Machine Learning Department
School of Computer Science
Carnegie Mellon University

Topic Modeling + Variational Inference

Matt Gormley
Lecture 22
Nov. 11, 2019

Reminders

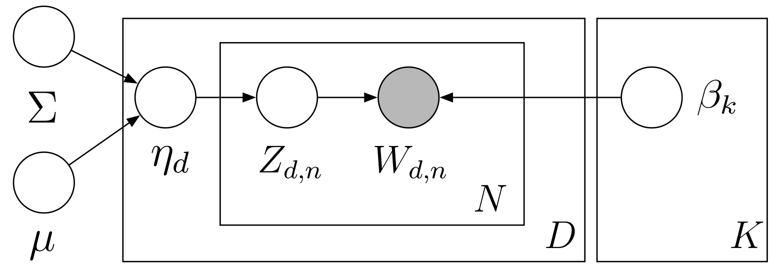
- Homework 4: Topic Modeling
  - Out: Wed, Nov. 6
  - Due: Mon, Nov. 18 at 11:59pm

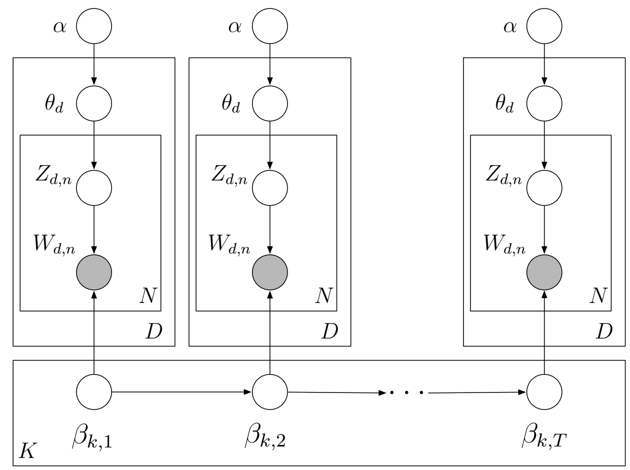
EXTENSIONS OF LDA

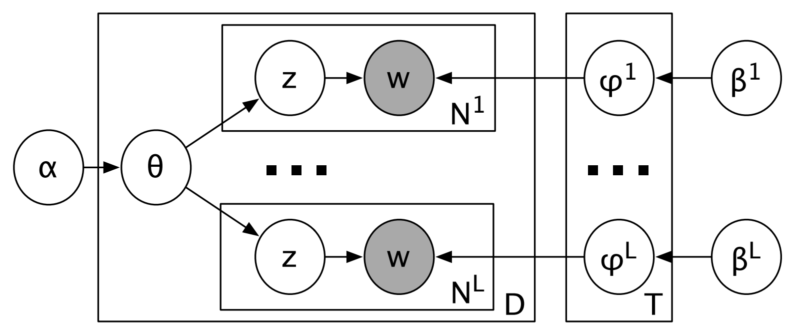
Extensions to the LDA Model

- Correlated topic models
  - Logistic normal prior over topic assignments
- Dynamic topic models
  - Learns topic changes over time
- Polylingual topic models
  - Learns topics aligned across multiple languages

…

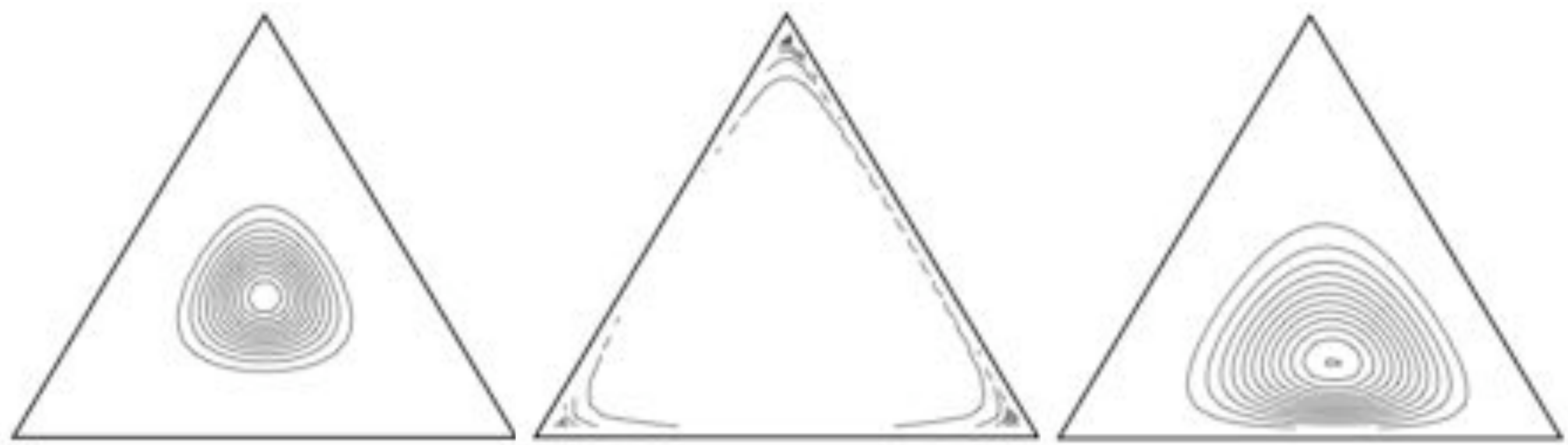






(Blei & Lafferty, 2004)

Correlated Topic Models

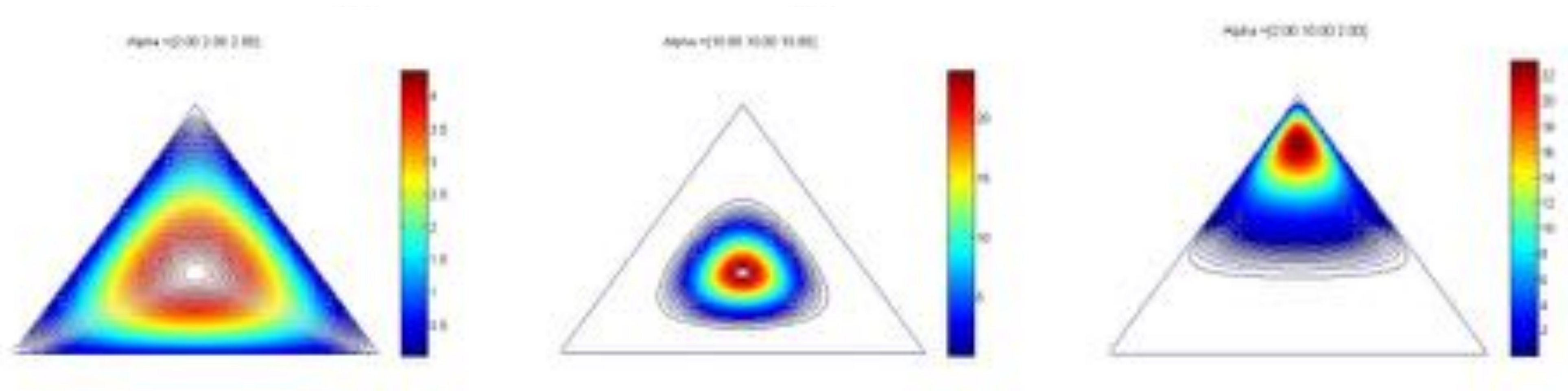


- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

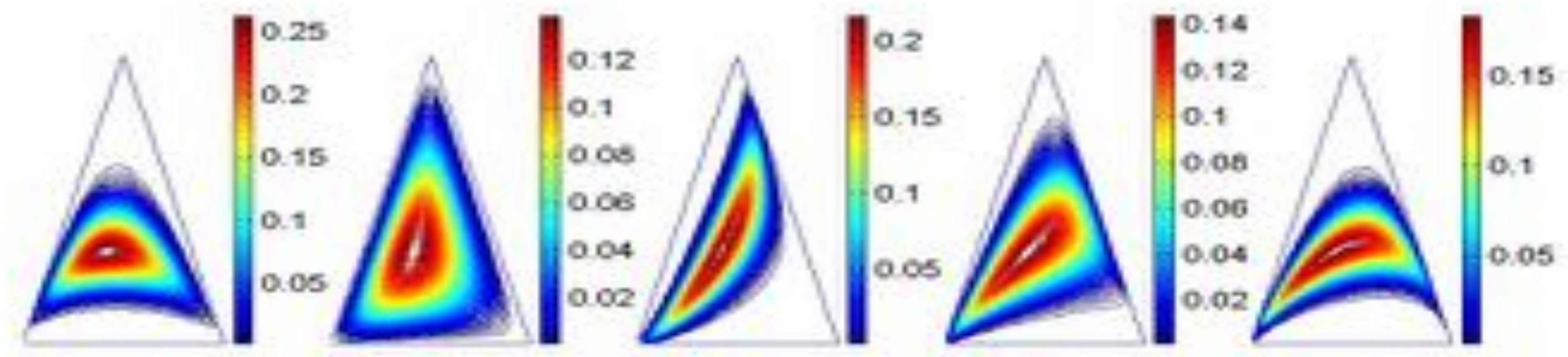


- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



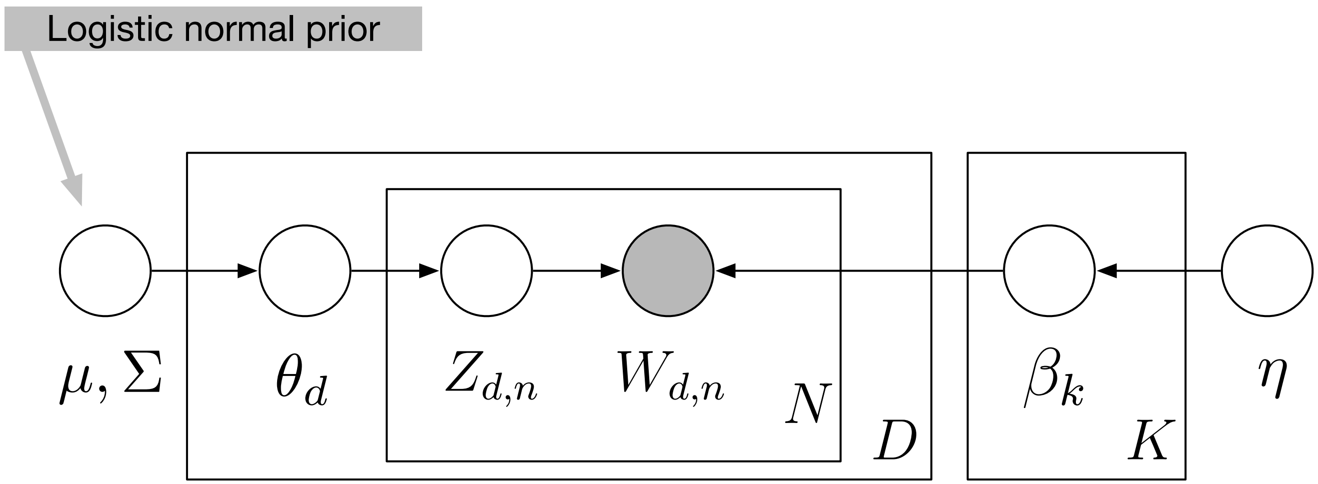
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$\begin{aligned} X &\sim \mathcal{N}_K(\mu, \Sigma) \\ \theta_i &\propto \exp\{x_i\}. \end{aligned}$$

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

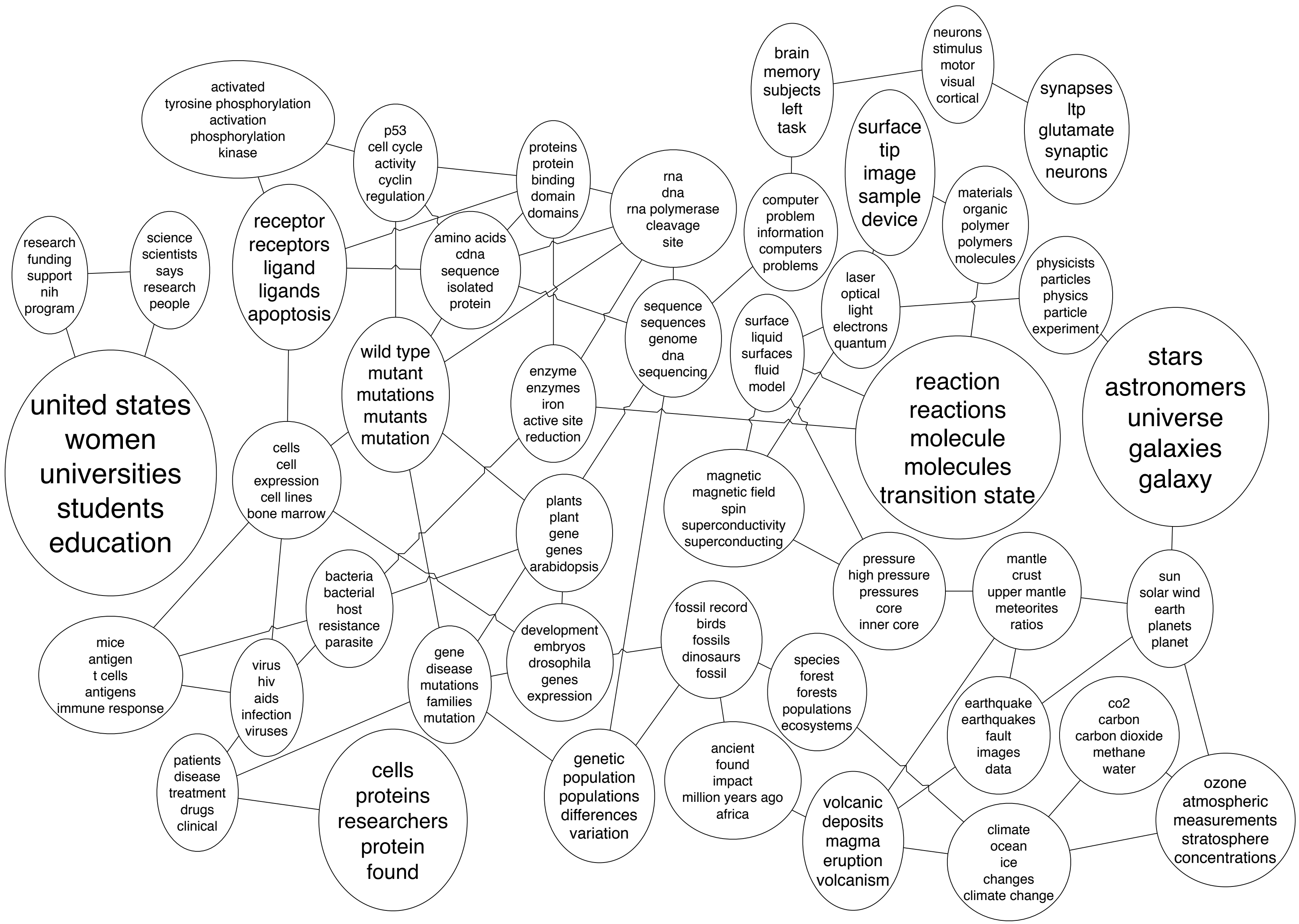


- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but computation is more complex

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



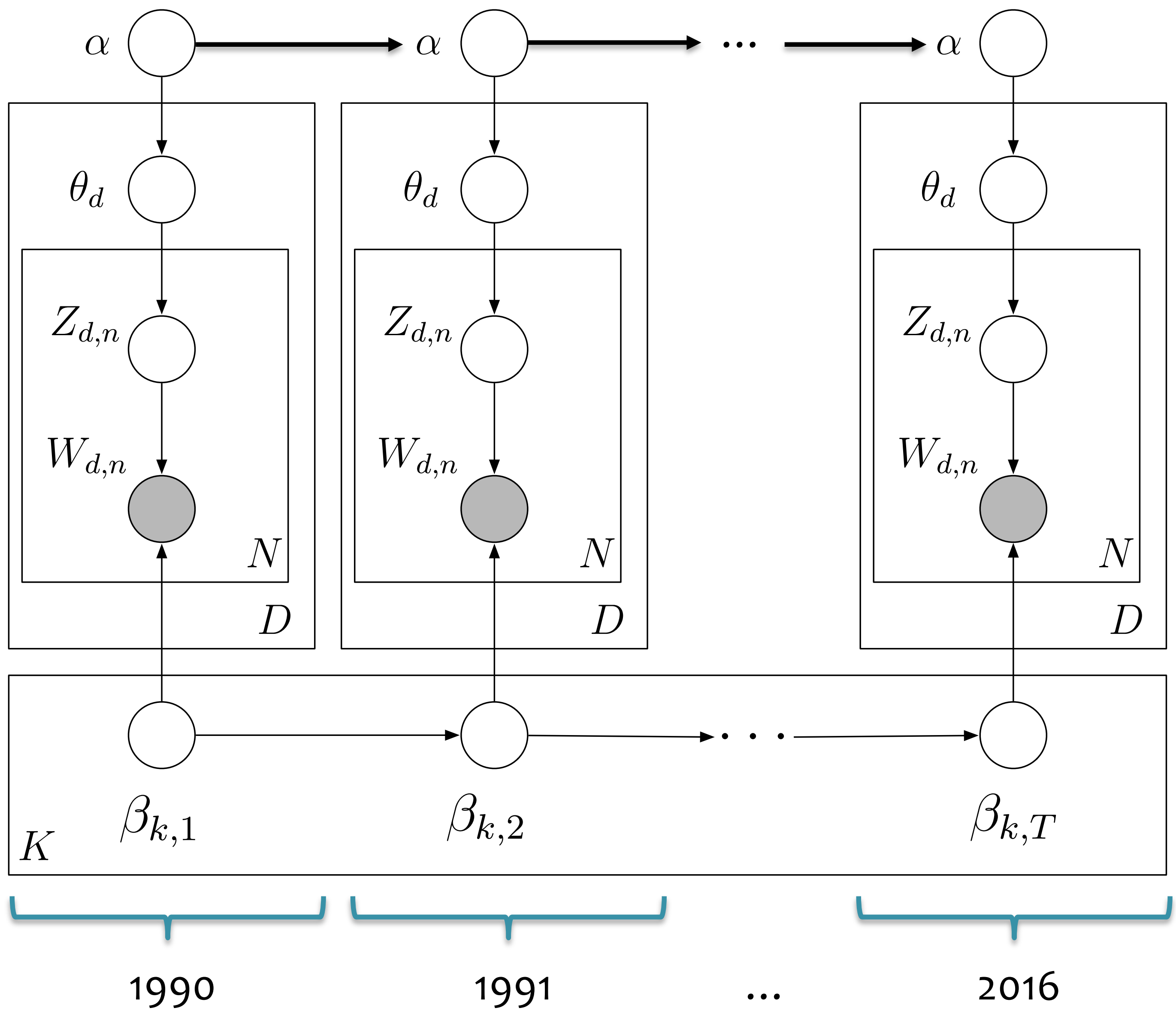
Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

High-level idea:

- Divide the documents up by year
- Start with a separate topic model for each year
- Then add a dependence of each year on the previous one



Dynamic Topic Models

1789 **2009**

Inaugural addresses

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics *drift* in a sequence.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

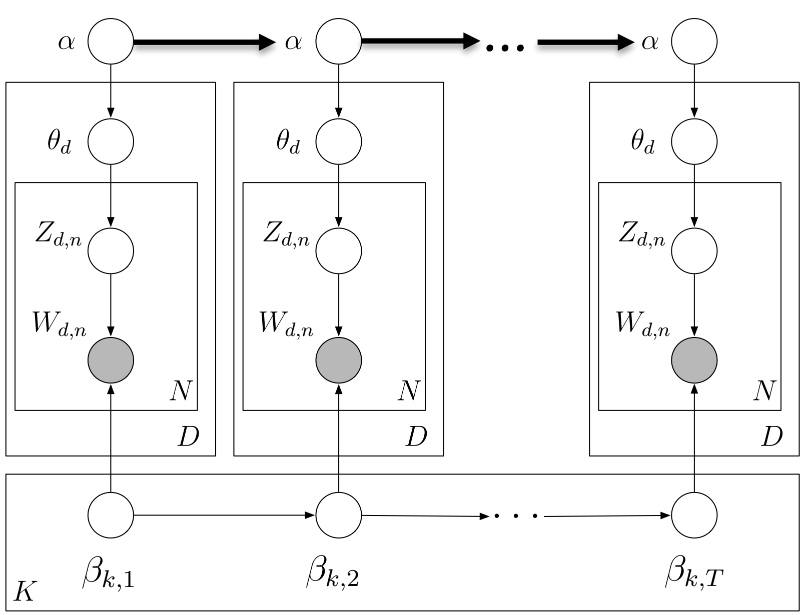
Generative Story

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim \mathit{Mult}(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim \mathit{Mult}(\pi(\beta_{t,z}))$.

Logistic-normal priors

The pi function maps from the natural parameters to the mean parameters:

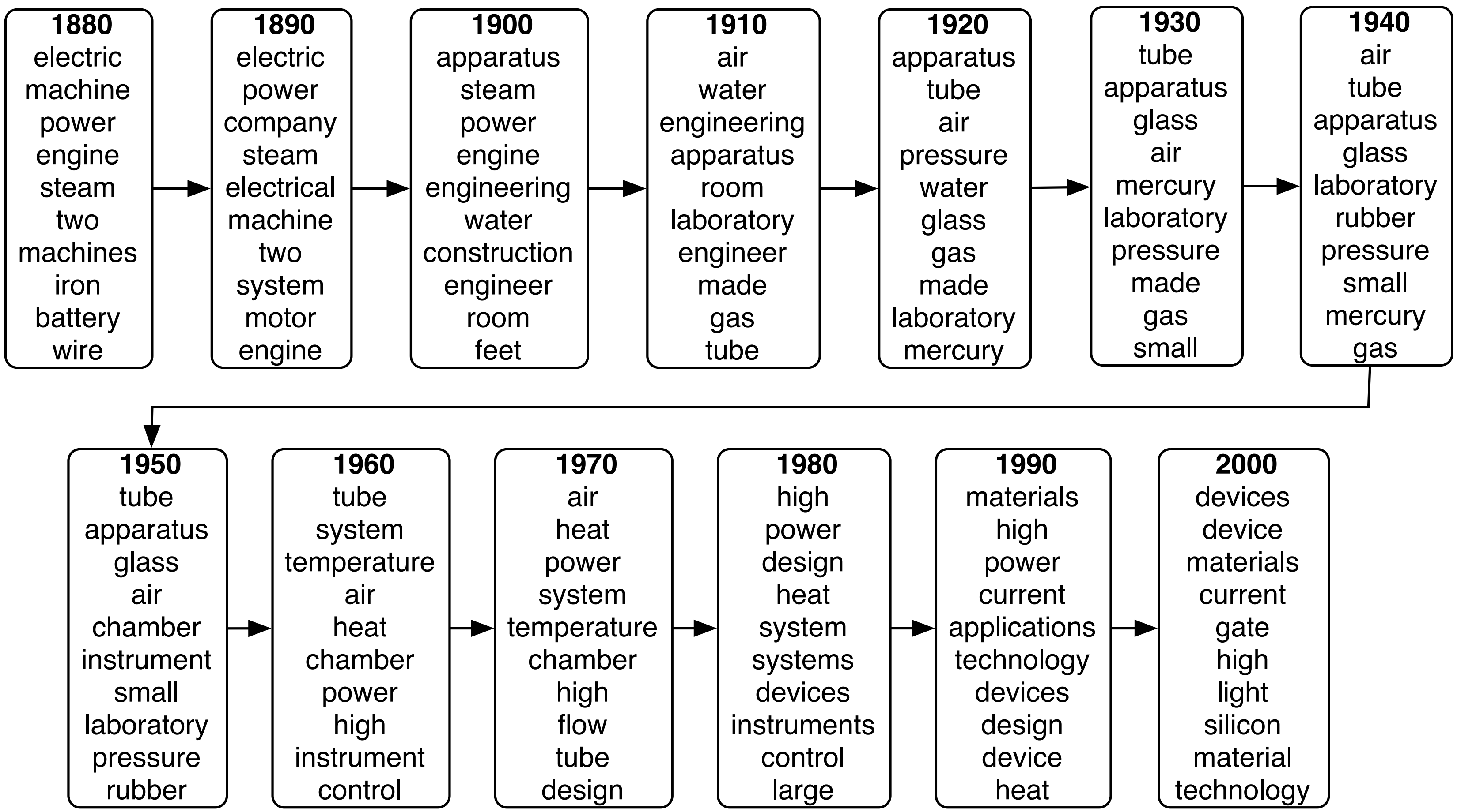
$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}.$$



(Blei & Lafferty, 2006)

Dynamic Topic Models

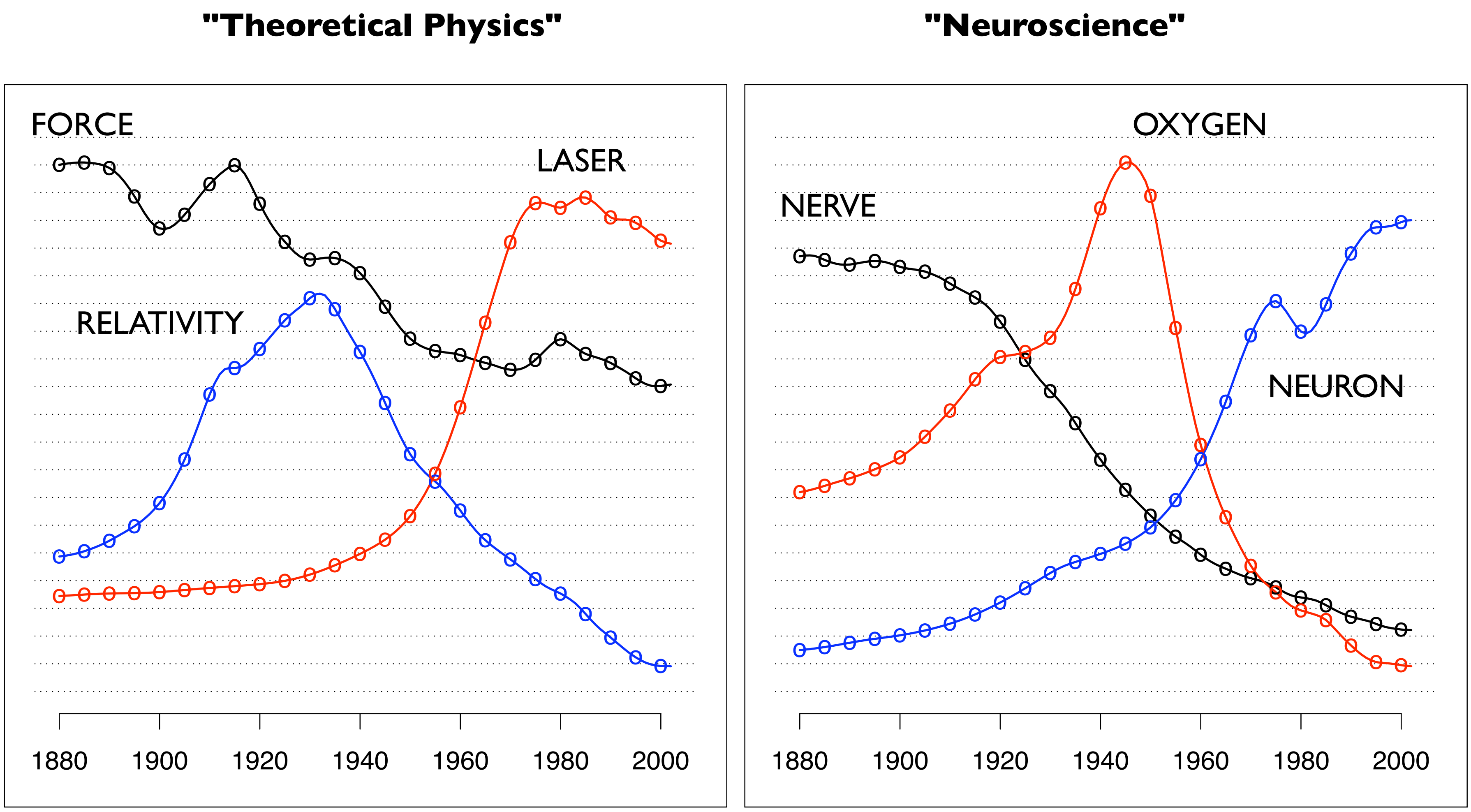
Top ten most likely words in a "drifting" topic shown at 10-year increments



(Blei & Lafferty, 2006)

Dynamic Topic Models

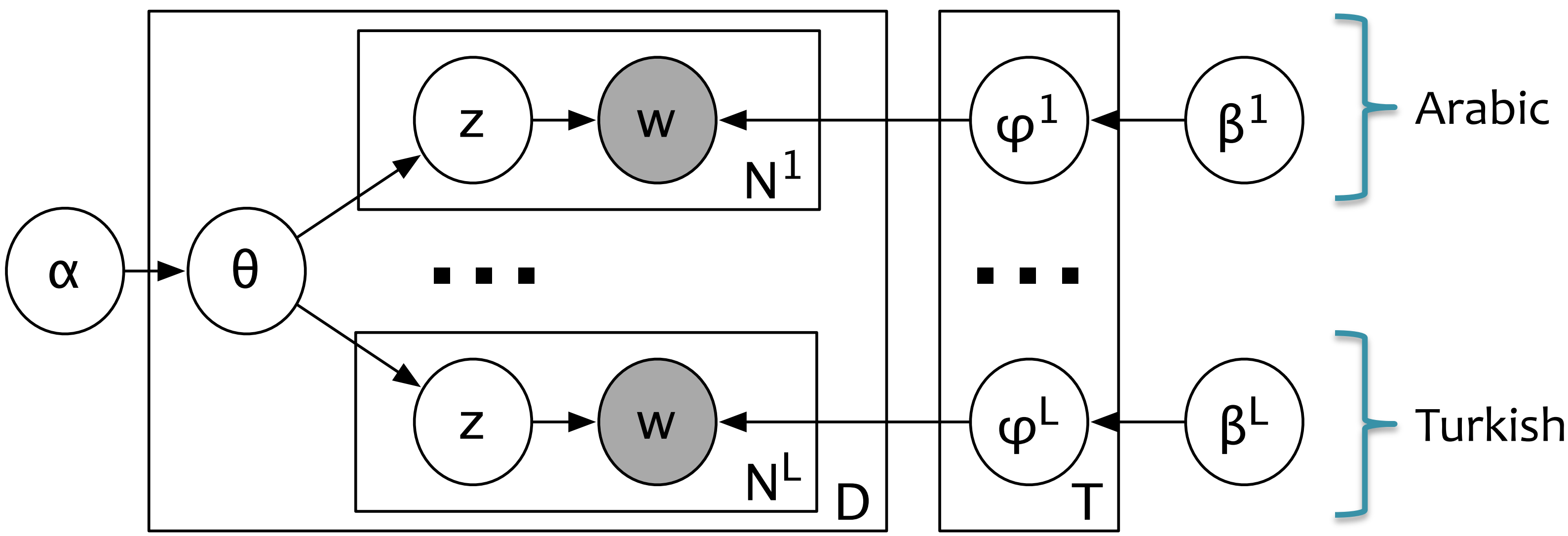
Posterior estimate of **word frequency as a function of year** for three words each in two separate topics:



(Mimno et al., 2009)

Polylingual Topic Models

- **Data Setting:** Comparable versions of each document exist in multiple languages (e.g. the Wikipedia article for "Barak Obama" in twelve languages)
- **Model:** Very similar to LDA, except that the topic assignments, z, and words, w, are sampled separately for each language.



(Mimno et al., 2009)

Polylingual Topic Models

Topic 1 (twelve languages)

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

(Mimno et al., 2009)

Polylingual Topic Models

Topic 2 (twelve languages)

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

(Mimno et al., 2009)

Polylingual Topic Models

Topic 3 (twelve languages)

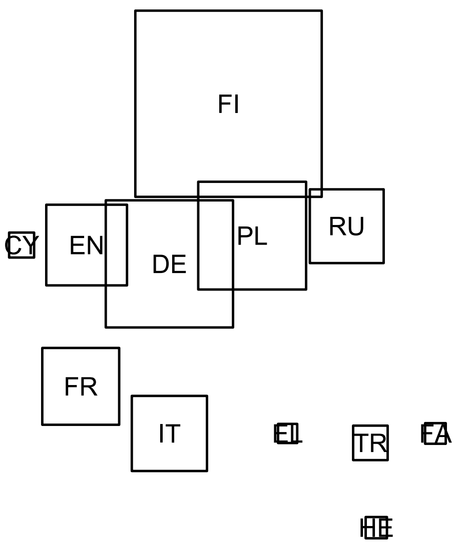
CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

(Mimno et al., 2009)

Polylingual Topic Models

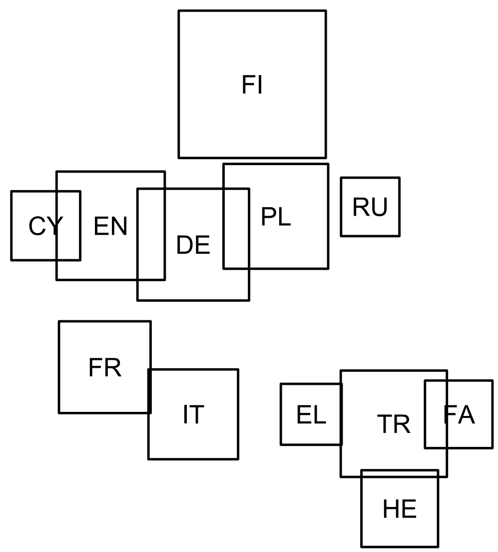
Size of each square represents proportion of tokens assigned to the specified topic.

world
ski
km
won



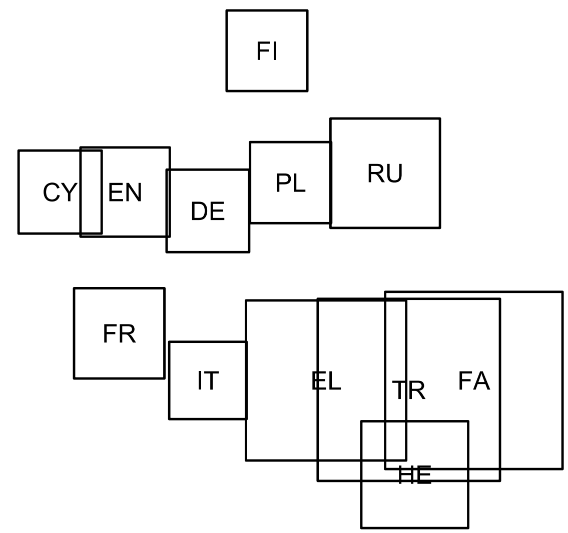
Analysis: mostly Nordic countries

actor
role
television
actress



Analysis: uniform across countries

ottoman
empire
khan
byzantine



Analysis: mostly countries near Istanbul

Supervised LDA

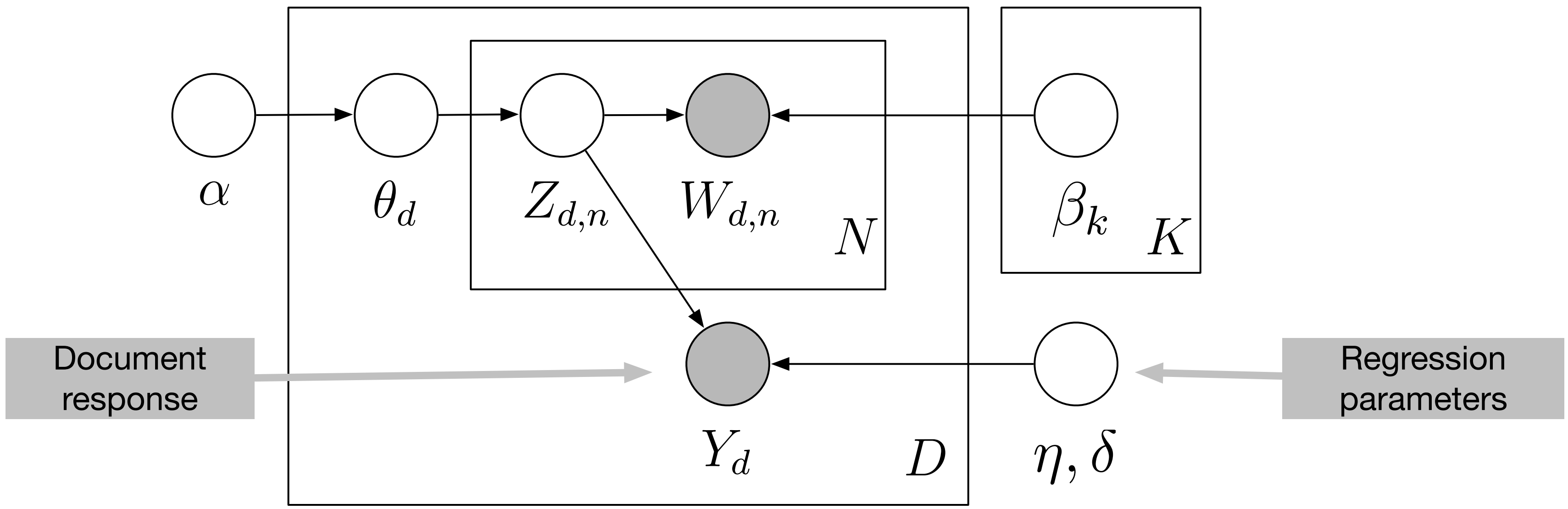
- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?

- Many data are paired with **response variables**.
    - User reviews paired with a number of stars
    - Web pages paired with a number of “likes”
    - Documents paired with links to other documents
    - Images paired with a category

- **Supervised LDA** are topic models of documents and responses. They are fit to find topics predictive of the response.

Slide from David Blei, MLSS 2012

Supervised LDA



1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Slide from David Blei, MLSS 2012

Gaussian LDA

a Topic Model with Word Embeddings

Key Idea:

Instead of generating words as discrete, generate a (pretrained) vector representation of each word.

Generative Story

1. for $k = 1$ to K
   (a) Draw topic covariance $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$
   (b) Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\kappa}\boldsymbol{\Sigma}_k)$
2. for each document d in corpus D
   (a) Draw topic distribution $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   (b) for each word index n from 1 to N_d
      i. Draw a topic $z_n \sim \mathrm{Categorical}(\boldsymbol{\theta}_d)$
      ii. Draw $\mathbf{v}_{d,n} \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n})$

Visualizing Topics in Word Embedding Space

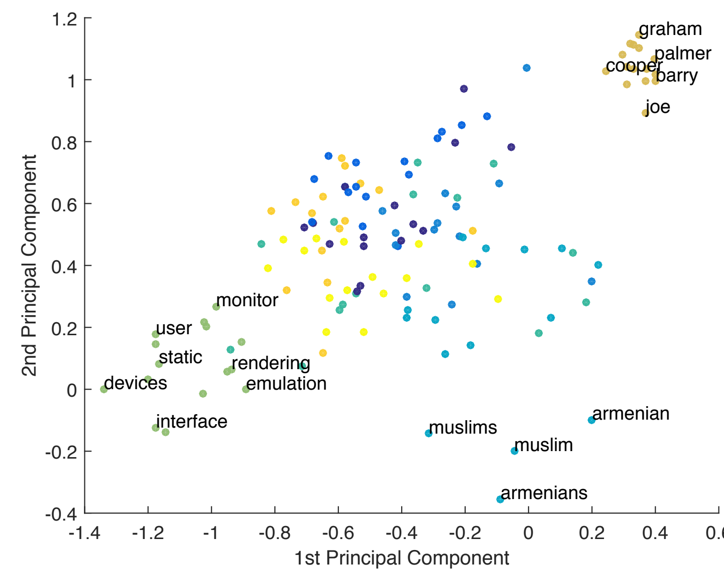


Figure 3: The first two principal components for the word embeddings of the top words of topics shown in Table 1 have been visualized. Each blob represents a word color coded according to its topic in the Table 1.

Summary: Topic Modeling

- **The Task of Topic Modeling**
  - Topic modeling enables the **analysis of large** (possibly unannotated) **corpora**
  - Applicable to more than just bags of words
  - Extrinsic evaluations are often appropriate for these unsupervised methods
- **Constructing Models**
  - LDA is comprised of **simple building blocks** (Dirichlet, Multinomial)
  - LDA itself can act as a building block **for other models**
- **Approximate Inference**
  - Many different approaches to inference (and learning) can be applied to the same model

What if we don't know the number of topics, K, ahead of time?

Solution: Bayesian Nonparametrics

- New modeling constructs:
  - Chinese Restaurant Process (Dirichlet Process)
  - Indian Buffet Process
- e.g. an **infinite number of topics** in a finite amount of space

Summary: Approximate Inference

- Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings, Gibbs sampling, Hamiltonion MCMC, slice sampling, etc.
- Variational inference
  - Minimizes KL(q||p) where q is a simpler graphical model than the original p
- Loopy Belief Propagation
  - Belief propagation applied to general (loopy) graphs
- Expectation propagation
  - Approximates belief states with moments of simpler distributions
- Spectral methods
  - Uses tensor decompositions (e.g. SVD)

HIGH-LEVEL INTRO TO VARIATIONAL INFERENCE

Variational Inference

Problem:

- For inputs **x** and outputs **z**, estimating the posterior p(**z** | **x**) is intractable
- For training data **x** and parameters **z**, estimating the posterior p(**z** | **x**) is intractable

Solution:

- Approximate p(**z** | **x**) with a simpler q(**z**)
- Typically q(**z**) has more independence assumptions than p(**z** | **x**) – fine b/c q(**z**) is tuned for a specific **x**
- **Key idea**: pick a single q(**z**) from some family Q that best approximates p(**z** | **x**)

Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Terminology:

- $q(\mathbf{z})$: the **variational approximation**
- Q: the **variational family**
- Usually $q_\theta(\mathbf{z})$ is parameterized by some θ called **variational parameters**
- Usually $p_\alpha(\mathbf{z} \mid \mathbf{x})$ is parameterized by some fixed α – we'll call them the parameters

Example Algorithms:

- mean-field approximation
- loopy belief propagation
- tree-reweighted belief propagation
- expectation propagation

Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Is this trivial?

- Note: We are not defining a new distribution simple $q_\theta(\mathbf{z} \mid \mathbf{x})$, there is one simple $q_\theta(\mathbf{z})$ for each $p_\alpha(\mathbf{z} \mid \mathbf{x})$
- Consider the MCMC equivalent of this:
  - you could draw samples $z^{(i)} \sim p(\mathbf{z} \mid \mathbf{x})$
  - then train some simple $q_\theta(\mathbf{z})$ on $z^{(1)}, z^{(2)}, \ldots, z^{(N)}$
  - hope that the sample adequately represents the posterior for the given $\mathbf{x}$
- How is VI different from this?
  - VI doesn't require sampling
  - VI is fast and deterministic
  - Why? b/c we choose an objective function (KL divergence) that defines which q_θ best approximates p_α, and exploit the special structure of q_θ to optimize it

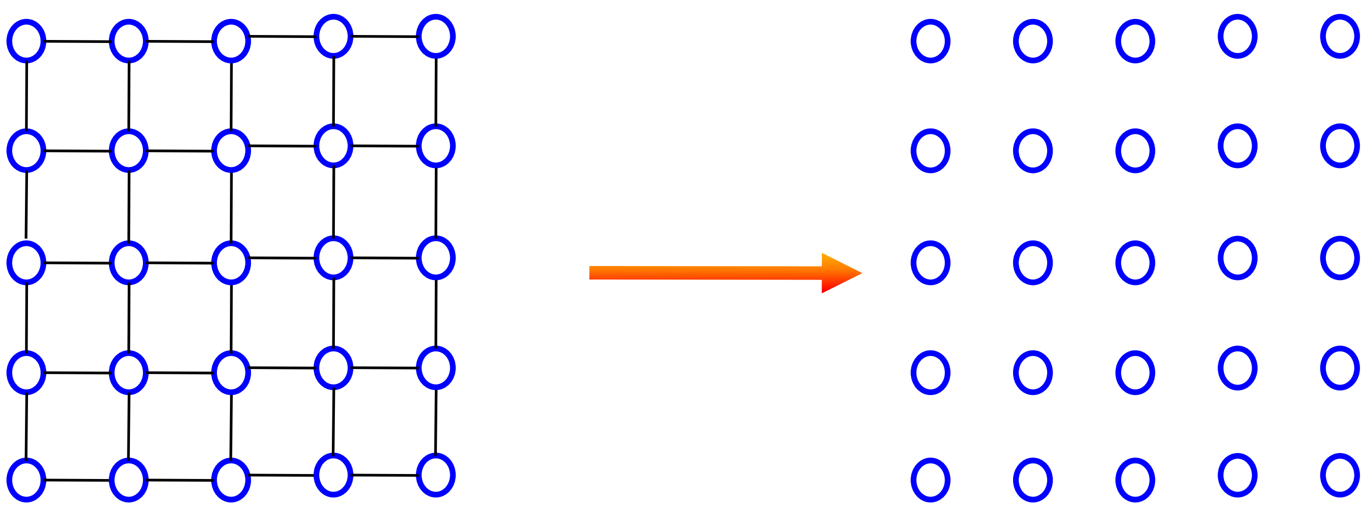
Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

EXAMPLES OF APPROXIMATING DISTRIBUTIONS

Mean Field for MRFs

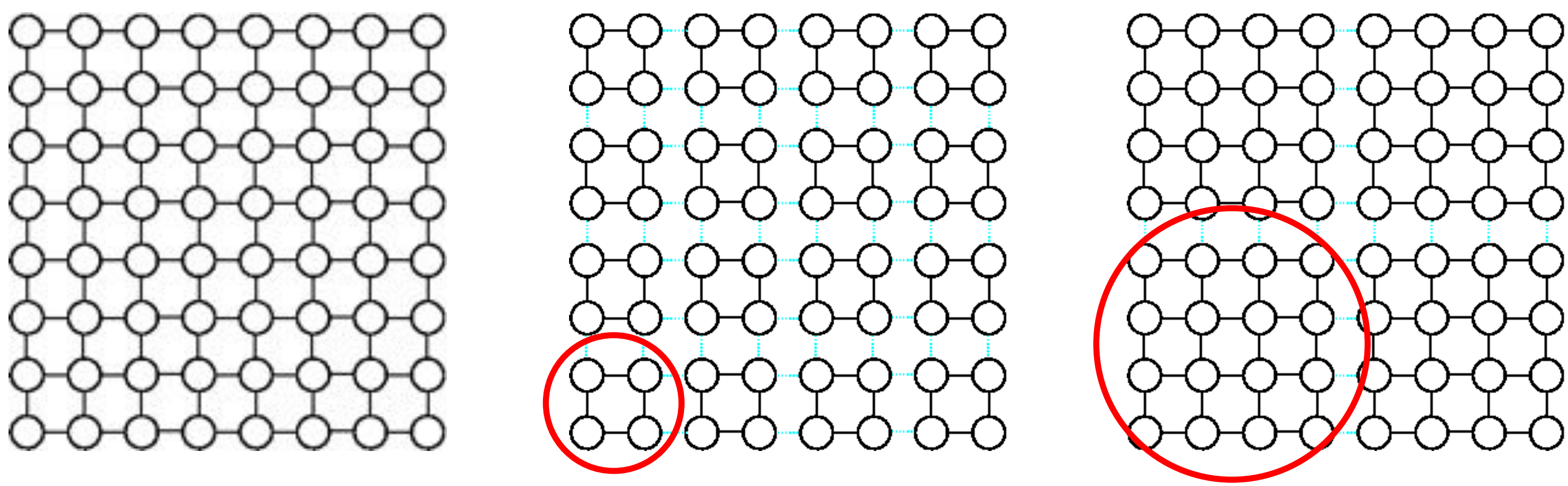
- Mean field approximation for Markov random field (such as the Ising model):

$$q(x) = \prod_{s \in V} q(x_s)$$



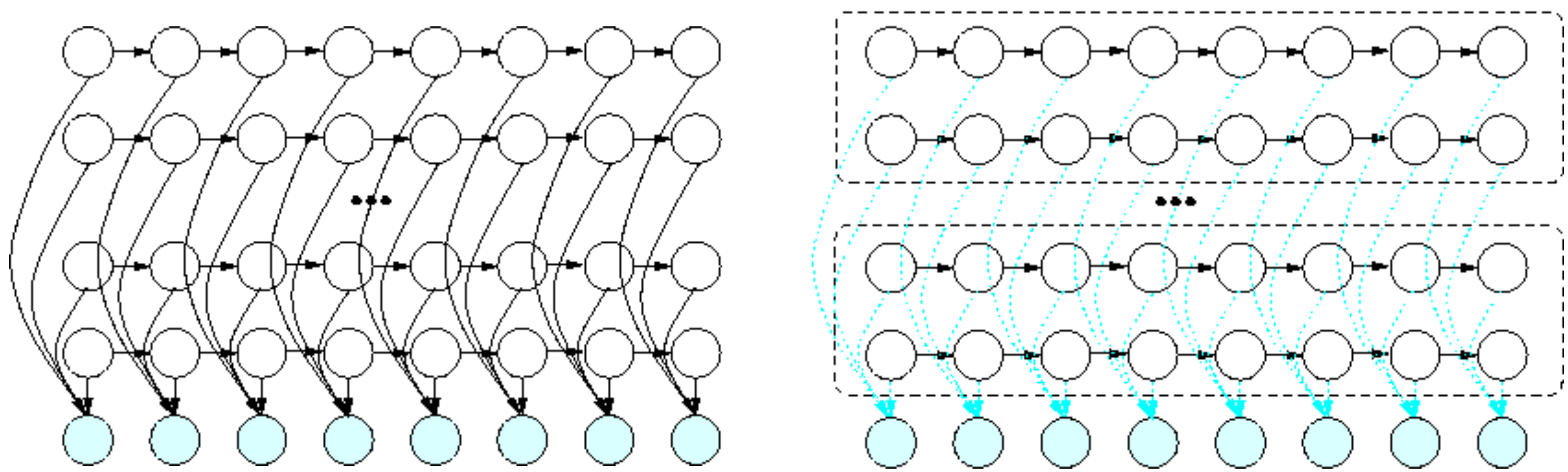
Mean Field for MRFs

- We can also apply more general forms of mean field approximations (involving clusters) to the Ising model:
- Instead of making all latent variables independent (i.e. naïve mean field, previous figure), clusters of (disjoint) latent variables are independent.



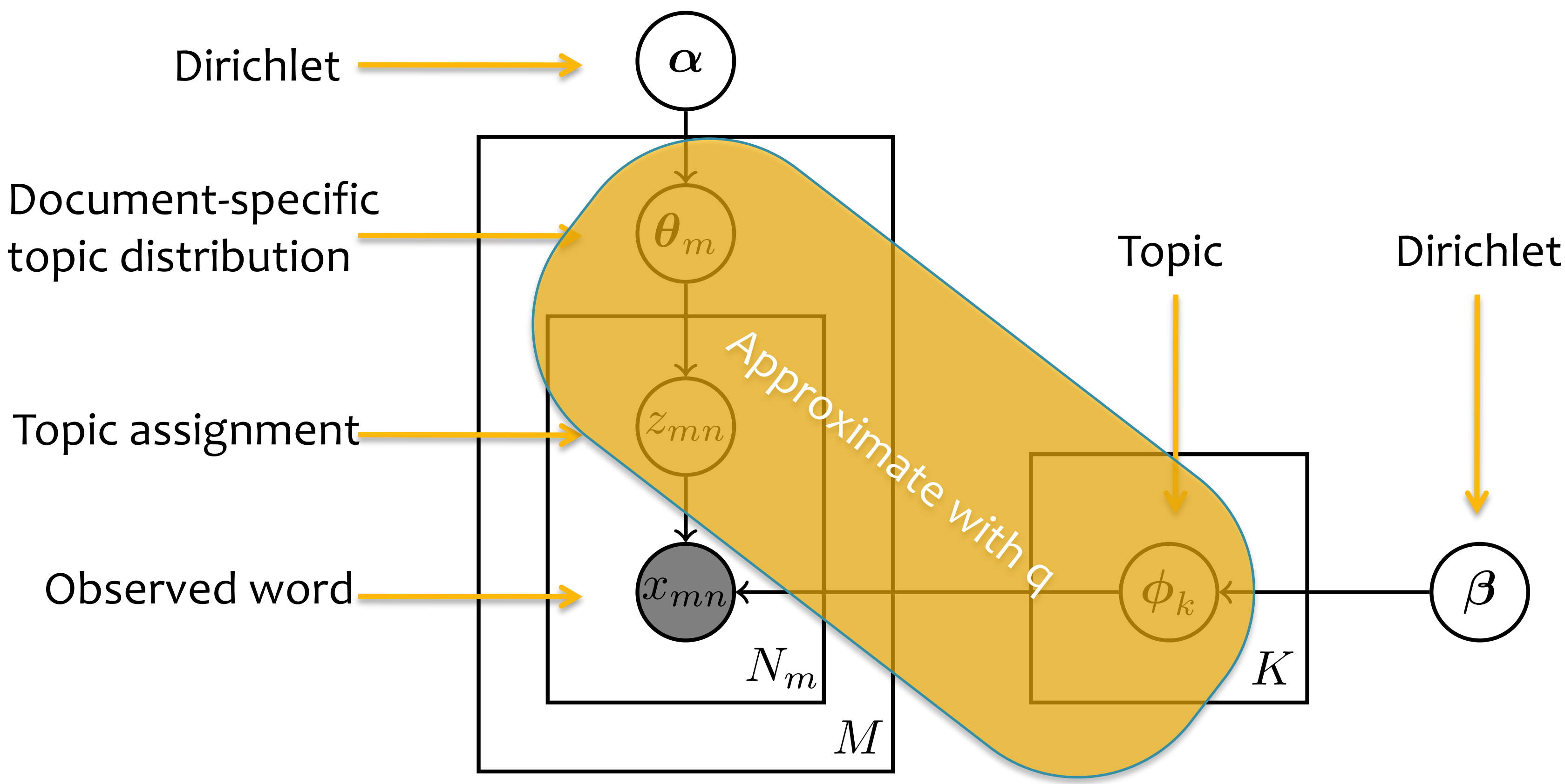
Mean Field for Factorial HMM

- For a factorial HMM, we could decompose into chains



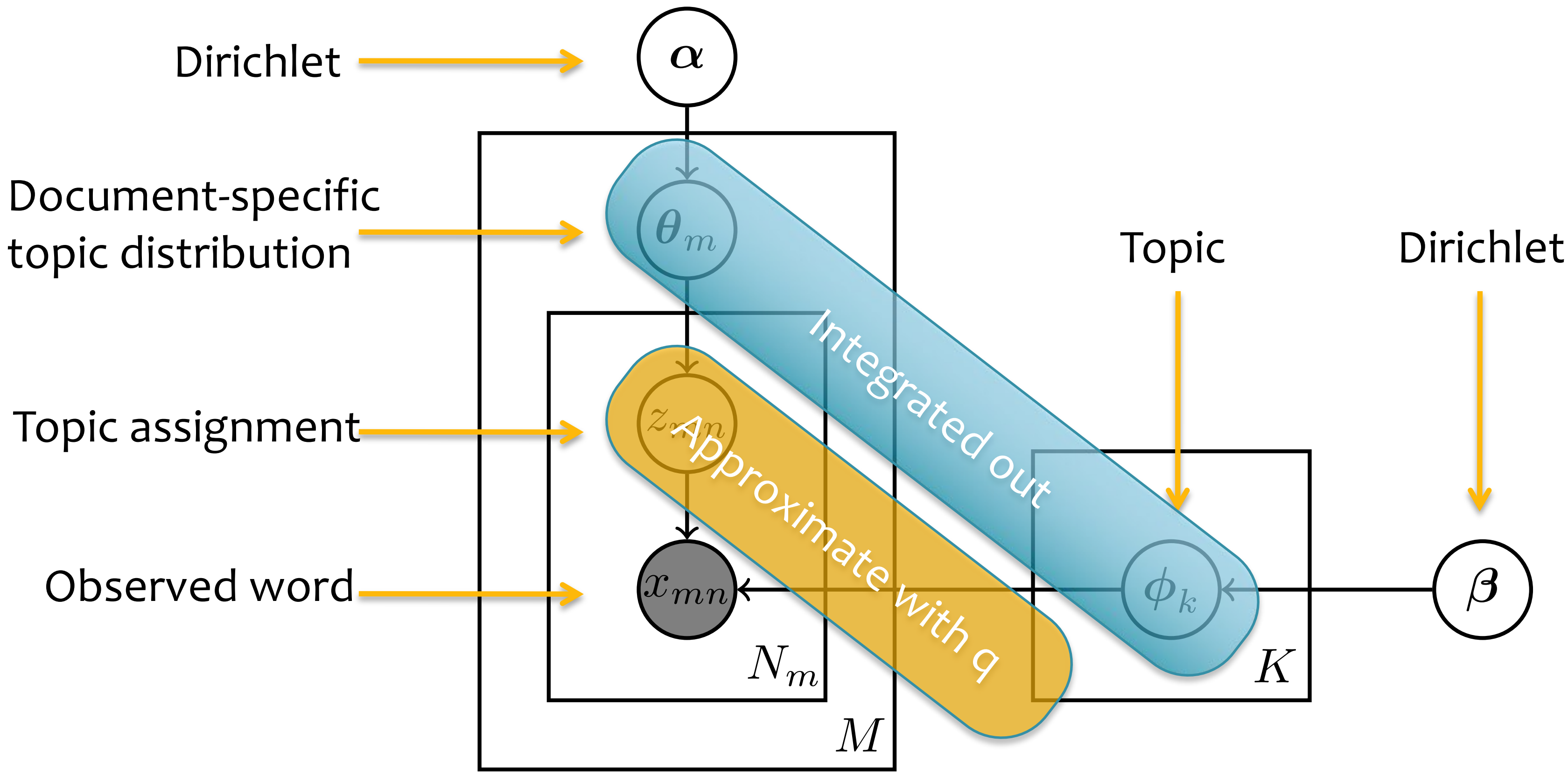
LDA Inference

- Explicit Variational Inference



LDA Inference

- Collapsed Variational Inference



MEAN FIELD VARIATIONAL INFERENCE

Variational Inference

Whiteboard

- Background: KL Divergence
- Mean Field Variational Inference (overview)
- Evidence Lower Bound (ELBO)
- ELBO's relation to log p(x)
- Mean Field Variational Inference (derivation)
- Algorithm Summary (CAVI)
- Example: Factor Graph with Discrete Variables

Variational Inference

Whiteboard

– Example: two variable factor graph
  - Iterated Conditional Models
  - Gibbs Sampling
  - Mean Field Variational Inference