Welcome To

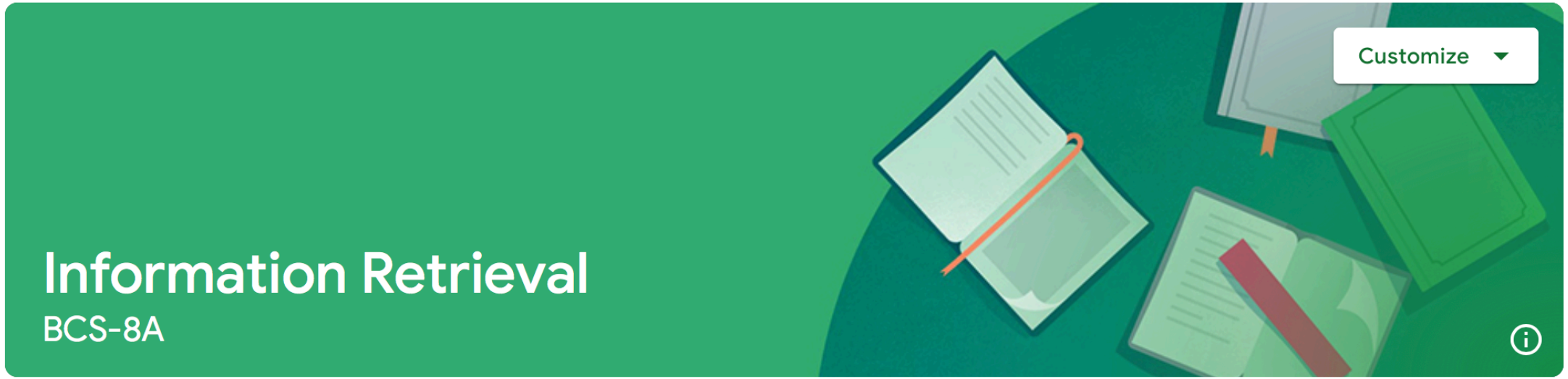
Information Retrieval

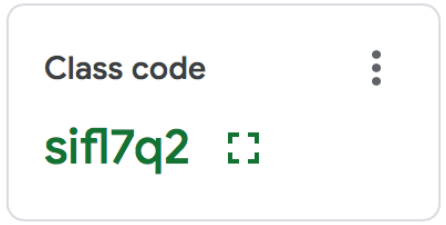
Lecture#1 - Introduction

Dr. Iqra Safder
Assistant Professor
FAST NUCES, Lahore

Google Classroom code

- Please join the classroom





sifl7q2

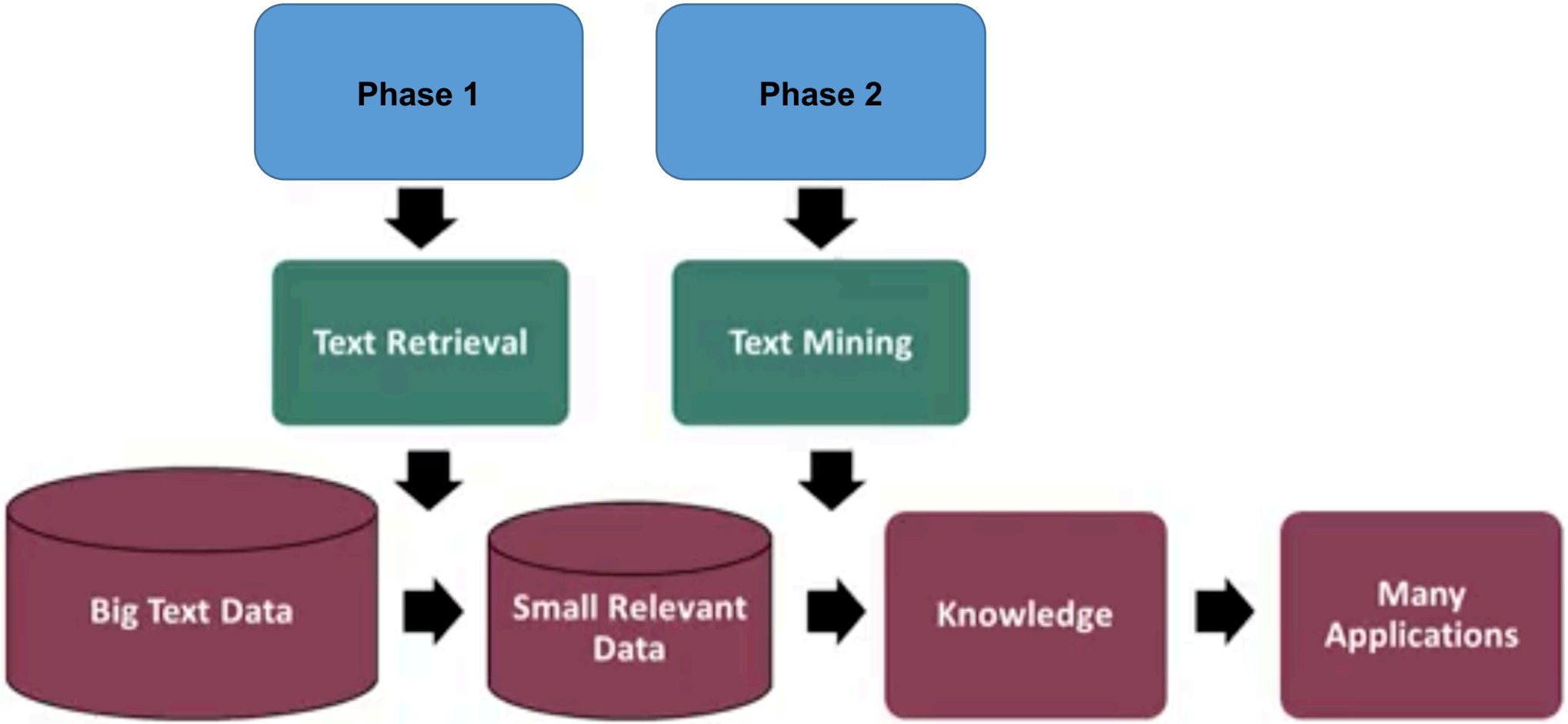
Motivation: Harnessing Big Data Text Data

- **Text data is produced by humans**
  - Contains people's opinions
  - Encodes human knowledge
  - Offers opportunities for discovering knowledge
- **Text is consumed by humans**
  - Need intelligent tools
  - Humans play an essential role in mining
- **What types of data we produce? Take a moment!**

Examples of Text Information System Applications

- **Search**
  - Google, Bing, Search Box on Lap top
- **Filtering/Recommendation**
  - News filters, Spam filters, Movie or Book Recommendations
- **Categorization**
  - Automatic Sorting of Emails, Product reviews, News Categories
- **Mining/Extraction**
  - Customer Complain Messages, Product Reviews, Systems to Read Literature
- **Many others**

Main Techniques for Harnessing Big Data: Information Retrieval + Text Mining



Course Objectives

- **Detailed** concepts and practical techniques in text retrieval
  - How search engines work
  - How to implement a search engine
  - How to evaluate a search engine
  - How to improve and optimize a search engine
  - How to build a recommender system
- **Hands-on experience** on
  - Creating a text collection for evaluating search engines
  - Experimenting with search engine algorithms
  - A term paper project (in group)

Prerequisites and Format

- **Prerequisites:**
  - Basic concepts of computer science (e.g. data structures)
  - Comfortable with programming, particularly with C++
  - (Optional) Knowledge of Python
- **Format**:
  - Lecture + quizzes + assignments + term paper project + mid term + final exam

Main Book

Text Data Management and Analysis: *A Practical Introduction to Information Retrieval and Text Mining*

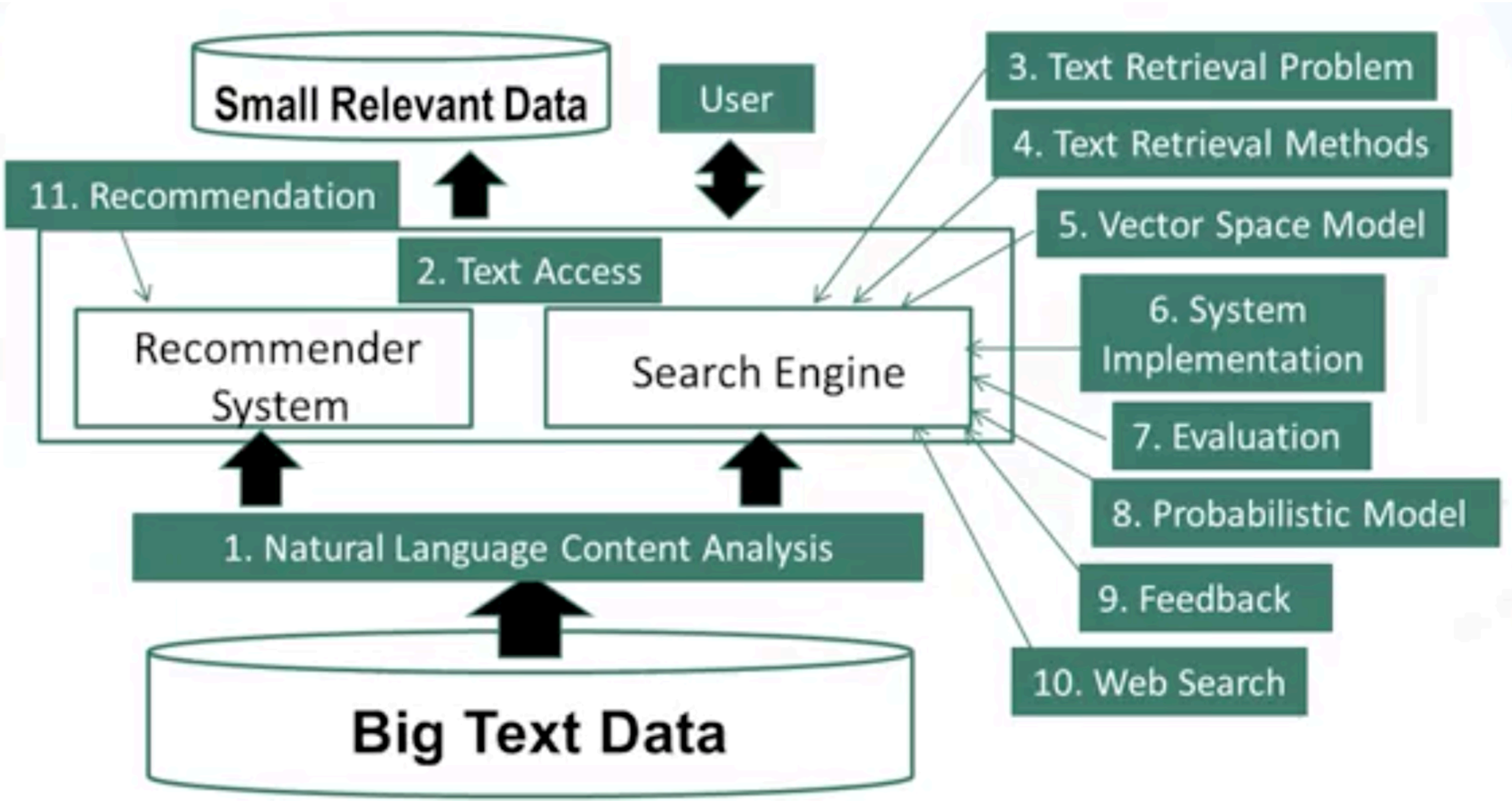
By: ChengXiang Zhai, Sean Massung

ISBN: 9781970001167 | PDF ISBN: 9781970001174

Hardcover ISBN:9781970001198

 | 471 Pages | Publication Date: July, 2016

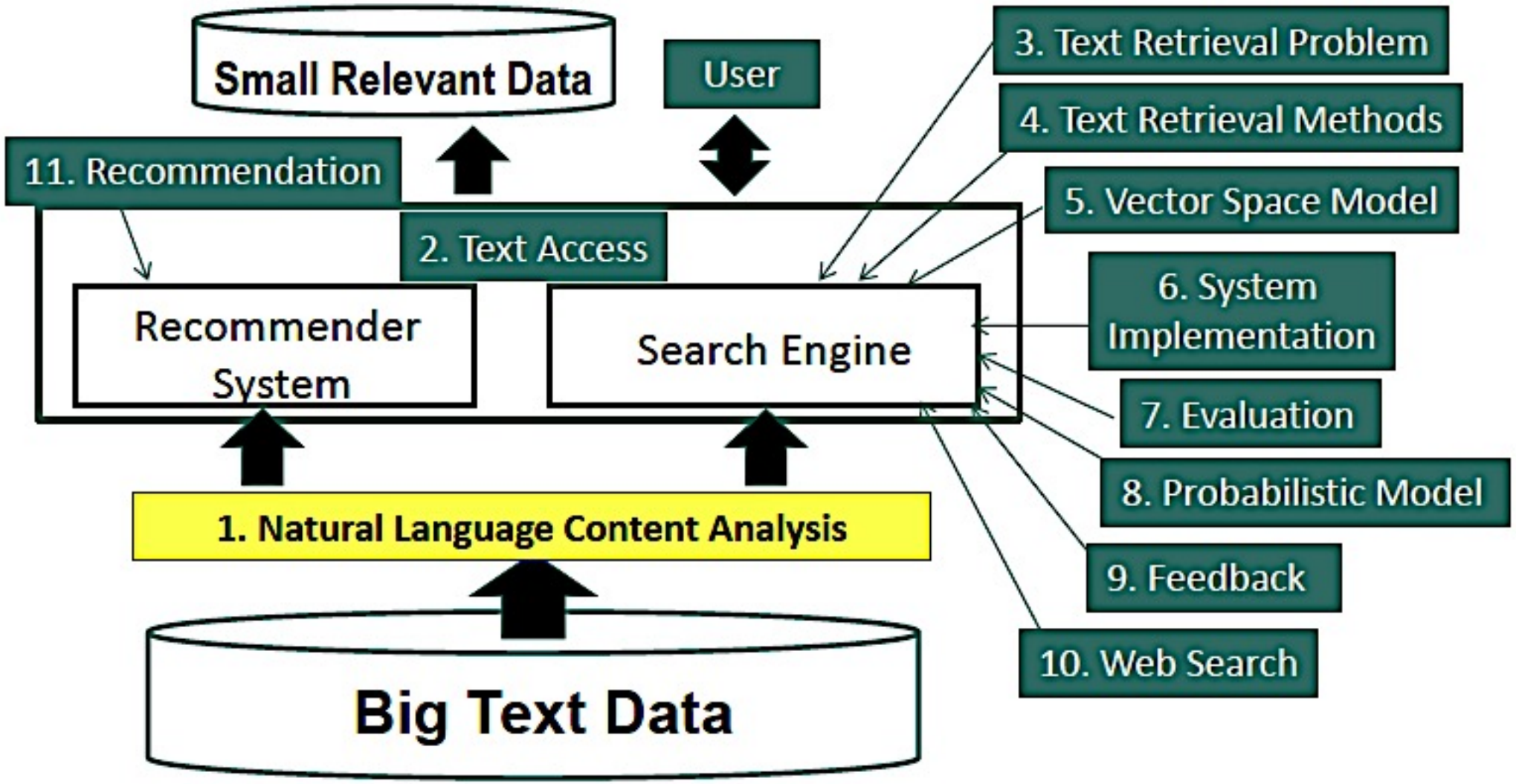
Infotmation Retrieval



Information Retrieval

Natural Language Content Analysis

Course Schedule

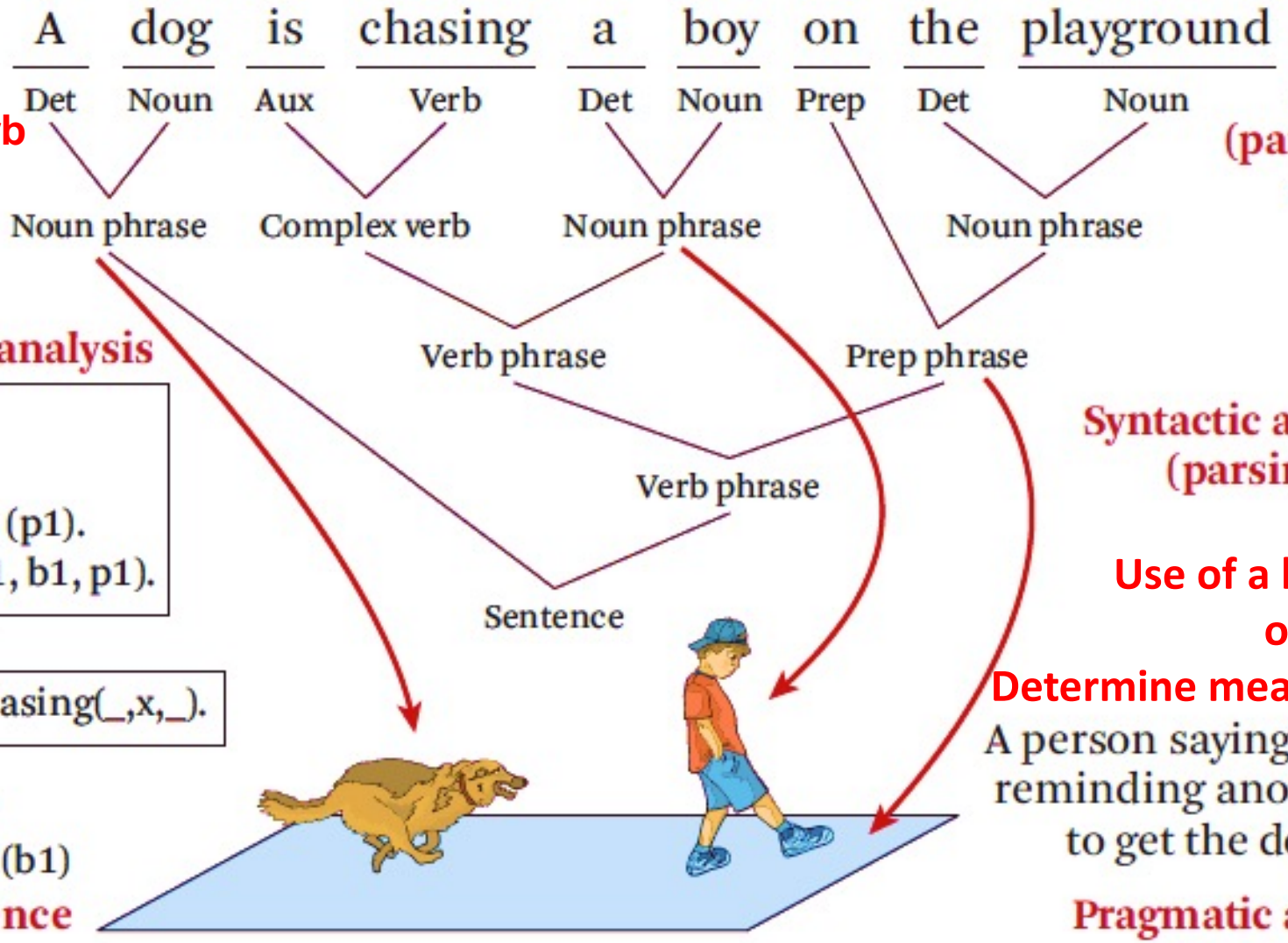


Overview

- What is Natural Language Processing (NLP)?
- State of the Art in NLP
- NLP for Text Retrieval

An Example of NLP

Det: determiner
Aux: auxiliary verb



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

Discourse analysis. Discourse analysis is needed when a large chunk of text with multiple sentences is to be analyzed; in such a case, the connections between these sentences must be considered and the analysis of an individual sentence must be placed in the appropriate context involving other sentences.

Examples of Challenges

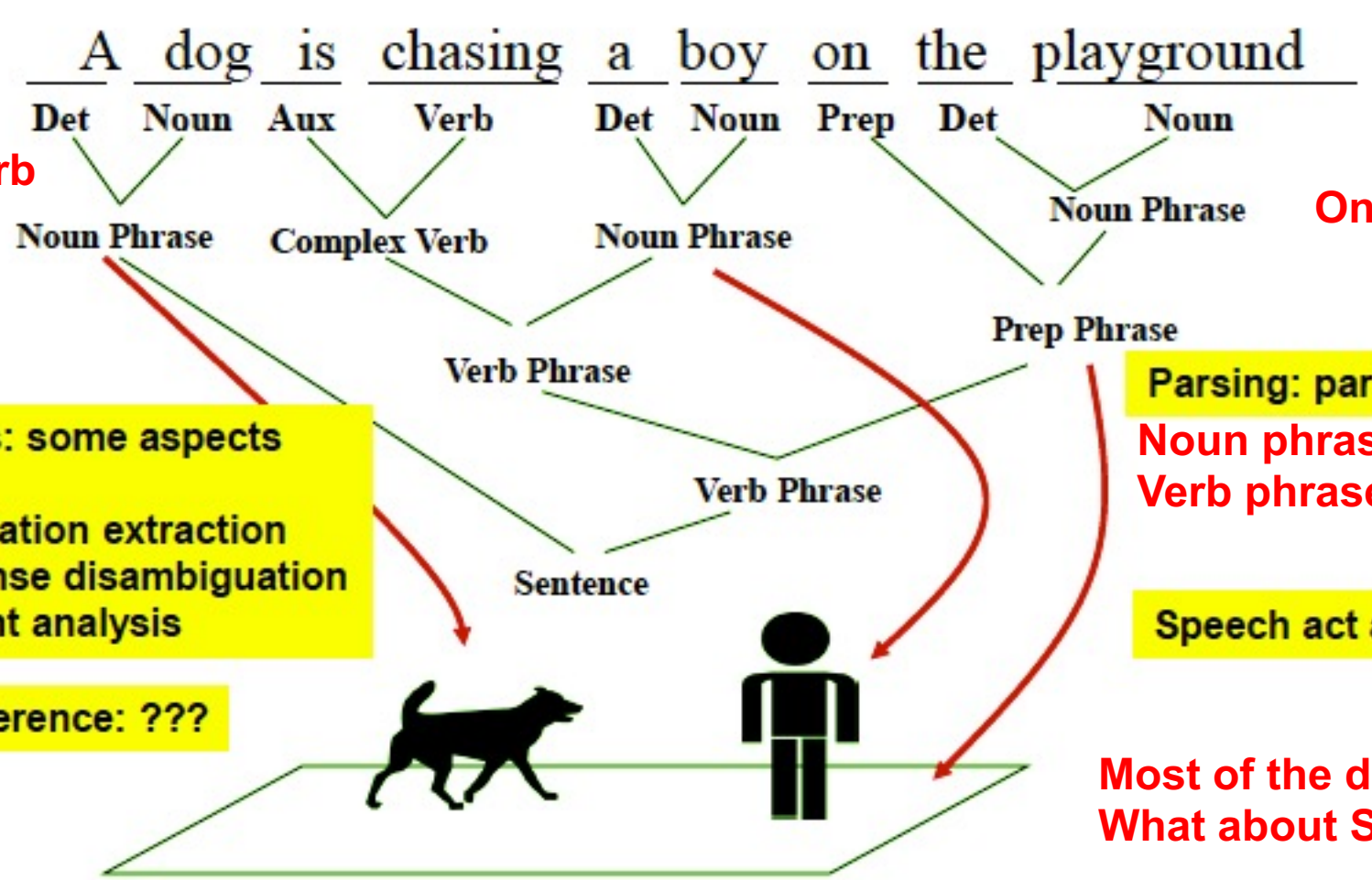
- Word-level ambiguity: E.g.,
  - "design" can be a noun or a verb (Ambiguous POS)
  - "root" has multiple meanings (Ambiguous sense)
- Syntactic ambiguity: E.g.,
  - "natural language processing" (Modification)
  - "A man saw a boy *with a telescope*." (PP Attachment)
- Anaphora resolution: "John persuaded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

Processing of Natural Language
Language Processing is Natural?

Who had the telescope?
Prepositional Phrase

Anaphora is the repetition of words or phrases in a group of sentences, clauses, or poetic lines.

The State of the Art



Det: determiner
Aux: auxiliary verb

POS Tagging: 97%

On a given dataset

Semantics: some aspects

- Entity/relation extraction
- Word sense disambiguation
- Sentiment analysis

Parsing: partial >90%(?)

Noun phrase structures
Verb phrase structures

Speech act analysis: ???

Inference: ???

We don't have complete understanding of NLP, in general.

Most of the datasets are News
What about Social Media Data?

Entity Extraction: Recognizing dimensions of people, locations, organizations, etc in text.

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."?

And in cases when the sentence is very long, imagine it has four or five prepositional phrases, and there are even more possibilities to figure out.

Robust & general NLP tends to be "shallow" while "deep" understanding doesn't scale up

These techniques may work well based on machine learning techniques on the data that are similar to the training data that the program has been trained on. But they generally wouldn't work well on the data that are very different from the training data.

NLP for Text Retrieval

- Must be general robust & efficient ➔ Shallow NLP
- **"Bag of words" representation tends to be sufficient** for most search tasks (but not all!)
- Some text retrieval techniques can naturally address NLP problems Feedback technique : Add more words with the query
  Q = java
  Q = java applet
- However, deeper NLP is needed for complex search tasks Question Answering Task
- **Google Knowledge Graph: Entities and their relations, which goes beyond BOW**

Summary

- What is Natural Language Processing (NLP)?
- State of the Art in NLP
- NLP for Text Retrieval

Additional Reading

Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.