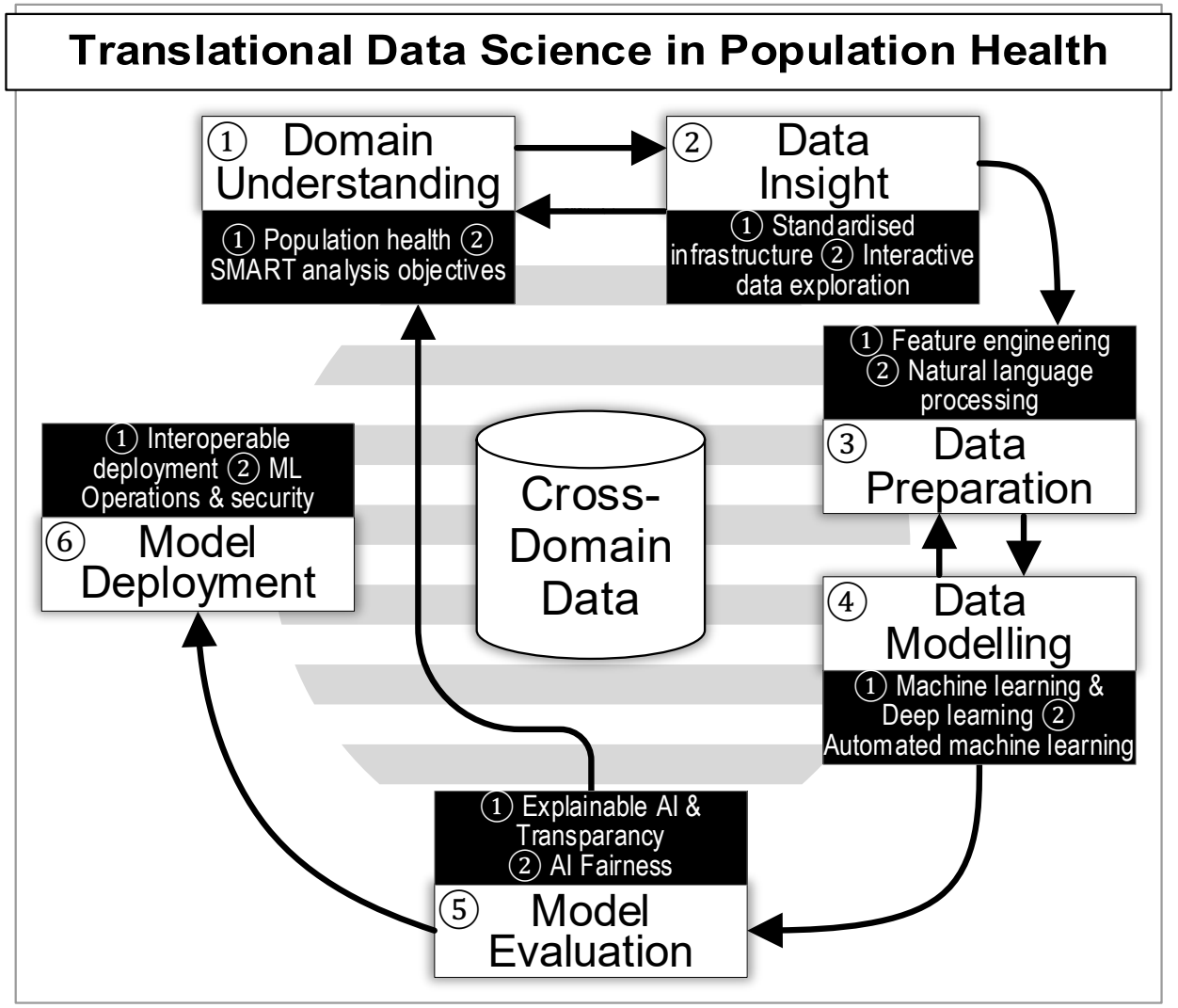
Translational Data Science

Data Science in Dutch Healthcare

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

Lecture: *Data Science Honours Class, 1 April 2022*



liacs Leiden Institute of Advanced Computer Science

Universiteit Leiden

LUMC Leiden University Medical Center

CAIRELab Summerschool

Artificial Intelligence: from theory to Value Based Healthcare practice (2021)

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

Title: *From theory to healthcare practice with the knowledge discovery process: A **translational data science** primer*

CAIRELAB

Leiden University Medical Center

About… Marco Spruit

As Engineer

1993
- Information Retrieval programmer
  - ZyLAB Europe BV

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

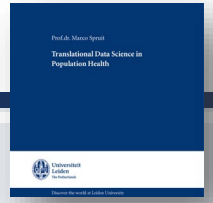
1997
- Product software developer/entrepreneur
  - Insertable Objects, Wizzer BV

As Researcher

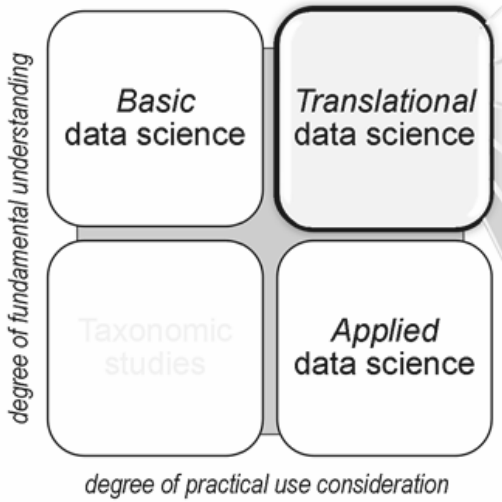
2003
- Ph.D. researcher in Computational Linguistics
  - University of Amsterdam

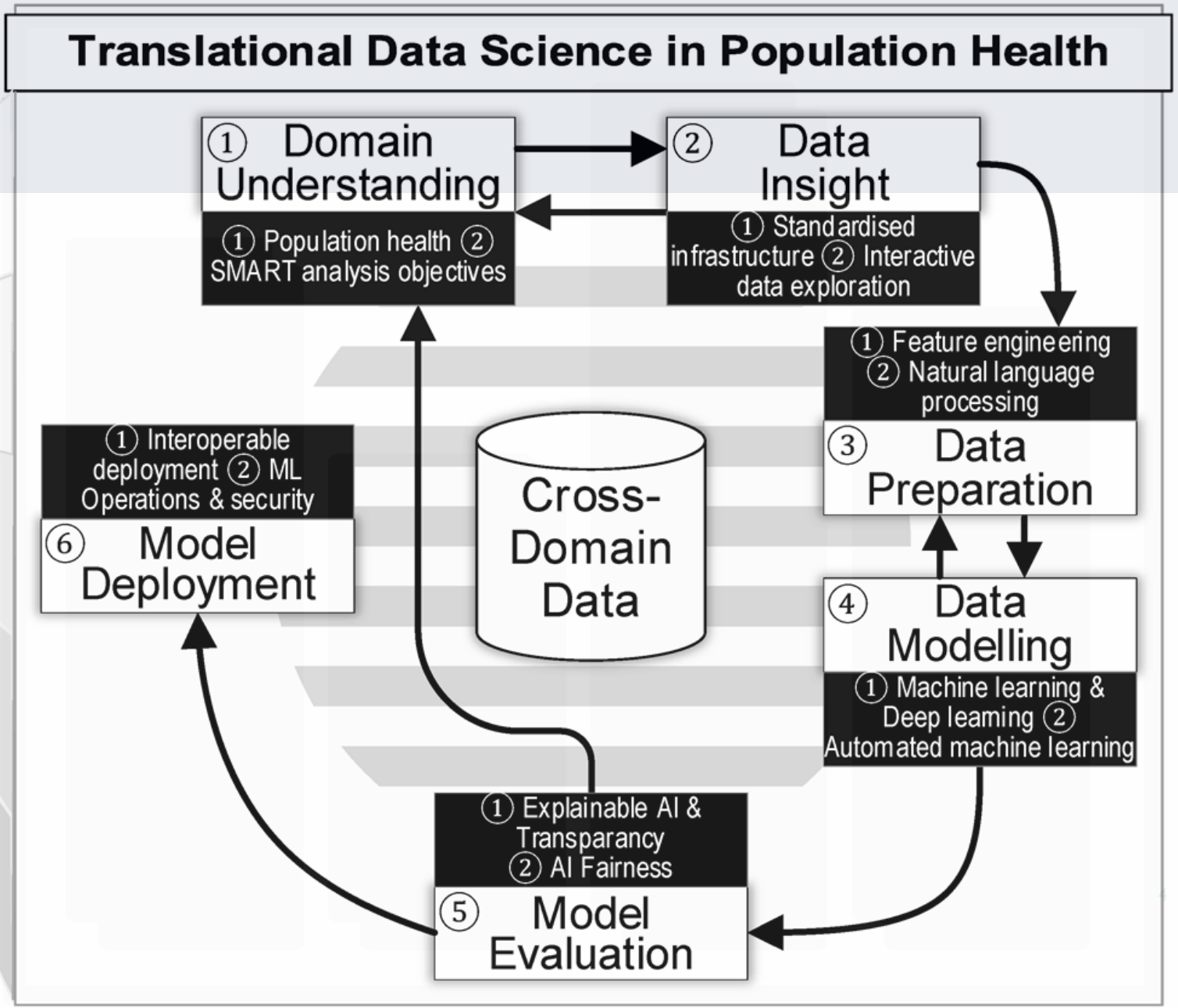
2007
- Assistant/Associate professor Information Science
  - Utrecht University >> Applied Data Science Lab

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



April Fools' Day



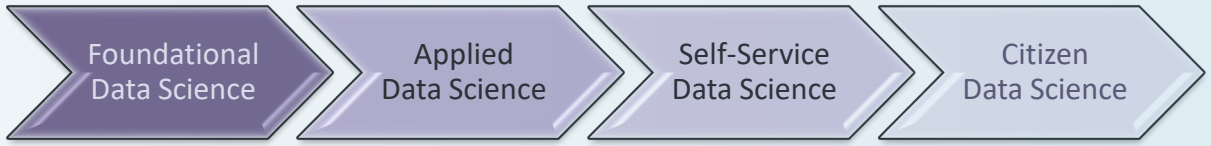


AGENDA

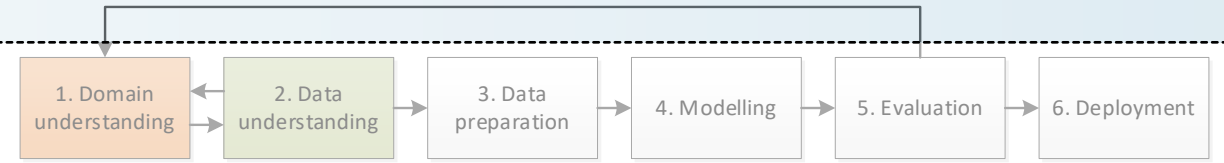
TDS

A. Setting the Scene

From Scientific Method to Translational Data Science



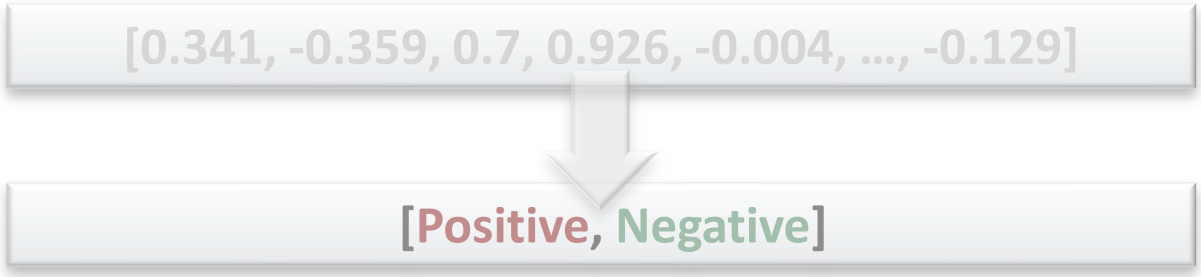
KDP



B. Case study 1

From Information Needs to Data Mining Goals in Long-term Care

NLP

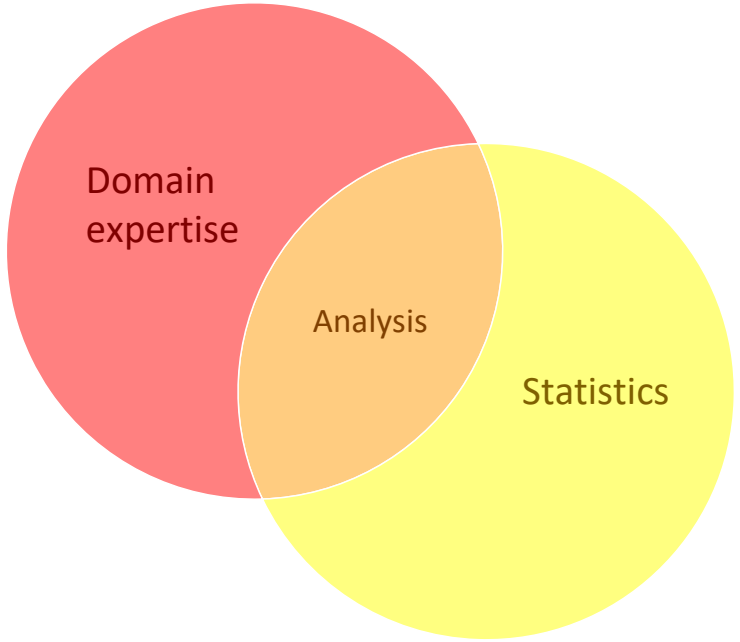


C. Case study 2

Natural Language Processing for Translational Data Science in Mental Healthcare

What is… Science?

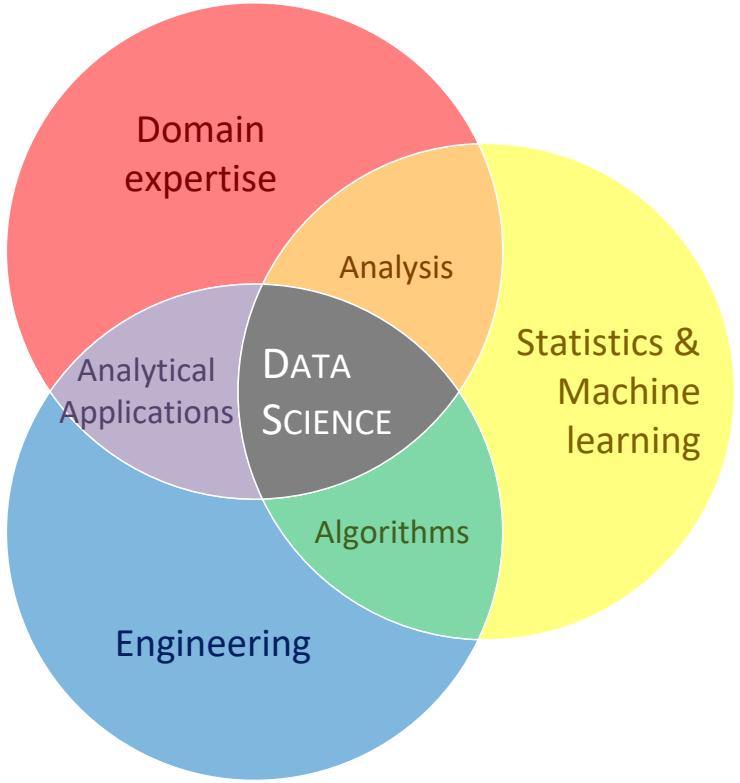
1. Domain expertise
2. Statistics



- Chang, W. L., & Grady, N. (2015). NIST big data interoperability framework: volume 1, big data definitions. [online]

What is… Data Science?

1. Domain expertise
2. Statistics & Machine Learning
3. Engineering
4. Analytical applications

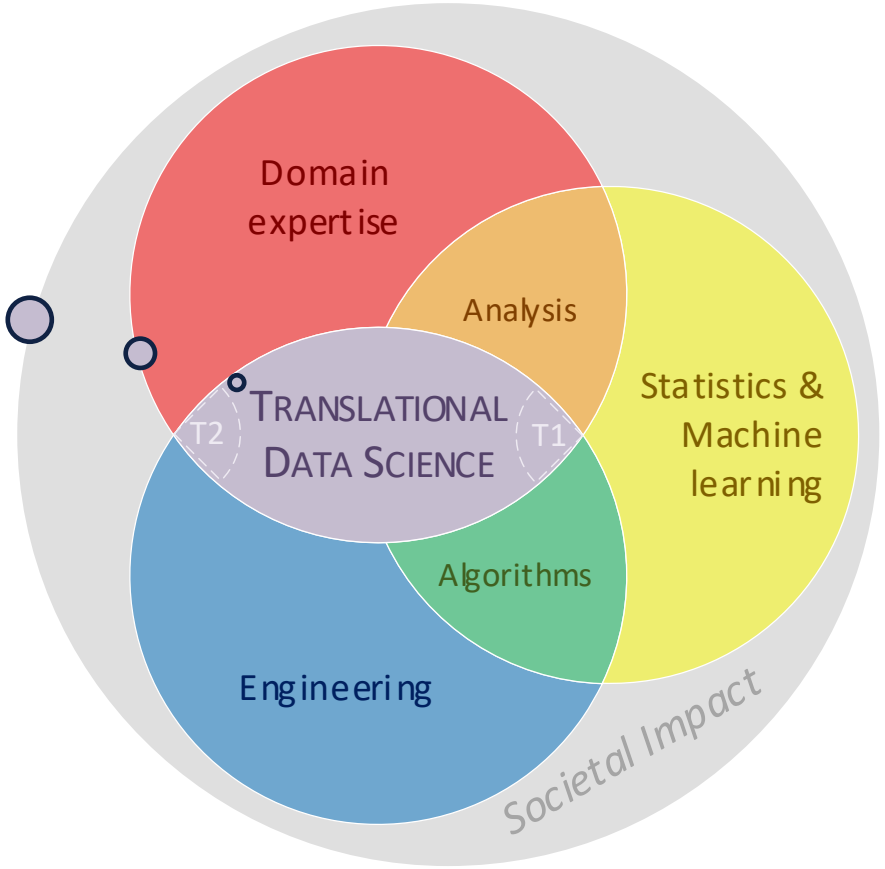


- Chang, W. L., & Grady, N. (2015). NIST big data interoperability framework: volume 1, big data definitions. [online]

What is… Translational Data Science?

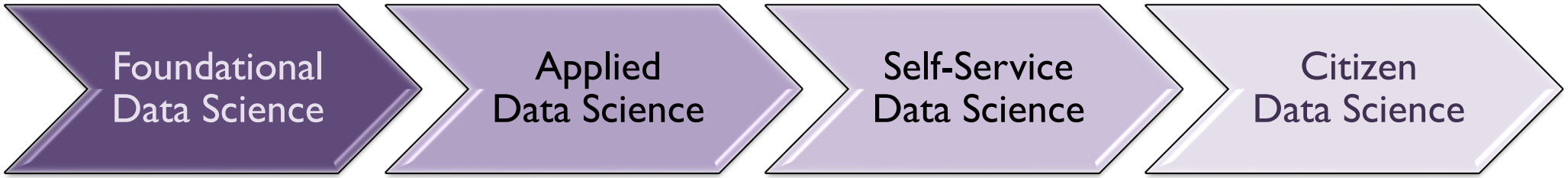
1. Domain expertise
2. Statistics & Machine Learning
3. Engineering
4. Analytical applications
5. Societal impact

CRISP -DM



- Spruit,M., & Jagesar,R. (2016). *Power to the People! Meta-algorithmic modelling in applied data science*. In Fred,A. et al. (Ed.), Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (pp. 400–406). KDIR 2016, November 11-13, 2016, Porto, Portugal: ScitePress. [pdf] [online]

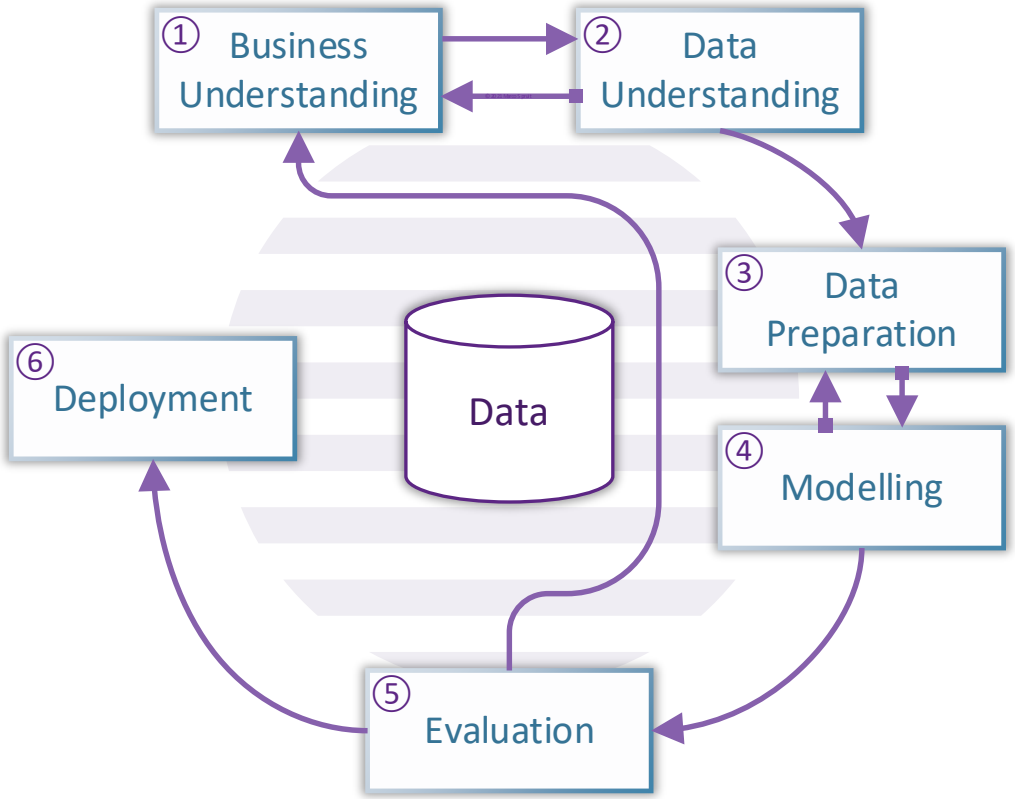
Translational Data Science



Foundational Data Science	Applied Data Science	Self-Service Data Science	Citizen Data Science	
• Basic	• Translational (T1)	• Translational (T2)	• Applied	*Research Type*
• Algorithms	• Meta-Algorithmic Models, Notebooks	• Adaptive Analytic Systems, Dashboards	• N/A	*Research Artifacts*
• Comput. Experiments	• Design Science, CRISP-DM	• Prototyping, CRISP-DM	• CRISP-DM	*Research Methods*
• Data Scientists	• Data Analysts	• Researchers	• Citizens	*Target Audience*

- 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 Process

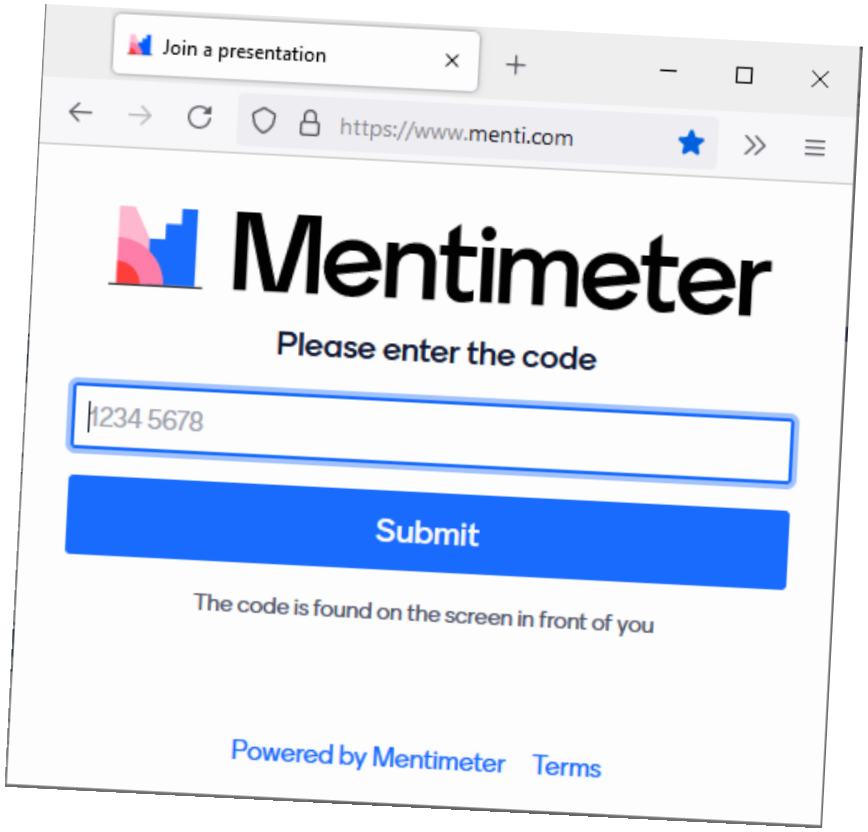


CRISP-DM

- CRoss-Industry Standard Process for Data Mining

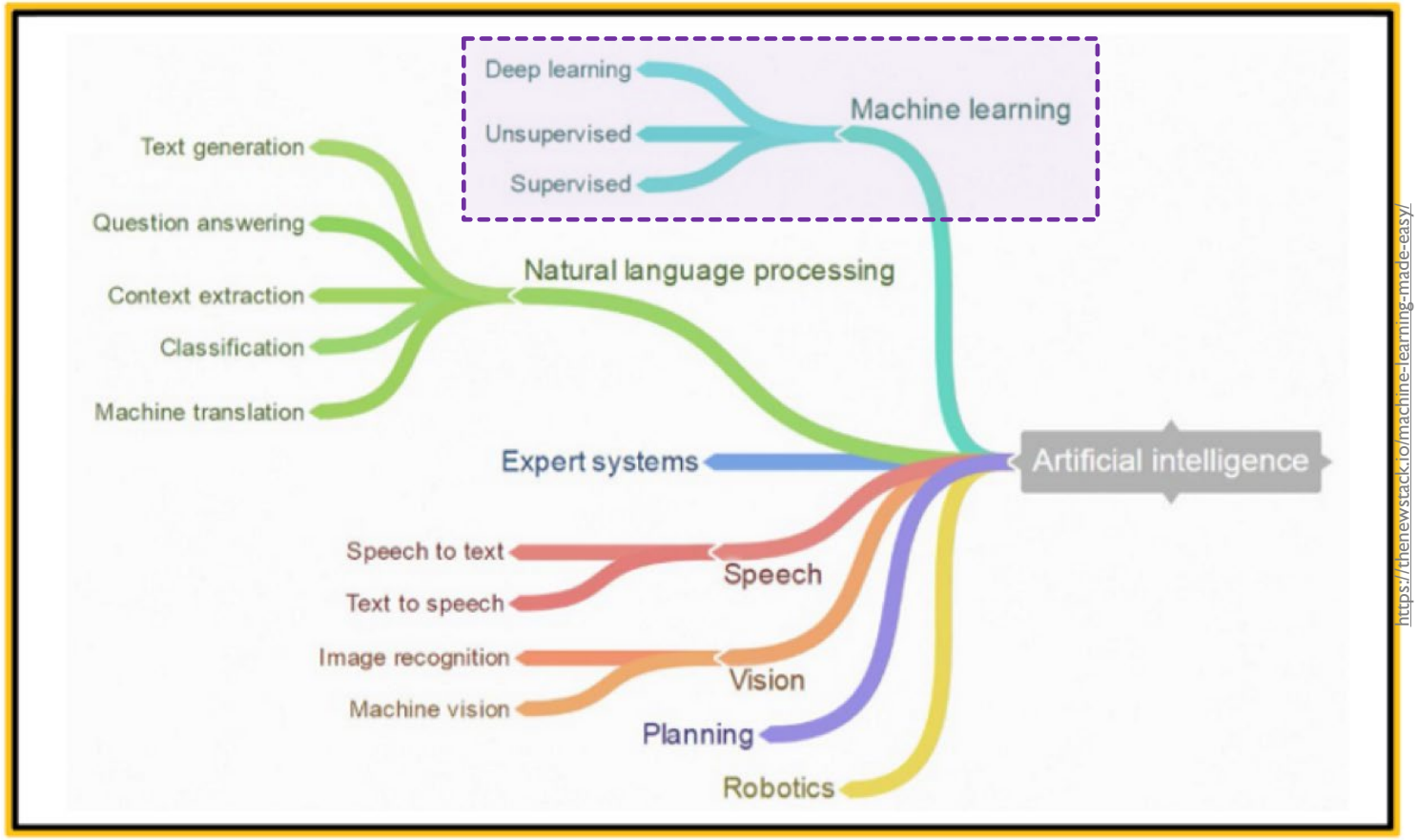
- Chapman, P. Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step Data Mining Guide*. [ftp]

What do you think? #1



What is the relationship between Data Science & Artificial Intelligence?

ANSWER #1: DATA SCIENCE VERSUS ARTIFICIAL INTELLIGENCE



https://thenewstack.io/machine-learning-made-easy/

"Machine Learning is an approach to Achieve Artificial Intelligence"

AGENDA

A. Setting the Scene

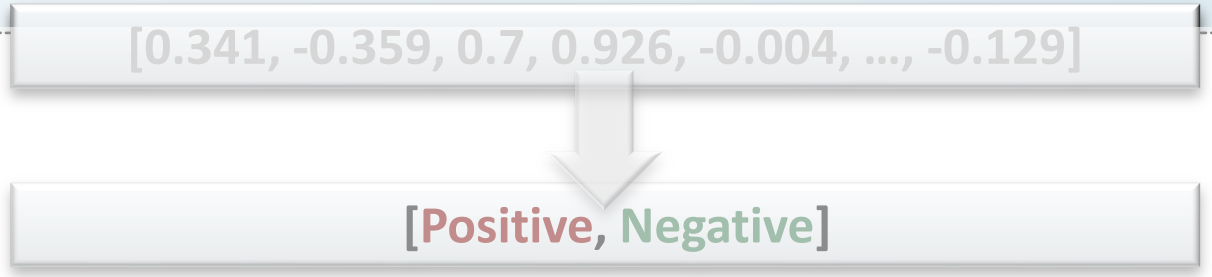
From Scientific Method to Translational Data Science

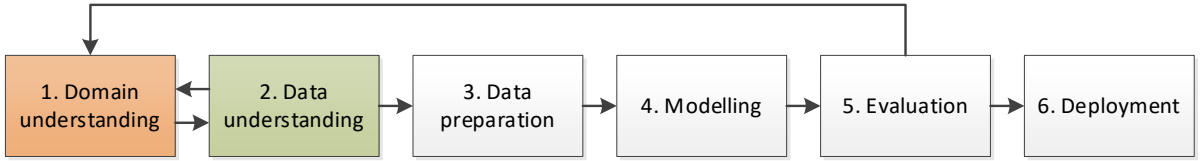
B. Case study 1

From Information Needs to Data Mining Goals in Long-term Care

C. Case study 2

Natural Language Processing for Translational Data Science in Mental Healthcare





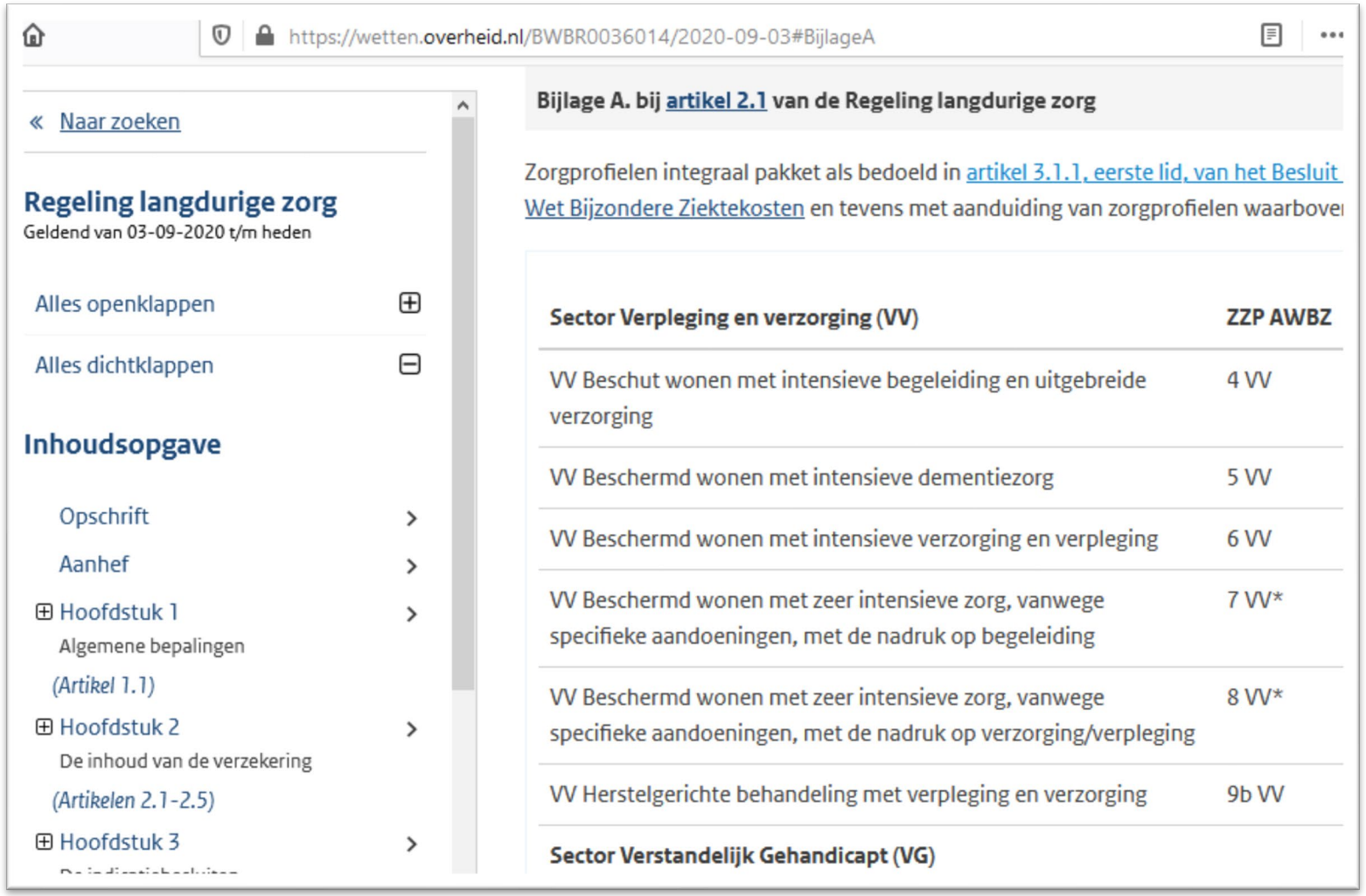
From Information Needs to Data Mining Goals in Long-term Care

CASE STUDY 1

Spruit, M., Vroon, R., & Batenburg, R. (2014). Towards healthcare business intelligence in long-term care: an explorative case study in the Netherlands. *Computers in Human Behavior, 30*, 698–707. [pdf] [online]

Introduction: Long-term care in the Netherlands

- Long-term care
  - Care for people with a long-term or chronic disorder
  - Relatively unexplored
- Main goals for long-term care
  - Care of good quality
  - Acceptable cost level
- One of the biggest expenses of Dutch government
  - 38% of the total healthcare budget
  - €14 billion in 2000 → €27 billion in 2012
- Care Intensity Package (ZZP)
  - Introduced in 2009
  - Different levels of care intensity
    - ZZP1: Extramural living with some guidance
    - ZZP8: Intramural living under full surveillance and 24/7 care
  - Operational cost depends on ZZP level
  - 2015: *Wet langdurige zorg* (Wlz) : ZZP ⇨ *Zorgprofiel*
- Electronic Client Record data
  - Large valuable dataset, but
  - Not fully exploited at this moment

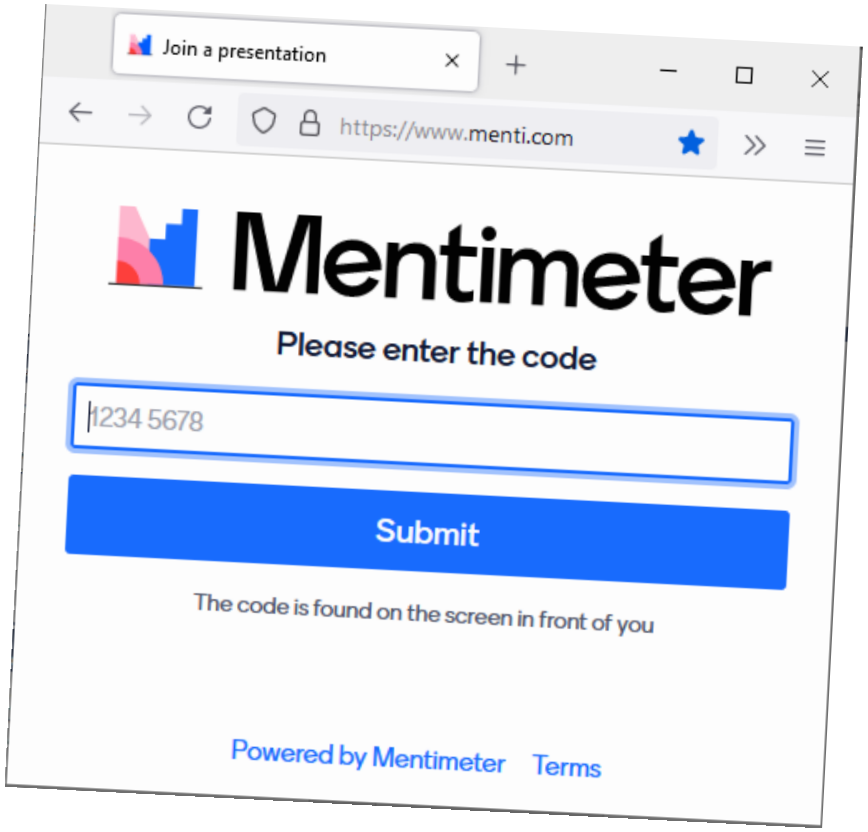


2015: Wet langdurige zorg (WLZ) ← Algemene Wet Bijzondere Ziektekosten (AWBZ)

Research Method: CRISP-DM

- CRoss Industry Standard Proces for Data Mining
- *Phase 1:* Business understanding
  - 18 unstructured in-depth interviews
    - 8 (board of) directors experts
    - 7 management experts
    - 7 experts from stakeholders perspective
      - (MinVWS, IGZ, Care insurer)
- 56 information needs
  - 33 related to quality of care
  - 23 related to financial state
- Information needs scored based on a valuation equation
  - Board members: 10
  - Managers: 6
  - Stakeholders: 3
- Score = $\sum_{Expert\ level} \frac{Times\ mentioned}{Number\ of\ interviews} \times Valuation$
- Score = $\left(\frac{8}{8} \times 10\right) + \left(\frac{4}{5} \times 6\right) + \left(\frac{3}{5} \times 3\right) = 16.6$

WHAT DO YOU THINK? #2

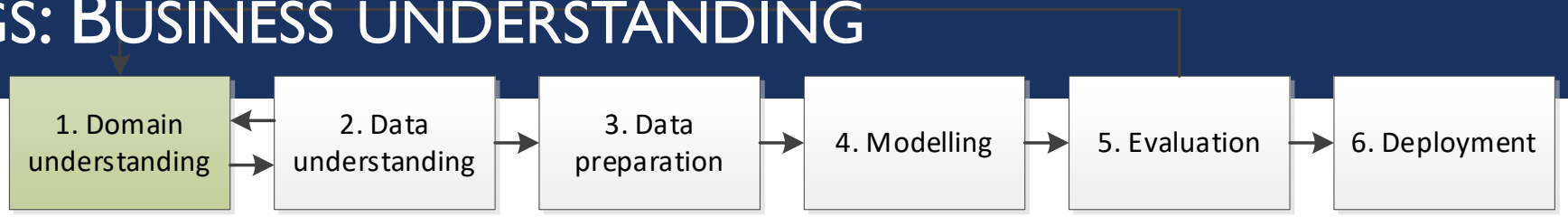


WHAT WAS BY FAR THE #1 INFORMATION NEED IN THE DUTCH LONG-TERM CARE SECTOR?

Answer #2: What do we need?

#	Type	Information need	Board	Mgmt	Stakeh.	Score
1	Q	**Customer experience**	**8**	**7**	**10**	**16.6**
2	F	Staffing with respect to ZZP-mix	7	4	2	14.8
3	F	ZZP-mix per business unit	7	4	0	13.6
4	F	ZZP-mix prognoses	7	4	0	13.6
5	F	Staffing with respect to operations	6	4	2	13.5
6	Q	Number of incidents occured	6	4	2	13.5
7	Q	Types of incidents occured	6	4	2	13.5
8	Q	Causes of occured incidents	6	4	2	13.5
9	F	Operations per ZZP	7	3	1	13.0
10	F	Production information (planned, realized, declared)	7	3	1	13.0

FINDINGS: BUSINESS UNDERSTANDING

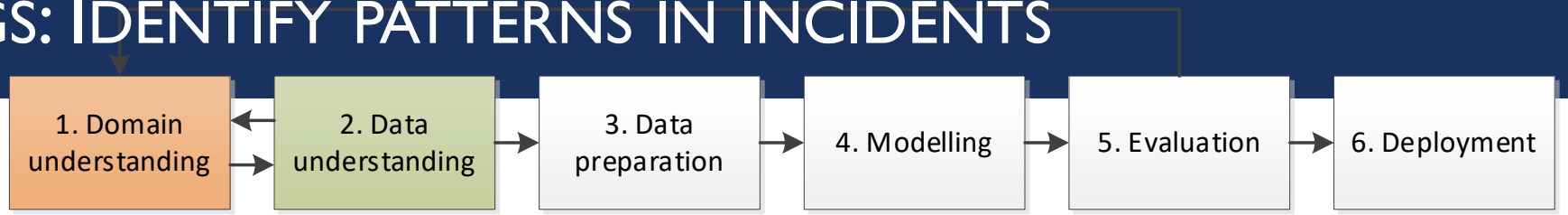


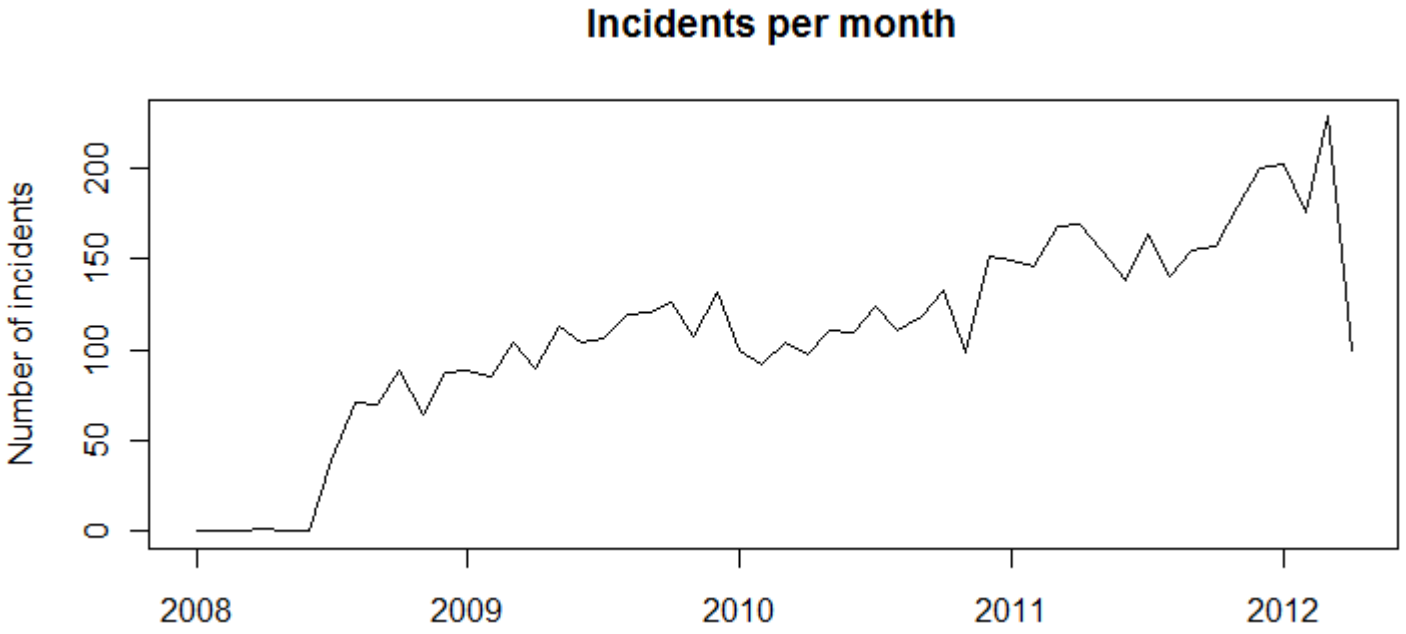
Information needs	*Data mining goals*
• Number of occured incidents • Types of occured incidents • Causes of the occured incidents • Patterns in occured incidents	• Identify patterns in incidents
• Number of clients at an increased risk • Types of risk the clients run	• Identify relationships in risk assessment
• Progress of care-related measures	• Identify patterns in care-related measures
• Treatment goals (obtained & not-obtained) • Care plan information	• Identify patterns in obtained and not-obtained treatment goals
• Number of clients per demand for care • ZZP-mix • ZZP-mix prognosis	• Identify & predict the ZZP mix

FINDINGS: DATA UNDERSTANDING

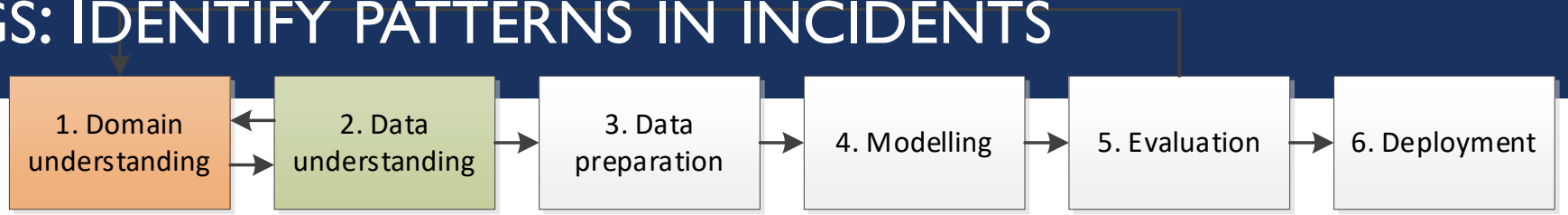
Data mining goals	*Available data*
• Identify patterns in incidents	• 5,692 records with incidents • 13 different incident types
• Identify patterns in risk assessment	• Depression risk: 2,129 records • Falling risk: 889 records • Incontinence risk: 877 records • Medication risk: 806 records • Problem behaviour risk: 0 records • Weight risk: 567 records
• Identify patterns in care-related measures	• 27,174 records with care-related measures
• Identify patterns in obtained and not-obtained treatment goals	• 20,725 records with treatment goals • 14.59% obtained goals • 85.41% not obtained goals
• Identify & predict the ZZP mix	• 1,831 records with historical delivered ZZP's

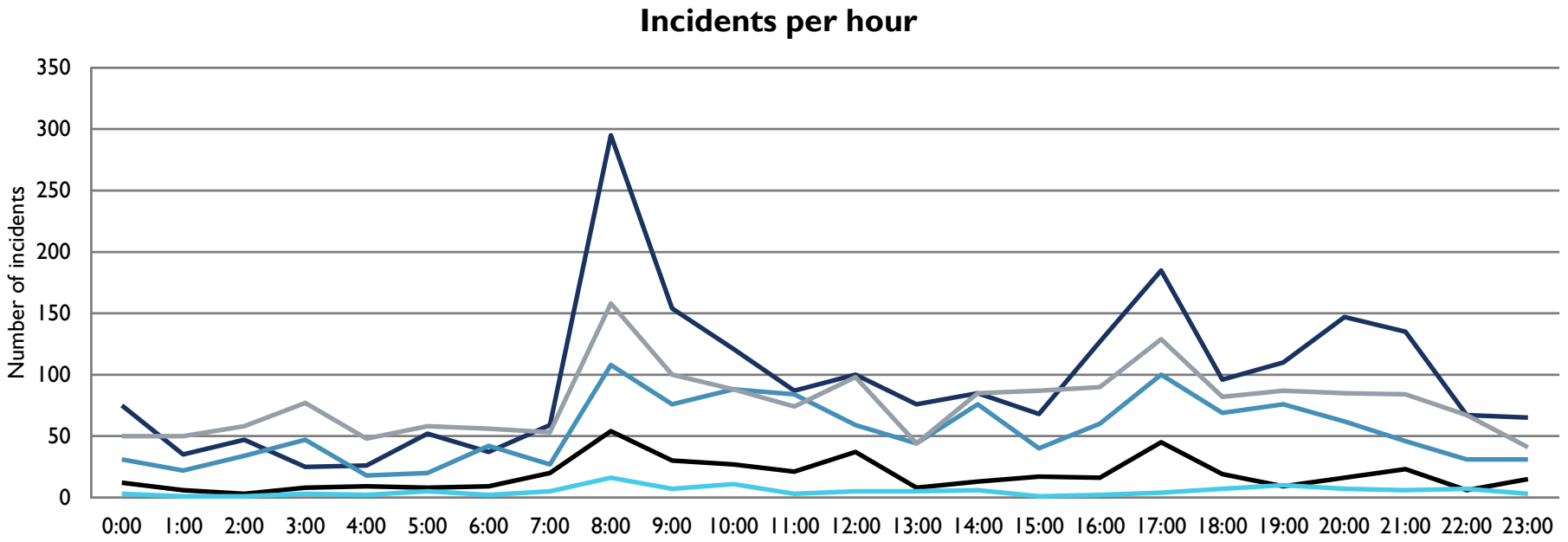
FINDINGS: IDENTIFY PATTERNS IN INCIDENTS





FINDINGS: IDENTIFY PATTERNS IN INCIDENTS

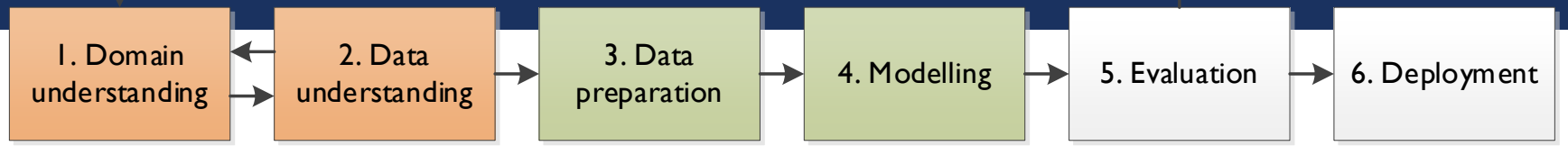




FINDINGS: IDENTIFY PATTERNS IN INCIDENTS

Incident location	*Loc1*	*Loc2*	*Loc3*	*Loc4*	*Loc5*	*Total*	*%*
Activities room	11	0	1	10	0	**22**	**0.4%**
Bathroom	126	77	152	43	17	**415**	**7.3%**
Corridor	101	72	92	9	3	**277**	**4.9%**
Kitchen	225	154	110	46	10	**545**	**9.6%**
Bedroom	524	278	575	38	31	**1,446**	**25.4%**
Toilet	62	45	54	5	8	**174**	**3.1%**
Living room	895	500	587	231	40	**2,253**	**39.6%**
Totals	**2,160**	**1,251**	**1,738**	**422**	**121**	**5,692**	**100%**

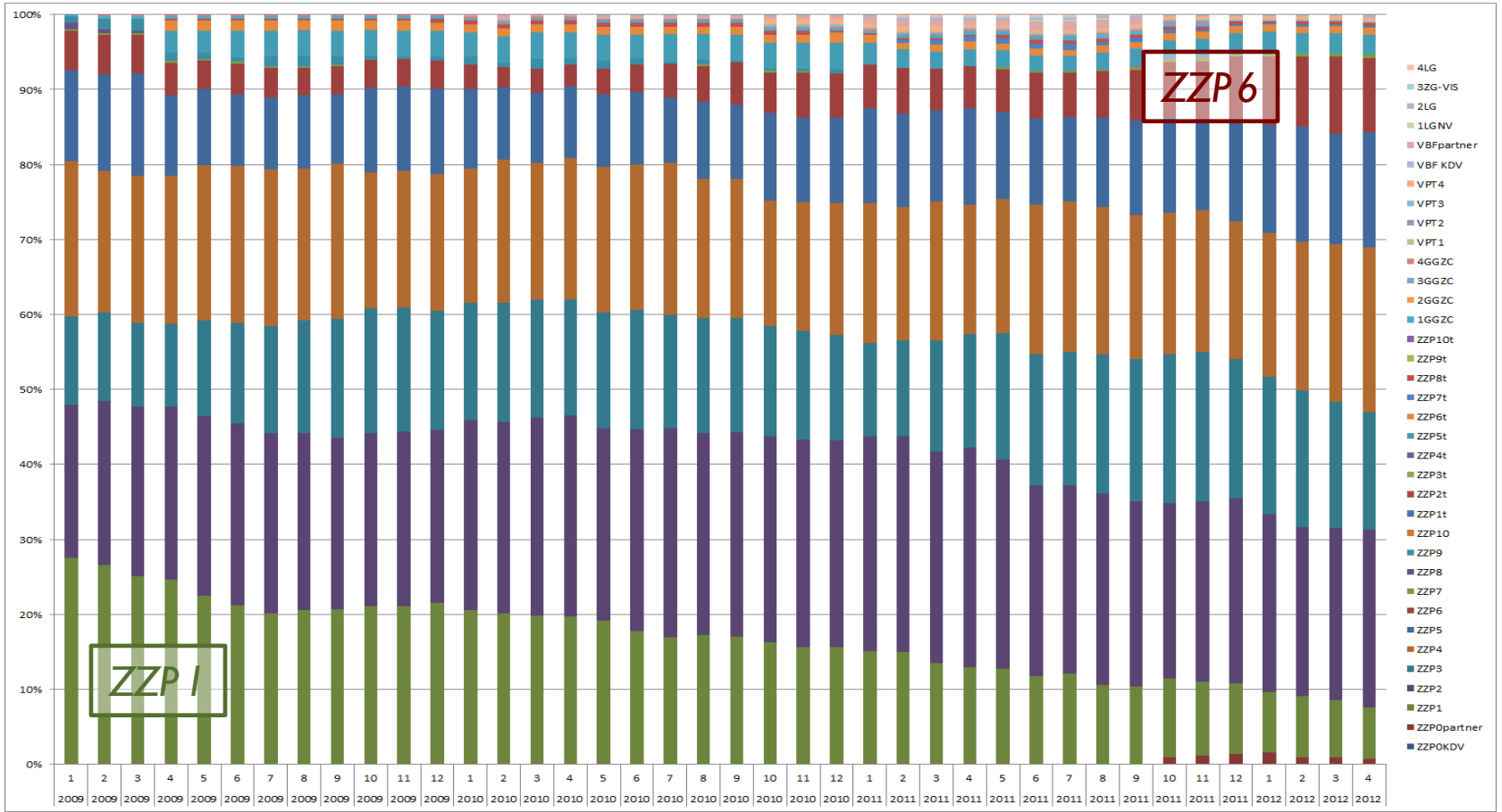
FINDINGS: IDENTIFY RELATIONSHIPS IN RISK ASSESSMENT



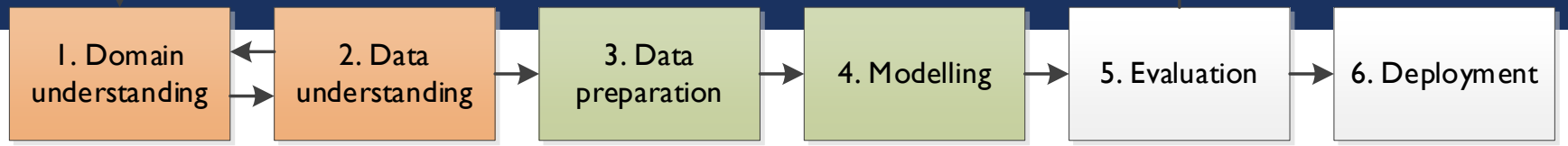
Rule	Support	Confidence	Lift
Incontinence, Medication → Falling	11.76%	90.63%	1.824
Falling, Medication → Incontinence	11.76%	95.08%	1.769
Incontinence, Weight intramural → Falling	11.56%	87.69%	1.765
Falling, Weight intramural → Incontinence	11.56%	93.44%	1.738
Depression, Incontinence → Falling	31.64%	82.11%	1.652

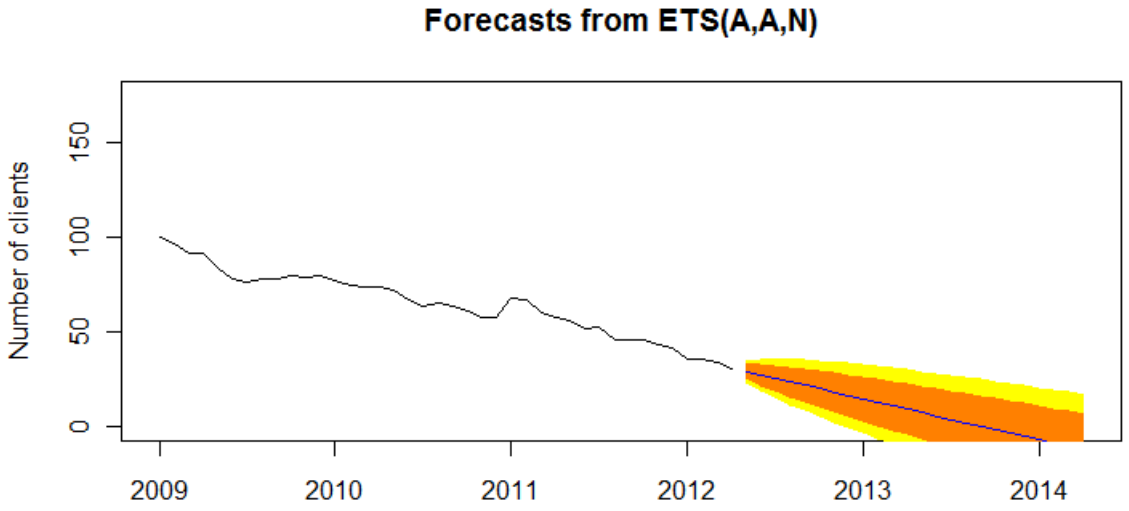
FINDINGS: PREDICT THE ZZP MIX

1. Domain understanding → 2. Data understanding → 3. Data preparation → 4. Modelling → 5. Evaluation → 6. Deployment

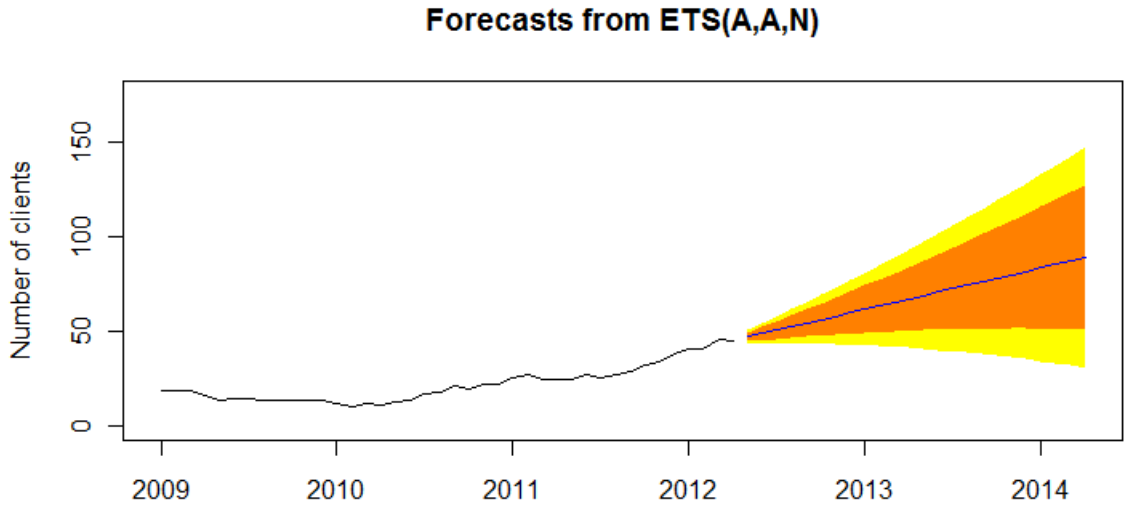


FINDINGS: PREDICT THE ZZP MIX



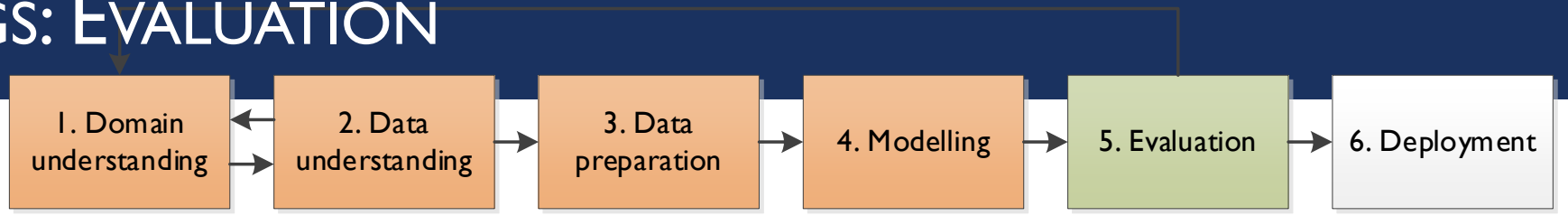


ZZP1



ZZP6

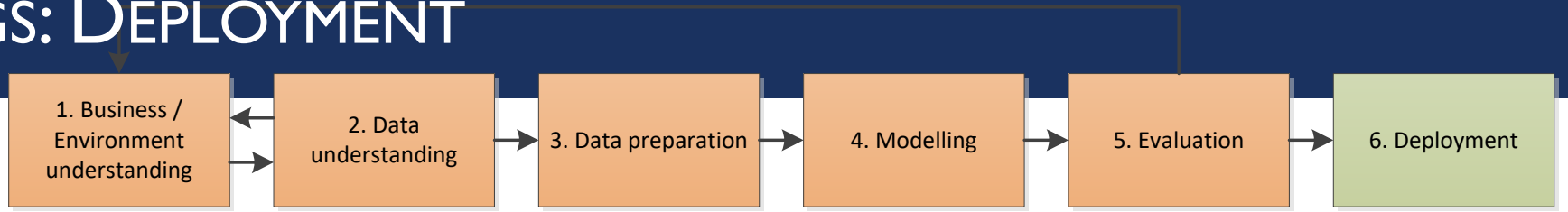
FINDINGS: EVALUATION



- Explorative research → no groundbreaking results
  - Information needs of multiple care institutions & stakeholders
- Lack of standardization
  - Heterogeneous data

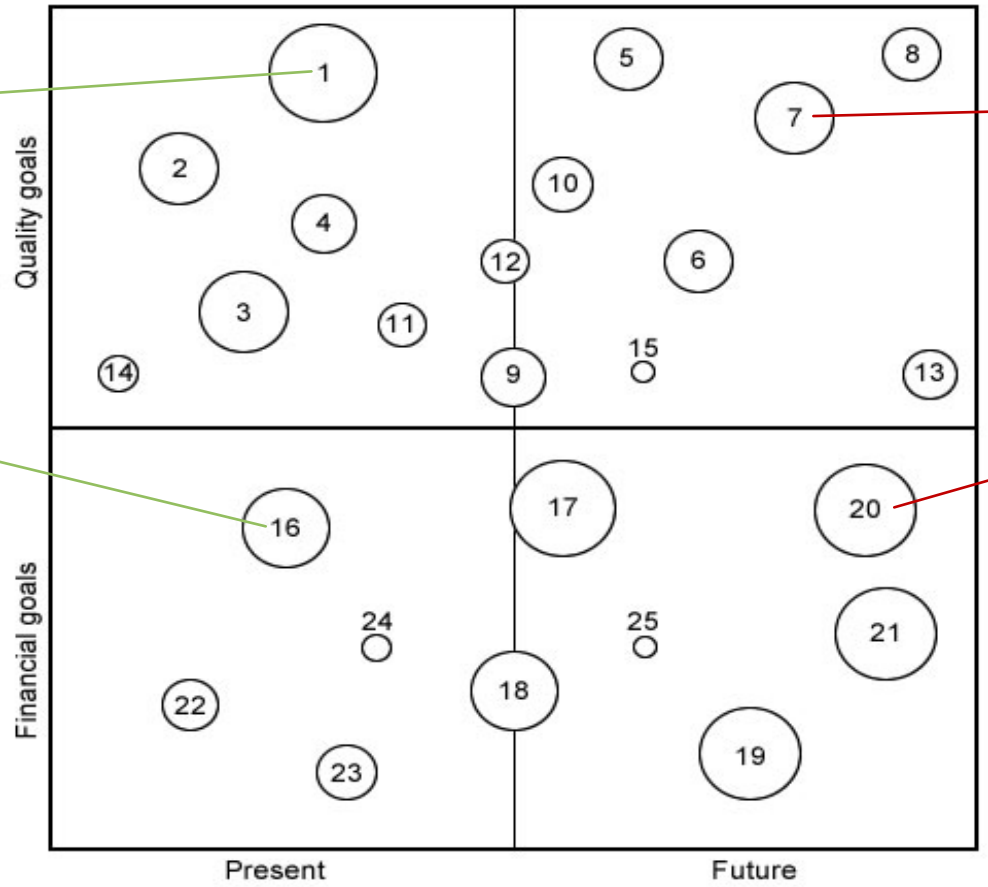
- Predictions are too dependent on environmental factors
  - Limited historical data
  - Too many dependent factors for the forecasts to be practically useful
    - Changes in laws and regulations have direct effect on analysis
- Qualitative data (e.g. free text input)
  - Complex to analyse

FINDINGS: DEPLOYMENT



Identify the patterns in incidents

Identify & predict the ZZP-mix



Relationship between care-related measures and incidents

Identify care within & outside ZZP indication (planned, realized)

AGENDA

A. Setting the Scene

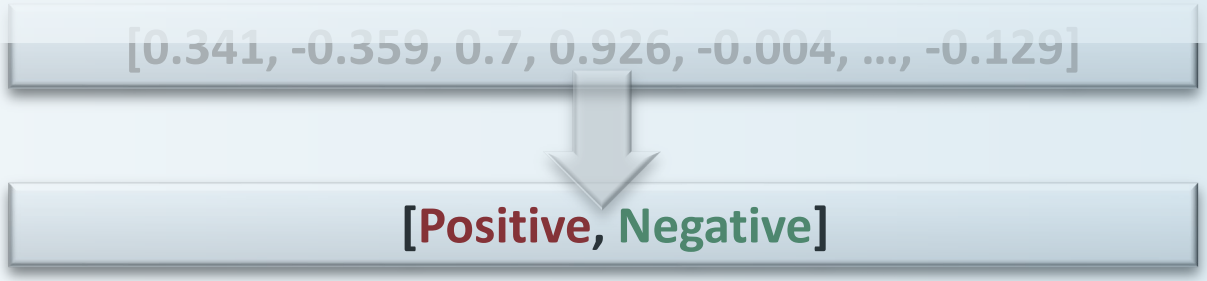
From Scientific Method to Translational Data Science

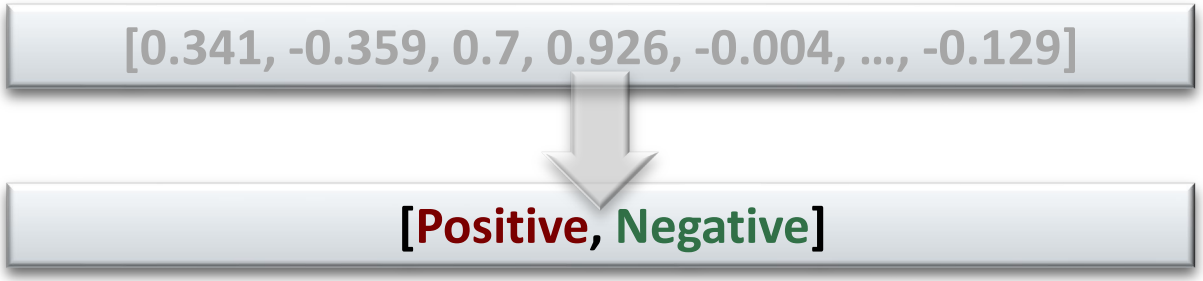
B. Case study 1

From Information Needs to Data Mining Goals in Long-term Care

C. Case study 2

Natural Language Processing for Translational Data Science in Mental Healthcare





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

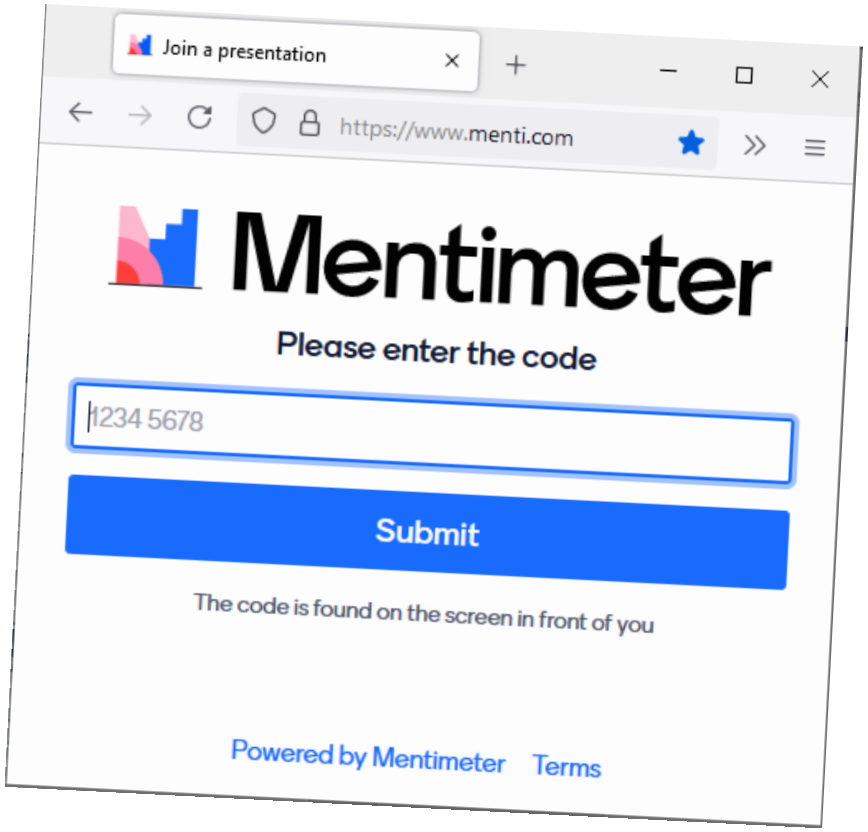
Characteristic	No. (%) Site 1	Site 2
Demographic characteristics		
Age, mean (SD), y	34.0 (16.6)	45.9 (16.6)
Men	1536 (48.2)	2097 (64.5)
Data set		
Admissions, No.	3189	3253
Unique patients, No.	2209	1919
Length of stay, median (IQR), d	16.0 (6.0-41.0)	15.0 (5.0-40.5)
No. of words in notes, median (IQR)	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."

?

WHAT DO YOU THINK? #3



WILL THIS CLINICAL NOTE RESULT IN A VIOLENCE INCIDENT WITHIN THE NEXT 4 WEEKS, YES OR NO?

[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

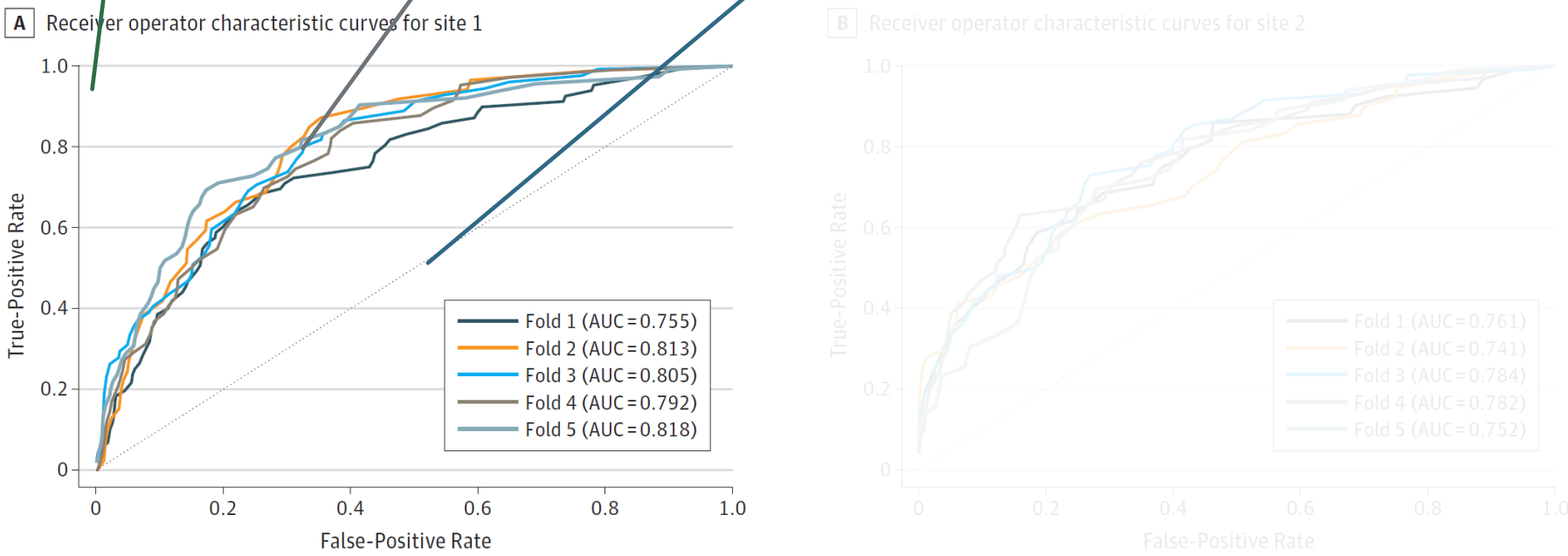
(SVM CLASSIFIER)

Perfect Score

Actual Score

Random Score

Figure. Receiver Operator Characteristic Curves for Internal Cross-validations



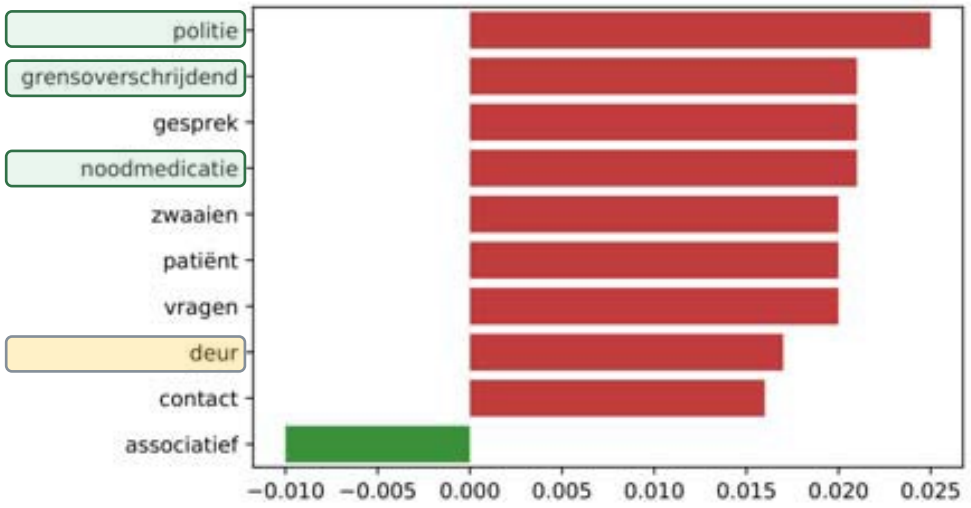
Receiver operator characteristic curves are shown for each fold, according to internal cross-validation in site 1 (A) and site 2 (B). Dashed diagonal lines denote an area under the curve (AUC) of 0.5, ie, predictive validity equivalent to chance. AUC indicates area under the curve.

[5/6]
Evaluation: Exploratory Analysis

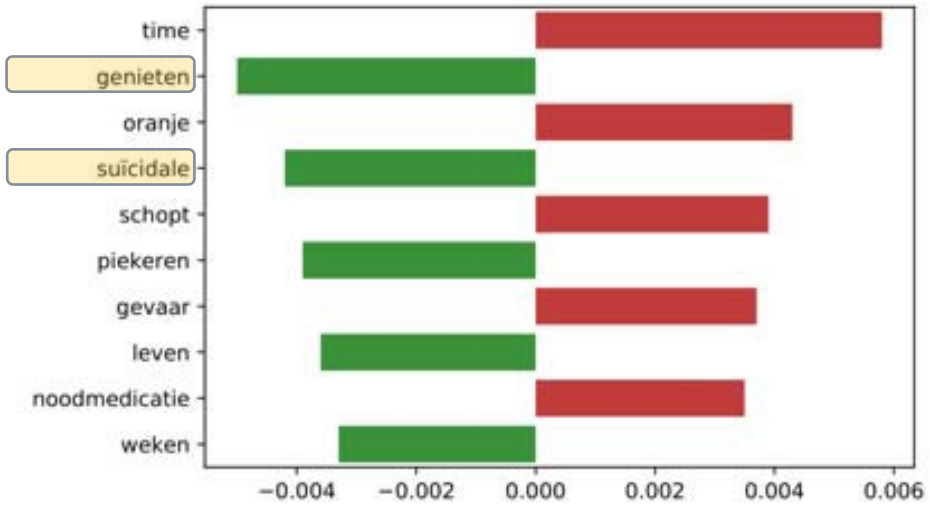
Table 3. Results of Exploratory Analysis

	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

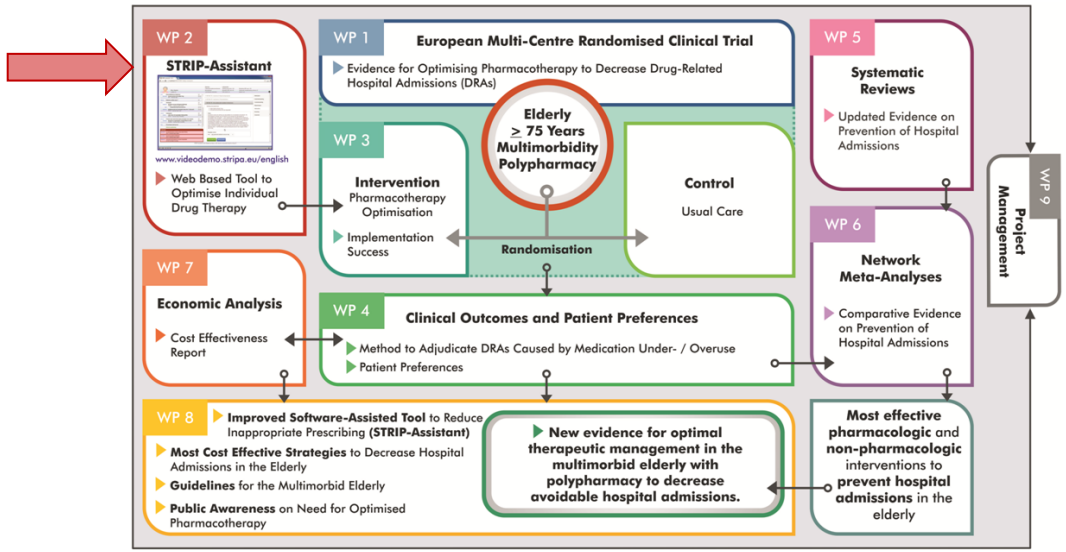
- DEDUCE

  "de-identification method for Dutch medical text"

  - pip install deduce
  - http://tdslab.nl/deduce

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