Text mining and natural language analysis

Jefrey Lijffijt

PART I:

Introduction to Text Mining

Why text mining

- The amount of text published on paper, on the web, and even within companies is inconceivably large
- We need automated methods to
  - **Find**, **extract**, and **link information** from documents

Main ‘problems’

- **Classification**: categorise texts into classes (given a set of classes)

- **Clustering**: categorise texts into classes (not given any set of classes)

- **Sentiment analysis**: determine the sentiment/ attitude of texts

- **Key-word analysis**: find the most important terms in texts

Main ‘problems’

- **Summarisation**: give a brief summary of texts
- **Retrieval**: find the most relevant texts to a query
- **Question-answering**: answer a given question
- **Language modeling**: uncover structure and semantics of texts

And **answer-questioning**?

Definitions from Wikipedia

Related domains

- **Text mining** […] refers to the process of deriving high-quality information from text.

- **Information retrieval (IR)** is the activity of obtaining information resources relevant to an information need from a collection of information resources.

- **Natural language processing (NLP)** is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.

- **Computational linguistics** is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective.

PART II:

Clustering & Topic Models

Main topic today

- Text clustering & topic models

- Useful to categorise texts and to uncover structure in text corpora

Primary problem

- How to **represent text**

  - What are the relevant features?

Main solution

- **Vector-space** (bag-of-words) **model**

	Word 1	**Word 2**	**Word 3**	**…**
Text 1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	
Text 2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	…
Text 3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	
…		…		

Simple text clustering

- Clustering with k-means algorithm and cosine similarity
- Idea: two texts are similar if the frequencies at which words occur are similar
- $s(w_1, w_2) = \frac{w_1 \cdot w_2}{||w_1||\ ||w_2||}$
- Score in [0,1] (since $w_{i,j} \geq 0$)
- Widely used in text mining

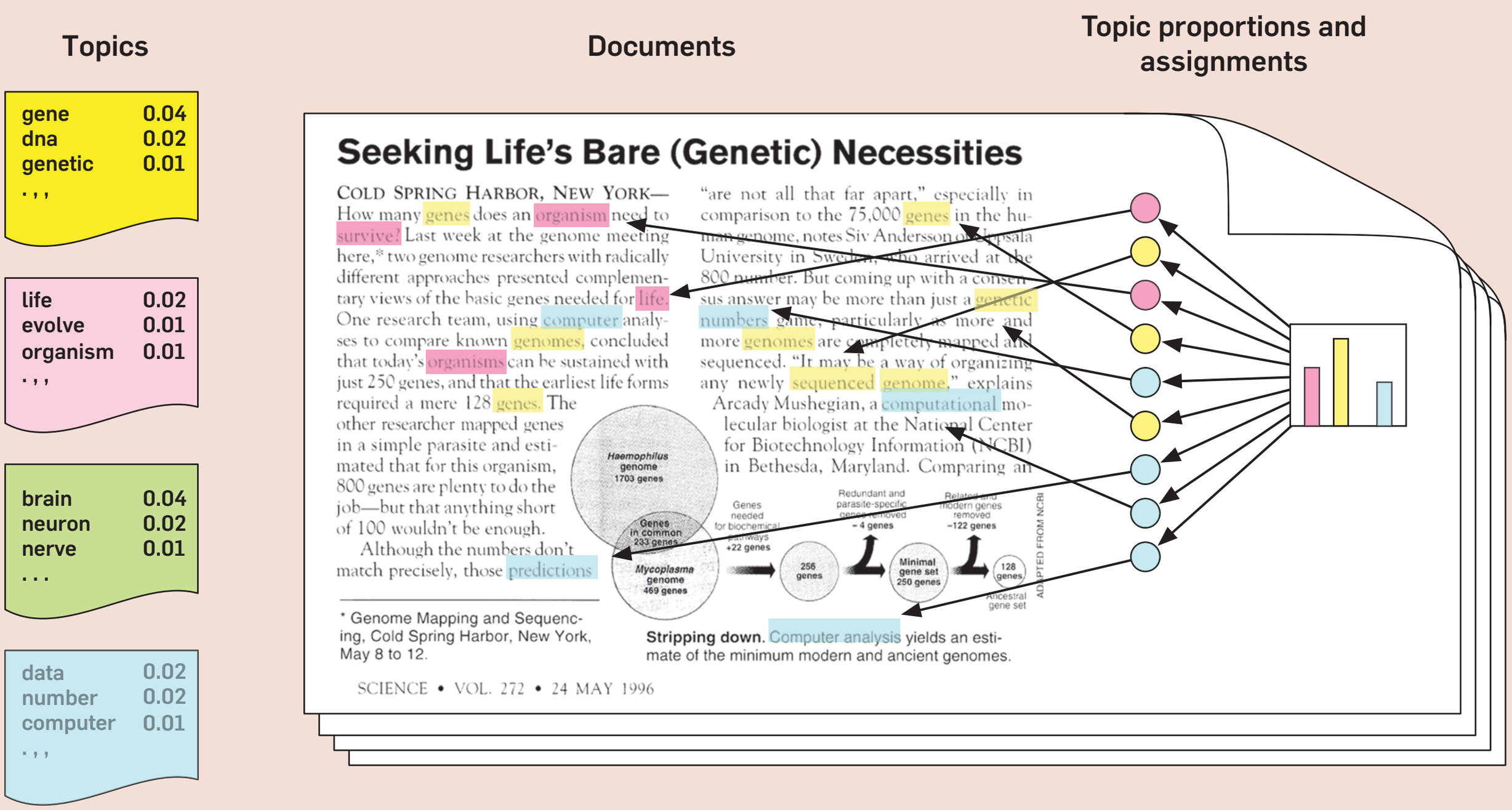
Demo

- Reuters-21578
- 8300 (categorised) newswire articles
- Clustering is a *single command* in Matlab
- Data (original and processed .mat):
  http://www.daviddlewis.com/resources/testcollections/reuters21578/
  http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

More advanced clustering

- Latent Dirichlet Allocation (LDA)
- Also known as *topic modeling*
- Idea: texts are a weighted **mix of topics**
- [The following slides are inspired by and figures are taken from David Blei (2012). Probabilistic topic models. *CACM* 55(4): 77–84.]

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.



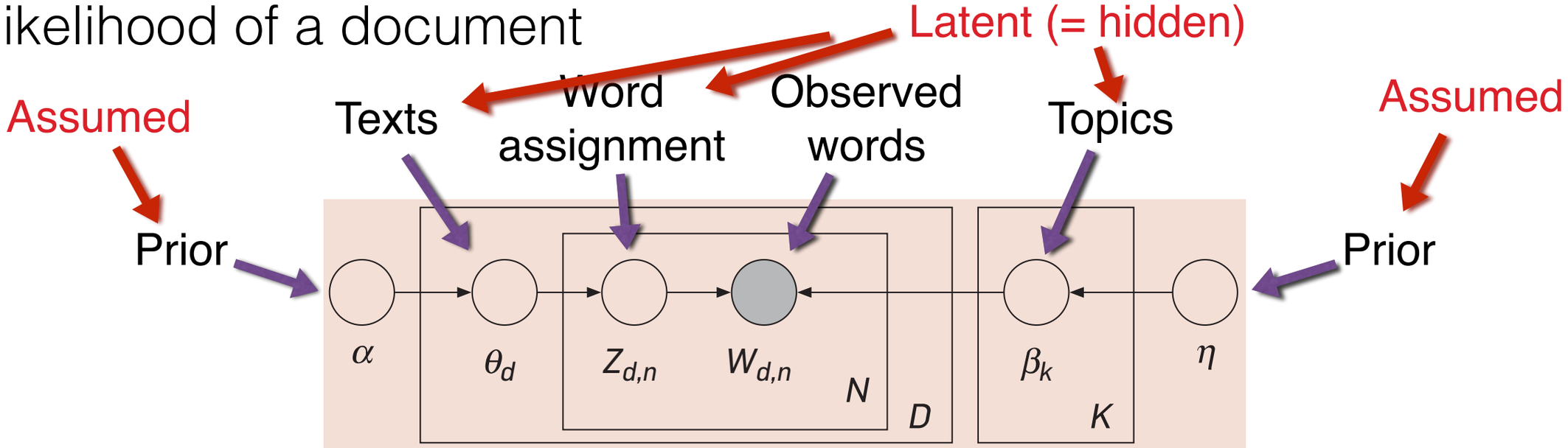
Technically (1/2)

- **Topics** are probability distribution over words
  - The distribution defines how often each word occurs, given that the topic is discussed
- **Texts** are probability distributions over topics
  - The distribution defines how often a word is due to a topic
- These are the *free parameters* of the model

Technically (2/2)

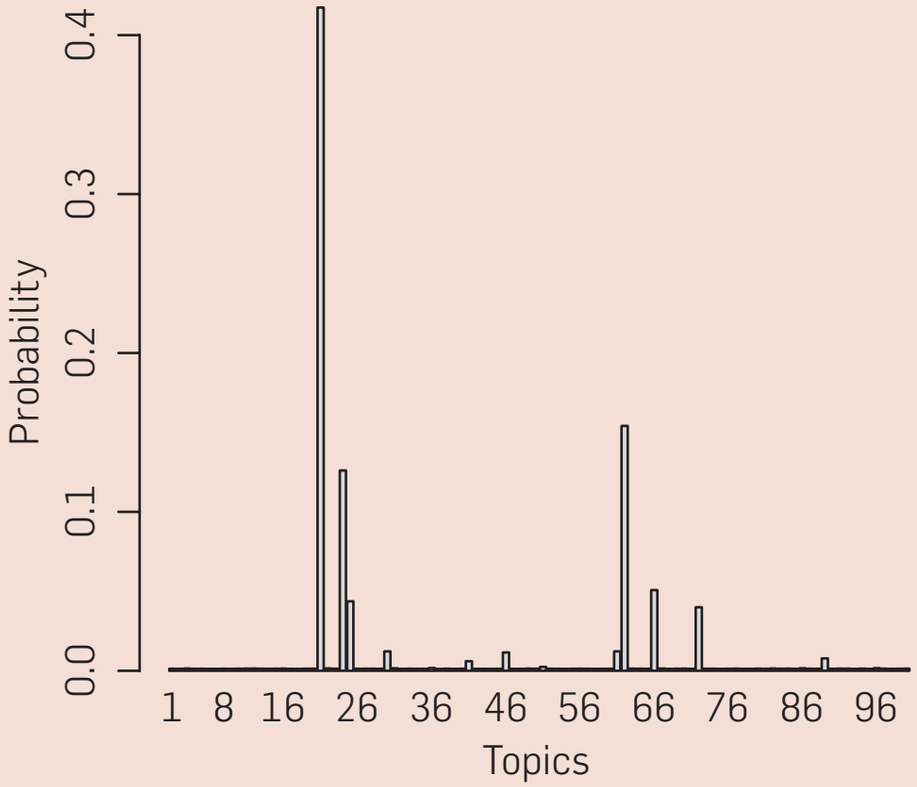
- For each word in a text, we can compute how probable it is that it belongs to a certain topic

- Given the topic probability and the topics, we can compute the likelihood of a document



- The optimisation problem is to find the posterior distributions for the topics and the texts (see article)

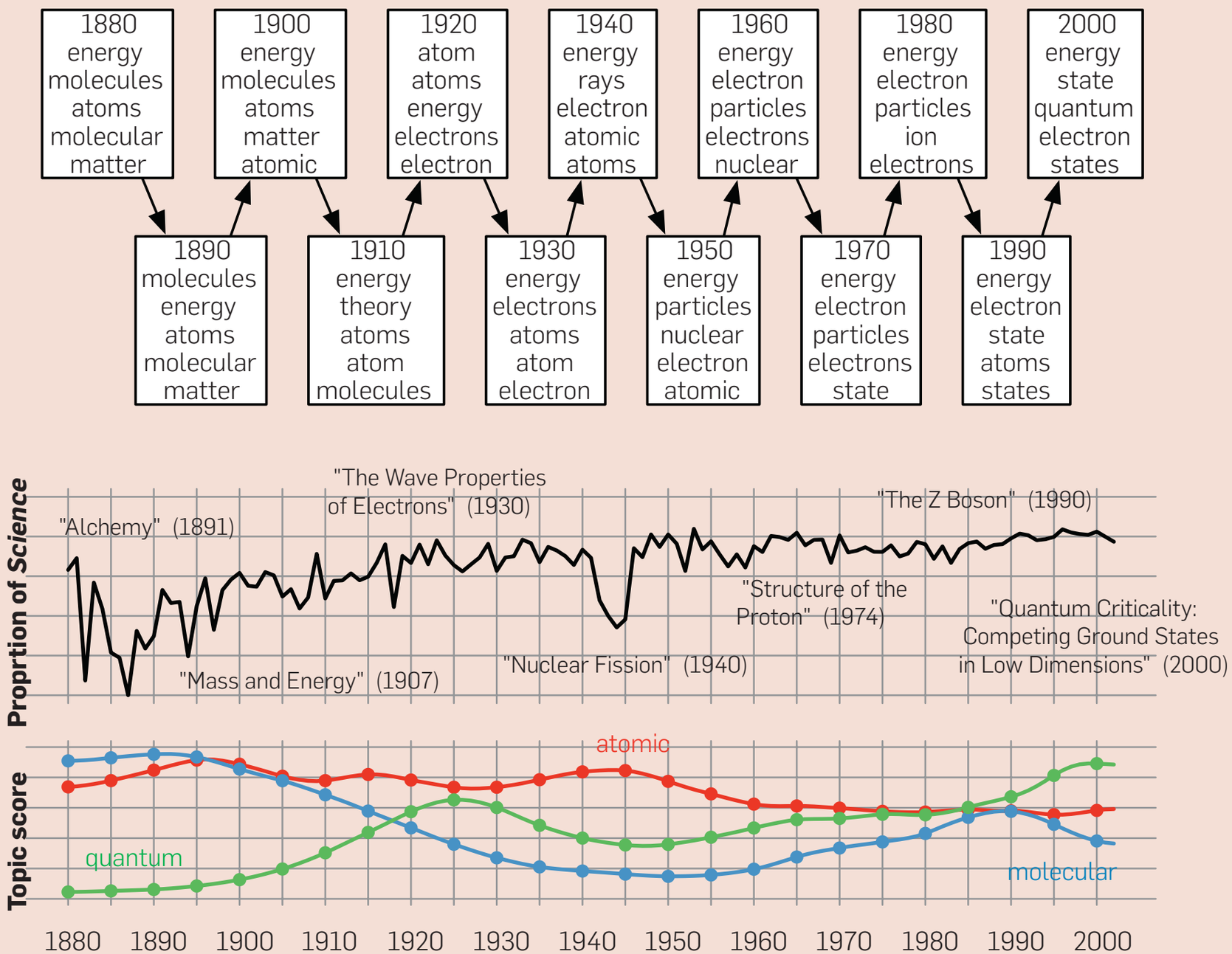
Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
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

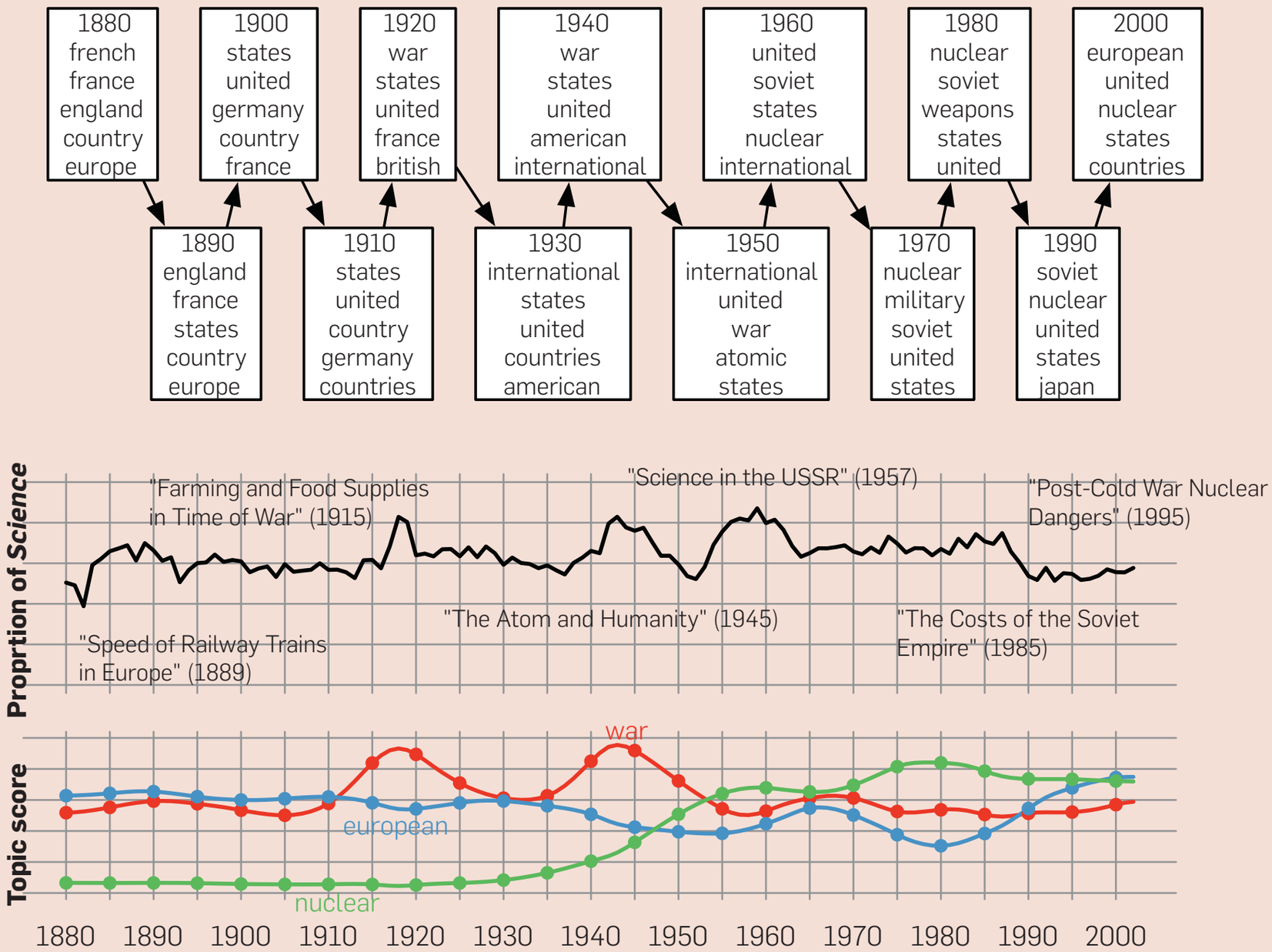
Active area of research: extensions of LDA

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.



Active area of research: extensions of LDA

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.



Summary

- Text mining is concerned with automated methods to
  - **Find**, **extract**, and **link information** from text
- Text clustering and topic models help us
  - **Organise** text corpora
  - **Find** relevant documents
  - **Uncover relations** between documents

Further reading

- David Blei (2012). Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.

- Christopher D. Manning & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA.