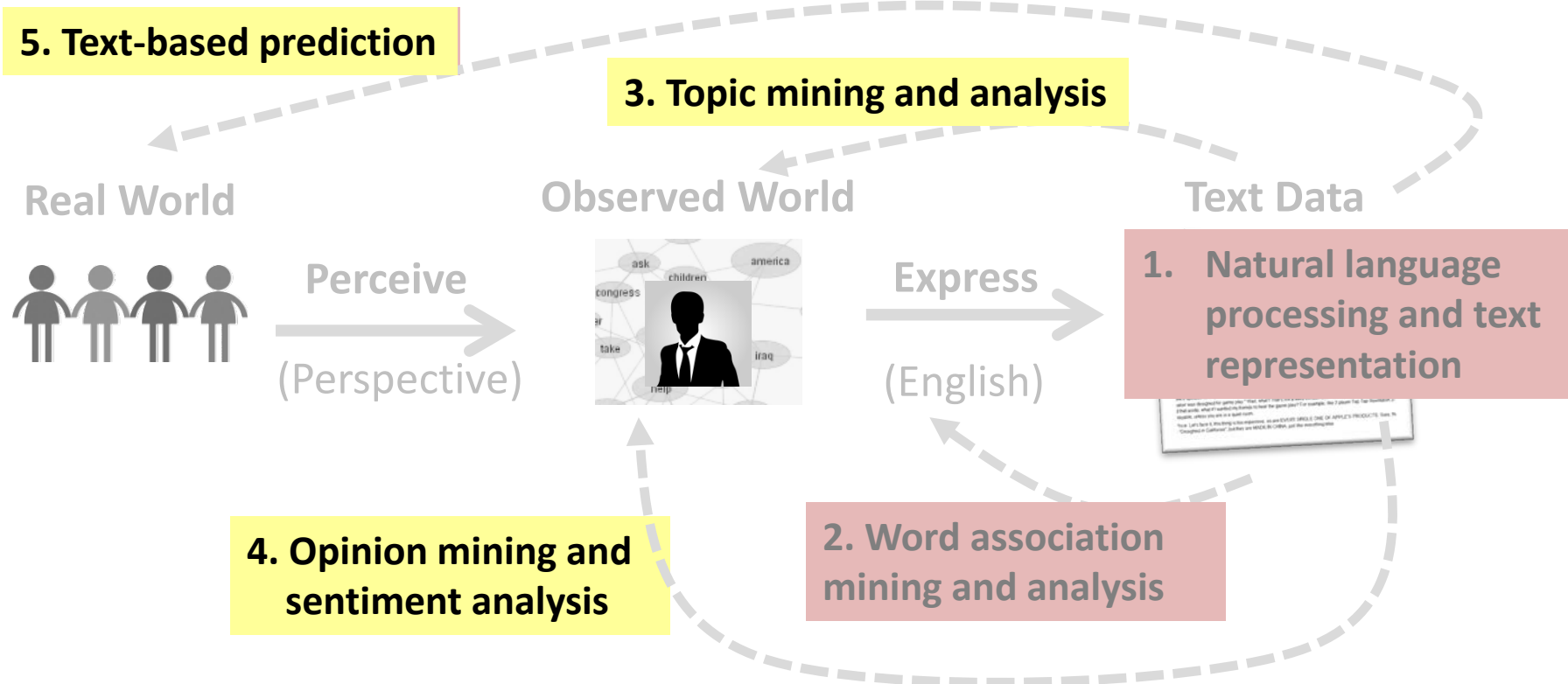


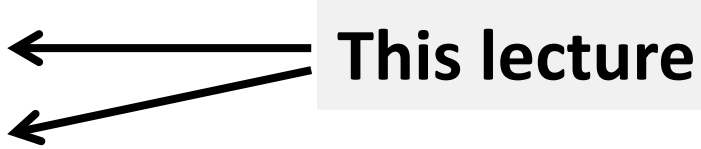
# Text Categorization: Motivation

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

# Text Categorization

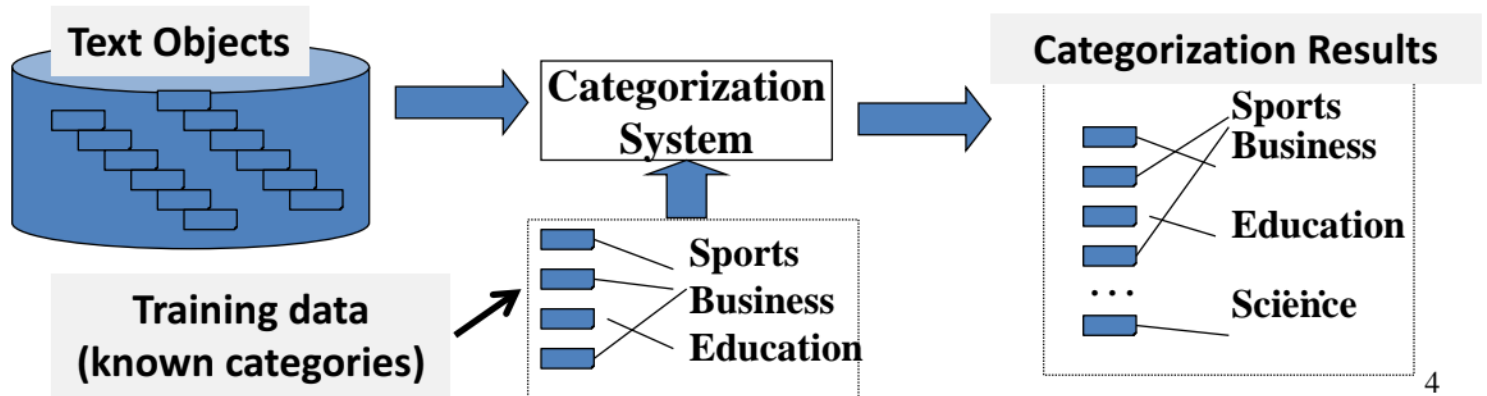


# Overview

- What is text categorization? ←
  - Why text categorization? ←
  - How to do text categorization?
    - Generative probabilistic models
    - Discriminative approaches
  - How to evaluate categorization results?
- This lecture**
- 

# Text Categorization

- Given the following:
  - A set of predefined categories, possibly forming a hierarchy *and often*
  - A training set of labeled text objects
- Task: **Classify** a text object into **one or more** of the **categories**



# Examples of Text Categorization

- **Text objects can vary** (e.g., documents, passages, or collections of text)
- **Categories can also vary**
  - “**Internal**” categories that characterize a text object (e.g., topical categories, sentiment categories)
  - “**External**” categories that characterize an entity associated with the text object (e.g., author attribution or any other meaningful categories associated with text data)
- Some **examples of applications**
  - News categorization, literature article categorization (e.g., MeSH annotations)
  - Spam email detection/filtering
  - Sentiment categorization of product reviews or tweets
  - Automatic email sorting/routing
  - Author attribution

# Variants of Problem Formulation

- **Binary** categorization: Only two categories
  - Retrieval: {relevant-doc, non-relevant-doc}
  - Spam filtering: {spam, non-spam}
  - Opinion: {positive, negative}
- **K-category** categorization: More than two categories
  - Topic categorization: {sports, science, travel, business,...}
  - Email routing: {folder1, folder2, folder3,...}
- **Hierarchical** categorization: Categories form a hierarchy
- **Joint** categorization: Multiple **related** categorization tasks done in a joint manner

Binary categorization can potentially support all other categorizations

# Why Text Categorization?

- To **enrich text representation** (more understanding of text)
  - Text can now be represented in multiple levels (keywords + categories)
  - Semantic categories assigned can be directly or indirectly useful for an application
  - Semantic categories facilitate aggregation of text content (e.g., aggregating all positive/negative opinions about a product)
- To **infer properties of entities** associated with text data (discovery of **knowledge about the world**)
  - As long as an entity can be associated with text data, we can always use the text data to help categorize the associated entities
  - E.g., discovery of non-native speakers of a language; prediction of party affiliation based on a political speech