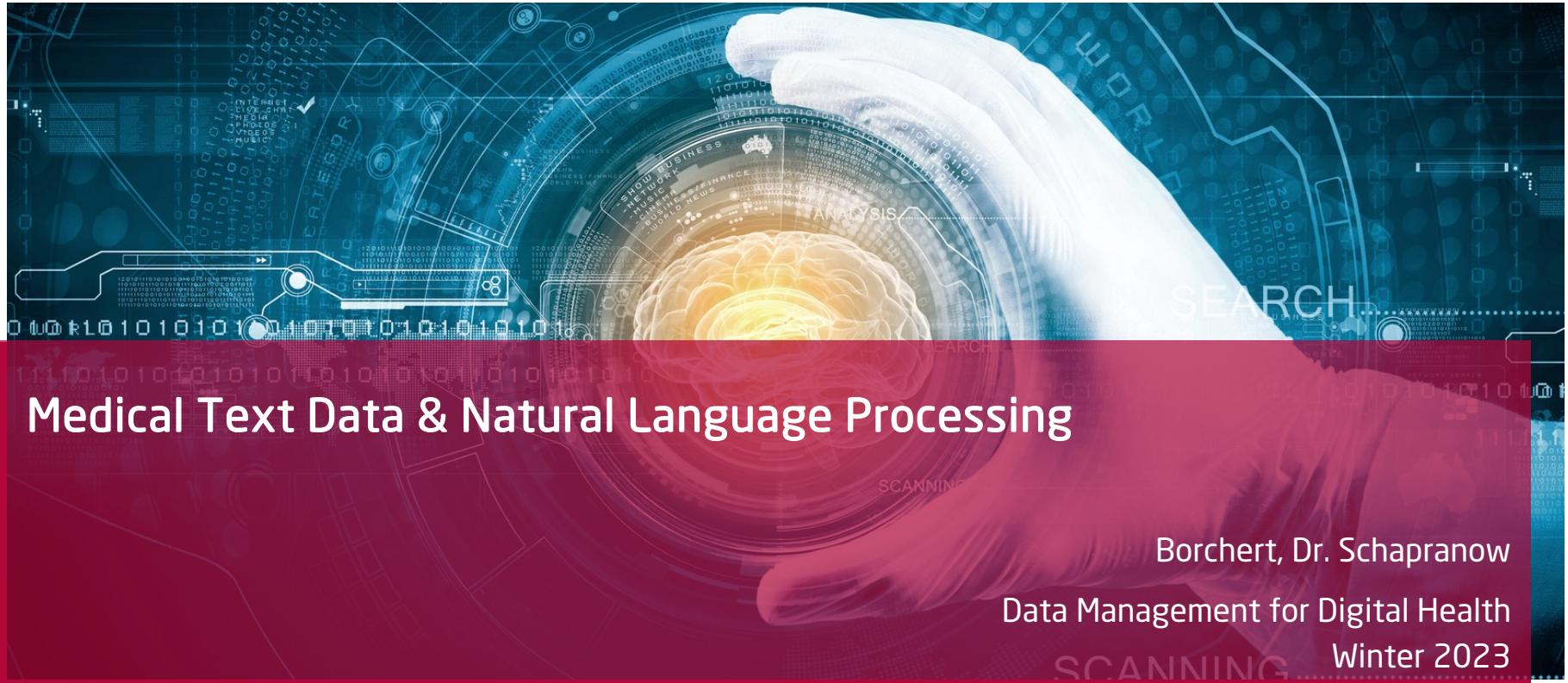
Hasso Plattner Institut
Digital Engineering · Universität Potsdam



Medical Text Data & Natural Language Processing

Borchert, Dr. Schapranow

Data Management for Digital Health
Winter 2023

Agenda
Pillars of the Lecture

Hasso Plattner Institut

Medical Use Cases

Biology Recap

Oncology

Nephrology

Infectious Diseases

Technology Foundation

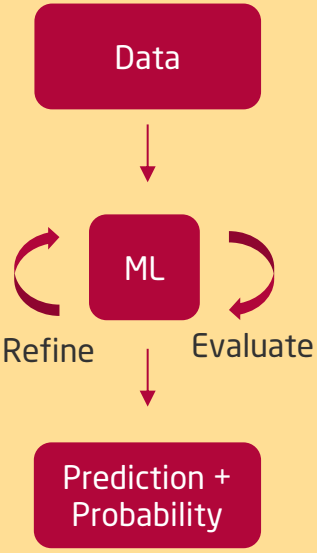
Data Sources

Data Formats

Processing and Analysis

Software Architectures

Machine Learning



Text Data & NLP

Data Management for Digital Health, Winter 2023

Agenda
Pillars of the Lecture

Hasso Plattner Institut

Medical Use Cases


Biology Recap


Oncology


Nephrology


Infectious Diseases

Technology Foundation

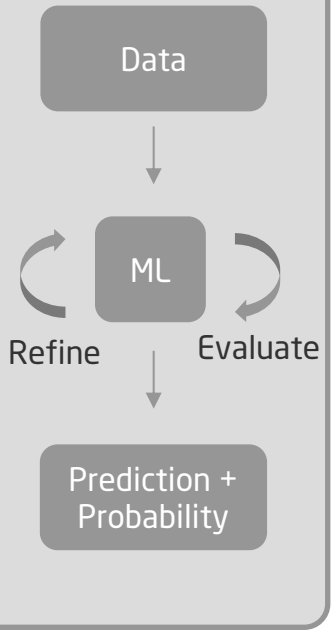

Data Sources


Data Formats

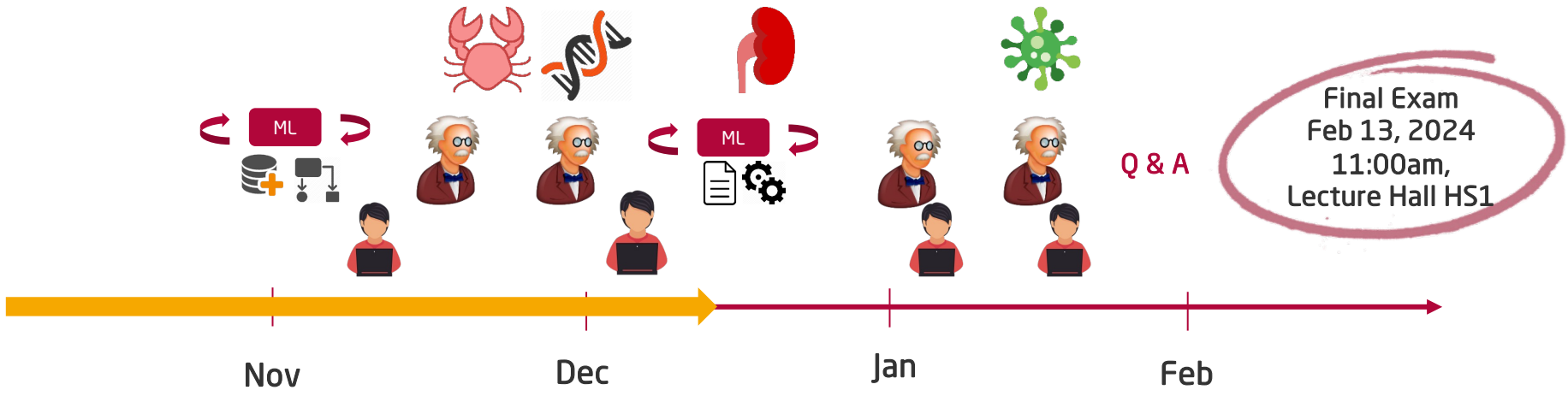

Processing and Analysis


Software Architectures

Machine Learning



Lecture Schedule



- Lecture Kickoff
- Actors in Healthcare
- Digital Health Data

- Machine Learning (ML) Foundations
- Use Case Oncology
- Biology Recap

- Natural Language Processing
- Use Case Nephrology & Intensive Care
- Supervised ML & Deep Learning

- Use Case Infectious Diseases
- Unsupervised ML

Agenda

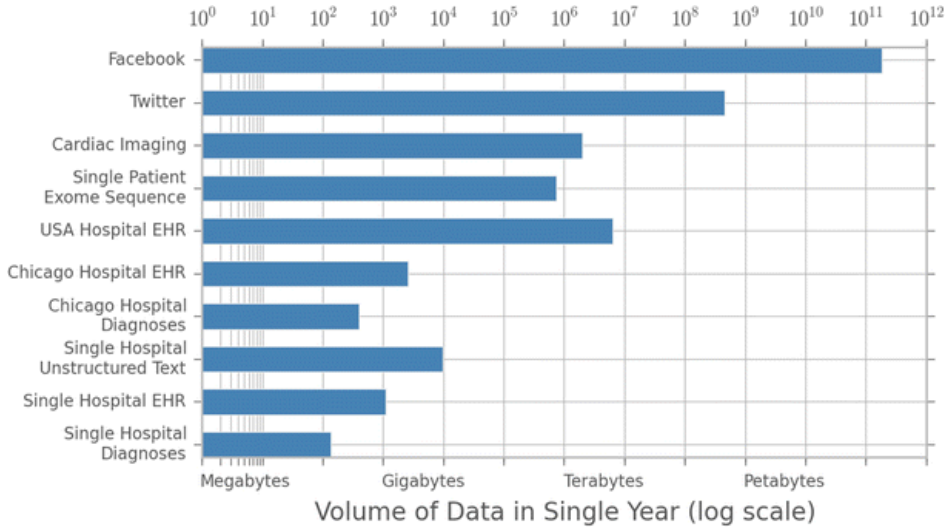
- Types of Medical Text
- Handling & Storing Strings
- Ontologies
- Introduction to Natural Language Processing
- Information Extraction

Volume of Textual and Unstructured Data

- Large proportion of medical data is unstructured

  → „80%“ is often cited, but hard to validate

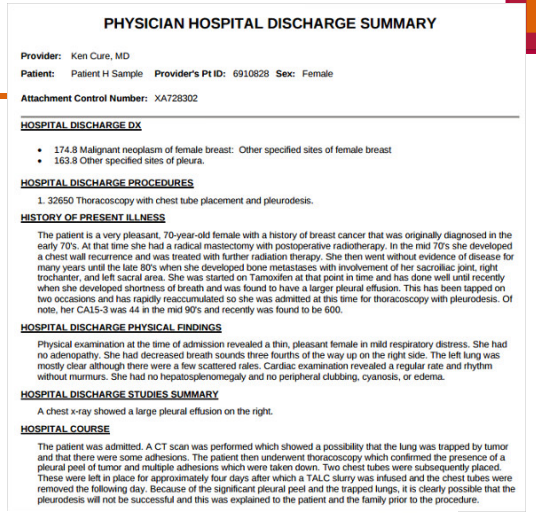
- According to Pah et. al (2015):
  - Unstructured text vs. structured EHR data in a single hospital: 1 order of magnitude
  - Imaging & sequencing data: many orders of magnitude larger



http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/#
Image: Pah, Adam Robert, et al. "Big data: what is it and what does it mean for cardiovascular research and prevention policy." *Current Cardiovascular Risk Reports* 9.1 (2015): 424.

Text-based Communication Between Physicians

- Clinicians with different specializations communicate via standardized documents
  - Referral to inpatient care: Admission note
  - Hospital to aftercare providers: Discharge summary
  - Radiologist to treating physician: Radiology report
  - Pathologist to treating physician: Pathology report
- Traditionally paper-based, but should be part of EHR
- Usually semi-structured:
  - Structured information
  - Free-text with (more or less) standardized structure
  - Classification systems



PHYSICIAN HOSPITAL DISCHARGE SUMMARY

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that was originally diagnosed in the early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid 70's she developed a chest wall recurrence and was treated with further radiation therapy. She then went without evidence of disease for many years until the late 80's when she developed bone metastases with involvement of her sacroiliac joint, right trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done well until recently when she developed shortness of breath and was found to have a larger pleural effusion. This has been tapped on two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy with pleurodesis. Of note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

HOSPITAL DISCHARGE PHYSICAL FINDINGS

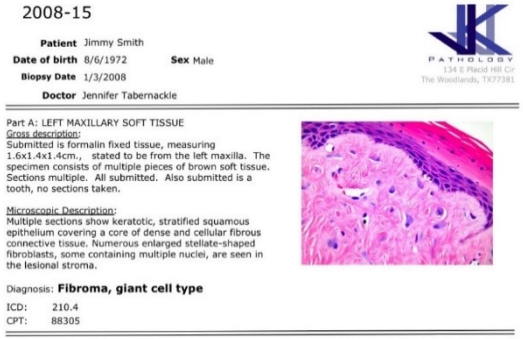
Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory distress. She had no adenopathy. She had decreased breath sounds three fourths of the way up on the right side. The left lung was mostly clear although there were a few scattered rales. Cardiac examination revealed a regular rate and rhythm without murmurs. She had no hepatosplenomegaly and no peripheral clubbing, cyanosis, or edema.

HOSPITAL DISCHARGE STUDIES SUMMARY

A chest x-ray showed a large pleural effusion on the right.

HOSPITAL COURSE

The patient was admitted. A CT scan was performed which showed a possibility that the lung was trapped by tumor and that there were some adhesions. The patient then underwent thoracoscopy which confirmed the presence of a pleural peel of tumor and multiple adhesions which were taken down. Two chest tubes were subsequently placed. These were left in place for approximately four days after which a TALC slurry was infused and the chest tubes were removed the following day. Because of the significant pleural peel and the trapped lungs, it is clearly possible that the pleurodesis will not be successful and this was explained to the patient and the family prior to the procedure.



2008-15

Patient Jimmy Smith
Date of birth 8/6/1972 **Sex** Male
Biopsy Date 1/3/2008
Doctor Jennifer Tabernackle

PATHOLOGY
134 E Placid Hill Cir
The Woodlands, TX77381

Part A: LEFT MAXILLARY SOFT TISSUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.

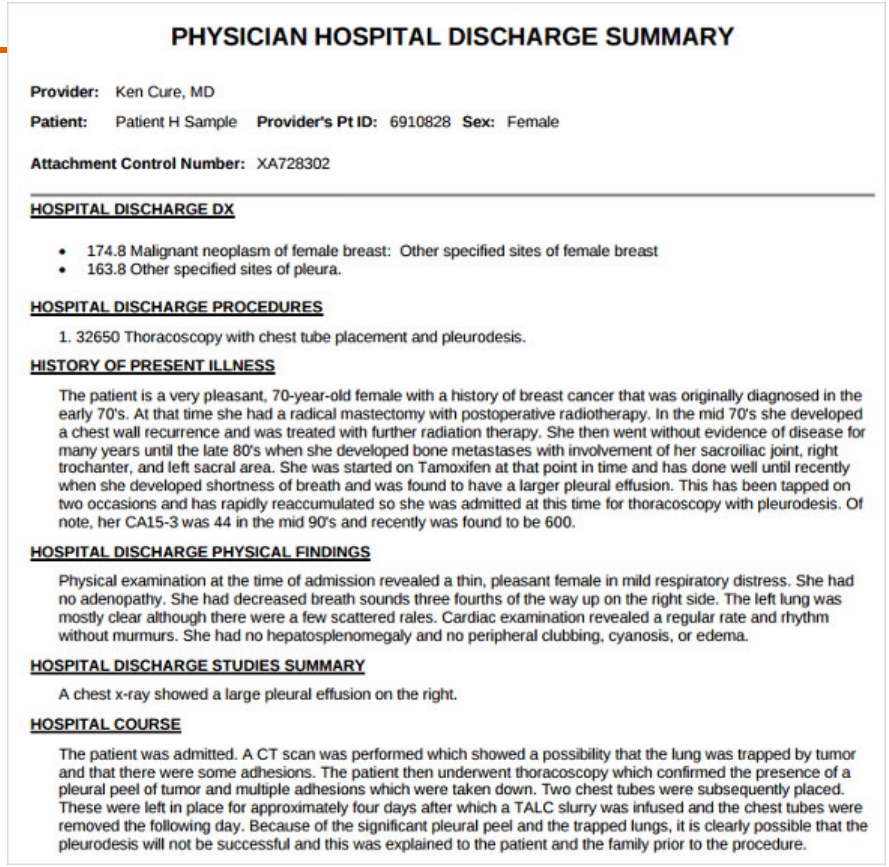
Microscopic Description:
Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

Discharge Summaries

- Provide information about hospital stay to aftercare providers
- Contents include:
  - Reason for hospitalization
  - Findings / diagnoses
  - Procedures / treatment
  - Patient's condition
  - Instructions for patients and families



PHYSICIAN HOSPITAL DISCHARGE SUMMARY

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that was originally diagnosed in the early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid 70's she developed a chest wall recurrence and was treated with further radiation therapy. She then went without evidence of disease for many years until the late 80's when she developed bone metastases with involvement of her sacroiliac joint, right trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done well until recently when she developed shortness of breath and was found to have a larger pleural effusion. This has been tapped on two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy with pleurodesis. Of note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

HOSPITAL DISCHARGE PHYSICAL FINDINGS

Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory distress. She had no adenopathy. She had decreased breath sounds three fourths of the way up on the right side. The left lung was mostly clear although there were a few scattered rales. Cardiac examination revealed a regular rate and rhythm without murmurs. She had no hepatosplenomegaly and no peripheral clubbing, cyanosis, or edema.

HOSPITAL DISCHARGE STUDIES SUMMARY

A chest x-ray showed a large pleural effusion on the right.

HOSPITAL COURSE

The patient was admitted. A CT scan was performed which showed a possibility that the lung was trapped by tumor and that there were some adhesions. The patient then underwent thoracoscopy which confirmed the presence of a pleural peel of tumor and multiple adhesions which were taken down. Two chest tubes were subsequently placed. These were left in place for approximately four days after which a TALC slurry was infused and the chest tubes were removed the following day. Because of the significant pleural peel and the trapped lungs, it is clearly possible that the pleurodesis will not be successful and this was explained to the patient and the family prior to the procedure.

https://www.sampletemplates.com/business-templates/discharge-summary.html

Pathology Reports

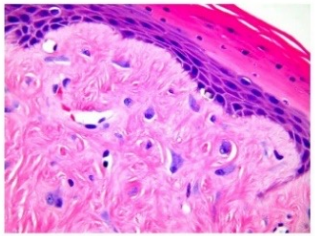
- Pathologists perform diagnoses for certain diseases (such as cancer) based on tissue samples
- Pathology report contains at least :
  - Identifying information
  - Gross description (size, shape, color)
  - Microscopic description (view with microscope)
  - Diagnosis (e.g., ICD code)
  - In oncology:
    - Grading (comparison with healthy / surrounding tissue)
    - Staging (extend of tumor growth and spread)

2008-15

Patient Jimmy Smith
Date of birth 8/6/1972 **Sex** Male
Biopsy Date 1/3/2008
Doctor Jennifer Tabernackle

PATHOLOGY
134 E Placid Hill Cir
The Woodlands, TX77381

Part A: LEFT MAXILLARY SOFT TISSUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.



Microscopic Description:
Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

EHR Data / Clinical Notes

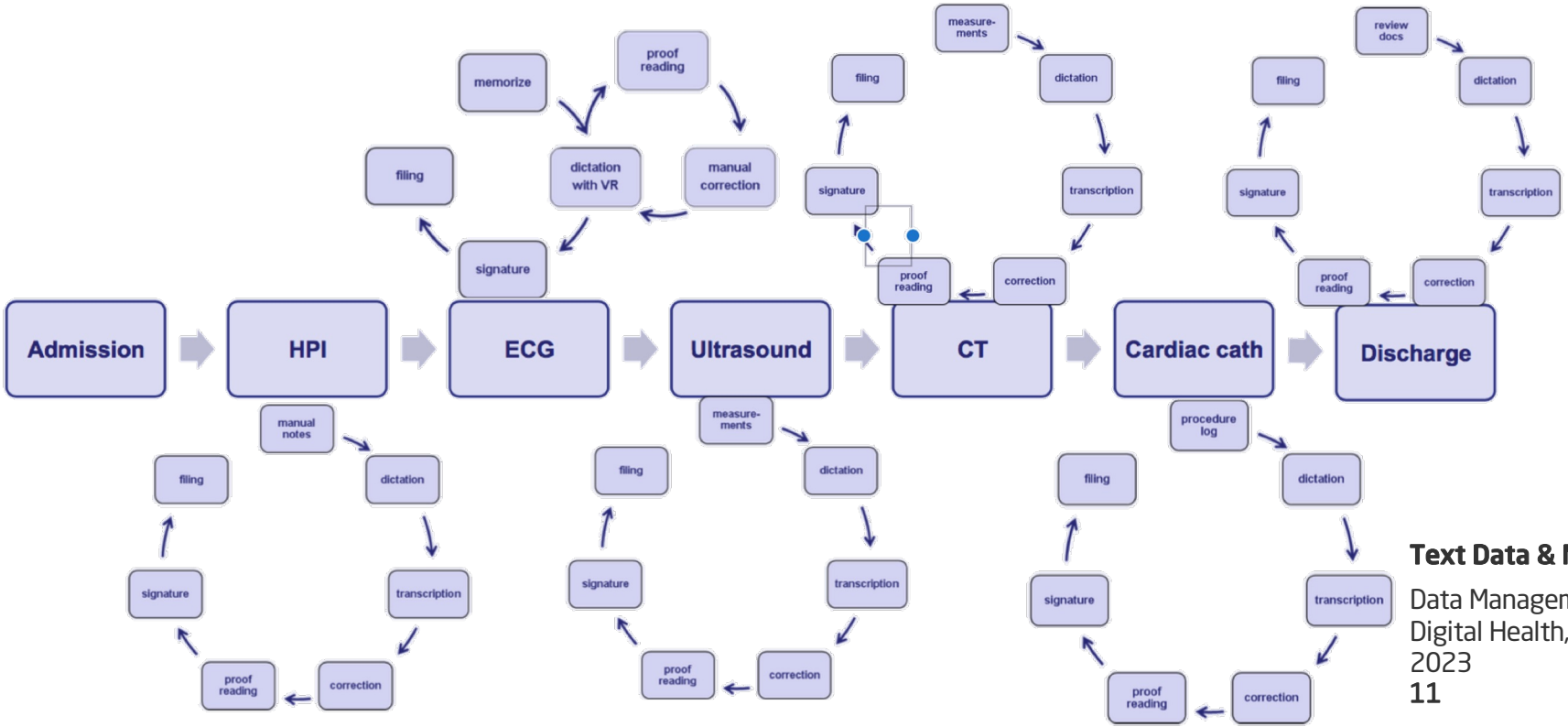
Clinical observations captured in free-text notes in EHR

Text Data & NLP

Data Management for Digital Health, Winter 2023

Image: https://www.emr-ehrs.com/gastrointestinal-electronic-health-records.php

Vicious Mandala of Document Generation



Text Data & NLP

Data Management for Digital Health, Winter 2023

Scientific Publications & Clinical Guidelines

Research findings are mainly disseminated via scientific publications (> 34M indexed on PubMed) and synthesized in clinical practice guidelines to enable evidence-based medicine

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Advanced

Format: Abstract | Send to

Lancet Oncol. 2012 Mar;13(3):239-46. doi: 10.1016/S1470-2045(11)70393-X. Epub 2012 Jan 26.

Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial.

Rosell R[1], Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, Palmero R, Garcia-Gomez R, Pallares C, Sanchez JM, Porta R, Cobo M, Garrido P, Longo F, Moran T, Insa A, De Marinis F, Corre R, Bover I, Illiano A, Dansin E, de Castro J, Milella M, Reguart N, Altavilla G, Jimenez U, Provencio M, Moreno MA, Terrasa J, Muñoz-Langa J, Valdivia J, Isla D, Domine M, Molinier O, Mazieres J, Baize N, Garcia-Campelo R, Robinet G, Rodriguez-Abreu D, Lopez-Vivanco G, Gebbia V, Ferrera-Delgado L, Bombaron P, Bernabe R, Bearz A, Artal A, Cortesi E, Rolfo C, Sanchez-Ronco M, Drozdowskyj A, Queralt C, de Aguirre I, Ramirez JL, Sanchez JJ, Molina MA, Taron M, Paz-Ares L; Spanish Lung Cancer Group in collaboration with Groupe Français de Pneumo-Cancérologie and Associazione Italiana Oncologia Toracica.

Author information

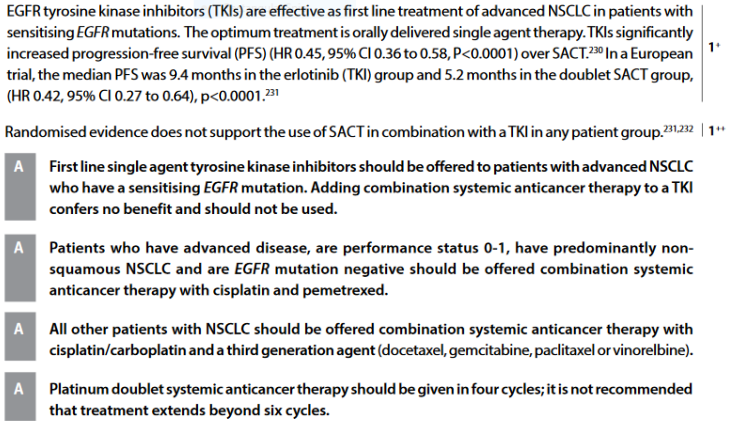
1 Catalan Institute of Oncology, Badalona, Spain. rrosell@iconcologia.net

Abstract

BACKGROUND: Erlotinib has been shown to improve progression-free survival compared with chemotherapy when given as first-line treatment for Asian patients with non-small-cell lung cancer (NSCLC) with activating EGFR mutations. We aimed to assess the safety and efficacy of erlotinib compared with standard chemotherapy for first-line treatment of European patients with advanced EGFR-mutation positive NSCLC.

METHODS: We undertook the open-label, randomised phase 3 EURTAC trial at 42 hospitals in France, Italy, and Spain. Eligible participants were adults (> 18 years) with NSCLC and EGFR mutations (exon 19 deletion or L858R mutation in exon 21) with no history of chemotherapy for metastatic disease (neoadjuvant or adjuvant chemotherapy ending ≥ 6 months before study entry was allowed). We randomly allocated participants (1:1) according to a computer-generated allocation schedule to receive oral erlotinib 150 mg per day or 3 week cycles of standard intravenous chemotherapy of cisplatin 75 mg/m(2) on day 1 plus docetaxel (75 mg/m(2) on day 1) or gemcitabine (1250 mg/m(2) on days 1 and 8). Carboplatin (AUC 6 with docetaxel 75 mg/m(2) or AUC 5 with gemcitabine

Clinical trial report on PubMed



EGFR tyrosine kinase inhibitors (TKIs) are effective as first line treatment of advanced NSCLC in patients with sensitising *EGFR* mutations. The optimum treatment is orally delivered single agent therapy. TKIs significantly increased progression-free survival (PFS) (HR 0.45, 95% CI 0.36 to 0.58, P<0.0001) over SACT.[230] In a European trial, the median PFS was 9.4 months in the erlotinib (TKI) group and 5.2 months in the doublet SACT group, (HR 0.42, 95% CI 0.27 to 0.64), p<0.0001.[231] 1+

Randomised evidence does not support the use of SACT in combination with a TKI in any patient group.[231,232] 1++

A **First line single agent tyrosine kinase inhibitors should be offered to patients with advanced NSCLC who have a sensitising *EGFR* mutation. Adding combination systemic anticancer therapy to a TKI confers no benefit and should not be used.**

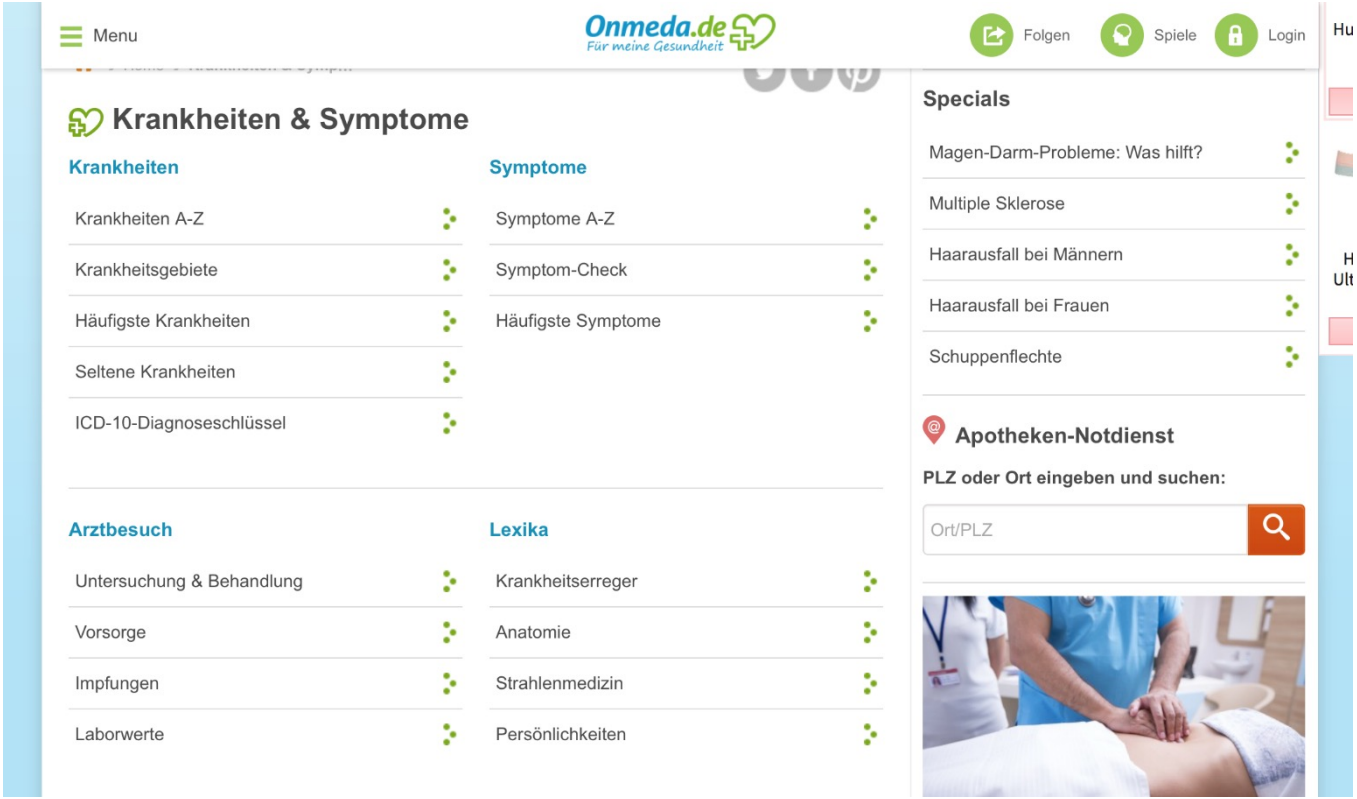
A **Patients who have advanced disease, are performance status 0-1, have predominantly non-squamous NSCLC and are *EGFR* mutation negative should be offered combination systemic anticancer therapy with cisplatin and pemetrexed.**

A **All other patients with NSCLC should be offered combination systemic anticancer therapy with cisplatin/carboplatin and a third generation agent** (docetaxel, gemcitabine, paclitaxel or vinorelbine).

A **Platinum doublet systemic anticancer therapy should be given in four cycles; it is not recommended that treatment extends beyond six cycles.**

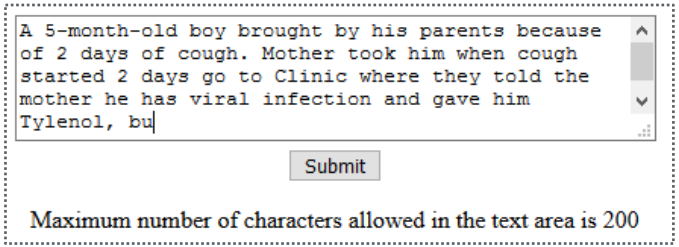
SIGN Guideline: Management of lung cancer (2014)

Encyclopedias, Forums, Social Media



Storing Text in Databases

- Text in databases is stored as strings
  - Fixed length (e.g., VARCHAR)
  - Variable length (e.g., CLOB)



- Encoding (mapping of characters to bits) must be specified for reading and writing, otherwise: `I%&/m{Jt`

Character	Unicode code point	Glyph
Latin A	U+0041	A
Latin sharp S	U+00DF	ß
Han for East	U+6771	東

- (Lossless) compression is challenging, as often no structure can be exploited

Searching Text Documents

- Sequential scan through documents only feasible for few / small documents (Ctrl+F, grep)
- Specialized database technology necessary for efficient storage and retrieval in large collections of text documents (e.g., full-text indices)
  - Elastic Search
  - SAP HANA Text Analytics
- More flexibility through
  - Fuzzy search for inexact matches
  - Regular expressions

Osimertinib or EGFR-TKIs/chemotherapy in patients with EGFR-mutated advanced nonsmall cell lung cancer: A meta-analysis.

Huang L[1], Huang H[1], Zhou XP[2], Liu JF[1], Li CR[1], Fang M[1], Wu JR[1].

Author information

1 Department of Clinical Laboratory, The Affiliated Tumor Hospital of Guangxi Medical University.
2 Department of Clinical Laboratory, The First Affiliated Hospital of Guangxi University of Chinese Medicine, Nanning, Guangxi, China.

Abstract

BACKGROUND: The aim of this meta-analysis is to investigate the impact of Osimertinib on treatment efficacy in advanced nonsmall cell lung cancer (NSCLC).

METHODS: Trials comparing Osimertinib against epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs)/chemotherapy in patients with NSCLC with an epidermal growth factor receptor (EGFR) mutation were included, and the pooled data for progression-free survival (PFS), overall survival (OS), overall response rate (ORR), disease control rate (DCR), and adverse events (AEs) were analyzed.

RESULTS: Analysis results based on 6 eligible trials showed that Osimertinib significantly improved the overall PFS (hazard ratio [HR] =0.38, 95% confidence interval [CI]=0.29-0.50), improved the OS (HR=0.66, 95% CI=0.48-0.89), increased the ORR (odds ratio [OR]=1.76, 95% CI=1.14-2.72), increased the overall DCR (OR=1.18, 95% CI=1.02-1.37), and reduced the grade 3 or greater AEs (relative ratio [RR]=0.50, 95% CI=0.33-0.75) in all subgroups except in the ORR in the Exon 19 deletion (Ex19del) and/or L858R subgroup. Compared to patients with Ex19del and/or L858R mutation, patients with the T790M mutation had the benefits of a greater PFS (41.7%), a greater ORR (80.0%), a greater DCR (71.2%), and fewer grade 3 or greater AEs (70.7%) (each P<.05). Race, sex, age, EGFR mutation, and smoking history may significantly predict additional benefits from Osimertinib, but there were no significant differences between subgroups stratified by these clinical characteristics.

lung cancer | Alle hervorheben | Groß-/Kleinschreibung | Ganze Wörter | 1 von 6 Übereinstimmungen

Raw Data | Distinct values | Analysis

lung cancer | 21 rows retrieved - 57 ms | Execute

DUI	TA_RULE	TA_COUNTER	TA_TOKEN	TA_SENTENCE	TA_OFFSET
D016159	Entity Extraction	13	LUNG CANCER	2	203
D016159	Entity Extraction	14	LUNG CANCER	2	203
D016159	Entity Extraction	15	LUNG CANCER	2	203
D017687	Entity Extraction	16	lung cancer	3	374
D017687	Entity Extraction	17	lung cancer	3	374
D017687	Entity Extraction	18	lung cancer	3	374

Handling Strings: Regular Expressions

- Formal syntax for describing sets of strings, e.g.:
  - Alternatives, e.g.: `(analyse|analyze)` or `analy[sz]e`
  - Wildcards, e.g.: `\d`

    (matches any digit)
  - Quantifiers, e.g.: `\d+(\.\d+)?`

    matches one or more (+) digits optionally (?) followed by a decimal point and one more other digits (= decimal numbers)
- Widely used for all kinds of text processing besides search
- Deeply rooted in theoretical computer science and formal language theory
- Can easily become very complex

Poll

Which set of strings is matched by this regular expression?

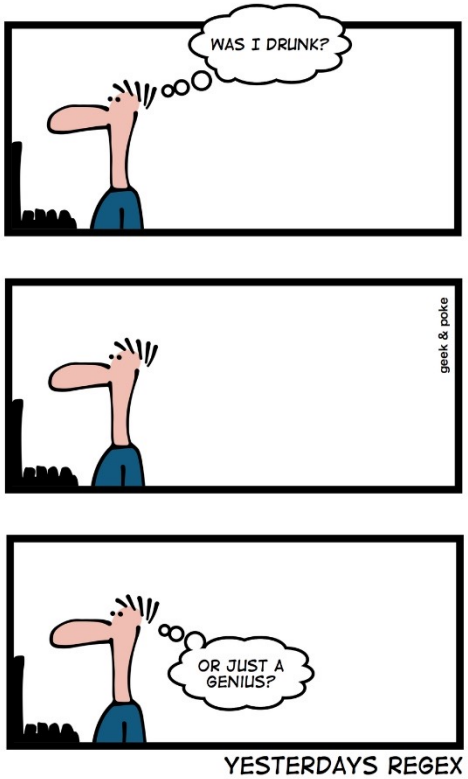
```
^[a-z0-9!#$%&'*+\/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+\/=?^_`{|}~-]+)*@(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?
```

A: The set of valid telephone numbers

B: The set of legal German person names

C: The set of valid e-mail addresses

D: The set of all ASCII characters
(the expression is completely useless)

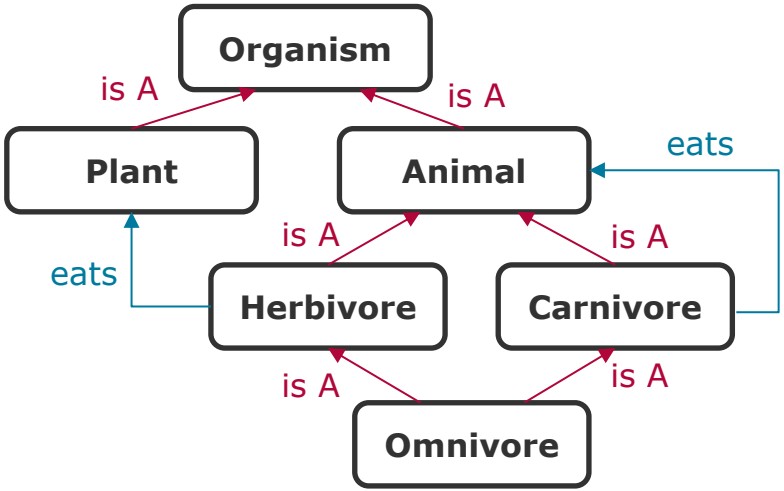


https://gist.github.com/gregseth/5582254

http://geek-and-poke.com/geekandpoke/2013/12/3/yesterdays-regex

Representing Meaning: Ontologies

- Strings have no “meaning” per se
- Ontologies = representation and formal naming of concepts in a domain and relations among them
- Process of creating an ontology is known as knowledge engineering
- Have a long history in philosophy and are used in Artificial Intelligence since 1970s

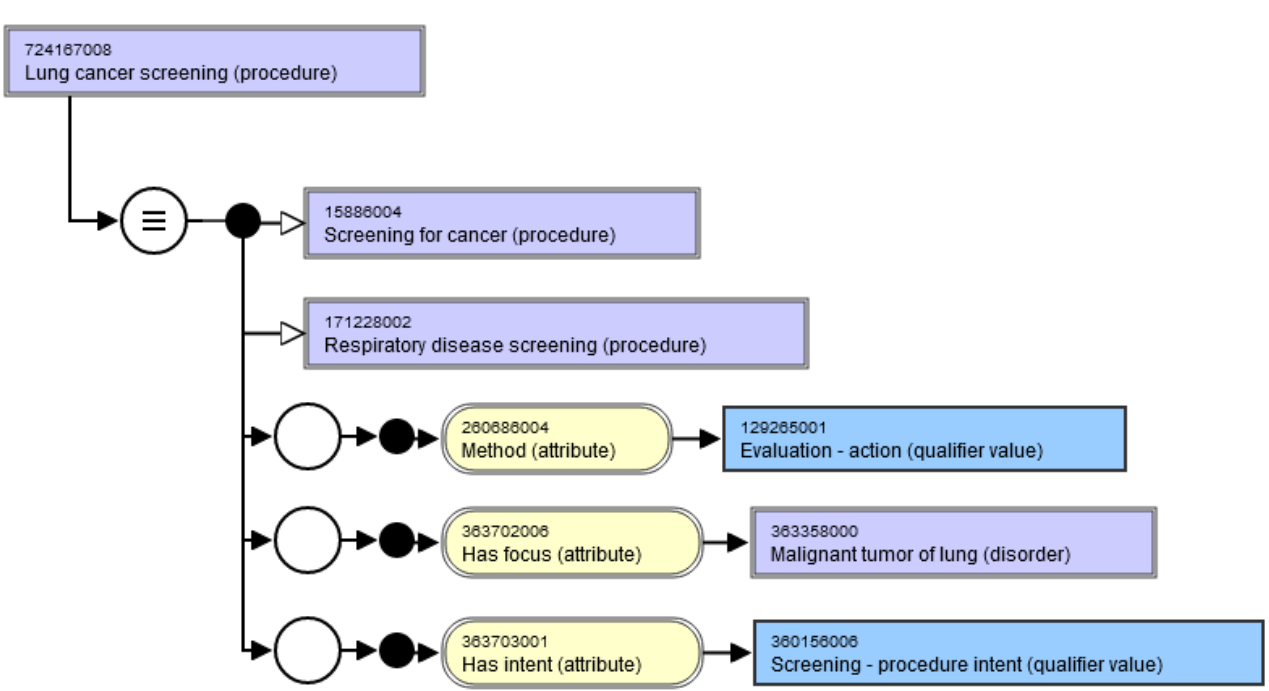


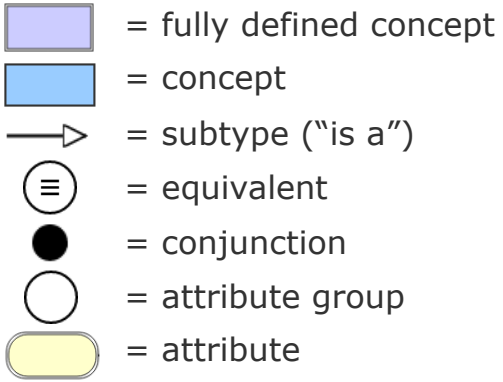
Ontologies & Controlled Vocabularies in Medicine

- Examples of widely used medical ontology-based systems:
  - Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT)
  - Domain-specific ontologies like GeneOntology
- Can be efficiently processed by algorithms and software and enable *semantic interoperability* between systems
- Controlled vocabularies have similar, but less rich semantics, e.g. Medical Subject Headings (MeSH) terms used by MEDLINE / PubMed
- Unified Medical Language System (UMLS): Compendium of >100 controlled vocabularies and semantic network

http://www.snomed.org/
https://www.nlm.nih.gov/research/umls/index.html
http://geneontology.org/
https://www.nlm.nih.gov/mesh/meshhome.html

SNOMED CT Example





https://browser.ihtsdotools.org

UMLS (Unified Medical Language System)

Hasso Plattner Institut

Concept: [C0281477] Screening for malignant neoplasm of lung

DA Date 1996-01-01 06:00:00.000000000

MR Major Revision Date 2017-11-23 06:00:00.000000000

ST Status R

Semantic Type

Diagnostic Procedure [T060]

Synonyms (13)

- Lung cancer screening
  - SNOMEDCT_US
- Lung cancer screening (procedure)
- Lung neoplasm screening
- Screening for Lung Cancer
  - NCI
- Screening for malignant neoplasm of lung
- Screening for malignant neoplasm of lung (procedure)
- cancer lung screening
- early detection of lung cancer
- lung cancer early detection
- lung cancer screen
- lung cancer screening
- screening for lung cancer
- screening lung cancer

Relations (57) REL | RELA | RSAB| String | CUI

[: 1 - 10 :]

CHD | isa | SNOMEDCT_US | Screening for malignant neoplasm of lung | C0281477
CHD | isa | SCTSPA | Screening for malignant neoplasm of lung | C0281477
PAR | inverse_isa | NCI | Screening for cancer | C0199230
PAR | inverse_isa | SCTSPA | Screening for cancer | C0199230
PAR | inverse_isa | SNOMEDCT_US | Screening for cancer | C0199230
PAR | inverse_isa | SNOMEDCT_US | Screening for malignant neoplasm of lung | C0281477
PAR | inverse_isa | SCTSPA | Screening for malignant neoplasm of lung | C0281477
PAR | inverse_isa | SNOMEDCT_US | Procedure for lung lesion | C0396558
PAR | inverse_isa | SCTSPA | Procedure for lung lesion | C0396558
PAR | inverse_isa | SNOMEDCT_US | Procedure for lung lesion | C0396558

The UMLS also contains around
270k German terms
vs. 10M English terms (< 3%)

https://uts.nlm.nih.gov/home.html

Making Sense of Text: Natural Language Processing & Text Mining

Hasso Plattner Institut

- "Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things." *Chowdhury (2005)*

- "Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. [...] There is a field called computational linguistics (also known as natural language processing) which is making a lot of progress in doing small subtasks in text analysis." *Hearst (2003)*

Why is Biomedical Language Processing hard?

Hasso Plattner Institut

- NLP in general is hard (and considered "AI complete"), because of:
  - Ambiguity: *"John kissed his wife, and so did Sam"*
  - Flexibility: *"Natalie ran out of the room"*
    "Natalie ran out of flour"
  - Implicitness: *"Who should drive to the party?"*
    "Susie's on antibiotics."
- Biomedical NLP is hard due to:
  - Domain-specific, technical, dense, and constantly evolving language:
    "Southern blot analysis was performed using EcoRI and methylation-sensitive EagI restriction enzymes followed by hybridization with StB12.3 probe targeting the FMR1 gene on chromosome Xq27."
  - Sensitive healthcare data

Aspects of Human Language Studied by Linguists

(Sample of) structural subfields of linguistics:

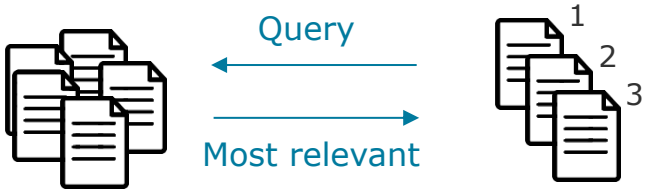
- Phonetics := study of sounds of human language
- Phonology := study of sound systems in human language
- Morphology := study of the formation and internal structure of words
- Syntax := study of the formation and internal structure of sentences
- Semantics := study of the meaning of sentences
- Pragmatics := study of the way sentences with their semantic meaning are used for particular communicative goals

Bender, Emily (2013): Linguistic Fundamentals for Natural Language Processing

Some Common NLP Tasks

Level: collections of text documents

- Information Retrieval



- Topic modeling

Some Common NLP Tasks

Level: single text documents

- Document Classification

- Summarization
  (abstractive or extractive)

- Machine Translation

Some Common NLP Tasks

Level: sentences / paragraphs

- Sentiment Analysis: „I feel a bit sad“ → Affective State: Polarity: ☹ Strength: 0.4
- Dialogue Systems: „I feel a bit sad“ → „Why do you feel sad?“

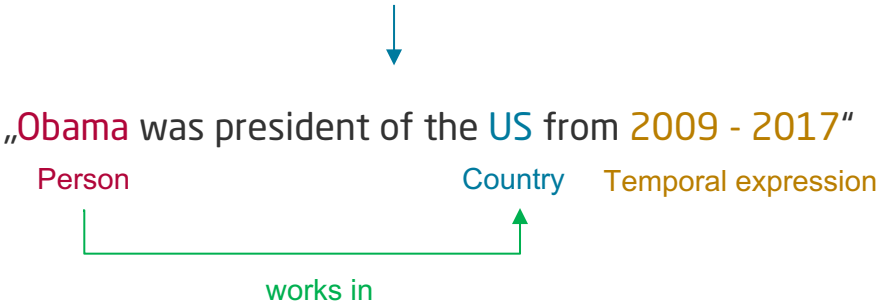
Some Common NLP Tasks

Level: word / tokens / text spans

- Language Modelling: "Natalie ran out of .... " →

  P("flour") = 0.01
  P("the") = 0.01
  P("plutonium") = 0.0001
  P("or") = 0.0000001

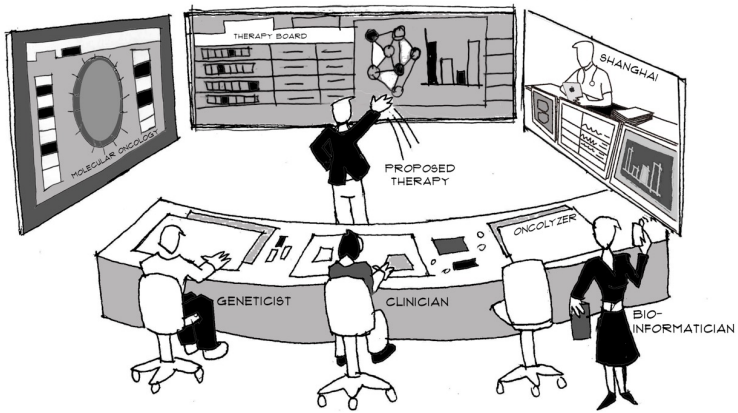
- Information Extraction: „Obama was president of the US from 2009 - 2017"



NLP Use Case: Molecular Tumor Board

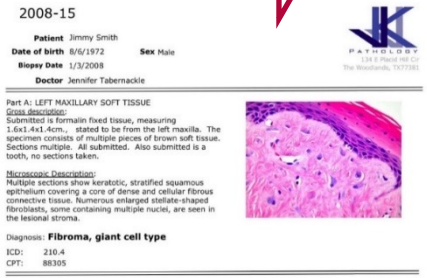
Patient data

- Genomic data
- Imaging data
- Clinical data
  - Patient history
  - Pathology reports
  - Radiology reports
  - …



Medical knowledge

- Variant / gene databases
- Drug databases
- Cellular pathways
- Clinical Trials
- Clinical Guidelines
- Research publications



2008-15

Patient Jimmy Smith
Date of birth 8/6/1972 Sex Male
Biopsy Date 1/3/2008
Doctor Jennifer Tabernackle

Part A: LEFT MAXILLARY SOFT TISSUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.

Microscopic Description:
Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

6.25.	Evidence-based Recommendation	2017
Grade of Recommendation **B**	After complete removal of a traditional serrated adenoma or sessile serrated adenoma, the follow-up should be the same as for classic adenomas.	
Level of Evidence **3b**	Sources: [576, 617, 618]	
	Consensus	

PubMed

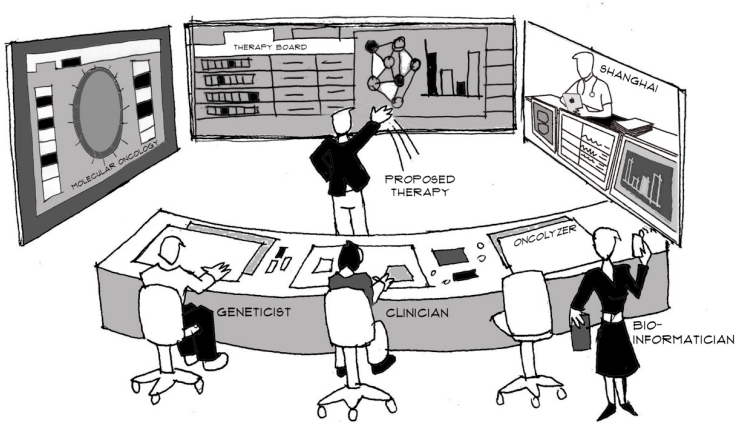
ML and Corpora

Data Management for Digital Health, Winter 2023

NLP Use Case: Molecular Tumor Board

Patient data

- Genomic data
- Imaging data
- Clinical data
  - Patient history
  - Pathology reports
  - Radiology reports
  - …



Medical knowledge

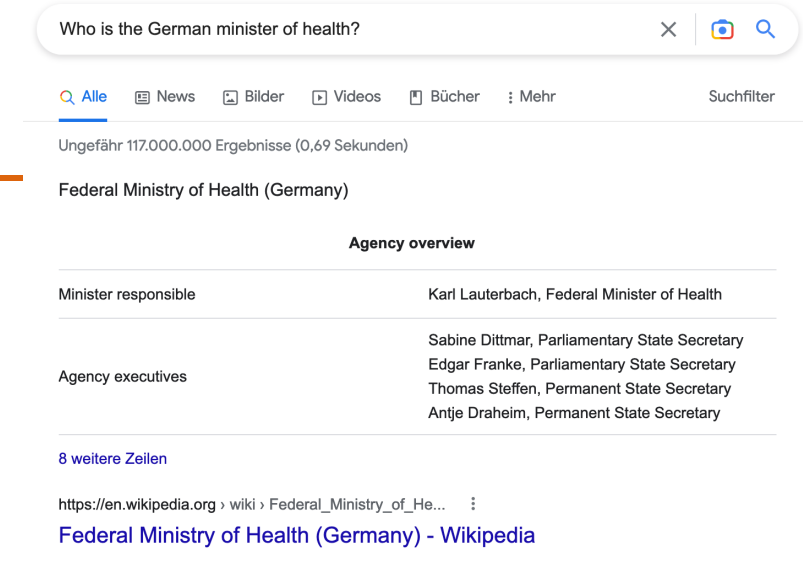
- Variant / gene databases
- Drug databases
- Cellular pathways
- Clinical Trials
- Clinical Guidelines
- Research publications

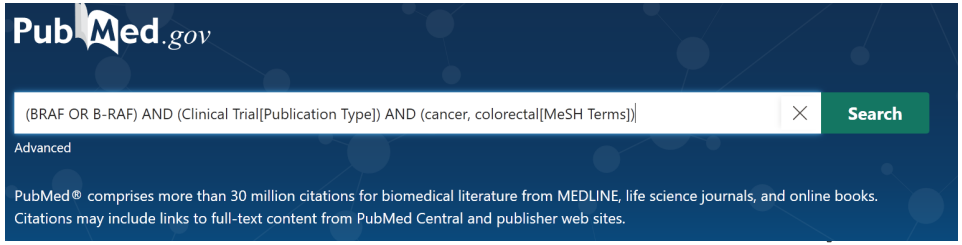
ML and Corpora

Data Management for Digital Health, Winter 2023

Information Retrieval

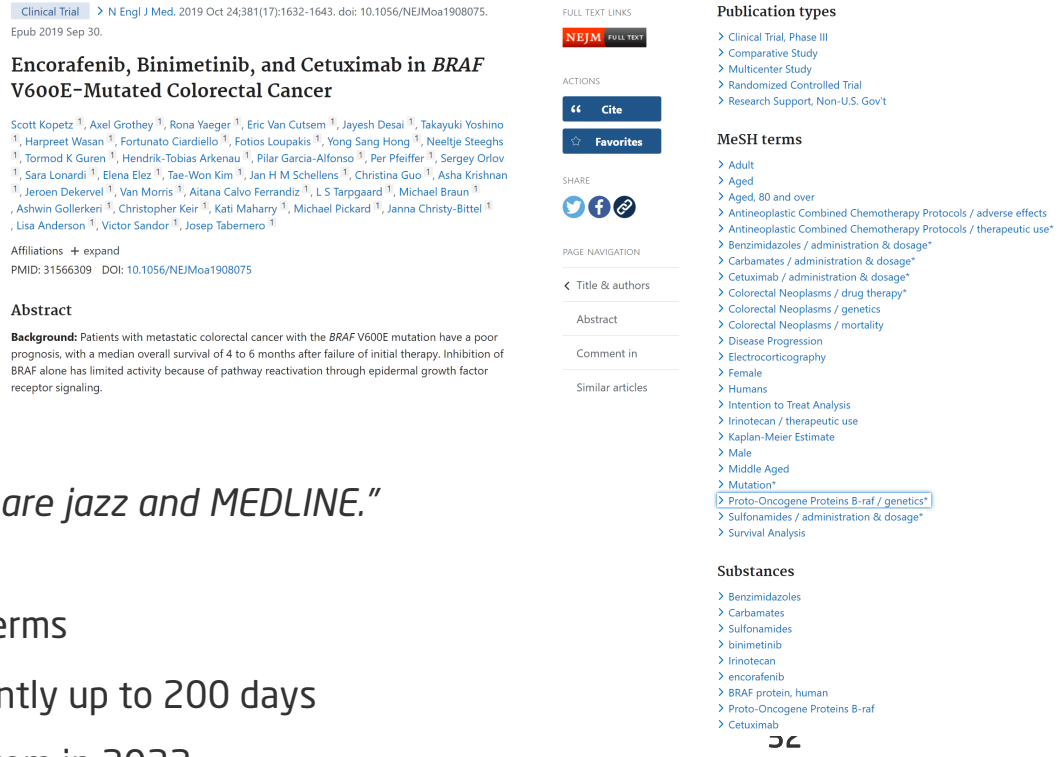
- From a (large) set of documents, find the most relevant ones given a user query
- User query can be:
  - in natural language (Question Answering)
  - structured (e.g., Boolean Search)
- Can be formulated as a document classification problem: P(Relevant | Document, User Query)





PubMed
https://pubmed.ncbi.nlm.nih.gov/

- PubMed is a search engine that provides access to:
  - MEDLINE (citations, abstracts)
  - PubMed Central (full-texts)
  - Some text books
- >30 M citations (and growing)



Clinical Trial › N Engl J Med. 2019 Oct 24;381(17):1632-1643. doi: 10.1056/NEJMoa1908075. Epub 2019 Sep 30.

Encorafenib, Binimetinib, and Cetuximab in *BRAF* V600E–Mutated Colorectal Cancer

Scott Kopetz [1], Axel Grothey [1], Rona Yaeger [1], Eric Van Cutsem [1], Jayesh Desai [1], Takayuki Yoshino [1], Harpreet Wasan [1], Fortunato Ciardiello [1], Fotios Loupakis [1], Yong Sang Hong [1], Neeltje Steeghs [1], Tormod K Guren [1], Hendrik-Tobias Arkenau [1], Pilar Garcia-Alfonso [1], Per Pfeiffer [1], Sergey Orlov [1], Sara Lonardi [1], Elena Elez [1], Tae-Won Kim [1], Jan H M Schellens [1], Christina Guo [1], Asha Krishnan [1], Jeroen Dekervel [1], Van Morris [1], Aitana Calvo Ferrandiz [1], L S Tarpgaard [1], Michael Braun [1], Ashwin Gollerkeri [1], Christopher Keir [1], Kati Maharry [1], Michael Pickard [1], Janna Christy-Bittel [1], Lisa Anderson [1], Victor Sandor [1], Josep Tabernero [1]

Affiliations + expand

PMID: 31566309 DOI: 10.1056/NEJMoa1908075

Abstract

Background: Patients with metastatic colorectal cancer with the *BRAF* V600E mutation have a poor prognosis, with a median overall survival of 4 to 6 months after failure of initial therapy. Inhibition of BRAF alone has limited activity because of pathway reactivation through epidermal growth factor receptor signaling.

FULL TEXT LINKS

NEJM FULL TEXT

ACTIONS

Cite

Favorites

SHARE

PAGE NAVIGATION

Title & authors

Abstract

Comment in

Similar articles

Publication types

› Clinical Trial, Phase III
› Comparative Study
› Multicenter Study
› Randomized Controlled Trial
› Research Support, Non-U.S. Gov't

MeSH terms

› Adult
› Aged
› Aged, 80 and over
› Antineoplastic Combined Chemotherapy Protocols / adverse effects
› Antineoplastic Combined Chemotherapy Protocols / therapeutic use*
› Benzimidazoles / administration & dosage*
› Carbamates / administration & dosage*
› Cetuximab / administration & dosage*
› Colorectal Neoplasms / drug therapy*
› Colorectal Neoplasms / genetics
› Colorectal Neoplasms / mortality
› Disease Progression
› Electrocorticography
› Female
› Humans
› Intention to Treat Analysis
› Irinotecan / therapeutic use
› Kaplan-Meier Estimate
› Male
› Middle Aged
› Mutation*
› Proto-Oncogene Proteins B-raf / genetics*
› Sulfonamides / administration & dosage*
› Survival Analysis

Substances

› Benzimidazoles
› Carbamates
› Sulfonamides
› binimetinib
› Irinotecan
› encorafenib
› BRAF protein, human
› Proto-Oncogene Proteins B-raf
› Cetuximab

- *"America's two greatest gifts to the world are jazz and MEDLINE."* (BMJ Editorial from 2001)
- MEDLINE citations are indexed via MeSH terms
  - until recently: manually with delay currently up to 200 days
  - switched to automated (NLP) based system in 2022

LitVar

FILTERS

TOP JOURNALS | PUBLICATION TYPE | PART OF PUBLICATION

BRAF V600E was resolved to **p.V600E / rs113488022 (BRAF)**.

Showing 1 to 15 of 12338 publications.

Page 1 of 823

Most co-occured entities

DISEASE
- Melanoma (5032)
- Neoplasms (3595)
- Colorectal Neoplasms (1708)
- Thyroid cancer, papillary (1336)
- Thyroid Neoplasms (754)

more

CHEMICAL
- vemurafenib (3075)
- dabrafenib (1006)
- trametinib (644)
- valine-valine-saquinavir (231)
- sorafenib (216)

more

1 **Light-controlled inhibition of BRAFV600E kinase.**

PMID:31252305 PubTator Central Eur J Med Chem 2019

HOORENS MWH, OURAILIDOU ME, RODAT T, VAN DER WOUDEN PE, KOBAURI P, KRIEGS M, PEIFER C, FERINGA BL, DEKKER FJ, SZYMANSKI W

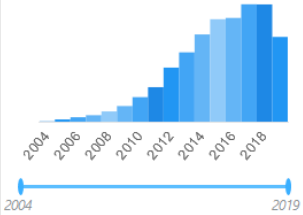
ABSTRACT Metastatic melanoma is amongst the most difficult types of cancer to treat, with current therapies mainly relying on the inhibition of the BRAFV600E mutant kinase.

ABSTRACT Here we show the development of BRAFV600E kinase inhibitors of which the activity can be switched on and off reversibly with light, offering the possibility to overcome problems of systemic drug activity by selectively activating the drug at the desired site of action.

ABSTRACT This research offers inspiration for the development of therapies for metastatic melanoma in which tumor tissue is treated with an active BRAFV600E inhibitor with high spatial and temporal resolution, thus limiting the damage to other tissues.

less

Results by year



rs113488022 *(BRAF V600E)*

Species: *Homo sapiens*
Position: 7:140753336
Clinical Significance: pathogenic

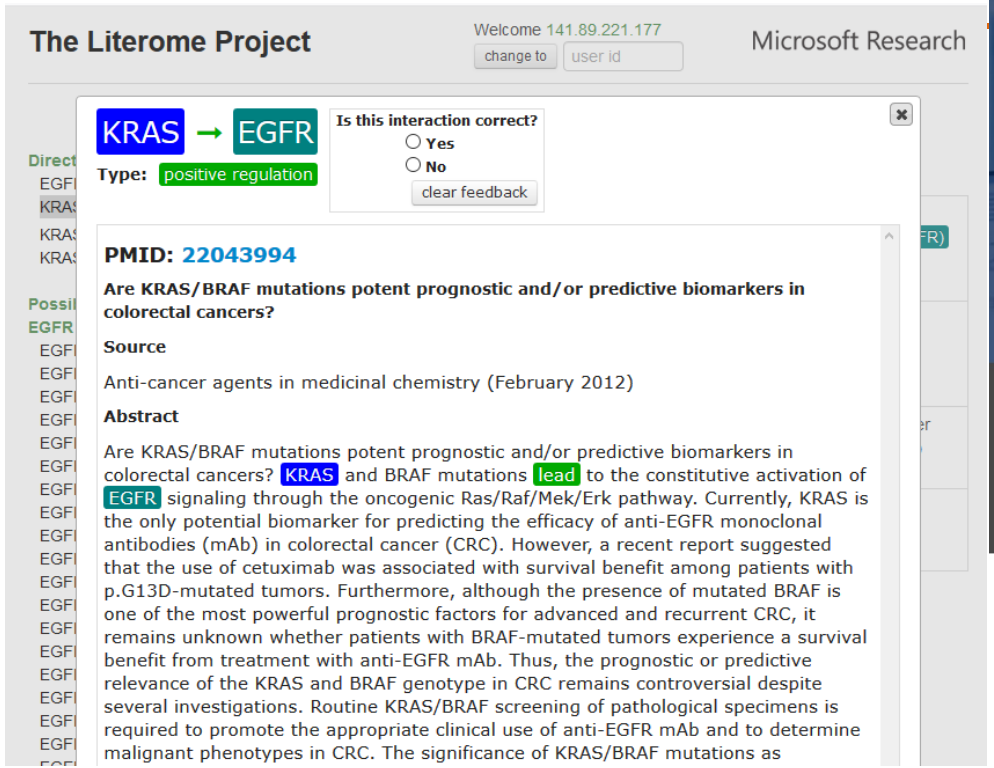
https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/#

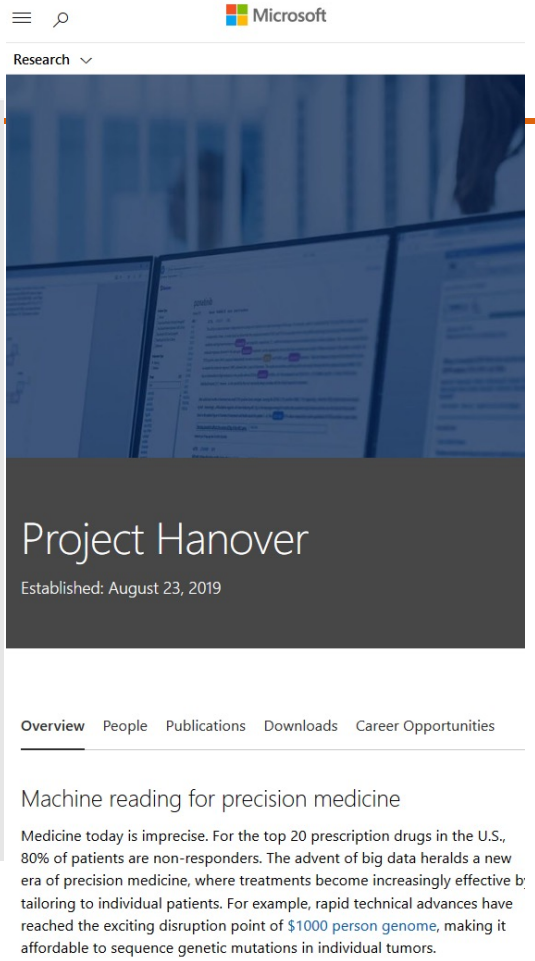
Hasso Plattner Institut

ML and Corpora

Data Management for Digital Health, Winter 2023

Precision Health NLP @ Microsoft





Hasso Plattner Institut

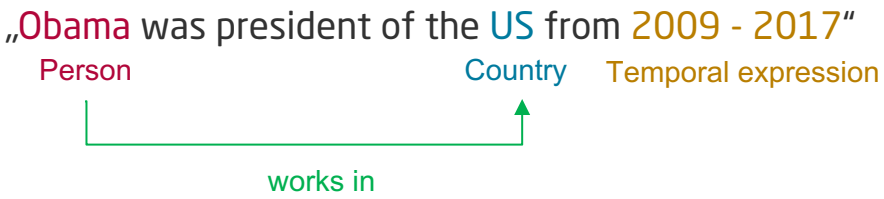
ML and Corpora

Data Management for Digital Health, Winter 2023

https://www.microsoft.com/en-us/research/project/project-hanover/
http://literome.azurewebsites.net

Information Extraction

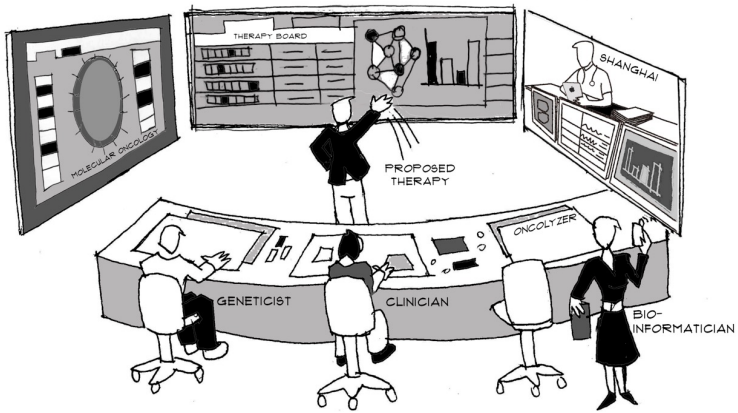
Extracting structured information from on unstructured text



NLP Use Case: Molecular Tumor Board

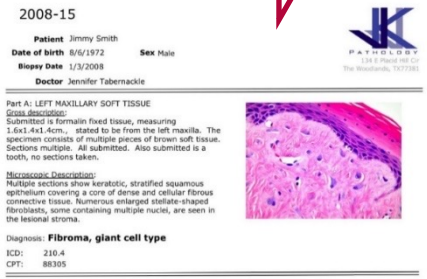
Patient data

- Genomic data
- Imaging data
- Clinical data
  - Patient history
  - Pathology reports
  - Radiology reports
  - …



Medical knowledge

- Variant / gene databases
- Drug databases
- Cellular pathways
- Clinical Trials
- Clinical Guidelines
- Research publications



2008-15

Patient Jimmy Smith
Date of birth 8/6/1972 Sex Male
Biopsy Date 1/3/2008
Doctor Jennifer Tabernackle

Part A: LEFT MAXILLARY SOFT TISSUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.

Microscopic Description:
Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

ML and Corpora

Data Management for Digital Health, Winter 2023

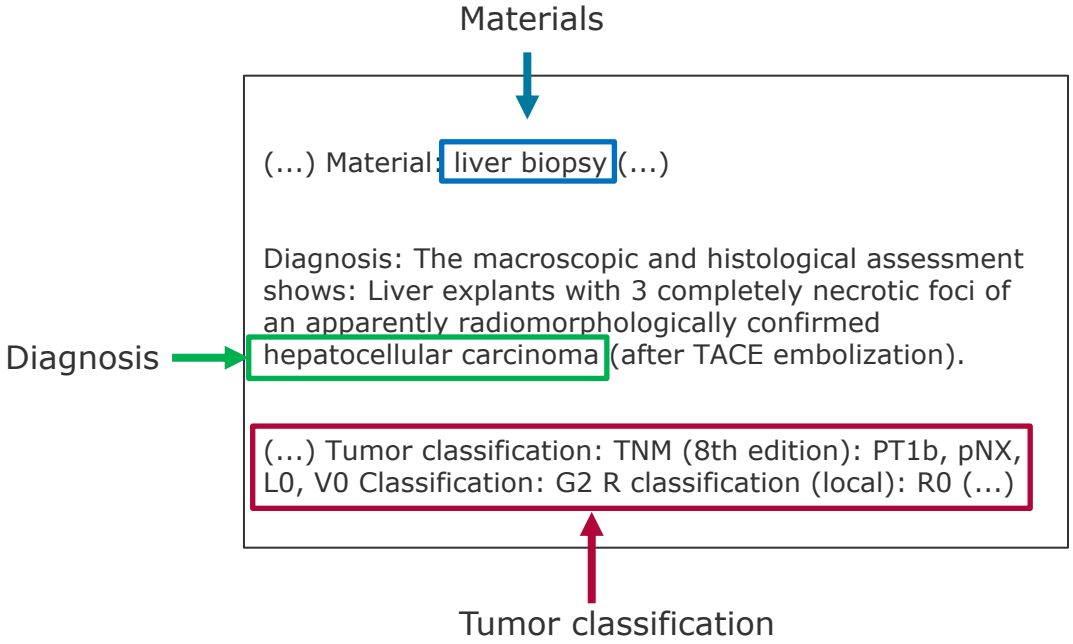
Information Extraction from Pathology Reports

(...) Material: liver biopsy (...)

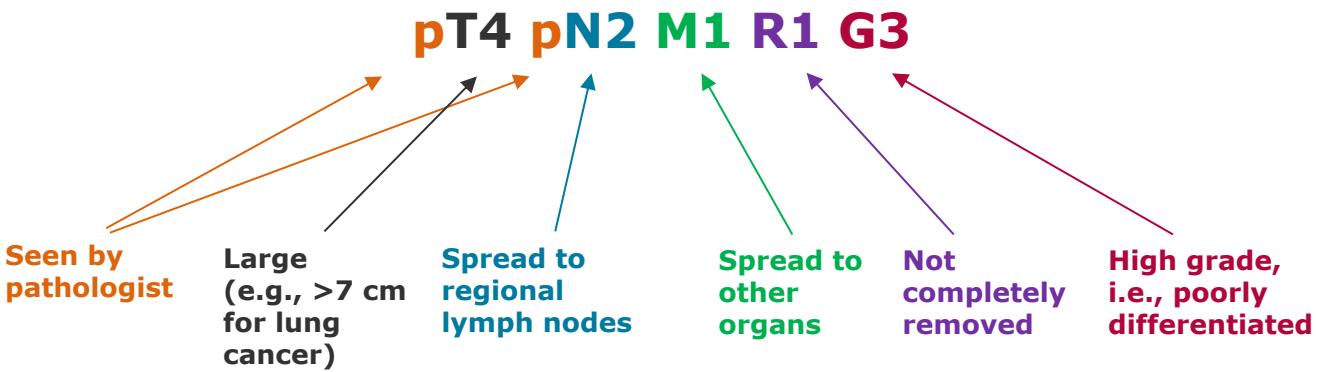
Diagnosis: The macroscopic and histological assessment shows: Liver explants with 3 completely necrotic foci of an apparently radiomorphologically confirmed hepatocellular carcinoma (after TACE embolization).

(...) Tumor classification: TNM (8th edition): PT1b, pNX, L0, V0 Classification: G2 R classification (local): R0 (...)

Information Extraction from Pathology Reports



Recap: TNM Classification Example



Rule-based TNM Extraction

Regular expressions

More regular expressions

Custom tokenizer (!)

Even more regular expressions

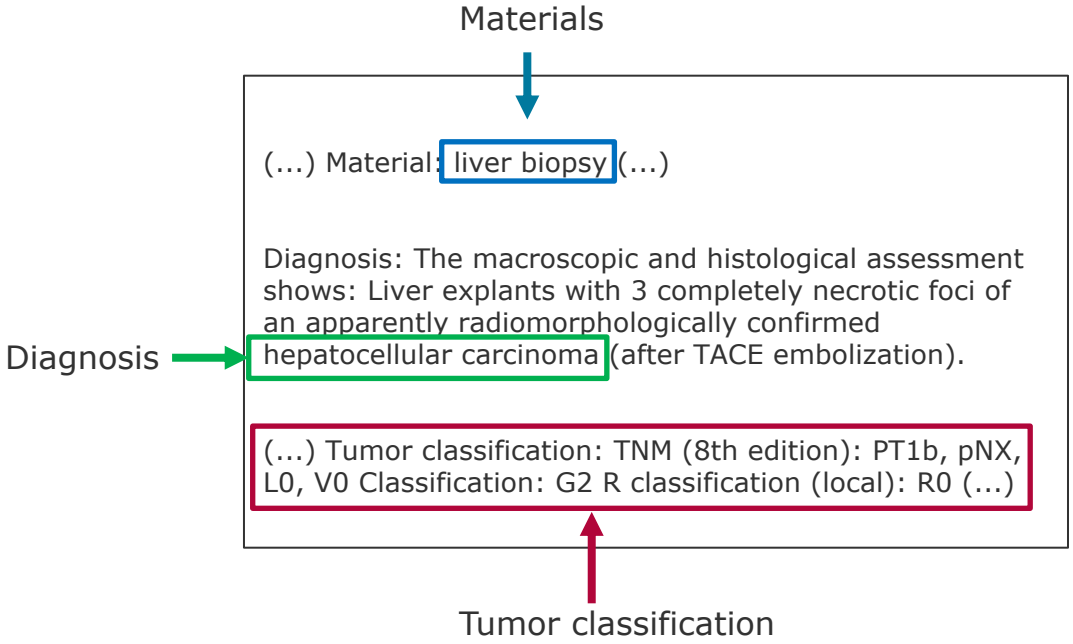
```
class RuleTNMExtractor():

    __tnm_rules = {
        'T' : r"[yr]?[ry]?[pc]?T([0-4][a-d]?|is|a|X|x)",
        'N' : r"[yr]?[ry]?[pc]?N([0-3][a-d]?|X|x)",
        'M' : r"[yr]?[ry]?[pc]?M([0-1][a-b]?|X|x)",
        'L' : r"[pc]?L[0-1Xx]",
        'V' : r"[pc]?V[0-2Xx]",
        'Pn': r"[pc]?Pn[0-1Xx]",
        'SX': r"[pc]?SX[0-3Xx]",
        'R' : r"[pc]?R[0-2][ab]?",
        'G' : r"G[1-4Xx]"
    }

    def __init__(self, language):
        self.nlp = load_spacy(language)
        rules = self.nlp.Defaults.tokenizer_exceptions

        infixes = list(self.nlp.Defaults.infixes)
        infixes.append(r'[\(\)-]')
        infixes = spacy.util.compile_infix_regex(tuple(infixes)).finditer

        prefixes = list(self.nlp.Defaults.prefixes)
        prefixes.extend(list(self.__tnm_rules.values()))
        prefixes.append(r'[-/"§\$&\\]')
        prefixes = spacy.util.compile_prefix_regex(tuple(prefixes)).search

        suffixes = list(self.nlp.Defaults.suffixes)
        suffixes = spacy.util.compile_suffix_regex(tuple(suffixes)).search

        self.nlp.tokenizer = Tokenizer(self.nlp.vocab, ···

        self.matcher = Matcher(self.nlp.vocab)

        def add_rule(k, v):
            self.matcher.add(k, None, [
                {"TEXT": {"REGEX" : '(?<![A-Za-z0-9])' + v}}
            ])
            self.matcher.add(k, None, [
                {"TEXT": {"REGEX" : '(?<![A-Za-z0-9])' + v}},
                {"TEXT": {"REGEX" : r'\s'}, "OP" : "*"},
                {"TEXT": '(' },
                {"TEXT": {"REGEX" : r'[^\(\)]'}, "OP": "+"},
                {"TEXT": ')' }
            ])
```

Hasso Plattner Institut

ML and Corpora

Data Management for Digital Health, Winter 2023

Information Extraction from Pathology Reports



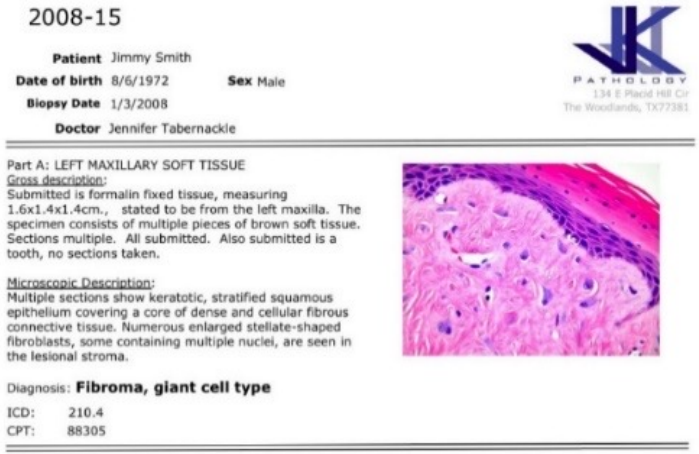
Characteristics of Clinical Text

- In clinical context, text data is created via:
  - Keyboard
  - OCR of printed documents
  - speech-to-text (common in radiology)
- Advantage: more expressive and flexible than structured forms
- But: flexibility of natural language can make interpretation of text hard, e.g.:
  - Abbreviations (BP – Blood Pressure, Pt. – Patient, etoh – Alcohol)
  - Synonyms (Aspirin / acetylsalicylic acid)
  - Homonyms ("dermatome" can refer to an area of the skin / a surgical instrument)
  - Errors in spelling, punctuation or grammar → no well-formed sentences

Image: https://www.healthcareitnews.com/news/fast-growing-medical-scribe-industry-poses-risks-patients-and-product-design

Clinical Practice Guidelines vs. Clinical Text

6.25.	Evidence-based Recommendation	2017
Grade of Recommendation **B**	After complete removal of a traditional serrated adenoma or sessile serrated adenoma, the follow-up should be the same as for classic adenomas.	
Level of Evidence **3b**	Sources: [576, 617, 618]	
	Consensus	



2008-15

Patient Jimmy Smith
Date of birth 8/6/1972 **Sex** Male
Biopsy Date 1/3/2008
Doctor Jennifer Tabernackle

Part A: LEFT MAXILLARY SOFT TISSUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.

Microscopic Description:
Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

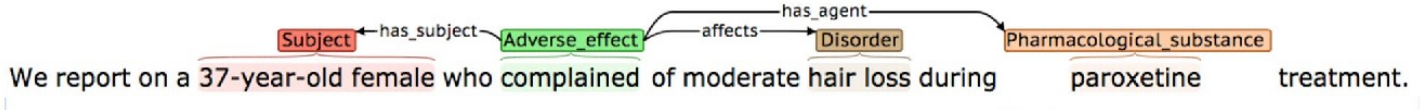
Both:

- Similar clinical terminology
- Created in many different national languages

Clinical Guidelines:

- Well-formed, scientific text
- No protected health information (PHI)

Tasks for Information Extraction in Clinical Context



Named Entity Recognition

Identifying instances of classes, e.g., *paroxetine* as a pharmacological substance

Named Entity Normalization

(aka Entity Linking)

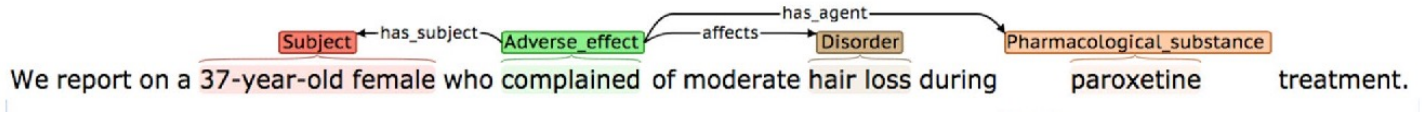
Mapping of entities to unique identity, e.g., *hair loss* to its ICD-10 code

(Temporal) Relation Extraction

Identifying *n*-ary relations between entities, e.g., *adverse_effect* (*paroxetine, hairloss*)

Thompson, Paul, et al. "Annotation and detection of drug effects in text for pharmacovigilance." *Journal of cheminformatics* 10.1 (2018): 37.

Extractable Information



Entity	Type	UMLS Concept	UMLS Semantic Type
37-year-old female	Subject	C0043210 Woman	Population Group
complained	Adverse Effect	C0277786 Chief complaint (finding)	Finding
hair loss	Disorder	C0002170 Alopecia	Disease or Syndrome
paroxetine	Pharmacological Substance	C0070122 paroxetine	Organic Chemical

Relation	Entity 1	Entity 2
Treatment	37-year-old female	paroxetine
Has Adverse Effect	paroxetine	hair loss

(Named) Entity Recognition

- Find and categorize mentions of entities (instances of certain classes)
- Definition of entity is task-specific, can also include quantities, durations, etc.
- IOB format used to describe tokens as Inside / Outside / Beginning of named entity
  - Input: The patient underwent a CT scan in April .
  - Output: O O O O B-PROC I-PROC O O O
  - Various flavors: IO, BIOES
  - Simple, but not very expressive (does not allow nesting)
- Text usually not given in properly tokenized format...

NER Pipeline
Preprocessing – Boundary Detection

Hasso Plattner Institut

Boundary Detection

The patient underwent a CT scan in April. | It did not reveal any abnormalities.

- Not trivial, e.g., because of nested punctuation (“Mr. Brown”) or indirect speech

NER Pipeline
Preprocessing – Tokenization

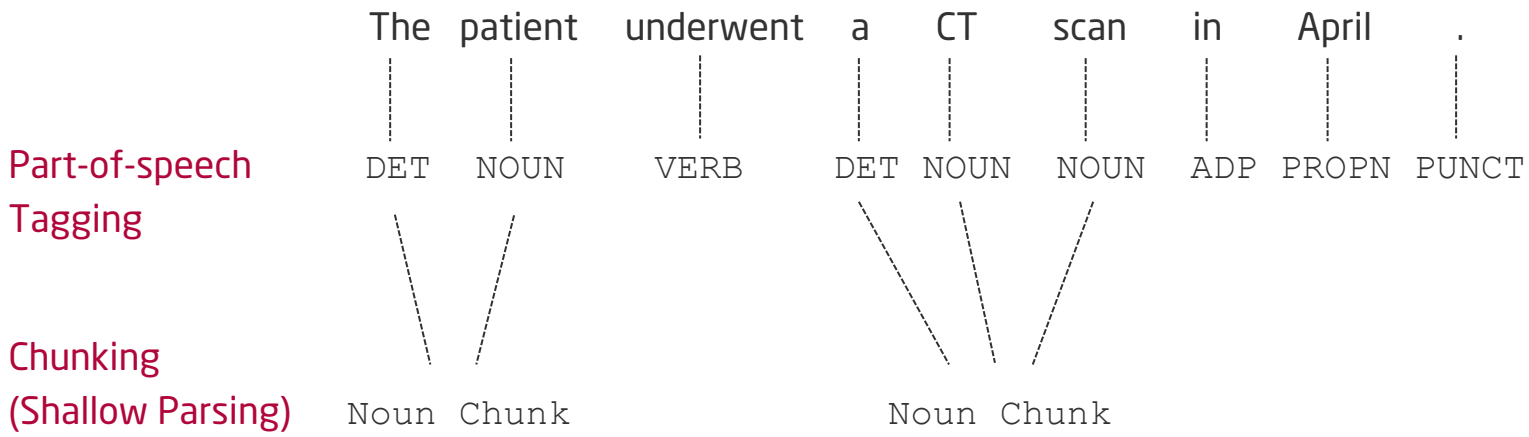
Hasso Plattner Institut

Tokenization

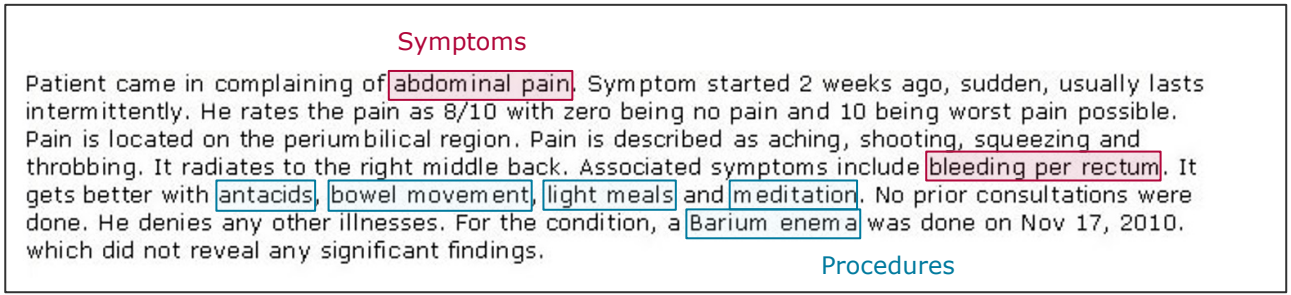
The | patient | underwent | a | CT | scan | in | April.

- Tokens are loosely equivalent to “words”, but details depend on the use case
  - aren’t | aren | t or are | n’t
  - 37-year-old or 37 | year | old
- Tricky in languages like German
- Really tricky in languages like Chinese

NER Pipeline
Preprocessing – Syntactic Analysis

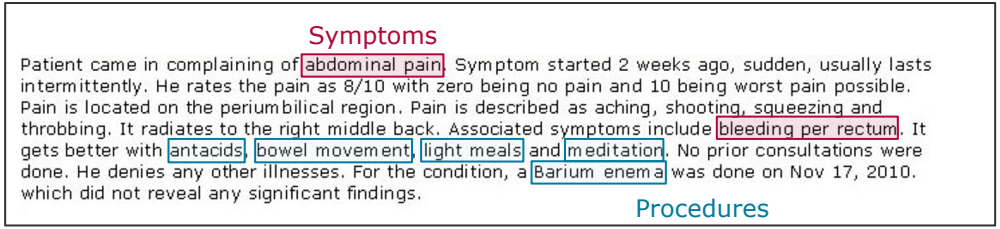


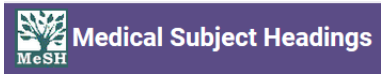
A More Complex Example



Dictionary-based Approach to NER

- Simple idea: use a dictionary (also called gazetteer) for Named Entity Recognition
- Ontologies and controlled vocabularies can be used to derive such lists (e.g., all terms from "Signs and Symptoms" category in MeSH)
- Fuzzy matching usually necessary to account for variations in spelling
- Named Entity Normalization for free
- Limitations:
  - For many applications, maintaining a complete dictionary hardly feasible (e.g., drug names)
  - Challenging to deal with homonyms, multi-token entities, co-reference and context

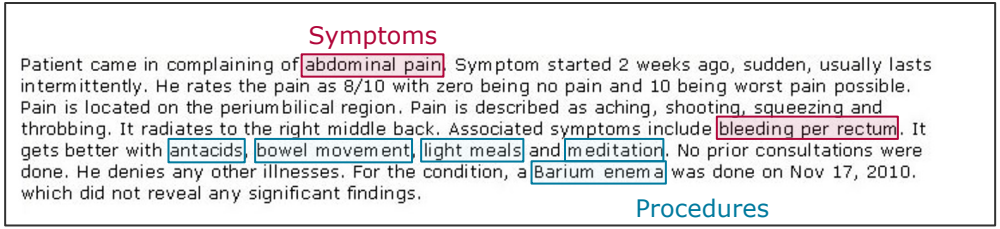




Pathological Conditions, Signs and Symptoms [C23]
Signs and Symptoms [C23.888]
Signs and Symptoms, Digestive [C23.888.821]
Abdominal Pain [C23.888.821.030]
Abdomen, Acute [C23.888.821.030.249]
Aerophagy [C23.888.821.061]
Anorexia [C23.888.821.108]

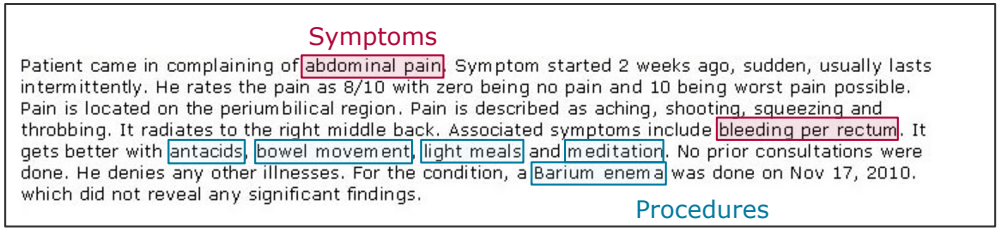
Rule-based NER

- Many systems used in practice make use of (large) sets of handwritten rules
- Advantage: Domain knowledge can be easily encoded
- Rules can be enhanced by dictionaries and linguistic information, e.g., only match noun phrases as named entities
- Elaborate rule-based systems can perform well, but are rather inflexible and costly to create and maintain



General vs. Specific Rules

- General rules → high recall
- Specific rules → high precision



- Some possible rules (pseudo code)

  matches: bleeding per (\w)+ AND POS tag: NOUN → symptom

  prefix: [Cc]omplain[\w]* (of)? AND matches: (\w)+ → symptom

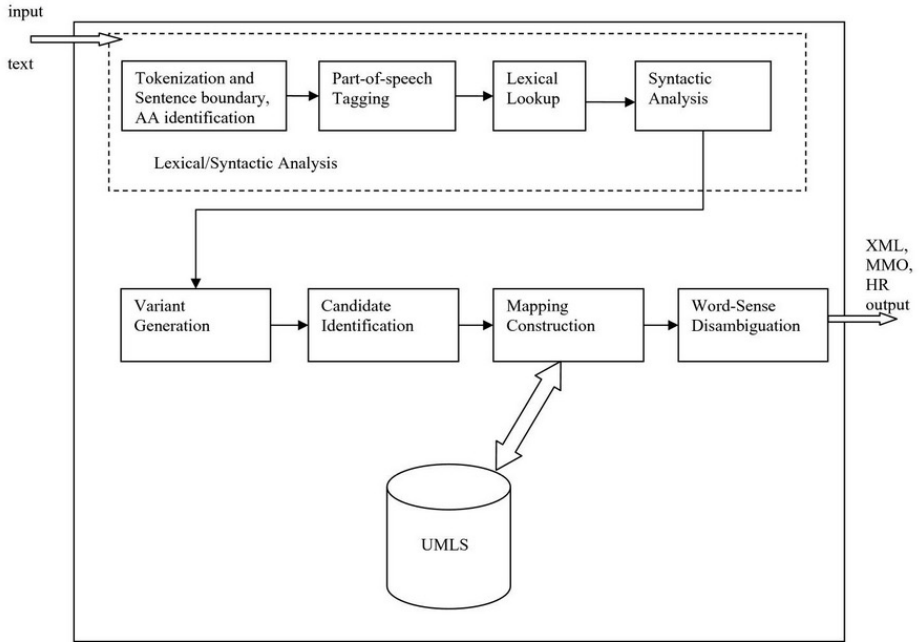
  chunk: NOUN phrase AND suffix: was done → procedure

  matches: (\w*acids) → procedure

  matches: (\w+ation) → procedure

Tools for (Bio-medical) NER: MetaMap

- Widely used tool to identify biomedical concepts in text
- Originally developed by NIH to provide link between UMLS and biomedical literature



https://metamap.nlm.nih.gov/

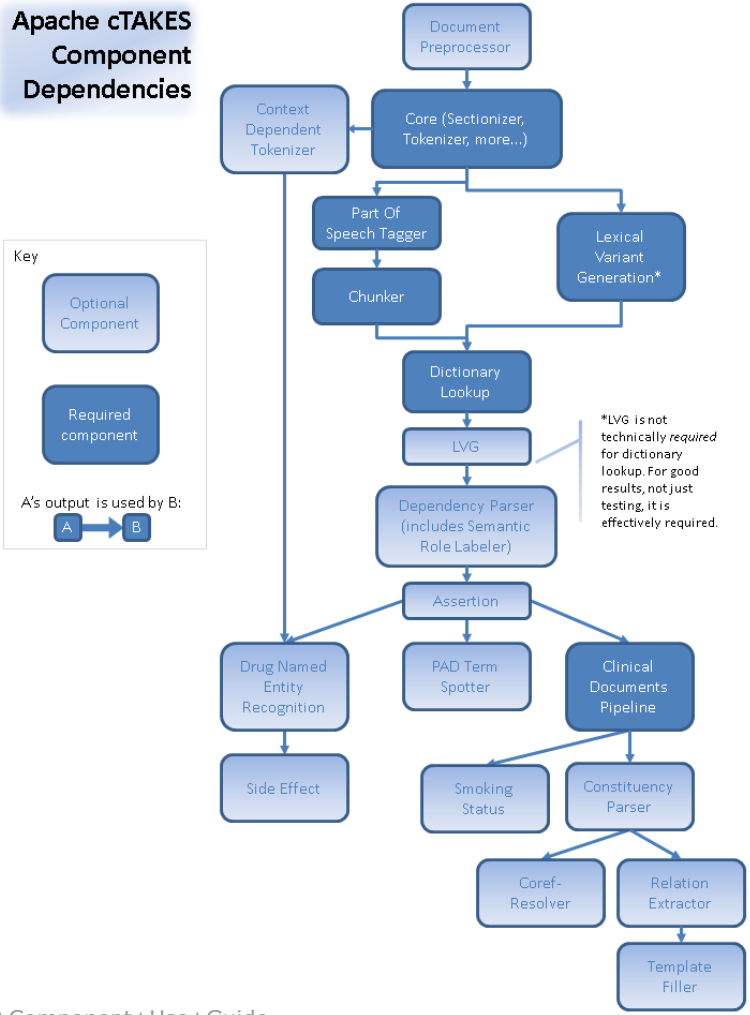
Tools for (Bio-medical) NER: MetaMap

```
Processing 00000000.tx.1: The patient has no signs of pneumonia.

Phrase: The patient
Meta Mapping (1000):
  1000   *^patient (Patients) [Patient or Disabled Group]
Meta Mapping (1000):
  1000   Patient (Abortion consent:Finding:Point in time:^Patient:Document) [Clinical Attribute]
Meta Mapping (1000):
  1000   Patient (Hysterectomy consent:Finding:Point in time:^Patient:Document) [Clinical Attribute]
Meta Mapping (1000):
  1000   Patient (Sterilization consent:Finding:Point in time:^Patient:Document) [Clinical Attribute]

Phrase: has

Phrase: no signs of pneumonia.
Meta Mapping (708):
   770   signs (Aspects of signs) [Functional Concept]
   604 N PNEUMONIA (Pneumonia) [Disease or Syndrome]
Meta Mapping (708):
   770   signs (Manufactured sign) [Manufactured Object]
   604 N PNEUMONIA (Pneumonia) [Disease or Syndrome]
Meta Mapping (708):
   770   SIGNS (Physical findings) [Finding]
   604 N PNEUMONIA (Pneumonia) [Disease or Syndrome]
```

Tools for (Bio-medical) NER: Apache cTakes

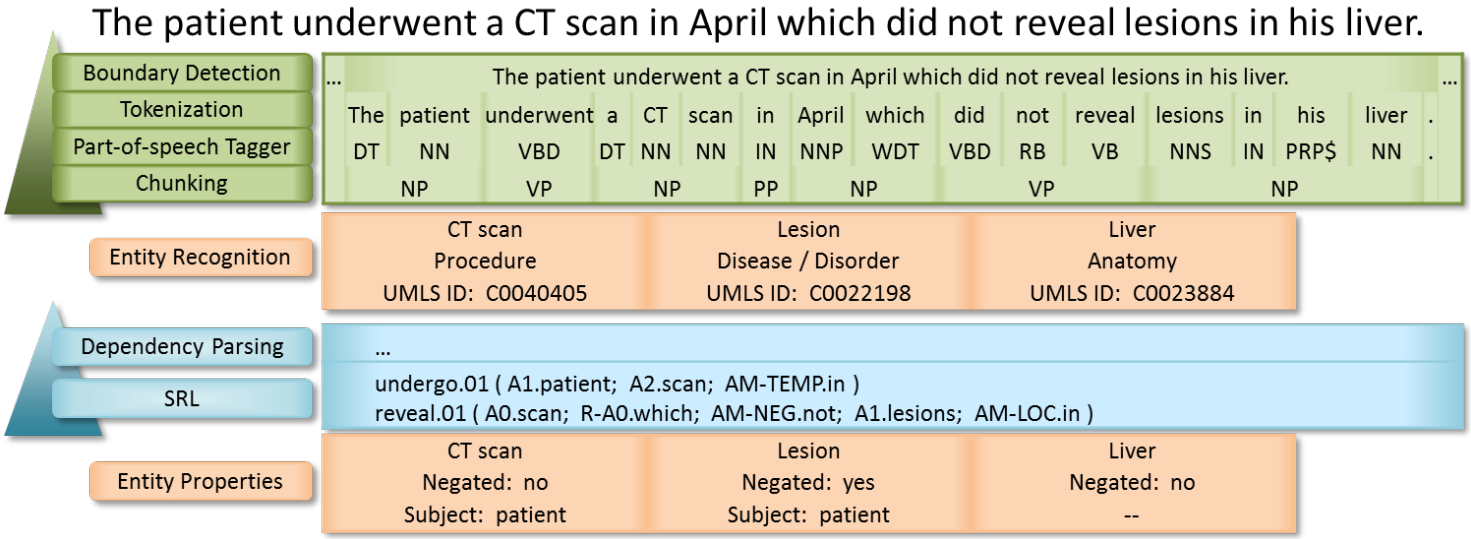
- clinical Text Analysis and Knowledge Extraction System
- Focus on information extraction from clinical text in Electronic Health Records
- Configurable pipeline concept based on UIMA (Unstructured Information Management Architecture)

Apache cTAKES Component Dependencies



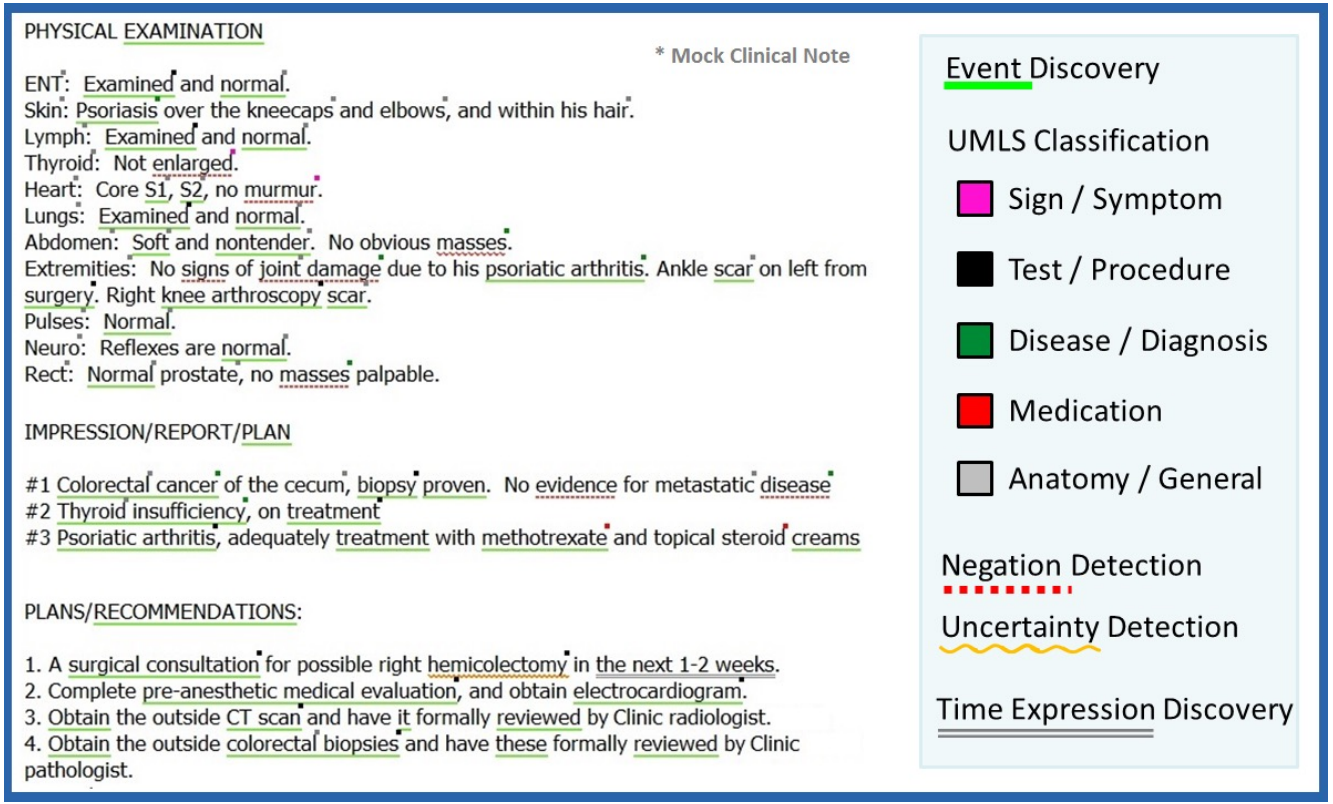
https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+4.0+Component+Use+Guide

cTakes Clinical Information Extraction Pipeline



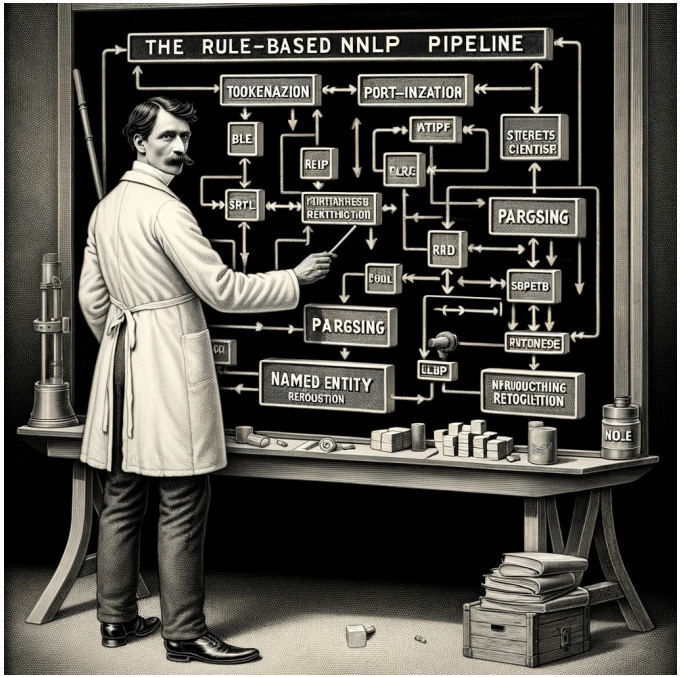
https://cwiki.apache.org/confluence/display/CTAKES/Default+Clinical+Pipeline

Tools for (Bio-medical) NER: Apache cTakes



https://ctakes.apache.org/

What to Take Home?



- Volume and types of biomedical text
- How to handle strings
- Ontologies & controlled vocabularies
- Challenges in processing medical text for humans and computers
- Aspects of human language & NLP tasks
- Dictionary- and rule-based information extraction

New Jupyter Notebook(s)!

(relevant for Exercise 3)

Generated by Dall-E 3