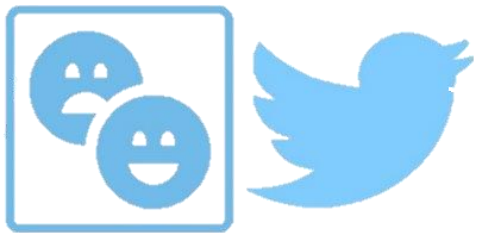
Part II-2: Topic Modeling



Outline

i. Overview
ii. Topic Modeling Approaches
iii. Structural Topic Model (STM)
iv. Keyword-Based Topic Extraction

Part II-2: Topic Modeling

Overview

Overview Goal of Topic Modeling

- **Goal:** discover latent semantic structures in a corpus & group documents into topical clusters
- **Exploratory** method that does not require prior knowledge
  → Unsupervised learning

as opposed to: topic classification

- Often particularly useful in **early phases** of text analysis
  - Getting a better feeling for the corpus at hand
  - Facilitating/enhancing downstream tasks (e.g., sentiment analysis)

Overview Terminology

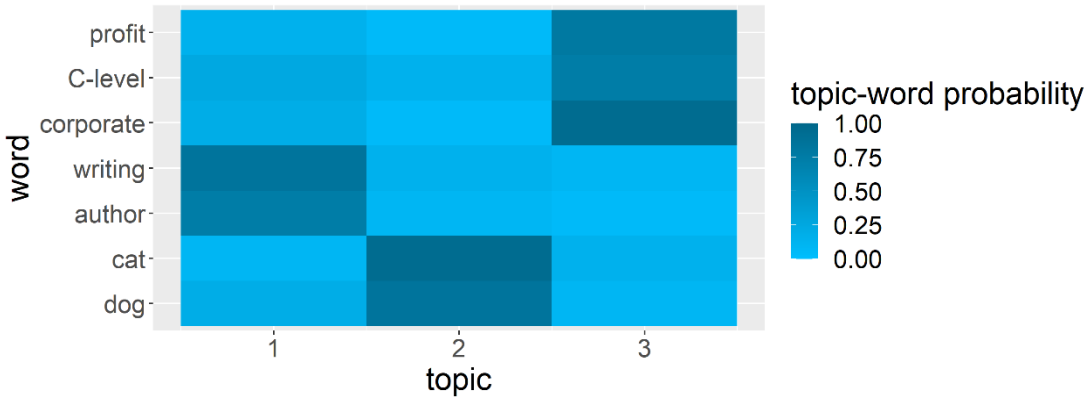
- **So, what exactly is a topic?**
  - Topic modeling revolves around the **probability** of words occurring in texts of a specific cluster.
  - Intuitively, we would expect some words to appear more frequently in documents about a certain topic than in others.

*e.g., the word **tasty** should be more likely to occur in a text about food than in one about stock markets*

  - In fact, a topic is just a **probability distribution** over a fixed vocabulary.

Overview Terminology

- **Topic-word distribution β_k:** probability distribution over vocabulary given topic *k*
  - Constant across documents
  - Characteristic of a topic

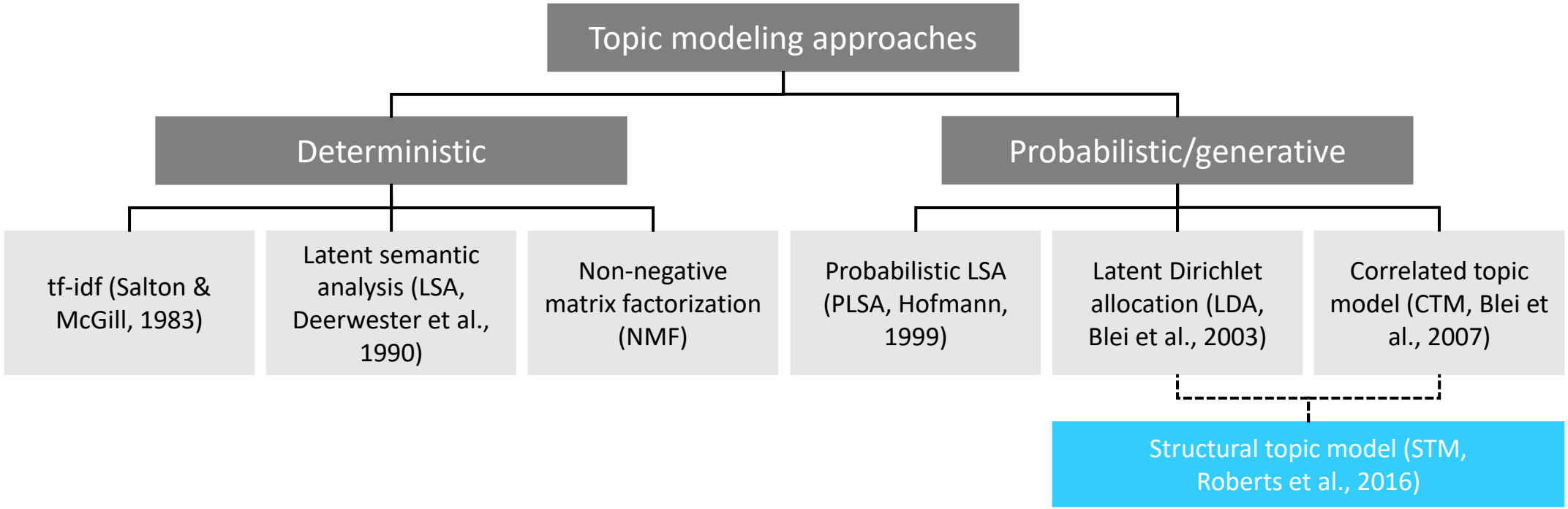


- **Topic proportions:** length-*K* vector of probabilities of a document belonging to a certain topic

Part II-2: Topic Modeling

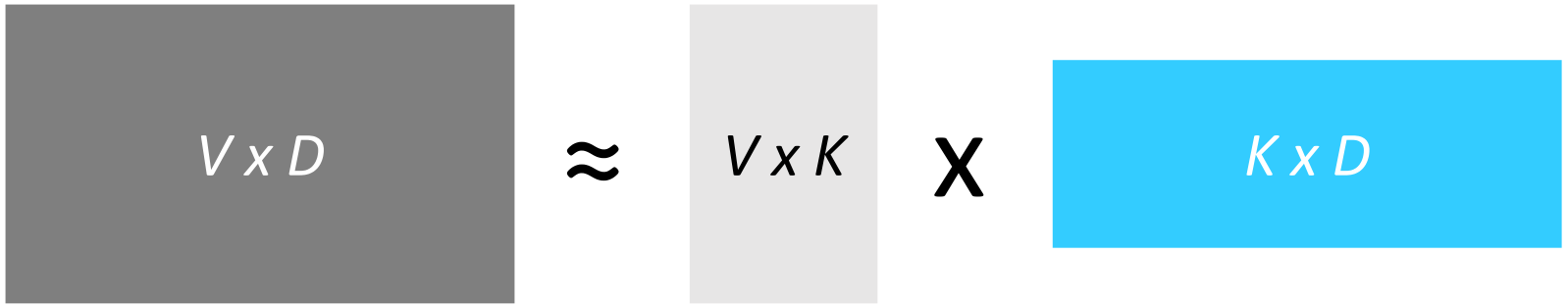
Topic Modeling Approaches

Approaches **Rough Taxonomy**



Approaches **Deterministic**

- **Deterministic approaches**
  - Term-by-document matrix
  - LSA, NMF: matrix factorization to identify latent topics



- **Problems**: inference & out-of-sample extension

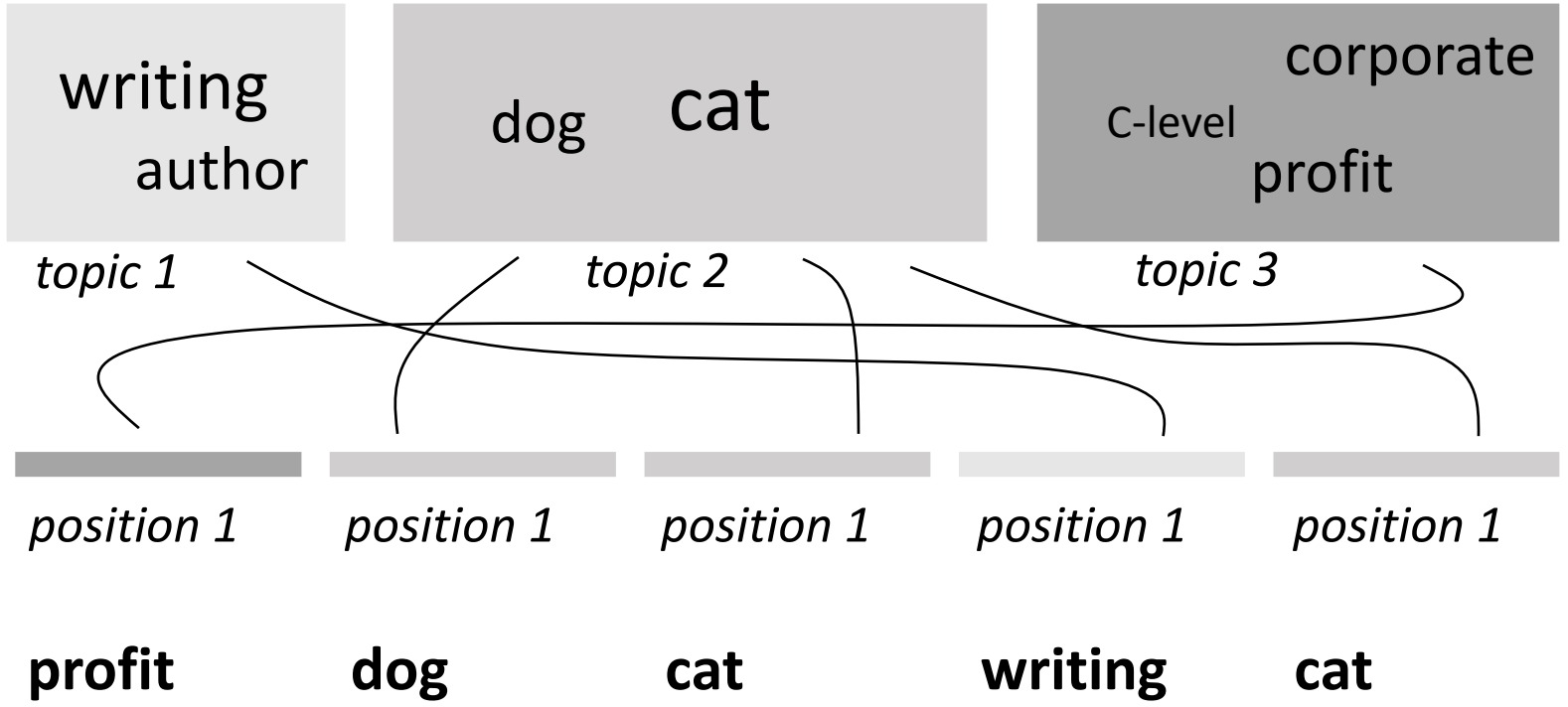
Approaches **Probabilistic/Generative**

- **Probabilistic/generative approaches**
  - Hierarchical Bayesian mixture models
  - **Idea**: reverse-engineer the imaginative process of document generation

1. *For each of document d within a corpus draw a vector of topic proportions from the assumed distribution*
2. *For each word position n within d*
   1. *draw a topic assignment from the assumed distribution*
   2. *draw a word from the assumed distribution*

Approaches **Probabilistic/Generative**

- **Example**



Overview **Challenges**

- Hyperparameters: most importantly, **number of topics**

- Extreme **brevity** of Twitter data
  - Problematic for most topic modeling approaches
  - Potential mitigation by **pooling**
  - Special models dedicated to short texts

Part II-2: Topic Modeling

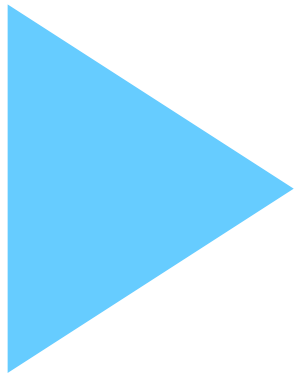
Structural Topic Model (STM)

STM Expert Talk

Expert Talk: STM

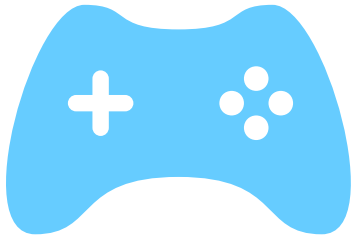
Patrick Schulze & Simon Wiegrebe: ***Twitter in the Parliament – A Text-based Analysis of German Political Entities***

STM Approach



Demo 7: STM

STM Exercise



Exercise 4: Topic Modeling

Part II-2: Topic Modeling

Keyword-Based Topic Extraction

Keyword-Based TE **Idea**

- **Situation**
  - (Statistical) topic modeling not always producing meaningful topics
  - Quite some human input required still
  - Also, unsupervised approach not always appropriate
- **Idea:** specify keywords & find related documents
- **Approach**
  1. Specify list of keywords
  2. Find similar words (both morphologically & semantically)
  3. Assign all documents using these words to the associated topic

Part II-2: Topic Modeling

Literature and References

Blei, D., Ng, A., and Jordan, M. (2003): Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, pp. 993-1022.

Blei, D., and Lafferty, J. (2007): A Correlated Topic Model of Science. *The Annals of Applied Statistics 1(1)*, pp. 17-35.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990): Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41(6),* pp. 391-407.

Hofmann, T. (1999): Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57.

Roberts, M., Stewart, B., Tingley, D., and Airoldi, E. (2013): The Structural Topic Model and Applied Social Science. *Advances in Neural Information Processing Systems Workshop on Topic Models.*

Salton, G., and McGill, M. (1983): Introduction to Modern Information Retrieval, McGraw-Hill.

Vayansky, I., and Kumar S.A.P. (2020): A Review of Ttopic Modeling Methods, *Information Systems*, doi: https://doi.org/10.1016/j.is.2020.101582.