Applied Text Mining in Python

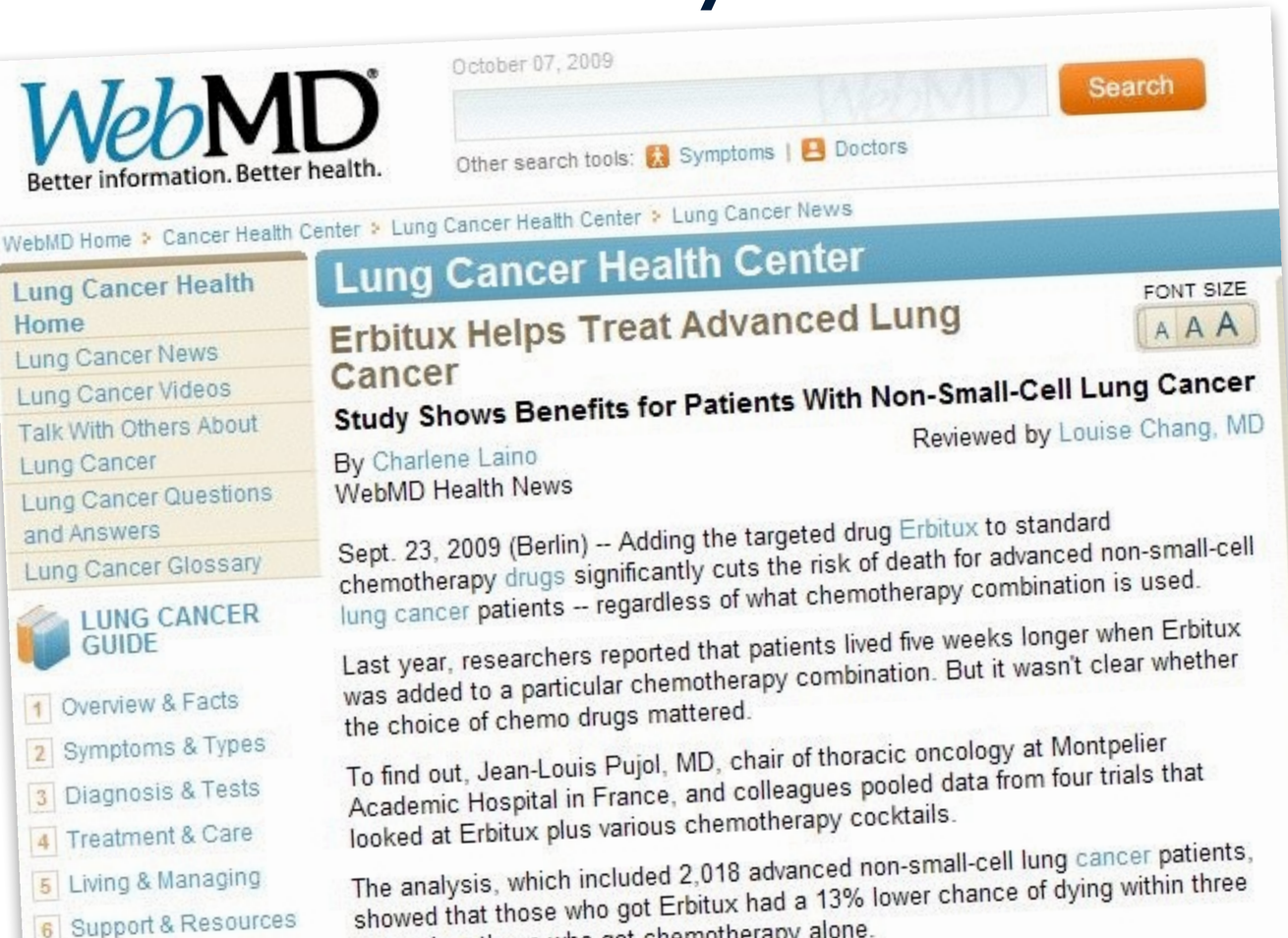
Information Extraction

Information is hidden in free-text

- Most traditional transactional information is structured
- Abundance of unstructured, freeform text
- How to convert unstructured text to structured form?

Information Extraction

- **Goal: Identify and extract fields of interest from free text**



Erbitux helps treat lung cancer

Author: Charlene Laino

Reviewer: Louise Chang, MD

Sept. 23, 2009 | Berlin | …

Fields of Interest

- Named entities
  - [NEWS] People, Places, Dates, …
  - [FINANCE] Money, Companies, …
  - [MEDICINE] Diseases, Drugs, Procedures, …
- Relations
  - What happened to who, when, where, …

Named Entity Recognition

- **Named entities**: Noun phrases that are of specific type and refer to specific individuals, places, organizations, …
- **Named Entity Recognition**: Technique(s) to identify all mentions of pre-defined named entities in text
  - Identify the mention / phrase: *Boundary detection*
  - Identify the type: *Tagging / classification*

Examples of Named Entity Recognition Tasks

The patient is a 63-year-old female with a three-year history of bilateral hand numbness and occasional weakness.

Within the past year, these symptoms have progressively gotten worse, to encompass also her feet.

She had a workup by her neurologist and an MRI revealed a C5-6 disc herniation with cord compression and a T2 signal change at that level.

Approaches to identify named entities

- Depends on kinds of entities that need to be identified

- For well-formatted fields like date, phone numbers: Regular expressions (Recall Week 1)
- For other fields: Typically a machine learning approach

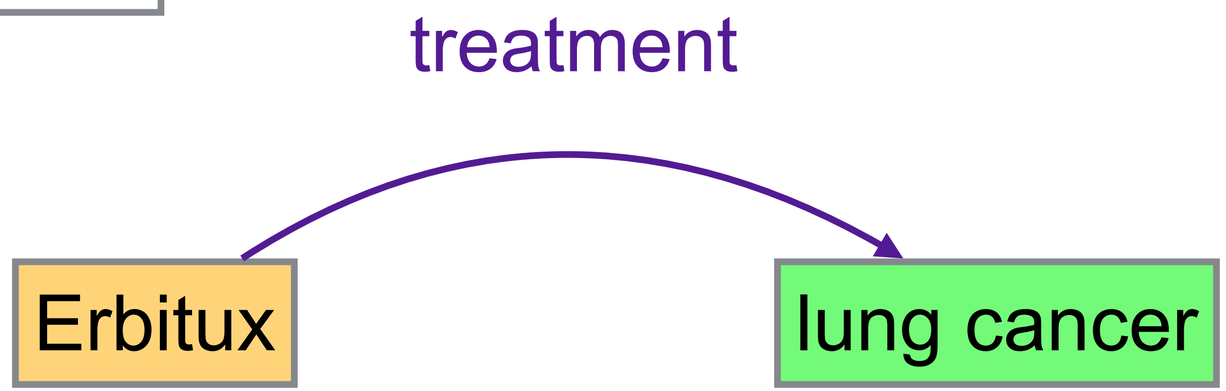
Person, Organization, Location/GPE

- Standard NER task in NLP research community
- Typically a four-class model
  - PER
  - ORG
  - LOC / GPE
  - Other / Outside (any other class)

Relation Extraction

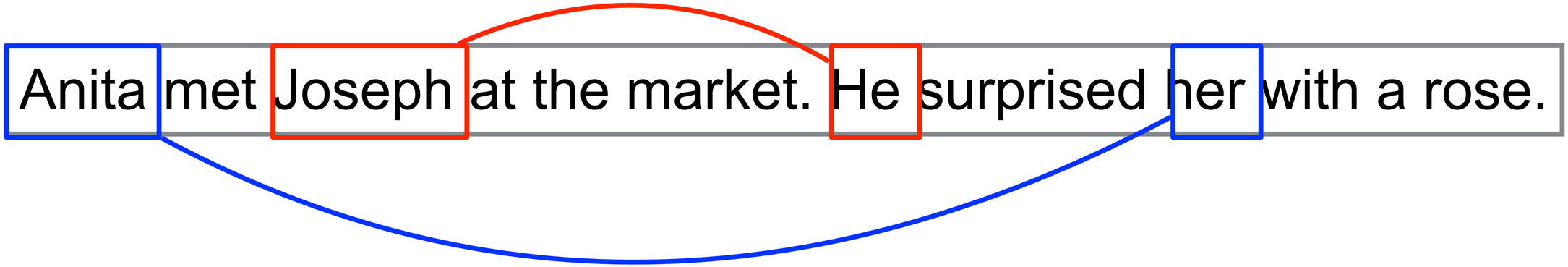
- **Identify relationships between named entities**

Erbitux helps treat lung cancer



Co-reference Resolution

- **Disambiguate mentions and group mentions together**



Question Answering

- Given a question, find the most appropriate answer from the text
  - What does Erbitux treat?
  - Who gave Anita the rose?
- Builds on named entity recognition, relation extraction, and co-reference resolution

Take Home Concepts

- Information Extraction is important for natural language understanding and making sense of textual data
- Named Entity Recognition is a key building block to address many advanced NLP tasks
- Named Entity Recognition systems extensively deploy supervised machine learning and text mining techniques discussed in this course