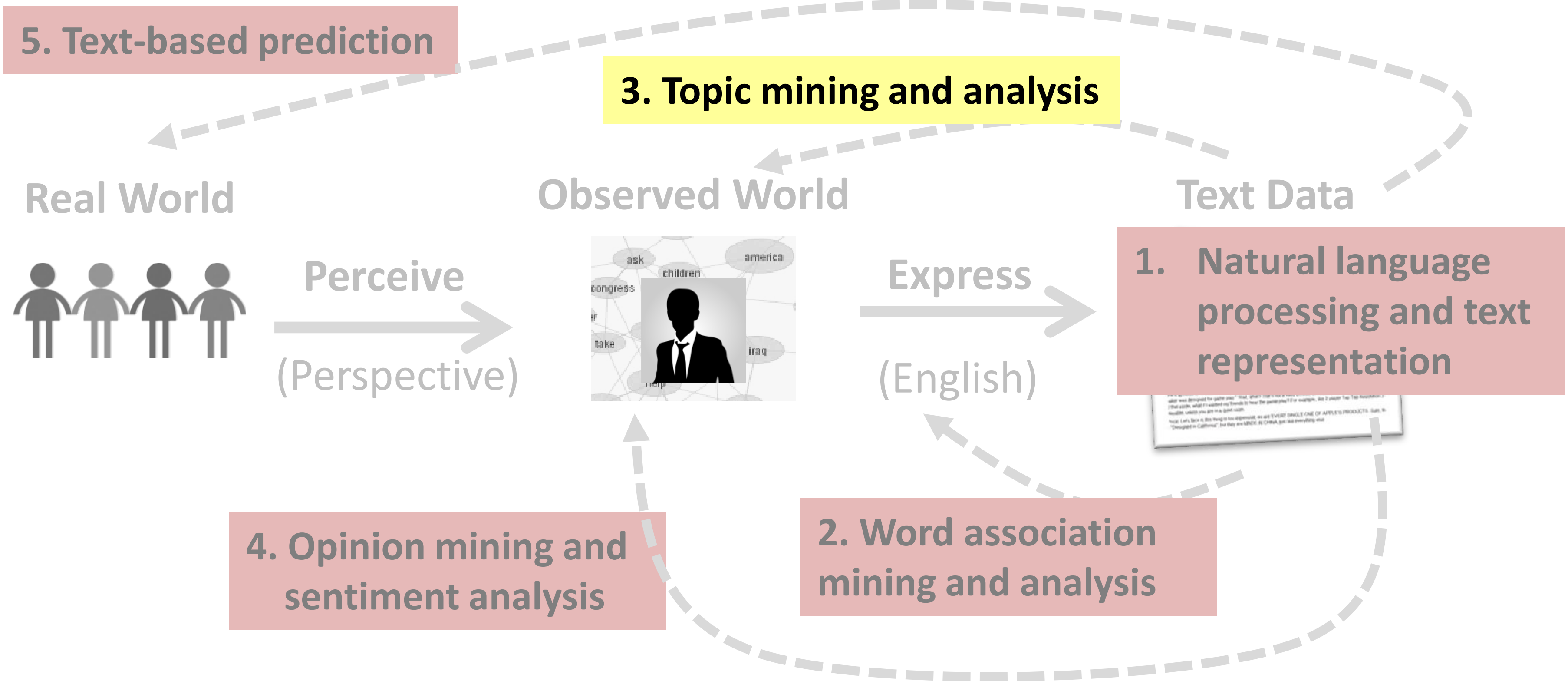
Text Clustering: Motivation

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

Text Clustering: Motivation



5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation

4. Opinion mining and sentiment analysis

2. Word association mining and analysis

Overview

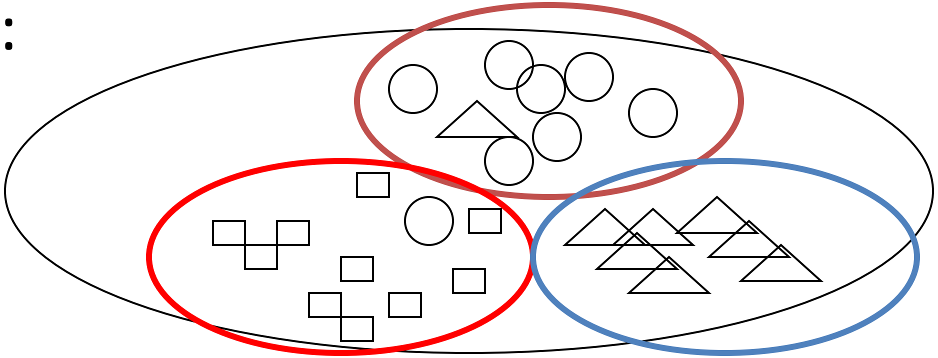
- What is text clustering?
- Why text clustering?

This lecture

- How to do text clustering?
  - Generative probabilistic models
  - Other approaches
- How to evaluate clustering results?

What Is Text Clustering?

- Discover “natural structure”
- Group similar objects together
- Objects can be documents, terms, passages, websites,…
- Example:

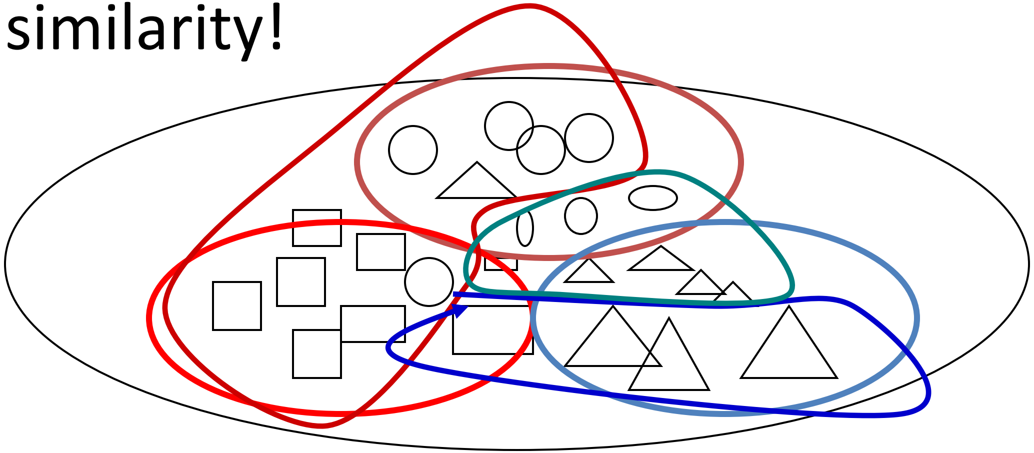


Not well defined! **What does “similar” mean?**

The “Clustering Bias”

- Any two objects can be similar, depending on how you look at them!
- Are “car” and “horse” similar?
- A user must define the **perspective** (i.e., a “**bias**”) for assessing similarity!

Basis for evaluation



Examples of Text Clustering

- Clustering of documents in the whole collection
- Term clustering to define “concept”/“theme”/“topic”
- Clustering of passages/sentences or any selected text segments from larger text objects (e.g., all text segments about a topic discovered using a topic model)
- Clustering of websites (text object has multiple documents)
- Text clusters can be further clustered to generate a hierarchy

Why Text Clustering?

- In general, very useful for text mining and <u>exploratory</u> text analysis:
  - ➔Get a sense about the overall content of a collection (e.g., what are some of the "typical"/representative documents in a collection?)
  - ➔Link (similar) text objects (e.g., removing duplicated content)
  - ➔Create a structure on the text data (e.g., for browsing)
  - ➔As a way to induce additional features (i.e., clusters) for classification of text objects
- Examples of applications
  - – Clustering of search results
  - – Understanding major complaints in emails from customers