# Topic modelling only works if you have the right document collection

Iris Hendrickx
Center for Language Studies, RU, Nijmegen

Radboud University

# Is the Neutrality of Technology Possible or Desirable?

**How neutral are Topic Models ?**

# Topic modeling   (Latent Dirichlet Allocation)

**Goal**: Determine the common themes or topics  in a large text collection

- **Unsupervised**
  - Topic labels are not given
  - The number of topics needs to be pre-specified

# Topic modeling: Generative probalistic model

Assumption: documents are generated from an underlying, (latent) topic distribution and each document is generated from a mixture of these topics that each have a different proportion in the document.

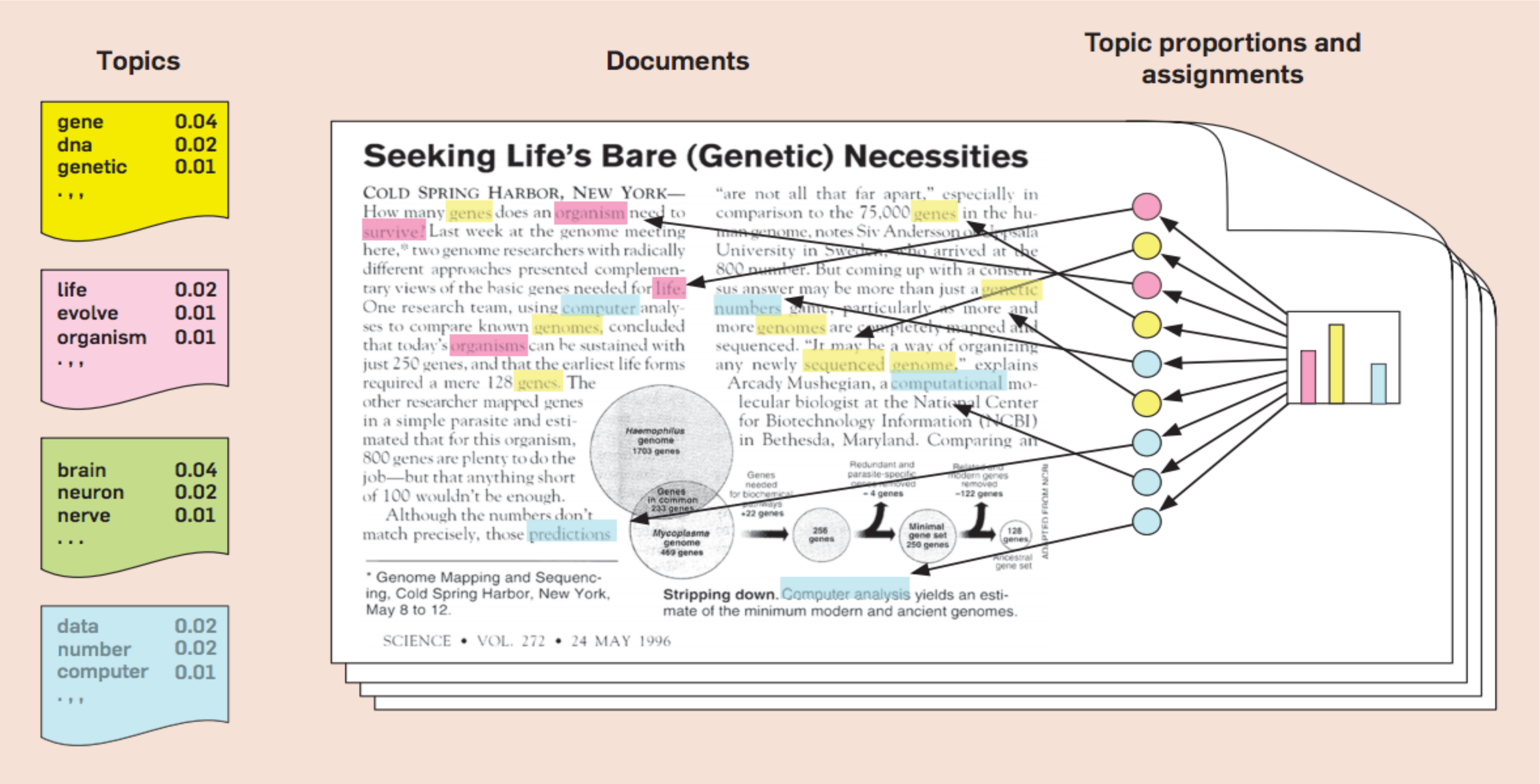
Topics defined as a distribution over words (a fixed vocabulary).

LDA uses an iterative process to estimate this underlying distribution based on the observed words in the text.

This model reflects the intuition that documents contain multiple topics.

Each document exhibits the topics in different proportion.

# Generative model



Topics

| | |
|---|---|
| gene | 0.04 |
| dna | 0.02 |
| genetic | 0.01 |
| . , , | |

| | |
|---|---|
| life | 0.02 |
| evolve | 0.01 |
| organism | 0.01 |
| . , , | |

| | |
|---|---|
| brain | 0.04 |
| neuron | 0.02 |
| nerve | 0.01 |
| . . . | |

| | |
|---|---|
| data | 0.02 |
| number | 0.02 |
| computer | 0.01 |
| . , , | |

Documents

Topic proportions and assignments

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

picture from David Blei (2012)

# Dream analysis

Despite the various research fields that study the meaning and purpose of dreams, such as psychiatry, psychology, neuroscience, and religious studies, a definitive explanation of the purpose of dreams is still lacking.

Psychologists and social scientists have studied dream content with quantitative methods for decades, working with the hypothesis that dreams reveal psychological information about the dreamer.

**Continuity hypothesis**: the content of dreams reflects a persons' daily life and personal concerns  (Domhoff, 1996)

75--80\% of dream content relates to everyday settings, characters, and activities

# topic modeling on dreambank.net (30k dream reports)

Random selection of the 50 topics:

44 money pay get give buy bank bill machine change
37 bathroom water toilet shower use clean bath floor sink
25 class school teacher students high test room classroom college
42 room door house see window open apartment go living
5 road hill tree see walking snow mountain trees people
28 love feel kiss make happy want man other hug
35 say says do see go man woman comes get
48 said did went came got told started saw looked asked

→ clear support for the **continuity hypothesis** as they reflect daily life events, characters and settings.

## Guidelines for using LDA

- LDA only works on **large** unstructured document collections (no matter how long each document is)

- **document length**:
  - **too short**: LDA does not work
  - **too long**: useless (fraction will give the same result)

- **number of topics**: too many leads to inefficiency /no-convergence

- LDA works best if:
  - **topics** consist of a few high probable words
  - **documents** consist of only a small set of high probable topics
  - (and rest has low probability)

- be aware: each run with LDA will result in slightly different topics (random start)

(Tang et al, 2014)

# How neutral are Topic Models?

**Neutral:**

- Unsupervised
- Reflect only the content collection and its word frequencies
- up to user to determine the value of the produced topics

**But:**

- Collection needs to adhere to certain prerequisites
- The assumptions and algorithmic implementation determine the usefulness of this method

# Thank you for your attention!

# References

- Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.
- Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In: *ICML*. 2014. p. 190-198.