Privacy-Preserving Techniques for Scanner and Mobile Phone Data Analysis

Liina Kamm, PhD
@liinakamm
Information Security Research Institute

About Cybernetica

- **Estonian ICT company**, founded in 1997
  - Successor of the Institute of Cybernetics of Estonian Academy of Sciences
- We develop and sell **mission-critical** e-government, information security, radio communications and surveillance **software products and systems**
- Our goal is to inspire **new areas of advancement through** interactions between **research and development**
- As of 2017 **we are 140 people** and **10% with PhD degrees**

Privacy Techniques Useful in Statistics

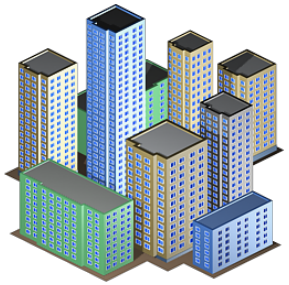
- **Computations on Encrypted Data** uses cryptography to make re-identification of data subjects infeasible and reduces the risks of insider attacks without reducing the accuracy of results
  - **Example technologies:** Secure Multi-party Computation, Homomorphic Encryption, Trusted Execution Environments
- **Anonymisation** adds noise makes re-identification of data subjects harder, but can also reduce the accuracy of results
  - **Example technologies:** Differential Privacy, k-anonymisation

Where to Use Privacy Technologies in Statistics

Input parties

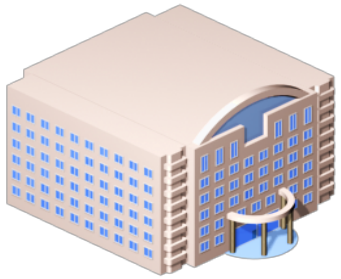
Computing parties

Result parties



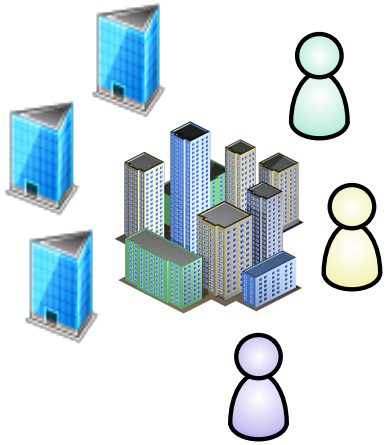
Organisations

Direct data collection (surveys, scraping etc) using Privacy Enhancing Technologies

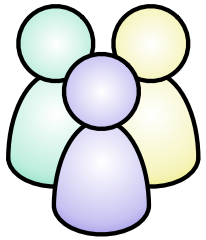


National Statistics Office

Secure, aggregation statistical analysis and result disclosure using Privacy Enhancing Technologies



Community

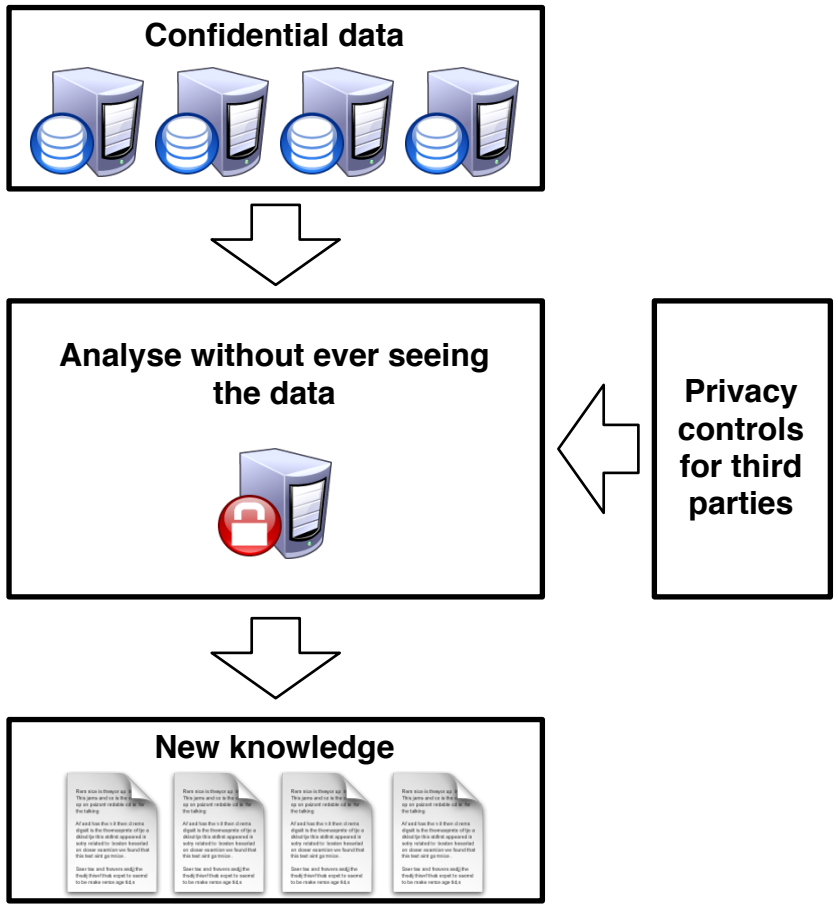


People

Data collection via intermediaries (telco data, payment data, observation data) using Privacy Enhancing Technologies

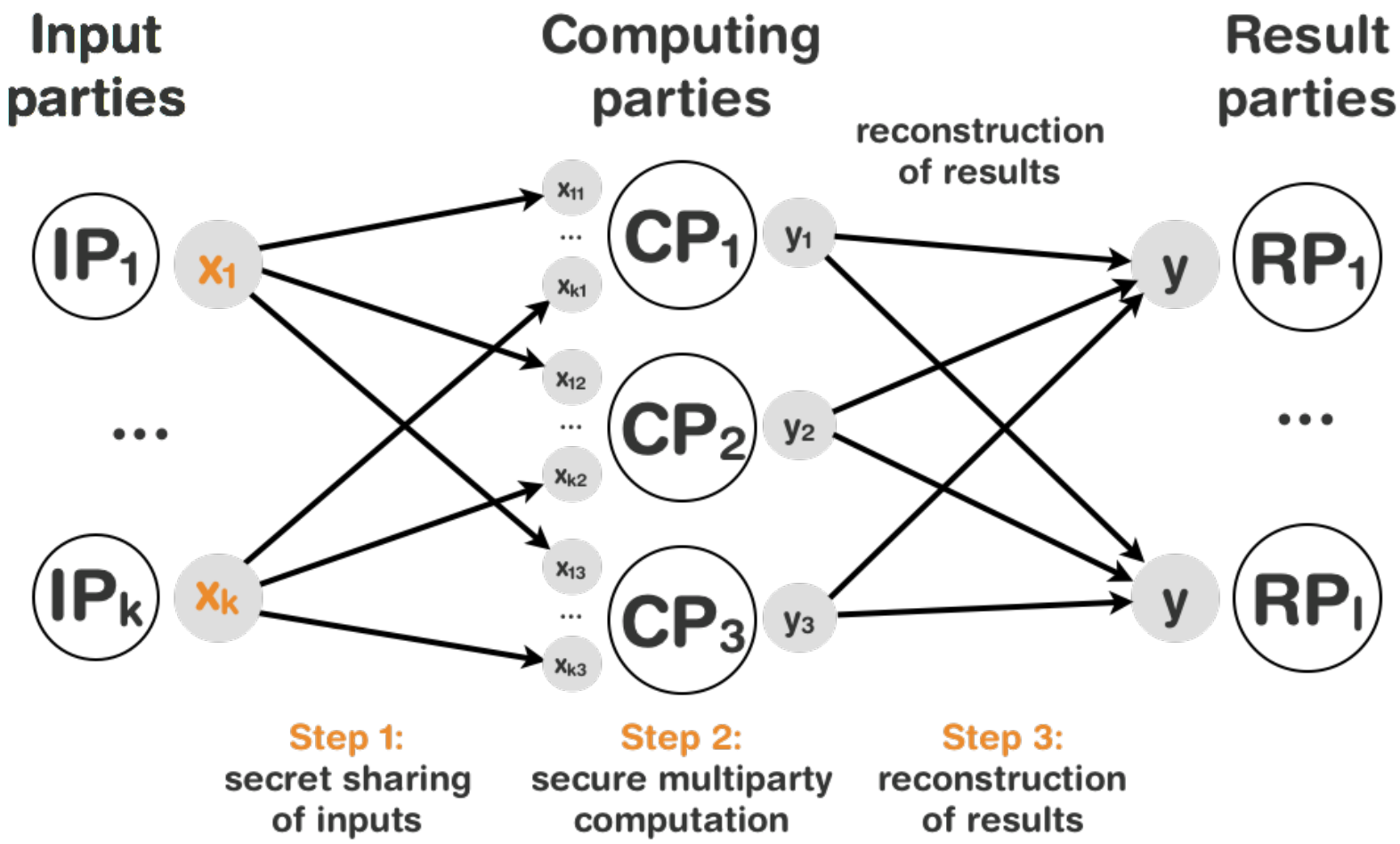
Other Organisations

Secure Multi-party Computation



- Data owners encrypt data on-site and upload it to the server(s).
- Data analysts build and run queries without accessing the data.
- The secure MPC platform processes the queries without removing the protection.
- Authorised users receive query results in an encrypted form which they can then decrypt.

Additive Secret Sharing



Concept of the Sharemind System

BUILD DATA-DRIVEN SERVICES WITH END-TO-END ENCRYPTION

DATA OWNERS

Encrypt data on site and upload to Sharemind

SHAREMIND

Runs queries without removing the protection

AUTHORISED USERS

Receive results in an encrypted form

SHAREMIND

Can give proofs of its activities to third parties

The Sharemind Model Has Two Implementations

	Sharemind MPC	Sharemind HI
Technology	Secure Multi-party Computation	Hardware Isolation using Trusted Execution Environments
Performance	Low to medium performance overhead	Minimal performance overhead
Deployment	Multi-party application server (three servers needed)	Single-node application server (one server with modern CPU)
Usage Model	Analytical tools and SDK available	Tailor-made applications only
Requirements	Deployable in any data centers or private/public clouds	Requires modern servers to run (available on some clouds)

The Rmind Tool

LXTerminal

File Edit Tabs Help

LXTerminal | LXTerminal

```
'citation()' on how to cite R or R packages in public
Type 'demo()' for some demos, 'help()' for on-line he
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> subject <- read.csv ("subject1000.csv", header = TR
> salary <- read.csv ("avg-salaries.csv", header = TR
> edu <- merge (subject, salary)
> age <- edu$age
> sal <- edu$avgSalary100
> plot(age, sal)
>
```

LXTerminal

File Edit Tabs Help

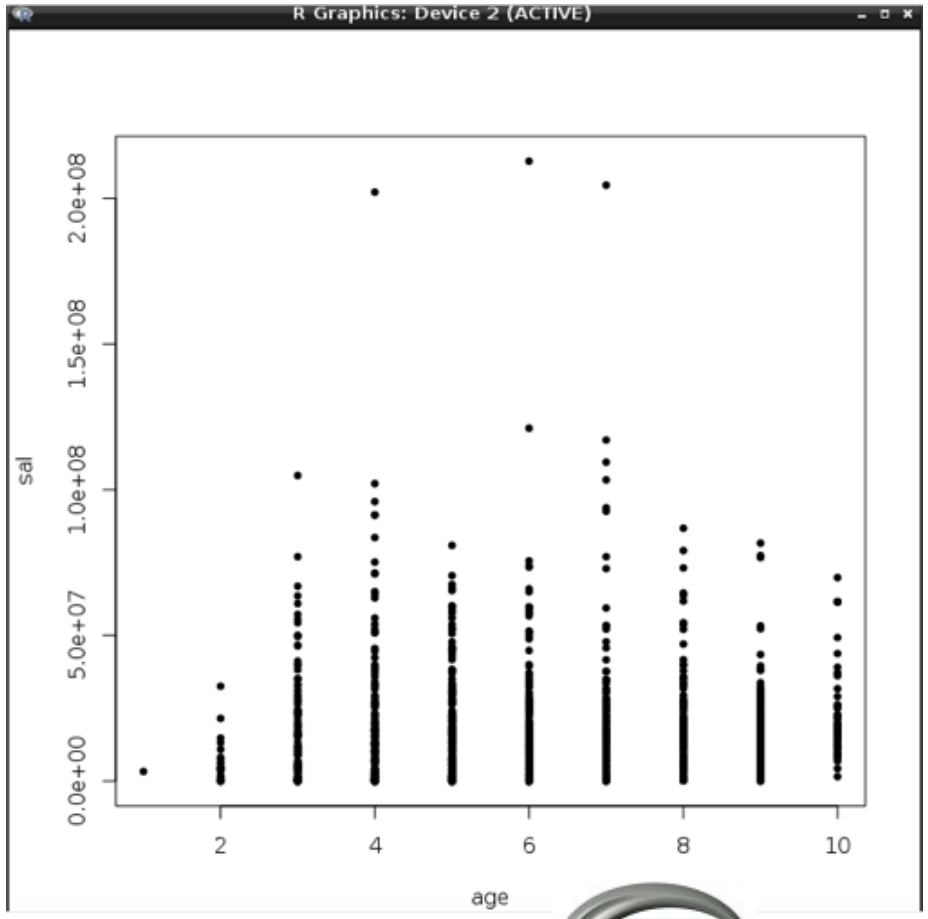
LXTerminal | LXTerminal

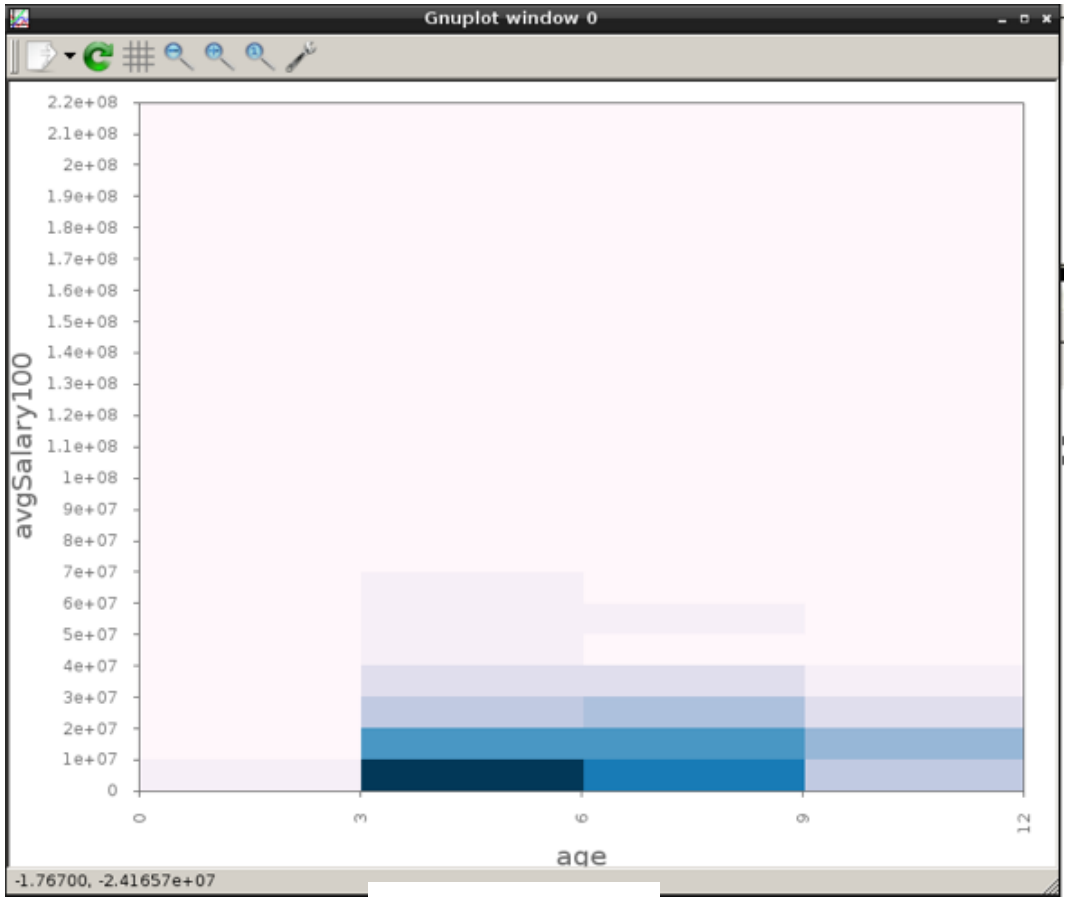
```
[sharemind@sm-build-vm rmind]$ ./rmind
Rmind
Copyright (C) Cybernetica AS
Type 'q()' to quit
Connecting to Sharemind...
Connected
> salary <- load("DS1", "salaries")
> subject <- load("DS1", "subjects")
> edu <- merge(subject, salary)
> age <- edu$age
> sal <- edu$avgSalary100
> heatmap (age, sal)
>
```

Rmind

CYBERNETICA

The Rmind Tool

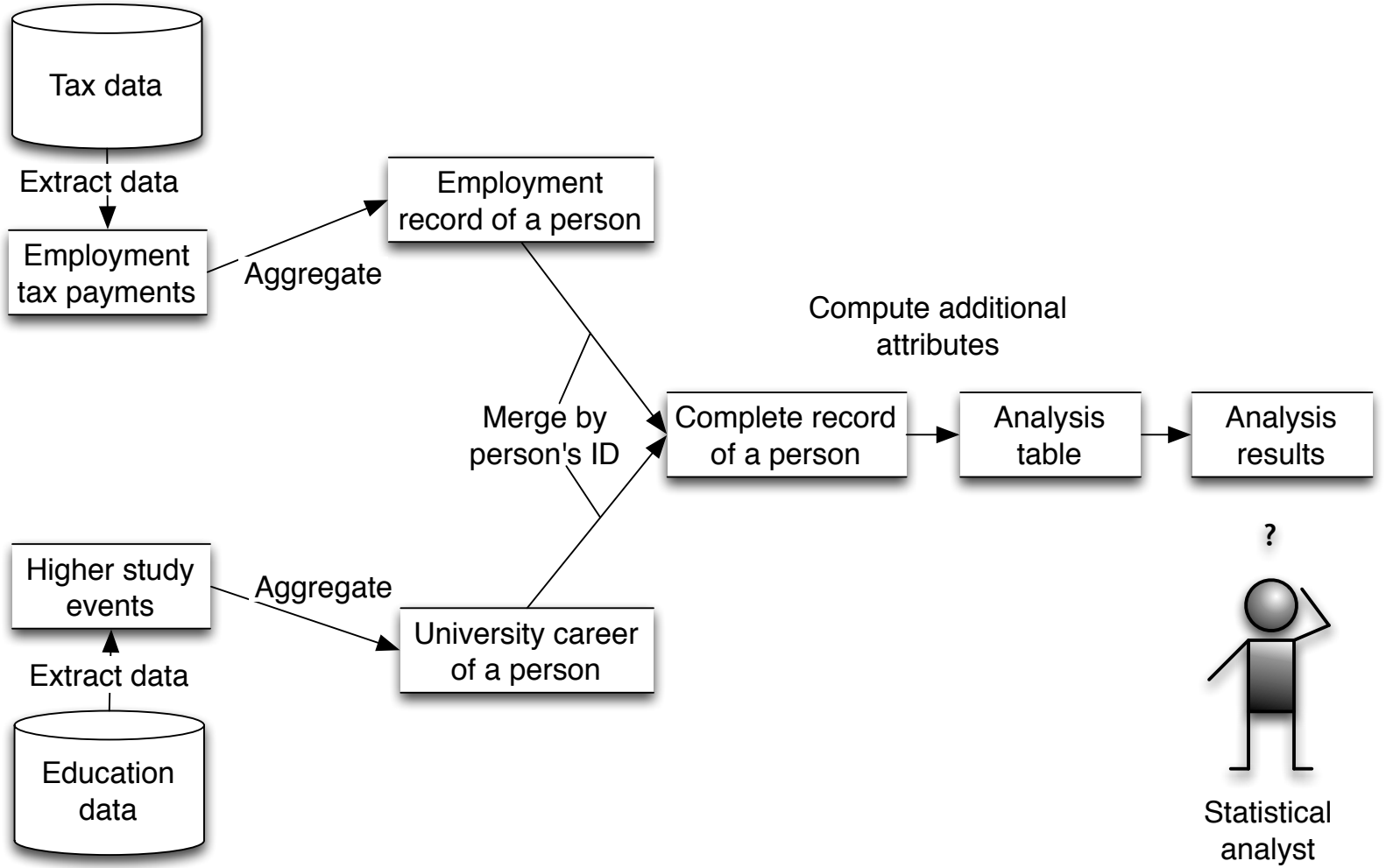




Rmind

CYBERNETICA

Data Analysis Workflow

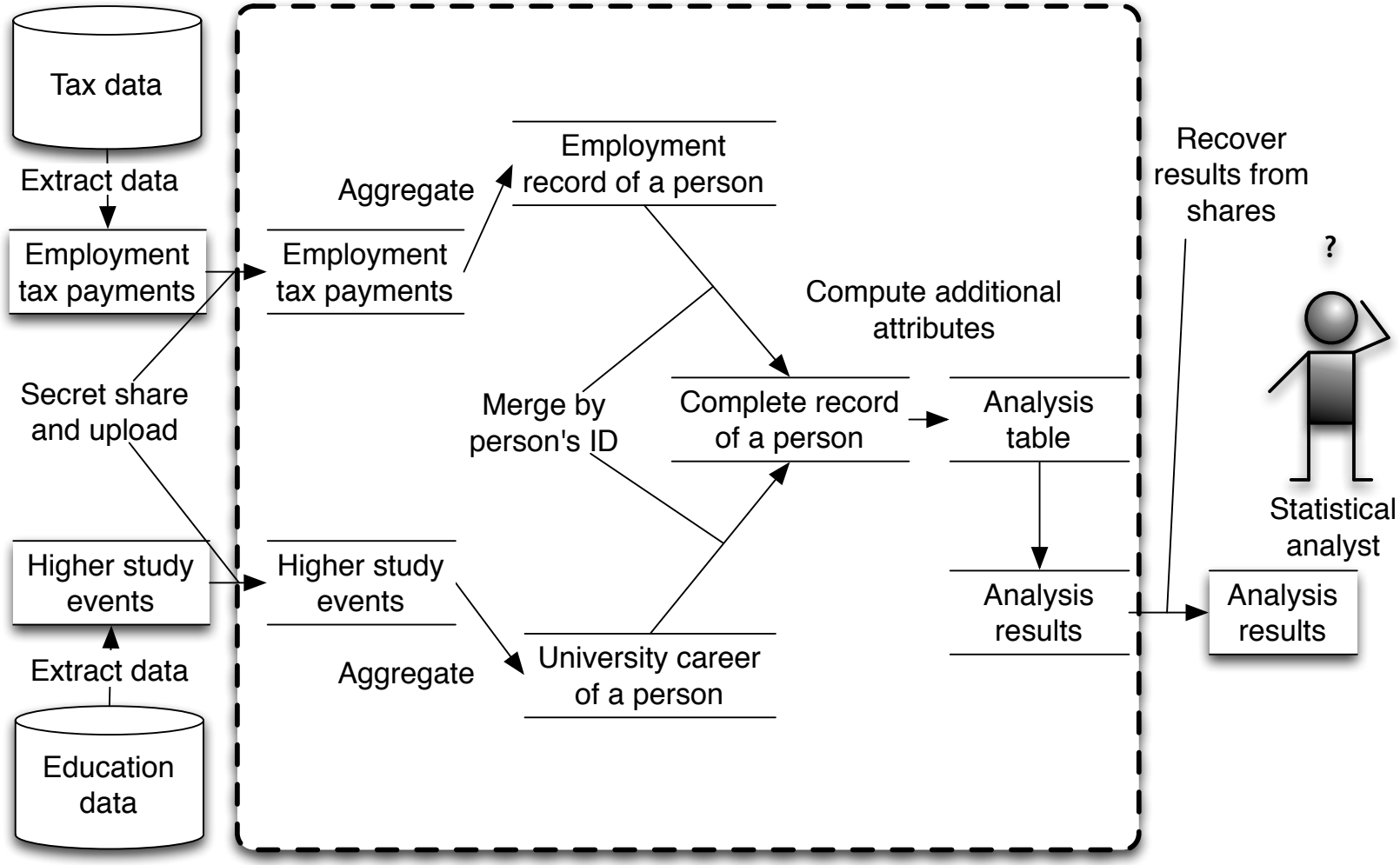


CYBERNETICA

Recipe for Converting a Workflow

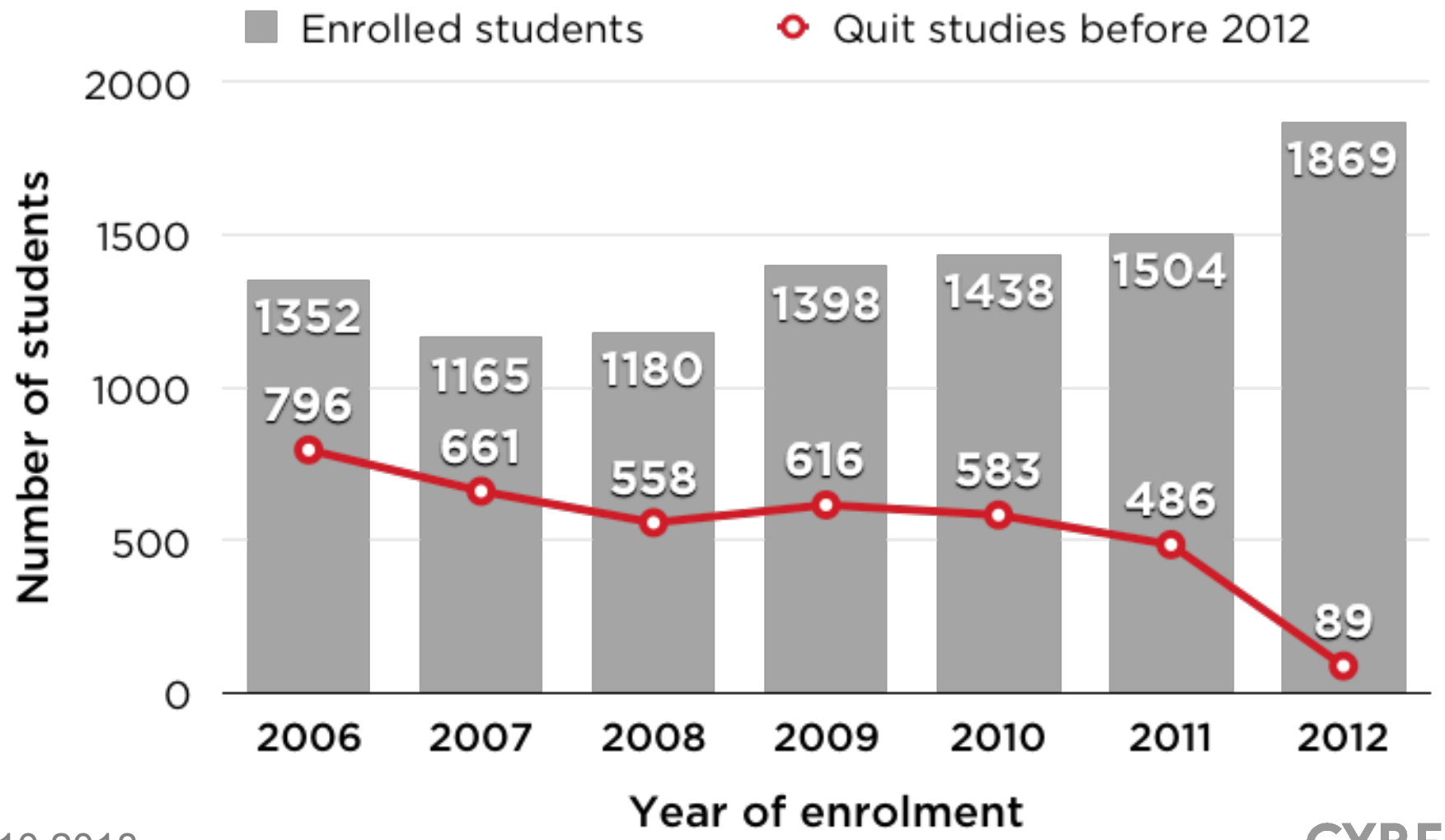
- Adapt algorithm to support secure computation
  - Optimise for the chosen secure computation technique
  - To avoid timing attacks
  - Reusable libraries exist
- Create data import tools
- Use existing or create custom query tools
  - Data analysts cannot view individual values

Data Analysis Workflow Using Secure MPC



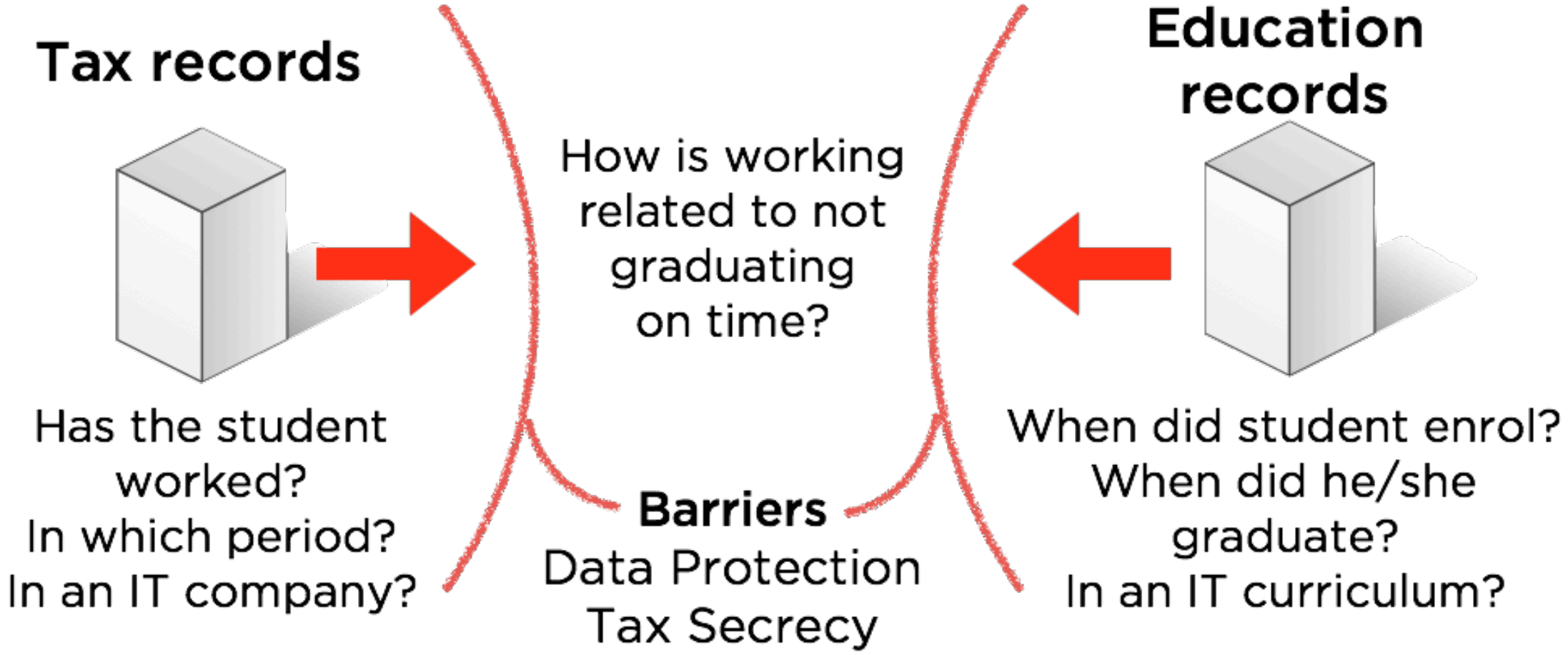
CYBERNETICA

Example: Linking Tax and Education Registries

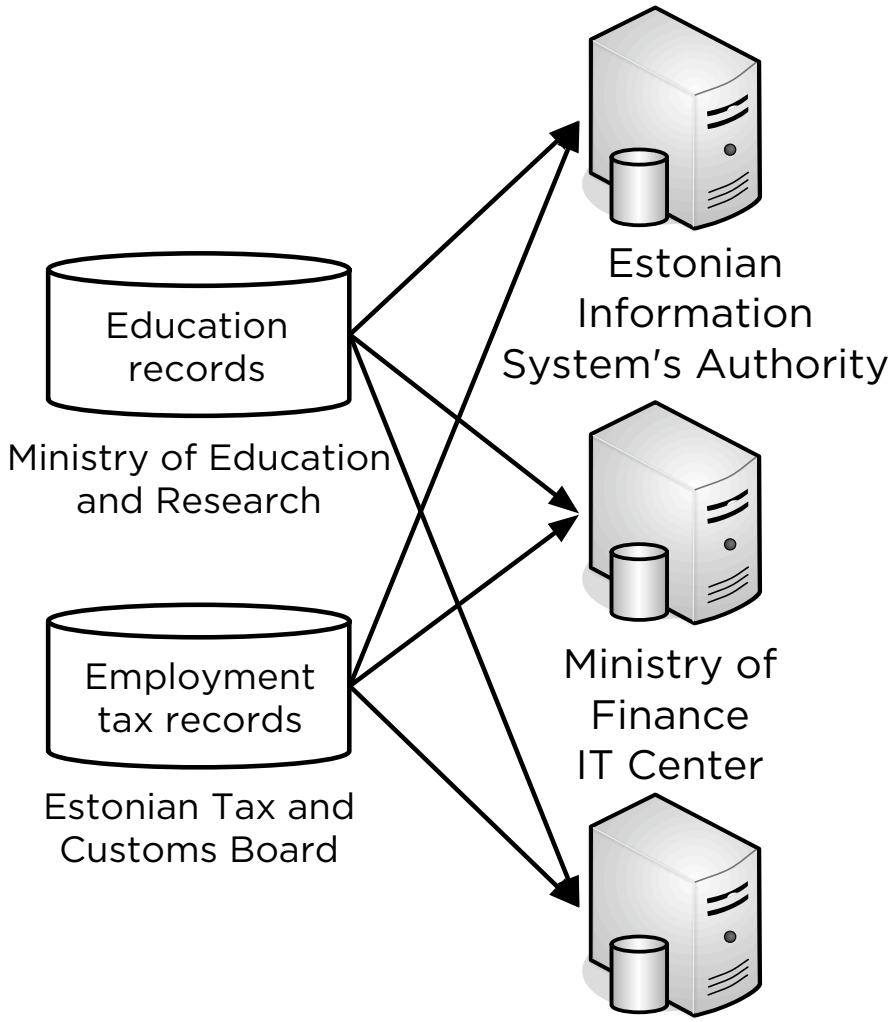


CYBERNETICA

Regulation Prevented a Data-Driven Answer

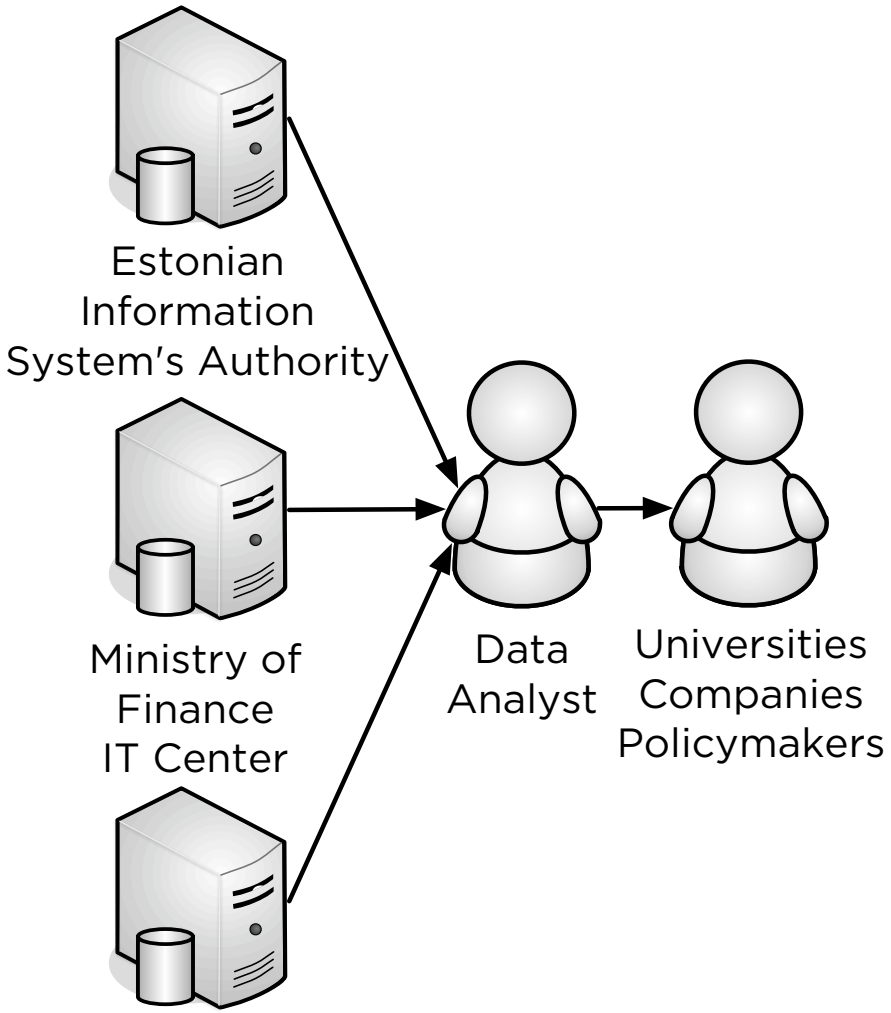


Privacy-Preserving Study of Students' Working Habits



- Source data:
  - 10 million tax records,
  - 600 000 education records.
- Sharemind hosted by government agencies and Cybernetica.
- Data owners used the Sharemind encryption tools to upload data.
- Data never existed outside the source in a decrypted state.

Privacy-Preserving Study of Students' Working Habits



- Data scientists used the Rmind tool to run the analysis.
- Sharemind prevented queries outside the study plan.
- Reports were given to industry, universities and the government.
- **Result**: no clear relation between working during studies and not graduating.

Regulatory precedents in Europe

- The Estonian Data Protection Agency stated that the combination of technology and processes ensured that **private data was not processed** and the requirements of the **Data Protection Act need not apply.**
  - Assumption: no identifiable records are published.
- The Internal Supervision of the Tax and Customs Board agreed to provide unmodified tax records after a code and process review.
- A German legal research team extended the precedent to work under the GDPR.
- We are now preparing for validation with other DPAs within new projects.

Example: Privacy-Preserving Mobile Data Analysis

LIVE DATA SOURCES

- Roaming phones in Estonia
- New data added every day
- Encrypted and uploaded with Sharemind HI tools

SECURE AGGREGATION

- Sharemind HI provides secure storage and processing
- 700 MB of data aggregated in a Trusted Execution Environment in under ten minutes

INTERACTIVE VISUALIZATION

- Explore inbound tourism data
- Statistical methodology provided by Positium

Privacy-Preserving Mobile Data Analysis Demo

Example: Privacy-Preserving Scanner Data Analysis

SYNTHETIC DATA

- Artificial price data provided by Stats NZ
- Encrypted and uploaded with Sharemind MPC tools

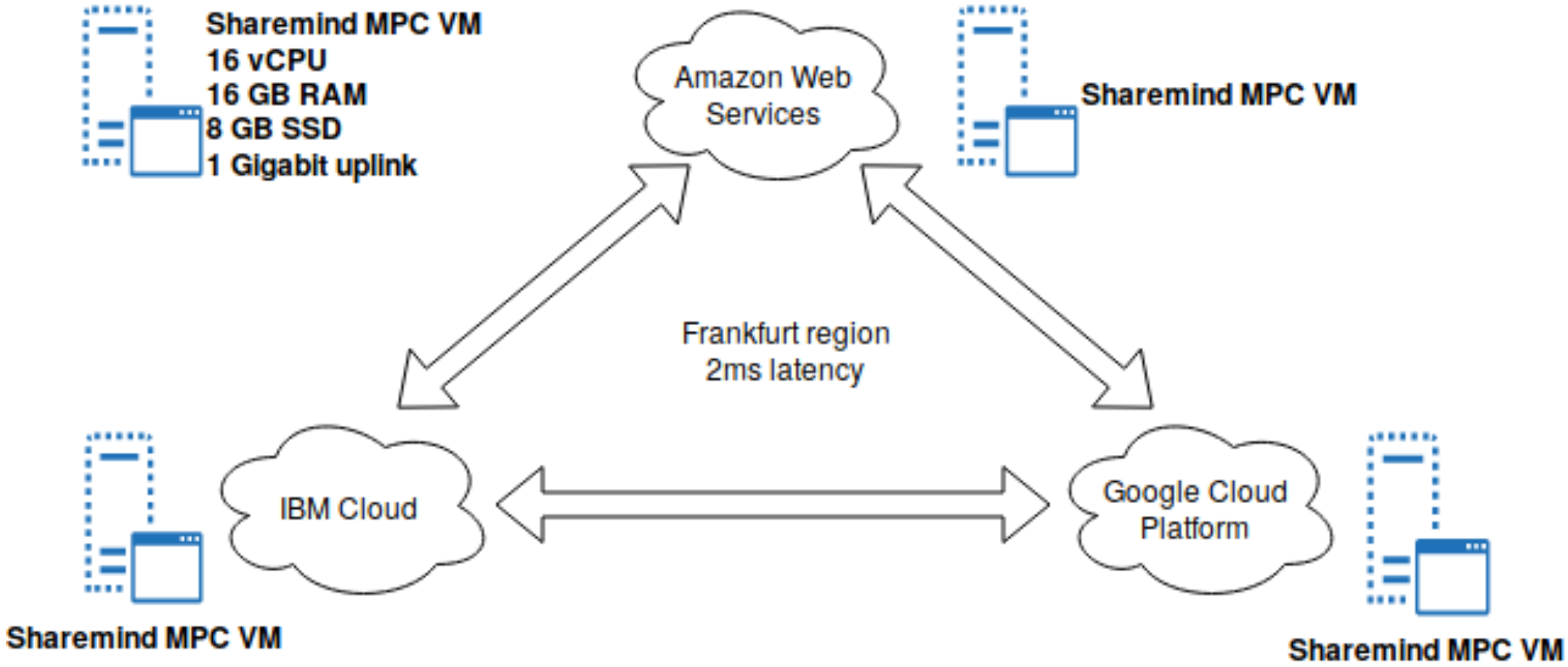
SECURE STATISTICS

- Sharemind MPC provides secure storage and processing
- Complete workflow performed in the privacy-preserving environment

CUSTOM INTERFACE

- Part of the UNGlobalPlatform
- Returns the price indices
- Methodology provided by Stats NZ

Privacy-Preserving Scanner Data Cloud Deployment



Privacy-Preserving Scanner Data Analysis

Data Profile	No. of Rows	No. of Sharemind Instances	Time Spent	Cloud Cost (Bandwidth)
230 products for 26 months	3843	2 (11 parallel regressions)	8m 57s	$31.50
230 products for 26 months	3843	4 (~5 parallel regressions)	7m32s	$31.50
460 products for 26 months	7654	4	22m 24s	$165

Find more information at
sharemind.cyber.ee


CYBERNETICA