Analyze personal data from multiple sources in a privacy-preserving manner

Chang Sun, PhD student
Institute of Data Science, Maastricht University

17/12/2019 Brussels

Center for Big Data Statistics

Maastricht University
Institute of Data Science

DE MAASTRICHT STUDIE

Take home message

Data Science for **Better** Decision:

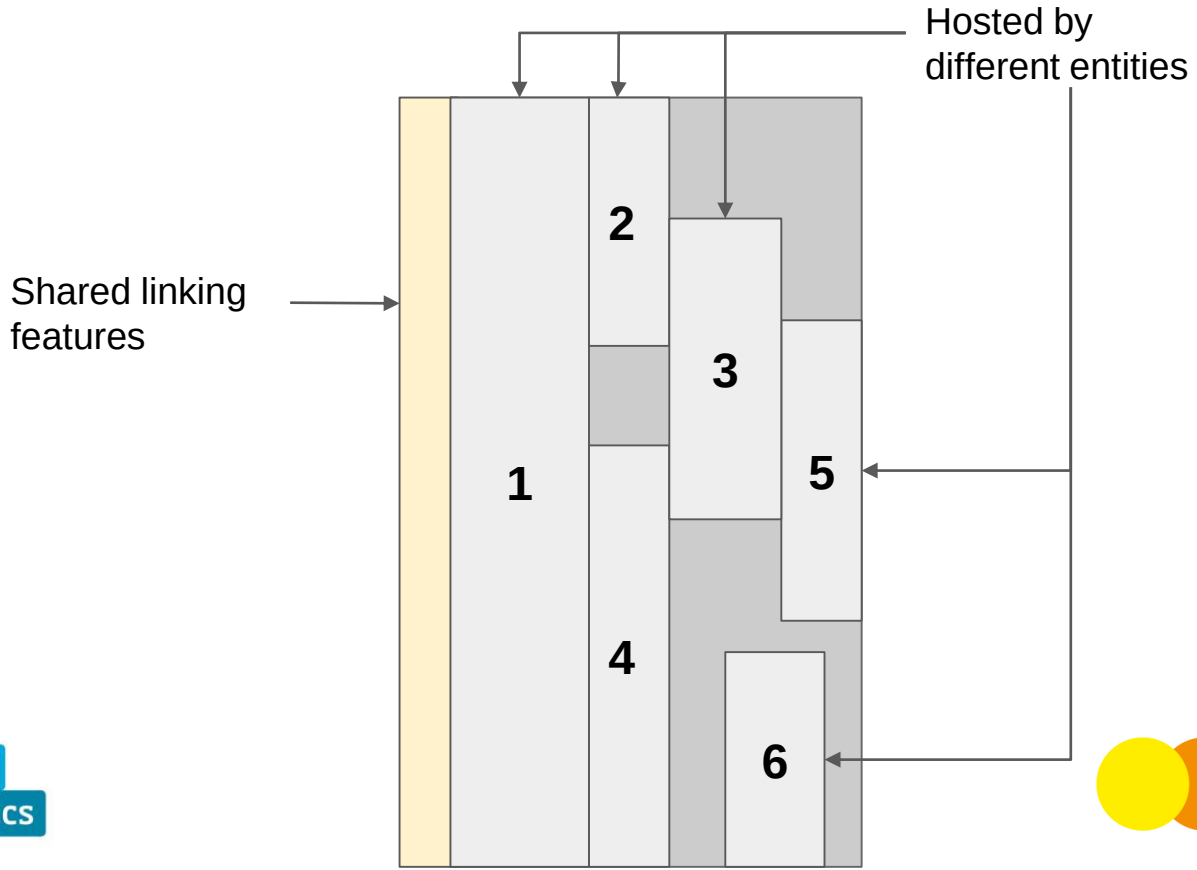
- Unlocking new possibilities by combining different data sources
- Privacy-preserving data analysis is a hot and moving topic
- Technical, legal, and ethical issues require multidisciplinary teams

Current machine learning models

+ High predictive performance
+ Data driven nature allows for unexpected discoveries
+ Personalized solutions/recommendations

- Black box
- Reliability of predictions
- Need large, diverse, and high-quality data

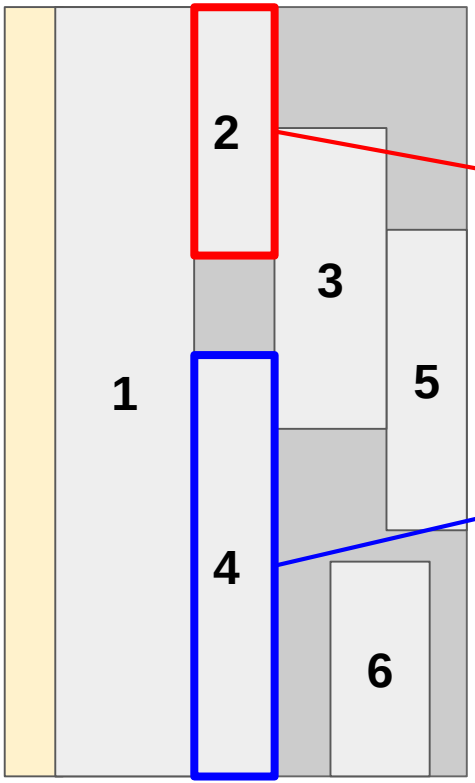
Data Distribution in reality



Center for
Big Data Statistics

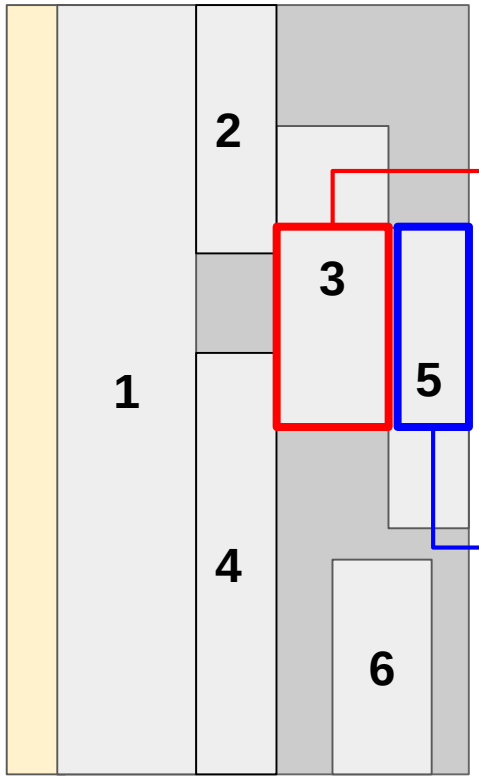
DE
MAASTRICHT
STUDIE

Horizontally Partitioned Data



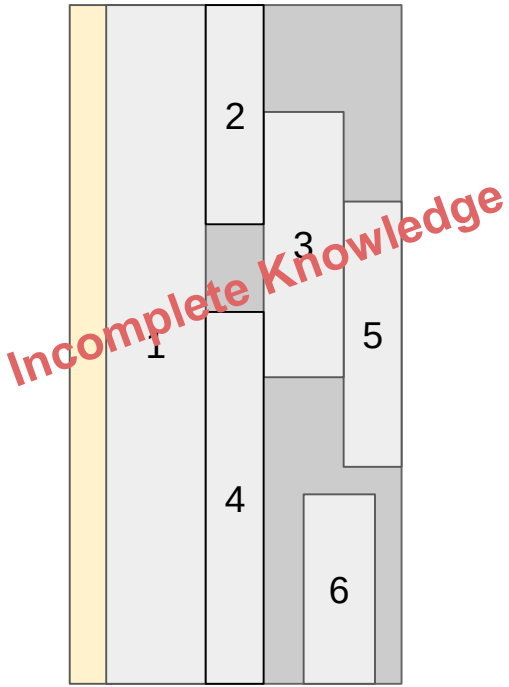
ID	Age	Gender	Education	Wellbeing	Diabetes
1	56	Male	University	Good	YES
2	25	Female	University	Medium	NO
3	31	Female	High School	Good	NO
4	45	Male	Primary school	Poor	YES
5	32	Male	Primary school	Good	NO
6	60	Female	High School	Poor	Yes
7	55	Male	University	Medium	NO

Vertically Partitioned Data

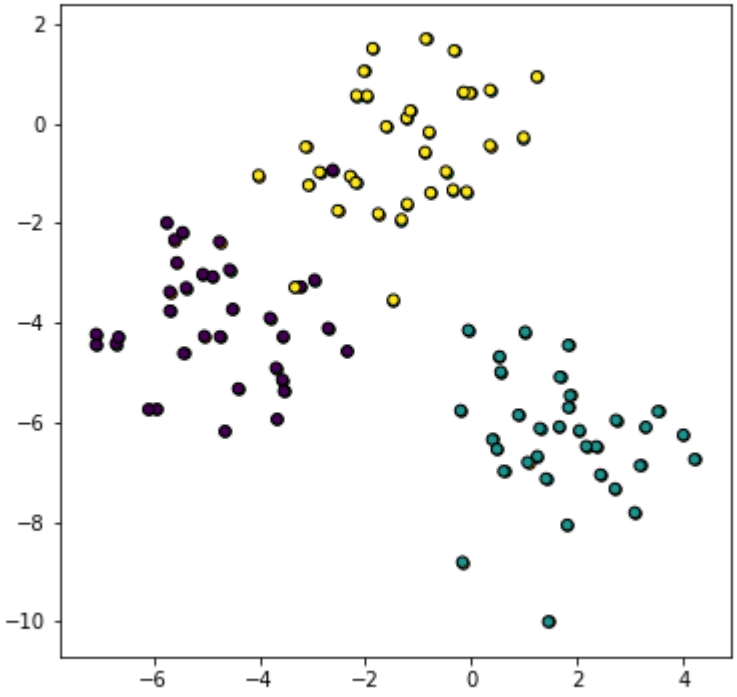


ID	Age	Gender	Education	Wellbeing	Diabetes
1	56	Male	University	Good	YES
2	25	Female	University	Medium	NO
3	31	Female	High School	Good	NO
4	45	Male	Primary school	Poor	YES
5	32	Male	Primary school	Good	No
6	60	Female	High School	Poor	Yes
7	55	Male	University	Medium	NO

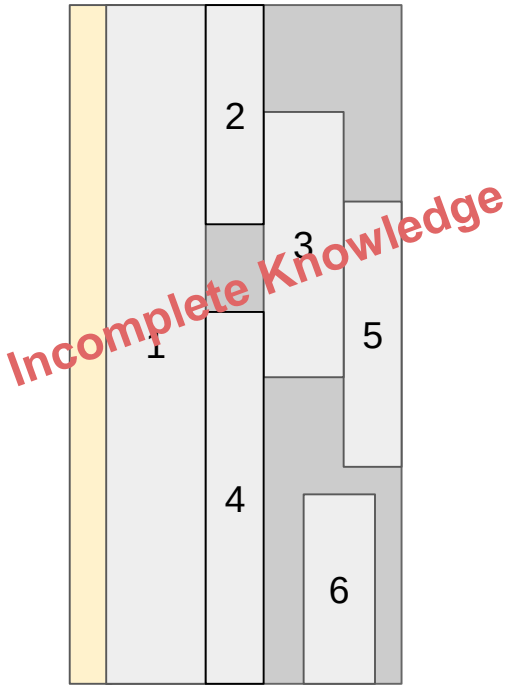
Incomplete knowledge learnt from partitioned data



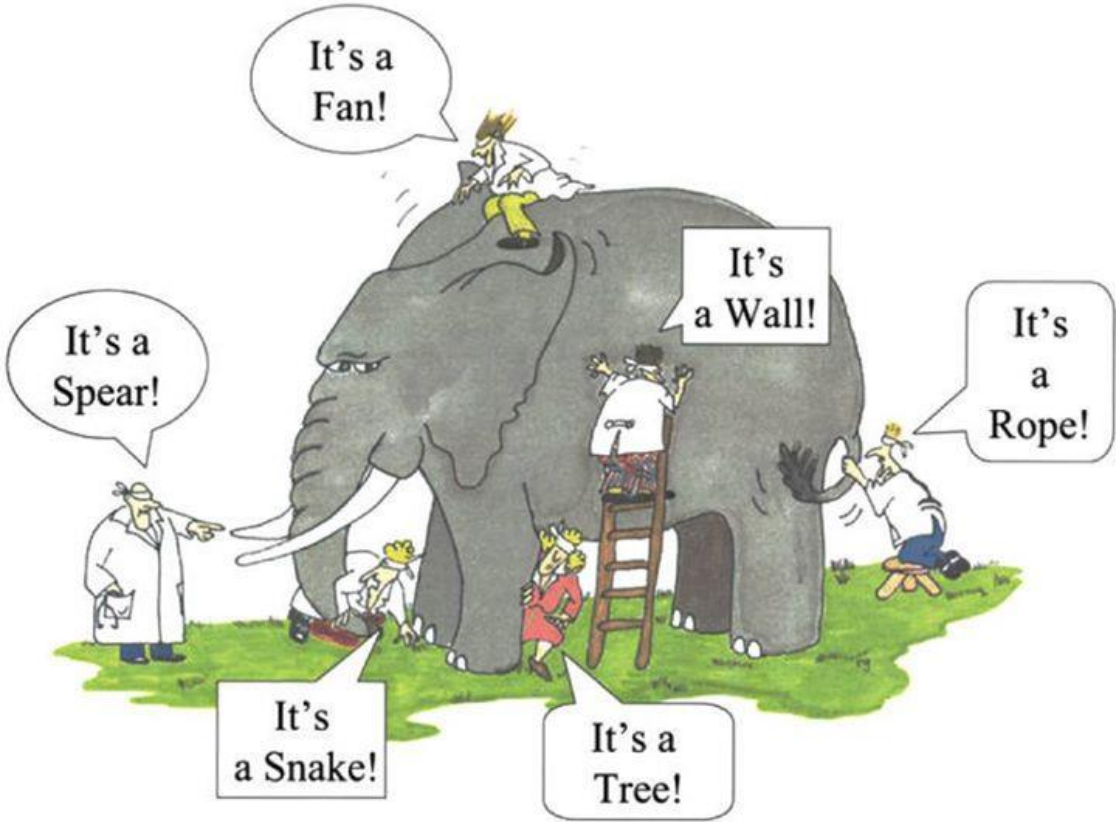
Horizontally:



Incomplete knowledge learnt from partitioned data



Incomplete Knowledge



Challenges of privacy-preserving machine Learning on distributed data

Legal

- Compliance to data governance and privacy laws
- Limitation of liability

Ethical

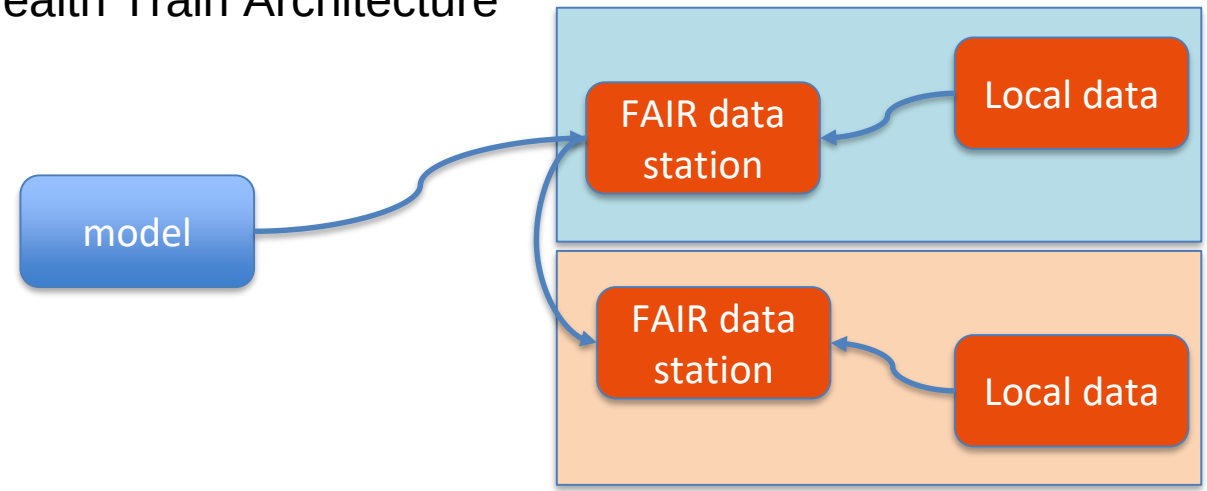
- Should organisations be able to gain access to any data?
- Should all stakeholders be aware of how data are used?

Technological

- How do we securely exchange information(data) and communicate among multiple sources?
- How do we guarantee the applied machine learning models are privacy-preserving?
- How do we monitor how data is being used?

Individual's privacy issues demand interdisciplinary solutions

Personal Health Train Architecture



F.A.I.R Principles - Summarized

Findable

- Globally unique, resolvable, and persistent **identifiers**
- Machine-readable **descriptions** to support structured search

Accessible

- Clearly defined **access and security protocols**
- **Metadata** is always accessible beyond the lifetime of the digital resource

Interoperable

- Extensible machine interpretable formats for data + metadata
- **Vocabularies** themselves must be FAIR
- **Linked** to other resources

Reusable

- **Provide** licensing, provenance, and use community-standards

Use case: Collaborative project between CBS and UM

We aim

1. To develop an innovative infrastructure, which supports secure analysis of personal health data from multiple providers with different governance policies in a privacy-preserving manner.

2. To use this infrastructure to predict differences in health care cost based on the Type 2 Diabetes Mellitus status, lifestyle, and mental health.

What we mean by privacy-preserving

There are many notions of privacy...

For this study, we define privacy-preserving as
i) a **restriction on which data elements** are available for analysis
ii) that **no data provider can directly access** data from another data provider
iii) the analysis may **only return results of analysis** (and not original data)

Vertically partitioned data hosted by UM and CBS

ID	DoB	Gender	HbA1C
1234	…	…	…
1235	…	…	…
1236	…	…	…

The Maastricht Study

Population study > 10.000 participants

Interest in Diabetes Mellitus Type 2

ID	DoB	Income	Edu	…
1234	…	…	…	
1235	…	…	…	
1236	…	…	…	
1237	…	…	…	
1238	…	…	…	
1239	…	…	…	

CBS

> 17 million citizens

All kinds of variables

Center for Big Data Statistics

Maastricht University

Institute of Data Science

DE MAASTRICHT STUDIE

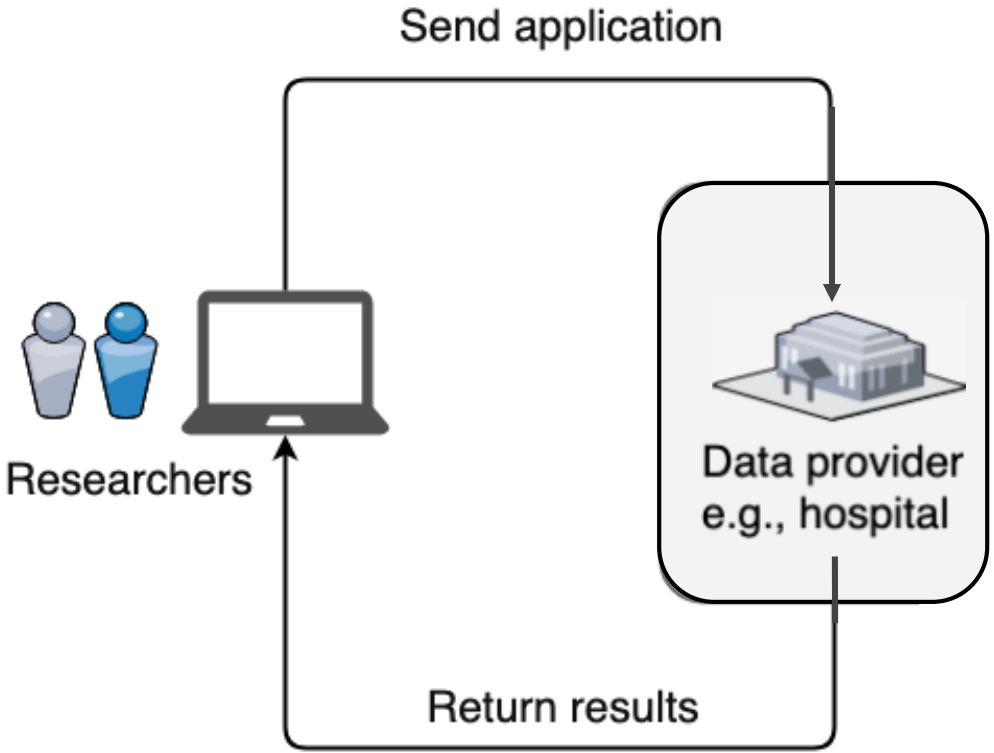
Research questions from multi-disciplines

Health: What is the association between diabetes and health care costs differs according to lifestyle and habits?

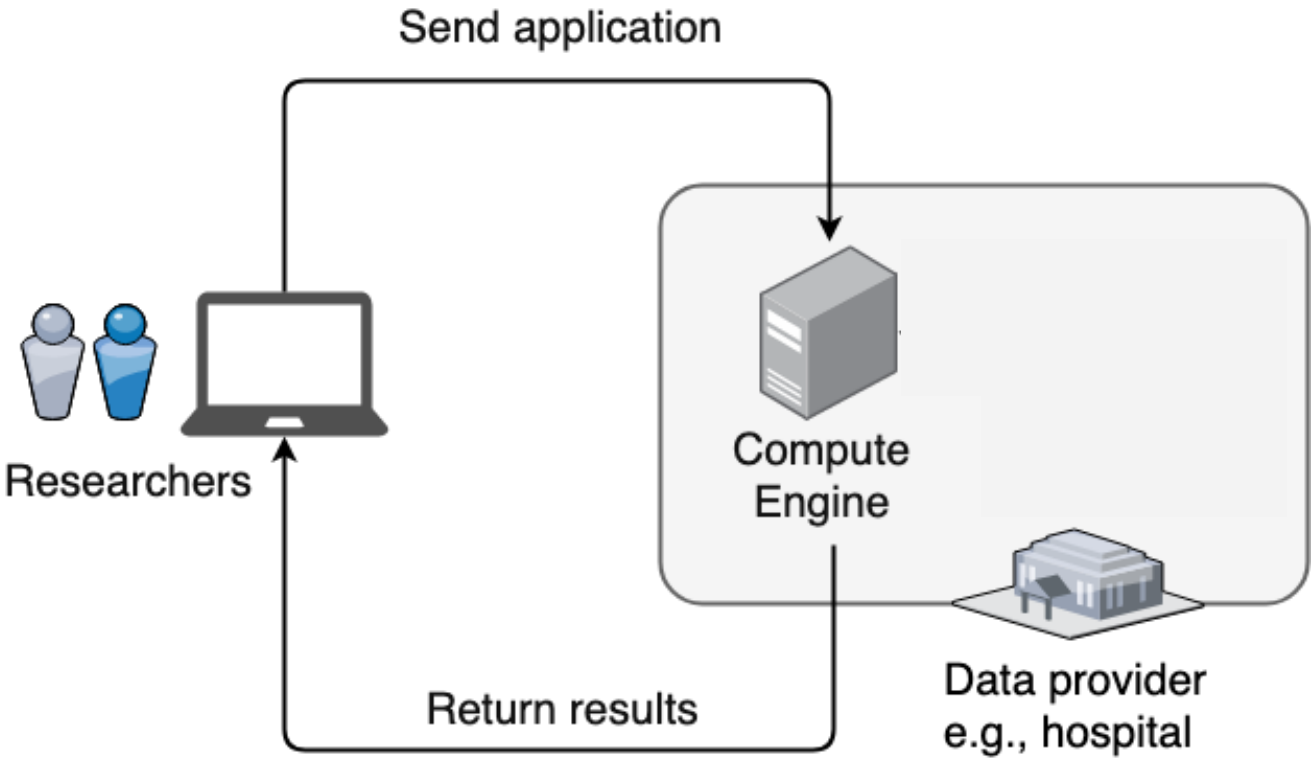
Technical: Can we use distributed machine learning to discover high confidence associations from vertically partitioned data located at multiple sites in a privacy-preserving manner?

Social/Legal/Ethical: Does our prior consent enable us to use these data for this experiment? Can we devise a consent framework that satisfies the GDPR and other (institutional/ethics/social) requirements?

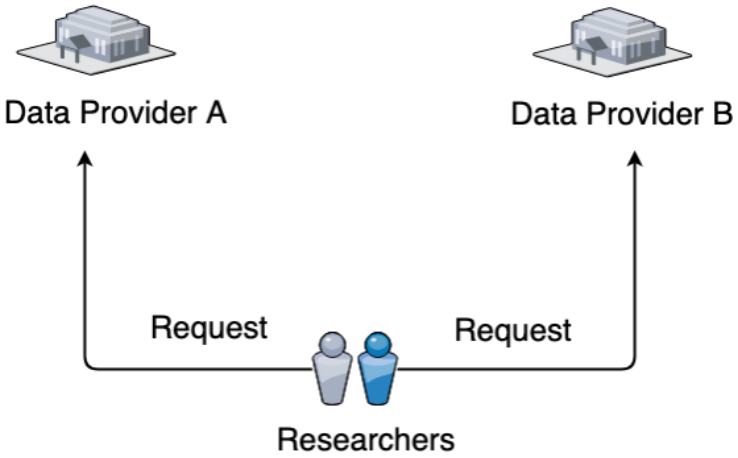
Current machine learning models using real data



What if we send the analysis to the data?

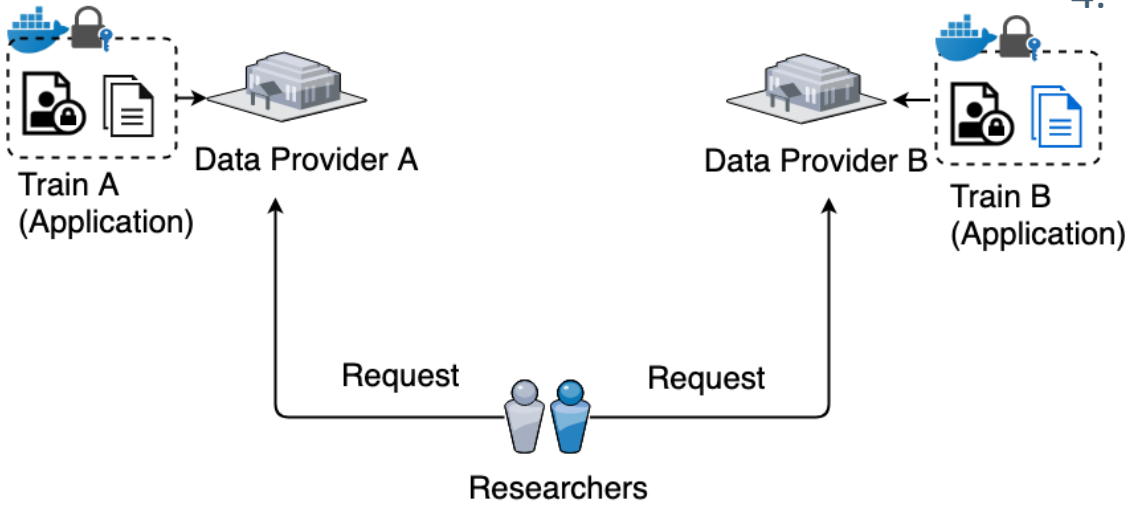


Researchers request access to the data

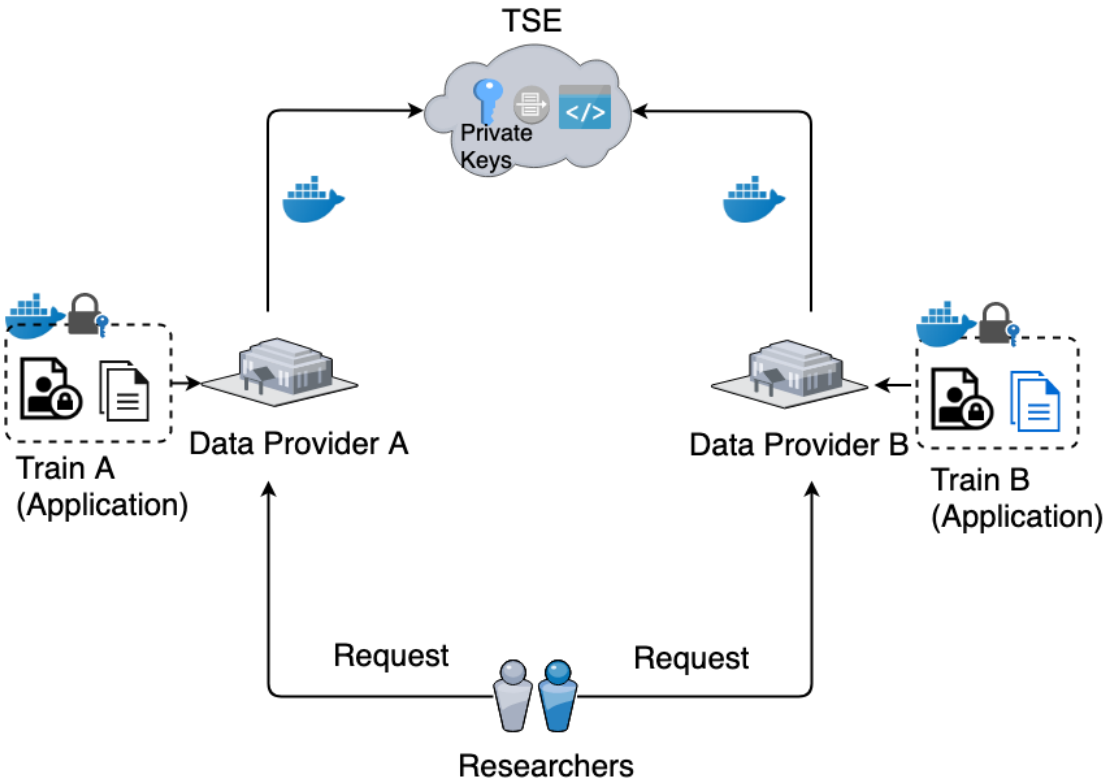


Data parties approve the request and prepare data

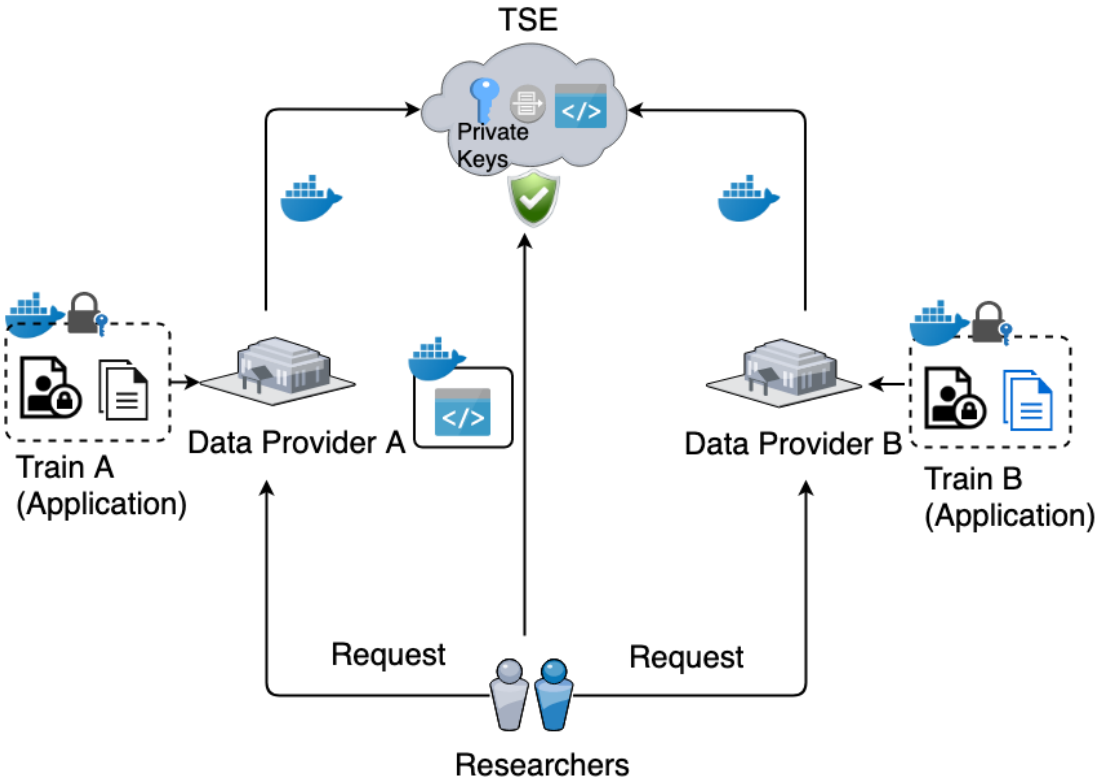
1. Legal conditions
2. FAIR Principles
3. Pseudonymization
4. Encryption



Encrypted data are securely sent to trusted secure environment (TSE)



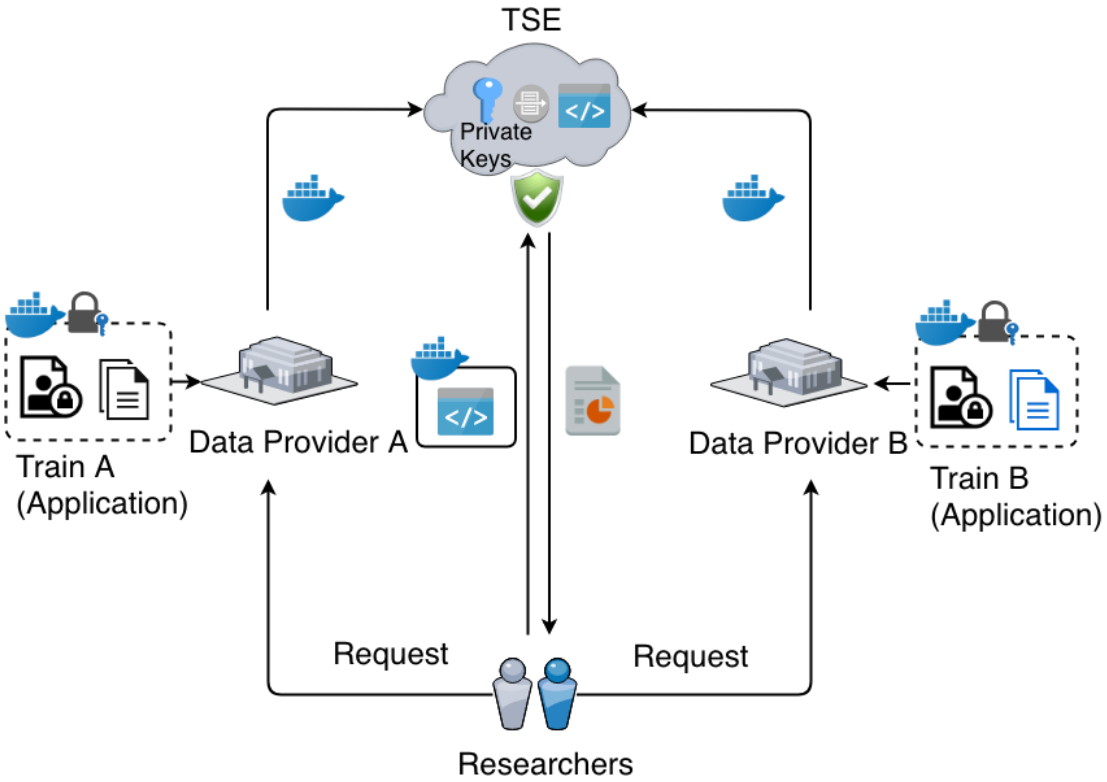
Application containing analysis triggers execution at TSE



Application contains machine learning pipeline

- Record Linkage
- Data description
- Normalization
- Feature selection
- Outlier detection
- Classification/Regression
- Cross-validation

Only agreed on results will be sent back



Linking(matching) features

Only certain situations allow for the use of the BSN. Unfortunately, we cannot use BSN in our case.

Therefore, we use:

- Date of birth (year, month, day)
- Gender
- Zip code
- House number (with/without suffix)

* Both parties formalized their linking features before matching two datasets

Linking(matching) methods - Hashing & encryption

DoB	Gender	Postal Code	H#	suf
15-11-1964	Male	6229 ET	12	1b
29-03-1973	Female	6221 HW	64	3-c
02-02-1972	Female	6213 NN	37	a

Merge columns

ID
15111964_M_6229ET_12_1b
29031973_F_6221HW_64_3-c
02021972_F_6213NN_37_a

One-way hash values

Encrypt & send to TSE

ID	Care act.	cost
3A79810E3A7ABA7E54646925D6057EC80F2D97AE17F1EBEF14B9DEE2D3D34DF6	…	…
6A23D4BA920B6262C8BA23016E4225358A1BA8E127537EA47210F4176EB8EF4F	…	…
8435B24CD6F90A9CD27C0BDEA1299316F7BE20F9EF087B90B61CB5081E13E598	…	…

Linking(matching) results - Accuracy

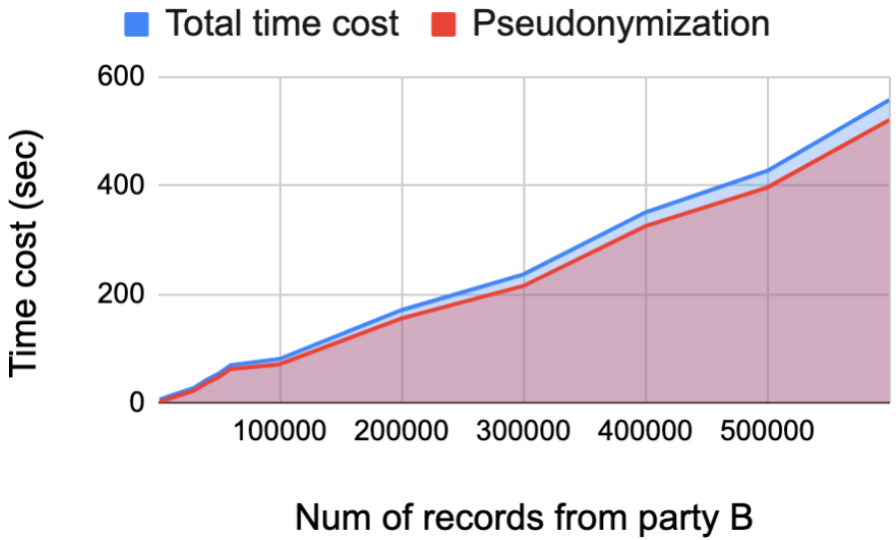
! Two participants withdraw their consents during our experiments

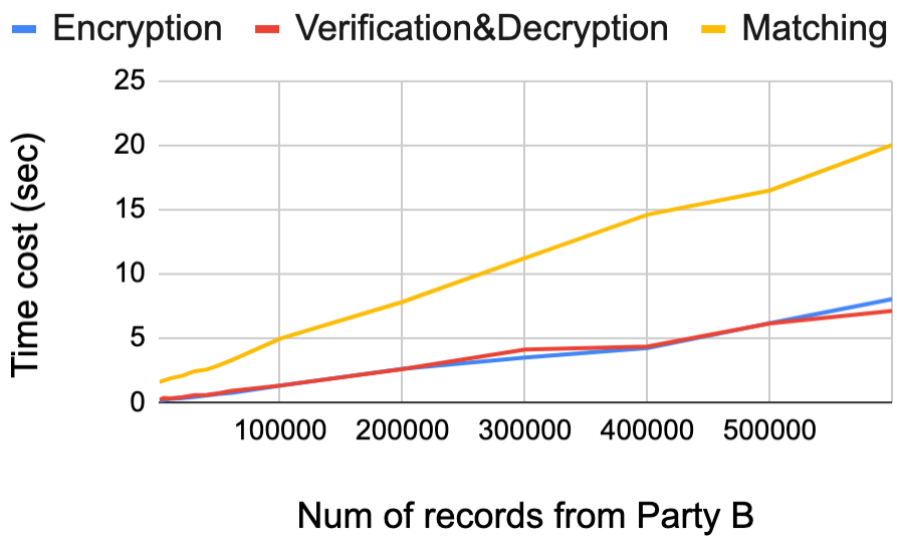
CBS	DMS	Results		
Num of Records = ~1 million	Num of records = **3283**	Unique match	Multiple match	No match
without using house number suffix		**3145**	17	121
using house number suffix		**3142**	0	141

Accuracy (unique match): **97%**

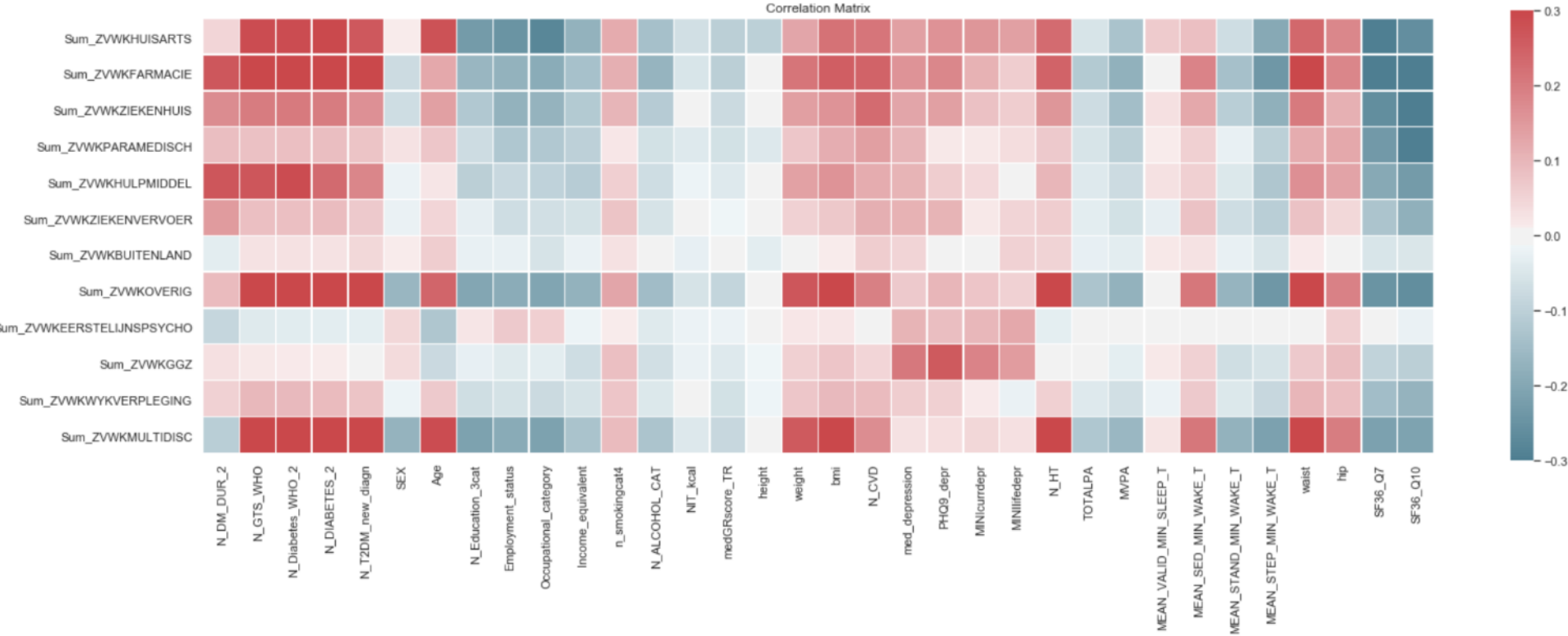
Linking(matching) results - efficiency

Time (second)					
	Num of rows	Pseudonymization	Encryption	Verification & decryption	Matching
CBS	~1 million	1838.749	47.459	17.208	197.267
DMS	3.283	9.964	0.179		





First analysis results - Correlation Matrix



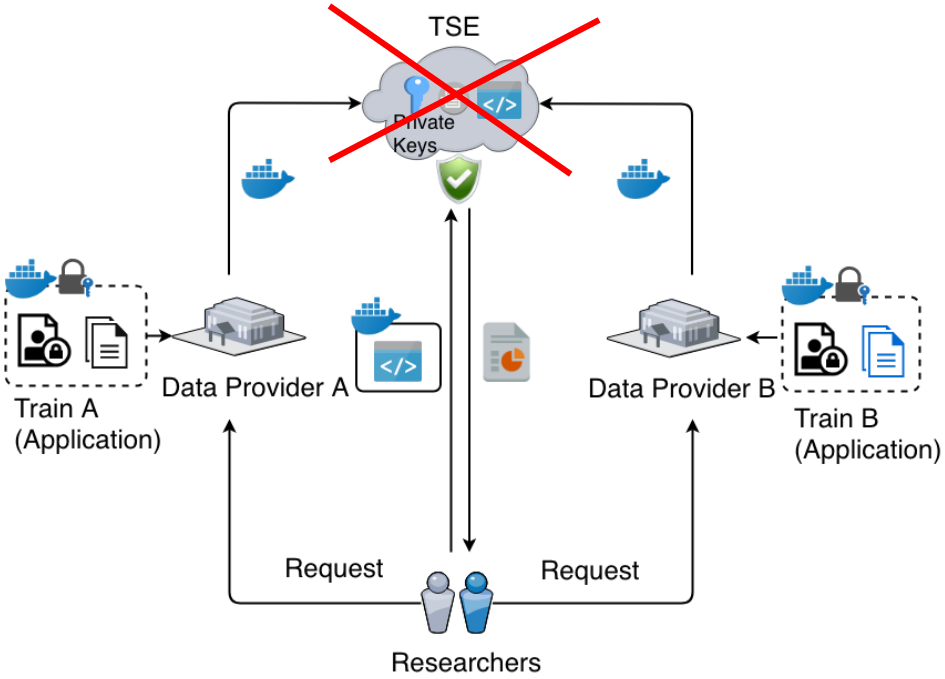
Center for Big Data Statistics

Maastricht University

Institute of Data Science

DE MAASTRICHT STUDIE

Next step: Remove TSE



Secure multiparty computation in machine learning models to replace TSE

- Only transfer models
- Communication among organizations needs to be efficient and secure
- Implemented in current infrastructure
- Running in demonstrator
  https://github.com/sunchang0124/PPDML

Publications

Publications:

- *Sun, Chang, et al. “A privacy-preserving infrastructure for analysing personal health data in a vertically partitioned scenario” - accepted by The 17th World Congress of Medical and Health Informatics (2019)*
- *van Kesteren, E.J., Sun, C., Oberski, D.L., Dumontier, M. and Ippel, L., 2019. Privacy-Preserving Generalized Linear Models using Distributed Block Coordinate Descent. arXiv preprint arXiv:1911.03183.*
- *van Soest, J., et al. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." Studies in health technology and informatics 247 (2018): 581.*
- *Sun, Chang, et al. "Analyzing Partitioned FAIR Health Data Responsibly." arXiv preprint arXiv:1812.00991 (2018).*

Public repository:

- Infrastructure with TSE [Repository: https://gitlab.com/CBDS/DataSharing]
- Trains/containers [Repository: https://github.com/sunchang0124/PPDML]
- Video demonstrator [https://www.youtube.com/watch?v=dagXZevwPV8&t=11s]

Wrap up

- Data Science for **Better** Decision:
  - Unlocking new possibilities by combining different data sources
  - Preserving privacy is essential, but difficult to prove
  - Technical, legal, and ethical issues require multidisciplinary teams

UM/Institute of Data Science


Michel Dumontier
Principal Investigator


Claudia van Oppen
Project manager


Chang Sun
PhD Student


Alexander Malic
Data Science Architect


Lianne Ippel
Postdoc


Seun Adekunle
Data scientist

UM/Maastro Clinic


Andre Dekker
Co-Investigator


Johan van Soest
PhD/PostDoc

Health, Ethics and Society


David Townend
Co-Investigator


Birgit Wouters
PhD Student

Maastricht Study


Annemarie Koster
Co-Investigator


Carla van der Kallen
Researcher

CB(D)S


Magchiel van Meeteren

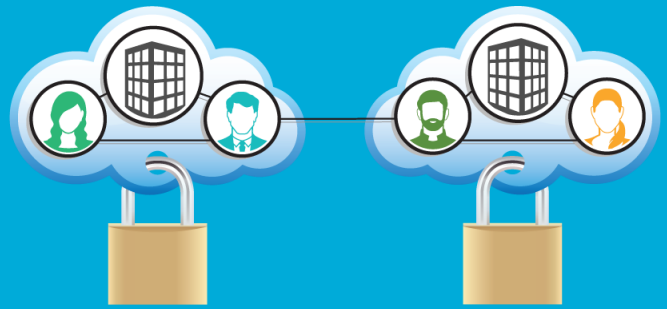

Bob van den Berg
CBS Project manager


Ole Mussmann
CBS Consultant


Susan van Dijk
CBS Researcher

Thank you for your attention!

Any questions?



Maastricht University

DE MAASTRICHT STUDIE

Center for Big Data Statistics