Latent Dirichlet Allocation

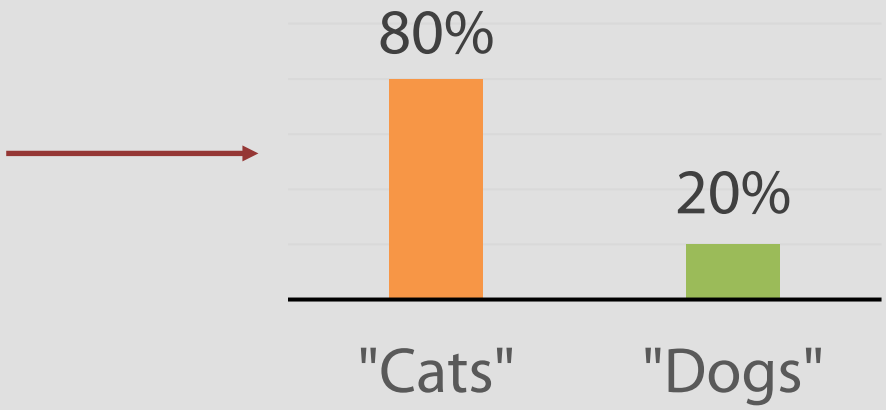
Topics

Document is a distribution over topics

Topics

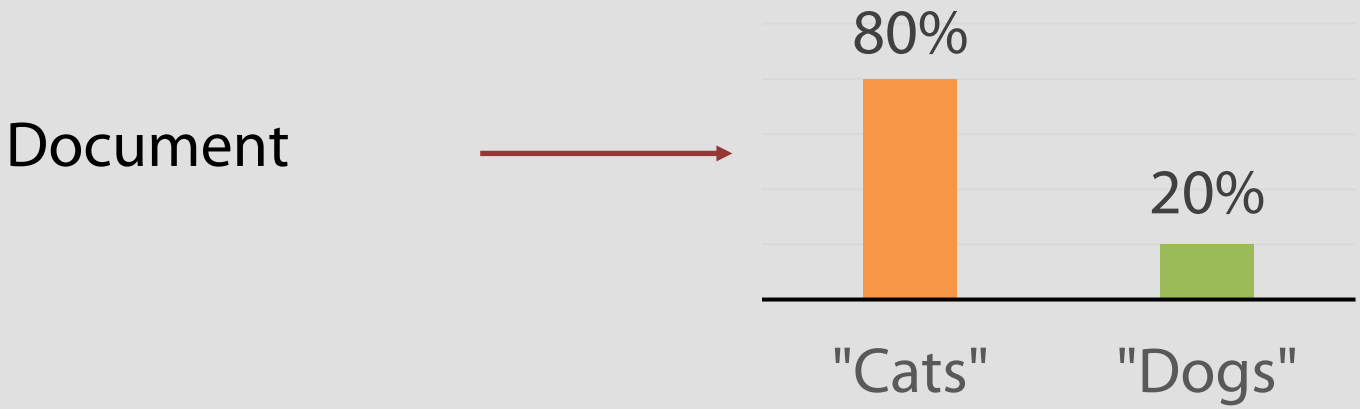
Document is a distribution over topics

Document



Topics

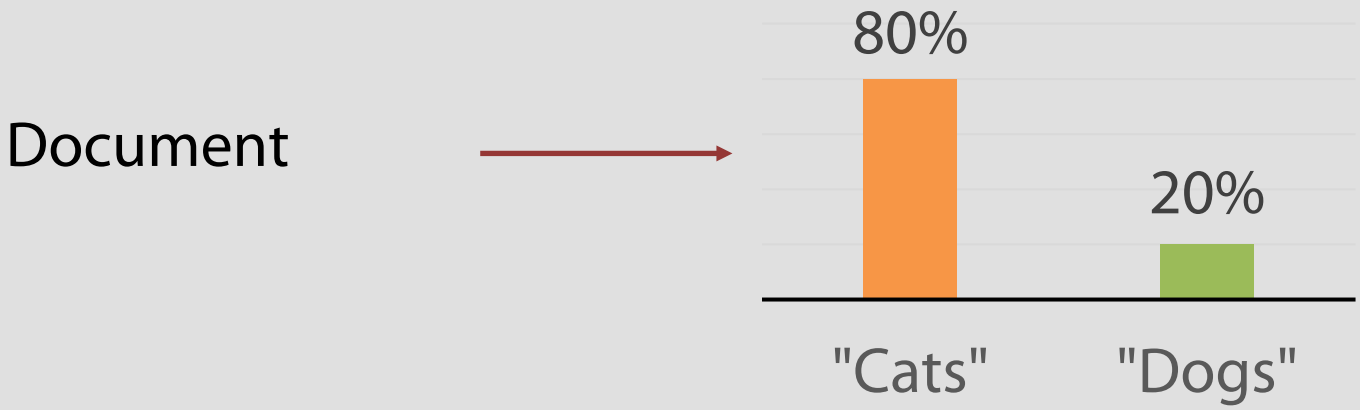
Document is a distribution over topics



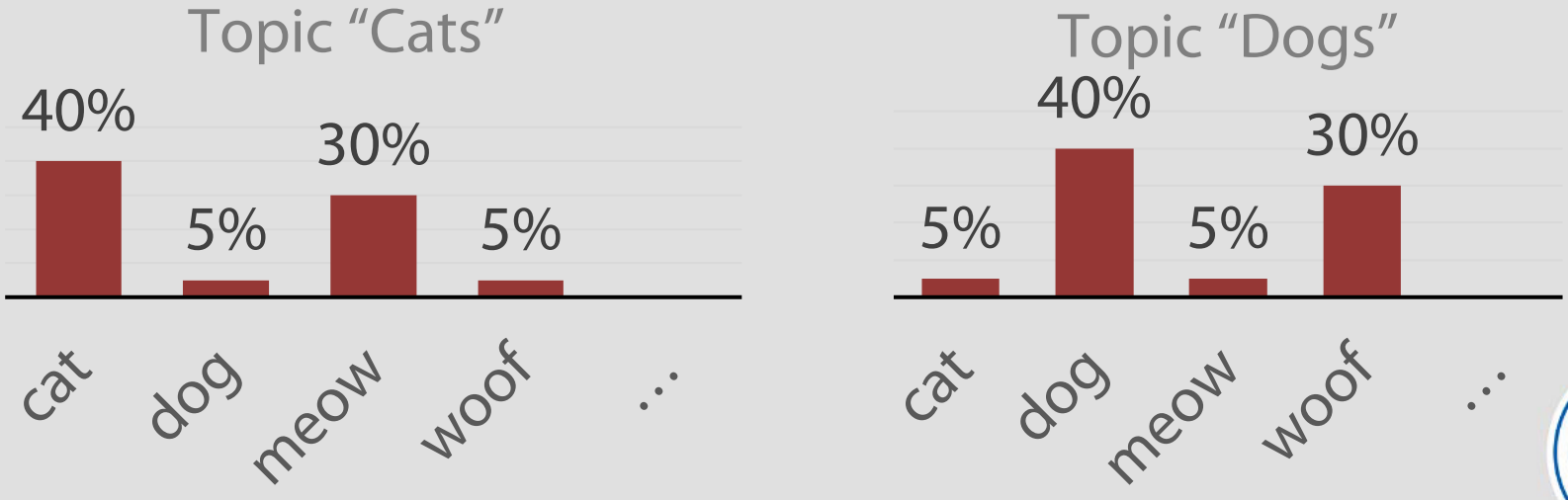
Topic is a distribution over words

Topics

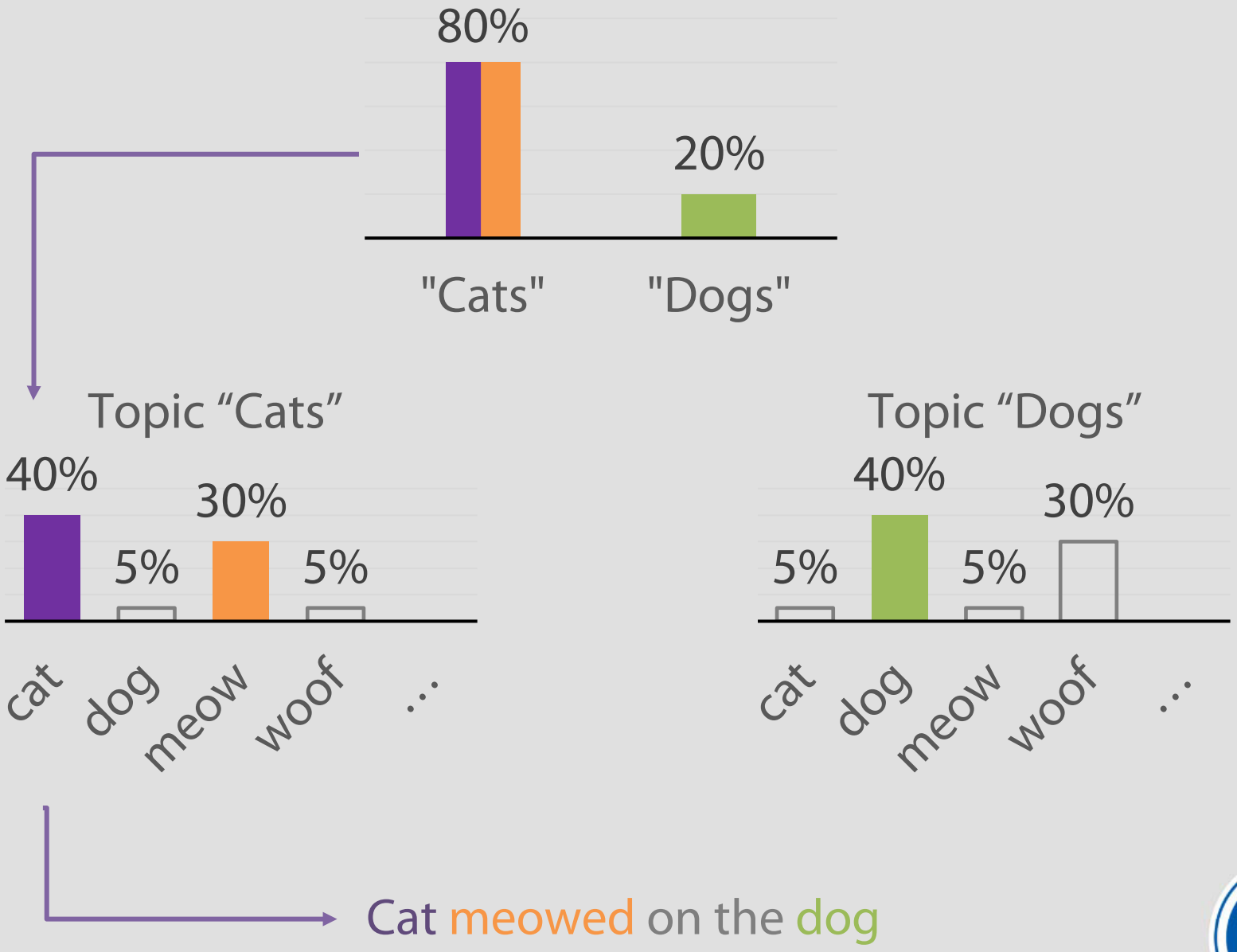
Document is a distribution over topics



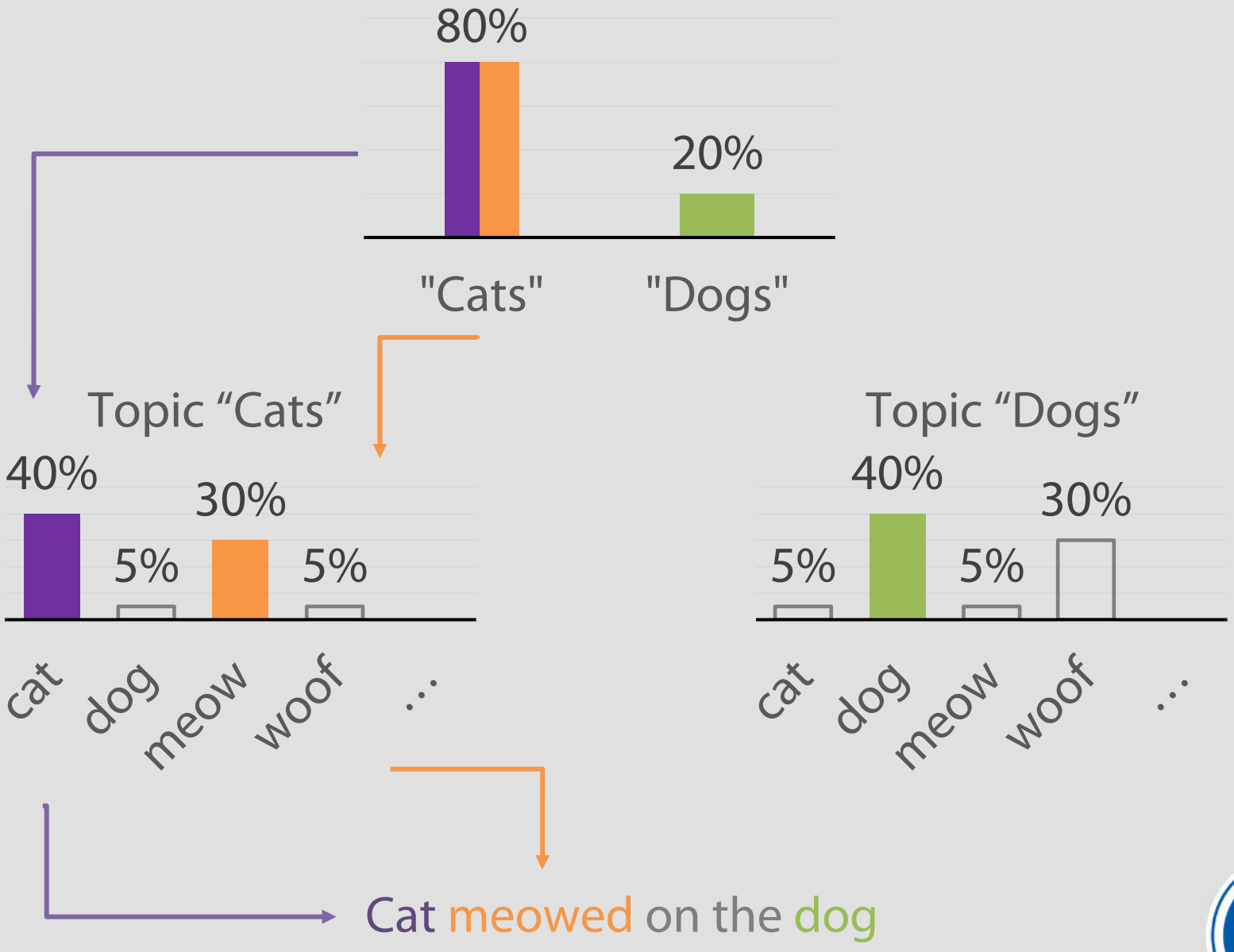
Topic is a distribution over words



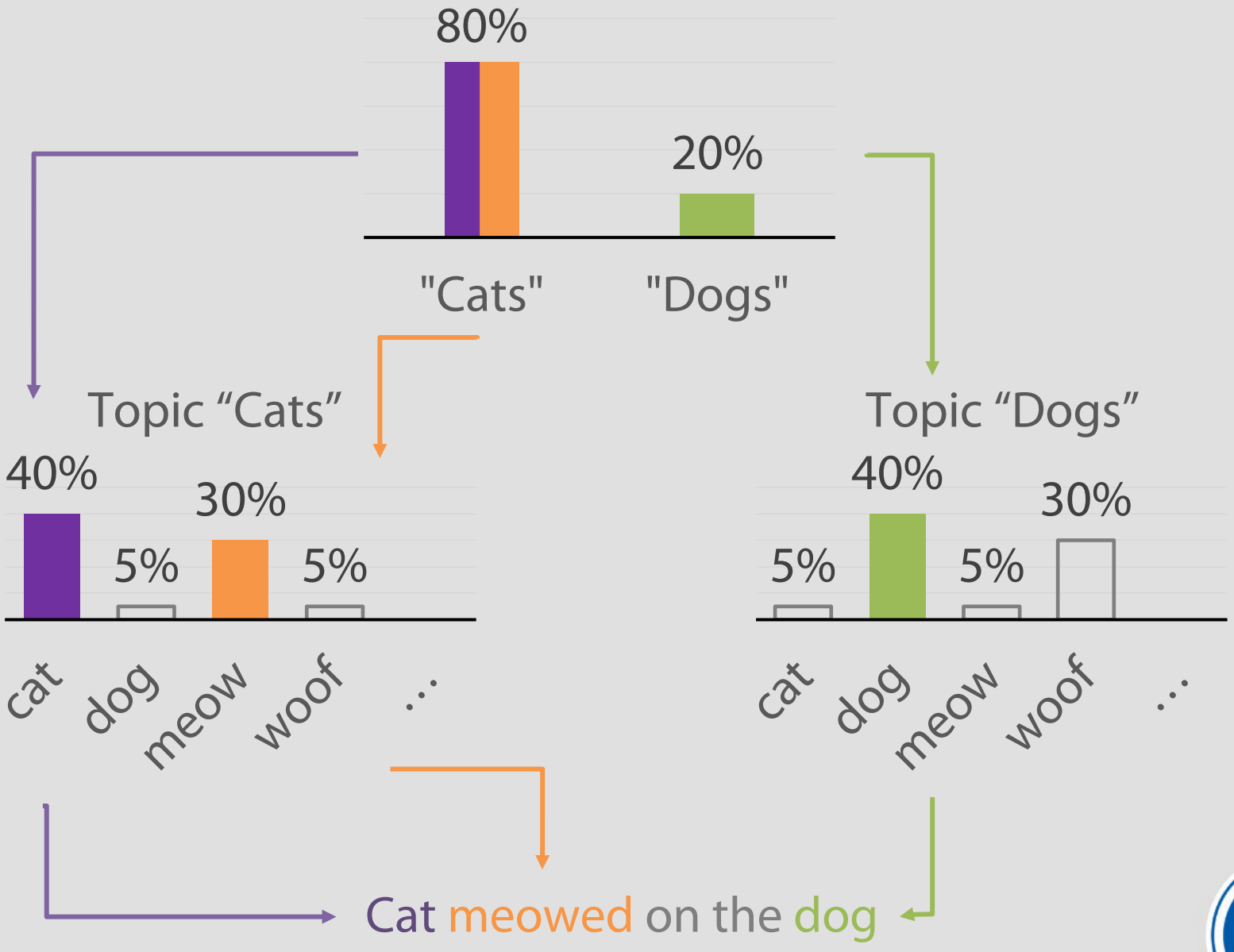
Text generation



Text generation

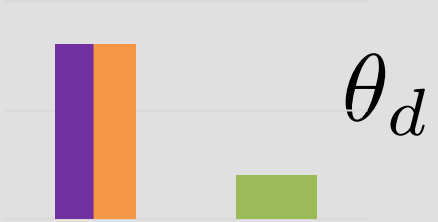


Text generation



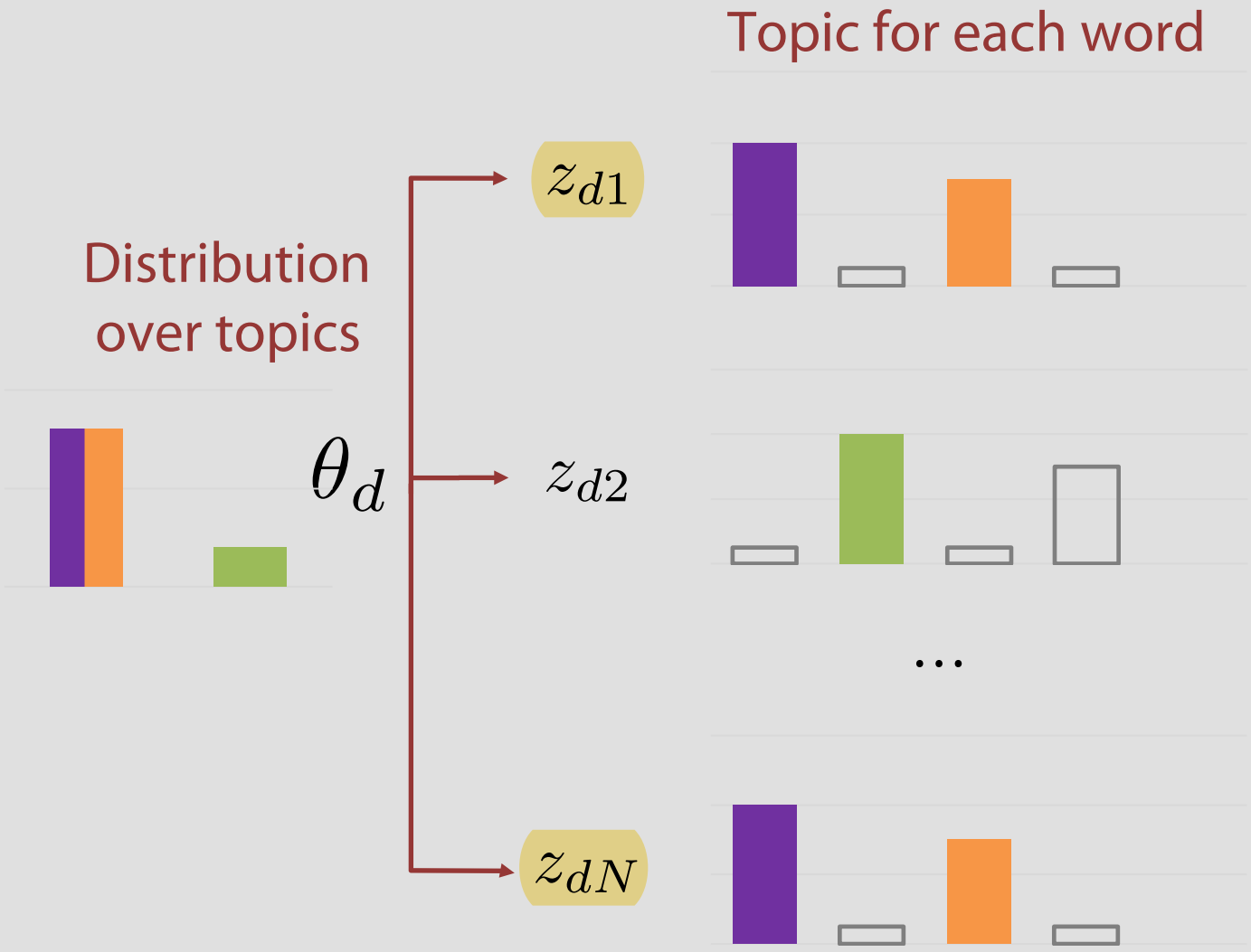
Model

Distribution
over topics

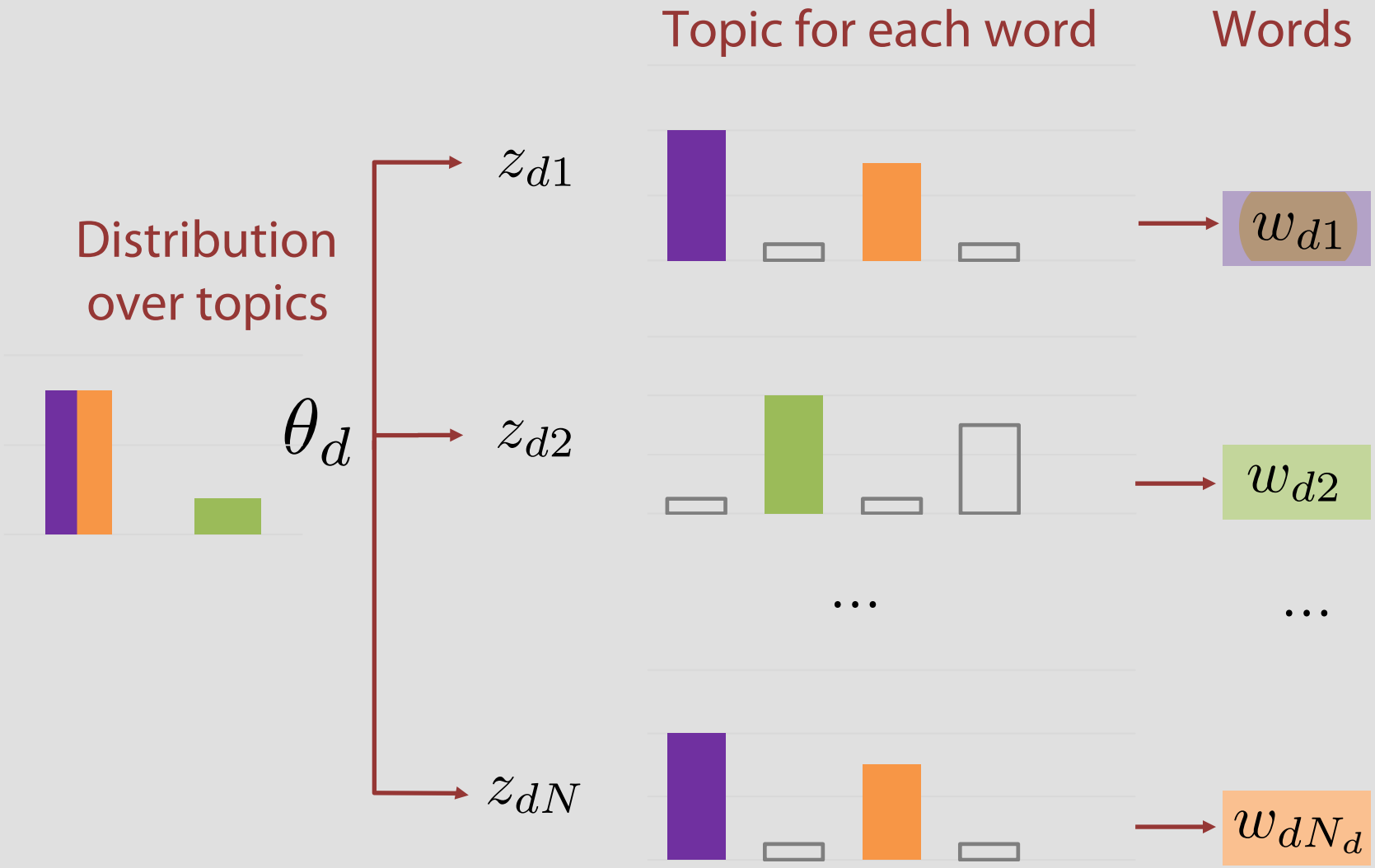


θ_d

Model

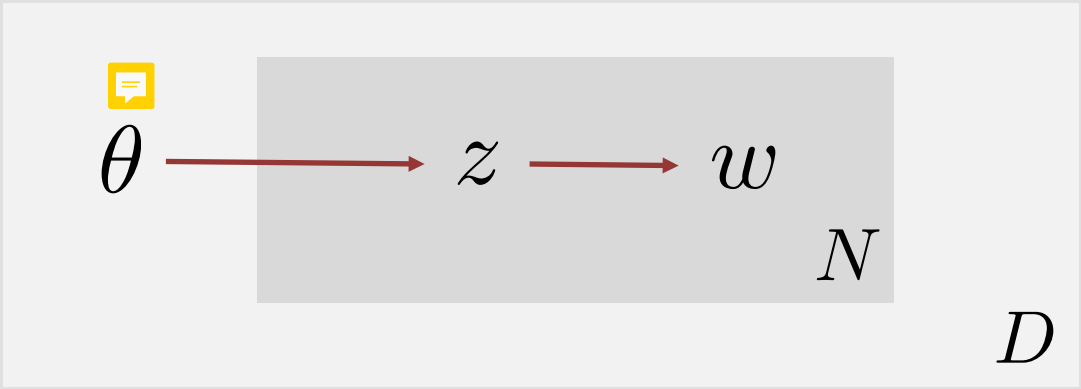

Topic for each word
Distribution over topics
θ_d
z_{d1}
z_{d2}
...
z_{dN}
$z_{dn} \in \{1..T\}$

Model



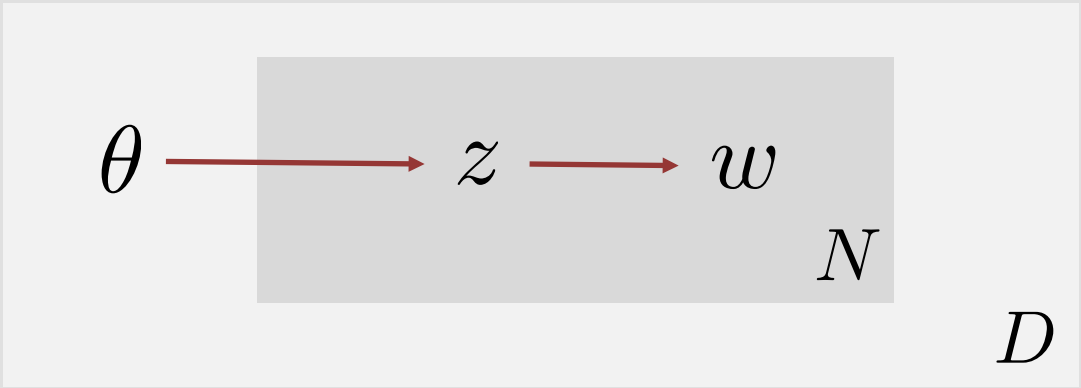
$$z_{dn} \in \{1..T\} \quad w_{dn} \in \{1..V\}$$

LDA Model



$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

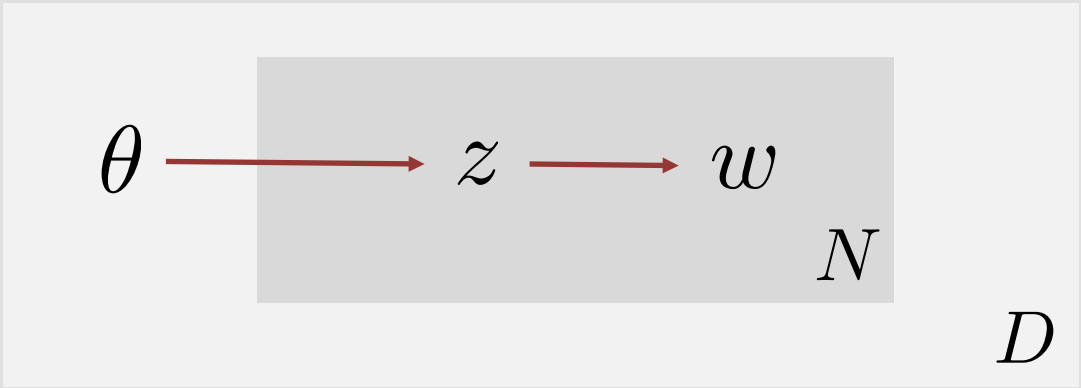
LDA Model



$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document

LDA Model

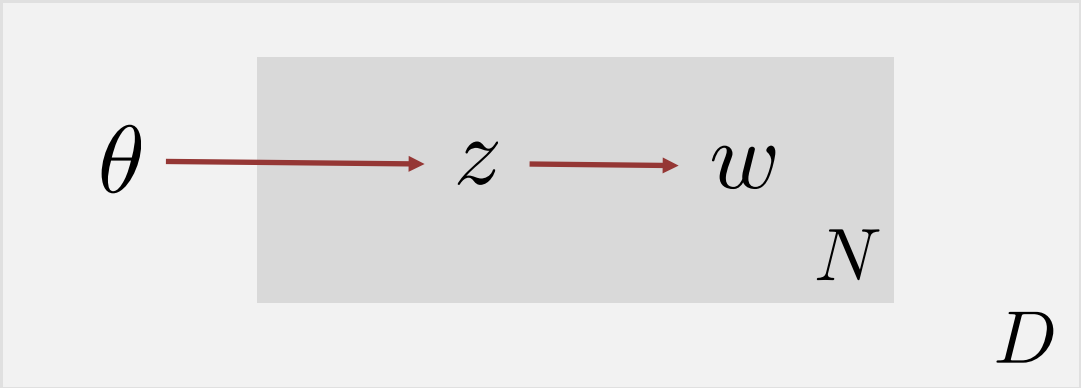


$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document

generate topic probabilities

LDA Model



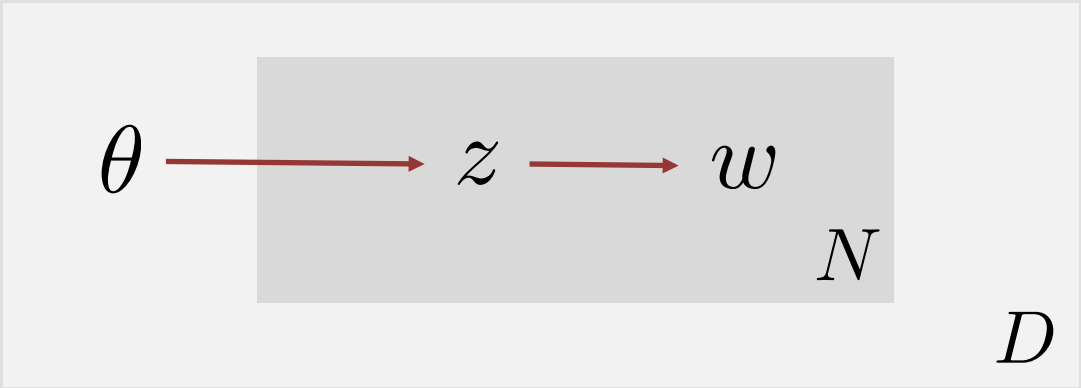
for each word

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document

generate topic probabilities

LDA Model



for each word

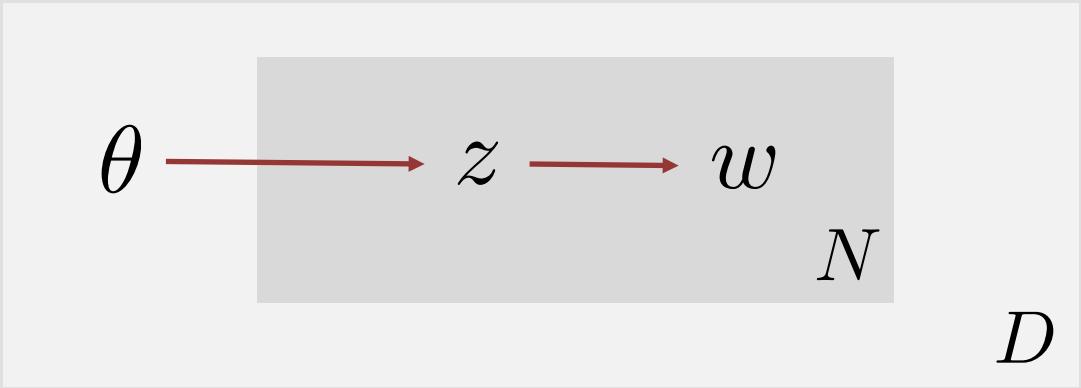
select topic

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document

generate topic probabilities

LDA Model



for each word

select topic

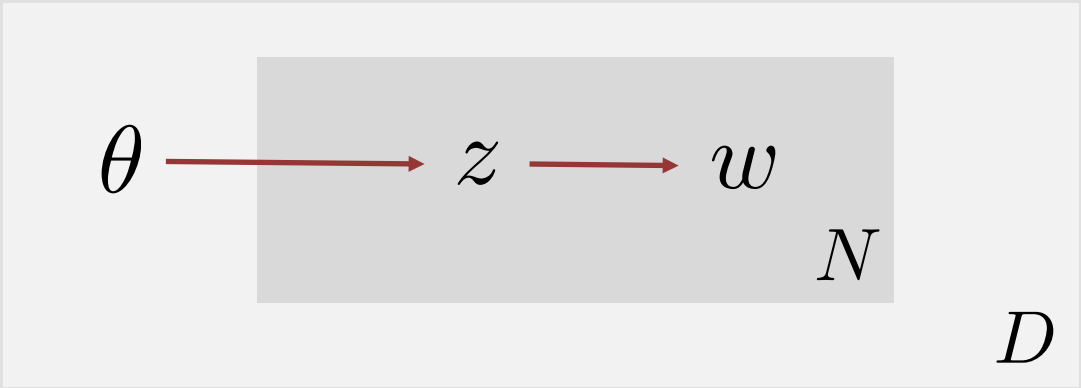
select word from topic

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document

generate topic probabilities

LDA Model



for each word

select topic

select word from topic

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document

generate topic probabilities

LDA Model

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

LDA Model

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

$$p(\theta_d) \sim \mathrm{Dir}(\alpha)$$

LDA Model

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

$$p(\theta_d) \sim \mathrm{Dir}(\alpha)$$

$$p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$$

LDA Model

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

$$p(\theta_d) \sim \mathrm{Dir}(\alpha)$$

$$p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$$

$$p(w_{dn}|z_{dn}) = \Phi_{z_{dn}w_{dn}}$$

LDA Model

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

$$p(\theta_d) \sim \mathrm{Dir}(\alpha)$$

$$p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$$

$$p(w_{dn}|z_{dn}) = \Phi_{z_{dn} w_{dn}}$$

Constraints:

$$\Phi_{tw} \geq 0$$

$$\sum_{w} \Phi_{tw} = 1$$

LDA Model

Known: W data

Unknown: Φ parameters, distribution over words for each topic

Unknown: Z latent variables, topic of each word

Unknown: Θ latent variables, distribution over topics for each document

ТЕХНИЧЕСКИЙ СЛАЙД (15 мин на доску)

- ВЫВОД ФОРМУЛ VAR. EM НА ДОСКЕ