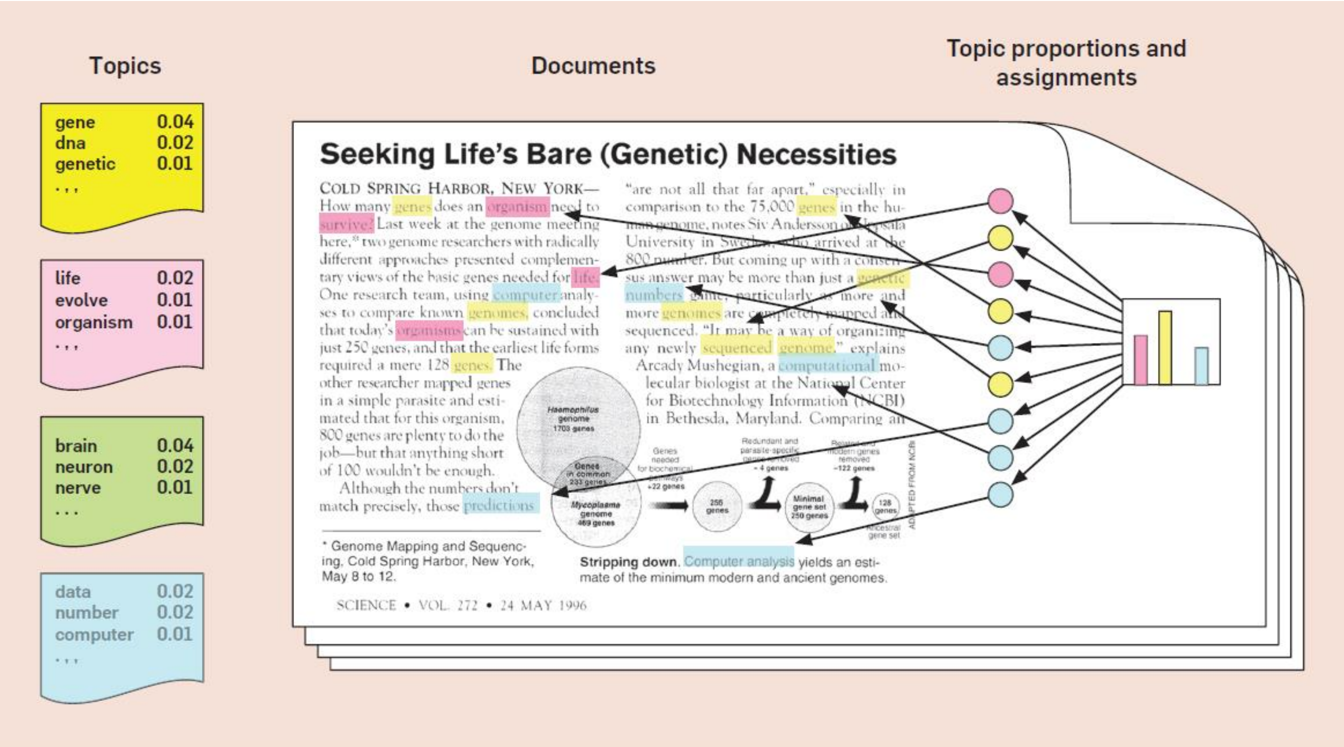


Lecture 7: Topic Modeling

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

LDA: Intuition

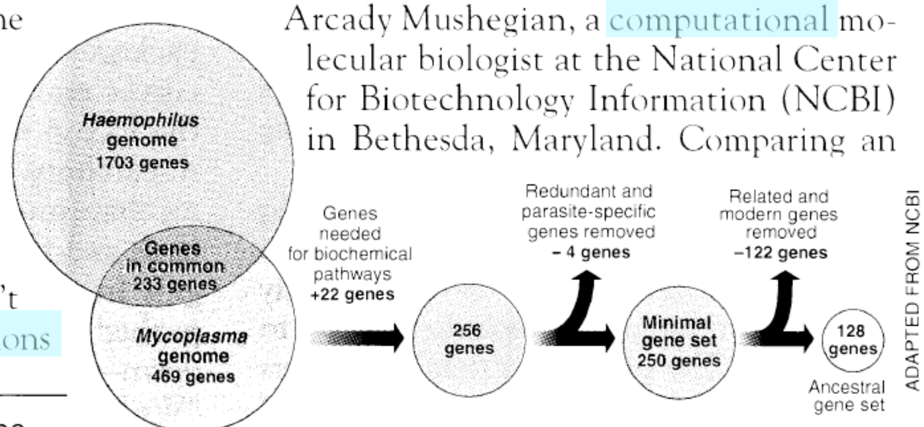
Blei (2012)

- Documents exhibit multiple topics

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



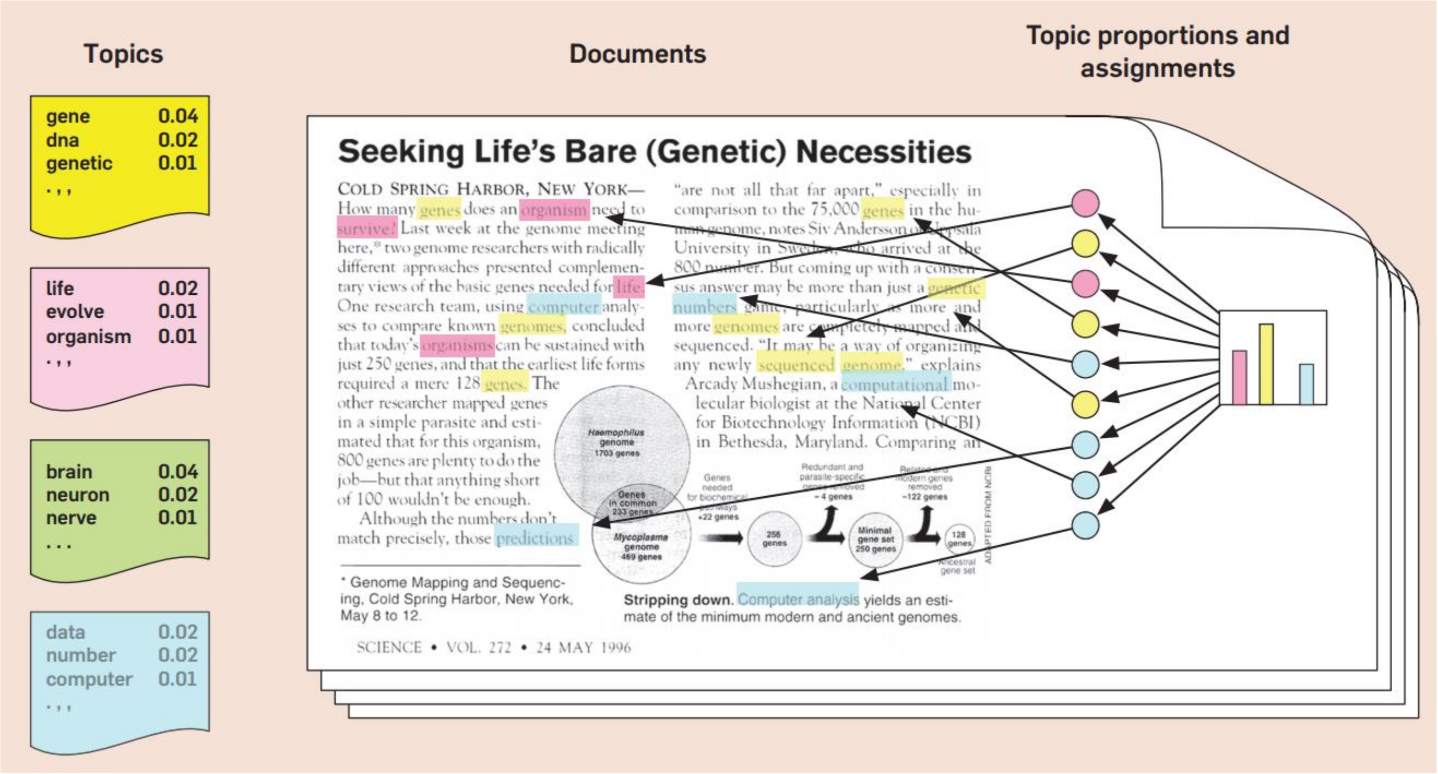
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

LDA: Intuition

Blei (2012)

- Documents exhibit multiple topics



✓ Each topic is a distribution over words

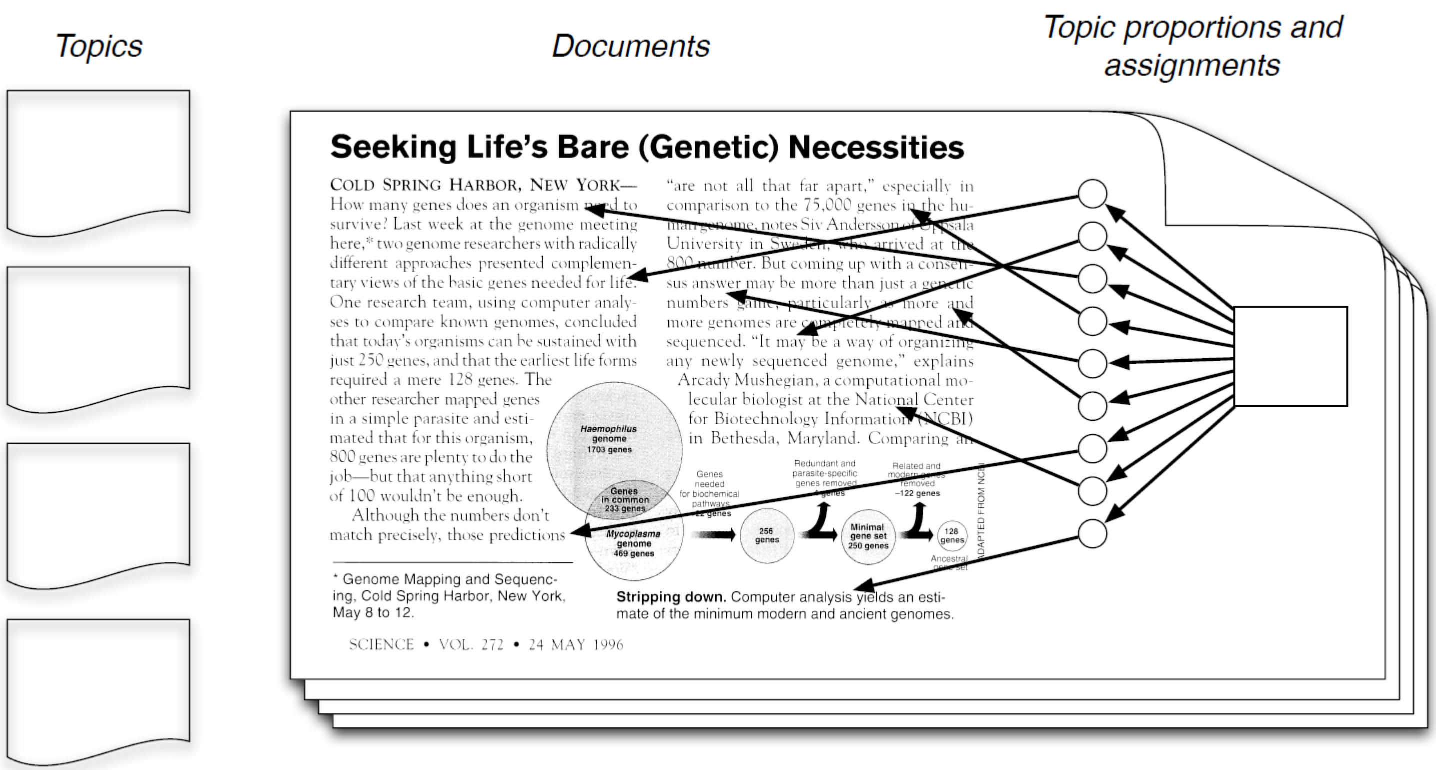
✓ Each document is a mixture of corpus-wide topics

✓ Each word is drawn from one of those topics

LDA: Intuition

Blei (2012)

- Documents exhibit multiple topics



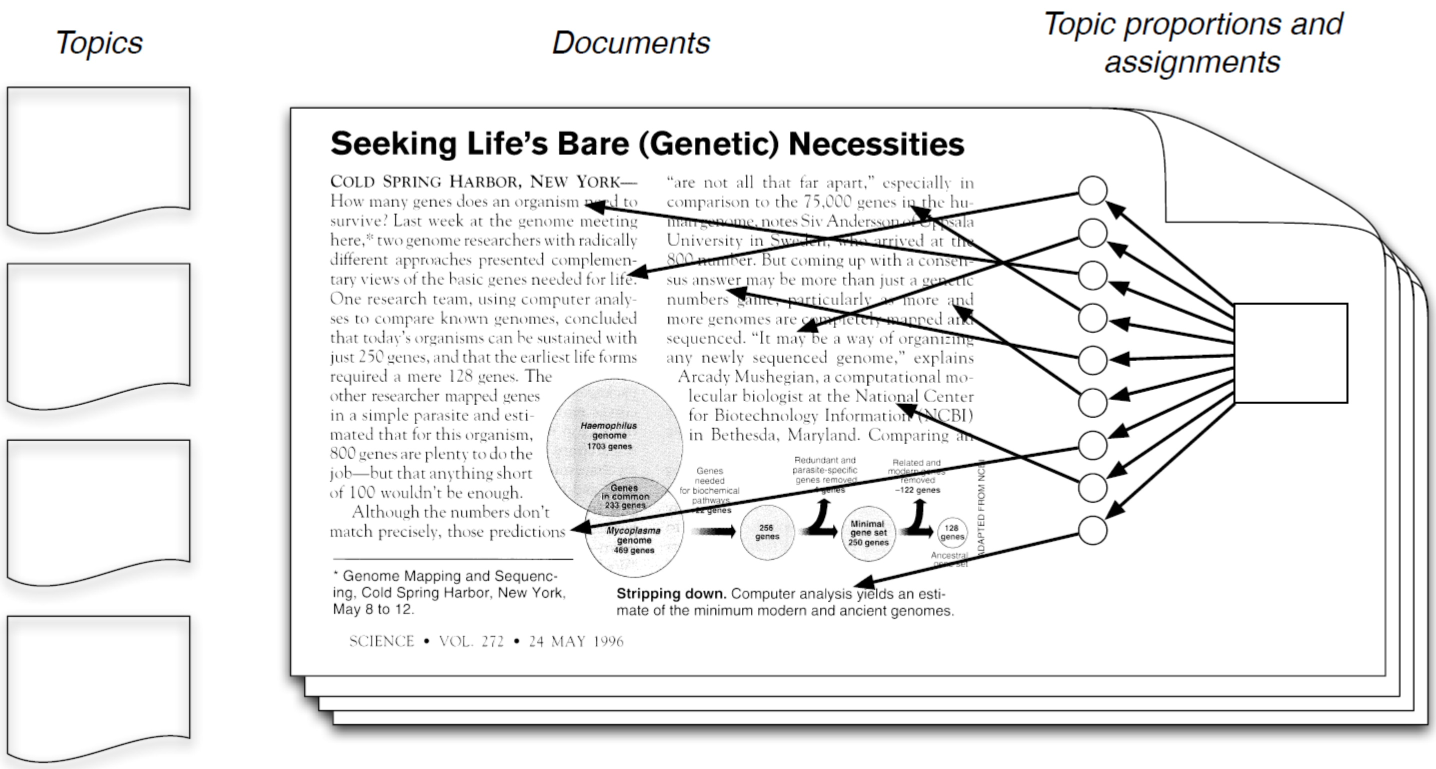
✓ In reality, we only observe the documents

✓ The other structure are hidden variables

LDA: Intuition

Blei (2012)

- Documents exhibit multiple topics



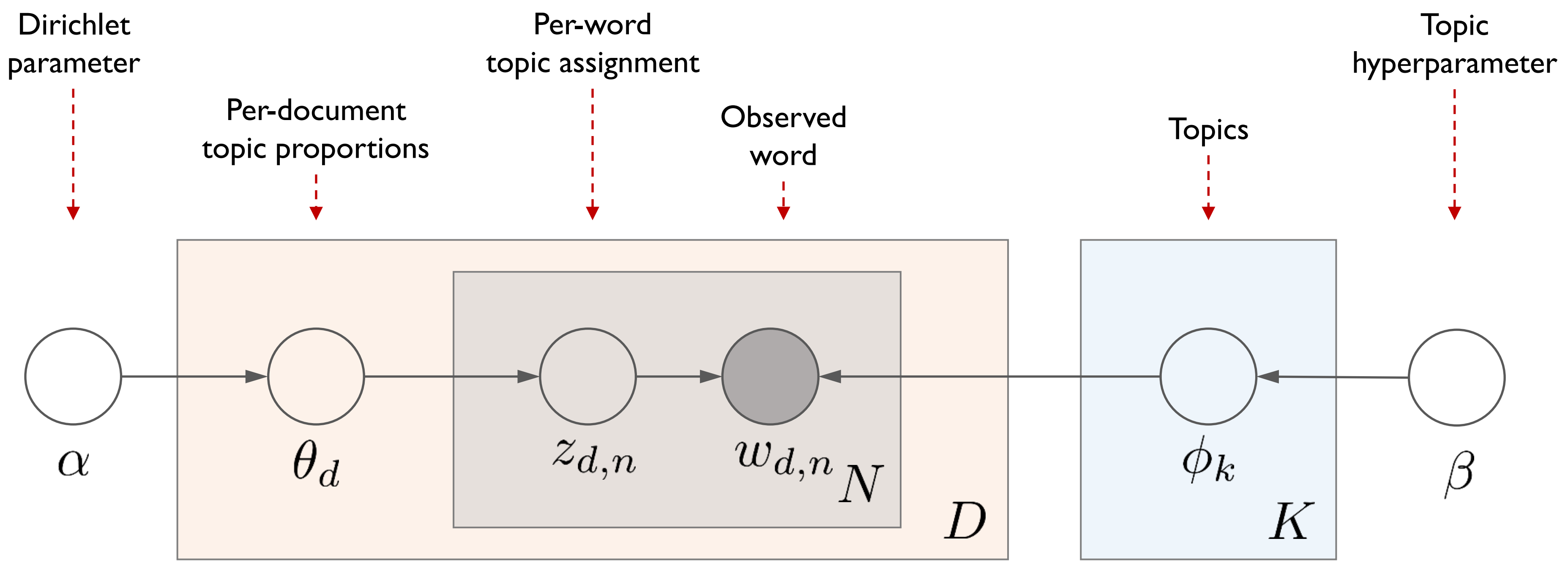
✓ The goal of LDA is to infer the hidden variables

✓ i.e. compute their distribution conditioned on the document

➔ p(topics, proportions, assignments | documents)

LDA Overview

- Documents exhibit multiple topics



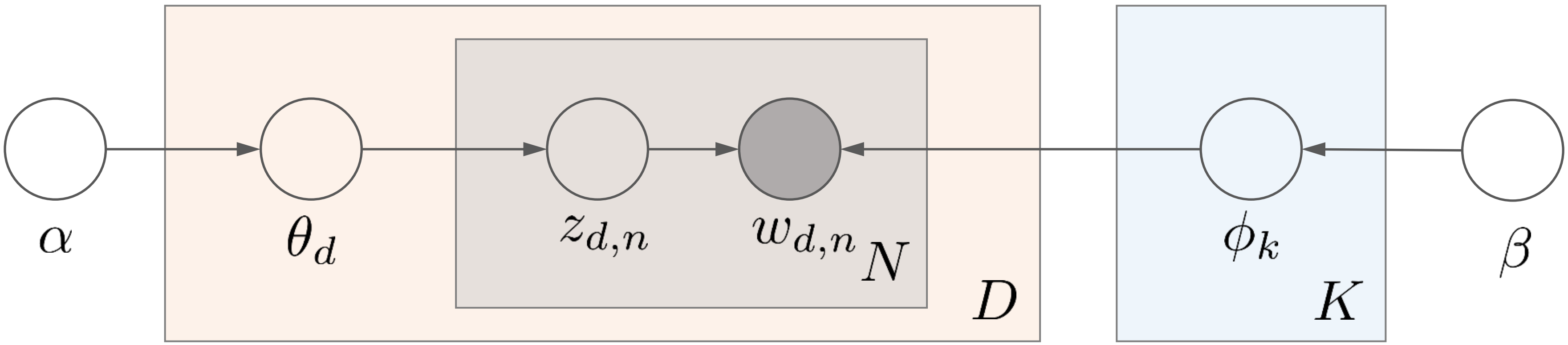
✓ Encode **assumptions**

✓ Define a **factorization** of the joint distribution

✓ Connect to **algorithms** for computing with data

LDA Overview

- LDA structure



✓ Nodes are random variables while edges indicate dependence

✓ Shaded nodes are observed

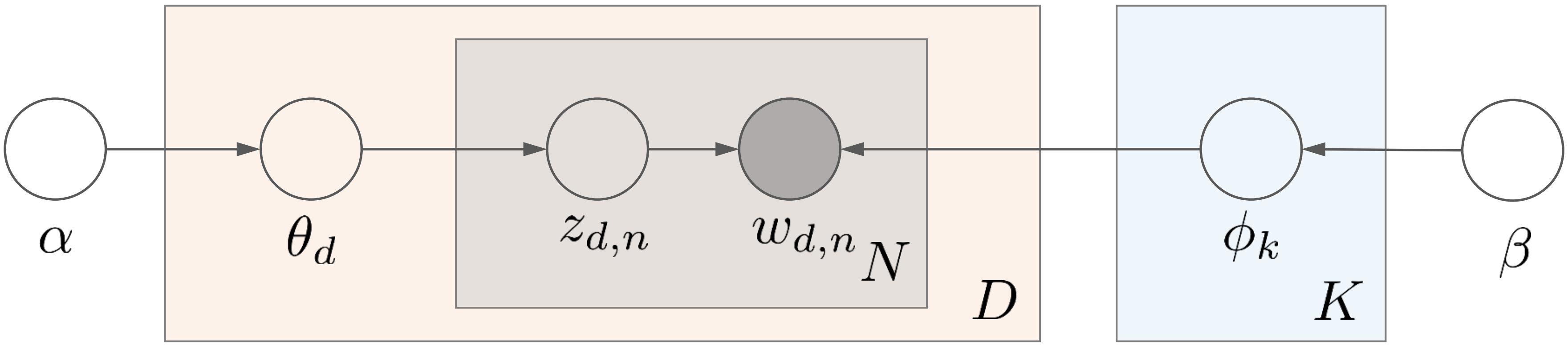
✓ Plates indicate replicated variables

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

$$\prod_{i=1}^{K} p(\phi_i|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\phi_{1:K}, z_{d,n}) \right)$$

LDA: Document generation process

- Document generation process



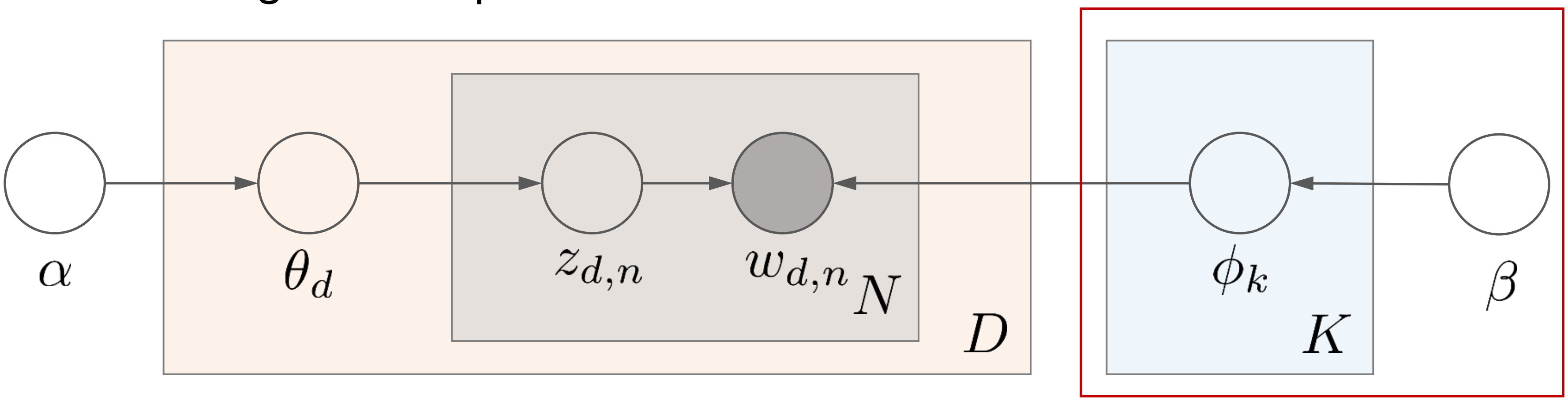
✓ Draw each topic $\phi_k \sim Dir(\beta)$ for $i \in \{1, ..., K\}$

✓ For each document

  - Draw topic proportions $\theta_d \sim Dir(\alpha)$
  - For each word
    - Draw $z_{d,n} \sim Multi(\theta_d)$
    - Draw $w_{d,n} \sim Multi(\phi_{z_{d,n},n})$

LDA: Document generation process

- Document generation process



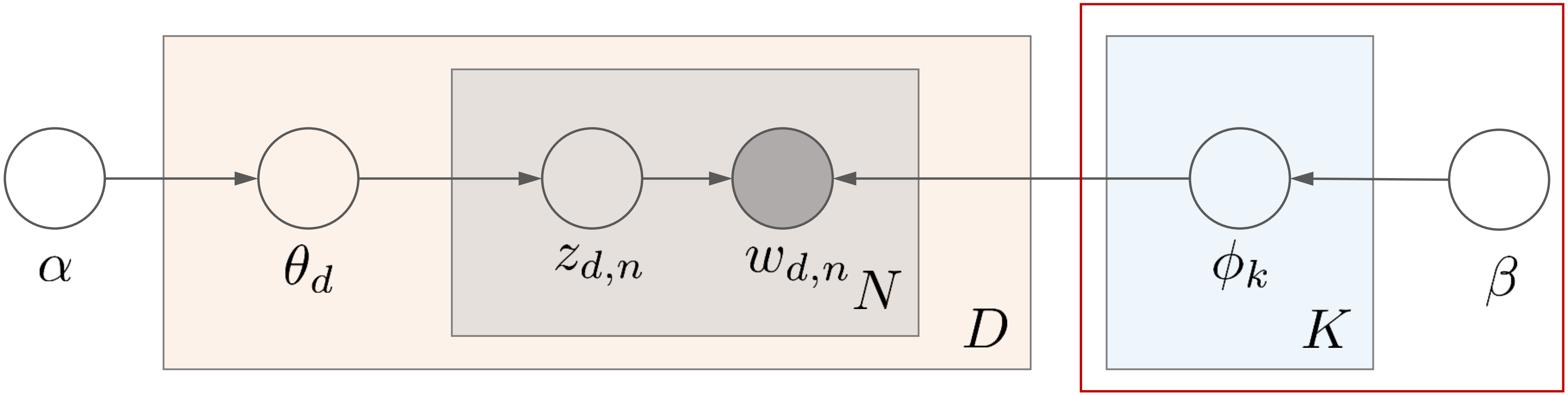
✓ Term distribution per topic

- Drawn from the Dirichlet distribution, given the Dirichlet parameter β, which is a V-vector with component $\beta_v > 0$

$$p(\phi|\beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}-1}$$

LDA: Document generation process

- Document generation process

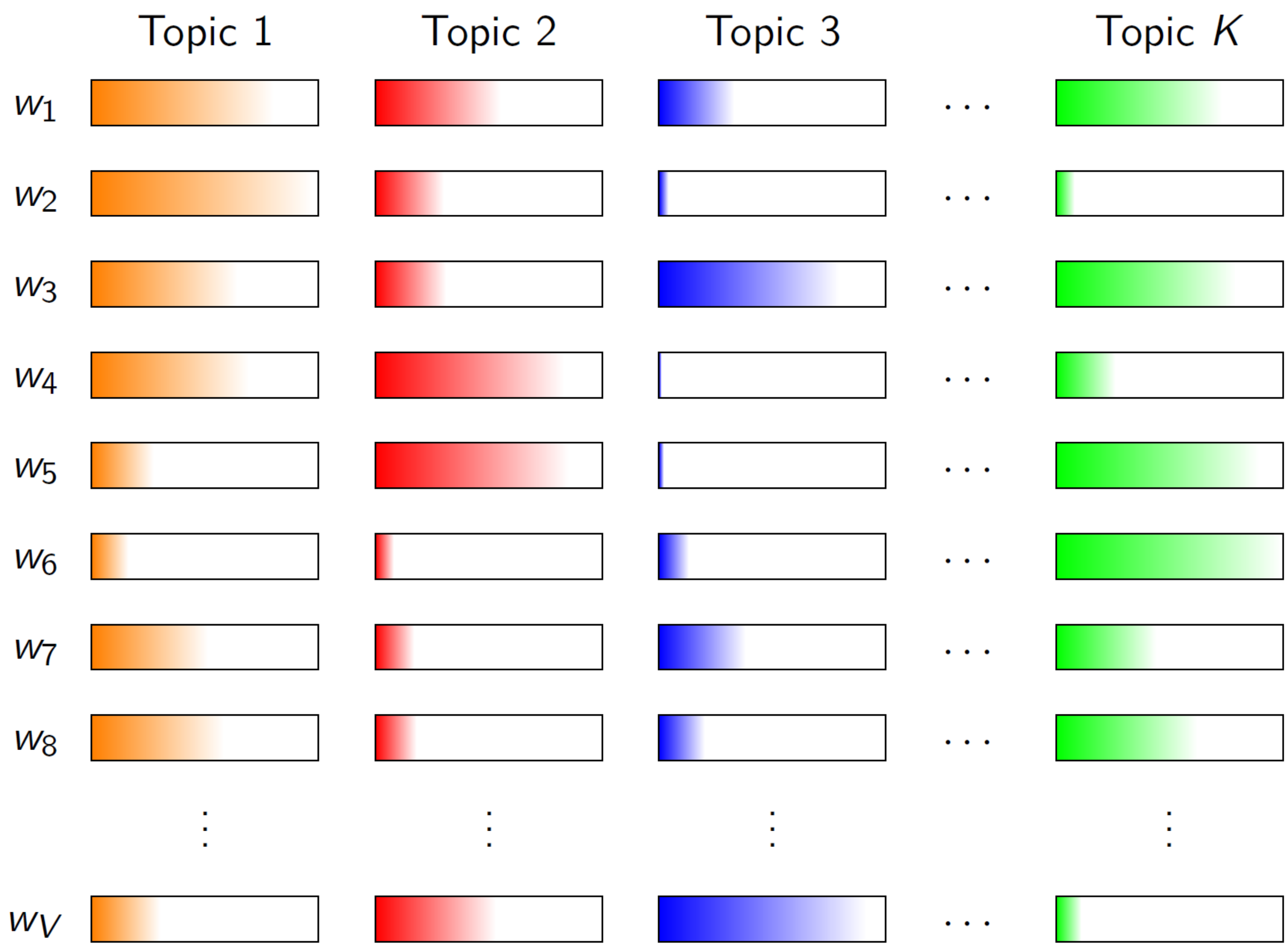


✓ Term distribution per topic

		Term 1	Term 2	Term 3	Term 4	Term 5=V
Topic 1	$\phi_{k=1}$	0.1	0.1	0	0.7	0.1
Topic 2	$\phi_{k=2}$	0.2	0.1	0.2	0.2	0.3
Topic 3	$\phi_{k=3}$	0.01	0.2	0.39	0.3	0.1
Topic 4	$\phi_{k=4=K}$	0.0	0.0	0.5	0.3	0.2

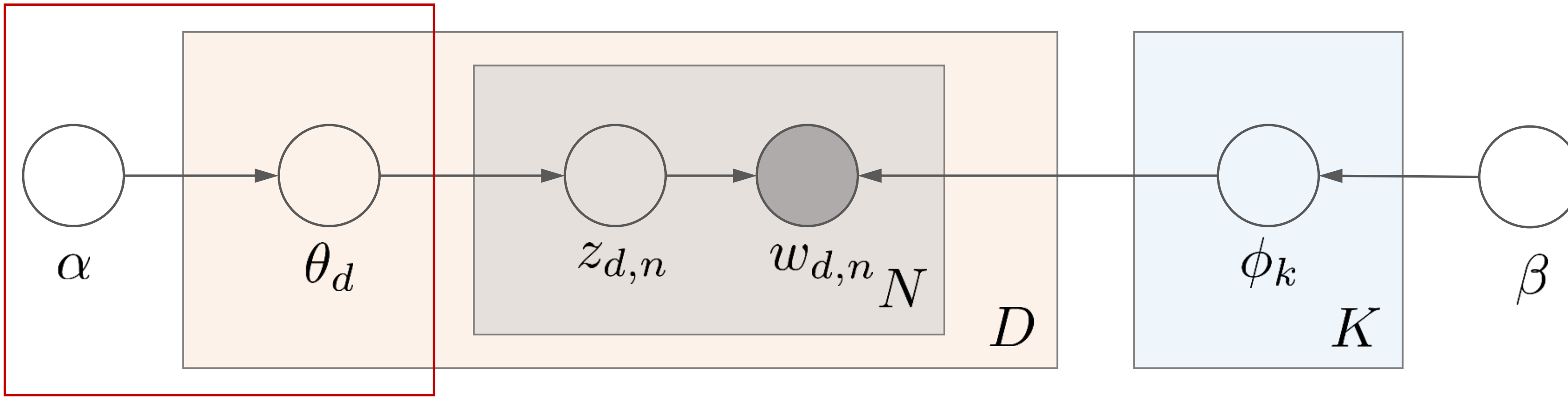
LDA: Document generation process

- Document generation process
  - ✓ Term distribution per topic



LDA: Document generation process

- Document generation process



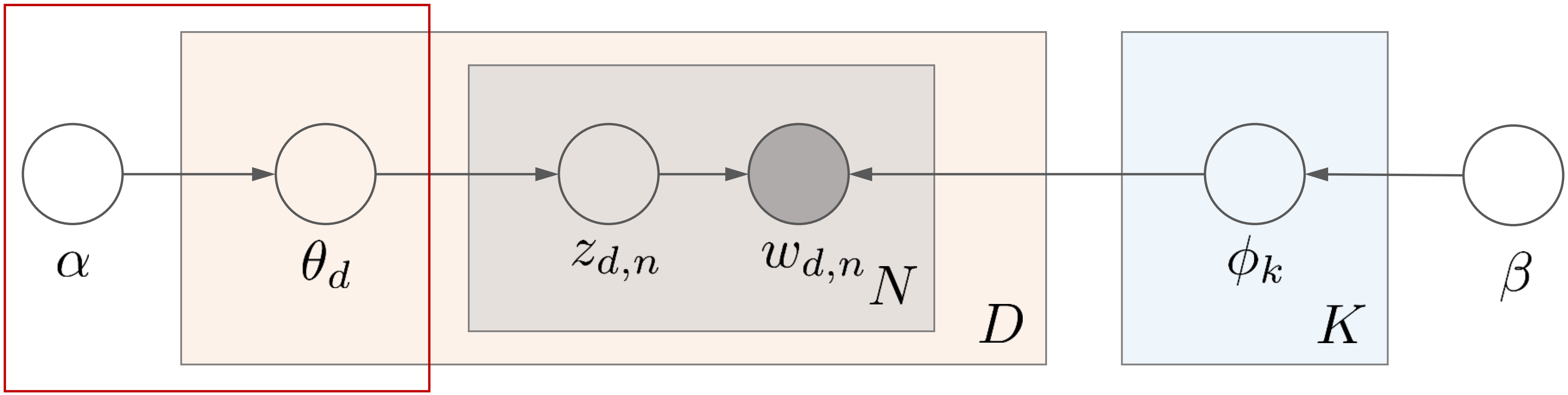
✓ Topic distribution per document

- Drawn from the Dirichlet distribution, given the Dirichlet parameter α, which is a K-vector with components $\alpha_k > 0$

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_k - 1} \cdots \theta_K^{\alpha_K - 1} = \frac{\Gamma(\alpha_.)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

LDA: Document generation process

- Document generation process

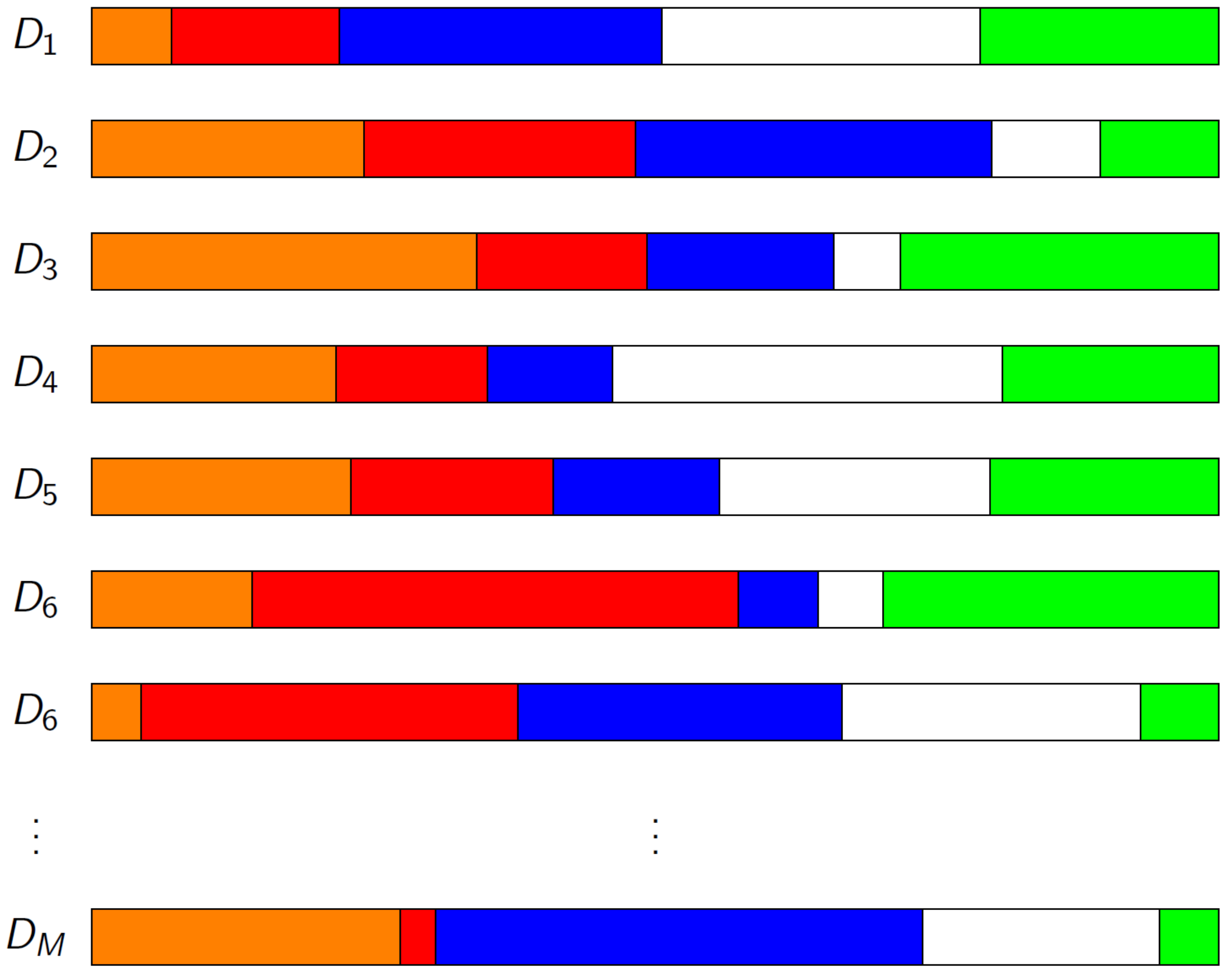


✓ Topic distribution per document

		Topic 1	Topic 2	Topic 3	Topic 4
Document 1	$\theta_{d=1}$	0.5	0.1	0.3	0.1
Document 2	$\theta_{d=2}$	0.0	0.9	0.1	0.0
Document 3	$\theta_{d=3=D}$	0.02	0.48	0.25	0.25

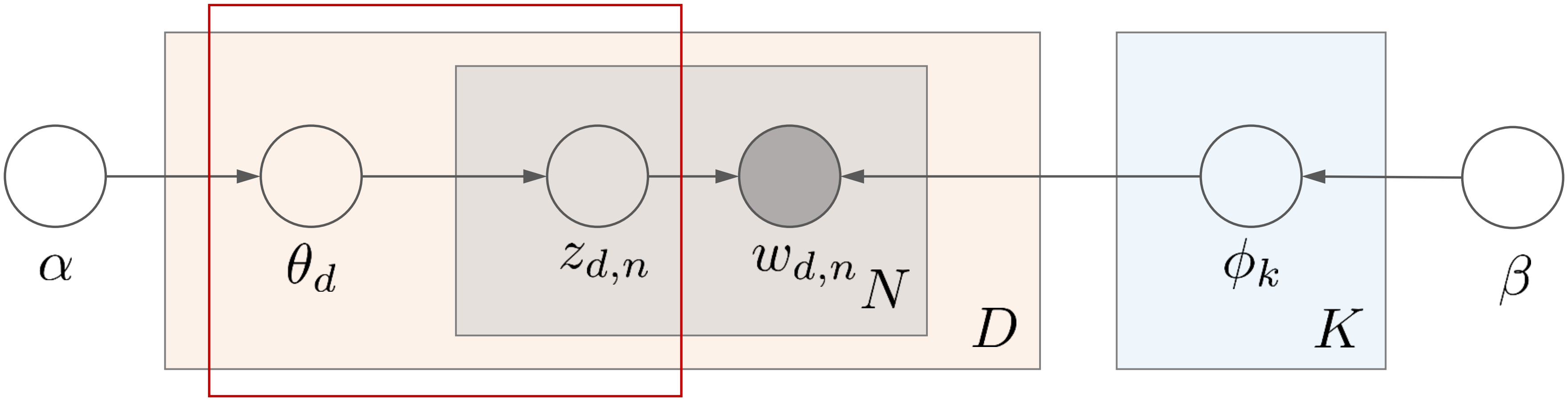
LDA: Document generation process

- Document generation process
  - ✓ Topic distribution per document



LDA: Document generation process

- Document generation process



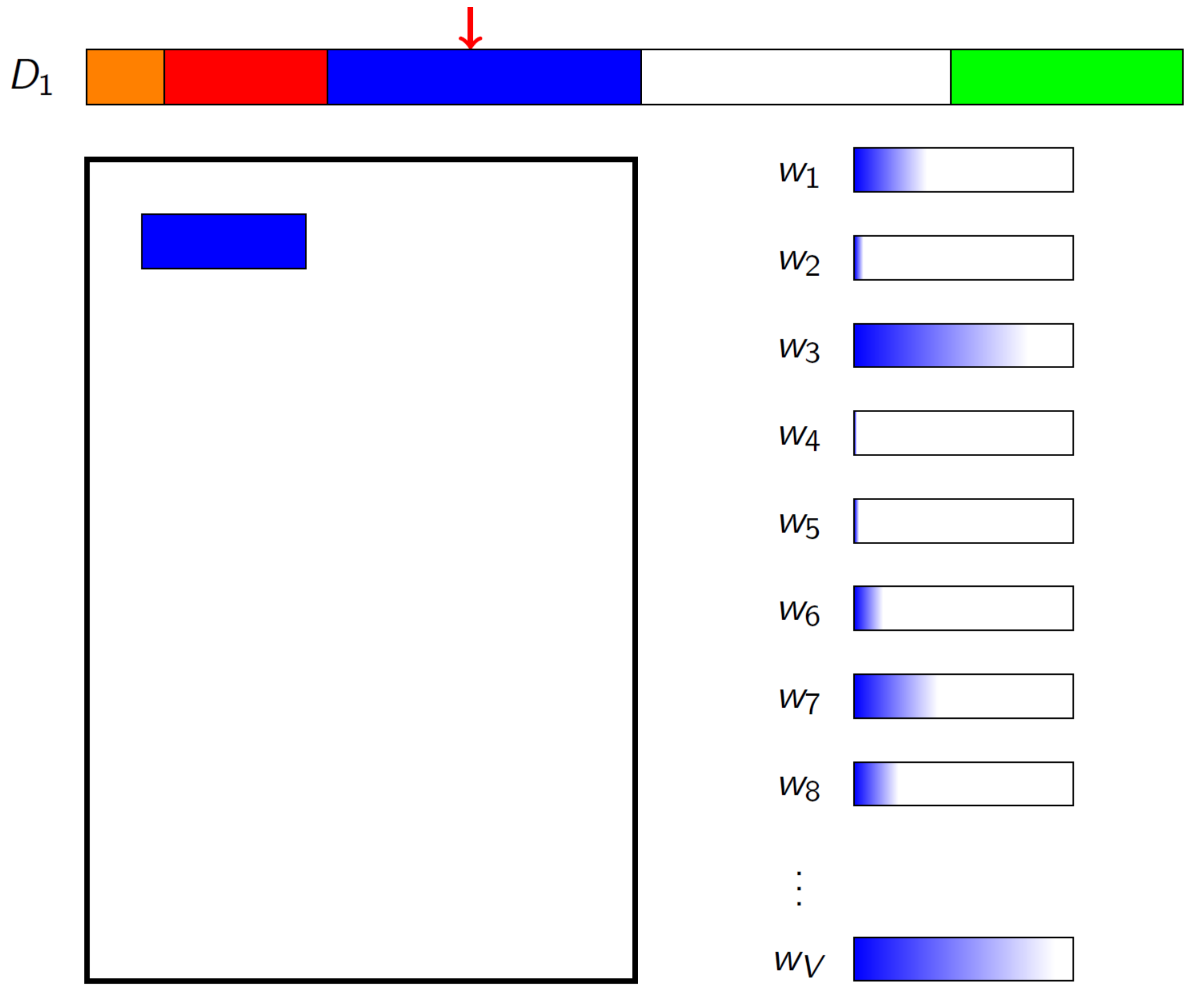
✓ Topic to words assignments

		Word w_1	Word w_2	Word w_3	Word w_4	Word w_5	Word w_6
Document 1	$z_{d=1}$	Topic k=2	Topic k=1	Topic k=1	Topic k=4	Topic k=3	Topic k=3
Document 2	$z_{d=2}$	Topic k=2	Topic k=3	Topic k=2	Topic k=2		
Document 3	$z_{d=3=D}$	Topic k=4	Topic k=2	Topic k=2	Topic k=4	Topic k=3	

$$p(z|\theta) = \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{n_{d,k,\cdot}}$$

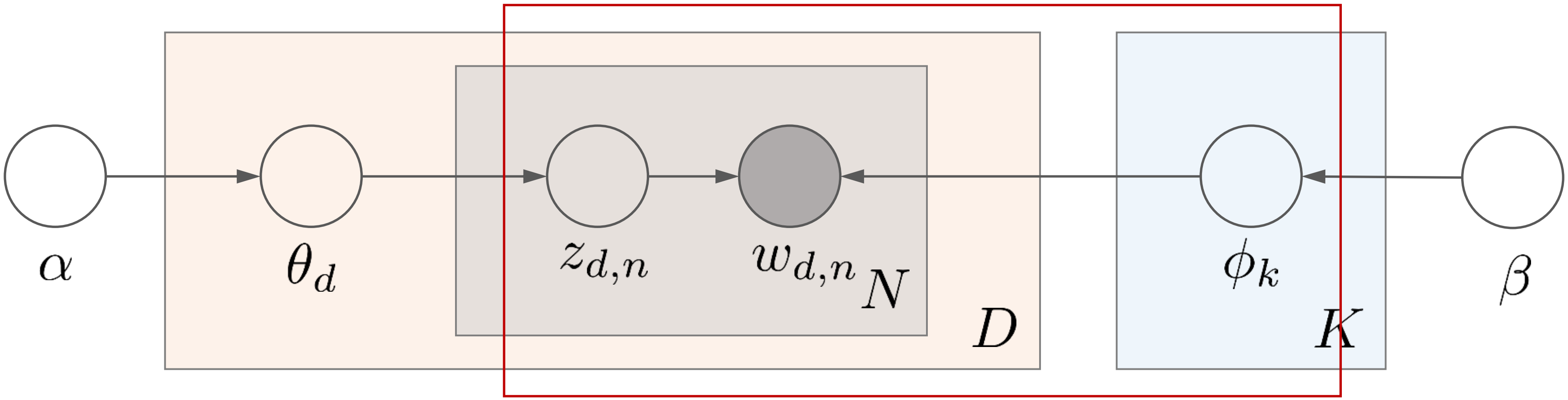
LDA: Document generation process

- Document generation process
  - ✓ Topic to words assignments



LDA: Document generation process

- Document generation process

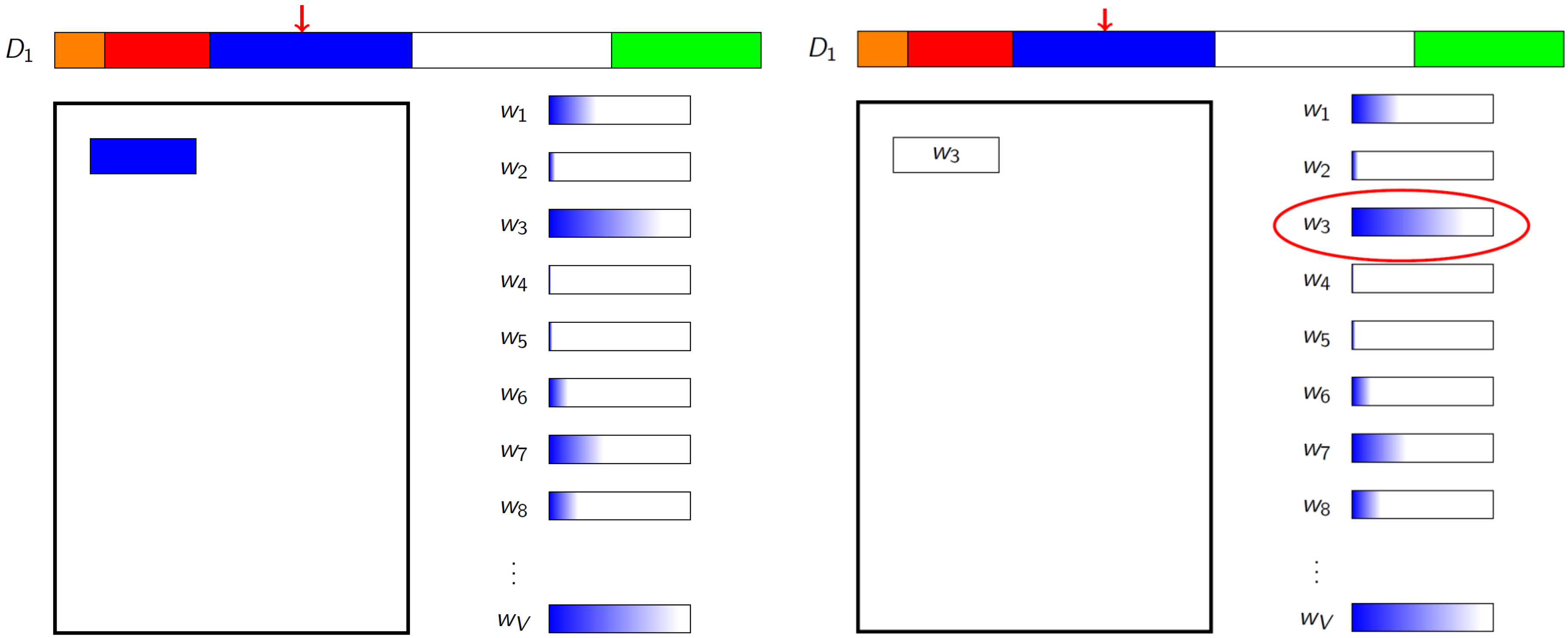


✓ Probability of a corpus

$$p(w|z,\phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{\cdot,k,v}}$$

LDA: Document generation process

- Document generation process
  - ✓ Word selection



LDA: Document generation process

- Document generation process
  - ✓ Word selection


ANY
questions?