Topic Modeling. LDA.

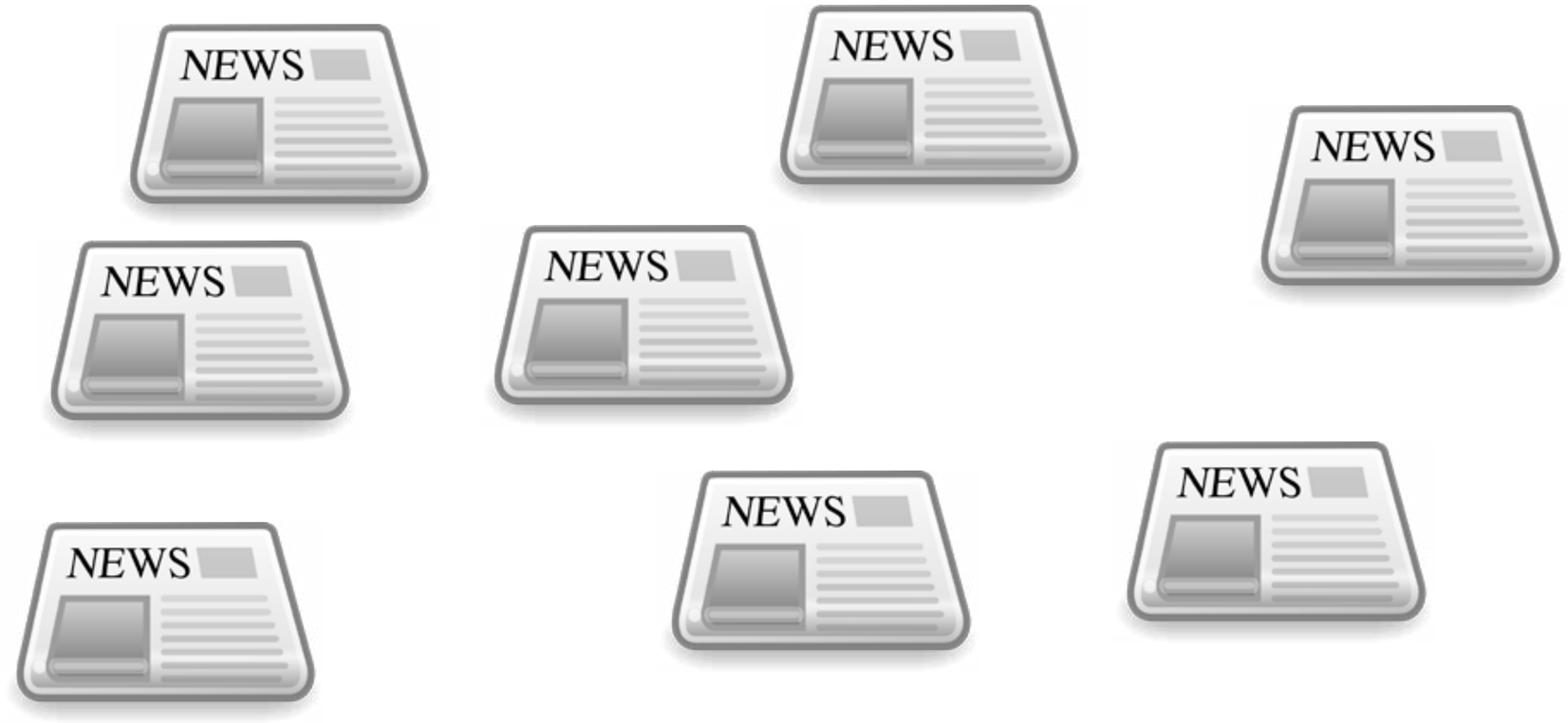
Topic modeling

- **Topic modeling** – unsupervised learning algorithm for discovering latent “topics” in a collection of documents.

- **Applications of topic modeling:**
  -Document classification
  -Document clustering and visualization
  -Recommender systems

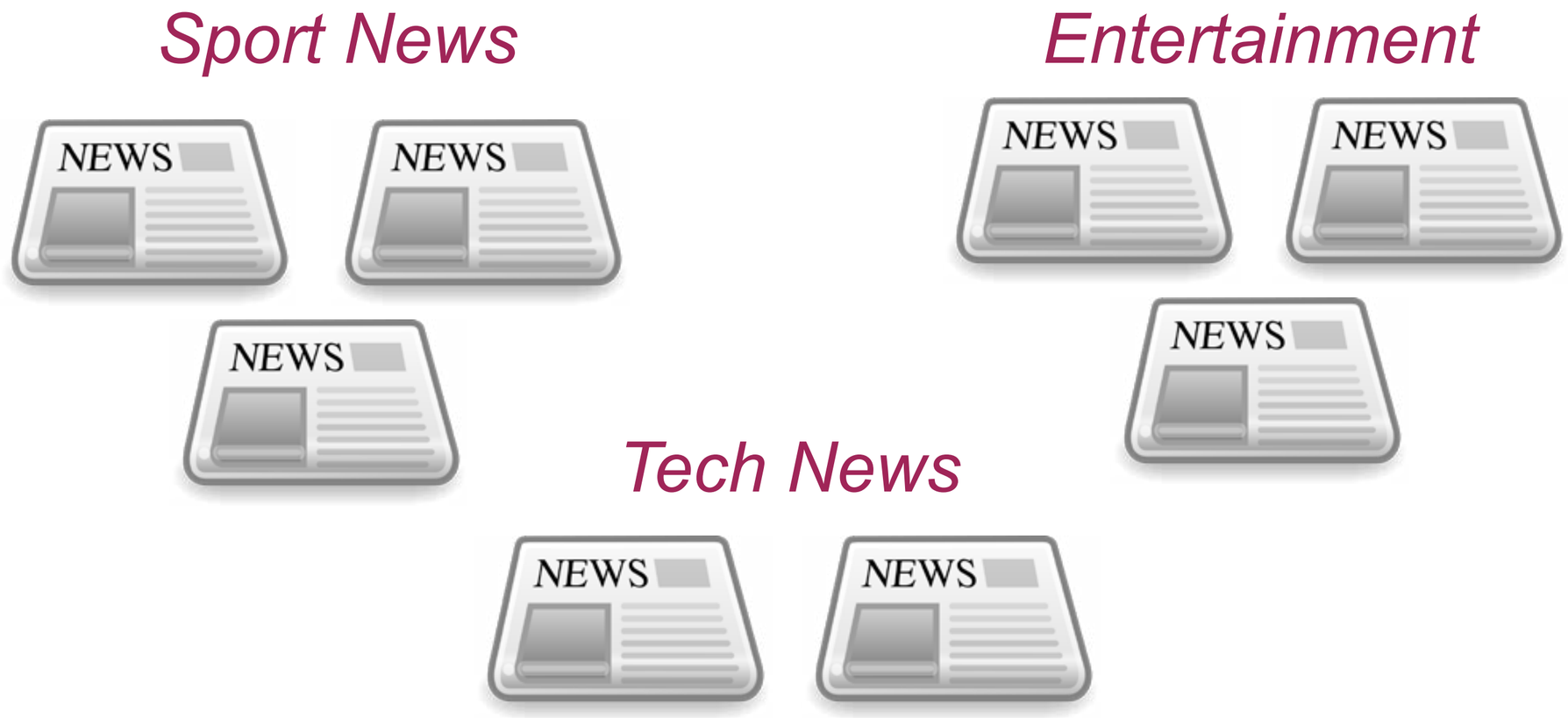
Application of topic modeling

- News clustering



Application of topic modeling

- News clustering



Assumptions:

- **w** – word, **d** – document, **t** – topic
- number of topics equals **T** and it is fixed
- Each topic **t** may generate a word **w** with probability **p**(**w**|**t**)
- Each document **d** has topic **t** with the probability **p**(**t**|**d**)

Topic Modeling Methods

- Latent Dirichlet allocation (LDA)

```
from pyspark.mllib.clustering import LDA, LDAModel
```

- Probabilistic latent semantic analysis (PLSA)

Topic Modeling Example

- Document collection – 17,000 articles of journal “Science”. Topic model (LDA) with 100 topics.

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

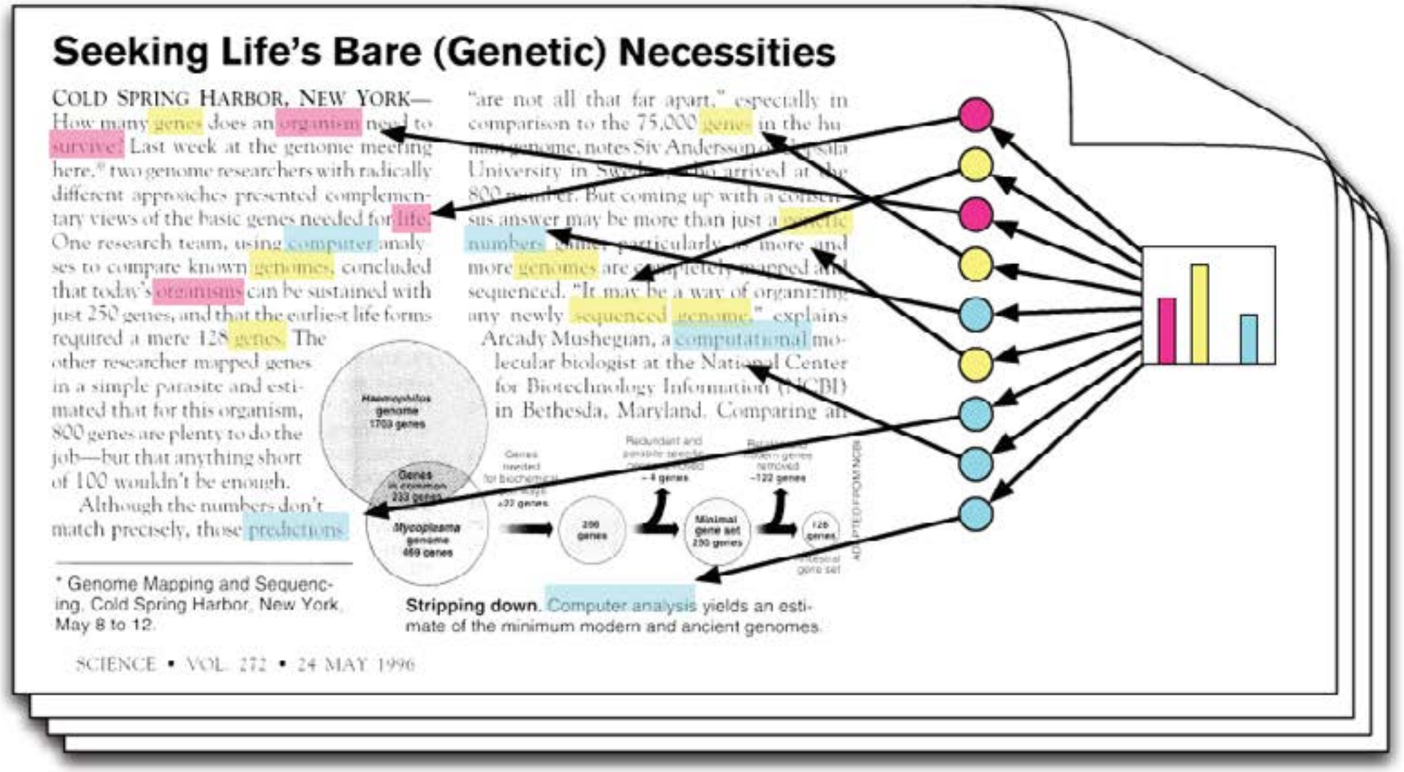
life	0.02
evolve	0.01
organism	0.01
...	

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

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

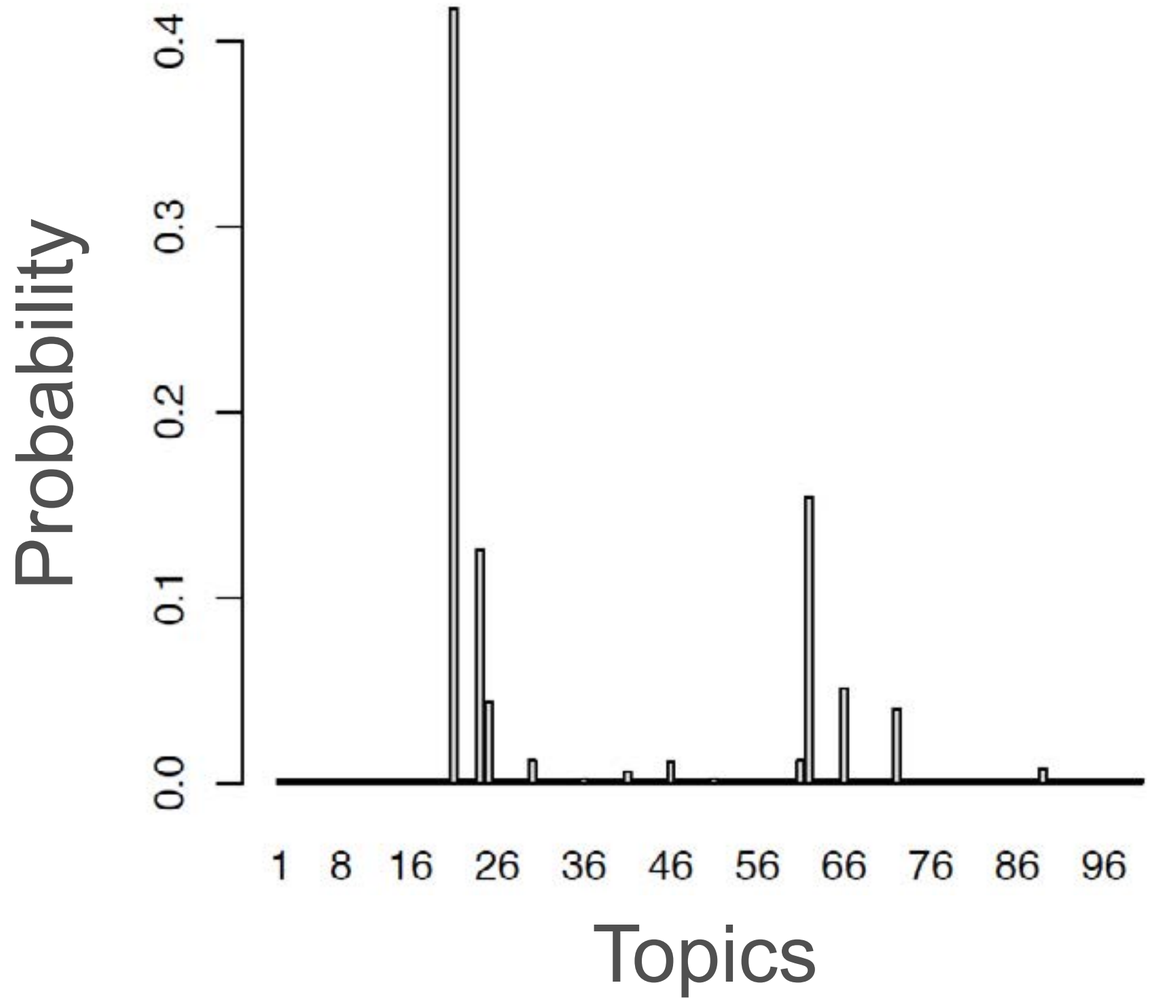
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Word Topics

For each topic top 5 words by **p**(**w**|**t**) are shown.

Topic 1	Topic 2	Topic 3	Topic 4
“Genetics”	**“Evolution”**	**“Disease”**	**“Computers”**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers

Document Topics



Topic Model Learning

- φ_{wt} = **p**(**w**|**t**) - a matrix **Φ**
- θ_{td} = **p**(**t**|**d**) - a matrix **Θ**
- n_{dw} = number of times which word **w** occurred in document **d**
- n_d = length of document **d**

$$\frac{n_{dw}}{n_d} \approx \sum_{t=1}^{T} \varphi_{wt}\, \theta_{td} = \Phi\Theta$$

Summary

- **Topic modeling** discovers latent “topics” in collections of documents
- Document labels are not required
- Two main algorithms for topic modeling are PLSA and LDA.
- Each topic has the list of most typical words
- Each document has the list of its topics