Latent Dirichlet Allocation

Topic Modelling

Today's Topic

- Topic modeling
- Topic modeling methods
- Introduction to LDA

Problem?

Topic Modelling

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.
It helps in:

- Uncover the hidden topical patterns that pervade the collection.
- Annotate the documents according to those topics.
- Use the annotations to organize, summarize, and search the texts.

Topic Modeling

The underlying assumption is that every document comprises a statistical mixture of topics, i.e. a statistical distribution of topics that can be obtained by “adding up” all of the distributions for all the topics covered. What topic modeling methods do is try to figure out which topics are present in the documents of the corpus and how strong that presence is.

Discover topics from a corpus

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Topic modeling methods

Some of the famous topic modeling methods are

- **Latent Dirichlet Allocation (LDA)**
- Non Negative Matrix Factorization (NMF)
- Latent Semantic Analysis (LSA)
- Parallel Latent Dirichlet Allocation (PLDA)

We will only discuss LDA today.

Overview of LDA (Latent Dirichlet Allocation)

A probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model,

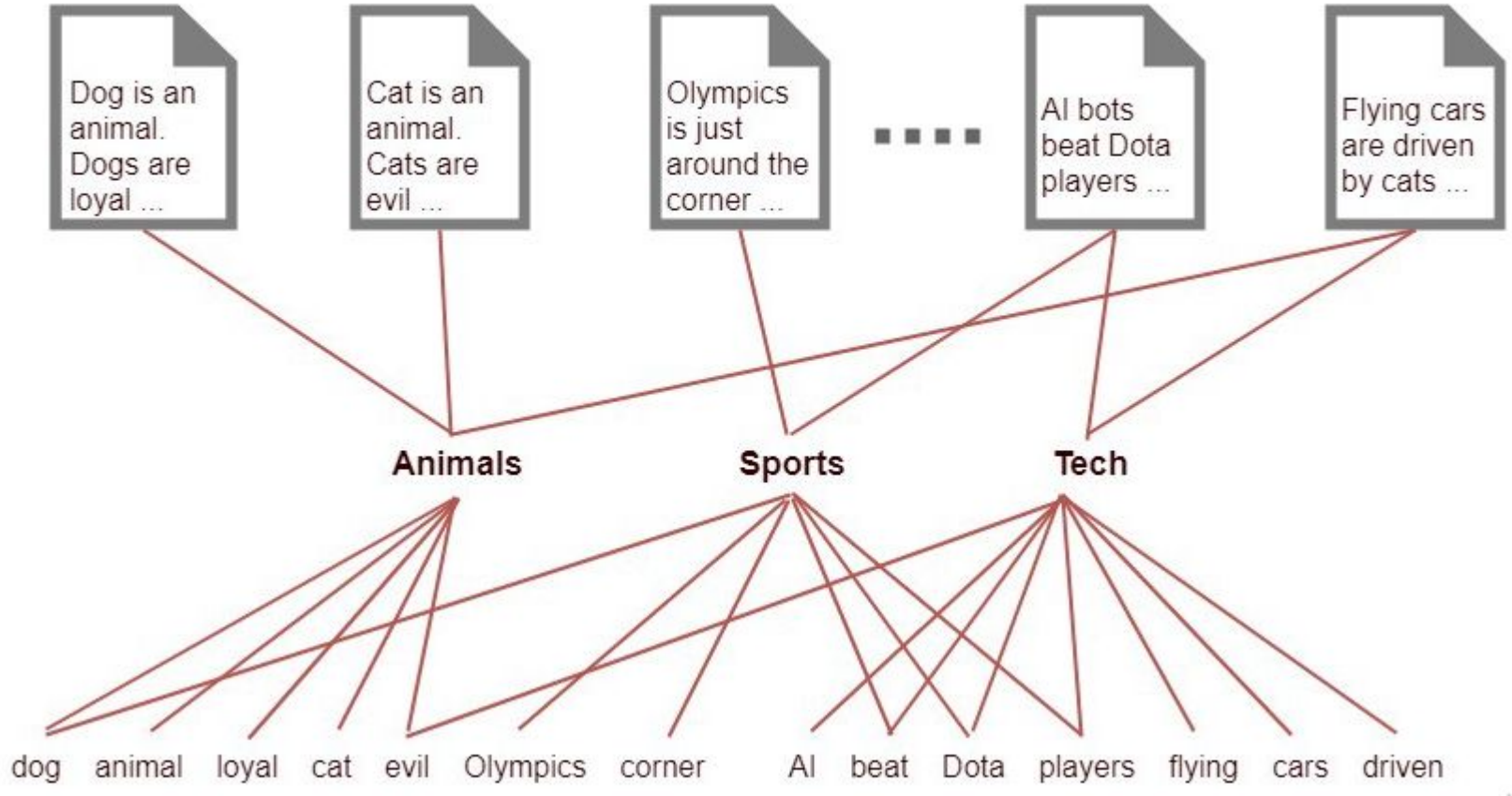
- In which each item of a collection is modeled as a finite mixture over an underlying set of latent topics.
- Each observed word originates from a topic that we do not directly observe.
- Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

Overview of LDA

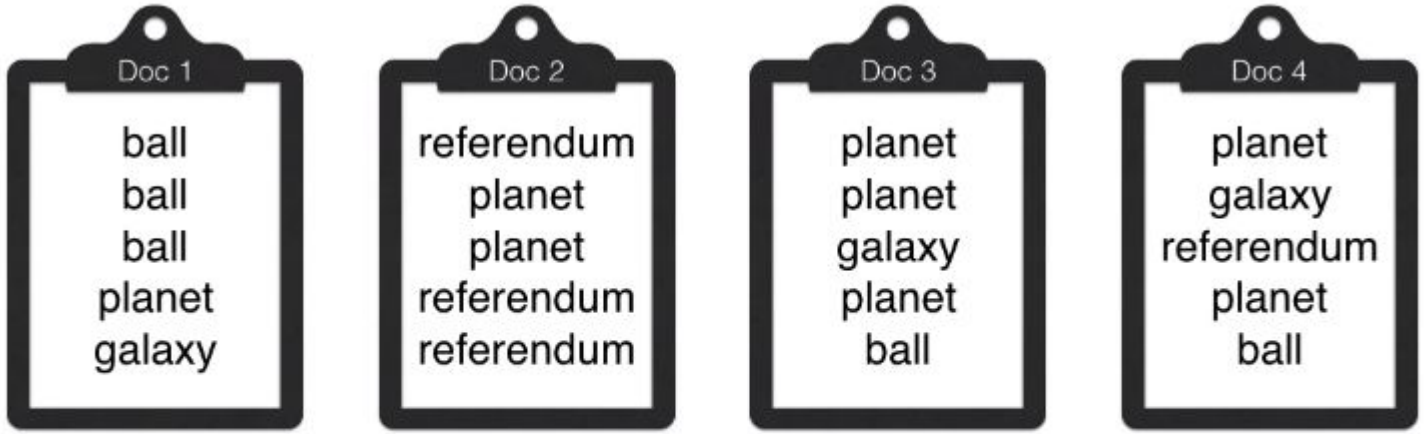
In other words this model follows the concept that each document can be described by the probabilistic distribution of topics and each topic can be described by the probabilistic distribution of words. Thus we can get a much clearer vision about how the topics are connected.

LDA Example

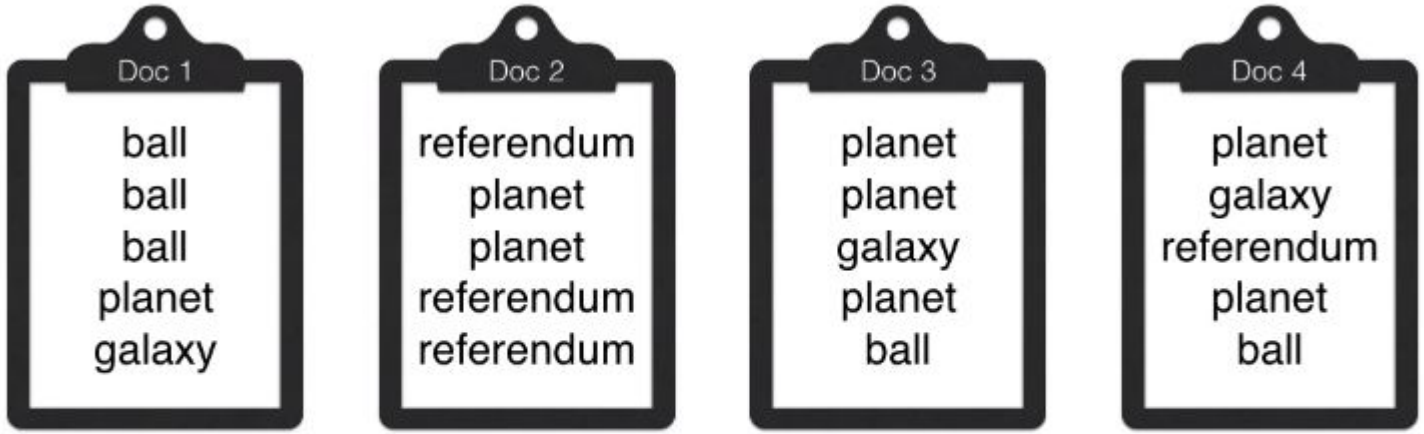
Consider you have a corpus of 1000 documents. After preprocessing the corpus, the bag of words consists of 1000 common words. By applying LDA, we can determine the topics which are related to each document. Thus it is made simple to obtain the extracts from the corpus of data.



The upper level represents the documents, the middle level represents the topics generated and the lower level represents the words. Thus it clearly explains the rule it follows that document is described a the distribution of topics and topics are described as the distribution of words.



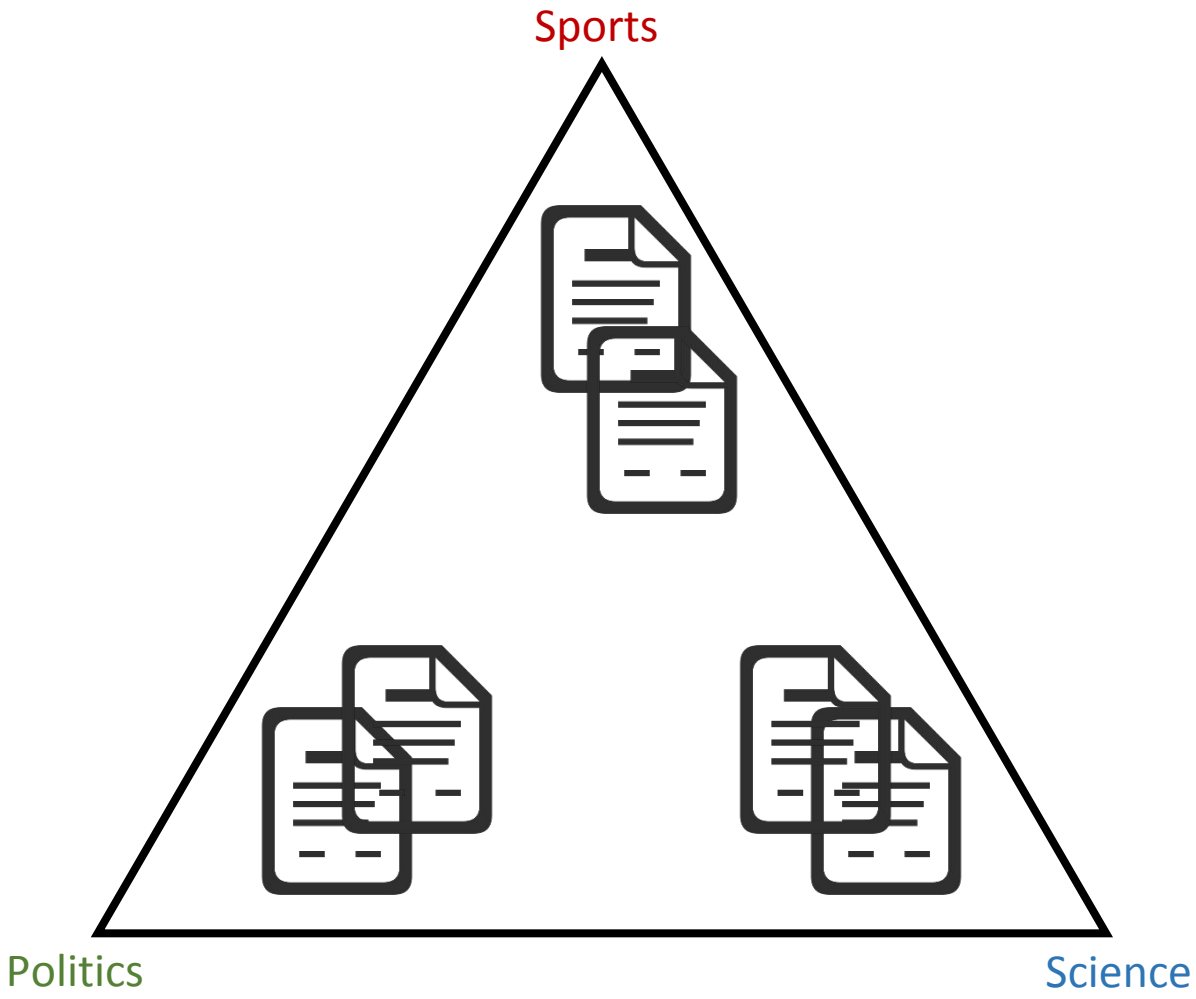
Sports **Politics** **Science**



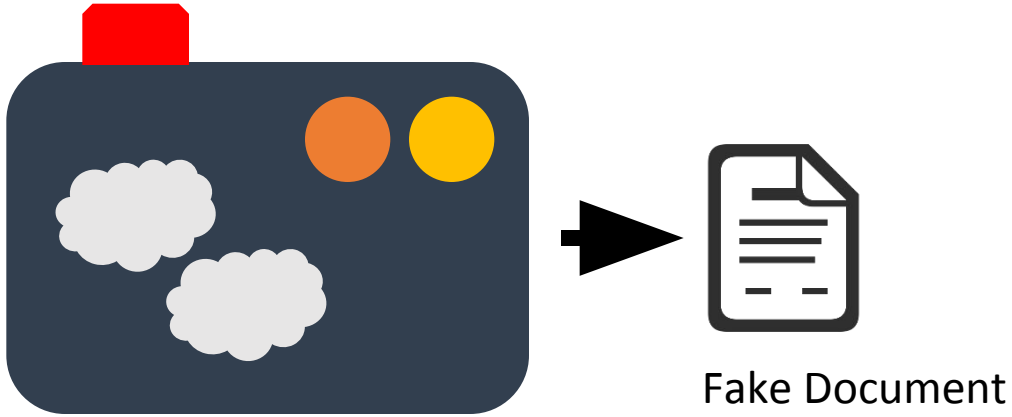
Sports Politics Science Science

Sports Politics Science

LDA



LDA: A machine that generates documents

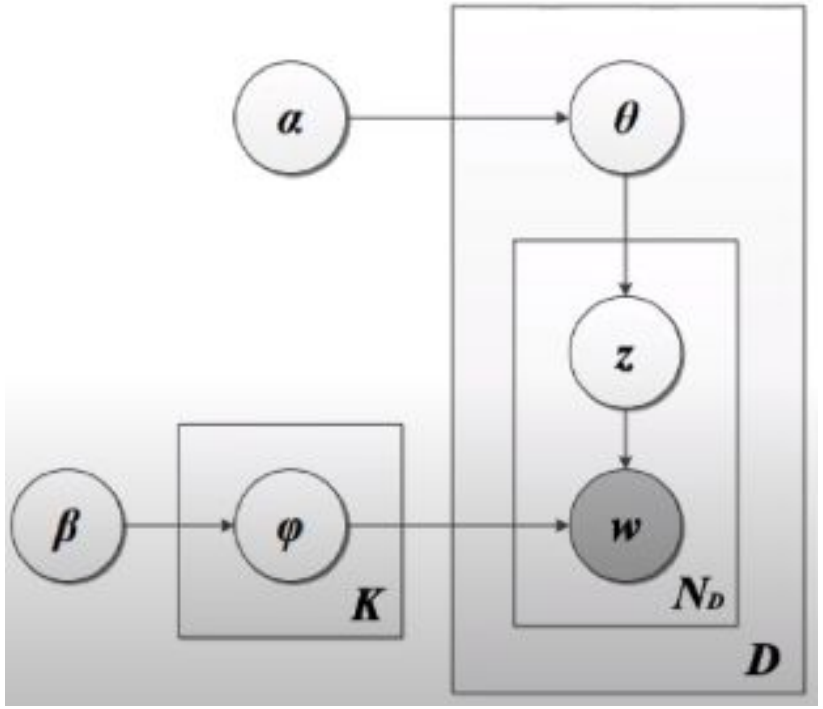


Best setting on the machine

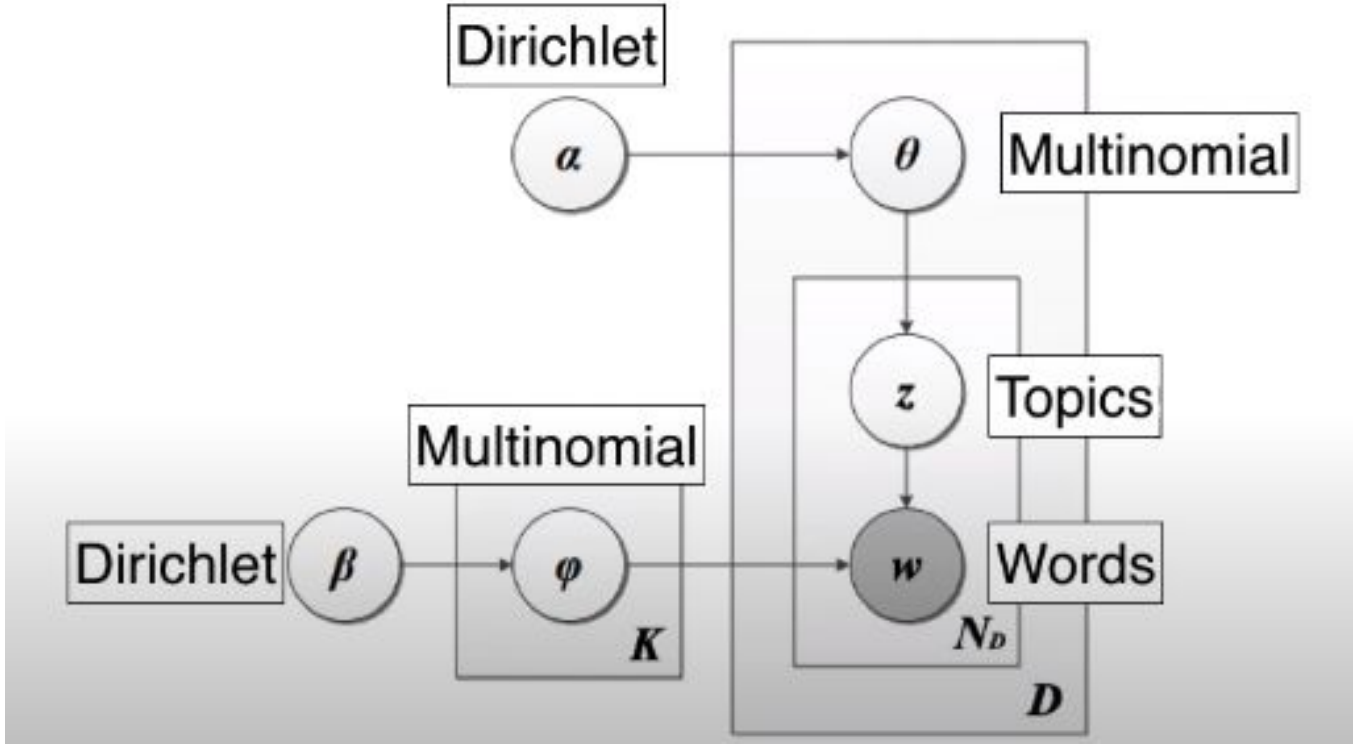
Best setting on the machine

Best Setting: Best machine

Blueprint of a LDA machine

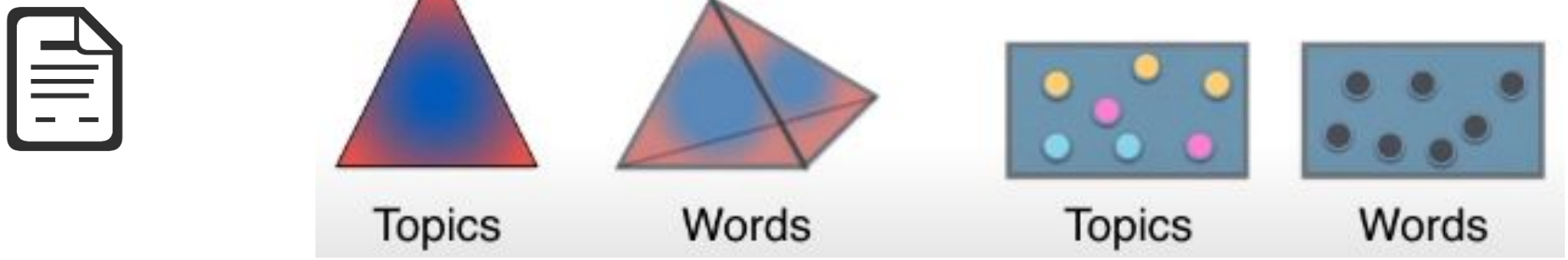


Blueprint of a LDA machine

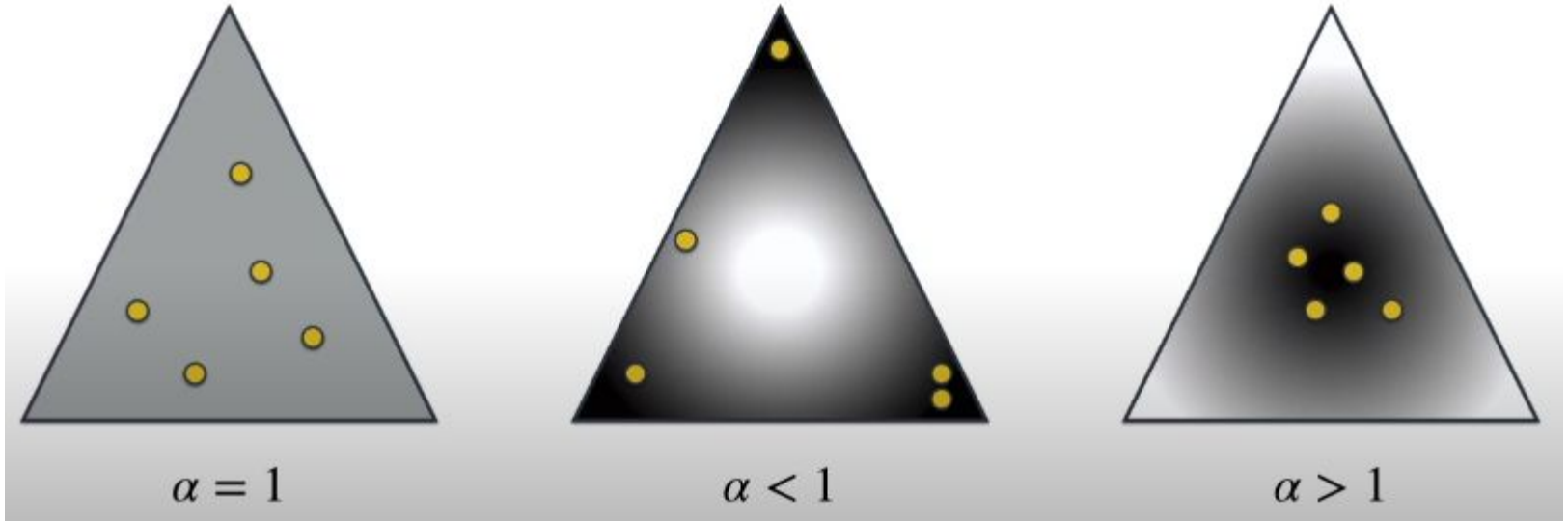


Probability of a document

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

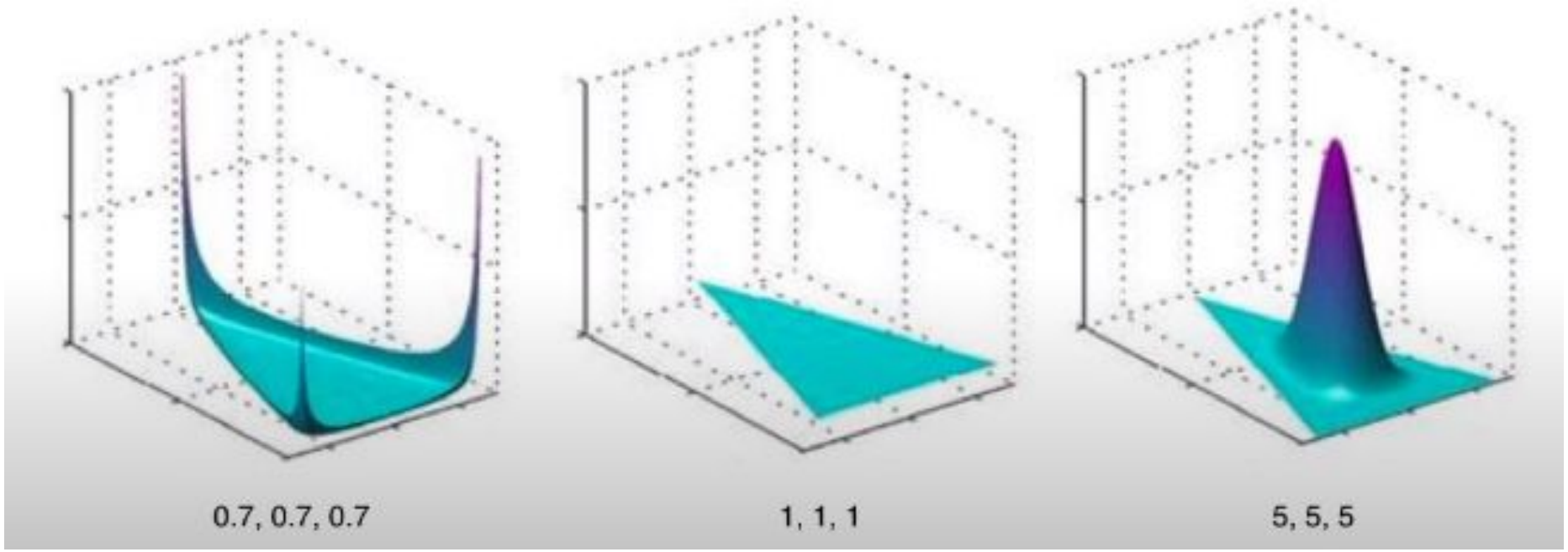


Dirichlet Distribution

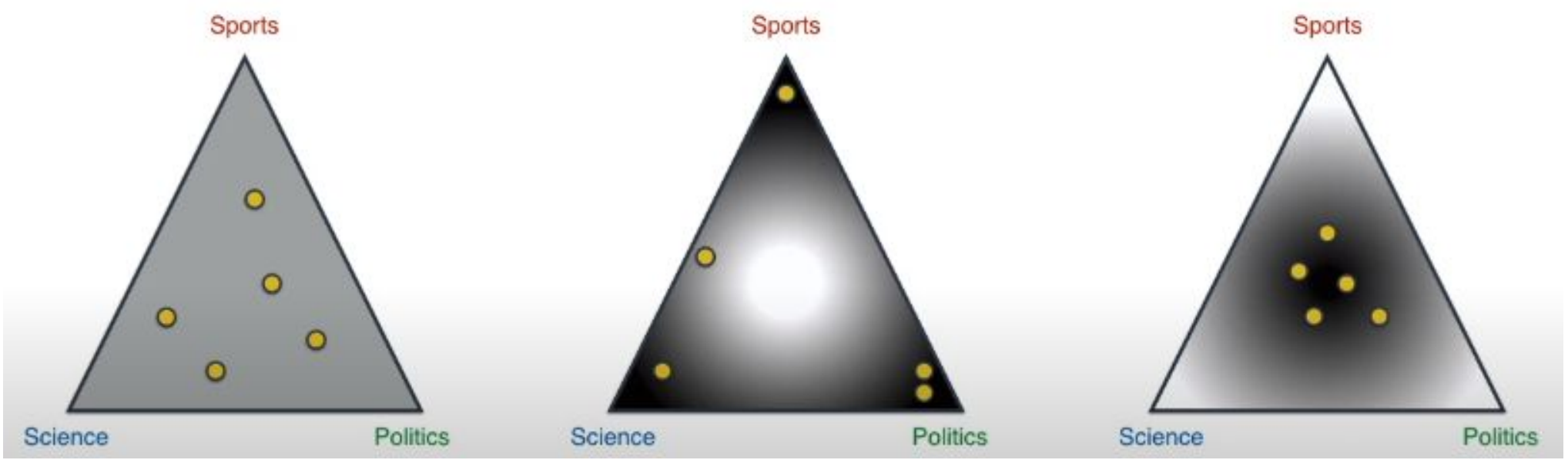


Dirichlet Distribution

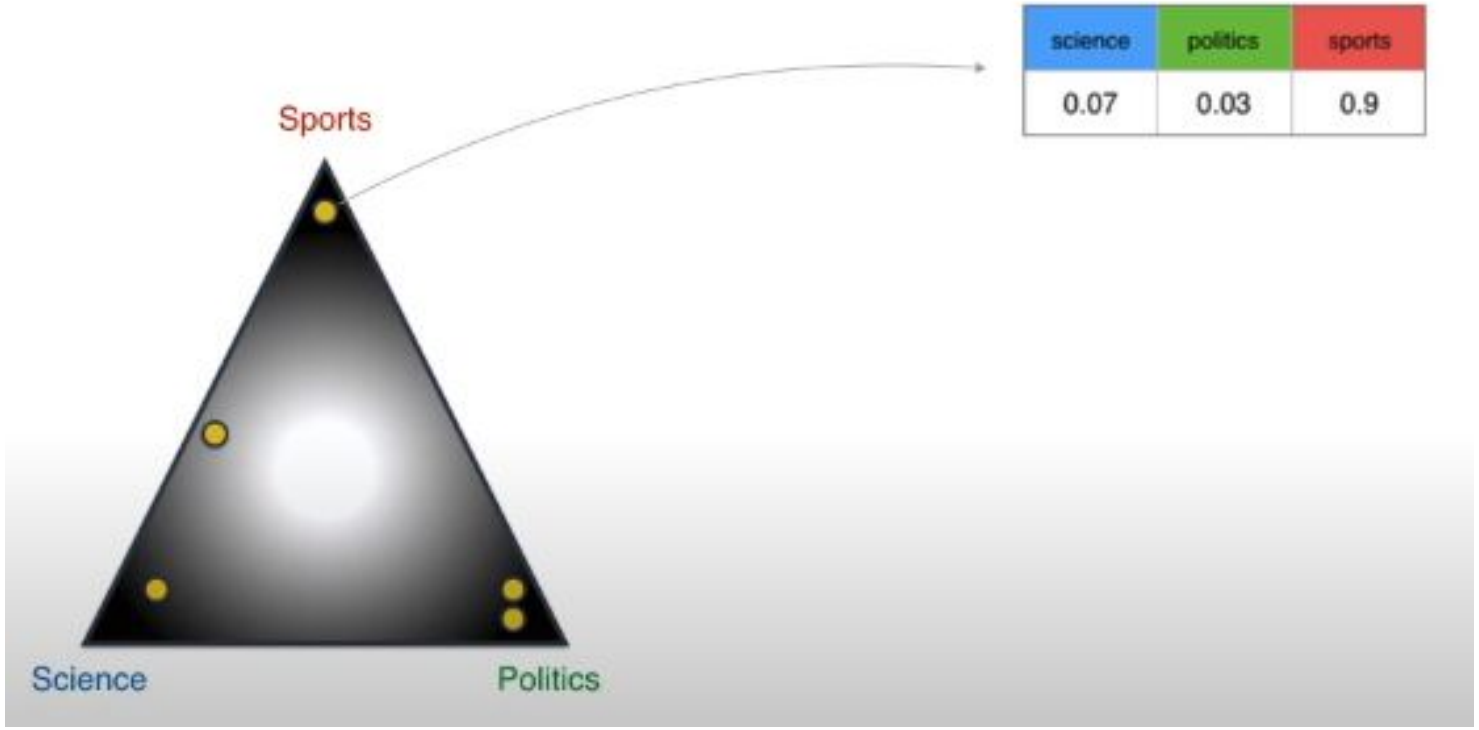
$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$



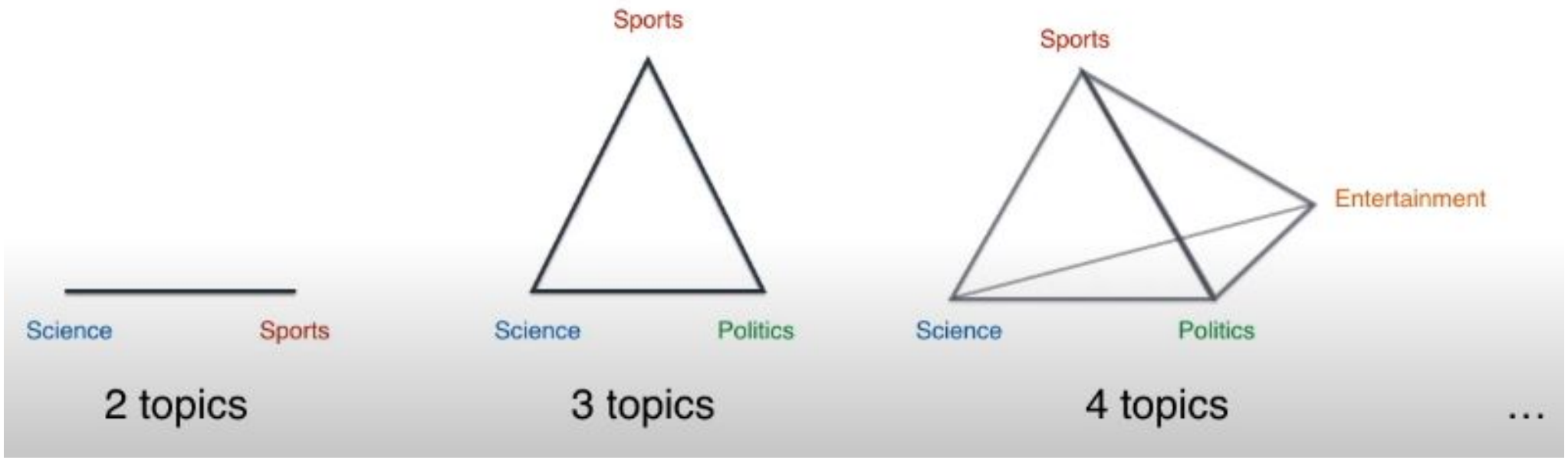
Which one for topics?



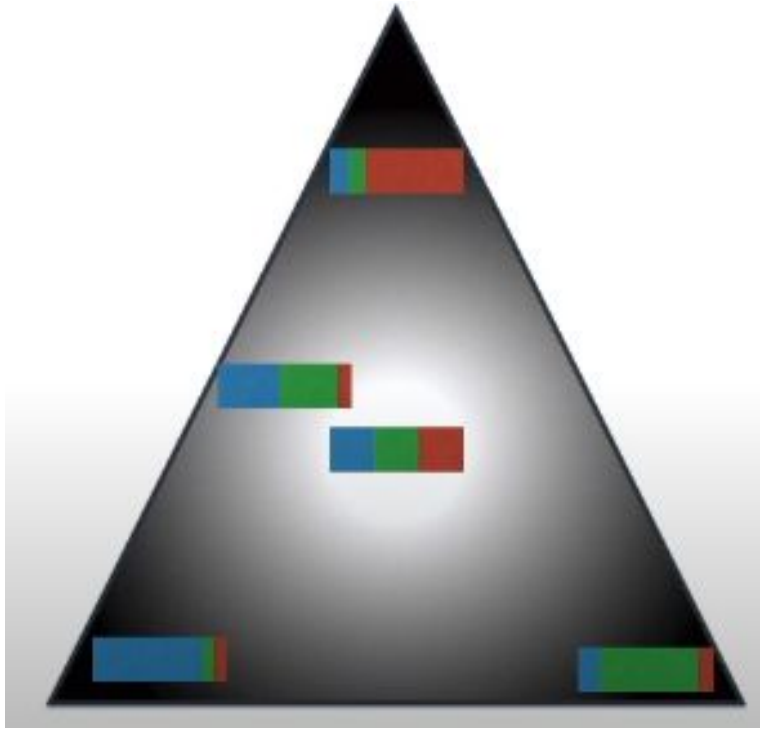
Which one for topics?



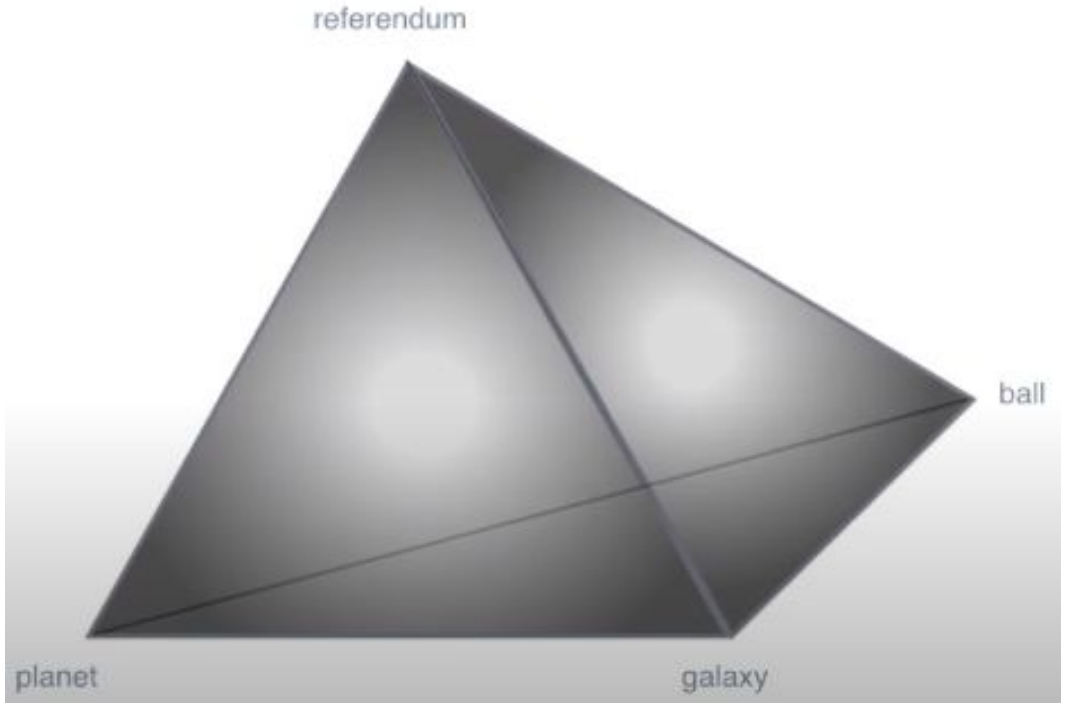
More topics? More Dimensions.



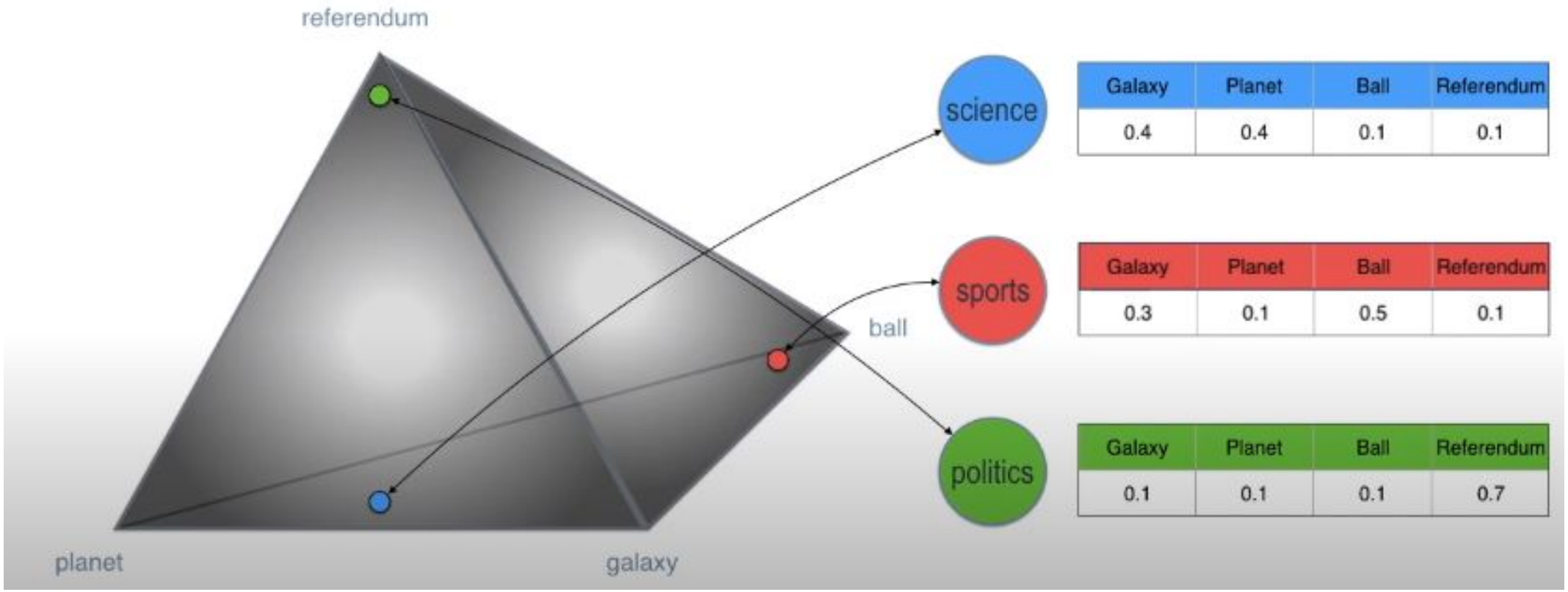
Distribution of distributions



Where to put the topics?



Where to put the topics?

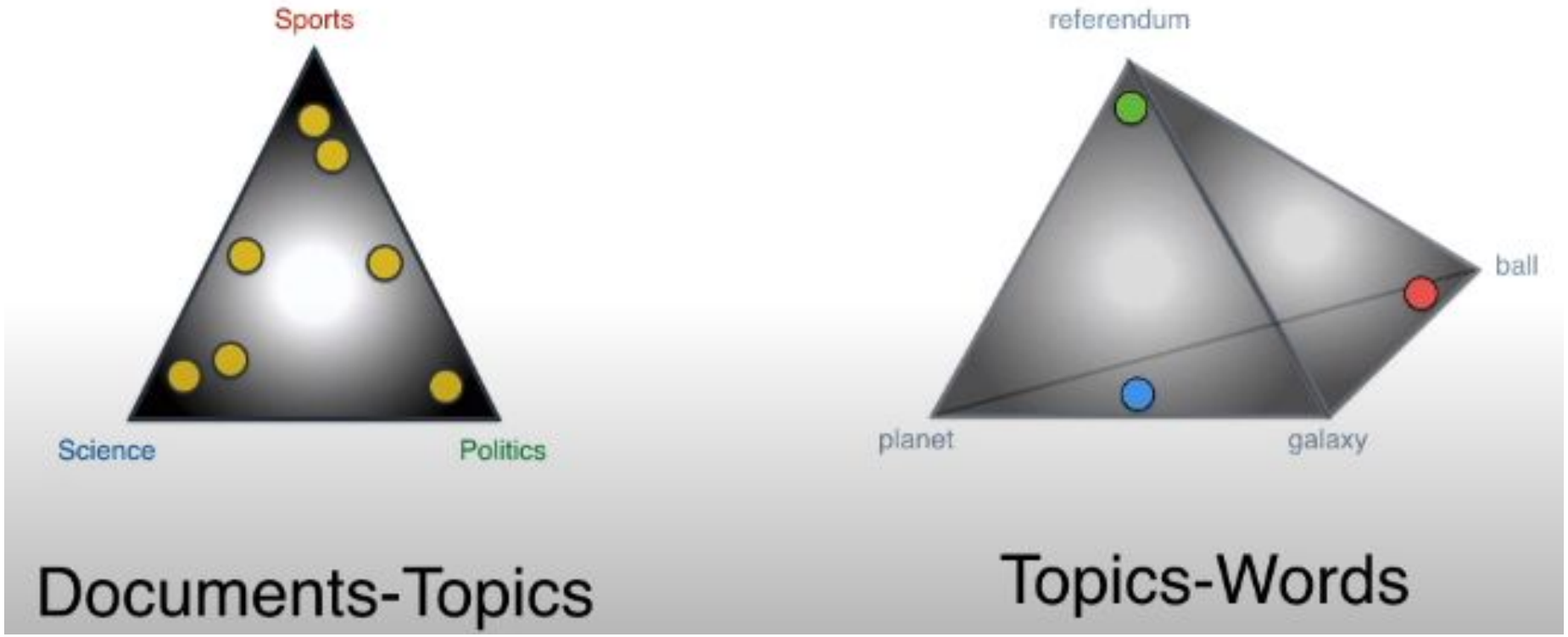


Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

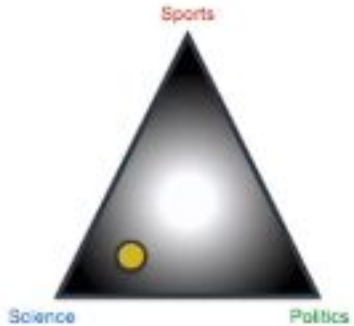
Two Dirichlet distributions



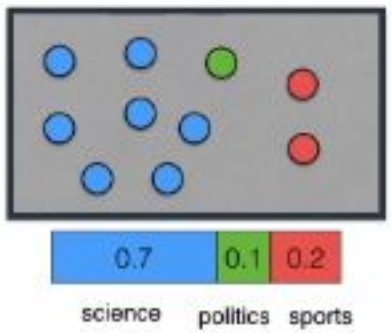
$$\prod_{j=1}^{M} P(\theta_j; \alpha)$$

$$\prod_{i=1}^{K} P(\varphi_i; \beta)$$

$$\prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \qquad P(W_{j,t} \mid \varphi_{Z_{j,t}})$$



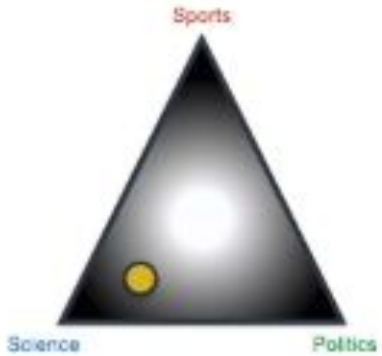
science	politics	sports
0.7	0.1	0.2



Topics

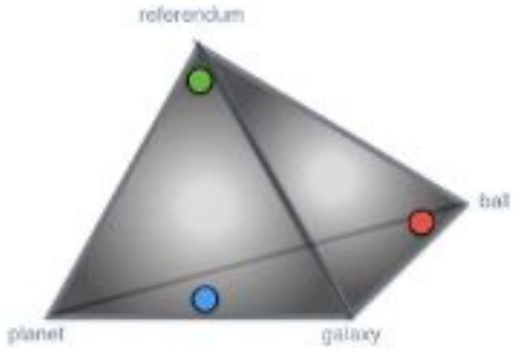
science
science
sports
science
science
politics
sports

$$\prod_{j=1}^{M} P(\theta_j; \alpha)$$



science	politics	sports
0.7	0.1	0.2

$$\prod_{i=1}^{K} P(\varphi_i; \beta)$$

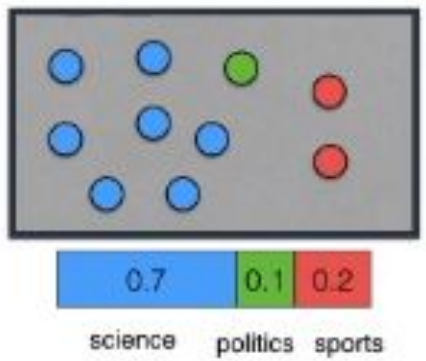


Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

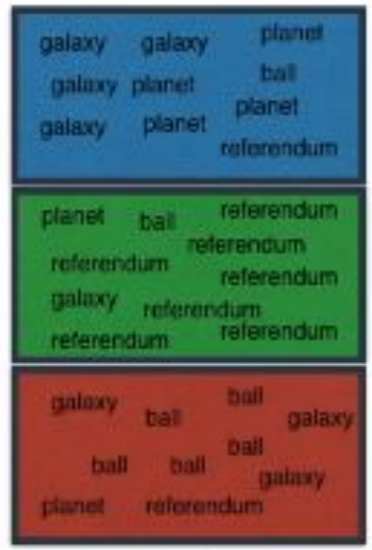
Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

$$\prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)$$



$$P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

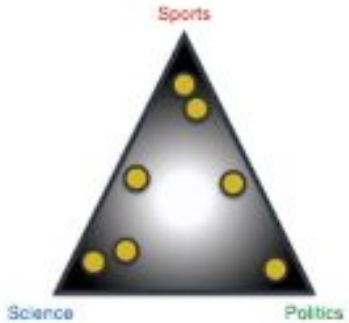


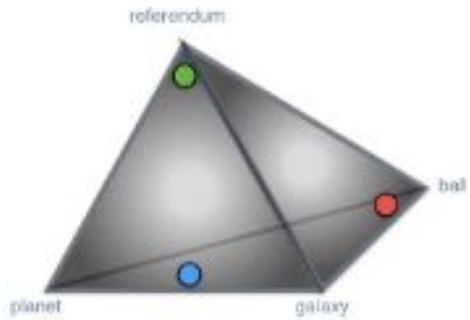
Topics		Words
science	→	planet
science	→	galaxy
sports	→	ball
science	→	planet
science	→	galaxy
politics	→	referendum
sports	→	galaxy

$$\prod_{j=1}^{M} P(\theta_j; \alpha)$$

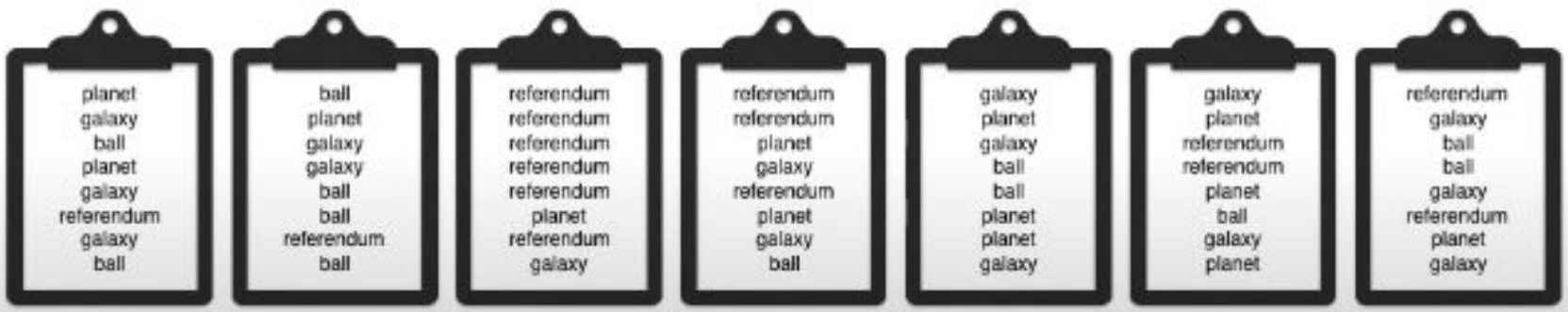
$$\prod_{i=1}^{K} P(\varphi_i; \beta)$$

$$\prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \qquad P(W_{j,t} \mid \varphi_{Z_{j,t}})$$





P(same articles) = low

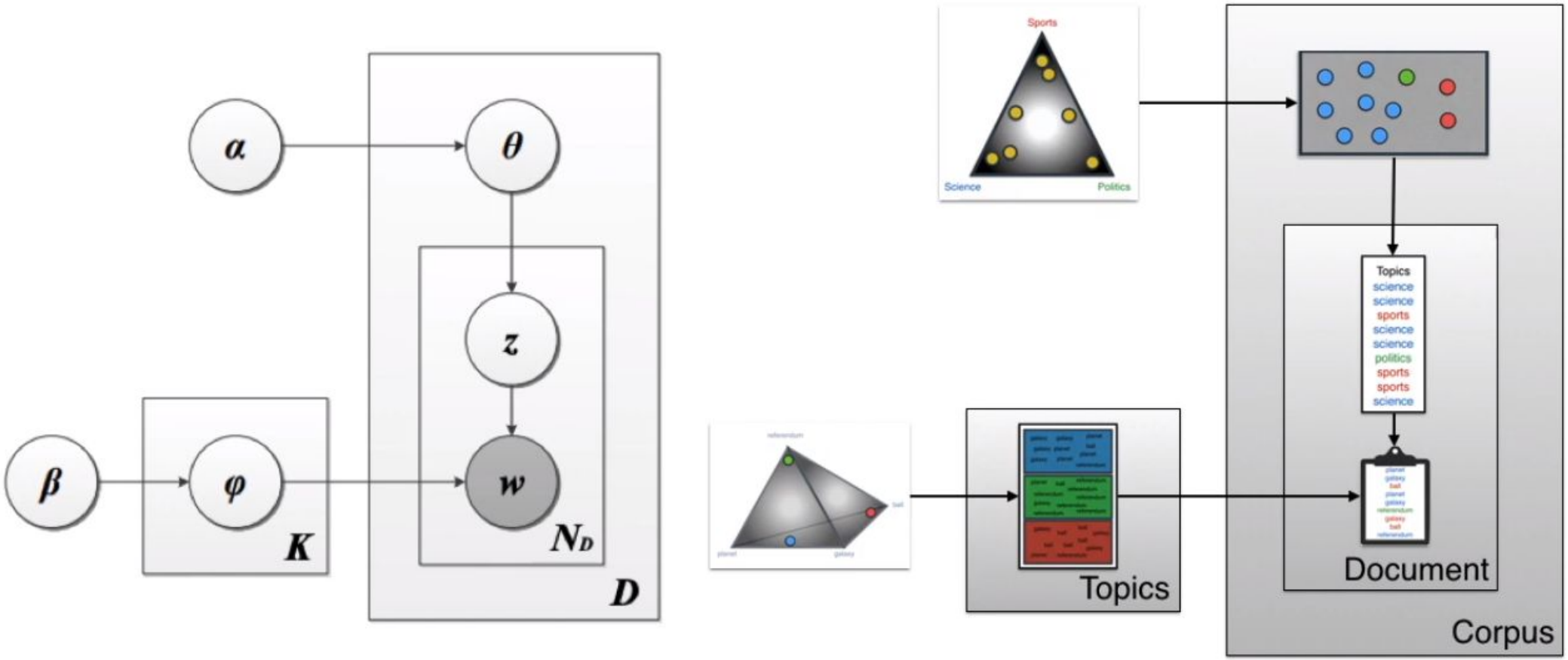


$$\prod_{j=1}^{M} P(\theta_j; \alpha) \qquad \prod_{i=1}^{K} P(\varphi_i; \beta) \qquad \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \qquad P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

P(same articles) = low

P(same articles) = very low

Latent Dirichlet Allocation



Training LDA? -> Gibbs Sampling

- Goal: You want LDA to learn the topic mix in each document, and the word mix in each topic
- Choose the number of topics you think there are in your corpus
  Example: K = 2
- Randomly assign each word in each document to one of 2 topics
  Example: The word 'banana' in Document #1 is randomly assigned to Topic B (animal-like topic)
- Go through every word and its topic assignment in each document. Look at (1) how often the topic occurs in the document and (2) how often the word occurs in the topic overall. Based on this info, assign the word a new topic.
  Example: It looks like (1) animals don't occur often in Document #1 and (2) 'banana' doesn't occur much in Topic B, so the word 'banana' probably should be assigned to Topic A instead
- Go through multiple iterations of this. Eventually, the topics will start making sense - interpret them.