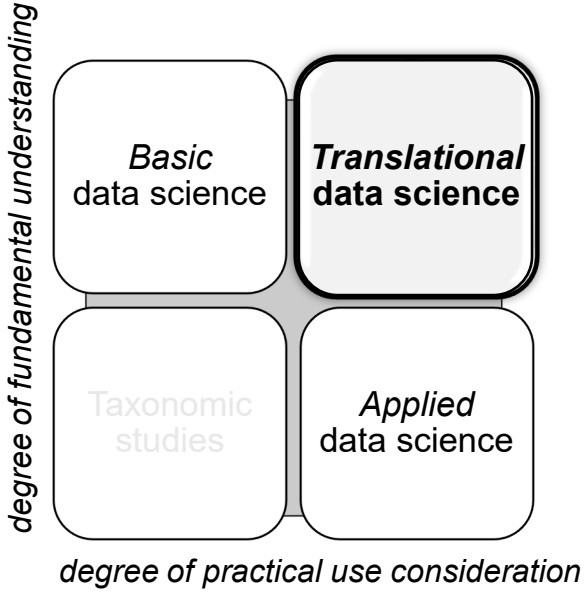
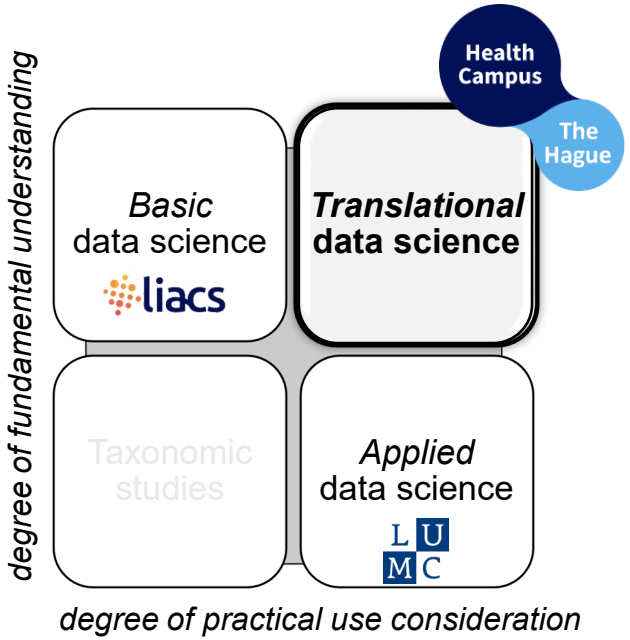
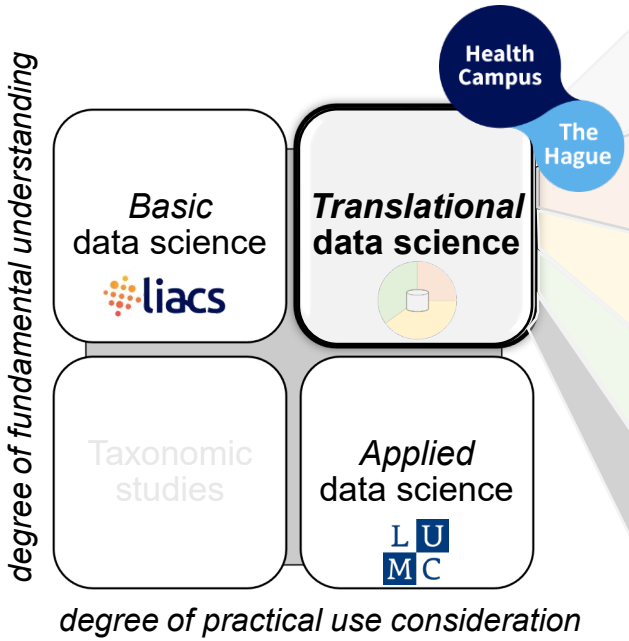
Natural language processing for enriching real world evidence from electronic health records

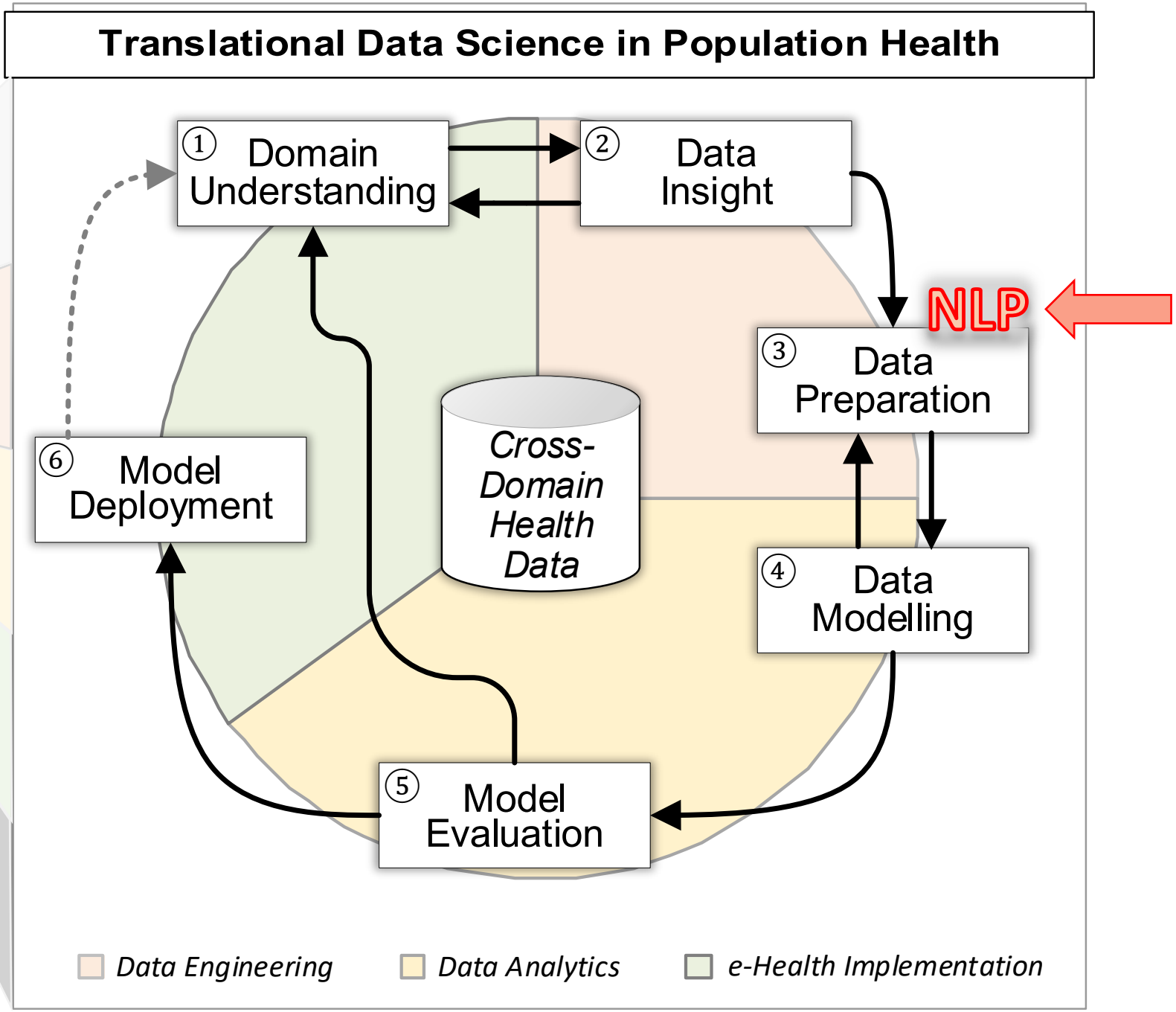
…AI @ Health Campus The Hague

3rd Leiden Drug Development Conference, @ECC Leiden, 19 September 2024, **Marco Spruit**
"LDDC-3: Artificial Intelligence in drug development, manufacturing and health care"

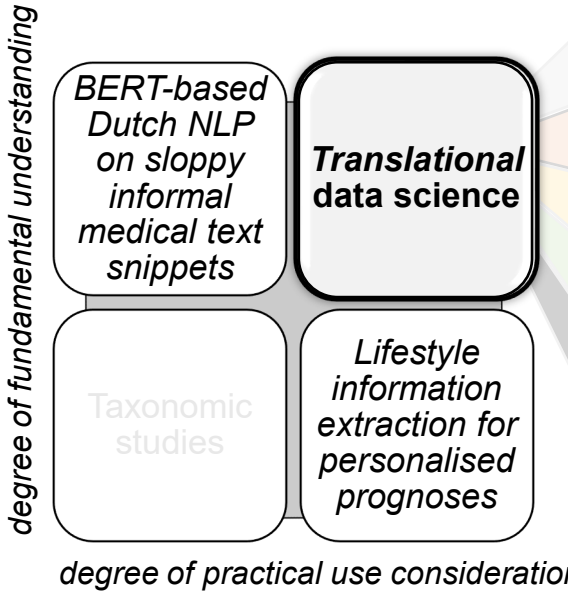

Basic
data science
Translational
data science
Taxonomic
studies
Applied
data science
degree of fundamental understanding
degree of practical use consideration


Health Campus
The Hague
Basic
data science
liacs
Translational
data science
Taxonomic
studies
Applied
data science
LUMC
degree of fundamental understanding
degree of practical use consideration





Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, *9*(13), 1-73. https://books.google.com/books/about/CRISP_DM_1_0.html



BMC Part of Springer Nature

Search Menu

BMC Medical Informatics and Decision Making

Home About Articles Submission Guidelines Collections Join The Board Submit manuscript

Download PDF

Research | Open access | Published: 03 June 2024

Extracting patient lifestyle characteristics from Dutch clinical text with BERT models

Hielke Muizelaar, Marcel Haas, Koert van Dortmont, Peter van der Putten & Marco Spruit

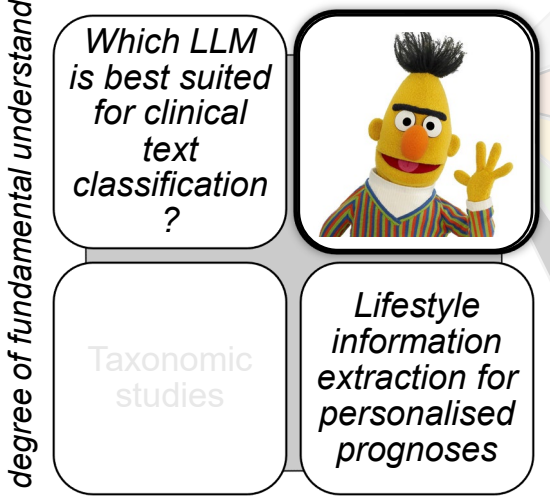
BMC Medical Informatics and Decision Making 24, Article number: 151 (2024) | Cite this article

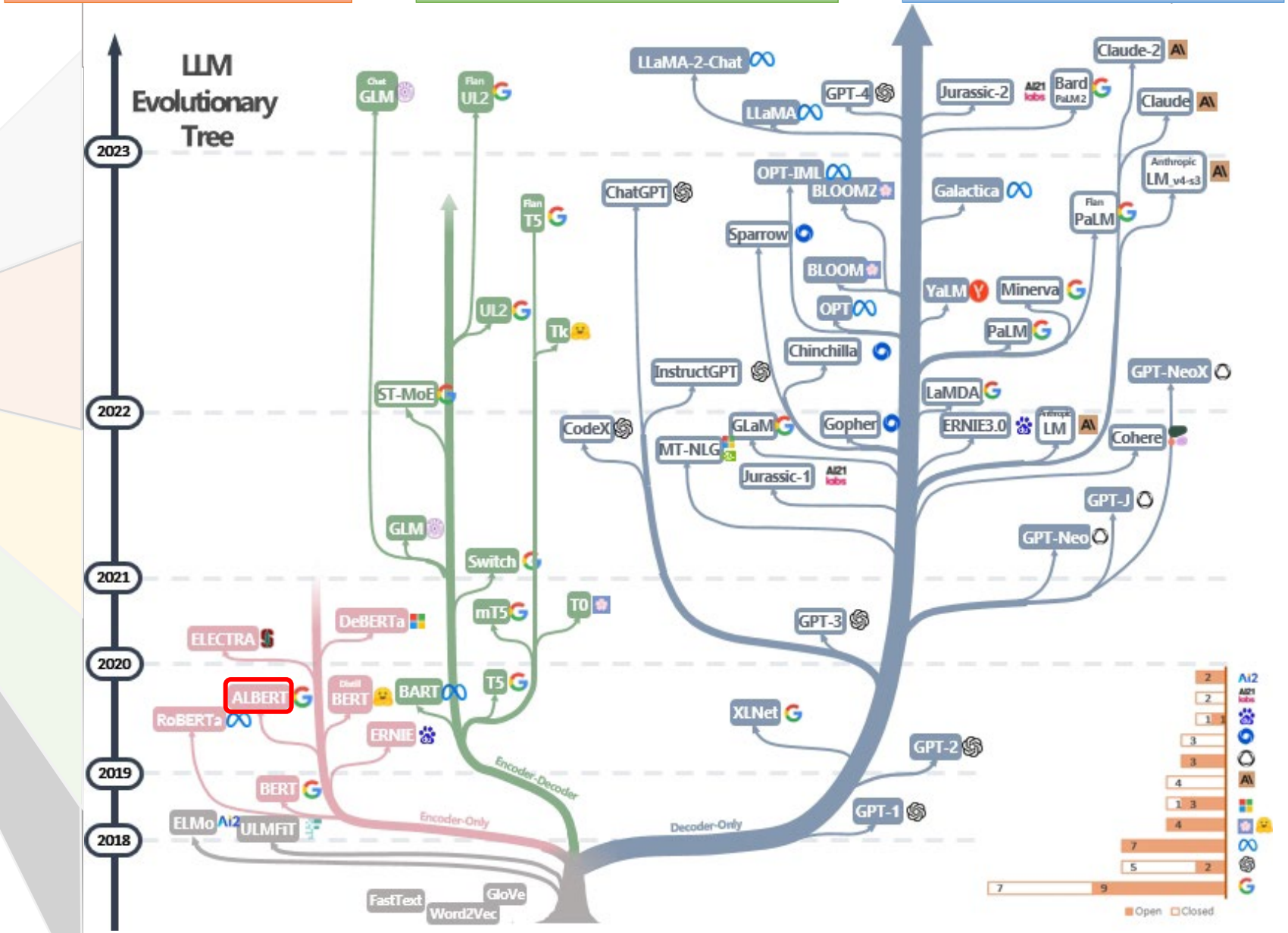
828 Accesses | 2 Citations | 2 Altmetric | Metrics

Abstract

Background

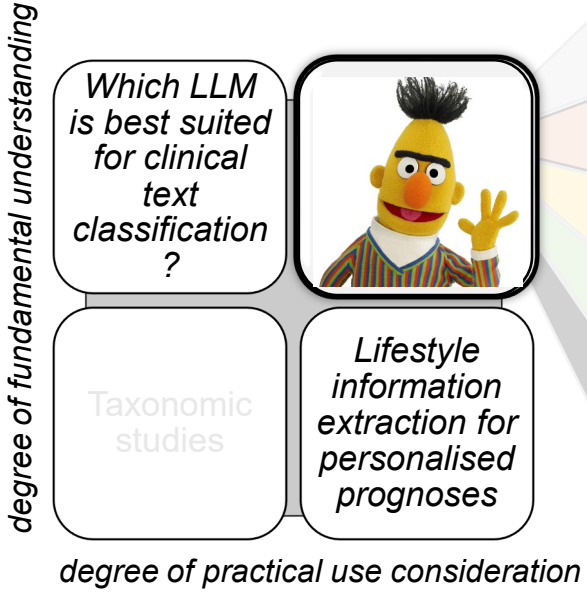
BERT models have seen widespread use on unstructured text within the clinical domain. However, little to no research has been conducted into classifying unstructured clinical notes on the basis of patient lifestyle indicators, especially in Dutch. This article

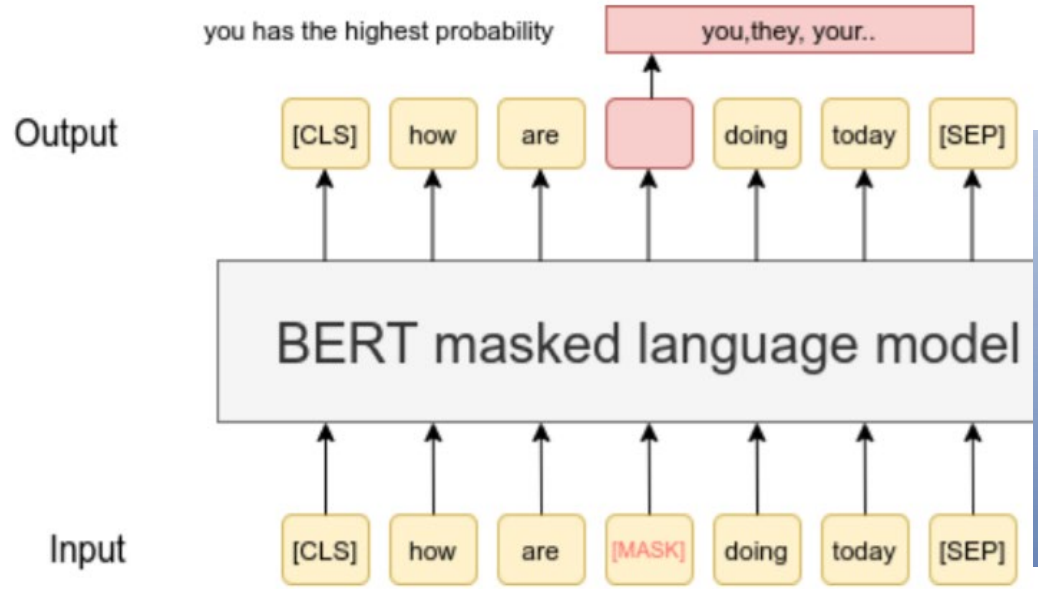


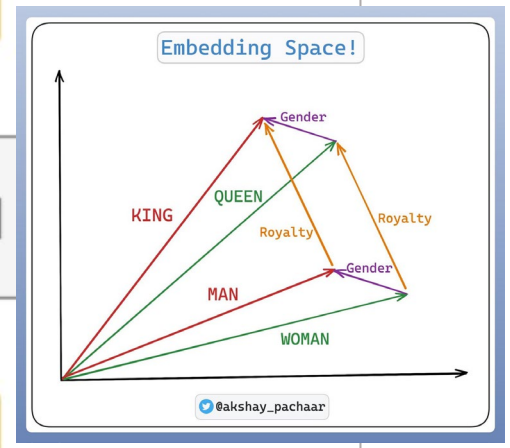


Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., ... & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*. https://doi.org/10.1145/3649506

BERT high-level architecture



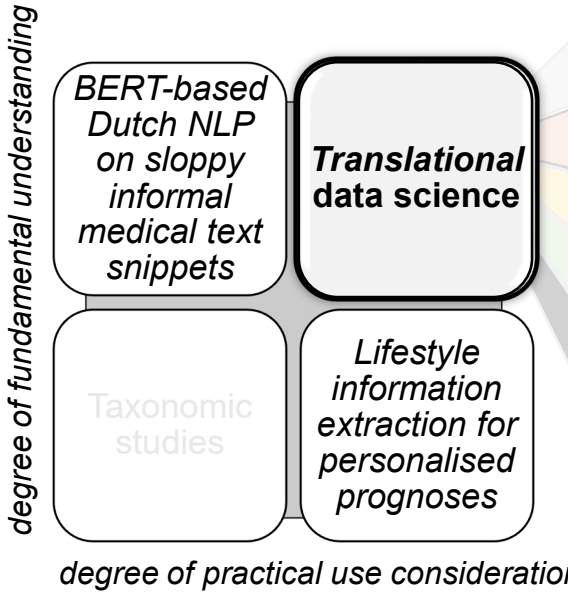




[MASK] : Masked Language Modelling (MLM) → Word prediction

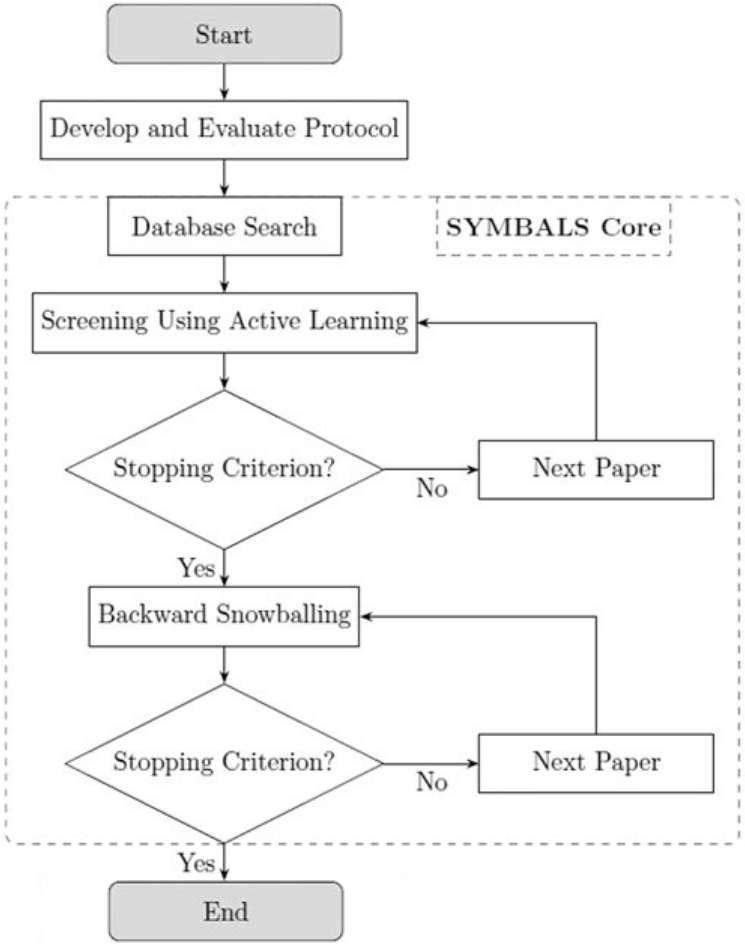
[CLS] : Classification for Next Sentence Prediction (NSP)

Domain understanding [1/6]: *Determining Objective*



HagaZiekenhuis

- Case study
- Previous research
- Lifestyle characteristics
- NLP task → BERT



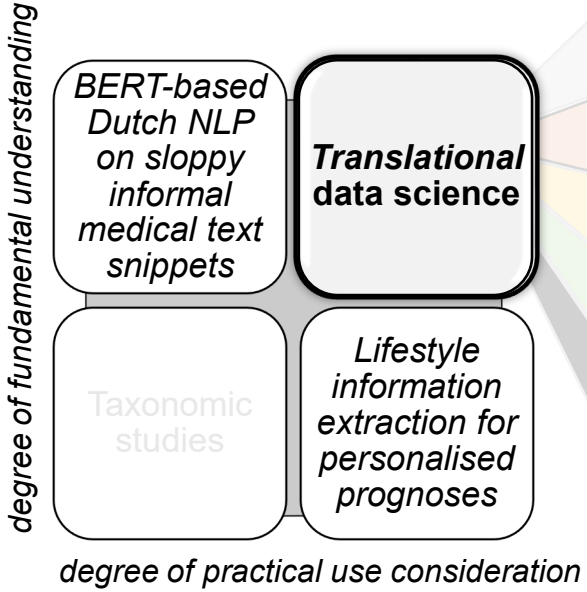
Systematic review

- Active learning
- Backward snowballing
- SYMBALS →
- 85 papers

Objectives

1. Pretraining BERT model from scratch
2. Pretraining on top of Dutch BERT models
3. Translation strategy

van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., & Spruit, M. (2021). SYMBALS: A systematic review methodology blending active learning and snowballing. *Frontiers in research metrics and analytics*, *6*, 685591. https://doi.org/10.3389/frma.2021.685591



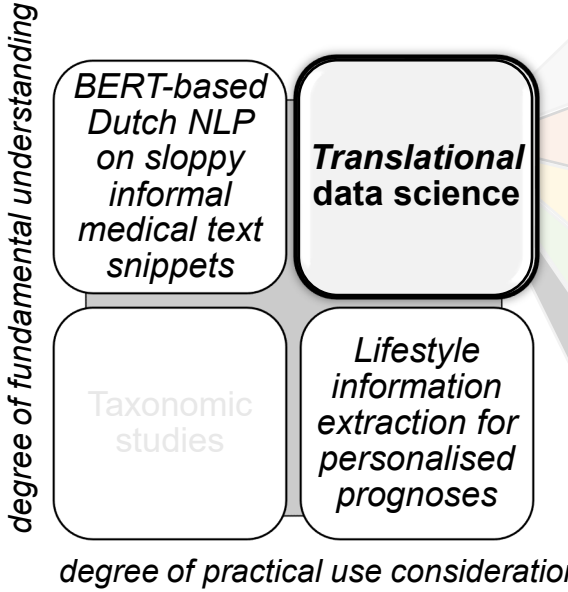
Data Insight [2/6]: *Data Description*

- CTcue for data extraction of clinical EHR texts (*with* labels)

Example text data	**Smoking**	**Alcohol**	**Drugs**
Patient smokes, does not drink or use drugs	Current user	Non-user	Non-user
Patient used to smoke, drinks 1 beer a day	Former user	Current user	Unknown
Patient used to smoke, uses marihuana daily	Former user	Unknown	Current user

- Descriptive statistics

Type of label	**#labelled texts**	**Current users**	**Former Users**	**Non-users**	**Unknown**
Smoking	148.768	7.015 (4.72%)	32.230 (21.66%)	44.677 (30.03%)	64.846 (43.59%)
Drinking	143.166	16.017 (11.25%)		39.119 (27.32%)	87.940 (61.43%)
Drugs	147.999	1.443 (0.98%)		53.005 (35.81%)	93.551 (63.21%)



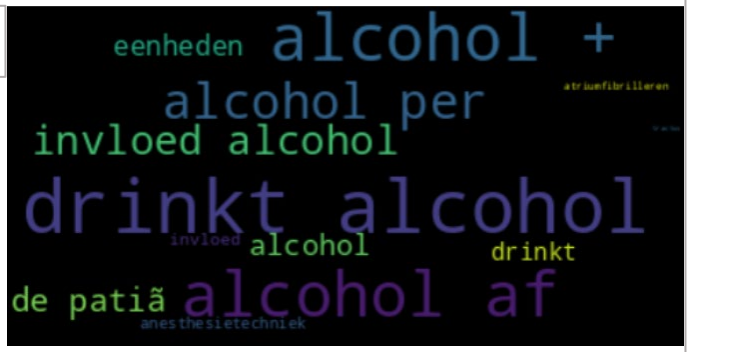
Data Insight [2/6]: *Exploratory Data Analysis*

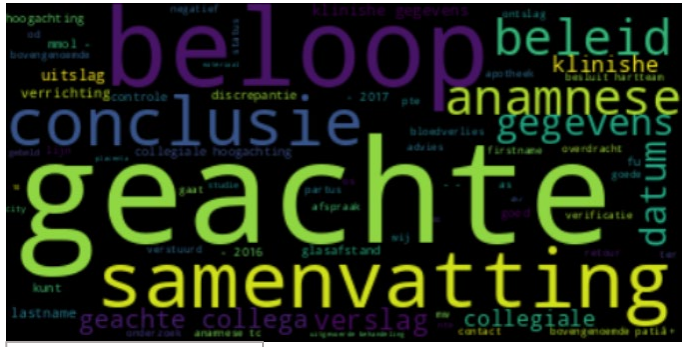
- Checking label quality after obtaining suspicious performance

Stochastic Gradient Descent class: *Alcohol Use*	**Unknown F1-score**	**Current F1-score**	**Non-user F1-score**	**Macro F1-score**
(Ngram 2, Stopwords kept)	1.00	0.99	0.99	0.99
(Ngram 2, No stopwords)	1.00	0.97	0.99	0.99
(Ngram 1, Less stopwords)	0.95	0.61	0.69	0.75

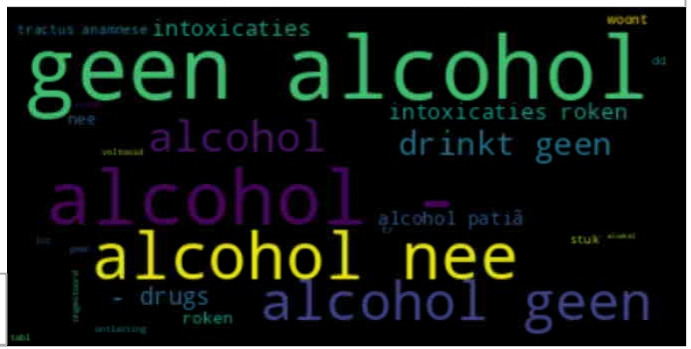
- No BERT necessary??

"Current user"

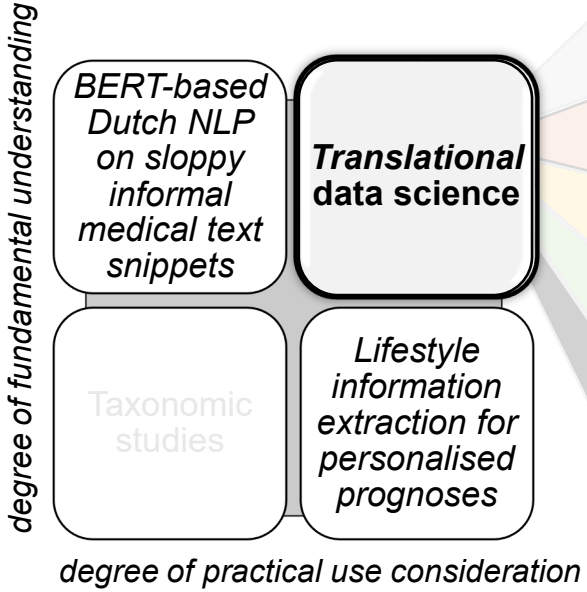




"Unknown"



"Non-user"



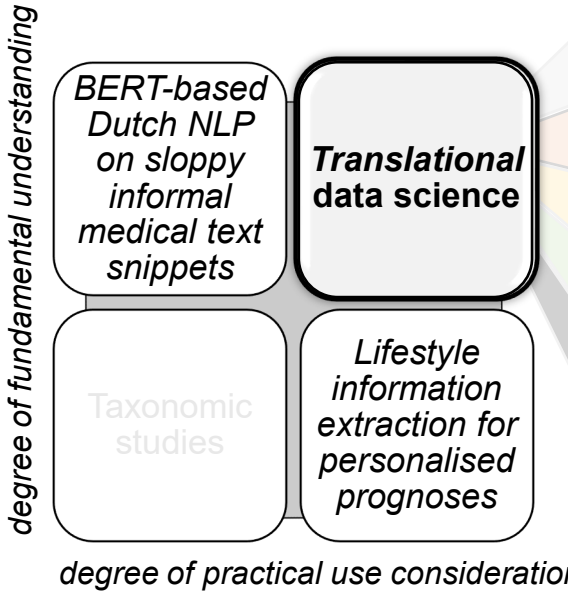
Data Insight [2/6]: *Exploratory Data Analysis*

- Exemplar *edge* cases from the Smoking task in clinical notes that were predicted as *Non-user* class but were misclassified.

Edge cases in Dutch clinical notes	***English translation***	**Truth**
... Intoxicaties: roken 2-4 sigaren per dag alleen in de zomer, ...	*Intoxications: smokes 2-4 cigars per day only in the summer,*	**Current user**
... roken (+ (20 packyears, is gestopt); ... Risicofactoren: Familieanamnese (+); roken (-) ...	*smoking (+ (20 packyears, has stopped); ... Risk factors: Family history (+); smoking (-)*	Former user
... Roken: 3-4 per dag al jaren lang, vroeger wel meer ...	*Smoking: 3-4 a day for years, in the past more*	**Current user**
... Roken: - (20 jaar geleden na 50 py), Cardiovasculaire risico-factoren: roken:+ ...	*Smoking:- (20 years ago after 50 py), Cardiovascular risk factors: smoking:+*	Former user

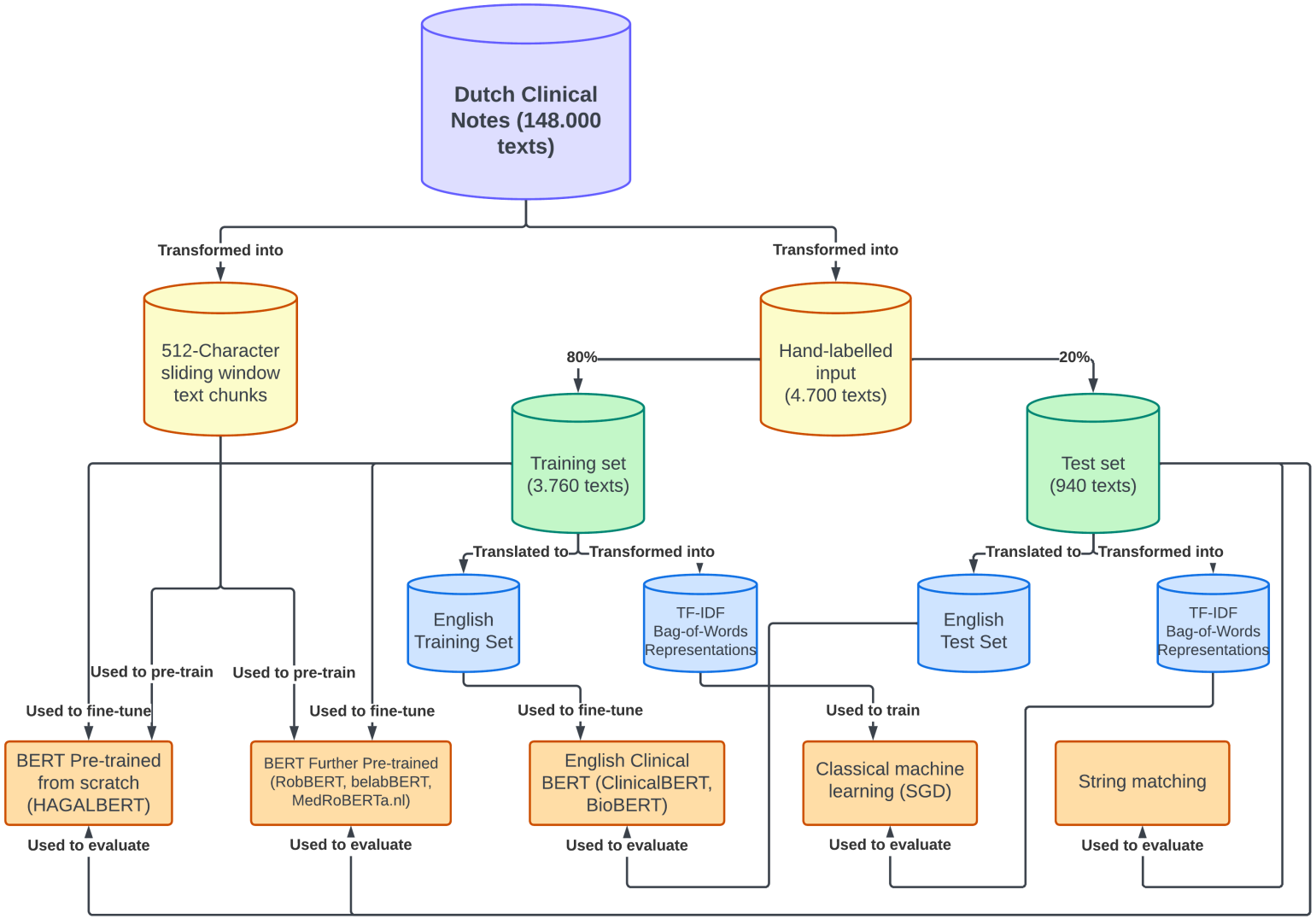
- Top 50 texts with highest scores for **Current user** class that were predicted *Non-user*
- *… Hand-labeling #4700 texts*

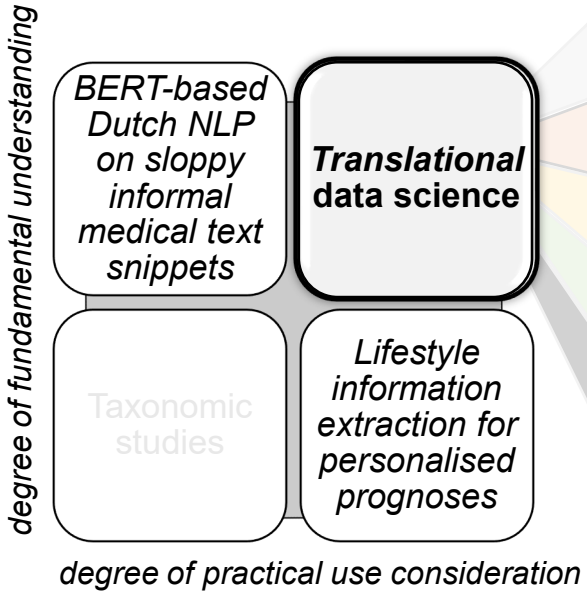
True label	#
Unknown	30/50
Current user	**5/50**
Non user	10/50
Former user	5/50



Computational experiments

Data preparation [3/6]

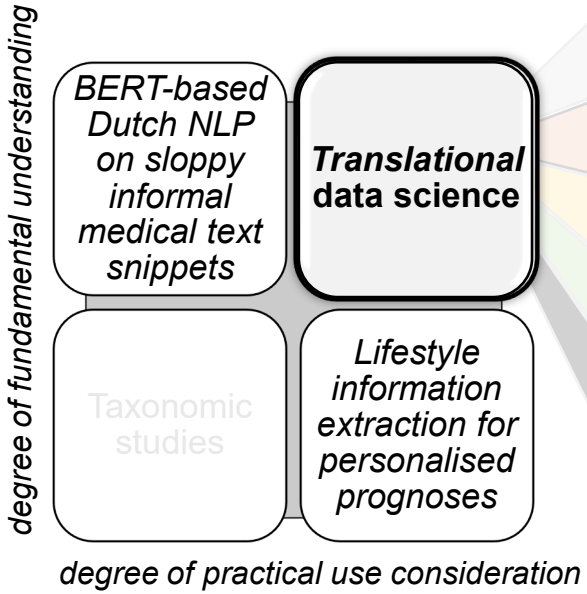




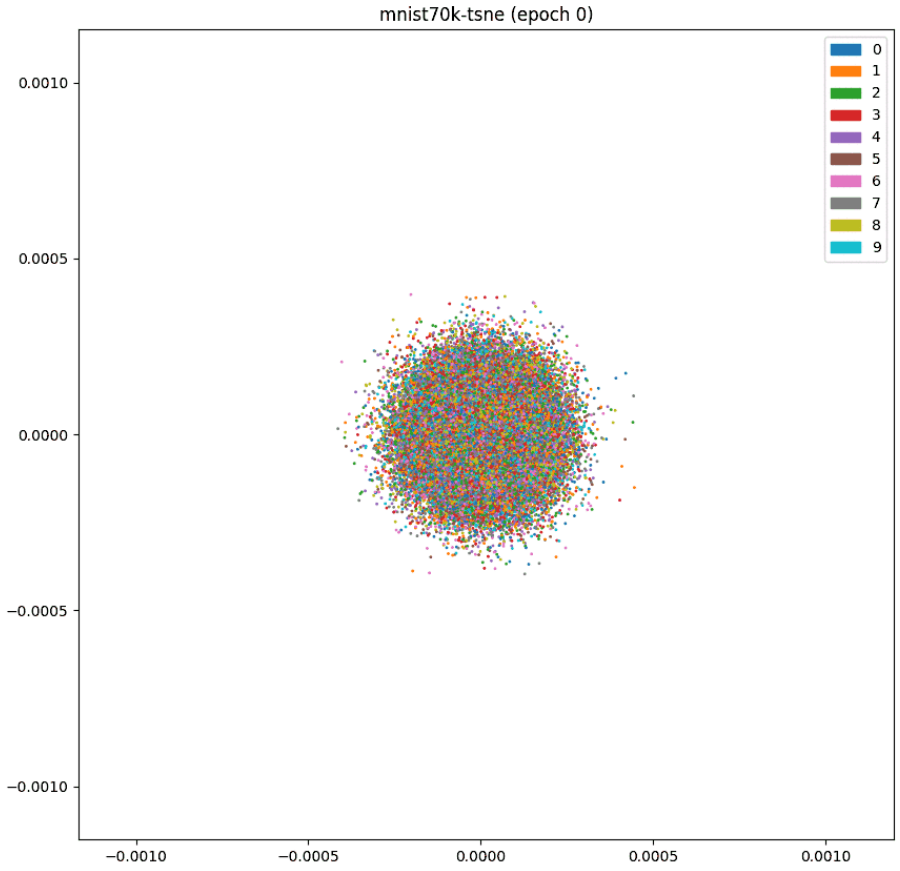
Data Modelling [4/6] –Experimental Results

Model	Smoking	Alcohol	Drugs
[a] Traditional models			
String Matching	0.84	0.74	0.68
Machine Learning (SGD)	0.85	0.71	0.60
[b] Trained from scratch			
HAGALBERT	0.66	0.54	0.43
[c] Fine-tuned models			
RobBERT-HAGA	0.87	0.71	0.63
belabBERT-HAGA	0.48	0.64	0.57
MedRoBERTa.nl-HAGA	0.93	0.79	0.77
[d] Large English models			
BioBERT (translated)	0.91	0.72	0.52
ClinicalBERT (translated)	0.92	0.80	0.61

Model Evaluation [5/6] - *t-SNE viz*

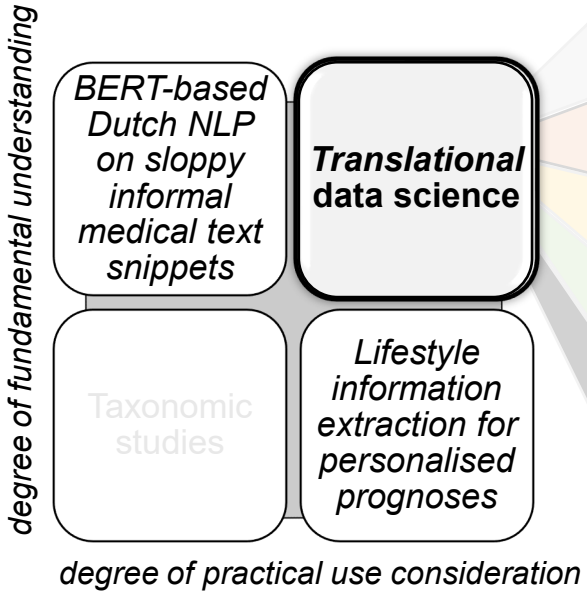


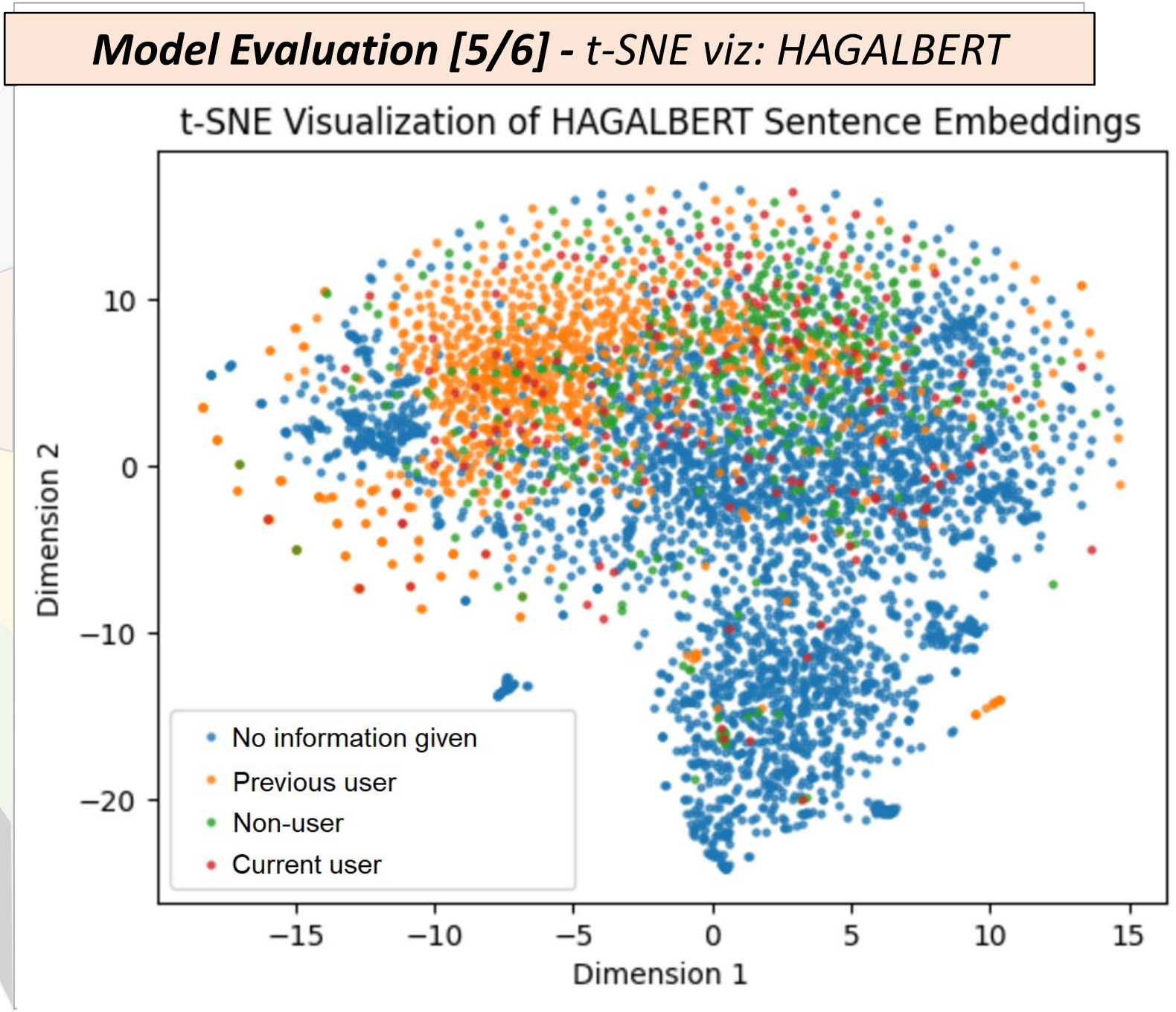
- t-SNE *(t-distributed Stochastic Neighbor Embedding)*: for nonlinear dimensionality reduction (*e.g.* PCA)



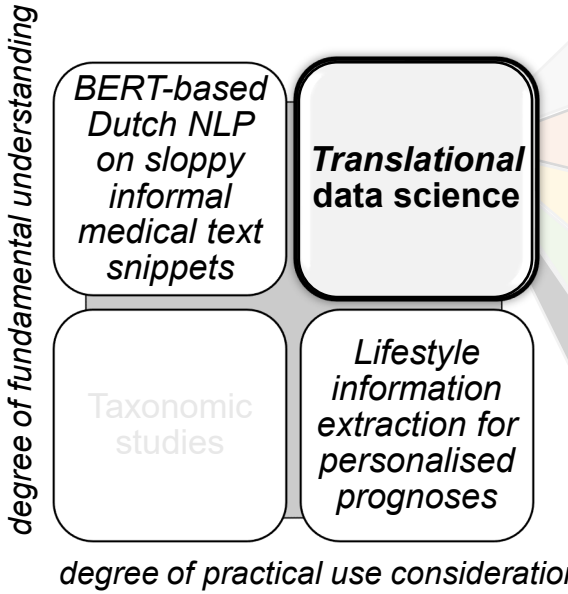
https://github.com/KellerJordan/tSNE-Animation

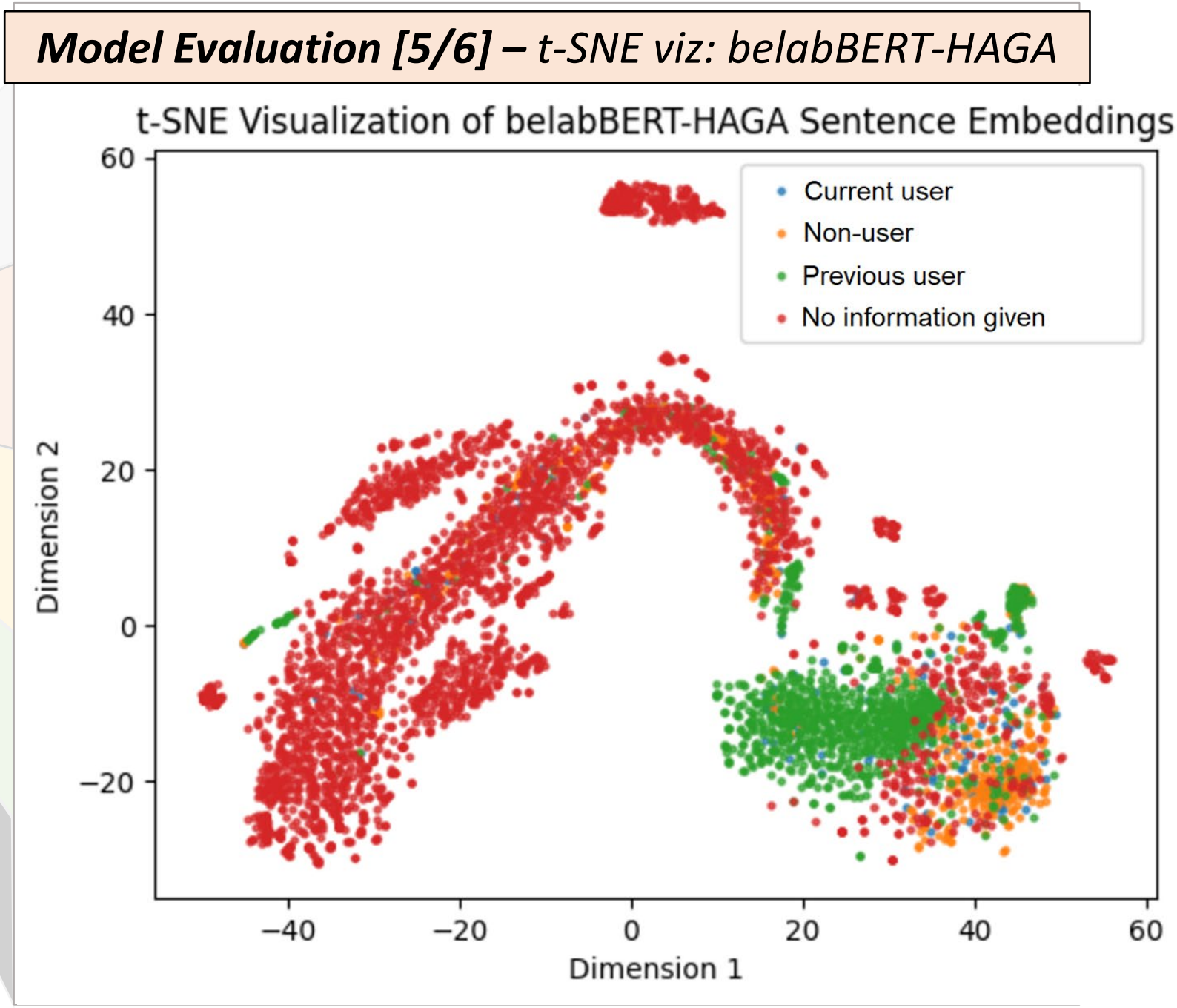
Model Evaluation [5/6] - *t-SNE viz: HAGALBERT*



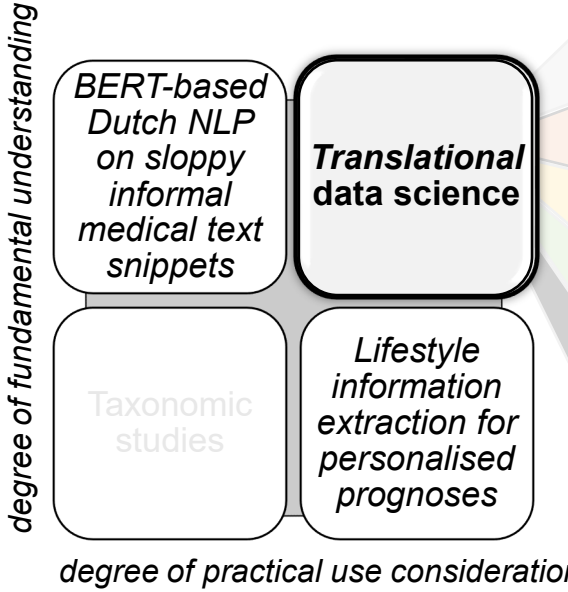


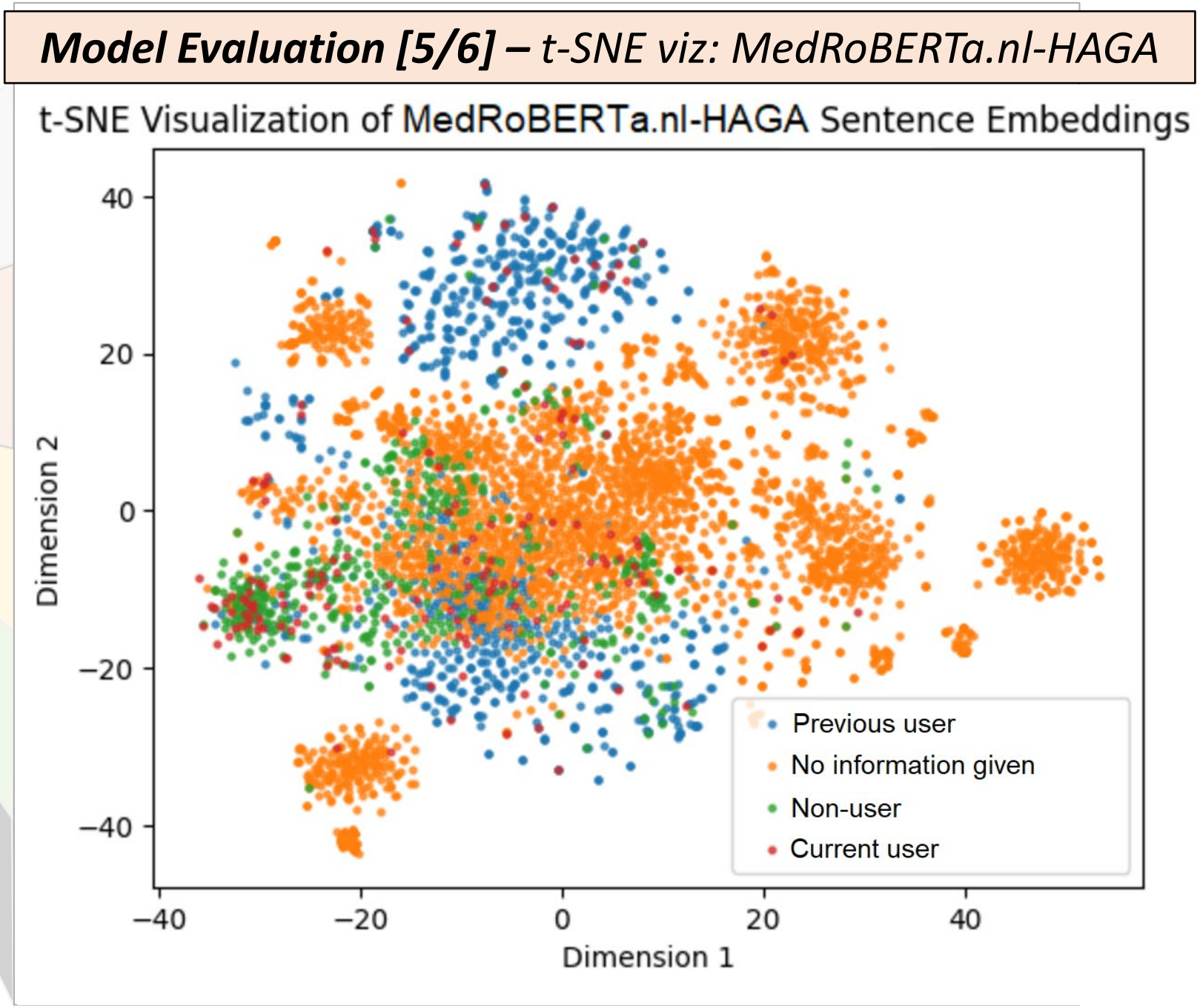
Model Evaluation [5/6] – *t-SNE viz: belabBERT-HAGA*



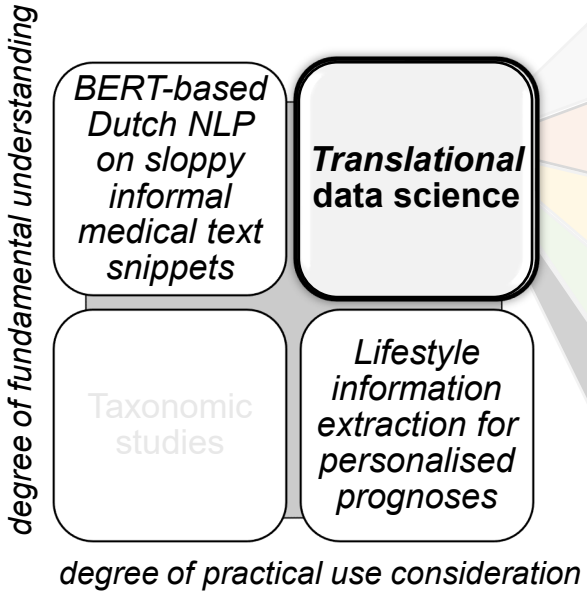


Model Evaluation [5/6] – *t-SNE viz: MedRoBERTa.nl-HAGA*

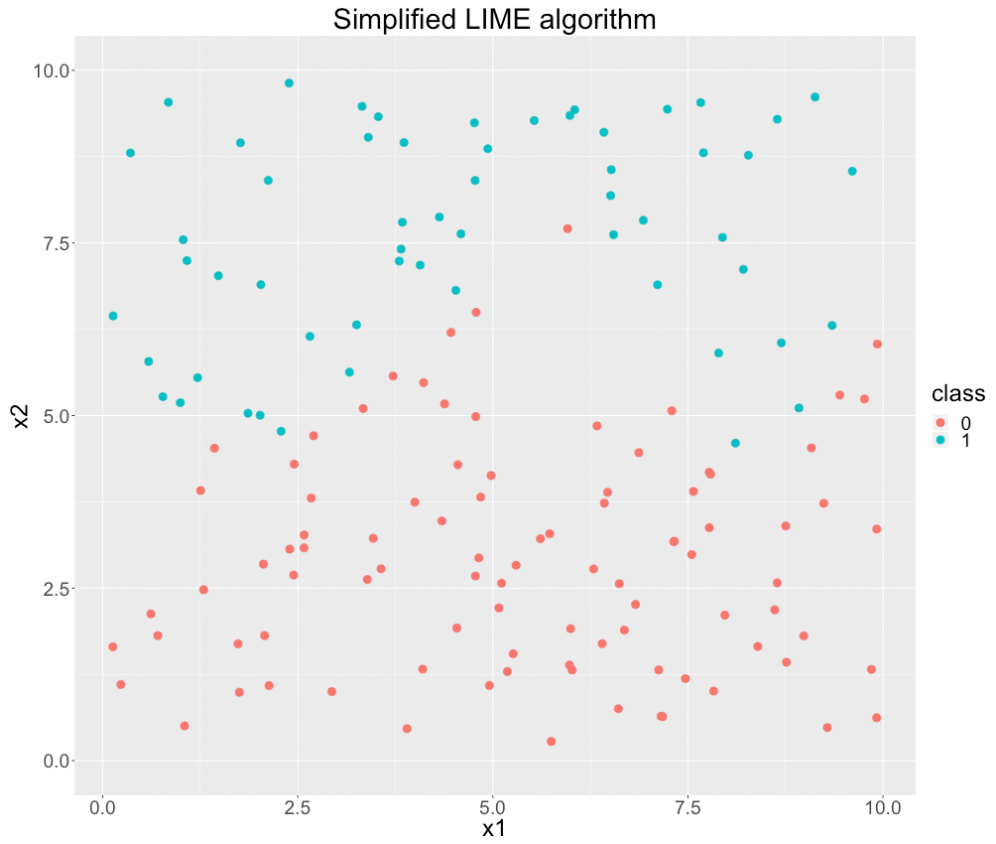




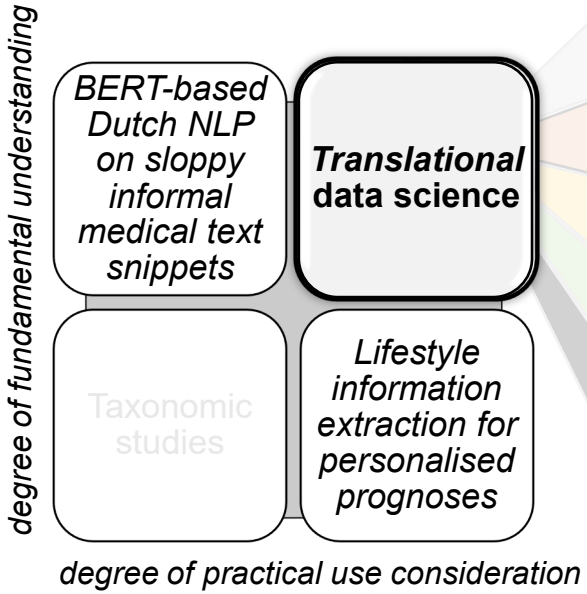
Model Evaluation [5/6] - LIME



- LIME: Local interpretable model-agnostic explanations, to train local *surrogate* models to explain individual predictions

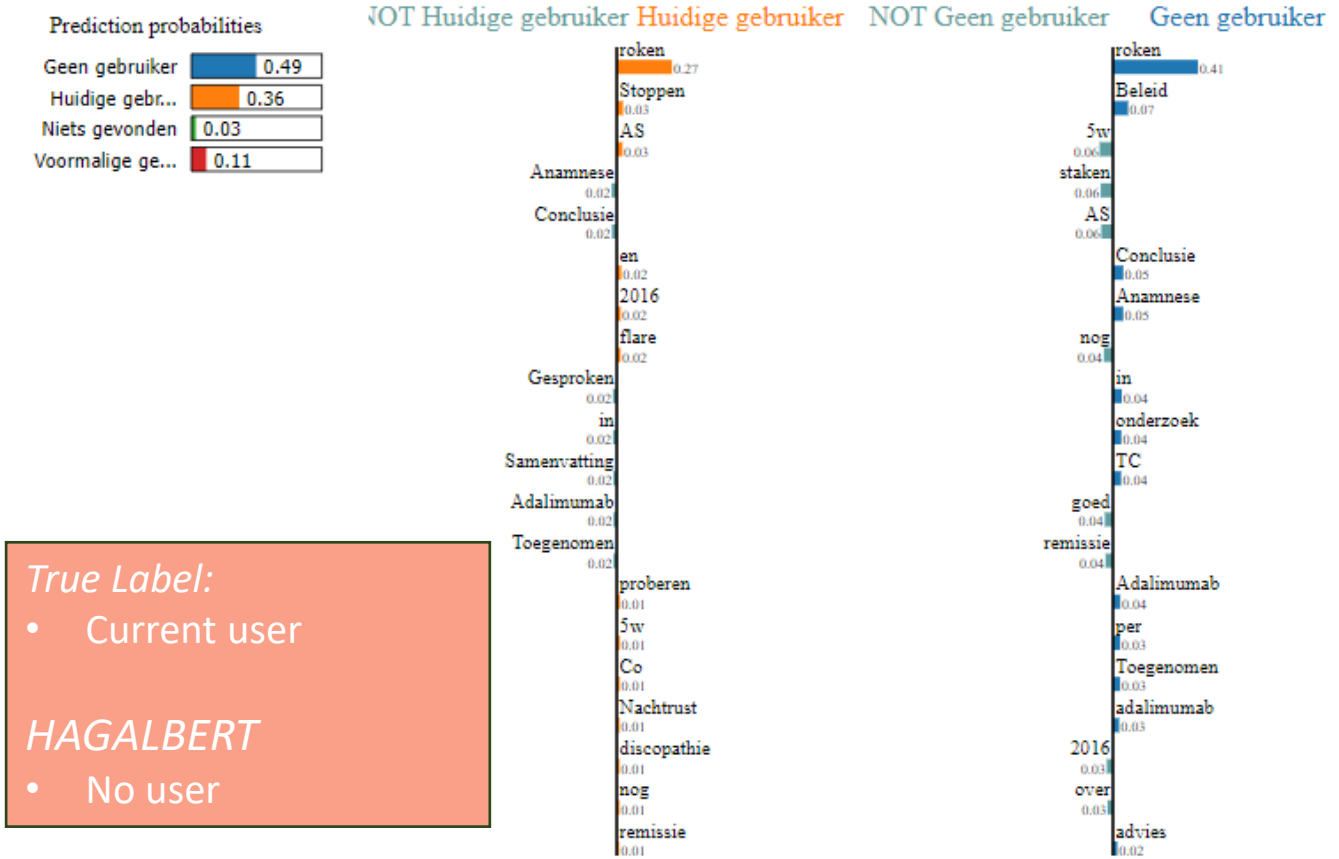


https://slds-lmu.github.io/iml_methods_limitations/lime.html



Model Evaluation [5/6] – *LIME: HAGALBERT*

- *"Summary: Course of radiology: 2014 Increased spondyloarthrosis and disc disease. 2016 flare after discontinuation. Case history: Talked about smoking, advice to stop. Sleeps well, also functions well in daily life. Adalimumab still per 4w. Physical examination: Hand osteoarthritis. Conclusion: Conclusion: AS in remission under adalimuma.b Policy: Policy: Co 1j (combi) + TC. Stop smoking. Try adalimumab per 5w."*

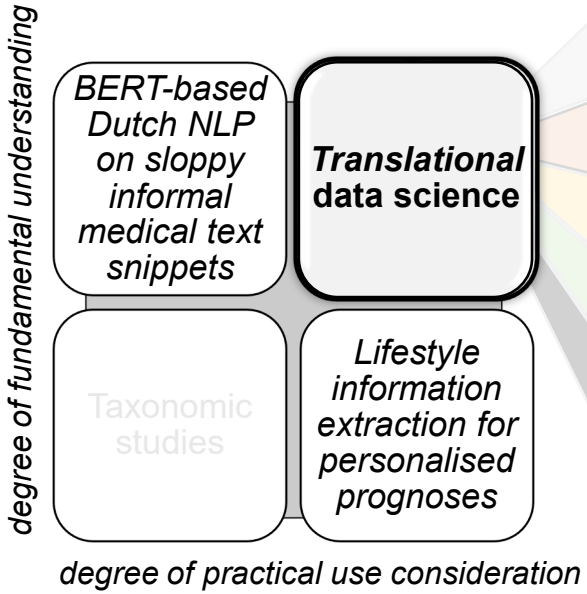


True Label:
- Current user

HAGALBERT
- No user

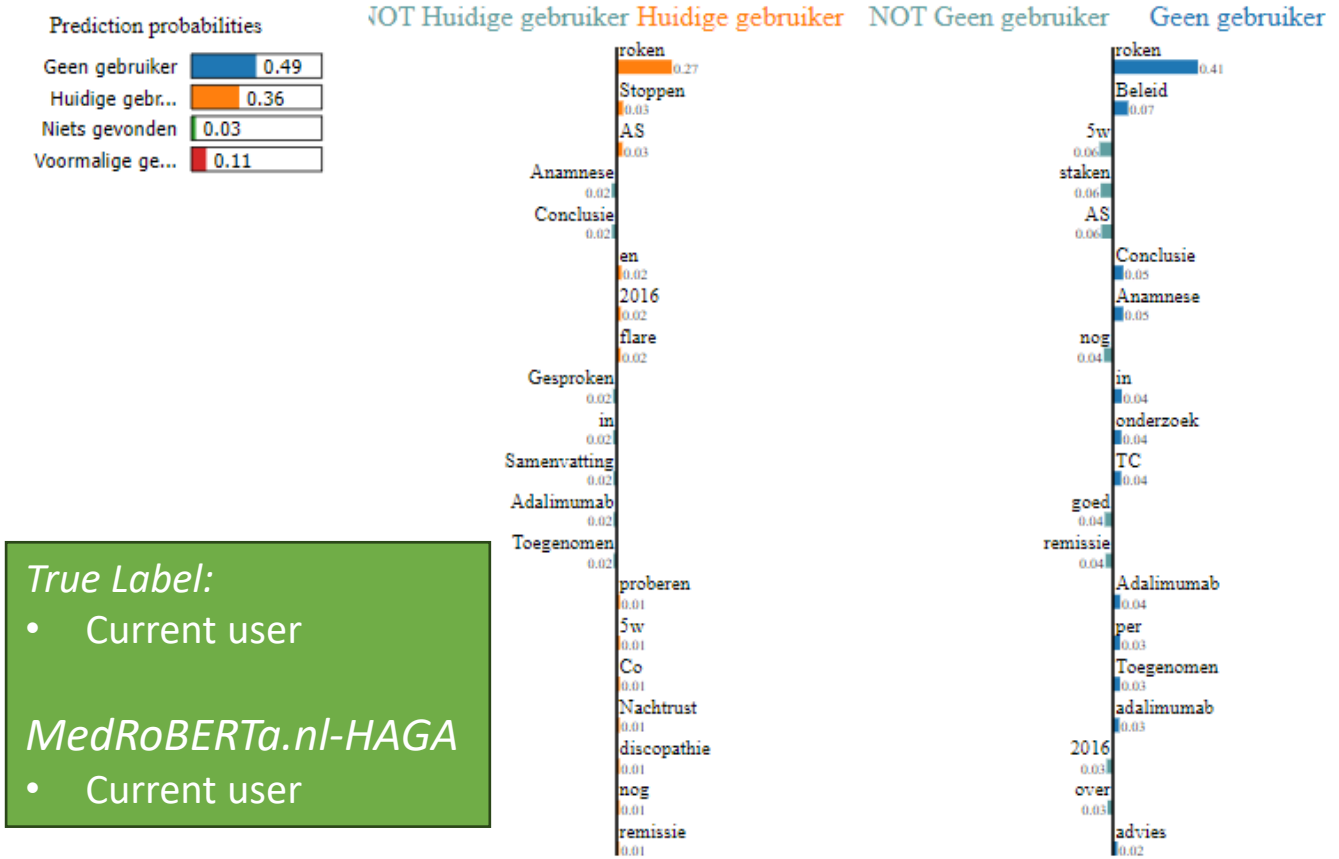
Text with highlighted words

Samenvatting: Beloop radiologie: 2014 Toegenomen spondylartrose en discopathie. 2016 flare na staken Anamnese: Gesproken over roken, advies staken Nachtrust goed, functioneert ook goed in dagelijks leven Adalimumab per 4w nog Lichamelijk onderzoek: Handartrose Conclusie: Conclusie: AS in remissie onder adalimumab Beleid: Beleid: Co 1j (combi) + TC Stoppen met roken Adalimumab per 5w proberen



LIME Model Evaluation [5/6] - MedRoBERTa.nl-HAGA

- *"Summary: Course of radiology: 2014 Increased spondyloarthrosis and disc disease. 2016 flare after discontinuation. Case history: Talked about smoking, advice to stop. Sleeps well, also functions well in daily life. Adalimumab still per 4w. Physical examination: Hand osteoarthritis. Conclusion: Conclusion: AS in remission under adalimuma.b Policy: Policy: Co 1j (combi) + TC. Stop smoking. Try adalimumab per 5w."*



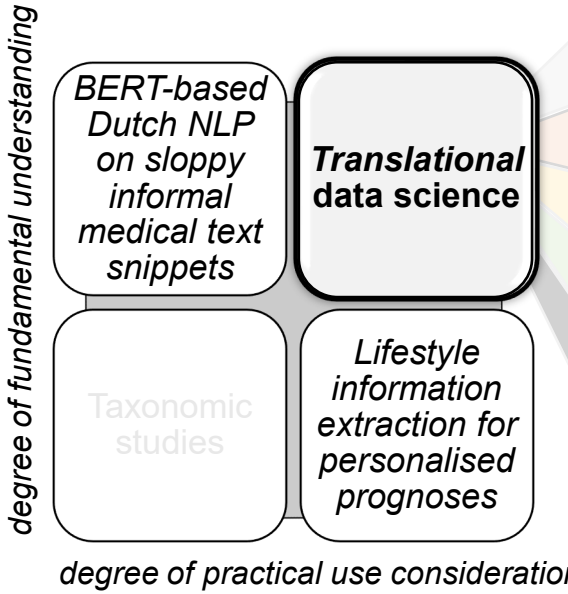
True Label:
- Current user

MedRoBERTa.nl-HAGA
- Current user

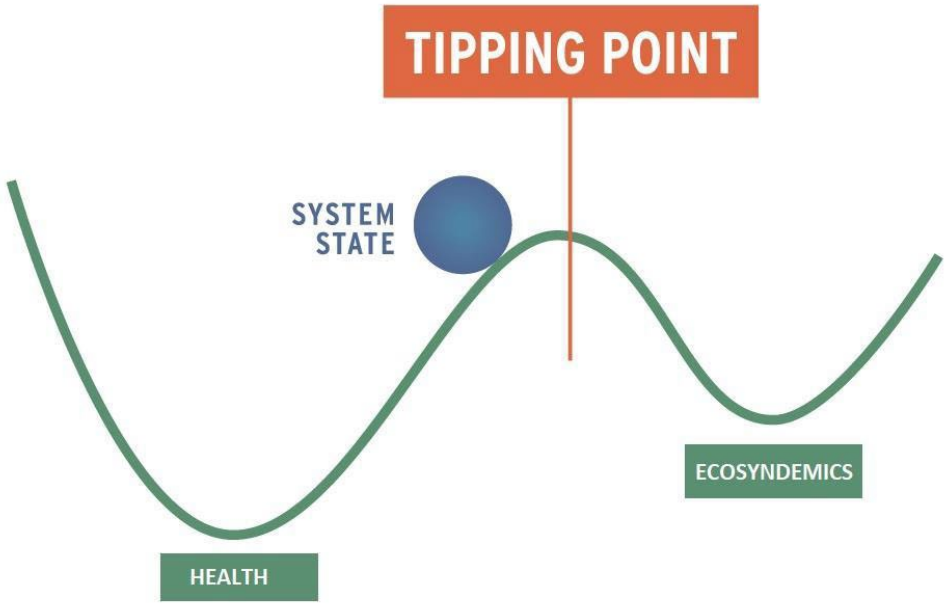
Text with highlighted words

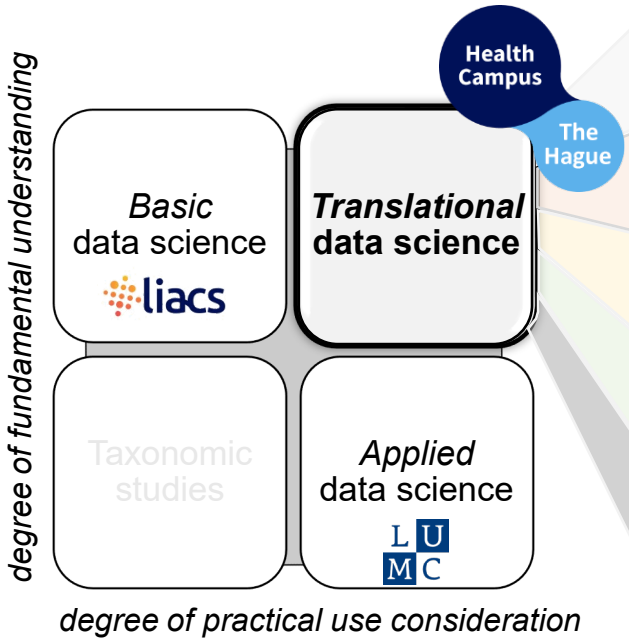
Samenvatting: Beloop radiologie: 2014 Toegenomen spondylartrose en discopathie. 2016 flare na staken Anamnese: Gesproken over roken, advies staken Nachtrust goed, functioneert ook goed in dagelijks leven Adalimumab per 4w nog Lichamelijk onderzoek: Handartrose Conclusie: Conclusie: AS in remissie onder adalimumab Beleid: Beleid: Co 1j (combi) + TC Stoppen met roken Adalimumab per 5w proberen

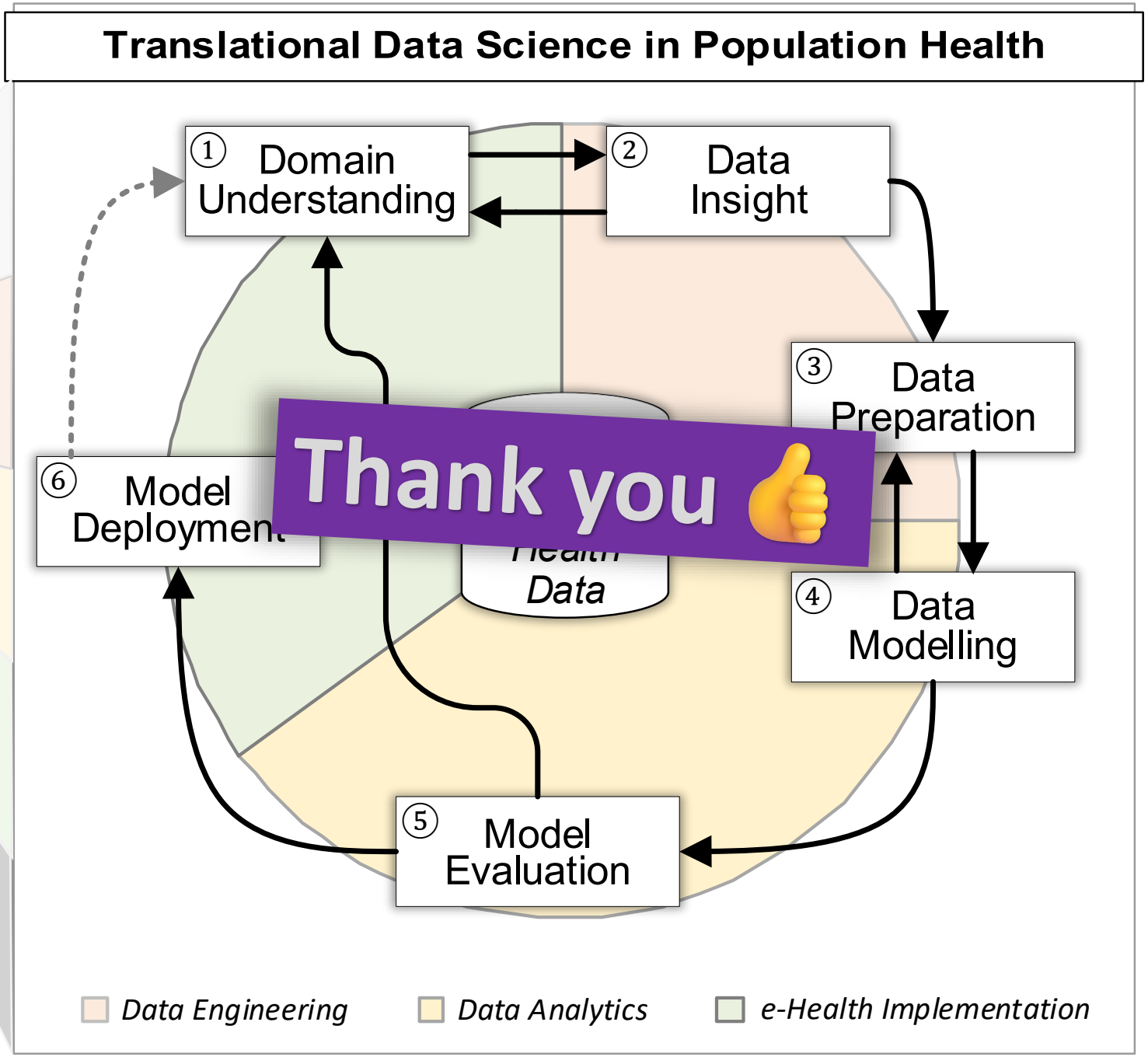
Model Deployment [6/6]

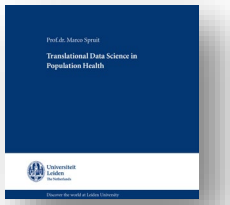


- Follow-up research grant for NWA ECOTIP: *Identifying tipping points of the effects of living environments on ecosyndemics of lifestyle-related illnesses*









Spruit, Marco. (2022). *Translational Data Science in Population Health* (p. 20). Inaugural lecture. Leiden University. https://doi.org/10.5281/zenodo.7665858

https://tdslab.liacs.nl