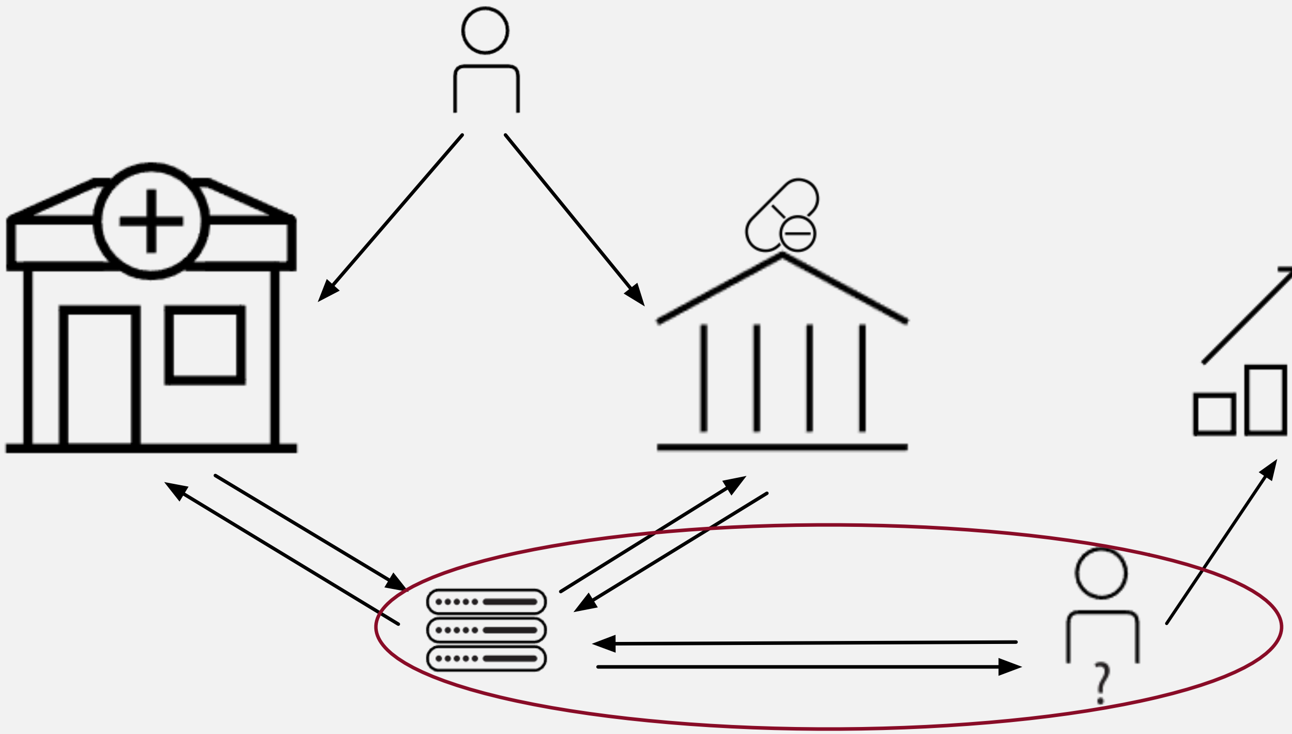
# Is Usable Privacy-Preserving Data Analysis an Oxymoron?

Liina Kamm

CYBERNETICA

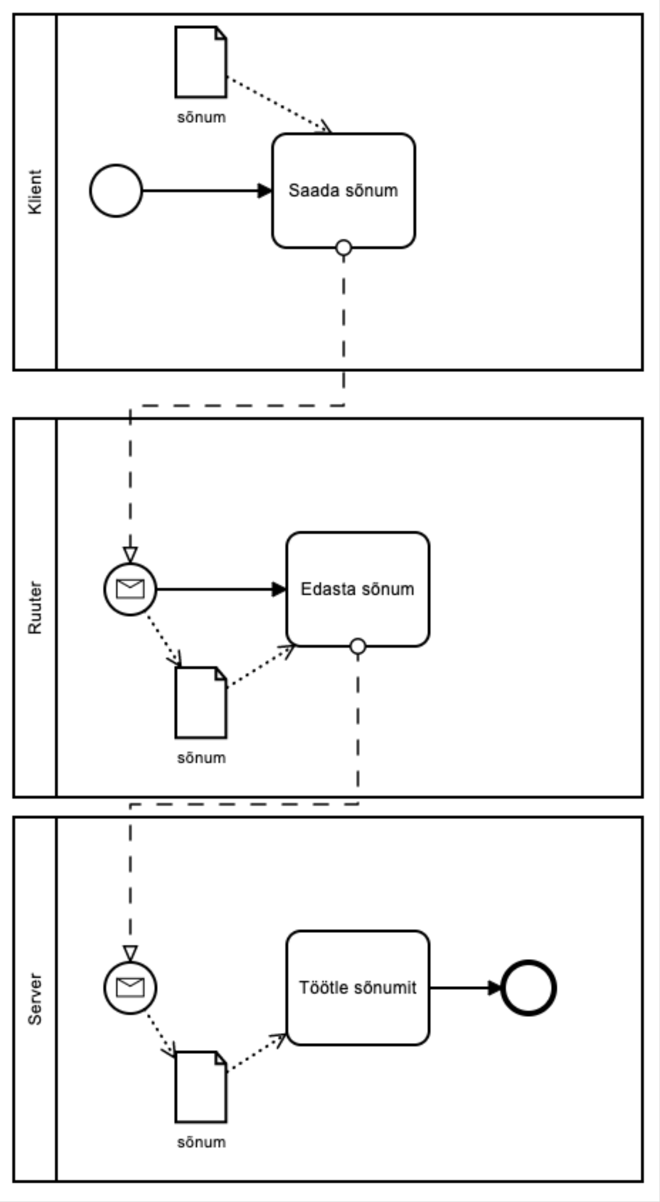
# Processing personal data

| ID | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| 49401223576 | Valli Vaarikas | 27 | F | 194 | 90 | 135 | 146 | 95 |

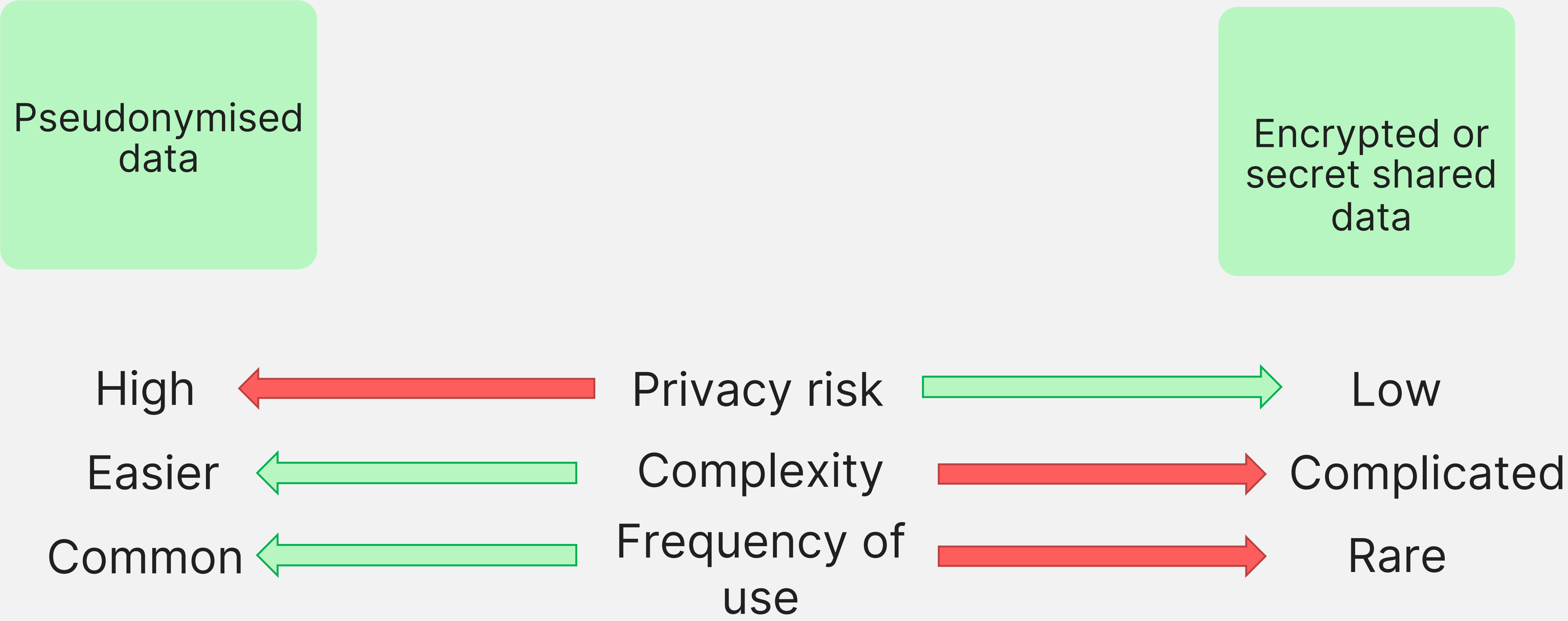


# Processing data

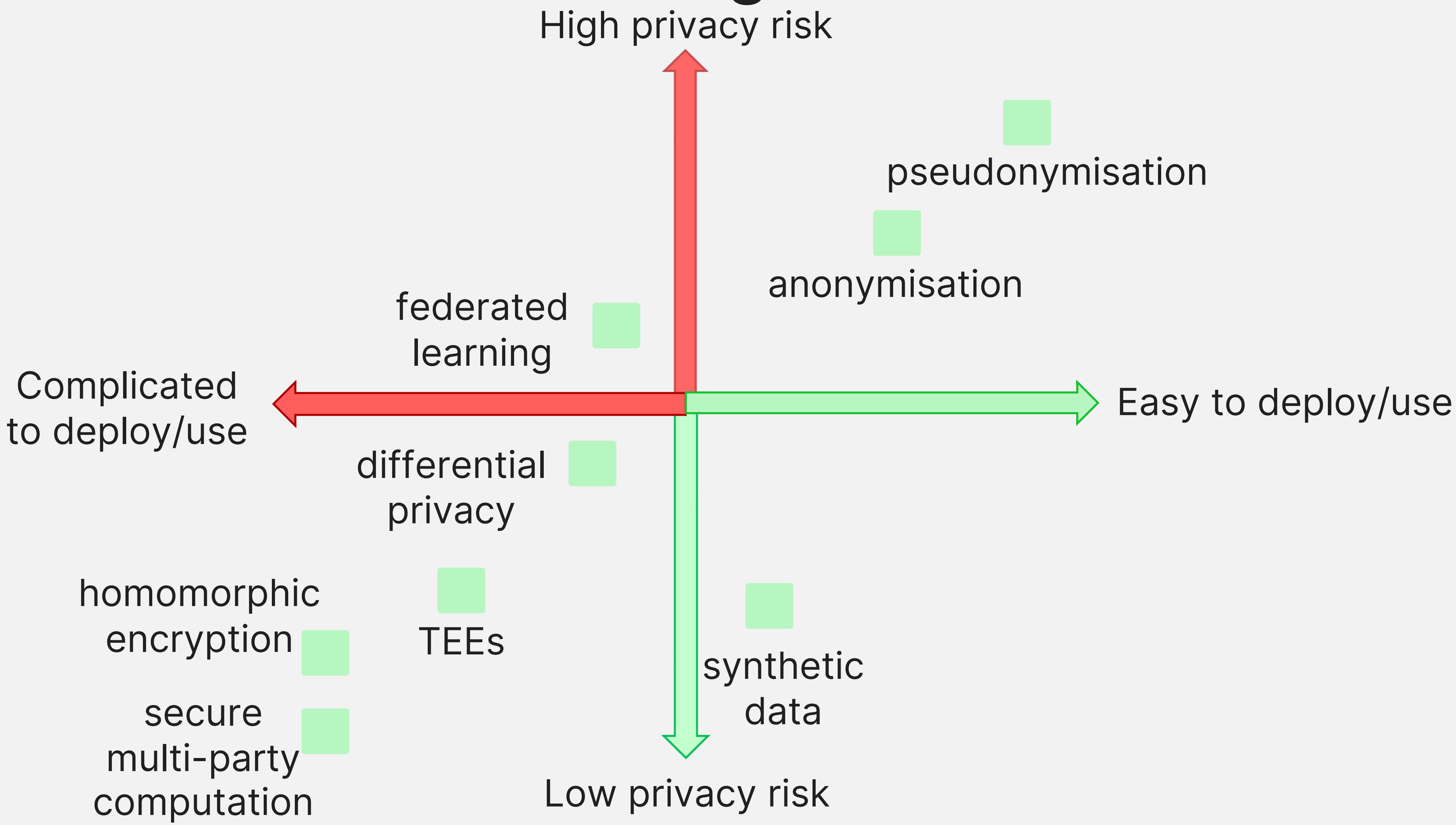
- What kind of data do I have?
- Where are they kept?
- What do I want to do with them?
- Where will they move in the process?
- Who is going to see them?



# Privacy-complexity trade-off



# Comparison of the technologies



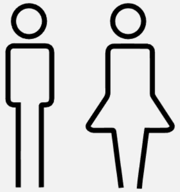

*Not all of these technologies are useful or relevant for all use-cases

# Pseudonymisation

**Pseudonymisation** is the **processing of personal data** in such a manner that the personal data **can no longer be attributed to a specific data subject without the use of additional information**, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

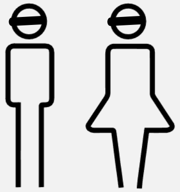
(GDPR, art. 4(5))

# Pseudonymisation



| ID | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| 49403136525 | Mari Maasikas | 27 | F | 165 | 75 | 110 | 122 | 75 |
| 39507082997 | Ants Õun | 25 | M | 165 | 85 | 162 | 132 | 88 |
| 49401223576 | Valli Vaarikas | 27 | F | 194 | 90 | 135 | 146 | 95 |
| ... | ... | | ... | ... | ... | ... | ... | ... |

↓ Pseudonymisation



| code | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| HYP_01 | | 27 | F | 165 | 75 | 110 | 122 | 75 |
| HYP_02 | | 25 | M | 165 | 85 | 162 | 132 | 88 |
| HYP_03 | | 27 | F | 194 | 90 | 135 | 146 | 95 |
| ... | | | ... | ... | ... | ... | ... | ... |

# Anonymisation

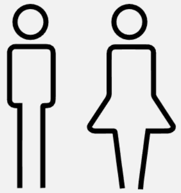
**Anonymisation** is a process by which **personal data is irreversibly altered** in such a way that a **data subject can no longer be identified directly or indirectly**, either by the data controller alone or in collaboration with any other party.

(ISO/TS 25237:2017)

# Anonymisation process

- In general it is not enough to simply remove an individual's identifiers
- **Quasi-identifiers** – combinations of attributes relating to an individual
- The process of anonymisation is final and it should not be possible to reverse this
- Whether an individual data item can be considered anonymous or not requires case-by-case evaluation

# Anonymisation



| ID | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| 49403136525 | Mari Maasikas | 27 | F | 165 | 65 | 110 | 122 | 75 |
| 39507082997 | Ants Õun | 25 | M | 165 | 85 | 162 | 132 | 88 |
| 49401223576 | Valli Vaarikas | 27 | F | 194 | 90 | 135 | 146 | 95 |
| ... | ... | | ... | ... | ... | ... | ... | ... |

Anonymisation

Anonymisation

Anonymisation

| age | sex | cm | kg | Hgb |
|---|---|---|---|---|
| 27 | F | 165 | 75 | 110 |
| 25 | M | 165 | 85 | 162 |
| - | - | - | - | - |
| 27 | F | 165 | 75 | 122 |
| 25 | M | 165 | 85 | 183 |
| ... | ... | ... | ... | ... |

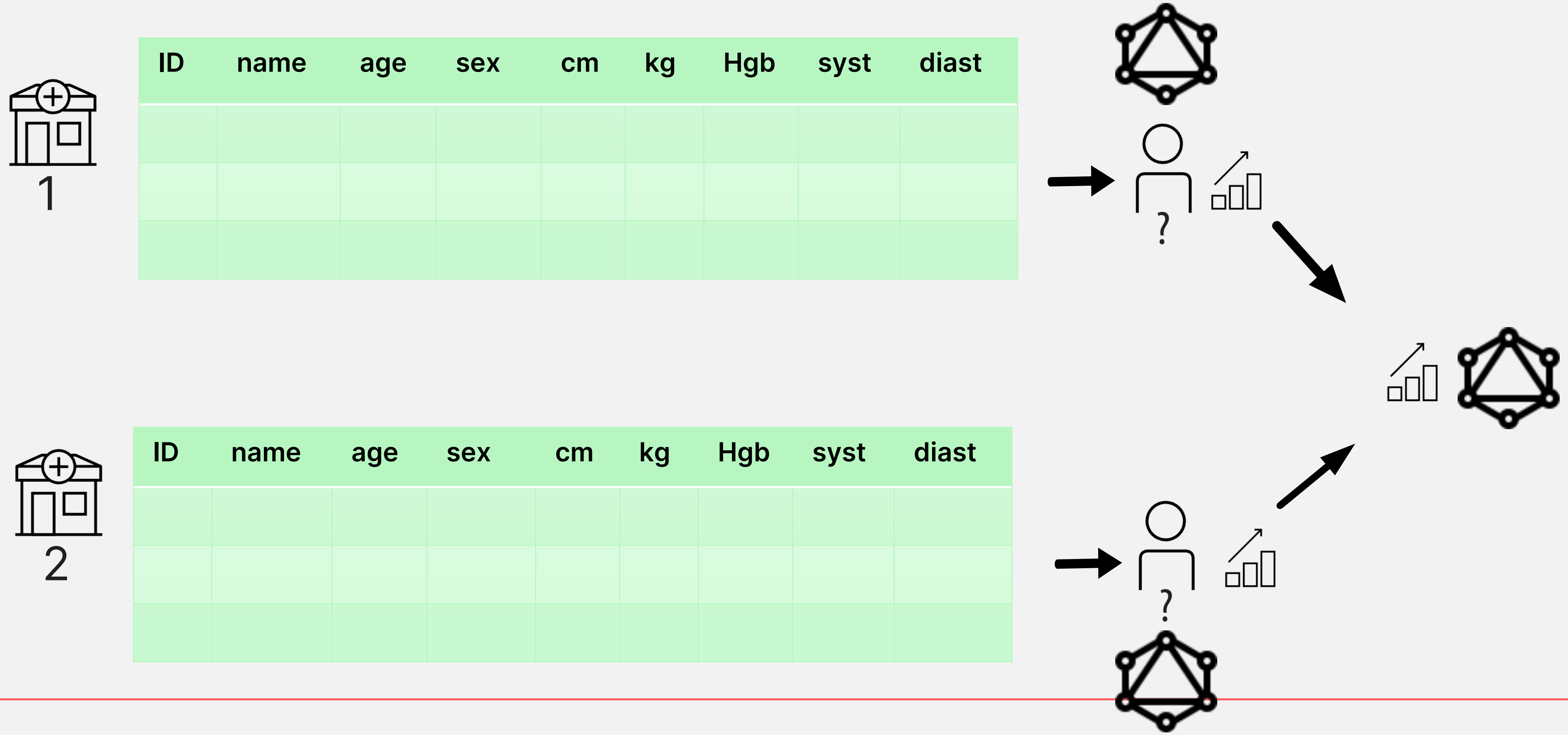
| sex | kg | Hgb | syst | d |
|---|---|---|---|---|
| F | 75 | 110 | 122 | 7 |
| M | 85 | 162 | 132 | 8 |
| F | 90 | 135 | 146 | 9 |
| F | 90 | 135 | 120 | 7 |
| M | 85 | 183 | 142 | 9 |
| F | 75 | 122 | 130 | 8 |
| ... | ... | ... | ... | ... |

| age | sex | BMI | Hgb | syst | diast |
|---|---|---|---|---|---|
| 27 | F | 23,9 | 110 | 122 | 75 |
| 25 | M | 31,2 | 162 | 132 | 88 |
| 27 | F | 23,9 | 135 | 146 | 95 |
| 27 | F | 23,9 | 135 | 120 | 70 |
| 25 | M | 31,2 | 183 | 142 | 99 |
| 27 | F | 23,9 | 122 | 130 | 85 |
| ... | ... | ... | ... | ... | ... |

# Pseudonymisation and anonymisation

- Pseudonymised data are not anonymised data
- Pseudonymisation: existence of an association between personal identifiers and pseudonyms. Re-identification **is** possible, data **is** personal data
- Anonymisation: such an association should not be available by any means, re-identification **is not** possible, data **is not** personal data
- Anonymised data do not qualify as personal data
- “Anonymous” in **common language** also describes cases where the identities of individuals are only hidden

# Federated statistics and federated learning



# Synthetic data generation

Synthetic data has been generated from real data and has the same statistical properties as real data.

Synthetic data is not real data.*

* Depends on the strength of the synthesis algrithm and from the perspective of legislation and data protection, it is not yet binding

Khaled El Emam, Lucy Mosquera, Richard Hoptroff “Practical Synthetic Data Generation”. O’Reilly 2020.

# Data synthesis using real data

| code | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| HYP_01 | | 27 | F | 165 | 75 | 110 | 122 | 75 |
| HYP_02 | | 25 | M | 165 | 85 | 162 | 132 | 88 |
| HYP_03 | | 27 | F | 194 | 90 | 135 | 146 | 95 |
| ... | | ... | ... | ... | ... | ... | ... | ... |

Model training

ML model

Data synthesis

| age | sex | age | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|
| 26 | M | 165 | 80 | 156 | 122 | 75 |
| 27 | F | 165 | 80 | 110 | 133 | 88 |
| 28 | F | 195 | 95 | 132 | 142 | 92 |
| | ... | ... | ... | ... | ... | ... |

# Trusted execution environments

In the central processing unit (CPU), trusted execution environments are secure subprocesses (enclaves), into which other processes cannot see

Intel Software Guard Extensions (SGX), ARM TrustZone, iPhone Secure Element

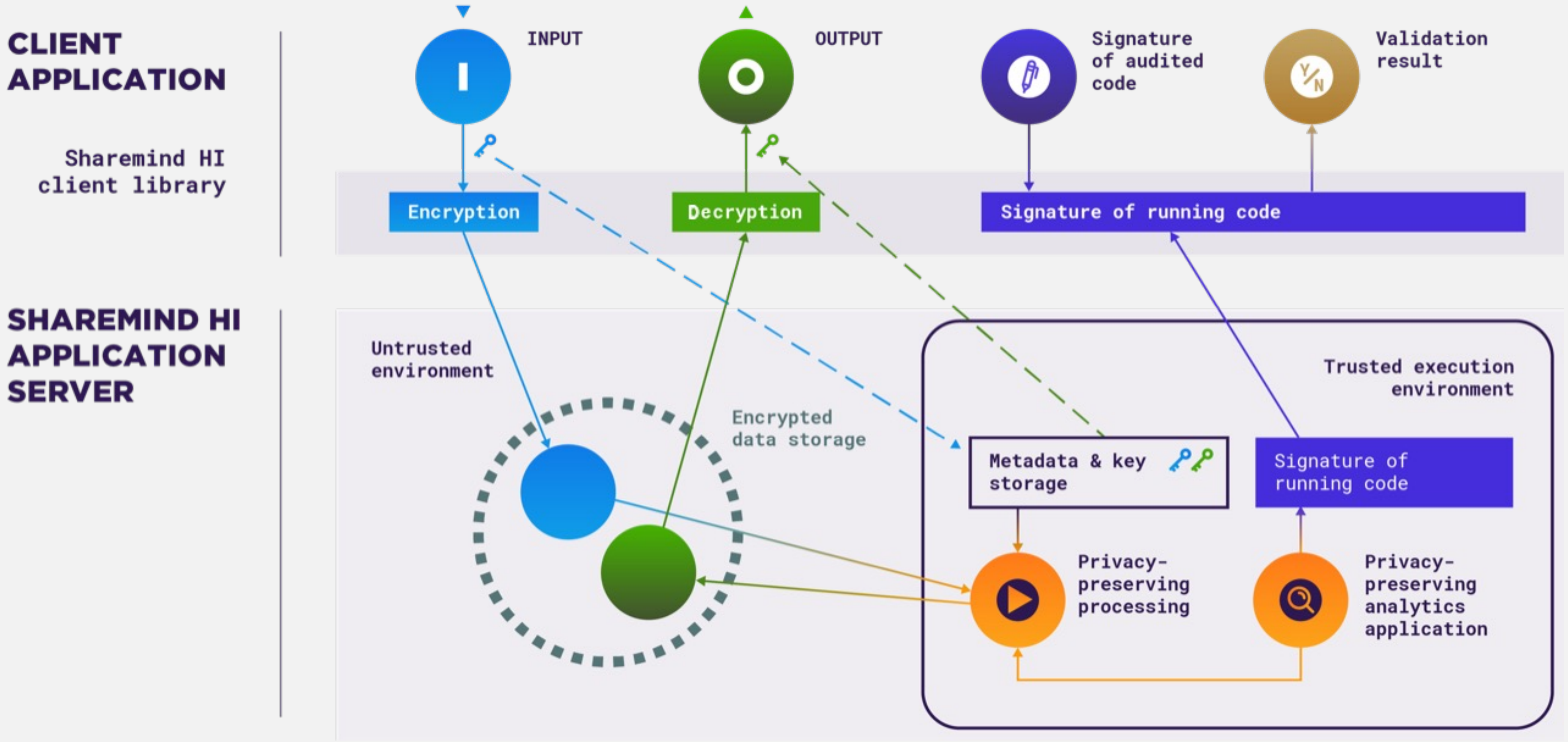
# Trusted execution environments

| ID | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| 49403136525 | Mari Maasikas | 27 | F | 165 | 75 | 110 | 122 | 75 |
| 39507082997 | Ants Õun | 25 | M | 165 | 85 | 162 | 132 | 88 |
| 49401223576 | Valli Vaarikas | 27 | F | 194 | 90 | 135 | 146 | 95 |
| ... | ... | | ... | ... | ... | ... | ... | ... |

Encryption

| ID | name | age | sex | cm | kg | Hgb | syst | diast |
|---|---|---|---|---|---|---|---|---|
| jablkhsjhbfd | lfknlekfnasyd | kjsfb | sdfkj | dbid | scbs | wpal | pfbm | anbf |
| djsbfkhbfsih | adphfguydla | dkjb | dkjb | sdnh | sklk | qlsk | qkdn | Stys |
| fljkshasdans | dlfkjbnabdua | dlfja | dkbd | nuyg | snby | qpkd | aknd | fhva |
| ... | ... | | ... | ... | ... | ... | ... | ... |

# Example of a TEE



# Secure multi-party computation

Several **independent** parties compute a function based on all input data, without knowing the input values of other parties. The output is revealed only to authorised parties.

# Secure multi-party computation

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 49401223576 | Valli Vaarikas | 27 | F | 194 | 90 | 135 | 120 | 80 |
| 1 | HYP_1 | | 3 | X | 76 | 15 | 122 | 10 | 12 |
| 2 | HYP_1 | | 17 | X | 76 | 10 | 10 | 100 | 8 |
| 3 | HYP_1 | | 7 | X | 42 | 65 | 3 | 10 | 60 |

# Secure multi-party computation and ML

- Possible to link databases
- Sufficient privacy protection measures
- Preserves privacy of individuals

- High computational overhead for certain methods
    - But what would the administrative overhead be?

# Secure multi-party ML in action

- Sharemind MPC has side-channel safe algorithms from linear regression to XGBoost
- User-friendly R-like interface (Rmind)
- Manageable overhead (except for XGBoost on large datasets)
- LASSO logistic regression shows promise
- Demonstrated for neural network evaluation
- Simple MPC is not feasible for neural network training
  - Could be done in a federated manner

# Case study: using Sharemind MPC for ML

Predicting the hospitalisation of chronically ill patients

More frequent checkups could reduce the number of hospitalisations

Medical data of 130k individuals (age, gender, clinical observations, procedures, measurements, doctor visits, prescription info); up to 25M entries in one table.

# Preprocessing and model training

**Preprocessing 1:** Select people with chronical illnesses (e.g., diabetes, hypertonia), people with cancer diagnoses are excluded.

**Preprocessing 2:** Link and clean data. Outputs a single table with around 150k rows and 500 columns.

**Normalising and splitting data:** Normalise data and split into training and test set.

**Training:** Train logistic regression and LASSO logistic regression models on the data

# Results and benchmarks

For training, we used different model algorithms and hyperparameters.
We experimented with floating-point and fixed-point numbers.

Preprocessing: ~80 hours
Best training (LASSO logistic regression): ~22 hours

Results (AUC):

| Sharemind MPC (LASSO Logistic Regression) | Without Sharemind MPC (XGBoost) | Without Sharemind MPC (LASSO logistic regression) |
|---|---|---|
| 0.709 | 0.727 | 0.687 |

# Thank you!

liina.kamm@cyber.ee

cybernetica

CyberneticaAS

cybernetica_ee

Cybernetica