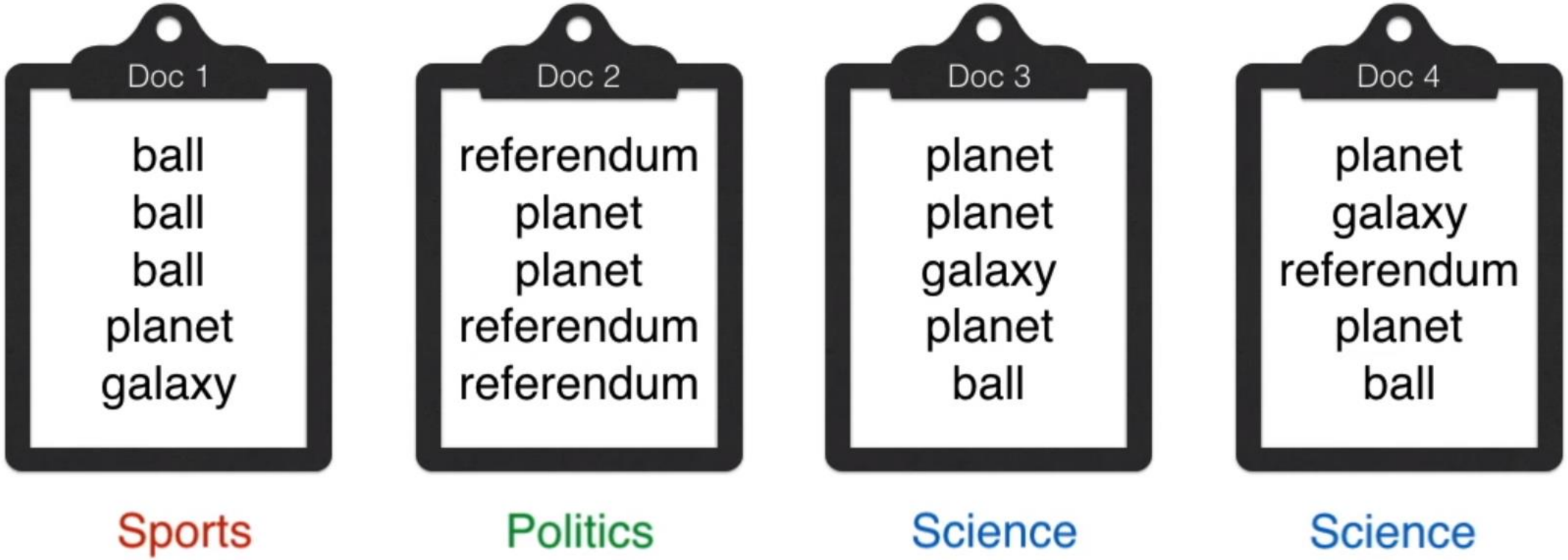
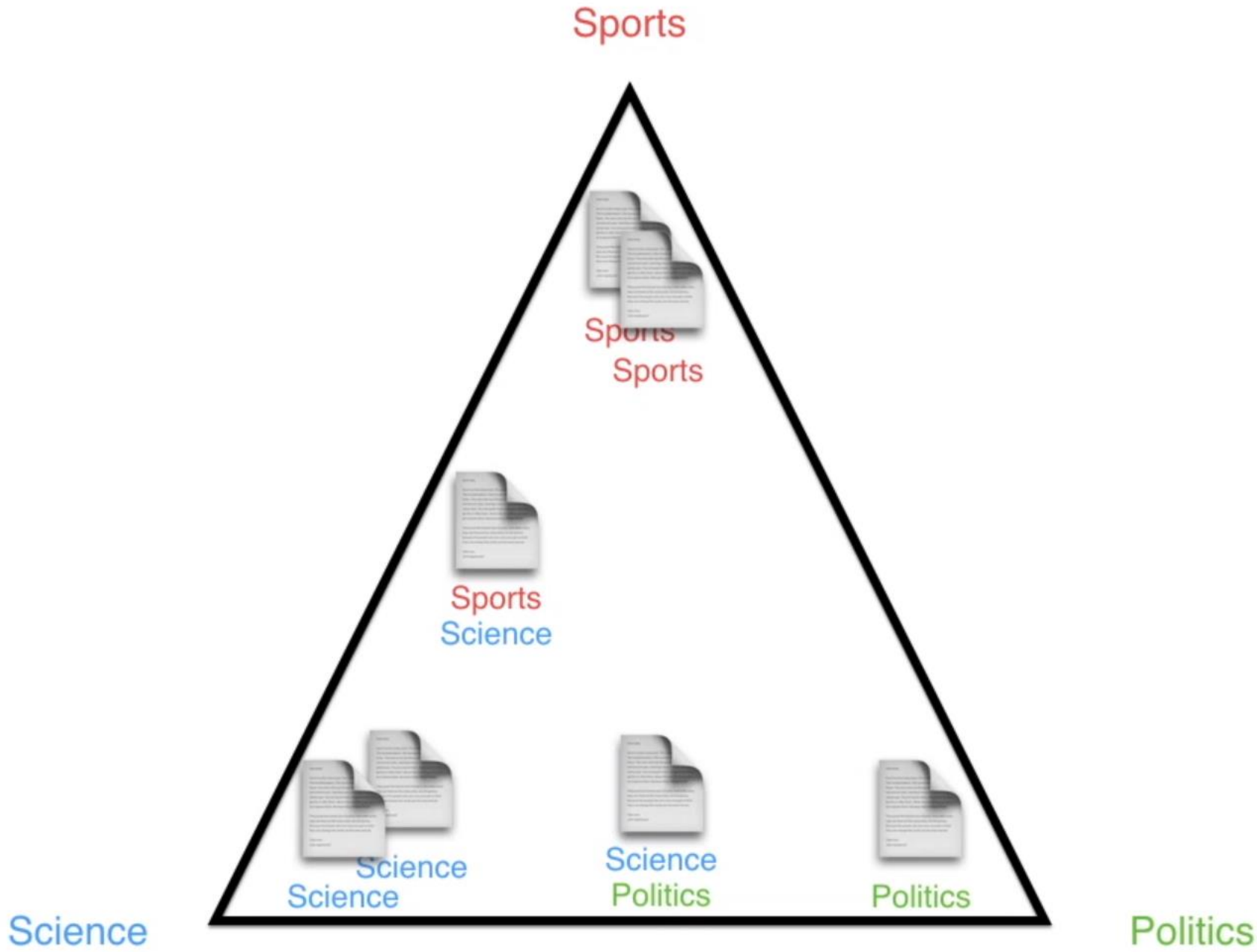
The problem


Science
Politics
Science
Sports
Sports
Science
Politics
Sports
Science

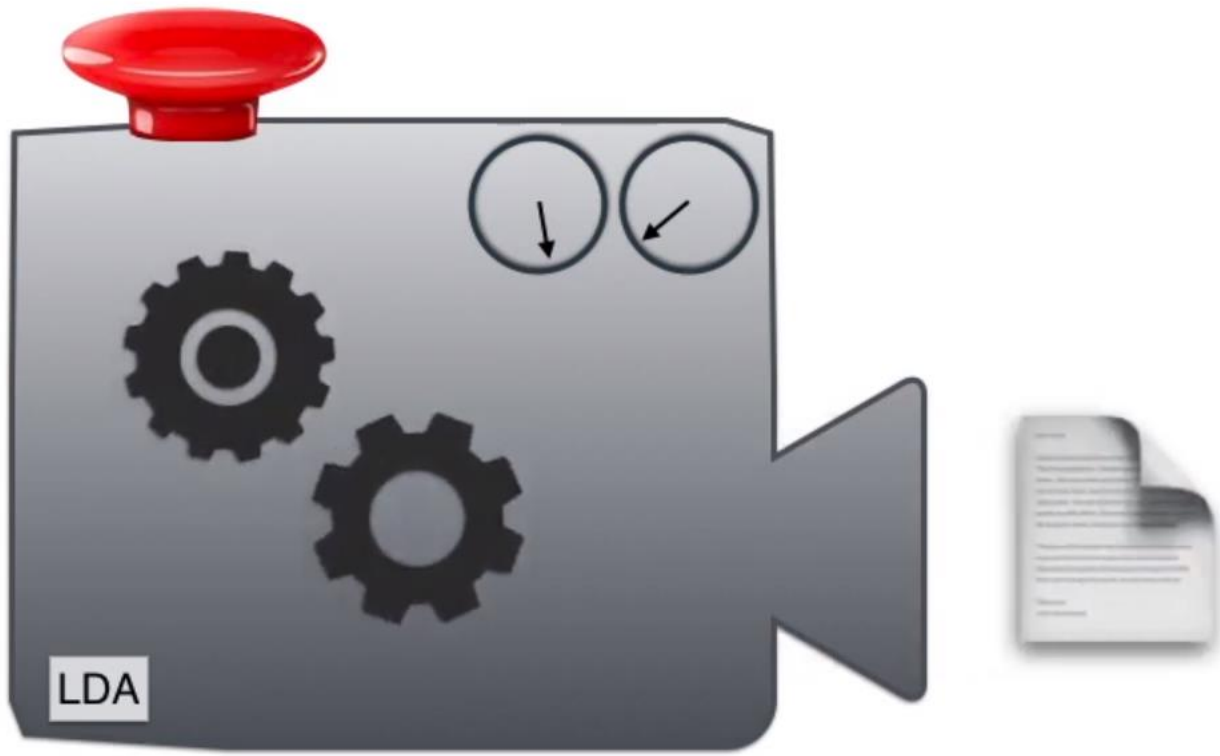


Science Politics Sports

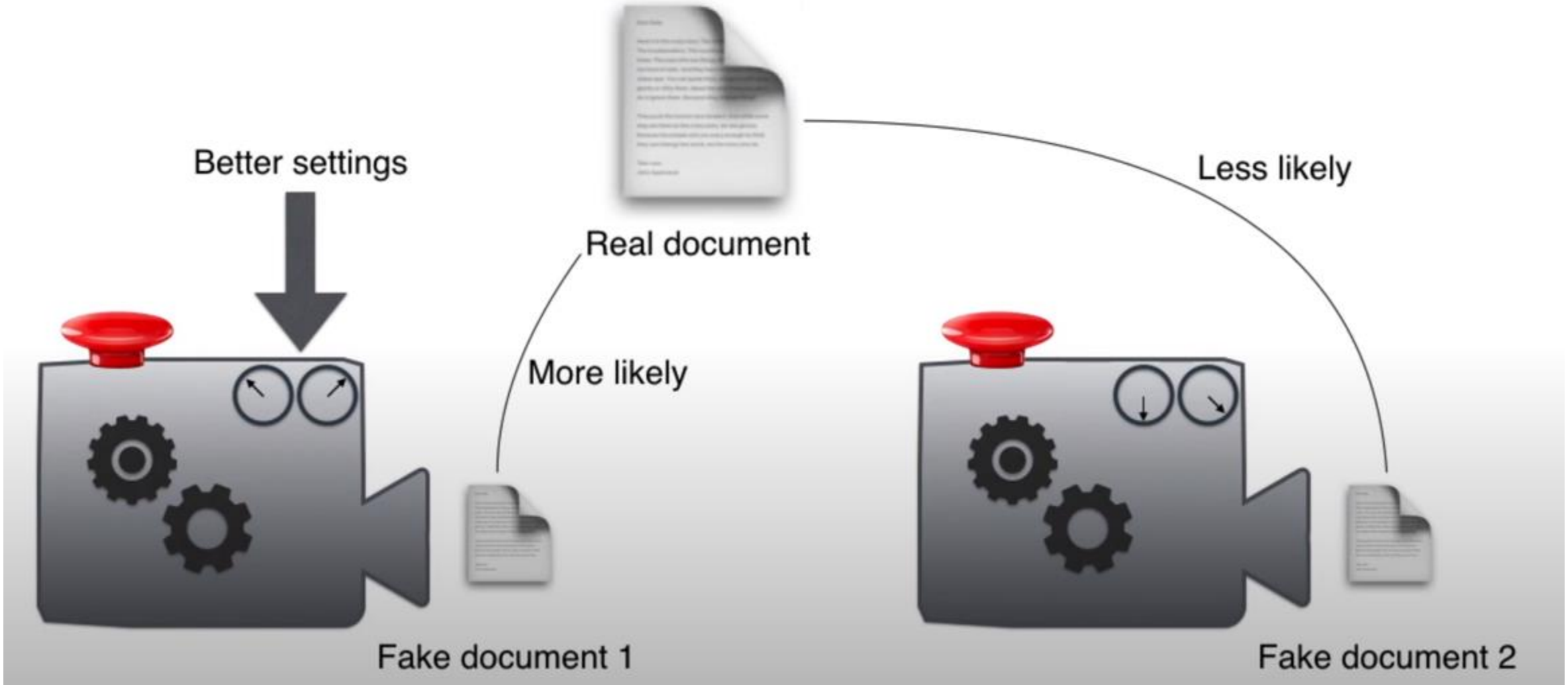
LDA


Sports
Sports
Sports
Sports
Science
Science
Science
Science
Politics
Politics
Science
Politics

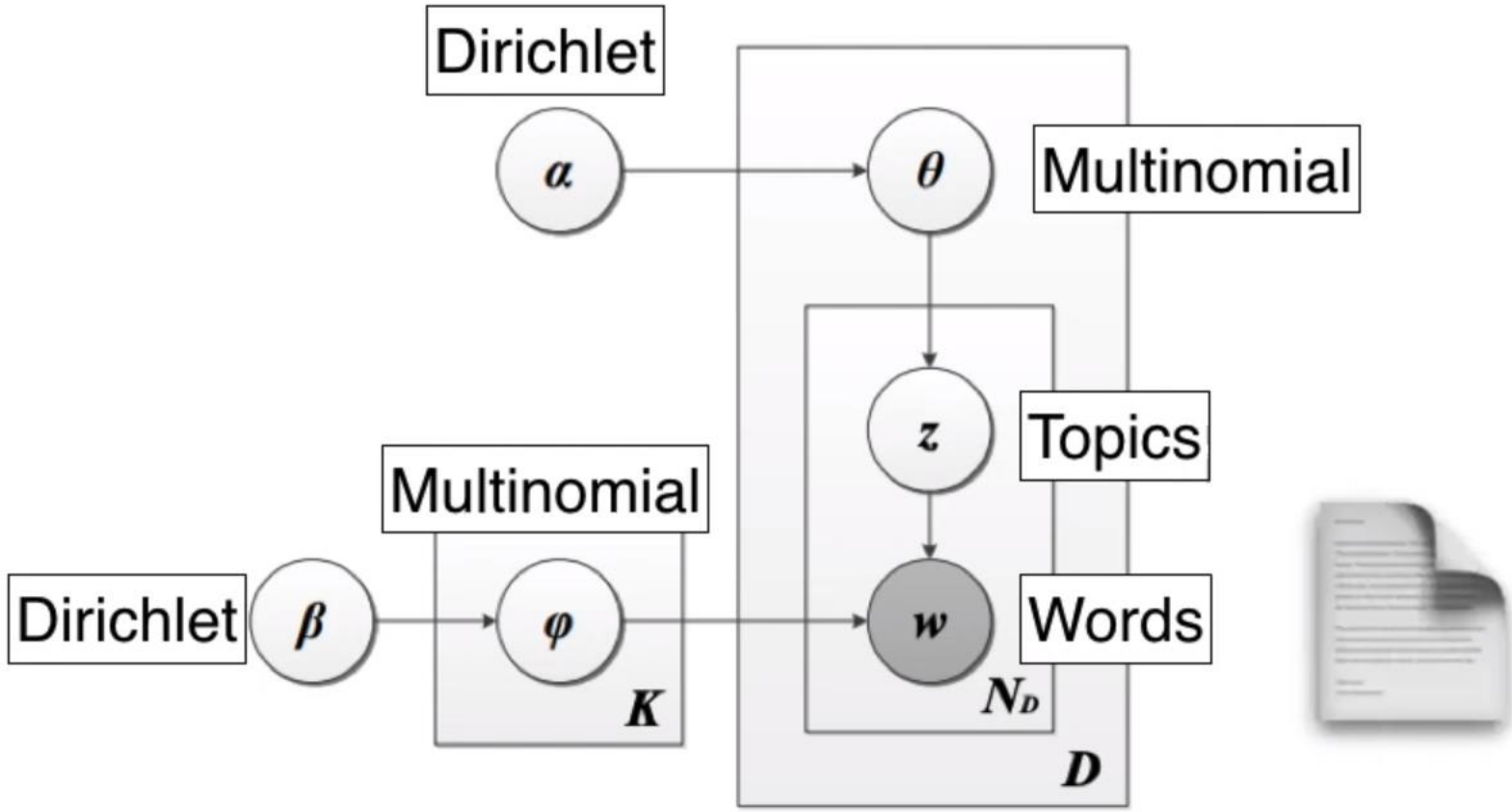
Machine that generates documents



Best settings on the machine

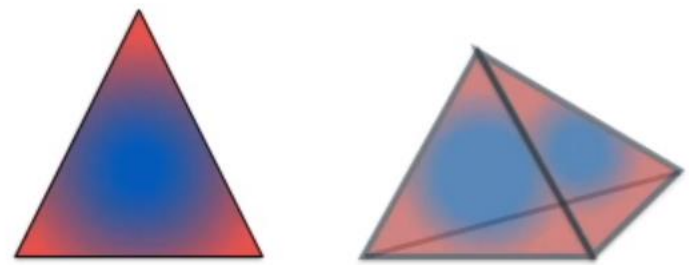


Blueprint for the LDA machine

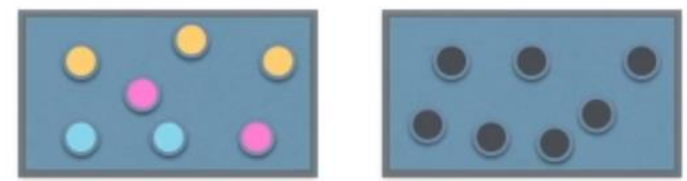


Probability of a document

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



Topics Words

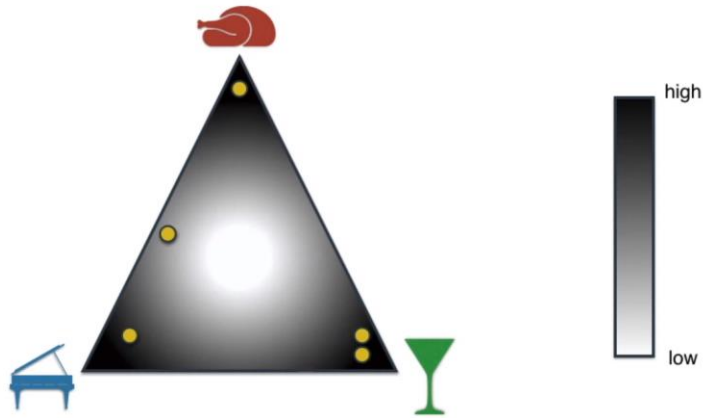


Topics Words

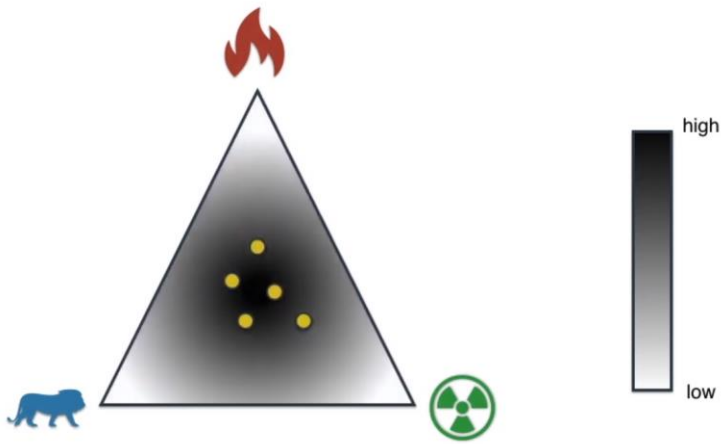
Dirichlet Distributions

Multinomial Distributions

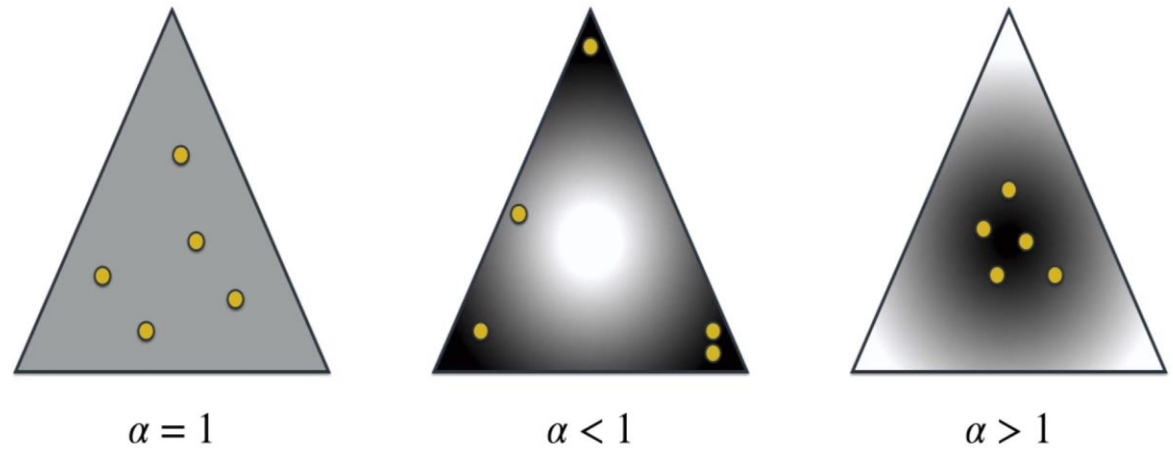
At a party



At a party

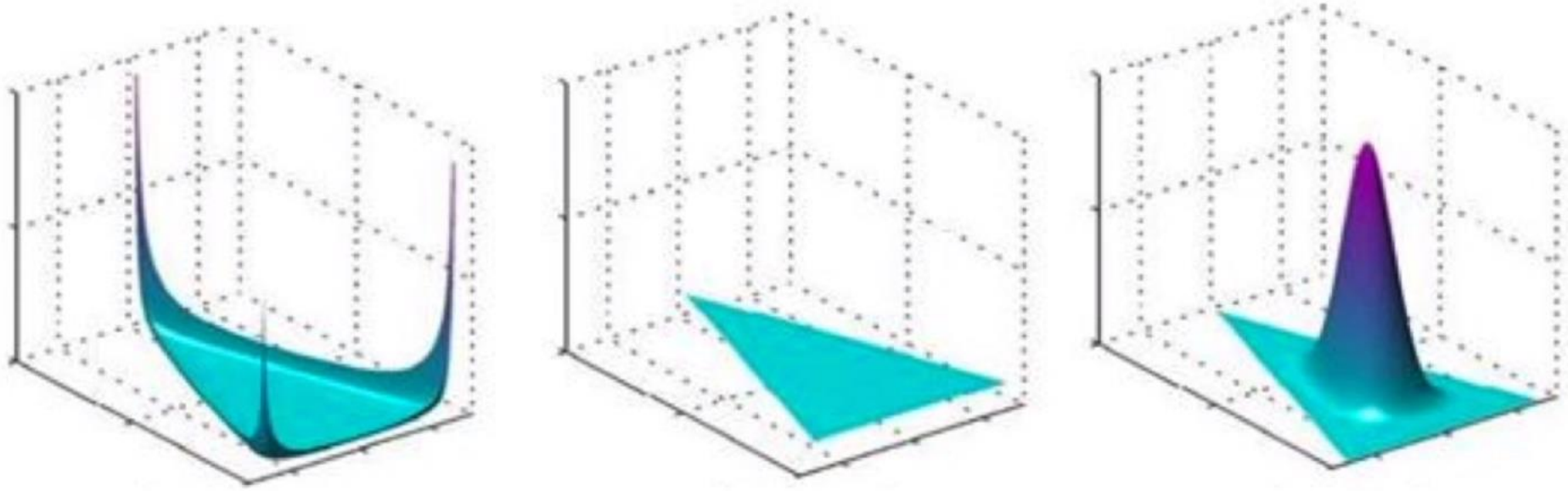


Dirichlet Distributions



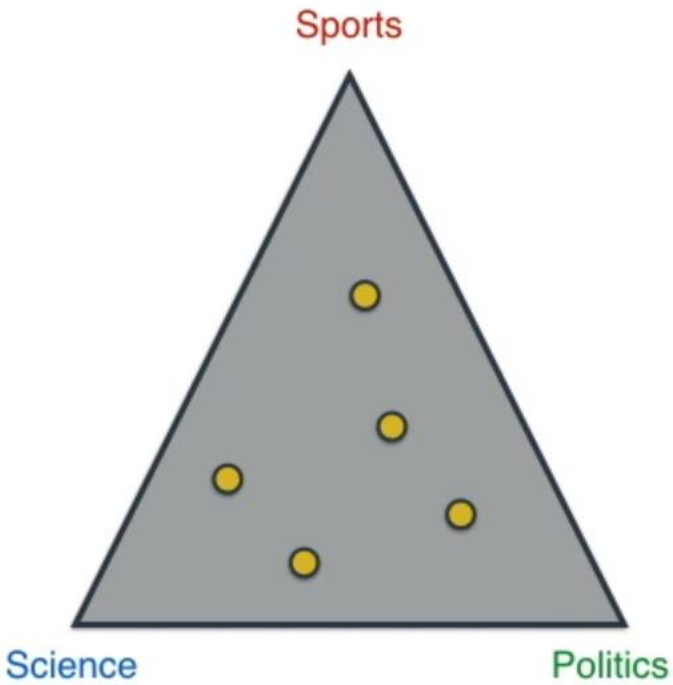
Dirichlet Distributions

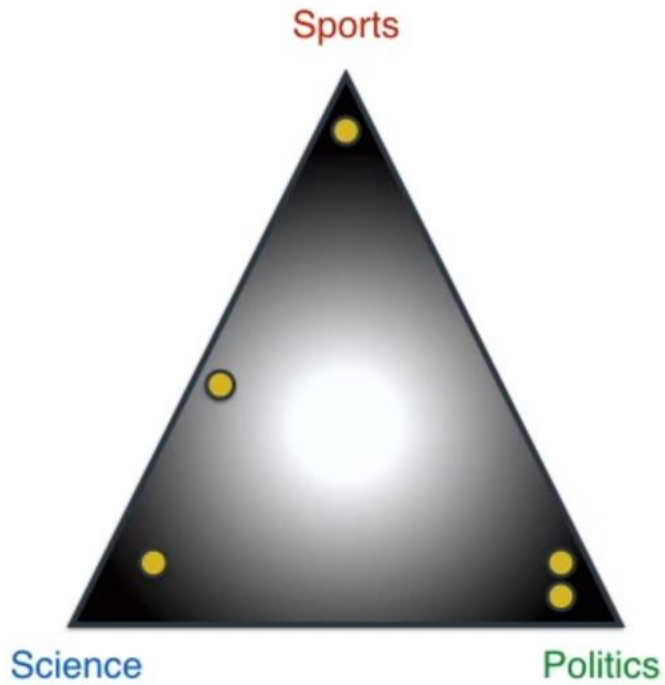
$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

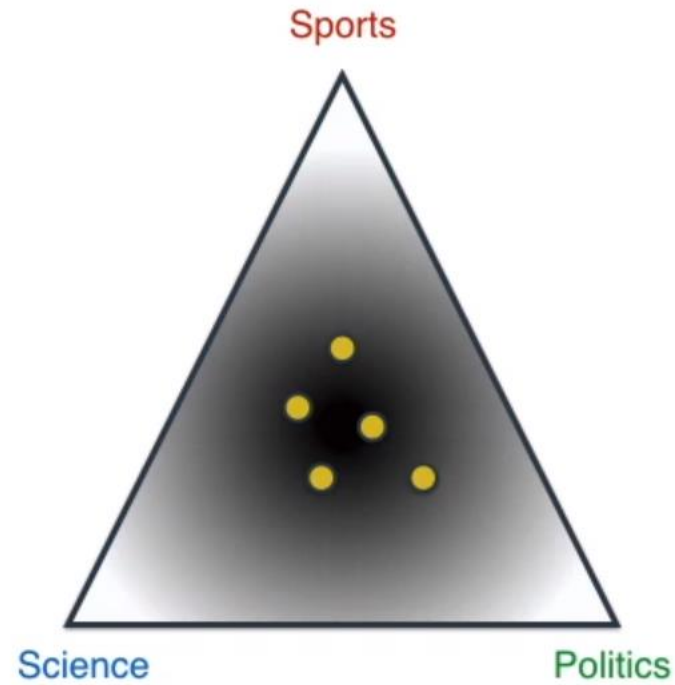


0.7, 0.7, 0.7 | 1, 1, 1 | 5, 5, 5

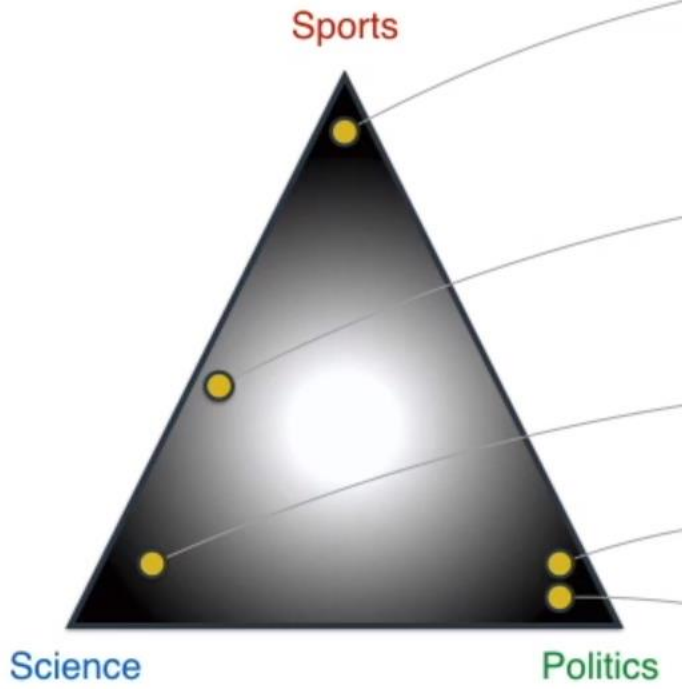
Quiz: Which one for topics?







Quiz: Which one for topics?



science	politics	sports
0.07	0.03	0.9

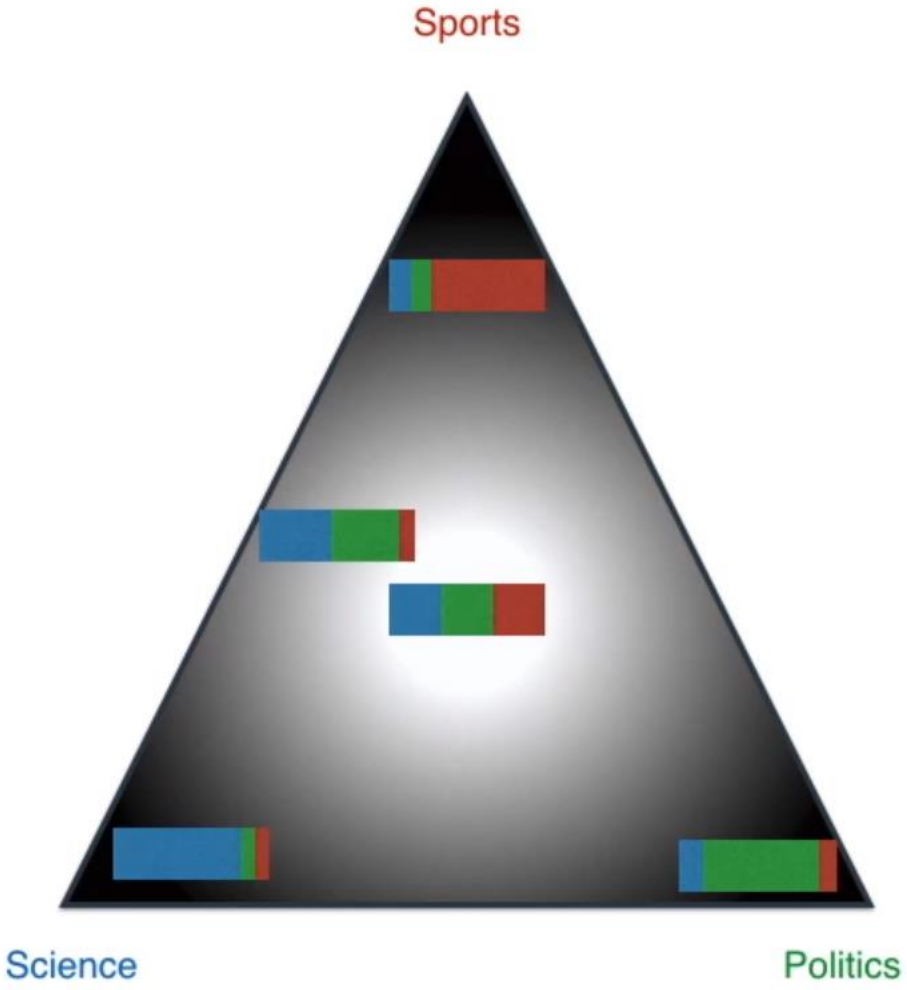
science	politics	sports
0.45	0.05	0.5

science	politics	sports
0.8	0.1	0.1

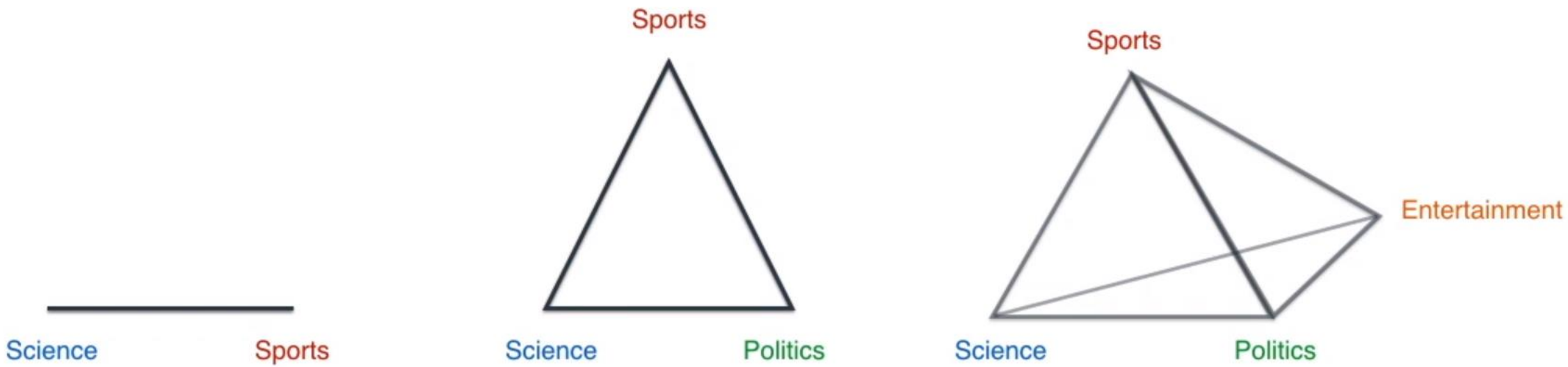
science	politics	sports
0.05	0.8	0.15

science	politics	sports
0.05	0.9	0.05

A distribution of distributions



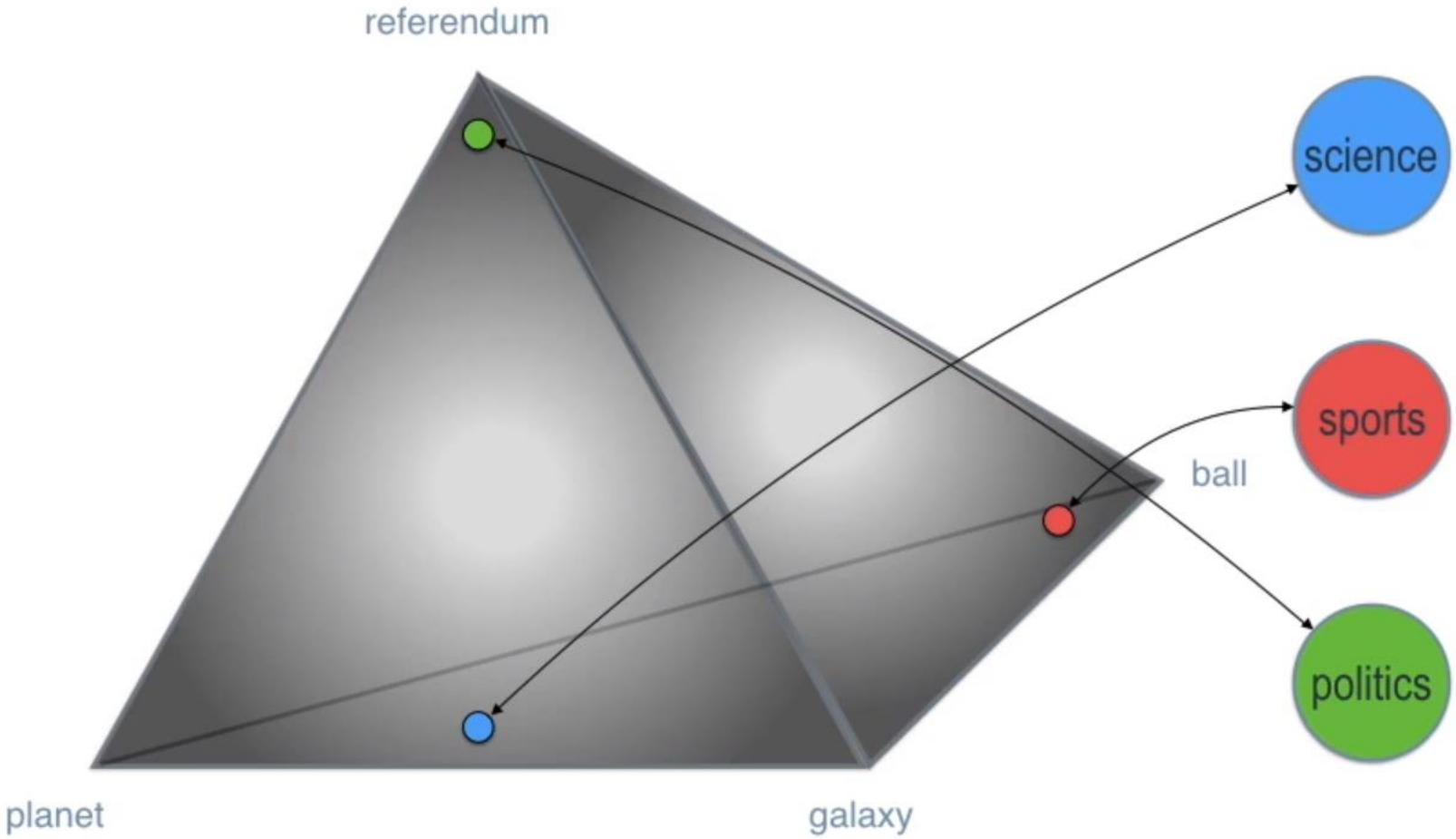
More topics? More dimensions



2 topics 3 topics 4 topics …

n-dimensional simplex

Quiz: Where to put the topics?

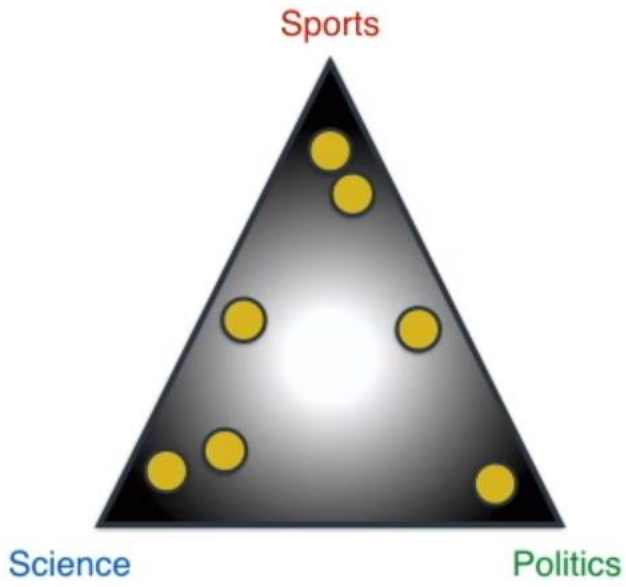


Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

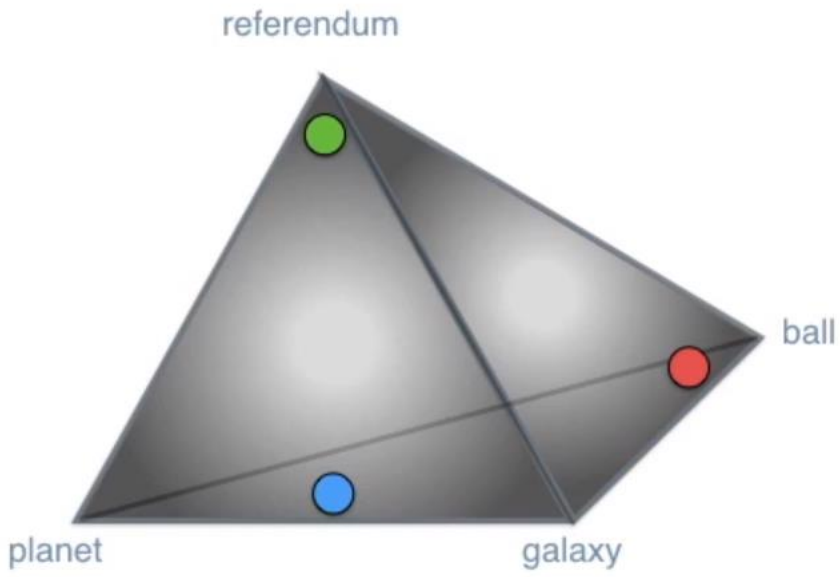
Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Two Dirichlet distributions

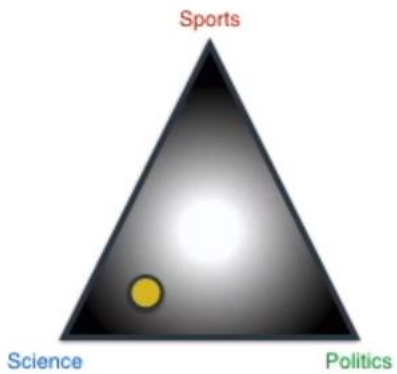


Documents-Topics



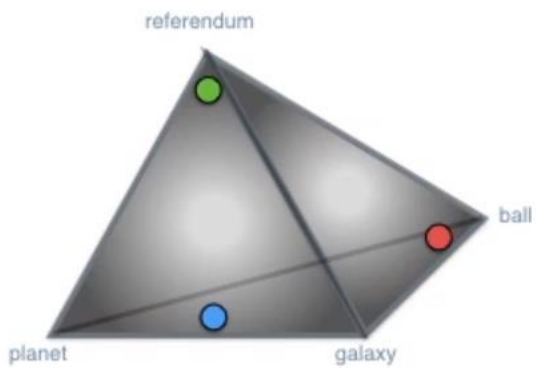
Topics-Words

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



science	politics	sports
0.7	0.1	0.2

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

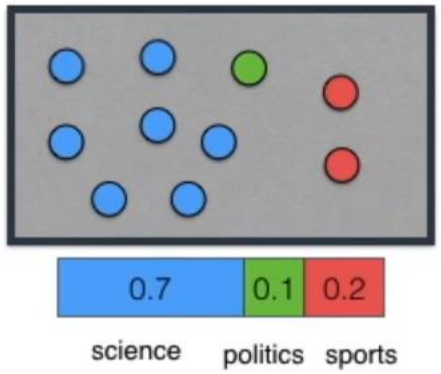


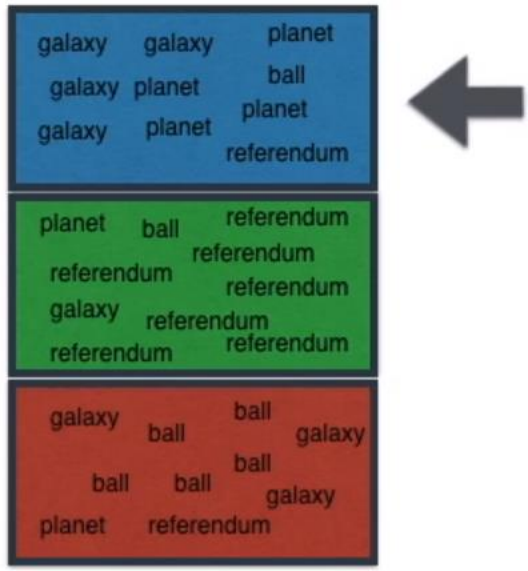
Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

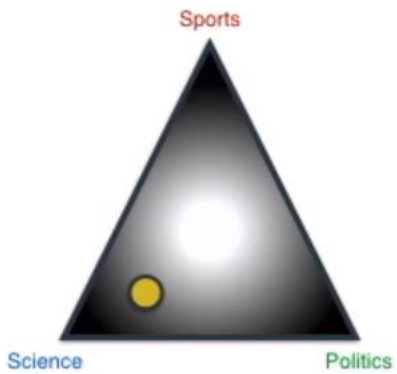
$$\prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \qquad P(W_{j,t} \mid \varphi_{Z_{j,t}})$$





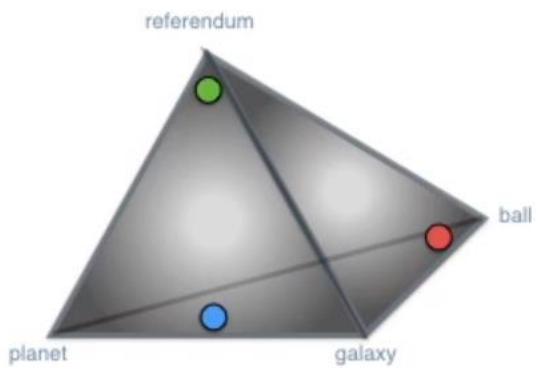
Topics		Words
science	→	planet
science	→	galaxy
sports	→	ball
science	→	planet
science	→	galaxy
politics	→	referendum
sports	→	galaxy
sports	→	ball
science	→	referendum

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



science	politics	sports
0.7	0.1	0.2

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

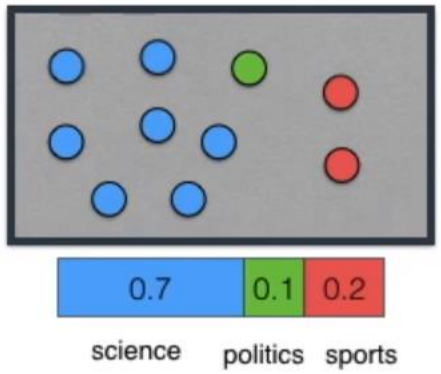


Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

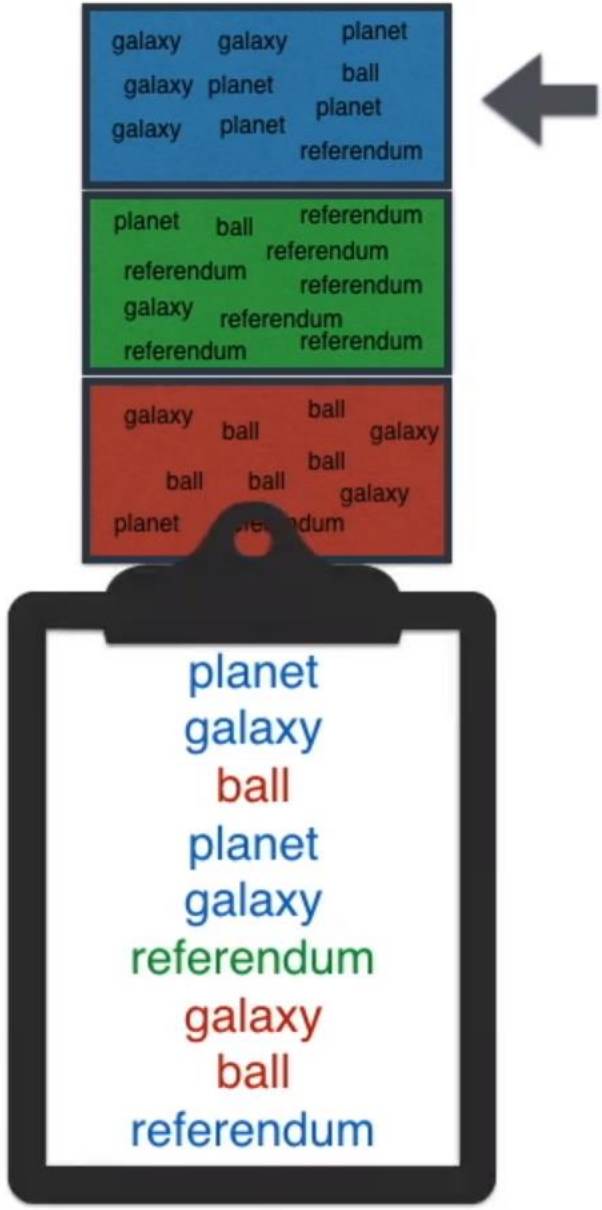
$$\prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)$$



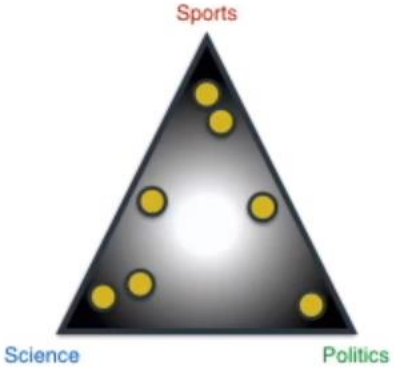
Topics

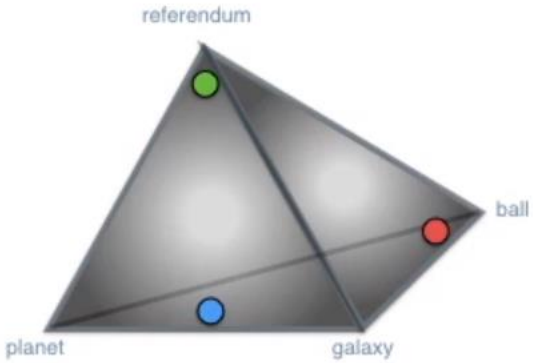
science
science
sports
science
science
politics
sports
sports
science

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

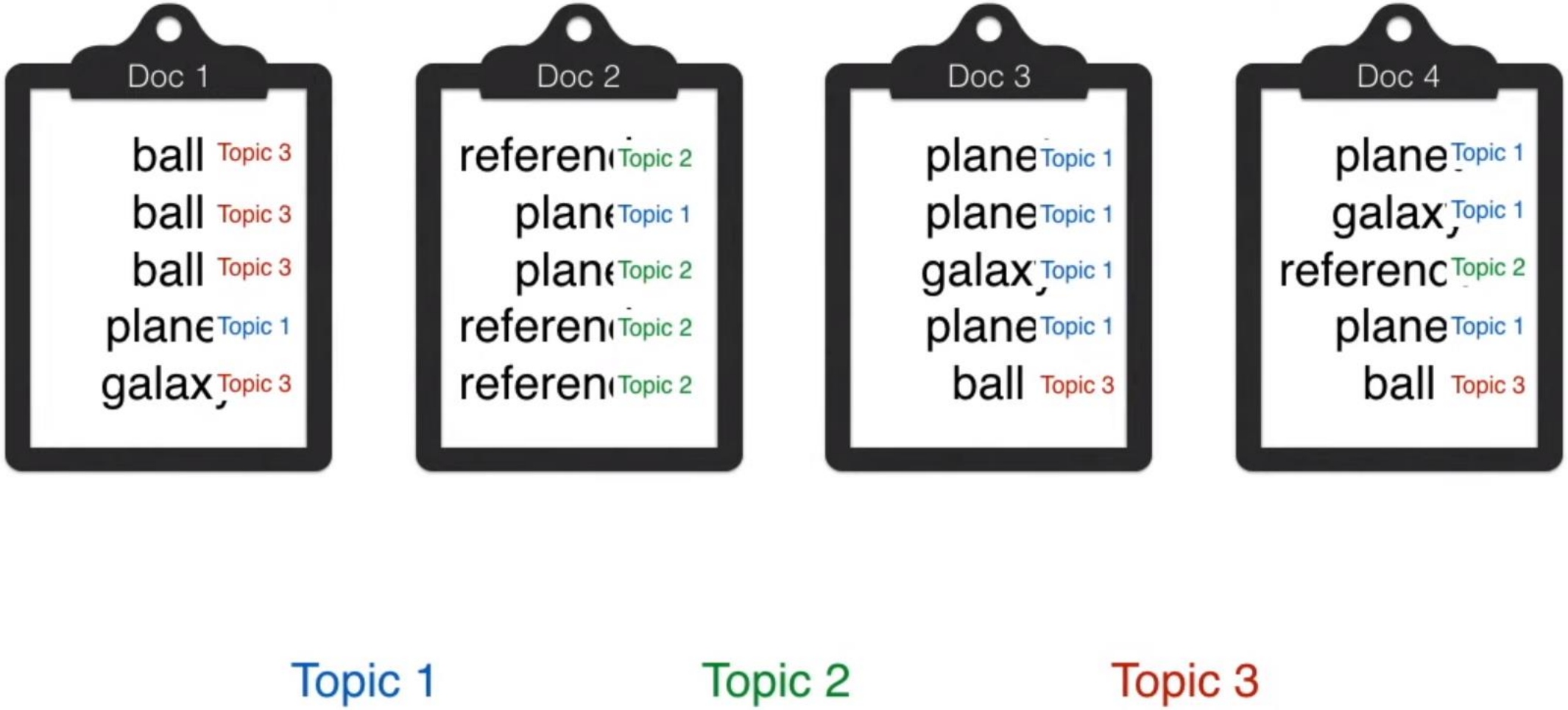


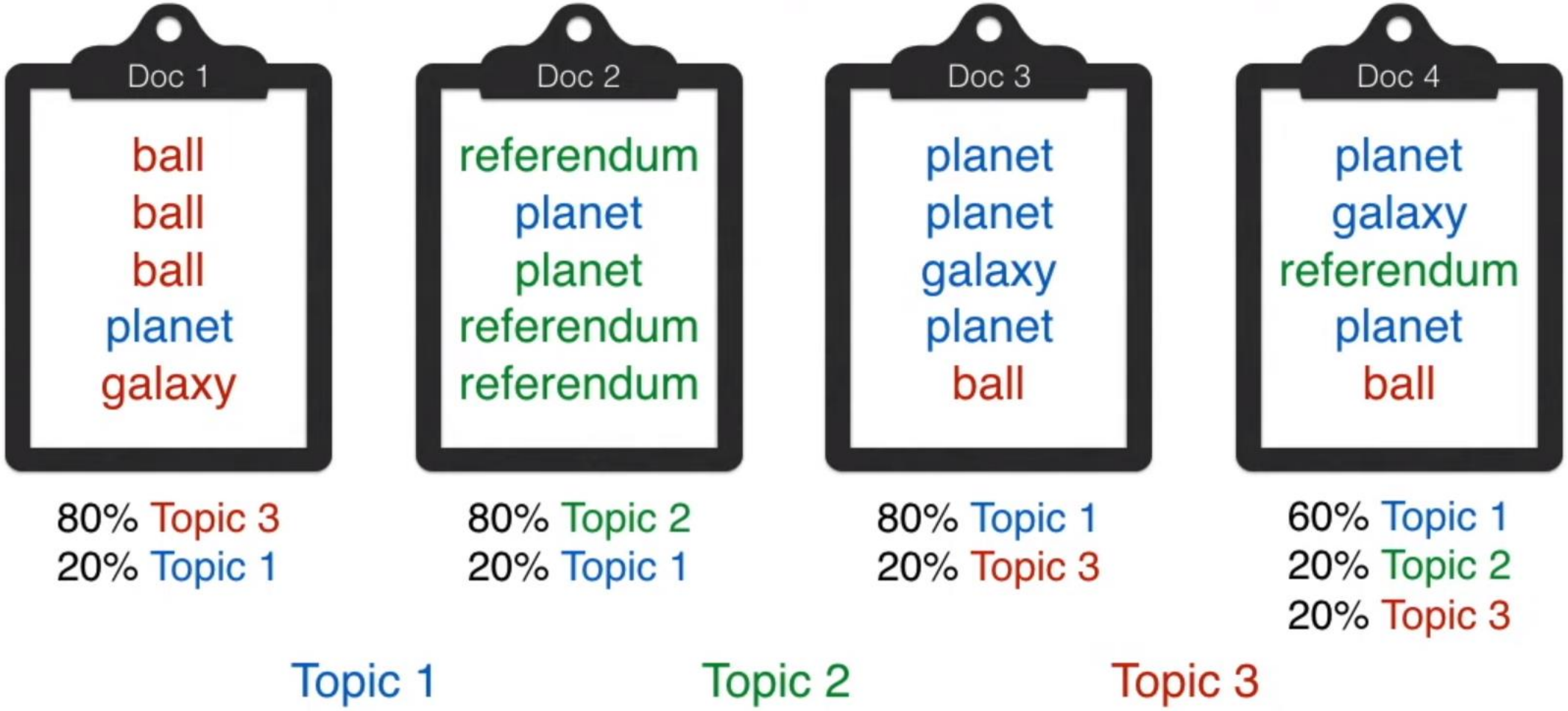
$$\prod_{j=1}^{M} P(\theta_j;\alpha) \qquad \prod_{i=1}^{K} P(\varphi_i;\beta) \qquad \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \qquad P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

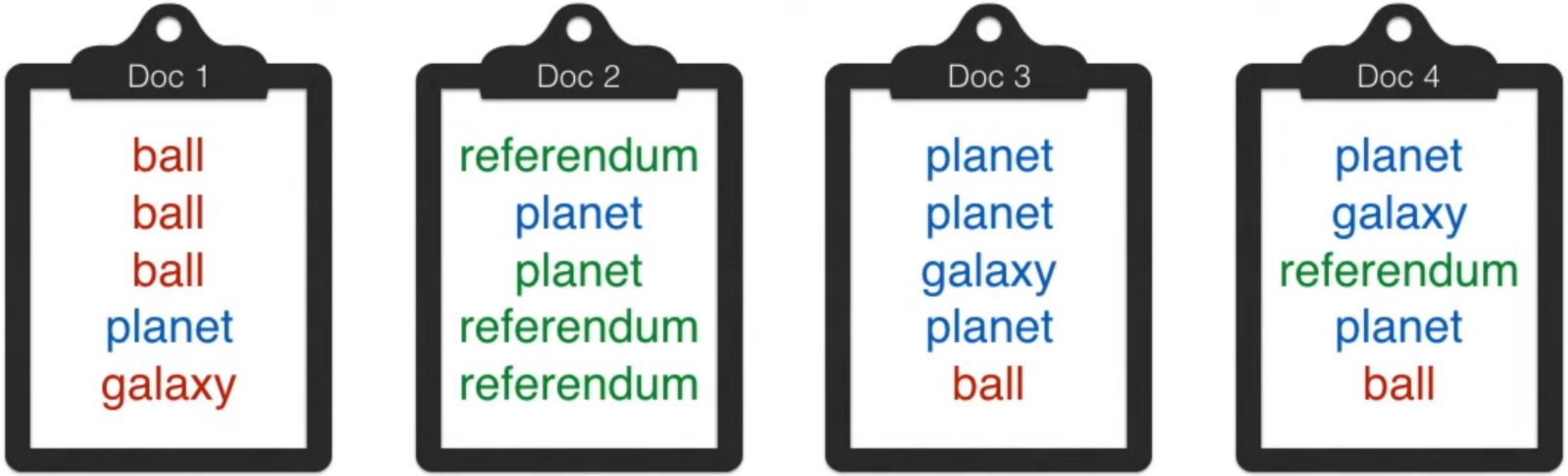




Gibbs Sampling

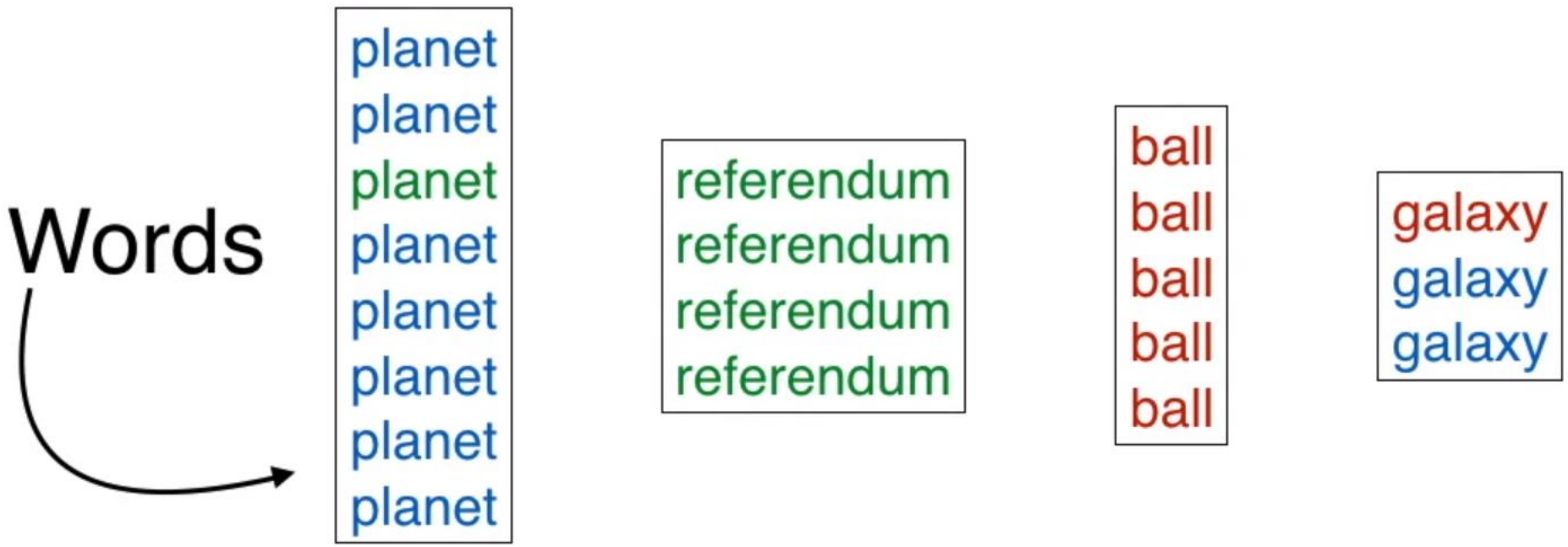

Doc 1
ball Topic 3
ball Topic 3
ball Topic 3
plane Topic 1
galax Topic 3
Doc 2
referen Topic 2
plane Topic 1
plane Topic 2
referen Topic 2
referen Topic 2
Doc 3
plane Topic 1
plane Topic 1
galax Topic 1
plane Topic 1
ball Topic 3
Doc 4
plane Topic 1
galax Topic 1
referenc Topic 2
plane Topic 1
ball Topic 3
Topic 1
Topic 2
Topic 3

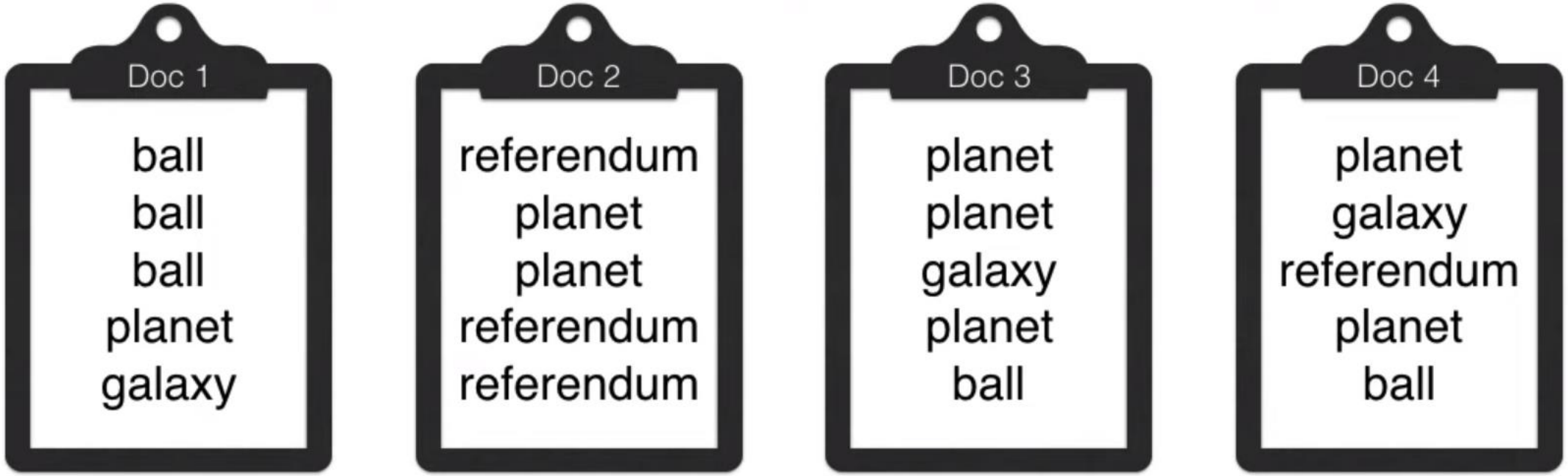

Doc 1
ball
ball
ball
planet
galaxy
Doc 2
referendum
planet
planet
referendum
referendum
Doc 3
planet
planet
galaxy
planet
ball
Doc 4
planet
galaxy
referendum
planet
ball
80% Topic 3
20% Topic 1
80% Topic 2
20% Topic 1
80% Topic 1
20% Topic 3
60% Topic 1
20% Topic 2
20% Topic 3
Topic 1
Topic 2
Topic 3



Property 1:
Articles are as monochromatic as possible

Property 2:
Words are as monochromatic as possible





Goal: Color each word with blue, green, red

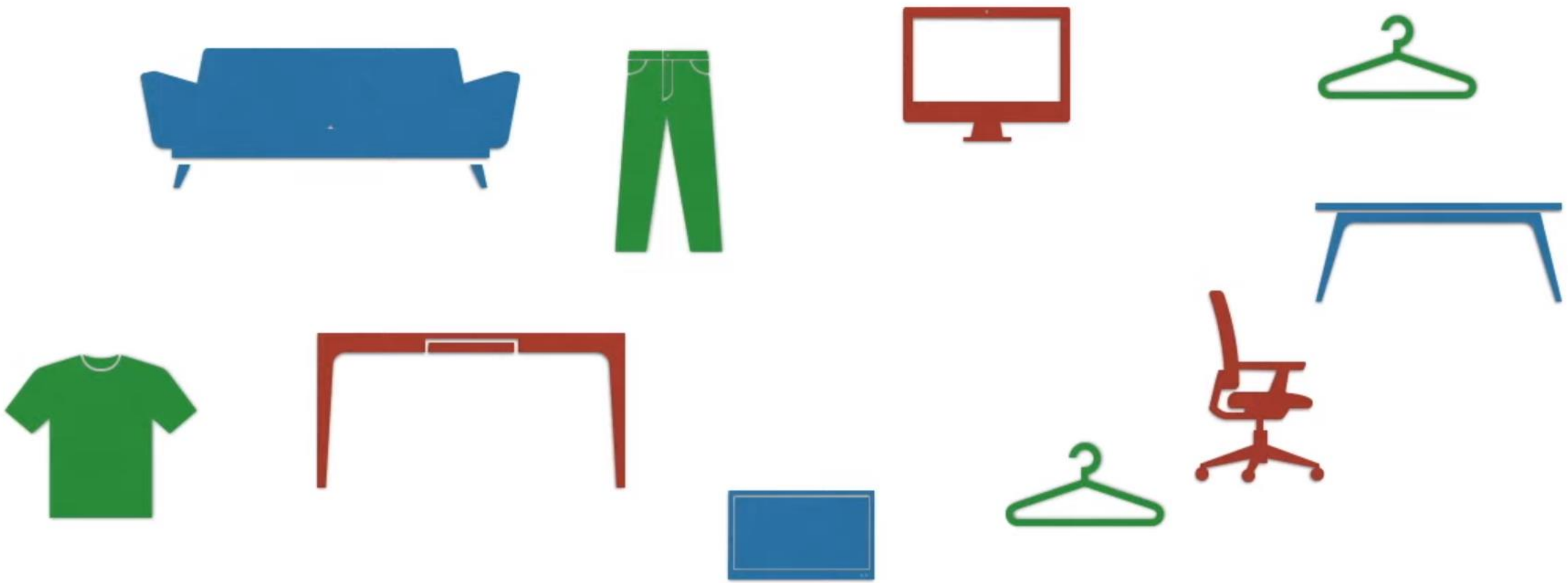
1. Each article is as monochromatic as possible
2. Each word is as monochromatic as possible

Goal: Color each word with blue, green, red

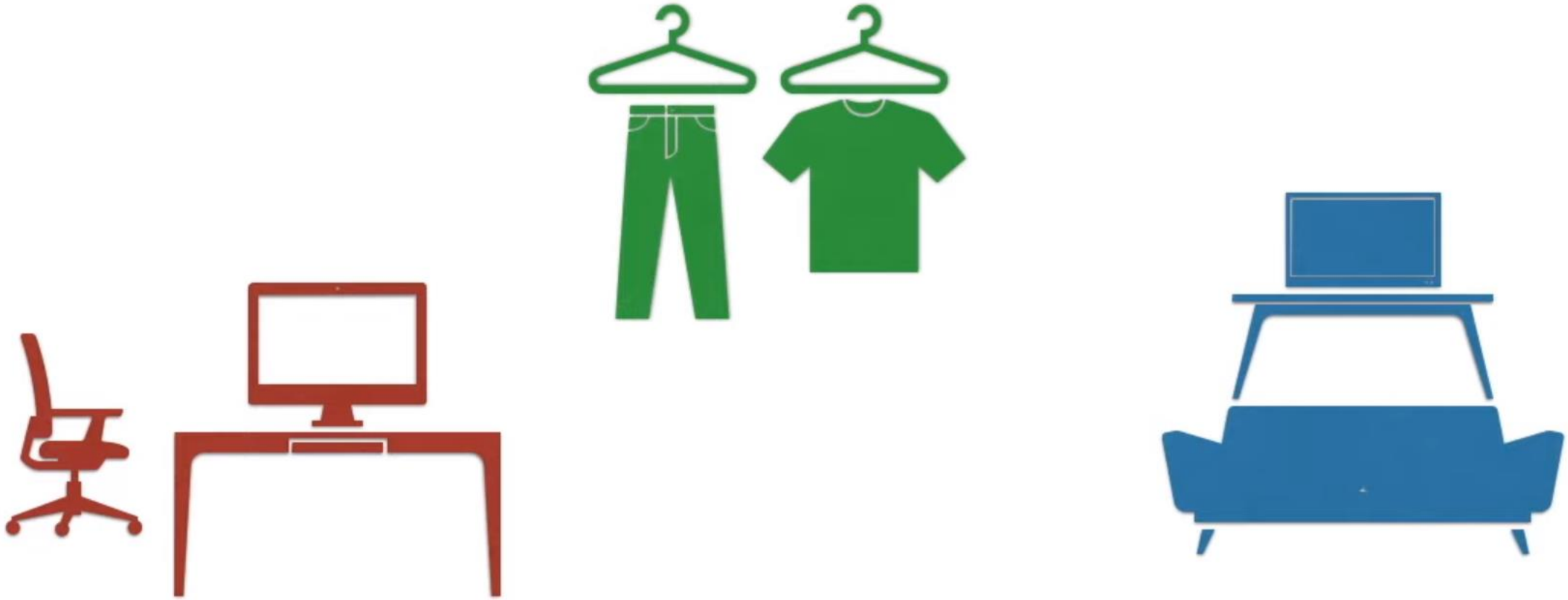
1. Each article is as monochromatic as possible
2. Each word is as monochromatic as possible


Doc 1
khuya
khuya
khuya
alliku
pachamama
Doc 2
yachay
alliku
alliku
yachay
yachay
Doc 3
alliku
alliku
pachamama
alliku
khuya
Doc 4
alliku
pachamama
yachay
alliku
khuya
alliku
alliku
alliku
alliku
alliku
alliku
alliku
alliku
yachay
yachay
yachay
yachay
khuya
khuya
khuya
khuya
khuya
pachamama
pachamama
pachamama

Gibbs sampling



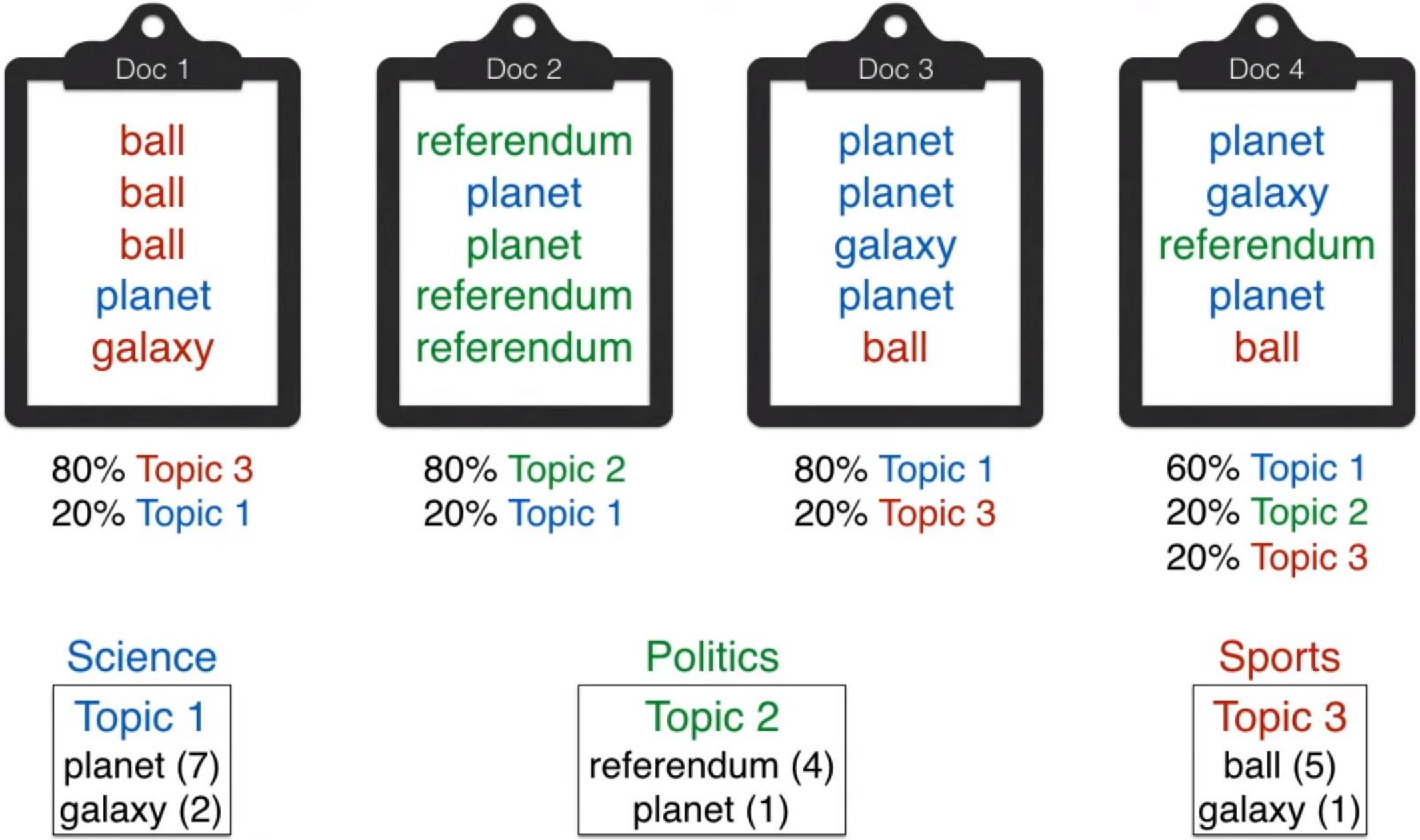
Gibbs sampling



Goal: Color each word with blue, green, red

1. Each article is as monochromatic as possible
2. Each word is as monochromatic as possible

Topic 1	Topic 2	Topic 3
How much is Topic 1 in Doc 1?	How much is Topic 2 in Doc 1?	How much is Topic 3 in Doc 1?
$2+\alpha$	$0+\alpha$	$2+\alpha$
How much is 'ball' in Topic 1?	How much is 'ball' in Topic 2?	How much is 'ball' in Topic 3?
$0+\beta$	$1+\beta$	$3+\beta$


Doc 1
ball
ball
ball
planet
galaxy
Doc 2
referendum
planet
planet
referendum
referendum
Doc 3
planet
planet
galaxy
planet
ball
Doc 4
planet
galaxy
referendum
planet
ball
80% Topic 3
20% Topic 1
80% Topic 2
20% Topic 1
80% Topic 1
20% Topic 3
60% Topic 1
20% Topic 2
20% Topic 3
Science
Topic 1
planet (7)
galaxy (2)
Politics
Topic 2
referendum (4)
planet (1)
Sports
Topic 3
ball (5)
galaxy (1)

Shashank Kapadia

Data Science Manager @Monster Building scalable and operationalized ML solutions for data-driven products. My articles on Medium don't represent my employer.

Follow

816

22

Eye Balling Models

- Top N words
- Topics / Documents

Intrinsic Evaluation Metrics

- Capturing model semantics
- Topics interpretability

Human Judgements

- What is a topic

Extrinsic Evaluation Metrics/Evaluation at task

- Is model good at performing predefined tasks, such as classification