Hasso Plattner Institut
Digital Engineering · Universität Potsdam

ML-Based NLP and Medical Text Corpora

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

Data

ML

Refine

Evaluate

Prediction + Probability

ML and Corpora

Data Management for Digital Health, Winter 2023

Agenda
Pillars of the Lecture

Hasso Plattner Institut

Medical Use Cases

Biology Recap

Oncology

Nephrology

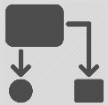
Infectious Diseases

Technology Foundation

Data Sources

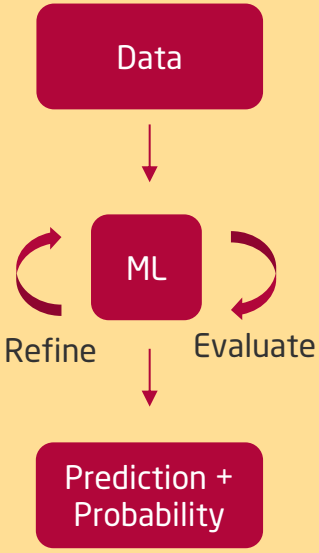
Data Formats

Processing and Analysis

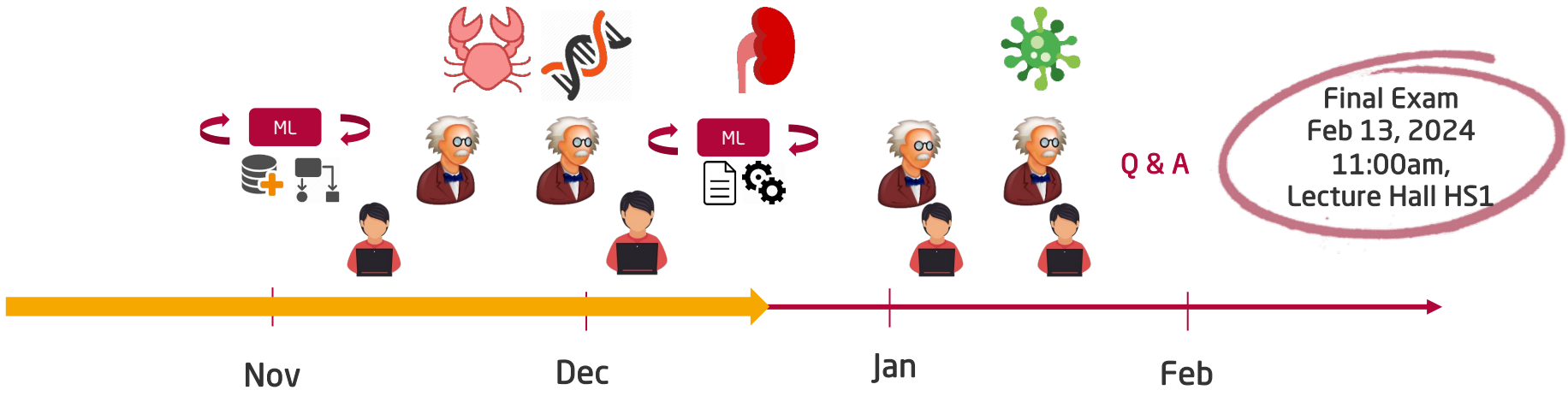


Software Architectures

Machine Learning



Lecture Schedule



- Lecture Kickoff
- Actors in Healthcare
- Digital Health Data

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

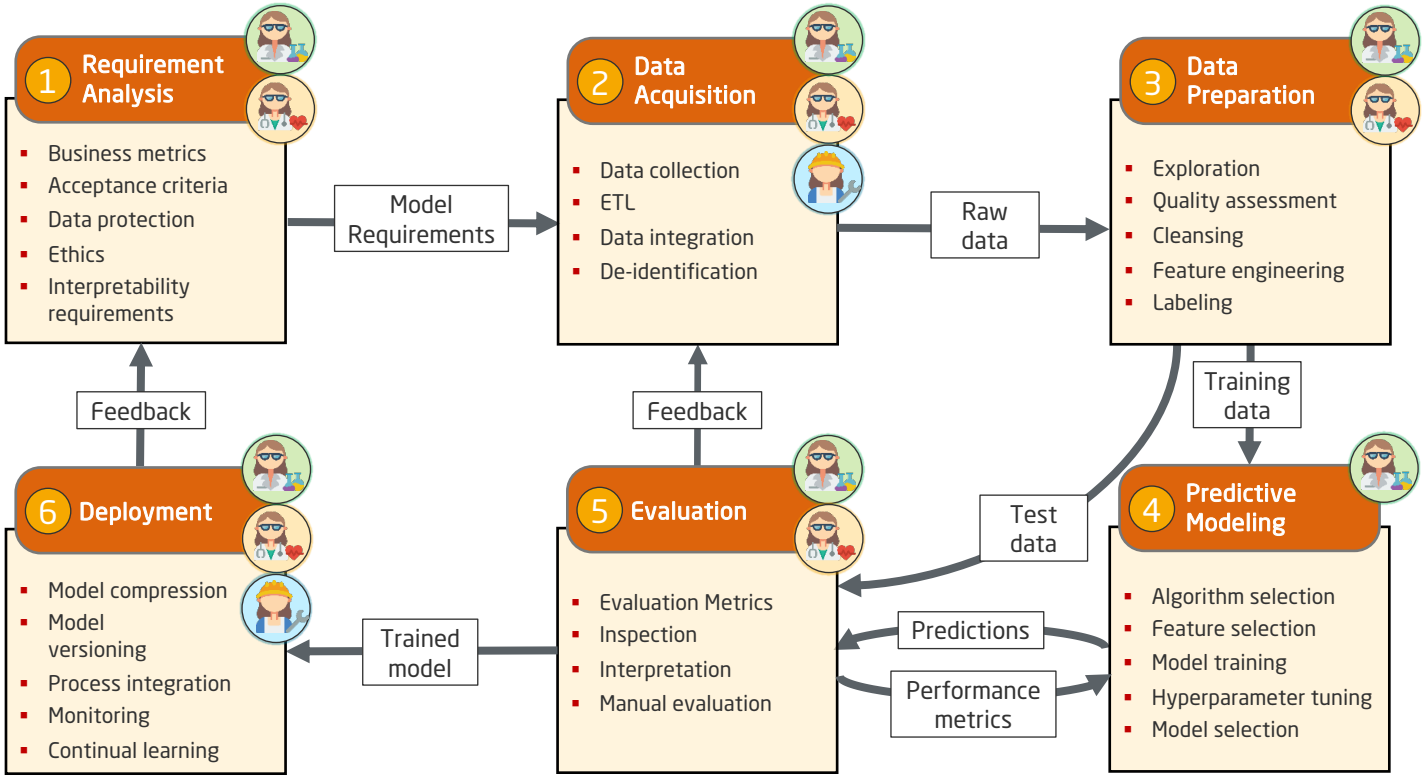
- Natural Language Processing
- Use Case Nephrology & Intensive Care
- Supervised ML & Deep Learning

- Use Case Infectious Diseases
- Unsupervised ML

Agenda

- Acquisition and Preparation of Text Corpora
- Feature Extraction from Texts
- Neural Networks
- NER as Sequence Tagging
- Evaluation of NER Models

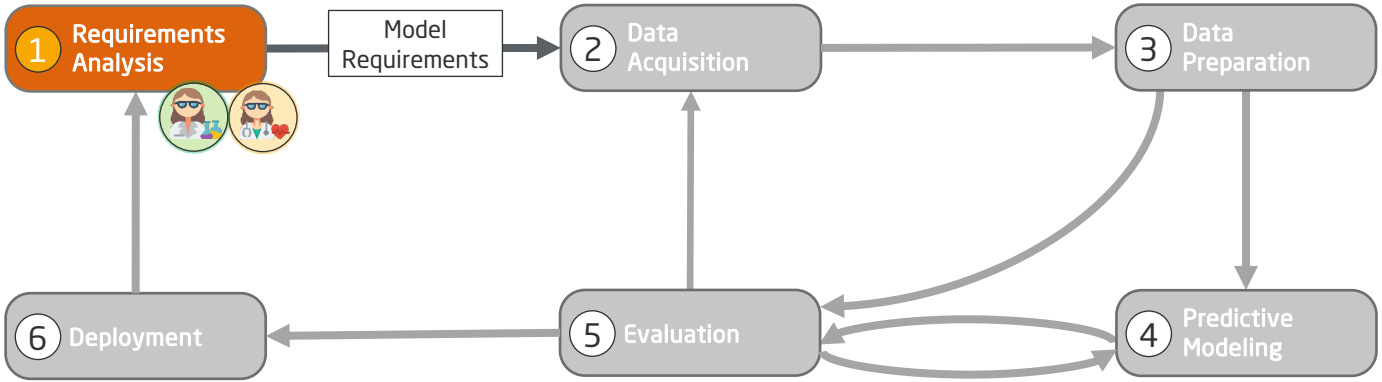
Revisiting the Process Model for ML in Digital Health



ML and Corpora

Data Management for Digital Health, Winter 2023

Requirements Analysis



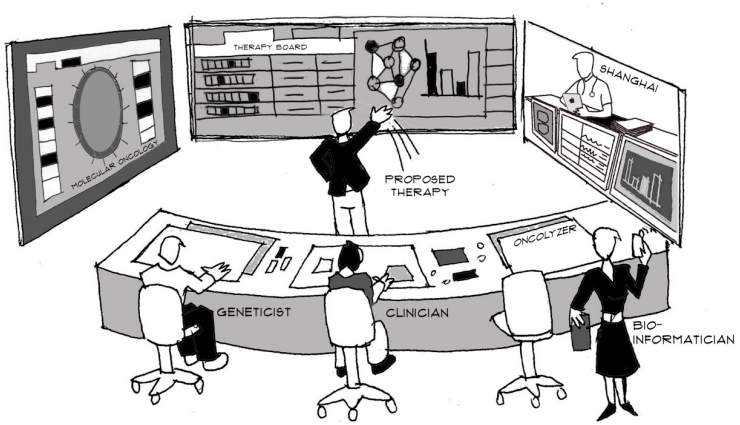
Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

NLP Use Case: Molecular Tumor Board

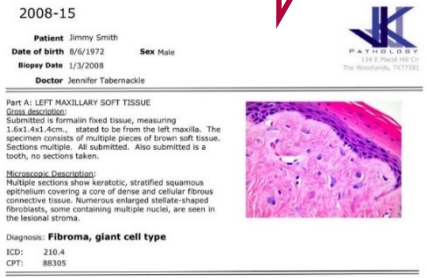
Patient data

- 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

Some Problems of Rule- and Dictionary-Based NER

- Ambiguity: e.g., gene names can be extremely weird (*a, white, swiss cheese, upregulated during skeletal muscle growth 5*)
- Emerging terminology: e.g., new drugs are approved all the time, have different trade names and only numeric codes when still experimental
- Complex entities: e.g., population and outcome statements come with a variety of modifiers and attributes that need to be detected

Disorders

Trial population and design

Genes / Proteins / Variants

Methods: In this open-label, phase 3 trial, we enrolled 665 patients with *BRAF* V600E-mutated metastatic colorectal cancer who had had disease progression after one or two previous regimens. Patients were randomly assigned in a 1:1:1 ratio to receive encorafenib, binimetinib, and cetuximab (triplet-therapy group); encorafenib and cetuximab (doublet-therapy group); or the investigators' choice of either cetuximab and irinotecan or cetuximab and FOLFIRI (folinic acid, fluorouracil, and irinotecan) (control group). The primary end points were overall survival and objective response rate in the triplet-therapy group as compared with the control group. A secondary end point was overall survival in the doublet-therapy group as compared with the control group. We report here the results of a prespecified interim analysis.

Drugs

Outcomes

Results: The median overall survival was 9.0 months in the triplet-therapy group and 5.4 months in the control group (hazard ratio for death, 0.52; 95% confidence interval [CI], 0.39 to 0.70; P<0.001). The confirmed response rate was 26% (95% CI, 18 to 35) in the triplet-therapy group and 2% (95% CI, 0 to 7) in the control group (P<0.001). The median overall survival in the doublet-therapy group was 8.4 months (hazard ratio for death vs. control, 0.60; 95% CI, 0.45 to 0.79; P<0.001). Adverse events of grade 3 or higher occurred in 58% of patients in the triplet-therapy group, in 50% in the doublet-therapy group, and in 61% in the control group.

Rule- or ML-based Information Extraction?

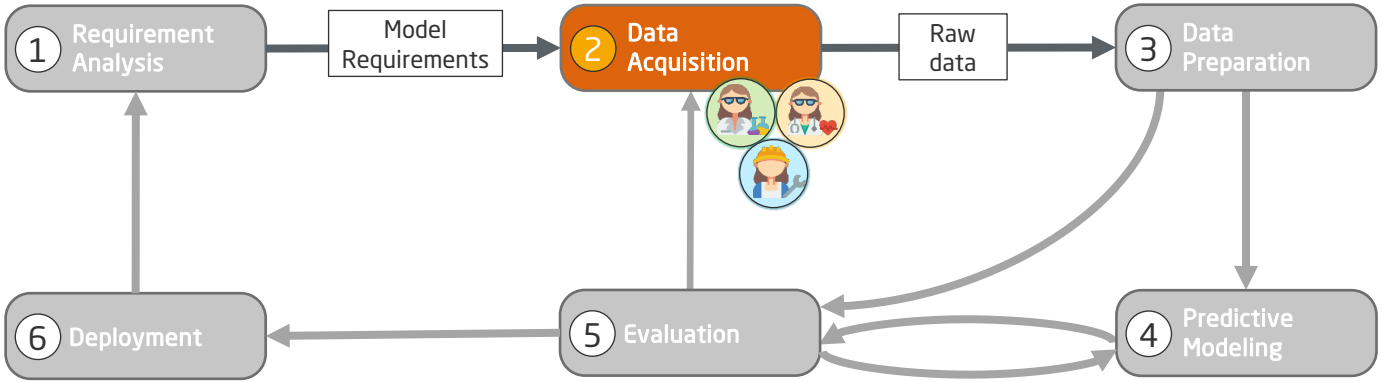
	Pros	Cons
Rule-based	• Declarative • Easy to comprehend • Easy to maintain • Easy to incorporate domain knowledge • Easy to trace and fix the cause of errors	• Heuristic • Requires tedious manual labor
ML-based	• Trainable • Adaptable • Reduces manual effort	• Requires labeled data • Requires retraining for domain adaptation • Requires ML expertise to use or maintain • Opaque

CAVE:

- from 2013 (ML-based back then != today)
- by IBM researchers (IBM sells / sold rule-based IE software)

Chiticariu, Laura et. Al (2013): Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!

Data Acquisition



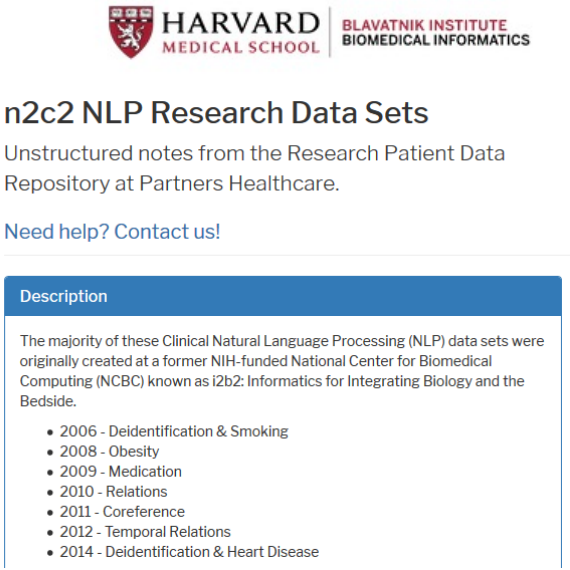
Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

Icons made by Smashicons from www.flaticon.com

Clinical Text Corpora

- Real-world gold standard text corpora necessary for development and evaluation of NLP methods
  - Available for general domains (Wikipedia, Twitter)
  - Many (biomedical) datasets based on MEDLINE abstracts
  - Few datasets (MIMIC, i2b2 / n2c2) of clinical text available for English language
- Issues:
  - Corpora for low-resource languages hardly available
  - Data protection for clinical text (every researcher is creating their own in-house corpus)
  - Licensing issues with full-text scientific articles
  - Tedious annotation process



https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Data Availability Problem for Non-English Language Communities

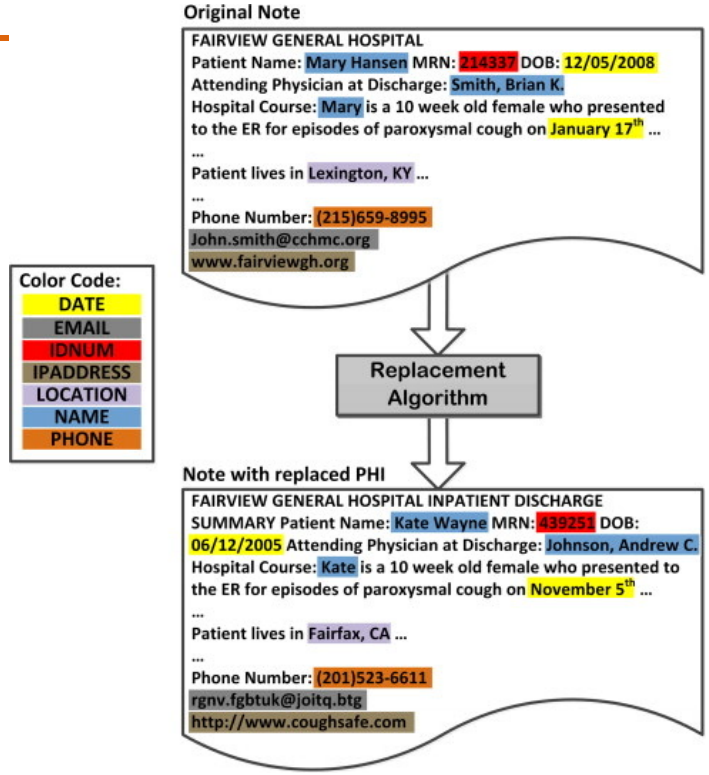
Hasso Plattner Institut

	Authentic Clinical Texts	De-identified / Synthetic Clinical Texts	Non-individual Medical Texts
English	• Discharge Summaries • Radiology Reports • Pathology Reports	i2b2 workbench, CLEF eHealth, MIMIC	PubMed
German	• Discharge Summaries • Radiology Reports • Pathology Reports	• Very few / small (e.g., GraSCCo, BRONCO)	• Guidelines • Textbooks

De-identification of Clinical Text

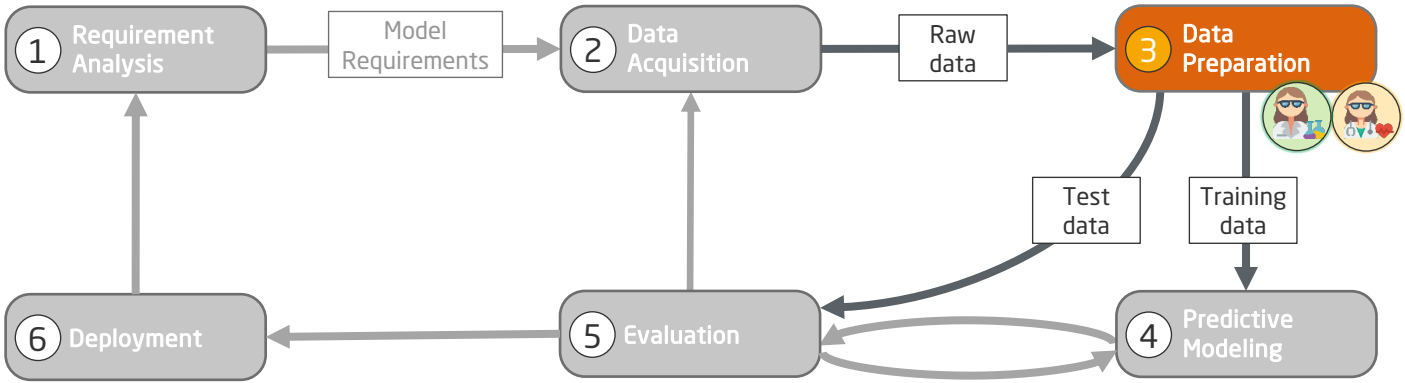
- De-identification of clinical text is tricky
  - US: 18 PHI (protected health information) categories shall be removed
  - EU: not well defined (GDPR)

- Automatic de-identification is challenging NLP task itself
  - e.g. i2b2 Challenge 2014
  - Similar problem to NER
  - Possible with relatively high accuracy (F1 scores > .90)
    → recall below 100% acceptable?



Deleger, Louise, et al. (2014) "Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research."

Data Preparation



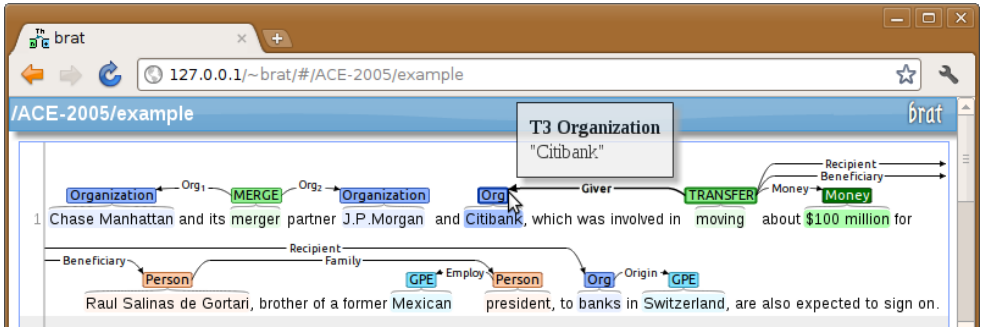
Roles

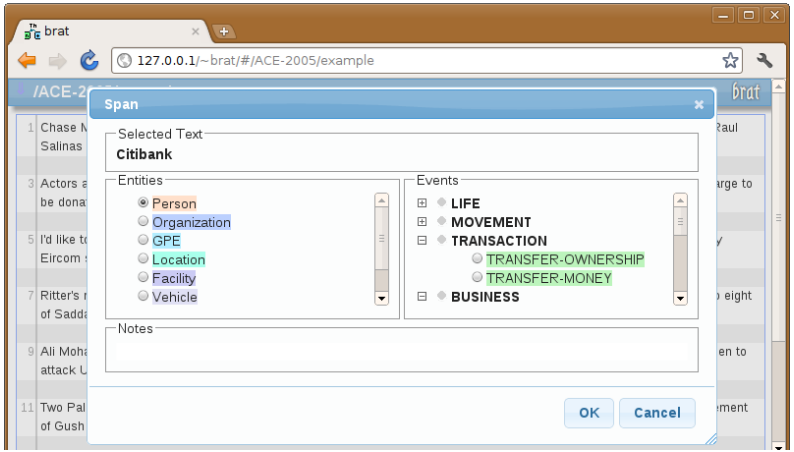
- Data Scientist
- Domain Expert
- (Data) Engineer

Icons made by Smashicons from www.flaticon.com

Annotation of Text Corpora

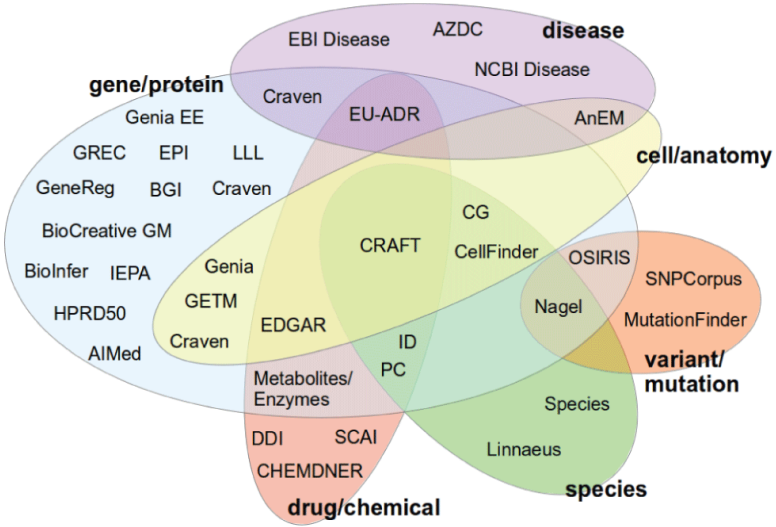
- Requires domain expertise
- Usually multiple annotators
- Need clearly defined annotation guidelines
- Crowdsourcing (Mechanical Turk, Upwork) is an option
  - Many noisy labels can be better for ML than few good labels
  - Gold standard evaluation data still necessary
- Popular visual online annotation tools: BRAT, INCEpTION





https://brat.nlplab.org/examples.html

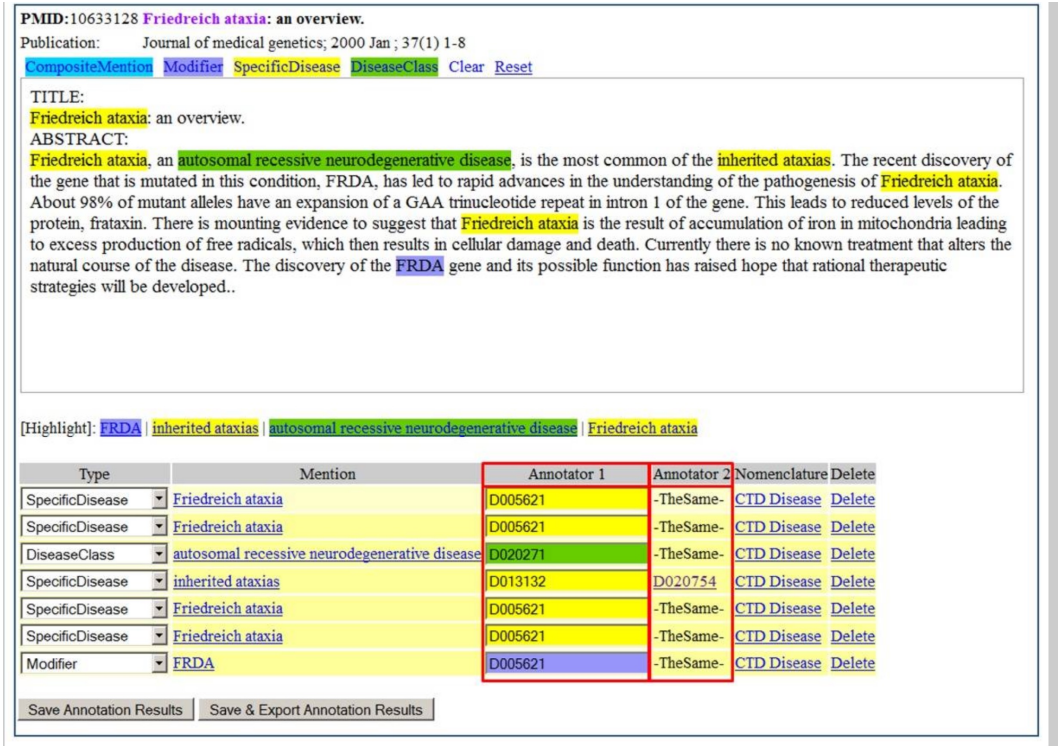
Biomedical Corpora



Neves, M. (2014). An analysis on the entity annotations in biological corpora. *F1000Research*, *3*.

NCBI Disease Corpus (Dogan et al. 2014)

- 793 MEDLINE abstracts
- Annotations:
  - Disease Mentions (6892)
  - MeSH / OMIM Concepts (790)
  - 14 annotators
  - Agreement F1-Score around 0.9
- Entity-level F1-Score:
  - upon publication: 63.7%
  - State-of-the-art: 89.1% (BioBERT)



PMID:10633128 **Friedreich ataxia: an overview.**

Publication: Journal of medical genetics; 2000 Jan ; 37(1) 1-8

CompositeMention Modifier SpecificDisease DiseaseClass Clear Reset

TITLE:
Friedreich ataxia: an overview.
ABSTRACT:
Friedreich ataxia, an autosomal recessive neurodegenerative disease, is the most common of the inherited ataxias. The recent discovery of the gene that is mutated in this condition, FRDA, has led to rapid advances in the understanding of the pathogenesis of Friedreich ataxia. About 98% of mutant alleles have an expansion of a GAA trinucleotide repeat in intron 1 of the gene. This leads to reduced levels of the protein, frataxin. There is mounting evidence to suggest that Friedreich ataxia is the result of accumulation of iron in mitochondria leading to excess production of free radicals, which then results in cellular damage and death. Currently there is no known treatment that alters the natural course of the disease. The discovery of the FRDA gene and its possible function has raised hope that rational therapeutic strategies will be developed..

[Highlight]: FRDA | inherited ataxias | autosomal recessive neurodegenerative disease | Friedreich ataxia

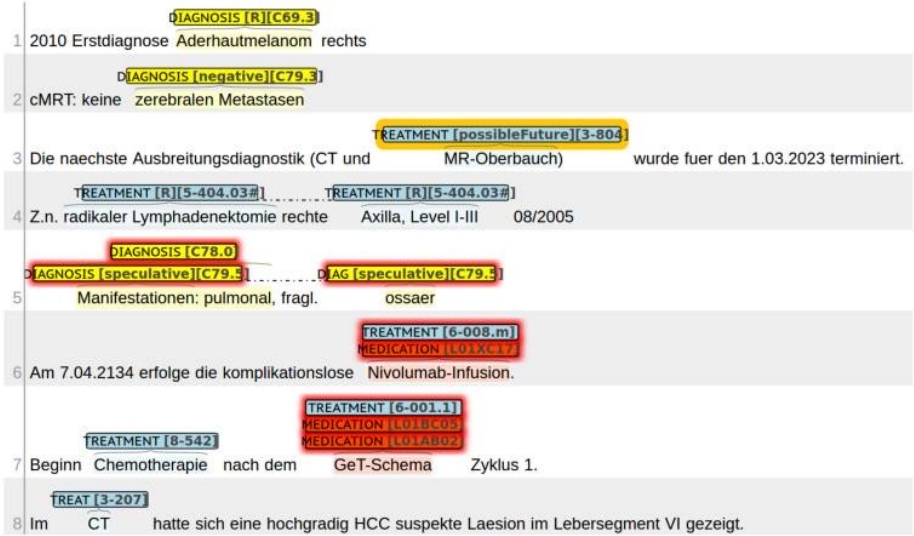
Type	Mention	Annotator 1	Annotator 2	Nomenclature	Delete
SpecificDisease	Friedreich ataxia	D005621	-TheSame-	CTD Disease	Delete
SpecificDisease	Friedreich ataxia	D005621	-TheSame-	CTD Disease	Delete
DiseaseClass	autosomal recessive neurodegenerative disease	D020271	-TheSame-	CTD Disease	Delete
SpecificDisease	inherited ataxias	D013132	D020754	CTD Disease	Delete
SpecificDisease	Friedreich ataxia	D005621	-TheSame-	CTD Disease	Delete
SpecificDisease	Friedreich ataxia	D005621	-TheSame-	CTD Disease	Delete
Modifier	FRDA	D005621	-TheSame-	CTD Disease	Delete

Save Annotation Results | Save & Export Annotation Results

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3951655/

German Clinical Text
BRONCO (Kittner et al. 2021)

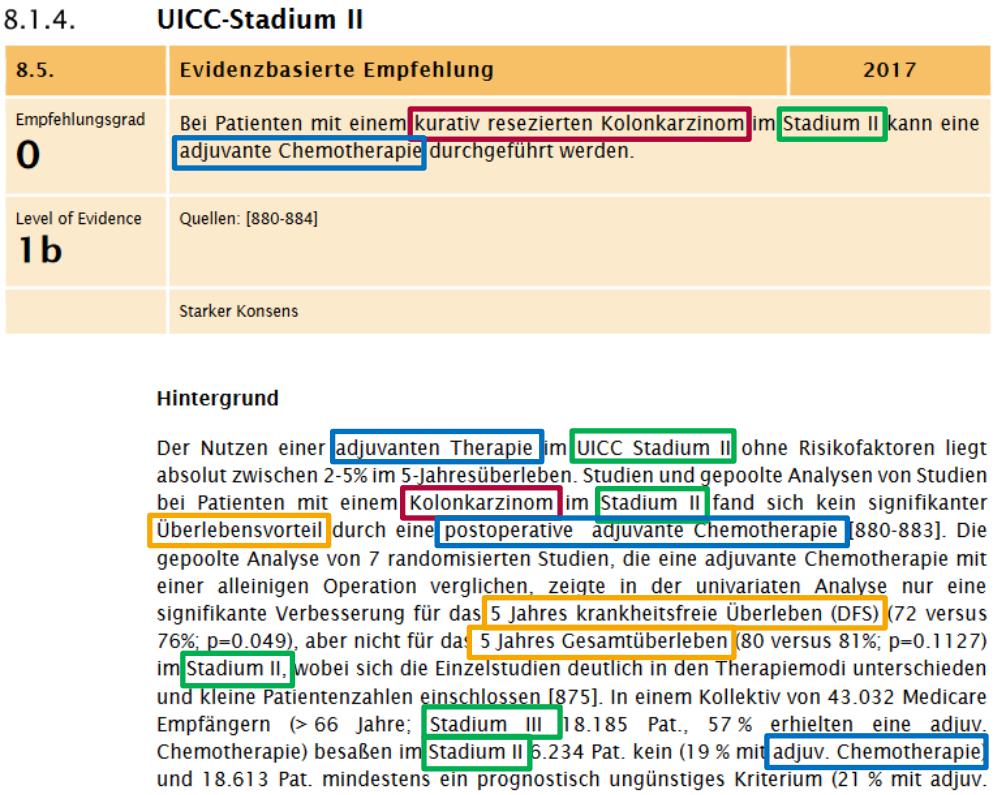
- First publicly available, de-identified German clinical text corpus (200 oncological discharge summaries)
- Interesting de-identification procedure:
  - "First, the corpus was completely manually deidentified. This process was confirmed by Charité and UKT data protection officers."
  - "Second, we only annotate and publish certain sections of the discharge summaries, avoiding all sections containing mostly biographic information"
  - "Third, we shuffled all sentences in the 2 subcorpora to blur their order and relationships."



Kittner et al. (2021), Annotation and initial evaluation of a large annotated German oncological corpus

GGPONC

- 1.87 M tokens from 30 German clinical guidelines in oncology
- Currently the largest, shareable corpus of German medical text (https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/)
- Available through Data Use Agreement (i.e., non-commercial use only)



8.1.4. **UICC-Stadium II**

8.5.	Evidenzbasierte Empfehlung	2017
Empfehlungsgrad **0**	Bei Patienten mit einem kurativ resezierten Kolonkarzinom im Stadium II kann eine adjuvante Chemotherapie durchgeführt werden.	
Level of Evidence **1b**	Quellen: [880-884]	
	Starker Konsens	

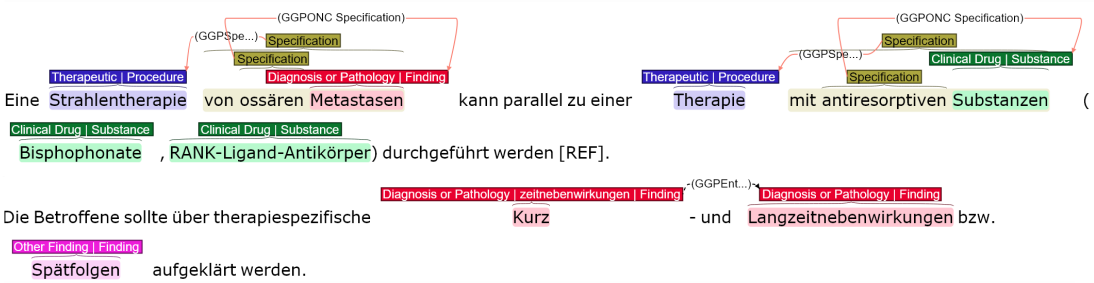
Hintergrund

Der Nutzen einer adjuvanten Therapie im UICC Stadium II ohne Risikofaktoren liegt absolut zwischen 2-5% im 5-Jahresüberleben. Studien und gepoolte Analysen von Studien bei Patienten mit einem Kolonkarzinom im Stadium II fand sich kein signifikanter Überlebensvorteil durch eine postoperative adjuvante Chemotherapie [880-883]. Die gepoolte Analyse von 7 randomisierten Studien, die eine adjuvante Chemotherapie mit einer alleinigen Operation verglichen, zeigte in der univariaten Analyse nur eine signifikante Verbesserung für das 5 Jahres krankheitsfreie Überleben (DFS) (72 versus 76%; p=0.049), aber nicht für das 5 Jahres Gesamtüberleben (80 versus 81%; p=0.1127) im Stadium II, wobei sich die Einzelstudien deutlich in den Therapiemodi unterschieden und kleine Patientenzahlen einschlossen [875]. In einem Kollektiv von 43.032 Medicare Empfängern (> 66 Jahre; Stadium III 18.185 Pat., 57 % erhielten eine adjuv. Chemotherapie) besaßen im Stadium II 6.234 Pat. kein (19 % mit adjuv. Chemotherapie und 18.613 Pat. mindestens ein prognostisch ungünstiges Kriterium (21 % mit adjuv.

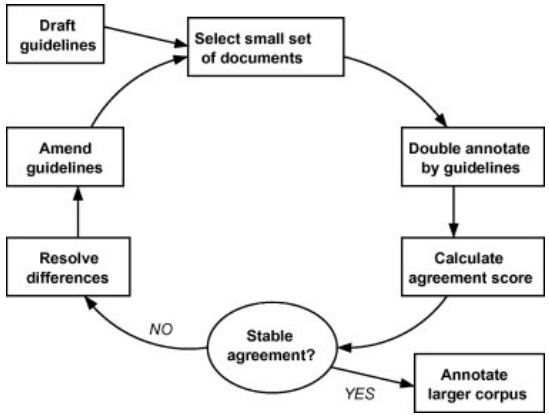
Borchert et al. (2022): GGPONC 2.0 — The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers

GGPONC
Annotation & Distribution

- Annotation team:
  - 7 annotators (medical students)
  - 1 curator (medical doctor)
- 6 months / 1200 hours
- > 200k entities



- Iterative guideline refinement until stable agreement was reached (ɣ = .94)



ML and Corpora

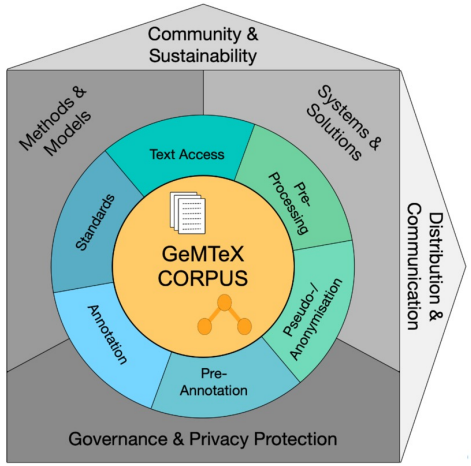
Data Management for Digital Health, Winter 2023

Roberts et al. (2009), Building a semantically annotated corpus of clinical texts

https://inception-project.github.io/use-cases/ggponc/

GeMTeX
German Medical Text Corpus

- Goal: Create a large annotated text corpus of German medical texts from routine patient care
- Building upon German Medical Informatics Initiative (MII) infrastructure:
  - Broad consent
  - Data integration centers (DIC) at six university hospitals
- Automated + manual de-identification
- Deep semantic annotation:
  - Entities and grounding (SNOMED CT)
  - Relations (e.g., temporal, causal)
  - Domain-specific annotations (oncology, cardiology, neurology, pharmacology)



ML and Corpora

Data Management for Digital Health, Winter 2023

Feature Engineering for ML-based Named-Entity Recognition

- Training data := Sequences of tokens
- Labels := Sequences of I/O/B tags
- Simple approach to ML-based NER
  - Turn token into feature vector v
  - Classify each token as I/O/B
    - $f(v(\text{“BRAF”})) = B$
    - $f(v(\text{“V600E”})) = I$
    - $f(v(\text{“mutation”})) = O$

$\mathbf{x} = (x_1, ..., x_n)$, e.g. $x_1 = \text{“BRAF”}$, $x_6 = \text{“mutation”}$

$\mathbf{y} = (y_1, ..., y_n)$, e.g. $y_1 = \text{“B”}$, $y_6 = \text{“O”}$

BRAF V600E is a driver mutation found in multiple tumor types

B I O O O O O O O O O

- What do you think could be useful features for NER on the token level?

Features: Word Identity

- Simple feature: Identity of current token
- One-hot-encoding : create a feature vector with length equal to size of vocabulary and set 1 if word == ith element else 0
- Vocabulary size is typically restricted to most common *k* tokens
- Special feature for unknown tokens (unknown tokens are often named entities!)
- Resulting feature vectors are very sparse (almost all elements are 0)

$$v(\text{“apple”}) \rightarrow \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad v(\text{“cancer”}) \rightarrow \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}$$

Word Shape Features

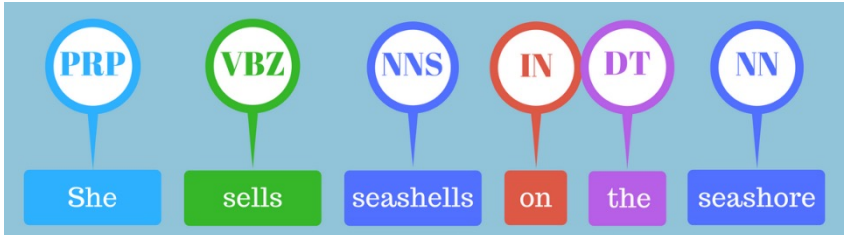
- Word shape features := abstract letter patterns
- Prefixes
- Suffixes (e.g., word ends with *"-itis"* → disease)
- character n-grams

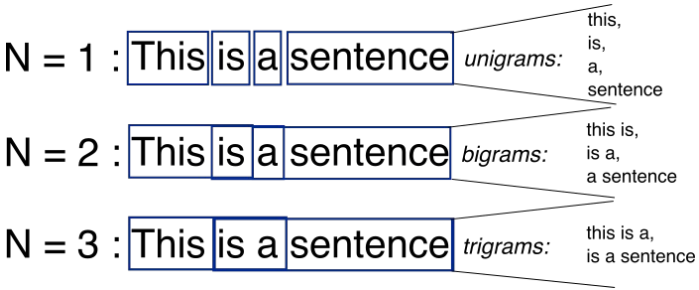
Feature	*About feature*	*Regular Expression*
Initcaps	First letter is in uppercase	[A-Z]\w+
Initcapsalpha	First letter is in uppercase. Second letter is in lowercase.	[A-Z][a-z]\w+
All caps	All letters are in uppercase	[A-Z]+
Caps mix	Mixture of uppercase and lowercase letters	[A-Za-z]+
Has digit	Protein name has a number in the middle	\w+[0-9]\w+
Single digit	Ranges from 0-9	[0-9]
Double digit	Two digit numbers	[0-9][0-9]
Natural numbers	Any natural number	[0-9]+
Real numbers	Decimal numbers /numbers with comma	[-0-9]+[.,]+[0-9.,]+
Has dash	Protein name with dash in middle	\w+ - \w+
Init dash	First character is a dash	- \w+
End dash	Last character is a dash	\w+ -
Alphanumeric (starts with alphabet)	Combination of alphabets and numbers. First character is an alphabet.	\w+ [A-Za-z] \w+ [0-9] \w+
Alphanumeric (starts with number)	Combination of alphabets and numbers. First character is a number.	\w+ [0-9] \w+ [A-Za-z] \w+
Roman letter	Any roman letter	[IVXDLCM]+
Has Roman	Any roman letter in the middle	\w+ [IVXDLCM]+ \w+
Greek	Any Greek letter	\w+ [aßYÖE]
Has Greek	Any Greek letter in middle	\w+ [aßYÖE] \w+
Punctuation	Protein name with punctuation	

Raja, Kalpana, Suresh Subramani, and Jeyakumar Natarajan. "A hybrid named entity tagger for tagging human proteins/genes." *International journal of data mining and bioinformatics* 10.3 (2014): 315-328.

More Token Features

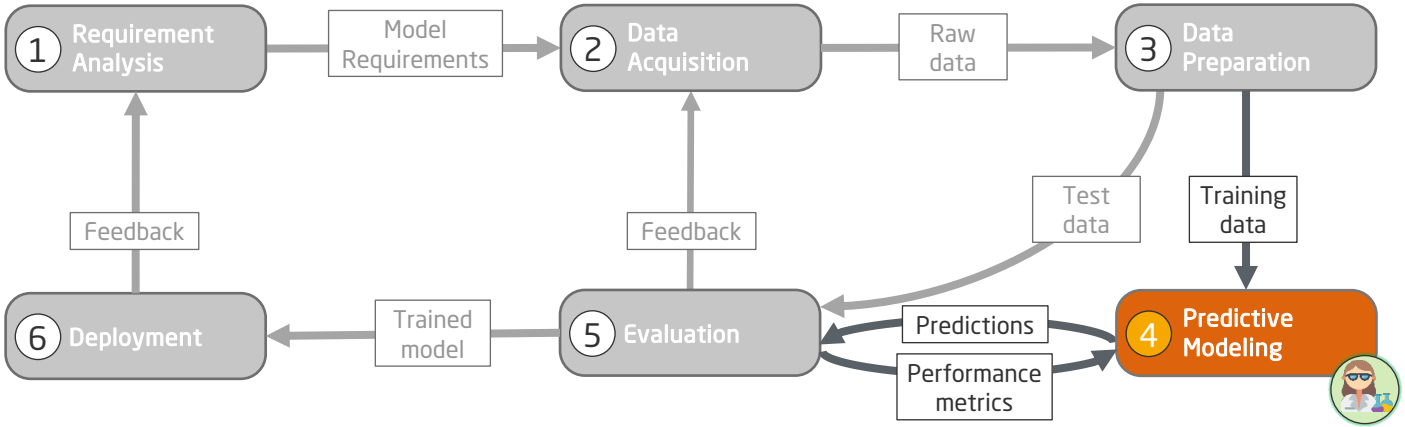
- Dictionary match
- Part-of-speech tags (e.g., is the word a proper noun)
- Context features
  - Distributional semantics: "a word is characterized by the company it keeps" (J.R. Firth, 1957)
    ➔ Surrounding tokens can be more important than token itself
  - word n grams (common: trigrams with current, left and right word)





https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31
https://stackoverflow.com/questions/18193253/what-exactly-is-an-n-gram

Predictive Modeling



Roles

Data Scientist

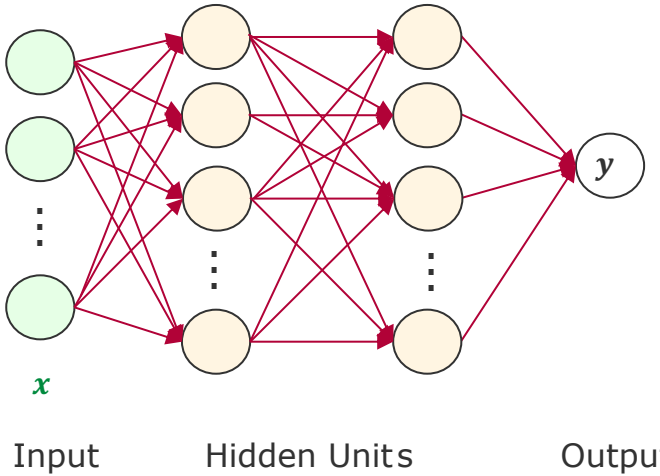
Domain Expert

(Data) Engineer

Icons made by Smashicons from www.flaticon.com

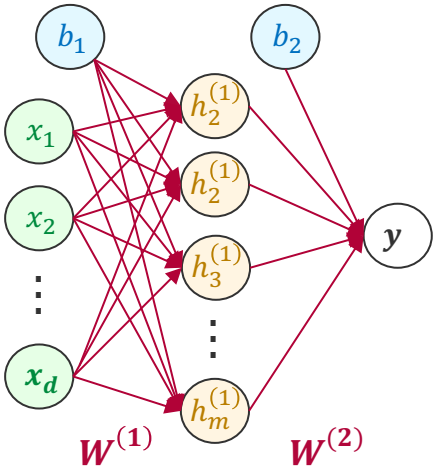
Neural Networks

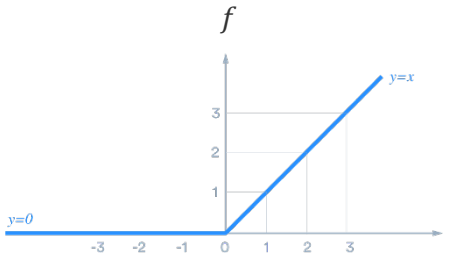
- Class of machine learning models that are loosely inspired by neurons in the brain
- Connections between units have learnable weights
- Hidden units can learn higher-level representations of the input



Feed Forward Neural Networks

- Simplest version of a NN (also called Multilayer Perceptron)
- Representations are learned in *hidden layers (h)*
  - $\boldsymbol{h^{(1)}} := f\ (\boldsymbol{W^{(1)\,T}x} + \boldsymbol{b_1})$
  - $\boldsymbol{y} := \sigma\ (\boldsymbol{W^{(2)\,T}h^{(1)}} + \boldsymbol{b_2})$ (binary classification)
  - $\boldsymbol{y} :=$ softmax $(\boldsymbol{W^{(2)T}h^{(1)}} + \boldsymbol{b_2})$ (multiclass classification)
- Note: Logistic Regression is equivalent to neural network without hidden layers!
- 2-Layer MLP is more expressive than Logistic Regression because of non-linear activation function f
- Common choice for f: Sigmoid or ReLU
- Deep Learning = Neural Networks with many layers

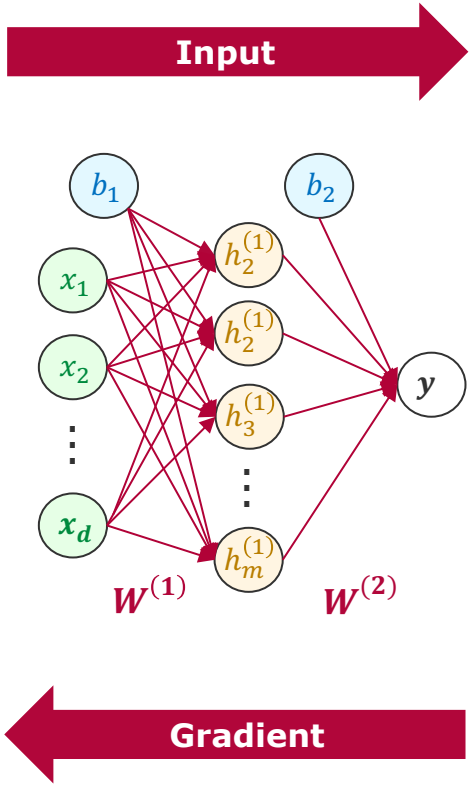




ReLU activation function

https://scikit-learn.org/stable/_images/multilayerperceptron_network.png
https://medium.com/@sonish.sivarajkumar/relu-most-popular-activation-function-for-deep-neural-networks-10160af37dda

Backpropagation



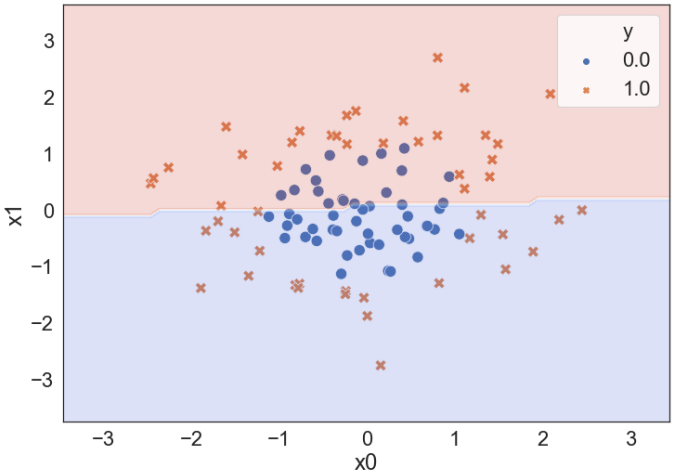
Training := Iteratively adapting weights such that error is decreased

Error

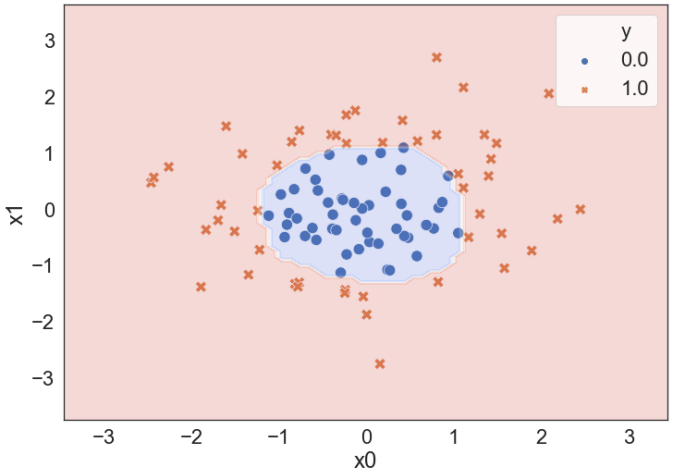
y_{true}

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

Neural Networks Can Learn Non-Linear Functions



Decision boundary learned by Logistic Regression

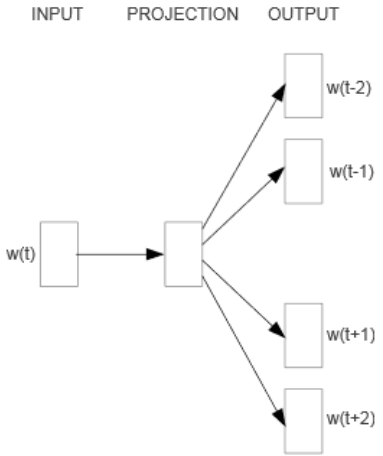


Decision boundary learned by Neural Network with 1 hidden layer

Word Embeddings

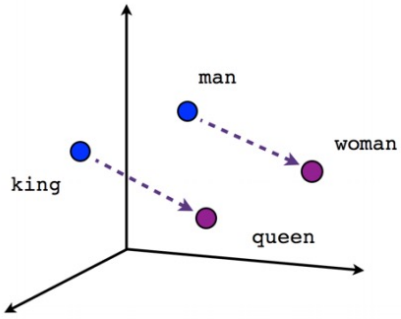
- One-hot-vectors have no notion of semantic similarity
- Mikolov et. al (2013) popularized Word Embeddings
- Idea:
  - Train a feed-forward NN that predicts *surrounding words* (distributional semantics)
  - Use hidden layer as dense representation (projection) of words in a feature space that perform well at this task
- self-supervised → no labels required
- Embeddings are used as feature vectors for downstream tasks, like NER (= transfer learning)

$$v(\text{“cancer”}) \rightarrow \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{pmatrix} \quad v(\text{“carcinoma”}) \rightarrow \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \\ 1 \\ \dots \\ 0 \end{pmatrix} \quad v(\text{“apple”}) \rightarrow \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

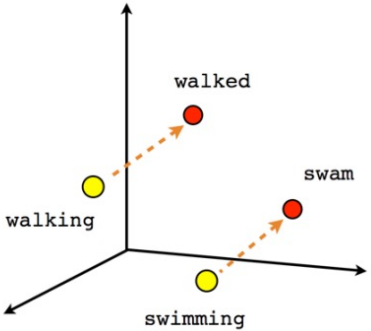


Mikolov et. al (2013): Efficient Estimation of Word Representations in Vector Space

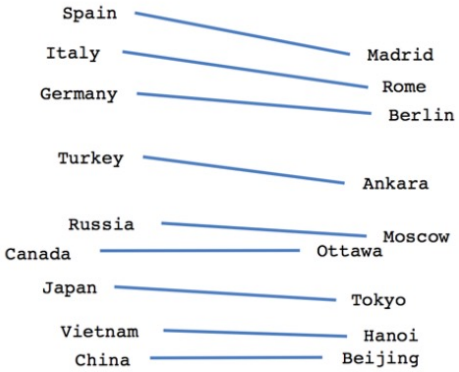
(Shallow) Semantics in Word Embedding Space



Male-Female



Verb tense

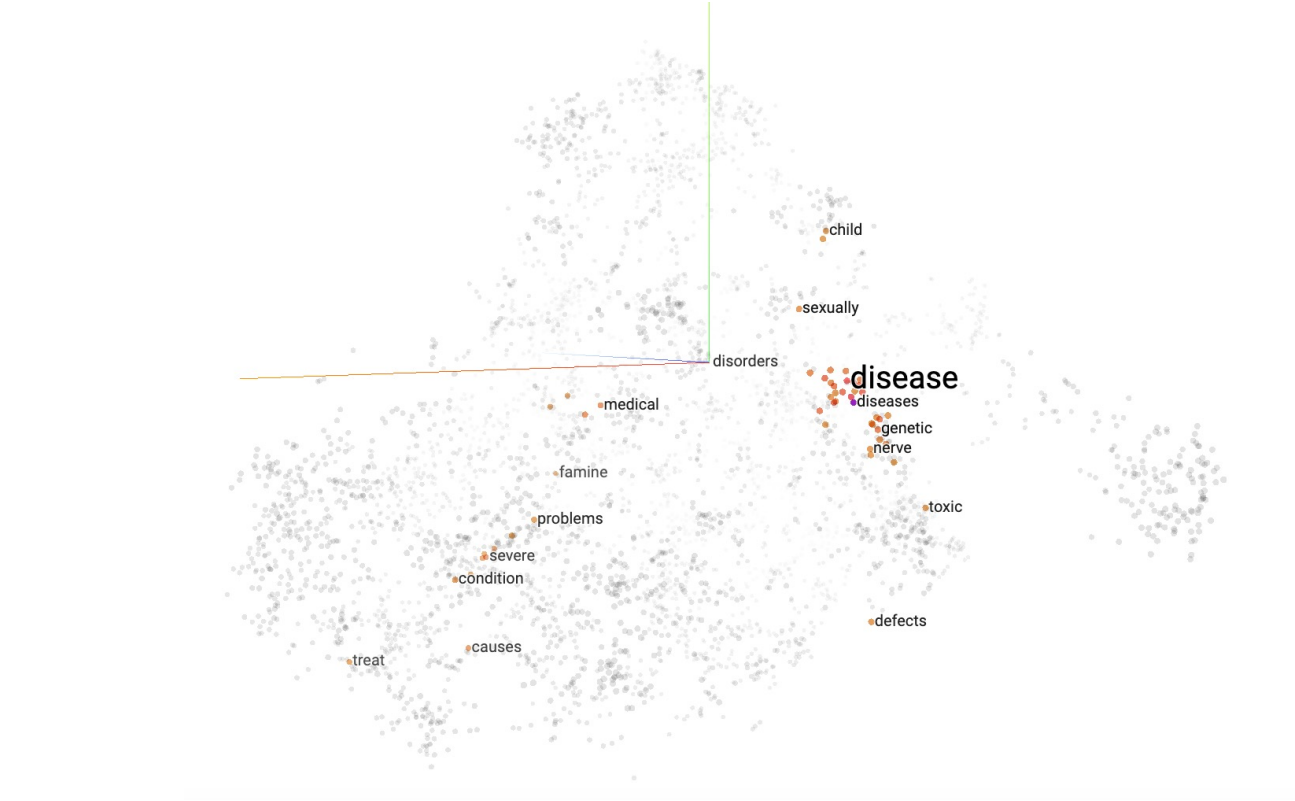


Country-Capital

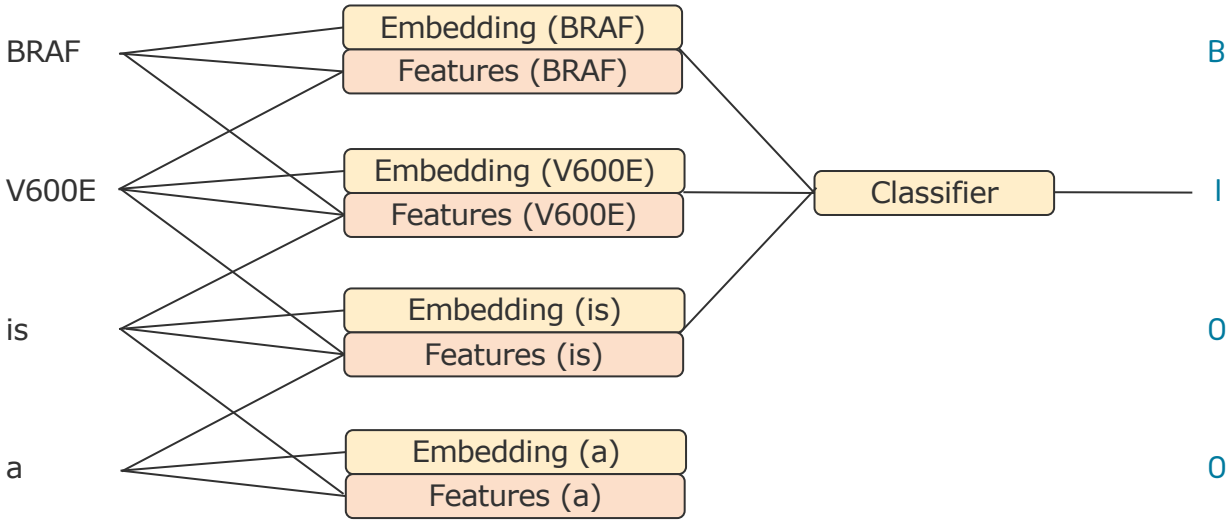
https://www.tensorflow.org/images/linear-relationships.png

Word Embedding Demo

https://projector.tensorflow.org/

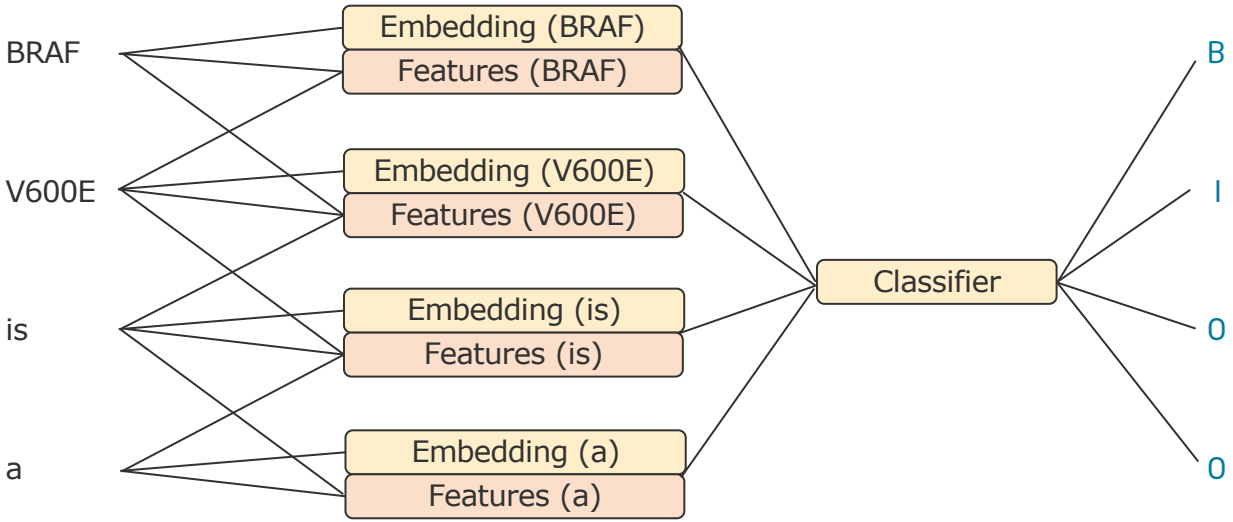


Word Embedding as Features



Structured Prediction for NLP

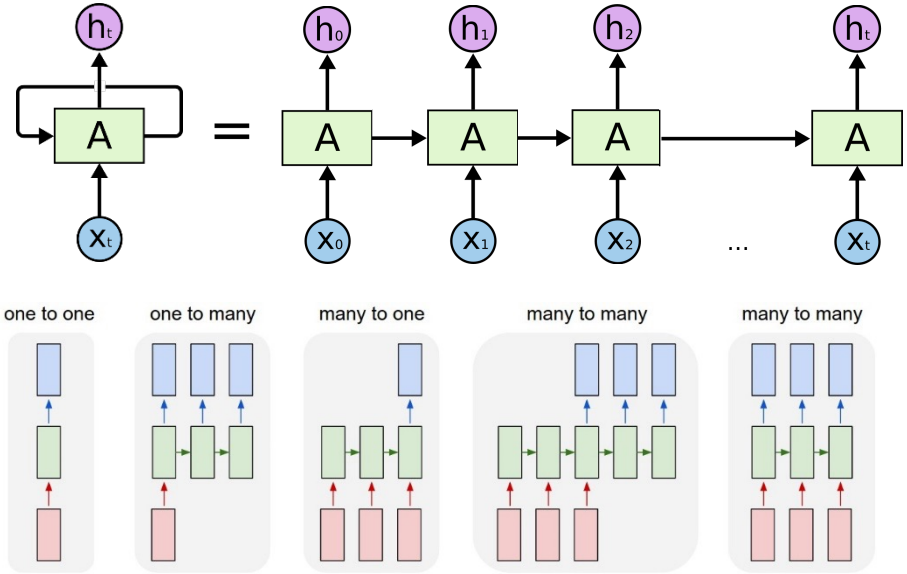
- Classifying single tokens assumes statistical independence of adjacent labels ➔ usually wrong
- Idea: predict whole sequence of labels **y** from sequence of inputs **x** (Structured Prediction)



- … in practice we need to impose some restrictions to keep the problem tractable

Recurrent Neural Networks

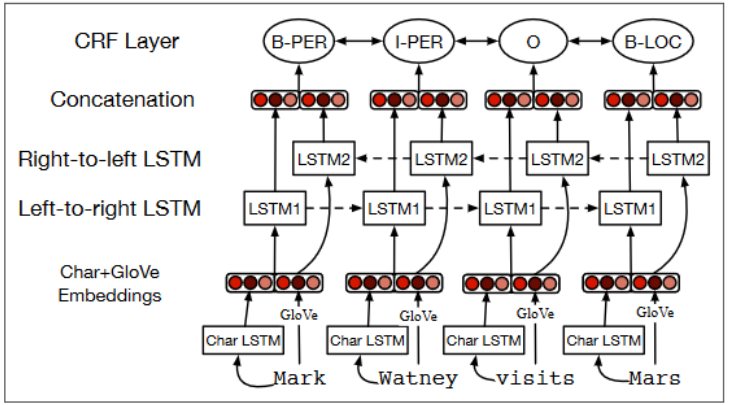
- Special type of neural network with recurrent connections → variable number of computational steps
- Various types of RNN cells exist (denoted by A in the picture)
- Enable end-to-end learning := directly map inputs (sentences) to outputs without explicitly solving intermediate steps (feature engineering, linguistic analysis)
- Computationally much more demanding than linear models



https://colah.github.io/posts/2015-08-Understanding-LSTMs/
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

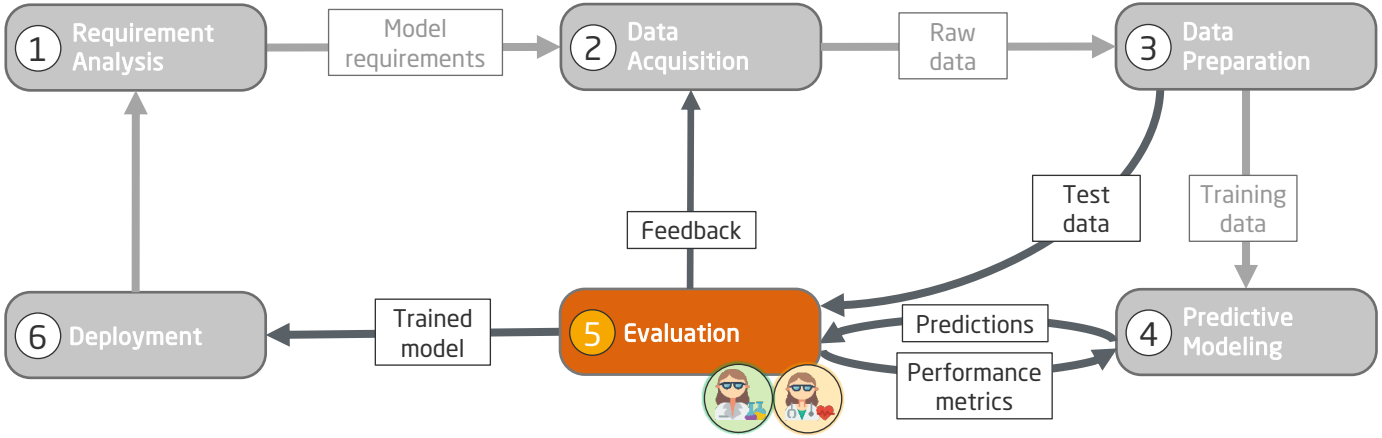
Deep Learning for Sequence-to-Sequence Problems

- State-of-the art approach for NER used to be bi-LSTM-CRF:
  - Word embeddings pretrained on large, unlabelled corpus (such as PubMed abstracts)
  - Bidirectional Long-Short-Term-Memory (LSTM) neural network (special form of recurrent neural network) for feature learning
  - CRF output layer to predict sequence of labels from LSTM representations
- Recent superseded by Transformer-based architectures, yet, still widely in use
- Sequence-to-sequence architectures also applicable to other problems, such as translation or summarization



Martin, James H., and Daniel Jurafsky. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 3rd ed. Draft https://web.stanford.edu/~jurafsky/slp3/*

Evaluation



Roles

Data Scientist

Domain Expert

(Data) Engineer

Icons made by Smashicons from www.flaticon.com

Evaluation of NER Results

- To evaluate performance of NER and other NLP systems, performance is compared against gold standard
- Typical evaluation measures precision, recall, F1 score
  → need to be adapted to account for partial matches

B I B B ← Ground Truth
The patient underwent a CT scan in April which did not reveal lesions in his liver.
B B B ← NER Result

- Strict evaluation (entity level): TP = 1, FP = 1, FN = 3, TN = 12
- Loose evaluation (token level): TP = 2, FP = 1, FN = 2, TN = 12
- Many possible ways to weight partial matches

Error Types in Strict Evaluation

Token	Paracetamol	for	migraine
True label	Medication	0	Diagnosis
Predicted label	Medication	Diagnosis	
Counted errors	2 = 1FP + 1FN		

Boundary Error

Token	Paracetamol	for	migraine
True label	Medication	0	Diagnosis
Predicted label	Medication	0	Medication
Counted errors	2 = 1FP + 1FN		

Labeling Error

Token	Paracetamol	for	migraine
True label	Medication	0	Diagnosis
Predicted label	Medication	Medication	Diagnosis
Counted errors	1FP		

False Positive

Token	Paracetamol	for	migraine
True label	Medication	0	Diagnosis
Predicted label	Medication	0	0
Counted errors	1FN		

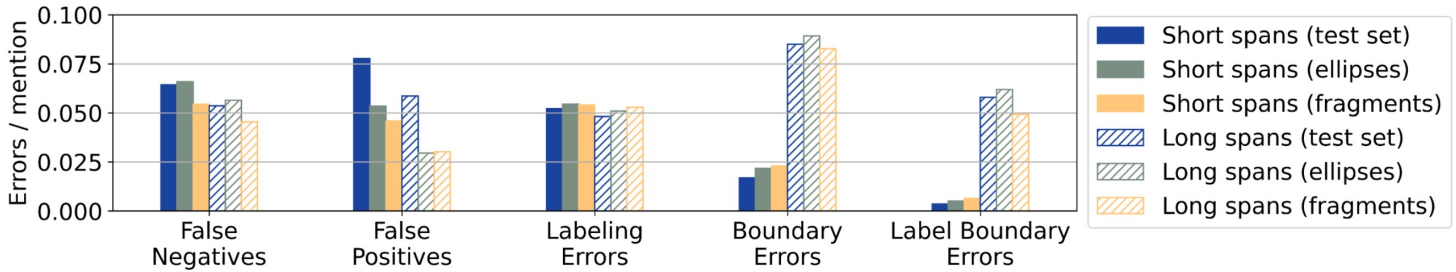
False Negative

- Traditional metrics punish labelling errors, boundary errors, and label boundary errors multiple times
- Alternative scores account for partial overlap, e.g., 1LE = 1BE = 1LBE = 0.5FP + 0.5FN (cf. Ortmann, 2022 and others)

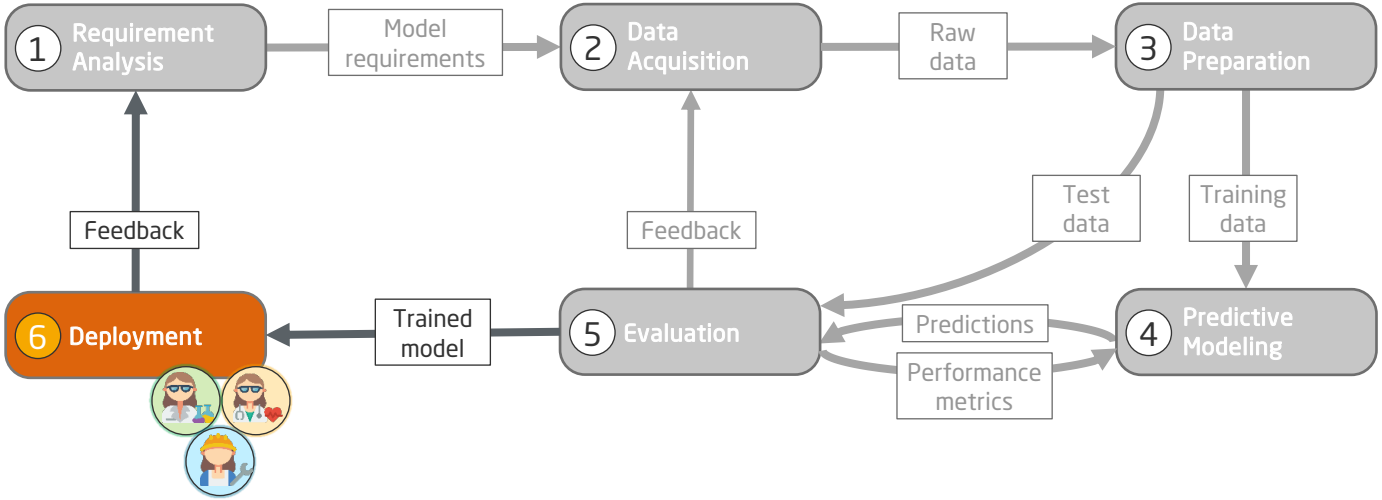
GGPONC: Evaluation and Error Analysis

- Good performance on short / coarse annotations
- Low F1 score for long / fine annotations
- Why? → Error Analysis

F1 Score	Coarse	Fine
Short	**.89**	.86
Long	.75	.72



Deployment

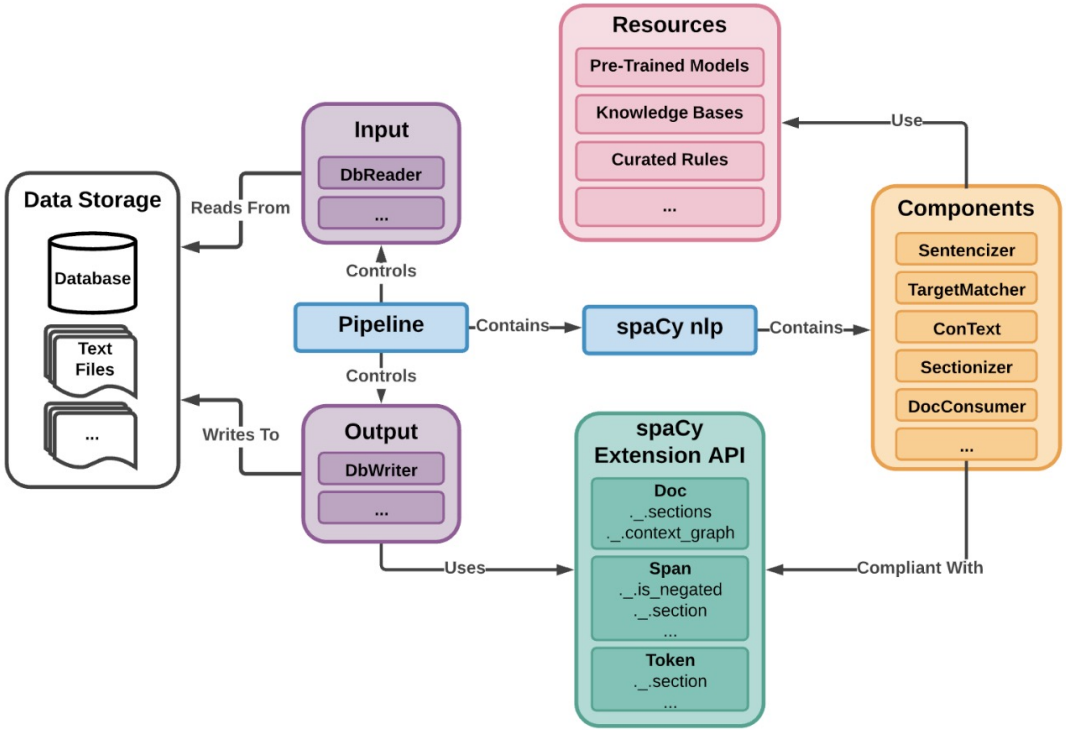


Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

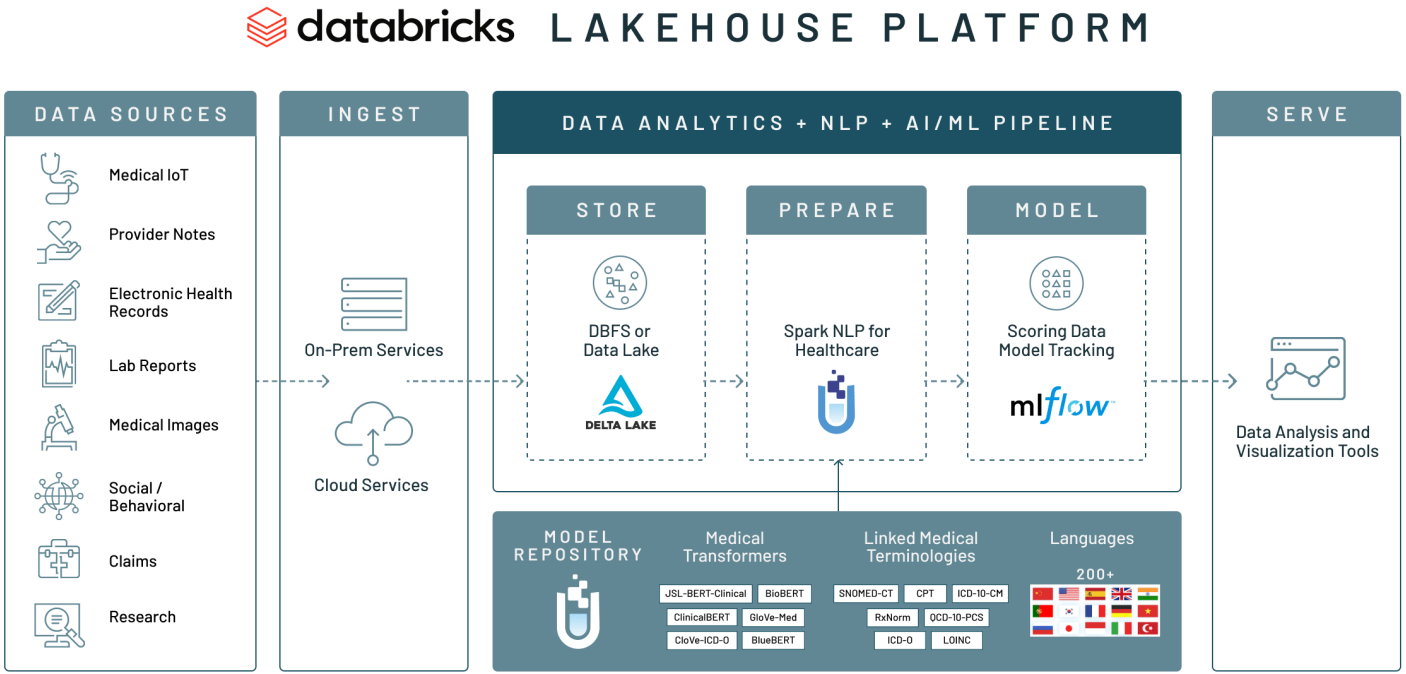
Icons made by Smashicons from www.flaticon.com

medspaCy



Eyre et al. (2021): Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python

Commercial Solutions for Large-scale Clinical NLP



ML and Corpora

Data Management for Digital Health, Winter 2023

https://www.databricks.com/blog/2021/07/01/applying-natural-language-processing-to-healthcare-text-at-scale.html

What to Take Home

- Clinical text corpora (de-identification, annotation, availability)
- Evaluation of NER results
- Features for ML-based NLP algorithms
- Neural networks and neural architectures for Named Entity Recognition

Generated by Dall-E 3