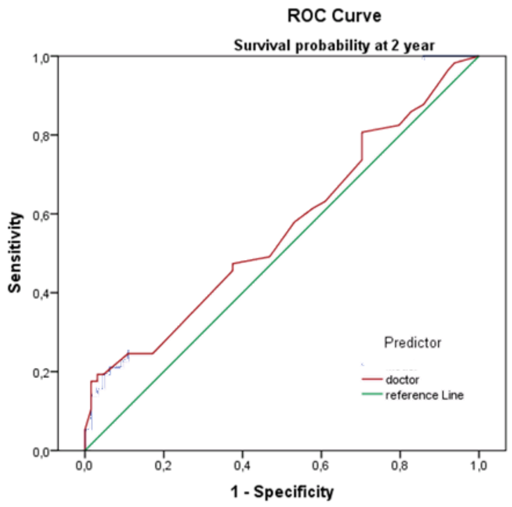
Maastricht UMC+

Perspectives in Health Science
Scientific Developments in Coming Decade

Andre Dekker
Medical Physicist, Professor of Clinical Data Science
Maastricht UMC+, Maastricht University, MAASTRO Clinic

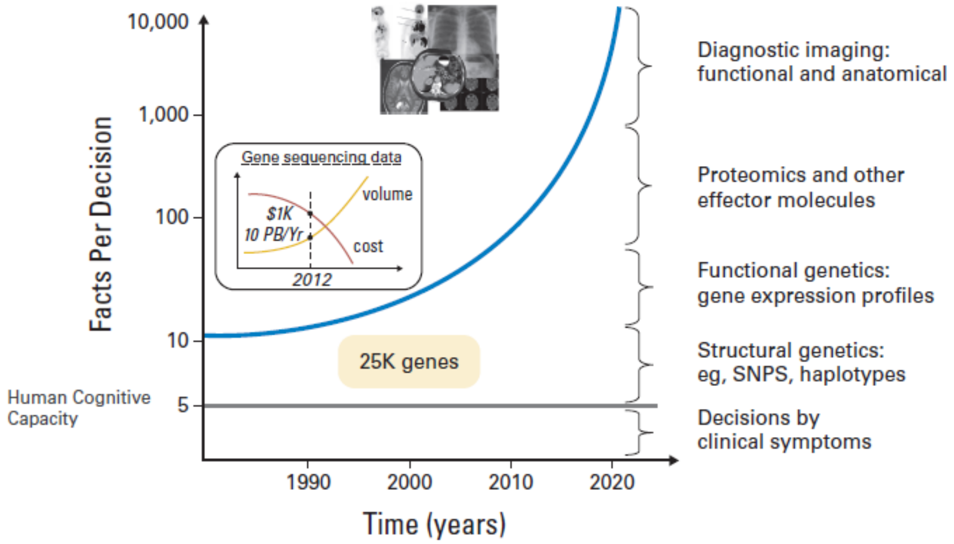
GEDE Workshop - Data Sharing a High Priority - Urgent Need to Act
April 14, 2020 Online Meeting https://zoom.us/j/902516996
09:55-10:15

Prediction of individual outcomes – we are drowning

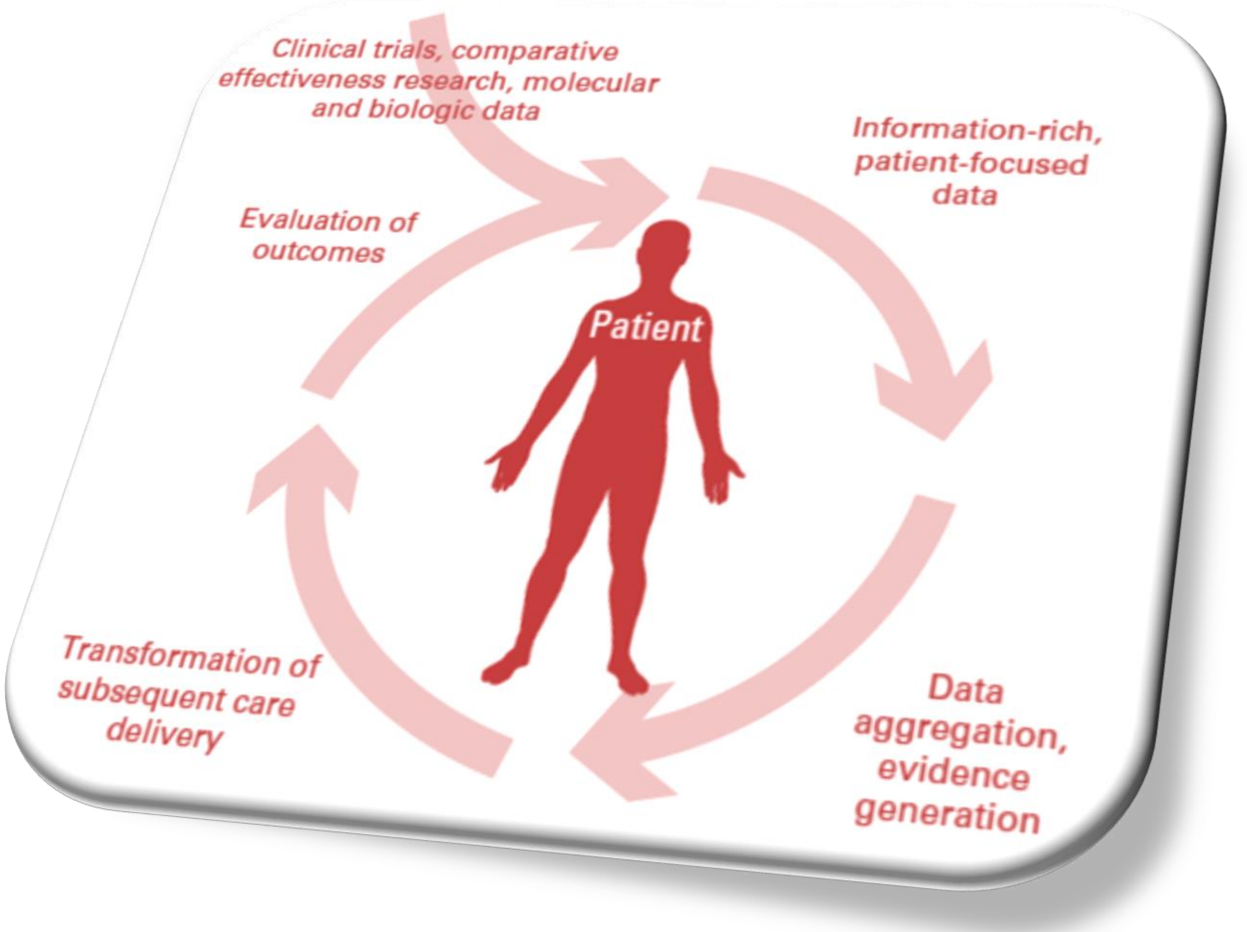


Lung Cancer
2 year survival
158 patients
5 MDs
Prospective
AUC: 0.56

- Explosion of data
- Explosion of decisions
- Explosion of 'evidence'
  - Too much to read
  - 3 % in trials, bias
  - Sharp knife

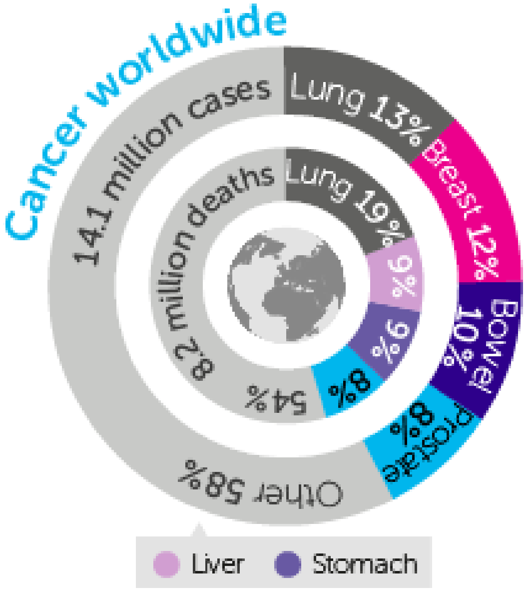


Learning Health System



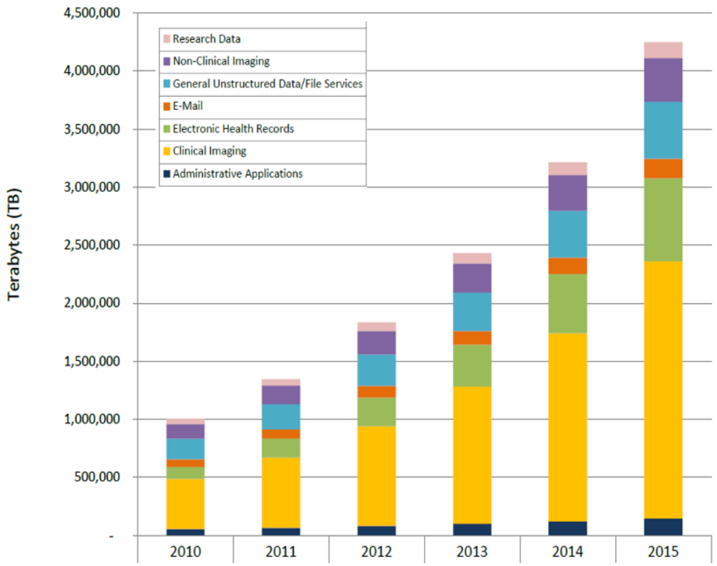
J Clin Oncol 2010;28:4268

A data example from cancer



Oncology
2007-2017
150M patients
0.1-10GB per patient

15-1500PB
80% unstructured



Hospitals
China: 25.000
India: 35.000
Germany: 2.000
France: 2.300
Italy: 1.100
USA: 5.500
Australia: 1.400

TOTAL ~100.000

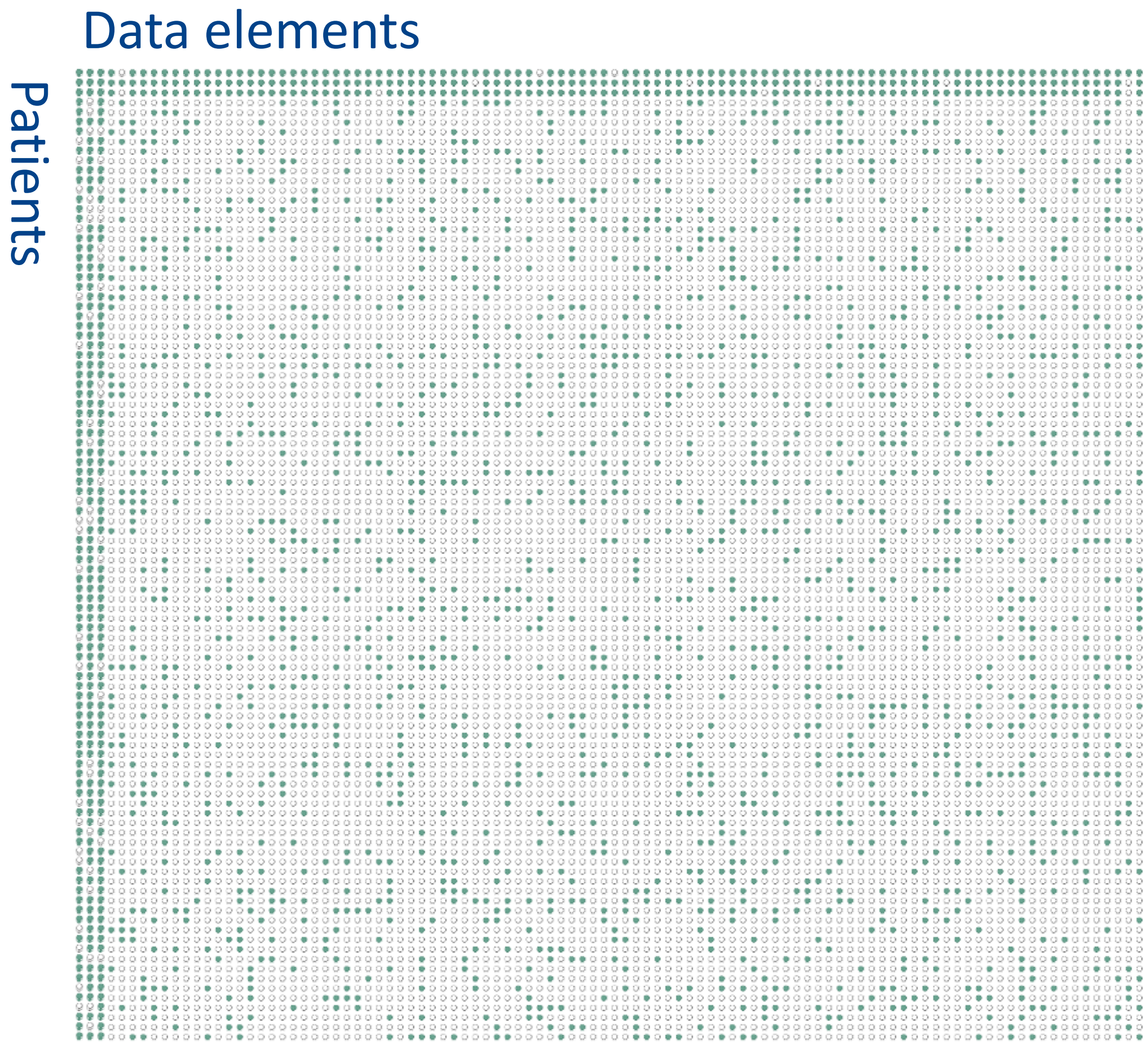
Barriers to sharing data

[..] the problem is not really technical [...]. Rather, the problems are **ethical, political, and administrative**.
Lancet Oncol 2011;12:933

1. Administrative (I don’t have the resources)
2. Political (I don’t want to)
3. Ethical (I am not allowed to)

4. Technical (I can’t)

Data landscape



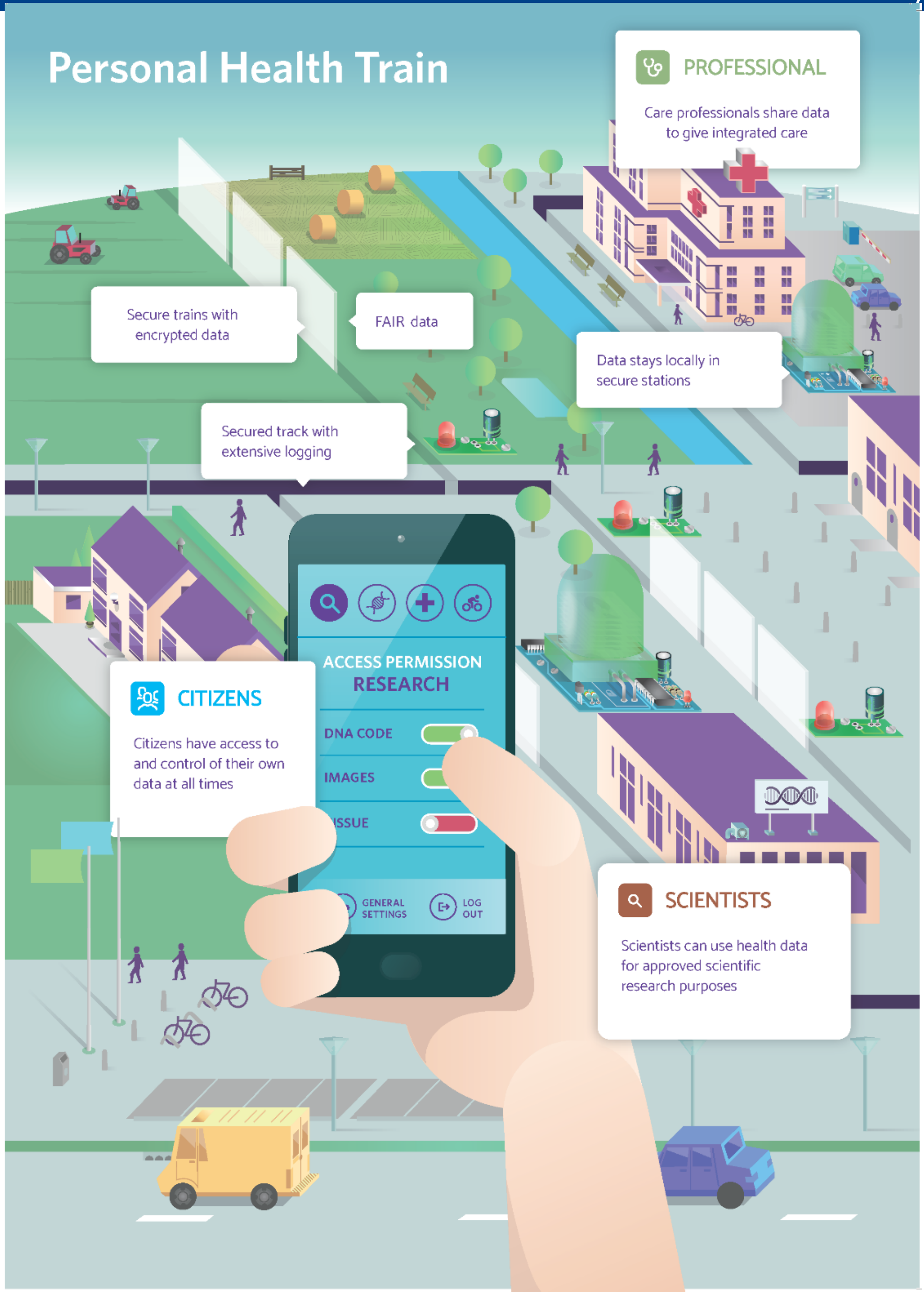
- Research
    - 3% of patients
    - 100% of features
    - 5% missing
    - 285 data points
- Registries
    - 100% of patients
    - 3% of features
    - 20% missing
    - 240 data points
- Routine
    - 100% of patients
    - 100% of features
    - 80% missing
    - **2000** data points

A different approach

- If sharing is the problem: Don't share the data

- If you can't bring the data to the research
- You have to bring the research to the data

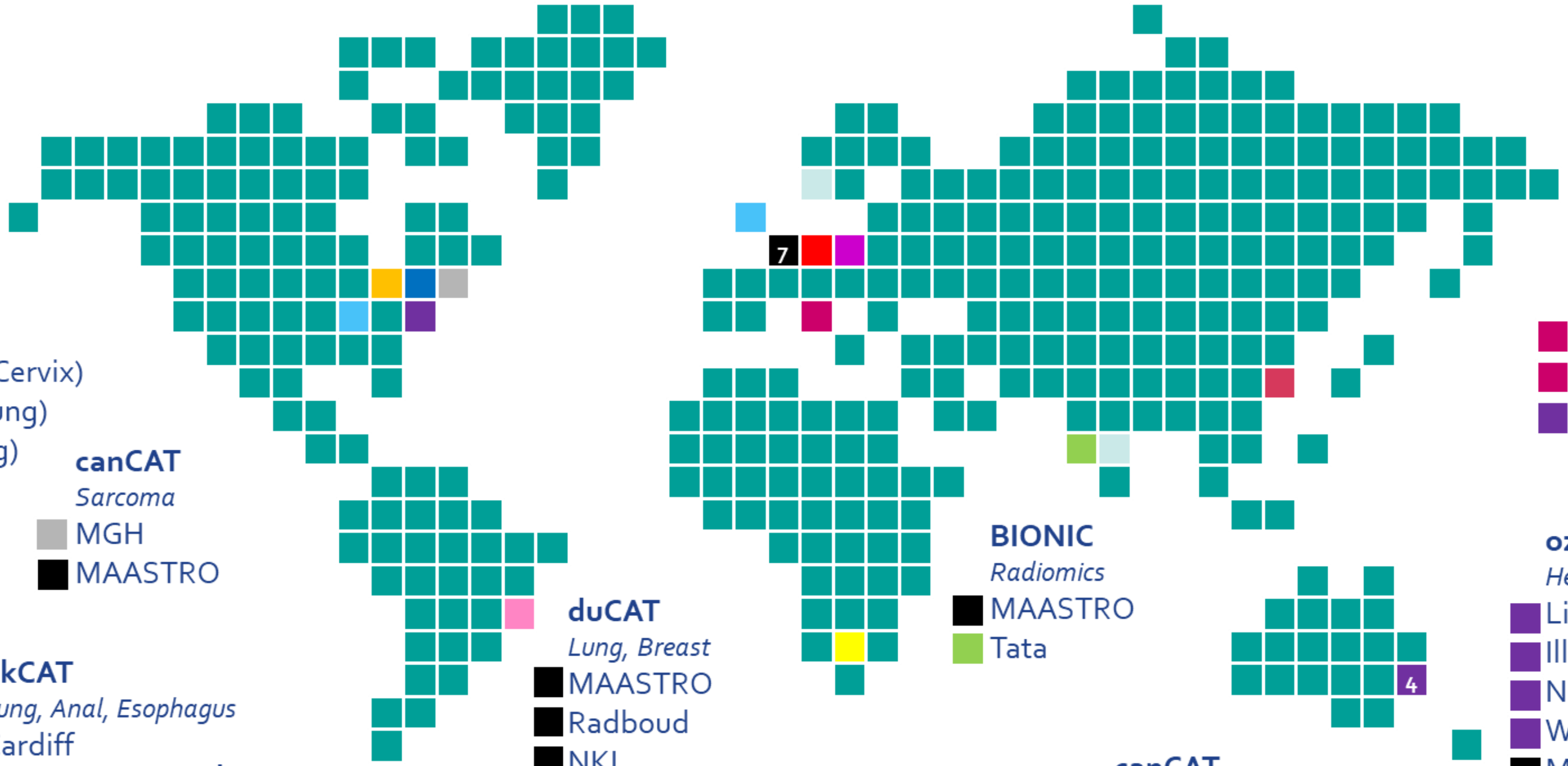
- Challenges
  - The research application has to be distributed (trains & track)
  - The data has to be understandable by an application (i.e. not a human) -> FAIR data stations



Maastricht UMC+

Wilkinson, DOI: 10.1038/sdata.2016.18

CORAL: Community in Oncology for RApid Learning



meerCAT
Lung
- U Michigan
- MAASTRO
- The Christie

Interested
- Oslo (Anal)
- Bloemfontein (Cervix)
- Odense (HN, Lung)
- Sao Paulo (Lung)
- McGill (Brain)

TRAIN
Radiomics
- MAASTRO
- Tata
- HGC

canCAT
Sarcoma
- MGH
- MAASTRO

ukCAT
Lung, Anal, Esophagus
- Cardiff
- Christie
- Leeds

poleCAT
Head&Neck
- Poznan
- MAASTRO

duCAT
Lung, Breast
- MAASTRO
- Radboud
- NKI
- Erasmus
- UMCU

cuCAT
Esophagus
- Cardiff
- MAASTRO

BIONIC
Radiomics
- MAASTRO
- Tata

canCAT
Lung
- Princess Margaret
- MAASTRO

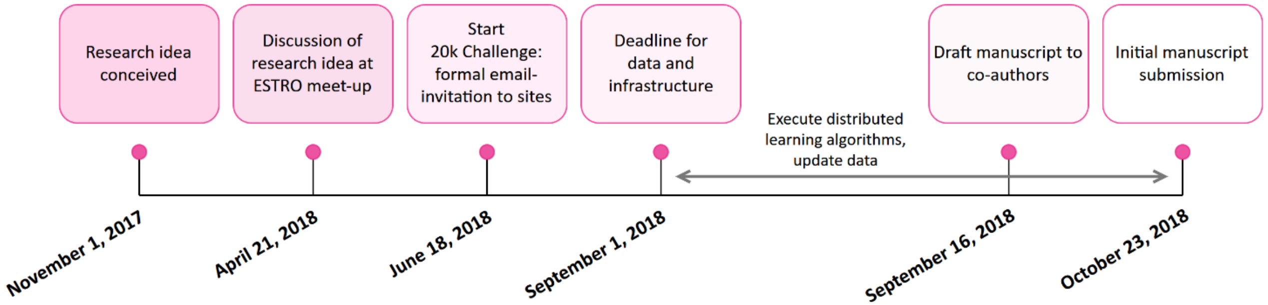
euroCAT
Lung
- UK Aachen
- LOC Hasselt
- Catharina
- MAASTRO
- CHU Liege

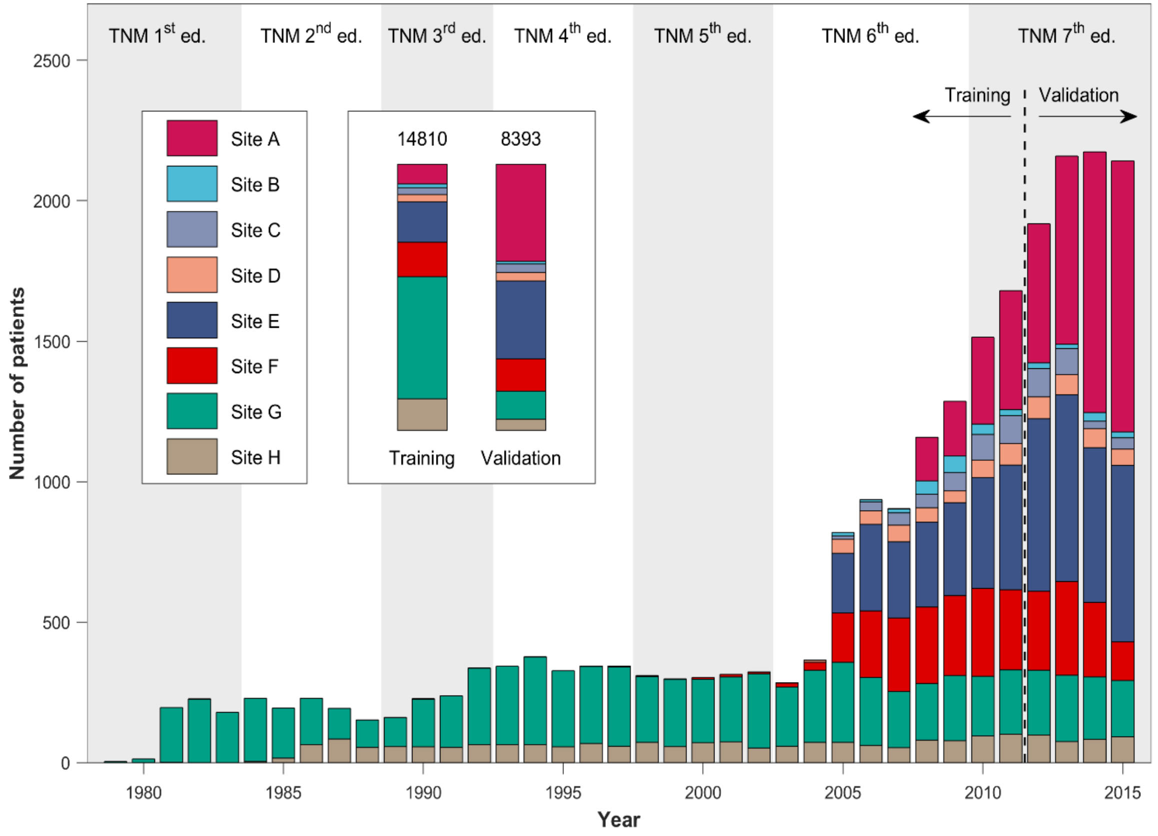
sinoCAT / VATE / SAGE
Rectum
- Fudan
- Rome
- RTOG/NRG

ozCAT
Head&Neck, Lung
- Liverpool
- Illawarra
- Newcastle
- Westmead
- MAASTRO
- RTOG/NRG

20k challenge

- Maastricht, Amsterdam, Cardiff, Nijmegen, Manchester, Rome, Rotterdam, Shanghai
- 37090 NSCLC patients





Deist, T. M. *et al.* Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. *Radiotherapy and Oncology* **144**, 189–200 (2020).

Discussion

How will data science develop in the coming decades anticipating the huge amounts of data being created and processed, and their inherent complexity given the many smart devices that are deployed everywhere?

- Federated learning on FAIR data sources

Are there already flagship projects in data science and/or data business that can indicate directions of developments? How will they evolve?

- Yes, DataShield, OHDSI, Personal Health Train

Who is this?

Pressure-Volume Loops in Cardiac Surgery

Proefschrift

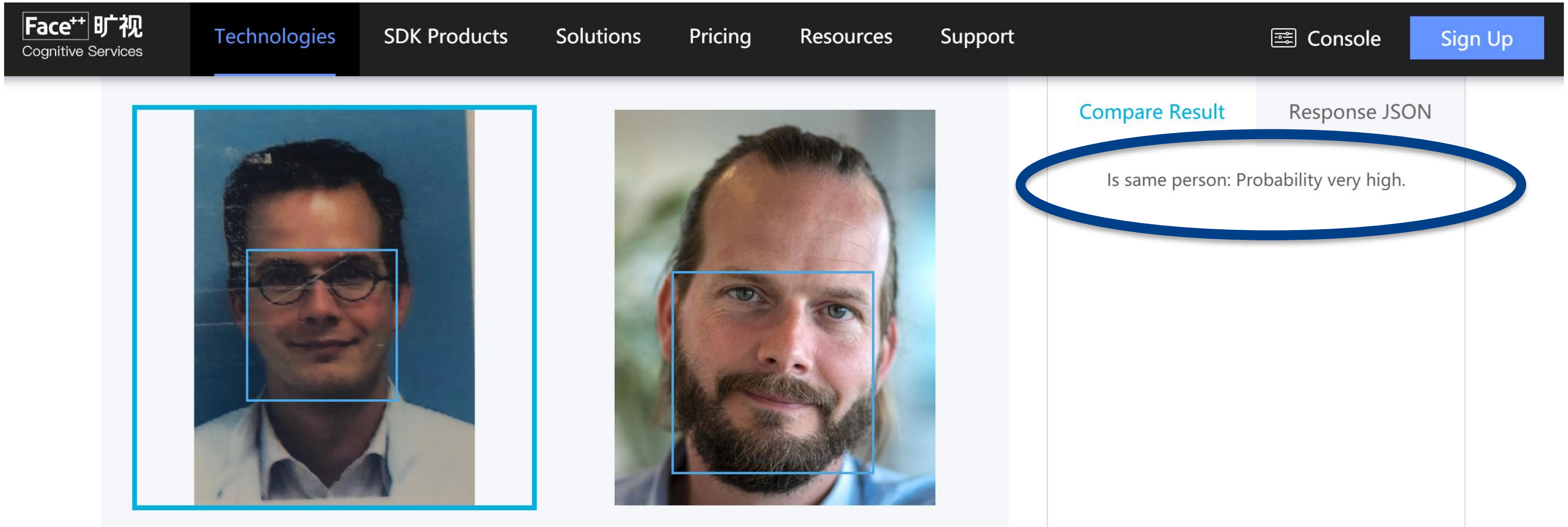
ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof.dr. A.C. Nieuwenhuijzen Kruseman,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen,
op vrijdag 12 september 2003 om 14:00 uur

door

André Dekker

UNIVERSITAIRE PERS MAASTRICHT

What is high quality data?

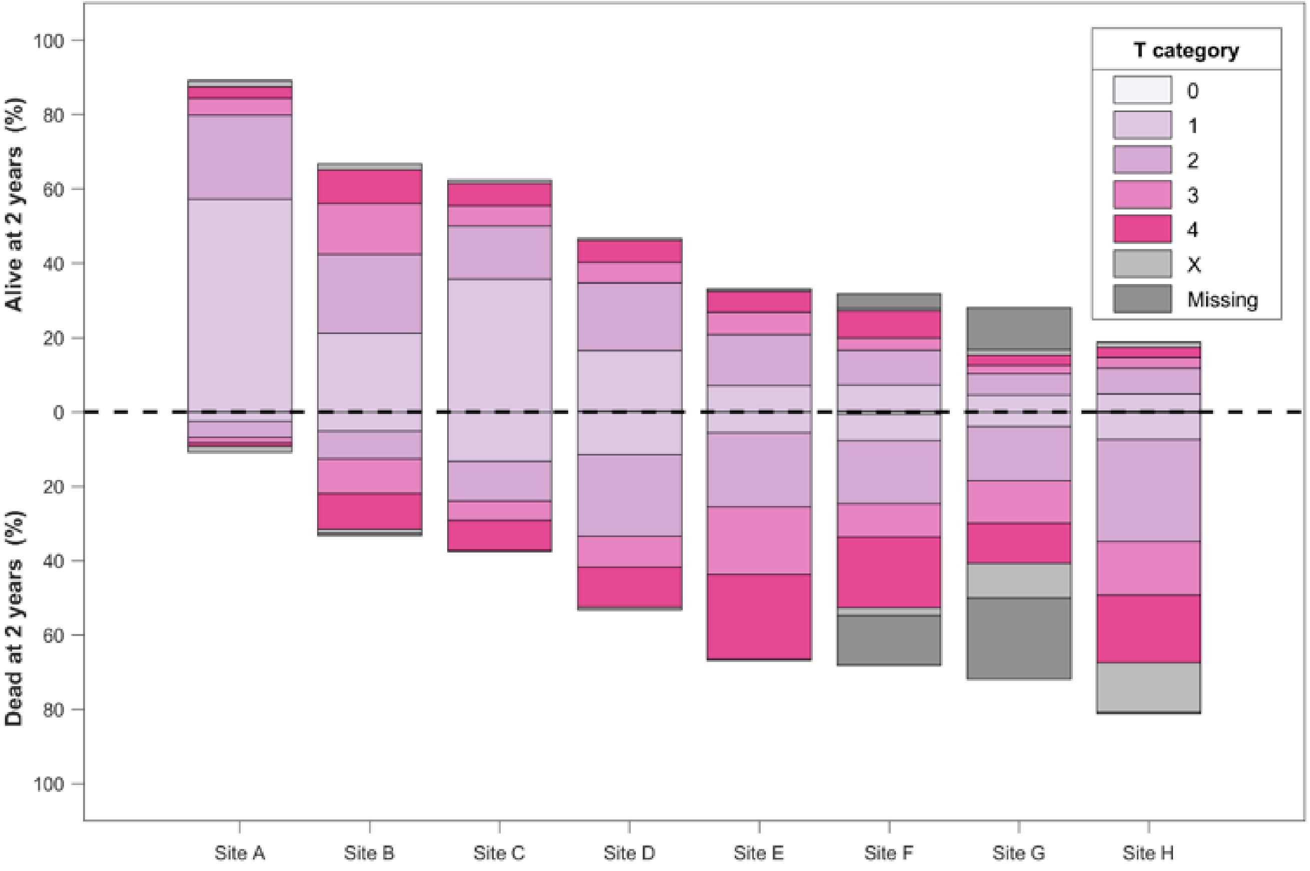


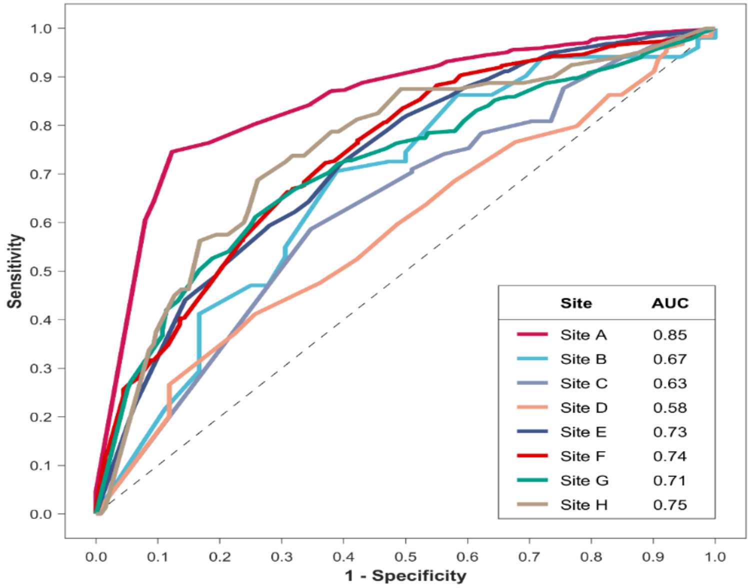
- What you think is high quality might be different for another person or for an AI
- Data quality is a characteristic of the question not of the data

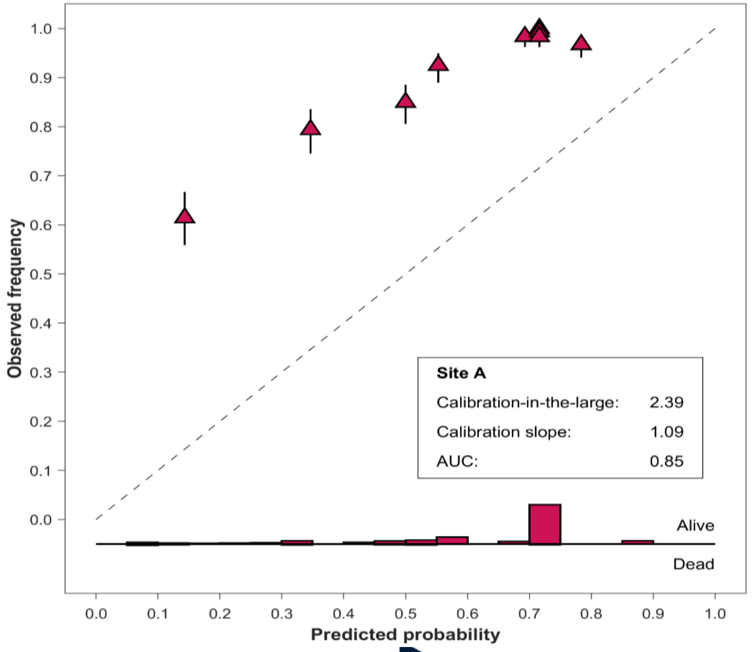
https://www.faceplusplus.com/face-comparing/

20k challenge (and some COVID-19)

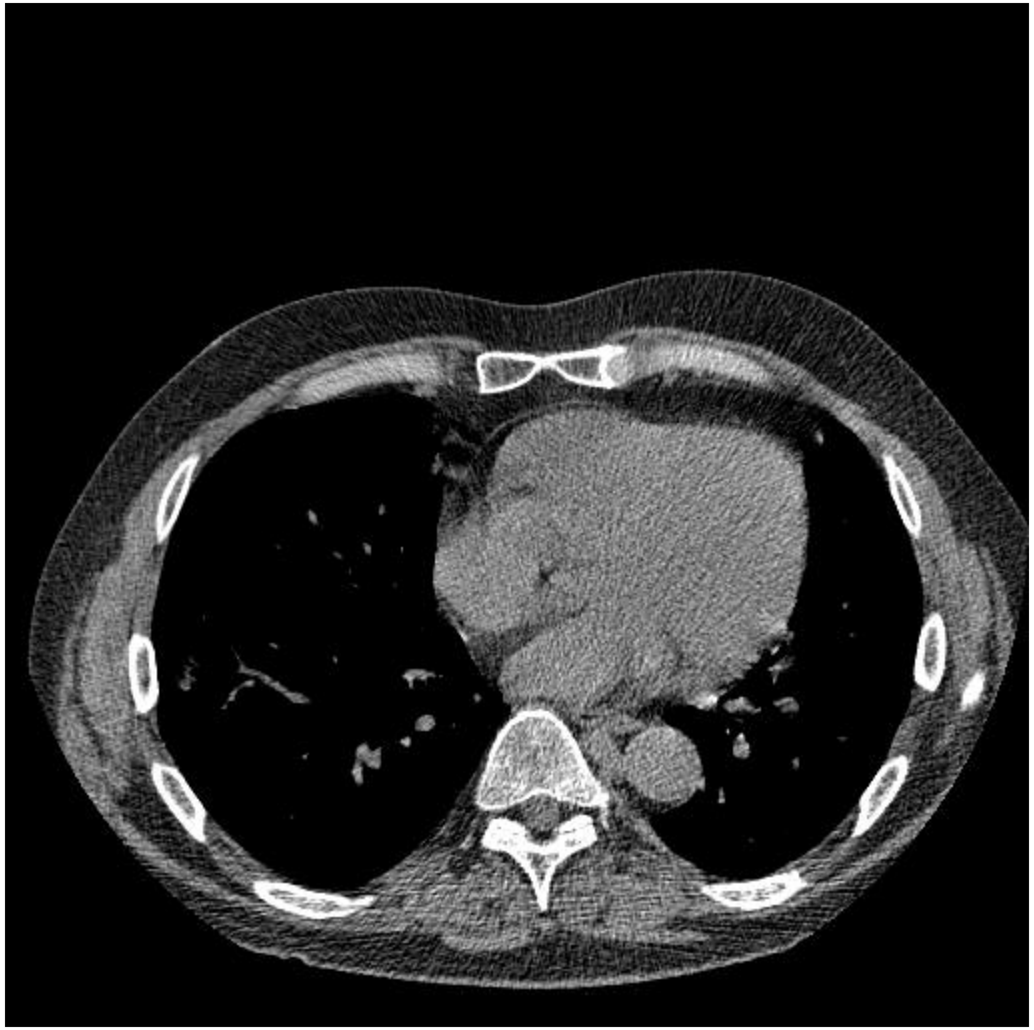
(a)

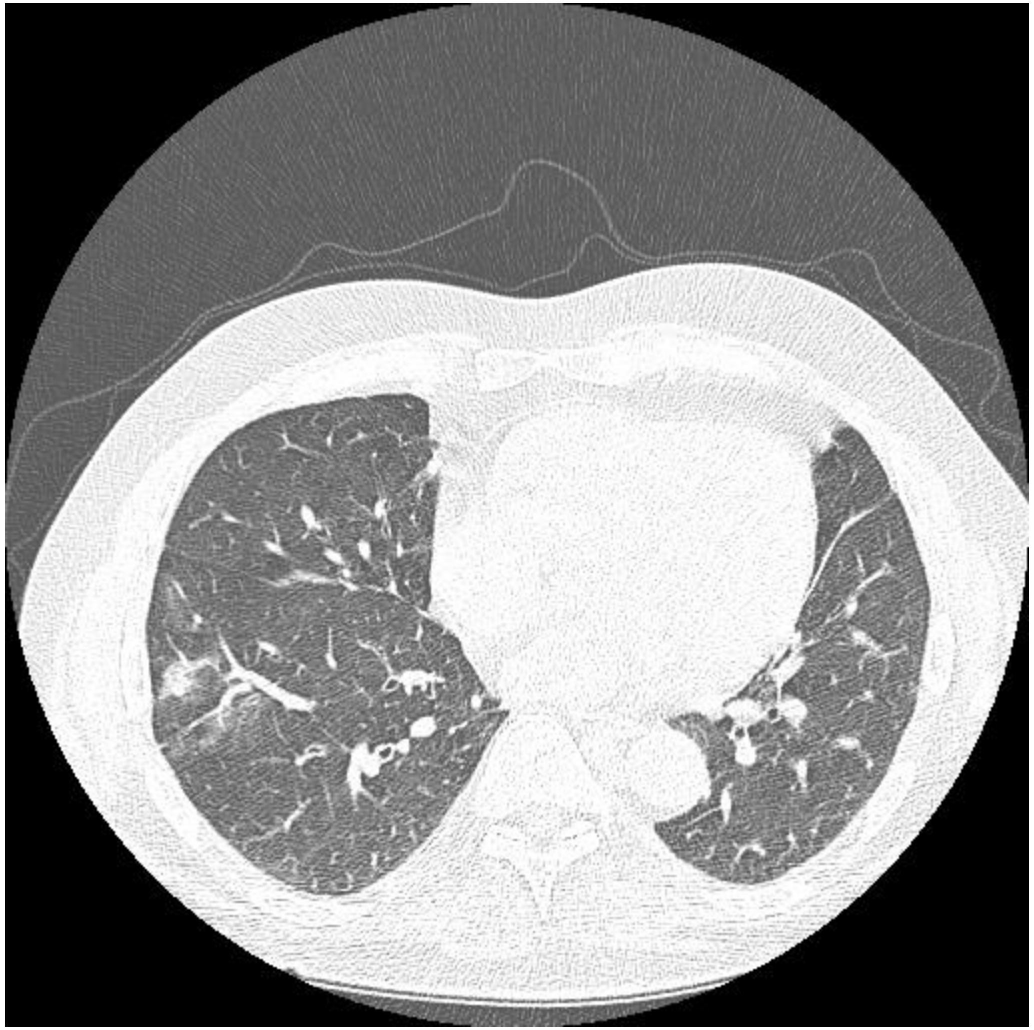






COVID-19





Discussion

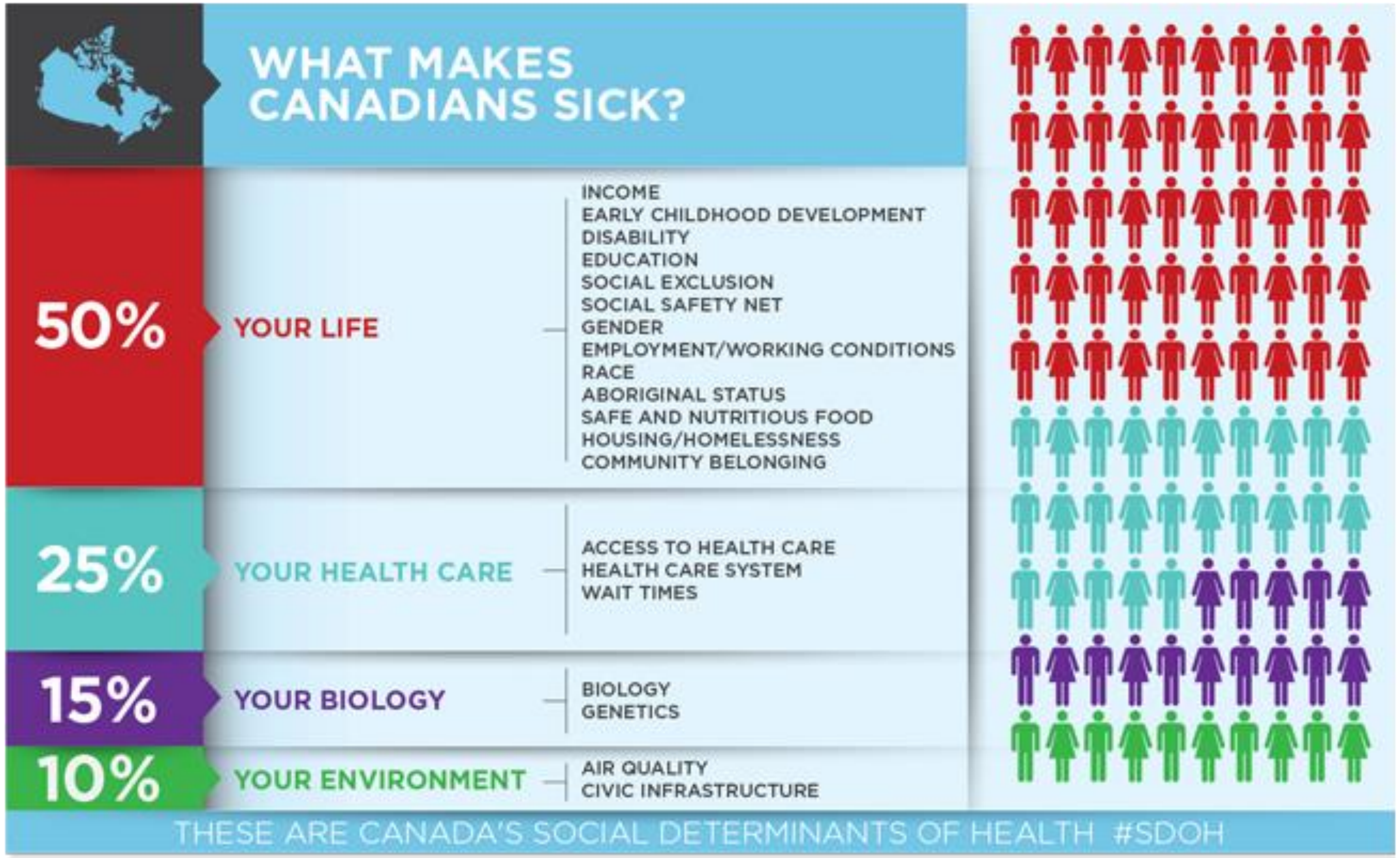
In such a scenario data quality and trust will play an important role. Which mechanisms will help to establish the required trust?

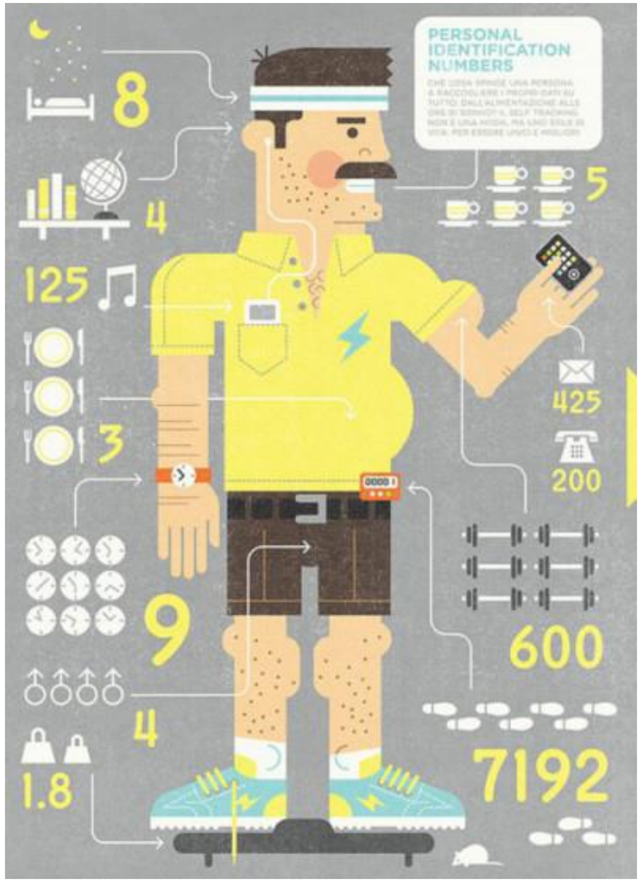
- Data quality and trust in data are a characteristic of the question NOT of the data

Trust in health data means more....

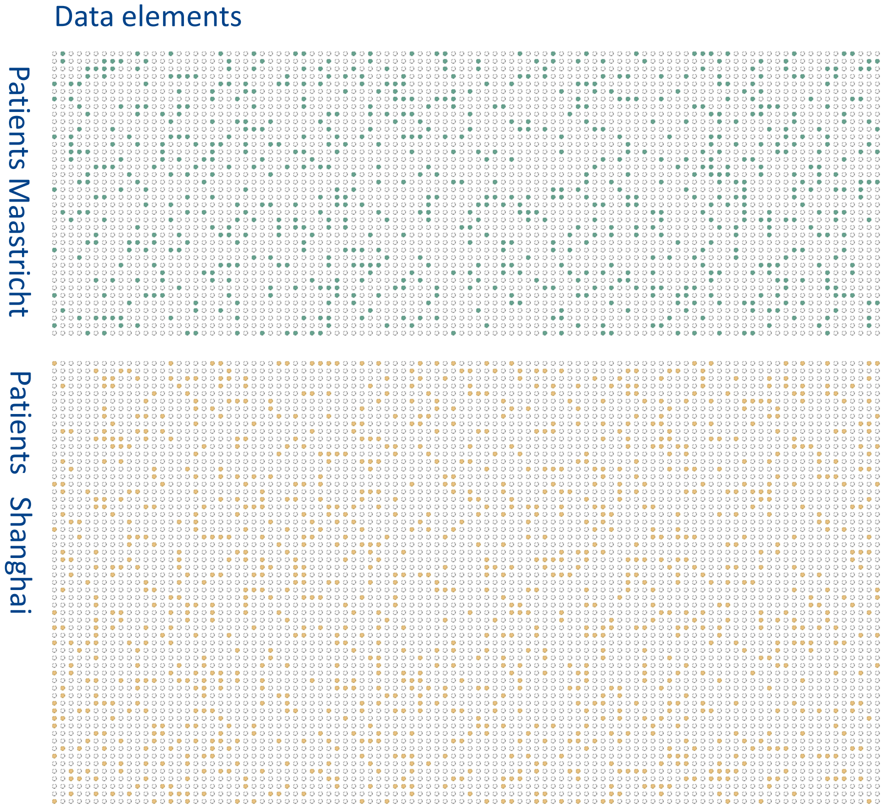
- Data Governance Technologies
- More attention to the A from FAIR
- FAIR trains

The most interesting data is probably outside of the hospital





Horizontal Partitions



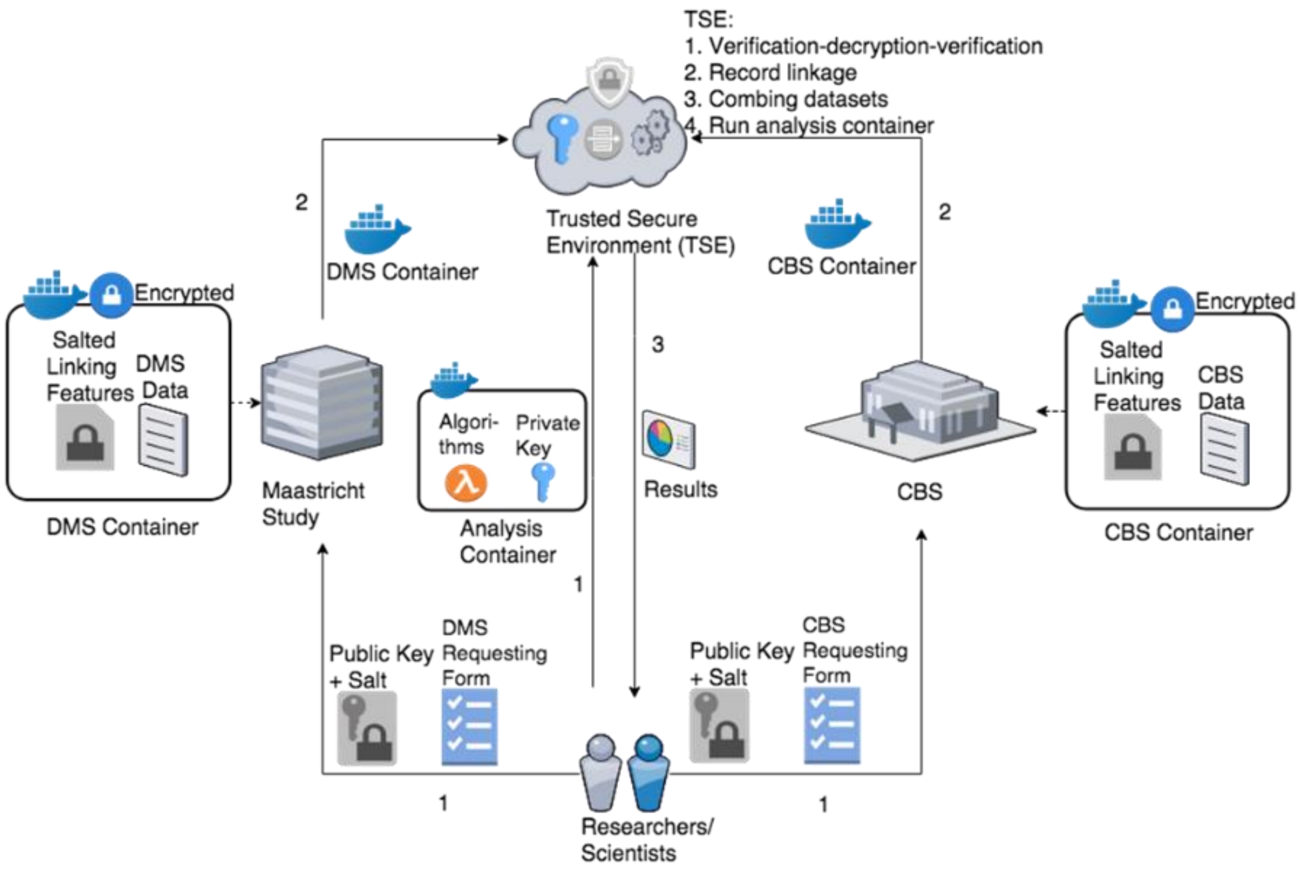
Dataset at Data Party A

ID	AGE	Employed	Type 2 Diabetes	Wellbeing	Education
1	56	NO	YES	GOOD	UNIVERSITY
2	25	YES	NO	MEDIUM	UNIVERSITY
3	31	YES	NO	GOOD	HIGH SCHOOL
4	45	NO	YES	POOR	PRIMARY SCHOOL

Dataset at Data Party B

ID	AGE	Employed	Type 2 Diabetes	Wellbeing	Education
5	32	YES	NO	VERY GOOD	SECONDARY SCHOOL
6	60	NO	YES	POOR	HIGH SCHOOL
7	55	YES	NO	MEDIUM	UNIVERSITY

Vertical Partitions



Dataset at Data Party A

ID	AGE	Employed	Type 2 Diabetes
1	56	NO	YES
2	25	YES	NO
3	31	YES	NO
4	45	NO	YES
5	32	YES	NO
6	60	NO	YES
7	55	YES	NO

Dataset at Data Party B

ID	Wellbeing	Education
1	GOOD	UNIVERSITY
2	MEDIUM	UNIVERSITY
3	GOOD	HIGH SCHOOL
4	POOR	PRIMARY SCHOOL
5	VERY GOOD	SECONDARY SCHOOL
6	POOR	HIGH SCHOOL
7	MEDIUM	UNIVERSITY

Van Soest et al., Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data, doi:10.3233/978-1-61499-852-5-581

Discussion

Will data science give impulses to cross-disciplinary/cross-silo work? If so, which mechanisms would be needed to carry out such data science efficiently?

- Yes, data determining health is everywhere
- Access mechanisms under control of citizens (e.g. Solid PODs)
- Data Governance Technologies
- Solving the vertical partitioning problem (linking subjects across sets)

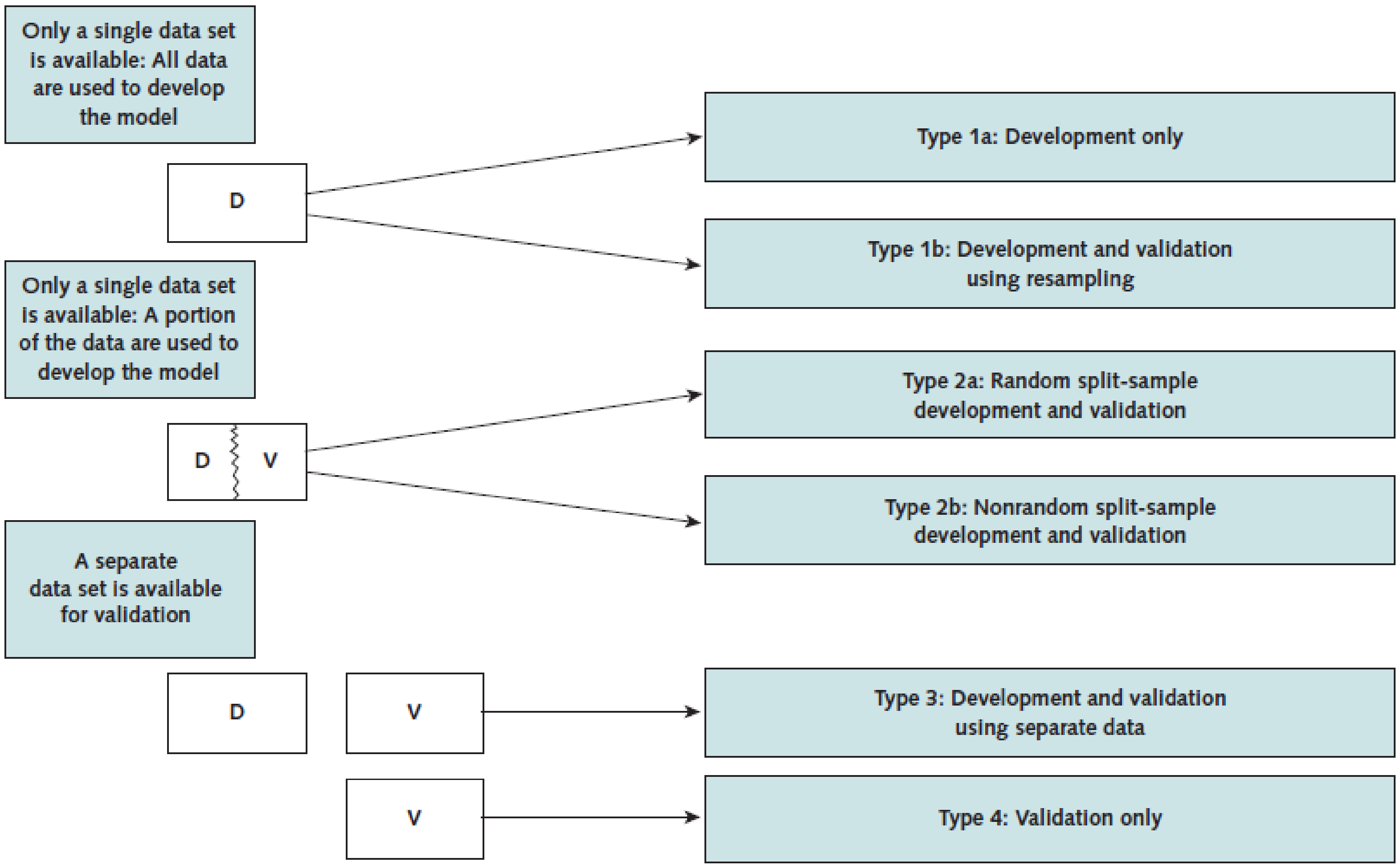
Performance vs Comprehensibility

Table 3 | Comparisons between human evaluations and different types of AI approaches

Approaches	Model comprehensibility	Performance	Reproducibility	Dependency on prior knowledge	Development and training costs[a]	Running costs	Around-the-clock availability	Update costs
Human evaluation	High	Moderate or high	Moderate	High	High	High	Low	High
Rule-based algorithms	High	Moderate or high	High	High	Moderate or high	Low	High	High
Feature-based machine-learning methods	Moderate or high	Moderate or high	High	Moderate[b]	Moderate	Low	High	Moderate[c]
Deep artificial neural networks	Low or moderate	High	High	Low	Moderate	Low	High	Low

TRIPOD

Figure 3. Types of prediction model studies covered by the TRIPOD Statement.



COVID-19 Example (do not share – unpublished work)

The promise

Table 5. The performance of deep COVID-19 classification model on test set.

	Accuracy	F1-Score	Precision	Recall	Specificity
COVID-19	0.961	0.951	0.927	0.977	0.950
Pneumonia	0.878	0.802	0.822	0.783	0.923
C&P	0.903	0.932	0.928	0.936	0.825

*C&P: COVID-19 & Pneumonia

The reality

	accuracy	F1	precision	recall	specificity	sensitivity
covid-19	0.7479	0.5490	0.7179	0.4444	0.9079	0.4444
pneumonia	0.2548	0.3399	0.2448	0.5556	0.0962	0.5556
^(pneumonia|covid-19)$	0.3534	0.5164	0.3481	1.0000	0.0126	1.0000

NPV	PPV
0.7561	0.7179
0.2911	0.2448
1.0000	0.3481

SCIENTIFIC DATA

OPEN ARTICLE

Distributed radiomics as a signature validation study using the Personal Health Train infrastructure

Zhenwei Shi[1,7*], Ivan Zhovannik[1,2,7], Alberto Traverso[1,6], Frank J. W. M. Dankers[1,2], Timo M. Deist[1,3], Petros Kalendralis[1], René Monshouwer[2], Johan Bussink[2], Rianne Fijten[1], Hugo J. W. L. Aerts[4,5], Andre Dekker[1] & Leonard Wee[1]

Discussion

Science and reproducibility are twins until now. Will this remain and if so, how can reproducibility be ensured, and which infrastructural mechanisms are required?

- Validation/generalizability is more important than reproducibility
- Distributed external, continuous validation needed

Acknowledgements

Netherlands
MAASTRO, Maastricht, Netherlands
Radboudumc, Nijmegen, Netherlands
Erasmus MC, Rotterdam, Netherlands
Leiden UMC, Leiden, Netherlands
Catharina Hospital, Eindhoven, Netherlands
Isala Hospital, Zwolle, Netherlands
NKI Amsterdam, Netherlands
UMCG, Groningen, Netherlands
IKNL, Utrecht, Netherlands

Europe
Policlinico Gemelli & UCSC, Roma, Italy
UH Ghent, Belgium
UZ Leuven, Belgium
Cardiff University & Velindre CC, Cardiff, UK
CHU Liege, Belgium
Uniklinikum Aachen, Germany
LOC Genk/Hasselt, Belgium
The Christie, Manchester, UK
State Hospital, Rovigo, Italy
St James Institute of Oncology, Leeds, UK
U of Southern Denmark, Odense, Denmark
Greater Poland Cancer Center, Poznan, Poland
Oslo University Hospital, Oslo, Norway

Africa
University of the Free State, Bloemfontein, South Africa

Asia
Fudan Cancer Center, Shanghai, China
CDAC, Pune, India
Tata Memorial, Mumbai, India
HGC Oncology, Bangalore, India

North America
RTOG, Philadelphia, PA, USA
MGH, Boston, MA, USA
University of Michigan, Ann Arbor, USA
Princess Margaret CC, Canada

South America
Albert Einstein, Sao Paulo, Brazil

Australia
University of Sydney, Australia
Westmead Hospital, Sydney, Australia
Liverpool and Macarthur CC, Australia
ICCC, Wollongong Australia
Calvary Mater, Newcastle, Australia
North Coast Cancer Institute, Coffs Harbour, Australia

Industry
Varian, Palo Alto, CA, USA
Philips, Bangalore, India
Sohard GmbH, Fuerth, Germany
Microsoft, Hyderabad, India
Mirada Medical, Oxford, UK
CZ Health Insurance, Tilburg, NL
Siemens, Malvern, PA, USA
Roche, Woerden, NL
Medical Data Works, Heerlen, NL

Acknowledgements

Netherlands
MAASTRO, Maastricht, Netherlands
Radboudumc, Nijmegen, Netherlands
Erasmus MC, Rotterdam, Netherlands
Leiden UMC, Leiden, Netherlands
Catharina Hospital, Eindhoven, Netherlands
Isala Hospital, Zwolle, Netherlands
NKI Amsterdam, Netherlands
UMCG, Groningen, Netherlands
IKNL, Utrecht, Netherlands

Europe
Policlinico Gemelli & UCSC, Roma, Italy
UH Ghent, Belgium
UZ Leuven, Belgium
Cardiff University & Velindre CC, Cardiff, UK
CHU Liege, Belgium
Uniklinikum Aachen, Germany
LOC Genk/Hasselt, Belgium
The Christie, Manchester, UK
State Hospital, Rovigo, Italy
St James Institute of Oncology, Leeds, UK
U of Southern Denmark, Odense, Denmark
Greater Poland Cancer Center, Poznan, Poland
Oslo University Hospital, Oslo, Norway

Africa
University of the Free State, Bloemfontein, South Africa

Asia
Fudan Cancer Center, Shanghai, China
CDAC, Pune, India
Tata Memorial, Mumbai, India
HGC Oncology, Bangalore, India

North America
RTOG, Philadelphia, PA, USA
MGH, Boston, MA, USA
University of Michigan, Ann Arbor, USA
Princess Margaret CC, Canada

South America
Albert Einstein, Sao Paulo, Brazil

Australia
University of Sydney, Australia
Westmead Hospital, Sydney, Australia
Liverpool and Macarthur CC, Australia
ICCC, Wollongong Australia
Calvary Mater, Newcastle, Australia
North Coast Cancer Institute, Coffs Harbour, Australia

Industry
Varian, Palo Alto, CA, USA
Philips, Bangalore, India
Sohard GmbH, Fuerth, Germany
Microsoft, Hyderabad, India
Mirada Medical, Oxford, UK
CZ Health Insurance, Tilburg, NL
Siemens, Malvern, PA, USA
Roche, Woerden, NL
Medical Data Works, Heerlen, NL

Discussion

What kind of common infrastructure and services are required to maintain a leading role for European data scientists? Which kinds of infrastructural support will be required to not load the data scientists which could better be done by data managers and stewards? What are the necessary timelines?

How will labs look like in about 10 years from now? What will be automated and what not? Which external services will be needed

- European data scientist do not have a leading role. We are lagging behind China and the USA
- We need distributed FAIR data infrastructures with especially a clearer path for Accessible sensitive data