Natural Language Processing

… from a Translational Data Science Perspective in Dutch Healthcare

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

Lecture: *Taaldiagnostiek, 13 April 2022*

Leiden University Medical Center

liacs Leiden Institute of Advanced Computer Science

Universiteit Leiden

About… Marco Spruit

As Engineer

1993
- Information Retrieval programmer
  - ZyLAB Europe BV

1995
- Big Data system developer
  - Dutch Military Intelligence and Security Service

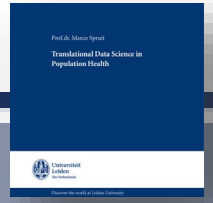
1997
- Product software developer/entrepreneur
  - Insertable Objects, Wizzer BV

As Researcher

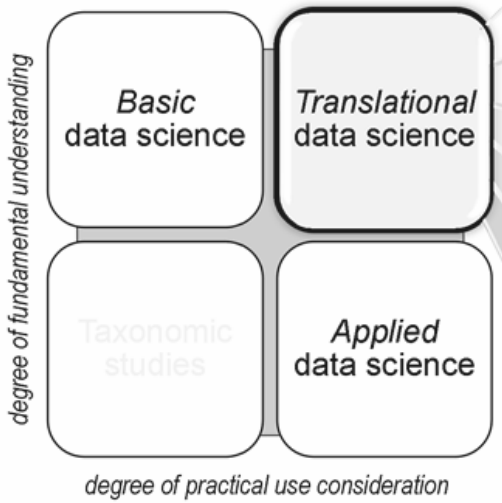
2003
- Ph.D. researcher in Computational Linguistics
  - University of Amsterdam

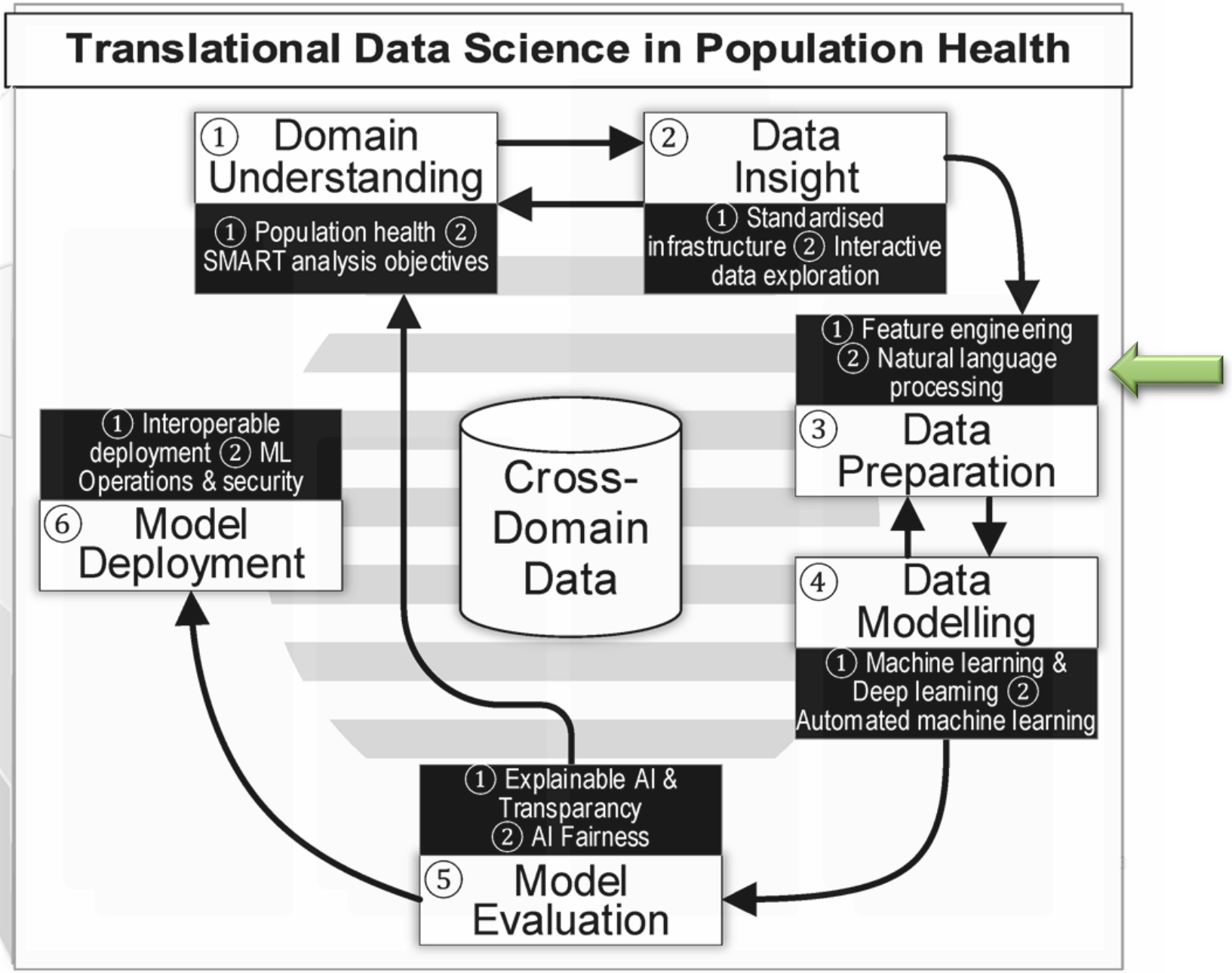
2007
- Assistant/Associate professor Information Science
  - Utrecht University >> Applied Data Science Lab

2020
- Professor Advanced Data Science in Population Health
  - LUMC/Leiden University
  - PH Living Lab, CAIRELab, TDS Lab, SIG Health Data Science



APRIL FOOLS' DAY



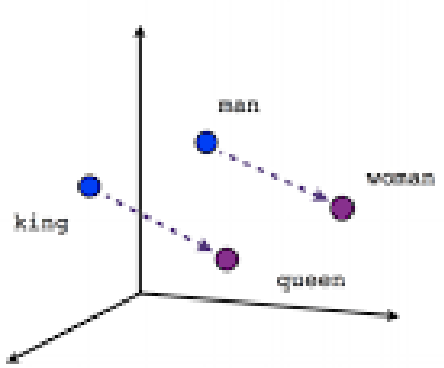


AGENDA

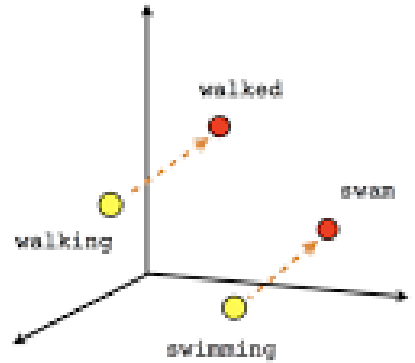
A. Setting the Scene

NLP 101: Symbolic NLP

Word embeddings: Probabilistic NLP



Male-Female



Verb tense

B. Case study

Natural Language Processing in Mental Healthcare

[0.341, -0.359, 0.7, 0.926, -0.004, ..., -0.129]

[Positive, Negative]

NLP 101

TEXT SIMILARITY → SYMBOLIC NLP

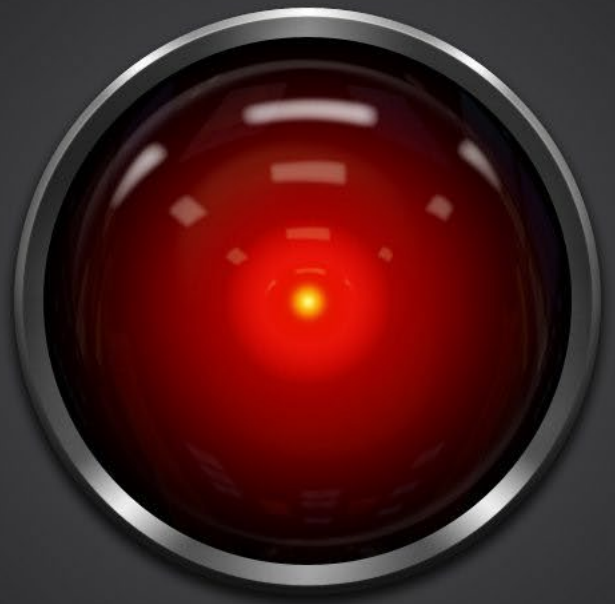
WORD EMBEDDINGS → PROBABILISTIC NLP

What is the difference between Linguistics and NLP?

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

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave. I'm afraid I can't do that.

WHERE IS THIS QUOTE FROM?

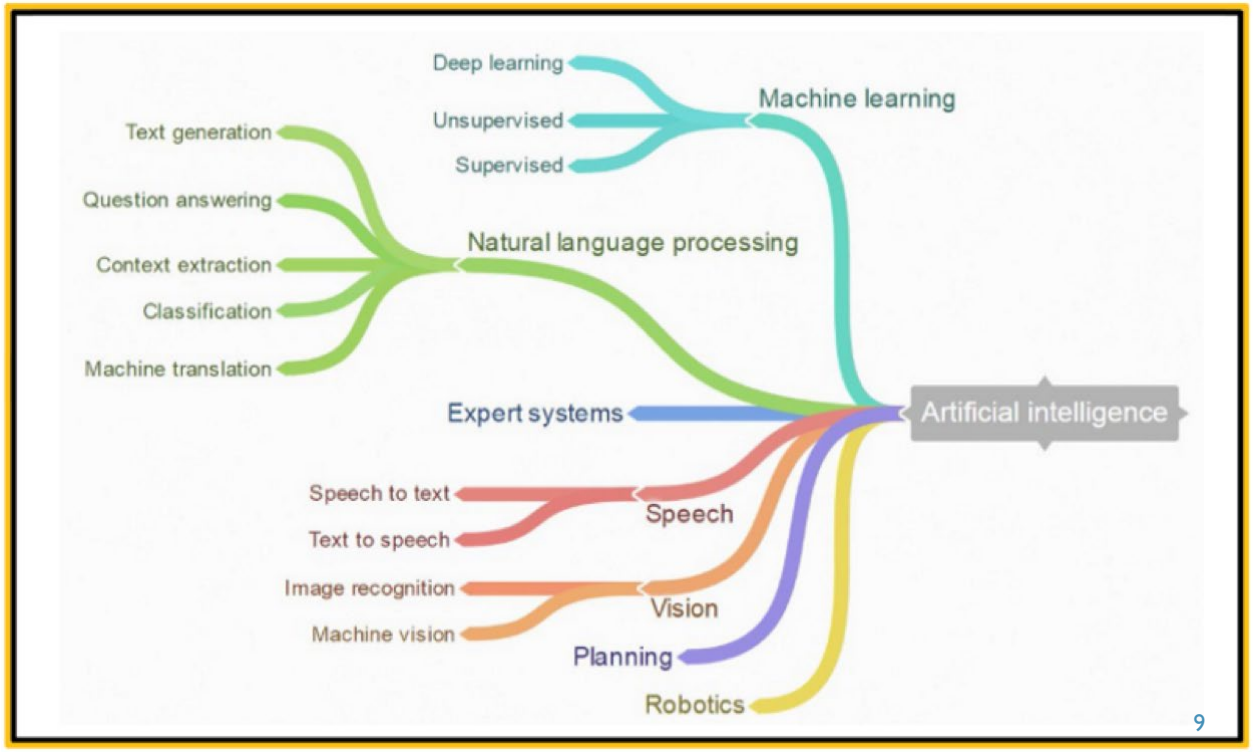


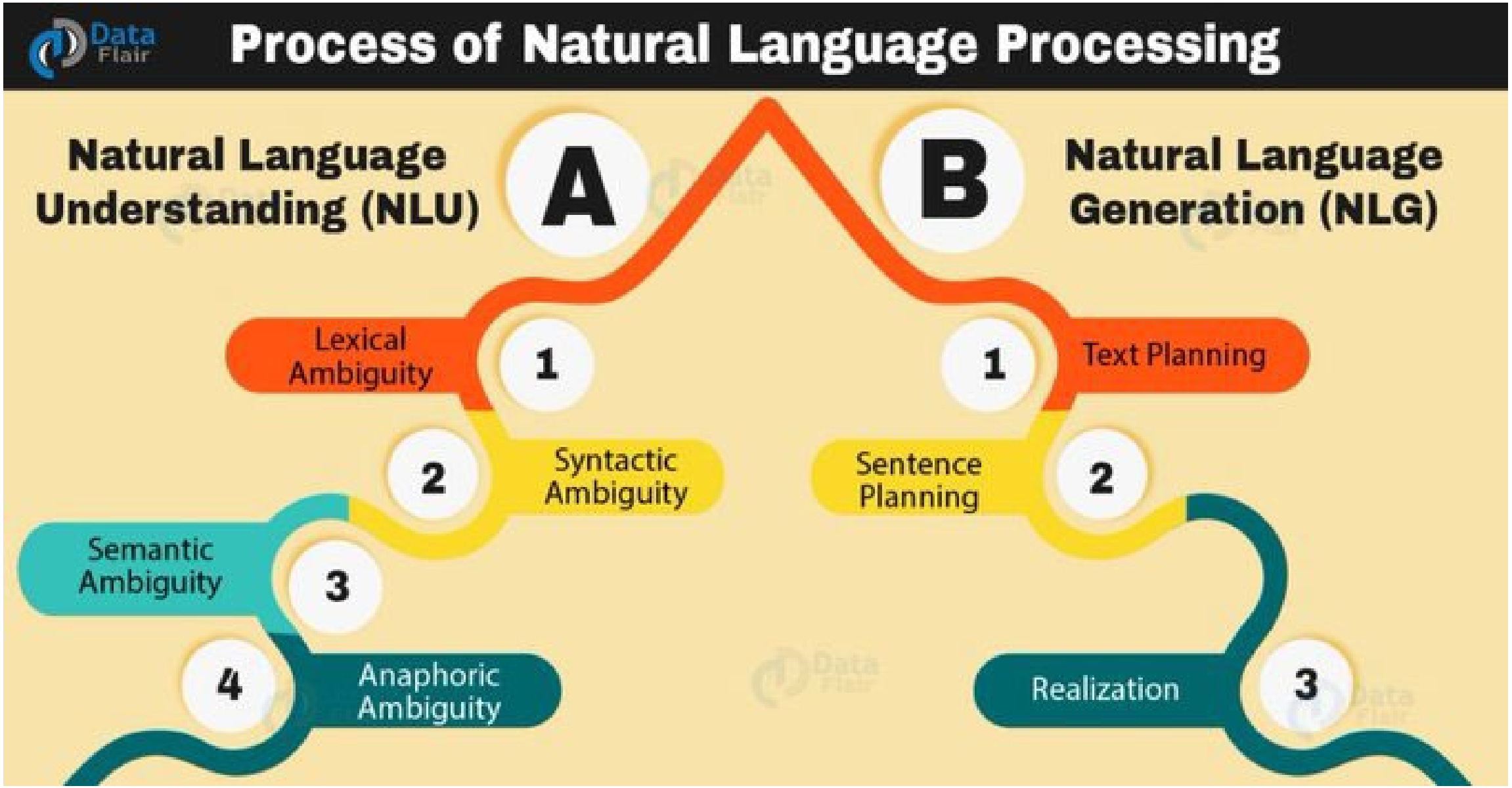
"2001: A Space Odyssey"

A 1968 film by Stanley Kubrick, based on a joint screenplay with Arthur C. Clarke.

A Multidisciplinary Wicked Problem

- 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.





https://twitter.com/DigitalTDaily/status/1480659757841145859

TEXT SIMILARITY

A FOUNDATIONAL CONCEPT IN SYMBOLIC NLP

Text Similarity and its Types

- People can express the same concept (or related concepts) in many different ways. For example,
  - "the plane leaves at 12pm" vs
  - "the flight departs at noon"
- Text similarity is a key component of Natural Language Processing
- If the user is looking for information about cats, we may want the NLP system to return documents that mention kittens even if the word "cat" is not in them.

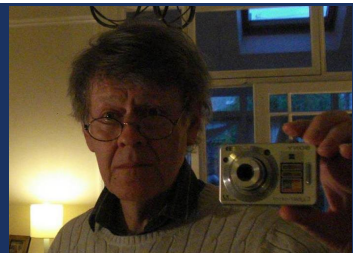
- Many types of text similarity exist:
  - **Morphological** similarity (*e.g.*, respect-respectful)
  - Spelling similarity (*e.g.*, theater-theatre)
  - Synonymy (*e.g.*, talkative-chatty)
  - Homophony (*e.g.*, raise-raze-rays)
  - Semantic similarity (*e.g.*, cat-tabby)
  - Sentence similarity (*e.g.*, paraphrases)
  - Document similarity (*e.g.*, two news stories on same event)
  - Cross-lingual similarity (*e.g.*, Dutch-Flemish-Afrikaans)

Morphological Similarity: The case of Stemming

- **Case:** Morphological Similarity
  - Words with the same root:
    - scan (base form)
    - scans, scanned, scanning (inflected forms)
    - scanner (derived forms, suffixes)
    - rescan (derived forms, prefixes)
    - rescanned (combinations)

- To **stem** a word is to reduce it to a base form, called the stem, after removing various suffixes and endings and, sometimes, performing some additional transformations
- Examples
  - scanned → scan
  - indication → indicate
- In practice, prefixes are sometimes preserved, so rescan will not be stemmed to scan

Porter's Stemming Method



- Porter's stemming method is a rule-based (i.e. symbolic) algorithm introduced by Martin Porter in 1980
- The input is an individual word. The word is then transformed in a series of steps to its stem
- The method is not always accurate
  - utilizes only suffix stripping, not addressing prefixes

An algorithm for suffix stripping
MF **Porter** - Program, **1980** - emerald.com
The automatic removal of suffixes from words in English is of particular interest in the field of information retrieval. An algorithm for suffix stripping is described, which has been ...
☆ Opslaan 99 Citeren Geciteerd door 12127 Verwante artikelen Alle 11 versies

Porter's Algorithm: Measure

- Example 1:
  - Input = computational
  - Output = comput
- Example 2:
  - Input = computer
  - Output = comput
- The two input words end up stemmed the same way
  - Stem is not (necessarily) the morphological root

- The *measure* of a word is an indication of the number of syllables in it
  - Each sequence of consonants is denoted by C
  - Each sequence of vowels is denoted as V
  - The initial C and the final V are optional
  - So, each word is represented as [C]VCVC ... [V], or **[C](VC){*m*}[V]**, where m is its measure

m = ? STREET?

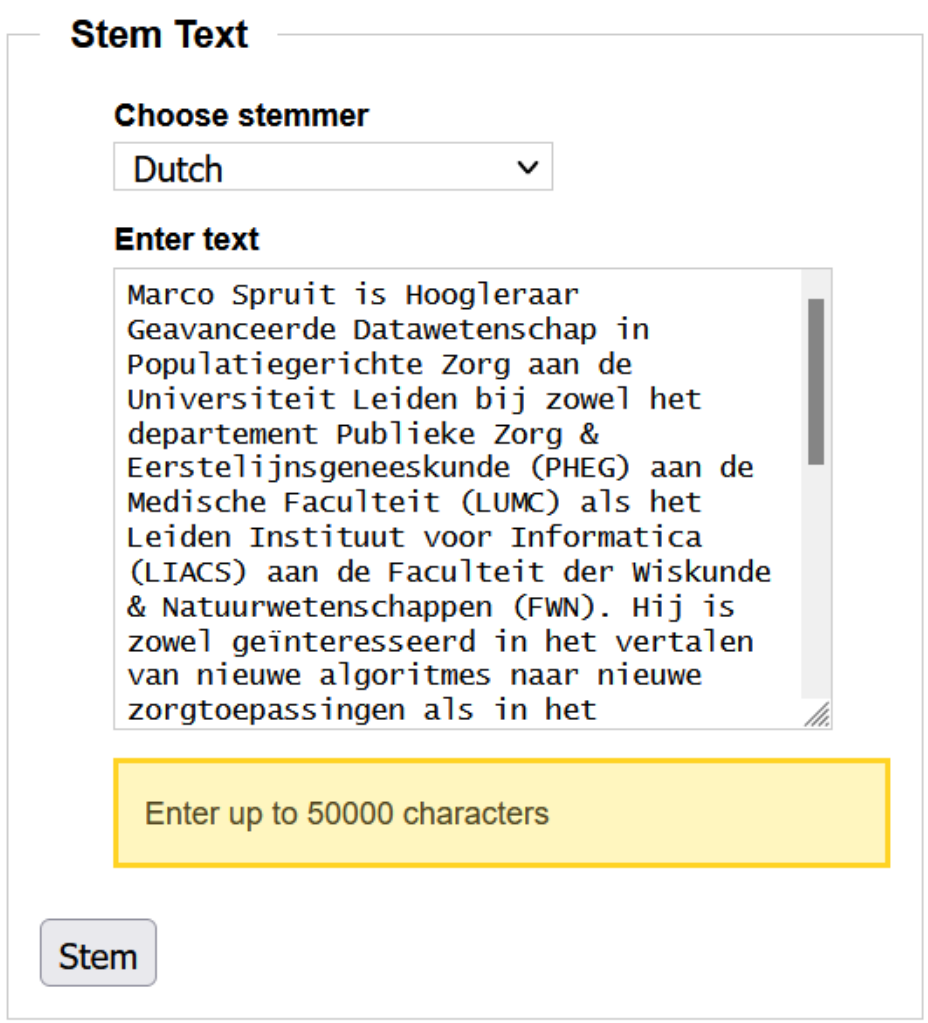
m = ? PRAGMATIC?

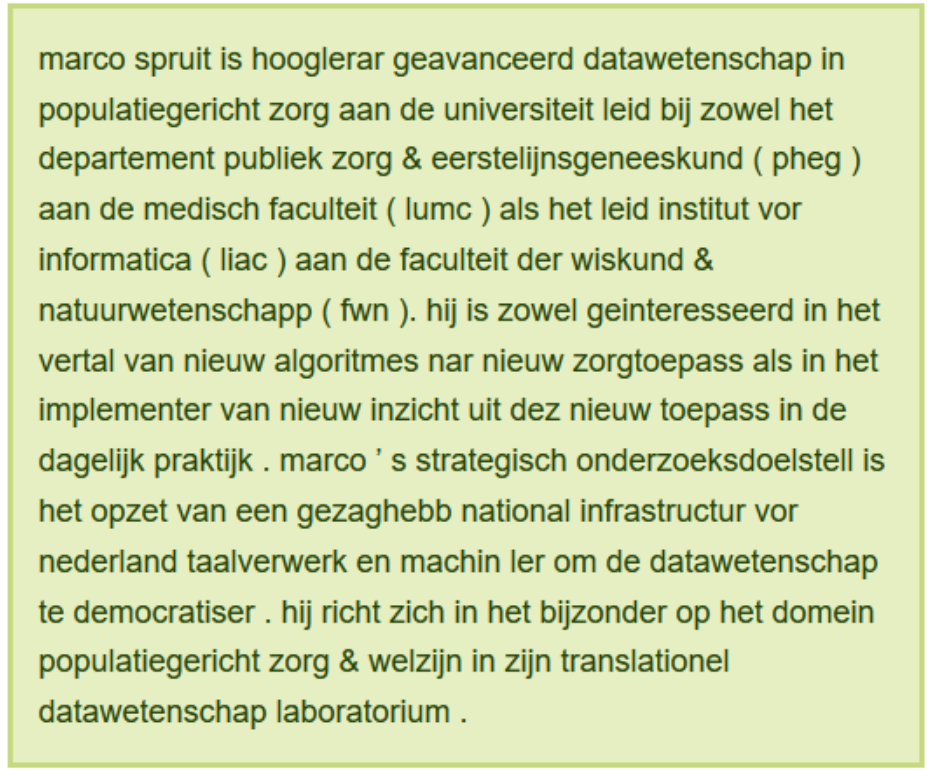
Porter's Algorithm: Transformation rules

- The initial word is then checked against a sequence of ~60 transformation patterns, in order.
- An example pattern is:
  - (m>0) ATION -> ATE (e.g. medication -> medicate)
- Whenever a pattern matches, the word is transformed and the algorithm restarts from the beginning of the list of patterns with the transformed word.
- If no pattern matches, the algorithm stops and outputs the most recently transformed version of the word.

- Example 1:
  - Input = computational
  - Step 2: replace ational with ate: computate
  - Step 4: replace ate with ø: comput
  - Output = comput
- • Example 2:
  - Input = computer
  - Step 4: replace er with ø: comput
  - Output = comput
- The two input words end up stemmed the same way...

ONLINE DEMO





- http://text-processing.com/demo/stem/

EXAMPLE #1 SYMBOLIC NLP: DEDUCE

- De-identification of Dutch medical text
- https://tdslab.nl/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 #2 SYMBOLIC NLP: 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	+

Word Embeddings

A FOUNDATIONAL CONCEPT FOR **TEXT REPRESENTATIONS** IN PROBABILISTIC NLP

Word Embeddings as Text Representations

- 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

- **word2vec** is a "predictive" model
  - Predictive models learn their vectors in order to improve their predictive ability of the loss of predicting the target words from the context words
    - capture co-occurrence one window at a time
- **GloVe** is a "count-based" model
  - Count-based models learn their vectors by doing dimensionality reduction on the co-occurrence counts matrix.
    - capture the counts of overall statistics how often it appears.

Word Embeddings: How and Why

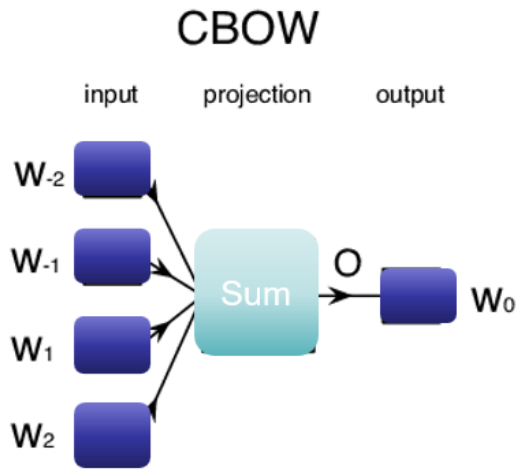
- Consider the following sentences:
  1. "Have a good day"
  2. "Have a great day"

  V={have, a, good, great, day}

- One-hot encoding: ("a 5 dimensional space")

have	1	0	0	0	0
a	0	1	0	0	0
good	0	0	1	0	0
great	0	0	0	1	0
day	0	0	0	0	1

- **Word-2-Vec** employs either of two methods (both involving Neural Networks):
  - *Continous Bag Of Words (CBOW):* takes the context of each word as the input and tries to predict the word corresponding to the context
  - *Skip Gram:* the inverse of CBOW



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

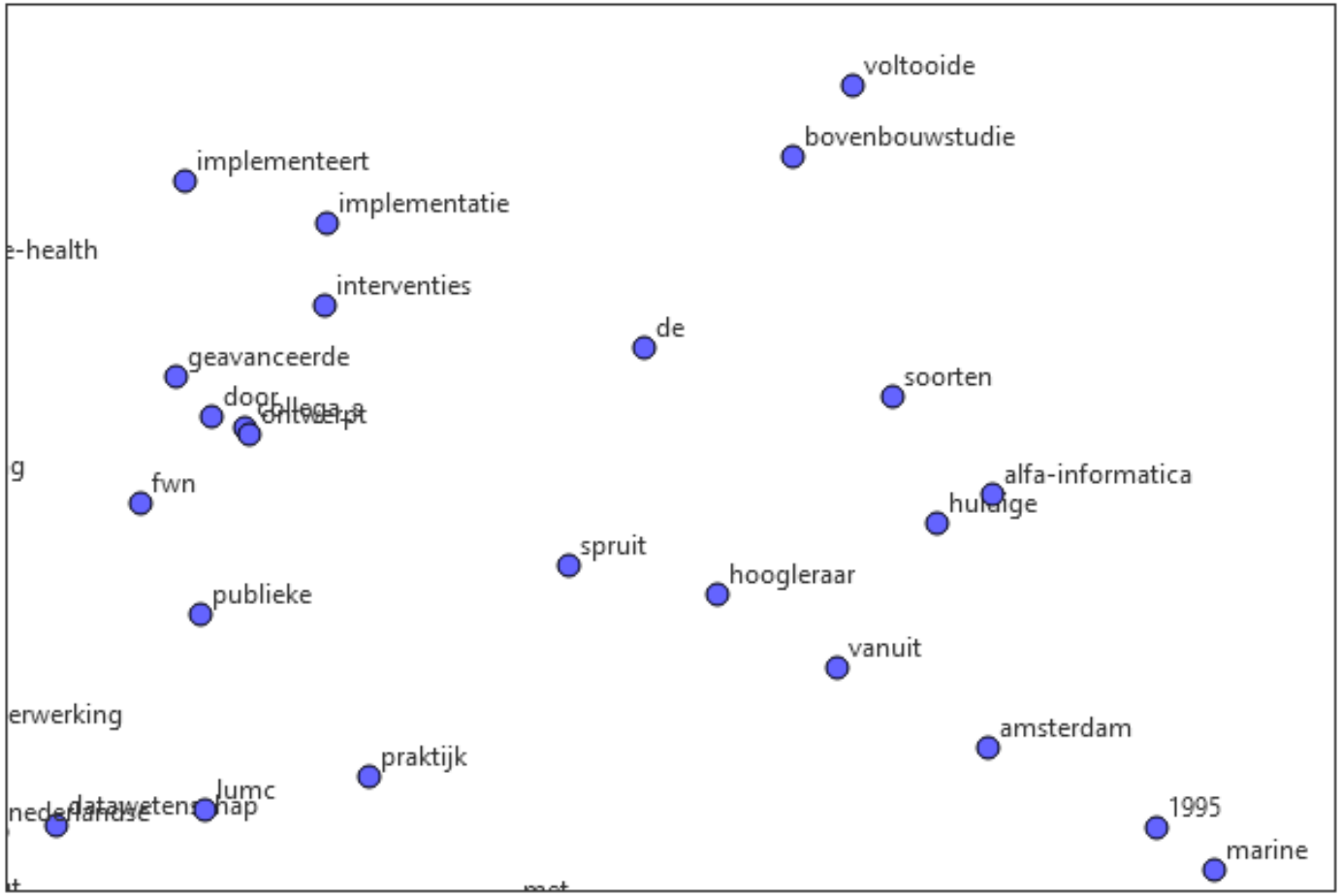
Training examples:

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

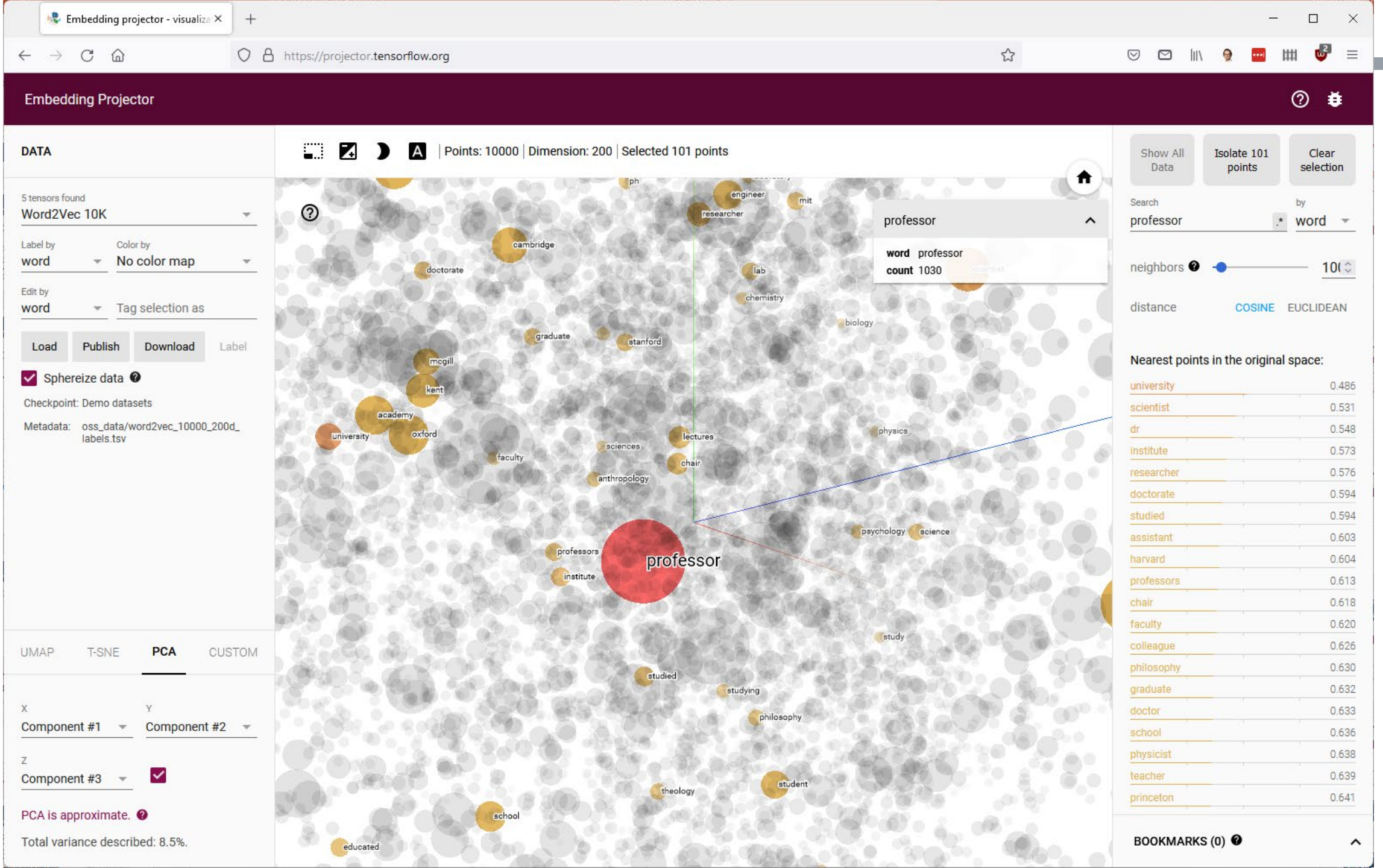
No. of training examples: 397

Word Embeddings as Text Representations

- After dimension reduction of the high-dimensional vector space:

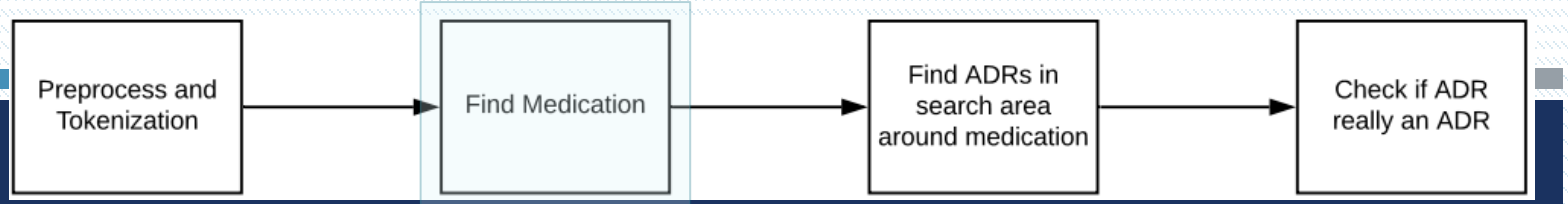


https://remykarem.github.io/word2vec-demo/



https://projector.tensorflow.org/

Siegersma (2022) *et al.*
Data Preparation: Find medication with Word Embedding Models



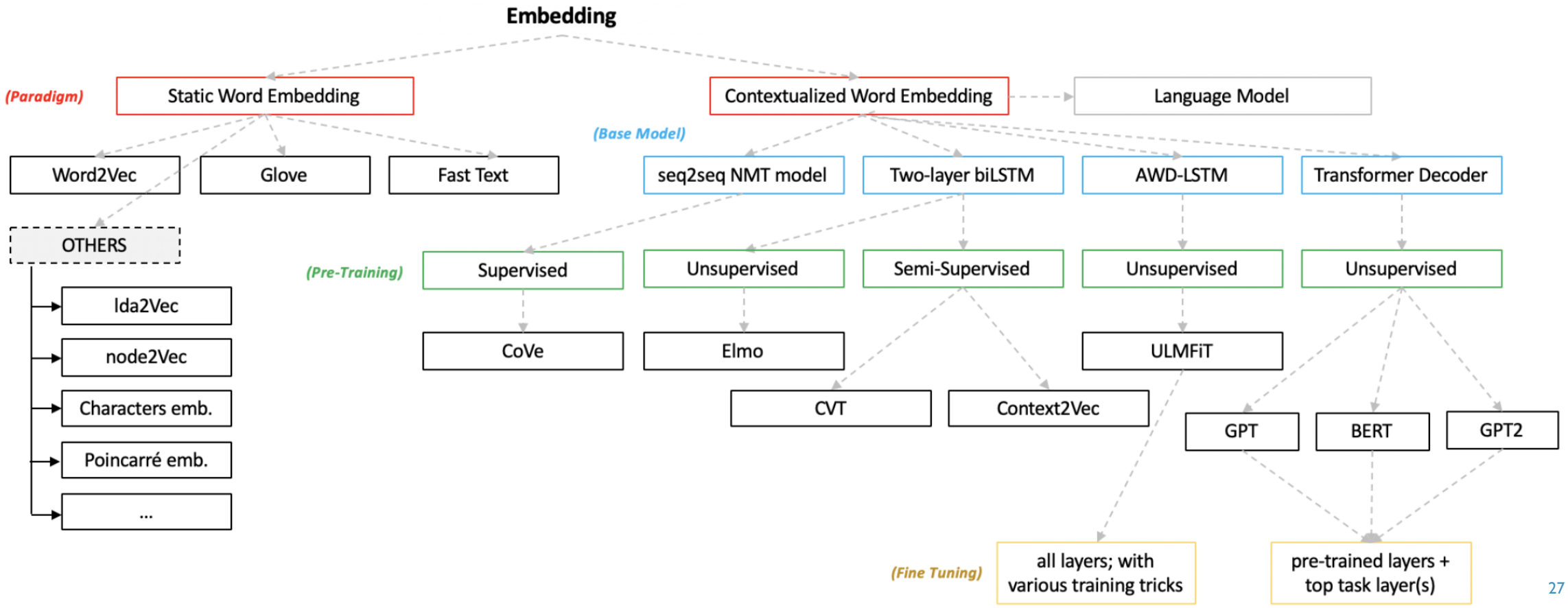
- *Idea:* Words that have similar neighbouring words are similarly shaped
  - Every word is represented in a numerical vector
  - Trained on all the 277.389 clinical notes
- The **Word2Vec** approach is used, which means that vectors are shaped based upon their neighbouring words
  - "King – Man + Woman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

```
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)]
```


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]

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



So… What can we do with Word Embeddings?...

Agenda

A. Setting the Scene

NLP 101: Symbolic NLP

Word embeddings: Probabilistic NLP

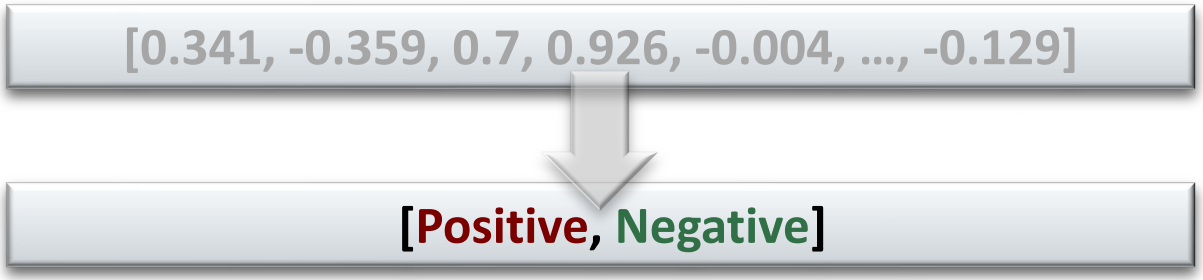
Male-Female Verb tense

B. Case study

Natural Language Processing in Mental Healthcare

[0.341, -0.359, 0.7, 0.926, -0.004, ..., -0.129]

[Positive, Negative]



PREDICTING INPATIENT VIOLENCE RISK WITH CLINICAL NOTES IN ELECTRONIC HEALTH RECORDS

CASE STUDY 2

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. [pdf] [online]

[1/6]
Domain Understanding: Objective

- "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"
  - Prediction task excludes incidents on Day 1 of admission
    - insufficient data available to make a prediction
  - 30 days interval chosen for sufficient specificity
    - majority of incidents included
    - mean duration of admission is 40.3 days
    - 81.9% of incidents happen during the first 30 days
- Area Under Curve (AUC) to report performance

[2/6] Data Understanding

- Site 1: UMC Utrecht
- Site 2: Antes, Parnassia Group, Rotterdam

Table 1. Descriptive Statistics of the Data Sets Obtained From the 2 Sites

Characteristic	No. (%) Site 1	Site 2
Demographic characteristics		
Age, mean (SD), y	34.0 (16.6)	45.9 (16.6)
Men	1536 (48.2)	2097 (64.5)
Data set		
Admissions, No.	3189	3253
Unique patients, No.	2209	1919
Length of stay, median (IQR), d	16.0 (6.0-41.0)	15.0 (5.0-40.5)
No. of words in notes, median (IQR)	2091 (1541-2981)	1961 (1160-3060)
Admissions with violent incidents	290 (9.1)	247 (7.7)
Incidents		
During admission, No.	962	652
During first 4 wk	658 (68.4)	318 (48.8)
During first 24 h	90 (9.4)	42 (6.4)
Staff Observation Aggression Scale-Revised score, median (IQR) [range]	12.0 (8.0-16.0) [2-21]	11.0 (7.0-14.0) [2-19]
Diagnostic and Statistical Manual		

[2/6]
DATA UNDERSTANDING

(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."

?

[2/6]
DATA UNDERSTANDING

(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."

[3/6] Data Preparation

Text representation

- Represent all clinical notes related to 1 admission as 1 vector (not words)

- *paragraph2vec*
- *SVM classifier*

[4/6]
MODELLING: PREDICTION PERFORMANCE

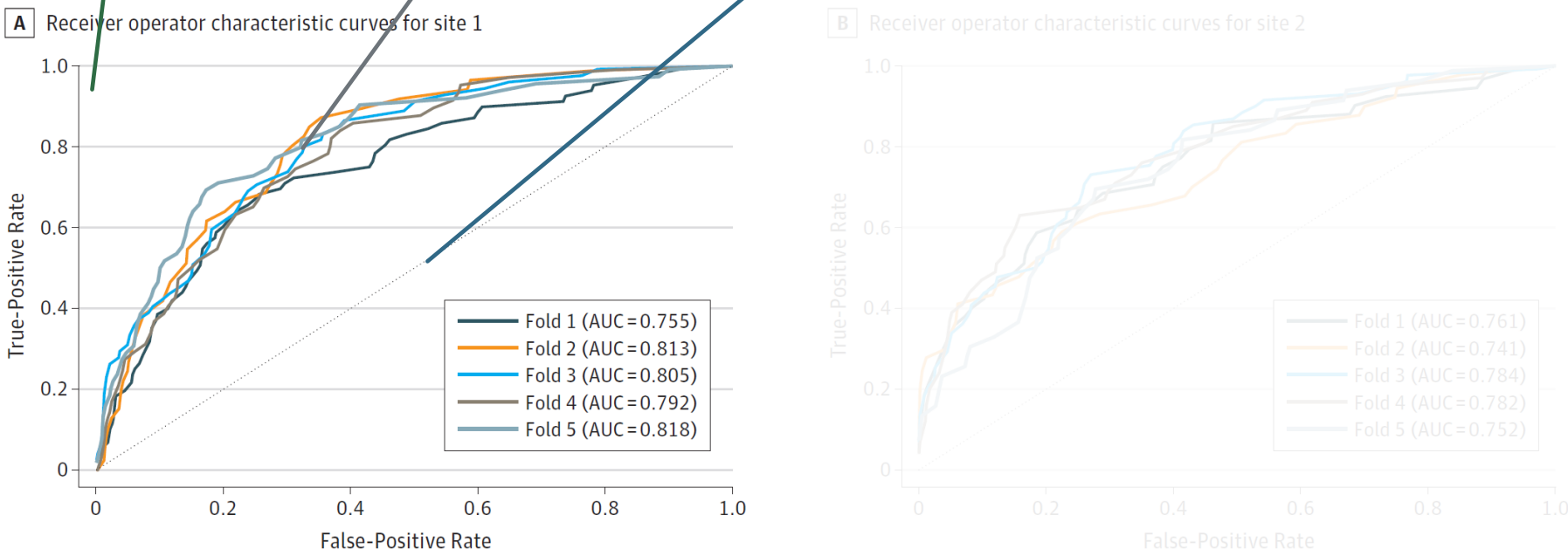
(SVM CLASSIFIER)

Perfect Score

Actual Score

Random Score

Figure. Receiver Operator Characteristic Curves for Internal Cross-validations



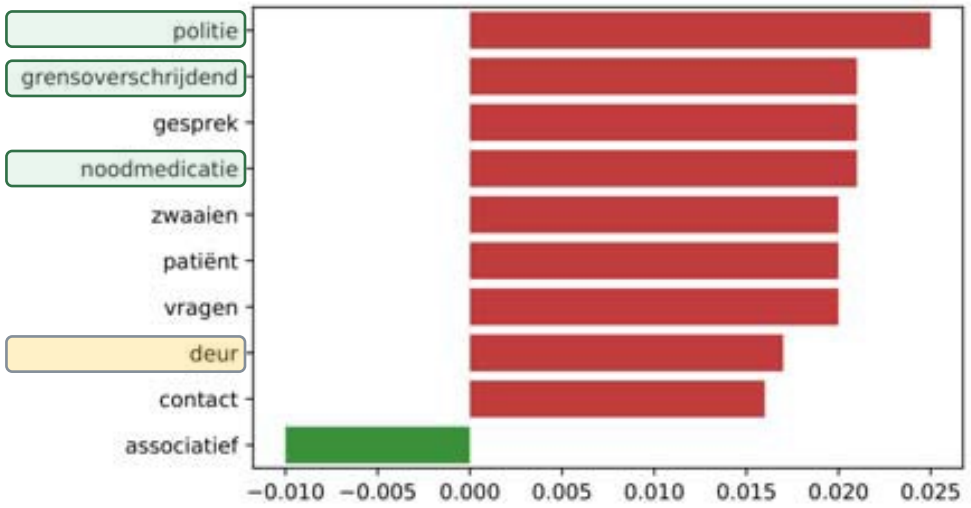
Receiver operator characteristic curves are shown for each fold, according to internal cross-validation in site 1 (A) and site 2 (B). Dashed diagonal lines denote an area under the curve (AUC) of 0.5, ie, predictive validity equivalent to chance. AUC indicates area under the curve.

[5/6]
Evaluation: Exploratory Analysis

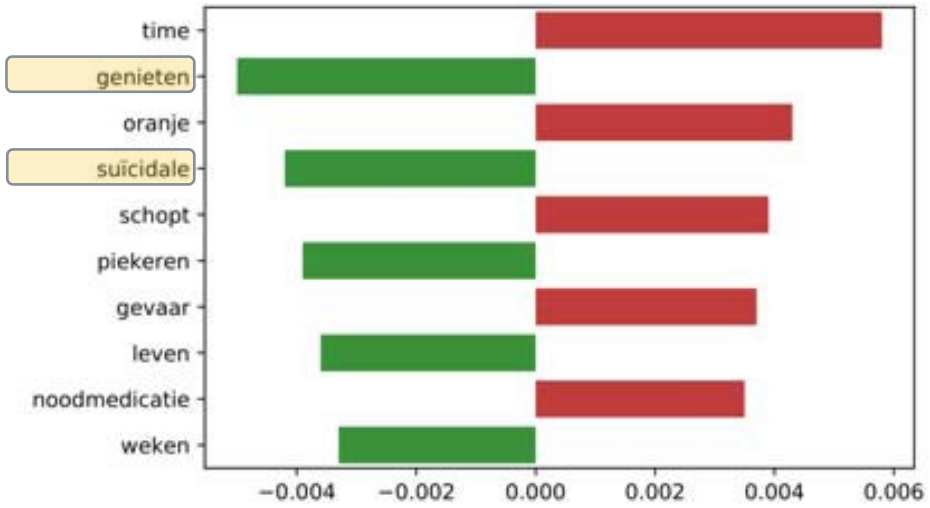
Table 3. Results of Exploratory Analysis

	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

[5/6]
Evaluation: Model Explainability



- Sample of Local Explanation predicting <u>high</u> risk of aggression



- Sample of Local Explanation predicing <u>low</u> risk of aggression

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

[6/6]
Deployment: GitHub Repository? Clinical intervention? Dashboard?

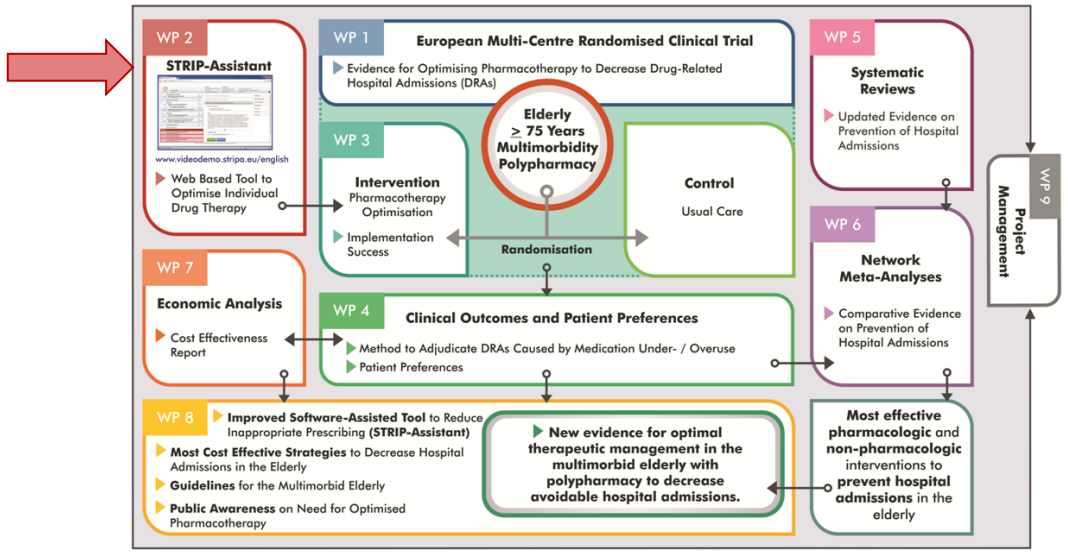
Startup options

- https://github.com/vmenger/violence-risk-assessment
- DEDUCE

  "de-identification method for Dutch medical text"
  - pip install deduce
  - http://tdslab.nl/deduce
- SNPcurator
  - http://tdslab.nl/snpcurator

Advanced options

- Availability in Dashboards
- Intervention instrument in RCT (e.g. STRIPA)
  - IMDD, Ethical committee approval



THANK YOU

PROF. DR. MARCO SPRUIT
contact: m.r.spruit@lumc.nl

https://www.universiteitleiden.nl/medewerkers/marco-spruit