

Health  
Campus

Den  
Haag

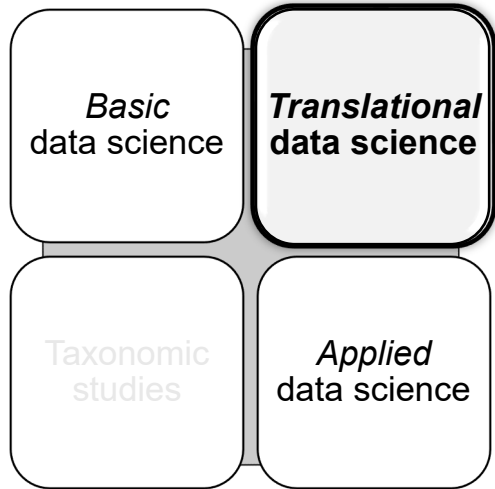
# Translational Data Science in Population Health

*...Data Techniques and Methodology for  
Violence as a Public Health Problem*

KIEM Pressure Cooker Workshop, 11 March 2024, Marco Spruit

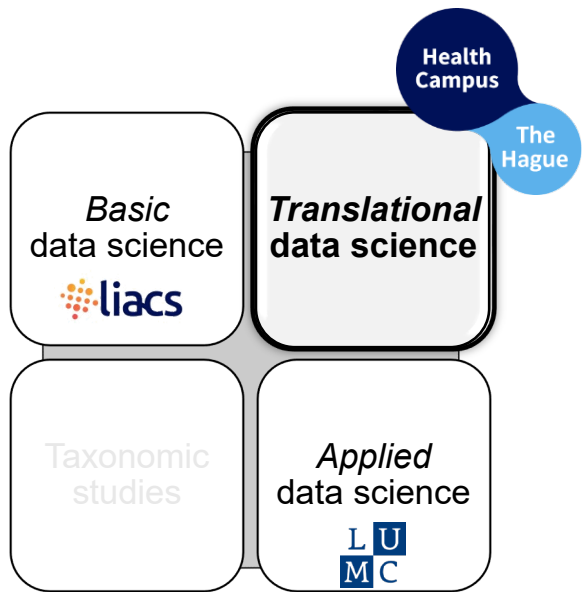


*degree of fundamental understanding*



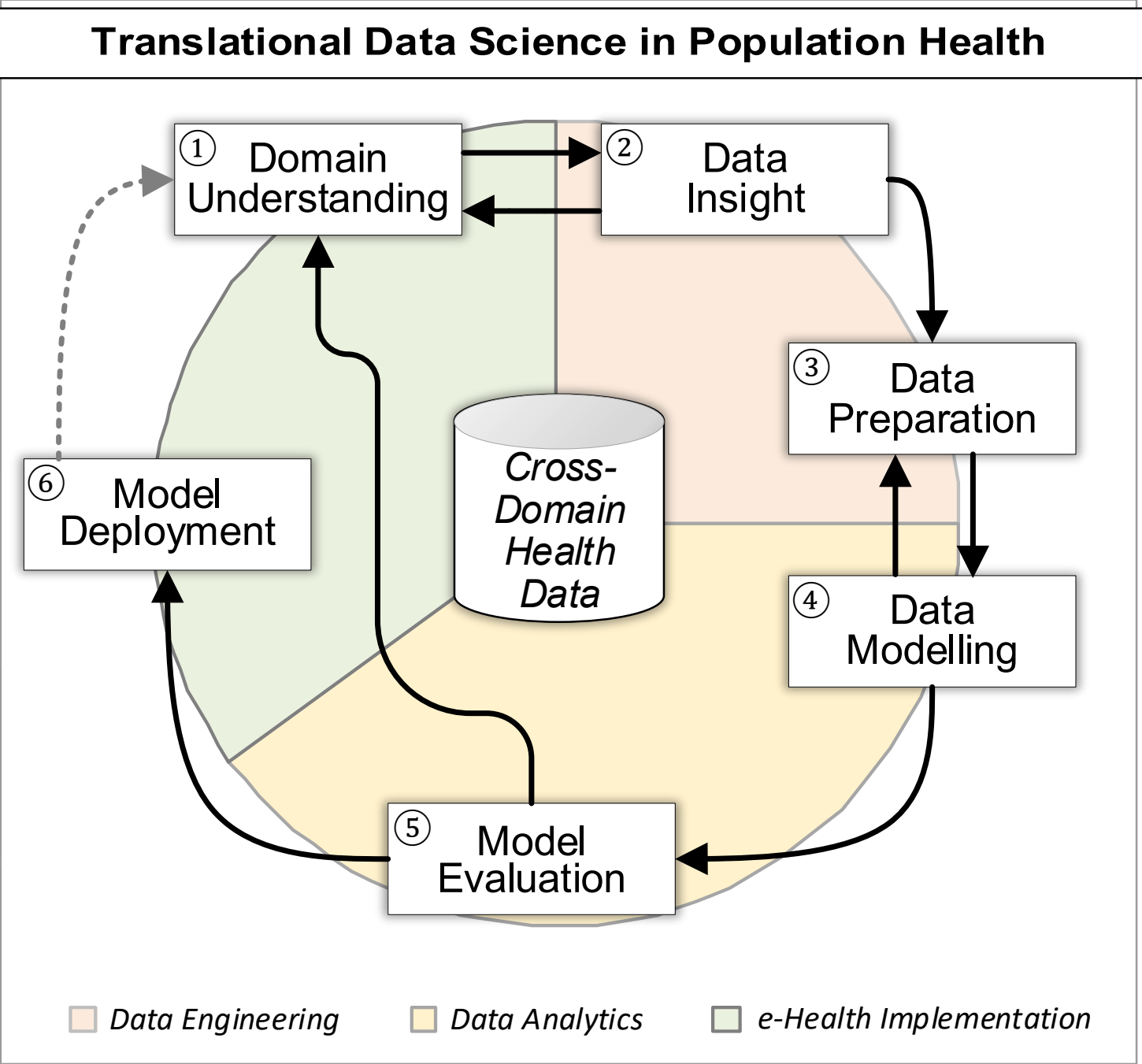
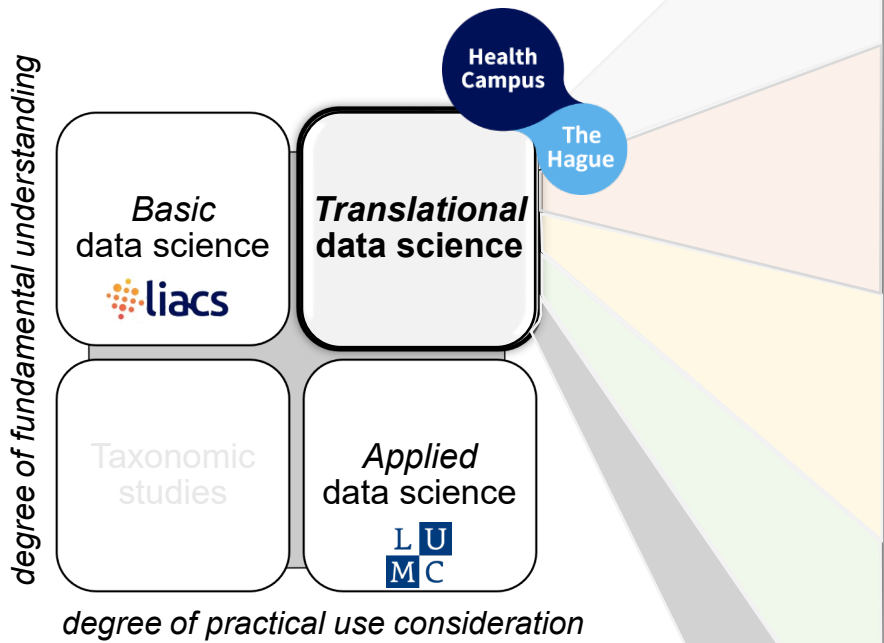
*degree of practical use consideration*

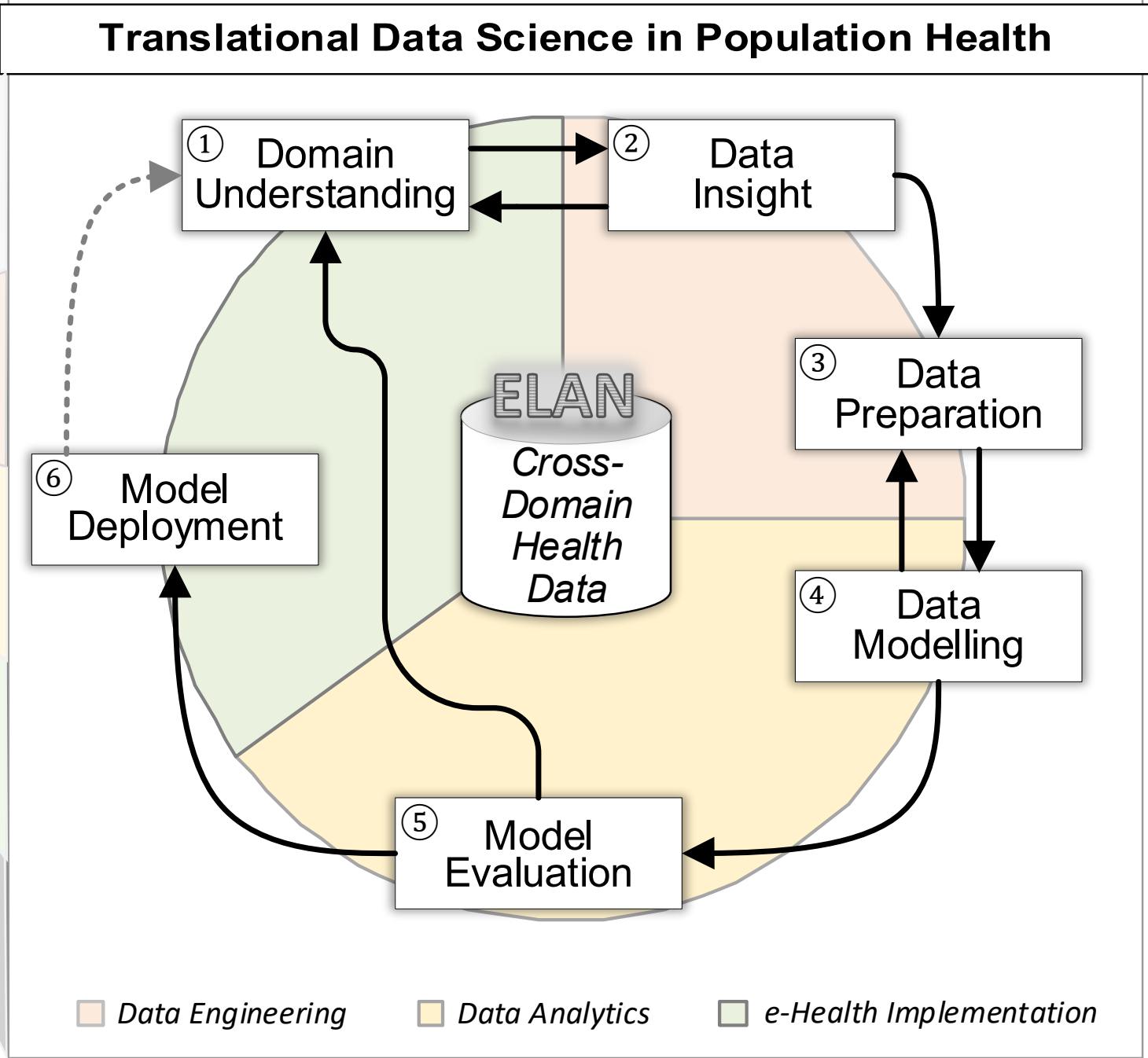
degree of fundamental understanding



degree of practical use consideration

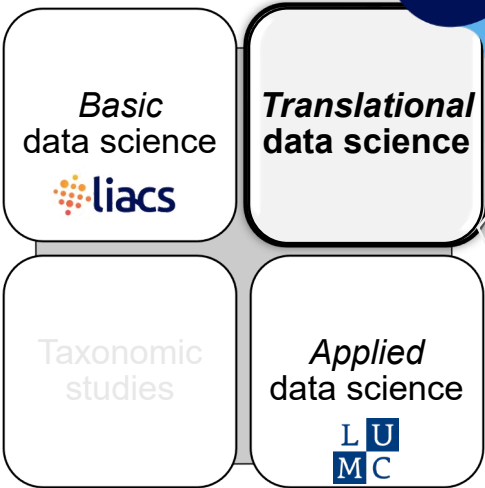
# Translational Data Science in Population Health

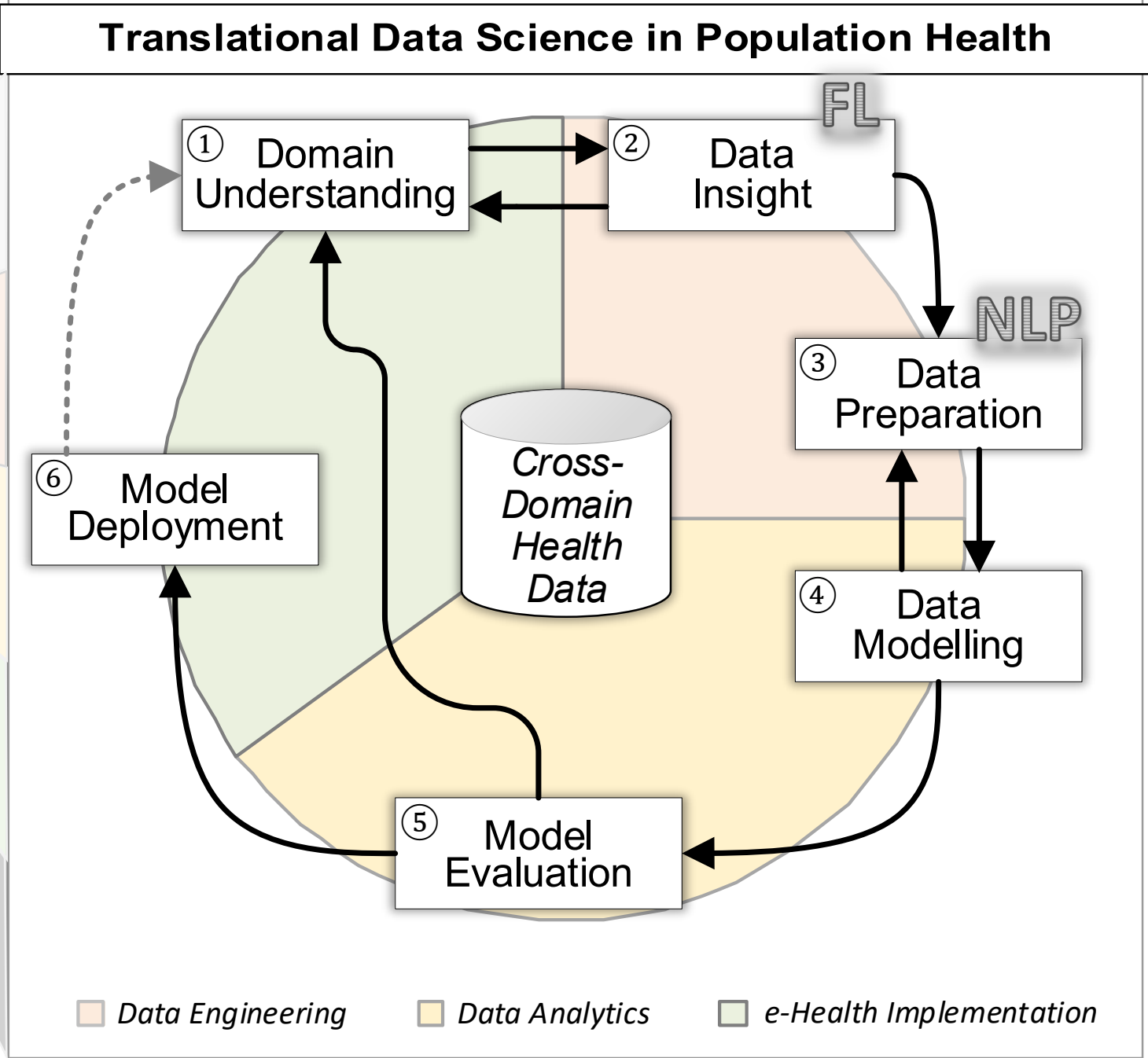




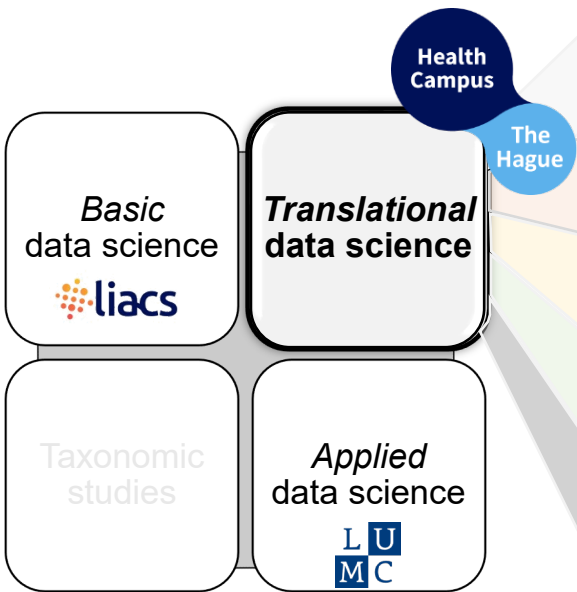
degree of fundamental understanding

degree of practical use consideration





degree of fundamental understanding



degree of practical use consideration

degree of fundamental understanding

BERT-based Dutch NLP on sloppy informal medical text snippets

**Translational data science**

Taxonomic studies

Lifestyle information extraction for personalised prognoses

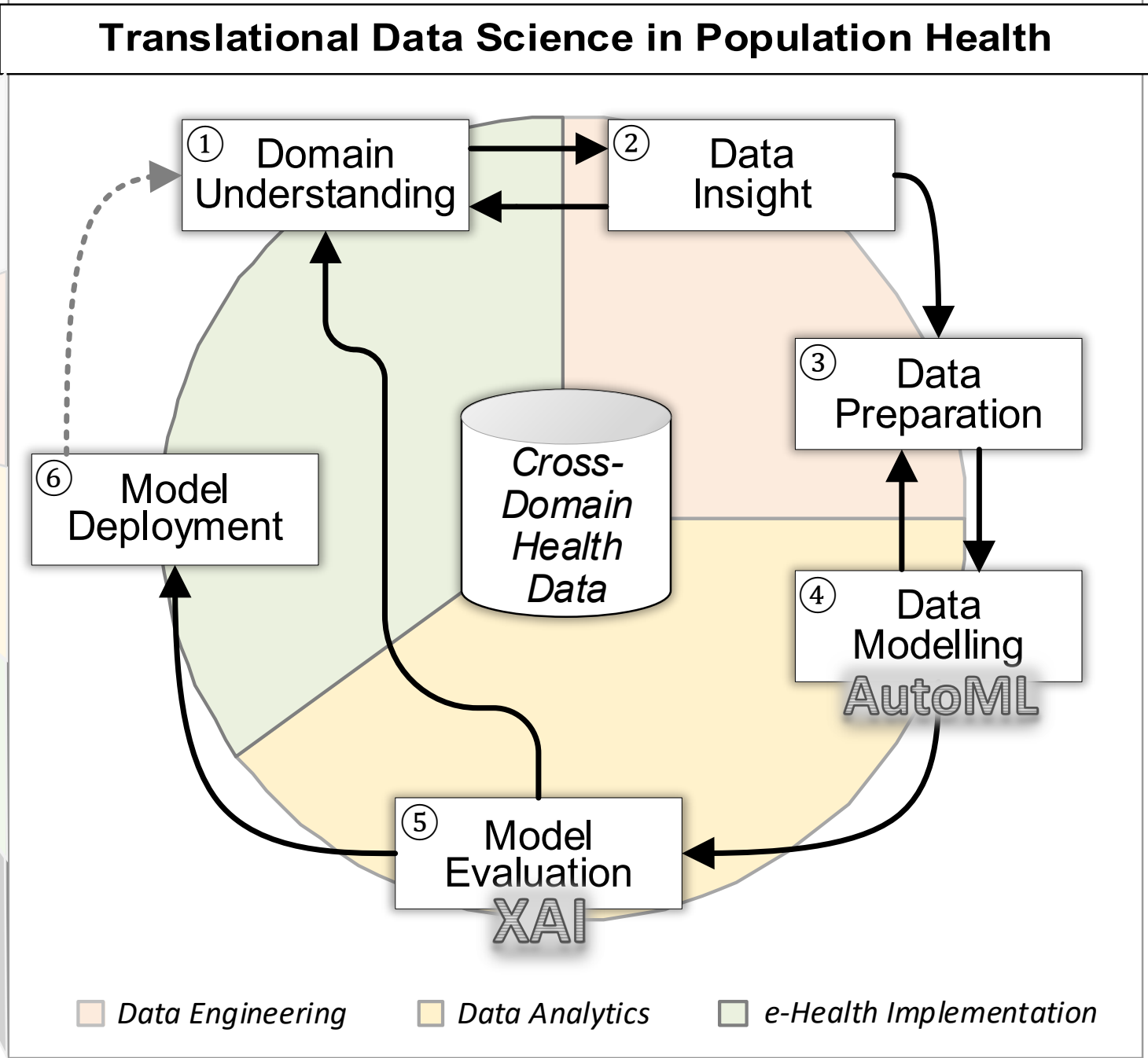
degree of practical use consideration

## Example: EHR Lifestyle Characteristics Extraction

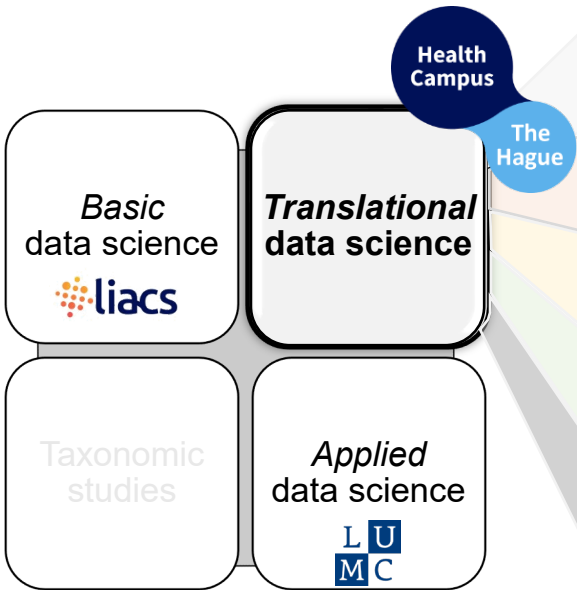
- Muizelaar, Haas, Dortmund, Putten & Spruit. Extracting Patient Lifestyle Characteristics from Dutch Clinical Text with BERT Models ([under review](#))

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

Model	Smoking	Alcohol	Drugs
String Matching	0.84	0.74	0.68
Machine Learning (SGD)	0.85	0.71	0.60
HAGALBERT	0.66	0.54	0.43
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
BioBERT (translated)	0.91	0.72	0.52
ClinicalBERT (translated)	0.92	0.80	0.61



degree of fundamental understanding



degree of practical use consideration



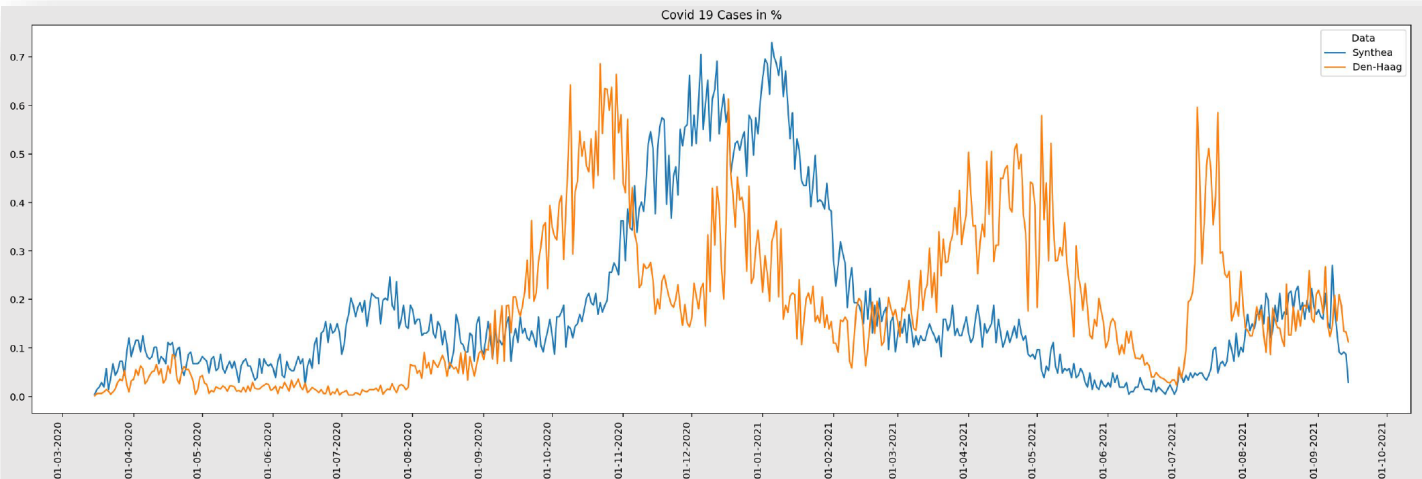
# Example: Virtual Patients and Population dataset

## ABM approach (Ammar Faiq)

- Synthea-based dataset  
An ELAN 'digital twin' is already being used in the PHM  
Fundamentals master course to let students analyse COVID outbreaks in The Hague region (see below)
- Joint research with Statistics Netherlands (CBS) & Syntho
- Workshop 'Guidance Ethics': many stakeholders, 50+ effects

## CGAN approach (Jim Achterberg)

- [Thesis](#)  
Evaluation Framework for synthetic EHR data (supporting heterogeneous types, time series, unpredictable quality)
  - tSNE extension
  - two-sample GoF test
  - evaluation metric for privacy risk through AiAs
- Horizon Europe, NWO OSF



degree of fundamental understanding

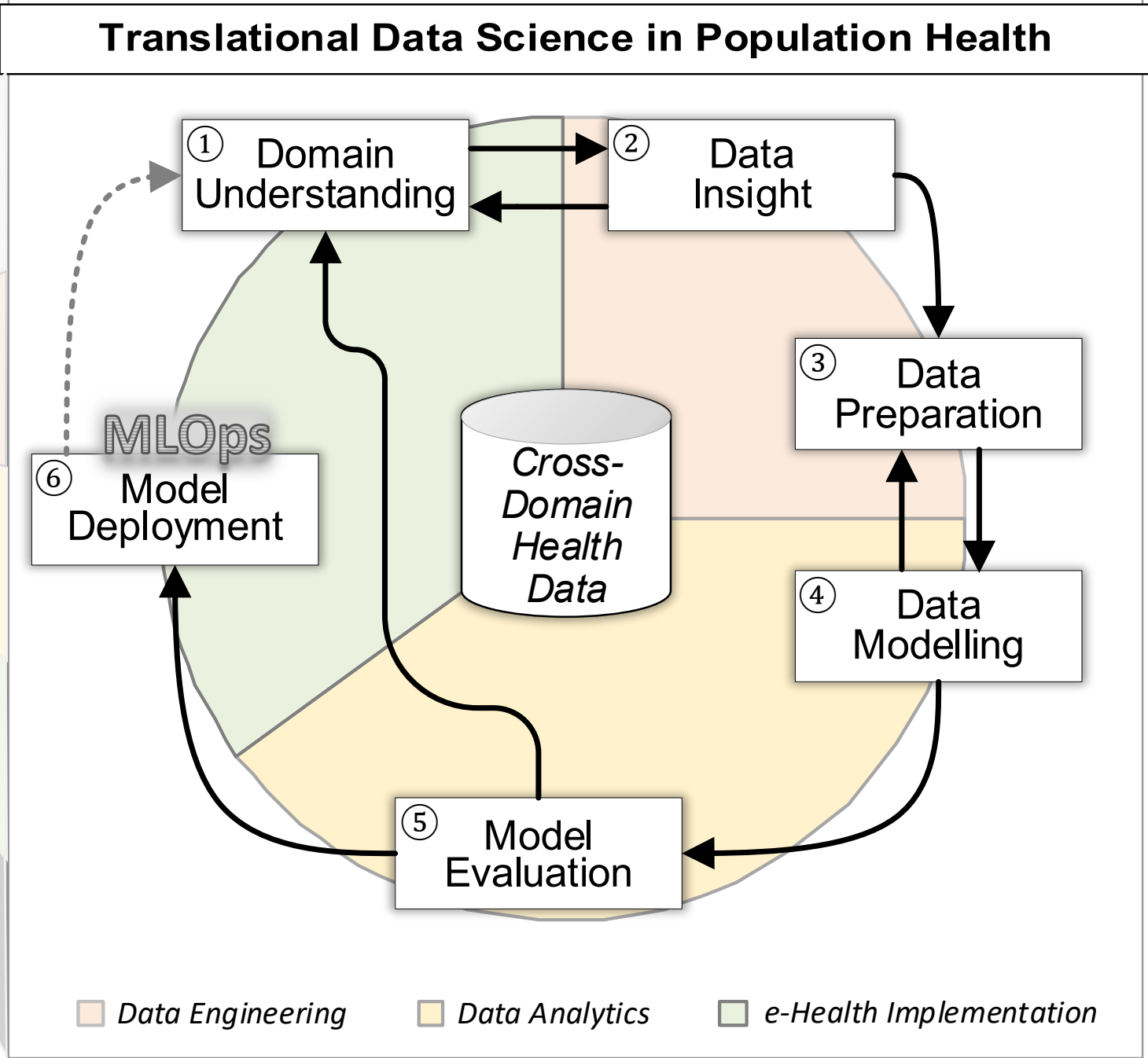
A CGAN vs ABM benchmark for synthetic EHR data

**Translational data science**

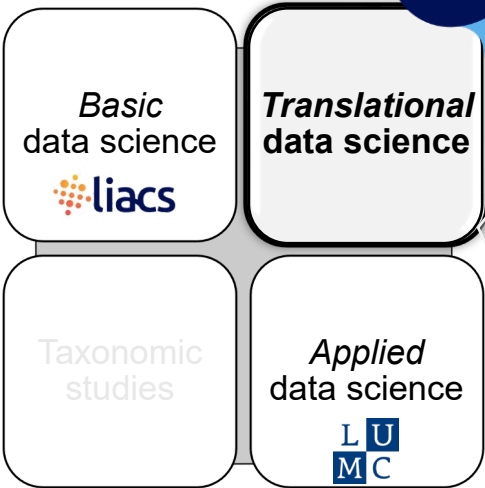
Taxonomic studies

A realistic data twin for faster research and better education

degree of practical use consideration



degree of fundamental understanding



degree of practical use consideration

# Example: Prediction Model Deployment for CDSS

Lisanne Wallaard

[GitHub](#)

[Thesis](#)

[Demo](#)

## Feature Selection



Race

American Indian/Alaskan Native

Sex

Female

Age category

18-24

BMI category

Normal weight (18.5 <= BMI < 25.0)

How many hours on average do you sleep?

7

How can you define your general health?

Excellent

## Heart Disease Prediction

Are you wondering about the condition of your heart? This app will help you to diagnose it!



I'll help you diagnose your heart health! - Dr. Logistic Regression

Predict

Did you know that machine learning models can help you predict heart disease pretty accurately? In this app, you can estimate your chance of heart disease (yes/no) in seconds!

Here, a logistic regression model using an undersampling technique was constructed using survey data of over 300k US residents from the year 2020. This application is based on it because it has proven to be better than the random forest (it achieves an accuracy of about 80%, which is quite good).

To predict your heart disease status, simply follow the steps bellow:

1. Enter the parameters that best describe you;
2. Press the "Predict" button and wait for the result.

**Keep in mind that this results is not equivalent to a medical diagnosis! This model would never be adopted by health care facilities because of its less than perfect accuracy, so if you have any problems, consult a human doctor.**

Author: Kamil Pytlak ([GitHub](#))

You can see the steps of building the model, evaluating it, and cleaning the data itself on my GitHub repo [here](#).

degree of fundamental understanding

Feasibility of ML model deployment from within a highly secure sandbox env

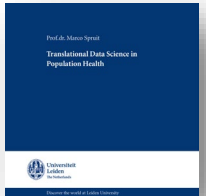
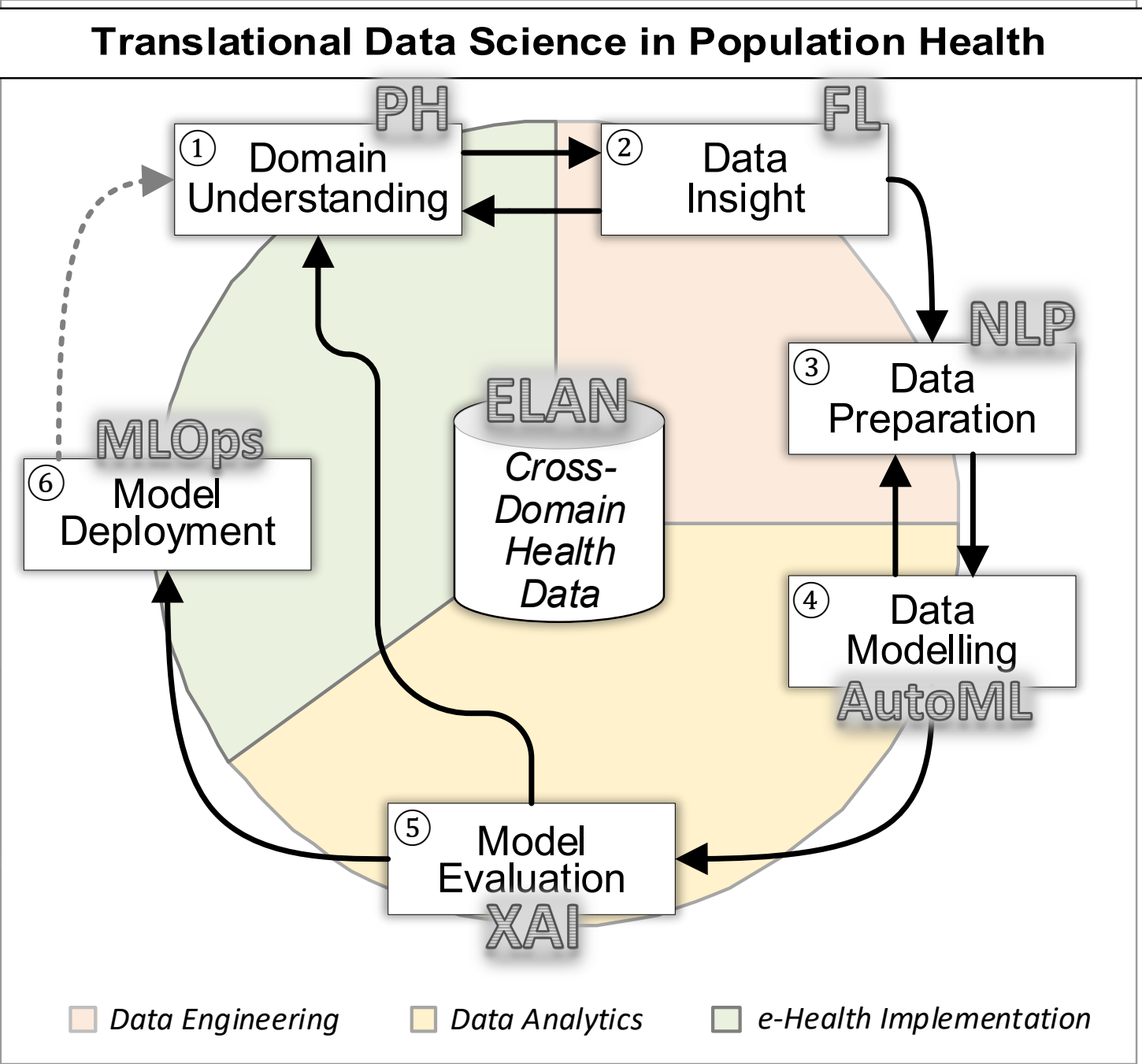
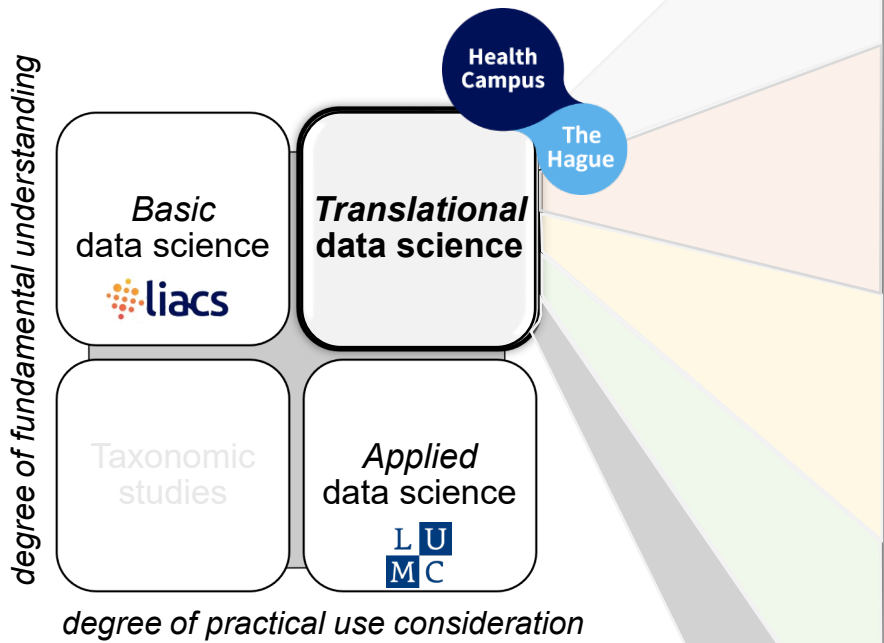
Translational data science

Taxonomic studies

Make ELAN ML models available on-demand to GP practices

degree of practical use consideration


# Translational Data Science in Population Health




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

# Translational Data Science LAB : members d.d. 14-12-2023


Dr. Armel Lefebvre  
Postdoc (LUMC/LIACS)  
*Research Data Management*




Friso van Dijk  
PhD candidate (UU)  
*Privacy Governance*




Els Roorda  
PhD candidate LUMC  
*Population Health Information Systems*





Prof.dr. Marco Spruit  
Full Professor  
*Translational Data Science (LUMC/LIACS)*




Dr. Marcel Haas  
Assistant prof. (LUMC)  
*Health Data Science*



Max van Haastrecht  
PhD candidate (LIACS)  
*Cybersecurity Risk Modelling & Validation*


Ian Shen  
PhD candidate (LIACS)  
*Healthcare and Open Science Engineering*



Samar Samir  
PhD candidate (LIACS)  
*Federated NLP in Mental health*




Bram van Dijk  
PhD candidate (LIACS)  
*Mindreading with NLP*




Emil Rijcken  
PhD candidate (TUE)  
*Topic Modelling in NLP*



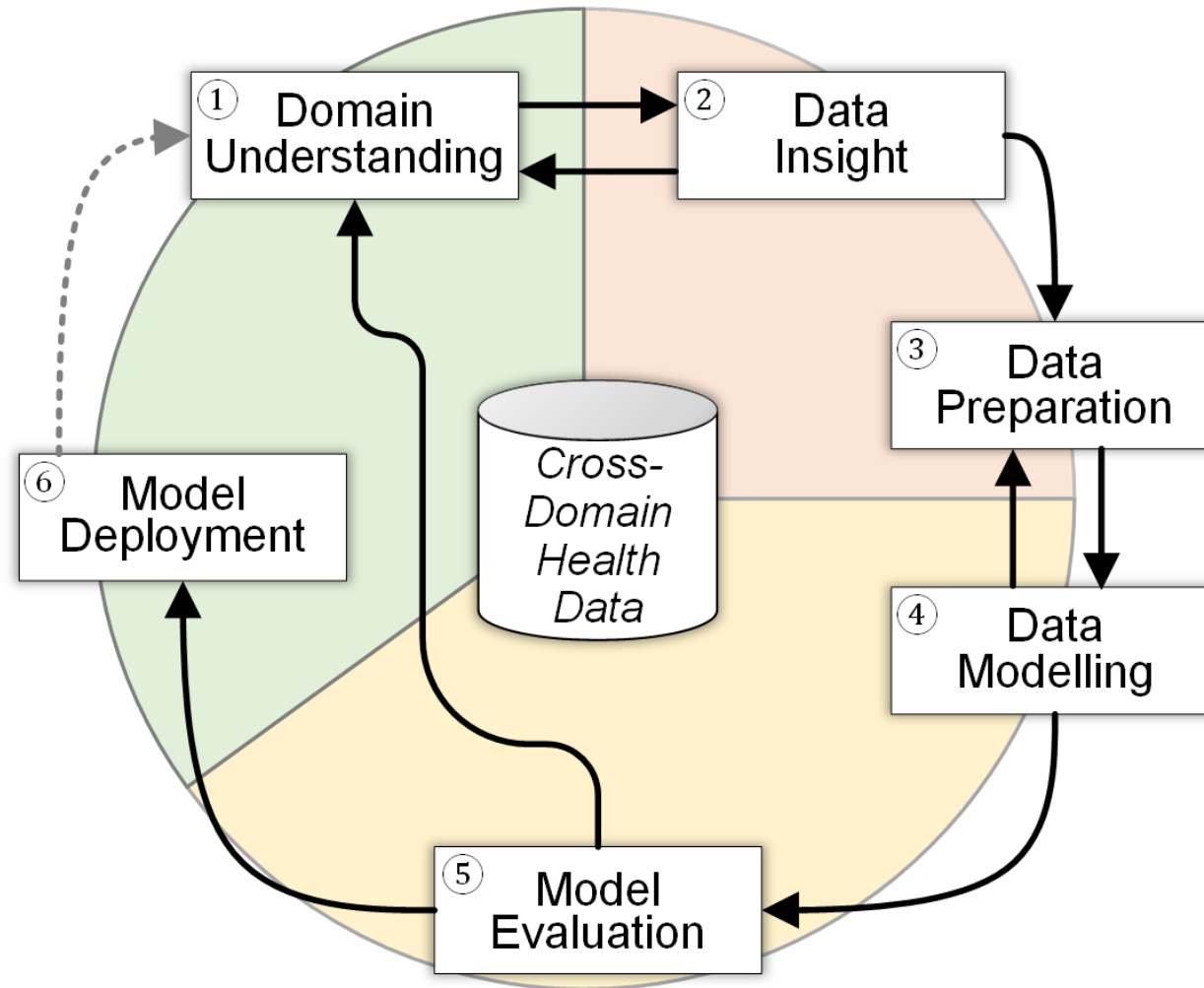
Hielke Muizelaar  
PhD candidate (LUMC)  
*NLP/ML for Patient Segmentation Modelling*



Sukainah Alfaraj  
PhD candidate (LUMC)  
*Risk Prediction of Diabetes progression*



Jim Achterberg  
PhD candidate (LUMC)  
*Synthetic HTA data Generation & Validation*



■ Data Engineering    
 ■ Data Analytics    
 ■ e-Health Implementation