NLP: Natural Language Processing

… Translational Data Science meets
AI in Dutch Healthcare

Speaker: *Prof. dr. Marco Spruit (LUMC/LIACS)*

Talk: *"Free text analysis in EHRs and clinical notes"*
workshop NHG Science Day, 10 June 2022

do's & don'ts

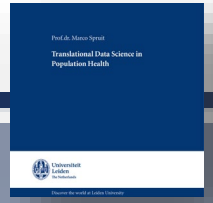
ZyLAB 1993 · 1995 · 1997 · 2003 · 2007 · 2020

Leiden University Medical Center

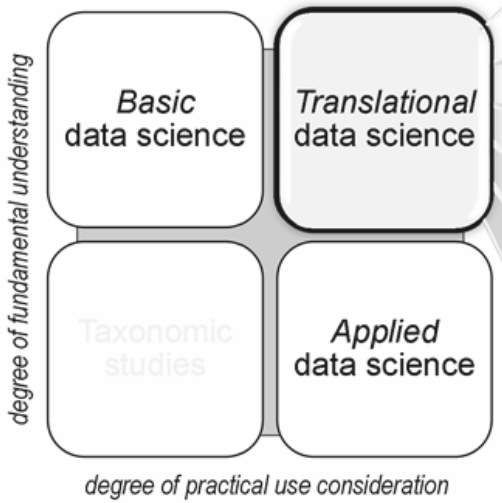
liacs Leiden Institute of Advanced Computer Science

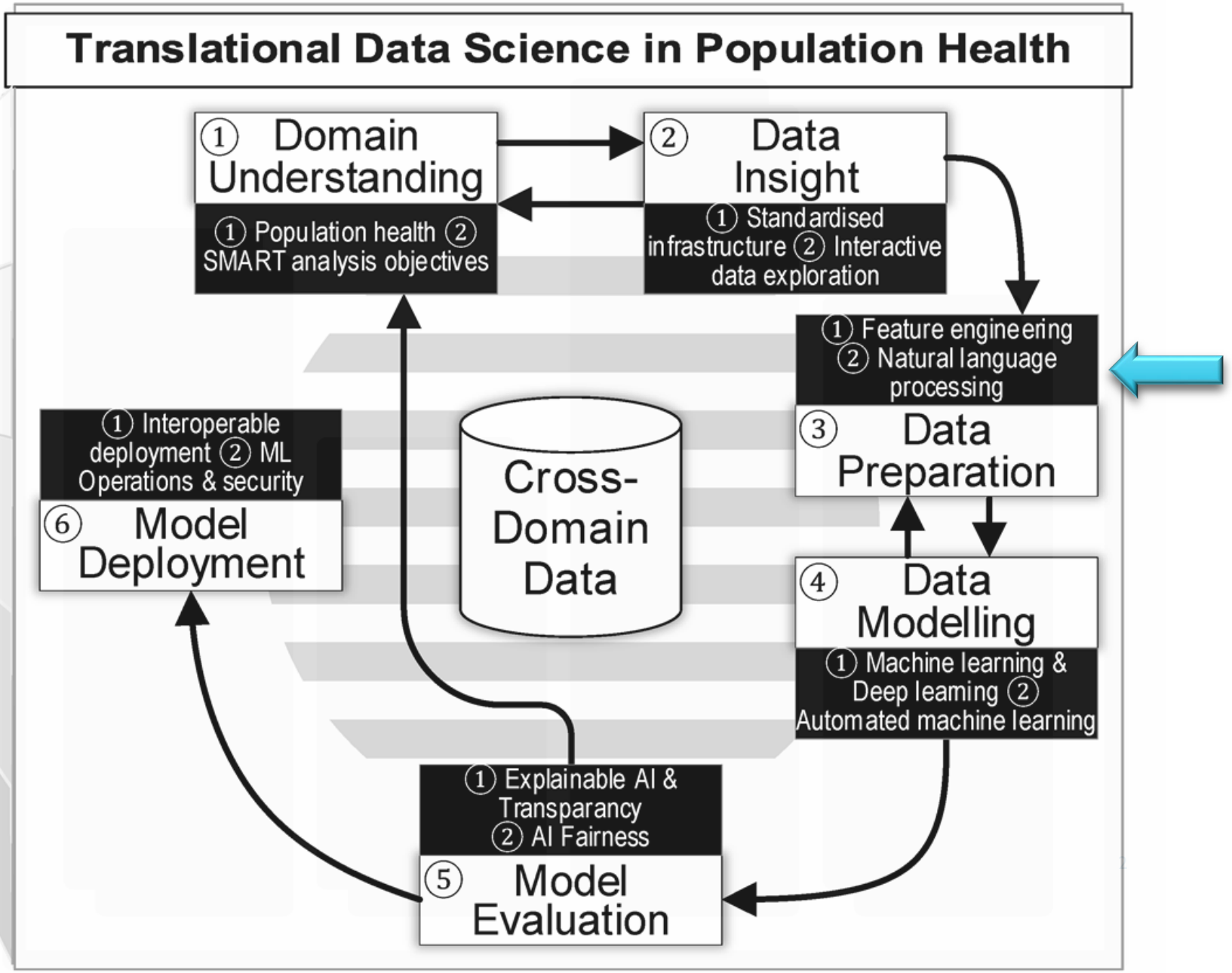
Universiteit Leiden

Leiden University Campus The Hague



APRIL FOOLS' DAY

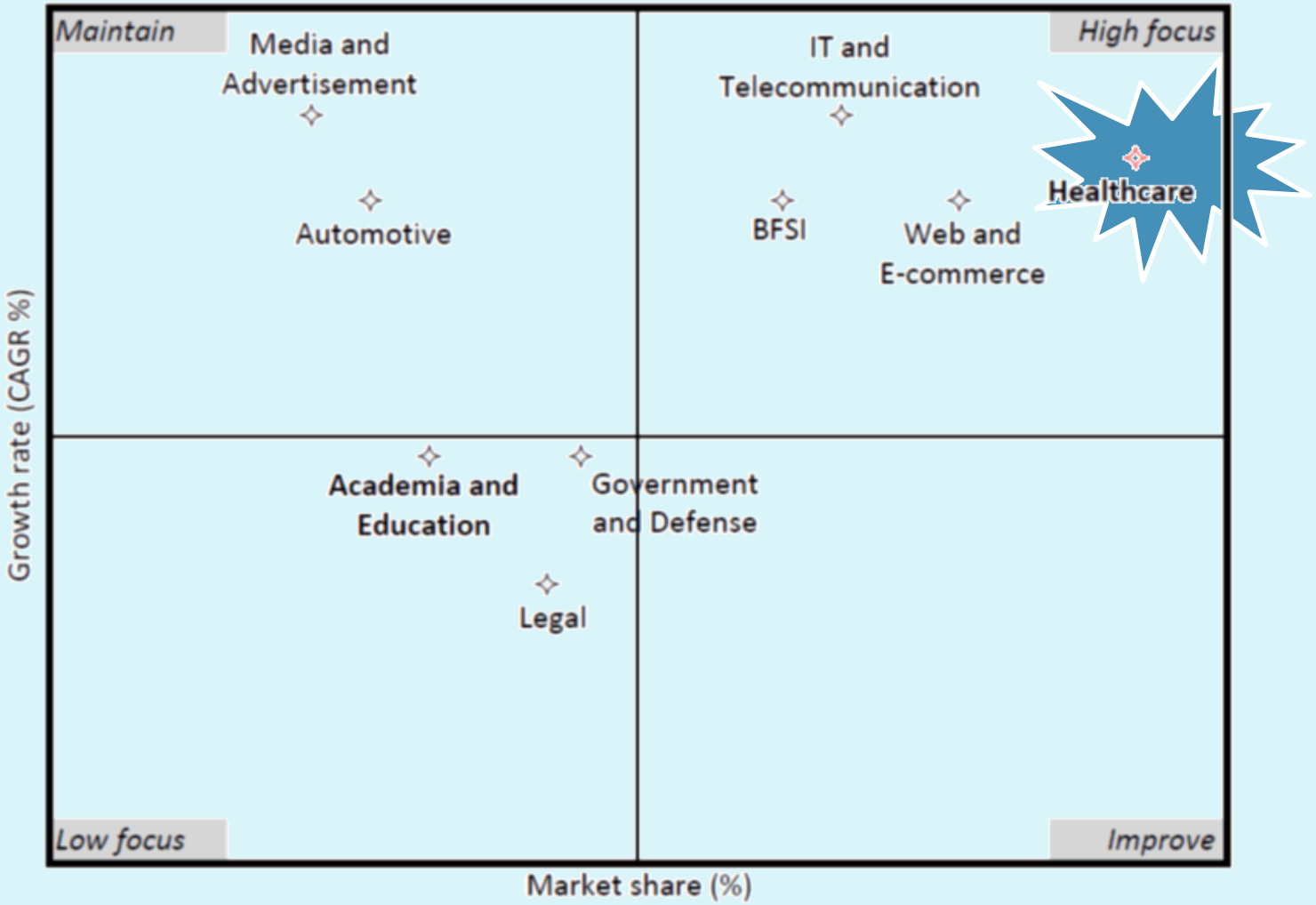




What is the difference between Linguistics and NLP?

Natural Language Processing (NLP) is the study of the computational treatment of natural (human) language.

Natural Language Processing: Here to Stay!

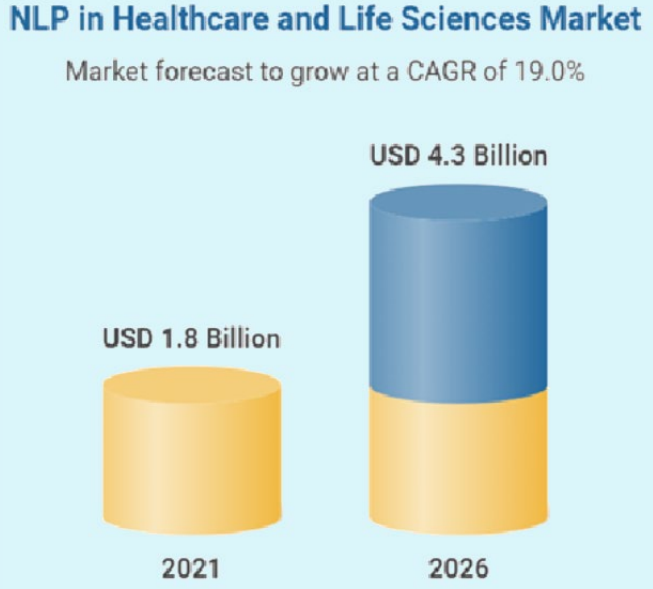


Market forecast (2015)

- *Worldwide:* $ 3.7B in 2013 ➔ $ 9.8B in 2018
- *Healthcare:* $ 1.1B in 2015 ➔ $ 2.7B in 2020

Market forecast (2021)

- *Worldwide:* $ 14.3B in 2021 ➔ $ 61B in 2027
- *Healthcare:* $ 1.8B in 2021 ➔ $ 4.3B in 2026



AGENDA: SETTING THE NLP SCENE WITH EXAMPLES IN HEALTHCARE

"Traditional" NLP 101 → Symbolic NLP

- De-identification & ADR extraction

"Modern" NLP → Probabilistic NLP

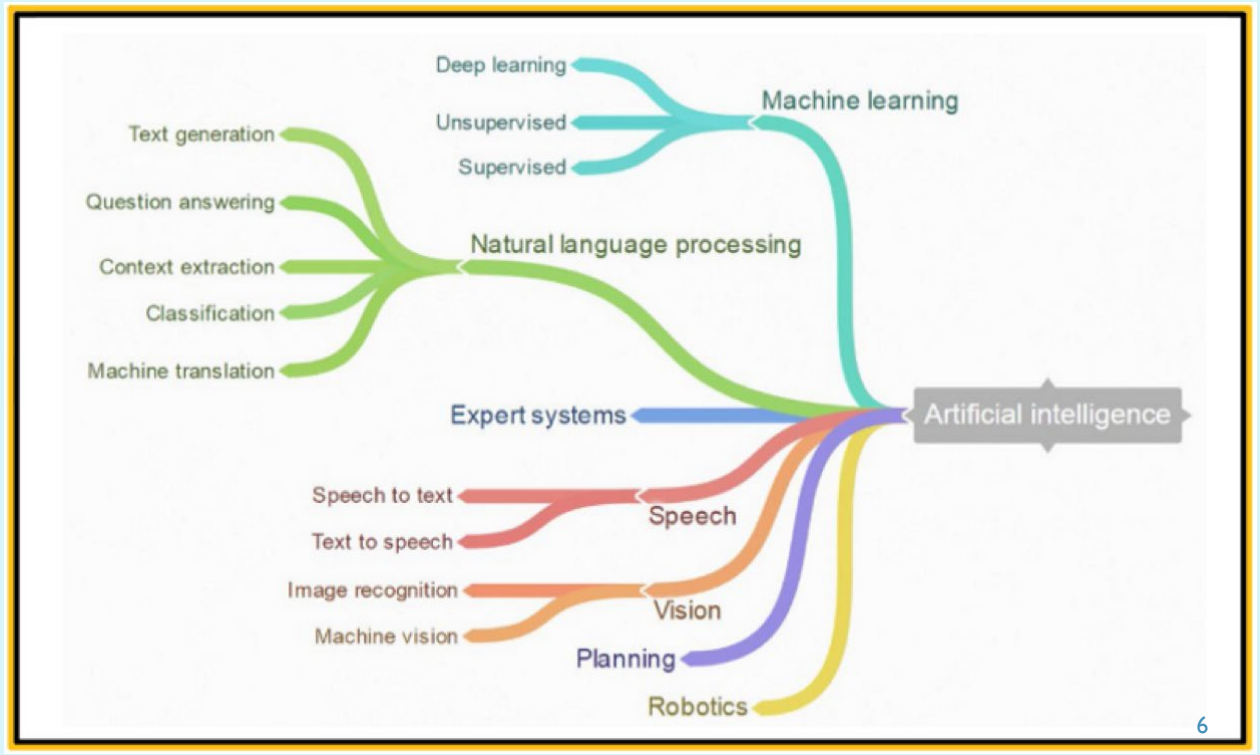
- Embeddings for Classification (VRA)

"Effective" NLP → Hybrid NLP

- Combining traditional and Modern approaches (e.g. ADRIN)

NLP is a Multidisciplinary Wicked Problem within AI

- Computers are confused by (human) language
  - Specific techniques are needed
  - NLP draws on research in
    - Linguistics,
    - Theoretical Computer Science,
    - Mathematics,
    - Statistics,
    - Artificial Intelligence,
    - Psychology,
    - etc.

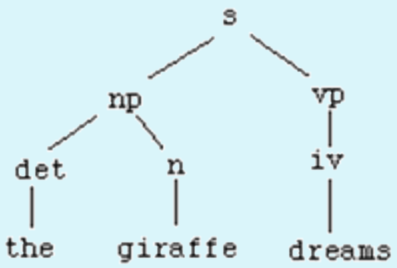


"TRADITIONAL" VERSUS "MODERN" NLP APPROACH

Derivation rules ('50s (Chomsky) →)

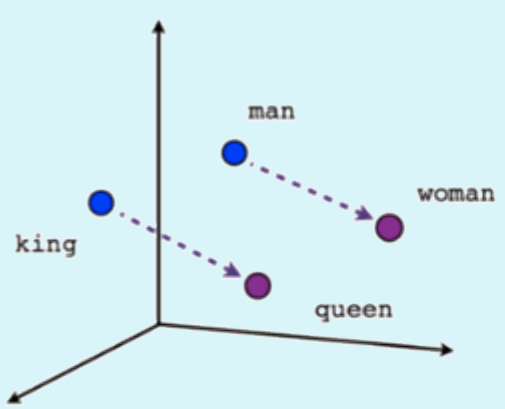
A grammar and a parse tree for "the giraffe dreams".

```
s    ⟶ np vp

np   ⟶ det n
vp   ⟶ tv np
     ⟶ iv

det  ⟶ the
     ⟶ a
     ⟶ an

n    ⟶ giraffe
     ⟶ apple

iv   ⟶ dreams
tv   ⟶ eats
     ⟶ dreams
```

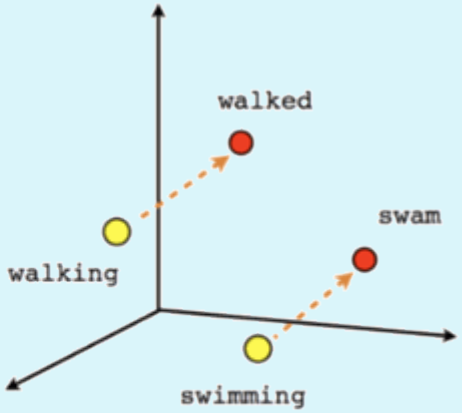


RegEx `^([0|\+[0-9]{1,5})?([0-9]{10})$`

Word embeddings (2013 (Google) →)



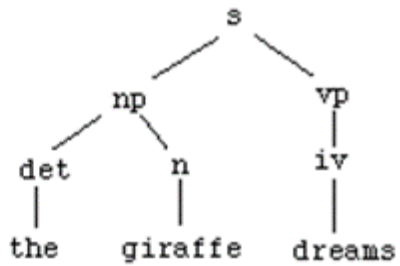
Male-Female



Verb tense

A grammar and a parse tree for "the giraffe dreams".

```
s    → np vp

np   → det n
vp   → tv np
     → iv

det  → the
     → a
     → an

n    → giraffe
     → apple

iv   → dreams
tv   → eats
     → dreams
```



"TRADITIONAL" NLP 101 → SYMBOLIC NLP

DE-IDENTIFICATION & INFORMATION EXTRACTION

Example #1: De-identification in Dutch

Menger,V., Scheepers,F., Wijk,L. van, & Spruit,M. (2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics, 35*(4), Patient Centric Healthcare, 727–736. [JIF: 6.182] [pdf] [online]

DEDUCE:

- De-identification of Dutch medical text
  - Information extraction (NER) of Protected Health Information (PHI) categories
- *Method:* Combines
  - Lookup tables, decision rules, and fuzzy string matching
- https://tdslab.nl/deduce
- `>>> pip install deduce`

[Legend: Patient Persoon Locatie Instelling Datum Leeftijd Patientnummer Telefoonnummer Url]

Annotated text

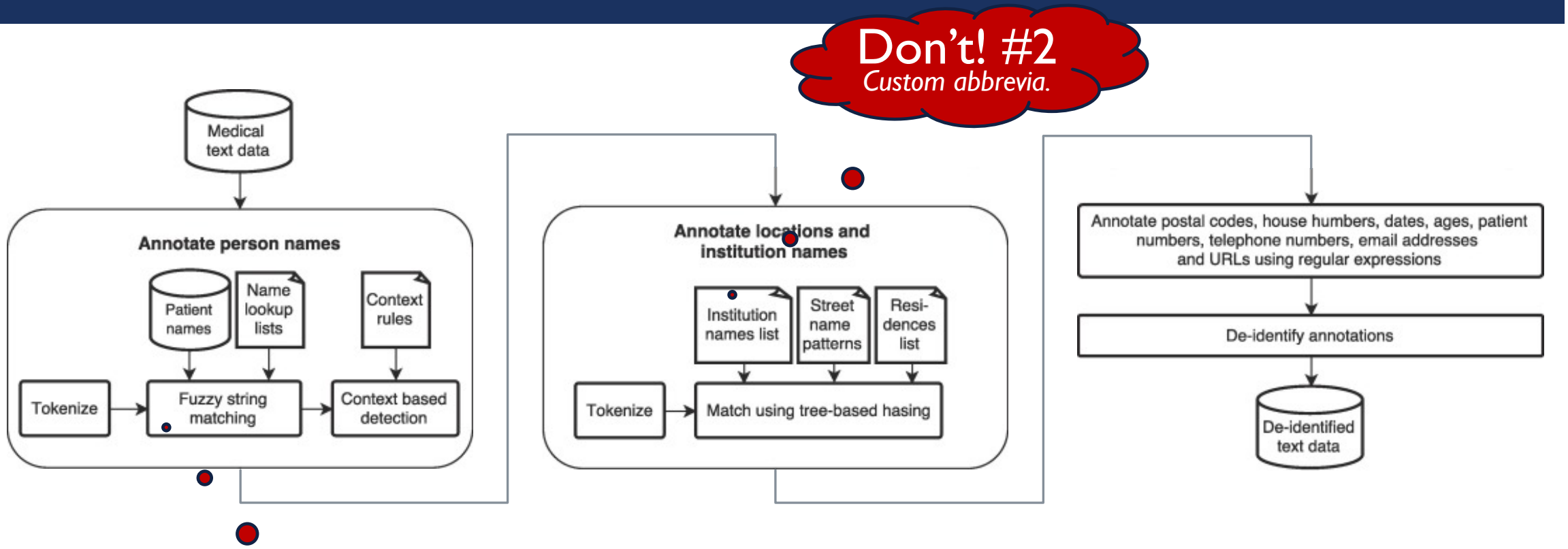
```
Intakegesprek met Jan Jansen (e:j.g.jsnen_1966@email.com,
t:0612345678, patnr:1243567). Het betreft een 51-jarige man
die van 14 maart t/m 31 juli op de polikliniek van het umcu
zal worden behandeld i.v.m. somberheidsklachten. Patient is
woonachtig aan de Voorstraat 45b in Utrecht en zal hier onder
behandeling komen te staan van Peter de Visser.
```

De-identified text

```
Intakegesprek met <PATIENT> (e:<URL-1>, t:<TELEFOONNUMMER-1>,
patnr:<PATIENTNUMMER-1>). Het betreft een <LEEFTIJD-1>-jarige
man die van <DATUM-1> t/m <DATUM-2> op de polikliniek van het
<INSTELLING-1> zal worden behandeld i.v.m.
somberheidsklachten. Patient is woonachtig aan de <LOCATIE-1>
in <LOCATIE-2> en zal hier onder behandeling komen te staan
van <PERSOON-1>.
```

EXAMPLE #1: DE-IDENTIFICATION - METHOD

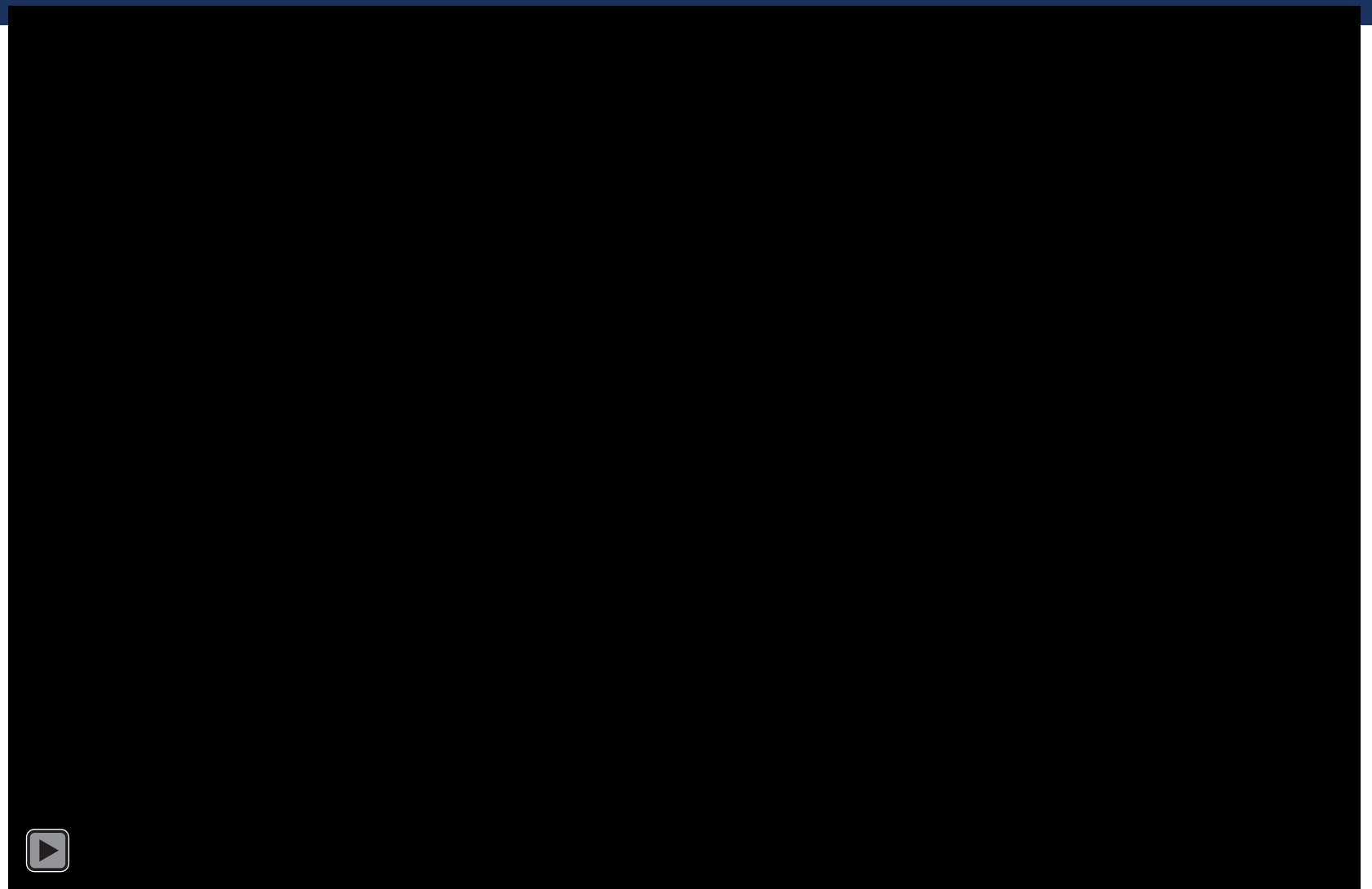
Menger,V., Scheepers,F., Wijk,L. van, & Spruit,M. (2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics, 35*(4), Patient Centric Healthcare, 727–736. [JIF: 6.182] [pdf] [online]



Example #1: De-identification - Demo

Menger,V., Scheepers,F., Wijk,L. van, & Spruit,M. (2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics, 35*(4), Patient Centric Healthcare, 727–736. [JIF: 6.182] [pdf] [online]

- https://tdslab.nl/deduce



EXAMPLE #2: ADR EXTRACTION

Shen,Z., & Spruit,M. (2021). Automatic Extraction of Adverse Drug Reactions from Summary of Product Characteristics. *Applied Sciences, 11*(6), Applications of Artificial Intelligence in Pharmaceutics, 2663. [JIF: 2.679] [pdf] [online]

- *"Automatic Extraction of Adverse Drug Reactions from Summary of Product Characteristics"*
- European Medicines Agency's…
  - The Electronic Medicines Compendium (EMC @UK)
    - >14,000 documents
    - ~ Structured Product Labels (US Food & Drug Admin. (FDA))
- *Aim:* CDSS support to safely use medicines, incl. ADRs
  - In Section 4.8 of the SoPC →
  - But, SoPCs still have a heterogeneous nature…
  - NLP for webscraping! → 647 medicines, in tablet form
  - But… What *is* an Adverse Drug Reaction? <*next page*>

- What *is* a Summary of Product Characteristics (SmPC) ?

https://www.medicines.org.uk/emc

4.8 Undesirable effects

Summary of the safety profile

Headache, abdominal pain, diarrhoea and nausea are among those adverse reactions that have been most commonly reported in clinical trials (and also from post-marketing use). In addition, the safety profile is similar for different formulations, treatment indications, age groups and patient populations. No dose-related adverse reactions have been identified.

Tabulated list of adverse reactions

The following adverse drug reactions have been identified or suspected in the clinical trials programme for esomeprazole and post-marketing. None was found to be dose-related. The reactions are classified according to frequency (very common > 1/10; common ≥1/100 to <1/10; uncommon ≥1/1000 to <1/100; rare ≥1/10000 to <1/1000; very rare <1/10000); not known (cannot be estimated from the available data).

Blood and lymphatic system disorders

Rare: Leukopenia, thrombocytopenia

Very rare: Agranulocytosis, pancytopenia

Immune system disorders

Rare: Hypersensitivity reactions e.g. fever, angioedema and anaphylactic reaction/shock

Metabolism and nutrition disorders

Uncommon: Peripheral oedema

Rare: Hyponatraemia

Not known: Hypomagnesaemia (see section 4.4); severe hypomagnesaemia can correlate with hypocalcaemia. Hypomagnesaemia [illegible] also be associated with hypokalaemia

Psychiatric disorders

Uncommon: Insomnia

Rare: Agitation, confusion, depression

Very rare: Aggression, hallucinations

Don't! #3
NO standardised reporting

EXAMPLE #2: ADR KNOWLEDGE BASE: MEDDRA®

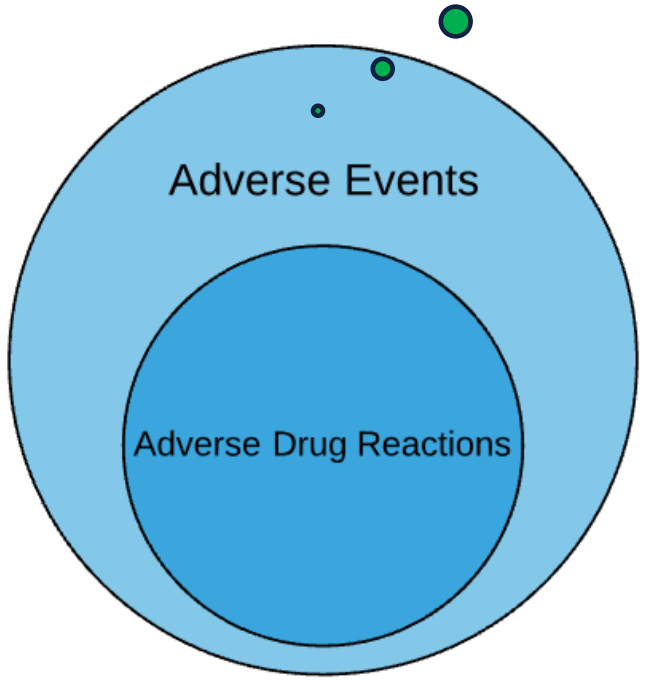
We'll meet MedDRA again later!

https://www.meddra.org/browsers

- **MedDRA:** Medical Dictionary for Regulatory Activities
  - Hierarchical, multilingual, standardised taxonomy of Adverse events

	Class	Term
1	**System Organ Class**	Cardiac disorders
2	High Level Group Term	Heart failures
3	High Level Term	Left ventricular failures
4	Preferred Term	Chronic left ventricular failure
5	Lowest Level Term	Chronic diastolic heart failure

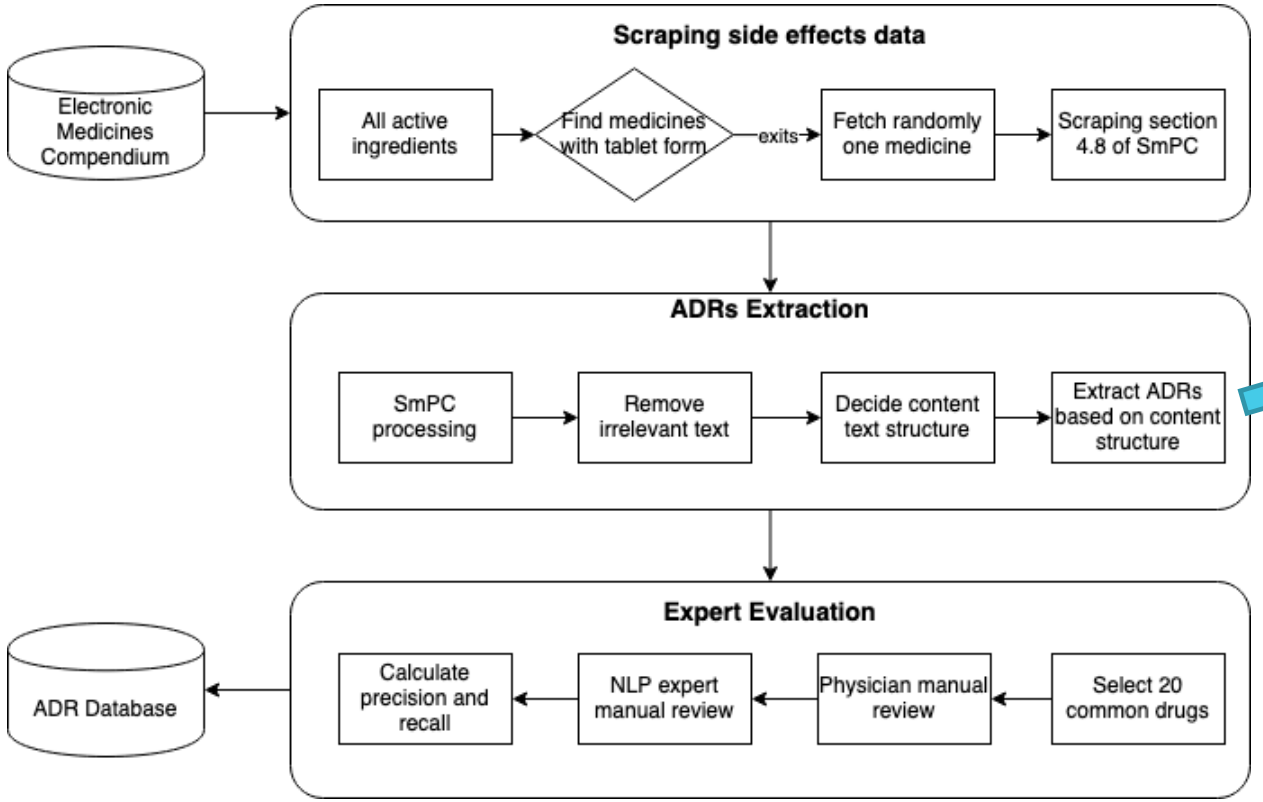
Do! #1
Standardise reporting



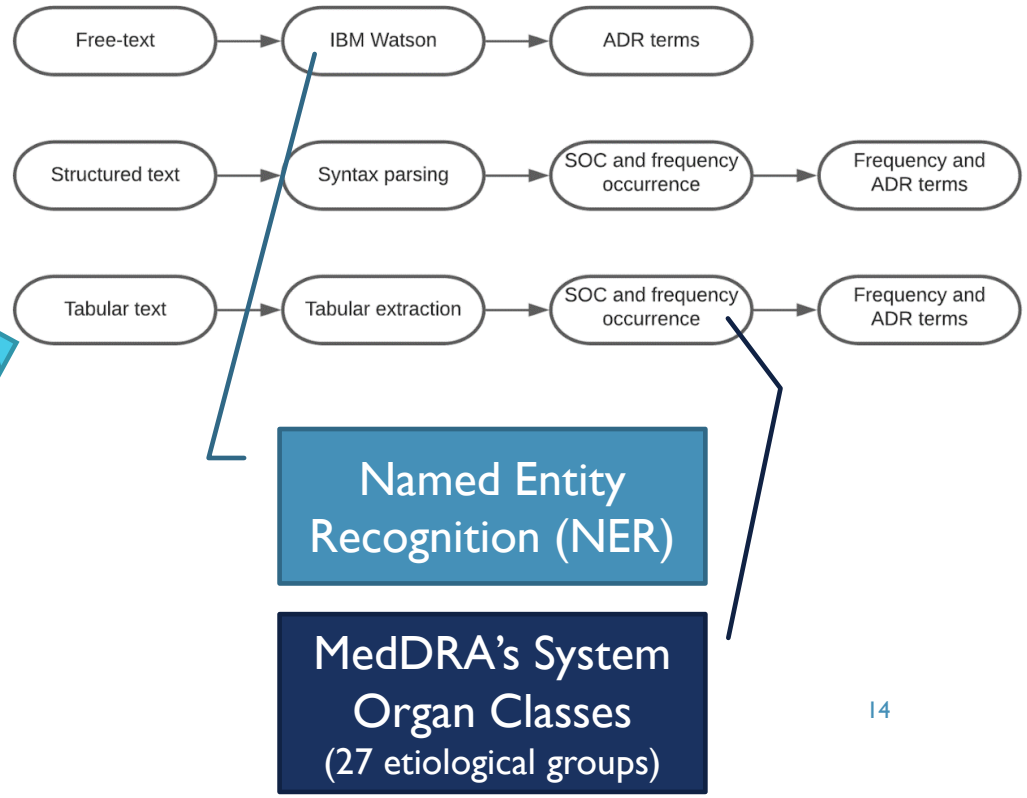
EXAMPLE #2: ADR EXTRACTION - METHOD

Shen,Z., & Spruit,M. (2021). Automatic Extraction of Adverse Drug Reactions from Summary of Product Characteristics. *Applied Sciences, 11*(6), Applications of Artificial Intelligence in Pharmaceutics, 2663. [JIF: 2.679] [pdf] [online]

- Adverse Drug Reactions (ADRs) extraction dev pipeline



- 3 flows within the ADR Extraction Pipeline



EXAMPLE #2: ADR EXTRACTION - PARSING

Shen,Z., & Spruit,M. (2021). Automatic Extraction of Adverse Drug Reactions from Summary of Product Characteristics. *Applied Sciences, 11*(6), Applications of Artificial Intelligence in Pharmaceutics, 2663. [JIF: 2.679] [pdf] [online]

- *Parsing:* An Adverse Drug Reactions (ADRs) extraction example for the structured text:
  https://www.medicines.org.uk/emc/product/2729#UNDESIRABLE_EFFECTS

Immune system disorders:

Very rare: Hypersensitivity reactions including urticaria, angio-oedema or anaphylactic reactions.

Psychiatric disorders:

Common: Decreased libido.

Uncommon: Increased libido.

Gastrointestinal disorders:

Very common: Diarrhoea.

Common: Abdominal pain, nausea, vomiting, flatulence.

Skin and subcutaneous tissue disorders:

Common: Pruritus, maculo-papular rash.

Not known: Vesiculo-bullous eruptions.

Reproductive system and breast disorders:

Common: Frigidity or impotence.

ADRs Extraction →


```
"atc_code": "N07BB03",
"adrs": {
  "very rare": [
    "hypersensitivity reactions including urticaria",
    "angio-oedema or anaphylactic reactions."
  ],
  "uncommon": [
    "increased libido."
  ],
  "common": [
    "decreased libido.",
    "abdominal pain",
    "nausea",
    "vomiting",
    "flatulence.",
    "pruritus",
    "maculo-papular rash.",
    "frigidity or impotence."
  ],
  "very common": [
    "diarrhoea."
  ],
  "unknown": [
    "vesiculo-bullous eruptions."
  ]
}
```

EXAMPLE #2: ADR EXTRACTION – EMC VS LAREB

Shen,Z., & Spruit,M. (2021). Automatic Extraction of Adverse Drug Reactions from Summary of Product Characteristics. *Applied Sciences, 11*(6), Applications of Artificial Intelligence in Pharmaceutics, 2663. [JIF: 2.679] [pdf] [online]

https://www.lareb.nl/nl/databank/result?atc=N07BB03&lang=nl&ref=FK

ACAMPROSAAT

Maagsapresistente tablet

Let op, de resultaten van de zoekopdracht hebben betrekking op de gehele groep geneesmiddelen die tot de werkzame stof in ACAMPROSAAT behoort. Klik *hier* voor een volledig overzicht van deze groep middelen.

Klik voor meer informatie over:

Bekende bijwerkingen

Behalve het gewenste effect kan dit medicijn bijwerkingen geven.

Als u epilepsie heeft en gaat stoppen met alcohol
Raadpleeg eerst uw arts. U kunt in de eerste 2 dagen na stoppen epileptische aanvallen krijgen. Uw arts zal u extra controleren.

De belangrijkste bijwerkingen zijn de volgende:

Soms (bij 10 tot 30 op de 100 mensen)
Maagdarmklachten, zoals diarree. Zelden buikpijn, misselijkheid en braken.

Als u dit medicijn inneemt tijdens het eten of met wat voedsel, heeft u er minder last van.

Zelden (bij 1 tot 10 op de 100 mensen)
Huiduitslag met jeuk. Zeer zelden wijst dit op overgevoeligheid (zie *Zeer zelden*).

Seksuele stoornissen, zoals impotentie, minder zin in vrijen of, zeer zelden, meer zin in vrijen.

Zeer zelden (bij minder dan 1 op de 100 mensen)
Overgevoeligheid voor dit medicijn. Dit merkt u aan huiduitslag, jeuk, galbulten, benauwdheid, duizeligheid of flauwvallen. Ook kunt u last krijgen van zwelling van het gezicht, lippen, mond, tong of keel. In deze gevallen moet u onmiddellijk een arts opzoeken of naar de Eerste-Hulpdienst gaan.

Als u overgevoelig bent voor dit medicijn, mag u het niet meer gebruiken. Geef aan de apotheker door dat u overgevoelig bent voor acamprosaat. Het apotheekteam kan er dan op letten dat u het in de toekomst niet meer krijgt.

Bron: Apotheek.nl

ADRs Extraction →


```
"atc_code": "N07BB03",
"adrs": {
  "very rare": [
    "hypersensitivity reactions including urticaria",
    "angio-oedema or anaphylactic reactions."
  ],
  "uncommon": [
    "increased libido."
  ],
  "common": [
    "decreased libido.",
    "abdominal pain",
    "nausea",
    "vomiting",
    "flatulence.",
    "pruritus",
    "maculo-papular rash.",
    "frigidity or impotence."
  ],
  "very common": [
    "diarrhoea."
  ],
  "unknown": [
    "vesiculo-bullous eruptions."
  ]
}
```

Tawfik,N., & Spruit,M. (2018). The SNPCurator: Literature mining of SNP disease association. *Database: The Journal of Biological Databases and Curation, 2018*(January), bay020. [JIF: 3.451] [online]

EXAMPLE #3: INFORMATION EXTRACTION - SNPCURATOR

- PubMed literature mining of enriched SNP-disease associations
- https://tdslab.nl/snpcurator

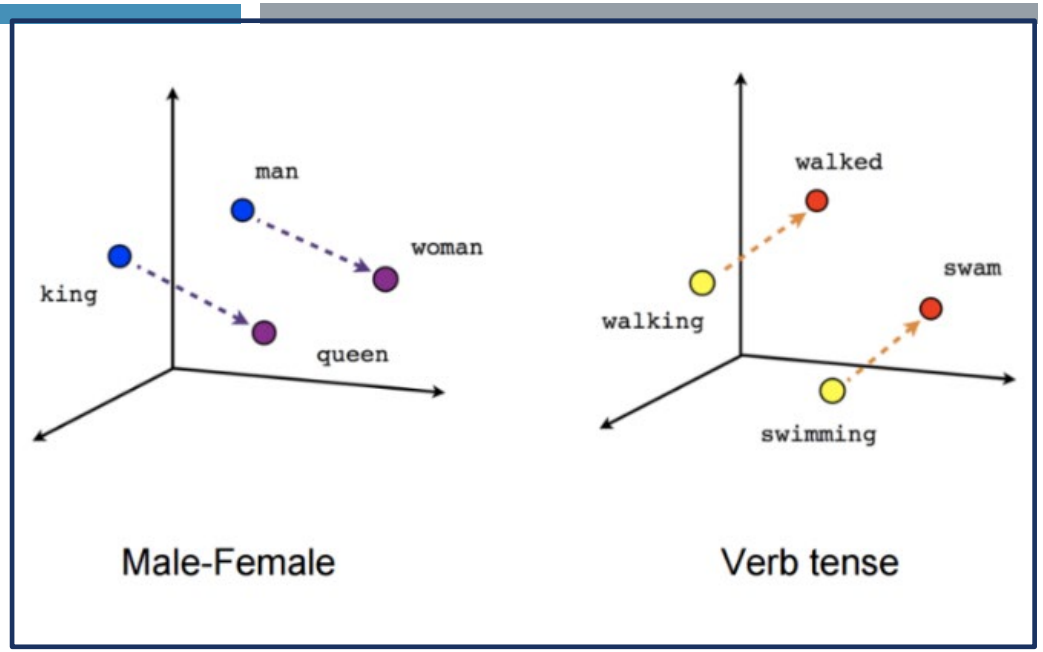
SNP Curator

Results for **ibd**:
- A total of **1535** articles were fetched by PubMed.
- The extracted list of abstracts was shortened to **280** via selecting those comprised of SNP mentions.
- **150** PubMed article(s) had statistical results reported withing the abstract text with a total of **524** SNP pairs.

Go back to home page Export Data to CSV File

SNP	PMID	Title	Date	Pvalue	ORvalue	Ethnicity	Patient group Size	Control group Size	Frequency	Text Evidence
rs61750370	29788244	Nonsynonymous Polymorphism in Guanine Monophosphate Synthetase Is a Risk Factor for Unfavorable Thiopurine Metabolite Ratios in Patients With Inflammatory Bowel Disease.		0.031		Caucasian	264		2	-
The SNP rs61750370 was significantly associated with 6-MMP:6-TGN ratios ≥100 odds ratio, 5.64; 95% confidence interval, 1.01-25.12; P < 0.031 in a subset of 264 Caucasian IBD patients.The GMPS SNP rs61750370 may be a reliable risk factor for extreme 6MMP preferential metabolism.										
rs61750370	29788244	Nonsynonymous Polymorphism in Guanine Monophosphate Synthetase Is a Risk Factor for Unfavorable Thiopurine Metabolite Ratios in Patients With Inflammatory Bowel Disease.		0.031		Caucasian	264		2	+
rs16969968	29688464	Smoking Interacts With CHRNA5, a Nicotinic Acetylcholine Receptor Subunit Gene, to Influence the Risk of IBD-Related Surgery.		0.05					4	+

"Mathematics with Language"



"MODERN" NLP → PROBABILISTIC NLP

EMBEDDINGS FOR CLASSIFICATION (VRA)

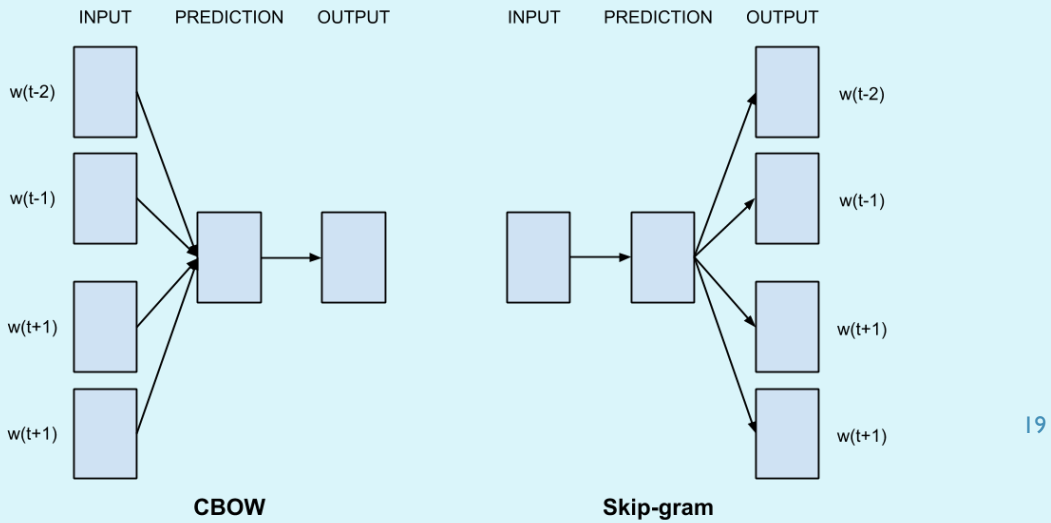
w[woman] - w[man] ≃ w[queen] - w[king]
→ *"man is to king as woman is to…. queen"*

Word Embeddings as Text Representations

vs GloVe
= count-based

- A **word embedding** is one of the most popular representations of *document vocabulary*.
- It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.
- Word embedding are simply : **vector representations** of a particular word.
  - Each word is mapped to one vector, and
  - the vector values are learned in a way that resembles a neural network (*i.e.* ML)
  - *Objective:* to have words with similar context occupy close spatial positions.

- **word2vec** is a "predictive" model using
  - *Continous Bag Of Words (CBOW):* takes the context of each word as the input and tries to predict the word corresponding to the context
    - captures co-occurrence one window at a time
    - *Skip-gram* is the inverse of CBOW (is better for rare words)



https://remykarem.github.io/word2vec-demo/

Place your text below:

Marco Spruit is Hoogleraar Geavanceerde Datawetenschap in Populatiegerichte Zorg aan de Universiteit Leiden bij zowel het departement Publieke Zorg & Eerstelijnsgeneeskunde (PHEG) aan de Medische Faculteit (LUMC) als het Leiden Instituut voor Informatica (LIACS) aan de Faculteit der Wiskunde & Natuurwetenschappen (FWN). Hij is zowel geïnteresseerd in het vertalen van nieuwe algoritmes naar nieuwe zorgtoepassingen als in het implementeren van nieuwe inzichten uit deze nieuwe toepassingen in de dagelijkse praktijk.

Specify model and parameters to generate dataset:

Model: CBOW

The Word2Vec (CBOW) model trains multiple words (in the form of bag-of-words) surrounding a target word.

Window size: 3

☐ Negative sampling?

Generate dataset

Tokens:

No.	Token	Freq
1	marco	5
2	spruit	1
3	is	3
4	hoogleraar	1
5	geavanceerde	1

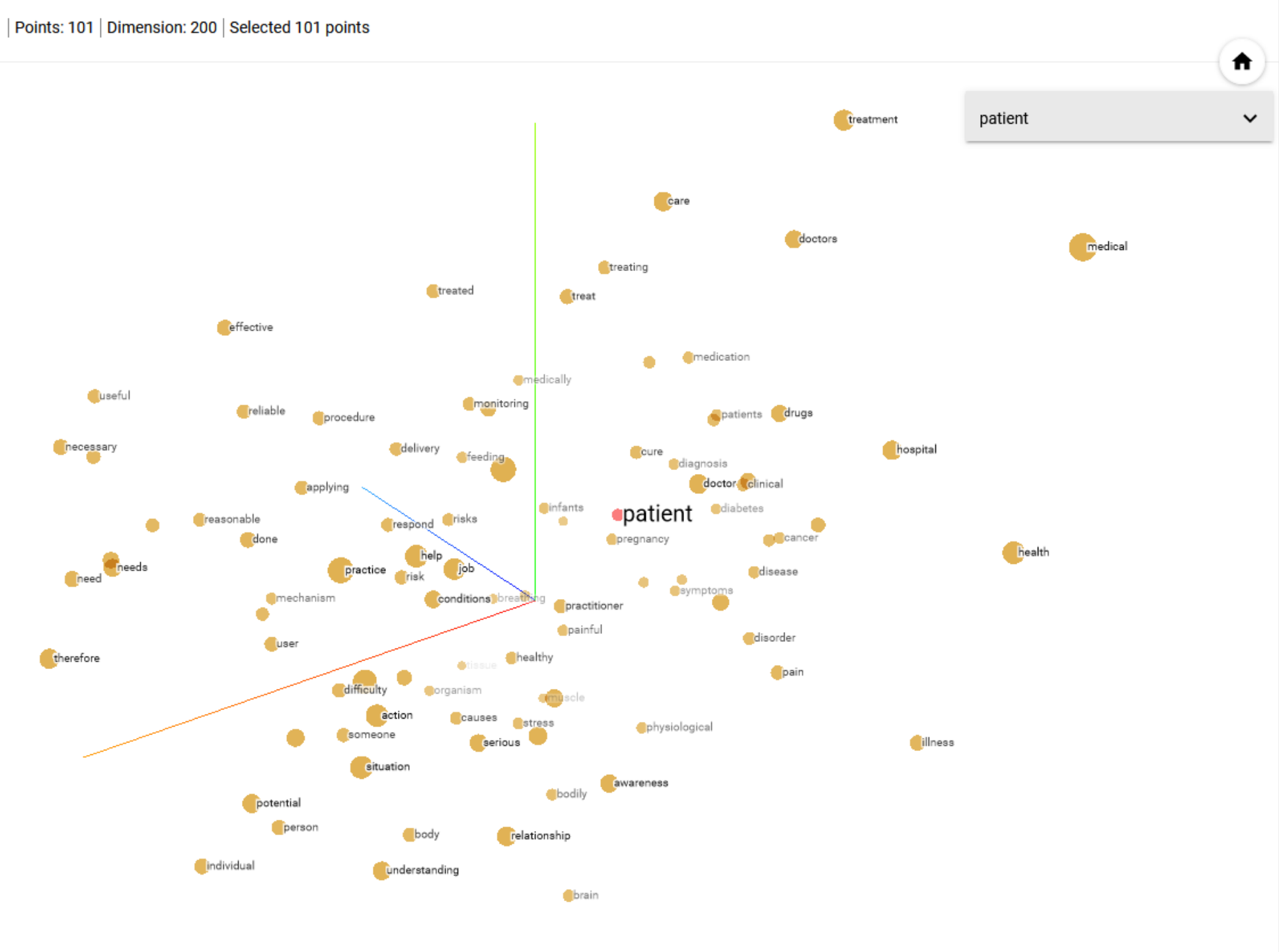
Vocabulary size: 217

Training examples:

4	marco spruit is geavanceerde datawetenschap in	hoogleraar
5	spruit is hoogleraar datawetenschap in populatiegerichte	geavanceerde
6	is hoogleraar geavanceerde in populatiegerichte zorg	datawetenschap
7	hoogleraar geavanceerde	in

No. of training examples: 397

https://remykarem.github.io/word2vec-demo/



Show All Data | Isolate 101 points | Clear selection

Search: patient .* by word

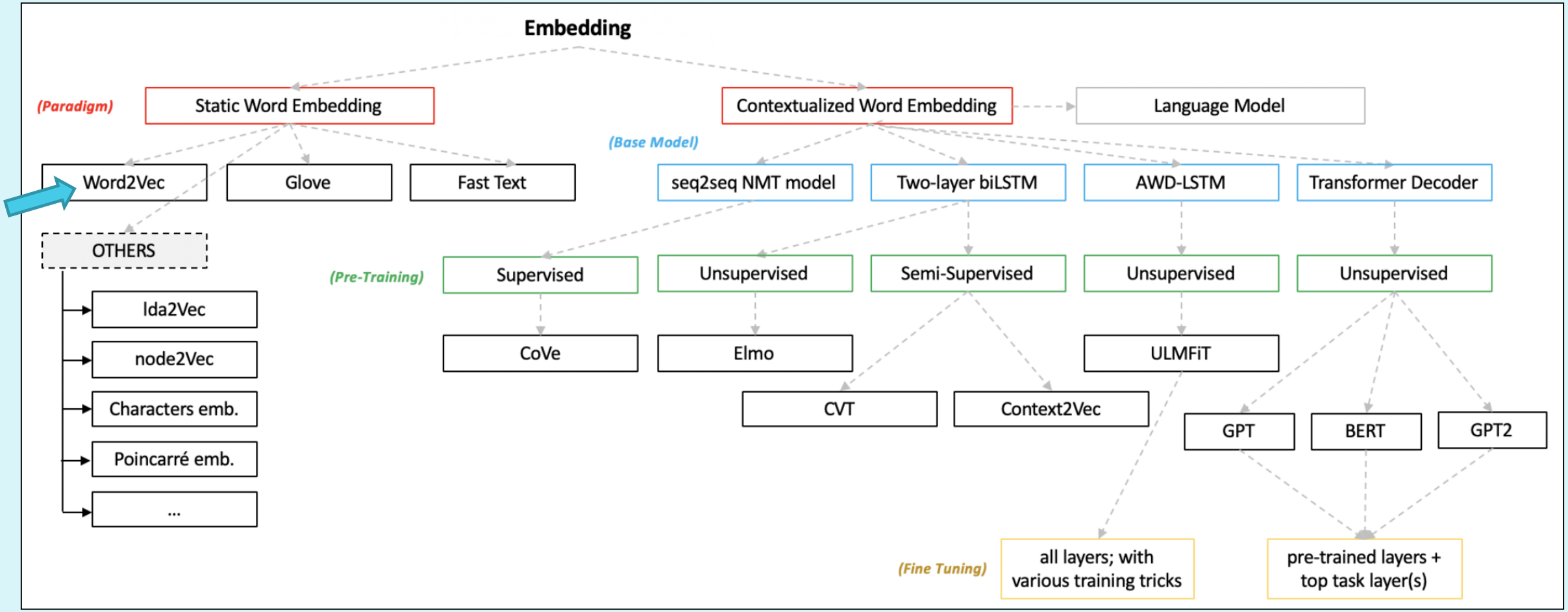
neighbors: 100

distance: COSINE EUCLIDEAN

Nearest points in the original space:

patients	0.740
diagnosis	0.775
clinical	0.811
symptoms	0.837
lung	0.845
doctor	0.847
hospital	0.855
treatment	0.869
cancer	0.873
breathing	0.882
diabetes	0.888
acute	0.897
practitioner	0.898
medically	0.898
medication	0.907
inducing	0.910
therapy	0.911
infants	0.914
someone	0.918
disease	0.922
treating	0.928
feeding	0.935

… Many state-of-the-art Text Representation Techniques!



EXAMPLE #4: VIOLENCE RISK PREDICTION

Menger,V., Spruit,M., Est,R. van, Nap,E., & Scheepers,F. (2019). Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Network Open, 2*(7), e196709. [JIF: 8.483] [pdf] [online]

- "Predict for which admissions a violence incident will occur in the first 30 days, **based on clinical texts** that are written up to and including the first day of admission"
  - 2*3200 admissions, 2000 words/note, 950+650 incidents
  - Prediction task excludes incidents on Day 1 of admission
  - 30 days interval chosen for sufficient specificity
- Internal and external validation (UMCU, Antes R'dam)
  - Area Under Curve (AUC) to report performance

(2012-07-29)
"Mw heeft matig geslapen, sliep van 1.00 uur tot 4.00 uur. Kwam toen uit bed, at koekjes en dronk thee. Nog geadviseerd medicatie te nemen en mijn zorgen geuit over evt. doorschieten in een manie. Mw was er niet gevoelig voor en reageerde geagiteerd. Mw had spreekdrang maar gaf aan dat wanneer zij zich goed voelt ook veel praat. Mw gaat vandaag naar <PERSOON-1> met haar zoon, ziet daar nu niet meer tegenop omdat de klachten die zij gisteren aan haar voeten ervaarde verdwenen zijn. Mw ging na 4.00 uur weer naar bed en kwam niet meer uit haar kamer tot de ochtend."

?

Do! #2

Discuss reporting guidelines

EXAMPLE #4: VIOLENCE RISK – DATA SAMPLE

Menger,V., Spruit,M., Est,R. van, Nap,E., & Scheepers,F. (2019). Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Network Open, 2*(7), e196709. [JIF: 8.483] [pdf] [online]

Text representation

- Represent all clinical notes related to 1 admission as 1 vector (*i.e.* not as words)

- *Representation:* paragraph2vec
- *Classification:* SVM

EXAMPLE #4: VIOLENCE RISK – PERFORMANCE

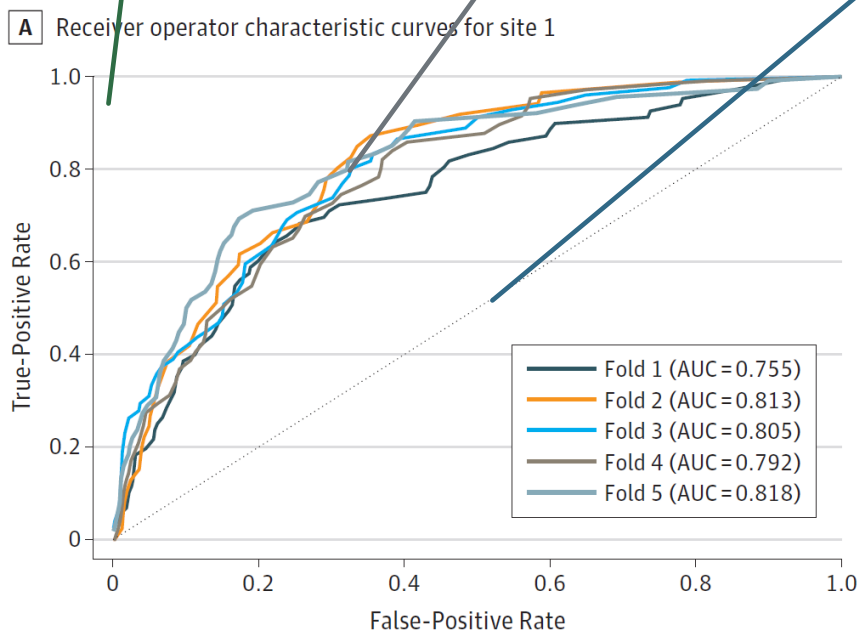
Menger,V., Spruit,M., Est,R. van, Nap,E., & Scheepers,F. (2019). Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Network Open, 2*(7), e196709. [JIF: 8.483] [pdf] [online]

Perfect Score

Actual Score

Random Score

Figure. Receiver Operator Characteristic Curves for Internal Cross-validations



Receiver operator characteristic curves are shown for each fold, according to internal cross-validation in site 1 (A) and site 2 (B). Dashed diagonal lines denote an area under the curve (AUC) of 0.5, ie, predictive validity equivalent to chance. AUC indicates area under the curve.

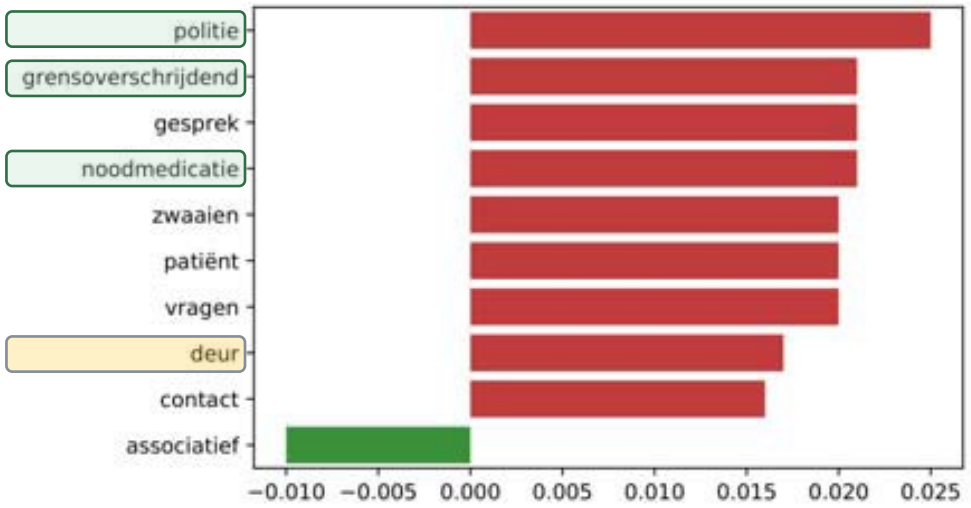
EXAMPLE #4: VIOLENCE RISK – INTERPRETATION OF REPRESENTATIVE WORDS

Table 3. Results of Exploratory Analysis

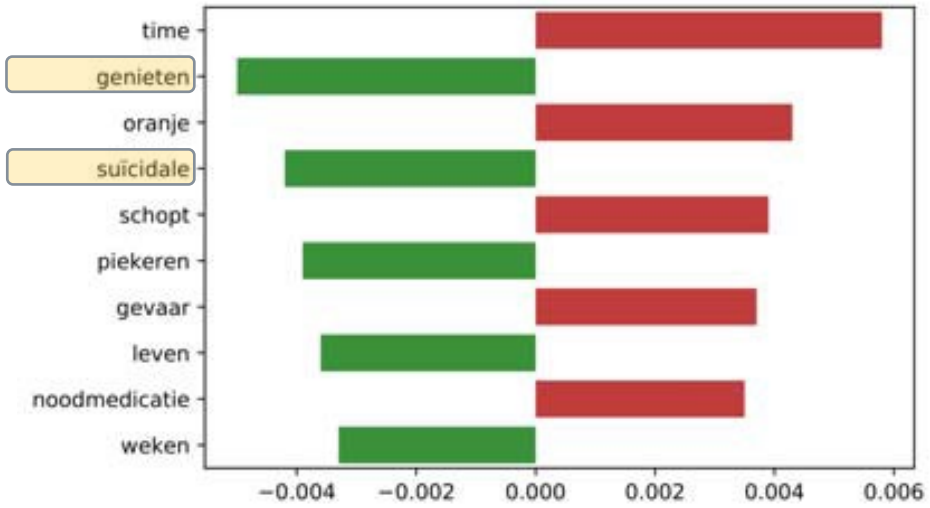
	Site 1				Site 2			
Rank[a]	Term (English Translation)[b]	Ratio	MCC (95% CI)[c]	*P* Value[d]	Term (English Translation)[b]	Ratio	MCC (95% CI)[c]	*P* Value[d]
1	Agressief (aggressive)	1.00	0.17 (0.13 to 0.21)	<.001	Verbaal (verbal)	1.00	0.14 (0.10 to 0.18)	<.001
2	Reageert (reacts)	1.00	0.15 (0.11 to 0.19)	<.001	Dreigend (threatening)	1.00	0.13 (0.08 to 0.16)	<.001
3	Aangeboden (offered)	1.00	0.14 (0.11 to 0.18)	<.001	Agressie (aggression)	1.00	0.15 (0.11 to 0.17)	<.001
4	Boos (angry)	1.00	0.16 (0.12 to 0.19)	<.001	Hierop ([up]on this)	1.00	0.13 (0.09 to 0.16)	<.001
5	Deur (door)	1.00	0.14 (0.10 to 0.18)	<.001	Kantoor (office)	1.00	0.12 (0.08 to 0.16)	<.001
6	Loopt (walks)	1.00	0.15 (0.11 to 0.18)	<.001	Personeel (staff)	1.00	0.12 (0.07 to 0.16)	<.001
7	Ibs (arrest)	1.00	0.14 (0.10 to 0.17)	<.001	Aangesproken (spoke to)	1.00	0.11 (0.08 to 0.15)	<.001
8	Aanbieden (offer)	1.00	0.12 (0.08 to 0.15)	<.001	Agressief (aggressive)	0.99	0.11 (0.08 to 0.15)	<.001
9	Noodmedicatie (emergency medication)	0.99	0.14 (0.10 to 0.17)	<.001	Gevaar agressie (danger aggression)	0.99	0.11 (0.07 to 0.15)	<.001
10	Liep (walked)	0.99	0.12 (0.08 to 0.16)	<.001	Agitatie (agitation)	0.99	0.11 (0.07 to 0.14)	<.001
11	Agressie (aggression)	0.99	0.13 (0.09 to 0.18)	<.001	Geirriteerd (irritated)	0.99	0.10 (0.06 to 0.14)	.001
12	Vraagt (asks)	0.99	0.13 (0.10 to 0.17)	<.001	Separeer (seclusion room)	0.99	0.10 (0.06 to 0.15)	<.001
13	Status vrijwillig (status voluntary)	0.99	−0.12 (−0.14 to −0.09)	<.001	Loopt (walks)	0.99	0.11 (0.08 to 0.14)	.02
14	Psychotisch (psychotic)	0.98	0.12 (0.09 to 0.16)	<.001	Grond (ground)	0.98	0.10 (0.06 to 0.14)	<.001
15	Collega (colleague)	0.98	0.11 (0.07 to 0.15)	<.001	Aanvang (commencement)	0.98	0.11 (0.08 to 0.14)	.01
16	Spreekt (speaks)	0.97	0.12 (0.08 to 0.15)	<.001	Mede (also)	0.98	0.10 (0.07 to 0.14)	.001
17	Gehouden (obliged)	0.97	0.11 (0.07 to 0.15)	<.001	Dhr wilde (Mr wanted)	0.98	0.10 (0.06 to 0.14)	.001
18	Beoordelen (judge), verb	0.96	0.11 (0.07 to 0.15)	<.001	Liep (walked)	0.98	0.10 (0.06 to 0.14)	.006
19	Momenten (moments)	0.96	0.12 (0.08 to 0.15)	<.001	Geagiteerd (agitated)	0.96	0.10 (0.06 to 0.14)	.01
20	Somber (dejected)	0.95	−0.14 (−0.17 to −0.11)	<.001	cvd (not available)	0.96	0.10 (0.06 to 0.14)	.004

EXAMPLE #4: MODEL INTERPRETATION

Menger,V., Spruit,M., Est,R. van, Nap,E., & Scheepers,F. (2019). Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Network Open, 2*(7), e196709. [JIF: 8.483] [pdf] [online]



- Sample of Local Explanation predicting high risk of aggression



- Sample of Local Explanation predicing low risk of aggression

The *"Linear Model-Agnostic Explanations"* (LIME) method

EXAMPLE #4: ALTERNATIVE MODELS

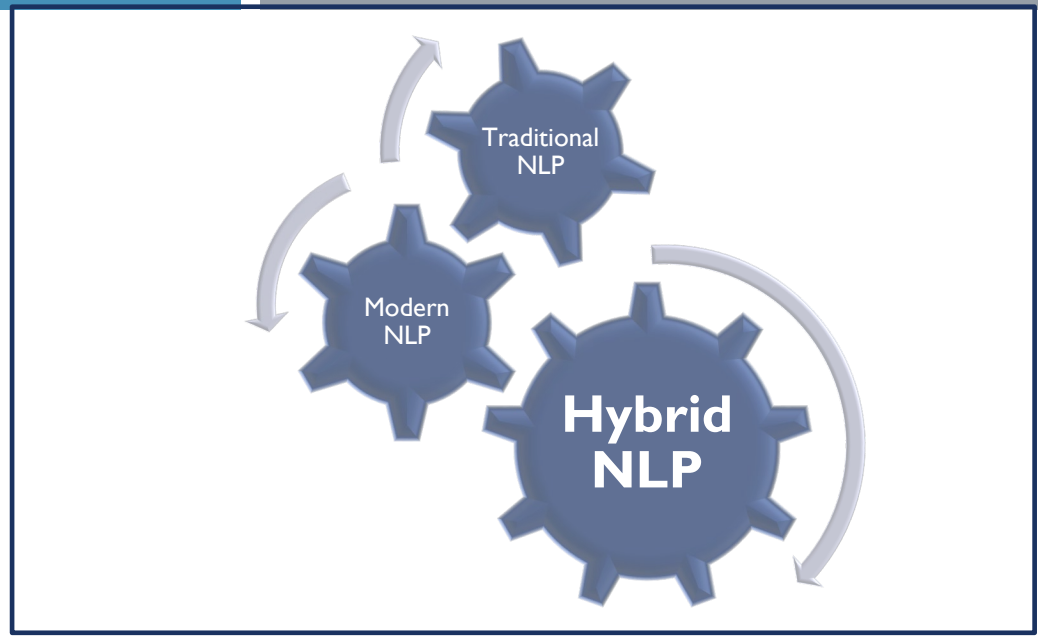
Menger,V., Scheepers,F., & Spruit,M. (2018). Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text. *Applied Sciences, 8*(6), Data Analytics in Smart Healthcare, 981. [JIF: 2.679]pdf] [online]

- In previous work, we determined SVM as an appropriate classifier for VRA, based on literature and experiments

Table 4. The performance for optimal hyperparameter values for each of the representations combined with the models, based on a 5-fold stratified cross validation. The performance is measured in AUC, along with its standard deviation. The best performance over different models is marked with an [a], the best performance over representations with a [b].

Model	Bag-of-Words Binary	Bag-of-Words tf-idf	Word Embeddings	Document Embeddings
RNN [1]	0.771 ± 0.018 [b]	0.753 ± 0.031	0.654 ± 0.043	0.788 ± 0.018 [a,b]
CNN [2]	0.729 ± 0.030	0.716 ± 0.038	0.684 ± 0.038	0.763 ± 0.024 [a]
NN [3]	0.727 ± 0.033	0.717 ± 0.038	0.751 ± 0.036 [a]	0.745 ± 0.022
NB [4]	0.686 ± 0.026	0.704 ± 0.034 [a]	0.700 ± 0.051	0.692 ± 0.046
SVM [5]	0.759 ± 0.040	0.756 ± 0.036 [b]	0.764 ± 0.024 [b]	0.770 ± 0.029 [a]
DT [6]	0.727 ± 0.018 [a]	0.719 ± 0.041	0.685 ± 0.041	0.665 ± 0.035

[1] Recurrent Neural Network; [2] Convolutional Neural Network; [3] Neural Network; [4] Naive Bayes; [5] Support Vector Machine; [6] Decision Tree.



"*Effective*" NLP → Hybrid NLP

Combining traditional and modern approaches (*e.g.* ADRIN)

EXAMPLE #5: ADR IDENTIFICATION, REVISITED

Siegersma,K., Evers,M., Bots,S., Groepenhoff,F., Appelman,Y., Hofstra,L., Tulevski,I., Somsen,A., Den Ruijter,H., Spruit,M., & Onland-Moret,C. (2022). Adverse Drug Reactions Identification in clinical Notes (ADRIN): Word embedding models and string matching. *JMIR Medical Informatics, 10*(1), e31063. [JIF: 2.96] [pdf] [online]

- *Hypothesis:* Information on ADRs is present in clinical notes, but is underreported in EHRs
- *Goal:* Method for recognising ADRs in Dutch clinical notes
- *Method:* Case study, using clinical notes from the UMCU for developing a prototype that automatically identifies ADRs in clinical notes
- *Dataset:* Cardiology Centre Netherlands (CCN)
  - 109.151 patients between 2007-2018
  - 277.389 unique clinical notes
  - In 36.533 clinical notes, why medication is stopped
    - 1.556 notes where the doctor noted an ADR
- Manual labelling
  - 3.156 clinical notes: validation; 1.000 notes: test

- What *is* a Dutch CCN clinical note ?

Hartfrequentie over het algemeen te hoog met 90 gemiddeld. Wat nu ook opvalt is dat de nierfunctie in korte tijd is verslechterd naar een klaring van 20 ml.\nDerhalve wordt de medicatie aangepast.

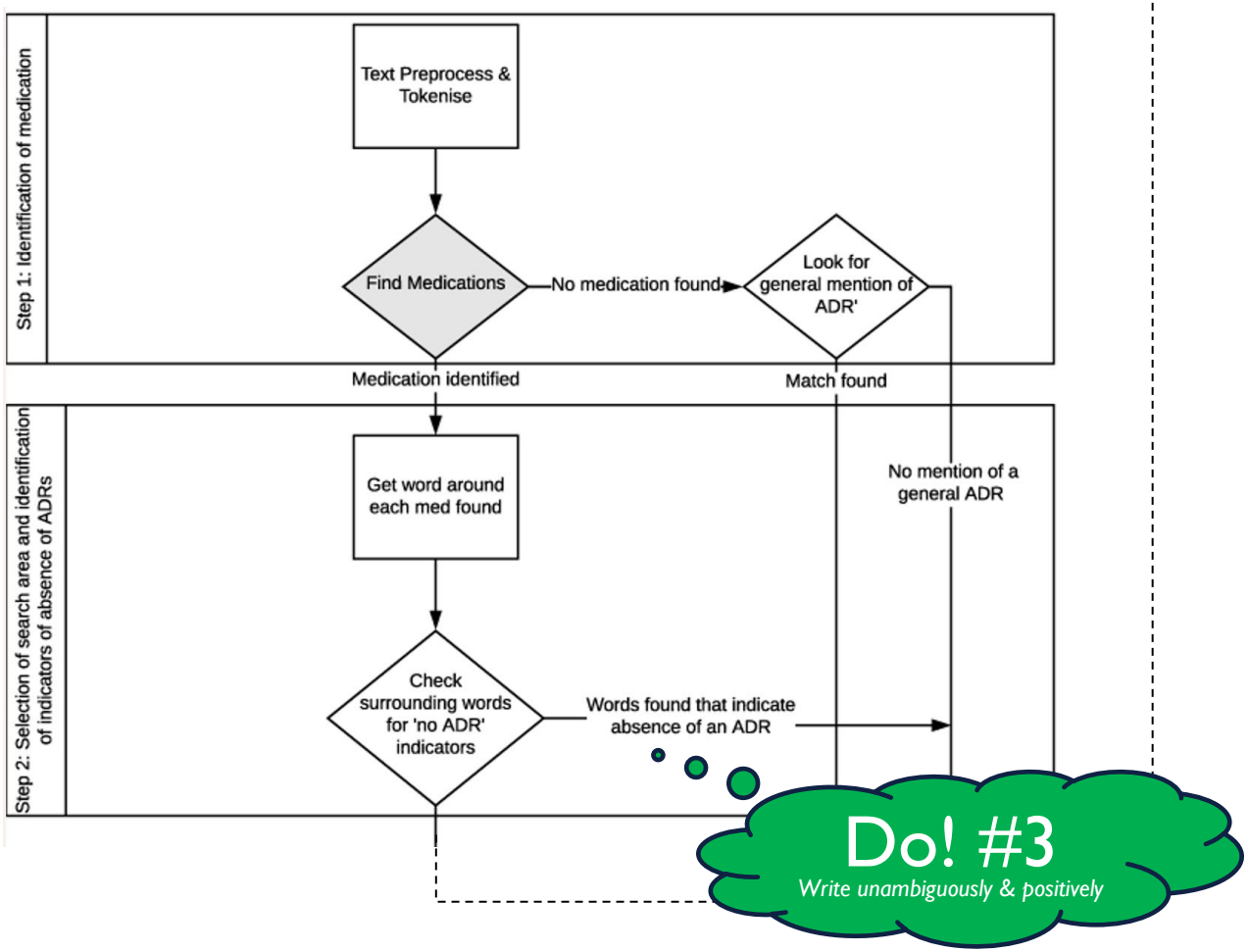
Met de patient gaat het goed. Verdraagt Monocedodard niet. Heeft afgelopen maand een keer druk op de borst gehad, in rust. Inspanning (lopen) gaat goed, 2 kg afgevallen.\n\nBeleid:\nContinueren\nControle over 6 maanden

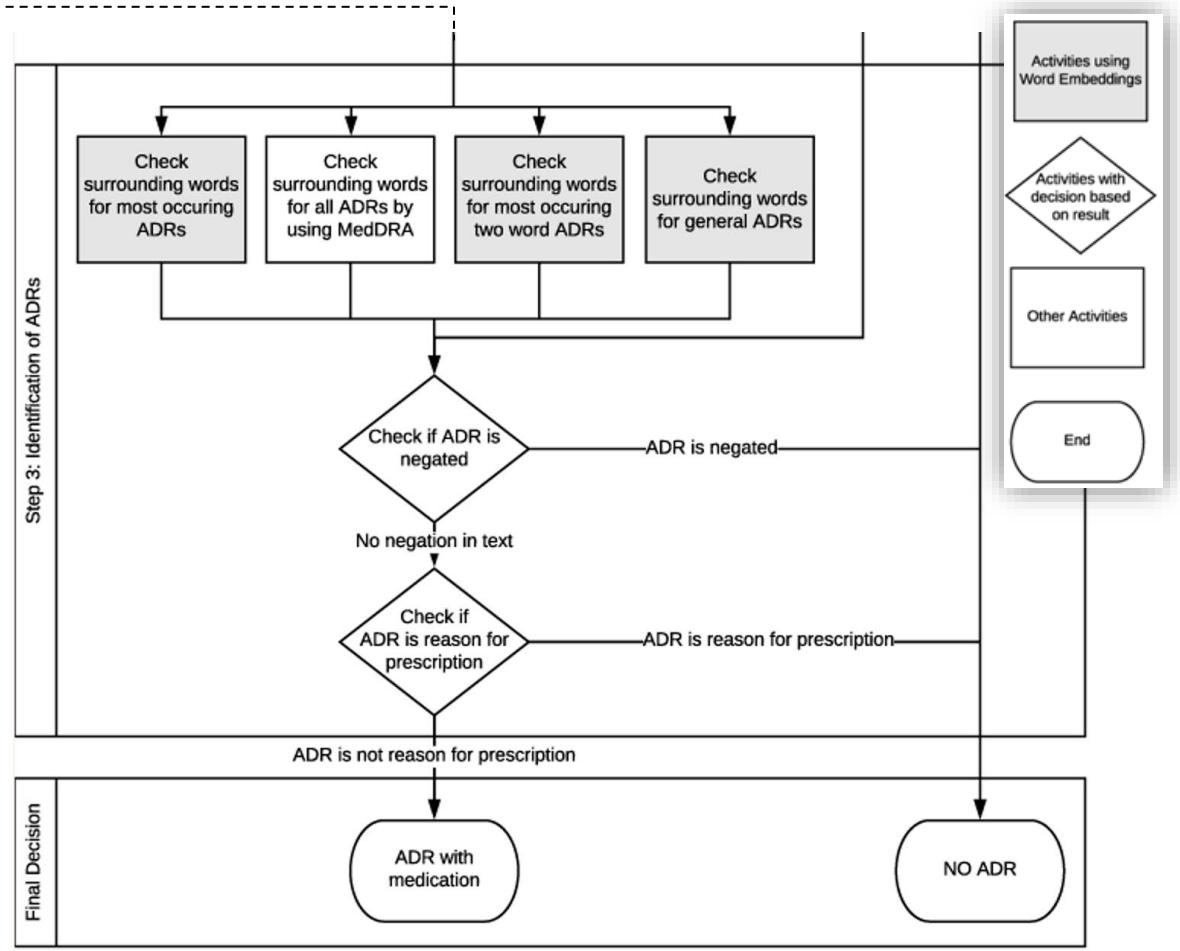
Bloeddruk blijft goed na halveren Olmetec. Echter zeer forse spierkrampen met name van de kuiten en fors haaruitval. Mogelijk bijwerking van de medicatie?\nB/ stop metoprolol plus controle 2 weken.

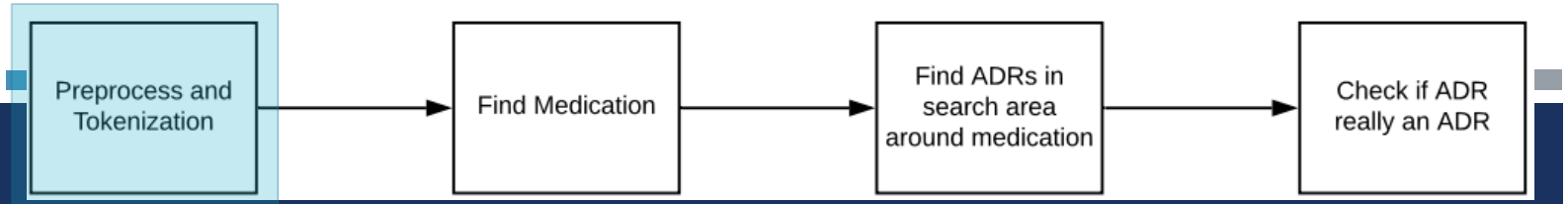
EXAMPLE #5: ADR IDENTIFICATION - METHOD

Adverse Drug Reactions Identification in clinical Notes (ADRIN)

Siegersma,K., Evers,M., Bots,S., Groepenhoff,F., Appelman,Y., Hofstra,L., Tulevski,I., Somsen,A., Den Ruijter,H., Spruit,M., & Onland-Moret,C. (2022). Adverse Drug Reactions Identification in clinical Notes (ADRIN): Word embedding models and string matching. *JMIR Medical Informatics, 10*(1), e31063. [JIF: 2.96] [pdf] [online]

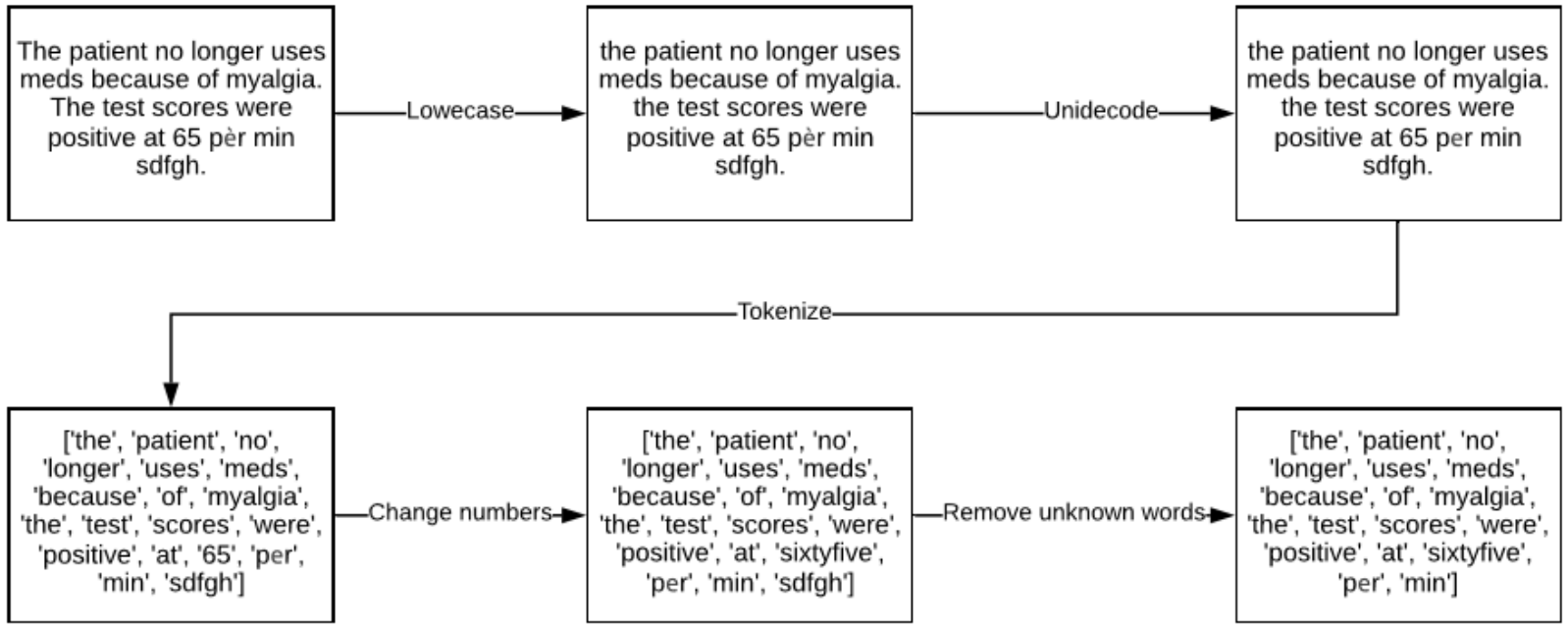


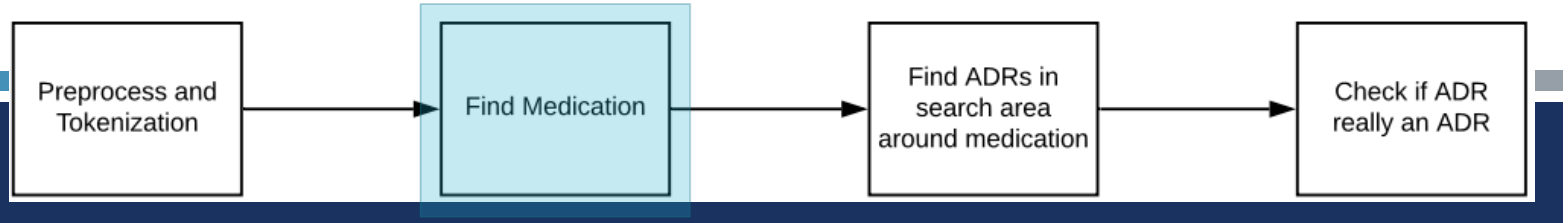




EXAMPLE #5: ADR IDENTIFICATION – METHOD – STEP 1

1. Lowercase
2. Unidecode
3. Tokenize
4. Change numbers
5. Remove unknown words



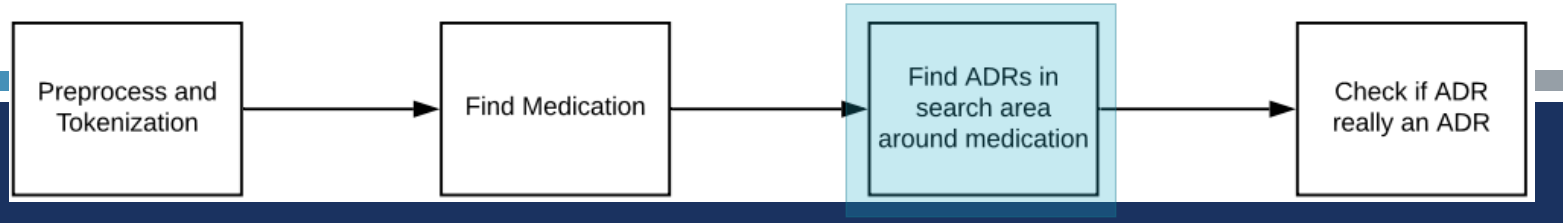


EXAMPLE #5: ADR IDENTIFICATION - METHOD – STEP 2

- *Idea:* Words that have similar neighbouring words are similarly shaped
  - Every word is represented in a numerical vector
  - Trained on all 277.389 clinical notes
- Because models are trained on domain specific text, domain specific results →
- The **Word2Vec** approach is used, *i.e.* vectors are shaped based upon their neighbouring words
  - "King – Man + Woman = Queen" →
  - **"Patient – Man + Woman = Patiente"**

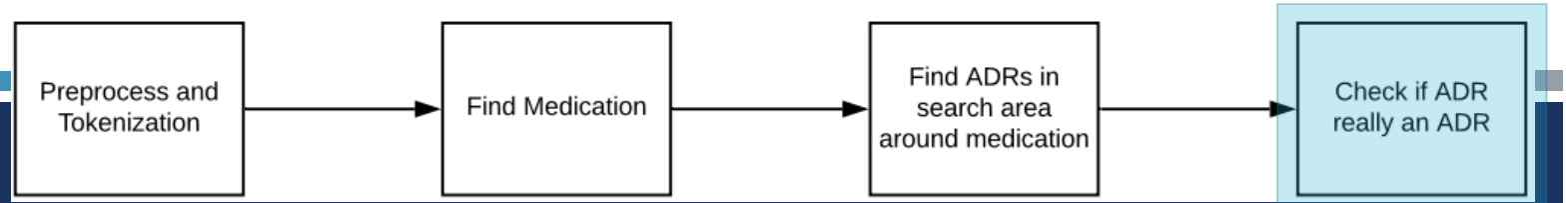
```
In [11]: model_all.wv.most_similar(['rood'])

Out[11]: [('jeukend', 0.7554248571395874),
          ('opgezwollen', 0.7433047890663147),
          ('jeukende', 0.7421249151229858),
          ('gezwollen', 0.7398363351821899),
          ('geirriteerd', 0.738010048866272),
          ('verkleuringen', 0.7336500287055969),
          ('verkleuring', 0.7277984023094177),
          ('zere', 0.7272820472717285),
          ('paars', 0.7250189185142517),
          ('verkleurd', 0.7249985933303833)]
```

```
In [79]: model_all.wv.most_similar(positive=['patient', 'vrouw'], negative=['man'], topn = 1)

Out[79]: [('patiente', 0.8561874032020569)]
```



EXAMPLE #5: ADR IDENTIFICATION - METHOD – STEP 3

- For ADR recognition with the word embedding models, for every word in the search area the similarity with predefined search words is computed.
- The predefined search words consist of the most occurring ADRs
- If the similarity is above a certain threshold, the word is identified as an ADR
- For the **MedDRA** search, a *Regular Expression* for each term in MedDRA is executed to see if it occurs in the search area
- Then, Check if ADRs *are* indeed ADRs
- When the prototype has found one or more [medication, ADR] combinations, two things are checked:
  1. Is the ADR not negated
  2. Is the ADR not the reason for prescription of medication



EXAMPLE #5: ADR IDENTIFICATION - METHOD – STEP 4

- *"The betablocker causes headaches and muscle pain. Because of these complaints the decision is made to stop bisoprolol and metoprolol."*

1. ['the', 'betablocker', 'causes', 'headaches', 'and', 'muscle', 'pain', 'because', 'of', 'these', 'complaints', 'the', 'decision', 'is', 'made', 'to', 'stop', 'bisoprolol', 'and', 'metoprolol']
2. ['the', '**betablocker**', 'causes', 'headaches', 'and', 'muscle', 'pain', 'because', 'of', 'these', 'complaints', 'the', 'decision', 'is', 'made', 'to', 'stop', '**bisoprolol**', 'and', '**metoprolol**']
3. Search area of 5 words before and 5 words after medication

   betablocker, ['the', 'betablocker', 'causes', 'headaches', 'and', 'muscle', 'pain',]

   bisoprolol, ['decision', 'is', 'made', 'to', 'stop', 'bisoprolol', 'and', 'metoprolol']

   metoprolol, ['made', 'to', 'stop', 'bisoprolol', 'and', 'metoprolol']

   → [betablocker, headaches]

   → [betablocker, muscle pain]
4. None of the ADRs are negated or the reason for prescription

EXAMPLE #5: ADR IDENTIFICATION - EVALUATION

Siegersma,K., Evers,M., Bots,S., Groepenhoff,F., Appelman,Y., Hofstra,L., Tulevski,I., Somsen,A., Den Ruijter,H., Spruit,M., & Onland-Moret,C. (2022). Adverse Drug Reactions Identification in clinical Notes (ADRIN): Word embedding models and string matching. *JMIR Medical Informatics, 10*(1), e31063. [JIF: 2.96] [pdf] [online]

- *Computational experiment:* Evaluate different versions of the prototype by varying the search area size that is used for identifying ADRs

Version	Search area	Sentences?	MedDRA?
1	All	No	Yes
2	10	No	Yes
3	5	No	Yes
4	All	Yes	Yes
5	10	Yes	Yes
5b	10	Yes	No
6	5	Yes	Yes

No MedDRA…

- *Evaluation tasks:* Four tasks for prototype evaluation:

1. Predict if a text contains one or more ADRs or none
   - 5b: F-score = 0.71
2. Find all present **[medication, ADR]** combinations
   - 5b: F-score = 0.59
3. Find all present ADRs, regardless of medication
   - 2: F-score = 0.67 (5b: 0.66)
4. Find all medication that triggers an ADR, regardless of which ADR
   - 5b: F-score = 0.69

EXAMPLE #6: SMOKING STATUS DETECTION

- *Title:* NLP for lifestyle extraction from discharge papers
- *Goal:* Distinguishing current smokers from past smokers
- *Dataset:* Haga Hospital, The Hague
  - Gathered from CTCue by searching for word "roken"
  - "Labels" gathered from field filled-in by med. specialist
    - "Mevrouw heeft een jaar geleden voor het laatst gerook"
    - "Rookt 30PY" or "30py"
- Some of the Challenges
  - *Subtle distinction:* small difference in words and context between smokers and past smokers
  - *Large documents* -> model needs to find the correct sentence that contains the needed information

- What *are* Haga discharge letter texts?

- "belang **stoppen** met **roken** besproken b/ 13-11 cap gas 14-11"
- "ologie intoxicaties: **roken**: >40j 10-15 sigaretten **per dag**, a"
- "enadviesbureau intox **roken** ja (**was gestopt**, nu met corona we"
- "s 2 jaar **gestopt** met **roken** (meer dan 20 **packyears**), geen alc"

- "lmatig last van zuurbranden, gebruikt ppi vooral bij gekruid eten, probeert dit te **minderen** intox: **roken**: **vroeger** na de koffie 1 e-sigaret. 15 py alc: sociaal: fa: allergie: aanvullend onderzoe"
  - past smoker?
- "eer is het goed. traplopen gaat goed. bij traplopen kort van adem. patient is aan het **minderen** met **roken**. allen bij opstaan hoesten. (bij weinig roken, dan ook minder last in de ochtedn). kno: geen"
  - smoker?

EXAMPLE #6: SMOKING STATUS DETECTION

- Manually determine rules to label, for example:
  - "roken+" = smoker
  - "gerookt" = past smoker
  - "roken" + "stoppen" = smoker
  - "gestopt" + !"was" = past smoker
  - "gestopt" + "was" = smoker

Range	Accuracy	Range	Accuracy
10/10	0.56	10/20	0.59
20/20	**0.60**	40/20	0.57
40/40	0.58	20/40	**0.60**
60/60	0.57	10/40	**0.60**
80/80	0.57	20/80	0.59
100/100	0.53	20/100	0.55

Table 2: Results for the exact string matching approach

- Haga discharge letters dataset characteristics

Total number of discharge papers	6560
Total number of patients	480
Currently smoking patients discharge papers	3861
Ever smoking patients discharge papers	2699
Currently smoking patients	254
Ever smoking patients	226
Number of sentences	221,349
Number of tokens/words	2,651,460
Number of unique tokens/words	115,645
Average token/word length	6.01
Average unique token/word length	9.42
Average sentence length	85.14

autoscriber

Do! #4
Go with the AI flow!

Recording of the consultation:

0:29 -2:30

Explore and edit extracted symptoms, properties and SNOMED codes below.

Once completed, click here to view and export your summary.

Generate summary

(Quickstart Tip: Skip the edits and edit the preliminary summary directly)

View and edit extracted symptoms

pijn op de borst	AS: Present	Highlight Filter	Remove
pijn	AS: Present	Highlight Filter	Remove
straalt	AS: Present	Highlight Filter	Remove
koorts	AS: Not_experience	Highlight Filter	Remove
hoesten	AS: Not_experience	Highlight Filter	Remove

undo

Transcript

Reset

00:01
Ja goedendag. Wat scheelt eraan?

00:04
Goedendag dokter. Ik ja ik heb wat last, pijn op de borst heb ik.

00:10
Oké, en hoe lang is dat er al?

00:13
[eh] nou dat is sinds gisteren.

00:16
Gisteren. Oké. Ooit eerder zoiets gehad?

00:19
Nee nee. Ja ik ben jong, nooit eerder gehad, nee.

00:24
Oké, en kan je een soort naam aan die pijn geven? Wat voor type pijn is het?

Do's and Don'ts: Recapitulation

Don'ts

1. Typing errors
2. Custom abbreviatns.
3. Nonstandardised, complex reporting

Do's

1. Standardise reporting
2. Discuss reporting guidelines
3. Write unambiguously & positively
4. Go with the AI flow!

What else??

```
>>> import stanza
>>> stanza.download('en')       # This downloads the English models for the neural pipeline
>>> nlp = stanza.Pipeline('en') # This sets up a default neural pipeline in English
>>> doc = nlp("Barack Obama was born in Hawaii.  He was elected president in 2008.")
>>> doc.sentences[0].print_dependencies()
```

THANK YOU FOR LISTENING

m.r.spruit@lumc.nl

https://www.universiteitleiden.nl/en/staffmembers/marco-spruit

Leiden University Medical Center

liacs Leiden Institute of Advanced Computer Science

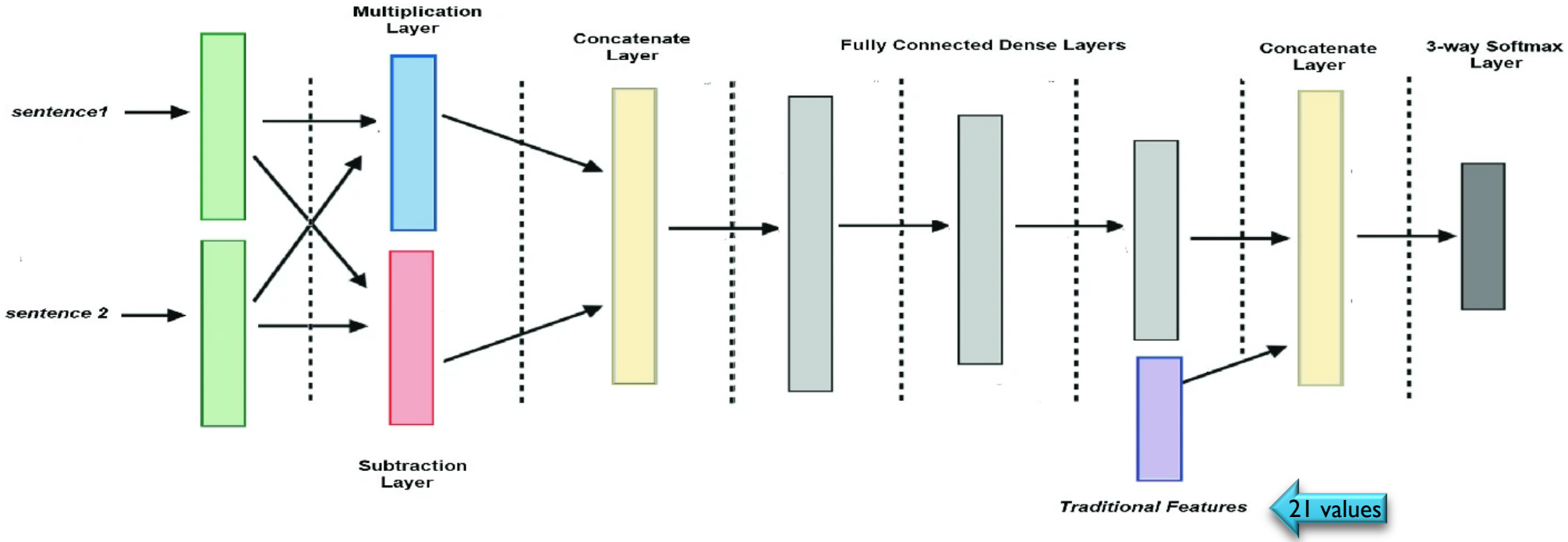
Universiteit Leiden

Leiden University Campus The Hague

EXAMPLE #7: SIMPLE INTEGRATION

Tawfik,N., & Spruit,M. (2019). Towards Recognition of Textual Entailment in the Biomedical Domain. In Métais, E. et al. (Eds.), *Lecture Notes in Computer Science 11608, NLDB 2019: International Conference on Applications of Natural Language to Information Systems* (pp. 368–375). NLDB 2019, University of Salford, MediaCityUK Campus, United Kingdom, 26–28 June 2019: Springer. [pdf] [online]

- *Computational experiment:* A *feature-assisted* neural network architecture for a Natural Language Inference (NLI) task.



String-Based Features (e.g. editDist)
Contradiction-Based Features (e.g. Negation)
Context-Based Features (e.g. embedSim)

Symbolic NLP	Probabilistic NLP	Hybrid NLP: Combined	Hybrid NLP: Integrated
Menger,V., Scheepers,F., Wijk,L. van, & Spruit,M. (2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics, 35*(4), Patient Centric Healthcare, 727–736. [JIF: 6.182] [pdf] [online]	Menger,V., Spruit,M., Est,R. van, Nap,E., & Scheepers,F. (2019). Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Network Open, 2*(7), e196709. [JIF: 8.483] [pdf] [online]	Siegersma,K., Evers,M., Bots,S., Groepenhoff,F., Appelman,Y., Hofstra,L., Tulevski,I., Somsen,A., Den Ruijter,H., Spruit,M., & Onland-Moret,C. (2022). Adverse Drug Reactions Identification in clinical Notes (ADRIN): Word embedding models and string matching. *JMIR Medical Informatics, 10*(1), e31063. [JIF: 2.96] [pdf] [online]	Tawfik,N., & Spruit,M. (2019). Towards Recognition of Textual Entailment in the Biomedical Domain. In Métais, E. et al. (Eds.), *Lecture Notes in Computer Science 11608, NLDB 2019: International Conference on Applications of Natural Language to Information Systems* (pp. 368–375). NLDB 2019: Springer. [pdf] [online]
Shen,Z., & Spruit,M. (2021). Automatic Extraction of Adverse Drug Reactions from Summary of Product Characteristics. *Applied Sciences, 11*(6), Applications of Artificial Intelligence in Pharmaceutics, 2663. [JIF: 2.679] [pdf] [online]	Rijcken,E., Kaymak,U., Scheepers,F., Mosteiro,P., Zervanou,K., & Spruit,M. (2022). Topic Modeling for Interpretable Text Classification from EHRs. *Frontiers in Big Data, 5*, Section Data Mining and Management, 846930. [pdf] [online]	Spruit,M., Verkleij,S., Schepper,C. de, & Scheepers,F. (2022). Exploring Language Markers of Mental Health in Psychiatric Stories. *Applied Sciences, 12*(4), Current Approaches and Applications in Natural Language Processing, 2179. [JIF: 2.679] [pdf] [online]	Zhou,J., Zhang,Z., Zhao,H., and Zhang,S. (2020). LIMIT-BERT : Linguistics Informed Multi-Task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, ACL. <? DEDUCE v2 (in progress) ?>