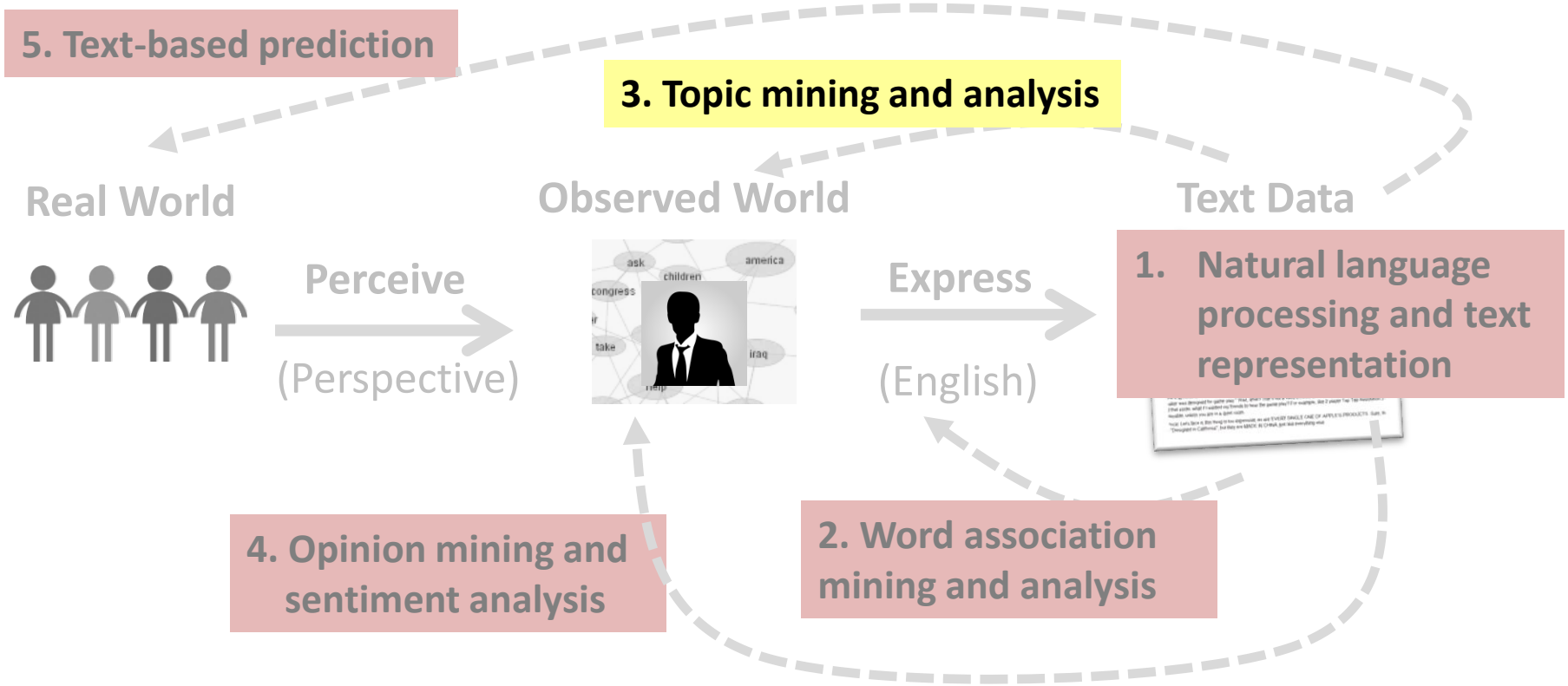
Text Clustering: Generative Probabilistic Models

Part 1

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Text Clustering: Generative Probabilistic Models (Part 1)



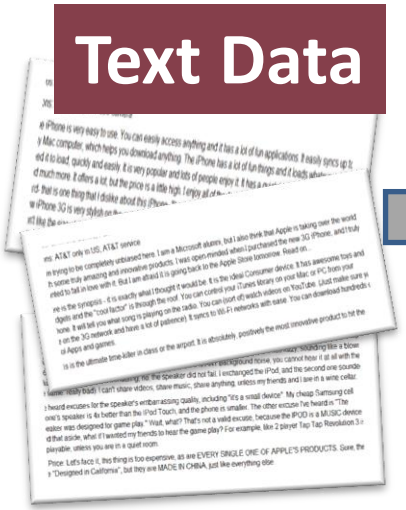
Overview

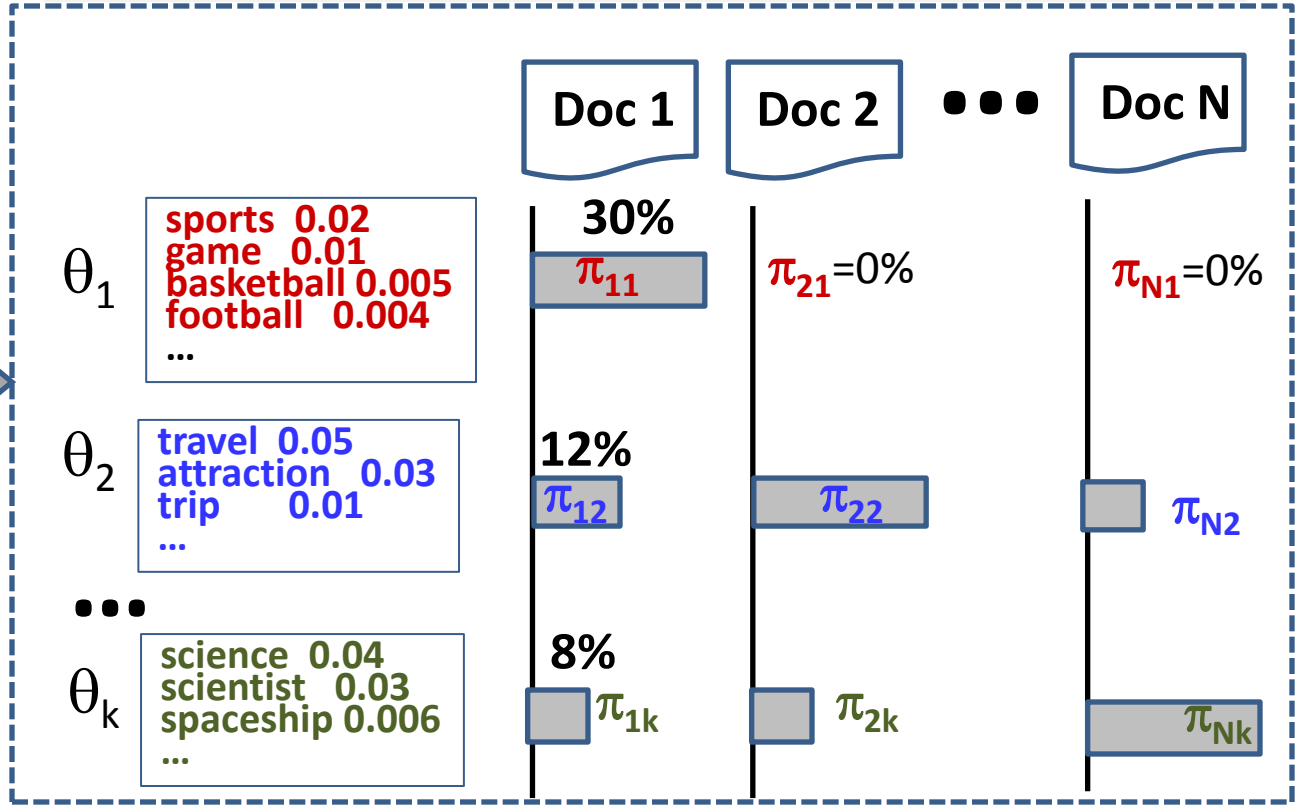
- What is text clustering?
- Why text clustering?
- How to do text clustering?
  - **Generative probabilistic models**
  - Similarity-based approaches
- How to evaluate clustering results?

Topic Mining Revisited

INPUT: C, k, V

OUTPUT: $\{\theta_1, \ldots, \theta_k\}, \{\pi_{i1}, \ldots, \pi_{ik}\}$



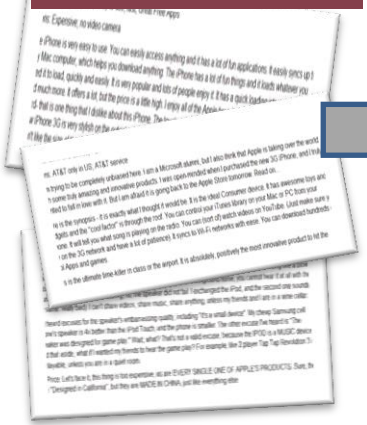


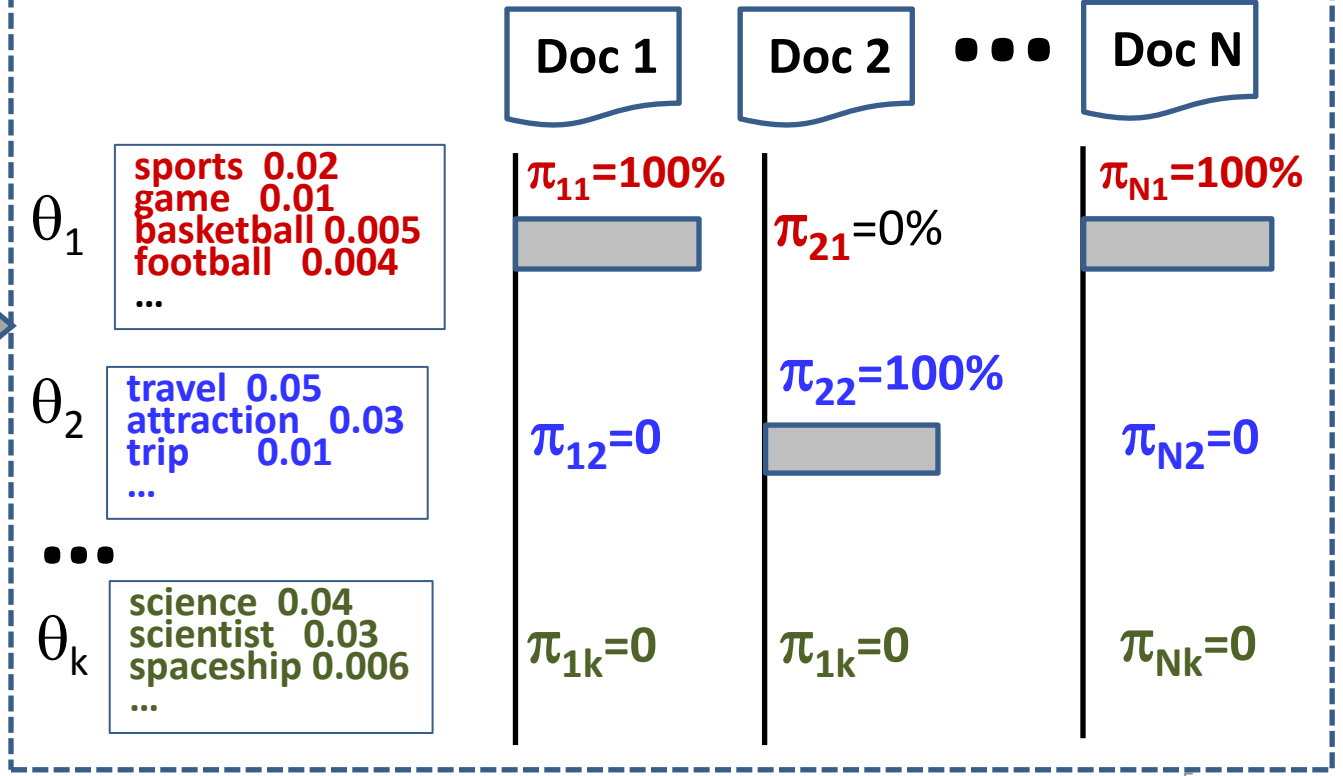
One Topic(=cluster) Per Document

INPUT: C, k, V

OUTPUT: $\{\theta_1, \ldots, \theta_k\}$, $\{c_1, \ldots, c_N\}$ $c_i \in [1,k]$

Text Data



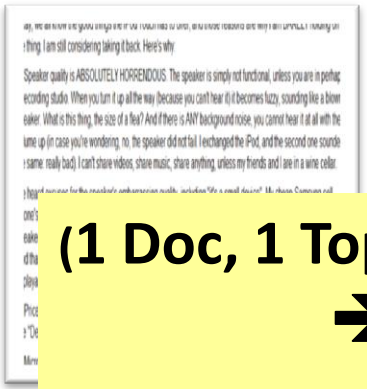


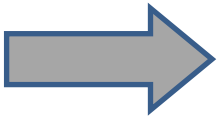
Mining One Topic Revisited

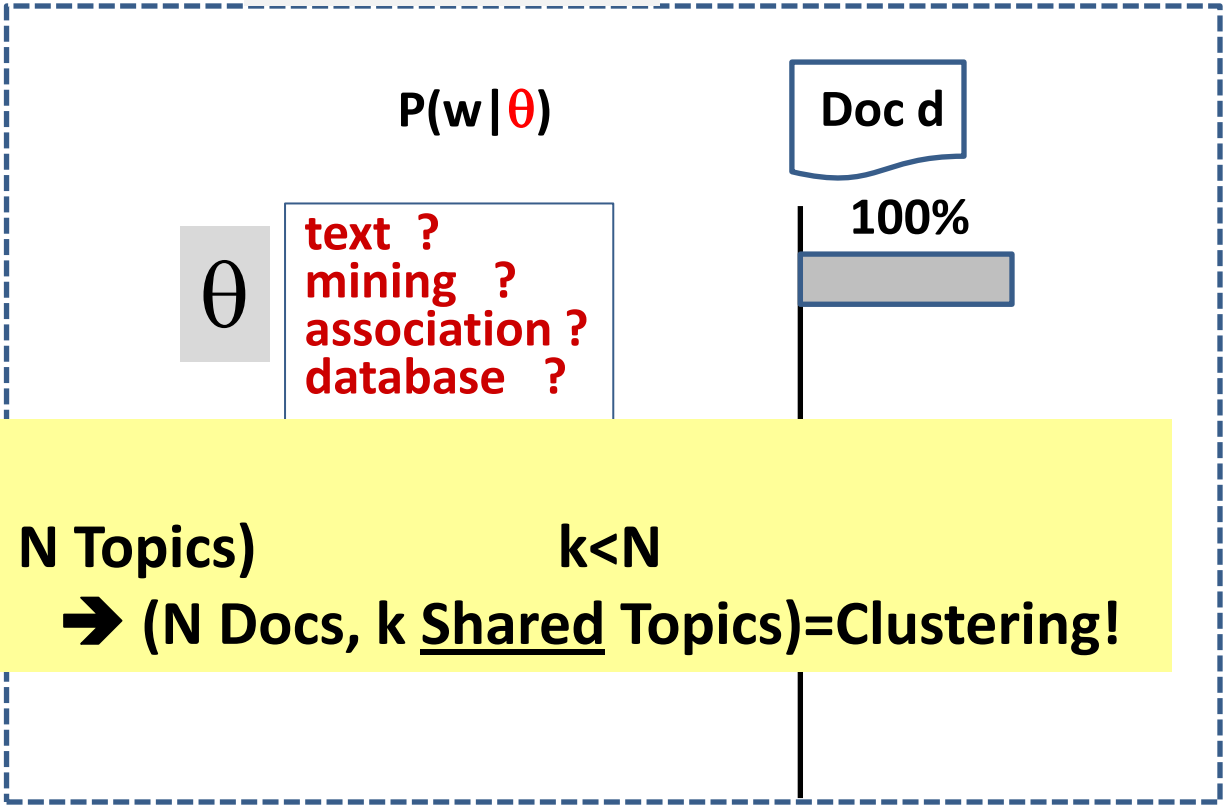
INPUT: C={d}, V

OUTPUT: { θ}

Text Data







(1 Doc, 1 Topic)

➔ (N Docs, N Topics) **k<N**

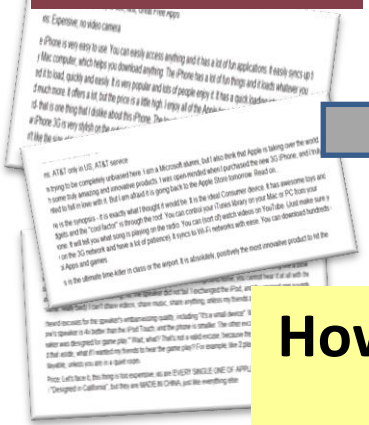
➔ (N Docs, k <u>Shared</u> Topics)=Clustering!

What Generative Model Can Do Clustering?

INPUT: C, k, V

OUTPUT: $\{\theta_1, \ldots, \theta_k\}$, $\{c_1, \ldots, c_N\}$ $c_i \in [1,k]$

Text Data



Doc 1 | Doc 2 | ••• | Doc N

θ_1: sports 0.02, game 0.01, basketball 0.005, football 0.004, …

π_{11}=100% π_{21}=0% π_{N1}=100%

θ_2: travel 0.05, attraction 0.03, trip 0.01

π_{12}=0 π_{22}=100% π_{N2}=0

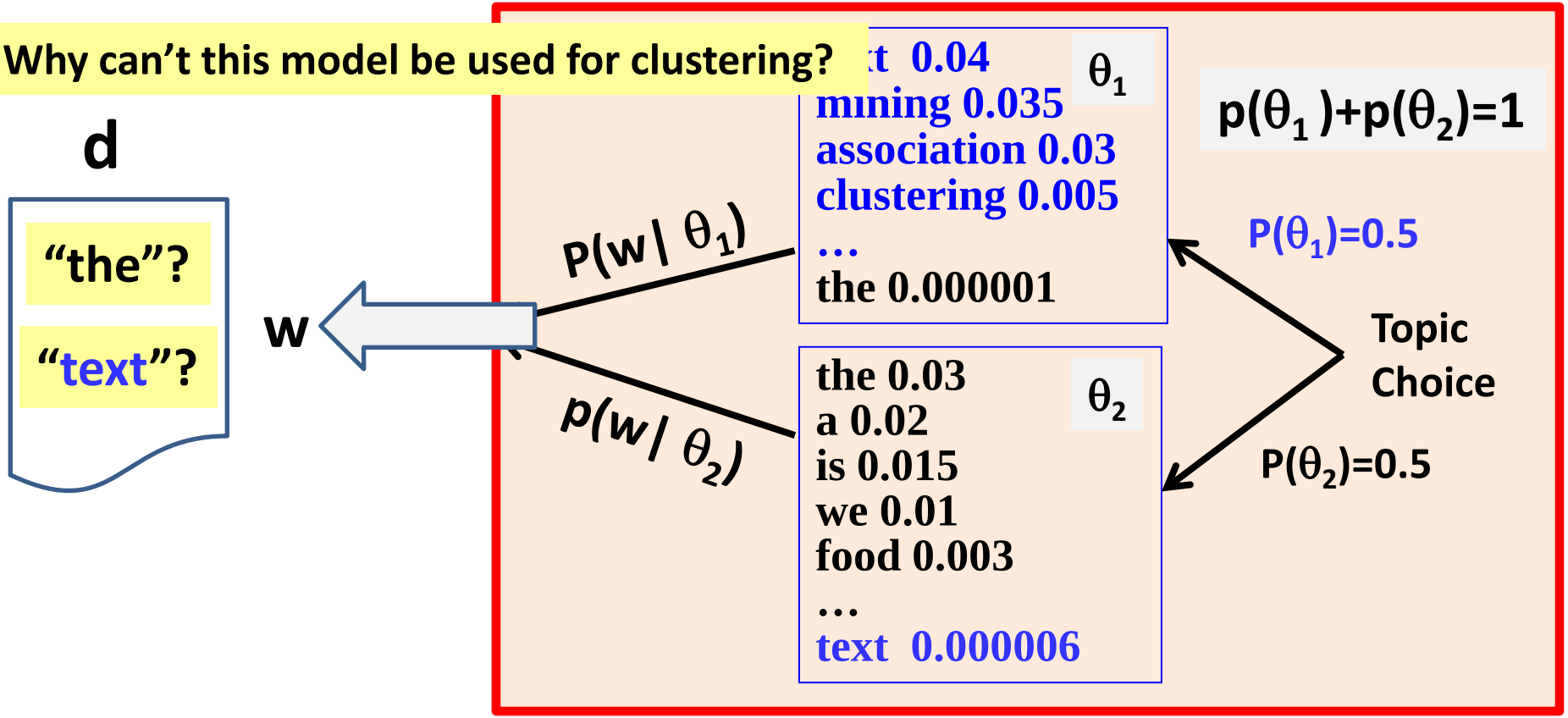
θ_k: spaceship 0.006, …

π_{1k} π_{1k} π_{Nk}=0

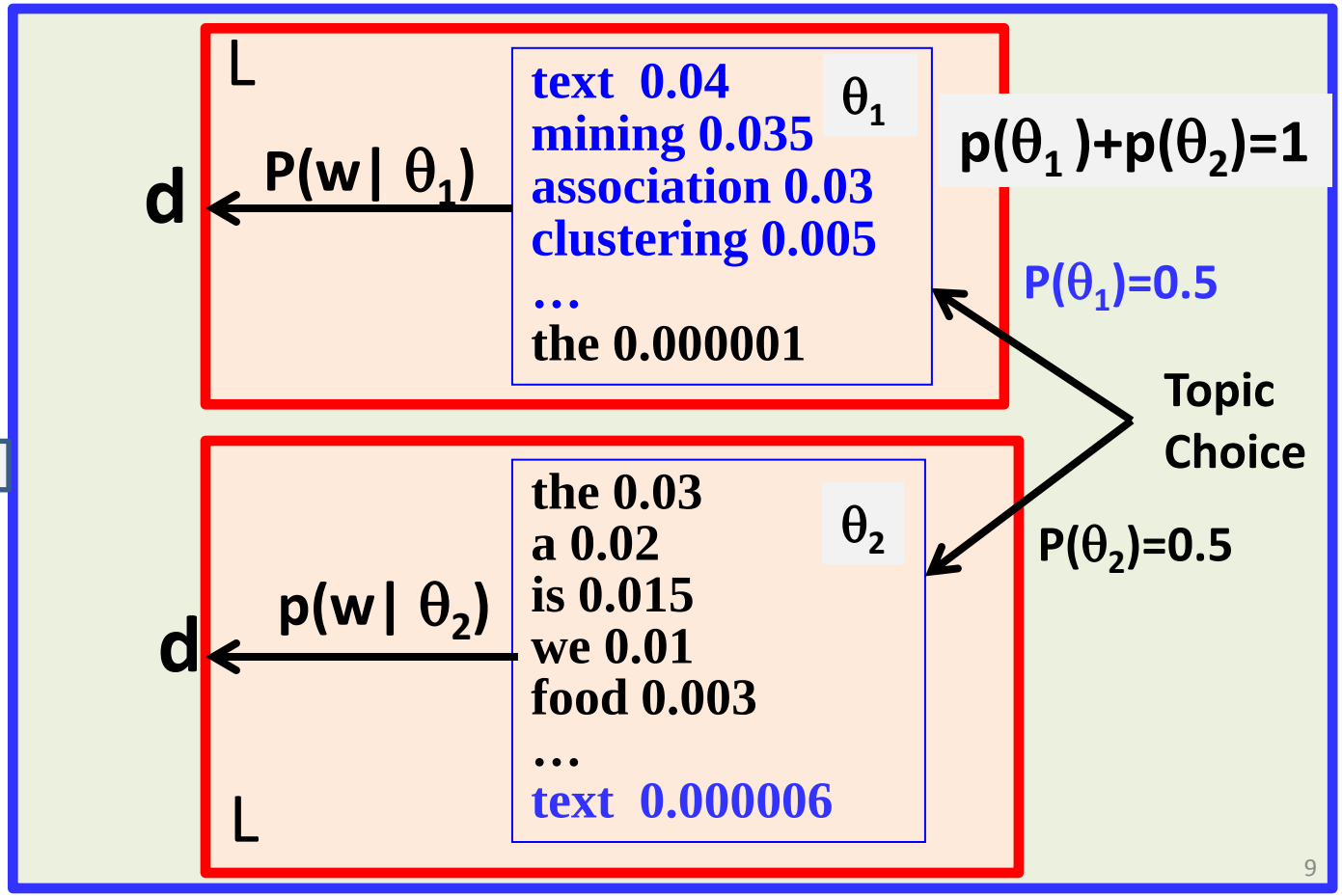
How can we force every document to be generated using one topic (instead of k topics)?

Generative Topic Model Revisited

Why can't this model be used for clustering?



Mixture Model for Document Clustering



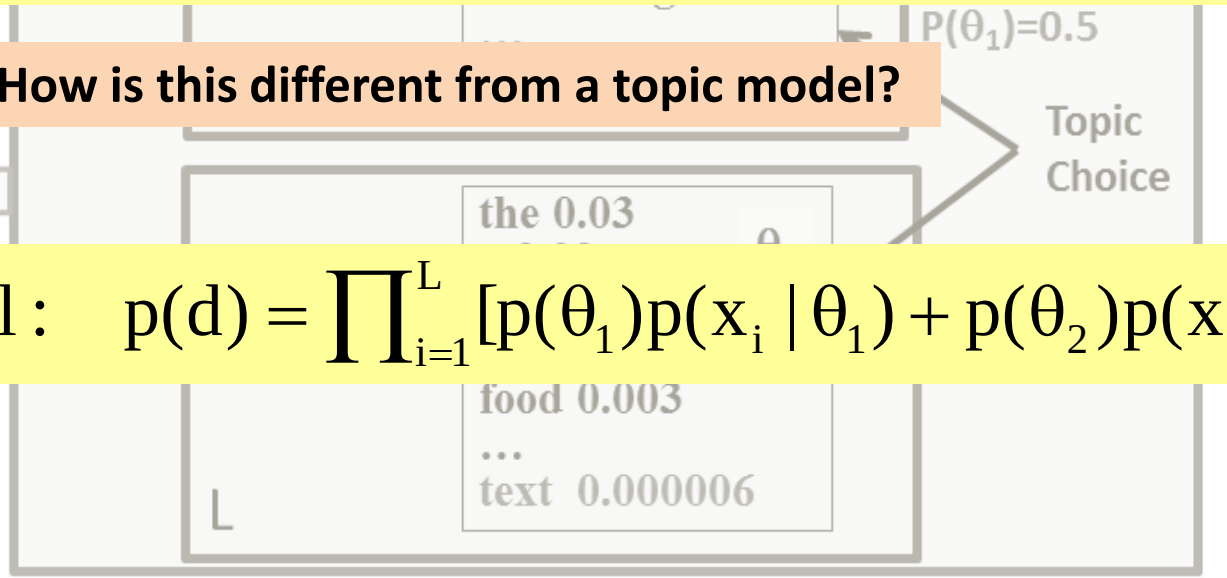
Likelihood Function: p(d)=?

$$p(d) = p(\theta_1)p(d \mid \theta_1) + p(\theta_2)p(d \mid \theta_2)$$

$$= p(\theta_1)\prod_{i=1}^{L} p(x_i \mid \theta_1) + p(\theta_2)\prod_{i=1}^{L} p(x_i \mid \theta_2)$$

d=x_1 x_2 … x_L

How is this different from a topic model?



$$\text{topic model}: \quad p(d) = \prod_{i=1}^{L} [p(\theta_1)p(x_i \mid \theta_1) + p(\theta_2)p(x_i \mid \theta_2)]$$