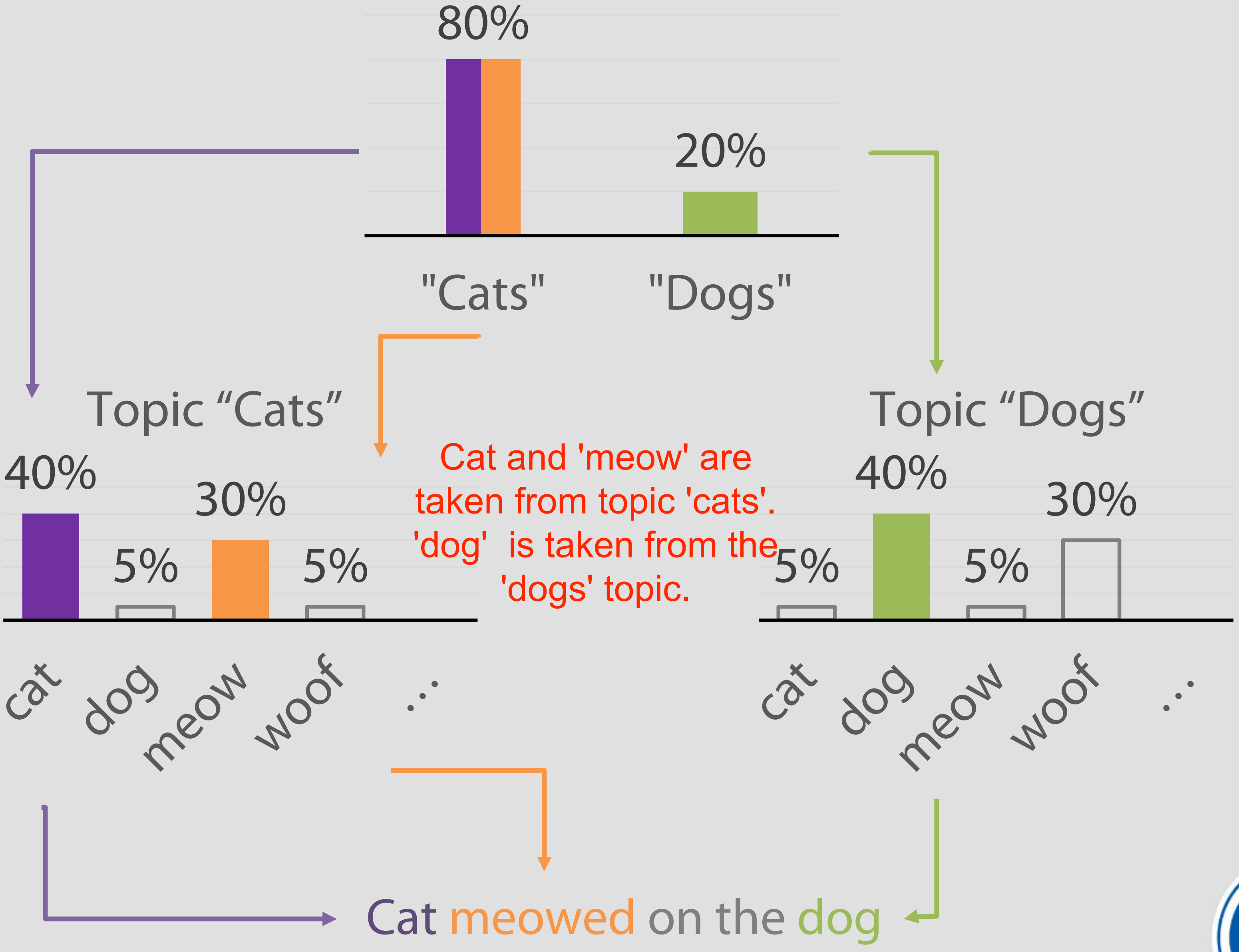
Text generation

How do we generate sentences?



Cat and 'meow' are taken from topic 'cats'. 'dog' is taken from the 'dogs' topic.

Model

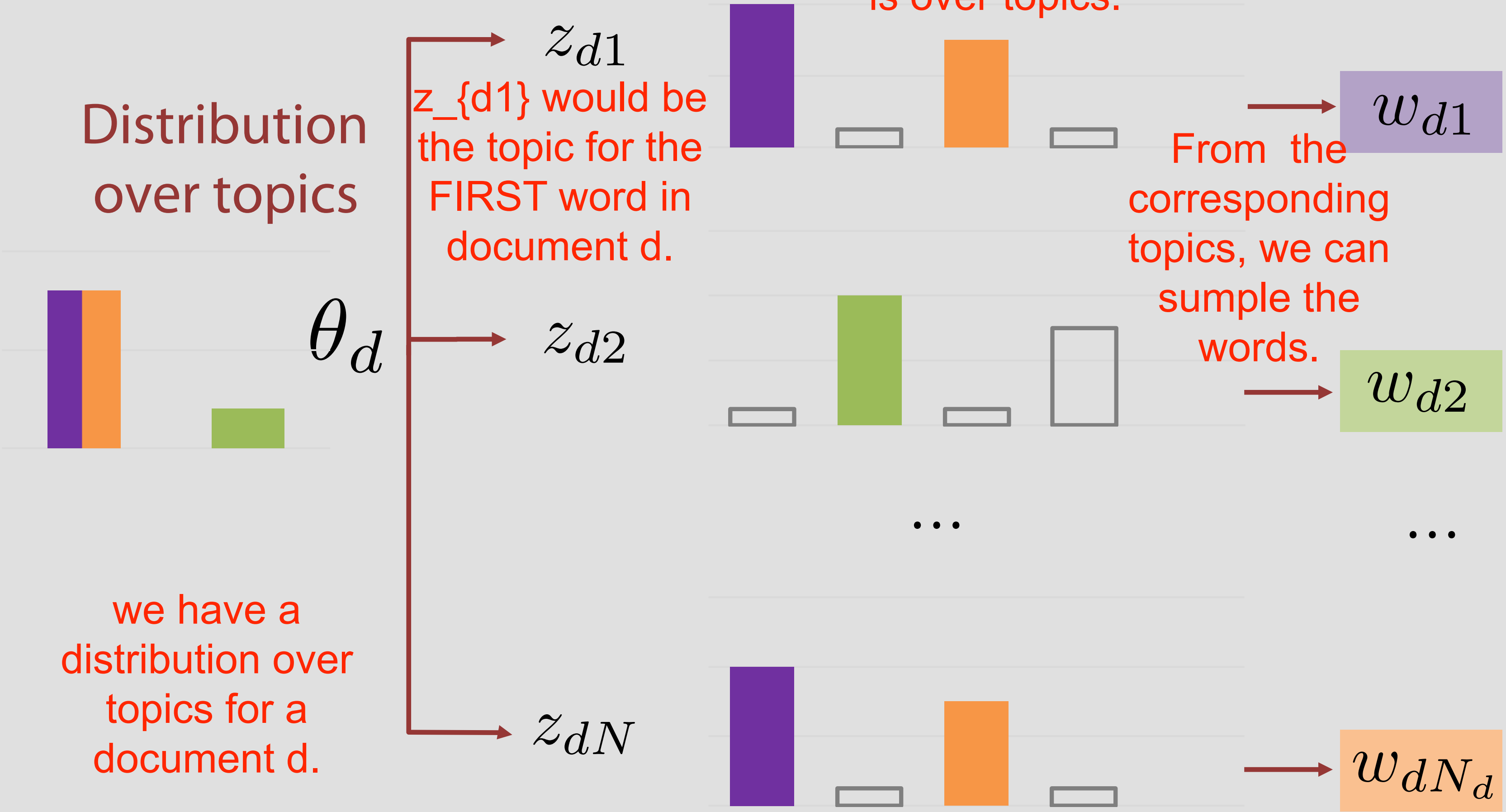
For each word in the document, we assign the topic.

Topic for each word

Words

this distribution is over topics.

Distribution over topics



z_{d1} would be the topic for the FIRST word in document d.

From the corresponding topics, we can sumple the words.

we have a distribution over topics for a document d.

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

There are T topics.

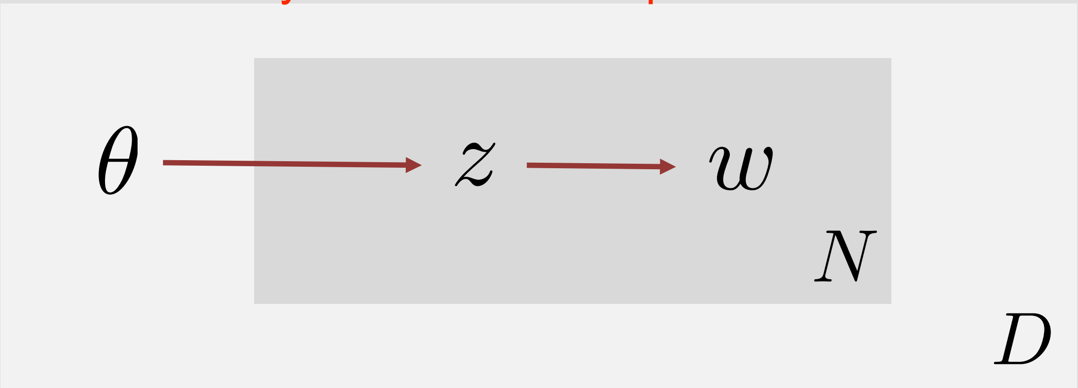
There are V words in the vocabulary.

LDA Model

What we did in the previous slide leads us to this:

A bayesian network in plate notation.

How to interpret plate notation: The theta here is in the D box, which means that we repeat this box for the D documents.
z->w is in the N box. This means that for each document, we have N of these (N words).



in the document,

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

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

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

sum of all theta_d sum up to 1. Why?
I suppose the k-simplex model seems pretty 'fitting'. A document can be really heavy on some topic, but may also touch on other topics too.

Constraints:

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

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

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

←

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

Row z_dn,
column w_dn.

TODO: WHAT is
tw here?

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