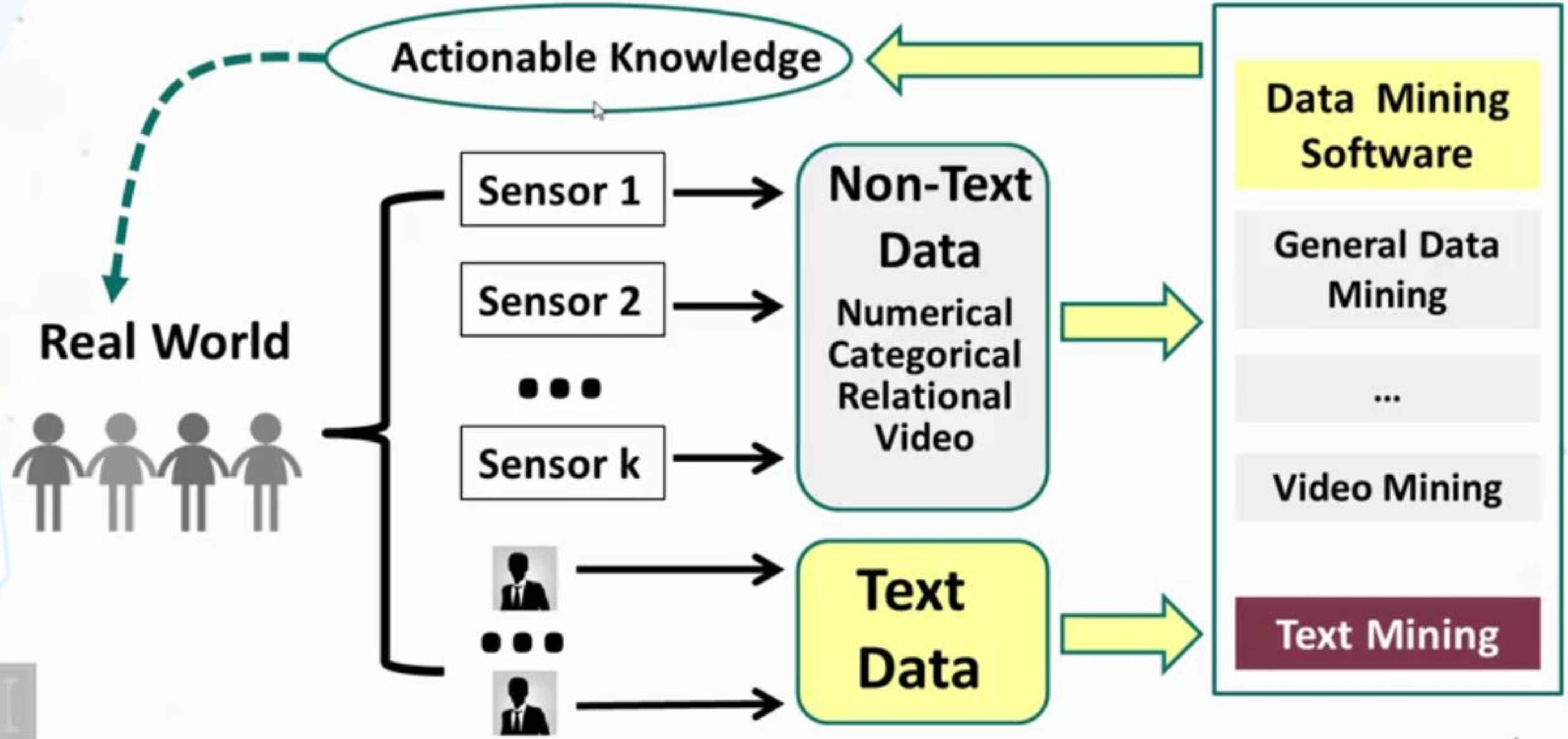
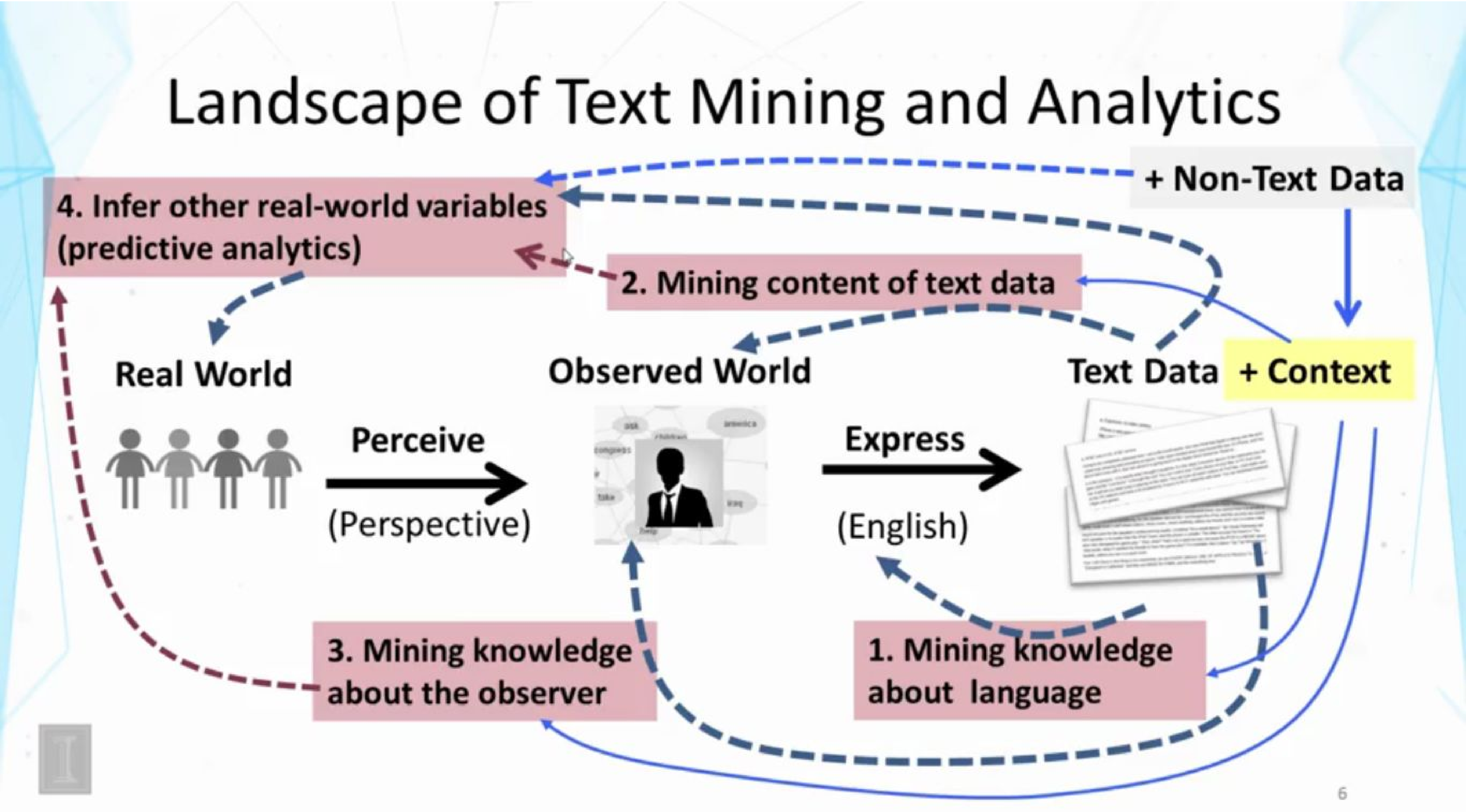
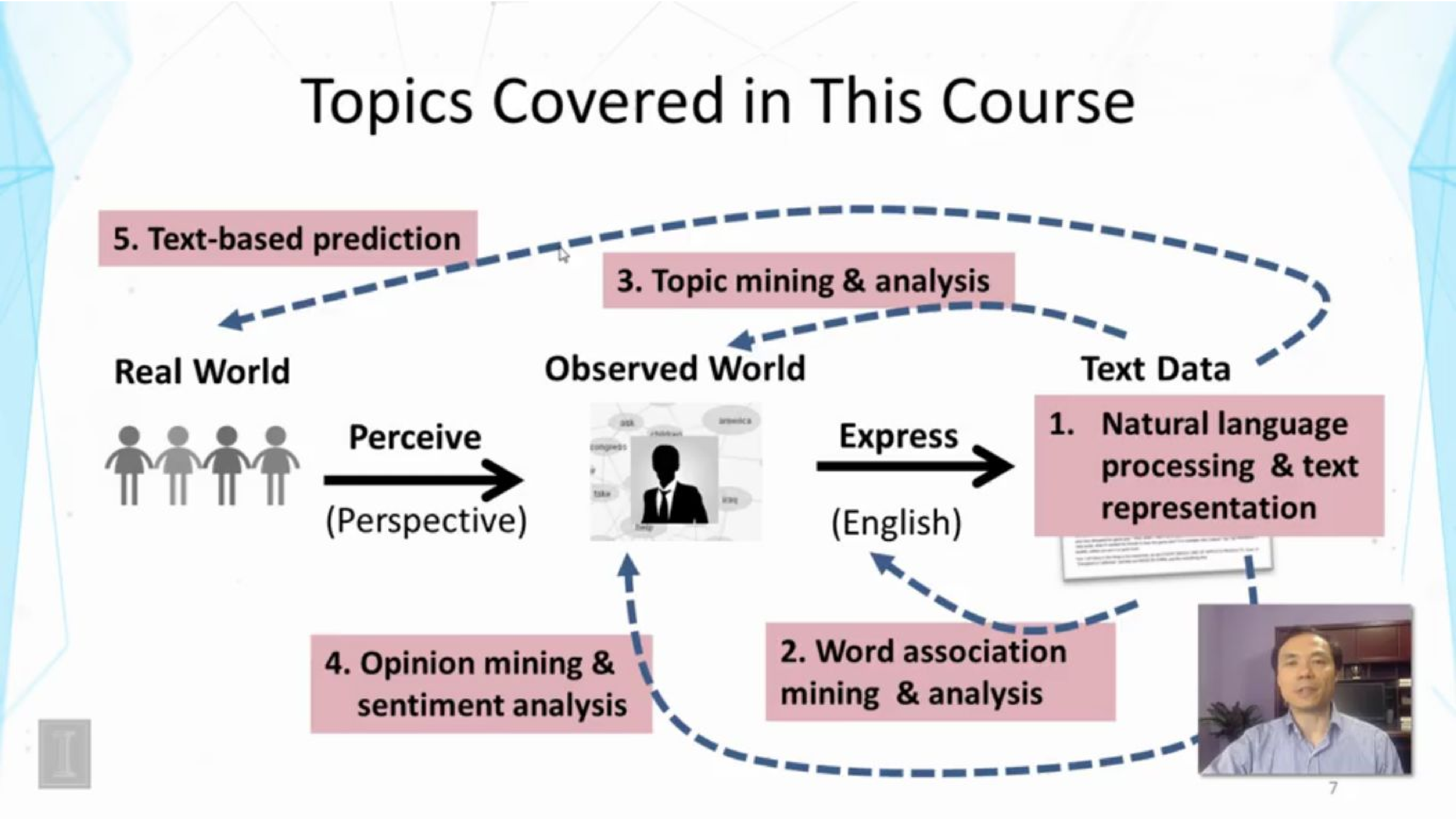
Text Mining and Analytics

- Text mining ≈ Text analytics
- Turn text data into **high-quality information** or **actionable knowledge**
  - **Minimizes human effort** (on consuming text data)
  - Supplies knowledge for **optimal decision making**
- Related to **text retrieval**, which is an essential component in any text mining system
  - Text retrieval can be a preprocessor for text mining
  - Text retrieval is needed for knowledge provenance

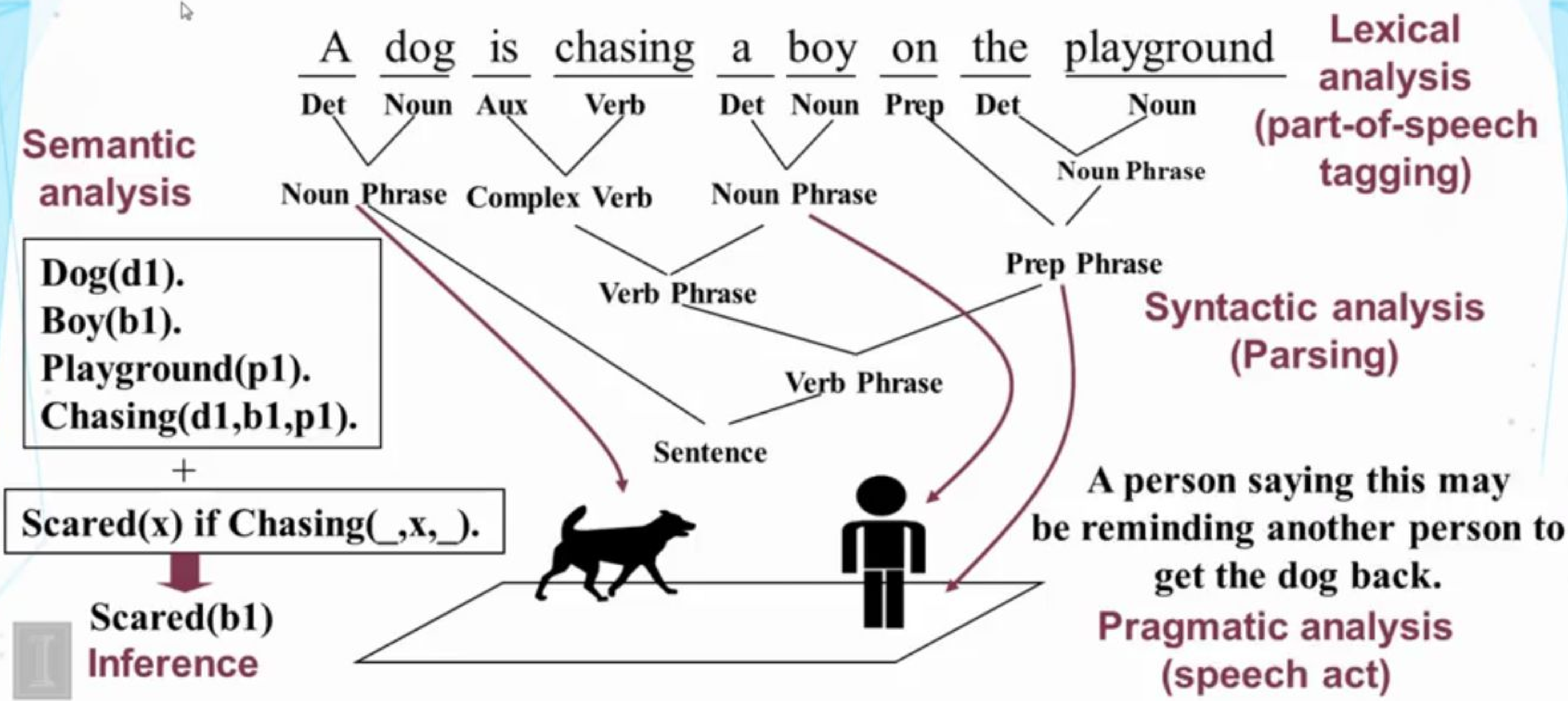
The General Problem of Data Mining




Landscape of Text Mining and Analytics
+ Non-Text Data
4. Infer other real-world variables (predictive analytics)
2. Mining content of text data
Real World
Observed World
Text Data
+ Context
Perceive
(Perspective)
Express
(English)
3. Mining knowledge about the observer
1. Mining knowledge about language
6


Topics Covered in This Course
5. Text-based prediction
3. Topic mining & analysis
Real World
Observed World
Text Data
Perceive
(Perspective)
Express
(English)
1. Natural language processing & text representation
4. Opinion mining & sentiment analysis
2. Word association mining & analysis
7

Basic Concepts in NLP



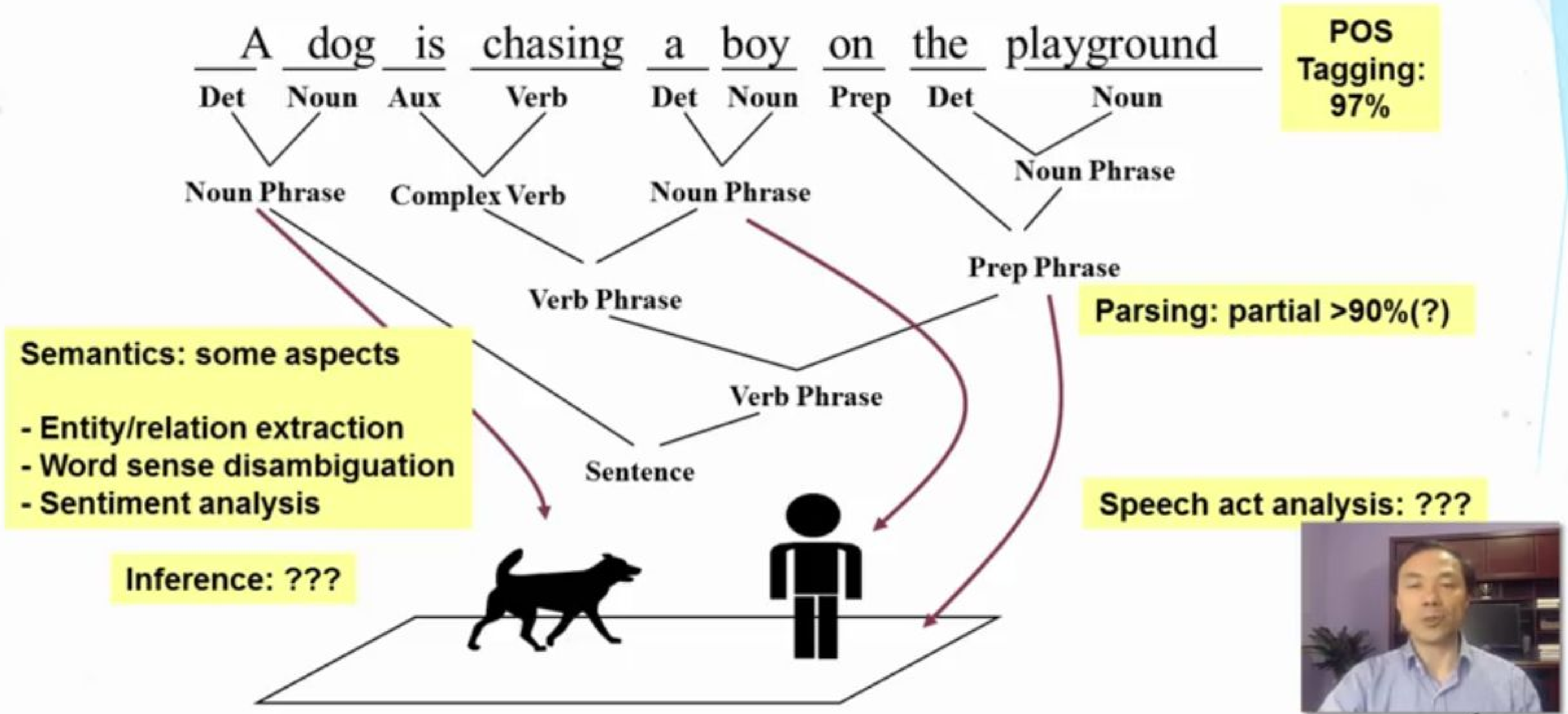
NLP Is Difficult!

- Natural language is designed to make human communication efficient. As a result,
  - we omit a lot of *common sense* knowledge, which we assume the hearer/reader possesses.
  - we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve.
- This makes EVERY step in NLP hard
  - Ambiguity is a *killer*!
  - Common sense reasoning is pre-required.

Examples of Challenges

- Word-level ambiguity:
  - "design" can be a noun or a verb (ambiguous POS)
  - "root" has multiple meanings (ambiguous sense)
- Syntactic ambiguity:
  - "natural language processing" (modification)
  - "A man saw a boy *with a telescope*." (PP Attachment)
- Anaphora resolution: "John persuaded Bill to buy a TV for *himself*." (himself = John or Bill?)
- Presupposition: "He has quit smoking" implies that he smoked before.

The State of the Art



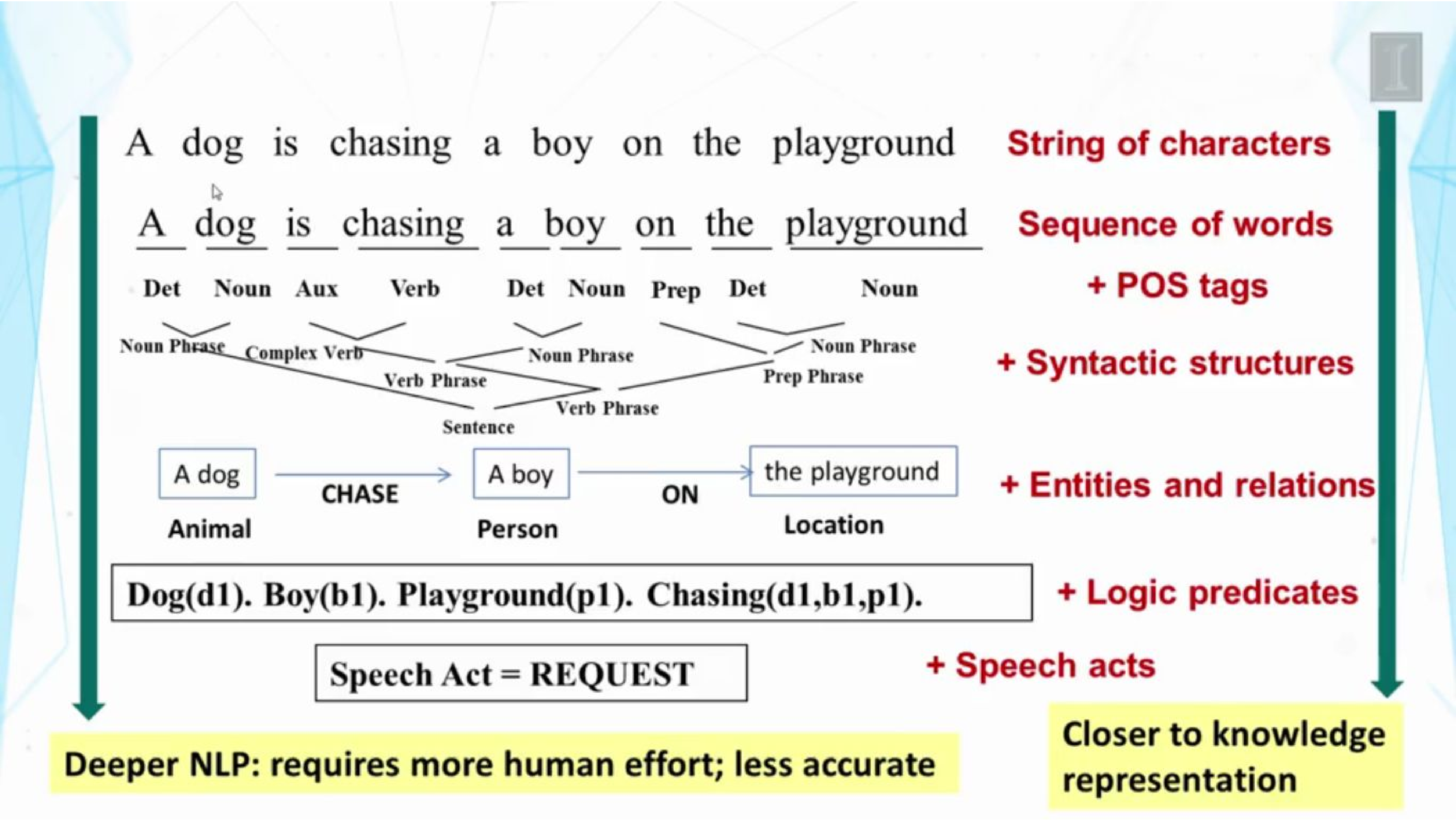
What We Can't Do

- 100% POS tagging
  - "He turned off the highway." vs "He turned off the fan."
- General complete parsing
  - "A man saw a boy with a telescope."
- Precise deep semantic analysis
  - Will we ever be able to precisely define the meaning of "own" in "John owns a restaurant"?

Robust and general NLP tends to be *shallow* while *deep* understanding doesn't scale up.

Summary

- NLP is the foundation for text mining
- Computers are far from being able to understand natural language
  - Deep NLP requires common sense knowledge and inferences, thus only working for very limited domains
  - Shallow NLP based on statistical methods can be done in large scale and is thus more broadly applicable
- In practice: statistical NLP as the basis, while humans provide help as needed


A dog is chasing a boy on the playground
String of characters
A dog is chasing a boy on the playground
Sequence of words
Det Noun Aux Verb Det Noun Prep Det Noun
+ POS tags
Noun Phrase
Complex Verb
Noun Phrase
Noun Phrase
Verb Phrase
Prep Phrase
Verb Phrase
Sentence
+ Syntactic structures
A dog
CHASE
A boy
ON
the playground
Animal
Person
Location
+ Entities and relations
Dog(d1). Boy(b1). Playground(p1). Chasing(d1,b1,p1).
+ Logic predicates
Speech Act = REQUEST
+ Speech acts
Deeper NLP: requires more human effort; less accurate
Closer to knowledge representation

Text Representation and Enabled Analysis

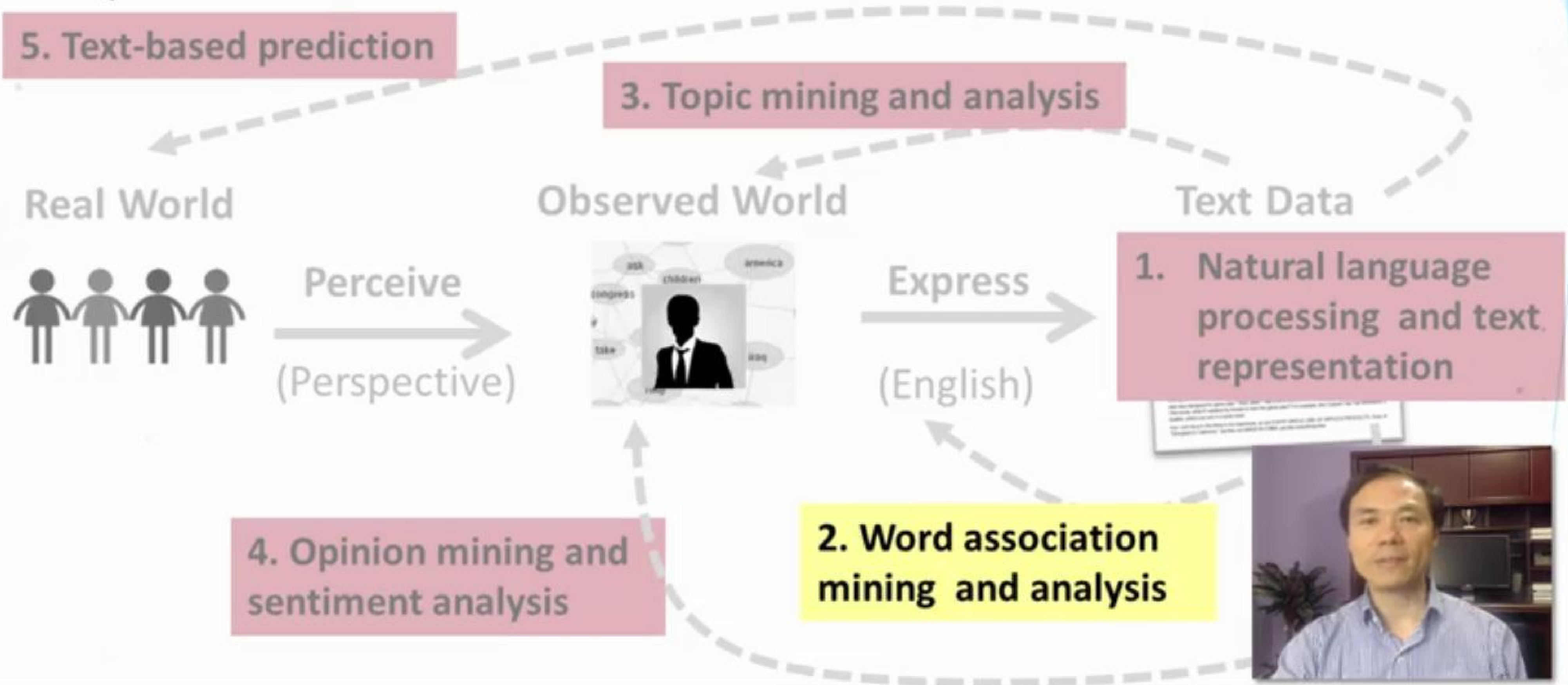
This course

Text Rep	Generality	Enabled Analysis	Examples of Application
String		String processing	Compression
Words		Word relation analysis; topic analysis; sentiment analysis	Thesaurus discovery; topic and opinion related applications
+ Syntactic structures		Syntactic graph analysis	Stylistic analysis; structure-based feature extraction
+ Entities & relations		Knowledge graph analysis; information network analysis	Discovery of knowledge and opinions about specific entities
+ Logic predicates		Integrative analysis of scattered knowledge; logic inference	Knowledge assistant for biologists

Summary

- Text representation determines what kind of mining algorithms can be applied
- **Multiple ways** of representing text are possible
  - string, words, syntactic structures, entity-relation graphs, predicates…
  - can/should be **combined** in real applications
- This course focuses on **word-based representation**
  - **General and robust**: applicable to any natural language
  - **No/little manual effort**
  - **“Surprisingly” powerful** for many applications (not all!)
  - **Can be combined** with more sophisticated representations

Word Association Mining & Analysis



Outline

- What is a word association?
- Why mine word associations?
- How to mine word associations?

Basic Word Relations: Paradigmatic vs. Syntagmatic

- Paradigmatic: A & B have paradigmatic relation if they can be substituted for each other (i.e., A & B are in the same class)
  - E.g., "cat" and "dog"; "Monday" and "Tuesday"
- Syntagmatic: A & B have syntagmatic relation if they can be combined with each other (i.e., A & B are related semantically)
  - E.g., "cat" and "sit"; "car" and "drive"
- These two basic and complementary relations can be generalized to describe relations of any items in a language

Why Mine Word Associations?

- They are useful for improving accuracy of many NLP tasks
  - POS tagging, parsing, entity recognition, acronym expansion
  - Grammar learning
- They are directly useful for many applications in text retrieval and mining
  - Text retrieval (e.g., use word associations to suggest a variation of a query)
  - Automatic construction of topic map for browsing: words as nodes and associations as edges
  - Compare and summarize opinions (e.g., what words are most strongly associated with “battery” in positive and negative reviews about iPhone 6, respectively?)

Mining Word Associations: Intuitions

Paradigmatic: similar context

My **cat** eats fish on Saturday
His **cat** eats turkey on Tuesday
My **dog** eats meat on Sunday
His **dog** eats turkey on Tuesday
...

cat:

My ___ eats fish on Saturday
His ___ eats turkey on Tuesday
...

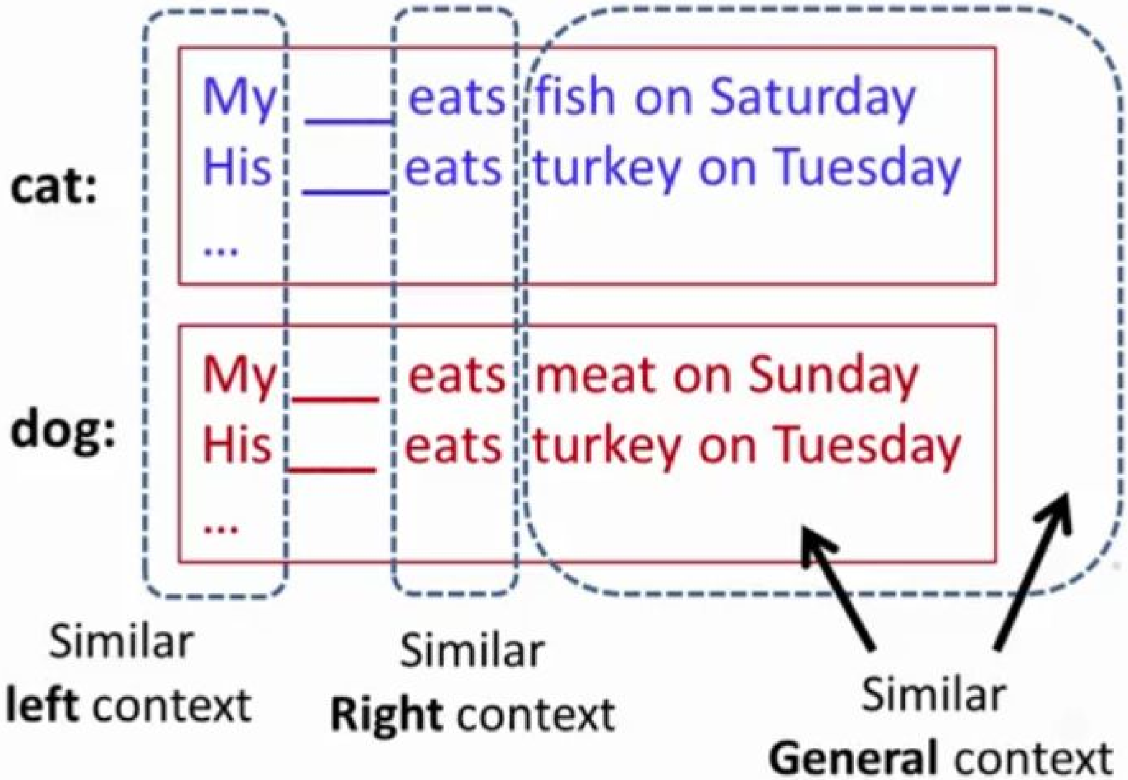
dog:

My ___ eats meat on Sunday
His ___ eats turkey on Tuesday
...

Mining Word Associations: Intuitions

Paradigmatic: similar context

My **cat** eats fish on Saturday
His **cat** eats turkey on Tuesday
My **dog** eats meat on Sunday
His **dog** eats turkey on Tuesday
...



How similar are context ("cat") and context ("dog")?
How similar are context ("cat") and context ("computer")?

Mining Word Associations: Intuitions

Syntagmatic: correlated occurrences

My cat **eats** fish on Saturday
His cat **eats** turkey on Tuesday
My dog **eats** meat on Sunday
His dog **eats** turkey on Tuesday
…

My __ **eats** __ on Saturday
His __ **eats** __ on Tuesday
My __ **eats** __ on Sunday
His __ **eats** __ on Tuesday
…

What words tend to occur to the **left** of **"eats"?**

What words to the **right?**

Whenever "**eats**" occurs, what **other words** also tend to occur?
How helpful is the occurrence of "**eats**" for predicting occurrence of "**meat**"?
How helpful is the occurrence of "**eats**" for predicting occurrence of "**text**"?