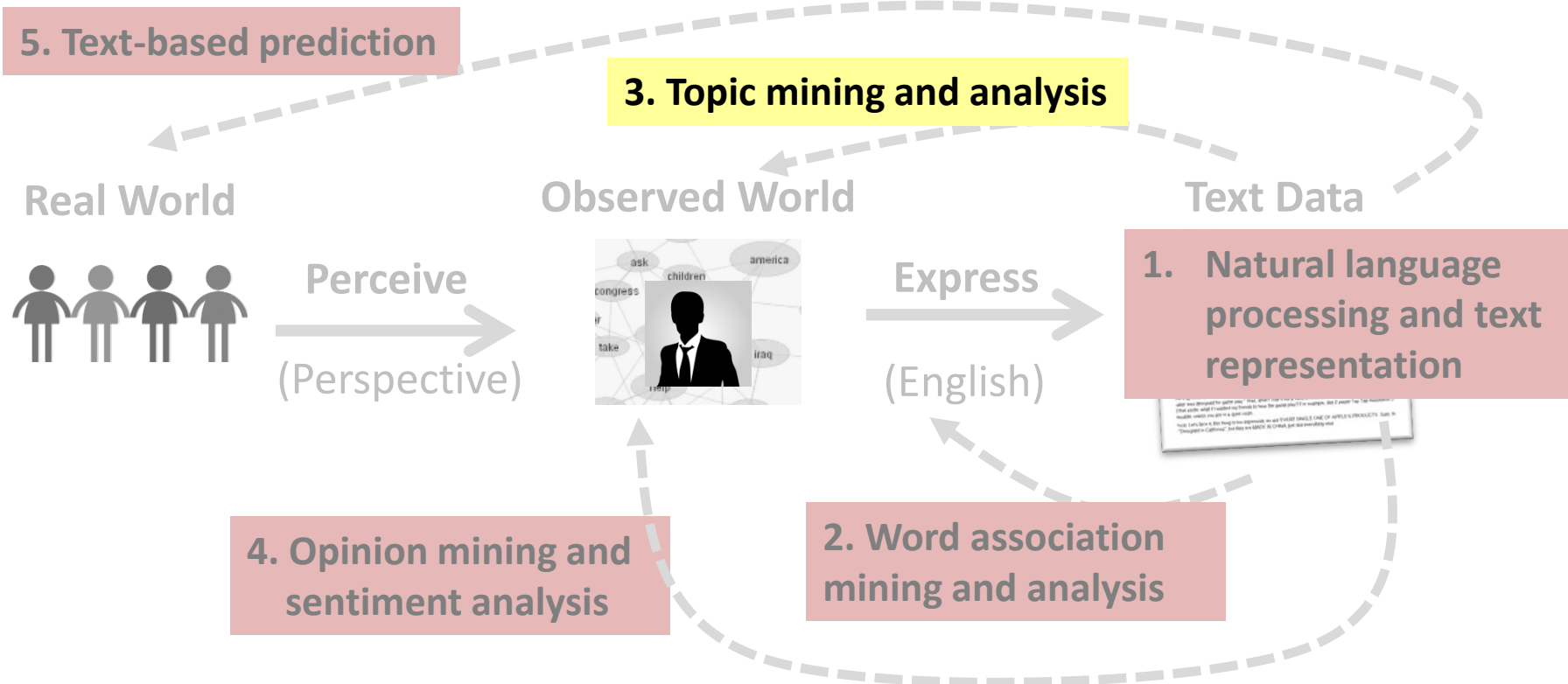


# Topic Mining and Analysis: Motivation and Task Definition

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

# Topic Mining and Analysis: Motivation and Task Definition



# Topic Mining and Analysis: Motivation

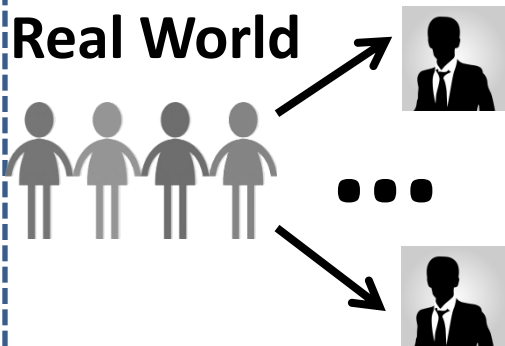
- Topic  $\approx$  main idea discussed in text data
  - Theme/subject of a discussion or conversation
  - Different granularities (e.g., topic of a sentence, an article, etc.)
- Many applications require discovery of topics in text
  - What are Twitter users talking about today?
  - What are the current research topics in data mining? How are they different from those 5 years ago?
  - What do people like about the iPhone 6? What do they dislike?
  - What were the major topics debated in 2012 presidential election?

# Topics As Knowledge About the World

Knowledge about the world

Non-Text Data

Real World



Text Data



+ Context  
Time  
Location  
...

Topic 1

Topic 2

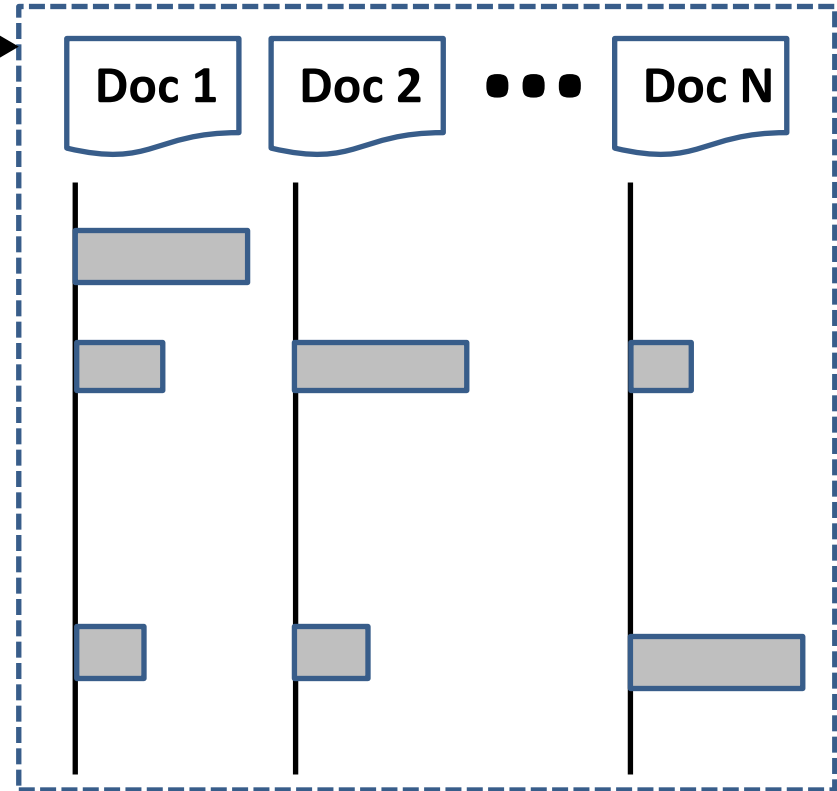
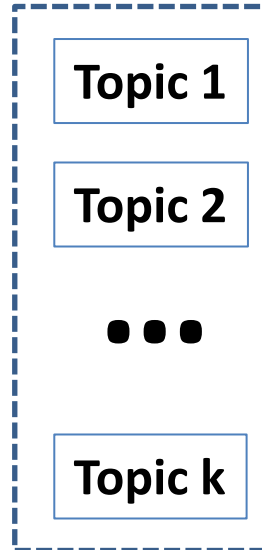
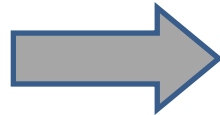
...

Topic k

# Tasks of Topic Mining and Analysis

**Task 2: Figure out which documents cover which topics**

**Text Data**



**Task 1: Discover k topics**

# Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents  **$C = \{d_1, \dots, d_N\}$**
  - **Number of topics:  $k$**
- Output
  - **$k$  topics:  $\{\theta_1, \dots, \theta_k\}$**
  - **Coverage of topics in each  $d_i$ :  $\{\pi_{i1}, \dots, \pi_{ik}\}$**
  - $\pi_{ij}$  = prob. of  $d_i$  covering topic  $\theta_j$

$$\sum_{j=1}^k \pi_{ij} = 1$$

**How to define  $\theta_i$  ?**