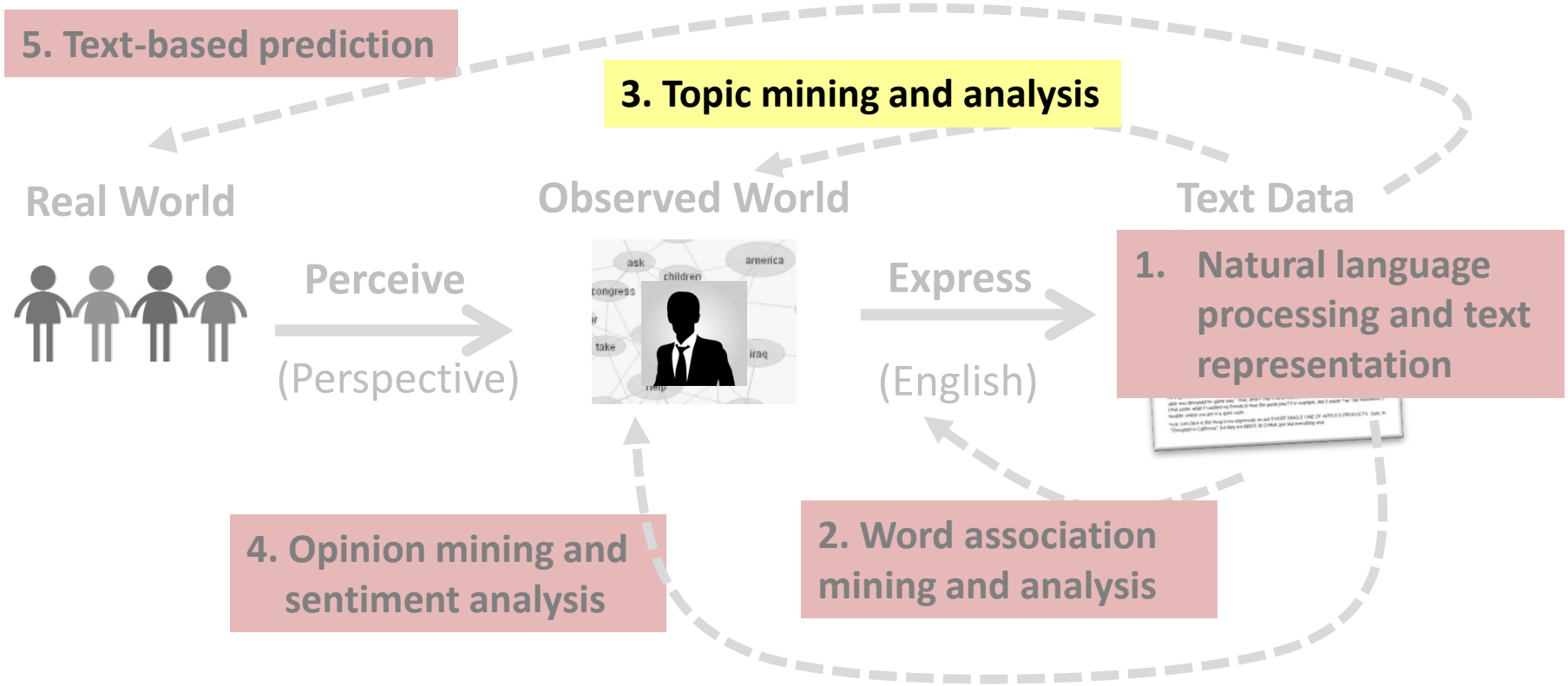
Text Clustering: Generative Probabilistic Models

Part 2

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

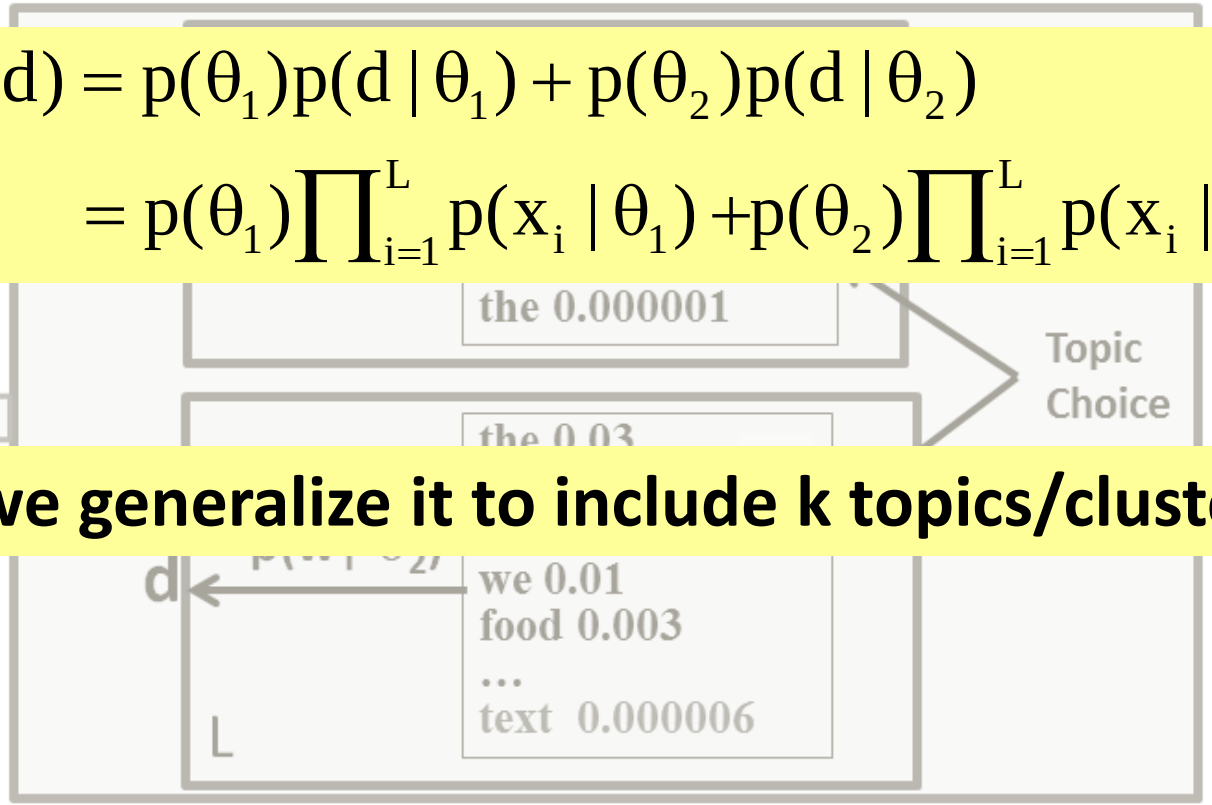
Text Clustering: Generative Probabilistic Models (Part 2)



Likelihood Function: p(d)=?

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

$d=x_1\ x_2 \ldots x_L$



How can we generalize it to include k topics/clusters?

Mixture Model for Document Clustering

- Data: a collection of documents $C=\{d_1, \ldots, d_N\}$
- Model: mixture of k unigram LMs: $\Lambda=(\{\theta_i\}; \{p(\theta_i)\})$, $i \in [1,k]$
  - To generate a document, first **choose a θ_i** according to $p(\theta_i)$, and then generate **all** words in the document using $p(w|\theta_i)$
- Likelihood:

$$p(d \mid \Lambda) = \sum_{i=1}^{k} [p(\theta_i) \prod_{j=1}^{|d|} p(x_j \mid \theta_i)]$$

$$= \sum_{i=1}^{k} [p(\theta_i) \prod_{w \in V} p(w \mid \theta_i)^{c(w,d)}]$$

- Maximum Likelihood estimate

$$\Lambda^* = \arg\max_{\Lambda} p(d \mid \Lambda)$$

Cluster Allocation After Parameter Estimation

- **Parameters** of the mixture model: $\Lambda=(\{\theta_i\}; \{p(\theta_i)\})$, $i \in [1,k]$
  - Each θ_i represents the **content of cluster i** : **$p(w|\ \theta_i)$**
  - **$p(\theta_i)$** indicates the **size of cluster i**
  - Note that unlike in PLSA, $p(\theta_i)$ doesn't depend on d!
- Which cluster should document d belong to? c_d=?
  - **Likelihood only**: Assign d to the cluster corresponding to the topic θ_i that most likely has been used to generate d

$$c_d = \arg\max_i p(d \mid \theta_i)$$

  - **Likelihood + prior $p(\theta_i)$ (Bayesian):** favor large clusters

$$c_d = \arg\max_i p(d \mid \theta_i)p(\theta_i)$$