Sharemind - practical privacy-preserving analytics

Sander Siim
Cybernetica AS
sander.siim@cyber.ee

CYBERNETICA

sharemind

About Sharemind

Sharemind uses MPC to analyse data that was not accessible before.

Sharemind resolves trust issues by removing centralised control and unwanted data access points.

Application Server paradigm

sharemind
interfaces

Java/JavaScript/C/C++/Haskell

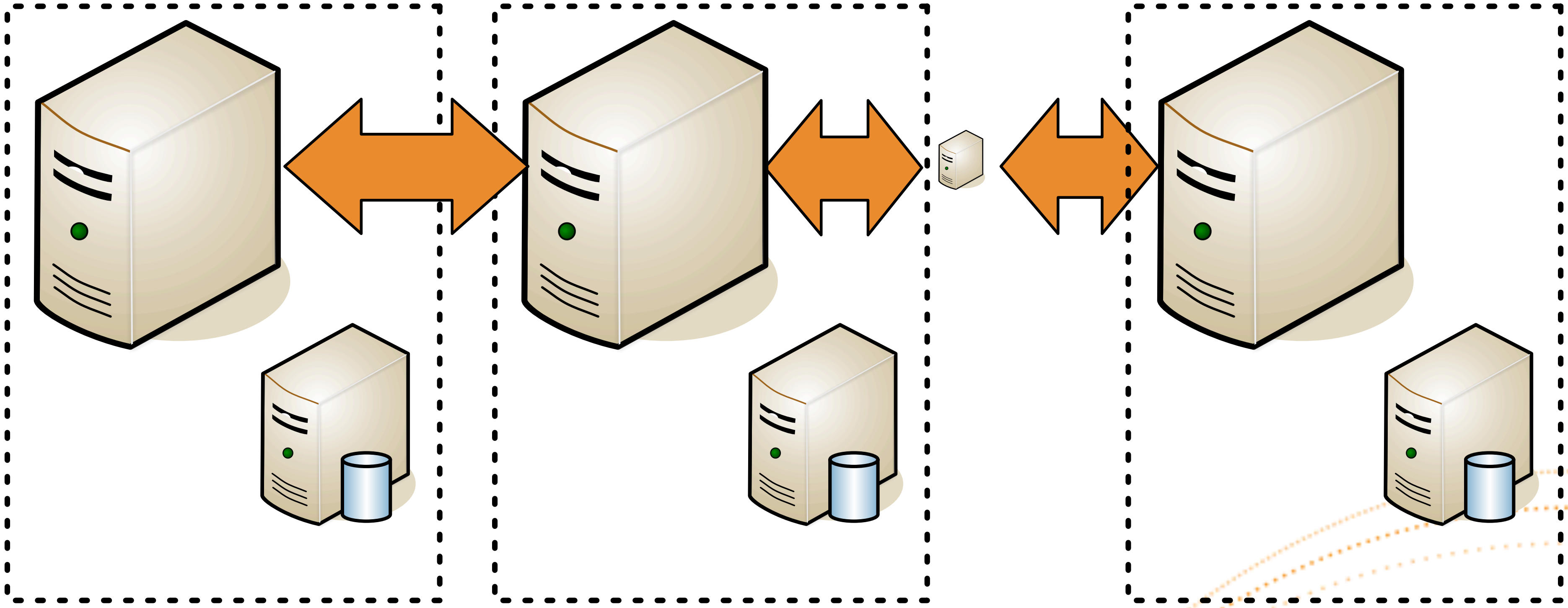
Mobile apps   Web apps   Desktop apps

SQL queries   Rmind statistics package

sharemind

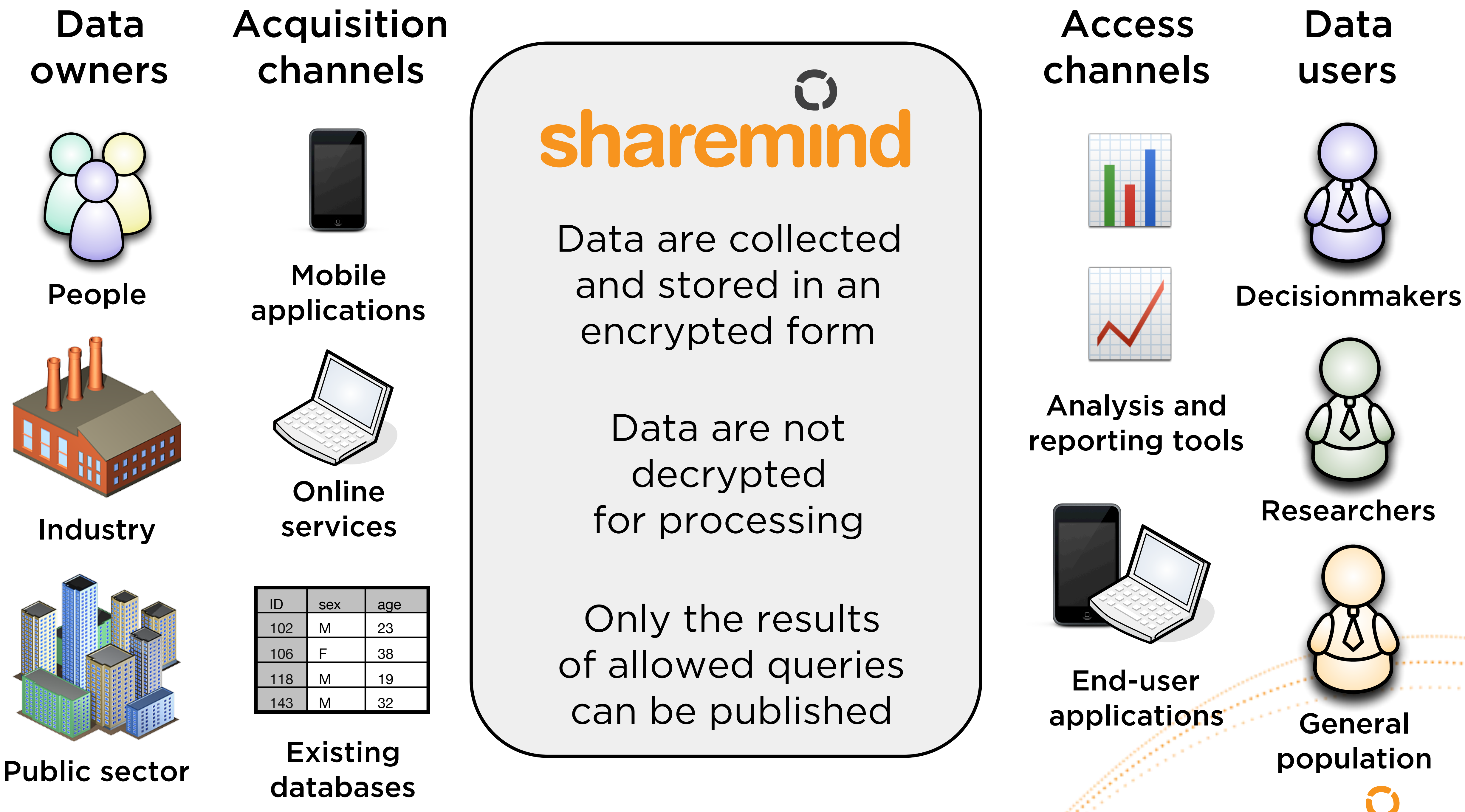
application servers

database backends



Host 1   Host 2   Host n

Encrypted computing



Data owners

People

Industry

Public sector

Acquisition channels

Mobile applications

Online services

ID	sex	age
102	M	23
106	F	38
118	M	19
143	M	32

Existing databases

sharemind

Data are collected and stored in an encrypted form

Data are not decrypted for processing

Only the results of allowed queries can be published

Access channels

Analysis and reporting tools

End-user applications

Data users

Decisionmakers

Researchers

General population

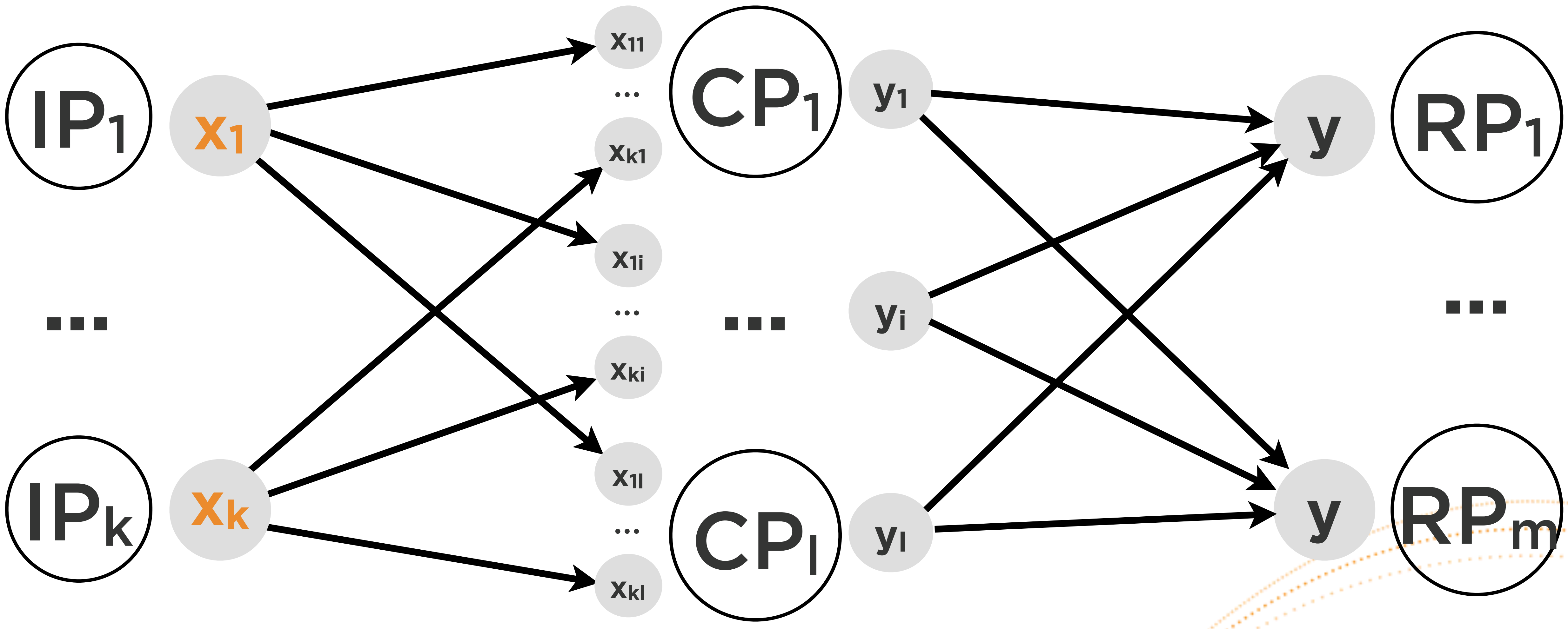
sharemind

Model of secure computing

Input parties

Computing parties

Result parties



Step 1:
upload and
storage of inputs

Step 2:
secure
computation

Step 3:
publishing
of results

sharemind

Secure computation cores

Name	num of input parties	num of computing parties	num of result parties	Technology	Status
shared3p	any	3	any	LSS MPC, (Yao)	In commercial use
shared2p	any	2	any	LSS MPC, (Yao)	Under development
sharednp	any	3 or more	any	LSS MPC	Under development

The shared3p core

- Storage: additive and bitwise secret sharing
- Computing: three-party MPC based on LSS
- Data types: 13 types (boolean, signed and unsigned integers, fixed point, floating point)
- Operations: 650 machine-optimized protocols

- Protocols developed by Cybernetica over the last 10 years, heavily tuned and optimized
- Powers all our commercial applications and most R&D prototypes

Protocol DSL and compiler

- Our newest and fastest protocols are implemented with a special-purpose compiler
- DSL(high-level description of π) = machine-code that runs π
- Easy to test and implement new protocols
- Optimizes protocol structure and communication — up to 40x speed-up
- Helps maintain our growing library of protocols
- Can use also in 2-party/n-party case

Peeter Laud and Jaak Randmets. A domain-specific language for low-level secure multiparty computation protocols. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-6, 2015, pages 1492–1503. ACM, 2015.

sharemind

Cores in development

shared2p

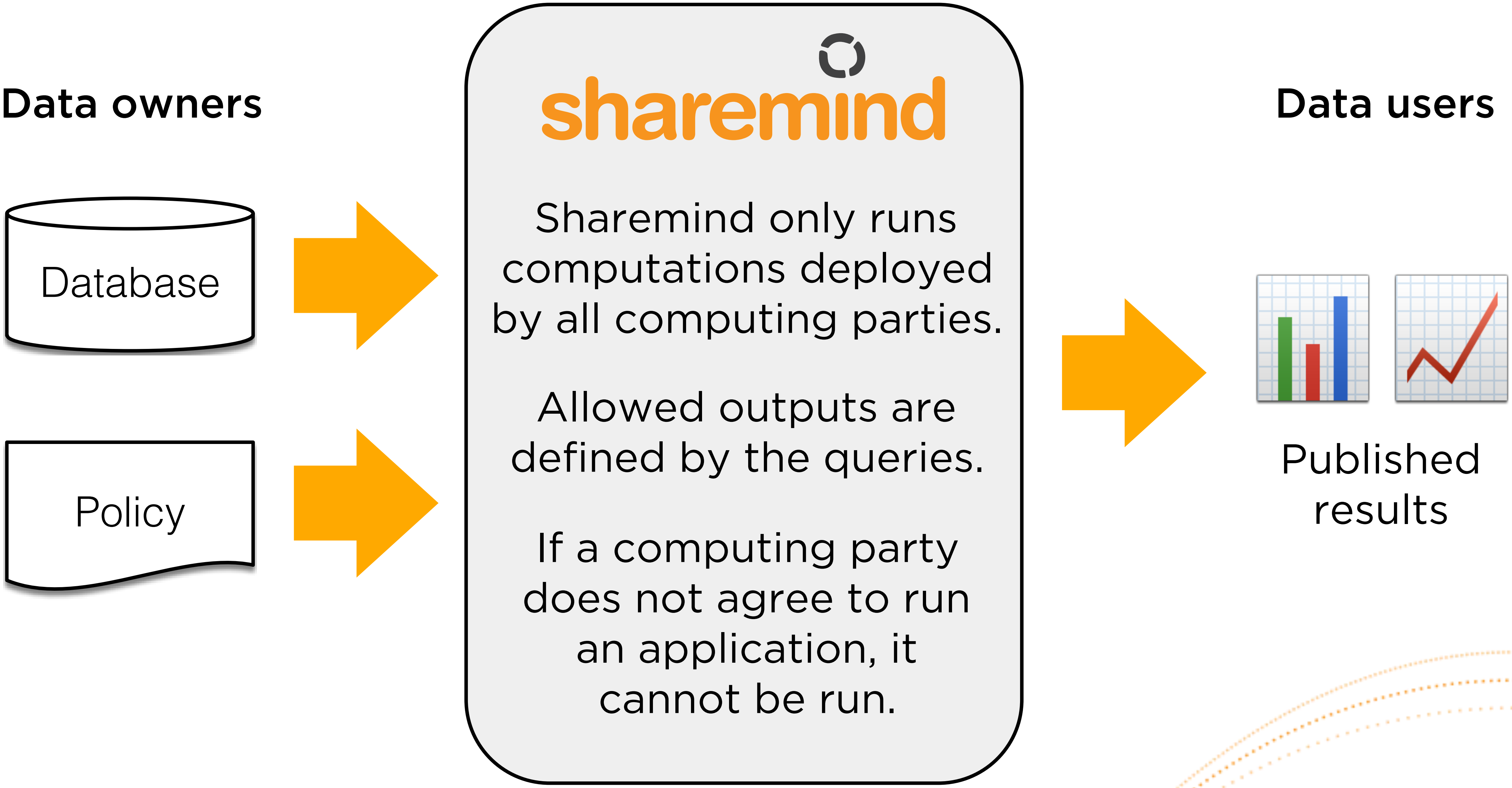
- Storage: additive and bitwise secret sharing
- Computing: two-party secure MPC
- Combination of shared3p techniques with Beaver triples

sharednp

- Storage: Shamir's secret sharing
- Computing: *n*-party secure MPC
- Classic Shamir protocols + custom designs

sharemind

Controlling computations



Data owners

Database

Policy

sharemind

Sharemind only runs computations deployed by all computing parties.

Allowed outputs are defined by the queries.

If a computing party does not agree to run an application, it cannot be run.

Data users

Published results

The SecreC language

```
// Import module for the secure protocol suite
import shared3p;
// Data in private domain is processed via MPC
domain private shared3p;

void main () {
  // Perform secure computations
  private int a = 2, b = 3;
  private int c = a * b;
  // Must explicitly declare publishing c
  print (declassify (c));
}
```

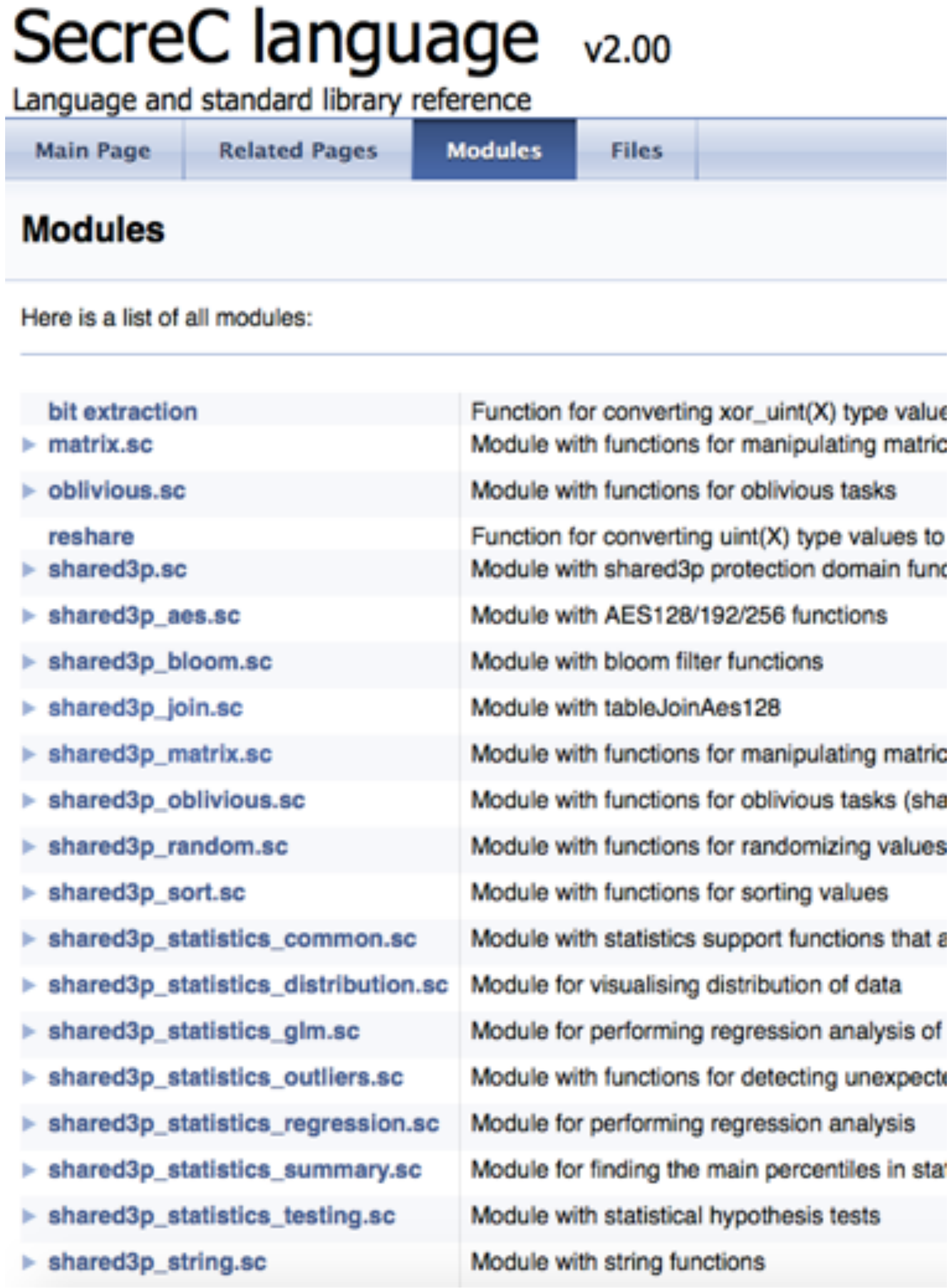
sharemind

Polymorphic functions

```
template <domain D>
D int scalarProd(D int[[1]] x, D int[[1]] y) {
  return sum(x*y);
}
domain private3 shared3p;
domain private2 shared2p;

void main () {
  private3 int[[1]] x3(100) = 2, y3(100) = 3;
  private2 int[[1]] x2(100) = 2, y2(100) = 3;
  print (declassify (scalarProd(x3, y3)));
  print (declassify (scalarProd(x2, y2)));
}
```

SecreC standard library



SecreC language v2.00

Language and standard library reference

Modules

Here is a list of all modules:

bit extraction	Function for converting xor_uint(X) type value
▸ matrix.sc	Module with functions for manipulating matric
▸ oblivious.sc	Module with functions for oblivious tasks
reshare	Function for converting uint(X) type values to
▸ shared3p.sc	Module with shared3p protection domain func
▸ shared3p_aes.sc	Module with AES128/192/256 functions
▸ shared3p_bloom.sc	Module with bloom filter functions
▸ shared3p_join.sc	Module with tableJoinAes128
▸ shared3p_matrix.sc	Module with functions for manipulating matric
▸ shared3p_oblivious.sc	Module with functions for oblivious tasks (sha
▸ shared3p_random.sc	Module with functions for randomizing values
▸ shared3p_sort.sc	Module with functions for sorting values
▸ shared3p_statistics_common.sc	Module with statistics support functions that a
▸ shared3p_statistics_distribution.sc	Module for visualising distribution of data
▸ shared3p_statistics_glm.sc	Module for performing regression analysis of
▸ shared3p_statistics_outliers.sc	Module with functions for detecting unexpecte
▸ shared3p_statistics_regression.sc	Module for performing regression analysis
▸ shared3p_statistics_summary.sc	Module for finding the main percentiles in sta
▸ shared3p_statistics_testing.sc	Module with statistical hypothesis tests
▸ shared3p_string.sc	Module with string functions

- A library of privacy-preserving algorithms.
- Array and matrix operations, oblivious access, statistical testing, sorting, linking, regression modelling, aggregation, etc.
- 15 000 lines of reusable SecreC code

Demo!
Prototype an MPC application in minutes

Sharemind SDK

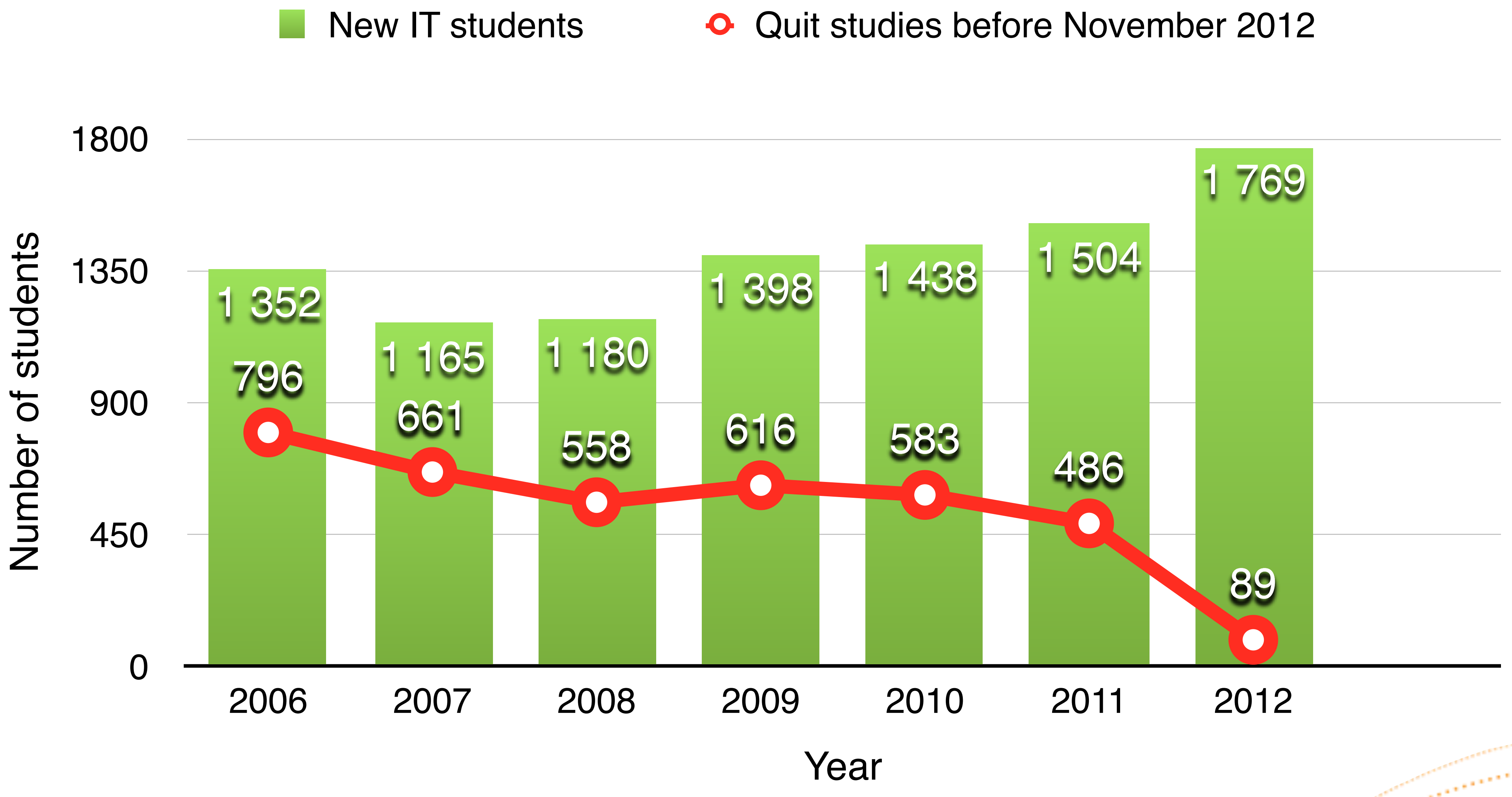
- Free open-source prototyping tools available:

  http://sharemind-sdk.github.io/

- Includes SecreC and the standard library
- An emulated Sharemind run-time that estimates online performance
- Excellent for quick prototyping

Case study:
Government data analytics

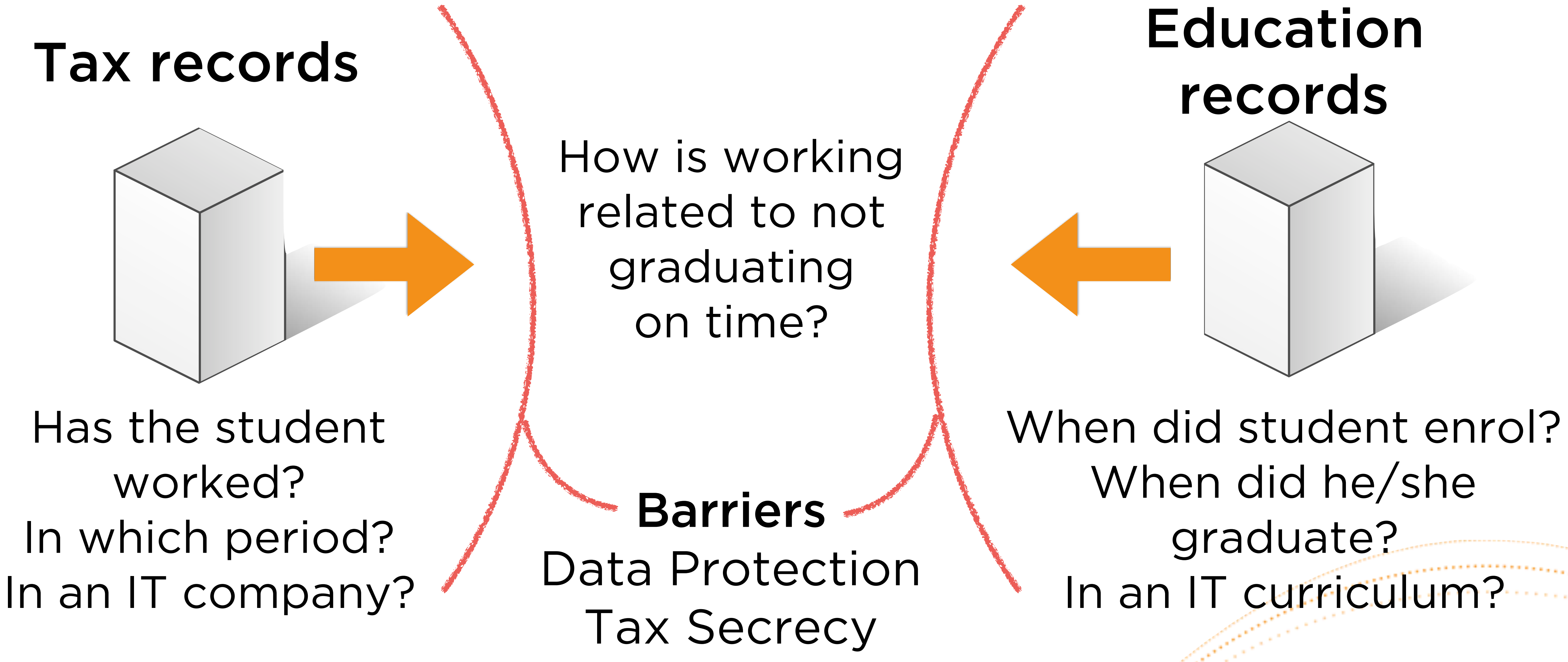
IT training has a failure rate



By 2012, a total of 43% of students enrolled in in the four largest IT higher learning institutions in Estonia during 2006-2012 had quit their studies. Source: Estonian Ministry of Education and Research, CentAR.

sharemind

Barriers for assessing the situation



Tax records

Has the student worked?
In which period?
In an IT company?

How is working related to not graduating on time?

Barriers
Data Protection
Tax Secrecy

Education records

When did student enrol?
When did he/she graduate?
In an IT curriculum?

Dan Bogdanov, Liina Kamm, Baldur Kubo, Reimo Rebane, Ville Sokk, Riivo Talviste. Students and Taxes: a Privacy-Preserving Social Study Using Secure Computation. In Proceedings on Privacy Enhancing Technologies, PoPETs, 2016 (3), pp 117–135, 2016.

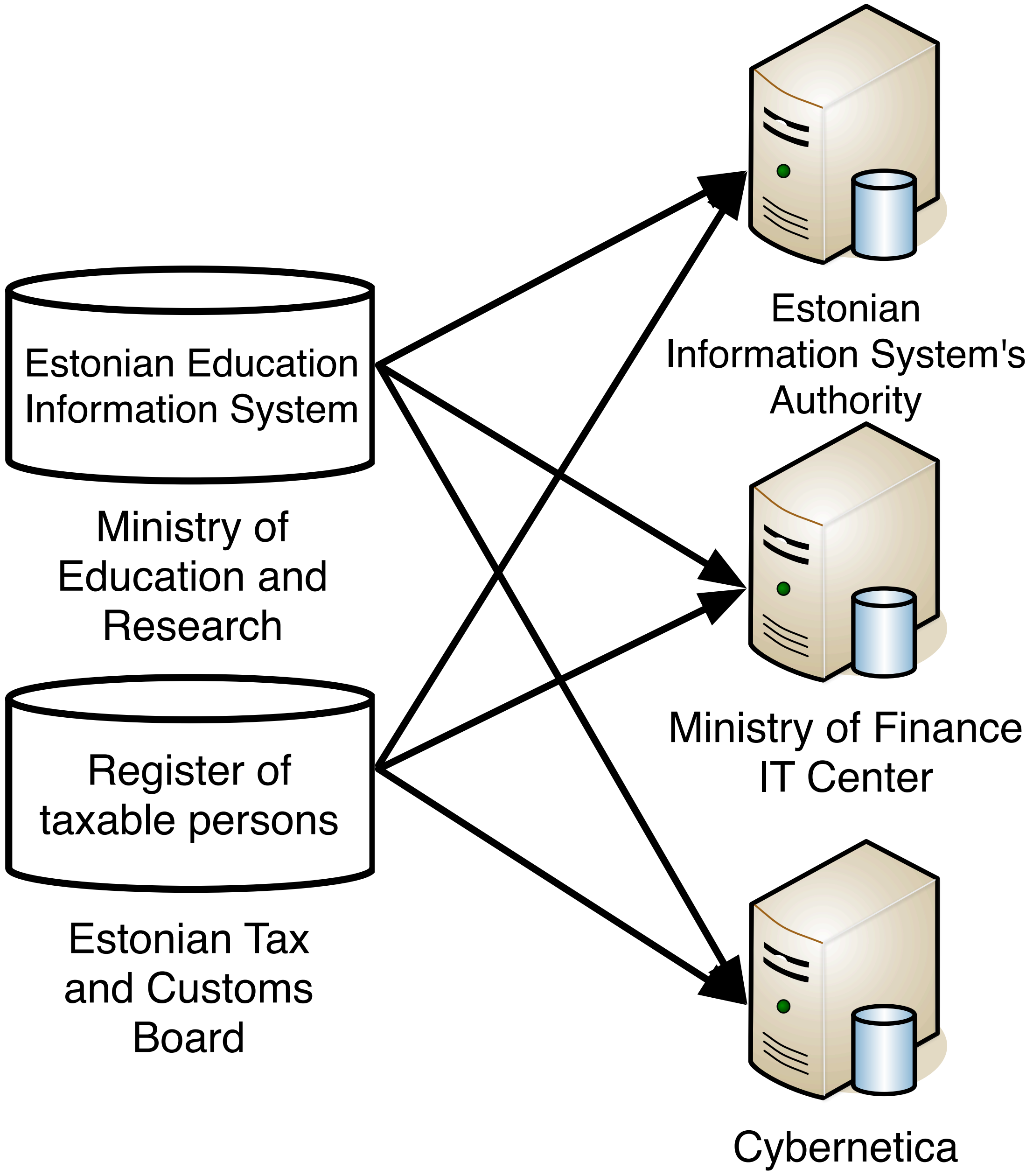
Legal breakthroughs

January 2014: Estonian Data Protection Agency declared that Sharemind technology and processes protect data so well that the Personal Data Protection Act doesn’t apply.

January 2015: after a code audit, the internal oversight at the Tax Board agreed to upload actual income tax records into the Sharemind-based analysis system.

February 2015: the Tax Board, Ministry of Education, Information Systems Authority, Ministry of Finance IT Center and Cybernetica signed the world’s first secure multi-party data analysis agreement.

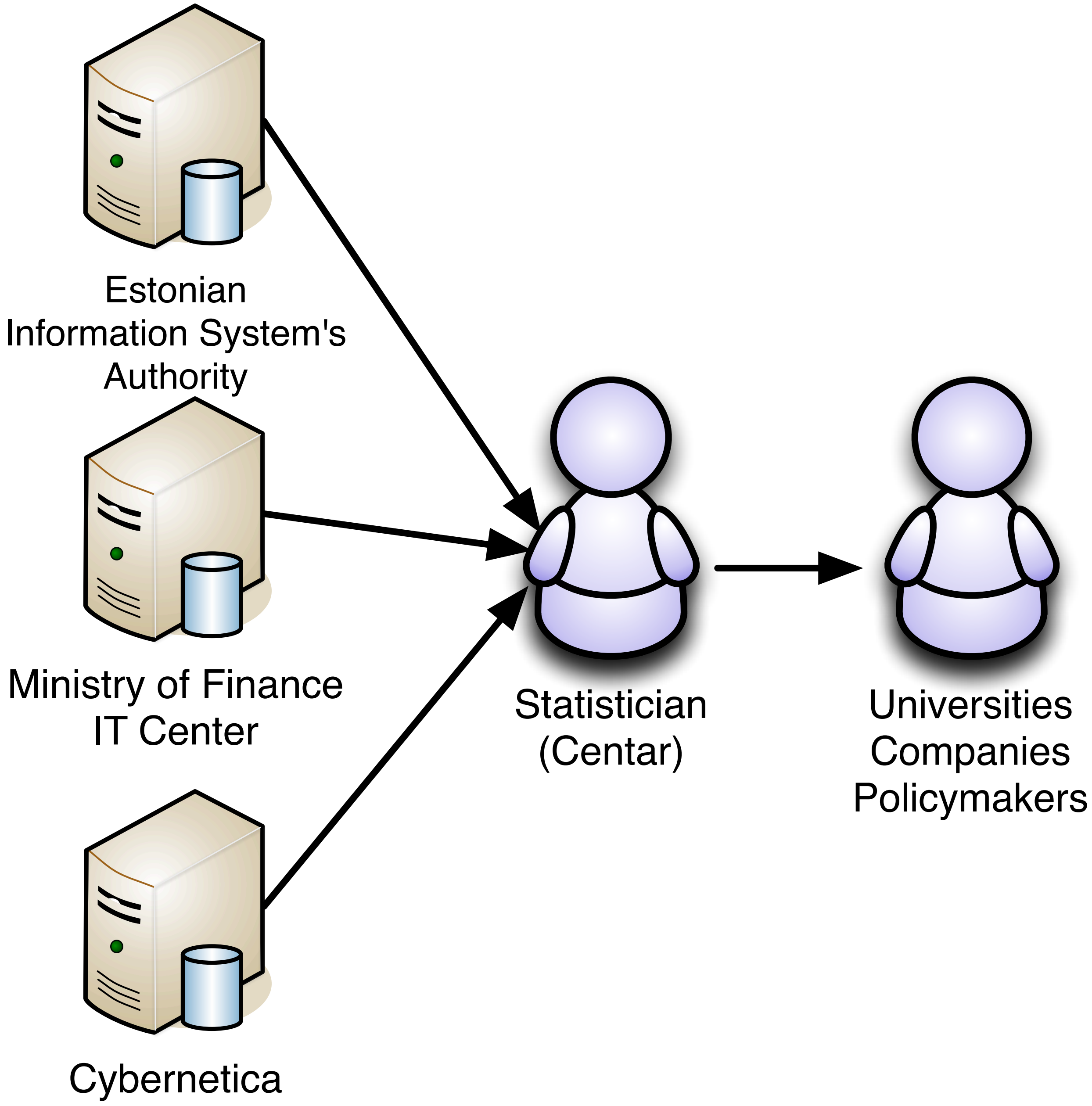
Step 1: Import data


Estonian Education Information System
Ministry of Education and Research
Register of taxable persons
Estonian Tax and Customs Board
Estonian Information System's Authority
Ministry of Finance IT Center
Cybernetica

- Data owners uploaded data with the Sharemind importer to a shared3p core.
- Each value was encrypted at the source, private data never left the data owner.
- Over 600 000 study records (100 MB) used.
- Over 10 million tax records (1 GB) used.
- Largest MPC application on real-world data.

sharemind

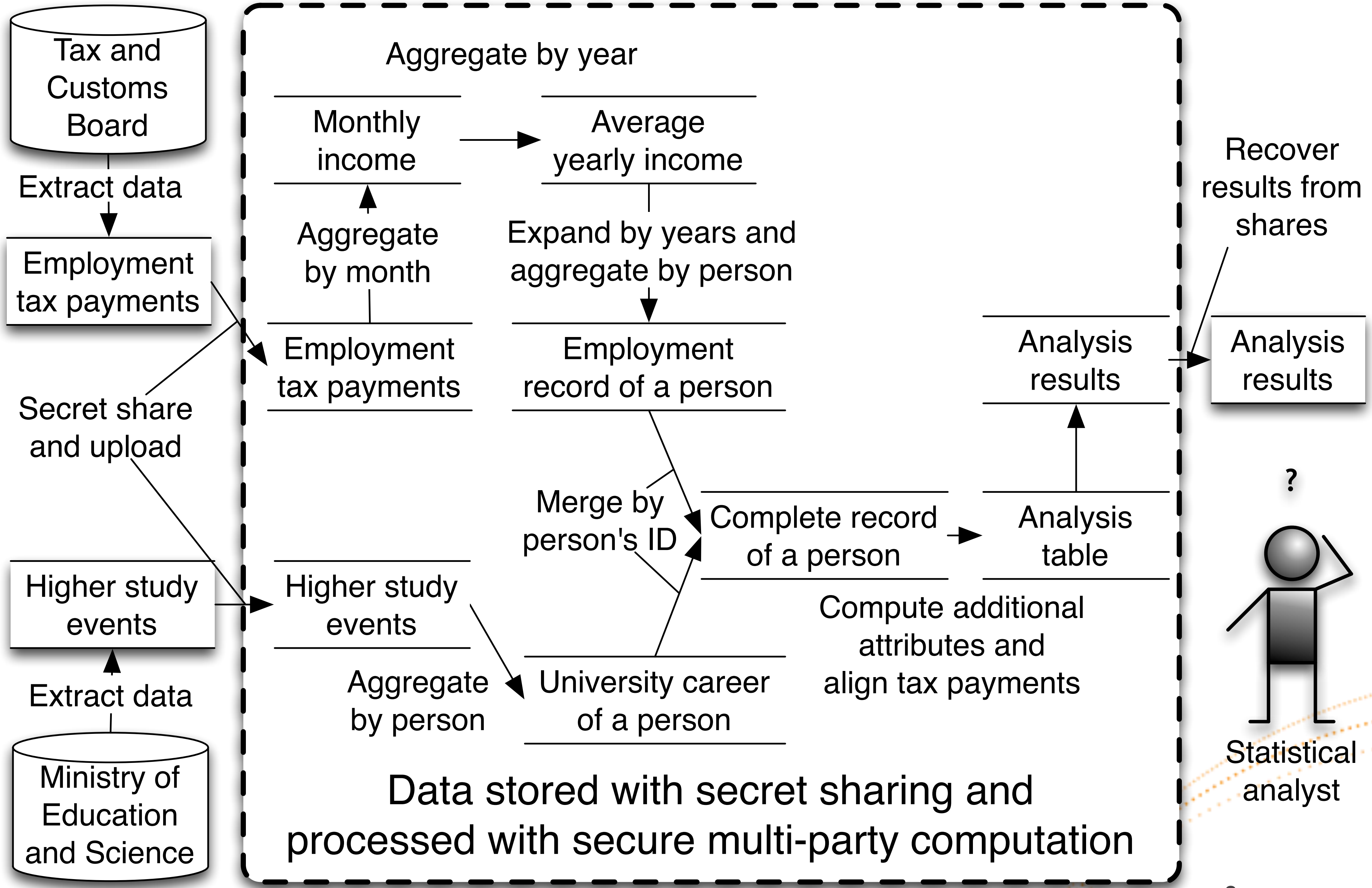
Step 2: Run the analysis



- Statisticians used Rmind to post queries.
- Sharemind ensured that only queries in the study plan were actually executed.
- Additional microdata protection controls were enforced.

sharemind

Operations performed



Tax and Customs Board

Extract data

Employment tax payments

Secret share and upload

Higher study events

Extract data

Ministry of Education and Science

Aggregate by year

Monthly income

Average yearly income

Aggregate by month

Expand by years and aggregate by person

Employment tax payments

Employment record of a person

Merge by person's ID

Complete record of a person

Analysis table

Analysis results

Higher study events

Aggregate by person

University career of a person

Compute additional attributes and align tax payments

Data stored with secret sharing and processed with secure multi-party computation

Recover results from shares

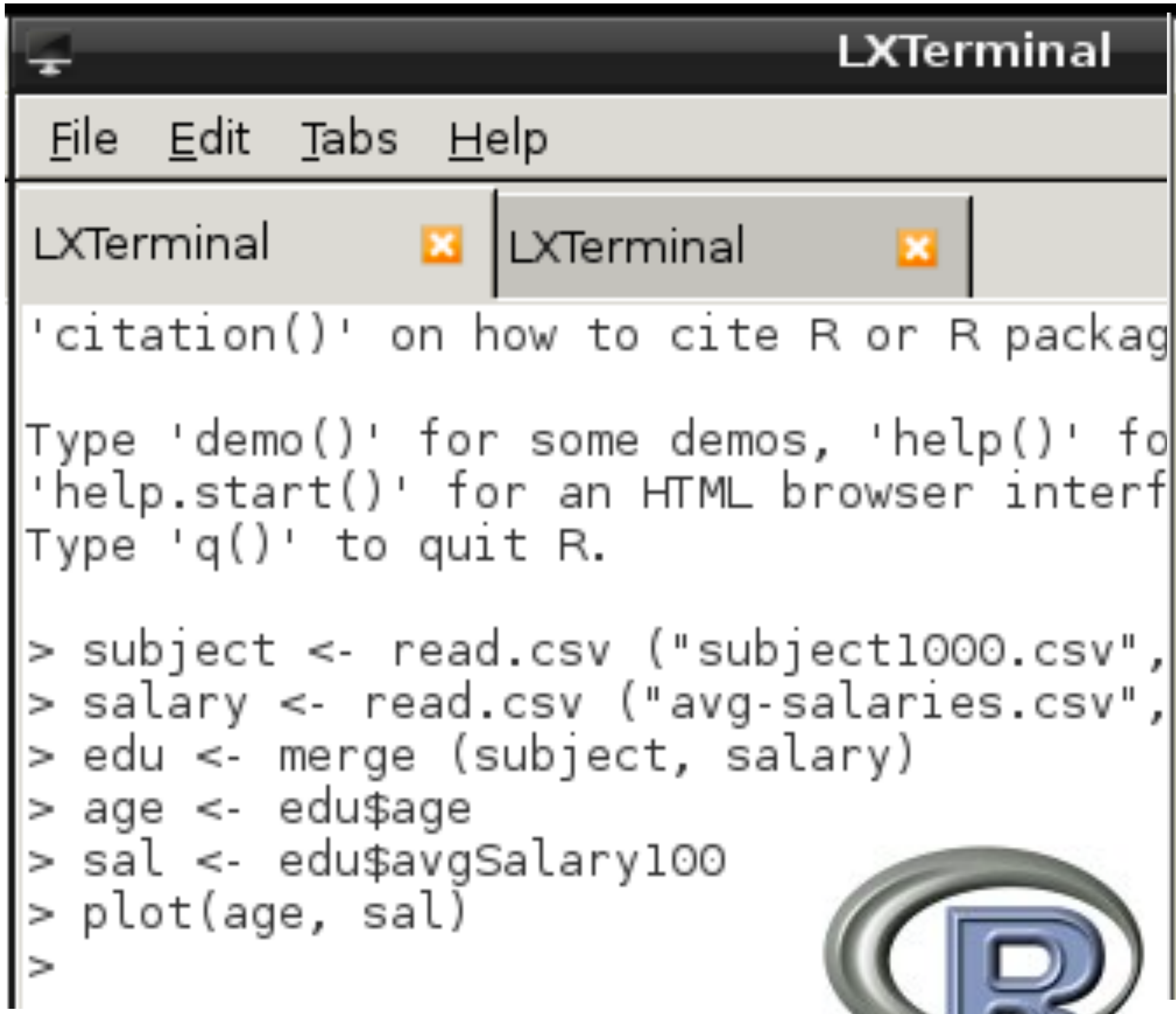
Analysis results

?

Statistical analyst

sharemind

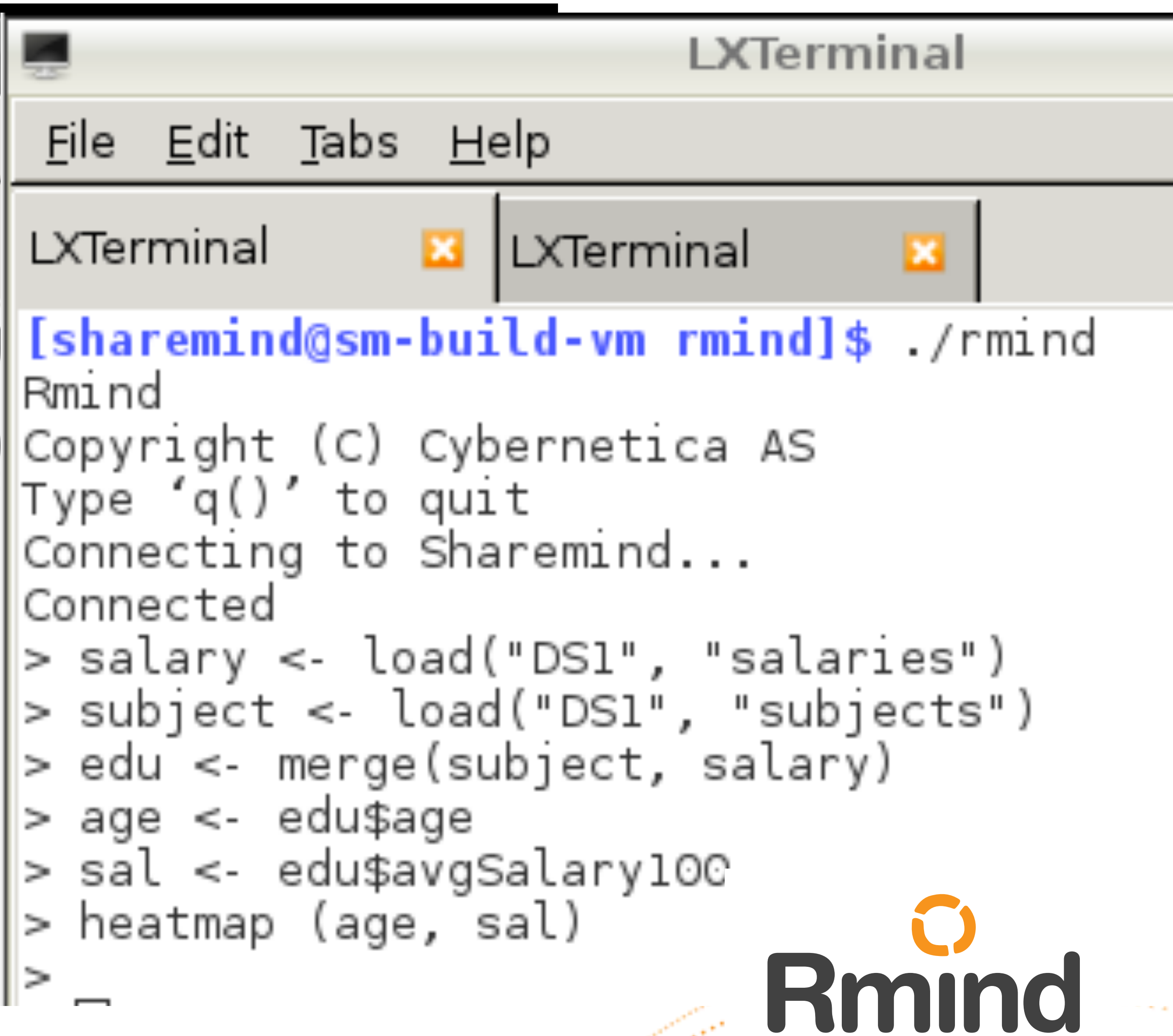
Sharemind Analytics Engine



```
'citation()' on how to cite R or R packag

Type 'demo()' for some demos, 'help()' fo
'help.start()' for an HTML browser interf
Type 'q()' to quit R.

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

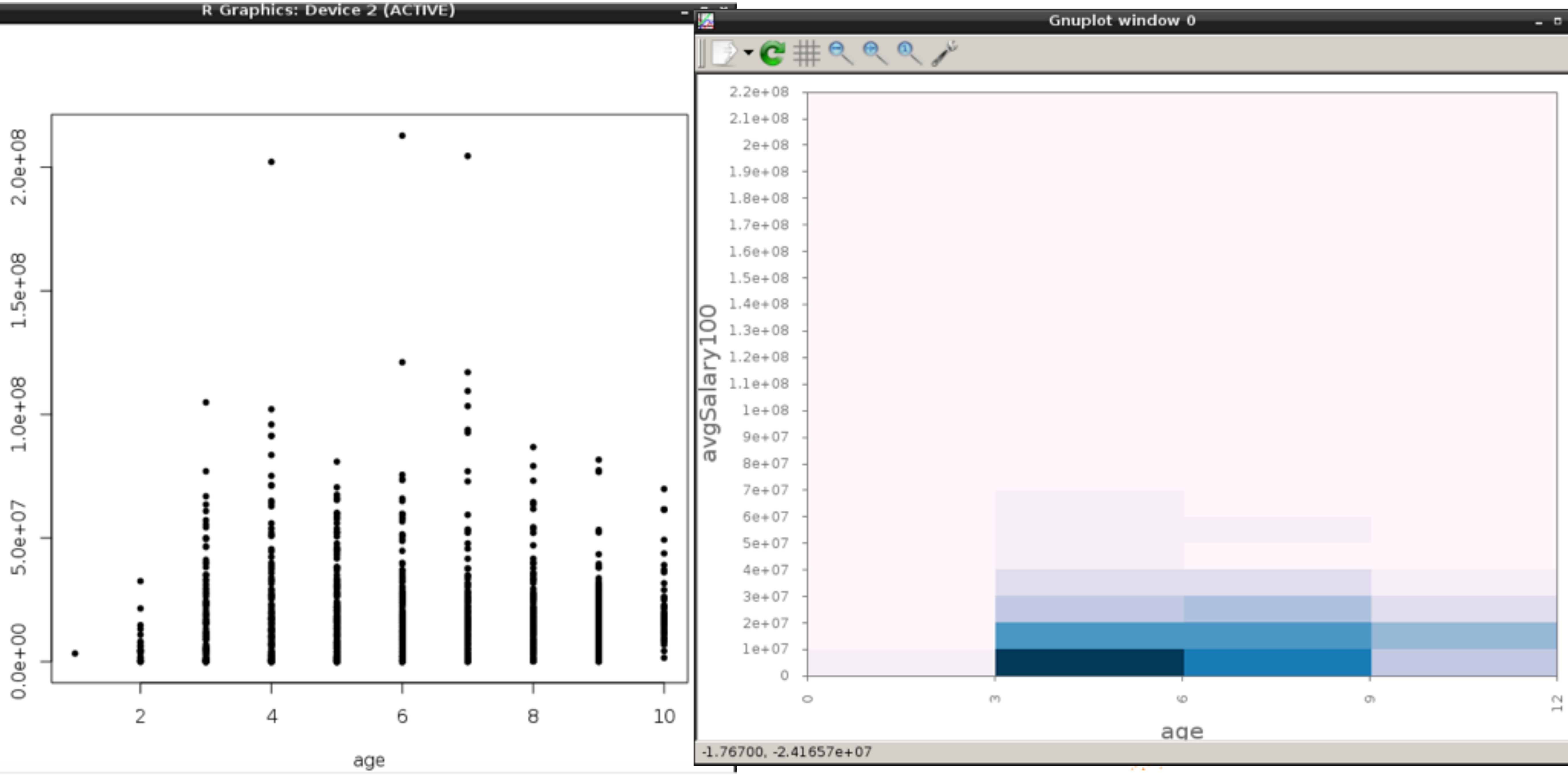


```
[sharemind@sm-build-vm rmind]$ ./rmind
Rmind
Copyright (C) Cybernetica AS
Type ‘q()’ to quit
Connecting to Sharemind...
Connected
> salary <- load("DS1", "salaries")
> subject <- load("DS1", "subjects")
> edu <- merge(subject, salary)
> age <- edu$age
> sal <- edu$avgSalary100
> heatmap (age, sal)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

Rmind

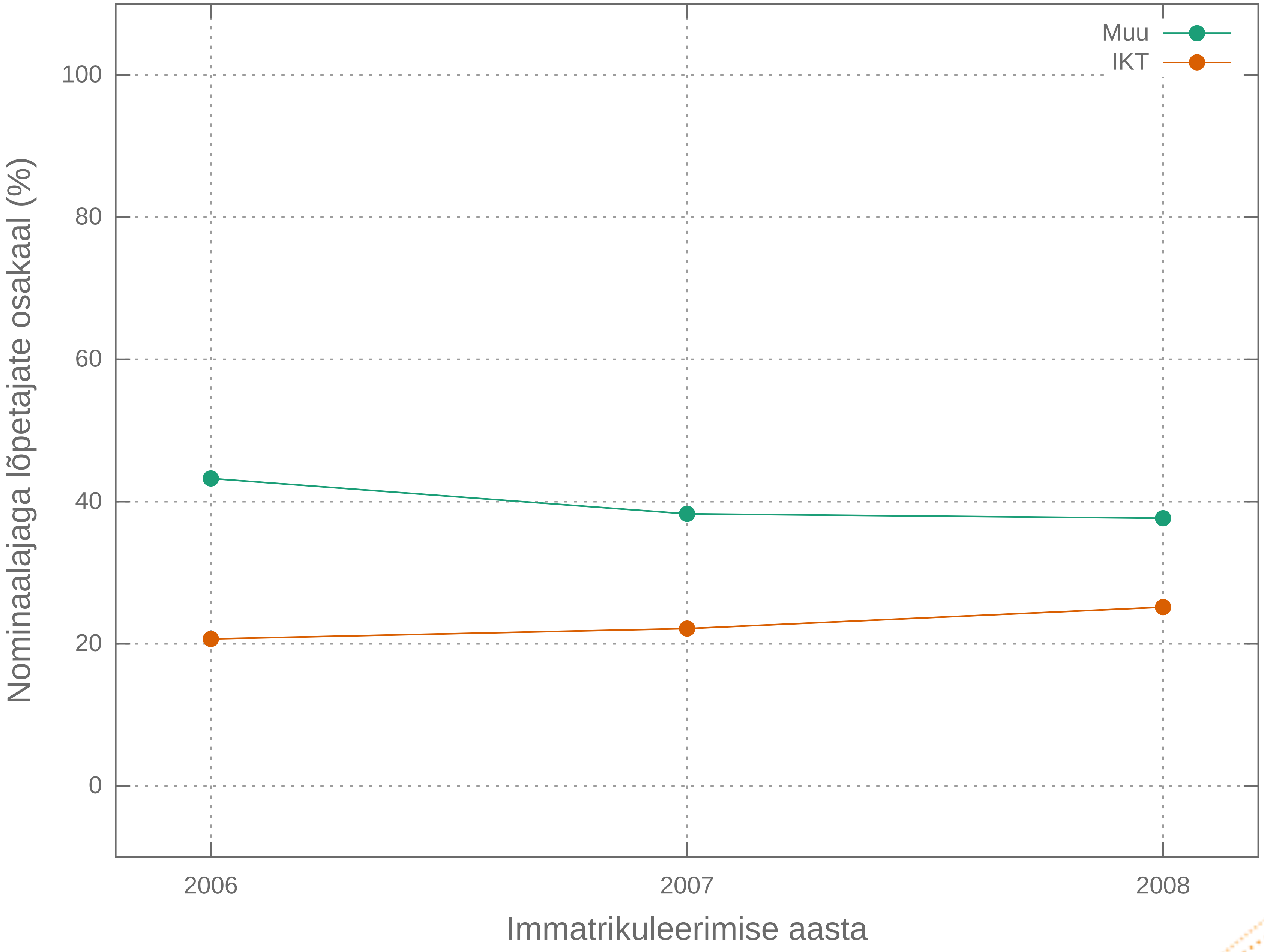
sharemind

Sharemind Analytics Engine



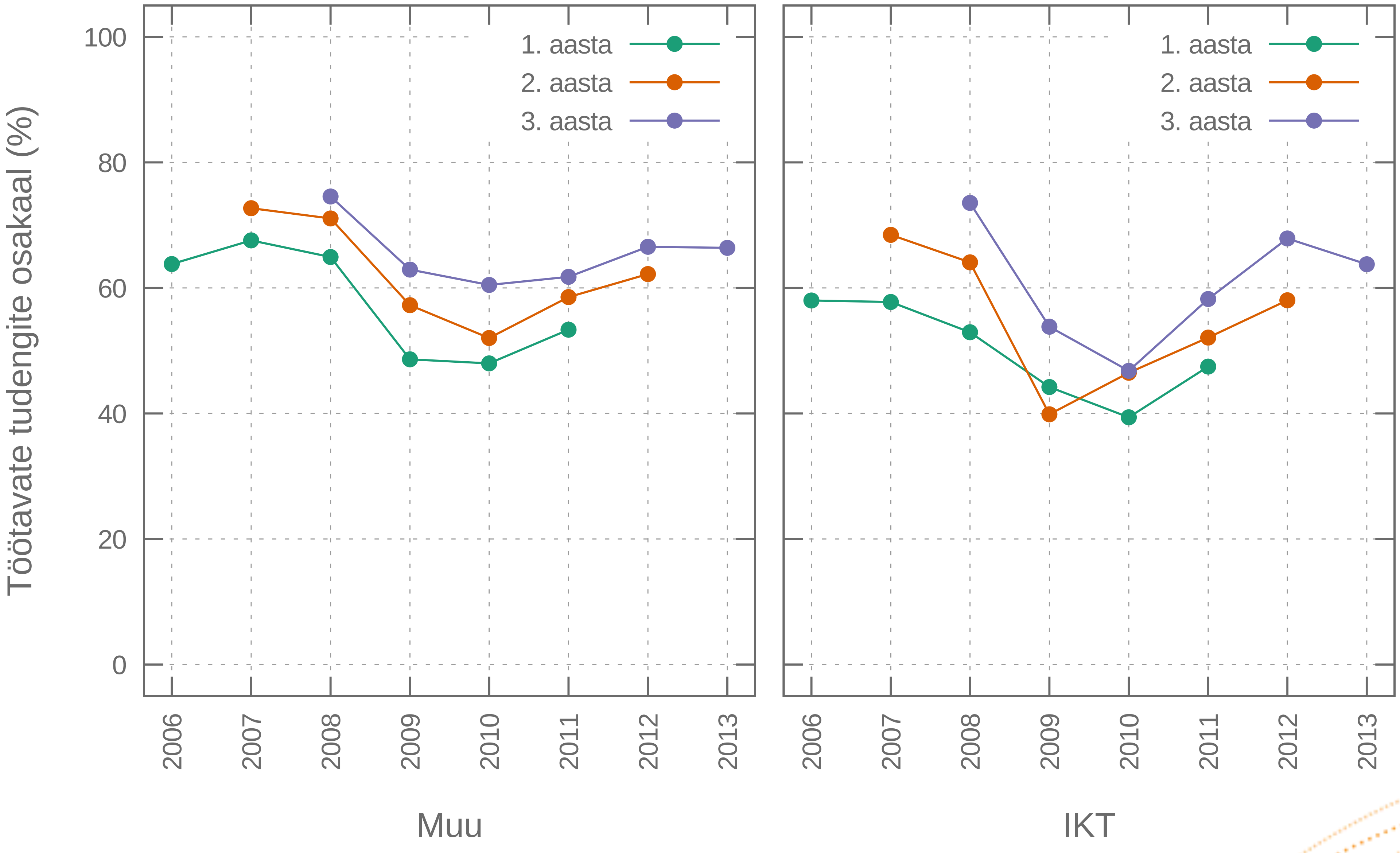
Rmind

IT is harder to graduate



Joonis 1. Nominaalajaga lõpetajate osakaal immatrikuleerimisaastate lõikes, IKT- ja mitte-IKT õppekavad, bakalaureuseõpe

All students are working



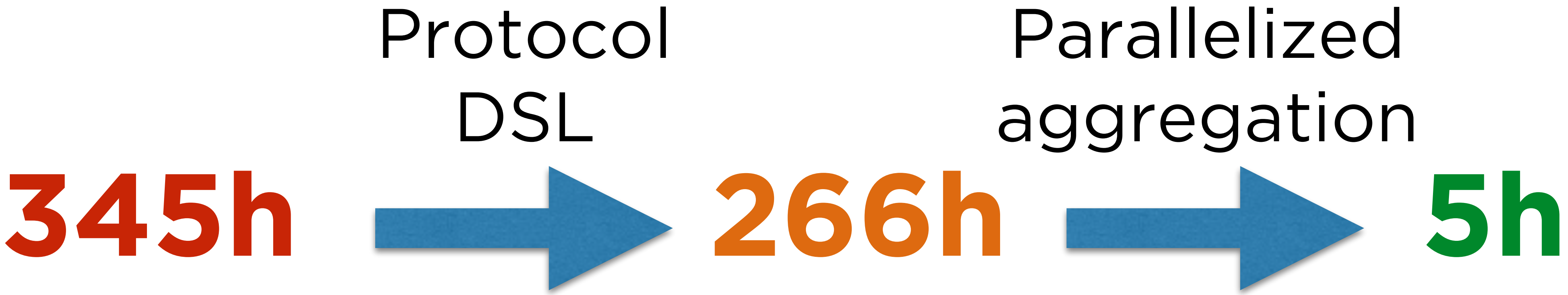
Joonis 4. Nominaalaja jooksul töötanud tudengite osakaal kõigist tudengitest aastati, IKT- ja mitte-IKT õppekavad, bakalaureuseõpe

sharemind

Practice makes perfect

- After successfully ending the project, we went back to the lab to see if we can do better
- The new protocol DSL gave a “conservative” 20% performance improvement
- It turned out we could significantly optimize the aggregation algorithms through better parallelization

Major speed-ups



345h

Protocol DSL

266h

Parallelized aggregation

5h

6 ms latency for one server, 1Gbps bandwidth

More gains from **high-level algorithm optimizations** than low-level protocols

Case study:
A privacy-preserving survey system

Privacy-preserving surveys

- Traditional survey systems do not hide individual answers from organizer/server
- Use MPC to remove centralised trusted service provider
- We built a secure survey system in the PRACTICE project together with Alexandra Institute and Partisia
- Has both Sharemind and Fresco/SPDZ back-ends

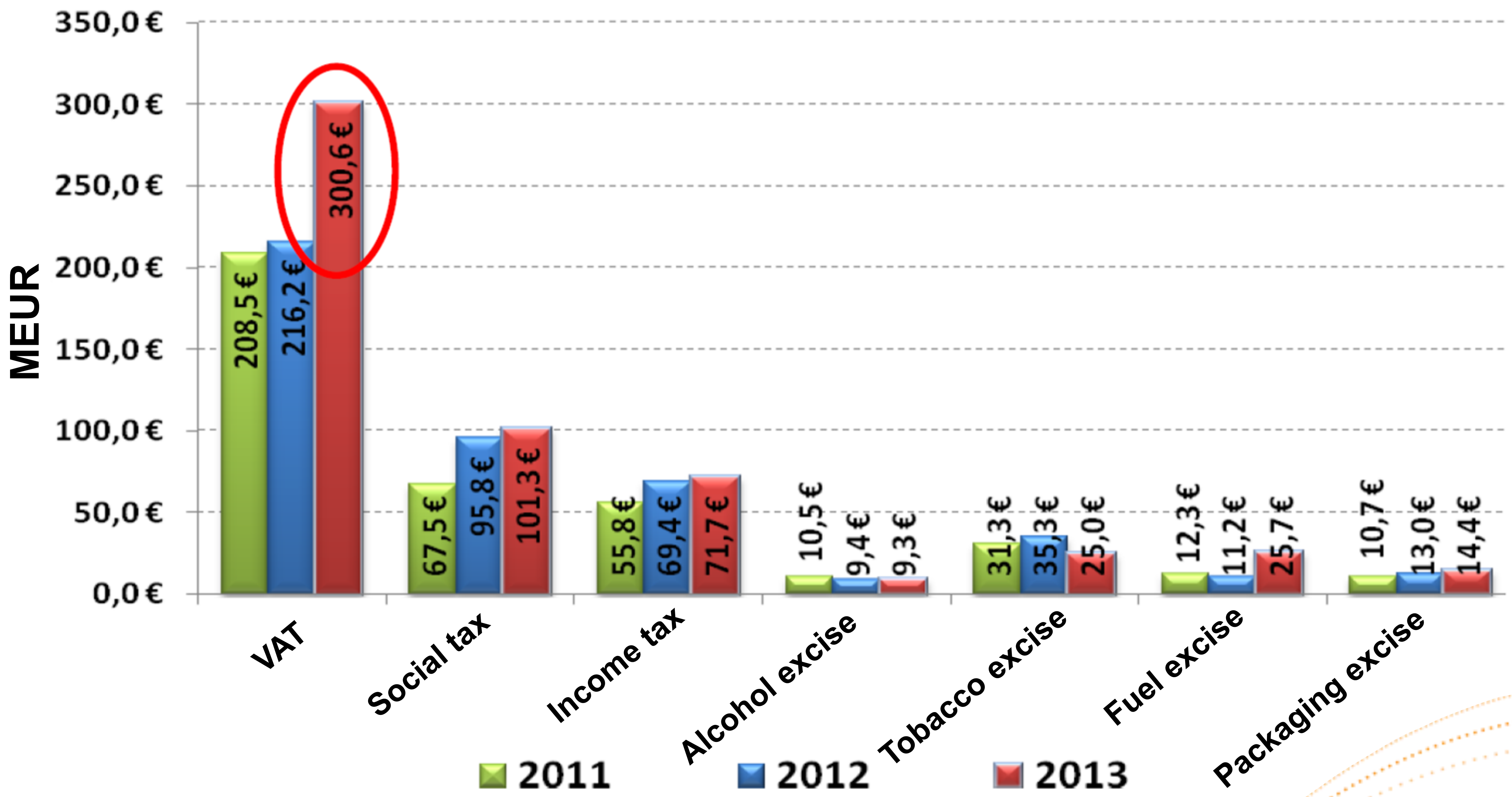
PRACTICE

sharemind

Demo!
A happy employee answering
a survey anonymously

Case study:
Tax fraud detection

Estimate of unpaid VAT

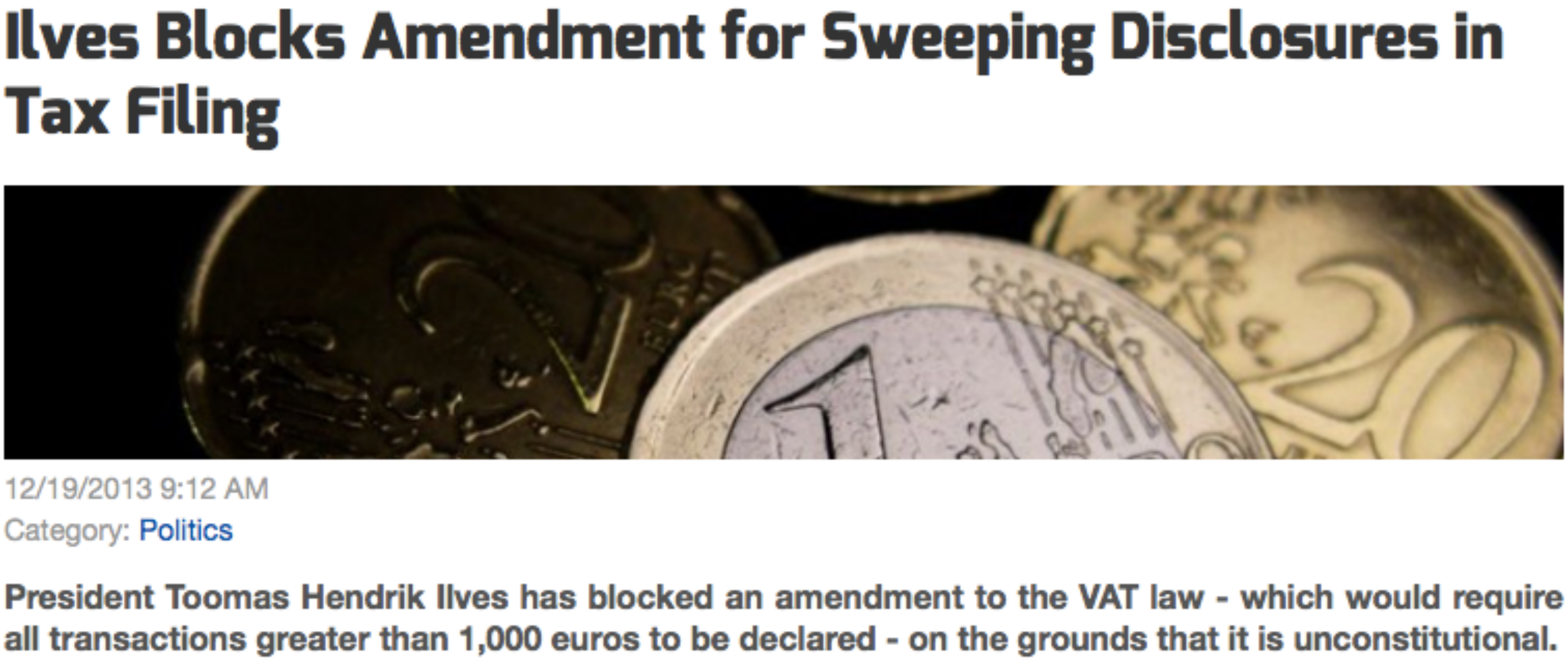


sharemind

Attempted fix to the gap

- In 2013, the Estonian parliament ratified the Value-Added Tax Act and the Accounting Act Amendment Act that would force enterprises to report transactions to the Tax and Customs Board (MTA).
- MTA would then match outgoing invoices to the incoming invoices reported by others and find companies trying to get refunds for fraudulently declared input VAT.

The story of the 1000 € law



Ilves Blocks Amendment for Sweeping Disclosures in Tax Filing

12/19/2013 9:12 AM

Category: Politics

President Toomas Hendrik Ilves has blocked an amendment to the VAT law - which would require all transactions greater than 1,000 euros to be declared - on the grounds that it is unconstitutional.

sharemind

Implementation using MPC

- The Tax Board was worried enough after the veto that they were willing to hear us out
- It also helped that Cybernetica was the company who won the tender to build the actual system.
- We agreed with the Tax Board that Cybernetica will build a research prototype that implements four risk analyses and will test its performance and that they will look at our results.
- We borrowed a systems analyst and an architect from our tax team to build the prototype.

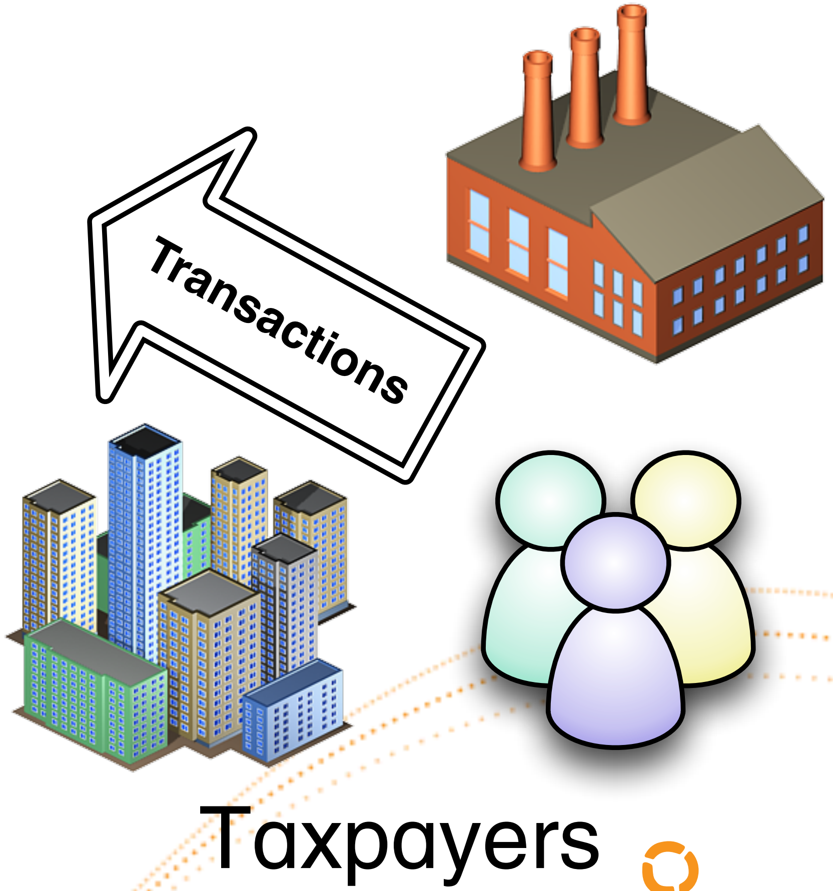
Secure implementation

<u>Benefits</u>

Encryption is applied on the data directly at the source.

The data is cryptographically protected during processing.

No need to unconditionally trust a single organization.



sharemind

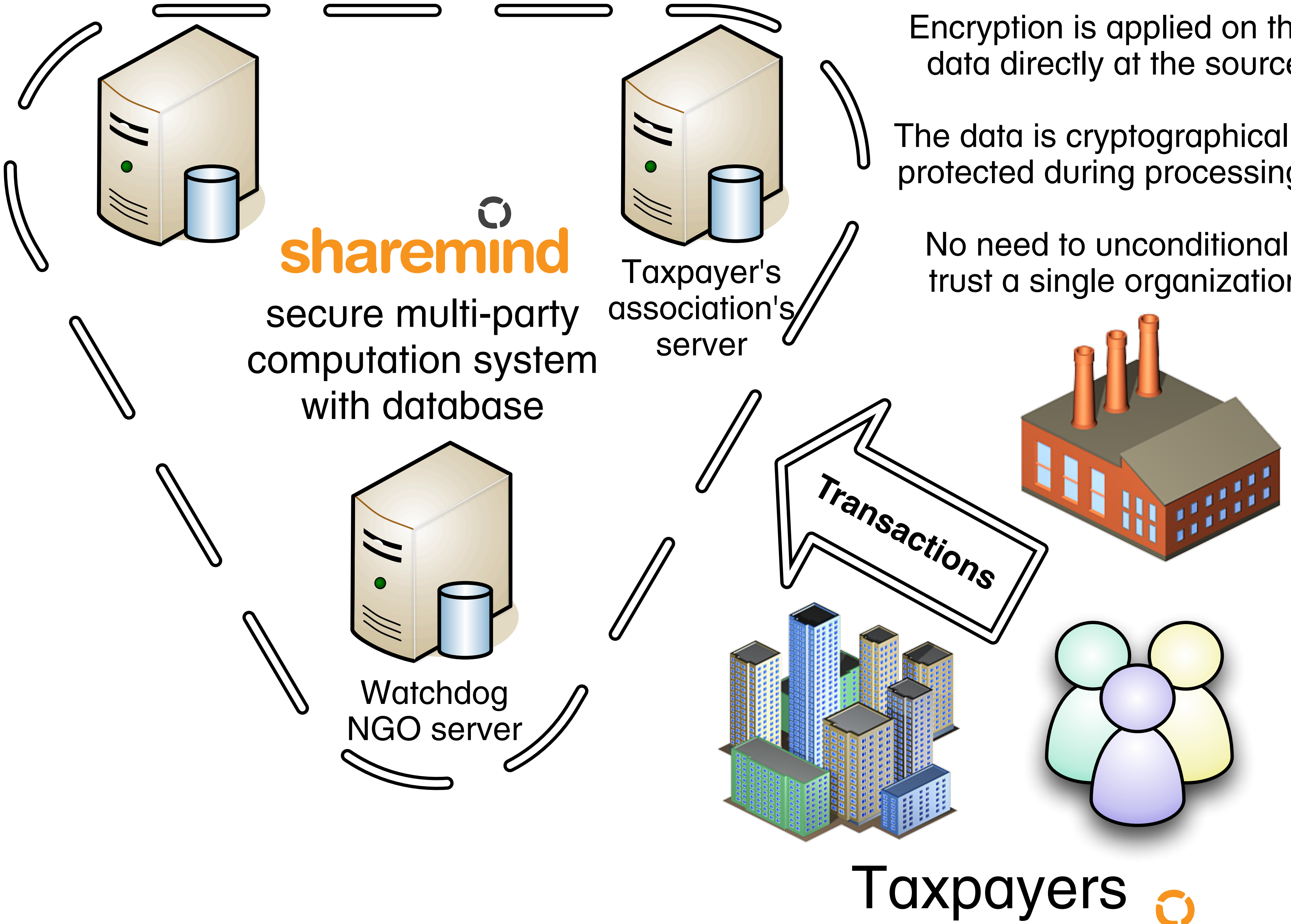
Secure implementation

Benefits

Encryption is applied on the data directly at the source.

The data is cryptographically protected during processing.

No need to unconditionally trust a single organization.


sharemind
secure multi-party computation system with database
Taxpayer's association's server
Watchdog NGO server
Transactions
Taxpayers

sharemind

Secure implementation

Benefits

Analyze, combine and build reports without decrypting data.

Confidentiality is guaranteed against all servers and against malicious hackers.

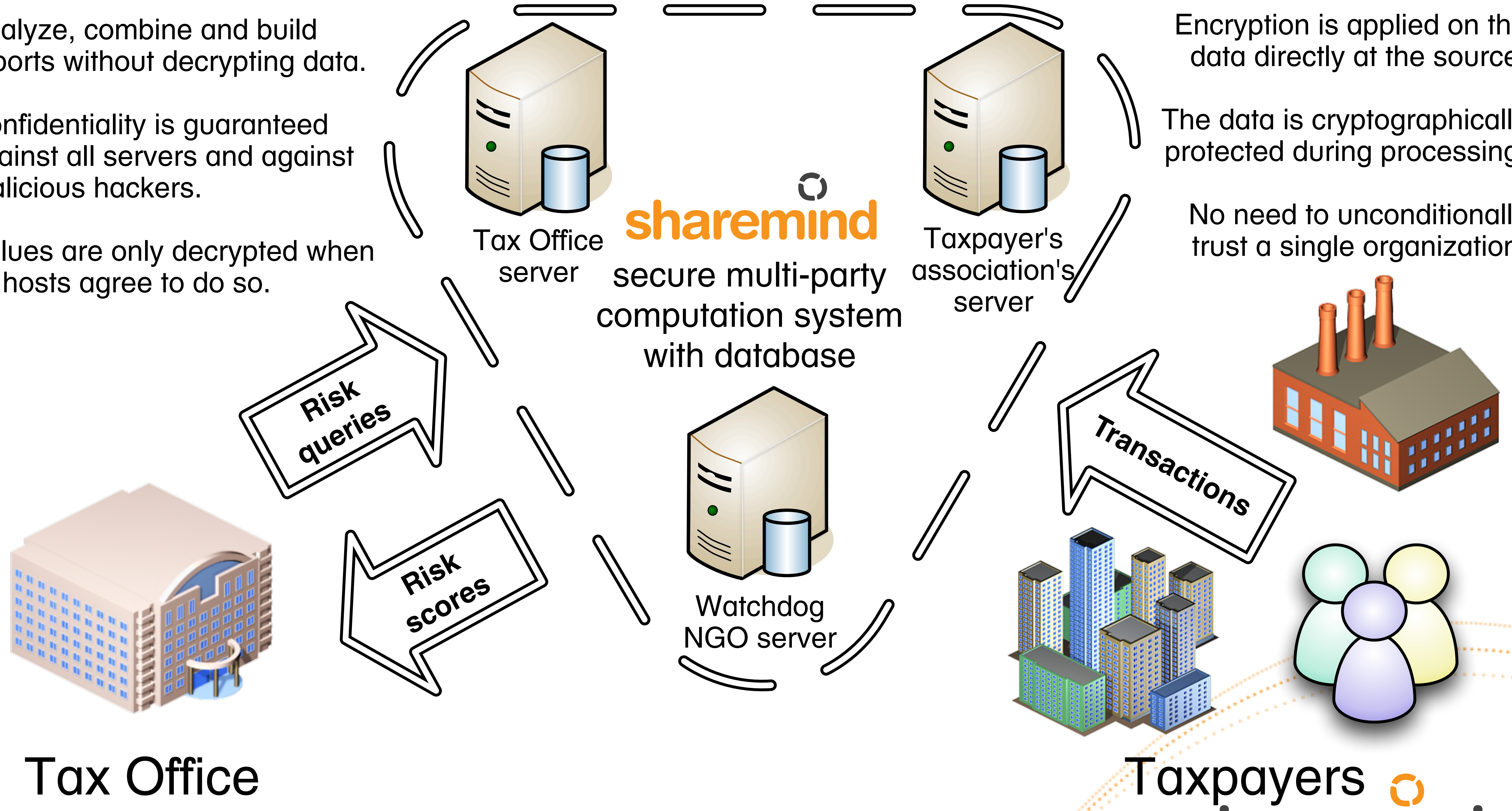
Values are only decrypted when all hosts agree to do so.

Benefits

Encryption is applied on the data directly at the source.

The data is cryptographically protected during processing.

No need to unconditionally trust a single organization.



Tax Office server

sharemind

secure multi-party computation system with database

Taxpayer's association's server

Watchdog NGO server

Risk queries

Risk scores

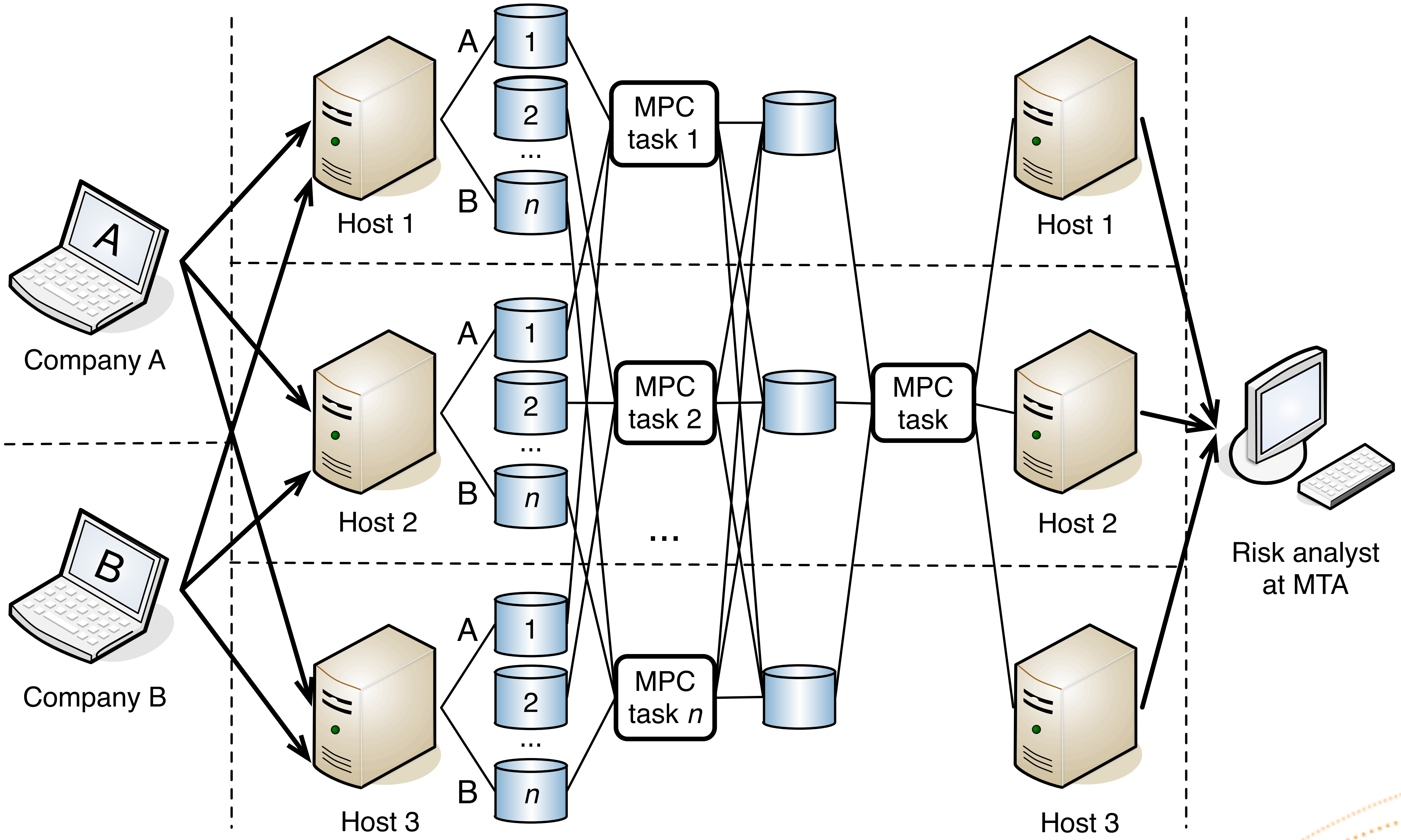
Transactions

Tax Office

Taxpayers

sharemind

Using fork-join parallelism



Secret sharing of transactions

Distribute inputs between tasks

Tasks aggregate transaction tables

Finalize aggregation and calculate scores

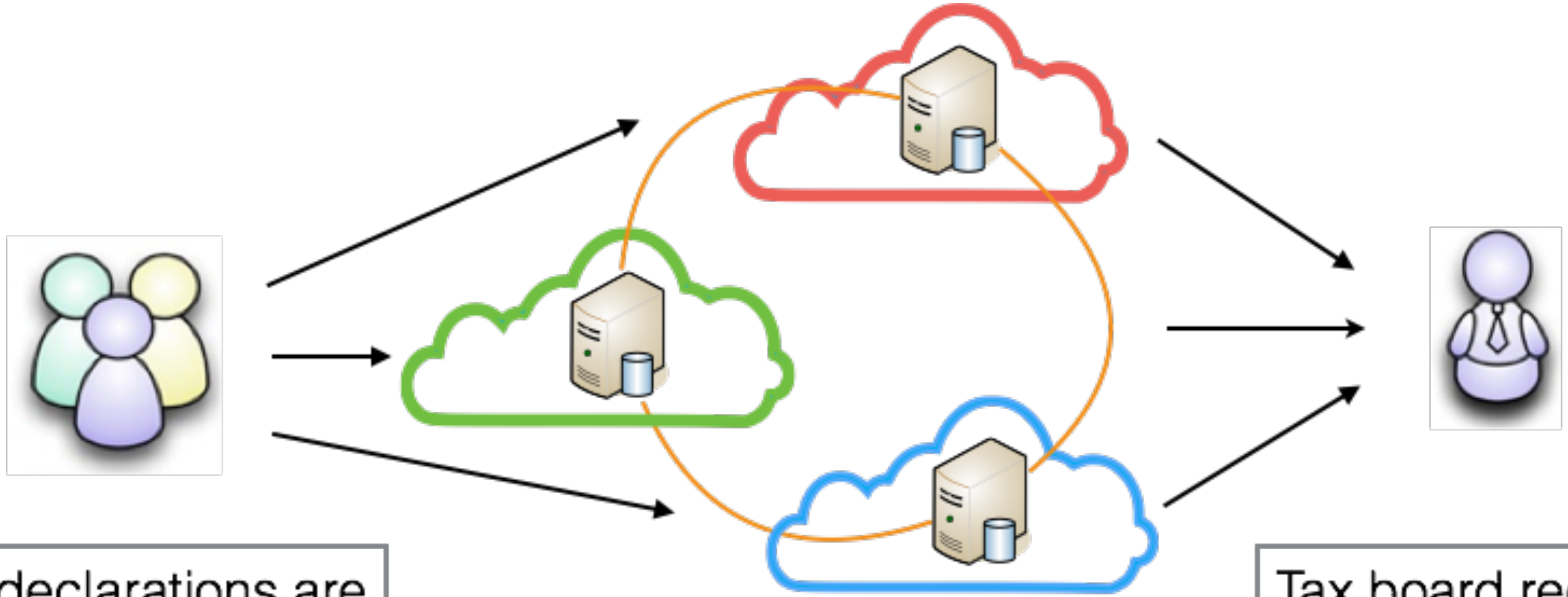
Send risk scores to analyst

- - - - organizational boundary

→ end-user communication with Sharemind

—— secure multi-party computation

Experiments on AWS cloud



Tax declarations are secret-shared

Servers in the cloud run secure computations

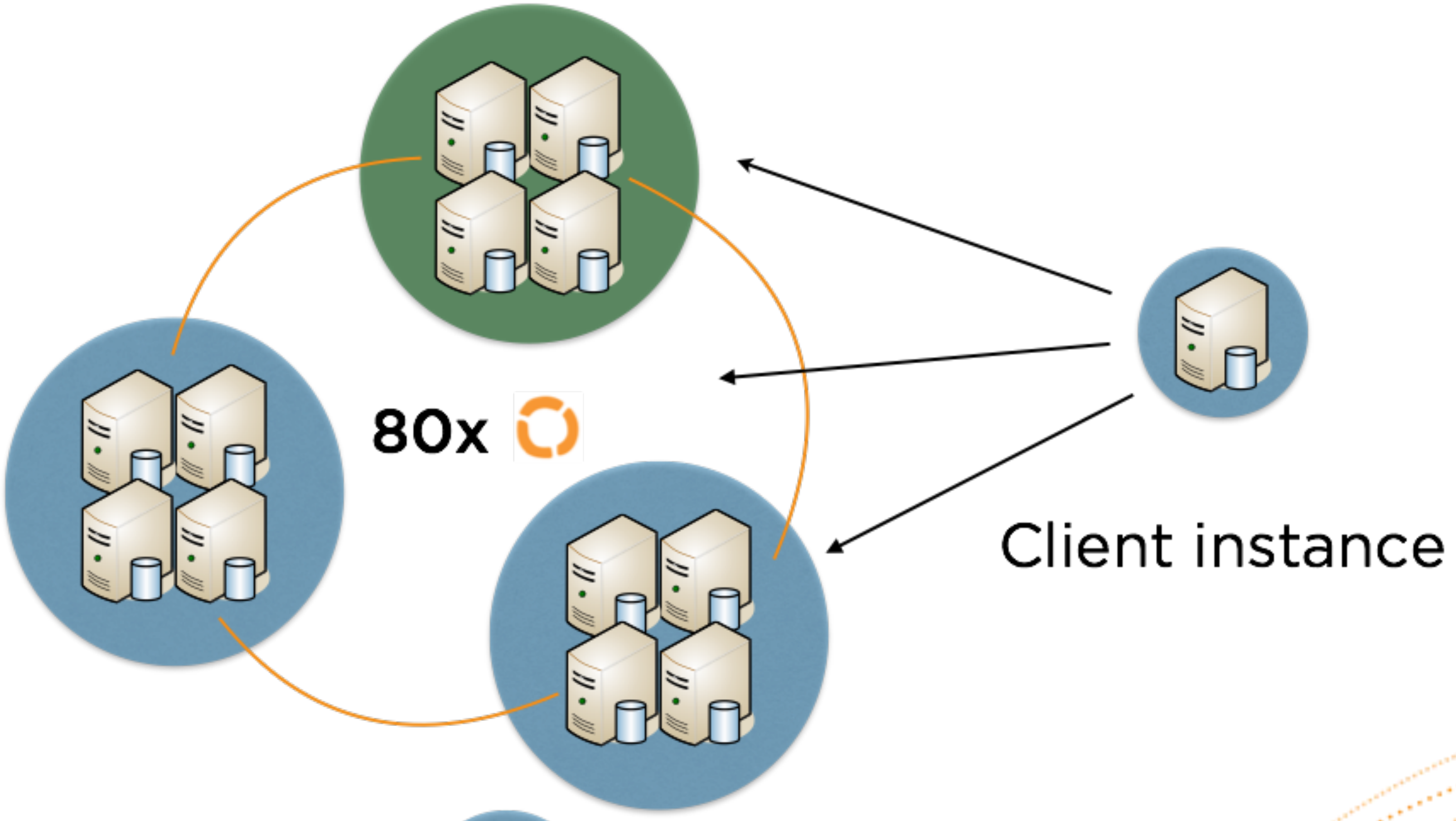
Tax board receives risk analysis results

Note: actual deployment should run on three different clouds.
However, we had a humble research grant from AWS.

sharemind

Much improved parallelism

12 computing nodes running
a total of 80 Sharemind processes



80x

Client instance

🟩 eu-central 🟦 eu-west

Computing environment

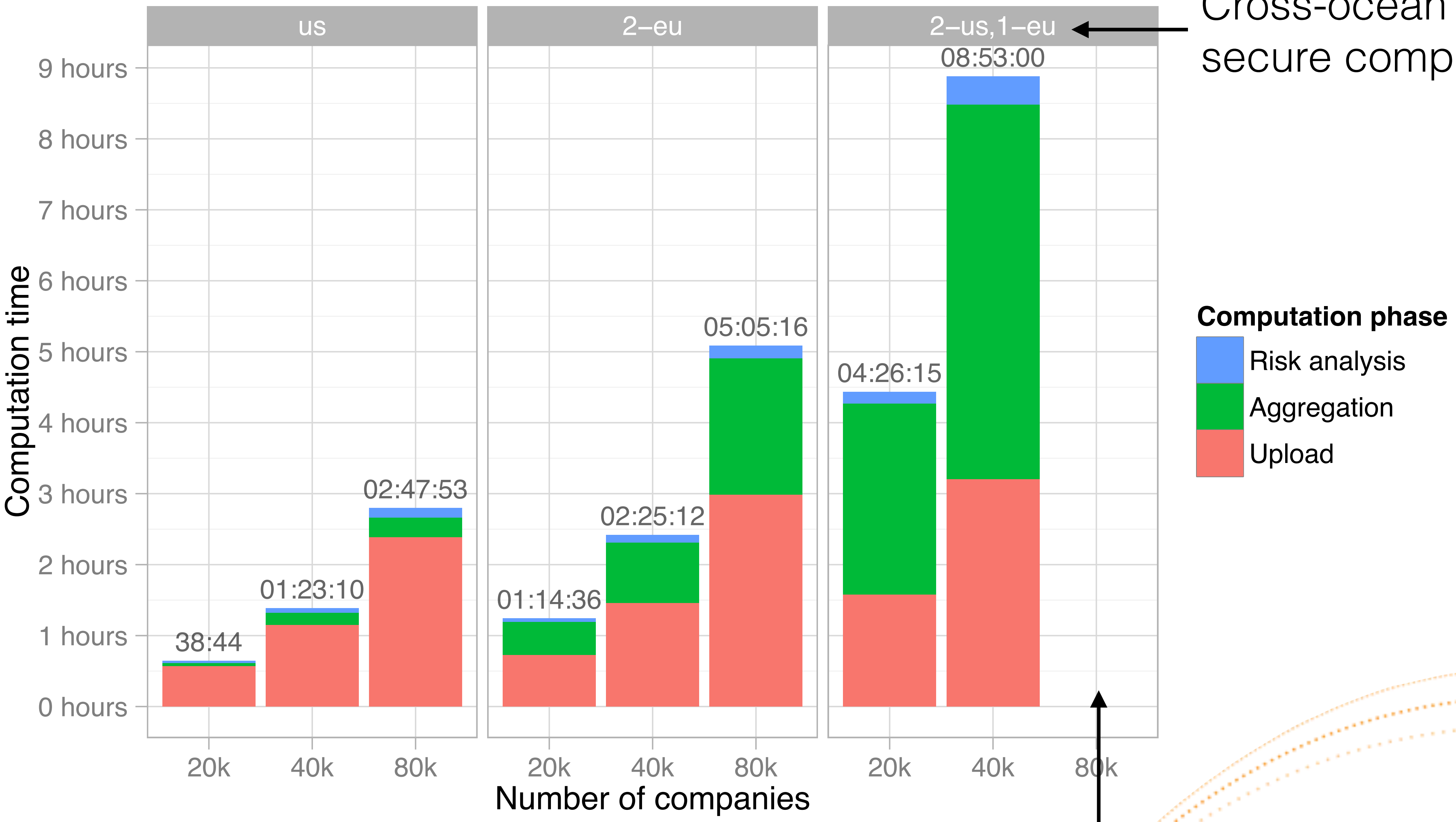
Setup	Client	Computing parties	Latency (round-trip)
1	us-east – c3.8xlarge	us-east – 12x c3.8xlarge	< 0.1ms between all nodes
2	eu-west – c3.8xlarge	eu-west – 8x c3.8xlarge eu-central – 4x c3.8xlarge	< 0.1ms inside eu-west 19ms (eu-west/eu-central)
3	us-east – c3.8xlarge	us-east – 4x c3.8xlarge us-west – 4x c3.8xlarge eu-west – 4x c3.8xlarge	77ms (us-east/us-west) 133ms (us-west/eu-west) 76ms (us-east/eu-west)

Realistic data sizes

No. of companies	No. of transaction partner pairs	Total no. of transactions
20 000	200 000	25 000 000
40 000	400 000	50 000 000
80 000	800 000	100 000 000

The source data for 100 000 000 transactions had a total size of 35 GB in XML format (about 1 GB in the secret-shared database).

Better running times

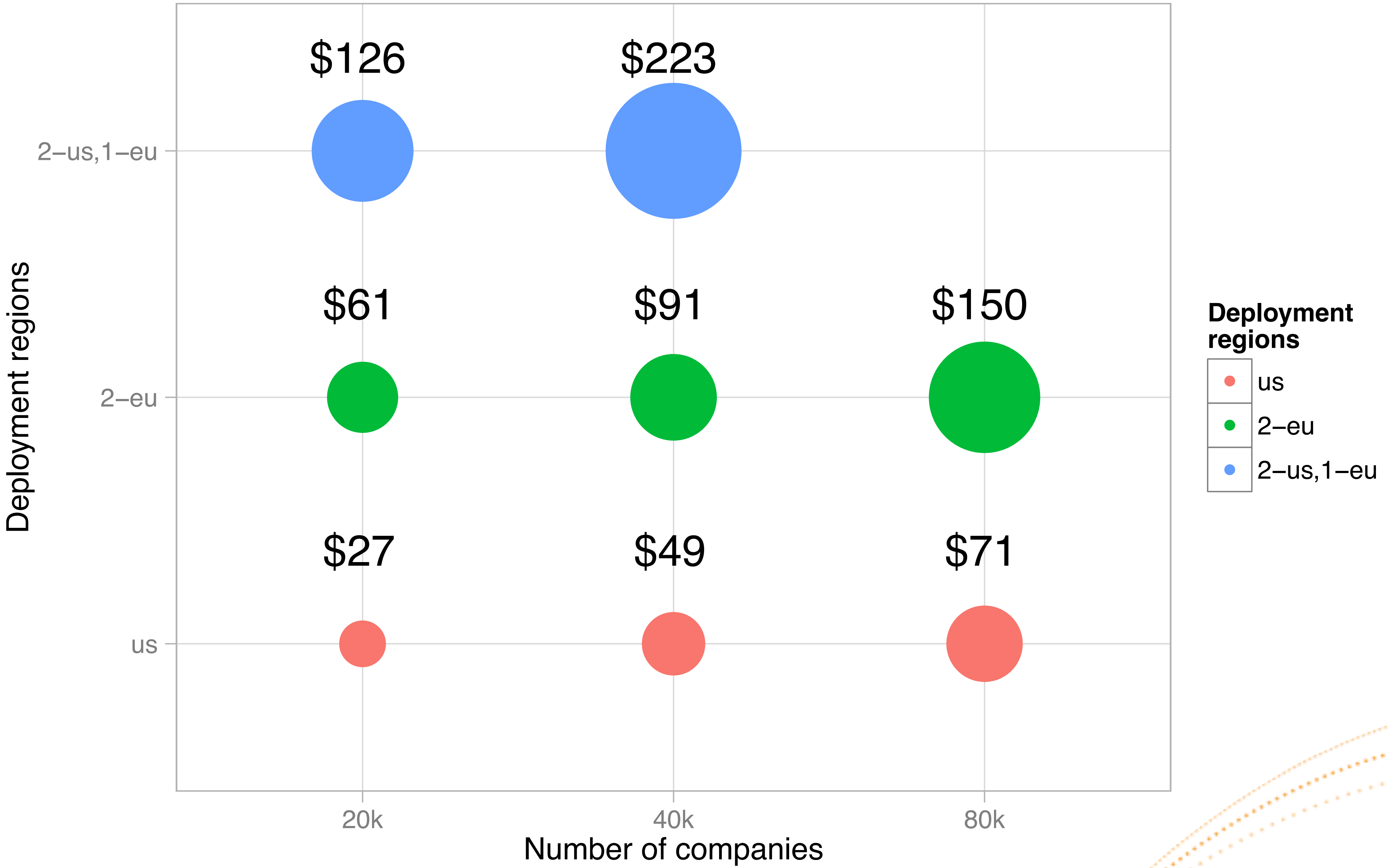


Cross-ocean secure computing!

Technical issues prevented the completion of this test and budgetary constraints did not allow for a repeat.

sharemind

Significantly lower price



sharemind

Conclusion

Our dream is to see MPC becoming an ubiquitous tool in applications where privacy is important

We can already demonstrate solving privacy issues for real-world users and organizations on a large scale

sharemind

We build applications

Learn about Sharemind and request an academic license

http://sharemind.cyber.ee/

Open source prototyping tools (under development)

http://sharemind-sdk.github.io/

Contact us for more information and collaborations

E-mail: sharemind@cyber.ee

Twitter: @sharemind

sharemind