

CS 6347

Lecture 18

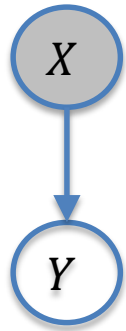
Topic Models and LDA

(some slides by David Blei)

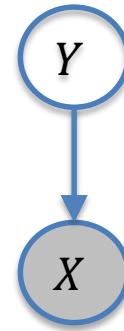
Generative vs. Discriminative Models



- Recall that, in Bayesian networks, there could be many different, but equivalent models of the same joint distribution



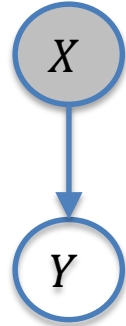
Discriminative



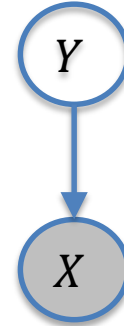
Generative

- Although these two models are equivalent (in the sense that they imply the same independence relations), they can differ significantly when it comes to inference/prediction

Generative vs. Discriminative Models



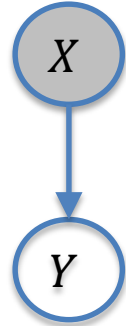
Discriminative



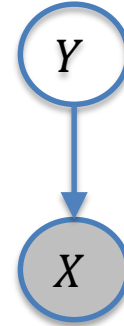
Generative

- Generative models: we can think of the observations as being generated by the latent variables
 - Start sampling at the top and work downwards
 - Examples?

Generative vs. Discriminative Models



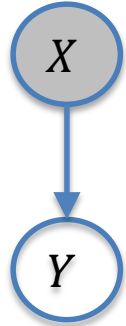
Discriminative



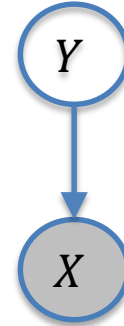
Generative

- Generative models: we can think of the observations as being generated by the latent variables
 - Start sampling at the top and work downwards
 - Examples: **HMMs, naïve Bayes, LDA**

Generative vs. Discriminative Models



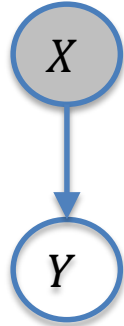
Discriminative



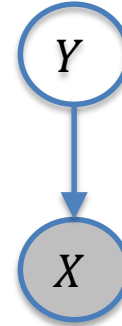
Generative

- Discriminative models: most useful for discriminating the values of the latent variables
 - Almost always used for supervised learning
 - Examples?

Generative vs. Discriminative Models



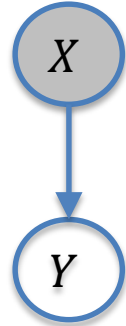
Discriminative



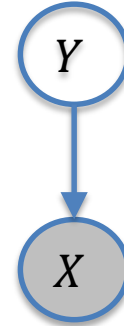
Generative

- Discriminative models: most useful for discriminating the values of the latent variables
 - Almost always used for supervised learning
 - Examples: **CRFs**

Generative vs. Discriminative Models



Discriminative



Generative

- Suppose we are only interested in the prediction task (i.e., estimating $p(Y|X)$)
 - Discriminative model: $p(X, Y) = p(X)p(Y|X)$
 - Generative model: $p(X, Y) = p(Y)p(X|Y)$

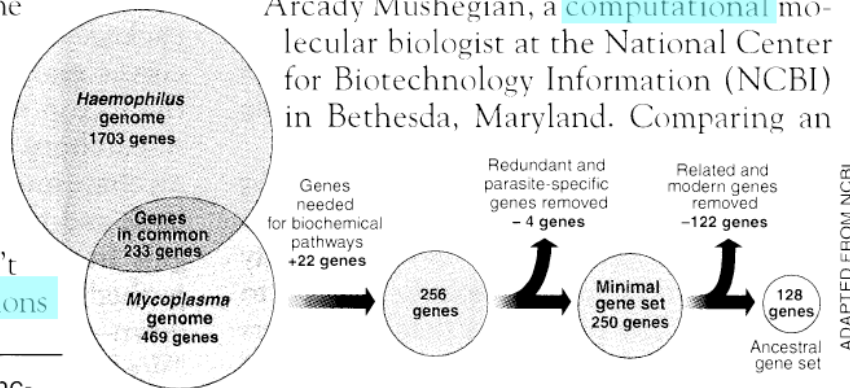
- The primary advantage of generative models is that they provide a model of the data generating process
 - Could generate “new” data samples by using the model
- Topic models (generative models of documents)
 - Methods for discovering themes (topics) from a collection (e.g., books, newspapers, etc.)
 - Annotates the collection according to the discovered themes
 - Use the annotations to organize, search, summarize, etc.

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

- Bag-of-words model: assume that the ordering of words in a document do not matter
 - This is typically false as certain phrases can only appear together
- Unigram model: all words in a document are drawn uniformly at random from categorical distribution
- Mixture of unigrams model: for each document, we first choose a topic z and then generate words for the document from the conditional distribution $p(w|z)$
 - Topics are just probability distributions over words

Latent Dirichlet Allocation (LDA)



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

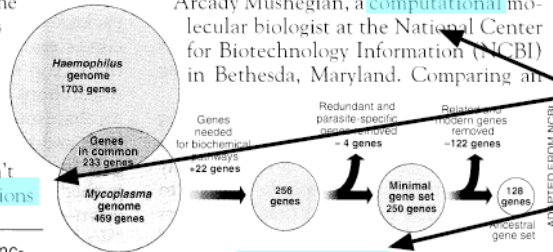
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

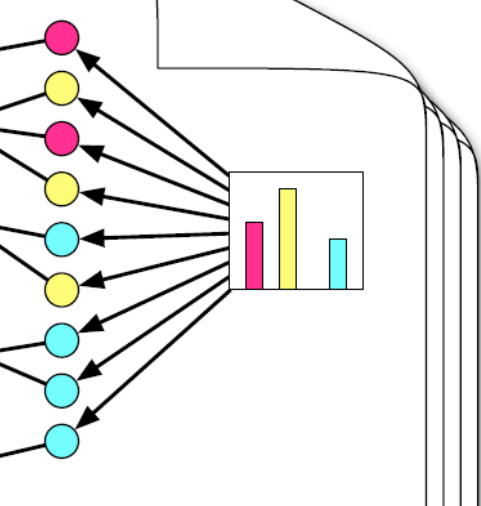


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

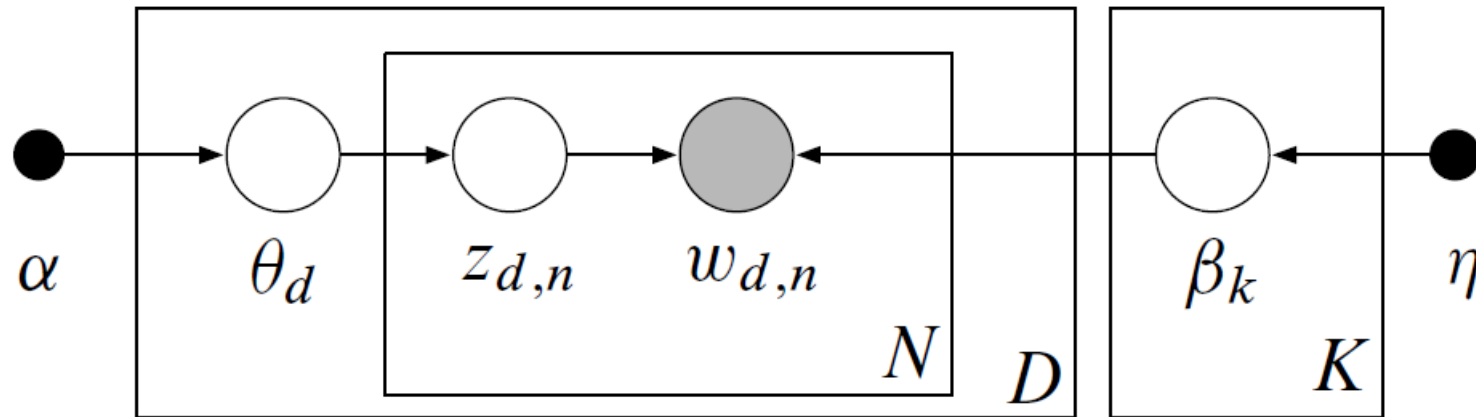
Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

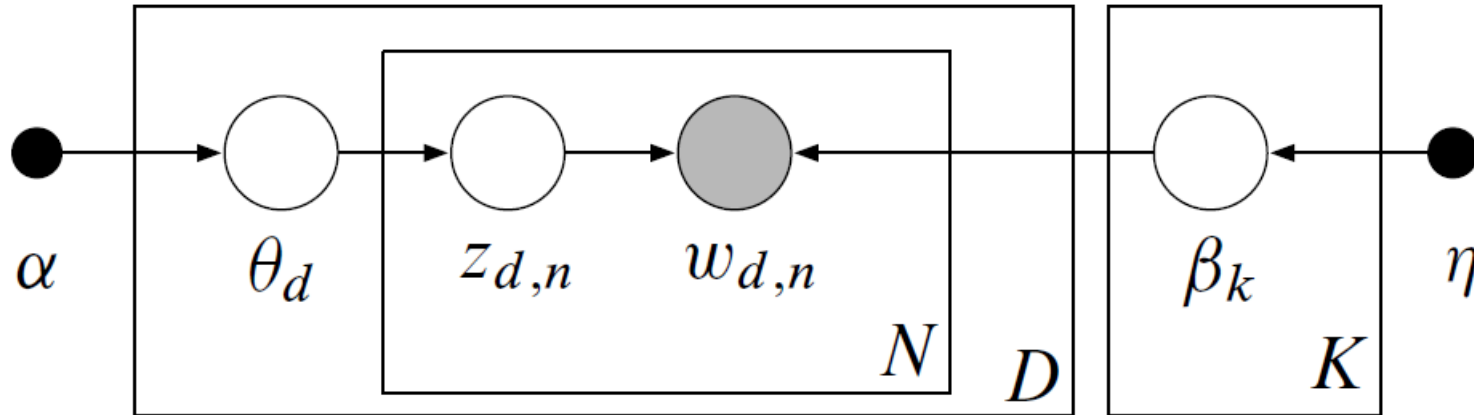


Latent Dirichlet Allocation (LDA)



- α and η are parameters of the prior distributions over θ and β
- θ_d is the distribution of topics for document d (real vector of length K)
- β_k is the distribution of words for topic k (real vector of length V)
- $z_{d,n}$ is the topic for the n th word in the d th document
- $w_{d,n}$ is the n th word of the d th document

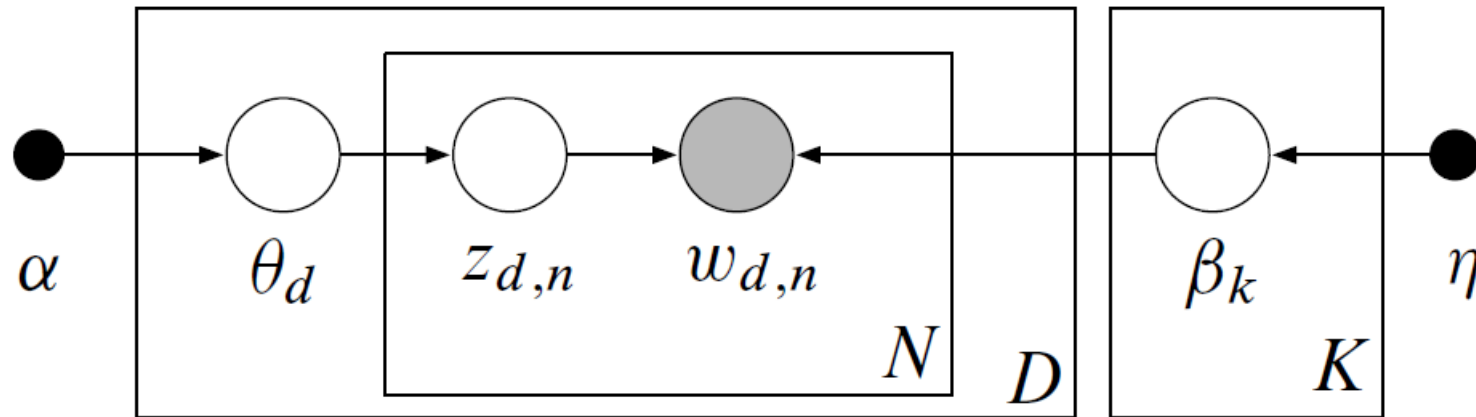
Latent Dirichlet Allocation (LDA)



- **Plate notation**

- There are $N \cdot D$ different variables that represent the observed words in the different documents
- There are K total topics (assumed to be known in advance)
- There are D total documents

Latent Dirichlet Allocation (LDA)



- The only observed variables are the words in the documents
 - The topic for each word, the distribution over topics for each document, and the distribution of words per topic are all latent variables in this model

- The model contains both continuous and discrete random variables
 - θ_d and β_k are vectors of probabilities
 - $z_{d,n}$ is an integer in $\{1, \dots, K\}$ that indicates the topic of the n th word in the d th document
 - $w_{d,n}$ is an integer in $\{1, \dots, V\}$ which indexes over all possible words

- $\theta_d \sim \text{Dir}(\alpha)$ where $\text{Dir}(\alpha)$ is the Dirichlet distribution with parameter vector $\alpha > 0$
- $\beta_k \sim \text{Dir}(\eta)$ with parameter vector $\eta > 0$
- Dirichlet distribution over x_1, \dots, x_K such that $x_1, \dots, x_K \geq 0$ and $\sum_i x_i = 1$

$$f(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) \propto \prod_i x_i^{\alpha_i - 1}$$

- The Dirichlet distribution is a distribution over probability distributions over K elements

- The discrete random variables are distributed via the corresponding probability distributions

$$p(z_{d,n} = k | \theta_d) = (\theta_d)_k$$

$$p(w_{d,n} = v | z_{d,n}, \beta_1, \dots, \beta_K) = (\beta_{z_{d,n}})_v$$

- Here, $(\theta_d)_k$ is the k th element of the vector θ_d which corresponds to the percentage of document d corresponding to topic k
- The joint distribution is then

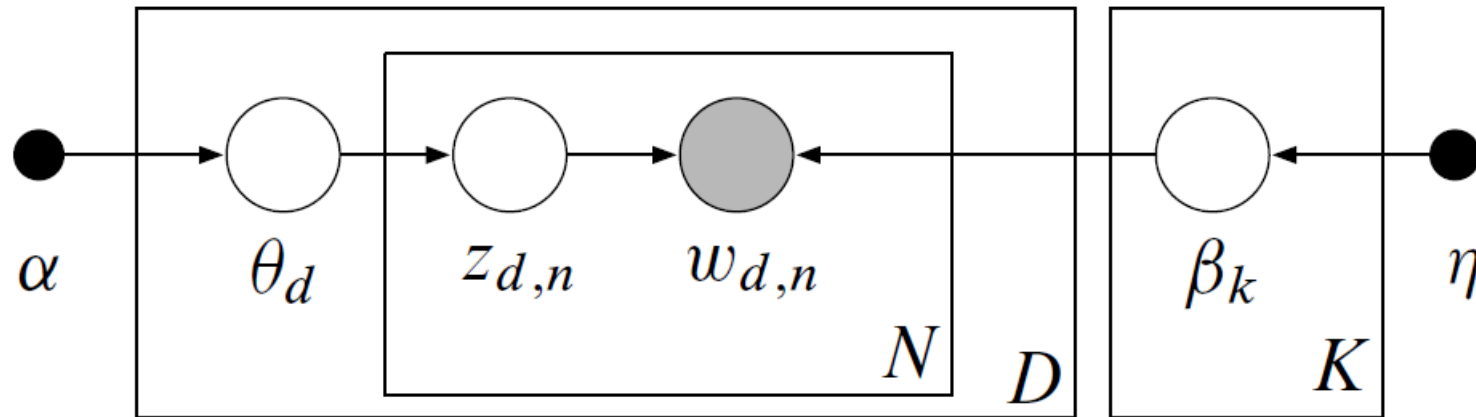
$$p(w, z, \theta, \beta | \alpha, \eta) = \prod_k p(\beta_k | \eta) \prod_d \left[p(\theta_d | \alpha) \prod_n p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta) \right]$$

Latent Dirichlet Allocation (LDA)



- LDA is a generative model
 - We can think of the words as being generated by a probabilistic process defined by the model
 - How reasonable is the generative model?

Latent Dirichlet Allocation (LDA)



- Inference in this model is NP-hard
- Given the D documents, want to find the parameters that best maximize the joint probability
 - Can use an EM based approach called **variational EM**

- Recall that the EM algorithm constructed a lower bound using Jensen's inequality

$$\begin{aligned}l(\theta) &= \sum_{k=1}^K \log \sum_{x_{mis_k}} p(x_{obs}^{(k)}, x_{mis}^{(k)} | \theta) \\&= \sum_{k=1}^K \log \sum_{x_{mis_k}} q_k(x_{mis}^{(k)}) \cdot \frac{p(x_{obs}^{(k)}, x_{mis}^{(k)} | \theta)}{q_k(x_{mis}^{(k)})} \\&\geq \sum_{k=1}^K \sum_{x_{mis_k}} q_k(x_{mis}^{(k)}) \log \frac{p(x_{obs}^{(k)}, x_{mis}^{(k)} | \theta)}{q_k(x_{mis}^{(k)})}\end{aligned}$$

- Performing the optimization over q is equivalent to computing $p(x_{mis}^{(k)} | x_{obs}^{(k)}, \theta)$
- This can be intractable in practice
 - Instead, restrict q to lie in some restricted set of distributions Q
 - For example, could make a mean-field assumption

$$q(x_{mis}^{(k)}) = \prod_{i \in mis_k} q_i(x_i^{(k)})$$

- The resulting algorithm only yields an approximation to the log-likelihood

$$p(w|\alpha, \eta) = \int \prod_k p(\beta_k|\eta) \int \sum_z \prod_d \left[p(\theta_d|\alpha) \prod_n p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta) \right] d\theta d\beta$$

- To apply variational EM, we write

$$\begin{aligned} \log p(w|\alpha, \eta) &= \log \int \int \sum_z p(w, z, \theta, \beta|\alpha, \eta) d\theta d\beta \\ &\geq \int \int \sum_z q(z, \theta, \beta) \log \frac{p(w, z, \theta, \beta|\alpha, \eta)}{q(z, \theta, \beta)} d\theta d\beta \end{aligned}$$

where we restrict the distribution q to be of the following form

$$q(z, \theta, \beta) = \prod_k q(\beta_k|\eta) \prod_d q(\theta_d|\alpha) \prod_n q(z_{d,n})$$

Example of LDA



“Arts”

NEW
FILM
SHOW
MUSIC
MOVIE
PLAY
MUSICAL
BEST
ACTOR
FIRST
YORK
OPERA
THEATER
ACTRESS
LOVE

“Budgets”

MILLION
TAX
PROGRAM
BUDGET
BILLION
FEDERAL
YEAR
SPENDING
NEW
STATE
PLAN
MONEY
PROGRAMS
GOVERNMENT
CONGRESS

“Children”

CHILDREN
WOMEN
PEOPLE
CHILD
YEARS
FAMILIES
WORK
PARENTS
SAYS
FAMILY
WELFARE
MEN
PERCENT
CARE
LIFE

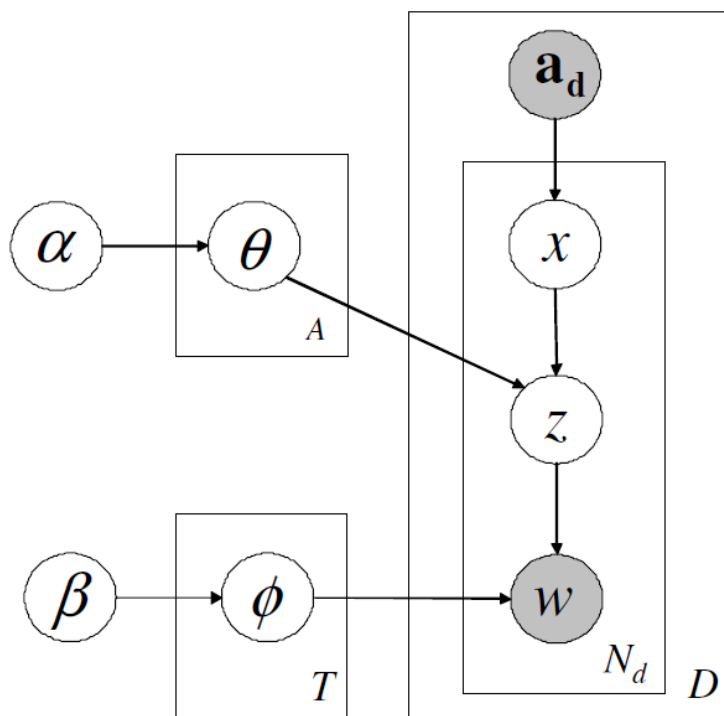
“Education”

SCHOOL
STUDENTS
SCHOOLS
EDUCATION
TEACHERS
HIGH
PUBLIC
TEACHER
BENNETT
MANIGAT
NAMPHY
STATE
PRESIDENT
ELEMENTARY
HAITI

Example of LDA



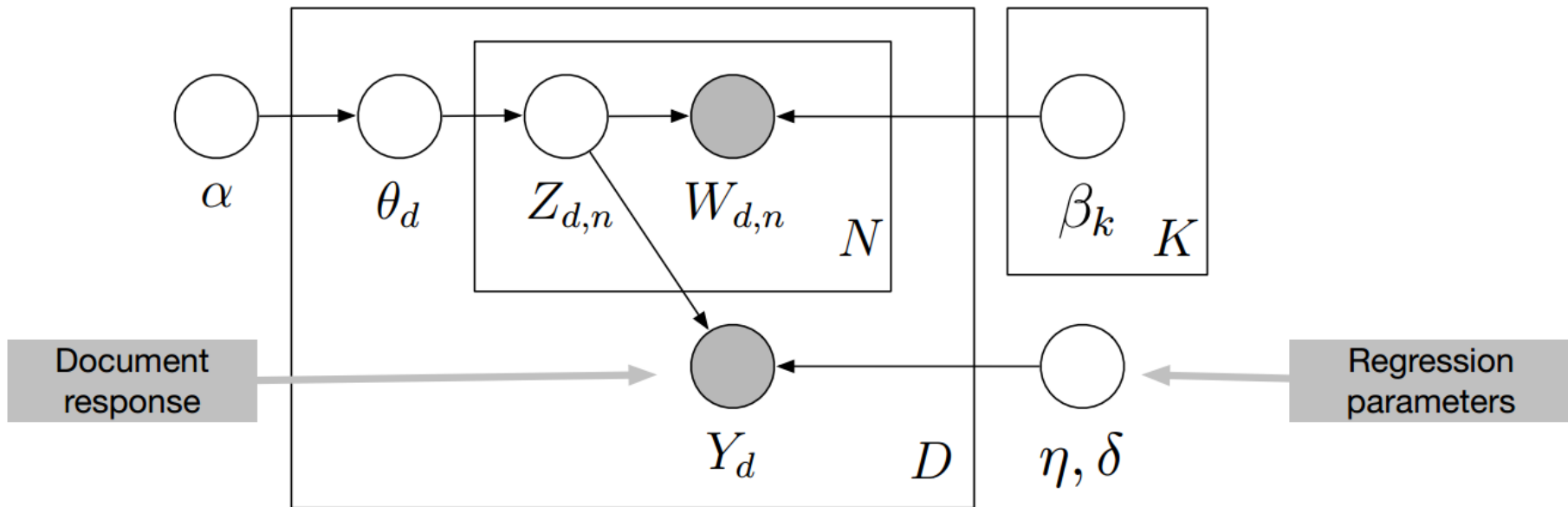
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as 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.



The Author-Topic Model for Authors and Documents
Rosen-Zvi et al.

- Author– Topic model
 - a_d is the group of authors for the d th document
 - $x_{d,n}$ is the author of the n th word of the d th document
 - θ_a is the topic distribution for author a
 - $z_{d,n}$ is the topic for the n th word of the d th document

Extensions of LDA



- Label Y_d for each document represents a value to be predicted from the document
 - E.g., number of stars for each document in a corpus of movie reviews