University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chenhao Tan

Lecture 20: Topic modeling

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Administrivia

- Poster printing (stay tuned!) Ai Dec 11 Dec 9
- HW 5 (final homework) is due on Friday!
- HW 4 grades
- Example questions
- Midpoint feedback

Learning Objectives

- Learn about latent Dirichlet allocation
- Understand plate notations
- Understand intuitions behind evaluations of topic models

Outline for CSCI 5622
We've already covered stuff in blue!

- Problem formulations: classification, regression
- Supervised techniques: decision trees, nearest neighbors, perceptron, linear models, neural networks, support vector machine, kernel methods
- Unsupervised techniques: clustering, linear dimensionality reduction, topic modeling
- "Meta-techniques": ensembles, expectation-maximization, variational inference
- Understanding ML: limits of learning, practical issues, bias & fairness
- Recurring themes: (stochastic) gradient descent

Outline

- Generative story for latent Dirichlet allocation
- Plate notations
- Evaluations of topic models

Outline

- Generative story for latent Dirichlet allocation
- Plate notations
- Evaluations of topic models

Topic models

- Discrete count data

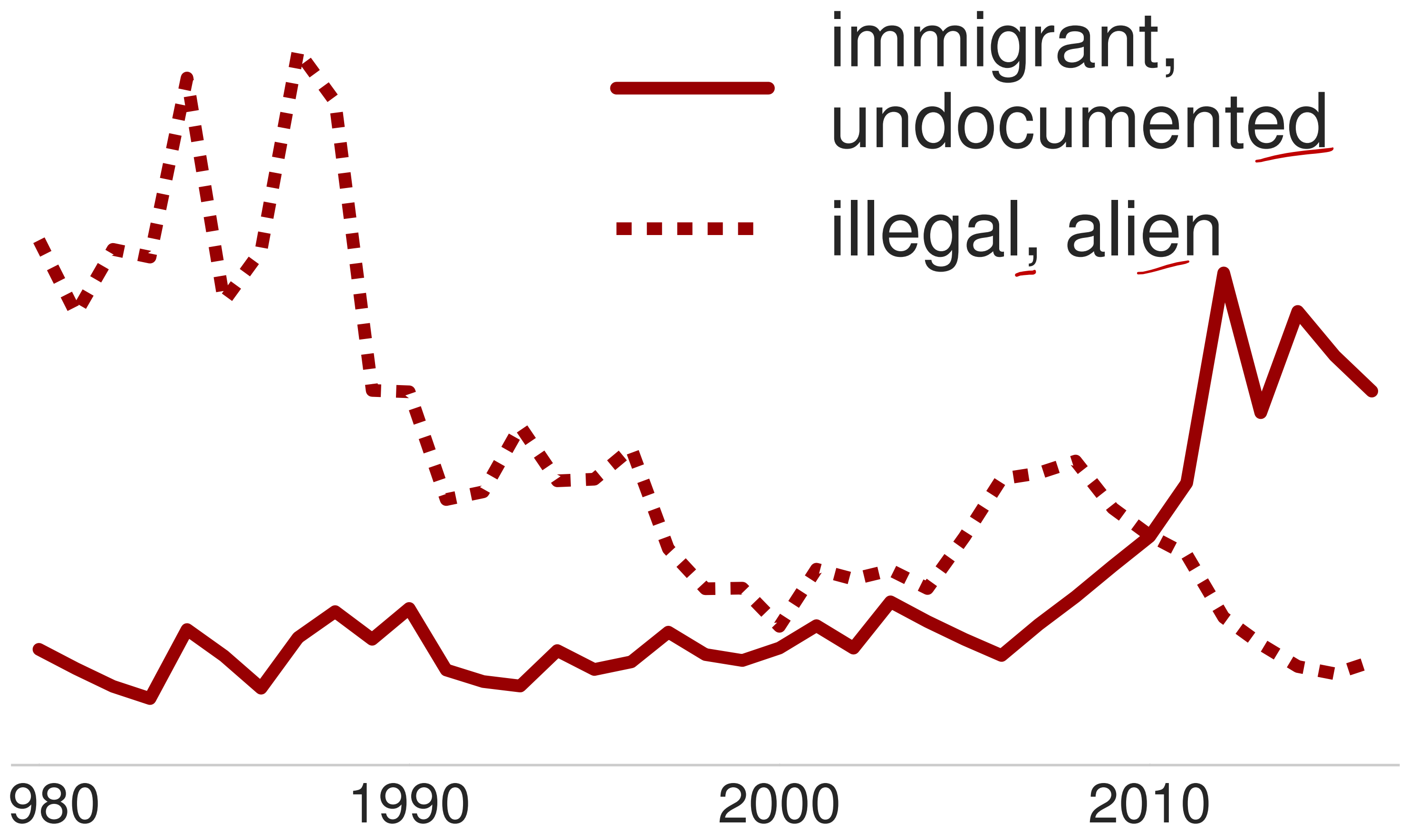
Topic models

- 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

Why should you care?

- Neat way to explore/understand corpus collections
  - E-discovery
  - Social media
  - Scientific data
- NLP Applications
  - Dimensionality reduction
  - Classification
- A general way to model count data and a general inference algorithm

Head-to-head
(anti-correlated, rarely cooccur)



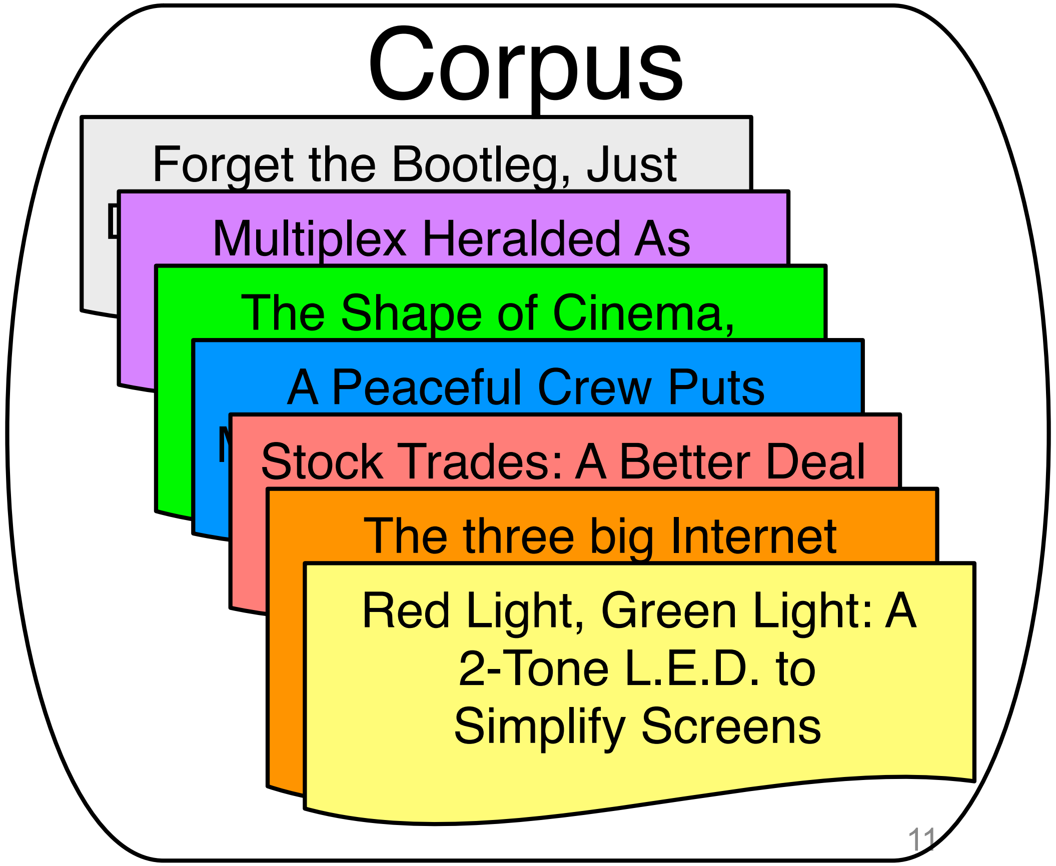
Tan, Card, Smith, 2017

Conceptual approach

- Input: a text corpus and number of topics *K*

|V|

- Output:
  - K topics, each topic is a list of words
  - Topic assignment for each document



Conceptual approach

- K topics, each topic is a list of words

TOPIC 1

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

IT

TOPIC 2

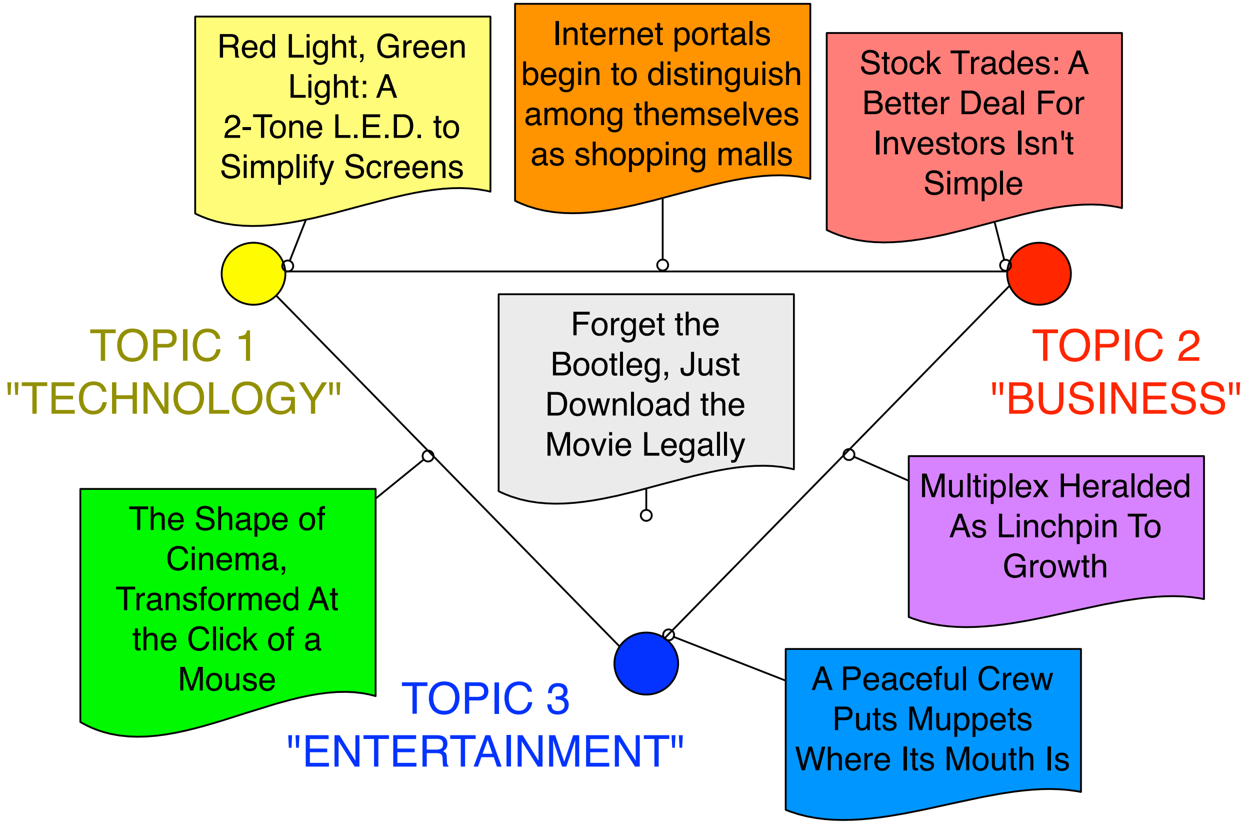
sell, sale, store, product, business, advertising, market, consumer

TOPIC 3

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

Conceptual approach

- Topic assignment for each 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

Topic models

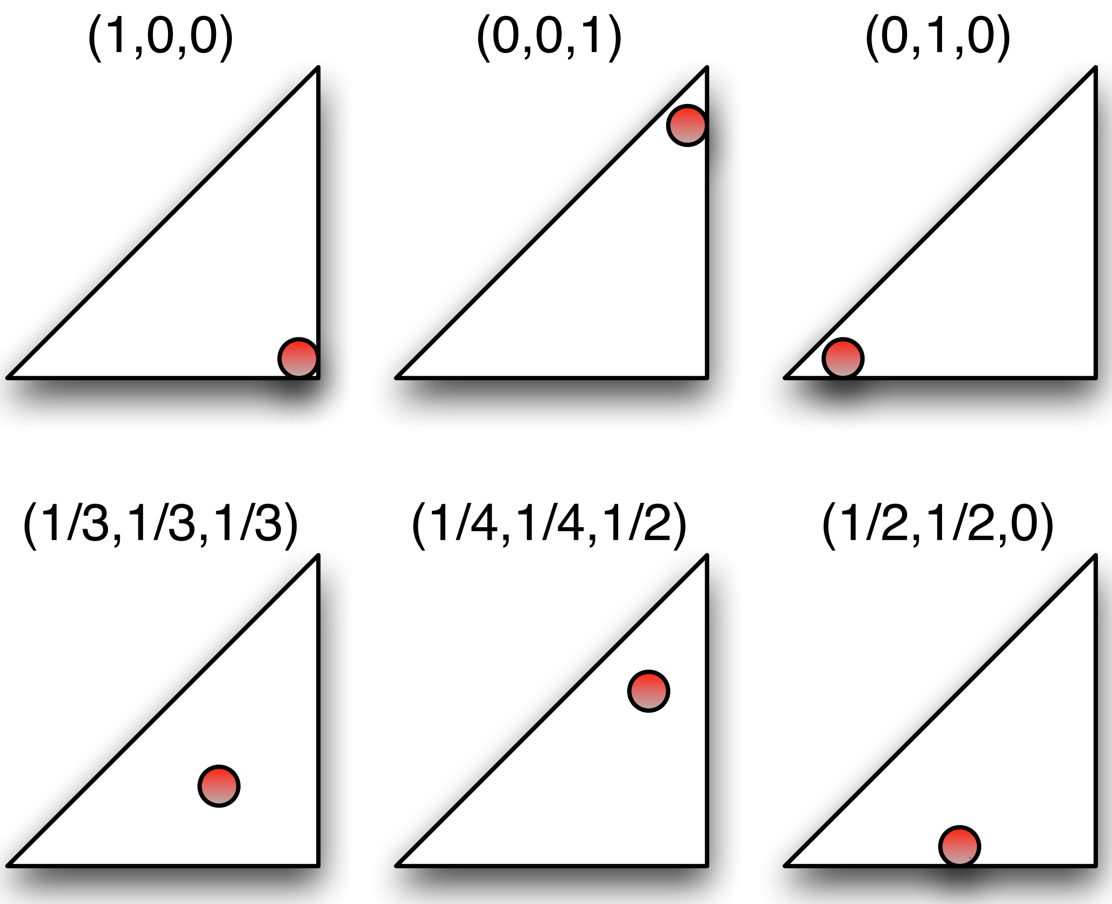
- Discrete count data
- Gaussian distributions are not appropriate

Generative model:
Latent Dirichlet Allocation

- Generate a document, or a bag of words
- Blei, Ng, Jordan. Latent Dirichlet Allocation. JMLR, 2003.

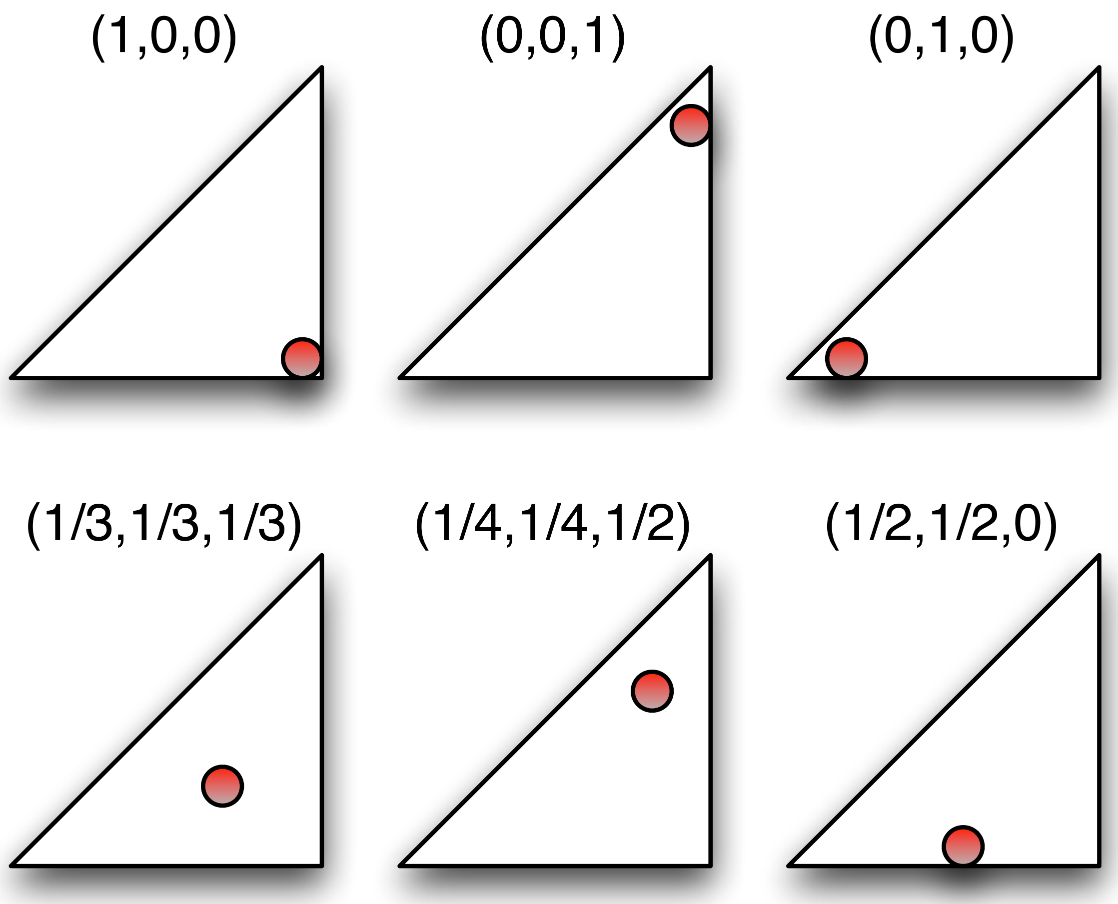
Generative model: Latent Dirichlet Allocation

- Generate a document, or a bag of words
- Multinomial distribution
  - Distribution over discrete outcomes
  - Represented by non-negative vector that sums to one
  - Picture representation



Generative model: Latent Dirichlet Allocation

- Generate a document, or a bag of words
- Multinomial distribution
  - Distribution over discrete outcomes
  - Represented by non-negative vector that sums to one
  - Picture representation
  - Come from a Dirichlet distribution



Generative story

TOPIC 1

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

$\sum_{w \in V} \beta_{1w} = 1$

$\beta_{1,\text{computer}} = 0.1$

$|V|$

TOPIC 2

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

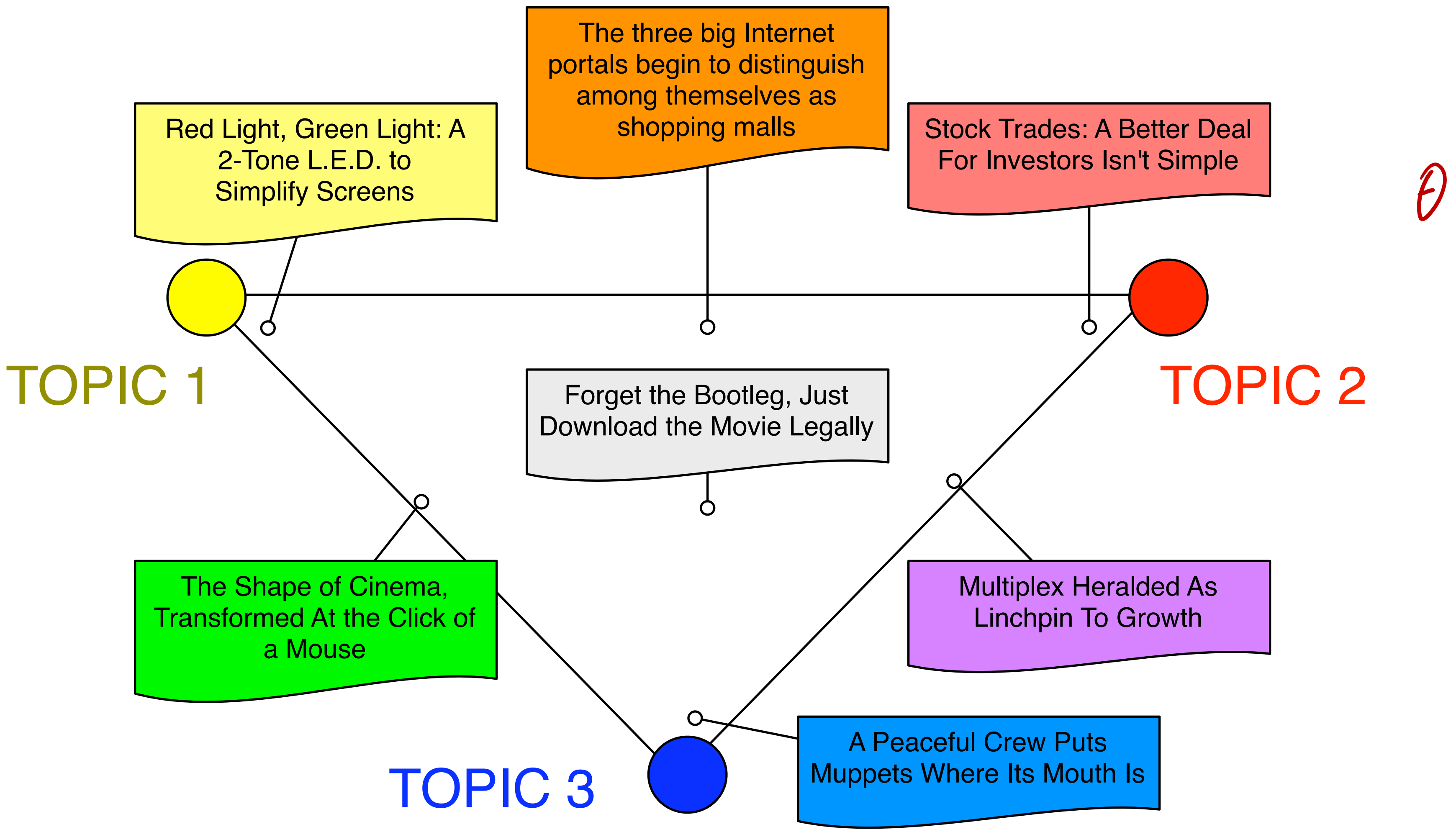
β_2

TOPIC 3

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

β_3

Generative story



Generative story

β_1 β_2 β_3 $P(w|t)$

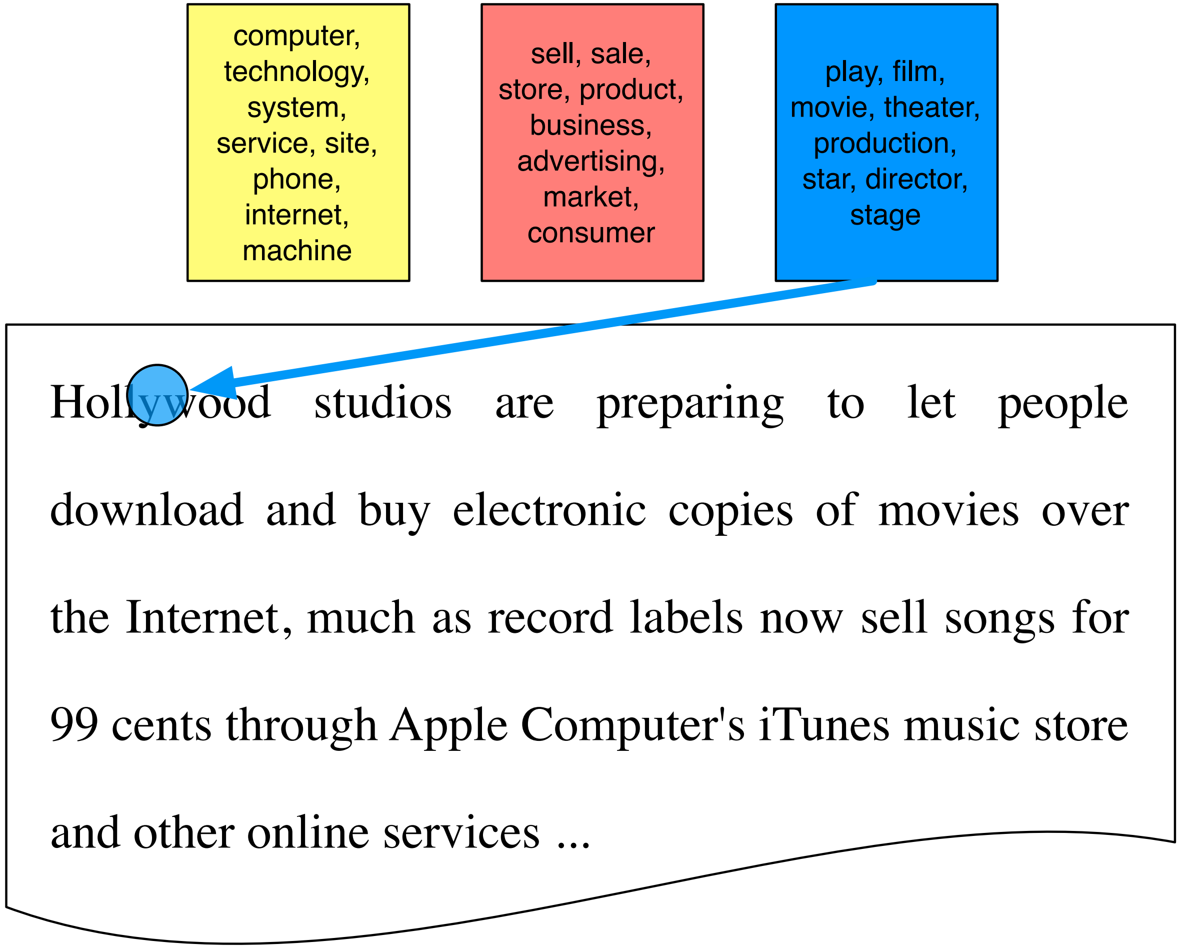
β_1	β_2	β_3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

$z=3$

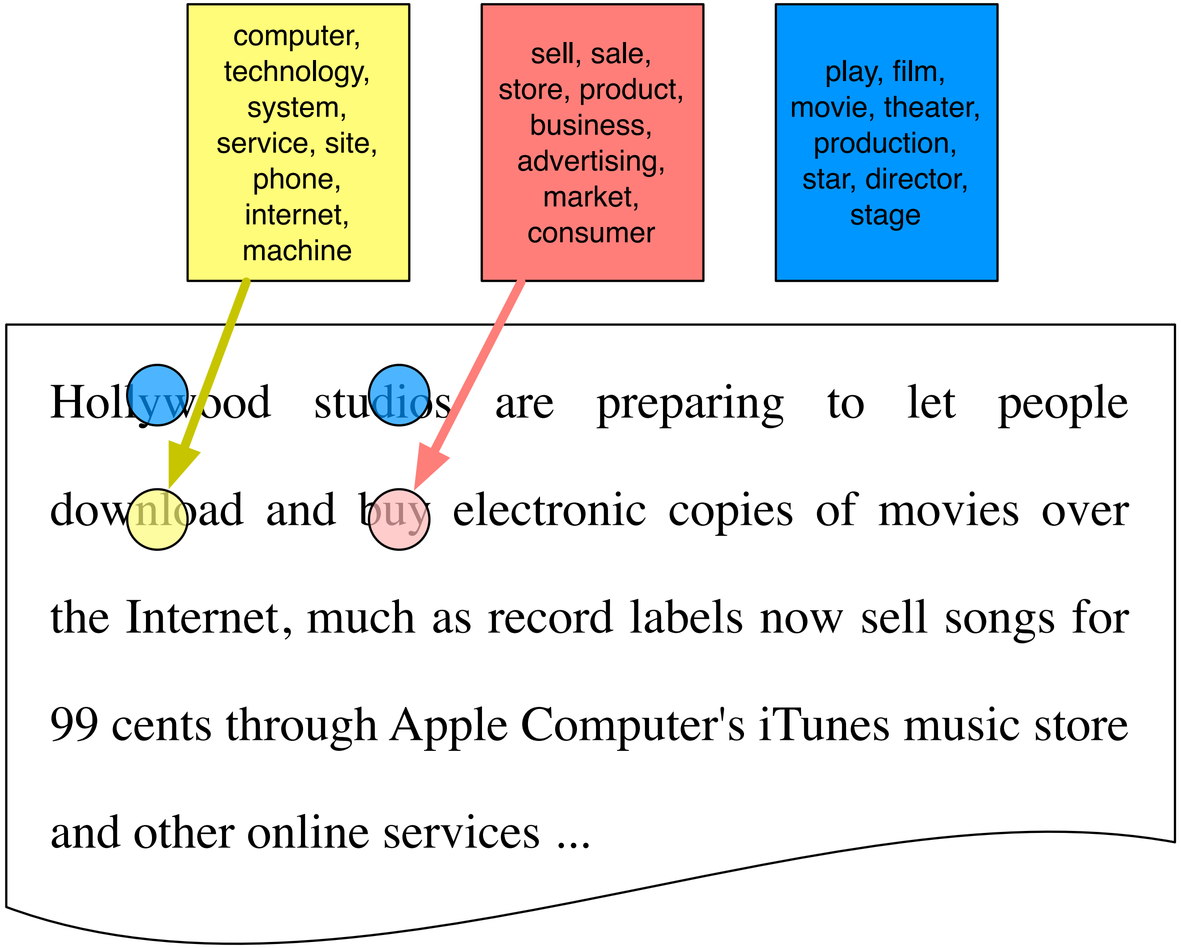
$\theta=(\theta_1,\theta_2,\theta_3)$

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 ...

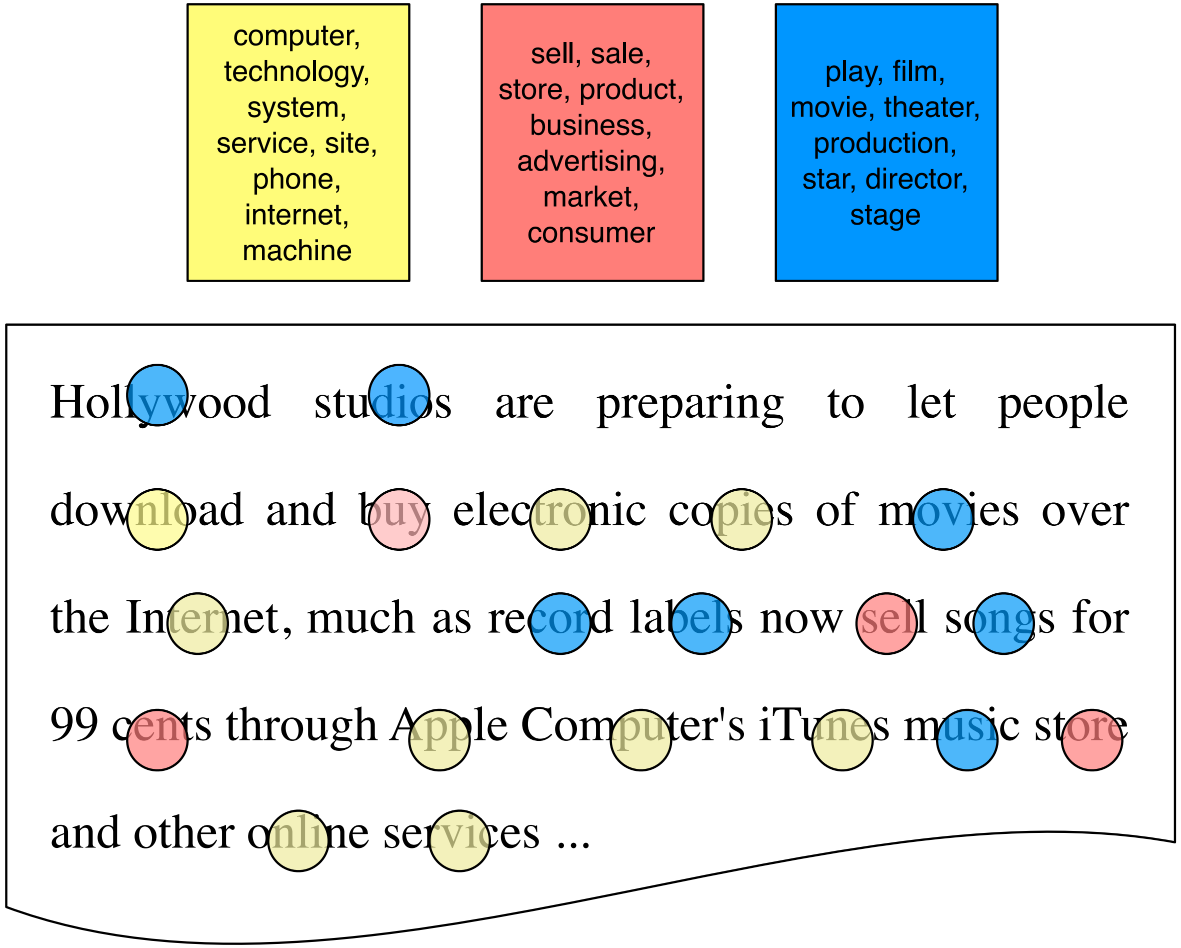
Generative story



Generative story



Generative story



Generative story

- Generate topic-word distribution for each topic β $|V|$
- For each document
    - Generate its document-topic distribution θ T
    - For each word
        - Choose a topic from the document-topic distribution
        - Choose a word from the topic's corresponding word-topic distribution

$$P(w|\theta,\beta) =$$

$$\Downarrow$$

$$P(\theta,\beta|w)$$

Missing component: how to generate a multinomial distribution

$\alpha m = (\frac{1}{50}, \ldots, \frac{1}{50})$ Dirichlet

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

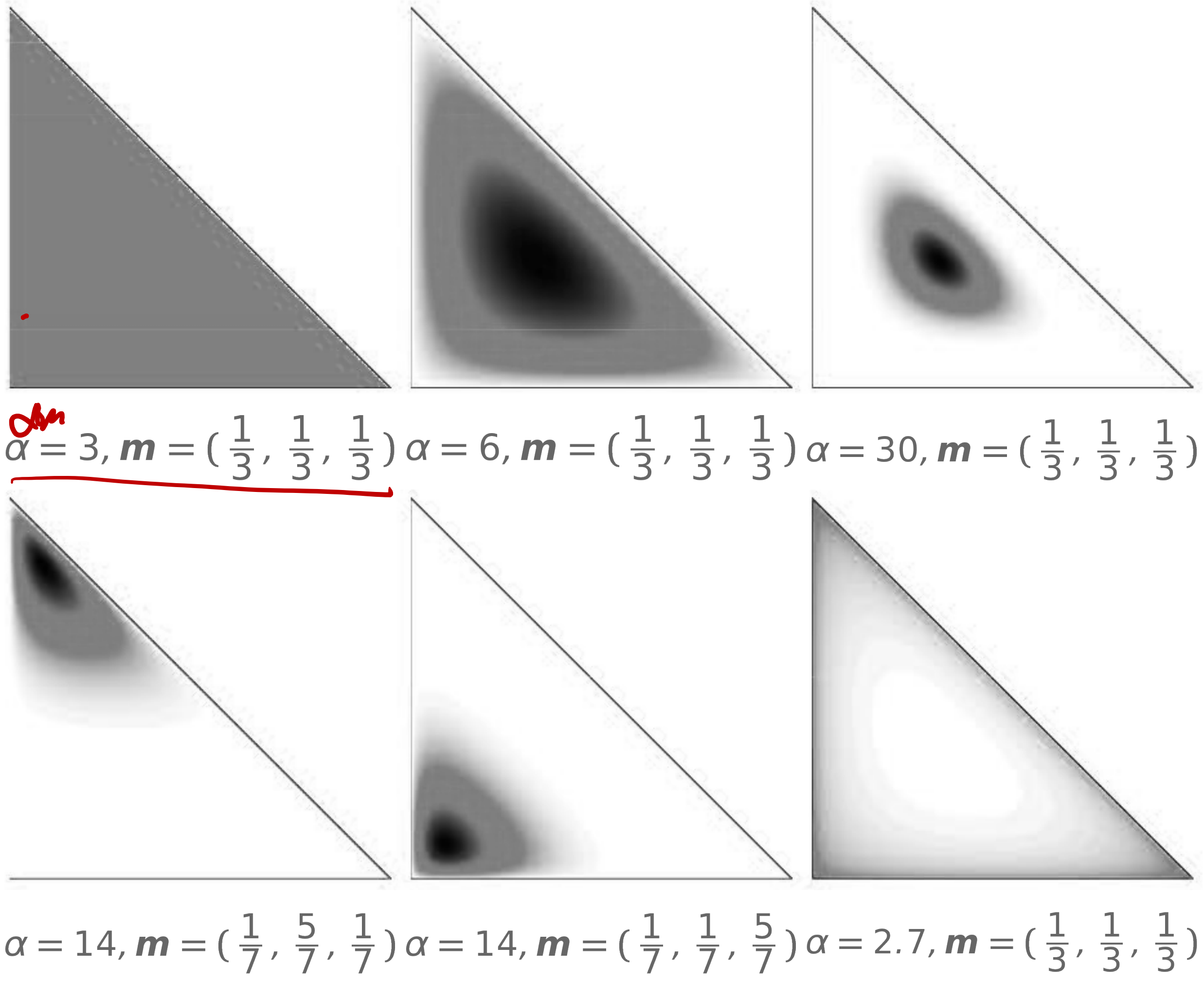
$\alpha m_k = d_k$

$\sum_k m_k = 1$

$P(p \mid \alpha m) \propto \prod_k p_k^{\alpha m_k - 1}$

Missing component: how to generate a multinomial distribution

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



Conjugacy of Dirichlet and Multinomial

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

$$p(\phi|\alpha, \boldsymbol{w}) \propto p(\boldsymbol{w}|\phi)p(\phi|\alpha) \tag{1}$$

$\alpha_w = \alpha m_w$

$$\propto \prod_w p_w^{c_w} \prod_w p_w^{\alpha_w - 1}$$

$$= \prod_w p_w^{c_w + \alpha_w - 1}$$

$(c_w + \alpha_w)$

$\phi = (\phi_1, \phi_2, \phi_3)$

$(1, 2)$ $\quad (1, 1)$ $\quad (2, 2, 3)$

$\phi_1 \phi_2$ $\quad \phi_1^2$ $\quad \phi_2^2 \phi_3$

Conjugacy of Dirichlet and Multinomial

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

$$
\begin{aligned}
p(\phi|\alpha, \boldsymbol{w}) &\propto p(\boldsymbol{w}|\phi)p(\phi|\alpha) & (1)\\
&\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} & (2)\\
&\propto \prod_k \phi^{\alpha_k + n_k - 1} & (3)
\end{aligned}
$$

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

Outline

- Generative story for latent Dirichlet allocation
- Plate notations
- Evaluations of topic models

Revisiting Gaussian mixture models

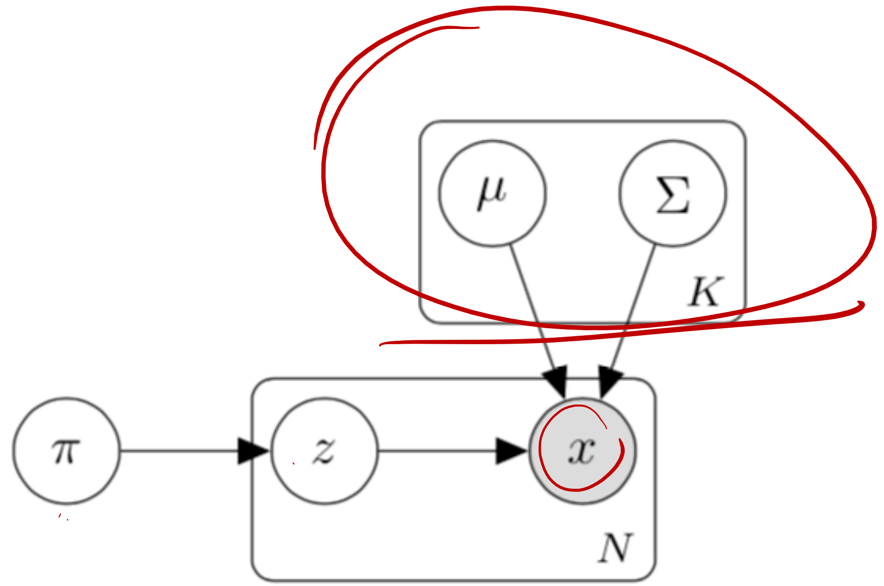
We assume each data point is generated in two steps:

1. Cluster assignment, z_i comes from a multinomial distribution (think of rolling a die);
2. Data comes from a Gaussian distribution, $p(\mathbf{x}_i \mid z_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$ (given a k, $\mathbf{x}_i$ is multivariate Gaussian).

Revisiting Gaussian mixture models

We assume each data point is generated in two steps:

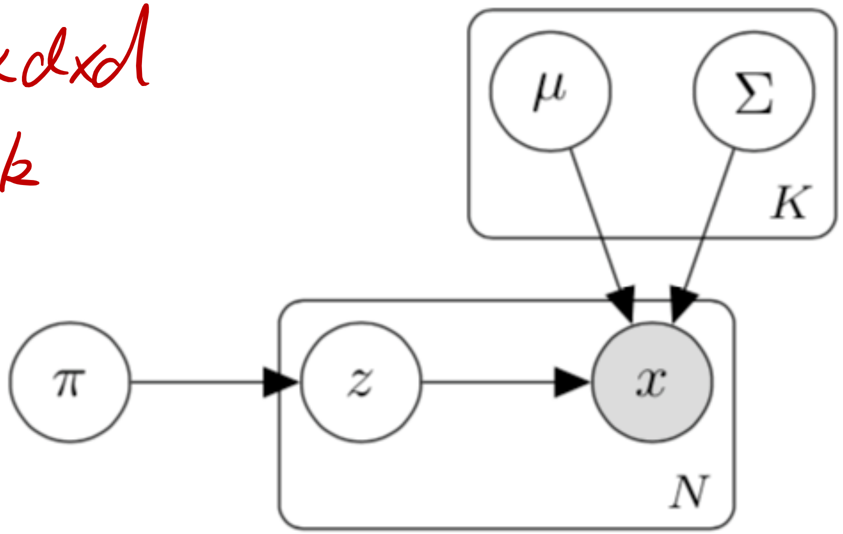
1. Cluster assignment, z_i comes from a multinomial distribution (think of rolling a die);
2. Data comes from a Gaussian distribution, $p(\mathbf{x}_i \mid z_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$ (given a k, $\mathbf{x}_i$ is multivariate Gaussian).



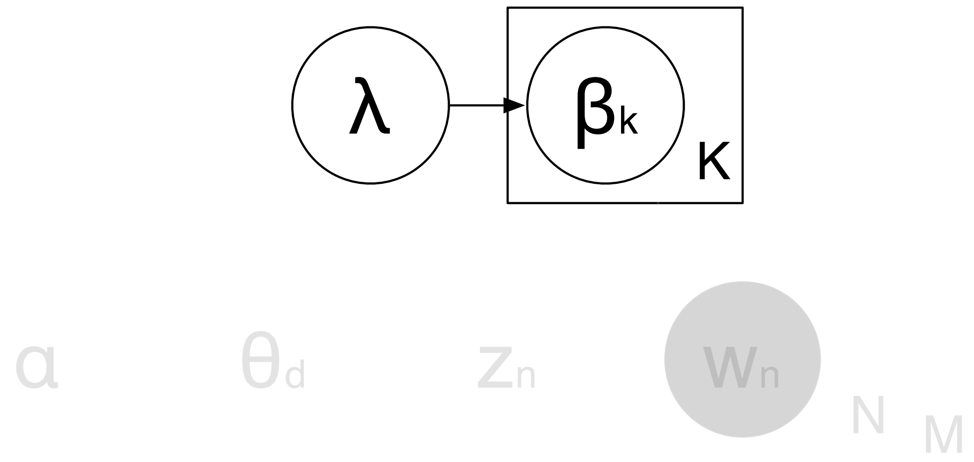
Revisiting Gaussian mixture models

- Given N data points in R^d, K clusters, how many parameters are there?

π K
μ $K \times d$
Σ $K \times d \times d$
z $N \times K$

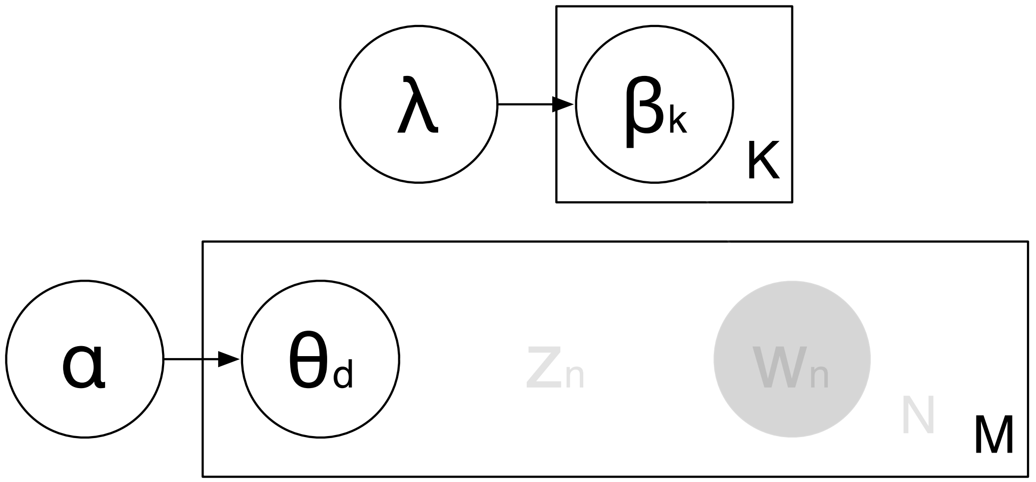


Making the generative story formal



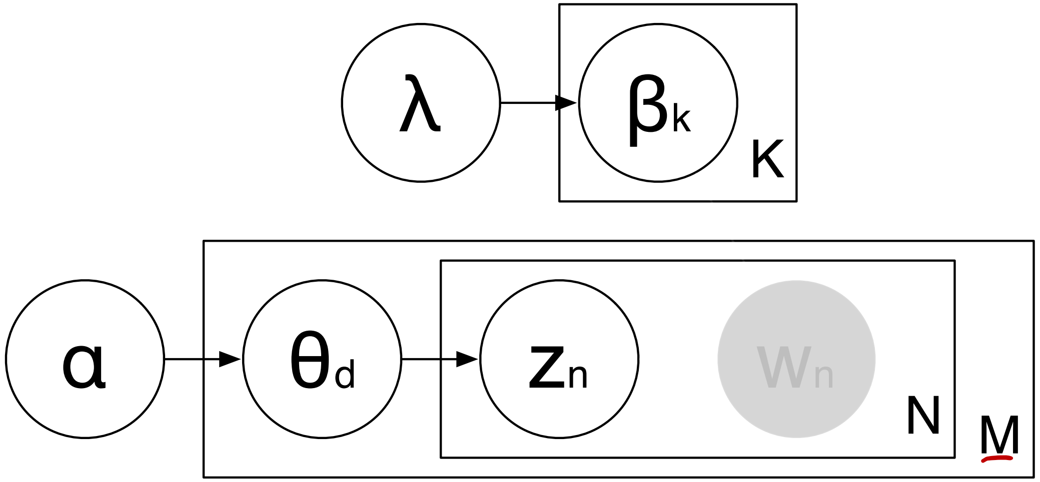
- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Making the generative story formal



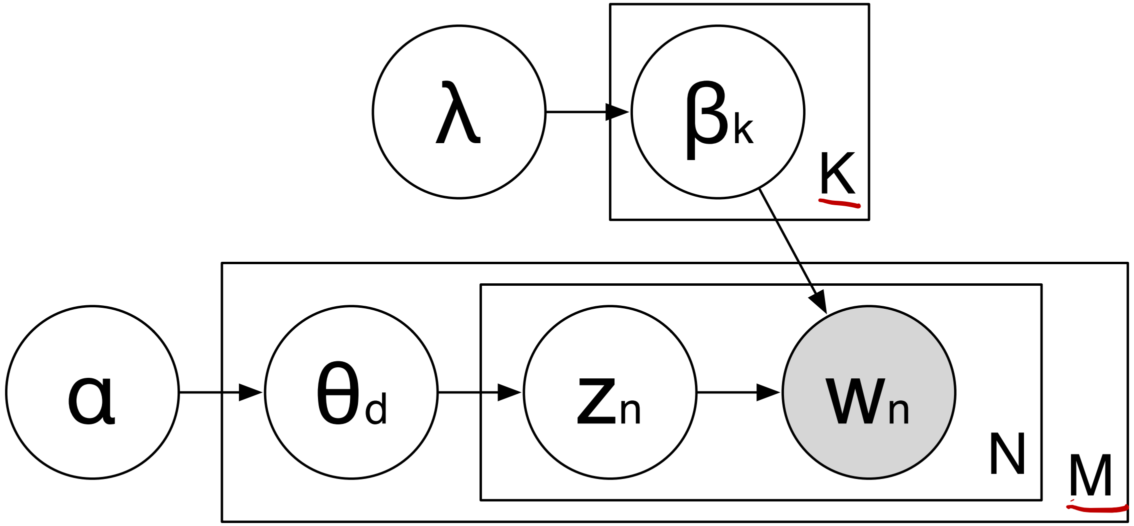
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Making the generative story formal



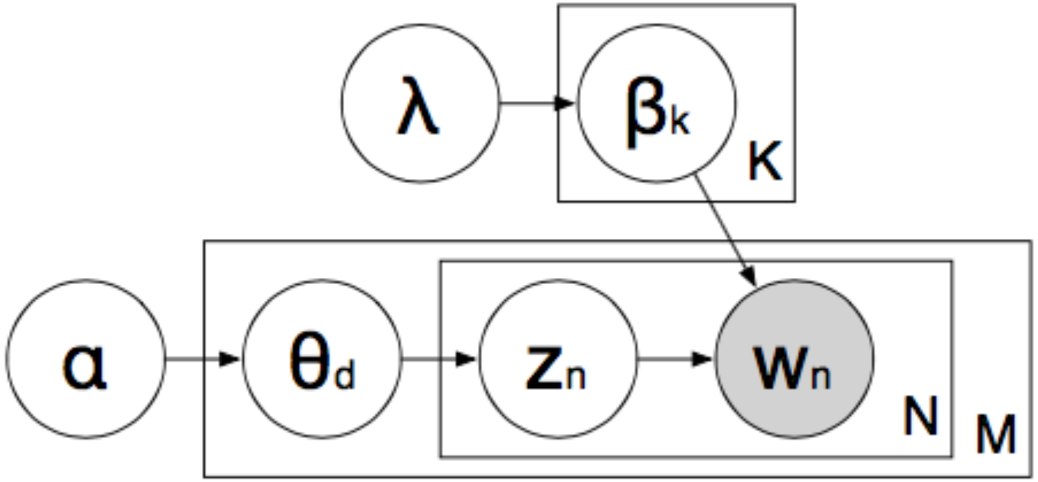
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ_d

Making the generative story formal



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}

Which variables are hidden?



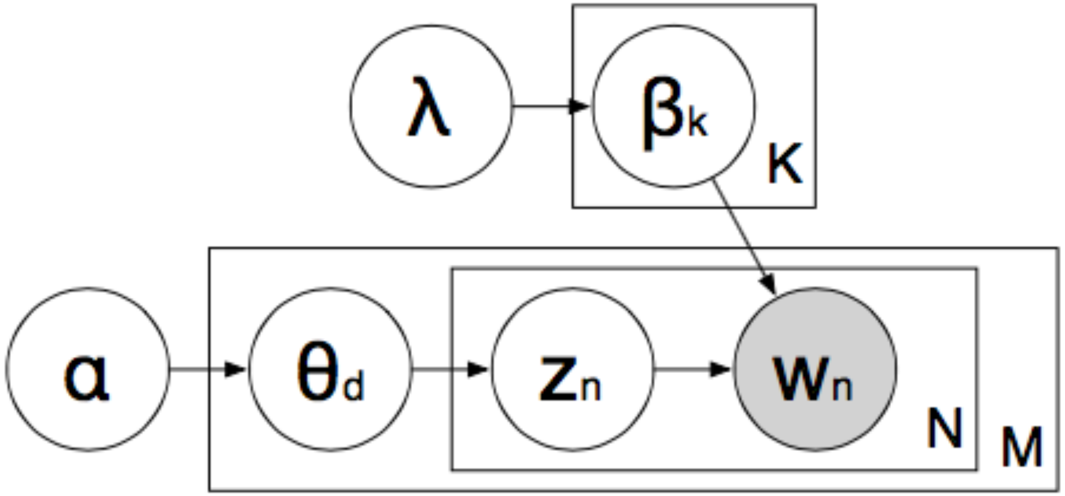
A. β

B. θ

C. β, θ

D. β, θ, z

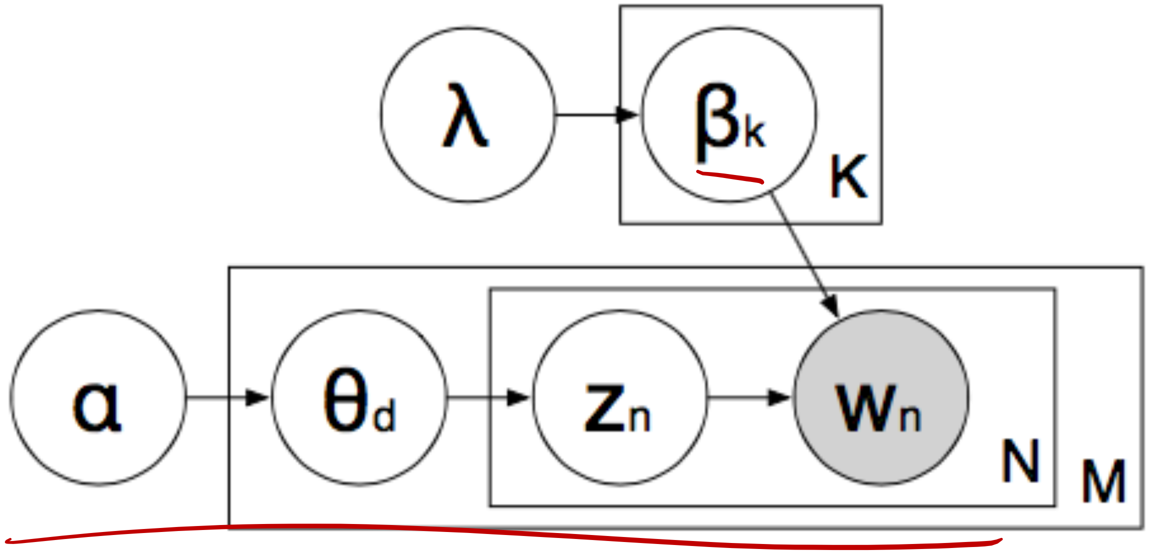
Size of Variable



Given M documents, each document N_d words, vocabulary size V, what is the size of the parameters?

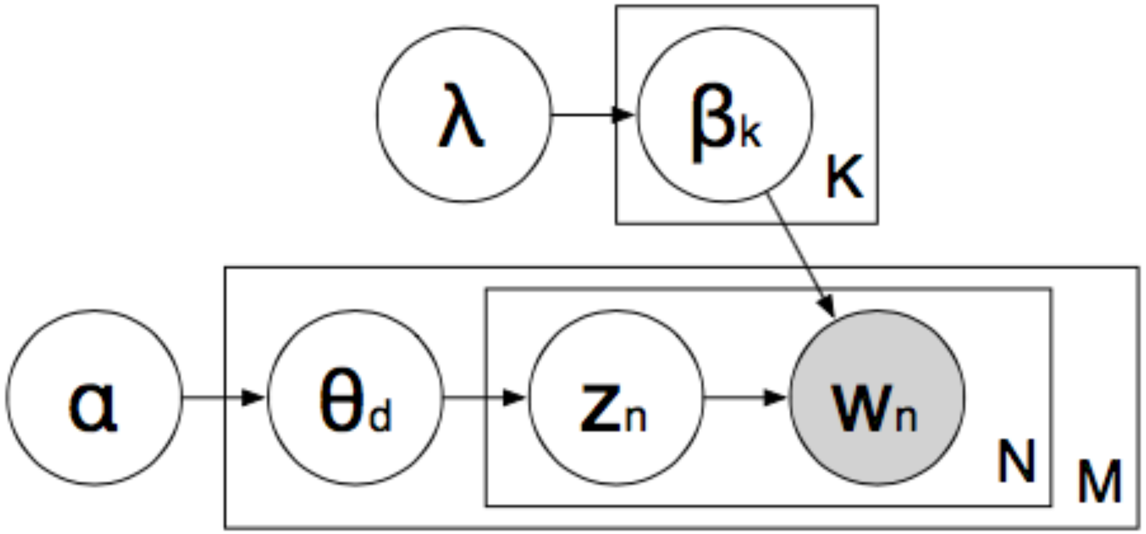
- β K×V
- θ M×K
- z M×N×K

Joint distribution



$$p(\theta, z, w \mid \underline{\alpha, \beta}) = \prod_d P(\theta_d \mid \alpha) \prod_w P(z_w \mid \theta_d) P(w \mid \beta_{z_w})$$

Joint distribution



$$p(\theta, z, w \mid \alpha, \beta) = \prod_{d} p(\theta_d \mid \alpha) \prod_{n} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Joint distribution

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

- $p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_k - 1}$ (Dirichlet)
- $p(z_{d,n} \mid \theta_d) = \theta_{d, z_{d,n}}$ (Draw from Multinomial)
- $p(w_{d,n} \mid \beta, z_{d,n}) = \beta_{z_{d,n}, w_{d,n}}$ (Draw from Multinomial)

Posterior distribution

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Posterior distribution:

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

$$p(w \mid \alpha, \beta) = \int_{\theta, z} p(\theta, z, w \mid \alpha, \beta)$$

$$= \prod_d \int_{\theta_d} p(\theta_d \mid \alpha) \prod_n \sum_{z_{d,n}} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

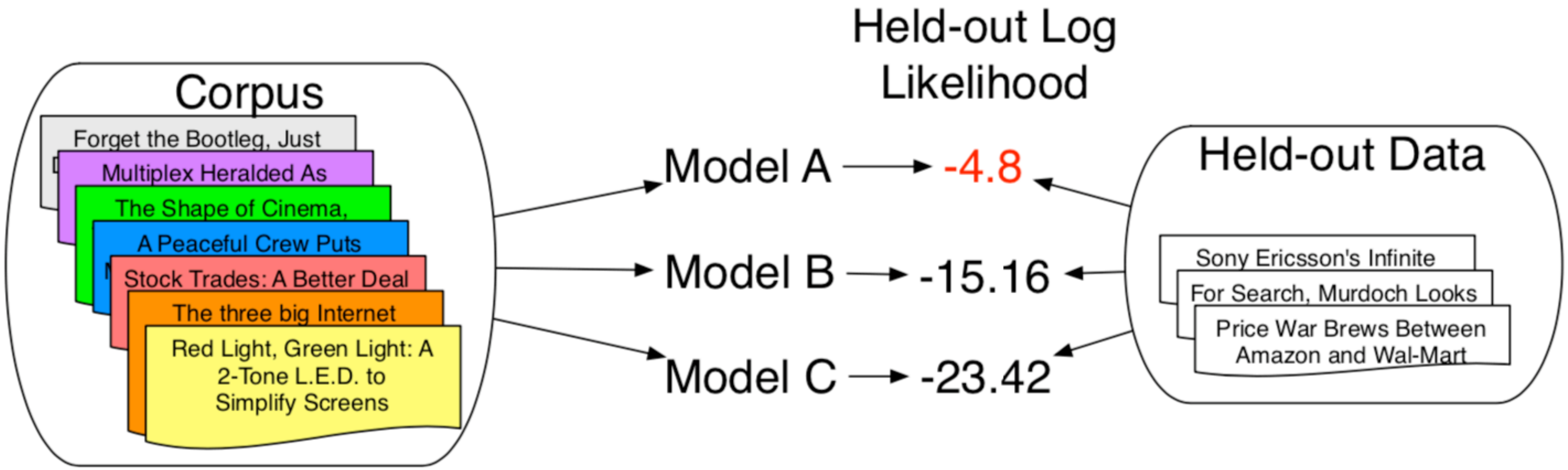
Outline

- Generative story for latent Dirichlet allocation
- Plate notations
- Evaluations of topic models

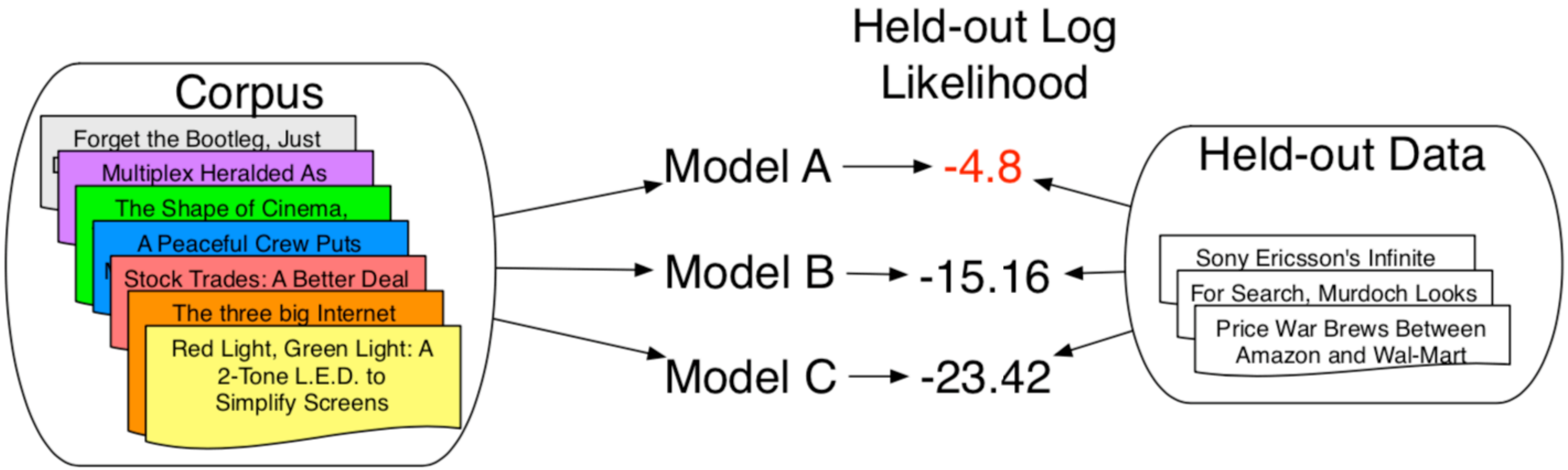
Evaluating Topic Models

- Held-out log likelihood
- Word intrusion

Held-out log likelihood



Held-out log likelihood



$$p(w \mid \alpha, \beta) = \int_{\theta, z} p(\theta, z, w \mid \alpha, \beta)$$

$$= \prod_{d} \int_{\theta_d} p(\theta_d \mid \alpha) \prod_{n} \sum_{z_{d,n}} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Word Intrusion

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	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

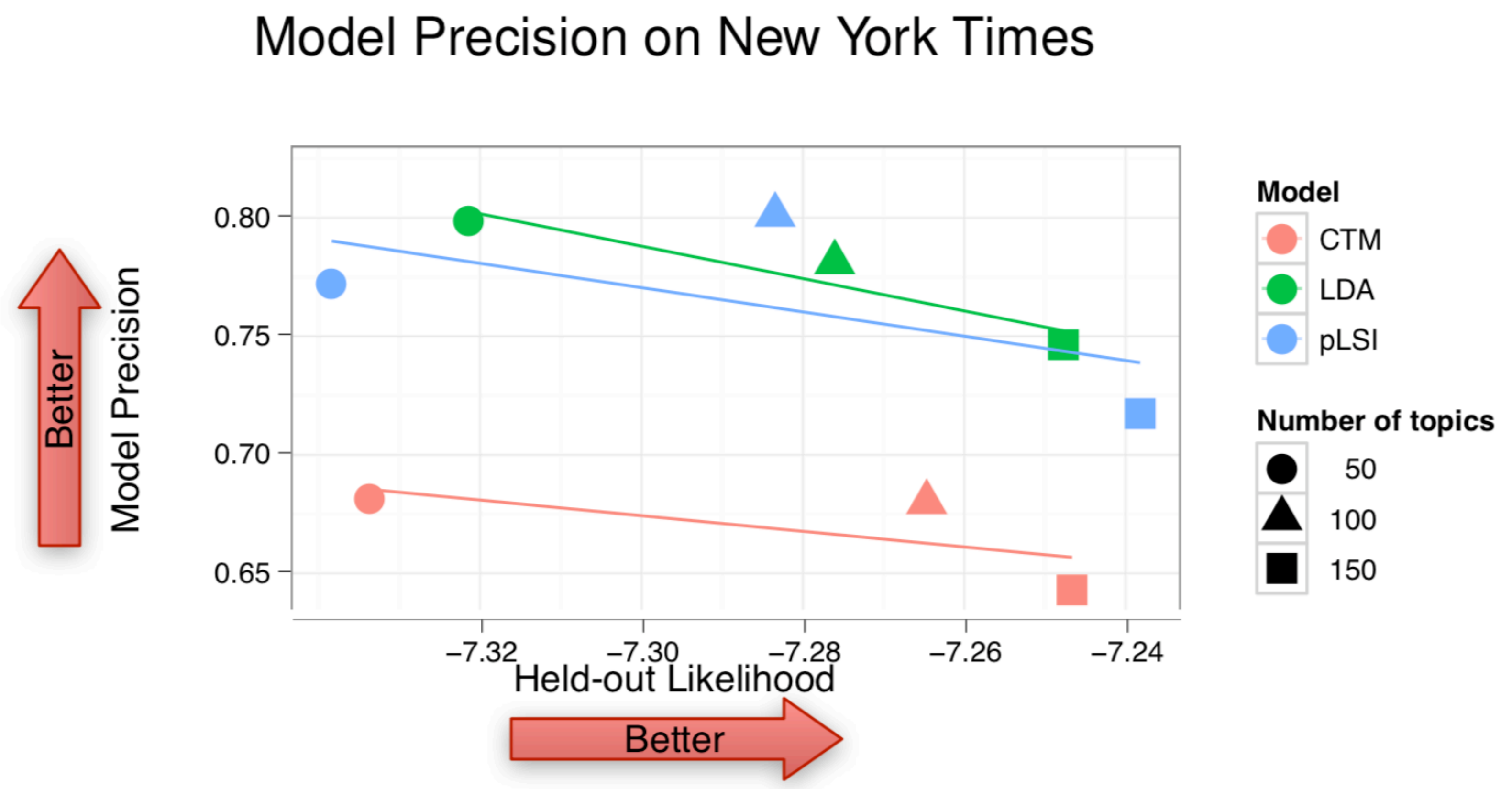
dog, cat, apple, horse, pig, cow

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

Hypothesis

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

Interpretability and likelihood



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

Evaluation takeaway

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

Topic models: What’s important

- Topic models (latent variables)
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Modeling & Algorithm
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces