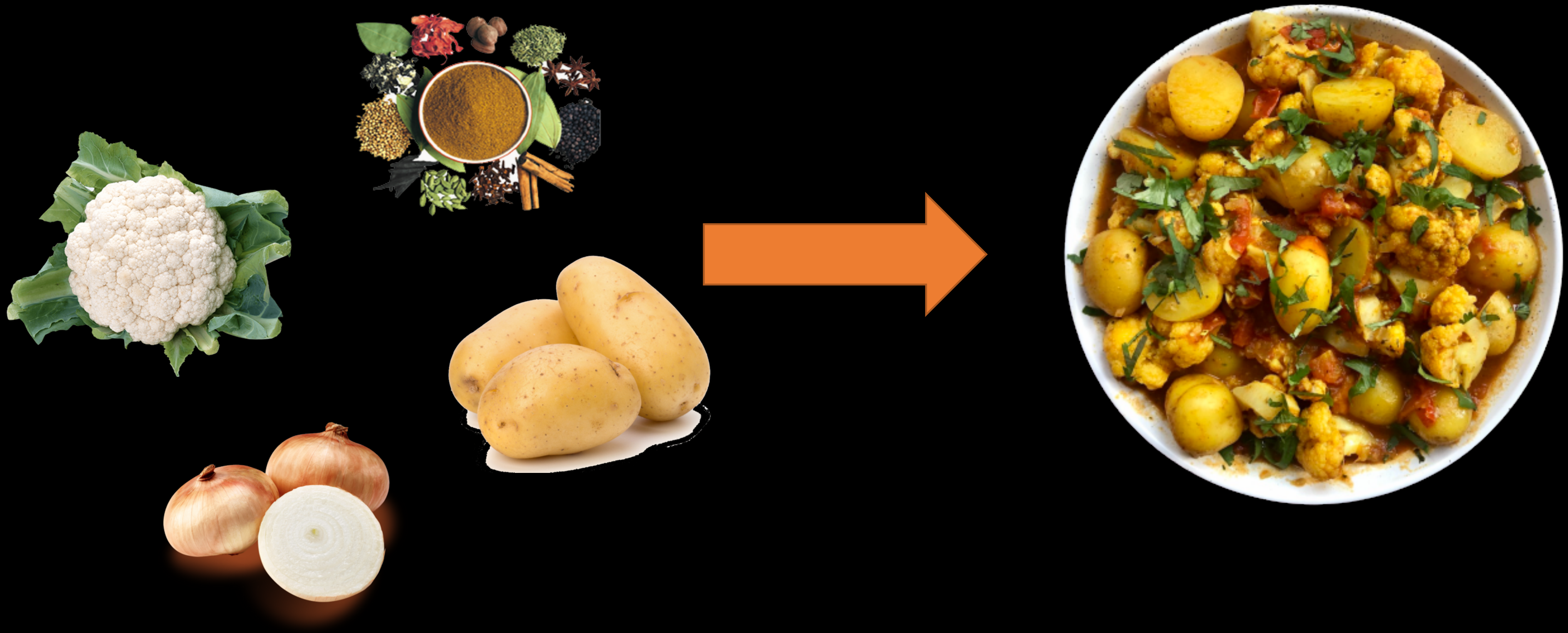
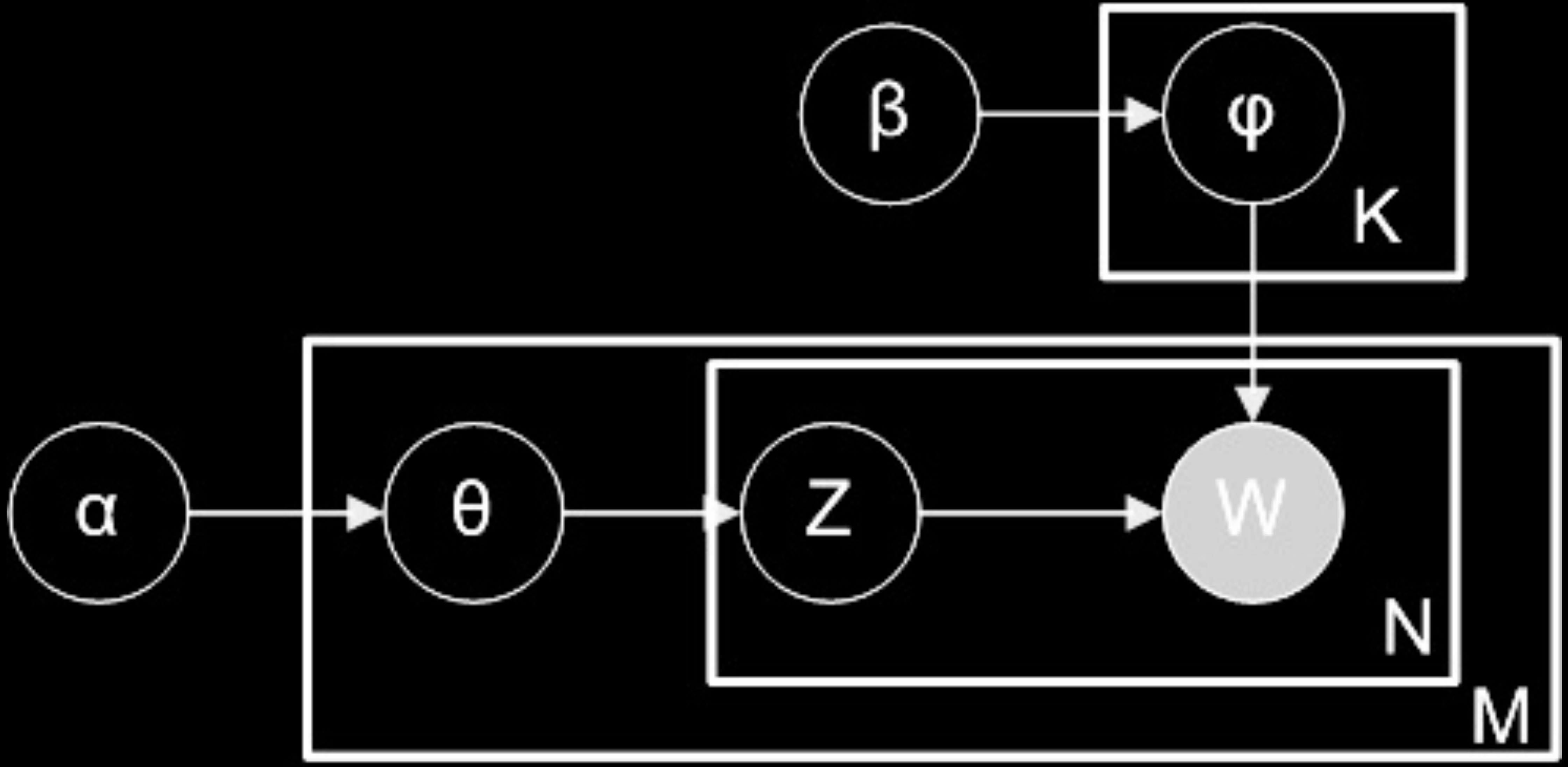
Topic Modelling

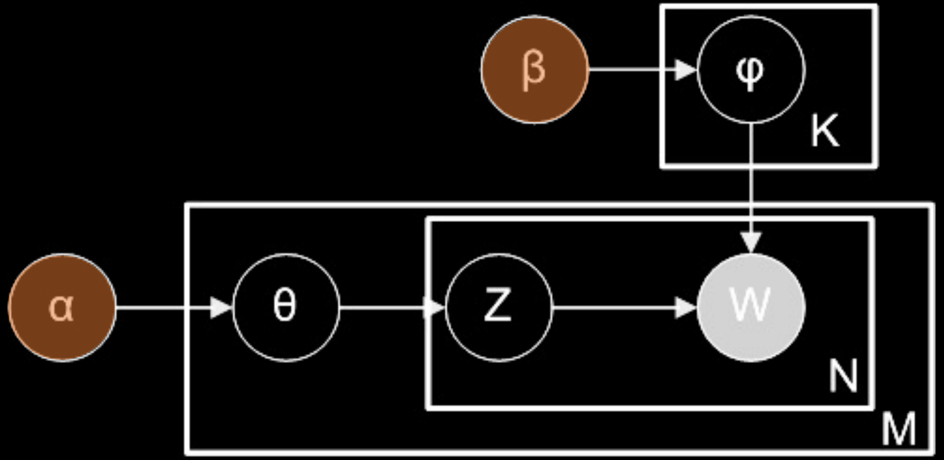
topics

document




β
φ
K
α
θ
Z
W
N
M

Level 1: The Corpus



α = the probability of each topic's appearing in the corpus

β = the probability of each word belonging to each topic

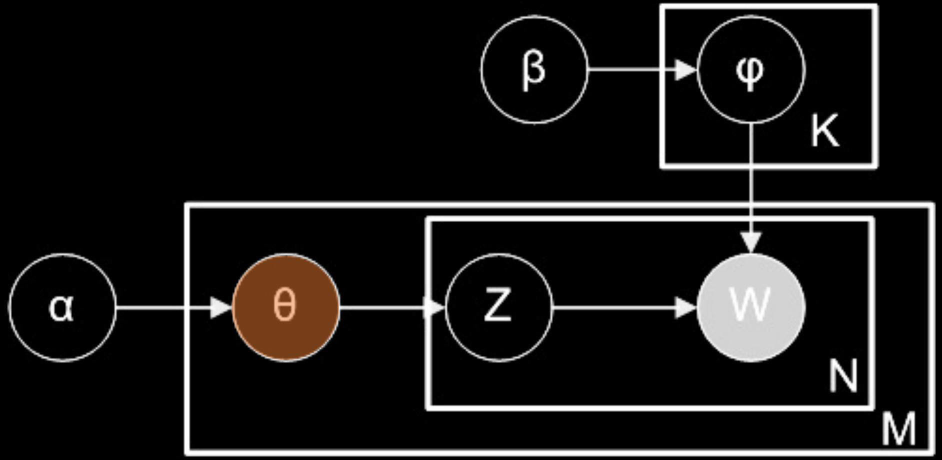
Let us assume that *k* = 10 (i.e. we want the computer sort all the words in all the documents into 10 different topics)

$\alpha = \{0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1\}$

β = *k* x V matrix

	apple	pear	tiger	nail
Topic 1	0.001	0.001	0.6	0.001
Topic 2	0.4	0.45	0.001	0.001
Topic 3	0.15	0.2	0.001	0.1
Topic 4	0.001	0.001	0.001	0.5

Level 2: The Document



θ = the probability of each topic appearing in each document

θ = *m* x *k* matrix

One column for each topic

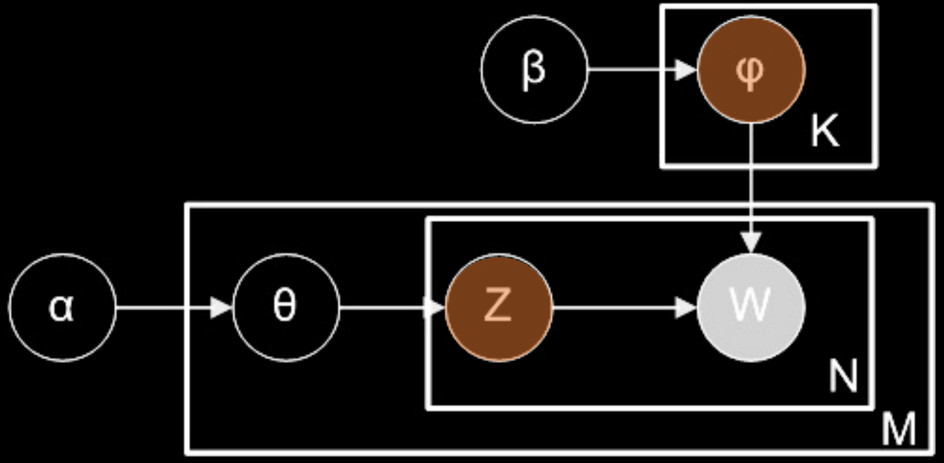
One row for each document

	Topic 1	Topic 2	Topic 3	Topic 4	…
Doc 1	0.3	0.5	0.001	0.00.1	…
Doc 2	0.001	0.001	0.9	0.001	…
Doc 3	0.1	0.3	0.001	0.001	…
Doc 4	0.001	0.001	0.001	0.001	…

Level 3: The Word

z = the topic of this word in the document (based on the topic mixture of the document)

φ = the word distribution for topic *z*



… , he continued, ‘I think we should plant an apple here under the wall.’ …

$\theta_1 =$

	Topic 1	Topic 2	Topic 3	Topic 4	…
Doc 1	0.3	0.5	0.001	0.001	…

$\text{sample}(\theta_1) \rightarrow 2$

$z = 2$

$\varphi = \beta_2 =$

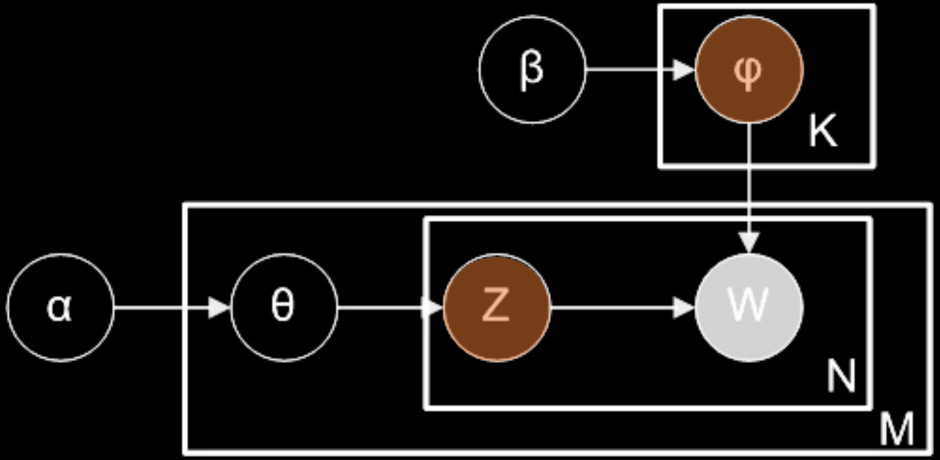
	apple	pear	tiger	nail
Topic 2	0.4	0.45	0.01	0.001

$\text{sample}(\beta_2) \rightarrow \text{apple}$

Level 3: The Word

z = the topic of this word in the document (based on the topic mixture of the document)

φ = the word distribution for topic *z*



… , he continued, ‘I think we should plant an apple here under the wall.’ …

Z = *n* x k matrix

One row for each individual word in the whole corpus (e.g. about 1 million rows for Shakespeare’s works)

	Topic 1	Topic 2	Topic 3	Topic 4	…
…	…	…	…	…	…
should	0	0	0	1	…
plant	0	1	0	0	…
apple	0	1	0	0	…

One column for each topic