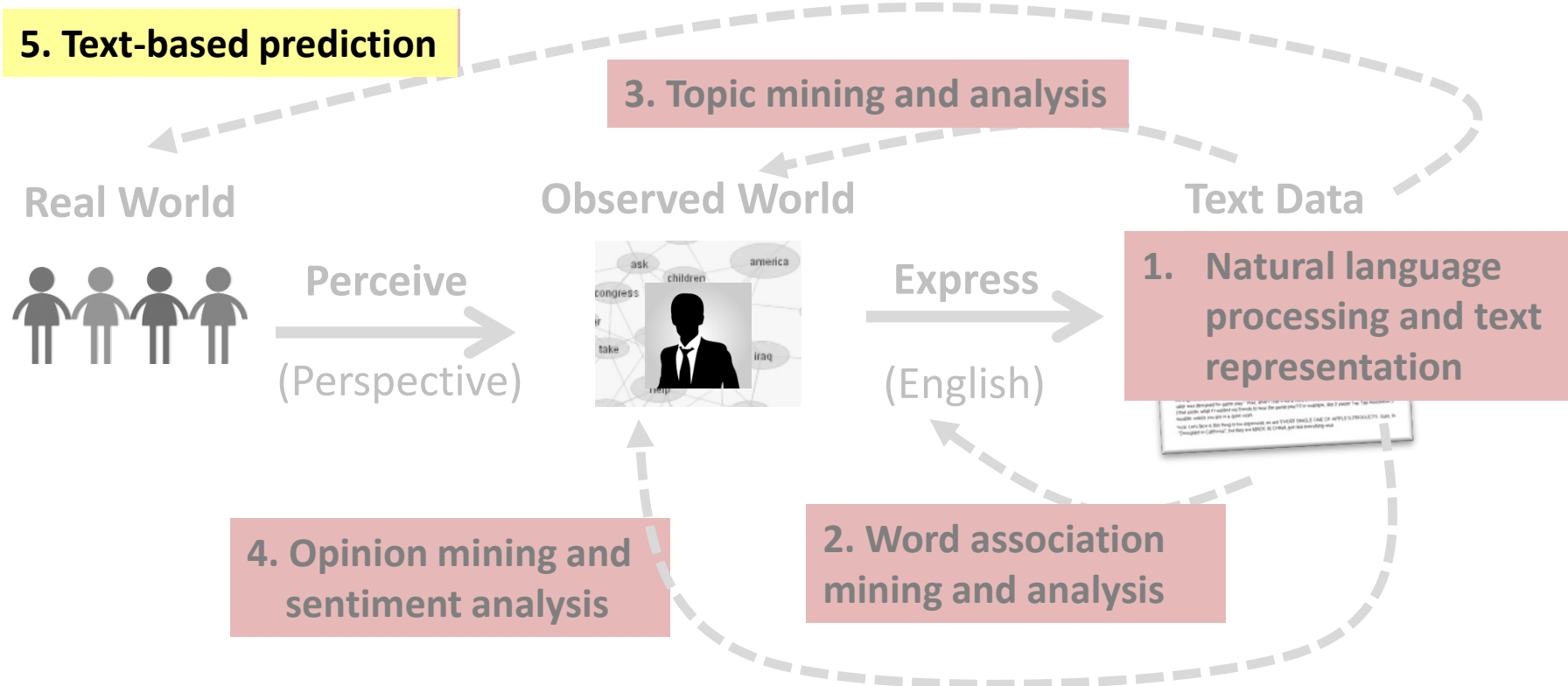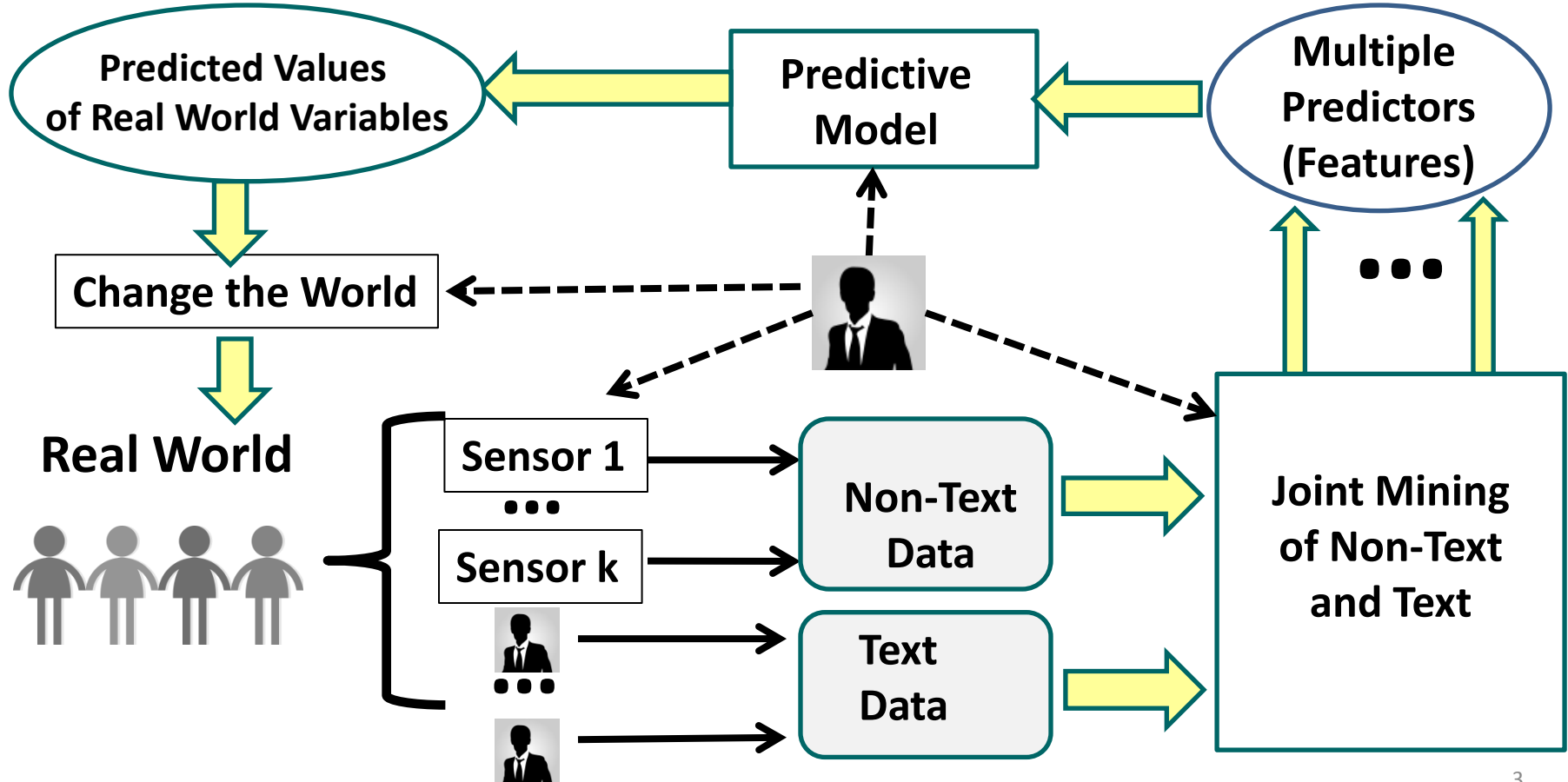# Text-Based Prediction

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

Text-Based Prediction
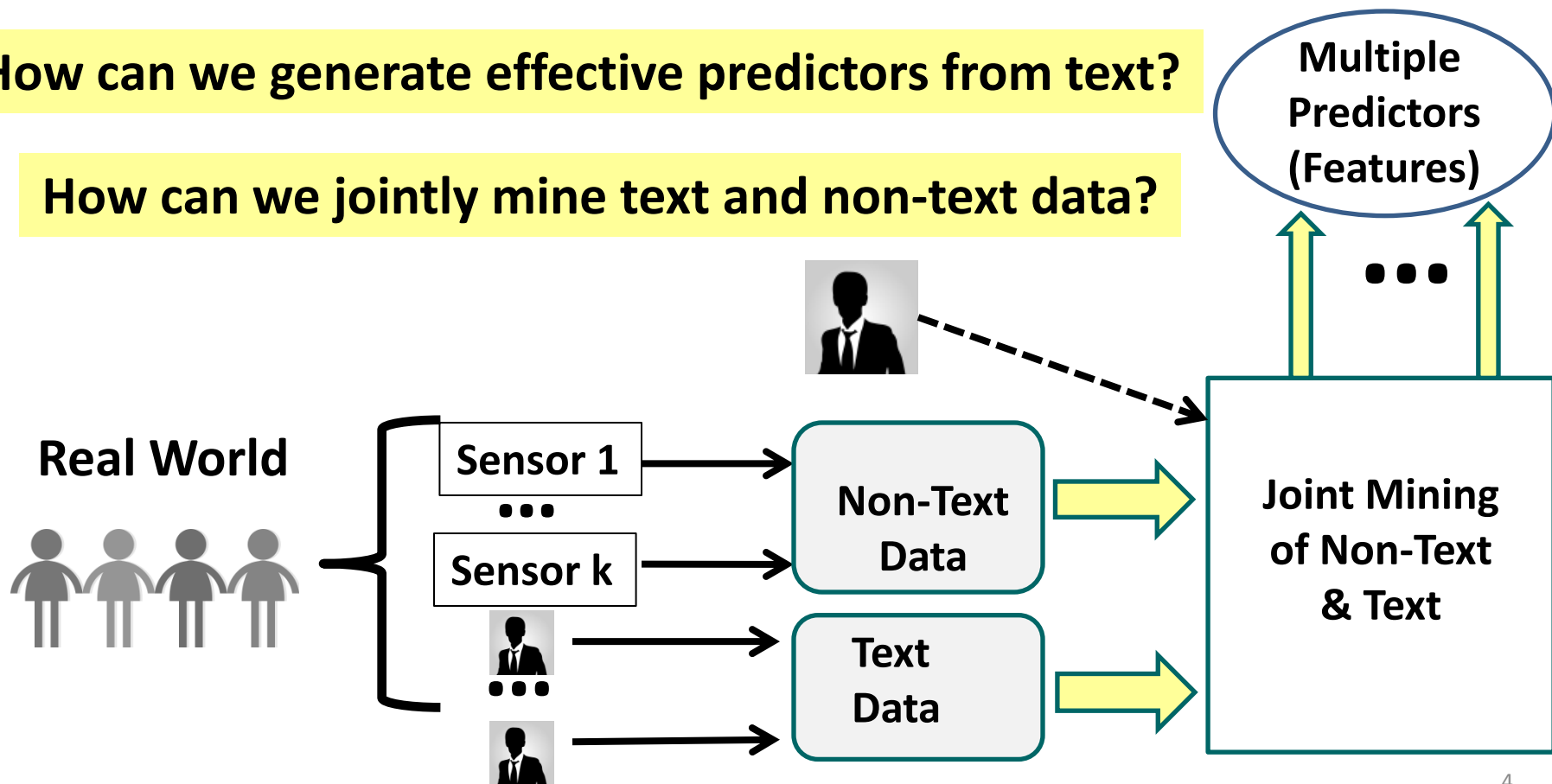
5. Text-based prediction

3. Topic mining and analysis

Real World   Observed World   Text Data

Perceive   Express   1. Natural language
                                                                processing and text
(Perspective)   (English)   representation

4. Opinion mining and sentiment analysis

2. Word association mining and analysis

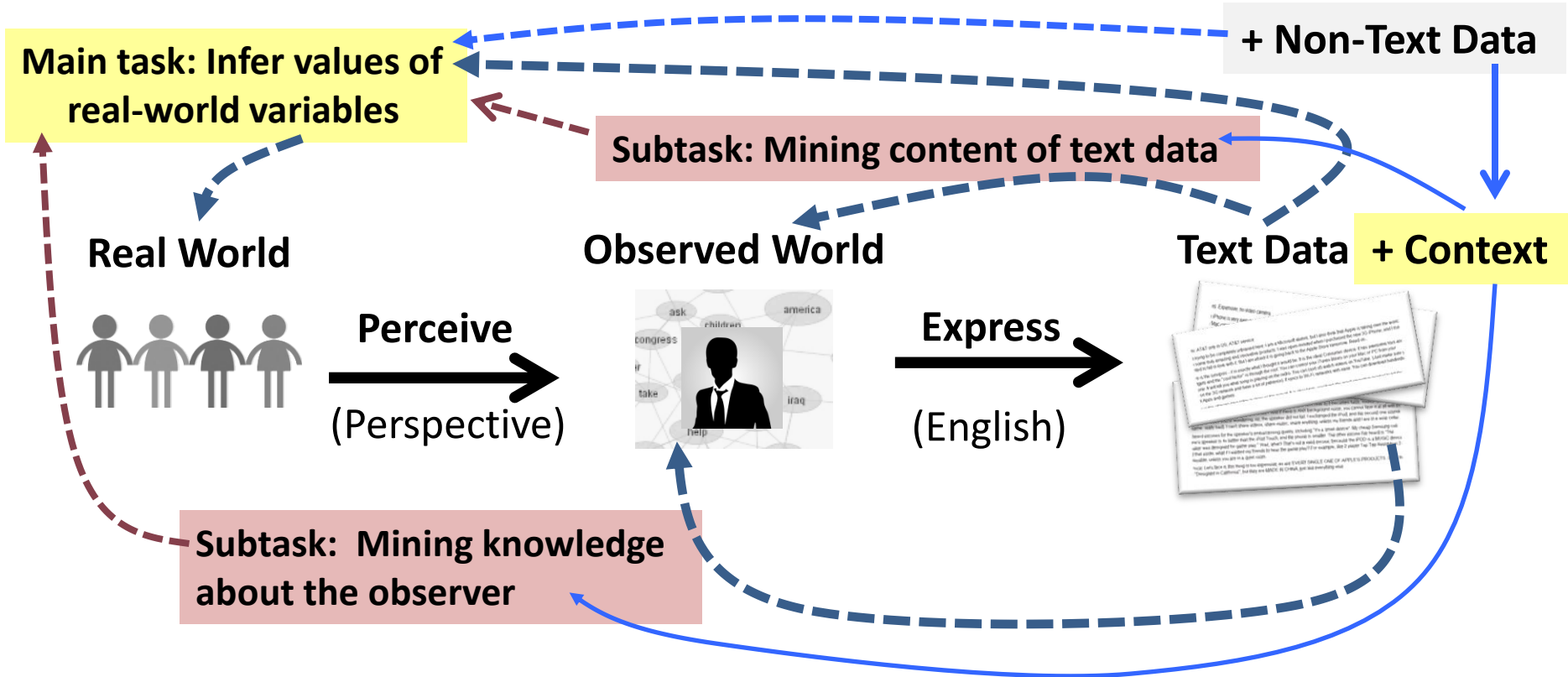# The Big Picture of Prediction: Data Mining Loop

Text-Based Prediction

How can we generate effective predictors from text?

How can we jointly mine text and non-text data?

Multiple Predictors (Features)

Real World

Sensor 1
···
Sensor k

Non-Text Data

Text Data

Joint Mining of Non-Text & Text

Text-Based Prediction =
a Unified View of Text Mining and Analysis

Main task: Infer values of real-world variables

+ Non-Text Data

Subtask: Mining content of text data

Real World

Observed World

Text Data **+ Context**

Perceive

(Perspective)

Express

(English)

Subtask: Mining knowledge about the observer

Joint Mining and Analysis of Text and Non-Text Data

- Non-text data help text mining
  - Non-text data provide context for mining text data
  - **Contextual Text Mining**: Mining text in the context defined by non-text data (see [Mei 2009] for a large body of work)
- Text data help non-text data mining
  - Text data help interpret patterns discovered from non-text data
  - **Pattern Annotation**: Using text data to interpret patterns found in non-text data (see [Mei et al. 2006] for detail)

Suggested Reading

- **[Mei et al. 2006]** Qiaozhu Mei, Dong Xin, Hong Cheng, Jiawei Han, and ChengXiang Zhai. 2006. Generating semantic annotations for frequent patterns with context analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD 2006). ACM, New York, NY, USA, 337-346. DOI=10.1145/1150402.1150441
- **[Mei 2009]** Qiaozhu Mei, Contextual Text Mining, Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2009. http://hdl.handle.net/2142/14707