbf{w}', \mathbf{z}', \alpha \mathbf{m}, \beta \mathbf{u})$$

How you compute it is important too (Wallach et al. 2009)

## Word Intrusion

### TOPIC 1

computer,  
technology,  
system,  
service, site,  
phone,  
internet,  
machine

### TOPIC 2

sell, sale,  
store, product,  
business,  
advertising,  
market,  
consumer

### TOPIC 3

play, film,  
movie, theater,  
production,  
star, director,  
stage

## Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

## Word Intrusion

1. Take the highest probability words from a topic

### Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

### Topic with Intruder

dog, cat, **apple**, horse, pig, cow

## Word Intrusion

1. Take the highest probability words from a topic

### Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

### Topic with Intruder

dog, cat, apple, horse, pig, cow

3. We ask users to find the word that doesn't belong

### Hypothesis

If the topics are interpretable, users will consistently choose true intruder

## Word Intrusion

1 / 10

crash      accident      board      agency      tibetan      safety

2 / 10

commercial      network      television      advertising      viewer      layoff

3 / 10

arrest      crime      inmate      pitcher      prison      death

4 / 10

hospital      doctor      health      care      medical      tradition

## Word Intrusion

**1 / 10****Reveal additional response**

crash

accident

board

agency

tibetan

safety

**2 / 10**

commercial

network

television

advertising

viewer

layoff

**3 / 10**

arrest

crime

inmate

pitcher

prison

death

**4 / 10**

hospital

doctor

health

care

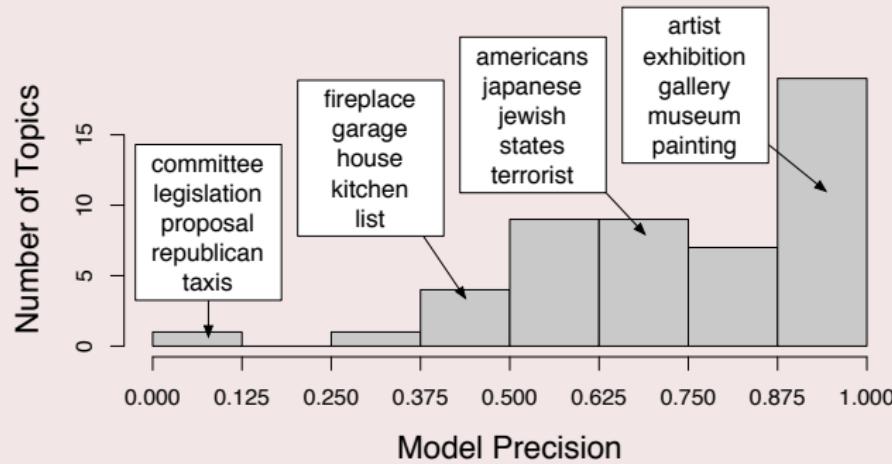
medical

tradition

- Order of words was shuffled
- Which intruder was selected varied
- Model precision: percentage of users who clicked on intruder

## Word Intrusion: Which Topics are Interpretable?

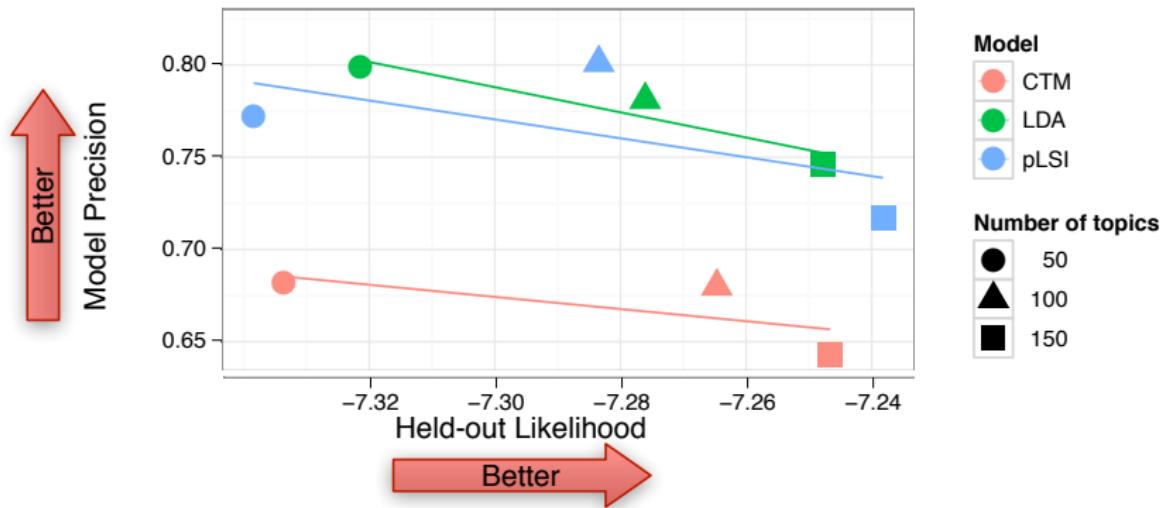
New York Times, 50 LDA Topics



Model Precision: percentage of correct intruders found

## Interpretability and Likelihood

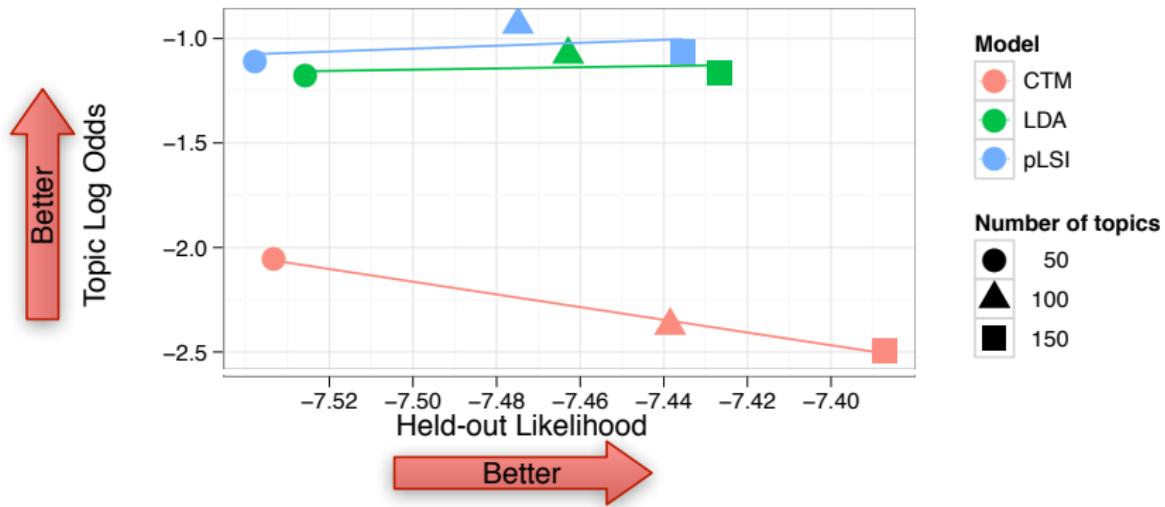
Model Precision on New York Times



within a model, higher likelihood  $\neq$  higher interpretability

## Interpretability and Likelihood

Topic Log Odds on Wikipedia



across models, higher likelihood  $\neq$  higher interpretability

## Evaluation Takeaway

- Measure what you care about
- If you care about prediction, likelihood is good
- If you care about a particular task, measure that

## Sampling Token A

### Assignments

Doc<sub>1</sub> :  $z_A = 1, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger dog iron  
pig

**Topic 2** :pig hamburger iron  
cat

**Topic 3** :dog iron cat

$$\blacksquare p(z_A = 1) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.333 \times 0.125 = 0.042 = 0.042$$

## Sampling Token A

### Assignments

Doc<sub>1</sub> :  $z_A = 1, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger dog iron

pig

**Topic 2** :pig hamburger iron

cat

**Topic 3** :dog iron cat

$$\blacksquare p(z_A = 1) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.333 \times 0.125 = 0.042 = 0.042$$

$$\blacksquare p(z_A = 2) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{4+5.000} \right) = 0.333 \times 0.111 = 0.037 = 0.037$$

## Sampling Token A

### Assignments

Doc<sub>1</sub> :  $z_A = 1, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger dog iron

pig

**Topic 2** :pig hamburger iron

cat

**Topic 3** :dog iron cat

$$\blacksquare \quad p(z_A = 1) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.333 \times 0.125 = 0.042 = 0.042$$

$$\blacksquare \quad p(z_A = 2) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{4+5.000} \right) = 0.333 \times 0.111 = 0.037 = 0.037$$

$$\blacksquare \quad p(z_A = 3) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{1+1.000}{3+5.000} \right) = 0.333 \times 0.250 = 0.083 = 0.083$$

## Sampling Token A

### Assignments

Doc<sub>1</sub> :  $z_A = 1, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger dog iron

pig

**Topic 2** :pig hamburger iron

cat

**Topic 3** :dog iron cat

$$\blacksquare \quad p(z_A = 1) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.333 \times 0.125 = 0.042 = 0.042$$

$$\blacksquare \quad p(z_A = 2) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{4+5.000} \right) = 0.333 \times 0.111 = 0.037 = 0.037$$

$$\blacksquare \quad p(z_A = 3) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{1+1.000}{3+5.000} \right) = 0.333 \times 0.250 = 0.083 = 0.083$$

## Sampling Token A

### Assignments

Doc<sub>1</sub> :  $z_A = 1, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger dog iron

pig

**Topic 2** :pig hamburger iron

cat

**Topic 3** :dog iron cat

$$\blacksquare p(z_A = 1) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.333 \times 0.125 = 0.042 = 0.042$$

$$\blacksquare p(z_A = 2) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{4+5.000} \right) = 0.333 \times 0.111 = 0.037 = 0.037$$

$$\blacksquare p(z_A = 3) = \left( \frac{1+1.000}{3+3.000} \right) \times \left( \frac{1+1.000}{3+5.000} \right) = 0.333 \times 0.250 = 0.083 = 0.083$$

New assignment for (0, 0): 3

## Sampling Token B

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron  
cat

**Topic 3** :dog dog iron cat

## Sampling Token B

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron cat

**Topic 3** :dog dog iron cat

$$\blacksquare p(z_B = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$$

## Sampling Token B

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron cat

**Topic 3** :dog dog iron cat

- $p(z_B = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_B = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$

## Sampling Token B

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron cat

**Topic 3** :dog dog iron cat

- $p(z_B = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_B = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_B = 3) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.500 \times 0.222 = 0.111 = 0.111$

## Sampling Token B

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron cat

**Topic 3** :dog dog iron cat

- $p(z_B = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_B = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_B = 3) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.500 \times 0.222 = 0.111 = 0.111$

## Sampling Token B

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 2, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron cat

**Topic 3** :dog dog iron cat

- $p(z_B = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_B = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_B = 3) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.500 \times 0.222 = 0.111 = 0.111$

New assignment for (0, 1): 3

## Sampling Token C

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

## Sampling Token C

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_C = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$

## Sampling Token C

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_C = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_C = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$

## Sampling Token C

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_C = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_C = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_C = 3) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.500 \times 0.222 = 0.111 = 0.111$

## Sampling Token C

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_C = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_C = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_C = 3) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.500 \times 0.222 = 0.111 = 0.111$

## Sampling Token C

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_C = 1) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.333 \times 0.125 = 0.042 = 0.042$
- $p(z_C = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_C = 3) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.500 \times 0.222 = 0.111 = 0.111$

New assignment for (0, 2): 3

## Sampling Token D

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

## Sampling Token D

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_D = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.167 \times 0.143 = 0.024 = 0.024$

## Sampling Token D

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_D = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.167 \times 0.143 = 0.024 = 0.024$
- $p(z_D = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.167 \times 0.250 = 0.042 = 0.042$

## Sampling Token D

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_D = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.167 \times 0.143 = 0.024 = 0.024$
- $p(z_D = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.167 \times 0.250 = 0.042 = 0.042$
- $p(z_D = 3) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{5+5.000}\right) = 0.667 \times 0.100 = 0.067 = 0.067$

## Sampling Token D

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_D = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.167 \times 0.143 = 0.024 = 0.024$
- $p(z_D = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.167 \times 0.250 = 0.042 = 0.042$
- $p(z_D = 3) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{5+5.000}\right) = 0.667 \times 0.100 = 0.067 = 0.067$

## Sampling Token D

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 1$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron pig

**Topic 2** :pig hamburger iron

**Topic 3** :dog dog iron cat cat

- $p(z_D = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.167 \times 0.143 = 0.024 = 0.024$
- $p(z_D = 2) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.167 \times 0.250 = 0.042 = 0.042$
- $p(z_D = 3) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{5+5.000}\right) = 0.667 \times 0.100 = 0.067 = 0.067$

New assignment for (0, 3): 3

## Sampling Token E

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron

**Topic 2** :pig hamburger iron

**Topic 3** :pig dog dog iron cat  
cat

## Sampling Token E

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron

**Topic 2** :pig hamburger iron

**Topic 3** :pig dog dog iron cat  
cat

$$\blacksquare p(z_E = 1) = \left( \frac{1+1.000}{2+3.000} \right) \times \left( \frac{1+1.000}{2+5.000} \right) = 0.400 \times 0.286 = 0.114 = 0.114$$

## Sampling Token E

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron

**Topic 2** :pig hamburger iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_E = 1) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.400 \times 0.286 = 0.114 = 0.114$
- $p(z_E = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$

## Sampling Token E

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron

**Topic 2** :pig hamburger iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_E = 1) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.400 \times 0.286 = 0.114 = 0.114$
- $p(z_E = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_E = 3) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{6+5.000}\right) = 0.400 \times 0.091 = 0.036 = 0.036$

## Sampling Token E

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron

**Topic 2** :pig hamburger iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_E = 1) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.400 \times 0.286 = 0.114 = 0.114$
- $p(z_E = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_E = 3) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{6+5.000}\right) = 0.400 \times 0.091 = 0.036 = 0.036$

## Sampling Token E

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 2, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger iron

**Topic 2** :pig hamburger iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_E = 1) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.400 \times 0.286 = 0.114 = 0.114$
- $p(z_E = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_E = 3) = \left(\frac{1+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{6+5.000}\right) = 0.400 \times 0.091 = 0.036 = 0.036$

New assignment for (1, 0): 1

## Sampling Token F

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger iron

**Topic 2** :pig iron

**Topic 3** :pig dog dog iron cat  
cat

## Sampling Token F

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger iron

**Topic 2** :pig iron

**Topic 3** :pig dog dog iron cat  
cat

$$\blacksquare p(z_F = 1) = \left( \frac{2+1.000}{2+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.600 \times 0.125 = 0.075 = 0.075$$

## Sampling Token F

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger iron

**Topic 2** :pig iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_F = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.600 \times 0.125 = 0.075 = 0.075$
- $p(z_F = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$

## Sampling Token F

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger iron

**Topic 2** :pig iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_F = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.600 \times 0.125 = 0.075 = 0.075$
- $p(z_F = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_F = 3) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{5+5.000}\right) = 0.200 \times 0.200 = 0.040 = 0.040$

## Sampling Token F

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger iron

**Topic 2** :pig iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_F = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.600 \times 0.125 = 0.075 = 0.075$
- $p(z_F = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_F = 3) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{5+5.000}\right) = 0.200 \times 0.200 = 0.040 = 0.040$

## Sampling Token F

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 3, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger iron

**Topic 2** :pig iron

**Topic 3** :pig dog dog iron cat  
cat

- $p(z_F = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.600 \times 0.125 = 0.075 = 0.075$
- $p(z_F = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_F = 3) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{5+5.000}\right) = 0.200 \times 0.200 = 0.040 = 0.040$

New assignment for (1, 1): 1

## Sampling Token G

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

## Sampling Token G

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

$$\blacksquare p(z_G = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.600 \times 0.250 = 0.150 = 0.150$$

## Sampling Token G

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

$$\blacksquare p(z_G = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.600 \times 0.250 = 0.150 = 0.150$$

$$\blacksquare p(z_G = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$$

## Sampling Token G

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_G = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.600 \times 0.250 = 0.150 = 0.150$
- $p(z_G = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_G = 3) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{5+5.000}\right) = 0.200 \times 0.100 = 0.020 = 0.020$

## Sampling Token G

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_G = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.600 \times 0.250 = 0.150 = 0.150$
- $p(z_G = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_G = 3) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{5+5.000}\right) = 0.200 \times 0.100 = 0.020 = 0.020$

## Sampling Token G

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_G = 1) = \left(\frac{2+1.000}{2+3.000}\right) \times \left(\frac{1+1.000}{3+5.000}\right) = 0.600 \times 0.250 = 0.150 = 0.150$
- $p(z_G = 2) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{2+5.000}\right) = 0.200 \times 0.143 = 0.029 = 0.029$
- $p(z_G = 3) = \left(\frac{0+1.000}{2+3.000}\right) \times \left(\frac{0+1.000}{5+5.000}\right) = 0.200 \times 0.100 = 0.020 = 0.020$

New assignment for (1, 2): 1

## Sampling Token H

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

## Sampling Token H

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger  
hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_H = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$

## Sampling Token H

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger  
hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_H = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_H = 2) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.500 \times 0.286 = 0.143 = 0.143$

## Sampling Token H

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger  
hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_H = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_H = 2) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.500 \times 0.286 = 0.143 = 0.143$
- $p(z_H = 3) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{5+5.000}\right) = 0.333 \times 0.200 = 0.067 = 0.067$

## Sampling Token H

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger  
hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_H = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_H = 2) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.500 \times 0.286 = 0.143 = 0.143$
- $p(z_H = 3) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{5+5.000}\right) = 0.333 \times 0.200 = 0.067 = 0.067$

## Sampling Token H

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 1, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog iron

**Topic 2** :pig iron

**Topic 3** :pig dog iron cat cat

- $p(z_H = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_H = 2) = \left(\frac{2+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{2+5.000}\right) = 0.500 \times 0.286 = 0.143 = 0.143$
- $p(z_H = 3) = \left(\frac{1+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{5+5.000}\right) = 0.333 \times 0.200 = 0.067 = 0.067$

New assignment for (2, 0): 2

## Sampling Token I

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron

**Topic 3** :pig dog iron cat cat

## Sampling Token I

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron

**Topic 3** :pig dog iron cat cat

- $p(z_I = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$

## Sampling Token I

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron

**Topic 3** :pig dog iron cat cat

$$\blacksquare \quad p(z_I = 1) = \left( \frac{0+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.167 \times 0.125 = 0.021 = 0.021$$

$$\blacksquare \quad p(z_I = 2) = \left( \frac{3+1.000}{3+3.000} \right) \times \left( \frac{2+1.000}{3+5.000} \right) = 0.667 \times 0.375 = 0.250 = 0.250$$

## Sampling Token I

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron

**Topic 3** :pig dog iron cat cat

- $p(z_I = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_I = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{2+1.000}{3+5.000}\right) = 0.667 \times 0.375 = 0.250 = 0.250$
- $p(z_I = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{4+5.000}\right) = 0.167 \times 0.111 = 0.019 = 0.019$

## Sampling Token I

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron

**Topic 3** :pig dog iron cat cat

- $p(z_I = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_I = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{2+1.000}{3+5.000}\right) = 0.667 \times 0.375 = 0.250 = 0.250$
- $p(z_I = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{4+5.000}\right) = 0.167 \times 0.111 = 0.019 = 0.019$

## Sampling Token I

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 3, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron

**Topic 3** :pig dog iron cat cat

- $p(z_I = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_I = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{2+1.000}{3+5.000}\right) = 0.667 \times 0.375 = 0.250 = 0.250$
- $p(z_I = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{4+5.000}\right) = 0.167 \times 0.111 = 0.019 = 0.019$

New assignment for (2, 1): 2

## Sampling Token J

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

## Sampling Token J

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

$$\blacksquare p(z_J = 1) = \left( \frac{0+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.167 \times 0.125 = 0.021 = 0.021$$

## Sampling Token J

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

$$\blacksquare \quad p(z_J = 1) = \left( \frac{0+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.167 \times 0.125 = 0.021 = 0.021$$

$$\blacksquare \quad p(z_J = 2) = \left( \frac{3+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.667 \times 0.125 = 0.083 = 0.083$$

## Sampling Token J

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

- $p(z_J = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_J = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.667 \times 0.125 = 0.083 = 0.083$
- $p(z_J = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.167 \times 0.222 = 0.037 = 0.037$

## Sampling Token J

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

- $p(z_J = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_J = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.667 \times 0.125 = 0.083 = 0.083$
- $p(z_J = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.167 \times 0.222 = 0.037 = 0.037$

## Sampling Token J

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

- $p(z_J = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_J = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.667 \times 0.125 = 0.083 = 0.083$
- $p(z_J = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{1+1.000}{4+5.000}\right) = 0.167 \times 0.222 = 0.037 = 0.037$

New assignment for (2, 2): 2

## Sampling Token K

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

## Sampling Token K

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

$$\blacksquare p(z_K = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$$

## Sampling Token K

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

$$\blacksquare \quad p(z_K = 1) = \left( \frac{0+1.000}{3+3.000} \right) \times \left( \frac{0+1.000}{3+5.000} \right) = 0.167 \times 0.125 = 0.021 = 0.021$$

$$\blacksquare \quad p(z_K = 2) = \left( \frac{3+1.000}{3+3.000} \right) \times \left( \frac{2+1.000}{3+5.000} \right) = 0.667 \times 0.375 = 0.250 = 0.250$$

## Sampling Token K

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

- $p(z_K = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_K = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{2+1.000}{3+5.000}\right) = 0.667 \times 0.375 = 0.250 = 0.250$
- $p(z_K = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{4+5.000}\right) = 0.167 \times 0.111 = 0.019 = 0.019$

## Sampling Token K

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

- $p(z_K = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_K = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{2+1.000}{3+5.000}\right) = 0.667 \times 0.375 = 0.250 = 0.250$
- $p(z_K = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{4+5.000}\right) = 0.167 \times 0.111 = 0.019 = 0.019$

## Sampling Token K

### Assignments

Doc<sub>1</sub> :  $z_A = 3, z_B = 3, z_C = 3, z_D = 3$

Doc<sub>2</sub> :  $z_E = 1, z_F = 1, z_G = 1$

Doc<sub>3</sub> :  $z_H = 2, z_I = 2, z_J = 2, z_K = 2$

### Topics

**Topic 1** :hamburger

hamburger dog

**Topic 2** :pig iron iron iron

**Topic 3** :pig dog cat cat

- $p(z_K = 1) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{3+5.000}\right) = 0.167 \times 0.125 = 0.021 = 0.021$
- $p(z_K = 2) = \left(\frac{3+1.000}{3+3.000}\right) \times \left(\frac{2+1.000}{3+5.000}\right) = 0.667 \times 0.375 = 0.250 = 0.250$
- $p(z_K = 3) = \left(\frac{0+1.000}{3+3.000}\right) \times \left(\frac{0+1.000}{4+5.000}\right) = 0.167 \times 0.111 = 0.019 = 0.019$

New assignment for (2, 3): 2