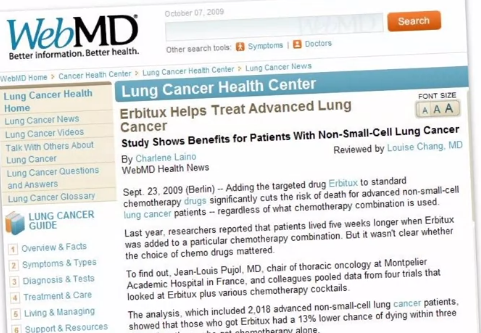
**Information Extraction:**

This document will look at how to use the topics addressed in this course in real applications.

The majority of information (80%) available in text format is **free text** which is unstructured and hard to work with. How do we extract the relevant information from this unstructured form?

**Information Extraction:**

Our goal in the following example is to extract relevant information from the text:



We might want to take the title and convert it to “Erbitux helps treat lung cancer”. We can drop the word “advanced” here because the title still makes complete sense without it. There is much more information on this webpage e.g. the author, the locations of publish, the data published, who reviewed the paper, and so on.

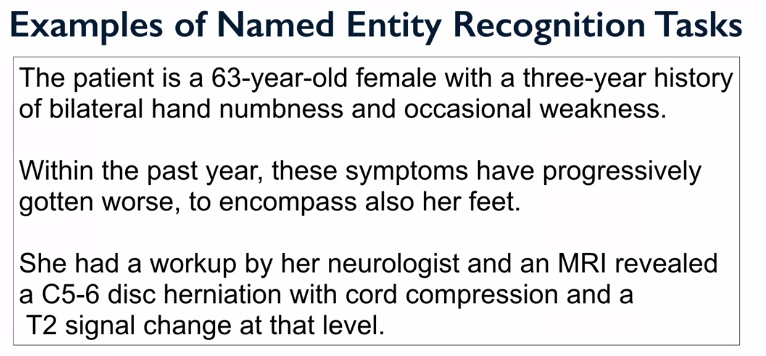
**Fields on Interest:**

**Named entities**: Noun phrases that are of a specific type and refer to specific individuals, places, organizations.

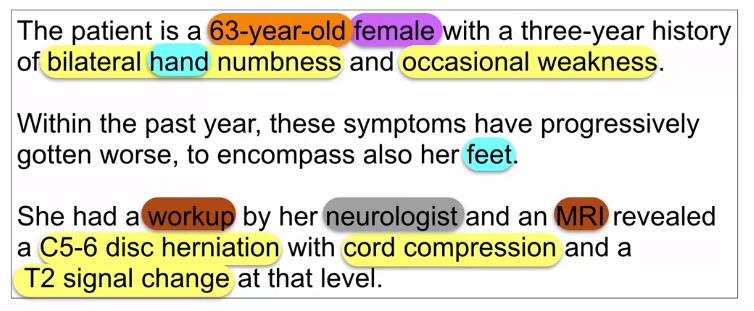
* News: People, Places, Dates, etc.
* Finance: Money, Companies, etc.
* Medicine: Diseases, Drugs, Procedures, etc.

It can be seen that the rules that apply to each of these entities is quite diverse. For example, when dealing with news its common to see places and people with capital letters, but with medicine words like cancer and lungs will not be capitalized.

With named entities we also need to understand where the entities start and end, and how to recognize them, this is known as **named entity recognition.** Once this entity has been found it should be labelled/classified into one of the groups. E.g. “Ramon Santiago” would be placed into the people group.



Named entities: age, diagnosis method, symptoms, diagnosis, gender, body parts.



We can see that some of the entities are subsets within other entities. How should we define the cut-off for the coarseness for out entities, e.g. why not C5-6 disc, or cord? These are the types of decision that need to be made when defining these named entity tasks.

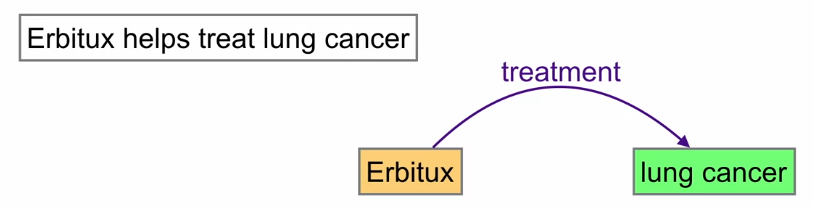
**Approaches to Identify Named Entities:**

* For well formatted fields like data, phone numbers, the use of regular expressions can be used.
* In machine learning these regular expressions are often used to produce features that can improve the model’s performance.

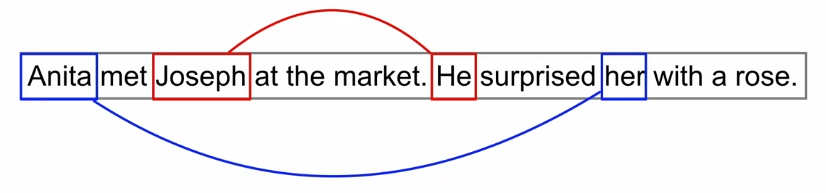
Often a combination of features will be used to make predictions, for example, the difference between phone numbers and fax number would be the numbers before the number or a word before it, this could be made into another feature. The next step would be to extract the relations.

**Relation Extraction**:

This step is used to identify the relationships between the named entities. E.g:



It is also important to be aware of **co-reference resolution** this is a way to understand future references to people or places with a semantic, e.g.



This type of relationship is important when we want to find an answer from a document, or at least the most appropriate answer. E.g. “what helps treat advanced lung cancer?” -> “Erbitux”.

