Text Analysis

Topic Models

Dirk Hovy

dirk.hovy@unibocconi.it

@dirk_hovy

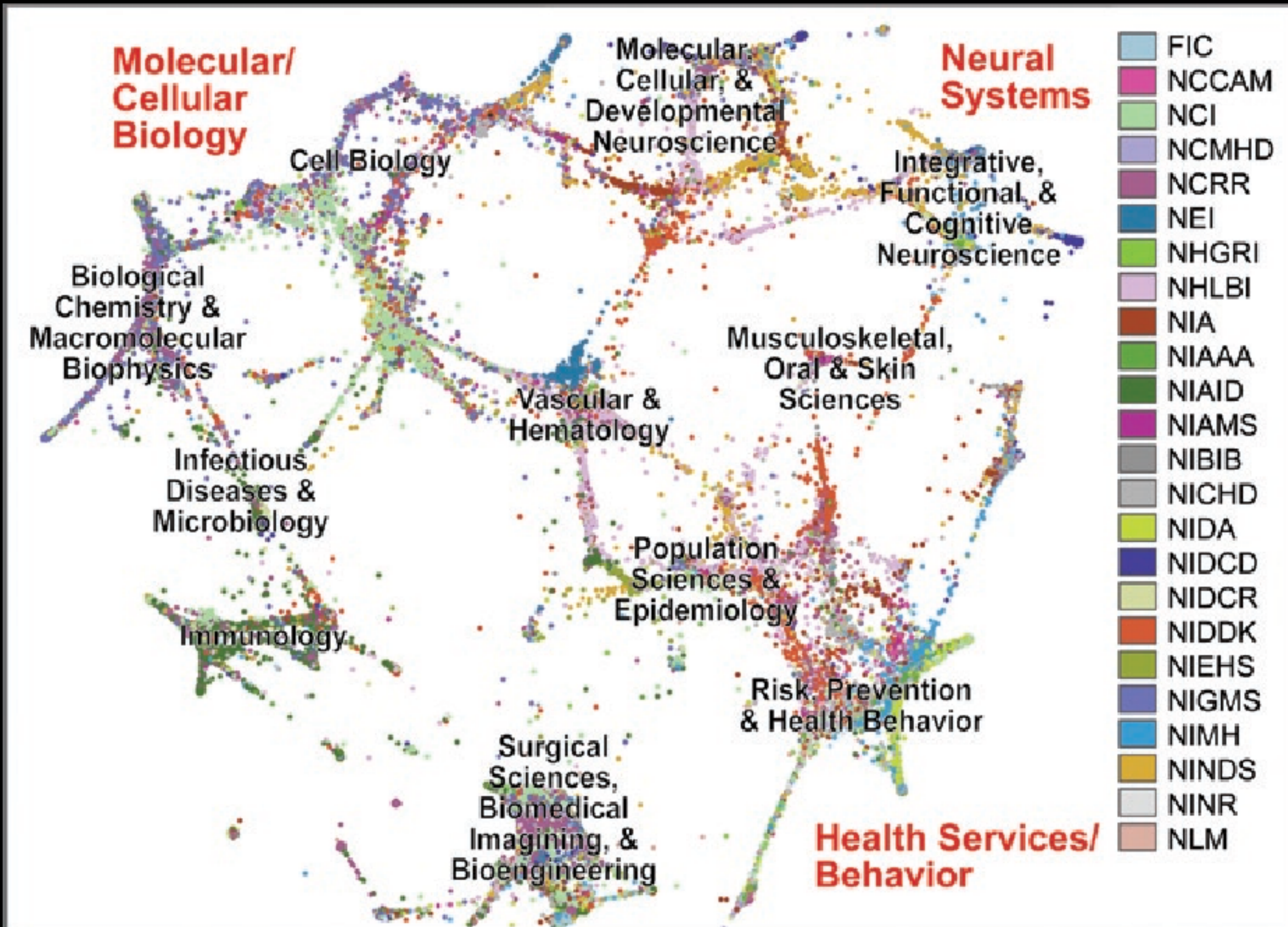
Bocconi

Goals for Today

- Understand what information **topic models** can and can not provide
- Learn about the **Latent Dirichlet Allocation (LDA)** model
- Understand the **parameters** influencing the output
- Learn about **evaluation** criteria

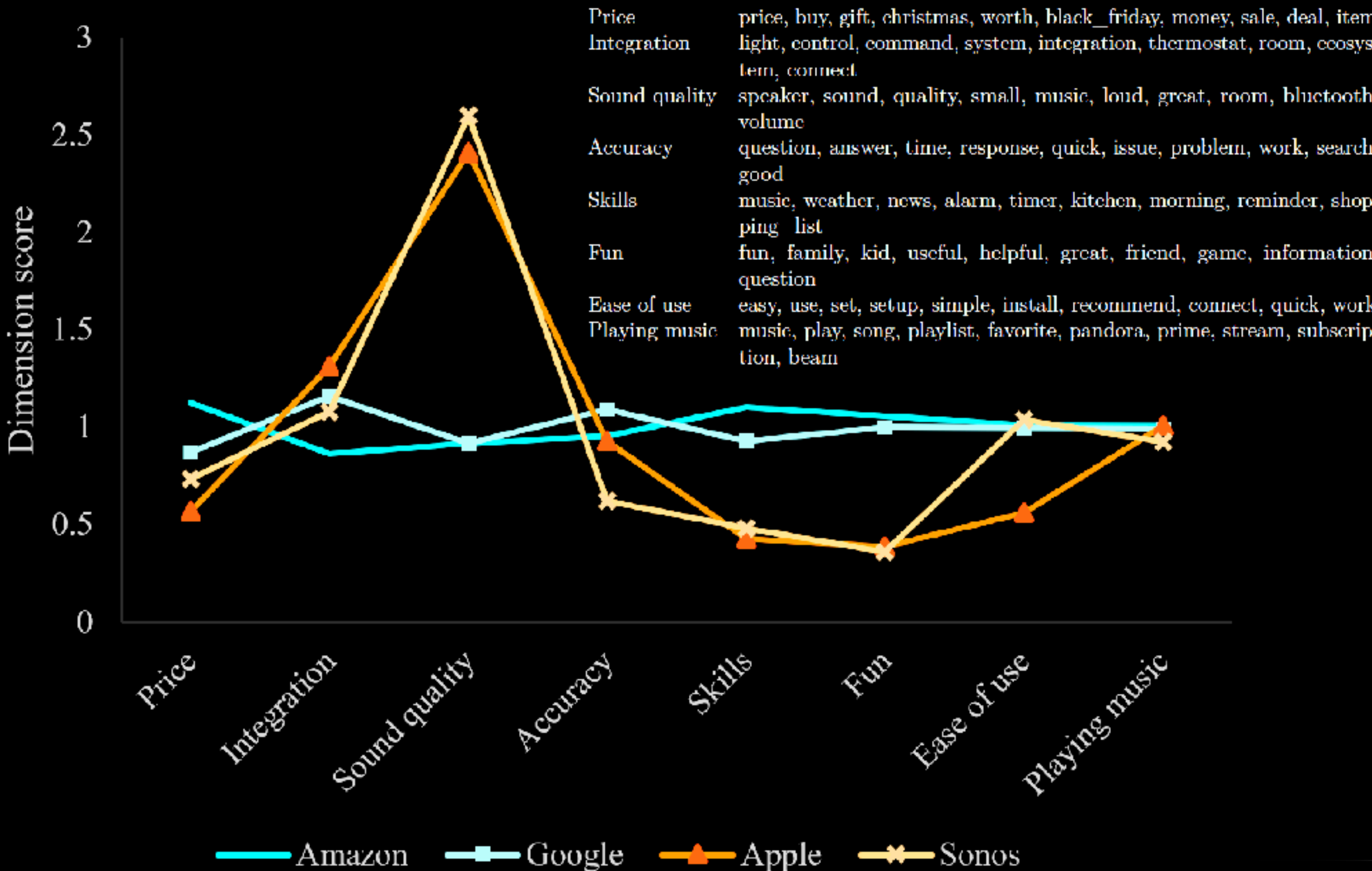
Talley et al., 2011

What Gets Funded?



Hanh & Hovy (2019)

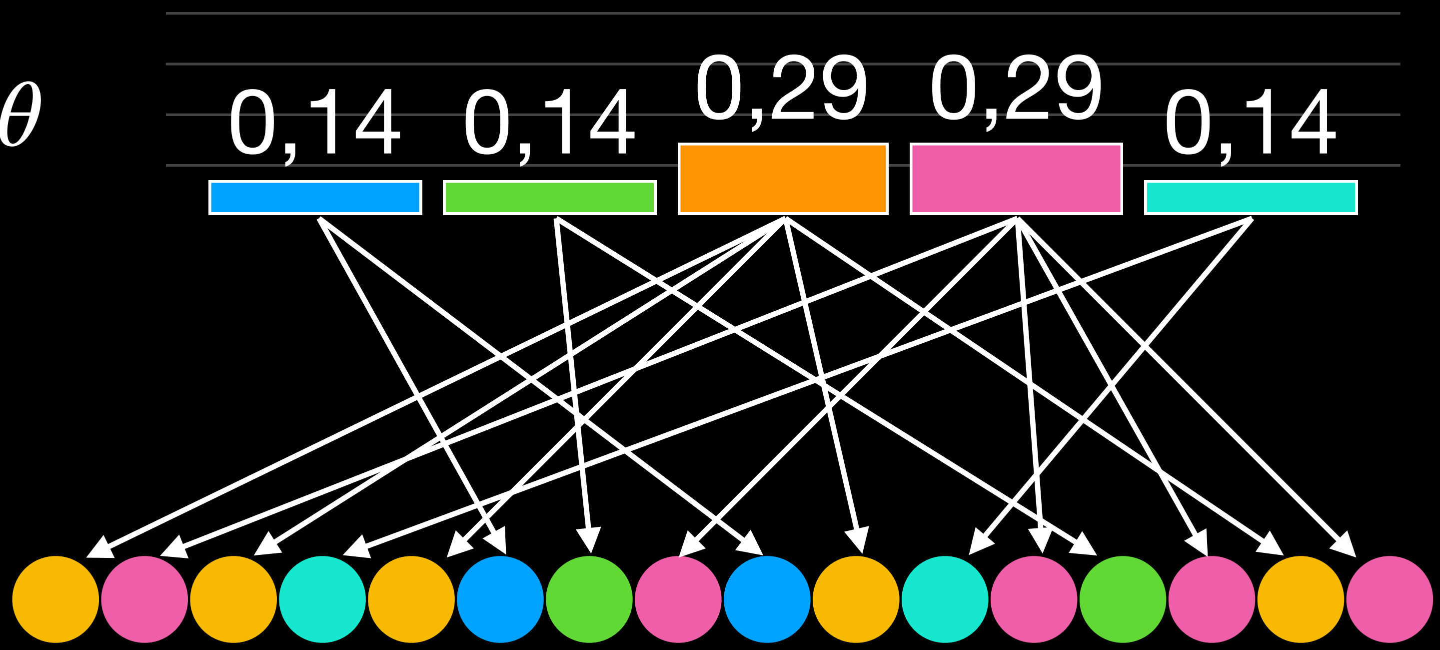
What do People Want in Smart Devices?



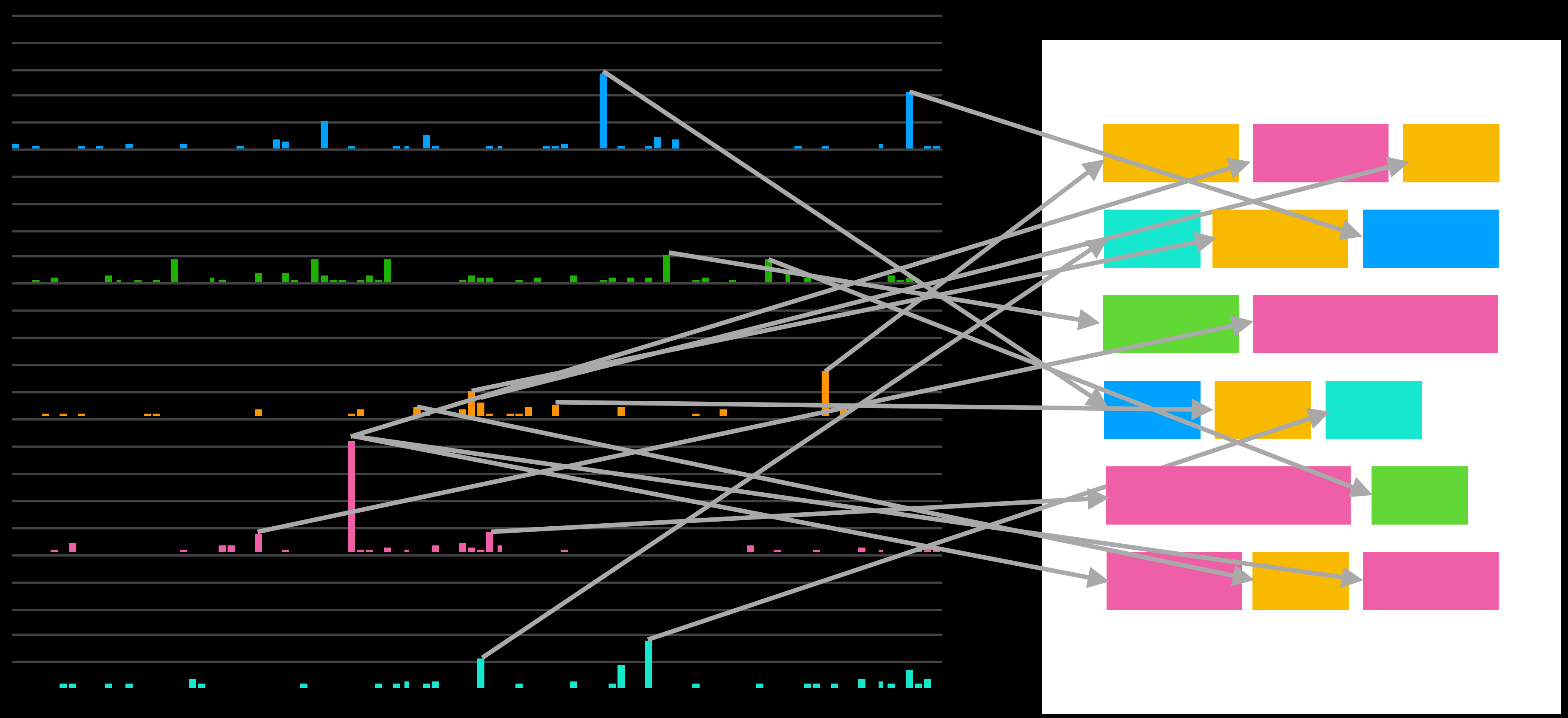
Latent Dirichlet Allocation

How to Generate Documents

- Draw a topic distribution θ



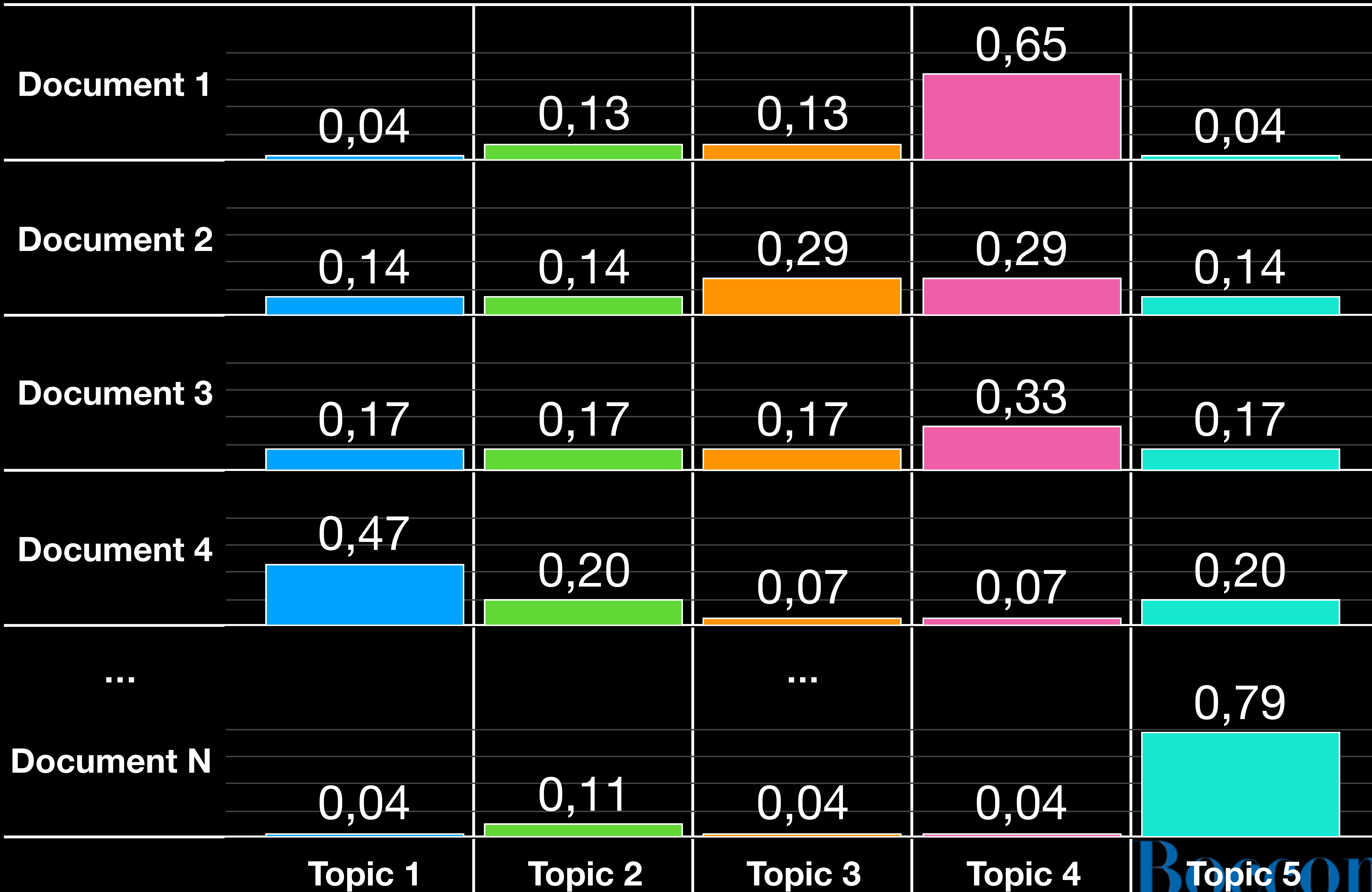
- For i in N:
  - Draw a topic from θ
  - Sample a word from the word distribution z



Bocconi

Topics per Document

$\theta = P(topic|document)$



Words per Topic

$z = P(word|topic)$

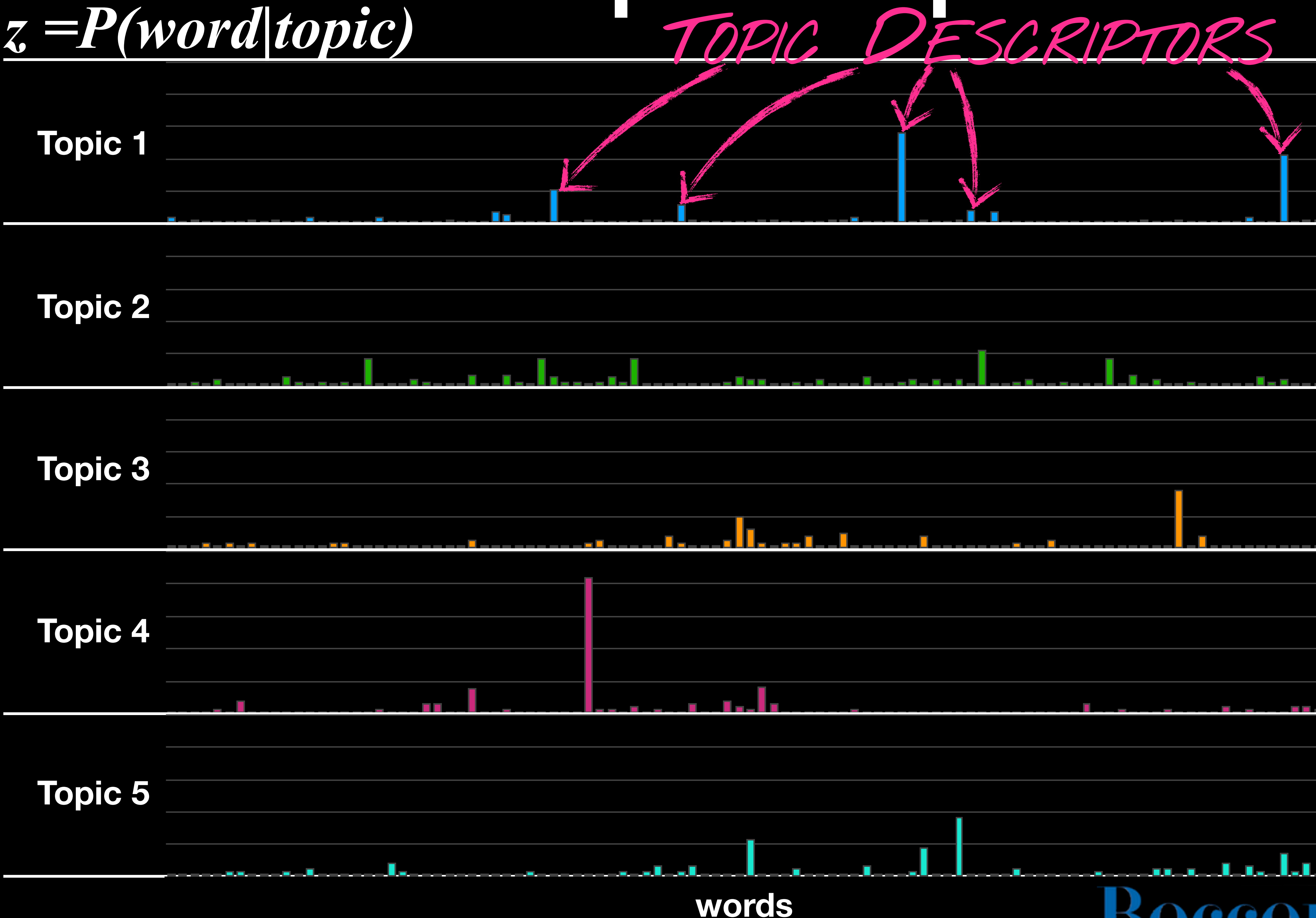
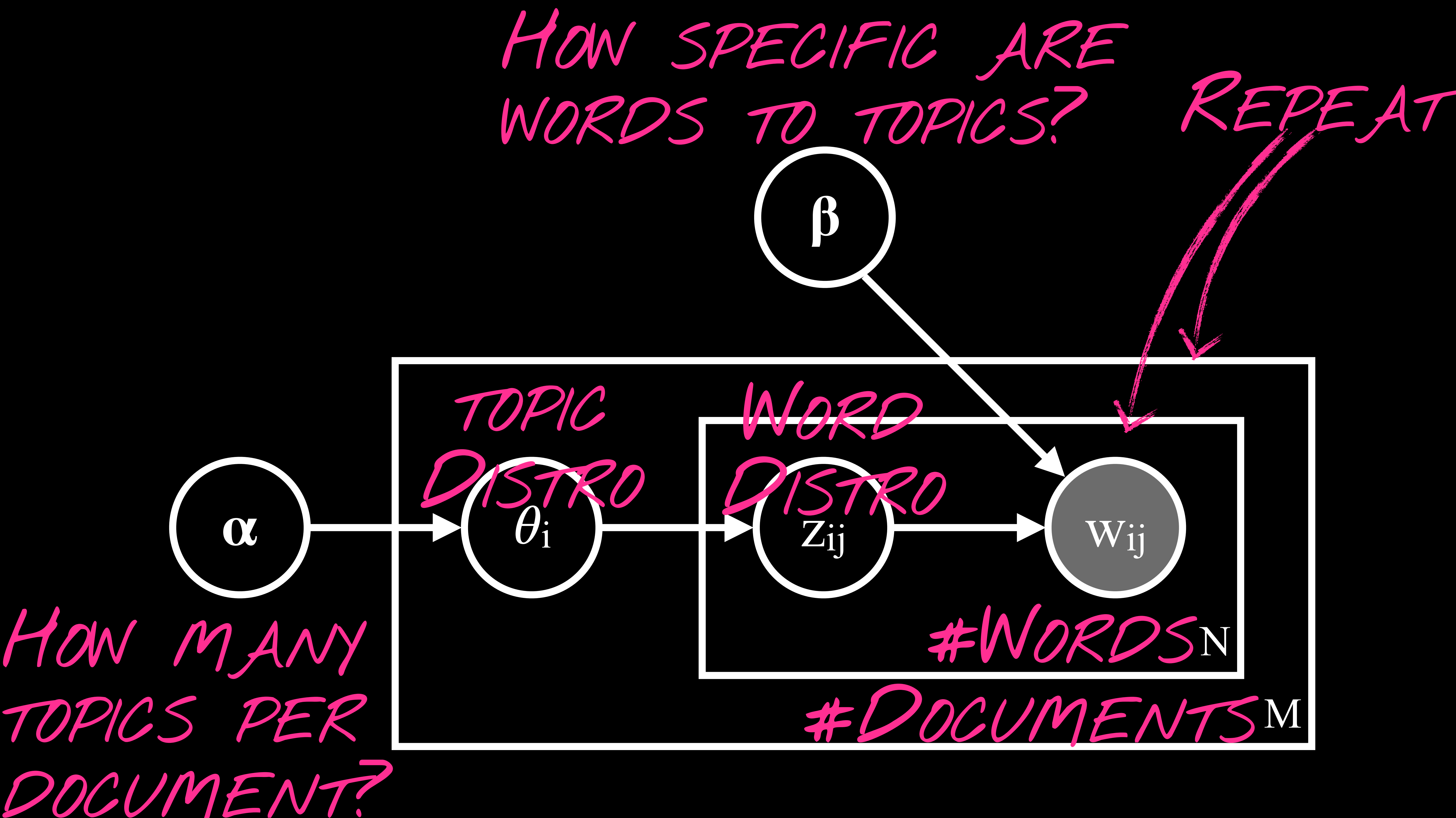
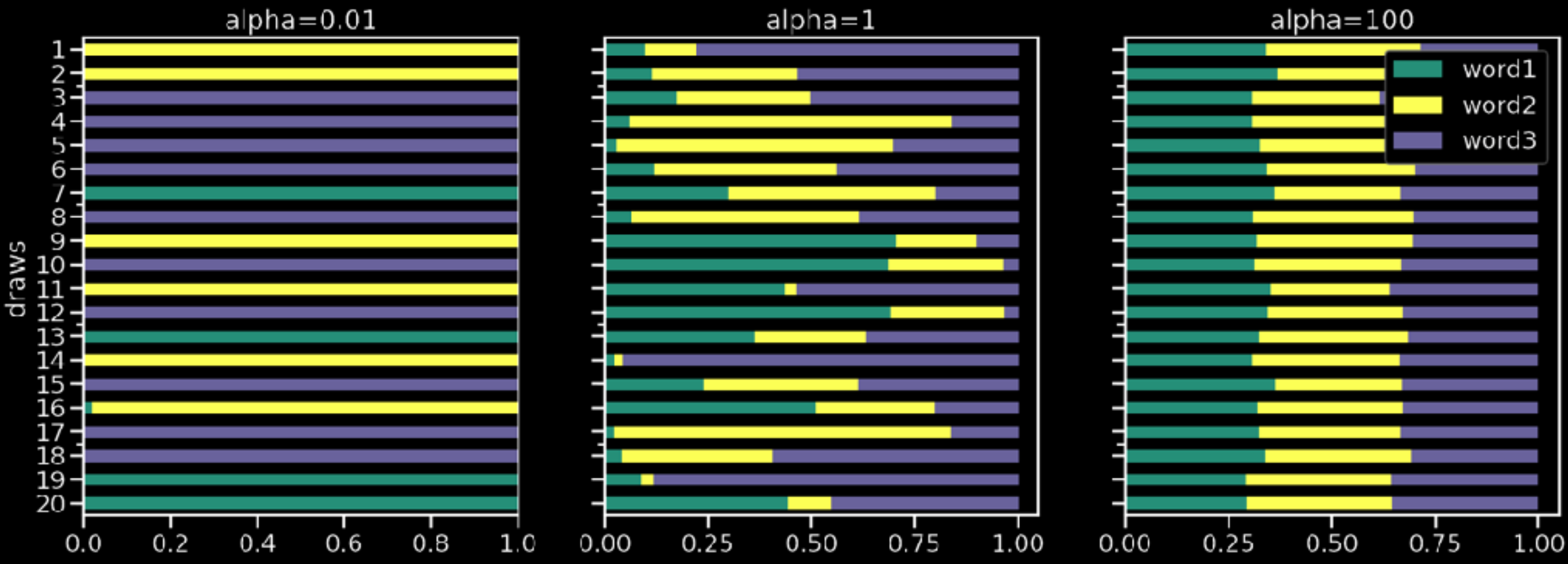


Plate Notation



Dirichlet Distributions

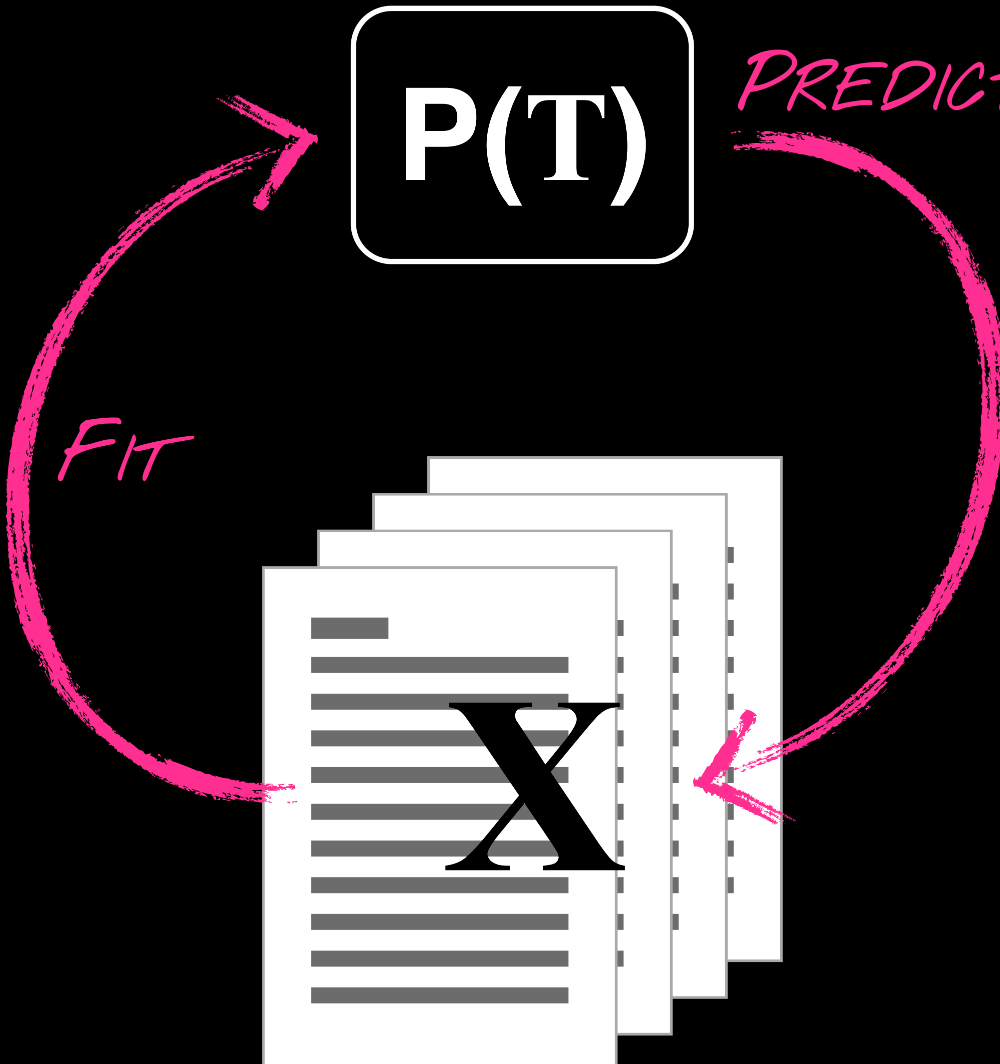
"DISTRIBUTION GENERATOR"



Bocconi

Evaluating LDA

Model-Inherent



$$= 2^{-\sum_x p(x) \log p(x)}$$

Perplexity

Content-Based

[apple, banana, pear, lime, orange]

[apple, banana, **foot**, lime, orange]

Which one's wrong?

Word Intrusion

Word and Topic Intrusion

Choose a word that is **not** related to others

- ○ loud
- ◉ time
- ○ music
- ○ sound
- ○ quality
- ○ speaker

WORD INTRUSION

TOPIC INTRUSION

Which group of words does **not** describe the following sentence:

I get my morning facts and news all in one easy to use system.

- ○ easy, use, setup, simple, install
- ○ control, command, system, integration, smart
- ○ music, weather, news, alarm, timer
- ◉ price, buy, sale, deal, item

Slide credit: Hanh Nguyen

Training and Parameters

Preprocessing

- Be aggressive:
  - lemmatization,
  - stopwords,
  - replace numbers/user names,
  - join collocations
  - use TFIDF
- use minimum document frequency 10, 20, 50, or even 100
- use maximum document frequency 50% – 10%

Training

Goal: Find distributions θ and z

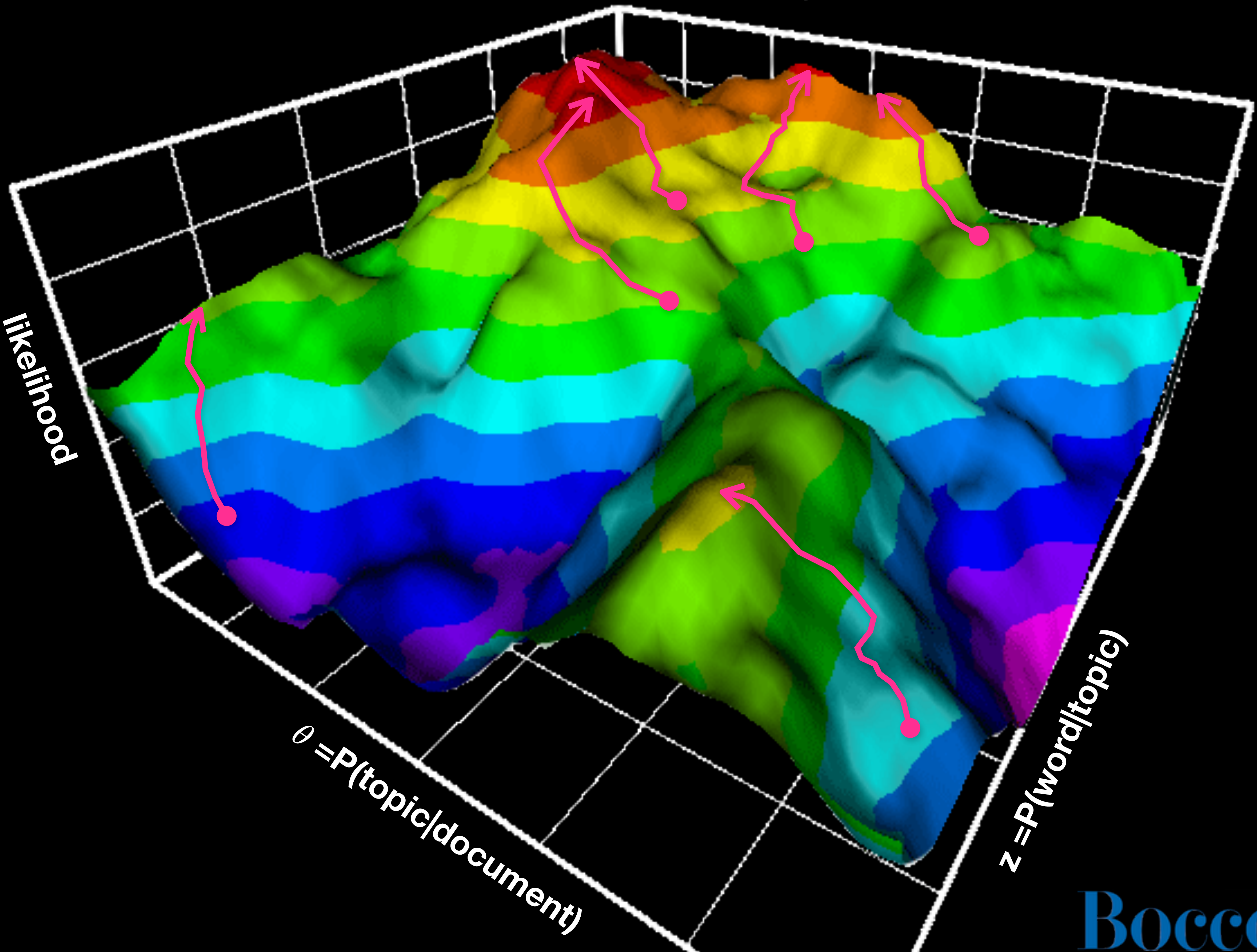
- In LM: use MLE (count and divide)
- In topic models: ??? (can't count what you don't see)

P(DATA) STOPS CHANGING

```
Initialize θ and z randomly
Repeat until convergence:
  "Hallucinate" topics from current θ and z
  Count hallucinated topics
  Normalize
```


Bocconi

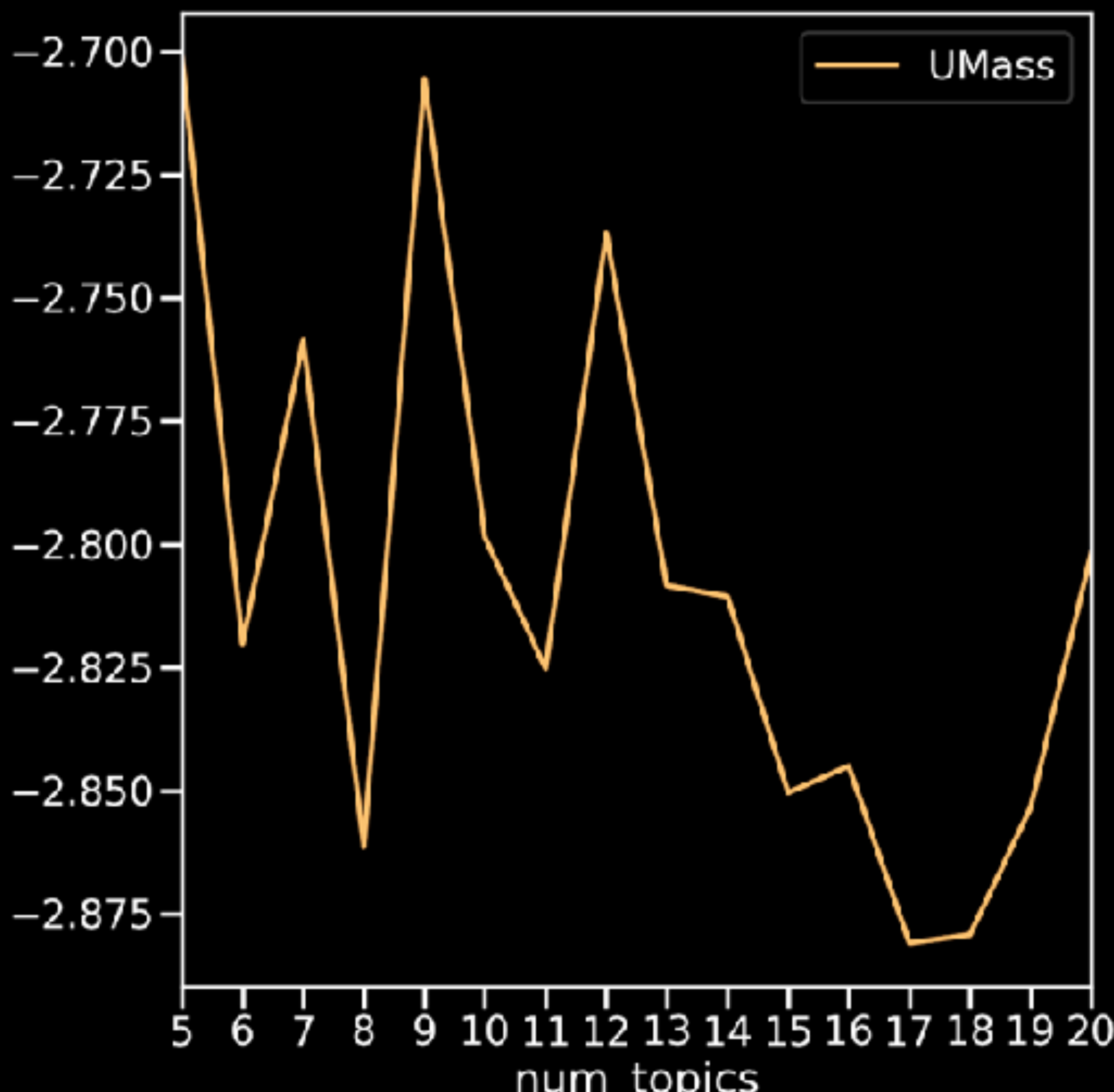
Training

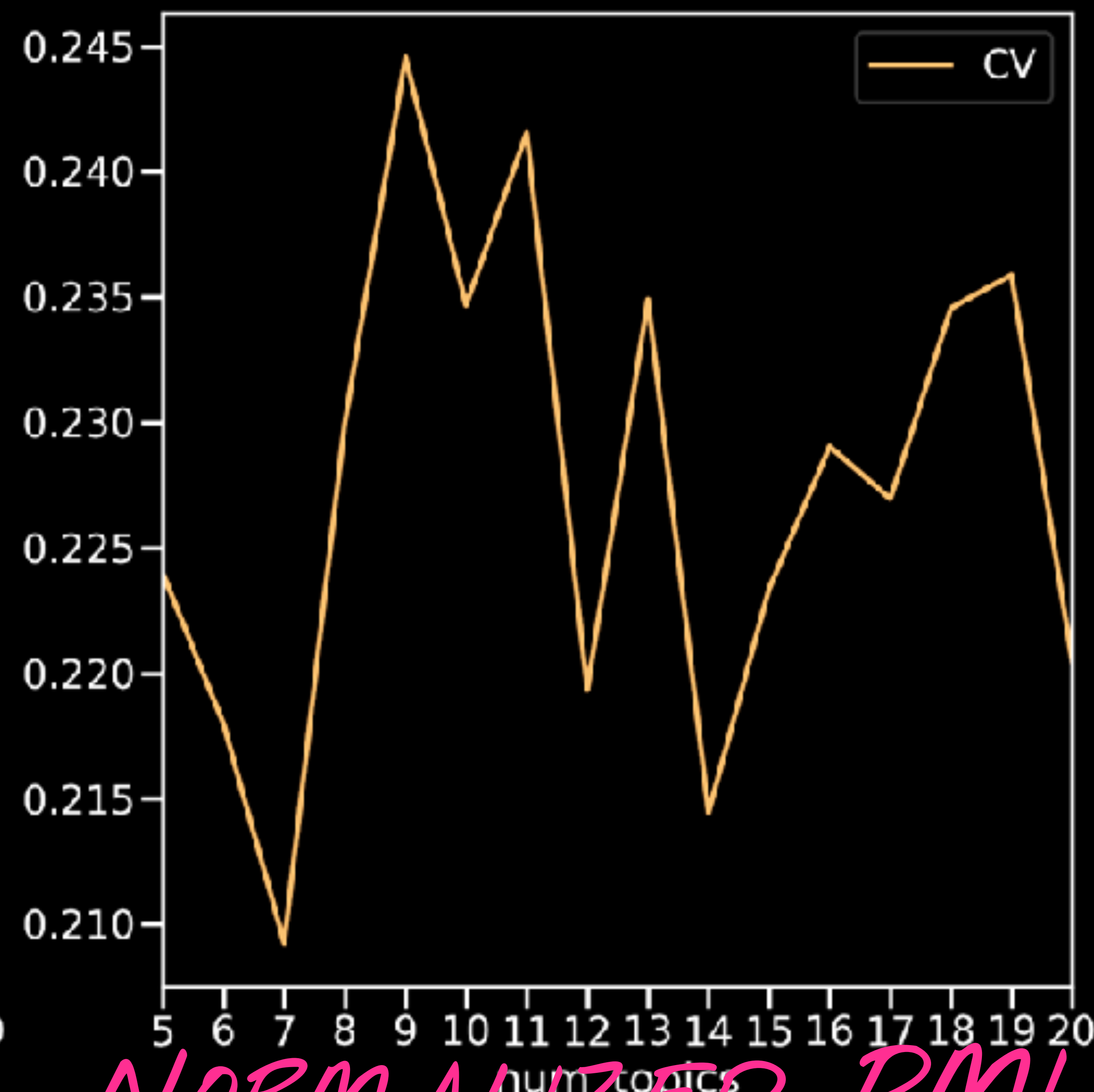


Bocconi

Parameters: K

Coherence Scores

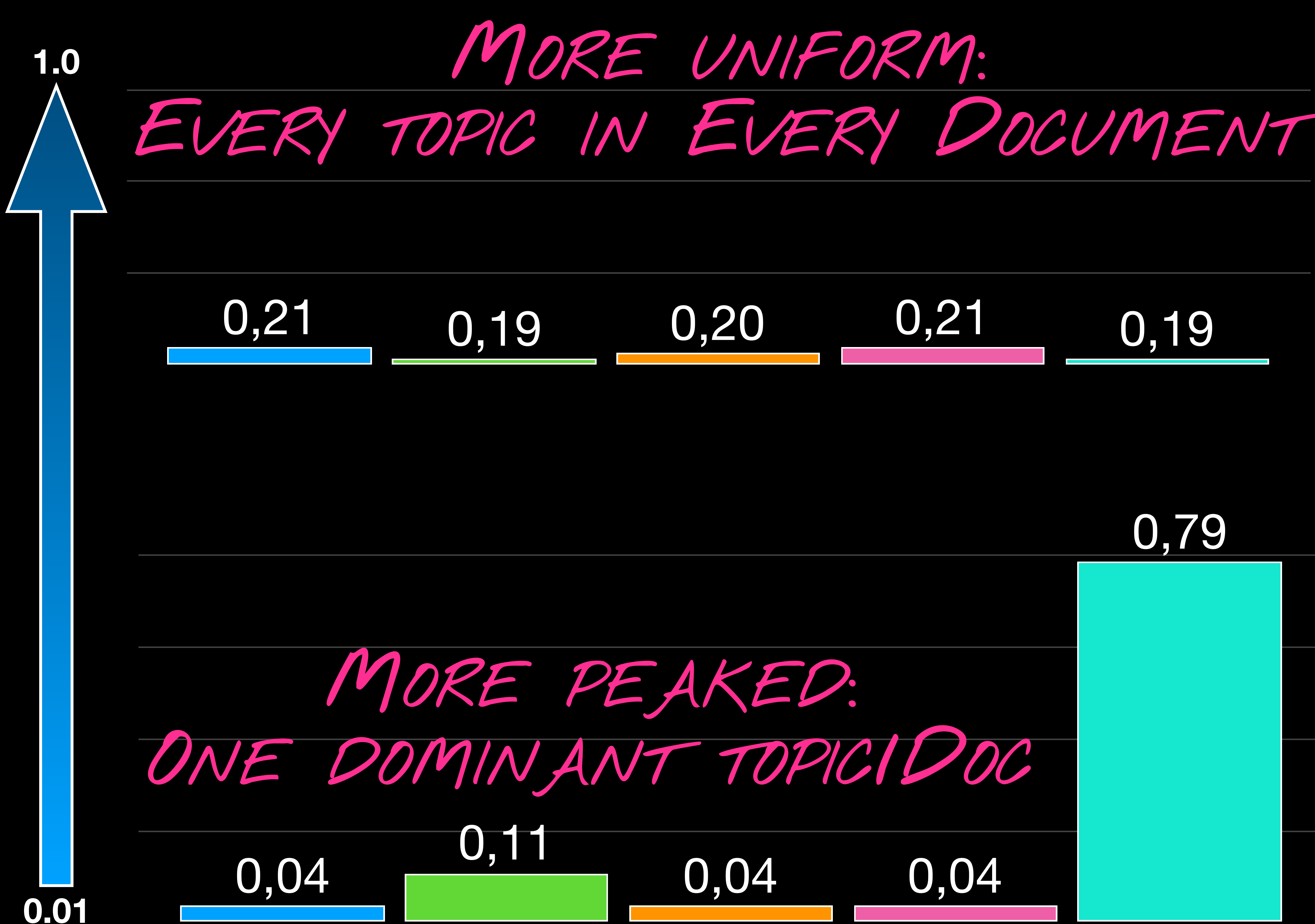




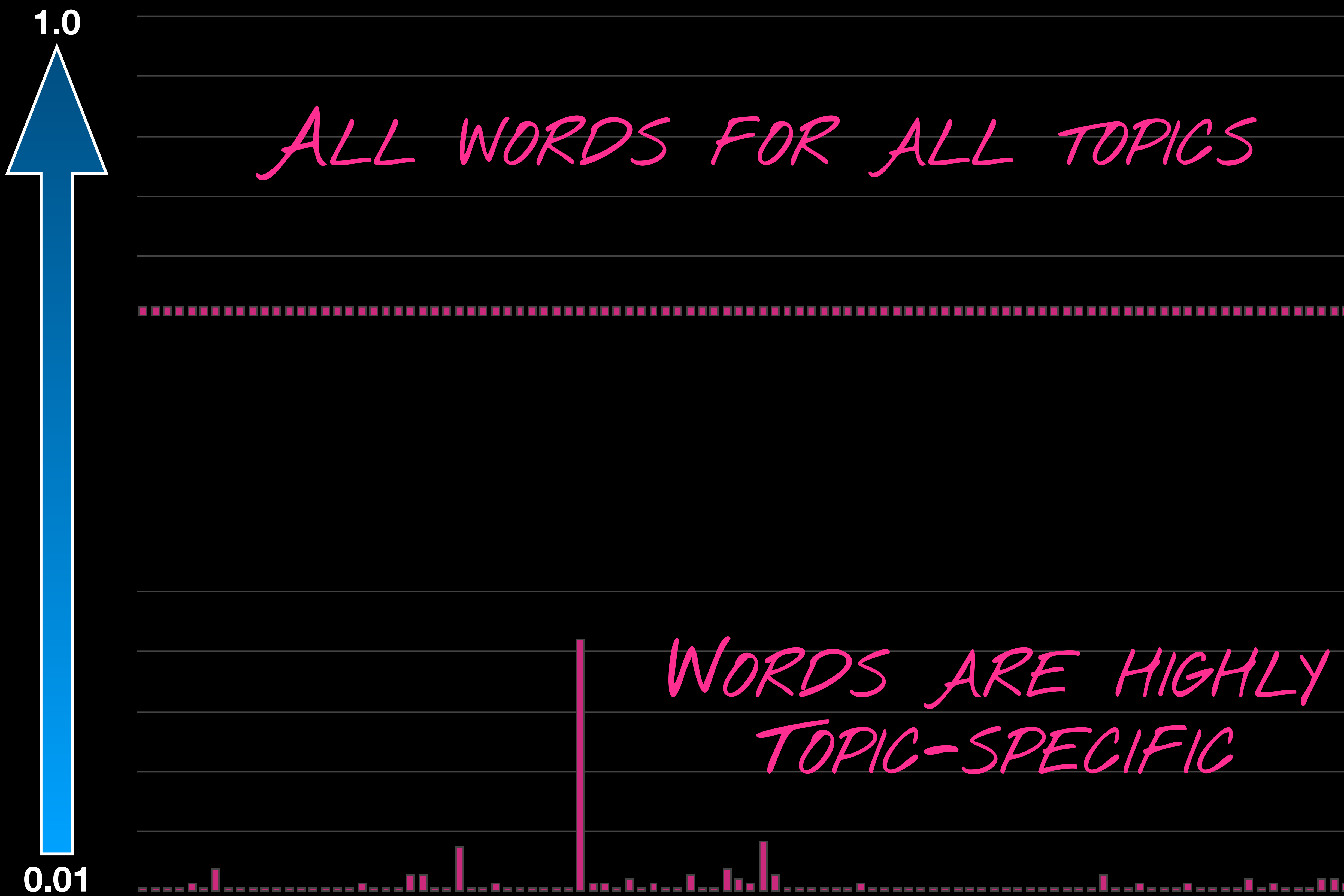
Log Prob of word Co-Occurrences

Normalized PMI and Cosine Similarity

Parameters: α



Parameters: β



Bocconi

Caveats!

Topic models ALWAYS need manual assessment, because:

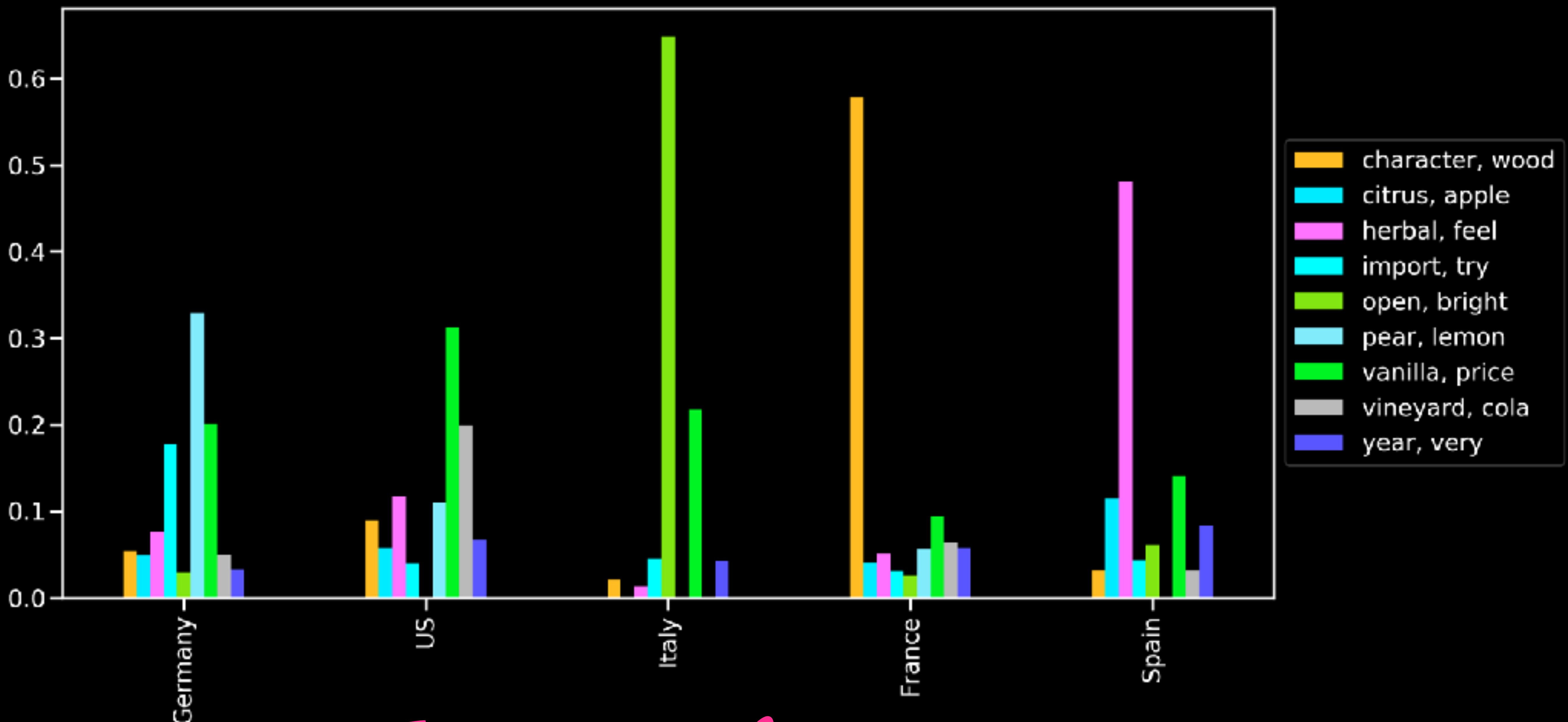
- Random initialization: no two models are the same!
- More likely models ≠ more interpretable topics
- "Interpretable" is subjective
- Topics are not stable from run to run

Topic or Not?

- "pasta, pizza, wine, sauce, spaghetti"
- "BLEU, Bert, encoder, decoder, transformer"

Author Topic Models

- Learn separate topic distribution for external factors



TOPICS BY COUNTRY

Wrapping Up

Take-Home Points

- **LDA** is one architecture for **topic models**
- Model document generation conditioned on latent topics
- Topic models are **stochastic**: each run is different
- **Preprocessing** and **parameters** influence performance
- Results need to be **interpreted**!