Topic models

Mate Akos

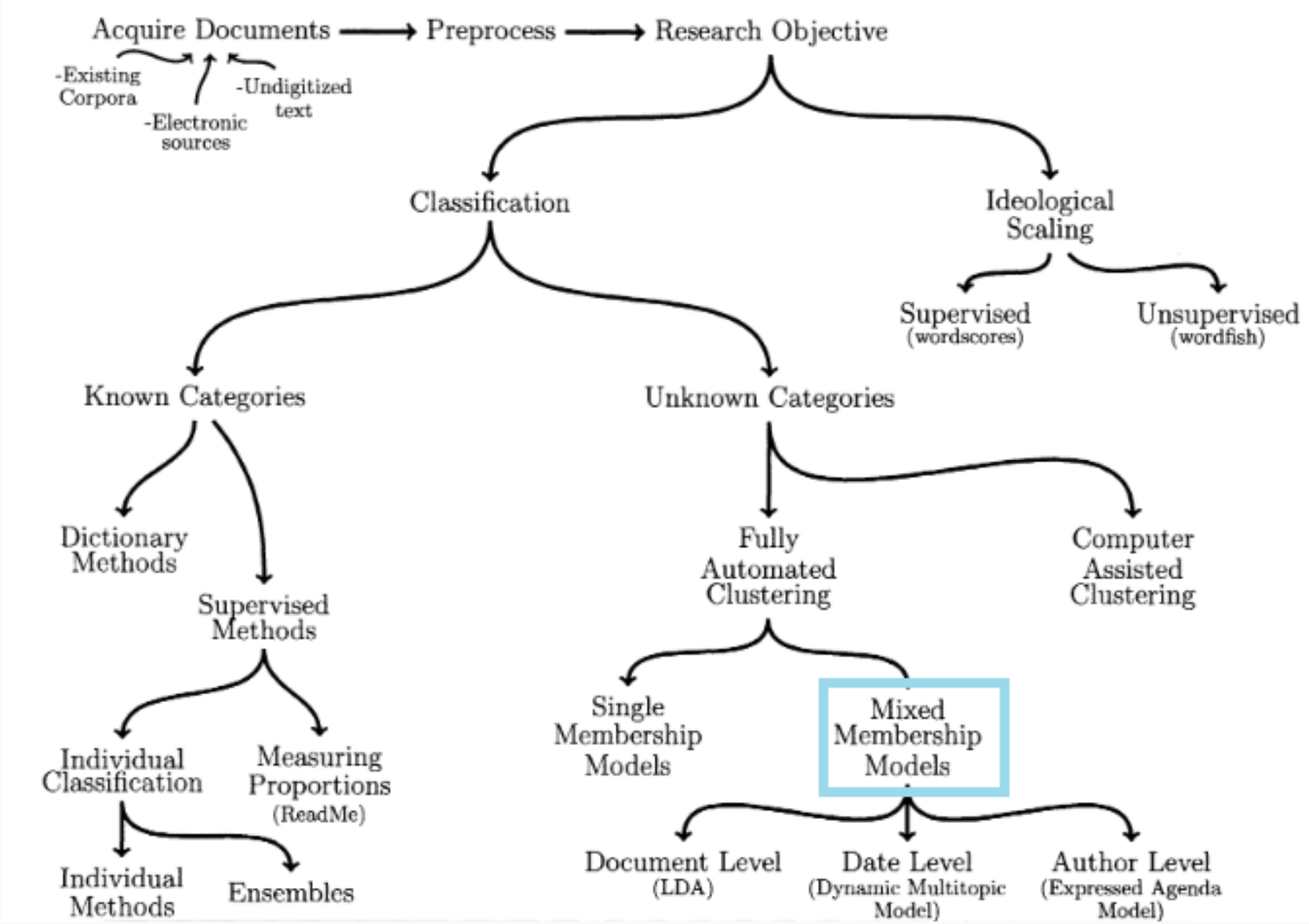
2019 November

TK PTI

For today

1. Where we are
2. Topic models
3. Latent Dirichlet Allocation
4. Structural Topic Models

Where we are


Acquire Documents
Preprocess
Research Objective
-Existing Corpora
-Electronic sources
-Undigitized text
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Fully Automated Clustering
Computer Assisted Clustering
Individual Classification
Measuring Proportions (ReadMe)
Single Membership Models
Mixed Membership Models
Individual Methods
Ensembles
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)

Topic models

Goal: identify the main topics in our corpus

- Discover main themes in our corpus
- Can be applied to massive text data
- Wide range of applicability (legal, social media, images, etc.)
- Great for exploratory analysis

Topic models: unsupervised clustering

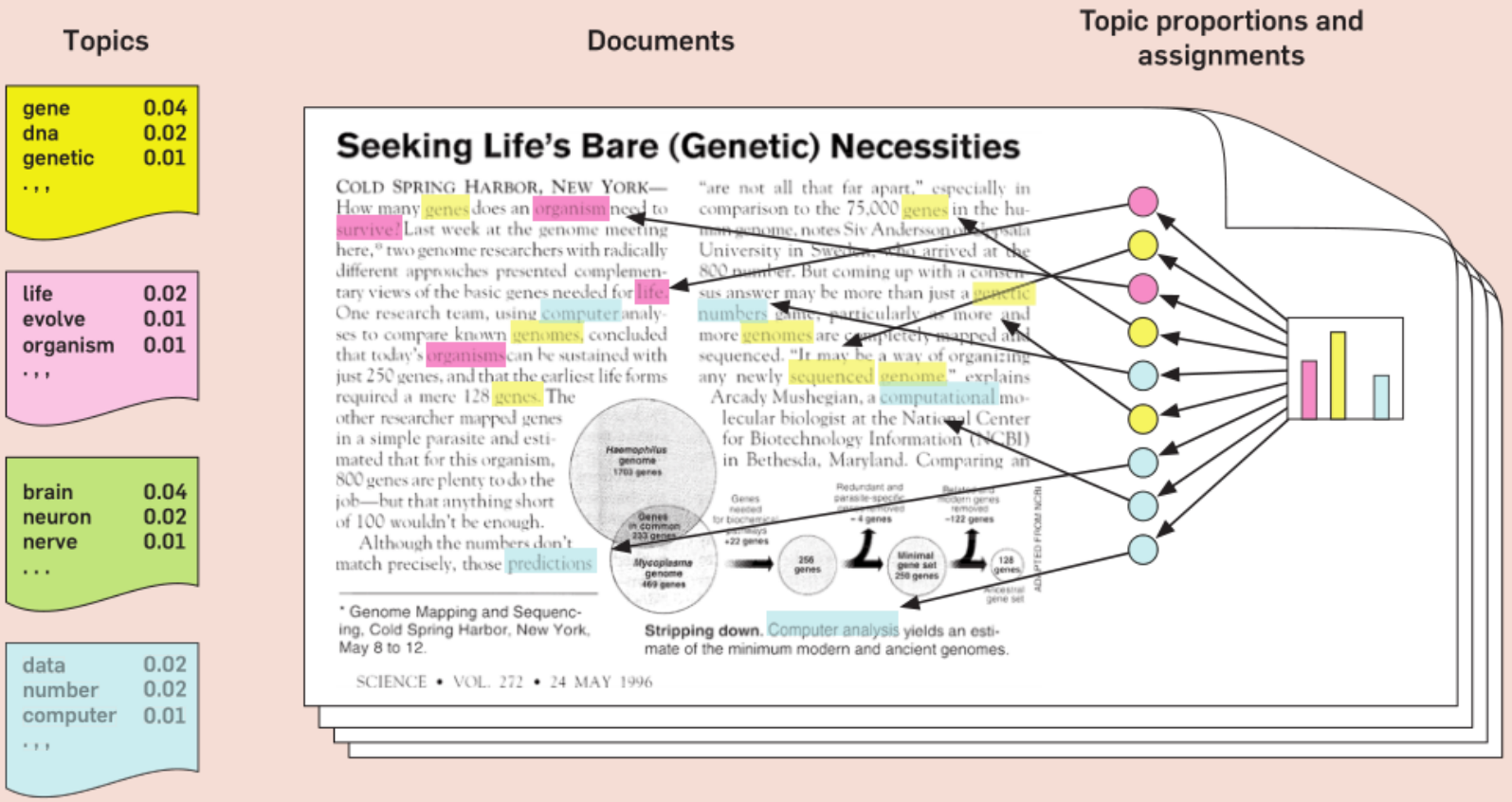
- We do not know how many ”topics” truly are in a corpus
- The models will find K models
- Well suited for explorative work
- most common model: Latent Dirichlet Allocation

LDA

- Each document contains multiple topics
- LDA offers a generative model for the texts
- Topics: a distribution over a fixed vocabulary (e.g.: nationalism topic would contain ”border”, ”sovereignity”, ”nation” words with high probability)
- Generating words for each document:
  1. Randomly choose a distribution over topics.
  2. For each word in the document
     - 2.1 Randomly choose a topic from the distribution over topics in step 1
     - 2.2 Randomly choose a word from the corresponding distribution over the vocabulary.
- Topic proportions for each document, and word probability for each topic

LDA

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



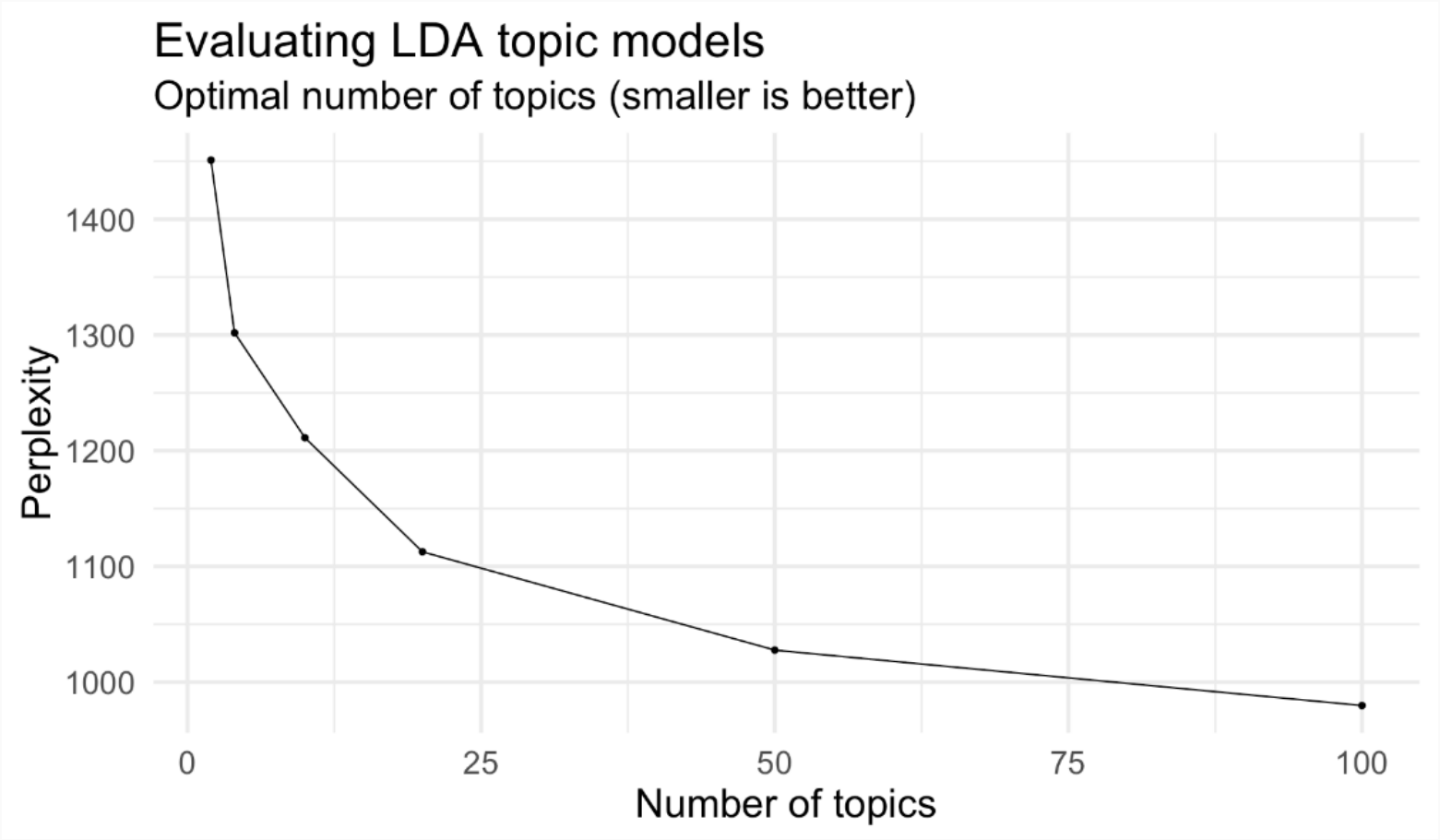
Honnnan a név?

- LDA assumes that the prior distribution of topics across documents and word distribution across topics follow the Dirichlet distribution.
- Interpretation: For example, in a two-topic model we could say “Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B.” (from www.tidytextmining.com)

Determining K

- Often the hardest choice
- Can be made based on ”substantive fit” (Grimmer and Stewart 2013)
- **perplexity** criterion is often used to evaluate models
- A statistical measure of how well a probability model predicts a sample
    - Estimate topic models with various values of k
    - Calculate perplexity score.
    - Choose topic model with lower perplexity

LDA


Evaluating LDA topic models
Optimal number of topics (smaller is better)
Perplexity
1400
1300
1200
1100
1000
0
25
50
75
100
Number of topics

Validating LDA

Based on Grimmer and Stewart (2013)

- Semantic validity: extent to which topics are coherent
- Predictive validity: how well does variation in topic usage correspond with predicted events
- Convergent validity: extent to which model output can be validated with other approaches

Beyond LDA - Structural Topic models

One recent innovation is **Structural Topic models**

- Similar to LDA but includes document level covariates
- The covariates can affect affect topical prevalence, topical content or both
- Documents with similar covariates tend to have similar topics
- Documents with similar covariates tend to use similiar words to refer to the same topics