Natural Language Processing for Translational Data Science in Mental Healthcare

Marco Spruit | Research Overview

21 June 2021

Leiden University
Campus The Hague

Agenda

- *Introductions*
  - Me, Translational data science, COVIDA

1. 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. [online]

2. Prediction understanding

Mosteiro,P., Rijcken,E., Zervanou,K., Kaymak,U., Scheepers,F., & Spruit,M. (2021). Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes. *Journal of Artificial Intelligence for Medical Sciences*. [online]

Translational Data Science

Natural Language Processing in Mental Healthcare

About... Marco Spruit

Engineer

1993

- Information Retrieval programmer, ZyLAB Europe BV

1995

- Big Data system developer, Dutch Military Intelligence & Security Service (MIVD)

1997

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

Researcher

2003

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

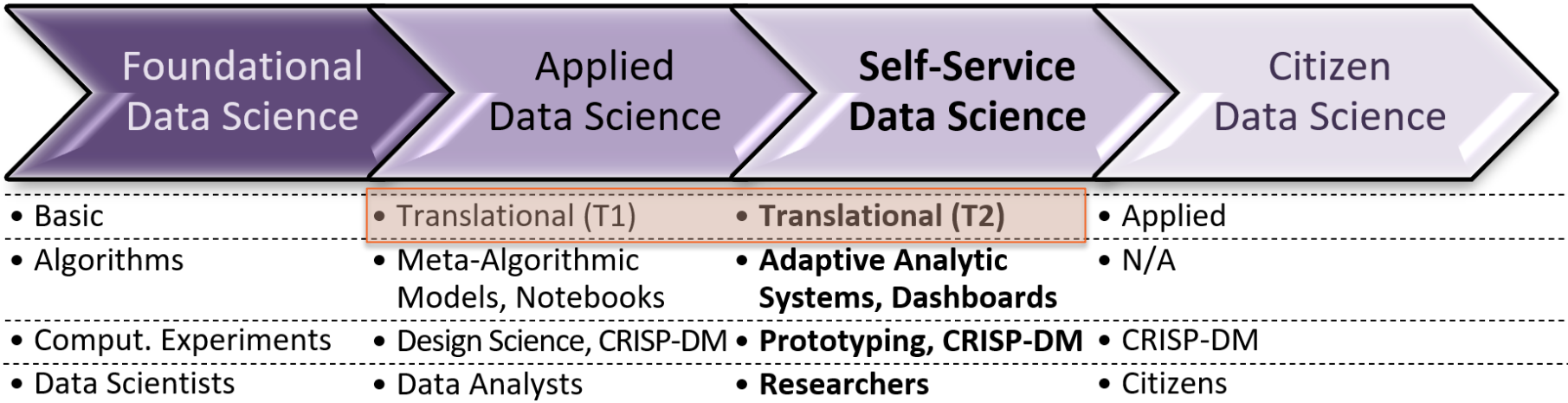
2007

- Assistant & Associate professor Information Science, Utrecht University
  - Applied Data Science Lab

2020

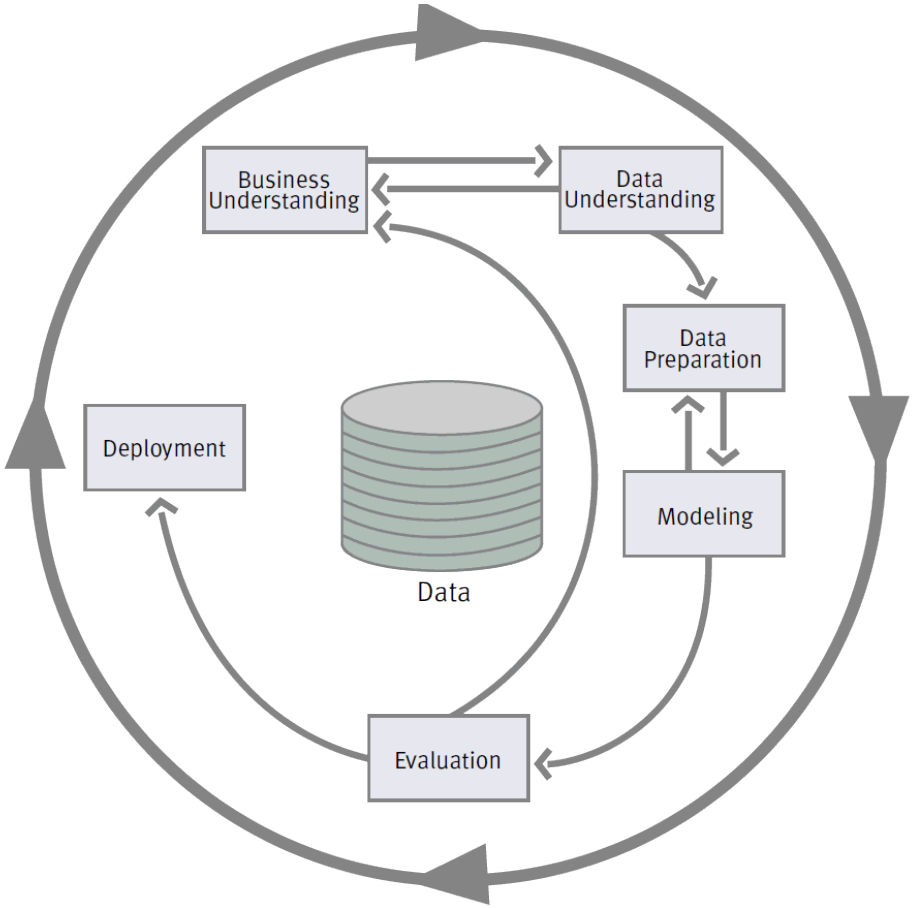
- Professor Advanced Data Science in Population Health, LUMC/Leiden University
  - PH Living Lab
  - CAIRE Lab
  - SIG Health Data Science

Translational Data Science



Spruit,M., & Vries,N. de (2021). Self-Service Data Science for Adverse Event Prediction in Electronic Healthcare Records. In Visvizi,A., Lytras,M., & Aljohani,N. (Eds.), *Springer Proceedings in Complexity*, *Research and Innovation Forum 2020: Disruptive Technologies in Times of Change* (pp. 517–535). RII 2020, Athens, Greece: Springer. [pdf] [online]

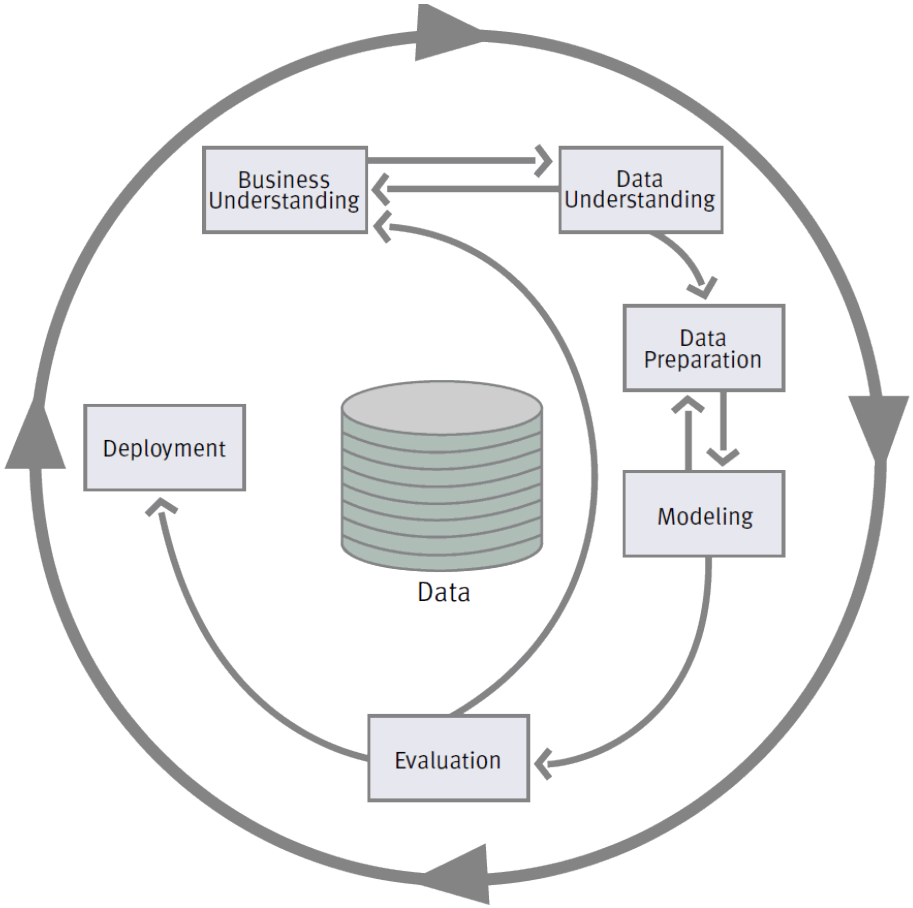
Translational Data Science



CRISP-DM

Chapman, P. Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step Data Mining Guide*. [ftp]

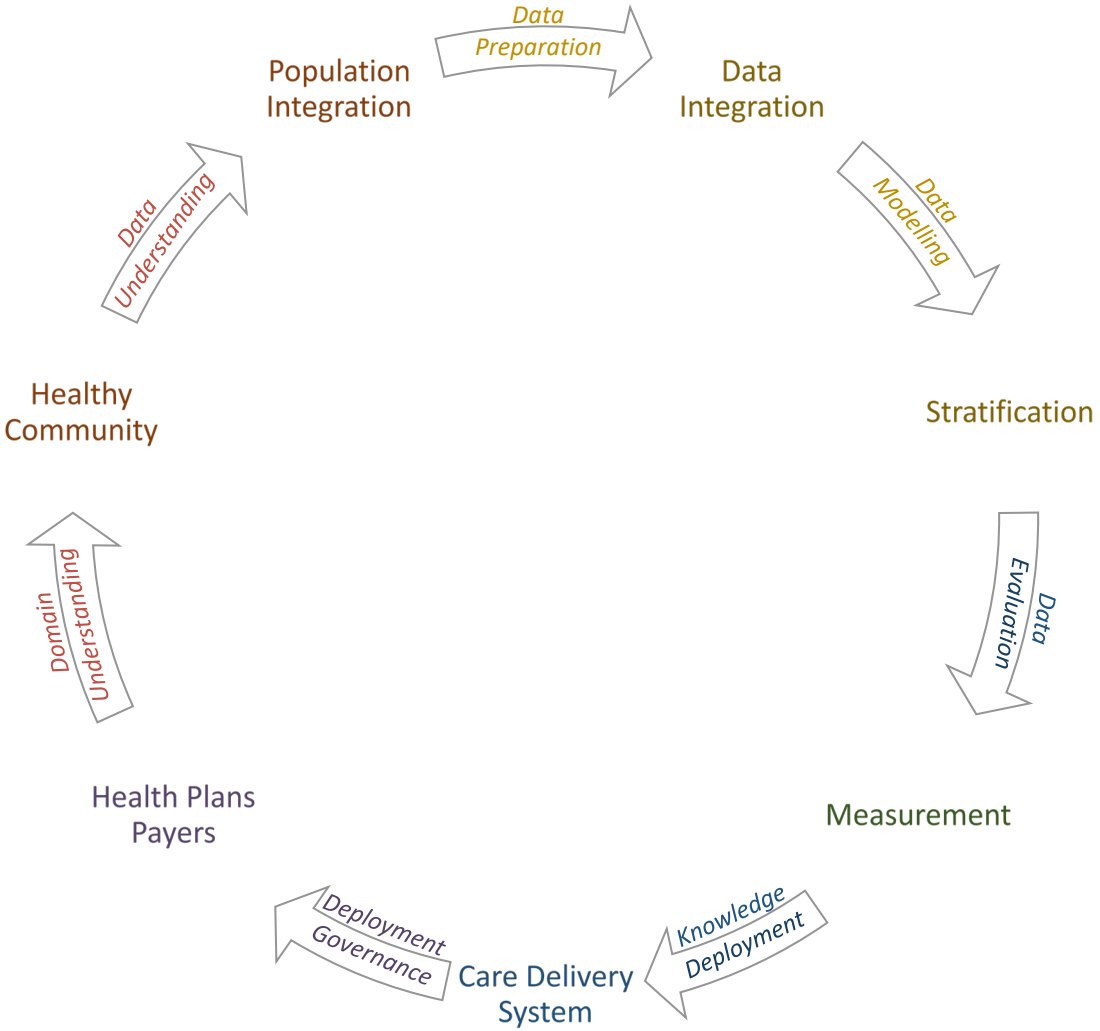
Translational Data Science... for Population Health



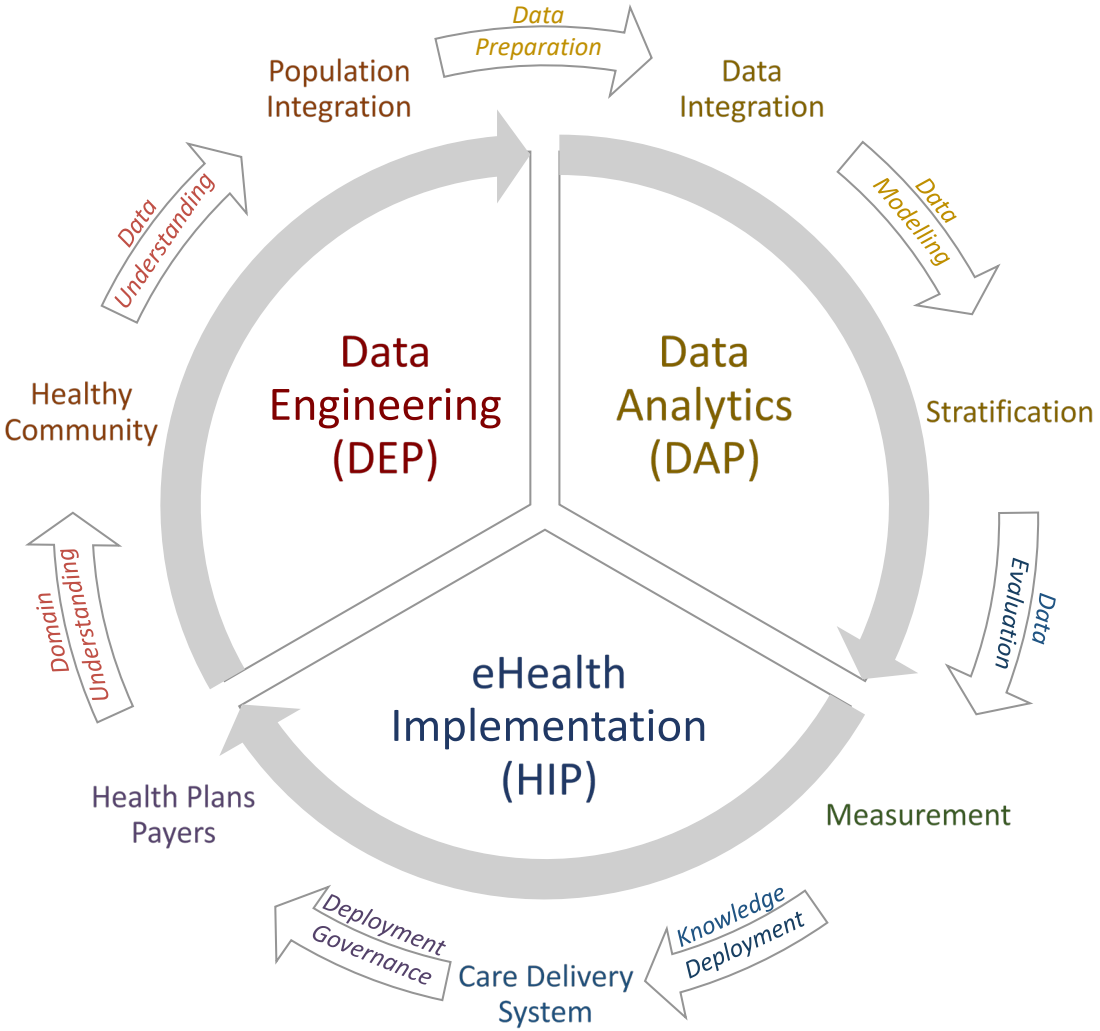
CRISP-DM

PHM

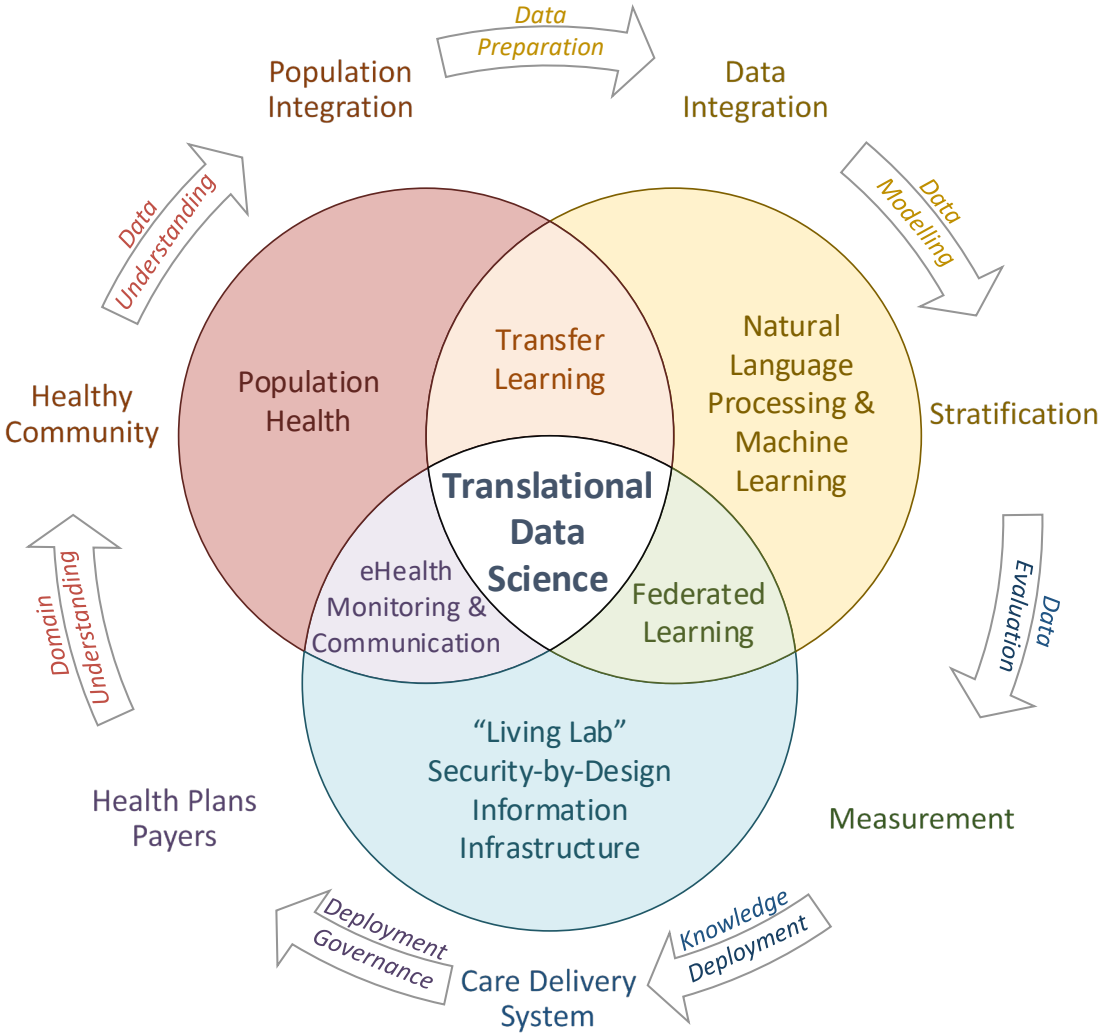
Translational Data Science for Population Health



Translational Data Science for Population Health

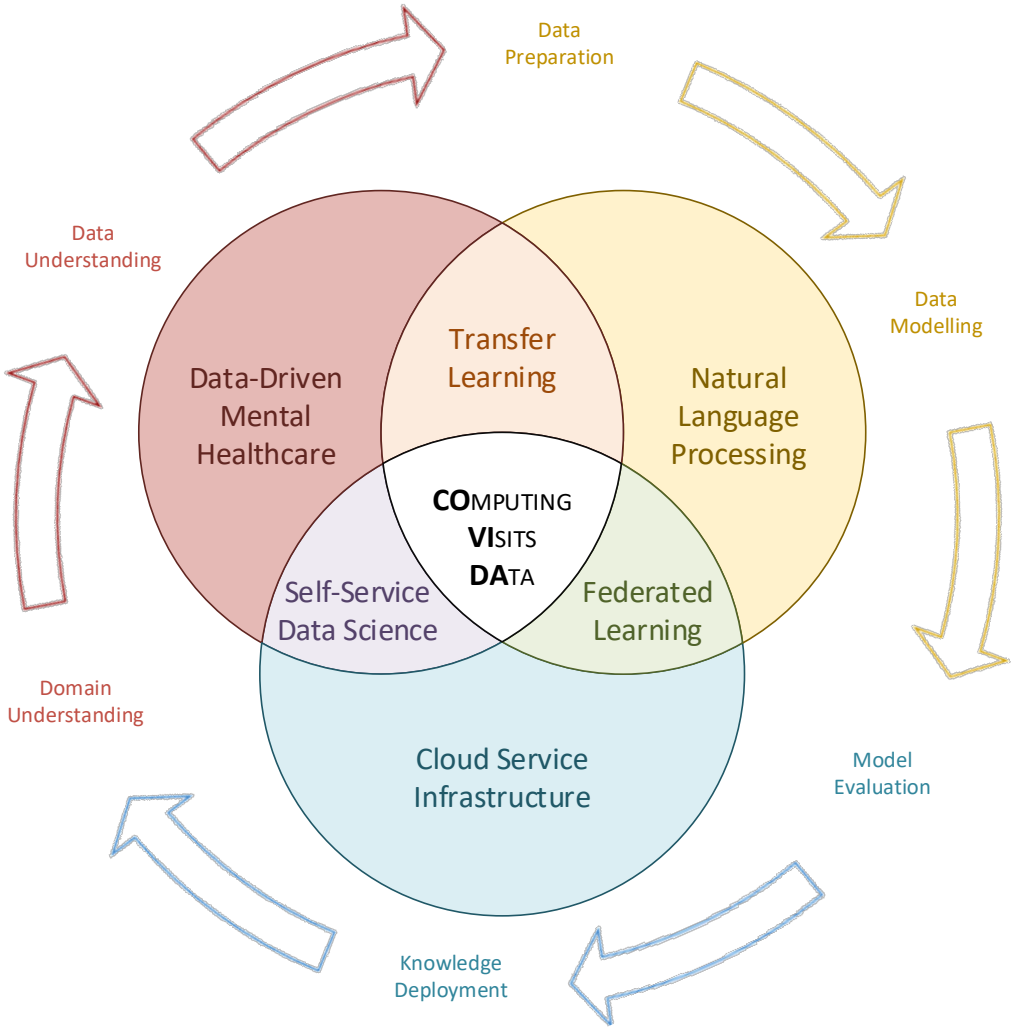


Translational Data Science for Population Health



Context: COVIDA

COVIDA investigates Transfer Learning for **Natural Language Processing** technologies with Federated Learning architectures for Self-Service Data Science in daily **Mental Healthcare** to enable medical practioners throughout the Dutch language area to safely and reliably reuse their daily clinical notes by nurses and doctors from patients' EHRs to predict inpatient violence risks, depression, and more.



url: https://bit.ly/covida-poster

Agenda

- *Introductions*
  - Me, Translational data science, COVIDA

1. 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. [online]

2. Prediction understanding

   Mosteiro,P., Rijcken,E., Zervanou,K., Kaymak,U., Scheepers,F., & Spruit,M. (2021). Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes. *Journal of Artificial Intelligence for Medical Sciences*. [url]

Translational Data Science

Natural Language Processing in Mental Healthcare

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

Data Under-standing

1. UMCU
2. Antes

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]

Data Under-standing

What is a Clinical Note?

(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."

Data Preparation

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

- *paragraph2vec*
- *SVM classifier*

- Data also included:
  - admissions
  - incidents

(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

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

ervaarde verdwenen zijn. Mw ging na 4.00 uur weer naar bed en kwa[illegible]niet meer uit haar kamer tot de ochtend."

[Positive, Negative]

van mijn been" […]

Modelling: Prediction performance

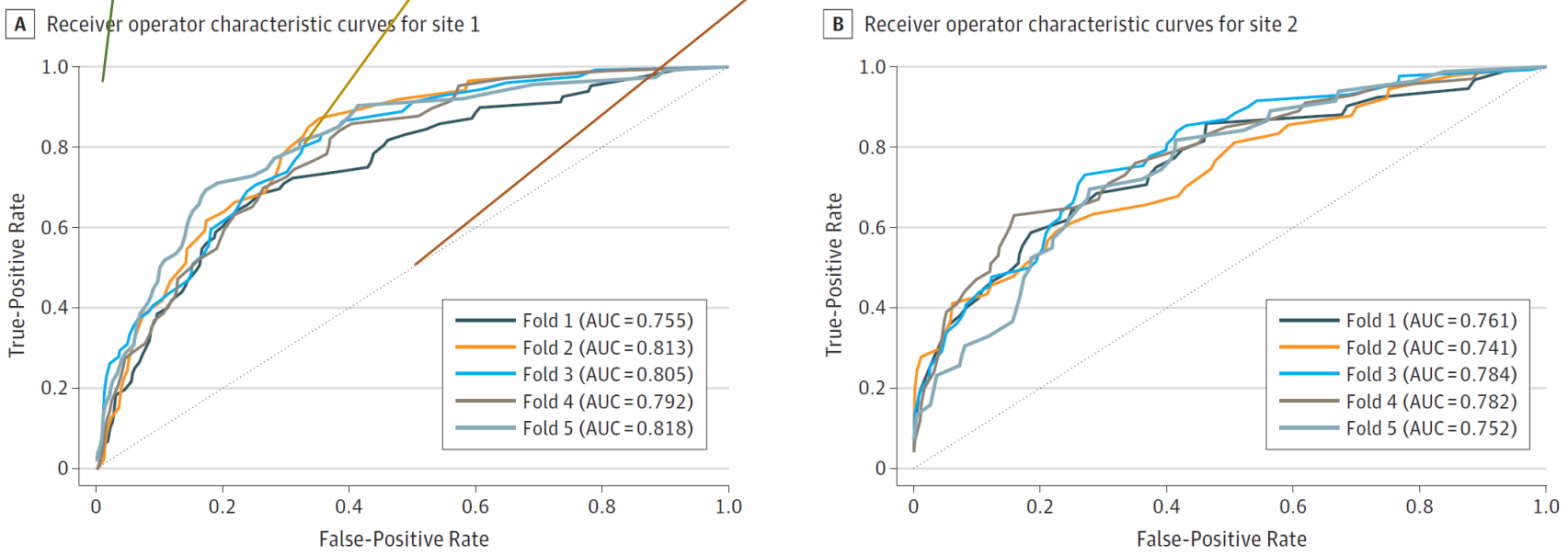
(SVM classifier)

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.

Evaluation: Exploratory Analysis

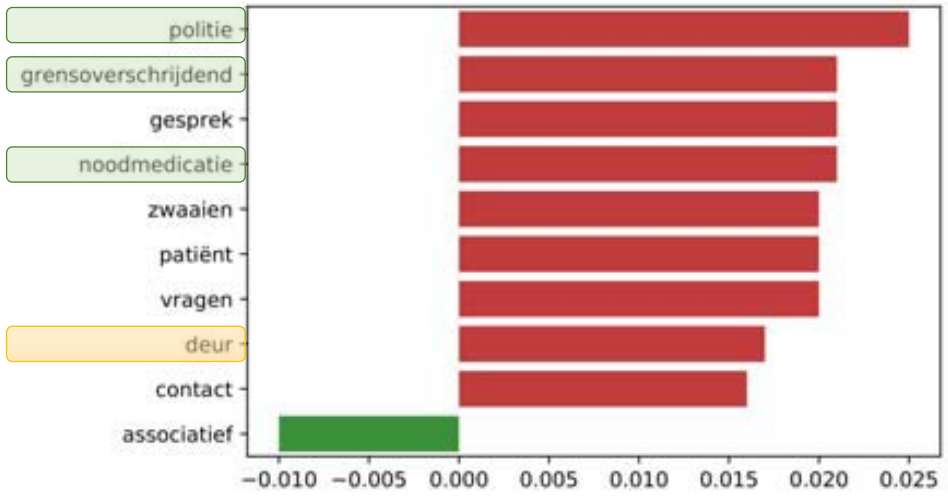
Table 3. Results of Exploratory Analysis

	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

Abbreviation: MCC, Matthews correlation coefficient.

[d] *P* values derived from χ^2 test, and a Holm-Bonferroni correction was applied to obtain

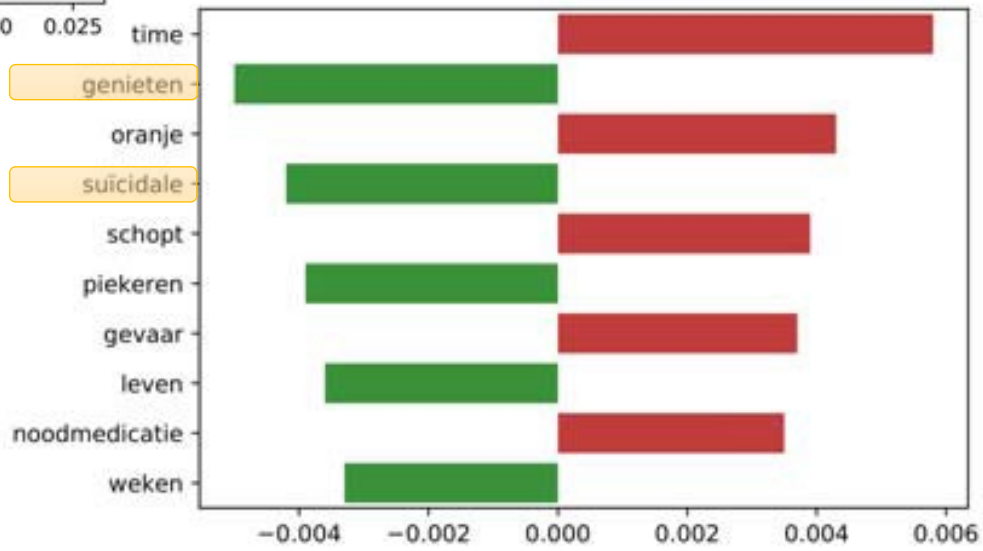
Evaluation: Model Explainability



Sample of Local Explanation predicting **high risk of aggression**

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

Sample of Local Explanation predicing **low risk of aggression**



Deployment: GitHub Repository...

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

Agenda

- *Introductions*
  - Me, Translational data science, COVIDA

1. 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. [online]

2. Prediction understanding

Mosteiro,P., Rijcken,E., Zervanou,K., Kaymak,U., Scheepers,F., & Spruit,M. (2021). Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes. *Journal of Artificial Intelligence for Medical Sciences*. [url]

Translational Data Science

Natural Language Processing in Mental Healthcare

Integrate Structured Data?

- Combining unstructured and structured EHR data

Words	Gender	Administrations	Prescriptions	Diagnosed	Age
1510	Male	10	9	0	56
1216	Female	11	16	1	42
1295	Female	10	14	0	46
1063	Male	0	3	0	17
1461	Female	0	3	0	20

Version	ROC-AUC
Baseline (BVC)	0.743 ± 0.109
Doc2Vec + SVM	0.792 ± 0.034
Doc2Vec + Random Forest	0.782 ± 0.030
Doc2Vec + Random Forest + Structured Var's	0.777 ± 0.026

Represent Text more interpretable?

- ~~doc2vec~~ → Latent Dirichlet Allocation (LDA) topic modeling

- A topic is a linear combination of words that together represent that topic
  - Topic3: [0.035 * patient + 0.029 * nurse + 0.015 * door + …]
- A clinical note is a linear combination of topics
  - [(0.028 * Topic1) + (0.012 * Topic2) + (0.143 * Topic3) + … + (0.031 * Topic25)]

Version	ROC-AUC
Baseline (BVC)	0.743 ± 0.109
Doc2Vec + SVM	0.792 ± 0.034
Doc2Vec + Random Forest	0.782 ± 0.030
Doc2Vec + Random Forest + Structured Var's	0.777 ± 0.026
LDA + Random Forest	0.785 ± 0.038

Employ  modelling techniques?

"*Dutch* is our problem" → BERTje, no ClinicalBERT for Dutch, etc.

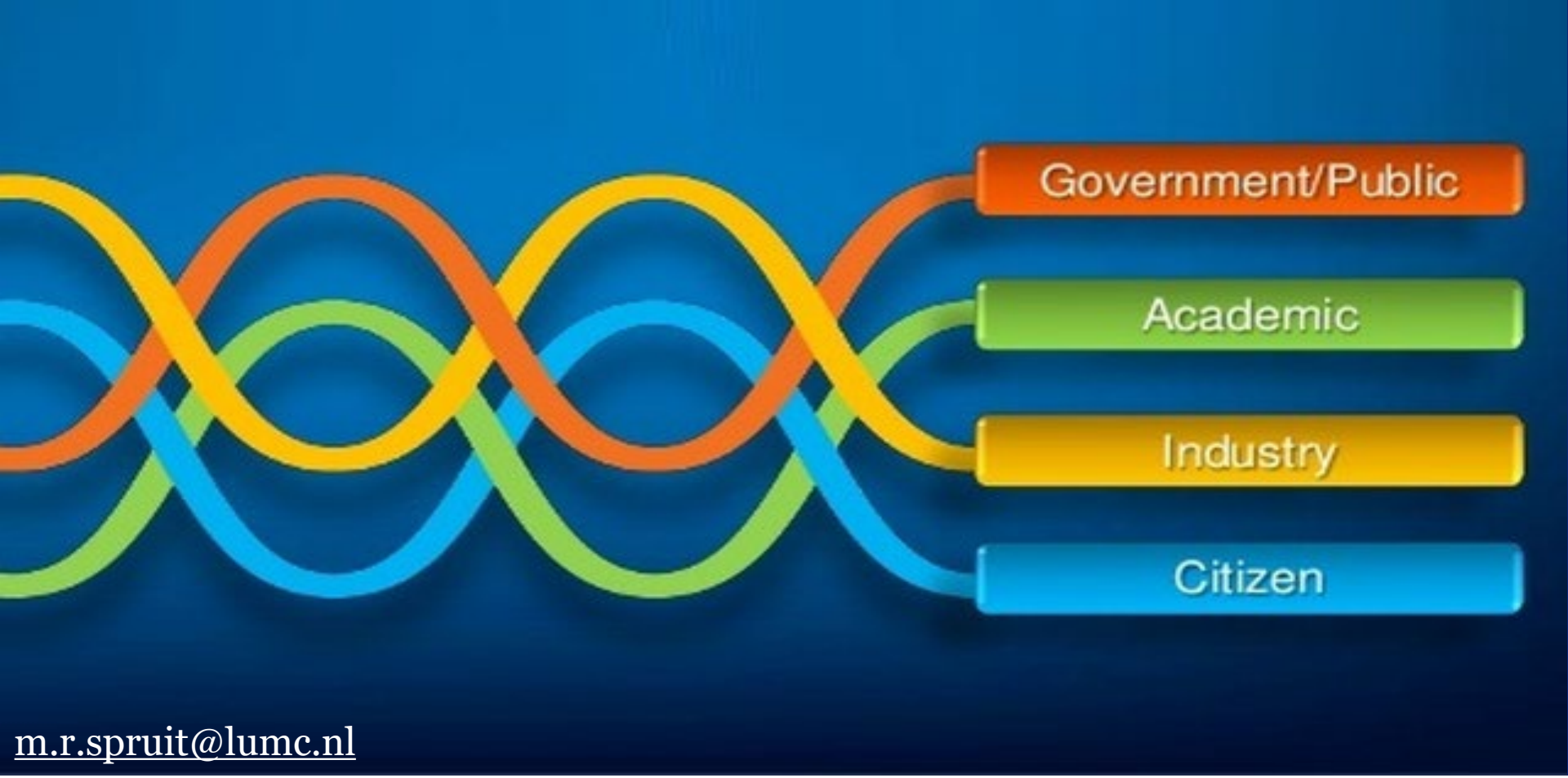
Version	ROC-AUC
Baseline (BVC)	0.743 ± 0.109
Doc2Vec + SVM	0.792 ± 0.034
Doc2Vec + Random Forest	0.782 ± 0.030
Doc2Vec + Random Forest + Structured Var's	0.777 ± 0.026
LDA + Random Forest	0.785 ± 0.038
BERTje	0.662 ± 0.039

Wrap up

- VRA with ML is promising, but
  - More data is needed
  - Need to fix labeling (violence, anonymization)
  - No improvement from using deep learning

- How to obtain more data?
  - Federated learning to cope with data scarcity

- Want to know more?
  - Attend Dr. Pablo Mosteiro's talk at the SIG Health Data Science seminar
  - on Thursday 26 June 13:15 CEST
  - https://www.universiteitleiden.nl/en/research/research-projects/data-science-research-programme/sig-health-data-science



m.r.spruit@lumc.nl

Leiden University
Campus The Hague

L U
M C
Leiden University
Medical Center

liacs
Leiden Institute of
Advanced
Computer
Science