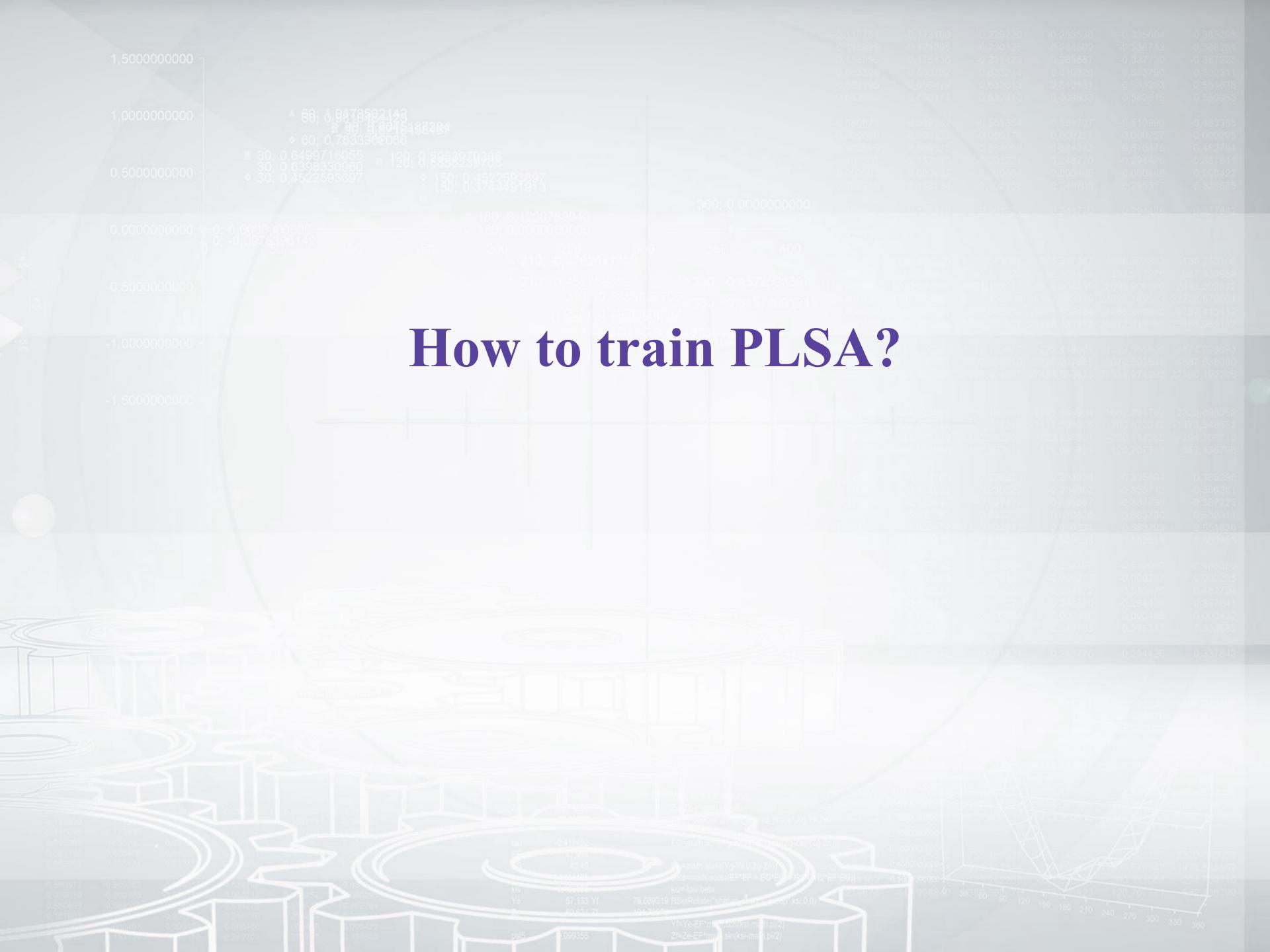


# How to train PLSA?



# How would you train the model?

## Probabilistic Latent Semantic Analysis:

$$p(w|d) = \sum_{t \in T} p(w|t) p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$

## Parameters of the model:

- $\phi_{wt}$  – probability of word  $w$  in topic  $t$
- $\theta_{td}$  – probability of topic  $t$  in document  $d$

# How would you train the model?

Log-likelihood optimization:

$$\log \prod_{d \in D} p(d) \prod_{w \in d} p(w|d)^{n_{dw}} \rightarrow \max_{\Phi, \Theta}$$



$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

Given non-negativity and normalization constraints:

$$\phi_{wt} \geq 0$$

$$\sum_{w \in W} \phi_{wt} = 1$$

$$\theta_{td} \geq 0$$

$$\sum_{t \in T} \theta_{td} = 1$$

# How would you train the model?

Log-likelihood optimization:

$$\log \prod_{d \in D} p(d) \prod_{w \in d} p(w|d)^{n_{dw}} \rightarrow \max_{\Phi, \Theta}$$



$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

Given non-negativity and normalization constraints:

$$\phi_{wt} \geq 0$$

$$\sum_{w \in W} \phi_{wt} = 1$$

$$\theta_{td} \geq 0$$

$$\sum_{t \in T} \theta_{td} = 1$$

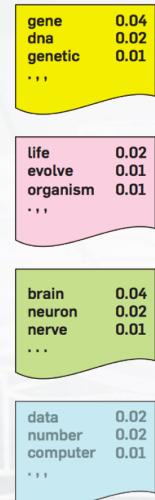
# We have just plain texts

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

# We have just plain texts

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

Topics



Documents

## 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.

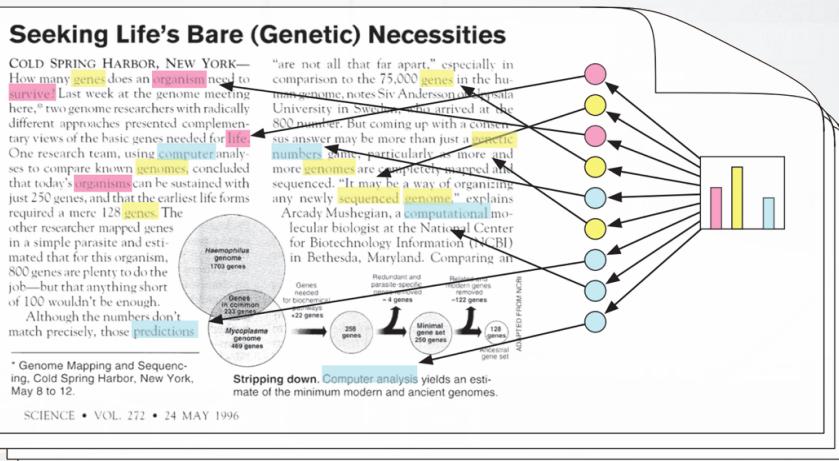
Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a scientist at Stockholm University in Sweden. "We arrived at the 800 number, but coming up with a consensus answer may be more than just a genetic number. Since, 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 all

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions  
and assignments



# If we knew topic assignments...

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

# If we knew topic assignments...

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

We would just count:

$$p(w = \text{sky} | \textcolor{red}{t}) = \frac{n_{w\textcolor{red}{t}}}{\sum_w n_{w\textcolor{red}{t}}} = \frac{1}{4}$$

# If we knew topic assignments...

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

We would just count:

$$p(w = \text{sky} | \textcolor{red}{t}) = \frac{n_{w\textcolor{red}{t}}}{\sum_w n_{w\textcolor{red}{t}}} = \frac{1}{4}$$

$$p(t = \textcolor{red}{t} | d) = \frac{n_{td}}{\sum_t n_{td}} = \frac{4}{54}$$

# But we have just plain texts

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

# But we have just plain texts

Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened again, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down...

**Idea! Let's estimate the topic assignment probabilities!**

$$p(t|d, w) = \frac{p(w, t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)}$$

*Bayes rule*      *Product rule*

# Put everything together: EM-algorithm

**E-step:**

$$p(t|d, w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}$$

**M-step:**

$$\phi_{wt} = \frac{n_{wt}}{\sum_w n_{wt}}$$

$$\theta_{td} = \frac{n_{td}}{\sum_t n_{td}}$$

$$n_{wt} = \sum_d n_{dw} p(t|d, w)$$

$$n_{td} = \sum_w n_{dw} p(t|d, w)$$