

# 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

The screenshot shows a WebMD article page. At the top, the WebMD logo is on the left, and a search bar with a 'Search' button is on the right. Below the search bar, there are links for 'Symptoms' and 'Doctors'. The main heading of the article is 'Erbixux Helps Treat Advanced Lung Cancer', with a subtitle 'Study Shows Benefits for Patients With Non-Small-Cell Lung Cancer'. The author is listed as 'Charlene Laino' and the reviewer as 'Louise Chang, MD'. The date is 'Sept. 23, 2009 (Berlin)'. The article text discusses the benefits of the drug Erbixux for advanced non-small-cell lung cancer patients, mentioning a study by Jean-Louis Pujol, MD, and a 13% lower chance of dying within three years for patients who received Erbixux compared to those who did not.

WebMD  
Better information. Better health.

October 07, 2009

Search

Other search tools: [Symptoms](#) | [Doctors](#)

WebMD Home > Cancer Health Center > Lung Cancer Health Center > Lung Cancer News

**Lung Cancer Health Center**

**Erbixux Helps Treat Advanced Lung Cancer**

**Study Shows Benefits for Patients With Non-Small-Cell Lung Cancer**

By [Charlene Laino](#)  
WebMD Health News

Reviewed by [Louise Chang, MD](#)

Sept. 23, 2009 (Berlin) -- Adding the targeted drug [Erbixux](#) to standard chemotherapy [drugs](#) significantly cuts the risk of death for advanced non-small-cell lung cancer patients -- regardless of what chemotherapy combination is used.

Last year, researchers reported that patients lived five weeks longer when Erbixux was added to a particular chemotherapy combination. But it wasn't clear whether the choice of chemo drugs mattered.

To find out, Jean-Louis Pujol, MD, chair of thoracic oncology at Montpelier Academic Hospital in France, and colleagues pooled data from four trials that looked at Erbixux plus various chemotherapy cocktails.

The analysis, which included 2,018 advanced non-small-cell lung cancer patients, showed that those who got Erbixux had a 13% lower chance of dying within three years compared to those who got chemotherapy alone.

Erbixux 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

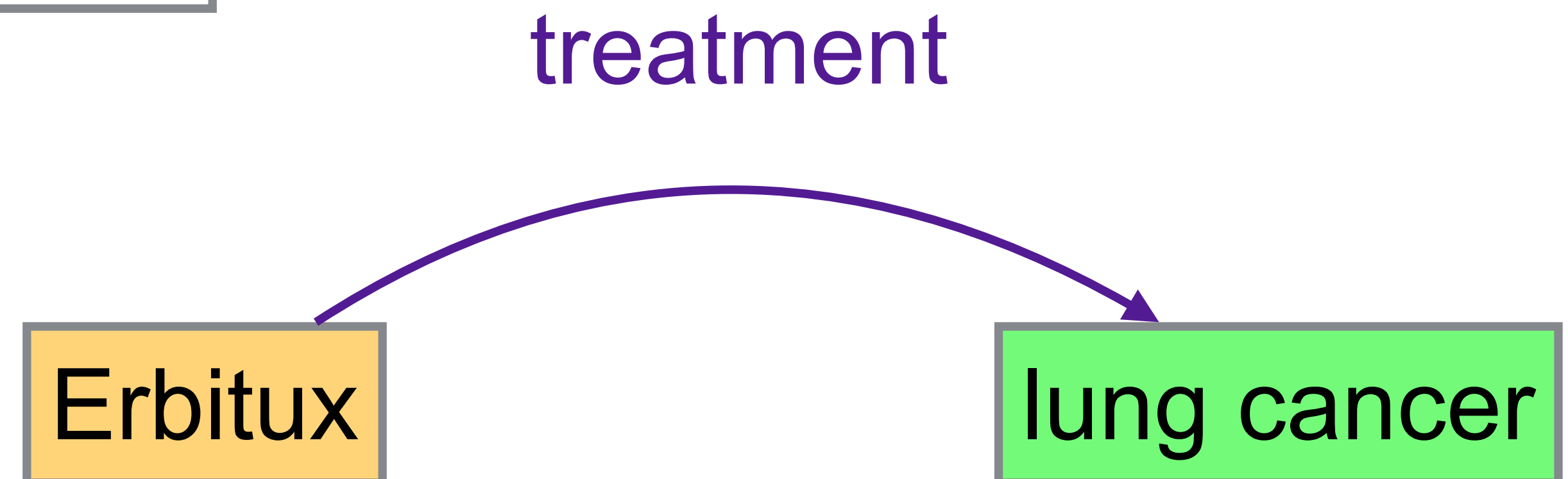
- Standard NER task in NLP research community
- Typically a four-class model
  - PER
  - ORG
  - LOC / GPE
  - Other / Outside (any other class)



# Relation Extraction

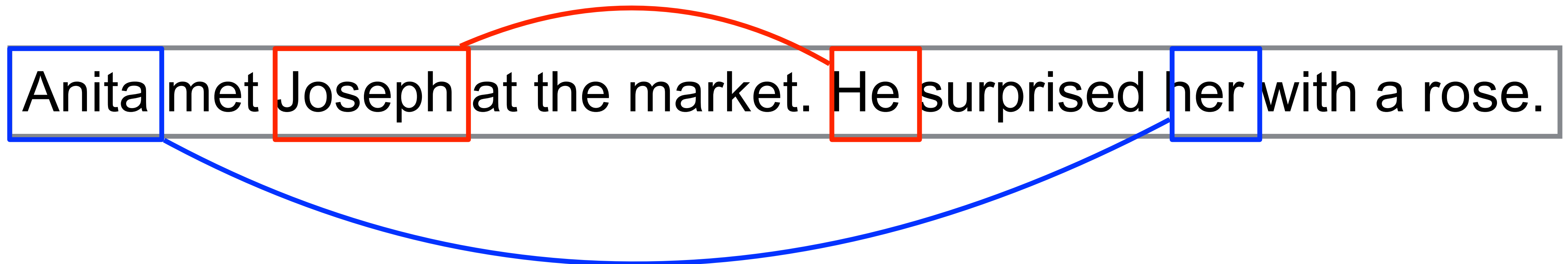
- Identify relationships between named entities

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