

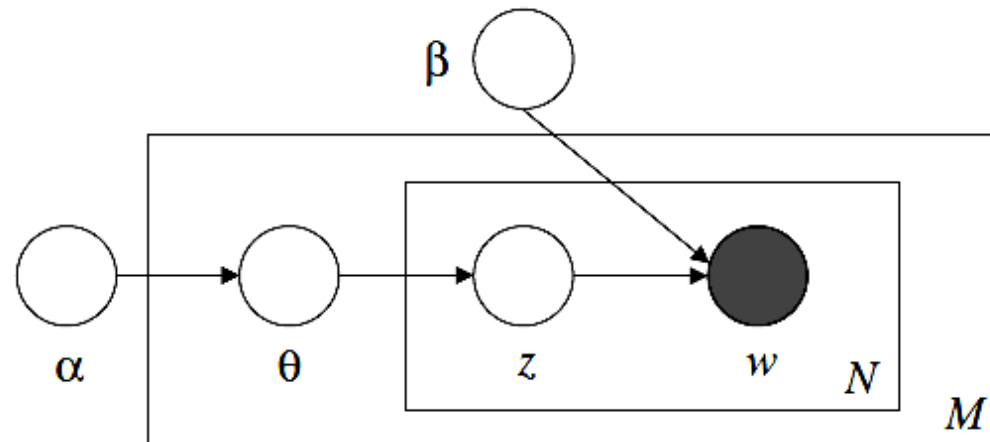
Latent Dirichlet Allocation

Blei, Ng and Jordan (2002)

Presented by Deepak Santhanam

What is Latent Dirichlet Allocation ?

- Generative Model for collections of discrete data
 - Data generated by parameters which can be learned and used to do inference.



Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

LDA is a hierarchical Bayesian Model

LDA and Document Modeling

A Document of a collection is modeled as a finite mixture over underlying topics.

Topics in turn are modeled as an infinite mixture over an underlying set of topic probabilities.

Topic probabilities are explicit representations of a document.

Find short descriptions of members while preserving statistical relations.

Document classification is easier with LDA

Previous Schemes for Document Modeling

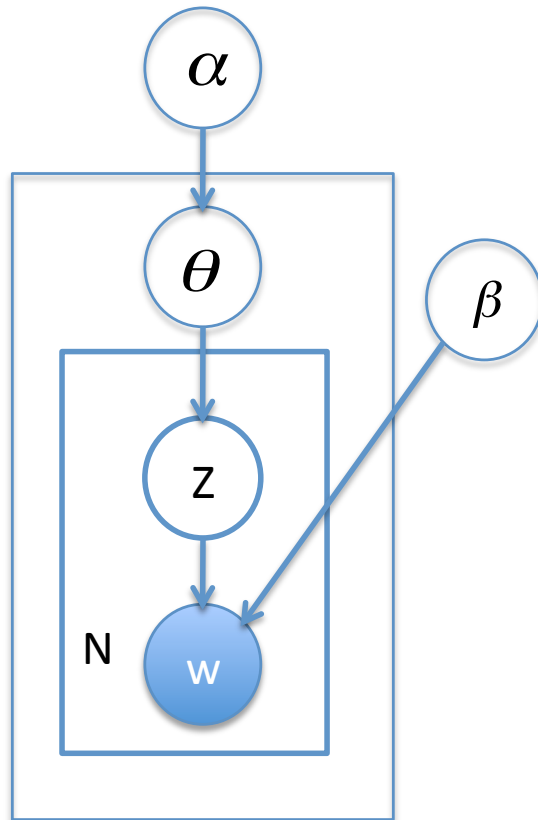
tf – idf scheme where counts are taken for each word and document is modeled.

Latent Semantic Indexing which uses SVD to capture tf-idf features which capture most of the variance.

pLSI – Each word in a document is a sample from a mixture model and generated from a single topic.

(Each document is represented as a mixing proportions of topics and there is not probabilistic model for these proportions)

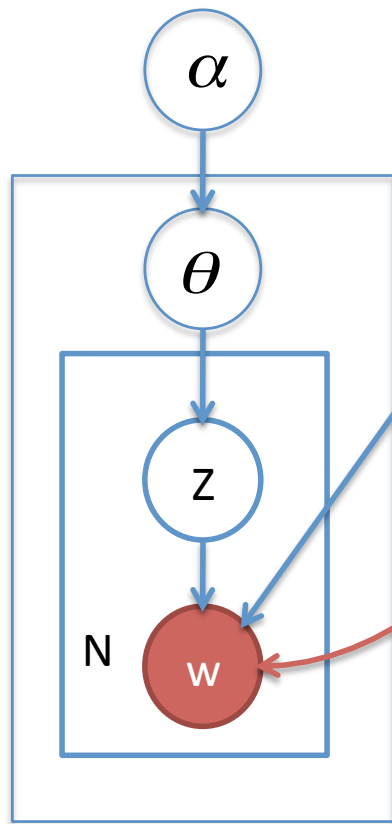
An Early Example..



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

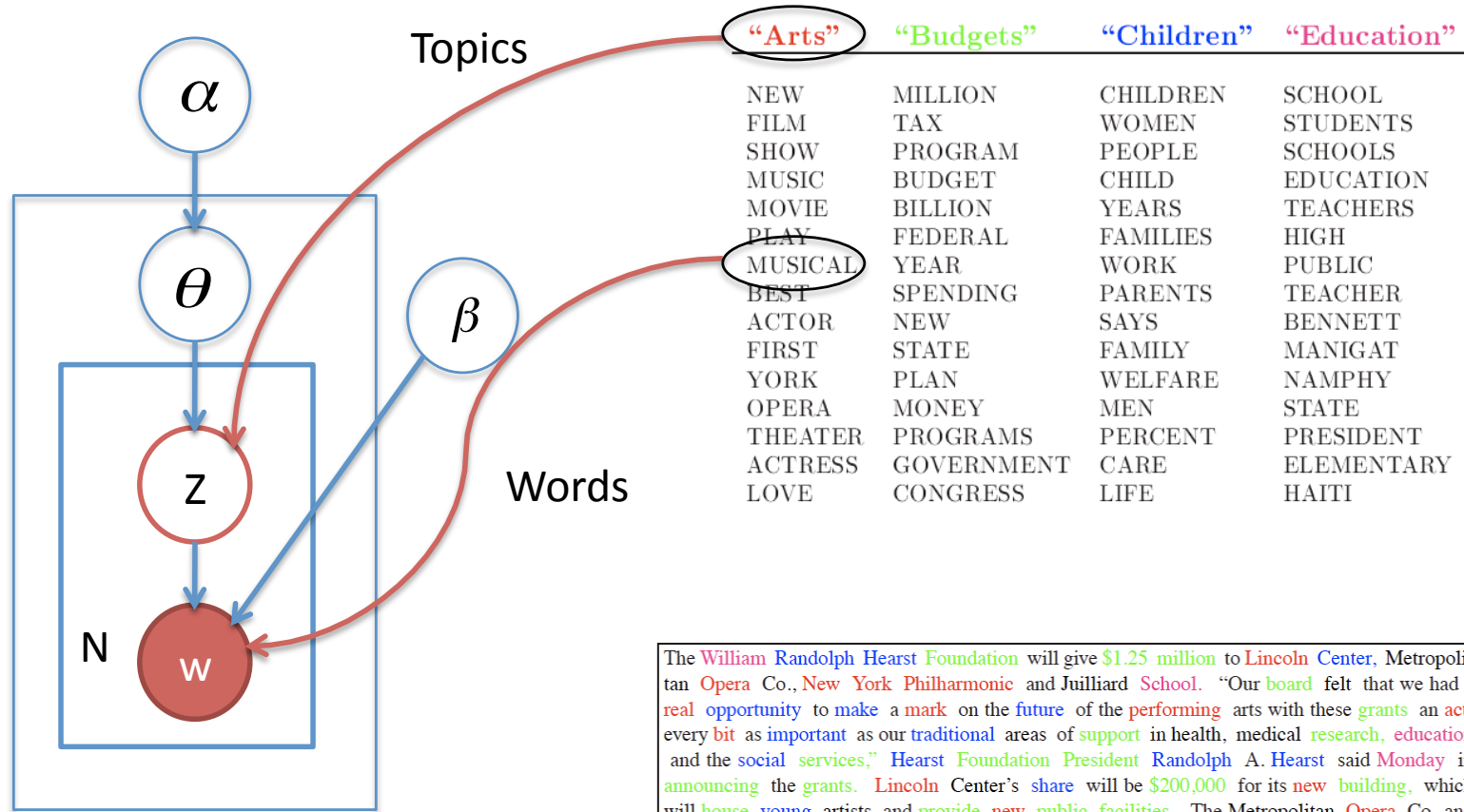
An Early Example..



“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

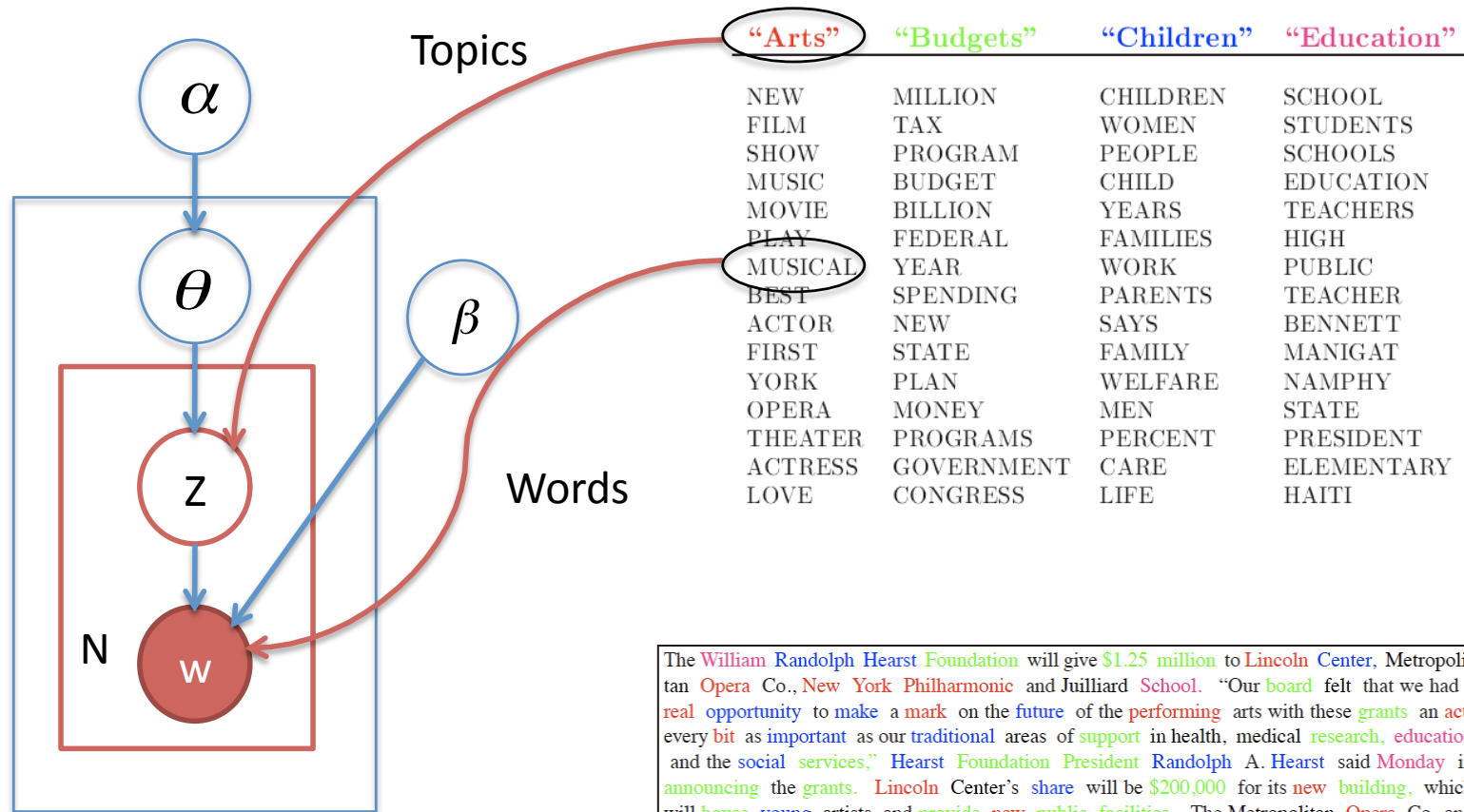
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

An Early Example..



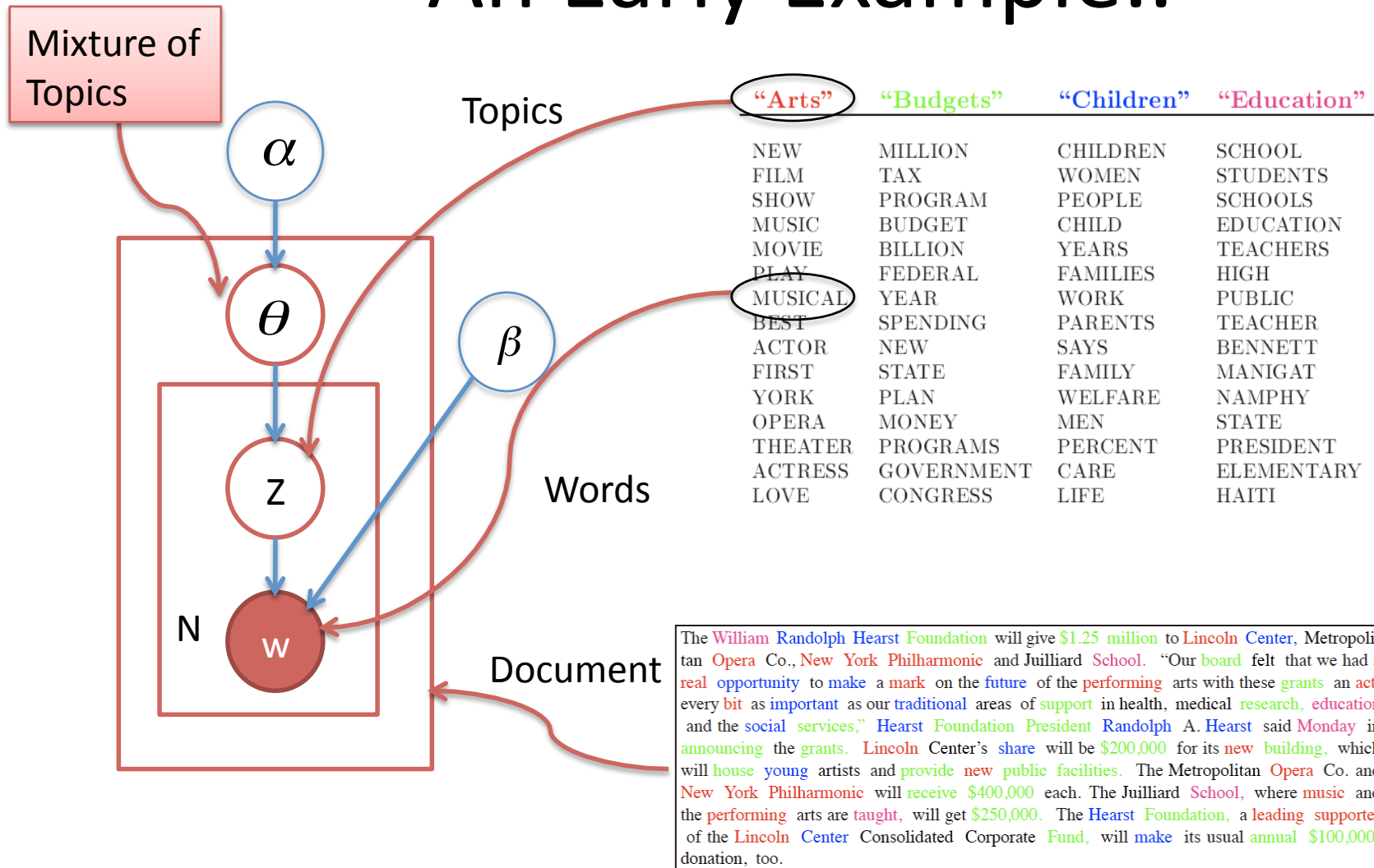
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

An Early Example..

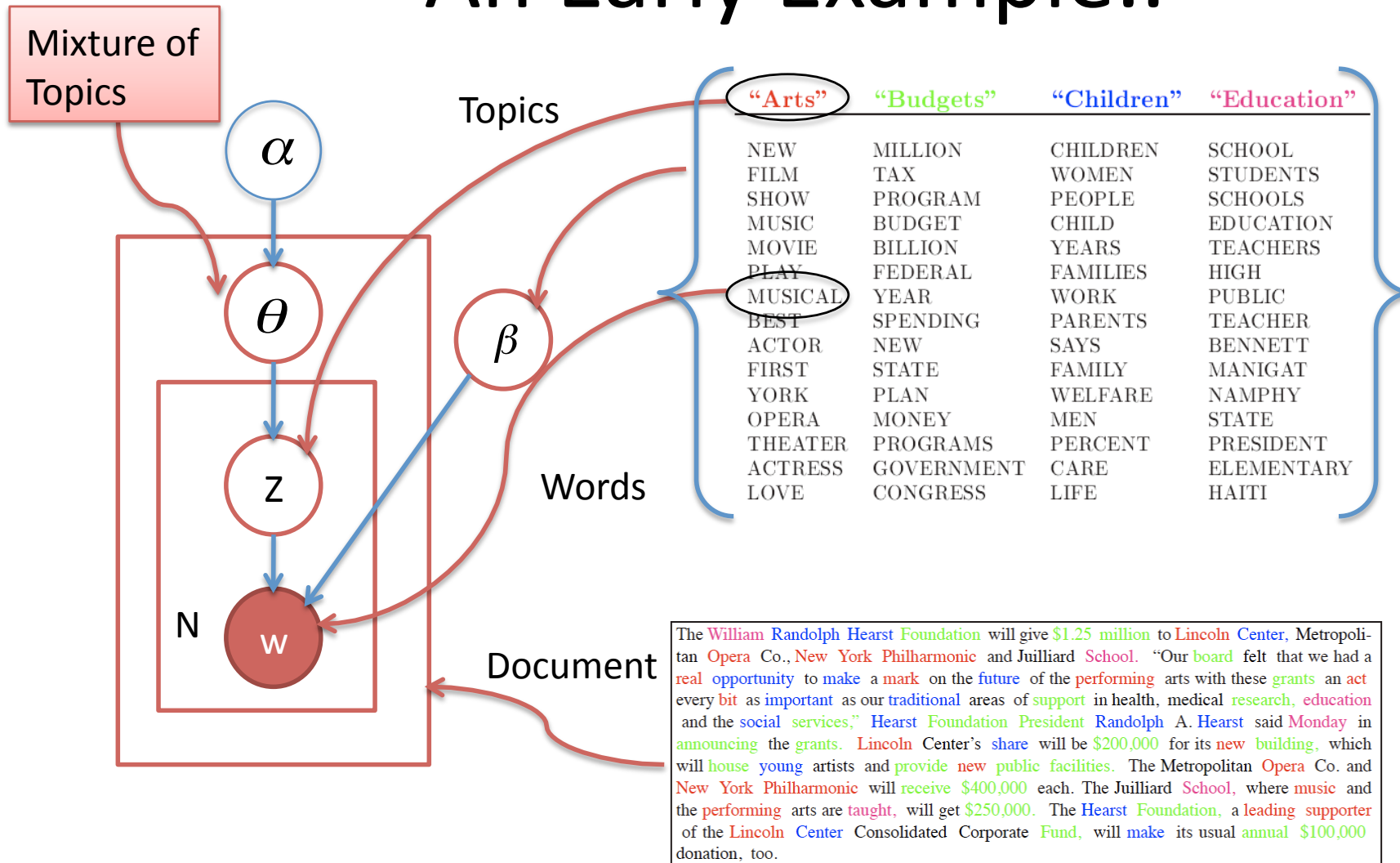


The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

An Early Example..



An Early Example..



Exchangeability and Bag of Words

Assumption that the order of words in the document can be neglected

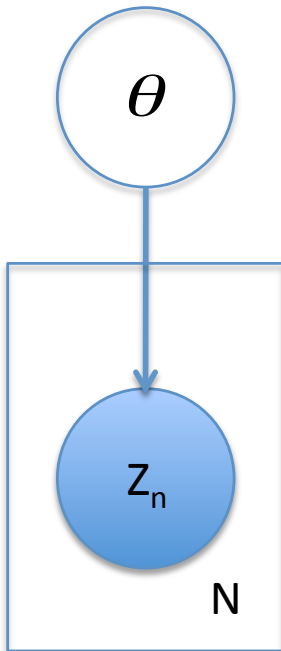
A finite set of Random Variables $\{x_1, \dots, x_N\}$ is exchangeable if the joint distribution is invariant to any permutation of these RVs. i.e. if σ is a permutation of 1 to N:

$$P(x_1, \dots, x_N) = P(x_{\sigma(1)}, \dots, x_{\sigma(N)})$$

e.g : Any weighted average of i.i.d sequences of random variables is exchangeable.

De Finetti's Theorem

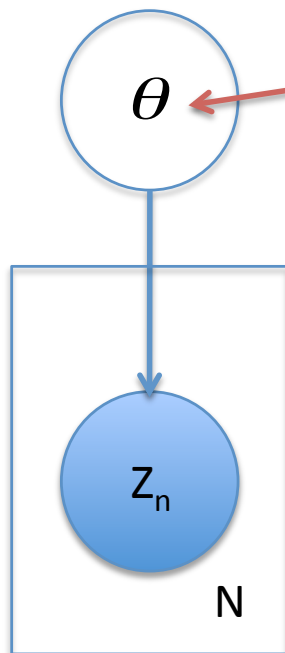
Can Rewrite the Joint of an infinitely exchangeable sequence of RVs by drawing a **random parameter** from **some distribution** and treating the RVs as **i.i.d** conditioned on that random parameter.



$$p(z_1, \dots, z_N) = \int p(\theta) \left(\prod_{n=1}^N p(z_n | \theta) \right) d\theta.$$

De Finetti's Theorem

Can Rewrite the Joint of an infinitely exchangeable sequence of RVs by drawing a **random parameter** from **some distribution** and treating the RVs as **i.i.d** conditioned on that random parameter.

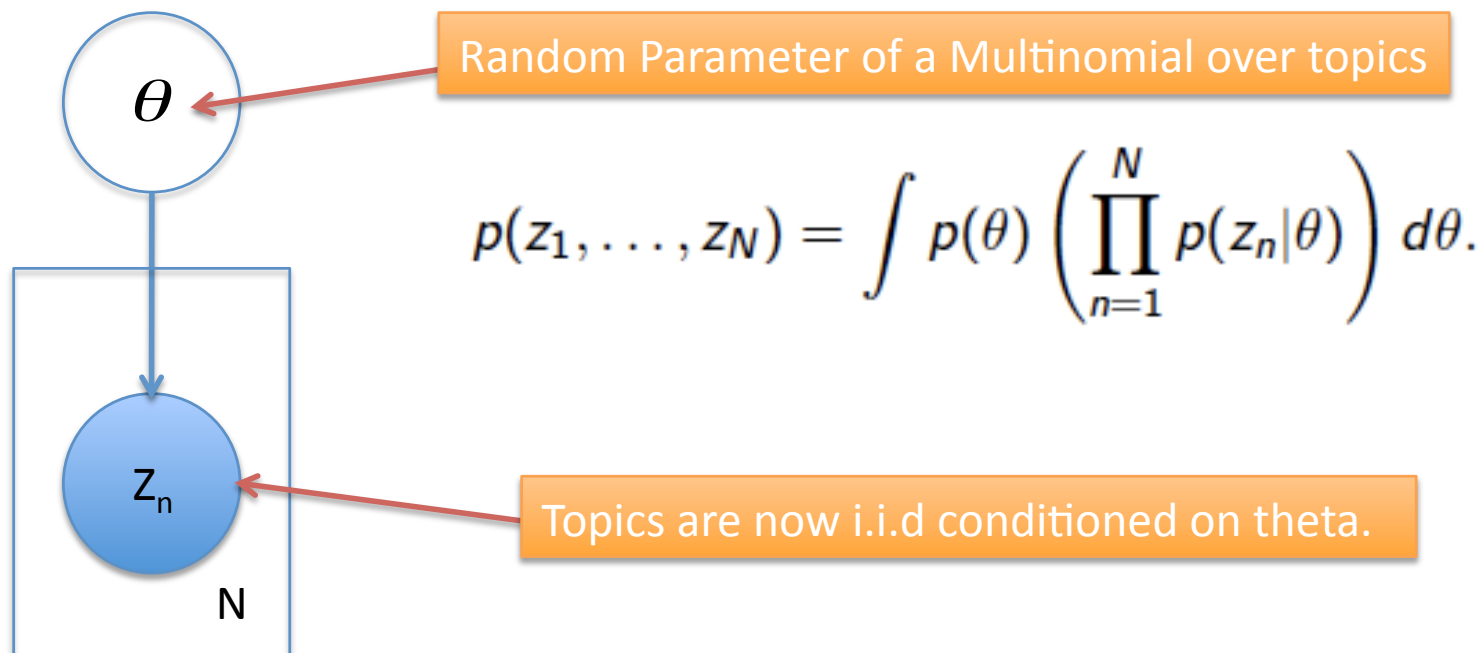


Random Parameter of a Multinomial over topics

$$p(z_1, \dots, z_N) = \int p(\theta) \left(\prod_{n=1}^N p(z_n | \theta) \right) d\theta.$$

De Finetti's Theorem

Can Rewrite the Joint of an infinitely exchangeable sequence of RVs by drawing a **random parameter** from **some distribution** and treating the RVs as **i.i.d conditioned** on that random parameter.



LDA and Exchangeability

- Words are generated by topics with a fixed conditional distribution
- Topics are infinitely exchangeable within a document.
- For a document $\mathbf{W} = (w_1, w_2, \dots, w_N)$ of N words and a corpus of \mathbf{M} documents $\mathbf{C} = \{W_1, W_2, \dots, W_M\}$ for k topics denoted by \mathbf{z} ,

$$p(w, z) = \int p(\theta) \left(\prod_{n=1}^N p(z_n | \theta) p(w_n | z_n) \right) d(\theta)$$

What type of distribution can be used to make it easy for inference ?

The Dirichlet Distribution

- A K-Dimensional Dirichlet RV θ can take values in the (k-1) simplex and has the following density on that simplex

$$p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1},$$

Where α is a k-vector with components greater than 0.

Dirichlet makes it easy for inference as it has finite dimensional sufficient statistics and is a conjugate to the Multinomial distribution.

Generative Process of LDA

- Choose $N \sim \text{Poisson}(\xi)$
- Choose $\theta \sim \text{Dir}(\alpha)$
- For each word w_n :
 - choose a topic $z_n \sim \text{Multinomial}(\theta)$
 - Choose a word w_n from $p(w_n | z_n, \beta)$ a multinomial probability conditioned on the topic z_n
 - Beta is a $k \times v$ Matrix and $\beta_{ij} = p(w^j = 1 | z^i = 1)$

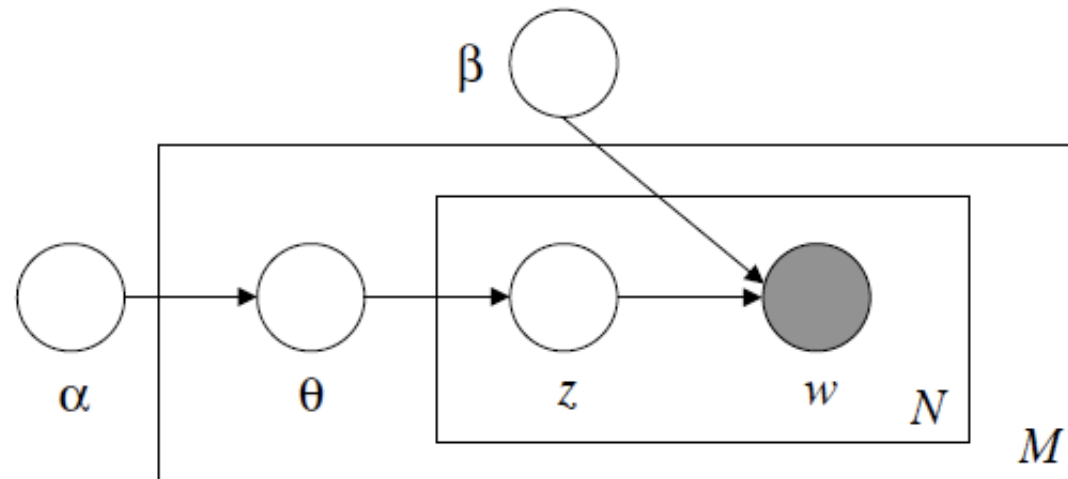
Graphical Model of LDA

The joint over the topics and words is given by,

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta),$$

↑
Sampled once per corpus

↓ ↓
Sampled once every document Sampled once every word




The Marginal of a Document and The Probability of the Corpus.

- Integrating over the topic mixtures and summing over the words gives the Marginal of a document.

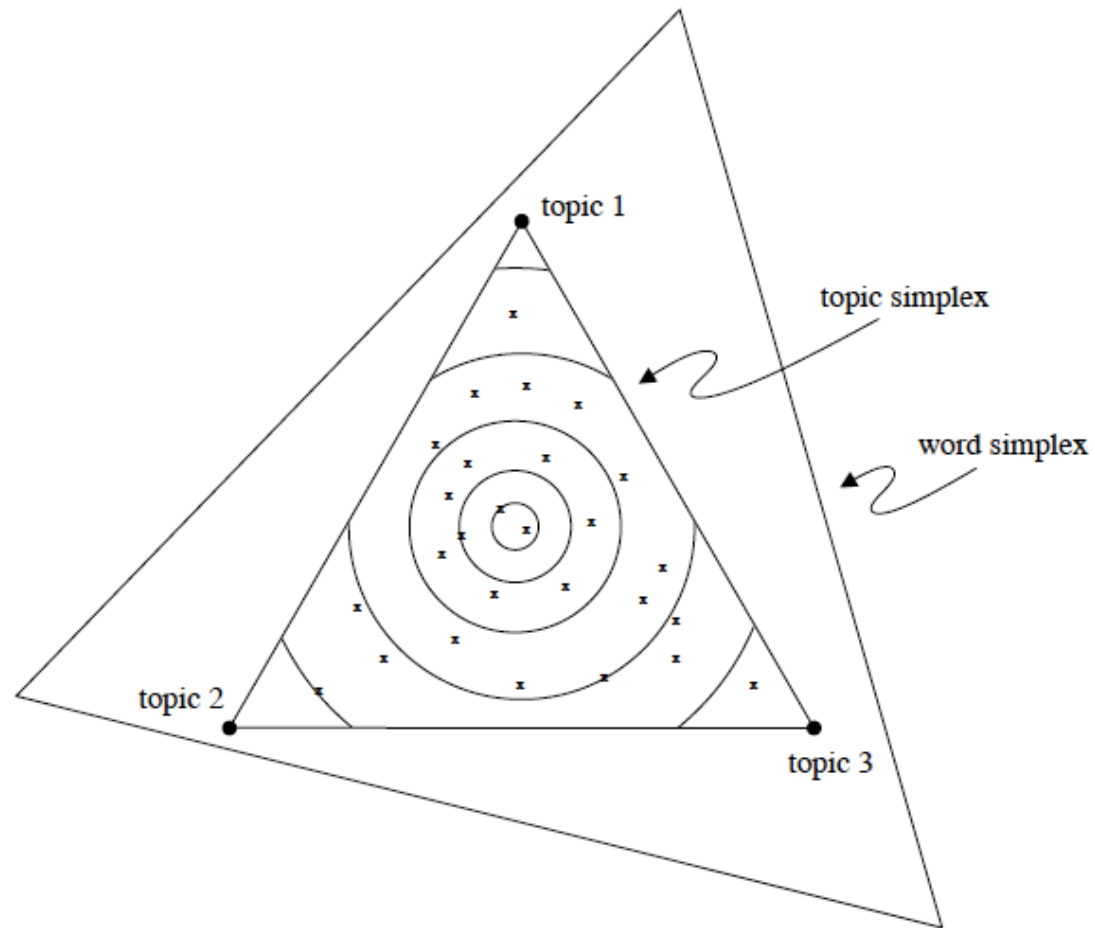
$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta.$$

- Product of the Marginals of all documents gives the probability of the corpus

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d.$$


Corpus level Document Level Word Level

Geometric Representation



Inference Problem

We have to find the Posterior of the latent variables of a document.

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}.$$

Intractable cause we need to marginalize over hidden variables.

$$p(\mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta,$$

Tight Coupling between two parameters

Use approximate inference like **MCMC** or **variational** methods.

Variational Inference

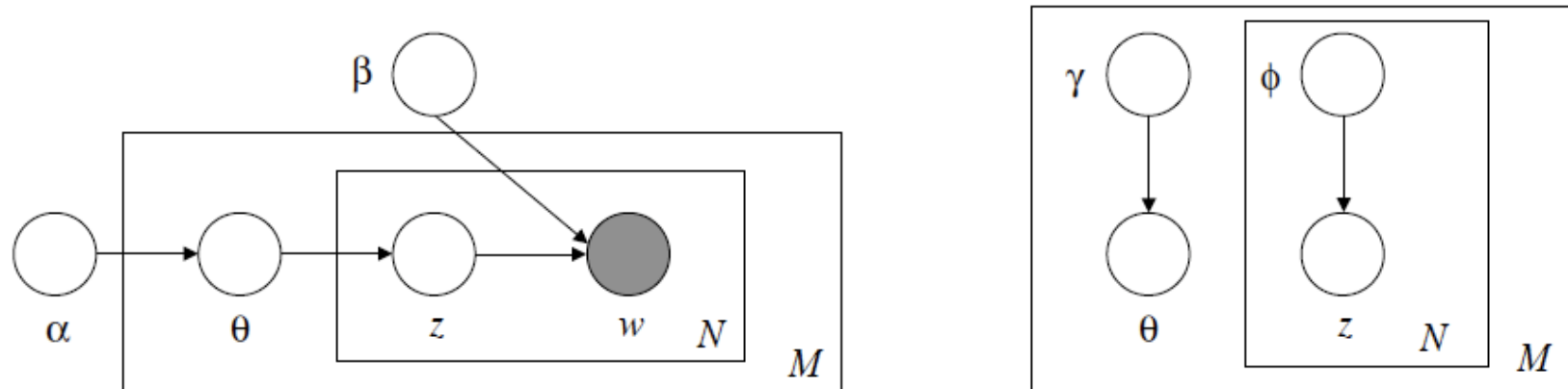


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

- Drop edges which cause the coupling in graphical model.
- Simplified graphical model with free variational parameters
- Problematic coupling not present in the simpler graphical model.

Variational Inference

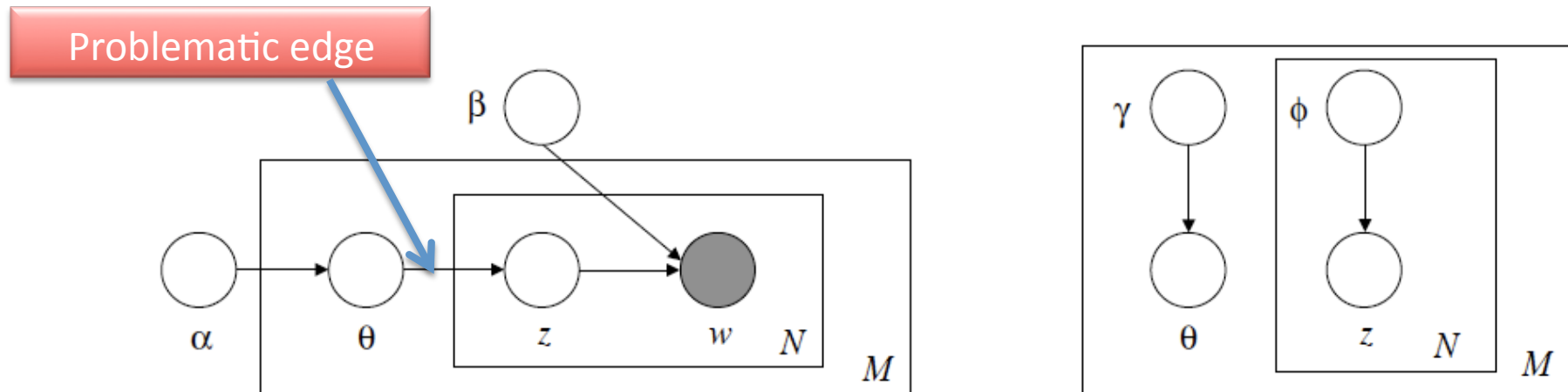


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

- Drop edges which cause the coupling in graphical model.
- Simplified graphical model with free variational parameters
- Problematic coupling not present in the simpler graphical model.

Variational Inference

Results in the following distribution :

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n),$$

Minimize the Kullback-Leibler divergence.

$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D(q(\theta, \mathbf{z} | \gamma, \phi) || p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)).$$

Equating derivatives of KL to zero, we get the update equations,

$$\begin{aligned} \phi_{ni} &\propto \beta_{i w_n} \exp\{E_q[\log(\theta_i) | \gamma]\} \\ \gamma_i &= \alpha_i + \sum_{n=1}^N \phi_{ni}. \end{aligned}$$

Variational Inference

Results in the following distribution :

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n),$$

Dirichlet Parameter

Multinomial Parameter

Minimize the Kullback-Leibler divergence.

$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D(q(\theta, \mathbf{z} | \gamma, \phi) || p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)).$$

Equating derivatives of KL to zero, we get the update equations,

$$\begin{aligned} \phi_{ni} &\propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) | \gamma]\} \\ \gamma_i &= \alpha_i + \sum_{n=1}^N \phi_{ni}. \end{aligned}$$

Parameter Estimation Using empirical Bayes

- Find the parameters which maximize the log likelihood of data.
- Intractable for same reasons.
- Variational inference provided a tight lower bound.

Alternating Variational Expectation Maximization:

- ***E-Step*** : for each document find optimizing values of **variational parameters** (γ, ϕ) .
- ***M-Step***: Maximize the lower bound on the likelihood with respect to the **model parameters** (α, β) .

Smoothing

- Likelihood of previously unseen documents is always zero.
- Smooth matrix β by considering its elements as RVs with a posterior conditioned on data.
- Do the whole inference procedure again for new model to get new update equations.

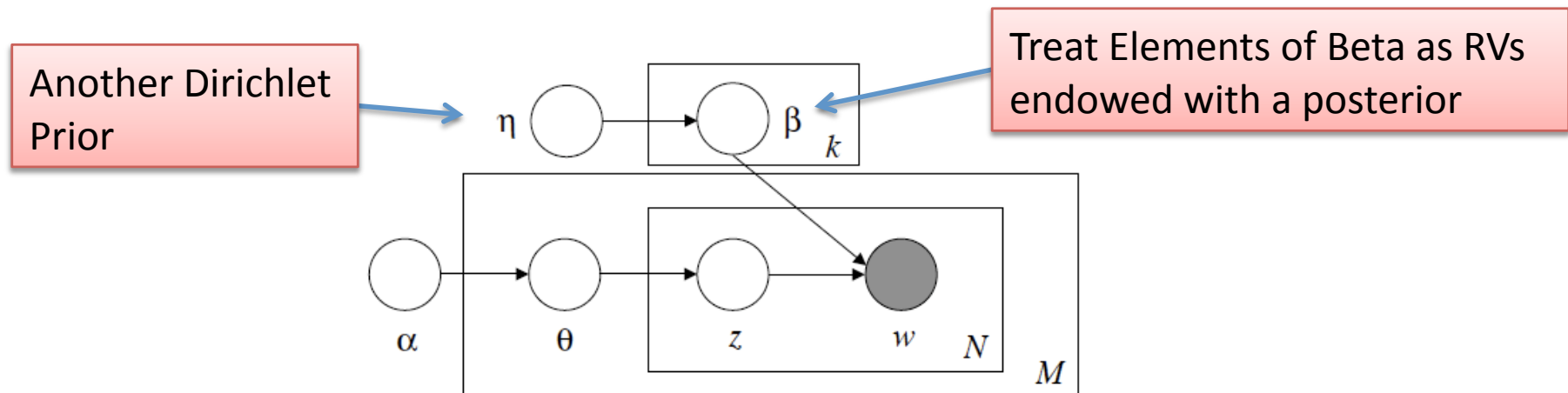


Figure 7: Graphical model representation of the smoothed LDA model.

Smoothing

- Likelihood of previously unseen documents is always zero.
- Smooth matrix β by considering its elements as RVs with a posterior conditioned on data.
- Do the whole inference procedure again for new model to get new update equations.

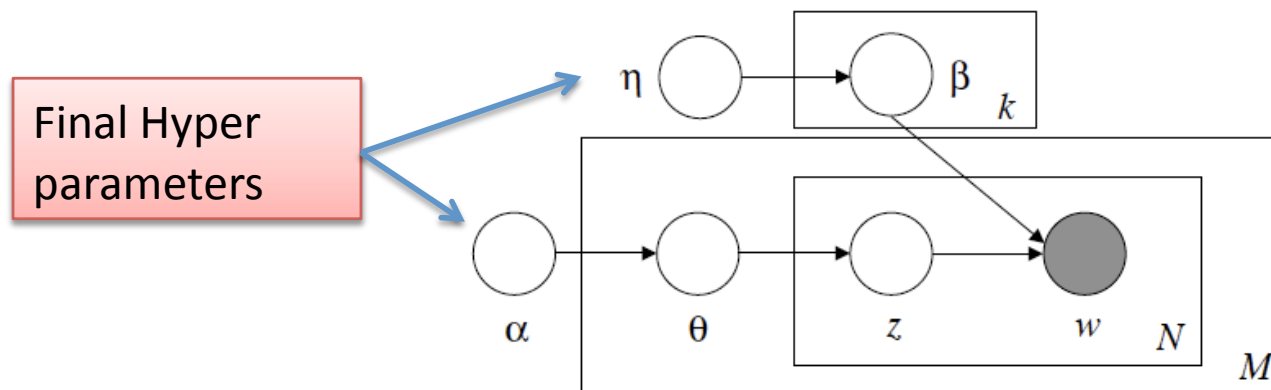


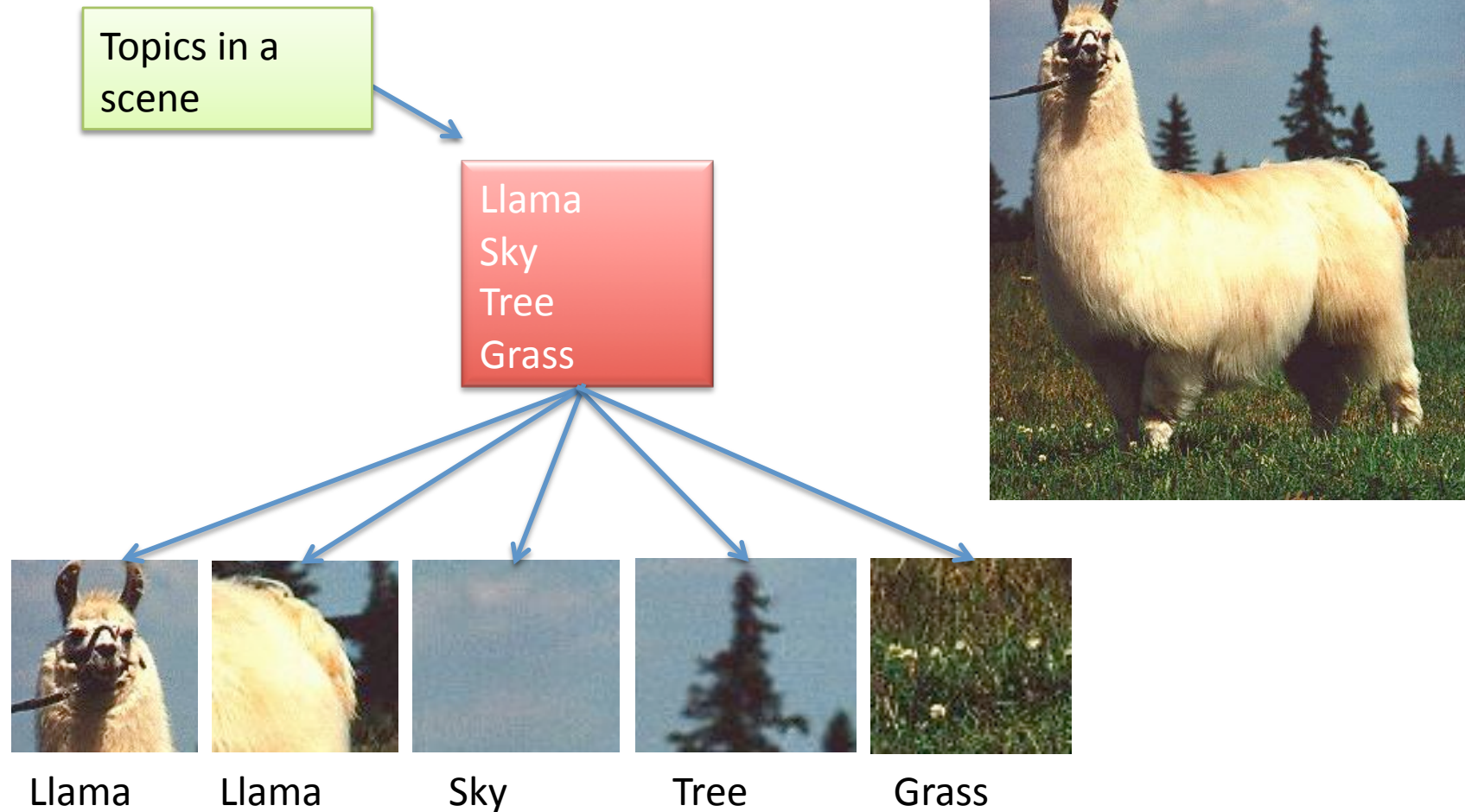
Figure 7: Graphical model representation of the smoothed LDA model.

Extending LDA

- Make a continuous variant using gaussians instead of multinomials.
- Particular form of clustering by having a mixture of Dirichlet distributions instead of one.
- What must be done to extend LDA to a more useful model?
- Can we use this LDA model in Computer Vision?

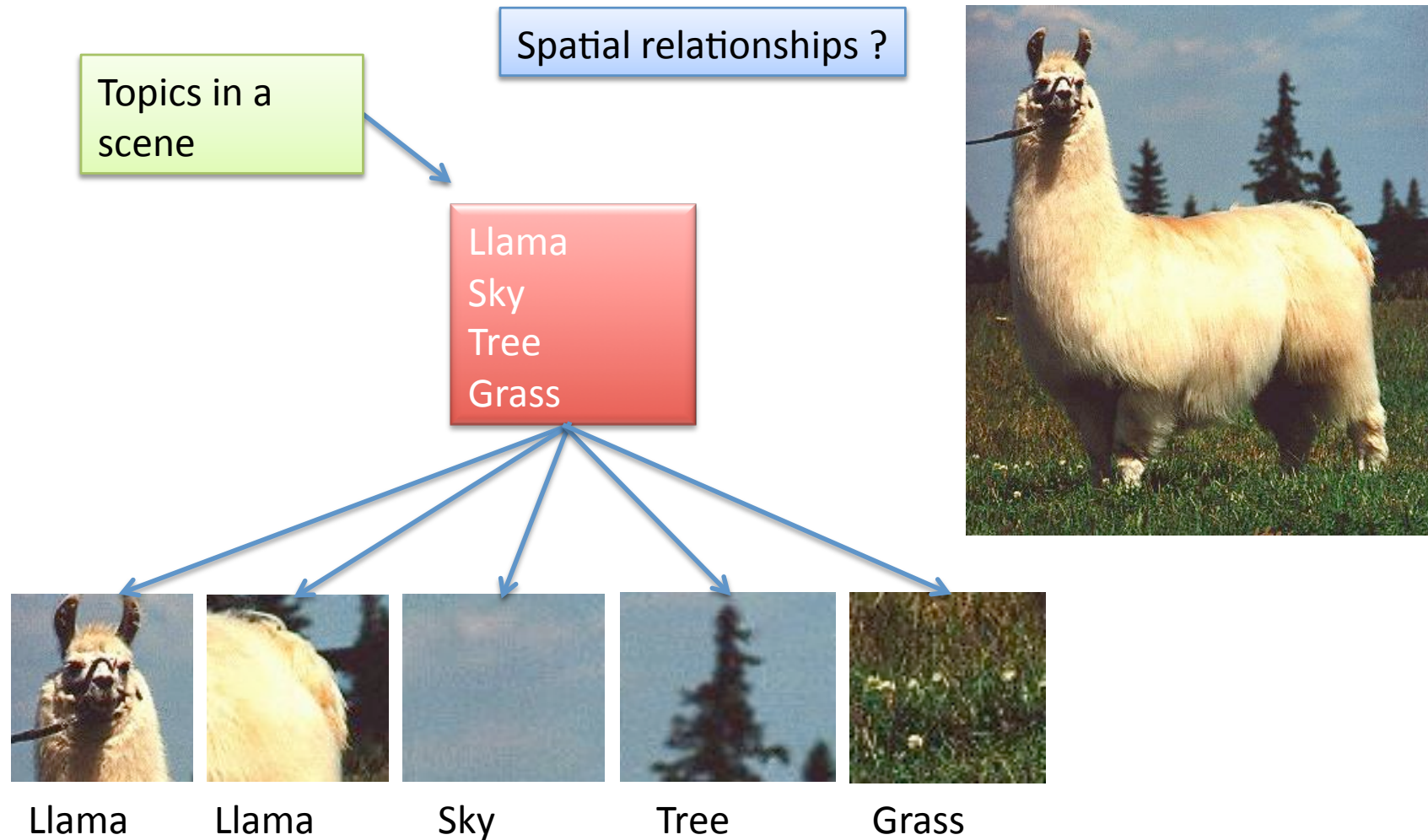
Application in Computer Vision

One of the methods extended in *Describing visual scenes (Sudderth et al 2005)*



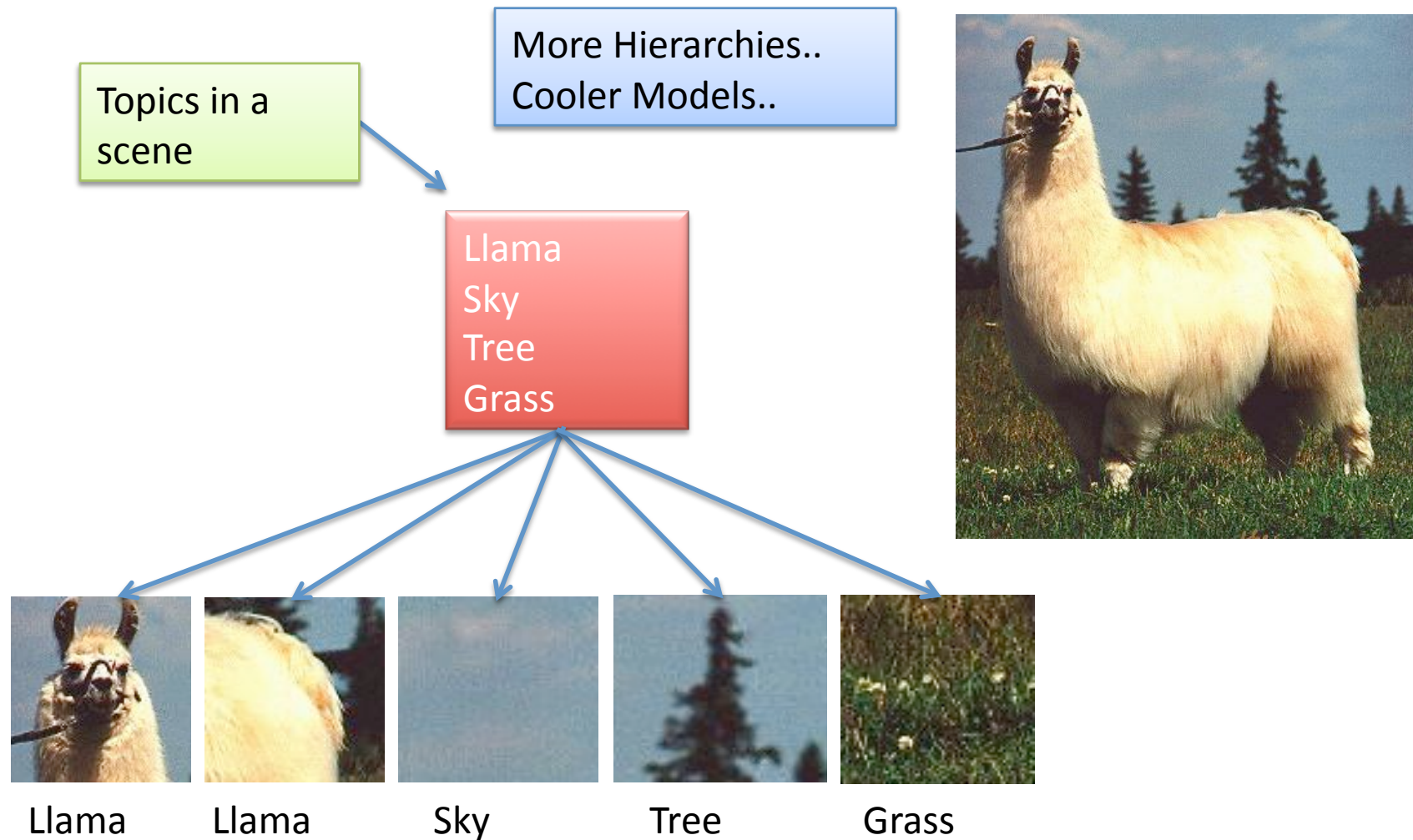
Application in Computer Vision

One of the methods extended in *Describing visual scenes (Sudderth et al 2005)*

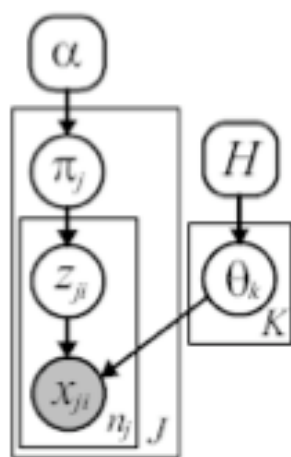


Application in Computer Vision

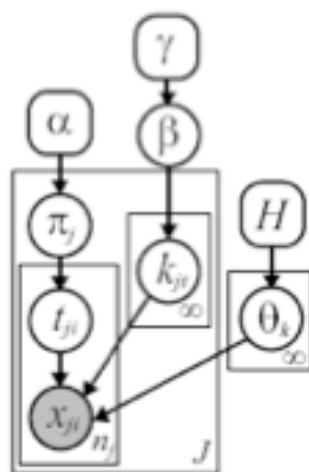
One of the methods extended in *Describing visual scenes (Sudderth et al 2005)*



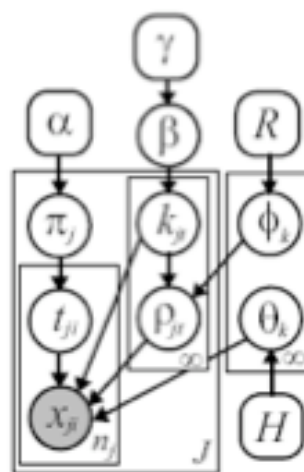
Even cooler models..!



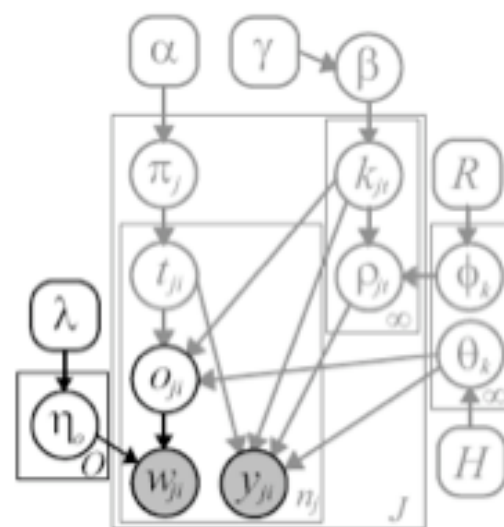
LDA Graphical Model



HDP Graphical Model



TDP Graphical Model



Visual Scene TDP Graphical Model

Take home message.

LDA illustrates how Probabilistic models can be scaled up.

With good inference techniques, we can solve hard problems in multiple domains which have a multiple hierarchies.

Generative models are modular and extensible easily.

Thank You !