Health Campus Den Haag

VIPP: Het Virtuele Patiënten en Populatie project

“Een digitale tweeling mét ELAN”

Symposium CBS, 1 juni 2023, Prof. Dr. Marco Spruit, LUMC/LIACS

“Synthetische data buiten het CBS en in het LUMC”

Agenda

1. Synthetische data in LUMC context? Over mij, HCDH en ELAN
2. Onderwijs met synthetische data? PRIMA 2020
3. Onderzoek met synthetische data? Drie benaderingen
4. Et cetera? Effecten en de nabije toekomst

Health Campus Den Haag

Synthetische data in LUMC context

Ikzelf, de Health Campus Den Haag, ELAN datawarehouse

Mattijs Numans Jeroen Struijs Dennis Mook

Over… Marco Spruit

Pretpakket VWO
Propedeuse Nederlands
Propedeuse Muziekwetenschap
Bovenbouwstudie α-Informatica

Als Ontwikkelaar

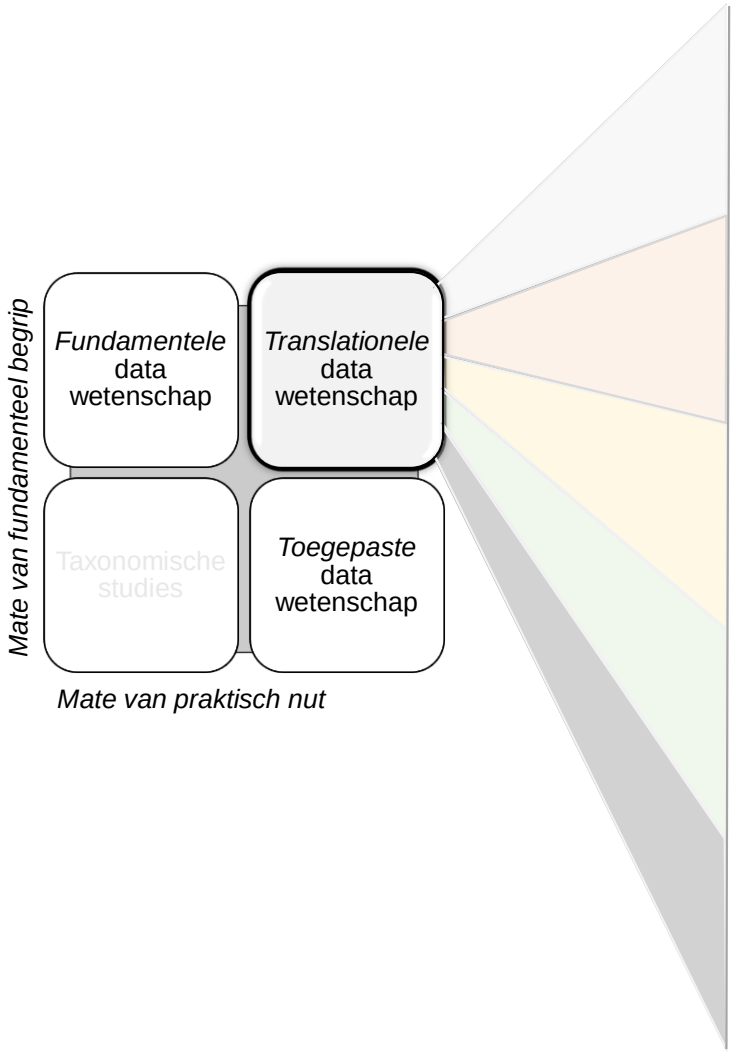
- 1993
  - *Information Retrieval* programmeur
    - ZyLAB Europe BV
- 1995
  - *Big Data* systeemontwikkelaar
    - Koninklijke Marine
- 1997
  - Productsoftware ontwikkelaar/ondernemer
    - Insertable Objects, Wizzer BV

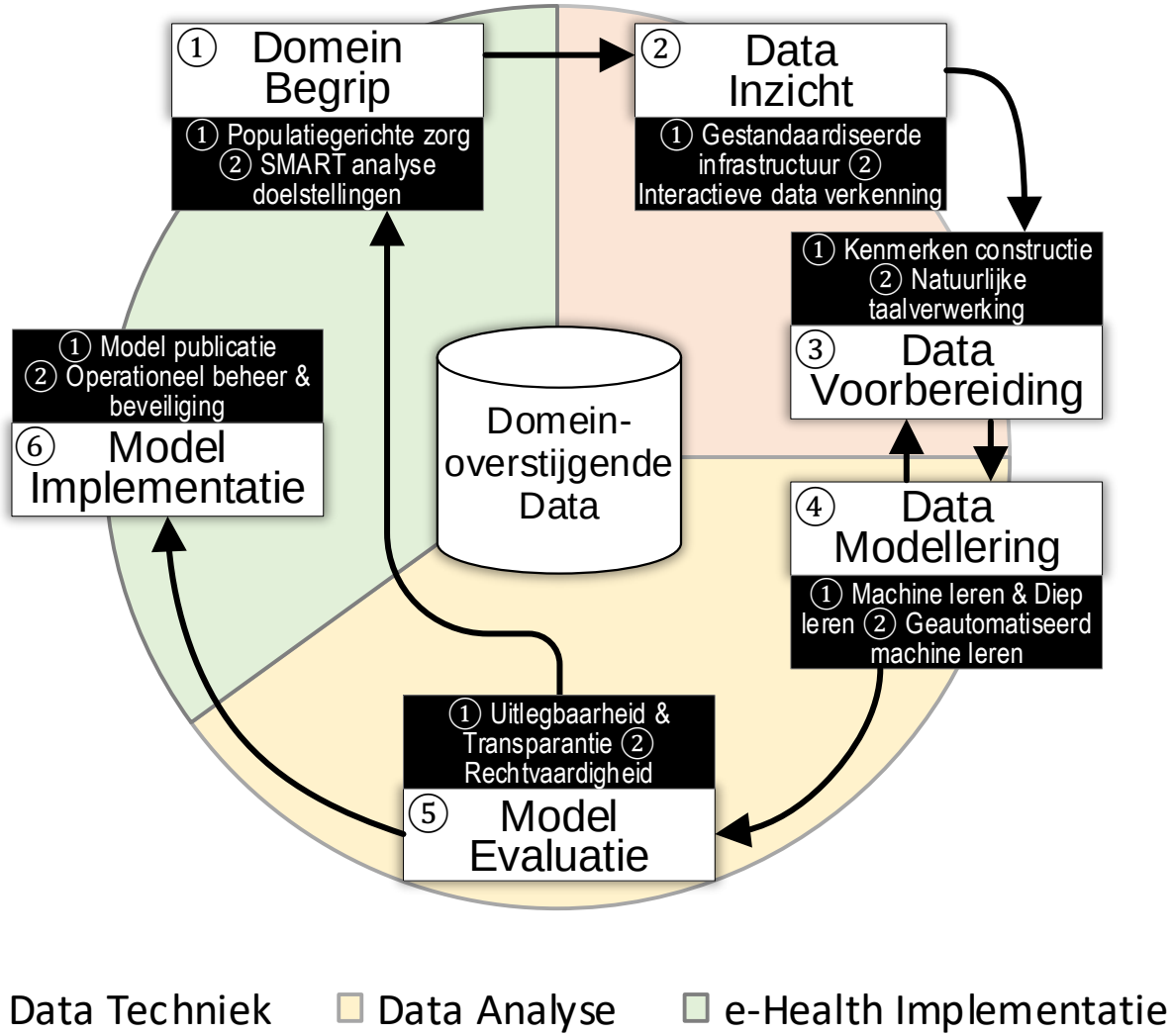
Als Wetenschapper

- 2003
  - OiO in Computationele Linguïstiek
    - Universiteit van Amsterdam (UvA)
- 2007
  - Universitair (hoofd)docent Informatiekunde
    - Universiteit Utrecht >> ADS Lab (2015)
- 2020
  - **Hoogleraar *Advanced Data Science in Population Health***
    - **LUMC/LIACS** @ Universiteit Leiden (HCDH)
    - PH Living Lab, CAIRELab, TDS Lab, SIG HDS

Health Campus Den Haag

Over… *Translationele* Datawetenschap





Health Campus Den Haag

Over… Health Campus Den Haag

1. Verkleinen gezondheidsverschil
2. Duurzame benadering
3. Breed gezondheidsperspectief

Leids Universitair Medisch Centrum

Populatiegericht zorgonderzoek

parnassia groep
Specialist in geestelijke gezondheid

Mentale gezondheid

Universiteit Leiden

Data science & AI

DE HAAGSE HOGESCHOOL

Zorg/Data-onderzoek

H+ MC

Ziekenhuis data

VREDE EN RECHT
Den Haag

oa WMO data

+

HagaZiekenhuis

Ziekenhuis data

Reinier de Graaf

Ziekenhuis data

hadoks

Huisarts data

GGD Haaglanden

oa Vaccinatie data

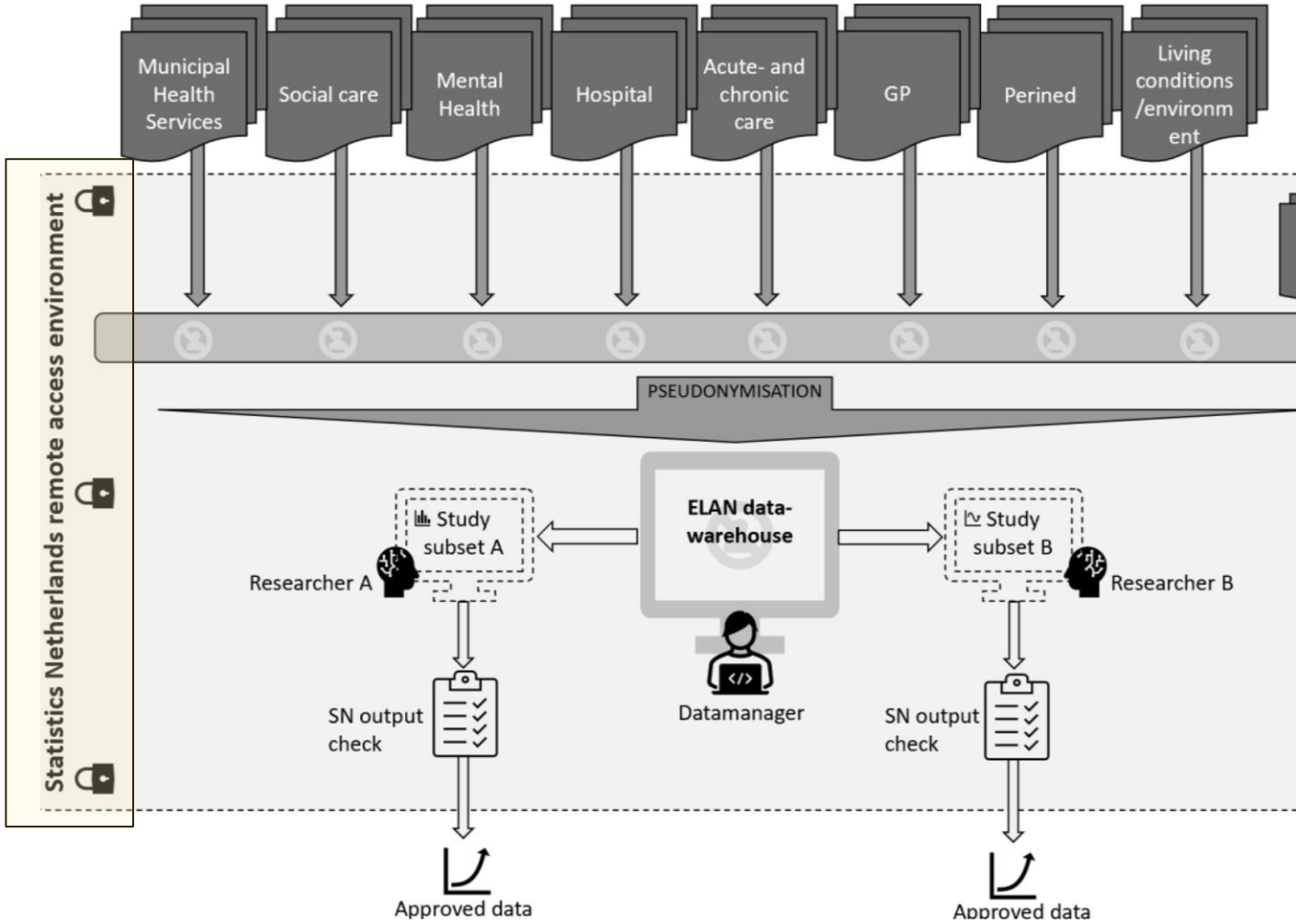
Health Campus Den Haag

Over… ELAN

- Het *“Extramurale LUMC Academische Netwerk”* (ELAN): CBS-RA

1. trans-disciplinair
2. domein-overstijgend
3. regionale data-infrastructuur

- inclusief bijbehorende beleidsstructuur, dat gezondheidsbeleid en onderzoek ondersteunt
- Totaal aantal patiënten:
  - ~1,2 miljoen records



Ardesch,F., Meulendijk,M., Kist,J., Vos,R., Vos,H., Kiefte-de Jong,J., Spruit,M., Bruijnzeels,M., Bussemaker,J., Numans,M., & Struijs,J. (2023). A data-driven population health management approach: The Extramural LUMC Academic Network data infrastructure. *Health Policy, 132*, 104769. [online]

Health Campus Den Haag

Onderwijs met synthetische data

LUMC Project Raamplan Implementatie Artsopleiding (PRIMA) 2020

NFU opdracht: *"Ontwikkelingen in technologie, waaronder informatie- en communicatietechnologie en kunstmatige intelligentie, leiden tot nieuwe mogelijkheden in preventie, diagnostiek en behandeling"*

Kans! AI-Tech met één revisie op constructieve en coherente wijze grondig integreren


Marco Spruit


Karoly Szuhai

To-be: De Leiden Virtuele Patiënten & Populatie dataset (VIPP)

- Vergroot het enthousiasme van zowel studenten als docenten voor Technologie, AI & Big Data
- Door een EPD-afgeleide dataset te ontwikkelen van ‘zo-goed-als-echte’ patiënten en lokale bevolking

Health Campus Den Haag

Health Campus Den Haag

Technieken voor synthetische data

Drie benaderingen voor onderzoek en toepassingen

A. **Regel-gebaseerd → Agent-Based Modelling (ABM)**
B. Neuraal netwerk-gebaseerd → Generative Adversial Networks (GAN)
C. Off-the-shelf → Syntho

Ammar Faiq Jim Achterberg Marcel Haas

A. Agent-Based Modelling (ABM)

- *Synthea:* an open-source Synthetic Patient Population Simulator
  - Doel: *"[…] simulate the lifespans of synthetic patients, modeling the 10 most frequent reasons for primary care encounters and the 10 chronic conditions with the highest morbidity* ***in the United States****"*
  - EPD waarden gecodeerd in standaard formaten (HL7, FHIR, CCDA, SNOMED)

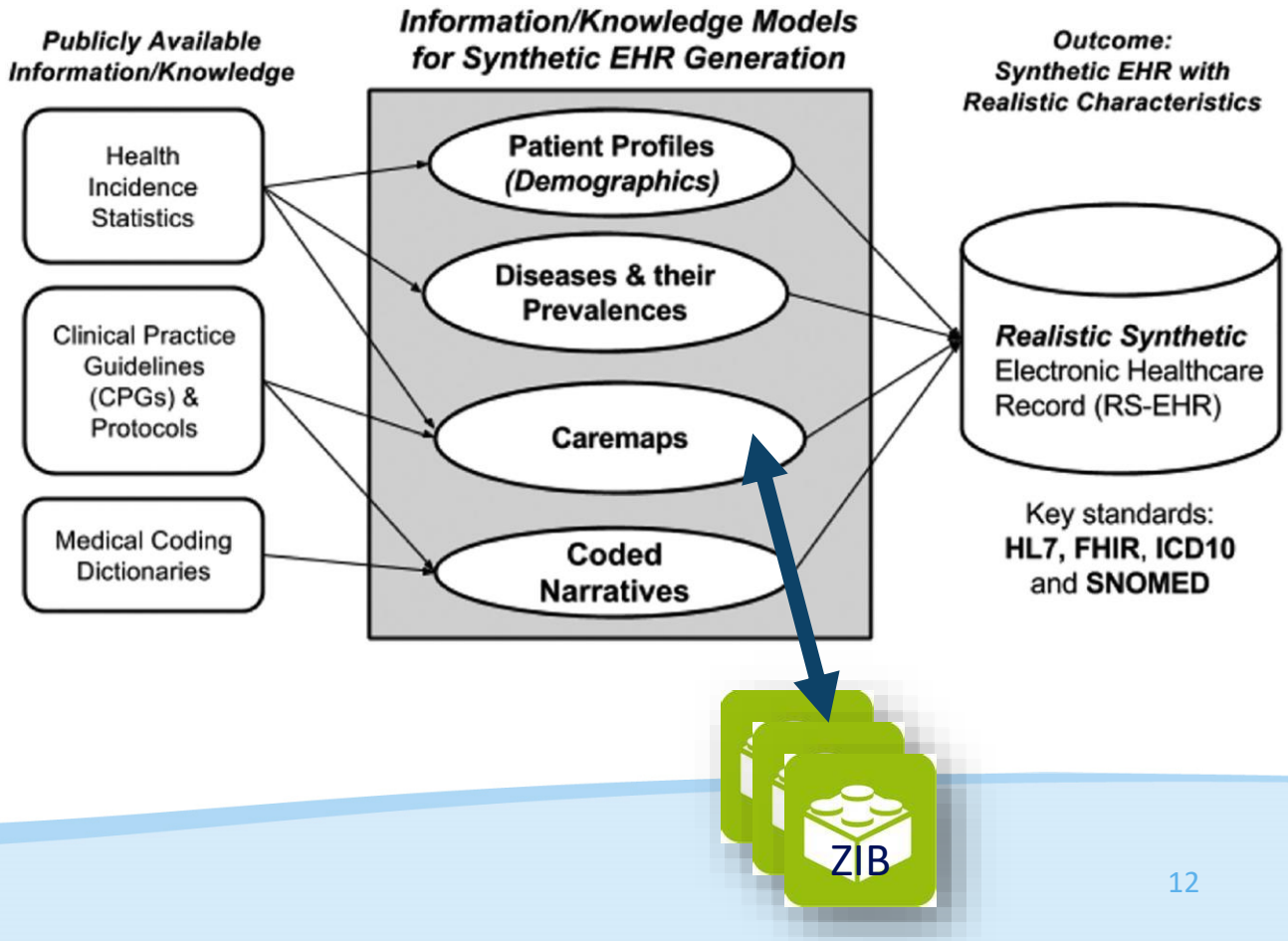
- https://github.com/synthetichealth/synthea
- Walonoski *et al.* (2018). [reeds 246 citaties]
- https://doi.org/10.1093/jamia/ocx079

Den Haag

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association, 25*(3), 230-238. [online]

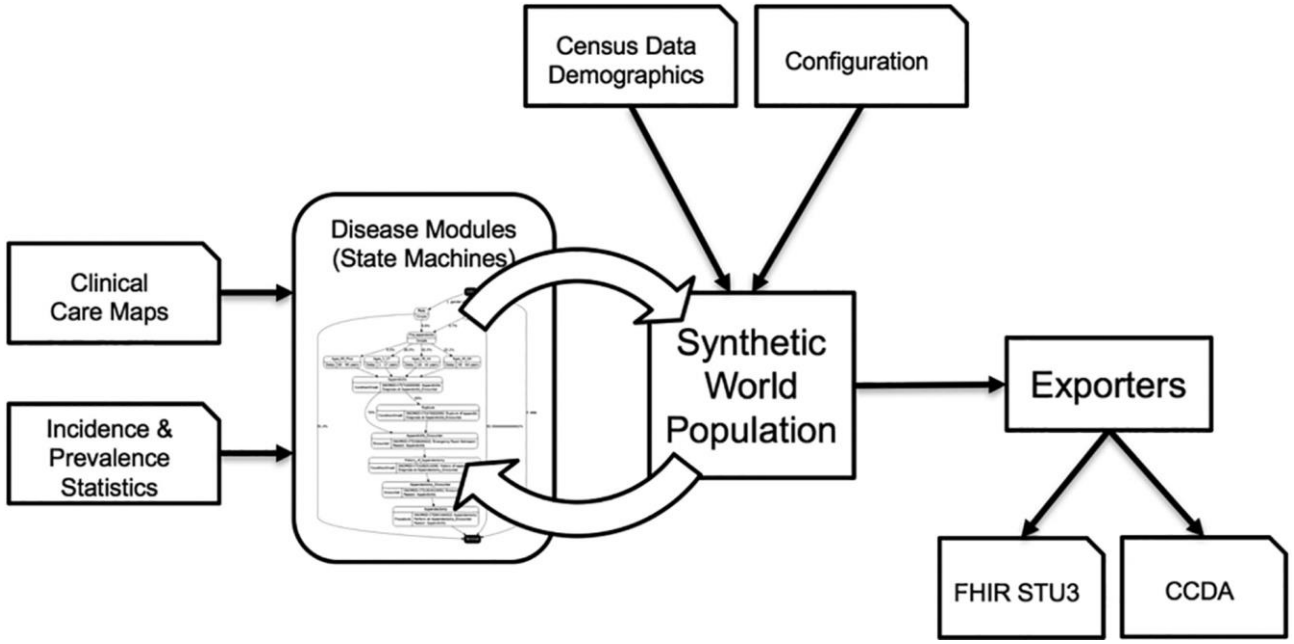
Benadering 1 “Anonymity vs Utility dilemma”

- Synthetische data (bewust) **NIET** obv originele EPD data (onwenselijk, onmogelijk), maar obv publiekelijk beschikbare data



Architectuur 1: Synthea componenten

- Prevalentie statistieken en patiënten demografie verschillen per land maar zijn bekend (CBS!)
- Klinische zorgpaden zijn in principe mondiaal
  - mbv land-statistieken
- Ziektebeelden
  - Grotendeels mondiaal



Health Campus Den Haag

Synthea Ziektebeeld

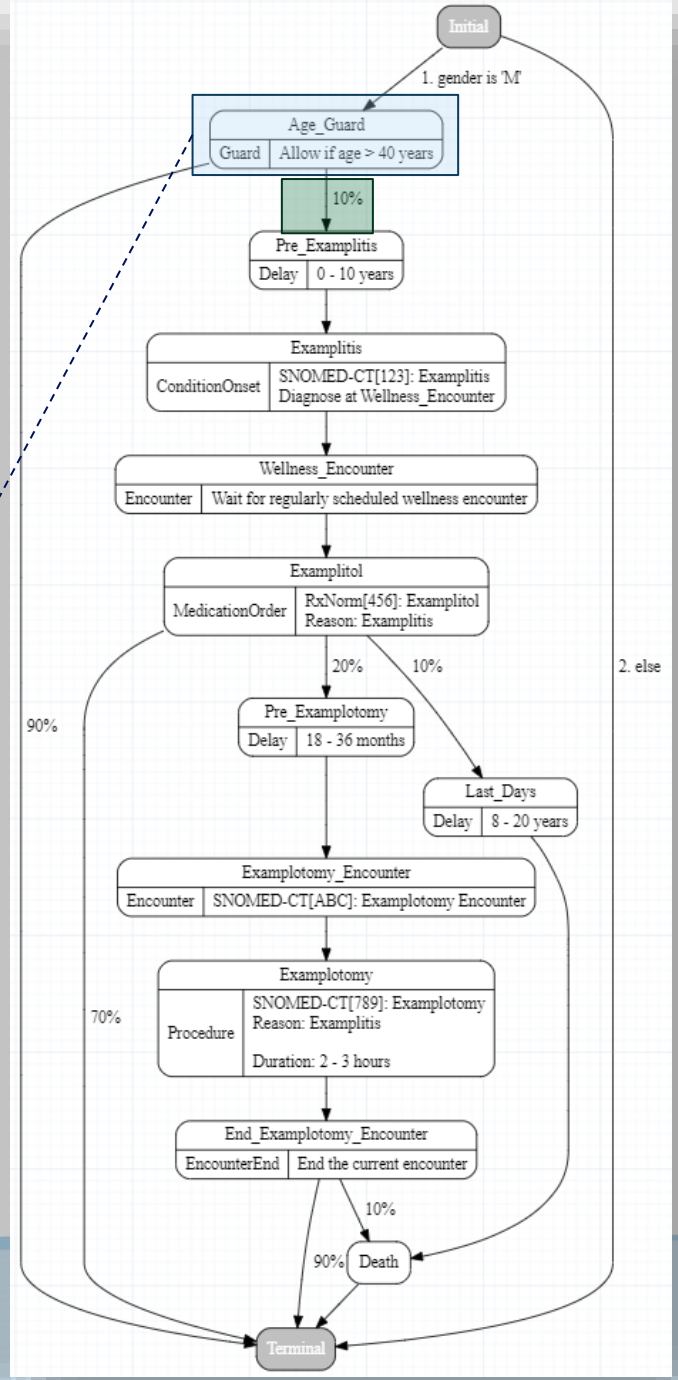
- FSM = (“wasmachine”)
- Synthea ziektebeeld = “stochastische automaat”

  oftewel een

  niet-deterministische eindigetoestanden-automaat
- Vanaf geboorte iedere week uitgevoerd voor iedere patiënt

Examplitis

```
{
  "name": "Examplitis",
  "remarks": [
    "Examplitis is a painful condition
      that affects only males. Most
      patients ",
    "can be cured with Examplitol or an
      Examplotomy but some never recover."
  ],
  "states": {
    "Initial": {
      "type": "Initial",
      "conditional_transition": [
        {
          "condition": {
            "condition_type": "Gender",
            "gender": "M"
          },
          "transition": "Age_Guard"
        },
        {
          "transition": "Terminal"
        }
      ],
      "name": "Initial"
    },
    "Age_Guard": {
      "type": "Guard",
      "allow": {
        "condition_type": "Age",
        "operator": ">",
        "quantity": 40,
        "unit": "years"
      },
      "distributed_transition": [
        {
          "distribution": 0.1,
          "transition": "Pre_Examplitis"
        },
        {
          "distribution": 0.9,
          "transition": "Terminal"
        }
      ],
      "name": "Age_Guard"
    },
    "Pre_Examplitis": {
      "type": "Delay",
```

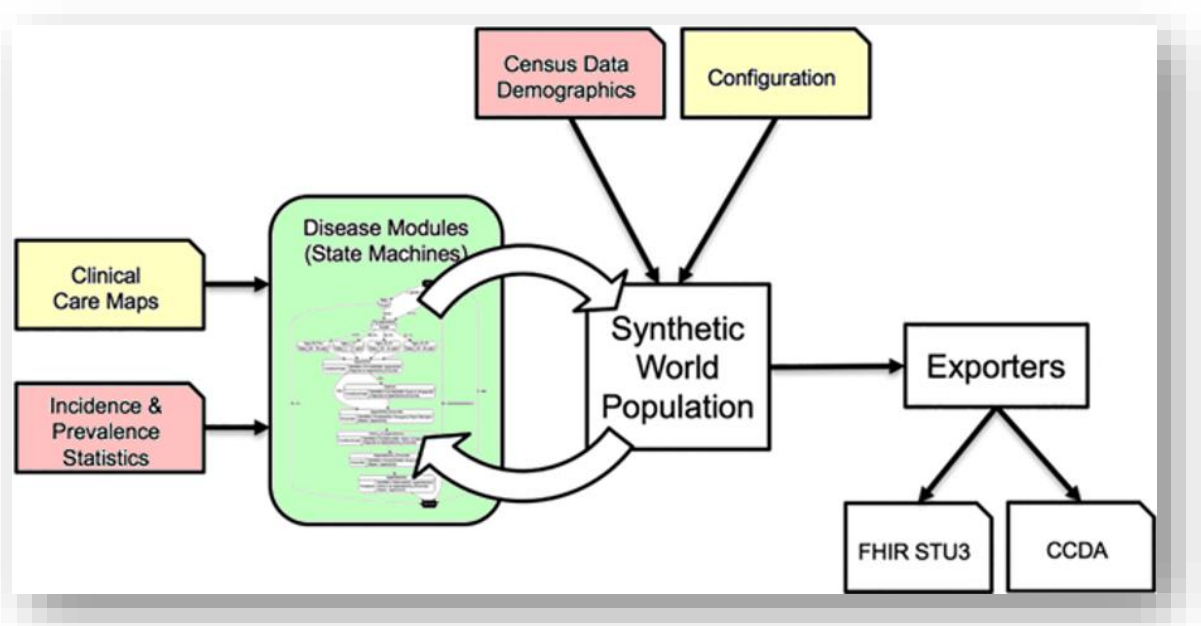


https://syntheticahealth.github.io/module-builder/#example_module

Health Campus Den Haag

Toepassing 1/2: VIPP → Synthea voor ELAN

- ELAN doel: *"[…] simulate the lifespans […]* ***in the ~~United States~~***"
  - Aanpassen: US prevalentie statistieken en patiënten demografie naar NL
  - *Tweaken*: zorgpaden & standaarden
  - *As-is*: ziektebeelden

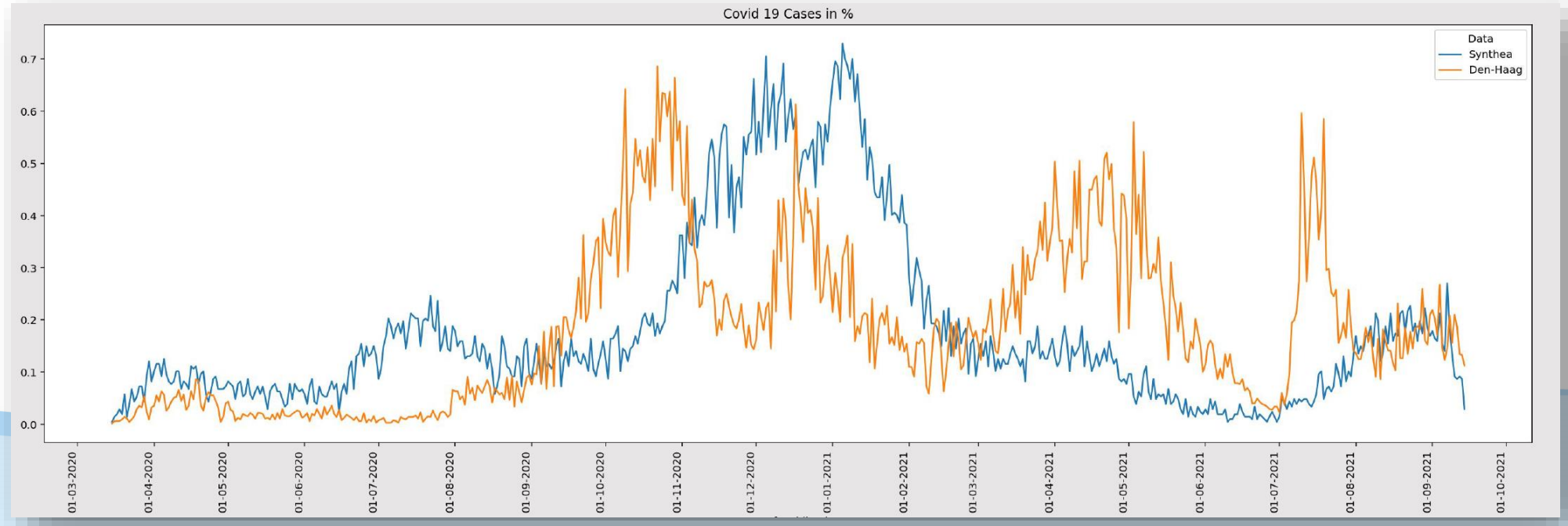


- Meer details bij onze VIPP poster presentatie!

Health Campus Den Haag

Toepassing 2/2: VIPP → Synthea voor PHM

- MSc *Fundamentals of Population Health Management* vakdoel
  - Praktijkopdracht data-analyse COVID19 voorspelmodel
  - Vereist integrale analyse van medische, sociale, regionale, etc data
- **PHM casus…** Meer details bij LUMC/VIPP poster presentatie!



Health Campus Den Haag

Health Campus Den Haag

Technieken voor synthetische data

Drie benaderingen

A. Regel-gebaseerd → Agent-Based Modelling (ABM)

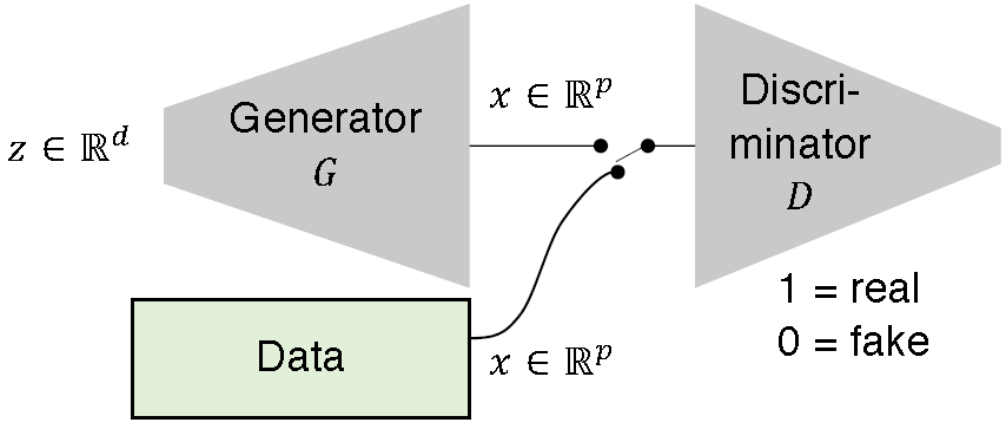
B. Neuraal netwerk-gebaseerd → Generative Adversial Networks (GAN)

C. Off-the-shelf → Syntho

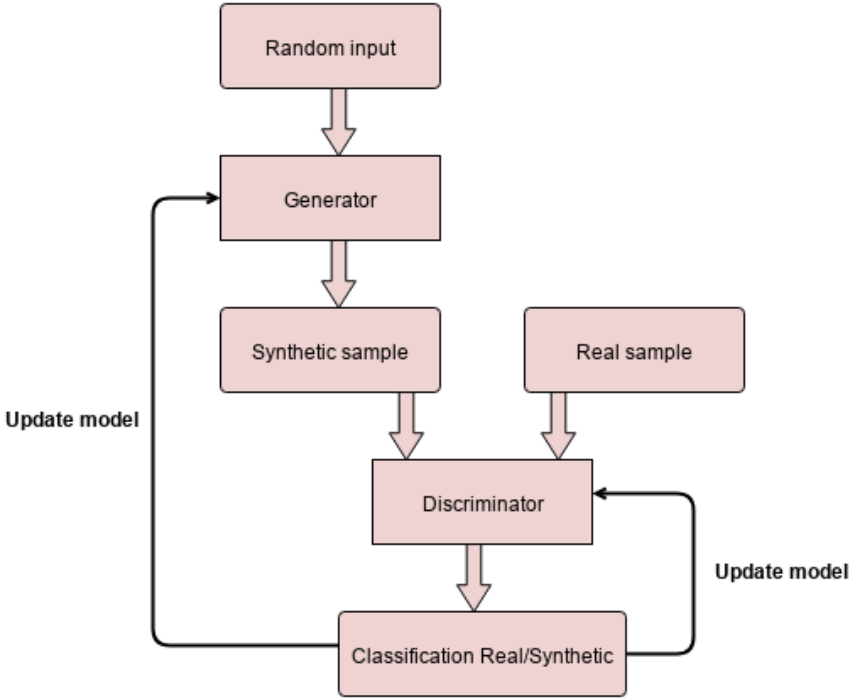
Ammar Faiq Jim Achterberg Marcel Haas

B. Generative Adversial Networks (GAN)

- "Het doel van een generatief model is het bestuderen van
  1. een verzameling trainingsvoorbeelden, en
  2. de waarschijnlijkheidverdeling die dit heeft voortgebracht."



- GAN trainingsproces



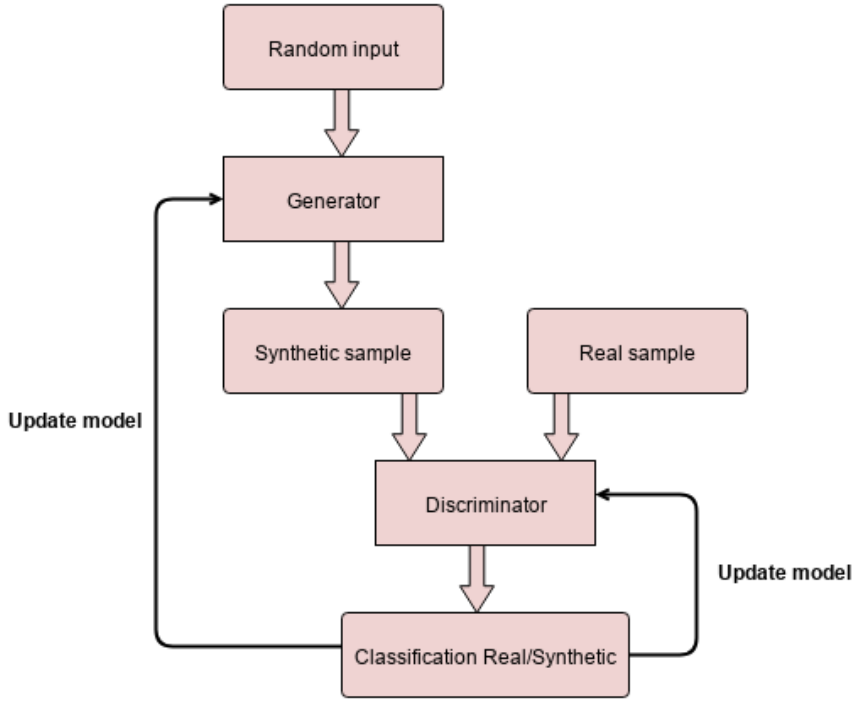
Health Campus Den Haag

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144. [online] → **60K citaties**!

B. Generative Adversial Networks (GAN)

- **GAN:** een spel tussen twee spelers: de Generator en de Discriminator.
- **Generator:** neuraal netwerk dat realistische synthetische data genereert uit willekeurige input.
  - Leert alleen van interactie met de discriminator
- **Discriminator:** neuraal netwerk dat de originele data en de synthetische data probeert te onderscheiden.
  - Ontvangt zowel originele data als synthetische data geproduceerd door de generator.
- De generator is succesvol als het adequaat de verdeling van de originele data leert via interactie met de discriminator.

- GAN trainingsproces



Health Campus Den Haag

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144. [online] → **60K citaties**!

DoppelGANger: een *Conditionele* GAN

- Ontworpen voor tijdreeksen data, en zowel continue als discrete attributen
- Adresseren tekortkomingen
  1. *Complexe* correlaties tussen tijdreeksen en de gerelateerde attributen
  2. *Lange-termijn* correlaties binnen tijdsreeksen
- https://github.com/fjxmlzn/DoppelGANger

Approach	Flexibility	Privacy	Fidelity
raw data	no	no	best
anonymized raw data	no	?	good
Markov model	yes	yes	bad
autoregressive model	yes	yes	bad
RNN	yes	yes	bad
GANs	**yes**	**yes**	**good**

Table 1: The potential of GANs to satisfy key desirable properties of synthetic datasets.

Health Campus Den Haag

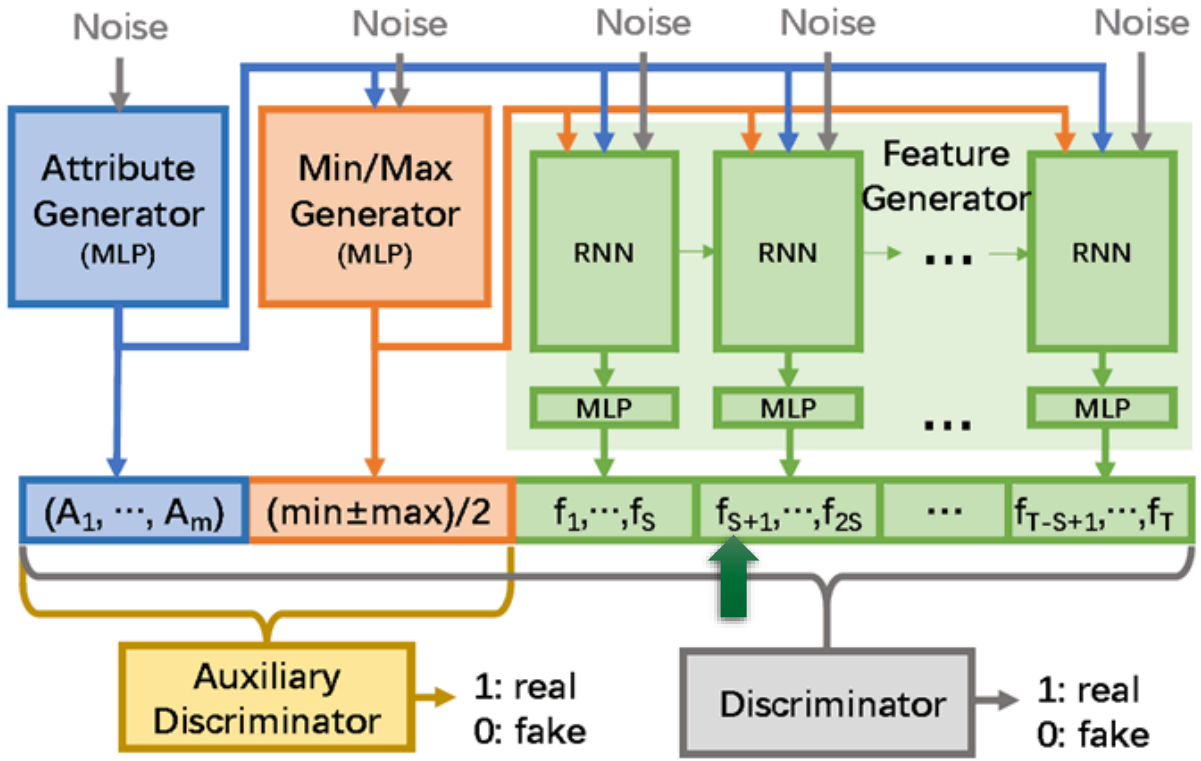
Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020, October). Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference* (pp. 464-483). [online]

DoppelGANger: Conditionele GAN

<klasse-specifieke distributies voor continue en discrete kenmerken>

- ***Top-3 kenmerken CGAN***

1. Ontkoppelde normalisatie (via realistische min/max limieten *'fidelity+ if divers.* → *'mode collapse'*)
2. Gebundelde samples generatie (om tijdsrelaties te versterken *'batch proc.'*)
3. Ontkoppelde attributen generatie (*'fidelity+ if lange/complexe tijdsreeks*)



Health Campus Den Haag

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020, October). Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference* (pp. 464-483). [online]

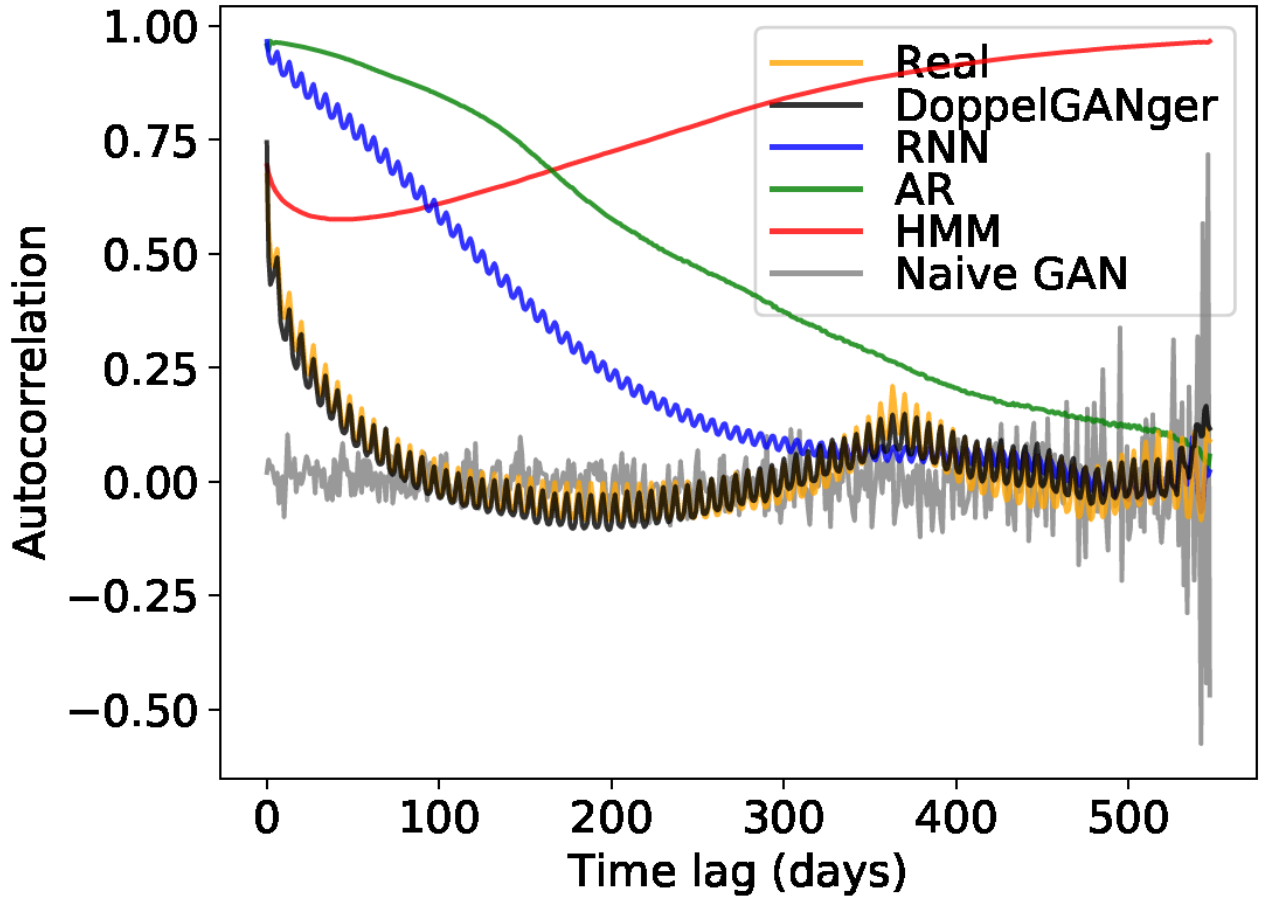


Figure 1: Autocorrelation of daily page views for Wikipedia Web Traffic dataset. DoppelGANger captures both weekly and annual correlation pattern.

Health Campus Den Haag

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020, October). Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference* (pp. 464-483). [online]

Toepassing: CGAN voor ELAN (…)

- Helaas nog niet gerealiseerd… Voorbereidingen wél vergevorderd!
- Voorstudie Jim Achterberg (EUR)
  - Scheppen kader voor *evaluatie* van synthetische *medische* data: combinatie van statische en tijdsgebonden variabelen van gemengde datatypes
    a) Uitbreiding *tSNE* algoritme tbv visualiseren tijdsreeksen van gemengde datatypes
    b) Verbetering afstandsmeting tussen twee punten mbv *dynamic time warping & Gower*
    c) Introductie nieuwe *two-sample goodness-of-fit test* obv classificatie-gebaseerd testen
    d) Introductie van interpreteerbare privacy-risico maat via *attribute inference attacks*


Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics and Management Science

An Evaluation Framework For Synthetic Medical Data

Author:
Jim Achterberg
481997

Supervisor:
O'Neill, E.P.

Second Assessor:
Gruber, L.

December 19, 2022

Abstract

This thesis presents a framework for evaluating synthetic medical data, which typically includes a combination of static and temporal variables of mixed data types. To assess the similarity of synthetic medical data to real data, we extend the tSNE algorithm to visualize sequences with mixed data types and use dynamic time warping and Gower distance to measure the distance between points. Additionally, we propose a new two-sample goodness-of-fit test for medical data that uses classification-based testing. We also introduce an interpretable metric for evaluating the privacy risk of synthetic data through attribute inference attacks. To demonstrate these evaluation methods, we apply them to synthetic medical data samples generated from two real world datasets using a generative adversarial network, and a novel application of a pre-existing autoregressive network.

The content of this thesis is the sole responsibility of the author and does not reflect the views of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

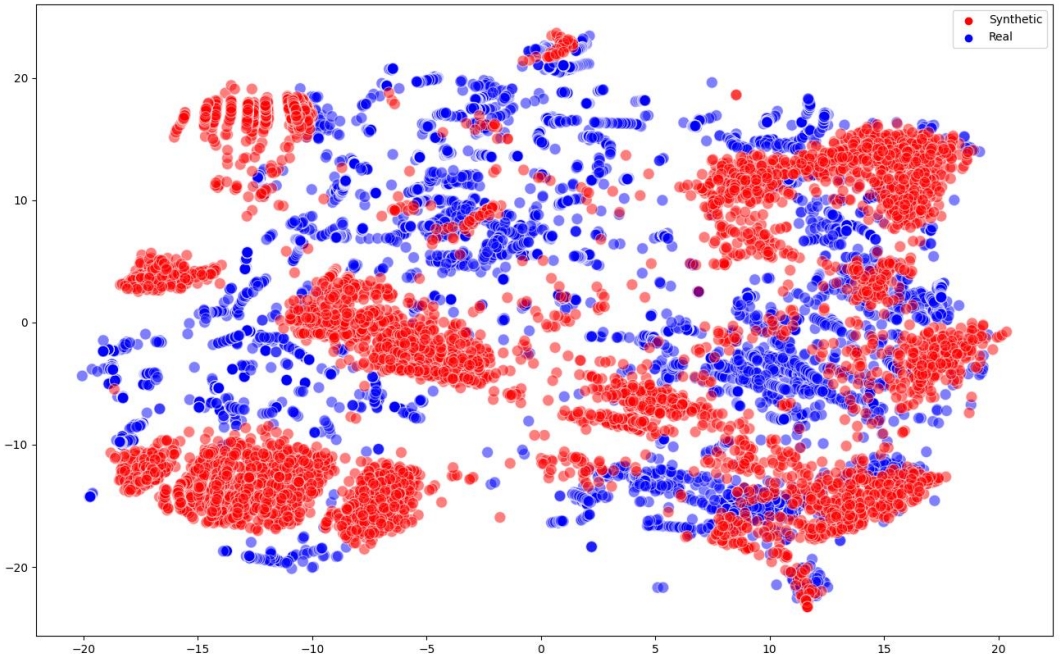
- Drie evaluatiematen
  - **Gelijkenis** *'fidelity'*, **Nut** *'utility'*, **Onthullingsrisico** *'privacy risk'*

Health Campus Den Haag

Achterberg, J. (2023, January 12). *An Evaluation Framework for Synthetic Medical Data.* Econometrie. Retrieved from http://hdl.handle.net/2105/65772

Component (a): tSNE visualisatie

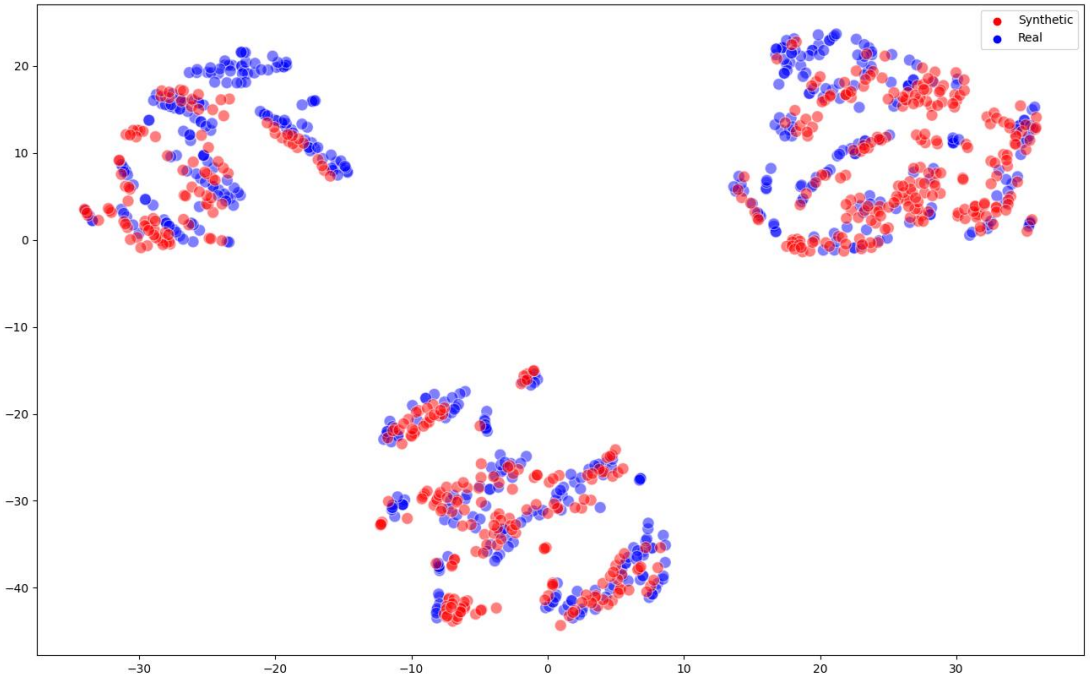
<dimensiereductie via t-distributed Stochastic Neighbor Embedding>
<met behoud van locale structuur per tijdstap>



Neuromuscular Monitoring (NM) dataset

- data structuur teveel 'puntenklodders'
  → te weinig data diversiteit *'mode collapse'*

- *Tijdsdimensie is impliciet, dus platgeslagen, dus tSNE toont hoe goed tijdsrelaties gegenereerd kunnen worden*



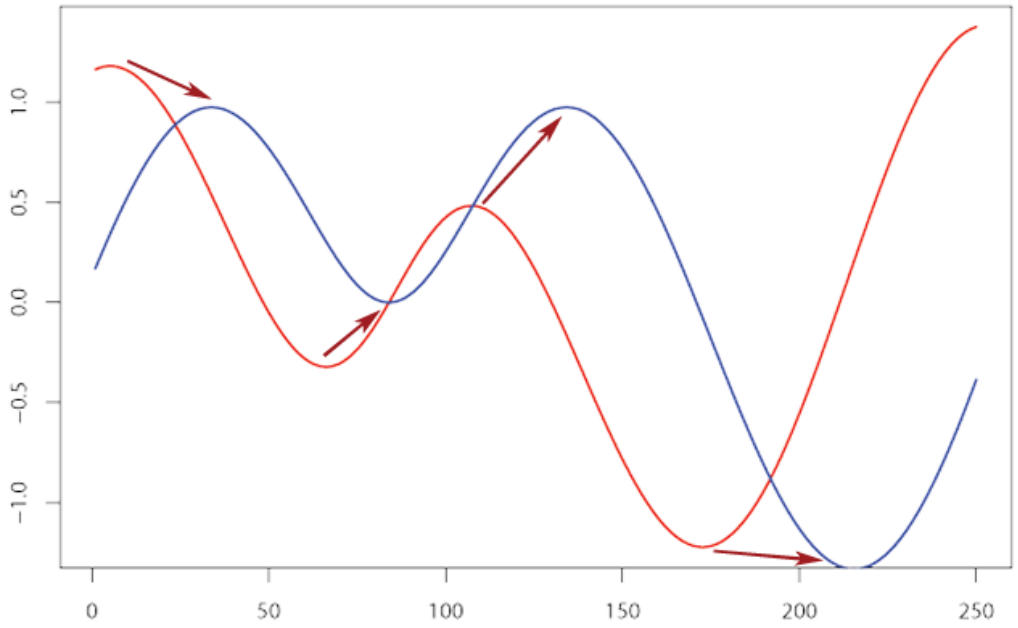
Cervical Dystonia (CD) dataset

- beter behouden data structuur, maar nog niet perfect

Health Campus Den Haag

Component (b): Tijdsreeks afstandsmaat

- tSNE obv *Dynamic Time Warping (DTW) algoritme:*
  - Berekent afstanden tussen reeksen na uitlijning op hun gelijke vormen
  - Euclidische → Gower maat
    - (vanwege gemengde datatypes)

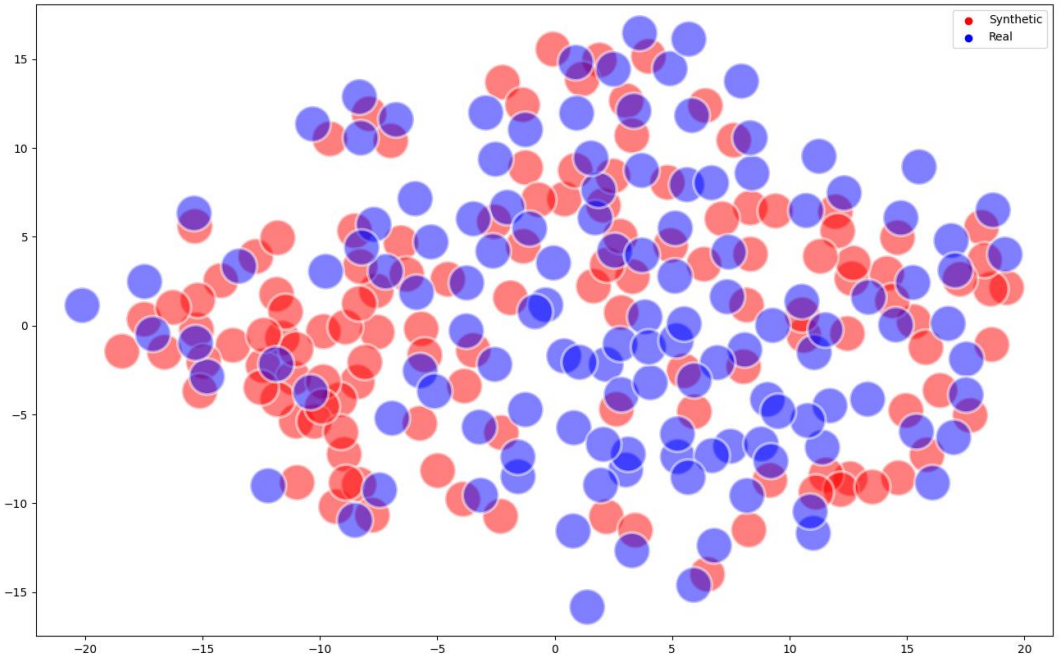


Health Campus Den Haag

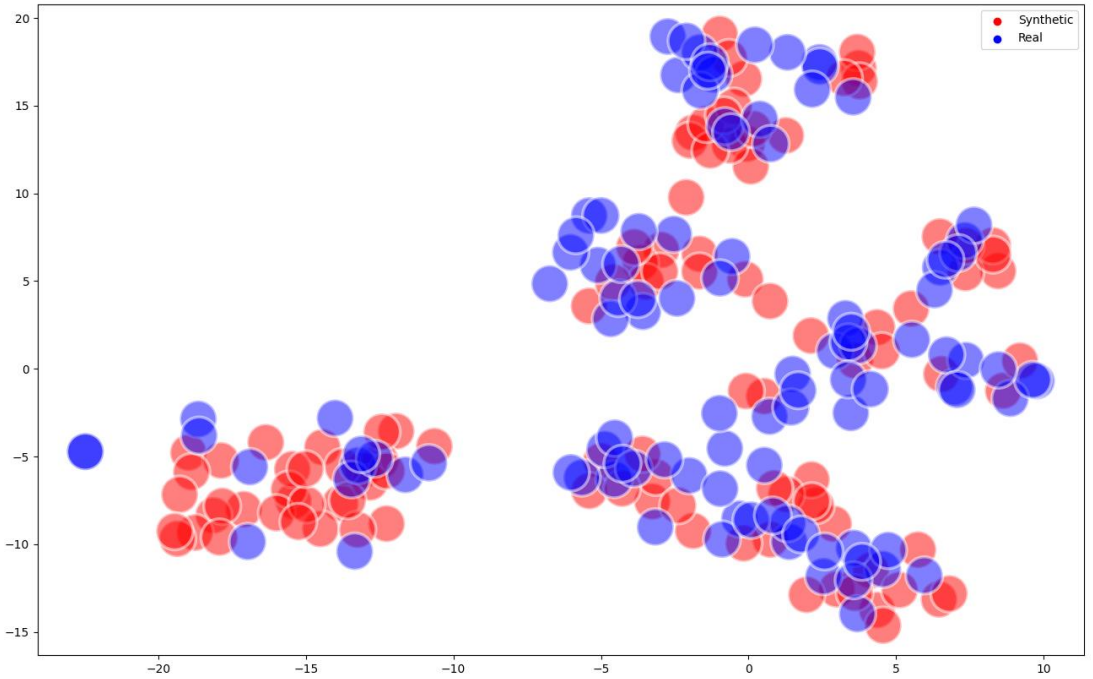
Senin, P. (2008). *Dynamic time warping algorithm review.* In: Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA 855.1-23, p. 40. [online]

Componenten (a+b): m-tSNE visualisatie

<dimensiereductie via t-distributed Stochastic Neighbor Embedding>
<met behoud van locale structuur per tijdsreeks (=nieuw)>



- *Visualisatie van gehele tijdsreeksen ipv afzonderlijke tijdstappen, door afstanden te berekenen via DTW + Gower*



Neuromuscular Monitoring (NM) dataset
- wederom enige 'puntenklodders'
  te weinig data diversiteit = *'mode collapse'*

→

Cervical Dystonia (CD) dataset
- ook enige 'puntenklodders'

Health Campus Den Haag

Component (c): Medische Goodness-of-Fit test

- Multi-variate *two-sample GOF* test:
  - Zijn twee distributies gelijk of niet?? Test uitvoeren!
  - Vereist wel een geschikt voorspelmodel, die samples van gemengde tijdseries en statische attributen kan beoordelen
    - Naar Friedman's (2003) intuïtie: **"als twee samples uit dezelfde distributie komen, dan kunnen ze op gelijke wijze gescoord worden door een voorspelmodel"**
    - Voorspelmodel kan hetzelfde RNN model zijn als voor evaluatie van discriminerende scores

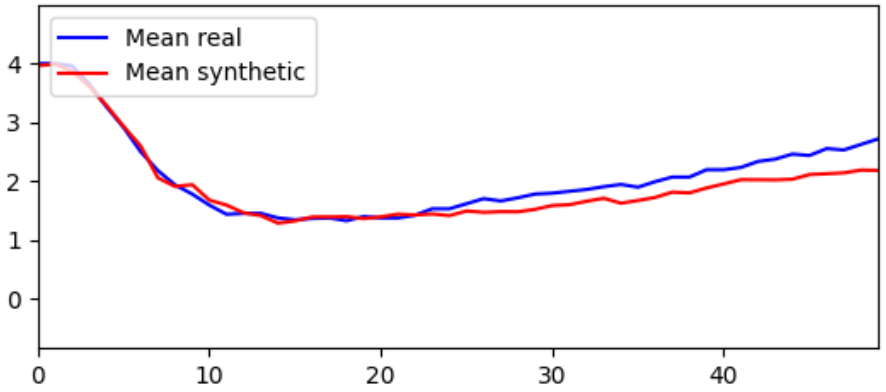
Health Campus Den Haag

Friedman, J. H. (2003). On multivariate goodness–of–fit and two–sample testing. *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, *1*, 311. [online]

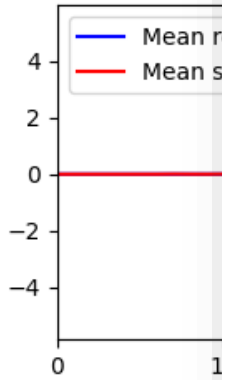
Evaluatiemaat: '*Fidelity*'

- ***Gelijkenis:*** Gelijkenis van distributies tov originele data
  - *i. Descriptieve statistiek:* bereik, gemiddelde, standaard deviatie
  - *ii. Twee-dimensionele visualisatie:* tSNE'
  - *iii. Medical Goodness-Of-Fit (MGOF)* test

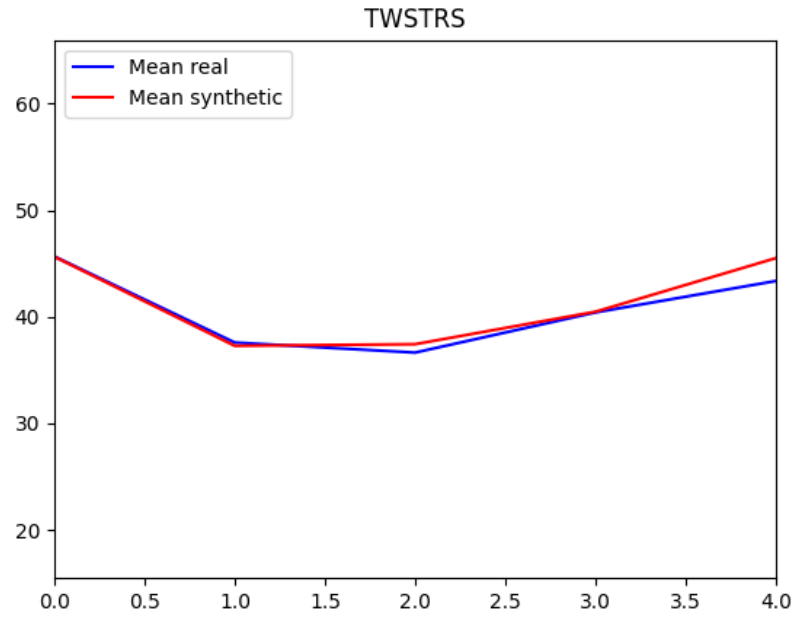
i) Aantal spiertrekkingen onder narcose



i) Hoeve



i) Gemiddelde ontwikkeling over de tijd



Health Campus Den Haag

Evaluatiemaat: ‘*Utility*’

- ***Nut:*** Nut bij gebruik in praktijktaken tov originele data
  - Voorspelmodel trainen op synthetisch vs originele data → vergelijken
  - *Groot nut =?* Synthetische data is bruikbaar in praktijk ipv originele data

  i. Train Synthetic, Test Real (TSTR)

  ii. Train Real, Test Real (TRTR)

- ***“Gouden standaard”:*** via reeds gepubliceerde ELAN data-analyses
  - Niet enkel vergelijken met originele data, ook met origineel voorspelmodel

Health Campus Den Haag

Evaluatiemaat: '*Privacy risk*'

- ***Onthullingsrisico:*** *Nut* bij gebruik in praktijktaken tov originele data
  - *'Attribute Inference Attack (AIA)':* Voorspelmodel trainen op synthetische data, om gevoelige attributen uit originele test-data te achterhalen

- Origineel=TRTR, GAN=TSTR
- Bij AIA accuratesse:
  - laag-laag: predict--, privacy++
  - hoog-laag: fidelity--,
  - hoog-hoog: predict++, privacy—
    - ~50% = willekeurige voorspelling

      → weinig risico op onthulling

Maat	Attribuut	Model	Dataset		
AIA Accuratesse	*Geslacht*			*NM*	*CD*
			Origineel	0.692	0.690
			GAN	0.558	0.55
Gemiddelde fout (MAPE)					
	Leeftijd			*NM*	*CD*
			Origineel	0.392	0.155
			GAN	0.695	0.193
	BMI			*NM*	
			Origineel	0.181	
			GAN	0.21	

Health Campus Den Haag

Health Campus Den Haag

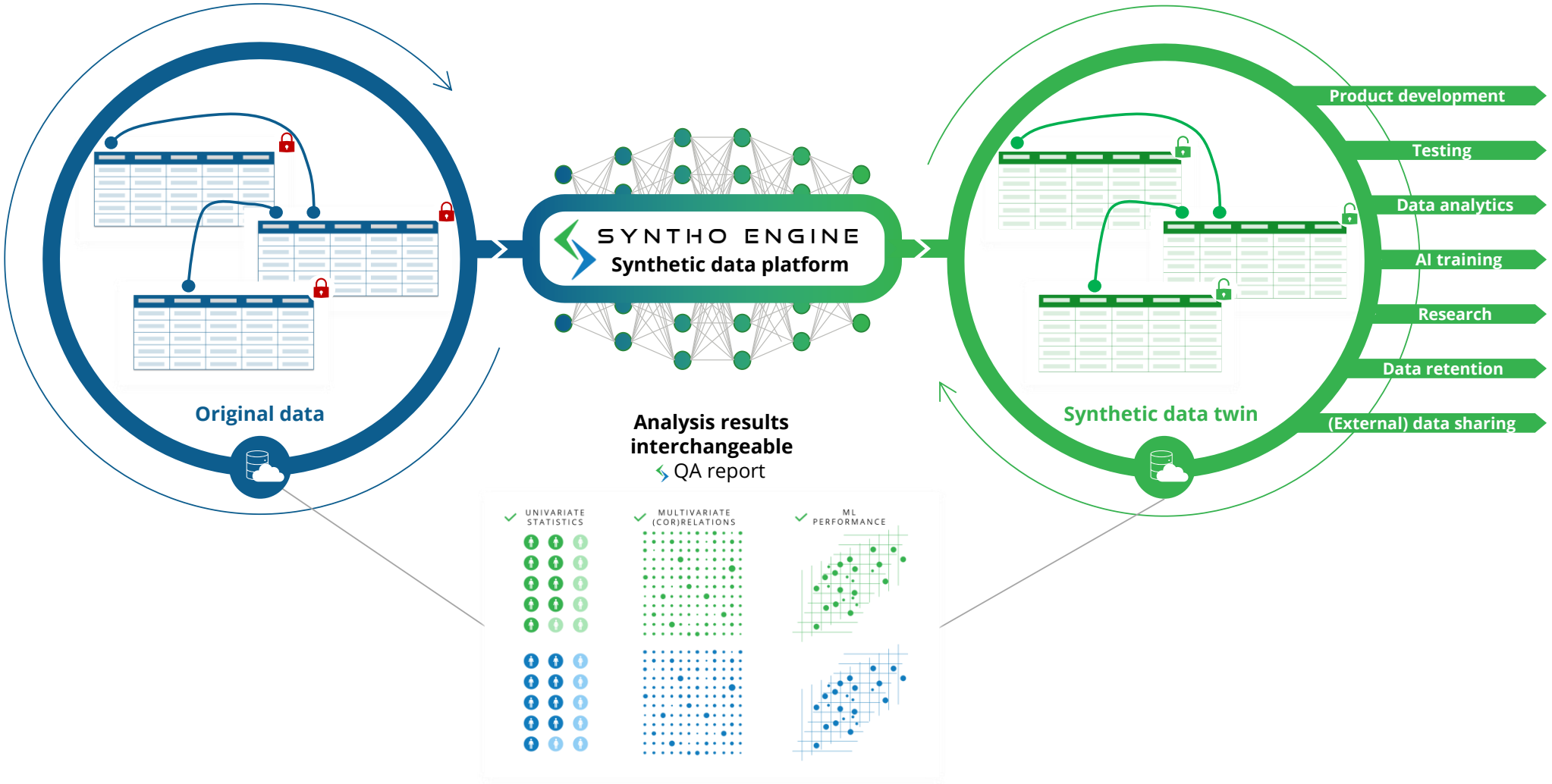
Technieken voor synthetische data

Drie benaderingen

A. Regel-gebaseerd → Agent-Based Modelling (ABM)
B. Neuraal netwerk-gebaseerd → Generative Adversial Networks (GAN)
C. Off-the-shelf → Syntho

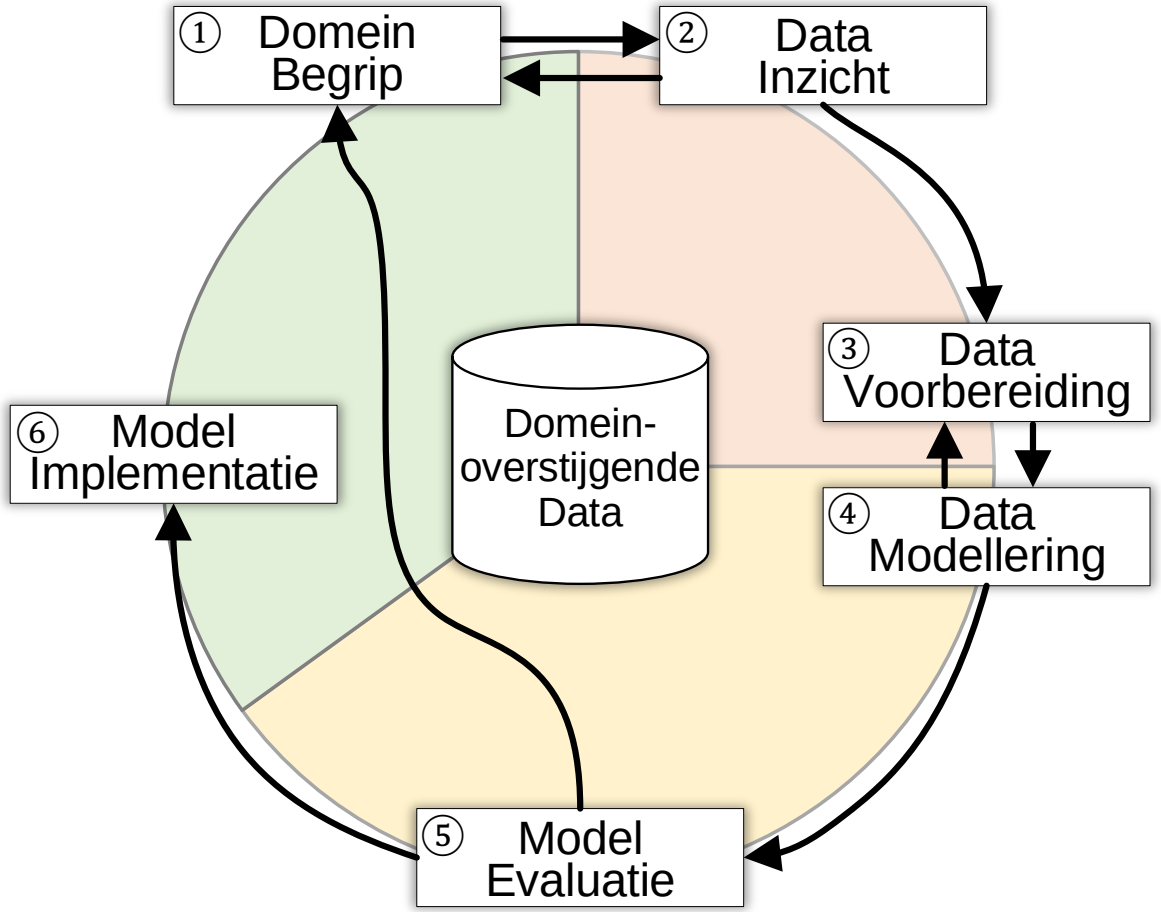

Ammar Faiq Jim Achterberg

C. Off-the-shelf (Syntho *platform*)



https://www.syntho.ai/syntho-engine/

C. Off-the-shelf (Syntho *platform*)



Health Campus Den Haag

Health Campus Den Haag

Et cetera

Effecten en de nabije toekomst

André Krom

Aanpak begeleidingsethiek

Fase 1
Casus

Fase 2
Dialoog

Fase 3
Handelingsopties

Technologie in context

Actoren

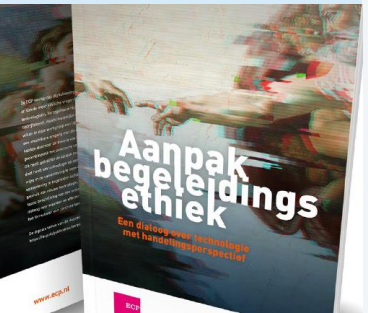
Waarden

Effecten

Technologie

Omgeving

Gebruik



Nabije toekomst

- **CPAN**: Probabilistische Autoregressieve Netwerken (PAN)
  - als regulier neuraal netwerk, maar met vertraagde invoer als tijdsmodel
- **Tekstdata** generatie: ABM+GAN = chatty-notes
  - Synthetische *FHIR Bundle* → episode extractie → ChatGPT prompt
  - Eerste GPT-2 poging met MI afstudeerder helaas voortijdig gestrand
- Samenwerking met promotieonderzoek Universiteit van **Murcia**
  - Onderzoek gaande op Madrileense EPD data (2 ziekenhuizen)
- **Nu:** projectsamenwerking met…

Health Campus Den Haag

SYNTHO

Health Campus Den Haag

Dank! m.r.spruit@lumc.nl

“Op naar een digitale tweeling mét ELAN”

VIPP-CBS is een samenwerkingsverband tussen CBS, LUMC, Universiteit Leiden, Syntho en Health Campus Partners