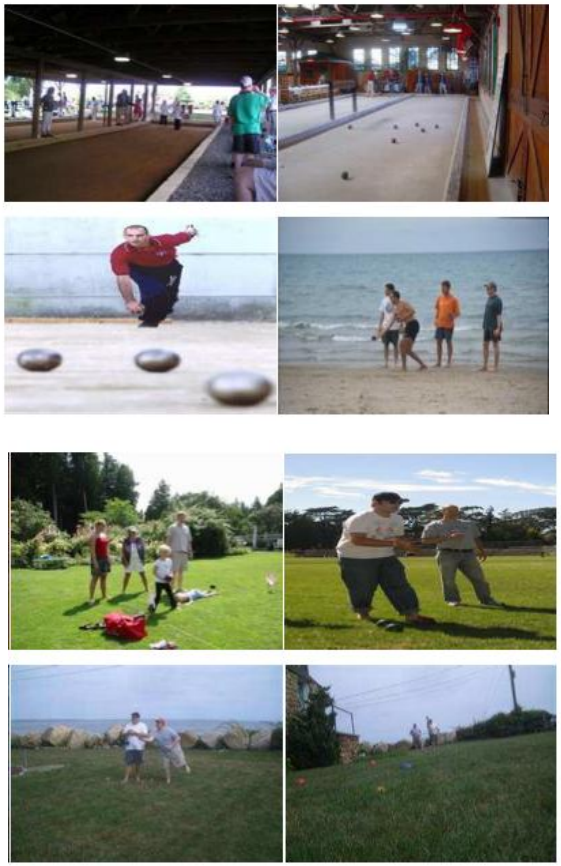
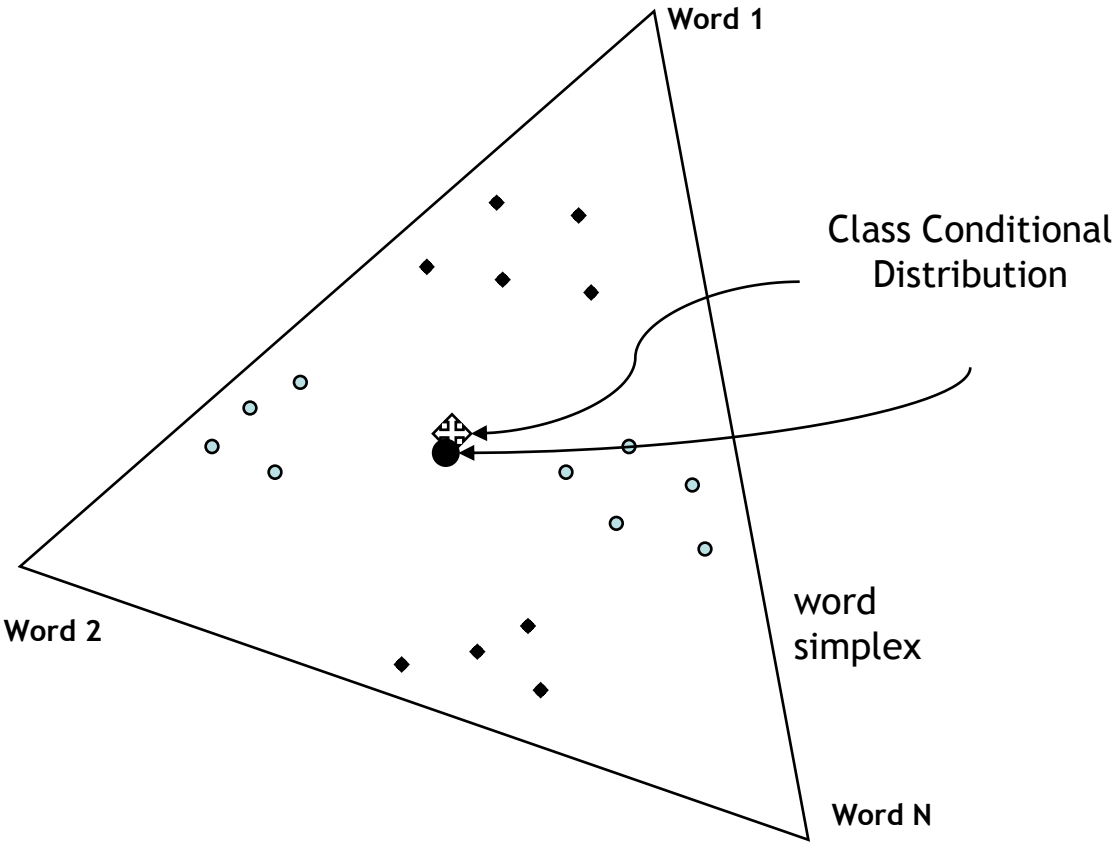
Problem of:

- **Intra-class variations.**





Topic Models

- Different words are not independent and usually there is a **semantic relationship between words**.
- Introduce "**topics**" to capture relationship between words
- A document may contain multiple "topics"

Note: For the next few slides we will not worry about classification and focus on "capturing the semantic relationship between words"

Intuition

1. Words are correlated
2. Document has many topics

Facebook

From Wikipedia, the free encyclopedia

Facebook is a social networking website that was launched on February 4, 2004. The website is owned and operated by Facebook, Inc., the parent company of the website and a privately held company. The free-access website allows users to join one or more networks, such as a school, place of employment, or geographic region to easily connect with other people in the same network. The name of the website refers to the paper facebooks depicting members of a campus community that some American colleges and preparatory schools give to incoming students, faculty, and staff as a way to get to know other people on campus.

Mark Zuckerberg founded Facebook while still a student at Harvard University. Website membership was initially limited to only Harvard students, but was later expanded to include any university student, then high school students, and finally to anyone aged 13 and over.

The website has more than 64 million active users worldwide.[3] From September 2006 to September 2007, the website's ranking among all websites, in terms of traffic, increased from 60th to 7th, according to Alexa.[4] It is also the most popular website for uploading photos, with 14 million uploaded daily.[3] Due to the website's popularity, Facebook has met with some criticism and controversy in its short lifespan because of privacy concerns, the political views of its founders, and censorship issues.

topic: Social network website

topic: education

topic: criticism

Intuition: Images

- Although it is arguable, but for the sake of understanding:
  - Document ⇔ Image
  - Topic ⇔ Objects
  - Words ⇔ Visual words

zebra

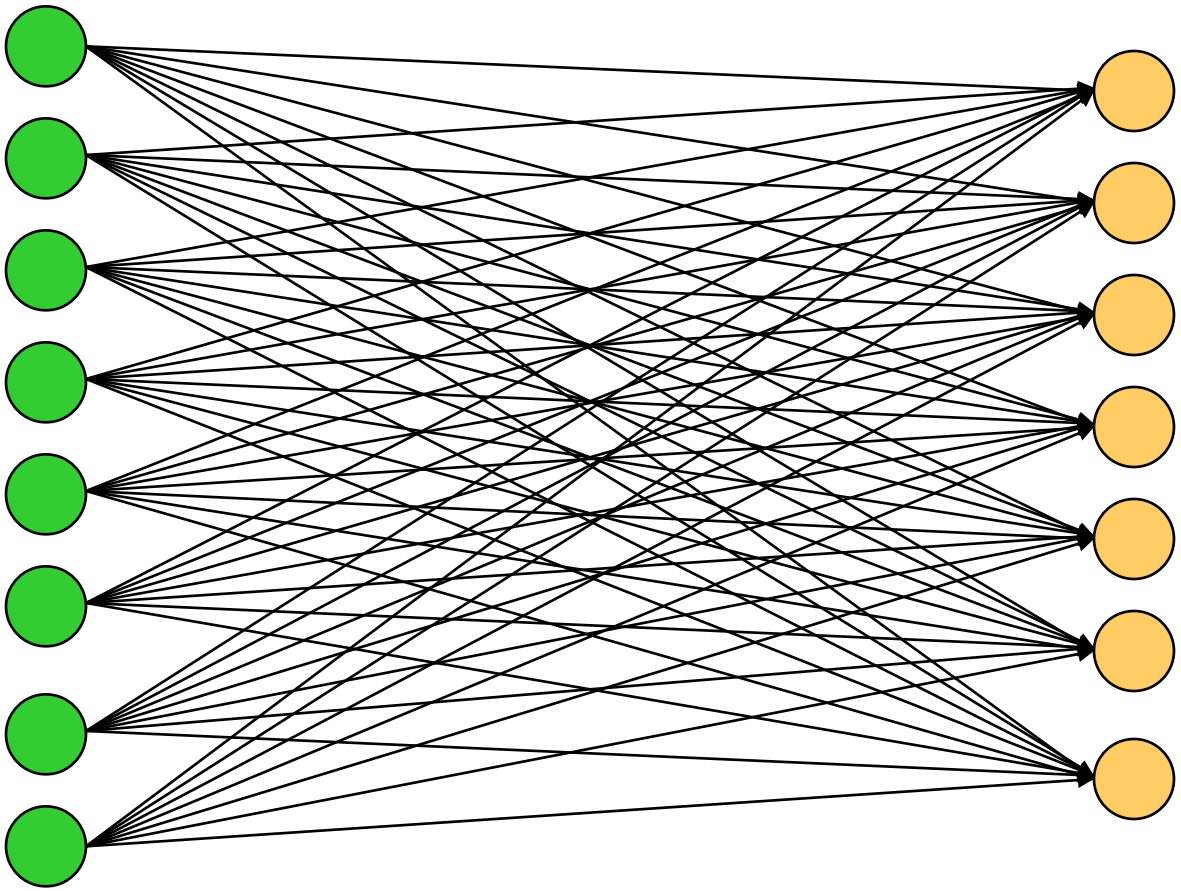
grass

tree

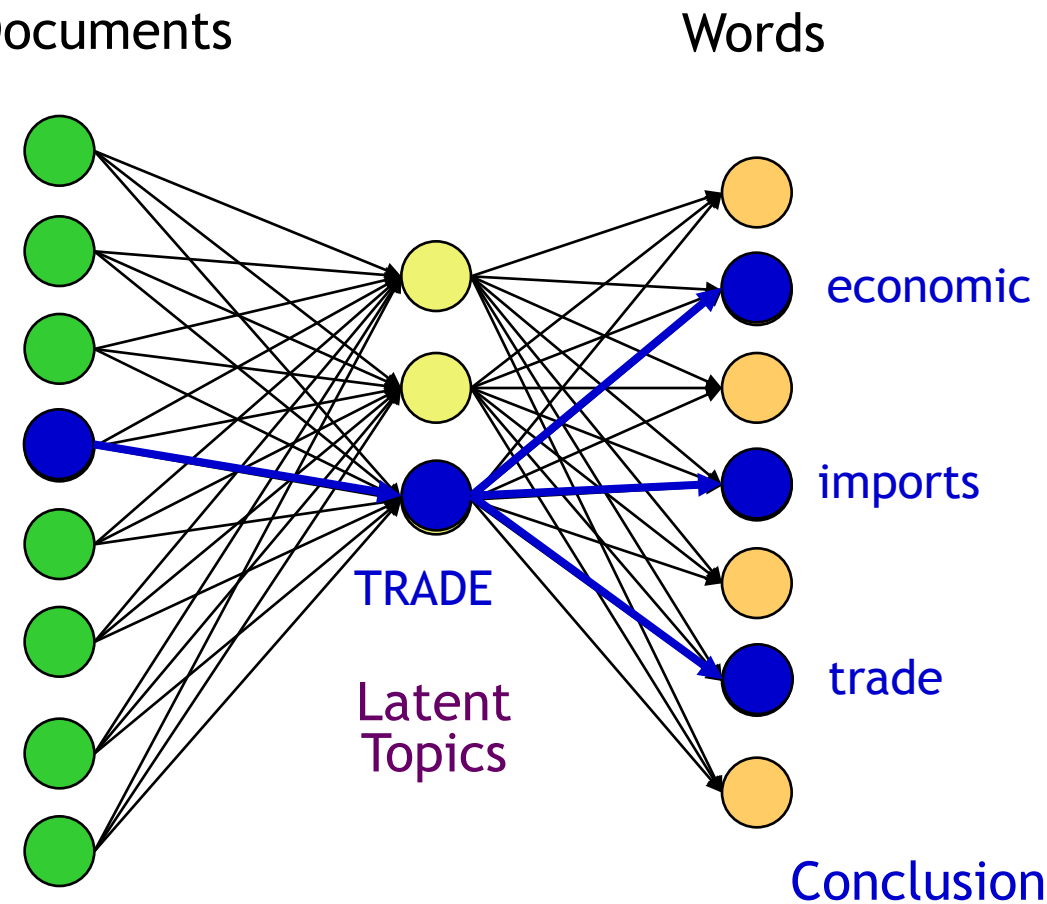
"visual topics"

Documents

Words



Topic Models



Topic probabilities are estimated based on all documents that are dealing with a topic.

"Unmixing" of superimposed concepts is achieved by statistical learning algorithm.

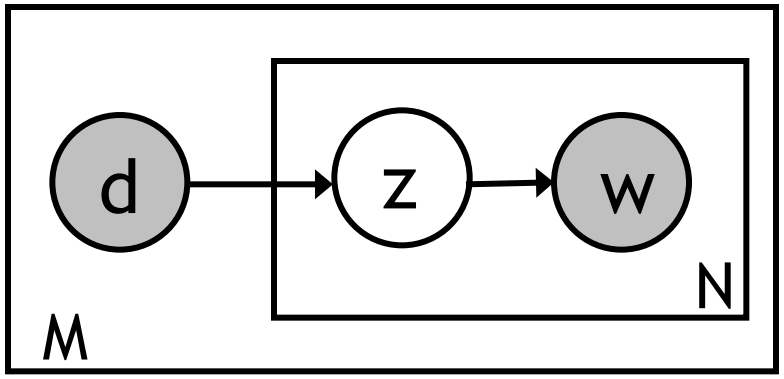
Conclusion: ⇒ No prior knowledge about concepts required, context and term co-occurrences are exploited

General Idea of Probabilistic Topic Models

- Cast this intuition into a generative probabilistic process
  - Each document is a mixture of corpus-wide topics
  - Each word is drawn from one of those topics
- Since we only observe the documents, need to figure out (Estimation/Inference)
  - What are the topics?
  - How are the documents divided according to those topics?
- Two basic models: PLSA and LDA

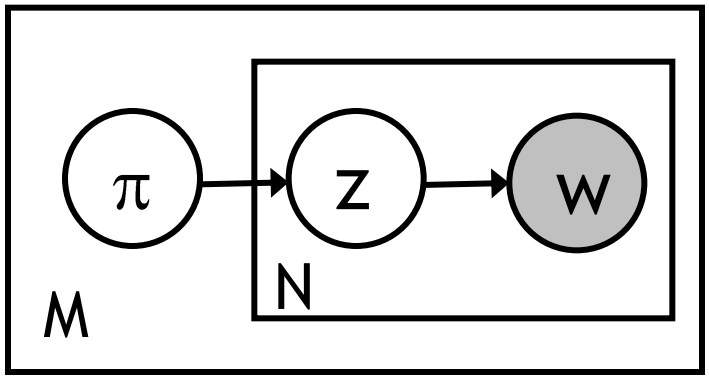
Probabilistic Topic Models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 1999

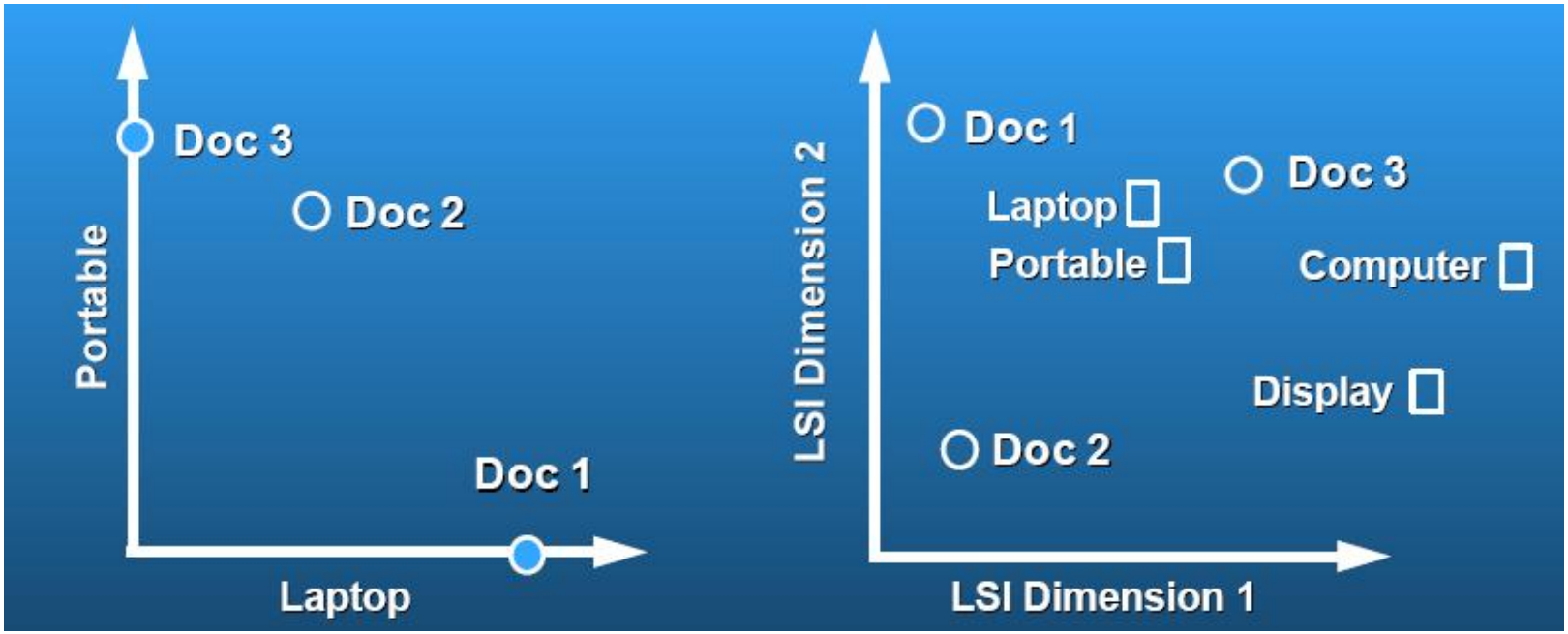
Latent Dirichlet Allocation (LDA)



Blei et al., 2001

Lets start slow…

- **Latent semantic Analysis**: illustrating example



courtesy of Susan Dumais
© Bellcore

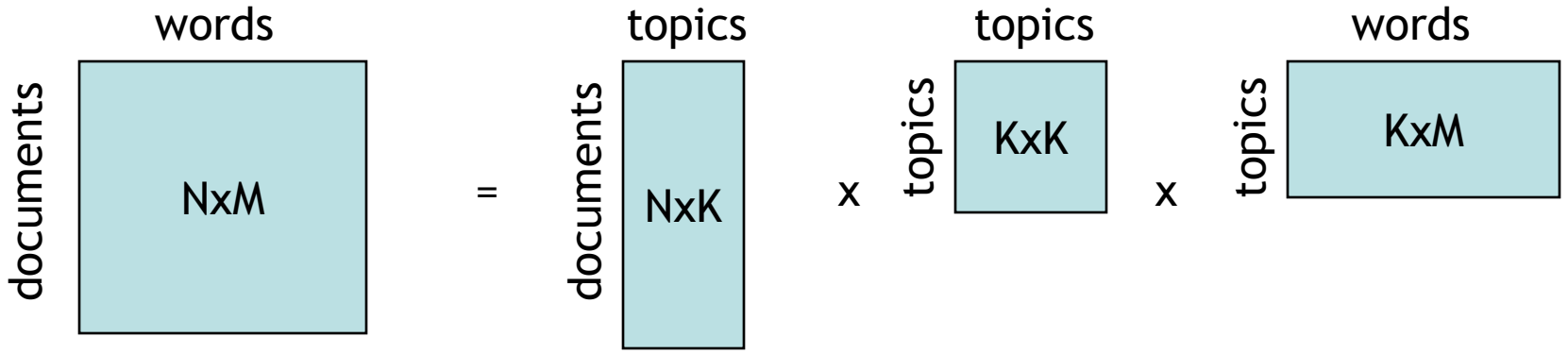
Latent Semantic Analysis

- 1990: Latent Semantic Analysis
  - Perform a **low-rank approximation** of **document-term matrix**
  - Map documents (and terms) to a **low-dimensional** representation.
  - Design a mapping such that the low-dimensional space reflects **semantic associations** (latent semantic space).

- Goals
  - Similar terms map to similar location in low dimensional space
  - Noise reduction by dimension reduction

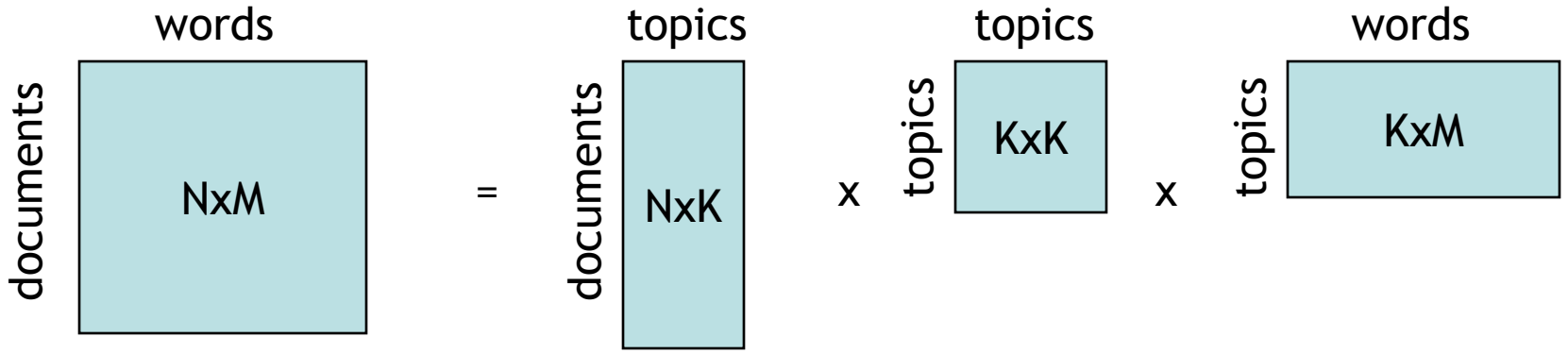
1990: Latent Semantic Analysis (LSA)

- $D = \{d_1,\ldots,d_N\}$ N documents
- $W = \{w_1,\ldots,w_M\}$ M words
- $N_{ij} = \#(d_i,w_j)$ NxM co-occurrence term-document matrix



What did we just do?

Singular Value Decomposition



Problems with LSA

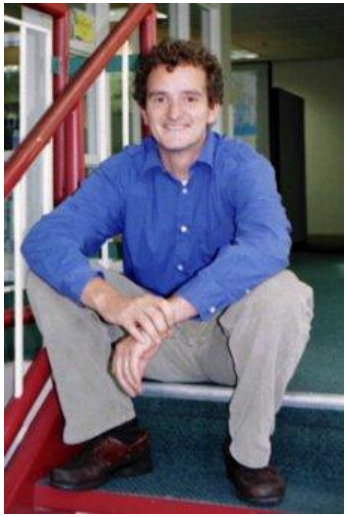
- LSA does not define a properly normalized **probability distribution**
- No obvious interpretation of the directions in the latent space
- Polysemy problem: LSI does not deal with the problem of polysemy.

Probabilistic Latent Semantic Analysis/Indexing [Hofmann 99]

Probabilistic Latent Semantic Analysis

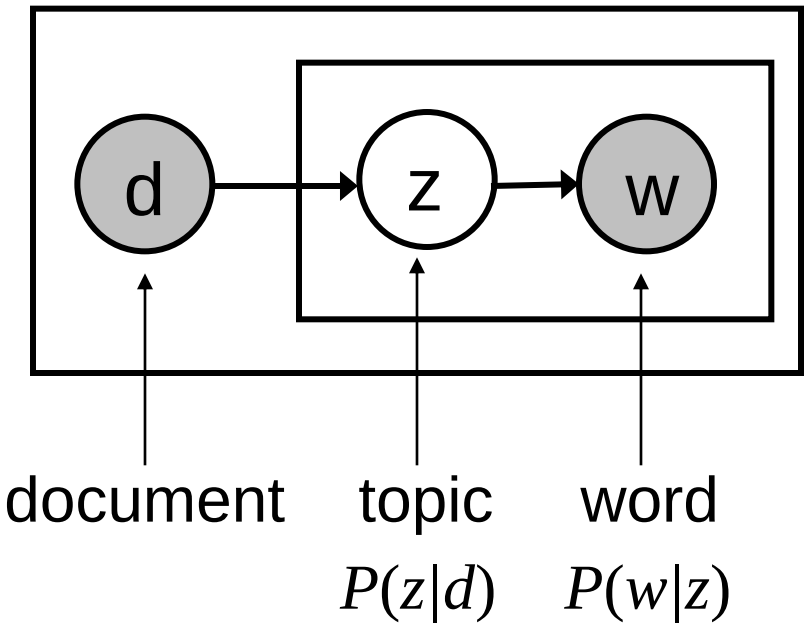
To appear in: Uncertainity in Artificial Intelligence, UAI'99, Stockholm

Thomas Hofmann
EECS Department, Computer Science Division, University of California, Berkeley &
International Computer Science Institute, Berkeley, CA
hofmann@cs.berkeley.edu



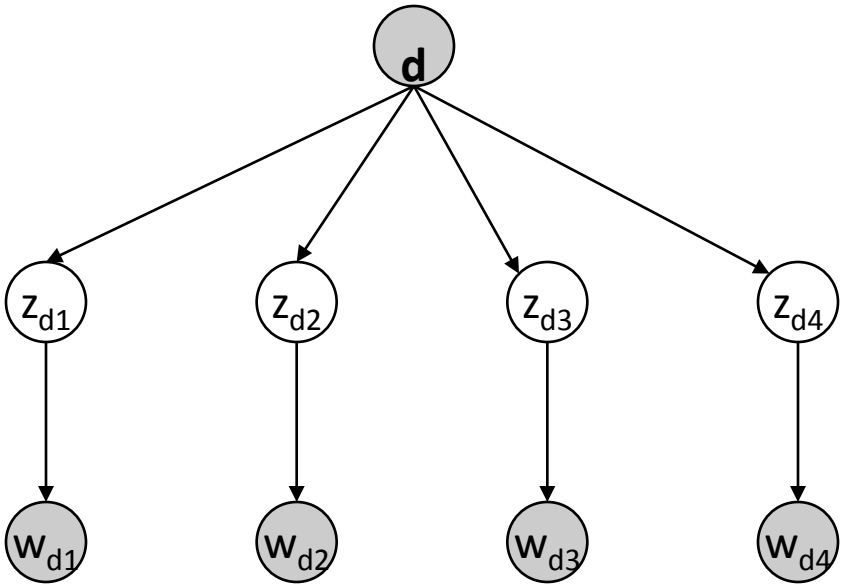
Probabilistic Latent Semantic Analysis

- Unsupervised technique
- **Two-level generative model**: a document is a mixture of topics, and each topic has its own characteristic word distribution



T. Hofmann, Probabilistic Latent Semantic Analysis, UAI 1999

The pLSA Model



Probabilistic Latent Semantic Indexing (pLSI) Model

For each word of document d in the training set,

- Choose a topic z according to a multinomial conditioned on the index d.

$$p(z|d)$$

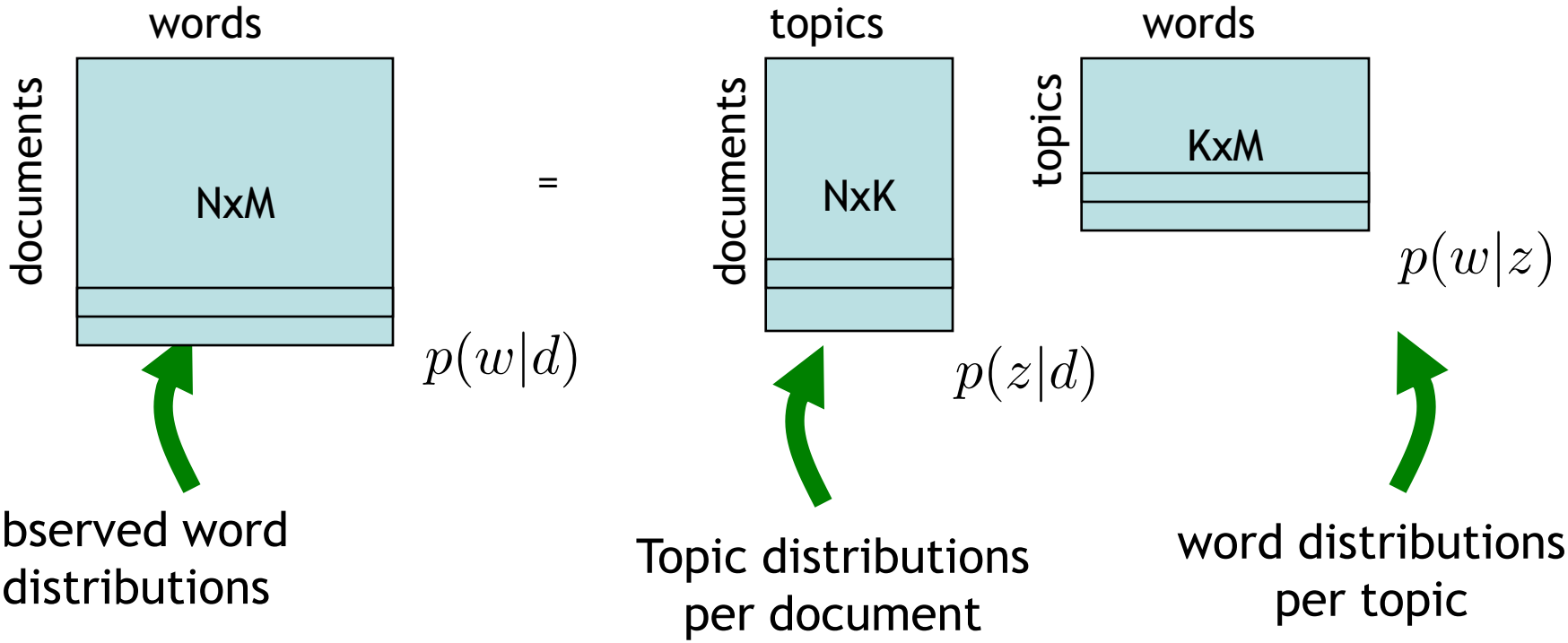
- Generate the word by drawing from a multinomial conditioned on z.

$$p(w|z)$$

pLSA to the rescue

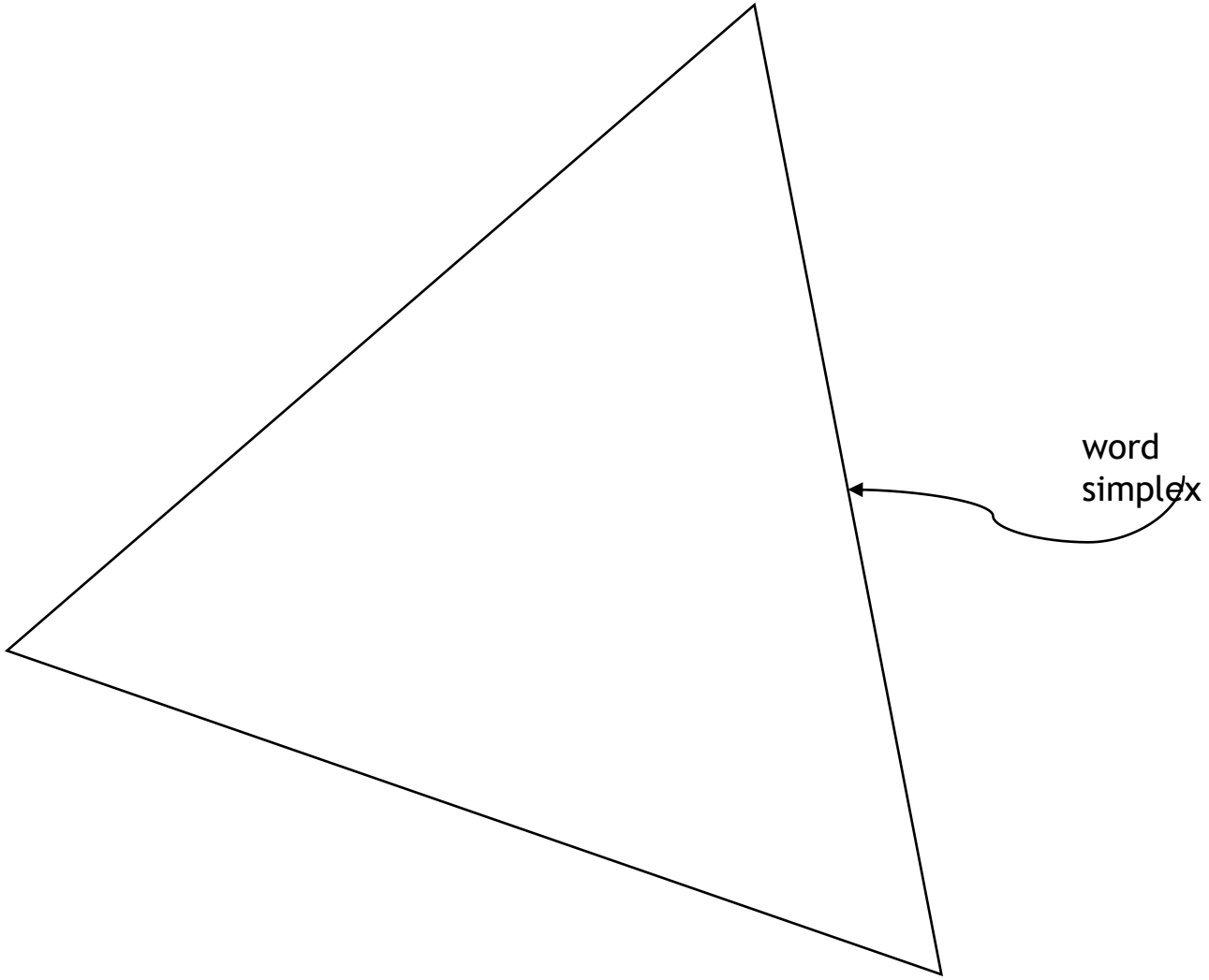
Decomposition into Probabilities!

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$

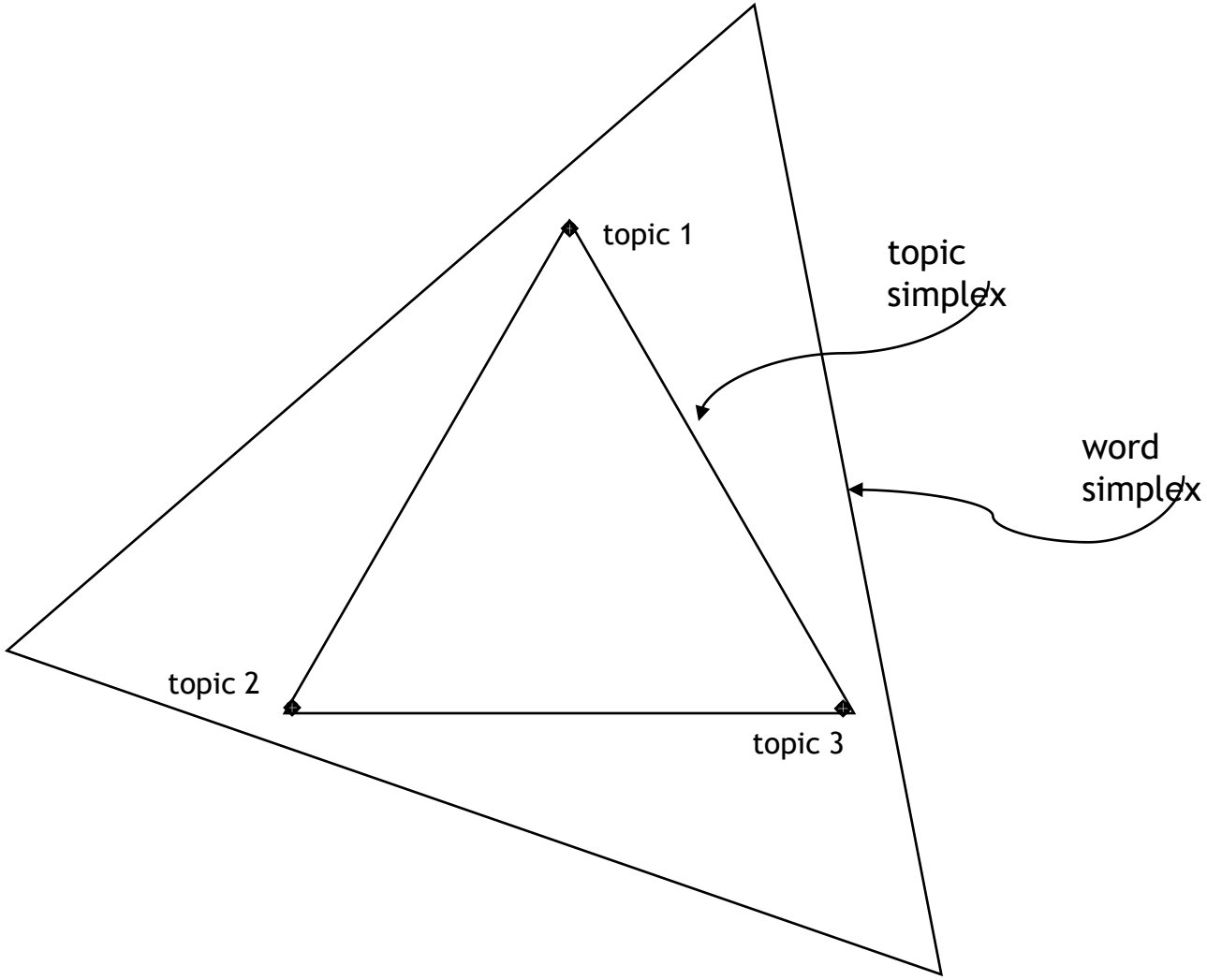


Slide credit: Josef Sivic

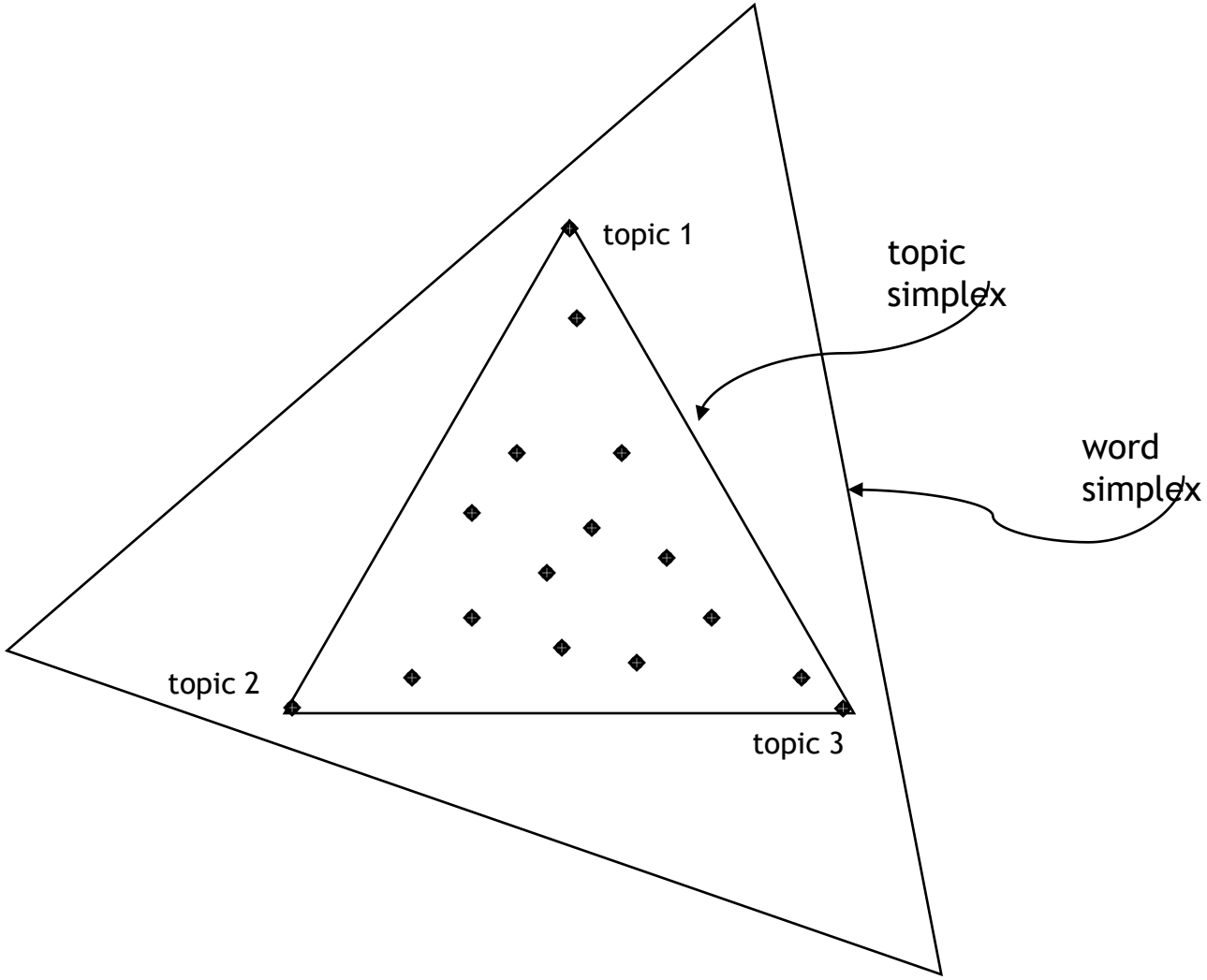
A geometric interpretation



A geometric interpretation



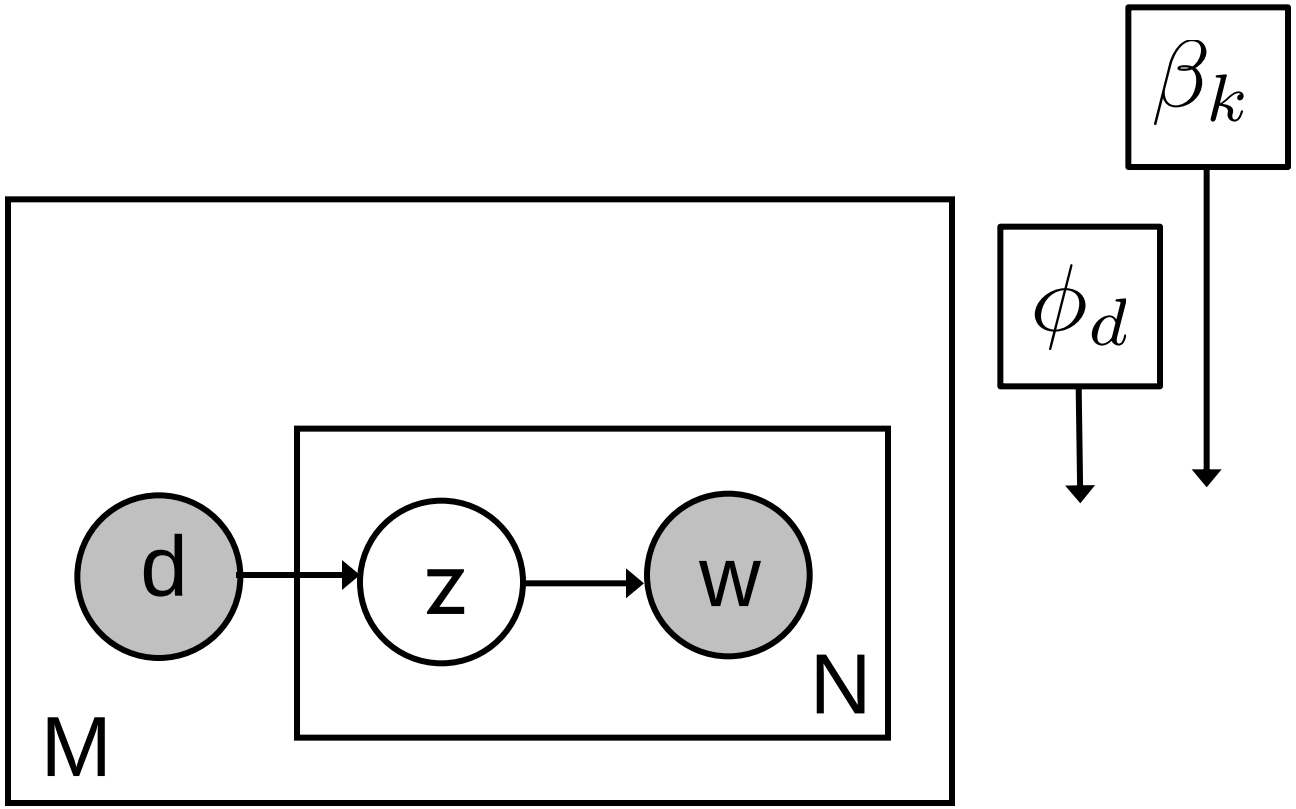
A geometric interpretation



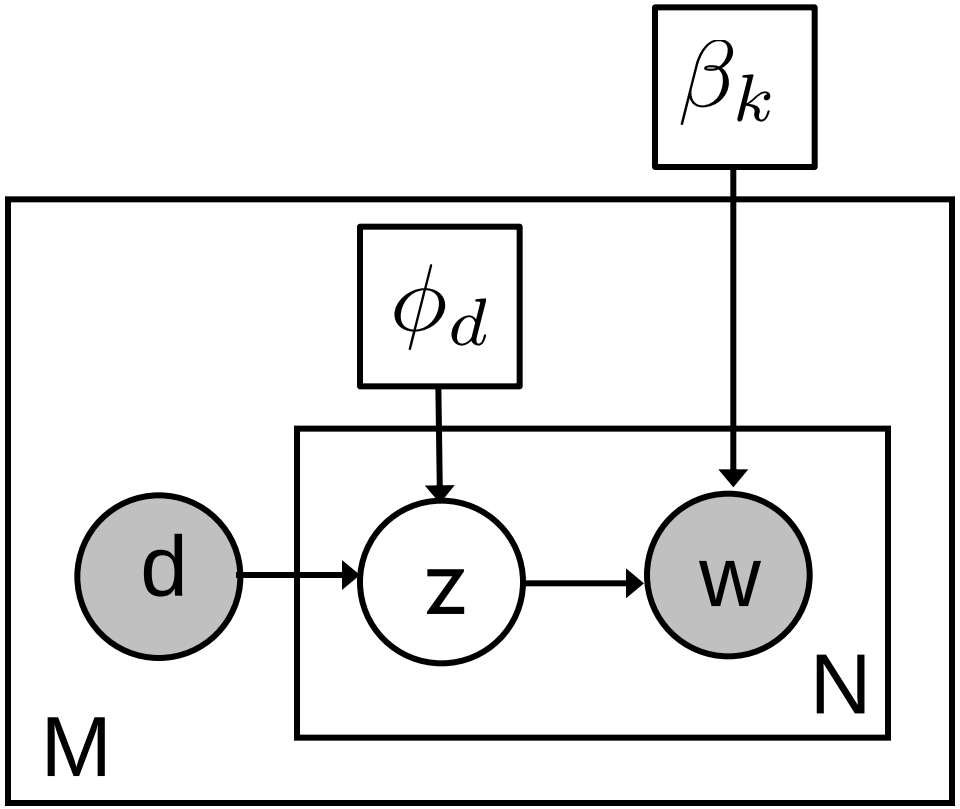
Full Graphical Model

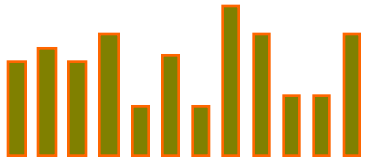
Parameters?

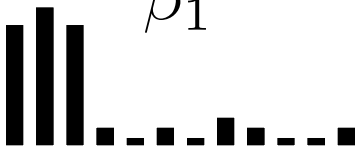
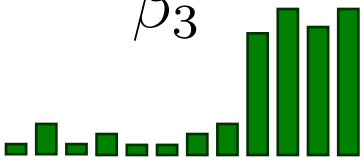
Where to place them?



Full Graphical Model





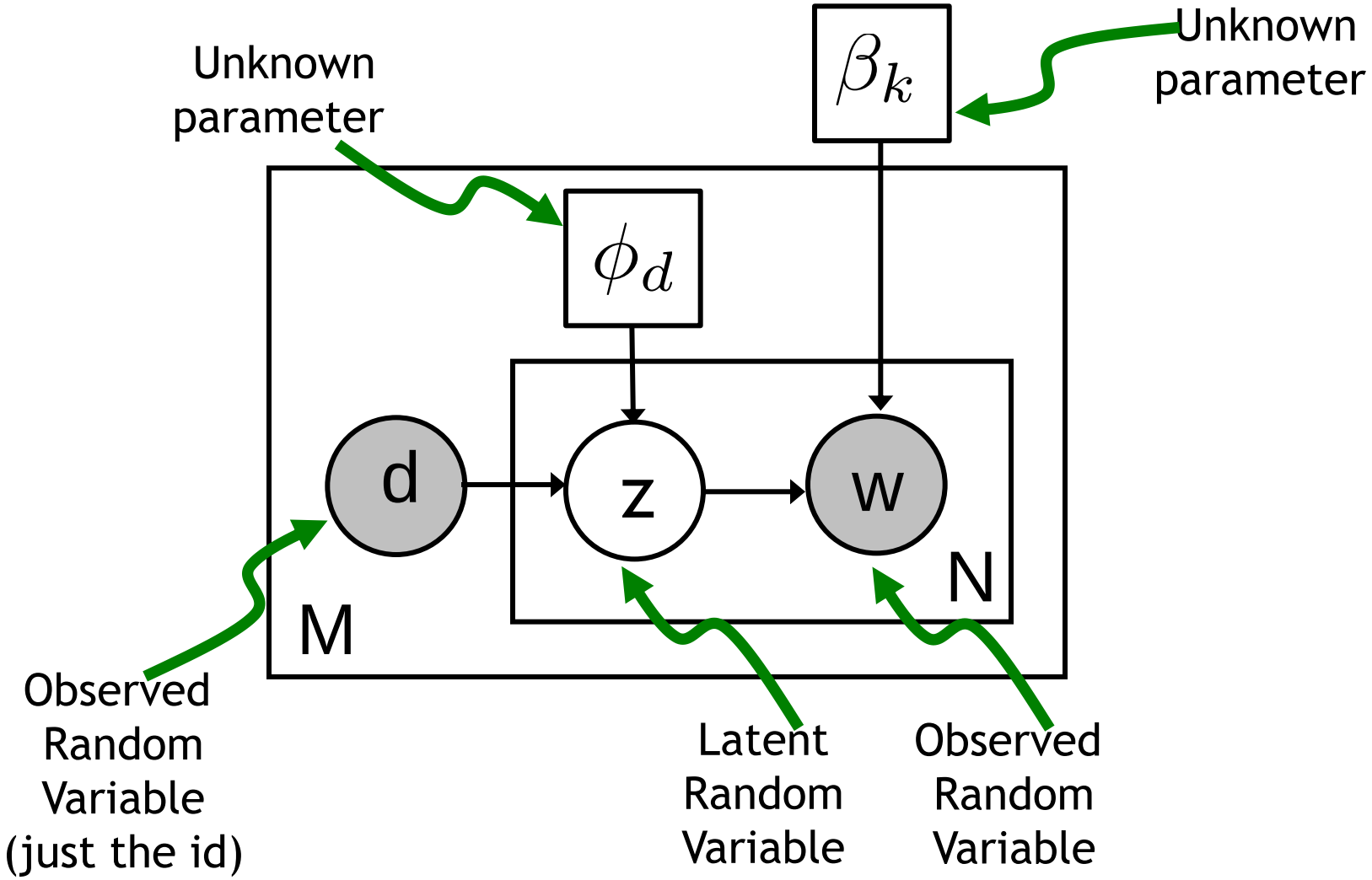
$= \phi_1^d$  β_1 $+ \phi_2^d$  β_2 $+ \phi_3^d$  β_3

Wildlife Safari = zebra + grass + tree

"visual topics"

T. Hofmann, Probabilistic Latent Semantic Analysis, UAI 1999

Graphical Model – Parameter Learning



Learning: Maximum likelihood estimation

Maximize (log)likelihood of data:

$$\log L = \sum_{m=1}^{M} \log p(\mathbf{w}_m, d_m)$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \log p(w_{mn}, d_m)$$

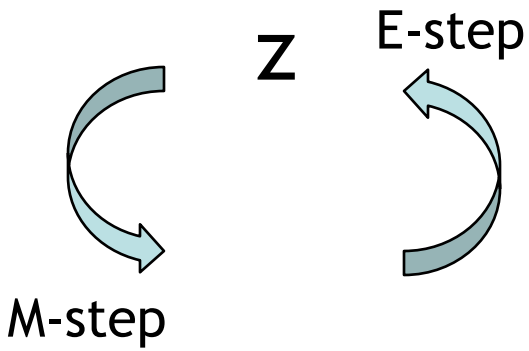
$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \log \sum_{k=1}^{K} p(w_{mn}|z_k) p(z_k|d_m) p(d_m)$$

$$= \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \log \sum_{k=1}^{K} p(w_{mn}|z_k) p(z_k|d_m)$$

The summation inside the log makes it difficult to directly optimize. Hence we use Expectation-Maximization.

EM for pLSA in a nutshell

- The latent variable z is a concern
- Expectation step:
  - Assume the parameters are known and estimate “expected” z
- Maximization step:
  - Given z, estimate the parameters.
- Repeat until convergence



EM for pLSA (training on a corpus)

- **E-step**: posterior probabilities for the latent variables

$$\hat{z}_{mvk} = \frac{\beta_{kv}\phi_{mk}}{\sum_{k=1}^{K}\beta_{lv}\phi_{ml}}$$

Probability that the occurence of term v in document m can be "explained" by concept z

- **M-step**: parameter estimation based on "completed" statistics

$$\phi_{mk} = \frac{\sum_v n(d_m, v)\hat{z}_{mvk}}{N_d}$$

$$\beta_{kv} = \frac{\sum_m n(d_m, v)\hat{z}_{mvk}}{\sum_{v'}\sum_m n(d_m, v')\hat{z}_{mv'k}}$$

- Concepts (10 of 128) extracted from Science Magazine articles (12K)

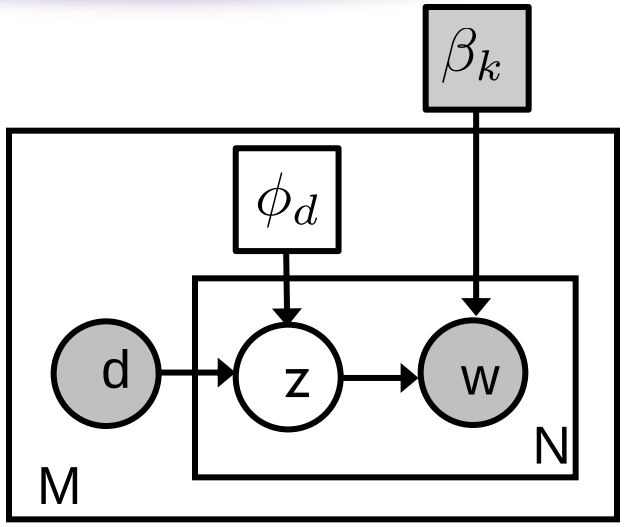
P(w|z)

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148

P(w|z)

bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824

Inference



- "Folding-in" Heuristic
- First train on Corpus to obtain

$$p(w|z)$$

- Now re-run same training EM algorithm, but estimate

$$p(z|d)$$

Problems with pLSA

- **Not a well-defined generative model** of documents; **d is a dummy index** into the list of documents in the training set (as many values as documents)
- **No natural** way to assign probability to a previously unseen document
- Number of **parameters to be estimated grows with size of training set**

Latent Dirichlet Allocation [Blei et al. 03]

Latent Dirichlet Allocation

David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

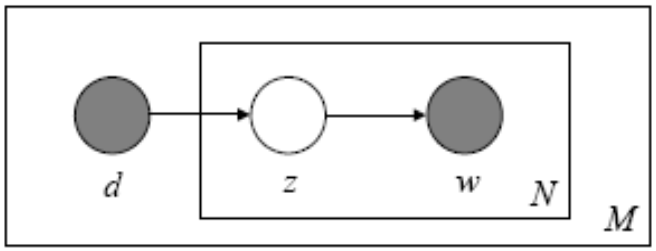
Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

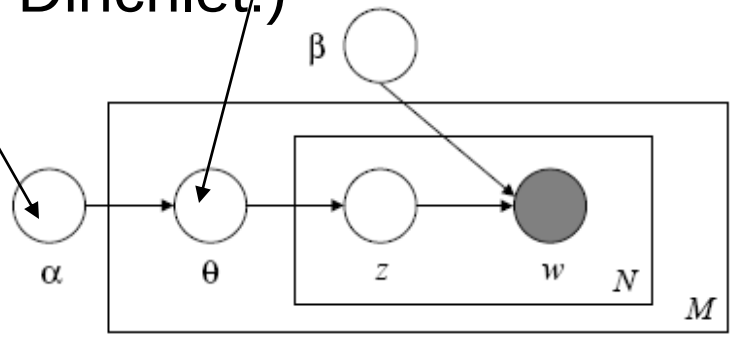
LDA to the rescue

- Latent Dirichlet Allocation treats the topic mixture weights as a k-parameter hidden random variable and places a Dirichlet prior on the multinomial mixing weights

- Dirichlet distribution is conjugate to the multinomial distribution (most natural prior to choose: the posterior distribution is also a Dirichlet!)

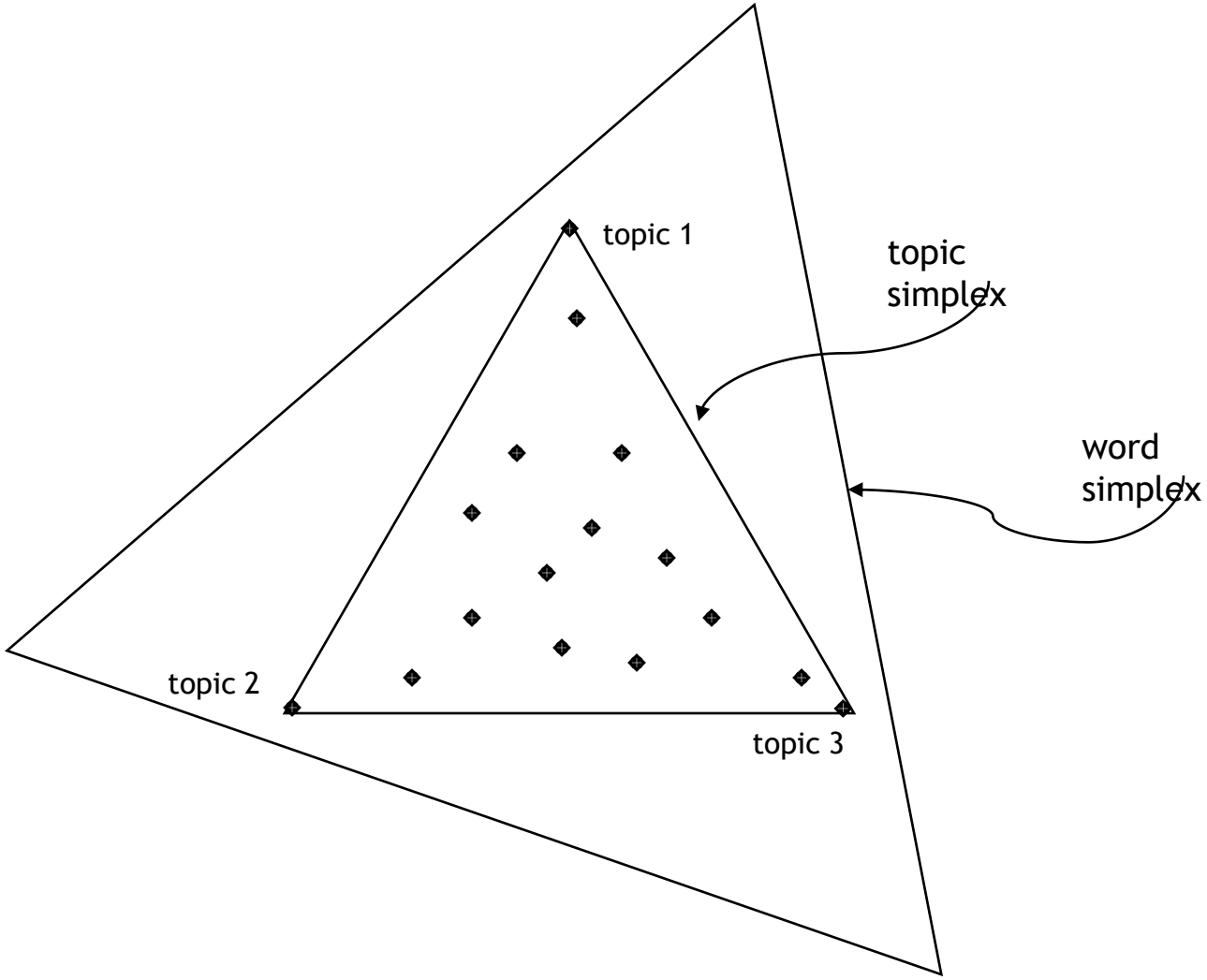


pLSA

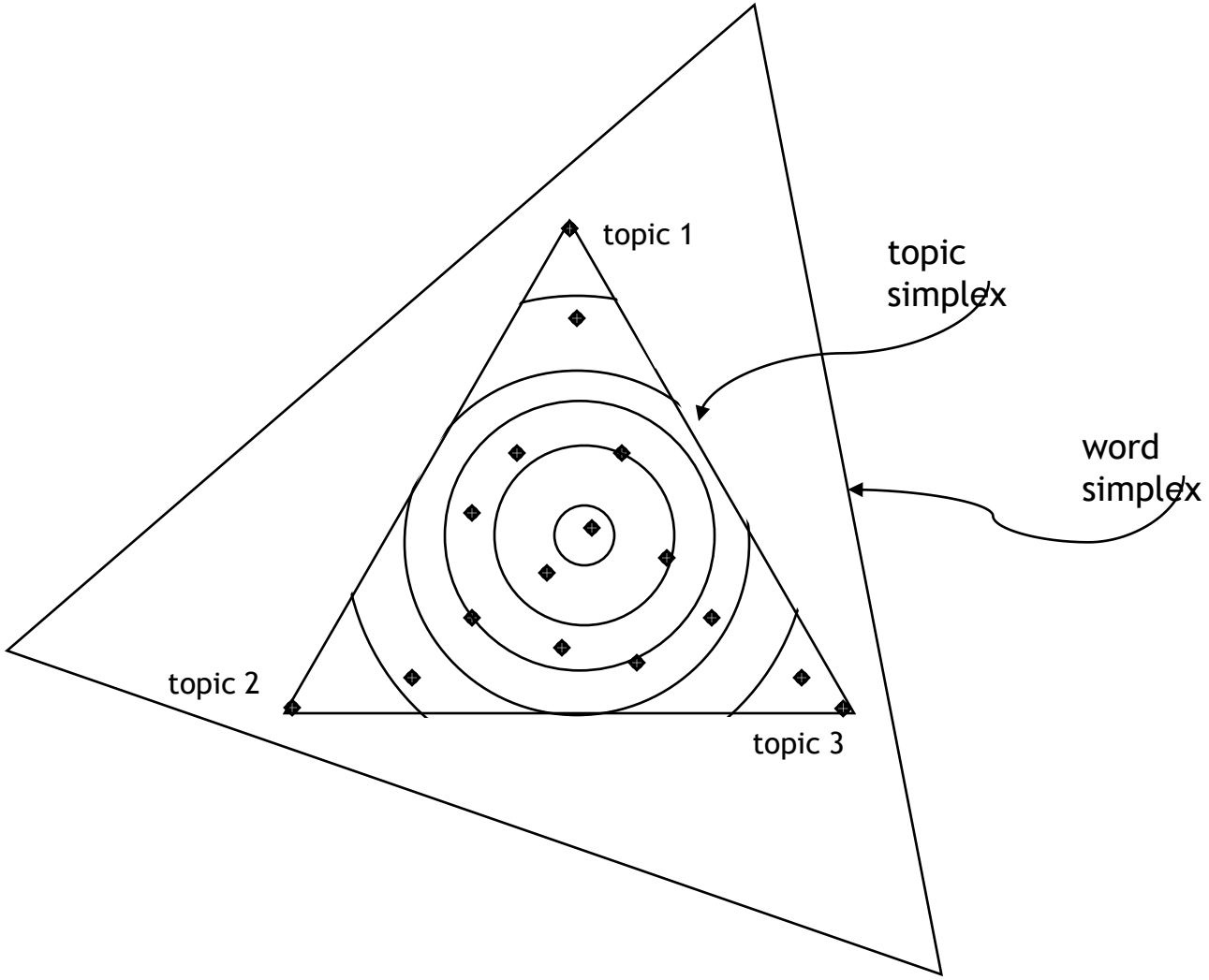


LDA

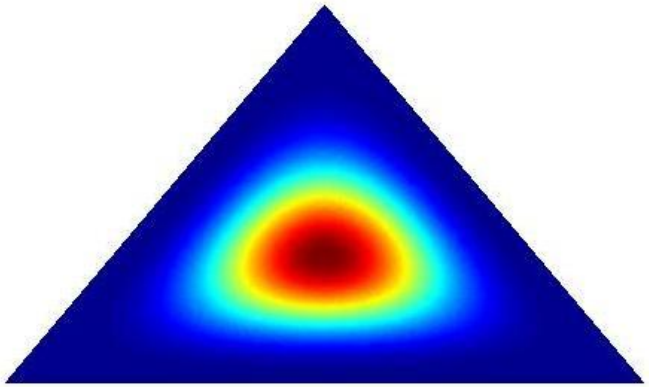
A geometric interpretation

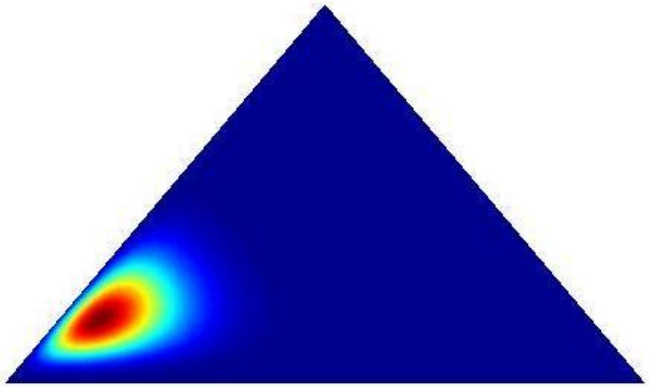


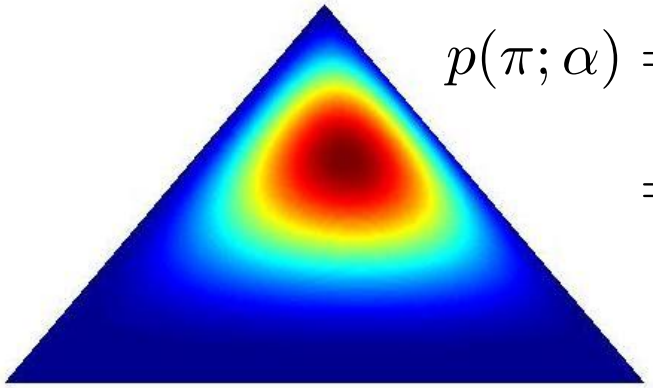
A geometric interpretation



Dirichlet Examples







$$p(\pi;\alpha) = Dirichlet(\pi;\alpha)$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod \pi_k^{\alpha_k - 1}$$

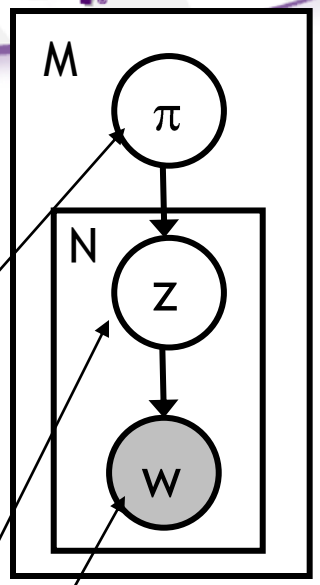
How to Generate an Image?

Given an image generate intermediate probability vector over 'topics'

For each word:

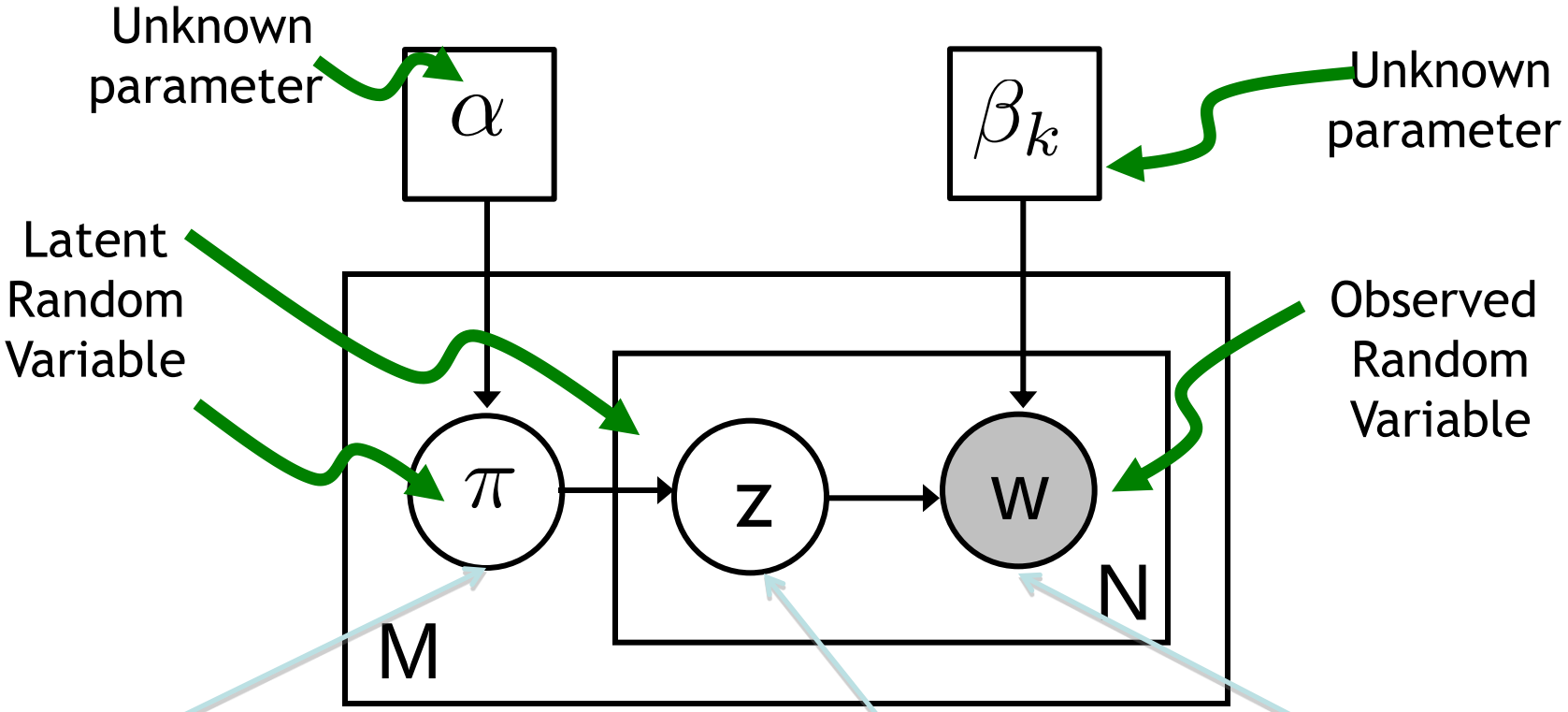
Determine current topic from mixture of topics

Draw a codeword from that topic



$$p(\mathbf{w}, \mathbf{z}, \pi; \alpha, \beta_k) = p(\pi; \alpha) \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi}) \; p(w_n|z_n; \beta_{z_n})$$

Graphical Model - Learning



$$p(\pi;\alpha) = Dirichlet(\pi;\alpha)$$
$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod \pi_k^{\alpha_k - 1}$$

$$p(z|\pi) = \pi_z \qquad p(w|z) = \beta_{zw}$$

Learning: Maximum likelihood estimation

Maximize (log)likelihood of data:

- Exact inference is intractable due to integral inside log
- Approximation techniques:
  - **Mean field variational methods** (Blei et al., 2001, 2003)
  - Expectation propagation (Minka and Lafferty, 2002)
  - Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
  - Collapsed variational inference (Teh et al., 2006)

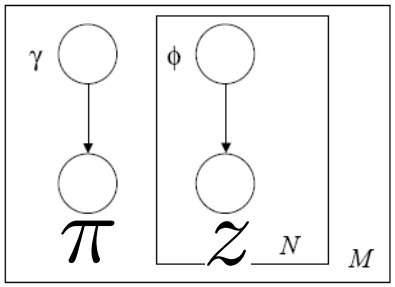
Variational Approximation for LDA

- Approximation the true posterior for latent variables

$$p(\pi, z_1, \ldots, z_n) \text{ by } q(\pi, z_1, \ldots, z_n)$$

- Assume independence for variational distribution

$$q(\pi, z_1, \ldots, z_n) = q(\pi; \gamma) \prod_n q(z_n; \phi_n)$$



Variational distribution

- Optimal Variational Parameters (image-specific) are obtained by minimizing the KL divergence between the variational distribution and the true posterior

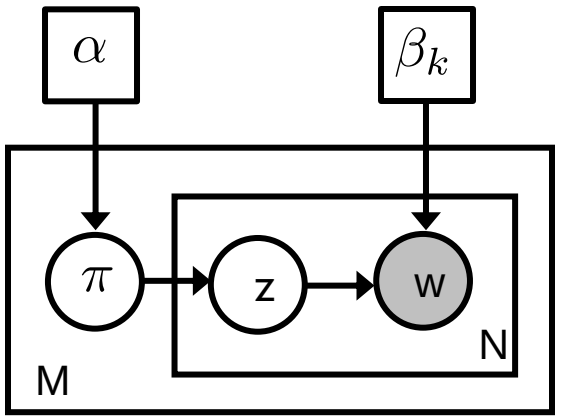
$$(\gamma^*, \phi_n^*) = \arg\min_{\gamma, \phi_n} KL(q(\pi, z_1, \ldots, z_n) || p(\pi, z_1, \ldots, z_n))$$

Parameter estimation

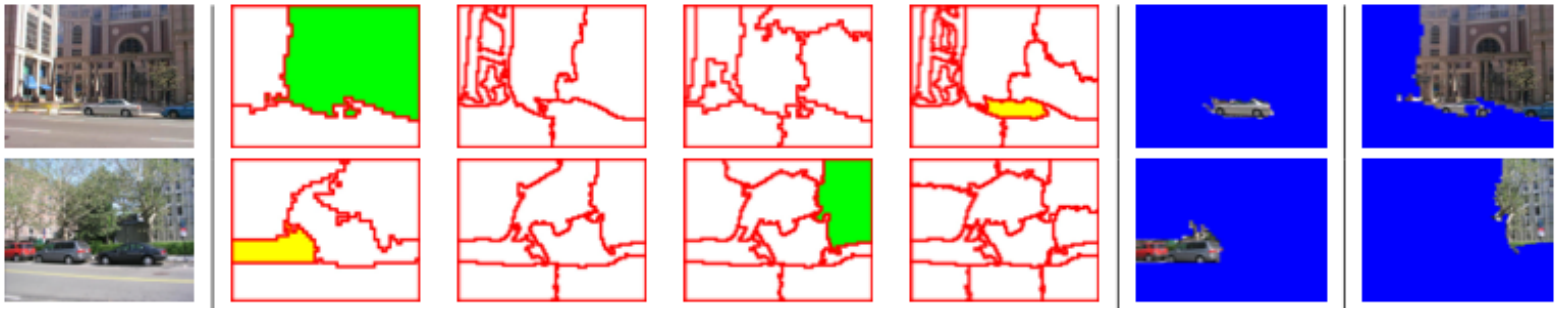
- Use **Expectation Maximization** where expectations are computed with respect to **variational distributions**.
- Variational EM
  - (E Step) For each document, find the optimizing values of the variational parameters (γ, φ) with α, β fixed.

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \varphi_{ni}$$

$$\varphi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right)$$



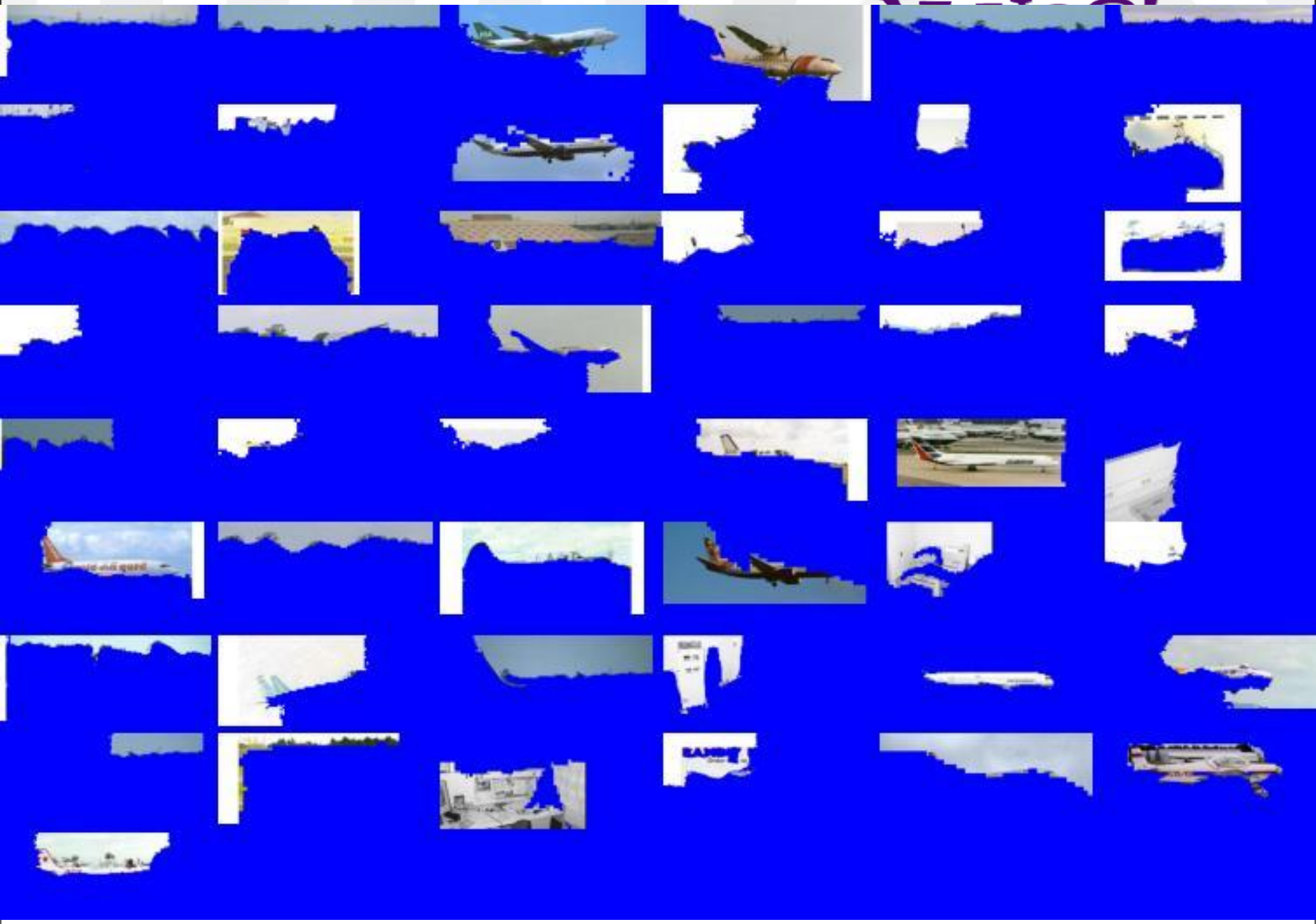
  - (M Step) Maximize variational distribution w.r.t. α, β for the γ and φ values found in the E step.
    - Using standard lagrangian approach.



Given a large, unlabeled collection of images:

1. For each image in the collection, compute multiple candidate segmentations, e.g. using Normalized Cuts [20] (section 2.1).
2. For each segment in each segmentation, compute a histogram of "visual words" [22] (section 2.2).
3. Perform topic discovery on the set of *all segments* in the image collection (using Latent Dirichlet Allocation [2]), treating each segment as a document (section 2.3).
4. For each discovered topic, sort *all segments* by how well they are explained by this topic (section 2.4).

B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, Using Multiple Segmentations to Discover Objects and their Extent in Image Collections, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, New York, June, 2006.



What about classification?

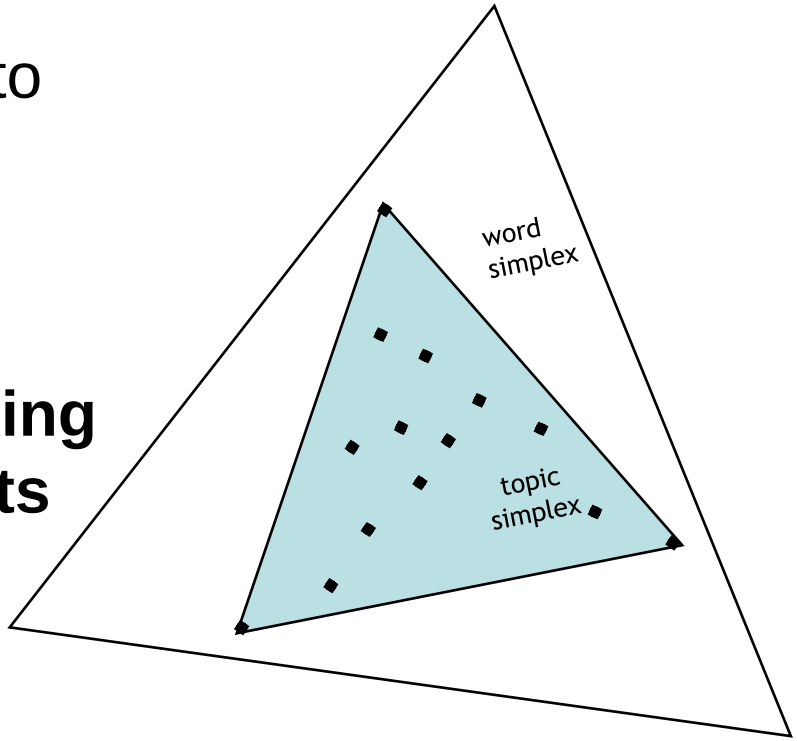
- Until now we have introduced pLSA and LDA to account for correlation between words.
  - How do we use these models for classification?

- Three approaches
  - First: Use discriminative approaches.
  - Second: Using “topic-supervision”
  - Third: Model class label into the generative process

1. Using Discriminative Approaches

- LDA/pLSA serves as **dimensionality reduction** techniques from word-simplex to topic-simplex

- Learn discriminative classifiers such as SVM**, using topic mixing probabilities of the documents as feature vectors**. [Blei'2003, Bosch'2008, Quelhas'2007, etc.]



- Not very interesting.

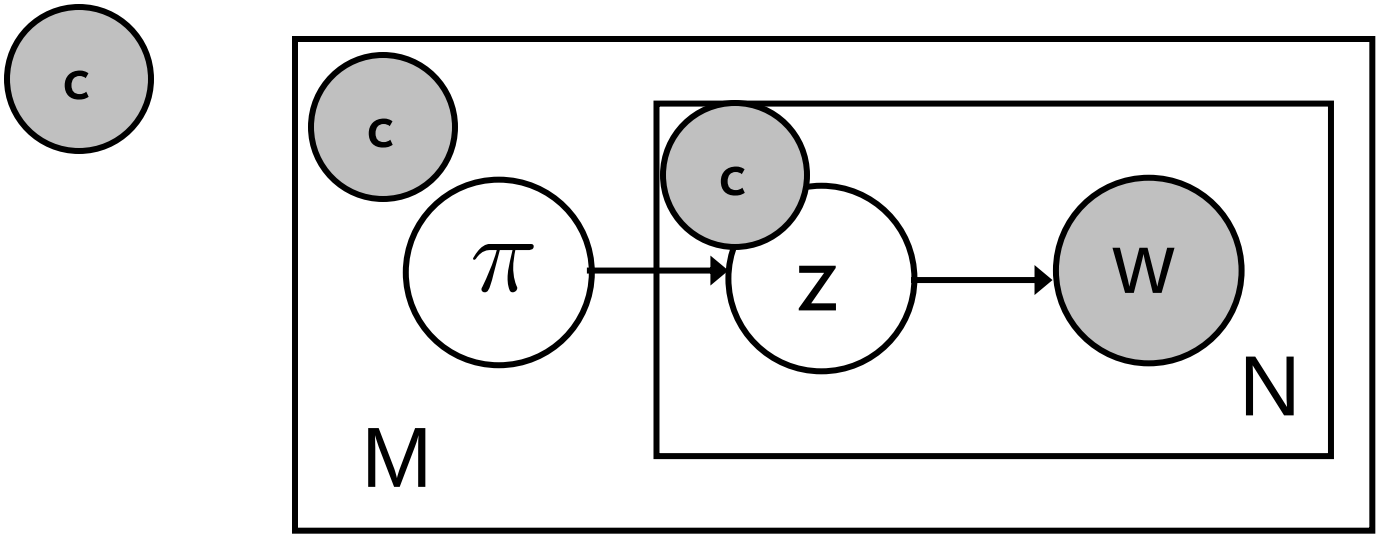
2. Using “topic supervision”

- Instead of “discovering” topics, set them to the class labels.
  - **The topic variables z are class labels.**
  - **The topic conditional distributions P(w|z) are learned in a supervised manner.**

- Labelled-LDA [Ramage’2009]
- Semi-LDA [Wang’2007]
- Prior-LDA, Dependency-LDA [Rubin’2011]
- Ts-cLDA, ts-sLDA [Rasiwasia’2012, under review]

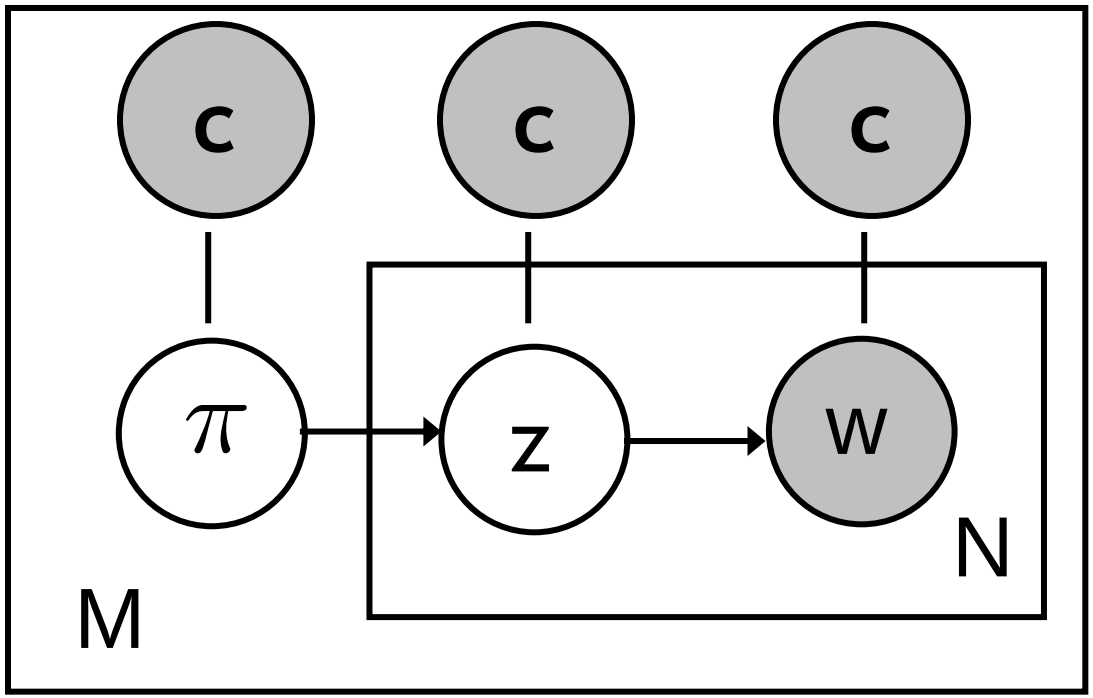
3. Model class label

- Should it be inside the box or outside?



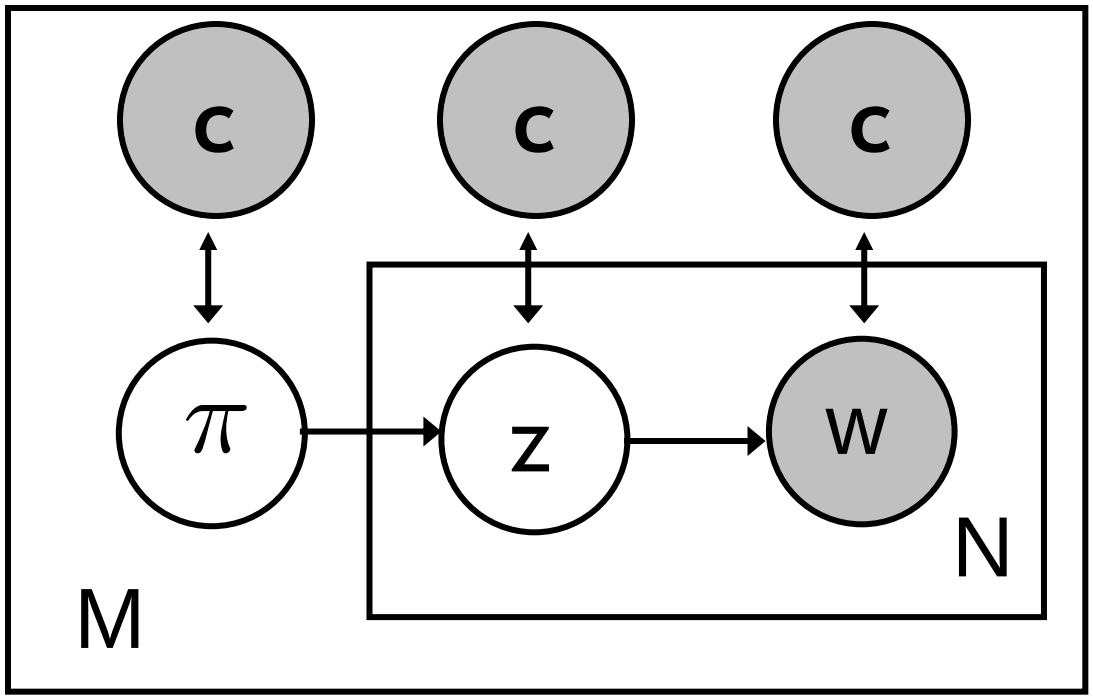
Where?

- Where should it be placed?

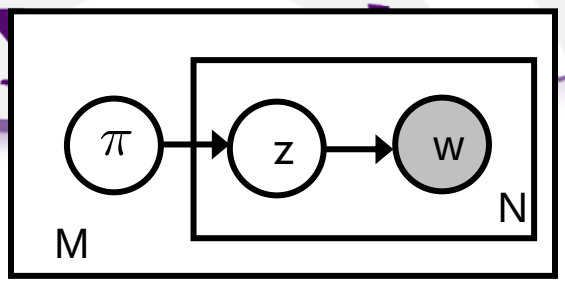


Where?

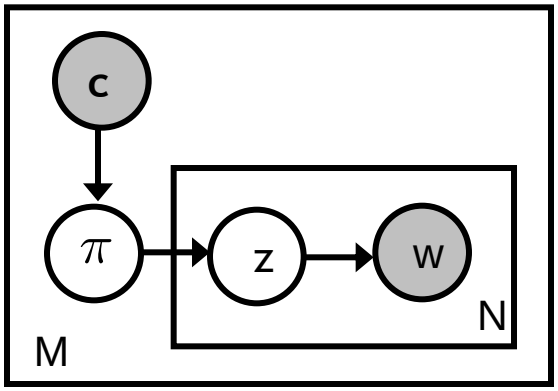
- How should the arrow look like?



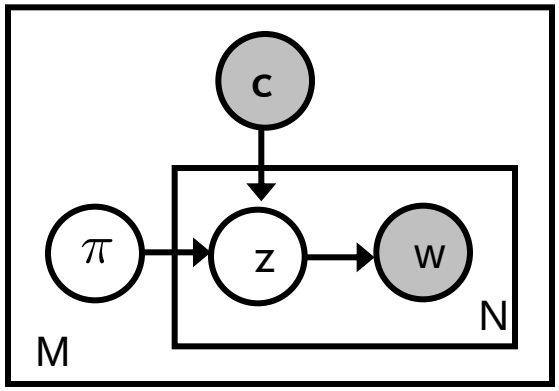
3. Model class label



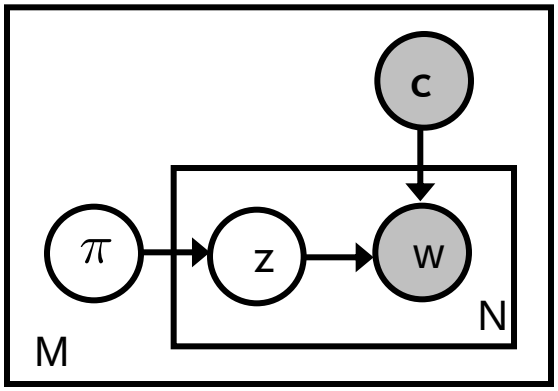
cLDA [L. Fei-Fei'05]



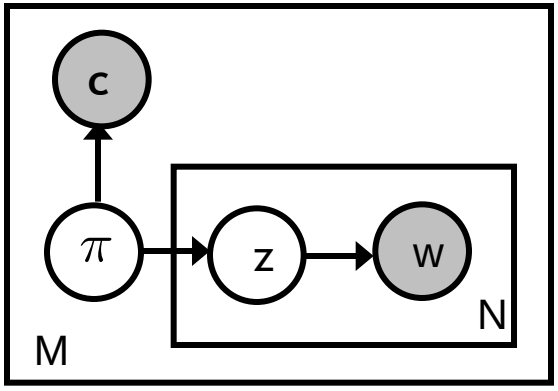
Similar to cLDA



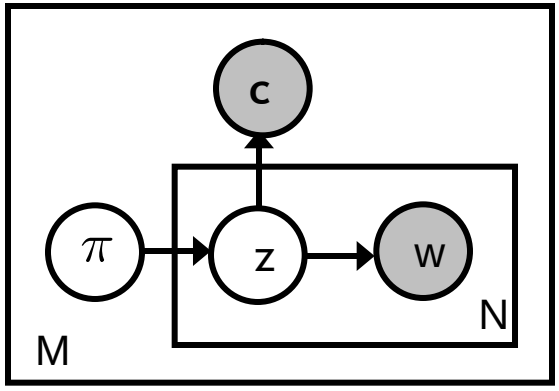
Css-LDA [Rasiwasia'12]



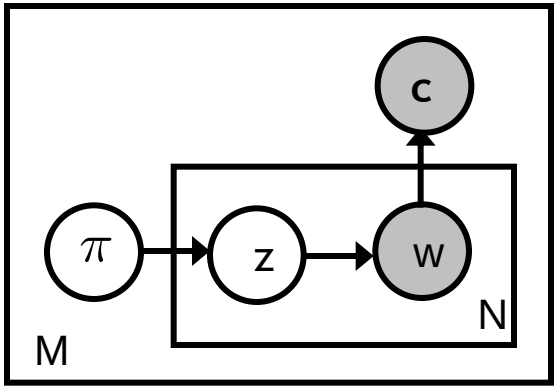
Similar to sLDA



sLDA [Blei'07]



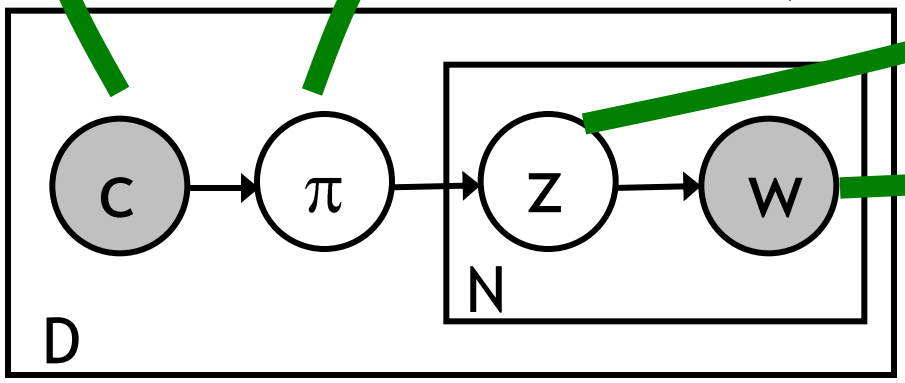
Does not make much sense



Class - Latent Dirichlet Allocation

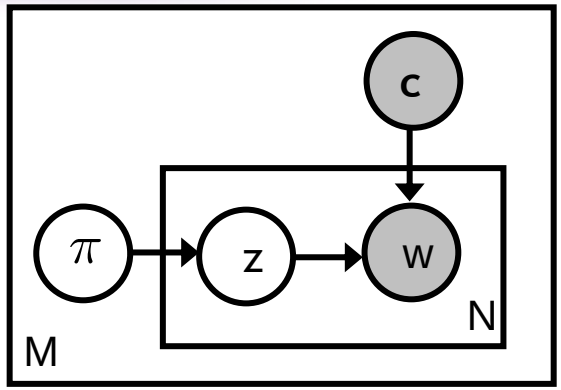
"beach"

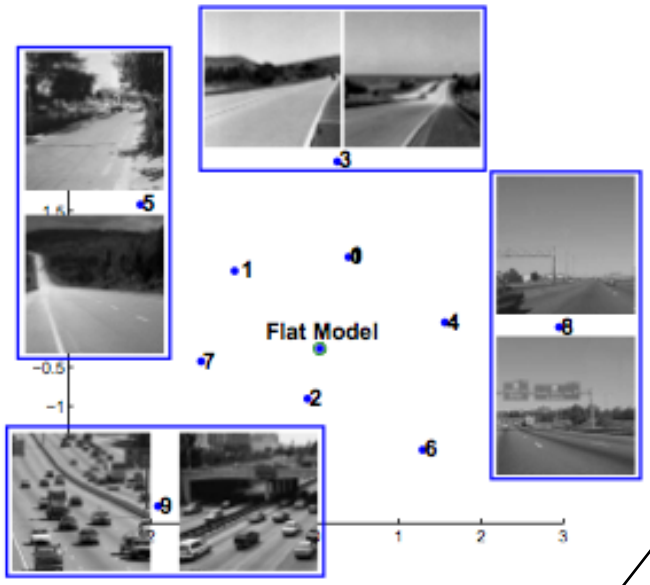
Latent Dirichlet Allocation (LDA)

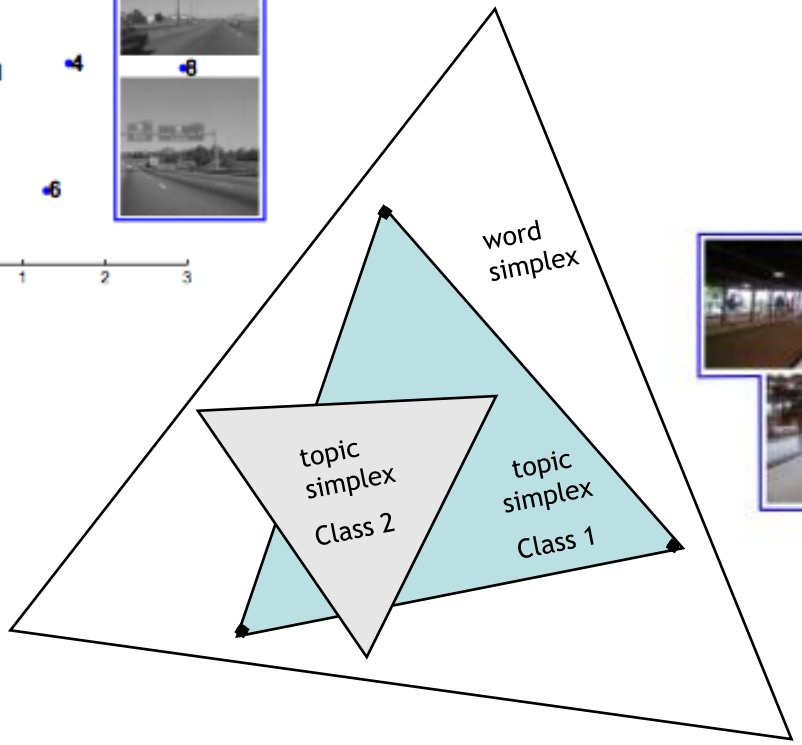


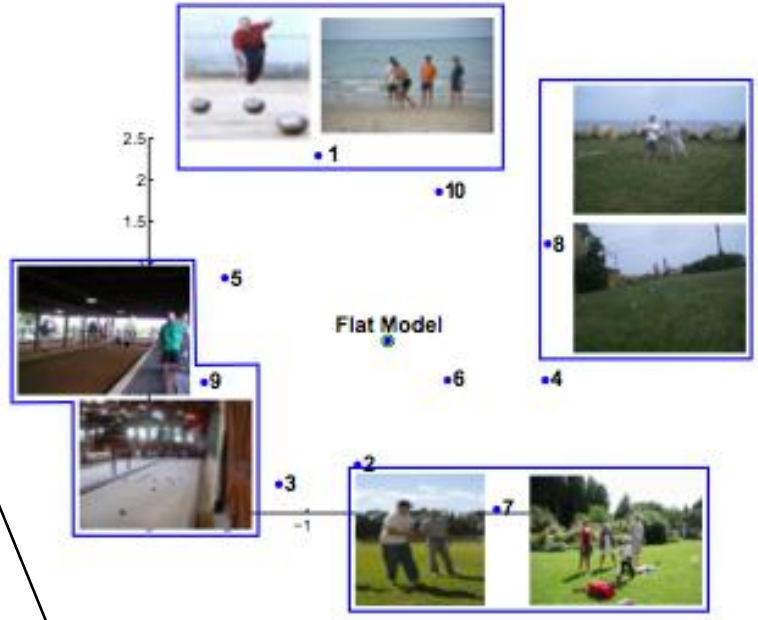
Fei-Fei et al. CVPR 2005

Class specific simplex Latent Dirichlet Allocation









Rasiwasia et. al. 2012

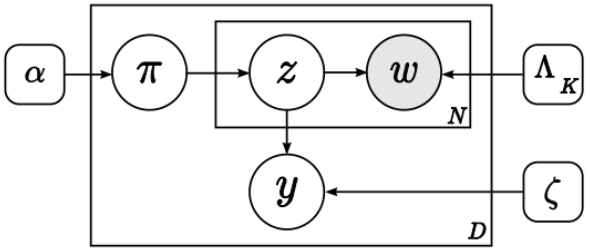
How do they compare!

	Dataset				
model	**N15**	**N13**	**N8**	**S8**	**C50**
css-LDA	**76.62 ± 0.32**	**81.03 ± 0.74**	**87.97 ± 0.84**	**80.37 ± 1.36**	**46.04**
flat	74.91 ± 0.38	79.60 ± 0.38	86.80 ± 0.51	77.87 ± 1.18	43.20
ts-sLDA	74.82 ± 0.68	79.70 ± 0.48	86.33 ± 0.69	78.37 ± 0.80	42.33
ts-cLDA	74.38 ± 0.78	78.92 ± 0.68	86.25 ± 1.23	77.43 ± 0.97	40.80
ts-LDA	72.60 ± 0.51	78.10 ± 0.31	85.53 ± 0.41	77.77 ± 1.02	39.20
medLDA [27]	72.08 ± 0.59	77.58 ± 0.58	85.16 ± 0.57	78.19 ± 1.05	41.89
sLDA [24]	70.87 ± 0.48	76.17 ± 0.92	84.95 ± 0.51	74.95 ± 1.03	39.22
cLDA [14]	65.50 ± 0.32	72.02 ± 0.58	81.30 ± 0.55	70.33 ± 0.86	34.33
LDA-SVM	73.19 ± 0.51	78.45 ± 0.34	86.82 ± 0.93	76.32 ± 0.71	45.46

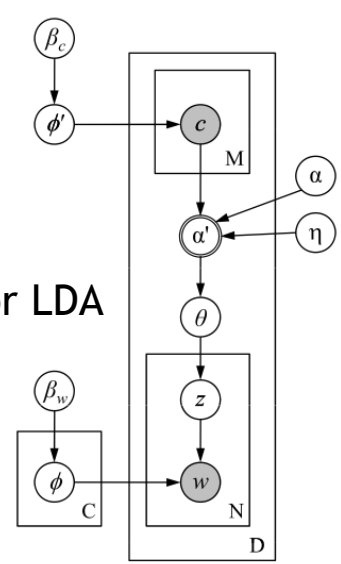
- LDA based models have not been very successful for classification
  - They underperform even the simple Naïve Bayes model (Flat)
- Recent advancements – cssLDA beats Naïve Bayes.
  - But does not beat discriminative models
- Recent advancement in text – Dependency LDA beats SVM
  - Needs to be tested on image datasets

What next?

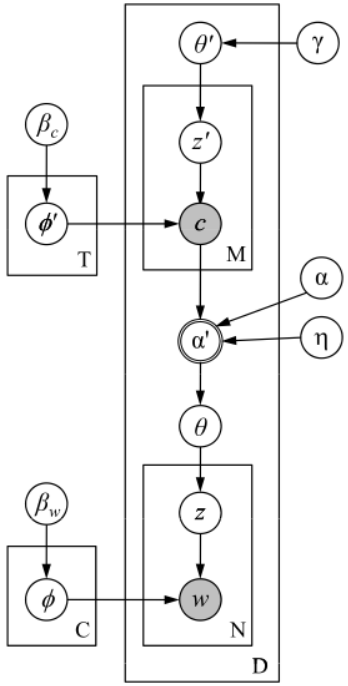
- Several advances have been made:



Supervised LDA

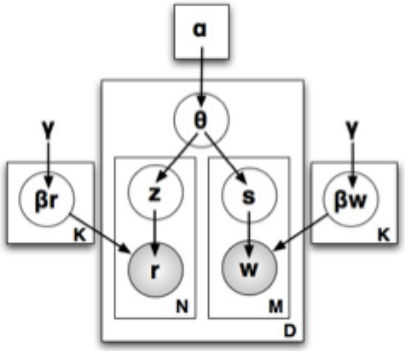


Prior LDA



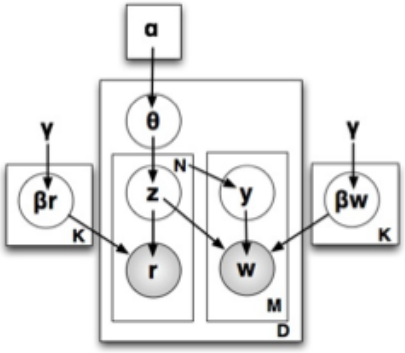
Dependency-LDA

Multimodal LDA



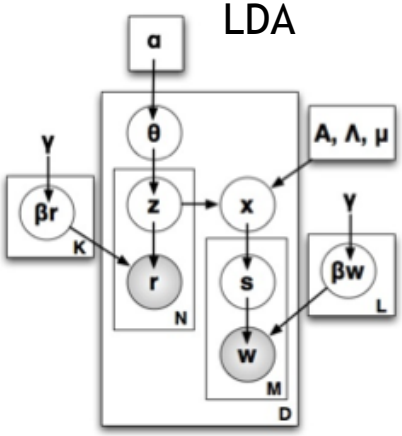
(b)

Correlation LDA



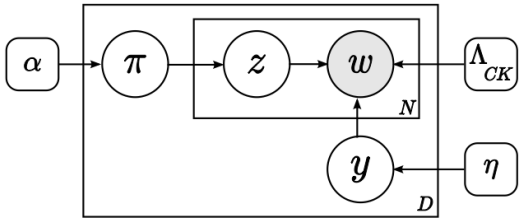
(c)

Topic Regression LDA



(d)

Class LDA



A handful references…

- D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR,* 2003.
- D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics,* 1(1):17–35, 2007.
- L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR,* 2005.
- Zhu, J., Ahmed, A., & Xing, E. P. (2009). MedLDA: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning, ICML'09* (pp. 1257–1264)
- Wang, Y., Sabzmeydani, P., & Mori, G. (2007). Semi-latent Dirichlet allocation: a hierarchical model for human action recognition. In *Proceedings of the 2nd conference on human motion: understanding, modeling, capture and animation*
- Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing,* Singapore, August 2009 (pp. 248–256).
- Mimno, D., Li, W., & McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. In *ICML'07: proceedings of the 24th international conference on machine learning* (pp. 633–640). New York: ACM.
- Lacoste-Julien, S., Sha, F., & Jordan, M. I. (2008). DiscLDA: discriminative learning for dimensionality reduction and classification. In *NIPS* (pp. 897–904).