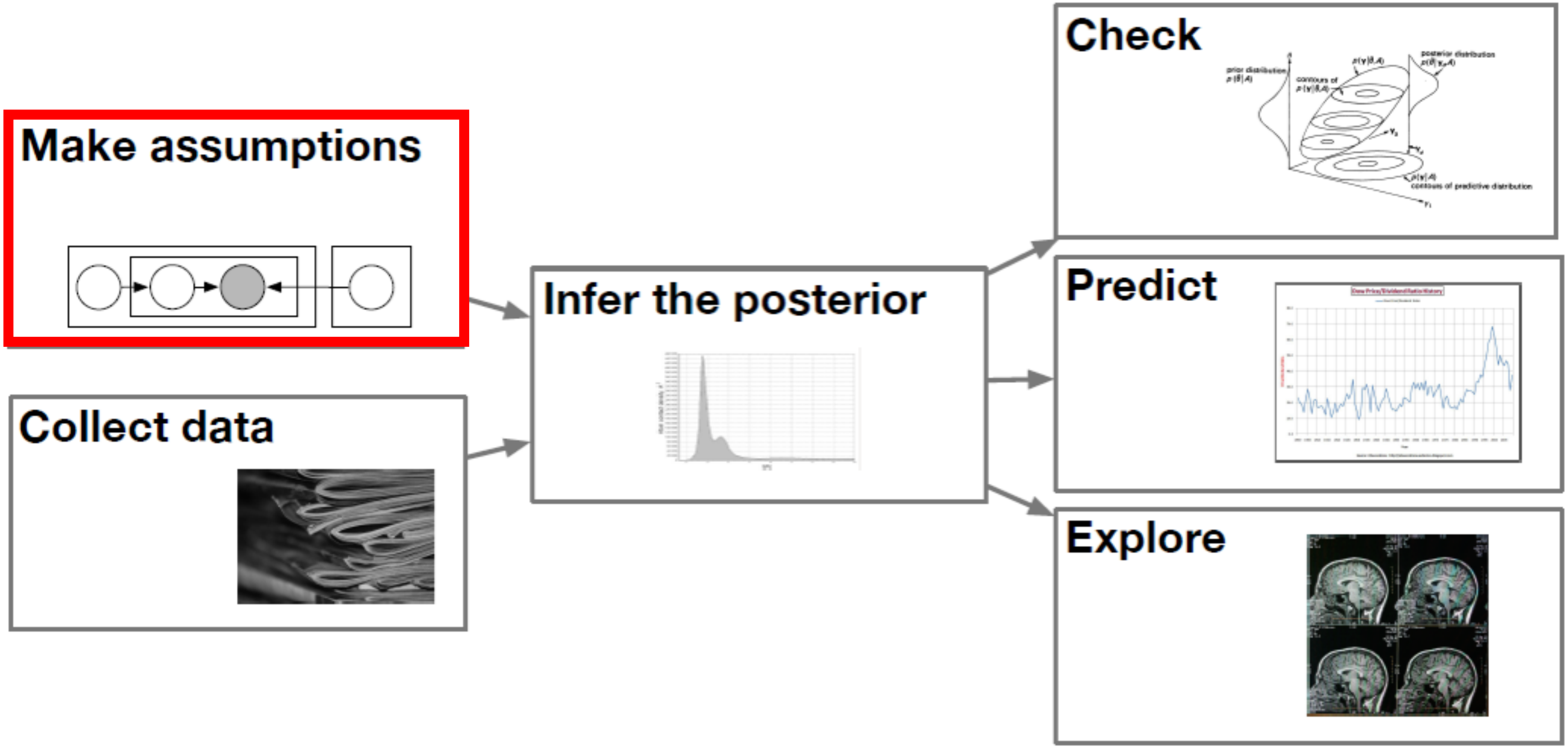
Formalizing the Model



From David Blie (2012) Probabilistic Topic Models, MLSS Slides

Objectives

Intuition

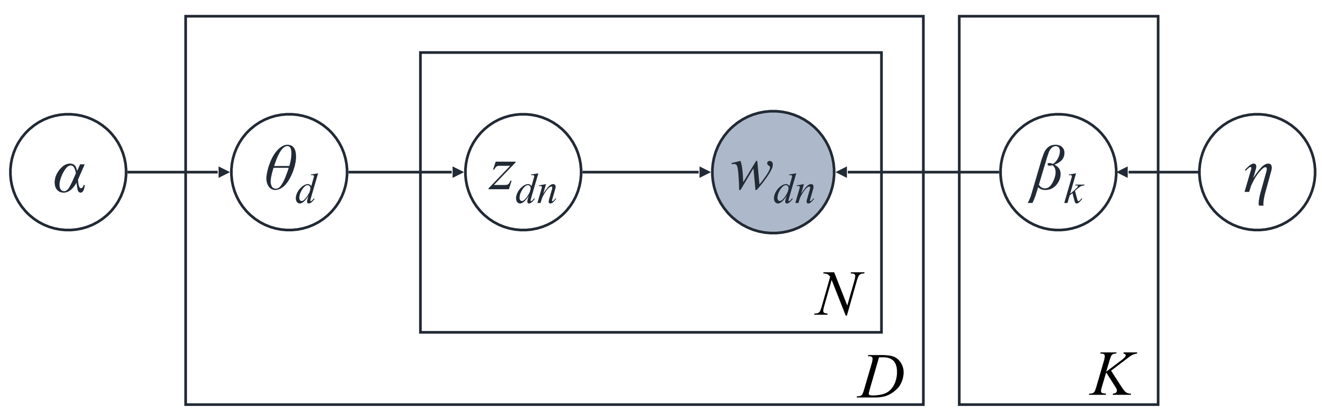
- There are topics
- Words are used differently (i.e., with different frequency) in different topics
- Documents can be about multiple topics

Assumptions

- Documents are bags of words
- Topics are fixed and finite
- Topics are independent
- Documents are independent

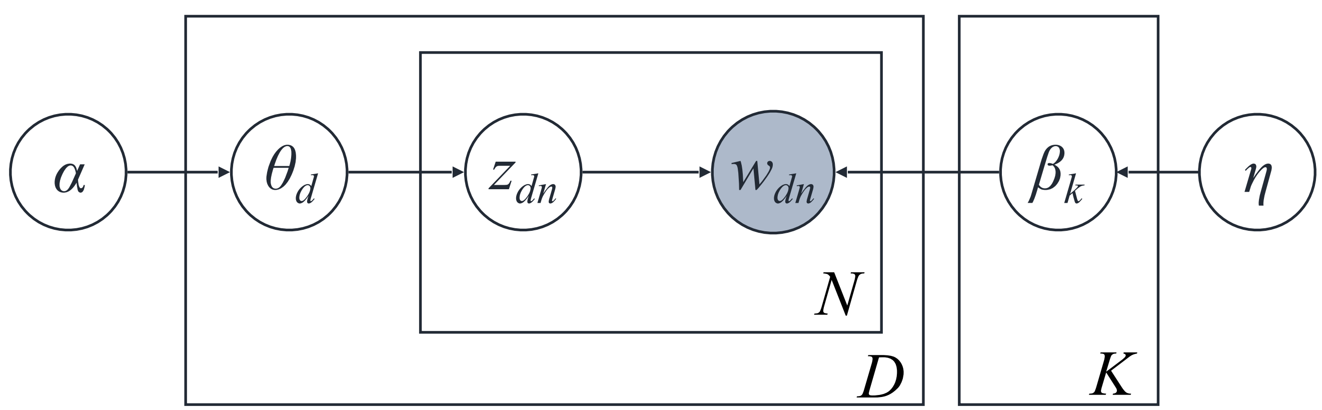
In the next few slides, we will walk through the generative statistical model behind LDA. Note that this is the model that is to be learned; we assume that we have already trained the model using a tool like Mallet.

The main purpose for walking through this detail is both to make the underlying math explicit and to show how that relates to our intuitions.



$$p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p\big(z_{d,n}|\theta_d\big)p\big(w_{d,n}|\beta_{1:K},z_{d,n}\big)\right)\left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)$$

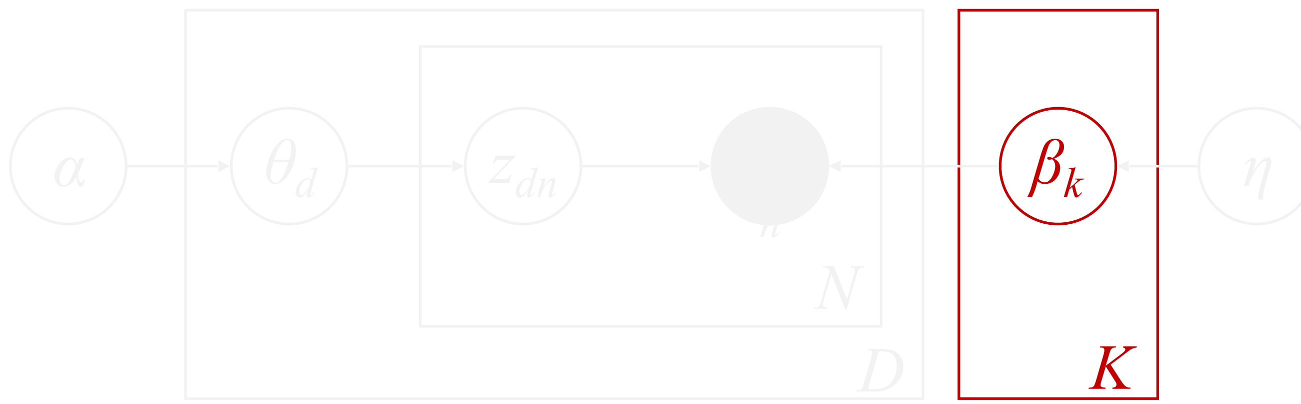
The LDA topic model is expressed as a probability function p. This is the joint probability (or likelihood) of a given set of topics ($\boldsymbol{\beta}$), a set of per-document topic distributions ($\boldsymbol{\theta}$), and the specific association of a topic ($\boldsymbol{z}$) for each word in each document ($\boldsymbol{w}$).



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p\left(z_{d,n} | \theta_d\right) p\left(w_{d,n} | \beta_{1:K}, z_{d,n}\right) \right) \left(\prod_{k=1}^{K} p(\beta_k | \eta) \right)$$

From the outset, we will assume that there is a universe of K topics, $\boldsymbol{\beta}$. We will refer to each individual topic by an index *k* as in β_k.

A topic is formally defined as a probability distribution over a vocabulary *V*. This corresponds to our intuition that certain words are more likely to occur in one rather than another (e.g. 'whale' is more likely to be associated with marine biology than space exploration.)

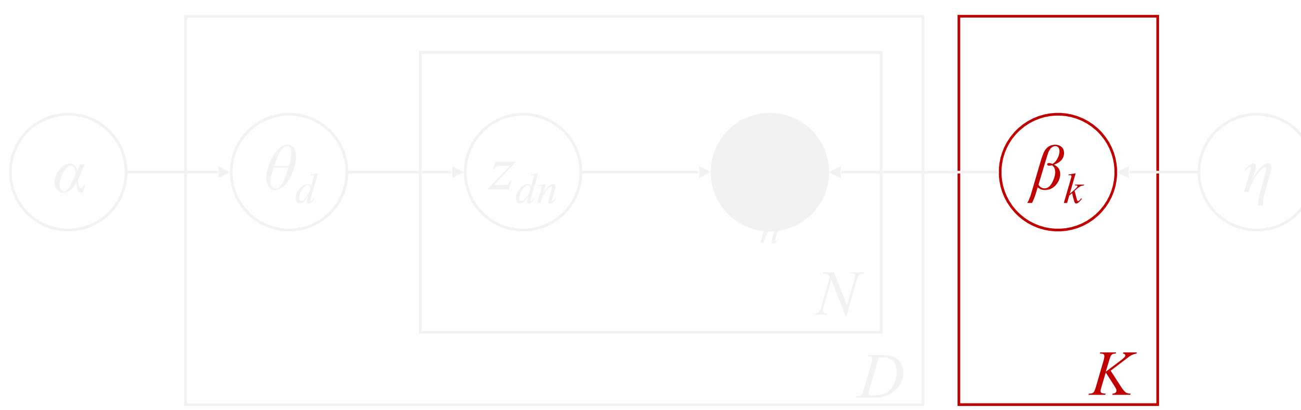


$$p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n})\right)\left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)$$

These aren't really topics. They are PDFs over a vocabulary. We can (and should) argue over what a "topic" really is in our model.

"We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words."

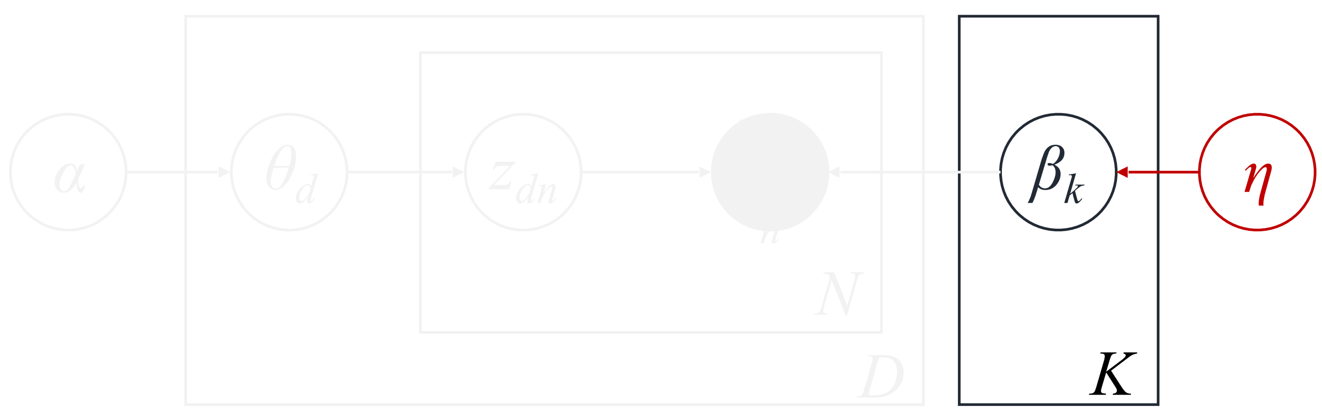
– Blei, 2003



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \left(\prod_{k=1}^{K} p(\beta_k|\eta) \right)$$

Some important statistical housekeeping. We need a way compute the likelihood of any given topic $p(\beta_k)$.

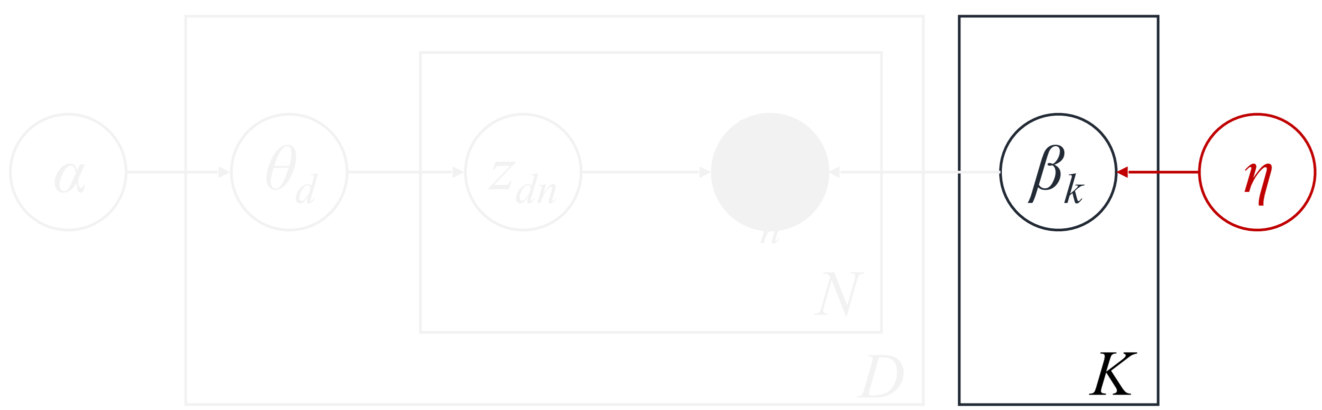
To do this, we will assume that β_k is drawn from a Dirichlet (hence Latent Dirichlet Allocation). A Dirichlet is a probability distribution that generates random probability distributions given (instead of, for example, drawing a single random value by flipping a coin). η is a parameter that governs the dispersion of that Dirichlet.



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \left(\prod_{k=1}^{K} p(\beta_k | \eta) \right)$$

In case you were wondering.

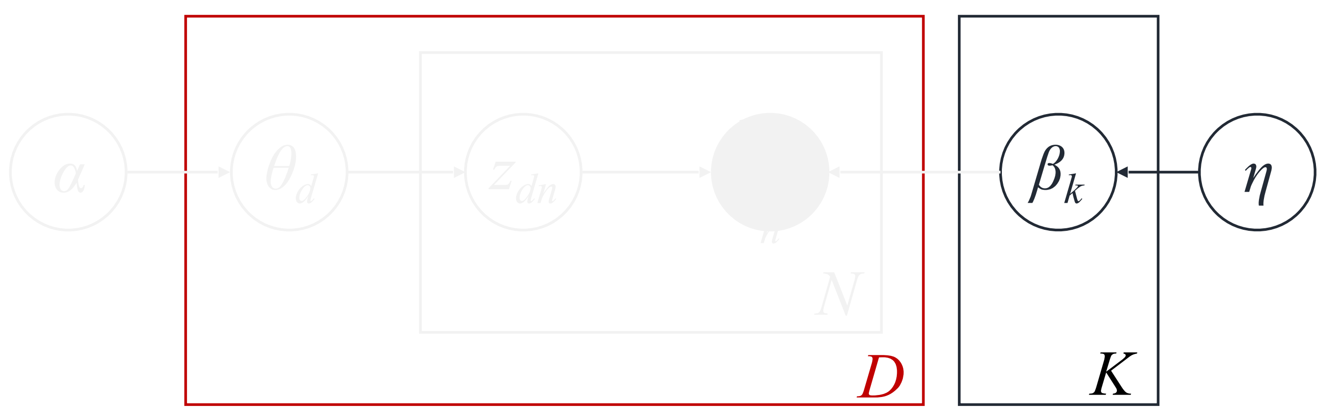
$$p(\beta_k|\eta) = \frac{\Gamma(\sum_{i=1}^{V}\eta_i)}{\prod_{i=1}^{V}\Gamma(\eta_i)}\beta_{k,1}^{\eta_1-1}\dots\beta_{k,V}^{\eta_V-1}$$



$$p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n})\right)\left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)$$

With this universe of topics, we can start generating documents.

For each document d

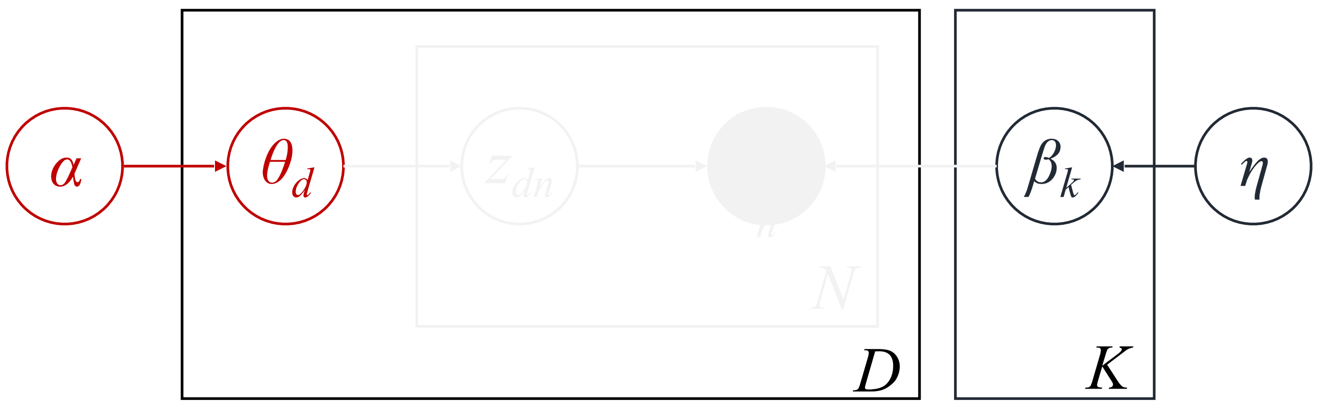


$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \left(\prod_{k=1}^{K} p(\beta_k | \eta) \right)$$

With this universe of topics, we can start generating documents.

For each document d

Step 1: Choose $\theta_d \sim \mathrm{Dir}(\alpha)$



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \left(\prod_{k=1}^{K} p(\beta_k | \eta) \right)$$

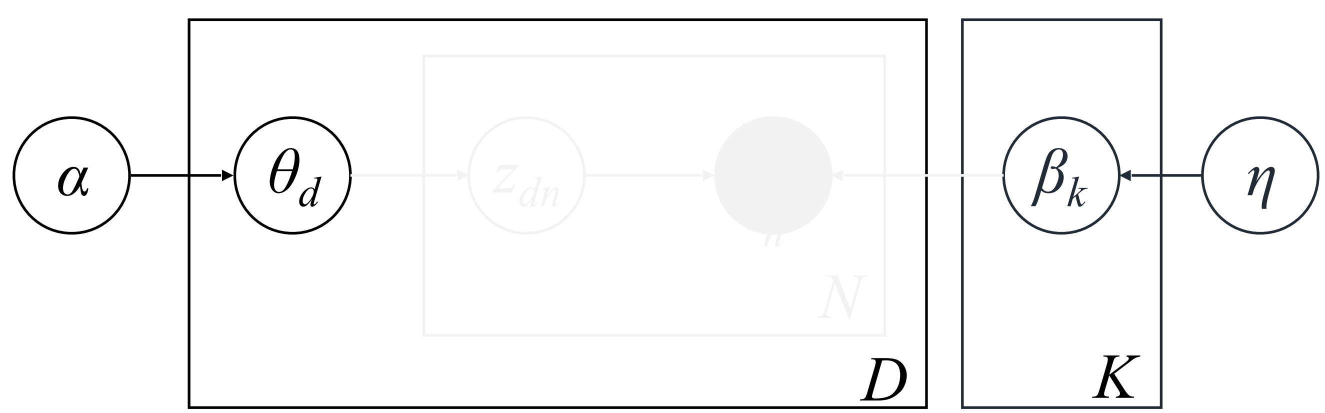
With this universe of topics, we can start generating documents.

For each document d

Step 1: Choose $\theta_d \sim \mathrm{Dir}(\alpha)$

θ_d is a PDF that describes how a document relates to each topic

α is a hyperparameter that controls about how θ_d is distributed



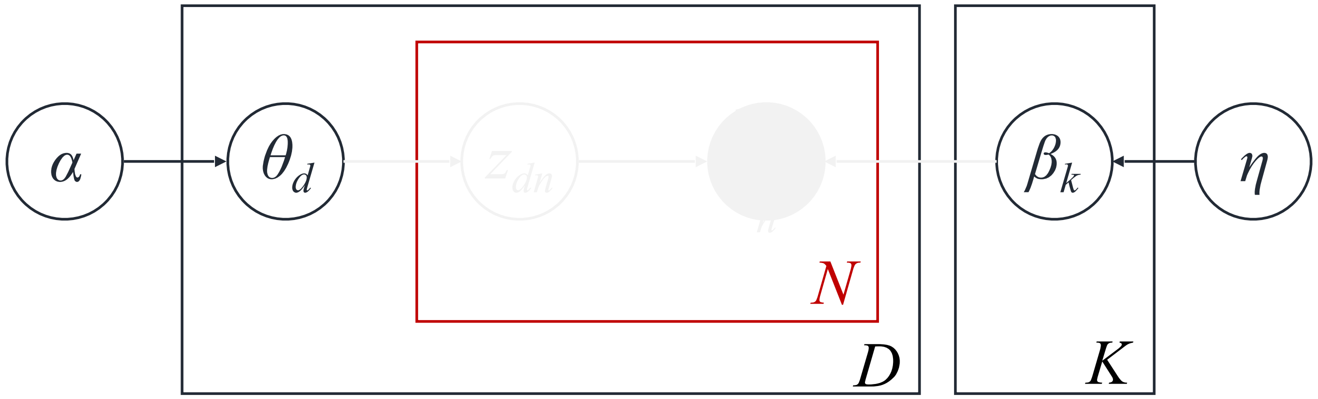
$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p\left(z_{d,n}|\theta_d\right) p\left(w_{d,n}|\beta_{1:K}, z_{d,n}\right) \right) \left(\prod_{k=1}^{K} p(\beta_k|\eta) \right)$$

With this universe of topics, we can start generating documents.

For each document d

Step 1: Choose $\theta_d \sim \mathrm{Dir}(\alpha)$

For each word w_{dn} in document d



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \left(\prod_{k=1}^{K} p(\beta_k | \eta) \right)$$

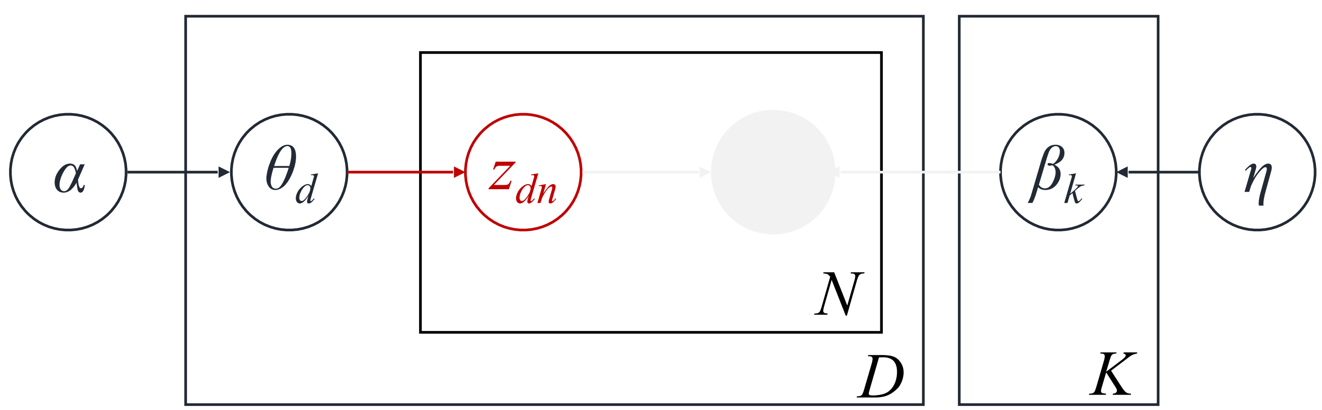
With this universe of topics, we can start generating documents.

For each document d

Step 1: Choose $\theta_d \sim \text{Dir}(\alpha)$

For each word w_{dn} in the document

Step 2: Choose a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})\right)\left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)$$

With this universe of topics, we can start generating documents.

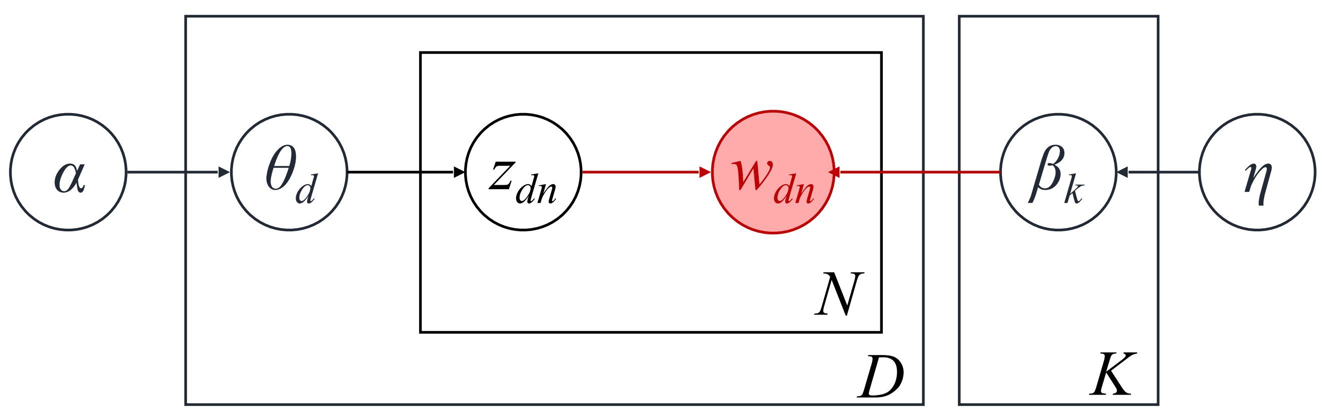
For each document d

Step 1: Choose $\theta_d \sim \mathrm{Dir}(\alpha)$

For each word w_{dn} in the document

Step 2: Choose a topic $z_{dn} \sim \mathrm{Multinomial}(\theta_d)$

Step 3: Choose a word $w_{dn} \sim p(w_{d,n} | \beta_{1:K}, z_{d,n})$



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \left(\prod_{k=1}^{K} p(\beta_k | \eta) \right)$$

Translation

- For each document d
  - **Step 1:** Choose $\theta_d \sim \mathrm{Dir}(\alpha)$
  - For each word w_{dn} in the document
    - **Step 2:** Choose a topic $z_{dn} \sim \mathrm{Multinomial}(\theta_d)$
    - **Step 3:** Choose a word $w_{dn} \sim p(w_{d,n} | \beta_{1:K}, z_{d,n})$

An author starts writing a document

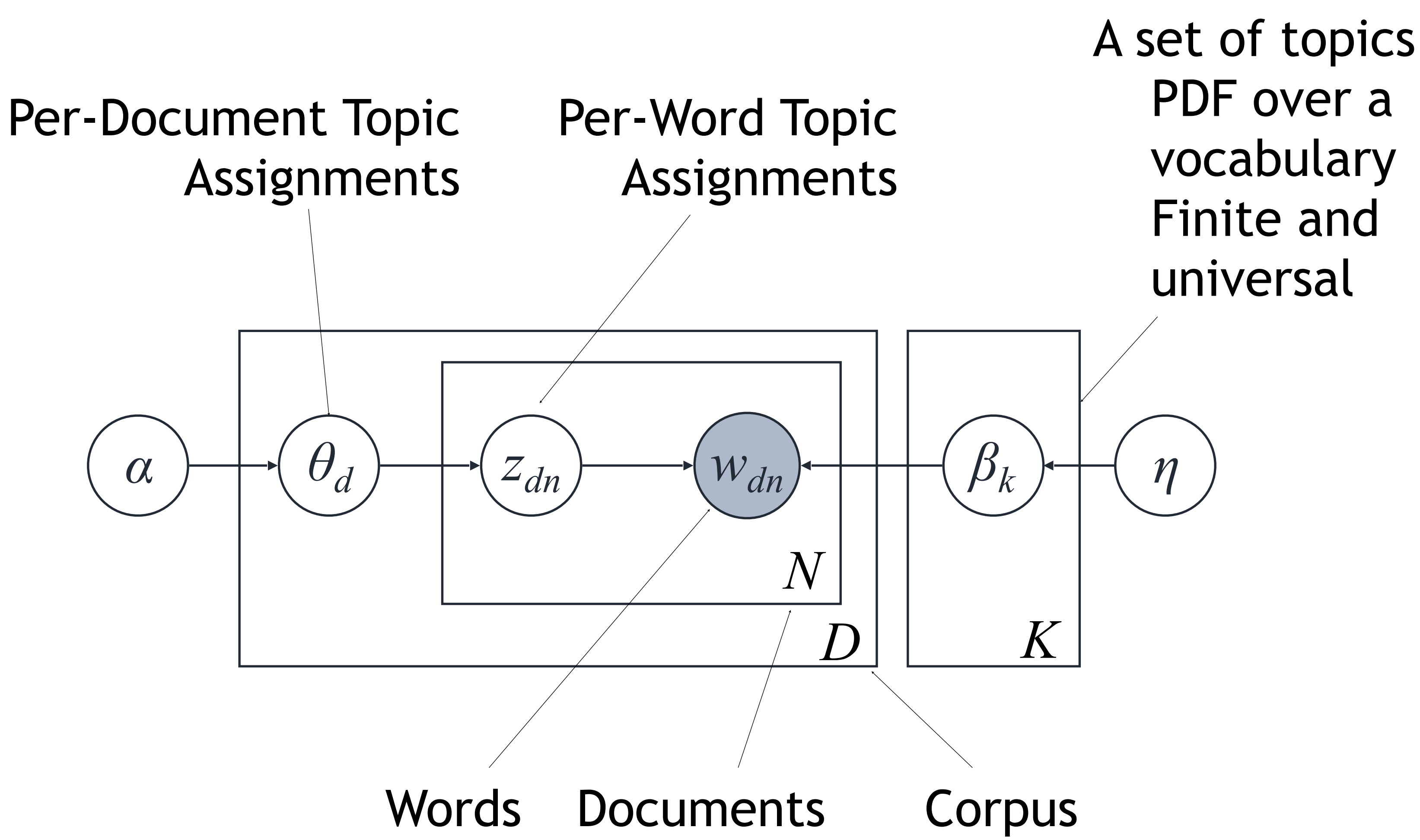
Step 1: She picks something to write about

She writes words on the page by

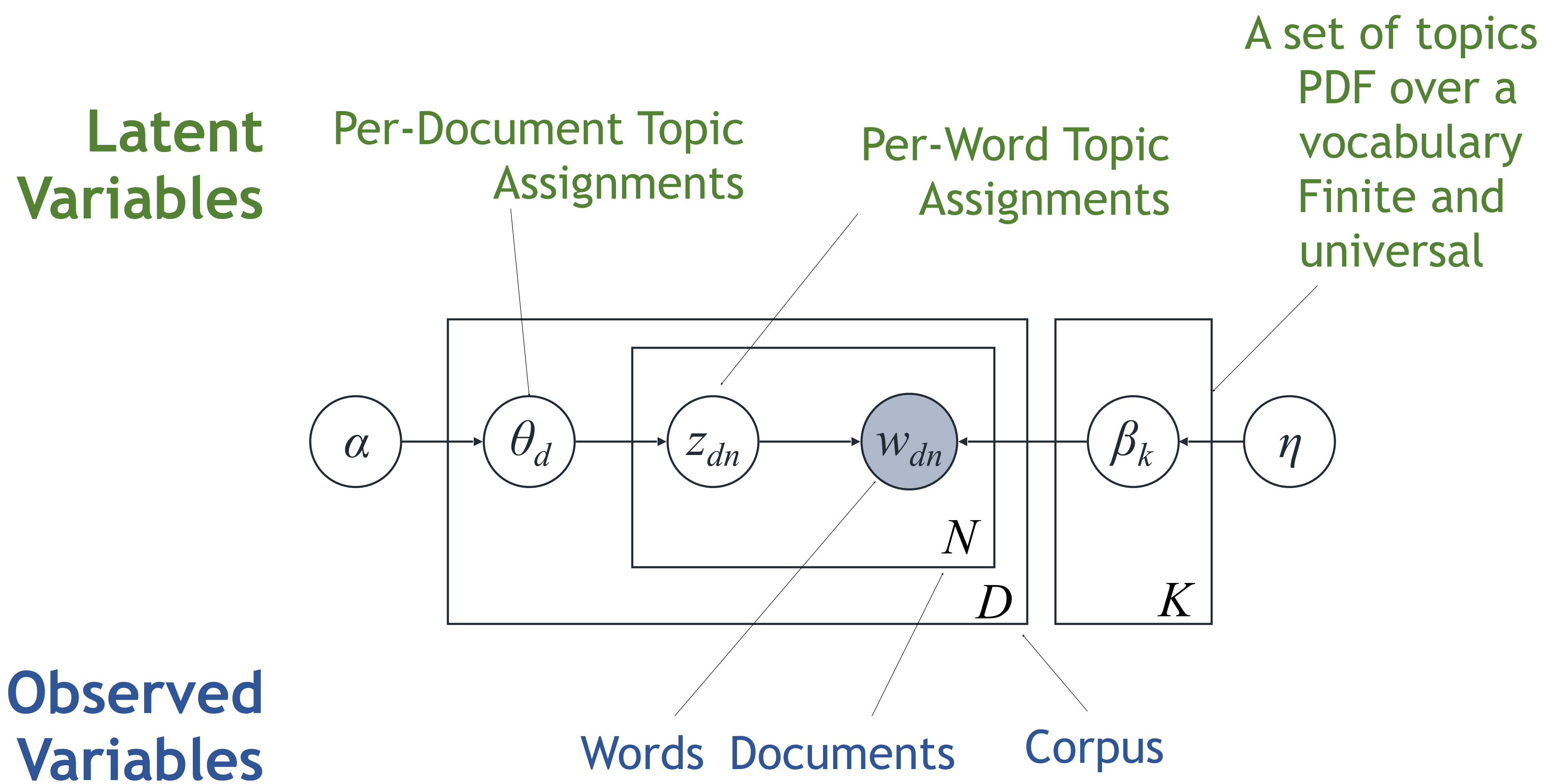
Step 2: Choosing the topic for this word

Step 3: Picking a word related to that topic

Our statistical model consists of:



The only data we actually have are the individual words and the documents they appear in. From that rather minimal starting point, we have to infer all the other parameters in the model.



We started by saying this modeled our assumptions about documents.

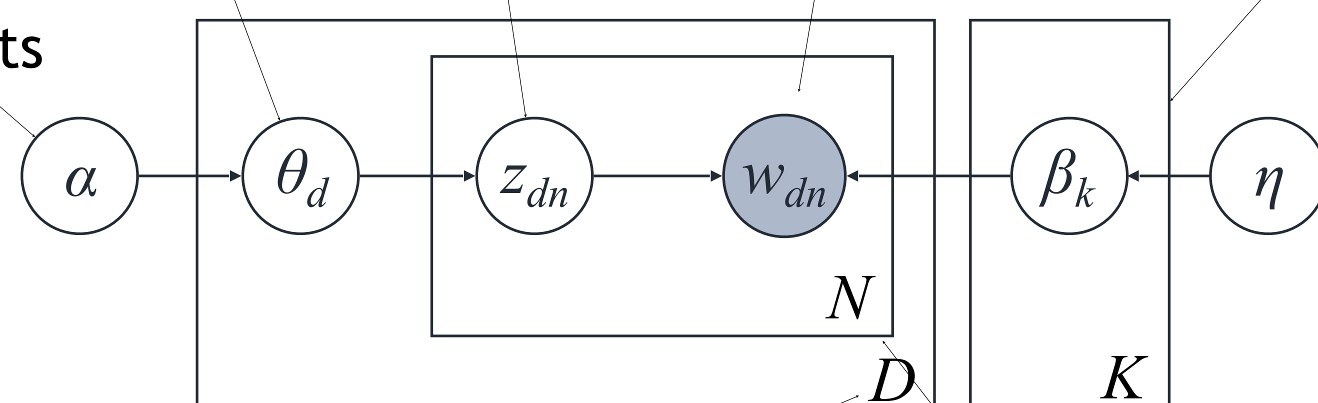
Documents can be about more than one topic, in different proportions

Each word relates to exactly one topic

Words are discrete units of information (stemming, stopping)

There are a fixed number of topics.
Topics are defined by word use (i.e., a given word is more likely to be associated with one topic than another)
Topics are fixed, finite and universal

How focused are the documents



How many words will have high-probability

The only relevant fact about documents is the words they contain. (ignore author, time period, journal, quality, etc.)

Documents are bags of words.