

The logo for Devoxx, featuring the word 'DEVOXX' in a stylized, white, blocky font on a black background. The 'X's are orange and white. A small 'TM' trademark symbol is at the top left.

# Privacy in Practice with Smart Pseudonymization

## Lessons from the Belgian Public Sector

Kristof Verslype  
Cryptographer (PhD.)  
Smals Research

**Physical  
masks**

**Written  
pseudonyms**

**Digital  
pseudonyms**

# Innovation @ Smals Research

# Smart Pseudonymisation

Conversion from citizen identifiers to pseudonyms

## Format-Preserving Pseudonymisation

Retroactive protection of personal data in TEST & ACC of legacy applications



## eHealth Blind Pseudonymisation

Proactive protection of personal data in applications  
Privacy by Design



## Oblivious Join

Non-trivial join & pseudonymise projects for research purposes  
Distributed & no integration



# Format-Preserving Pseudonymisation

- Problem statement
- Concept
- Experimental service
- Conclusion





# Format-Preserving Pseudonymisation

- **Problem statement**
- Concept
- Experimental service
- Conclusion



# Widespread use of personal data in non-prod environments

*“60% of organisations use raw production data in test environments”*

World Quality Report, 2020

# Security

## Data breaches from non-prod environments

2016

# UBER

Hacker exploited Uber's software development environments to break into the rideshare giant's cloud storage

2021

# T Mobile™

Hacker leveraged an unprotected router to gain access to T-Mobile's production, staging, and development servers, which compromised over 48 million social security numbers and other details.

2022

# LastPass... |

The hacker targeted the home computer of a LastPass senior DevOps engineer

**No negligible risk!**

# Compliance with GDPR

## Personal data in TEST/ACC

### ❖ *Legal basis*

- Informed and actively given consent?
- Legitimate interest (gerechtvaardigd belang) of organisation?
- Special categories of personal data  
Minors, medical data, sexual orientation, criminal data, ...
- Other legal basis?

### ❖ *Appropriate measures*

- Security TEST < PROD/ACC

## Pseudonymisation

- ❖ Encouraged by GDPR to protect personal data
- ❖ Some rules by GDPR more **relaxed**
- ❖ Could help become more compliant

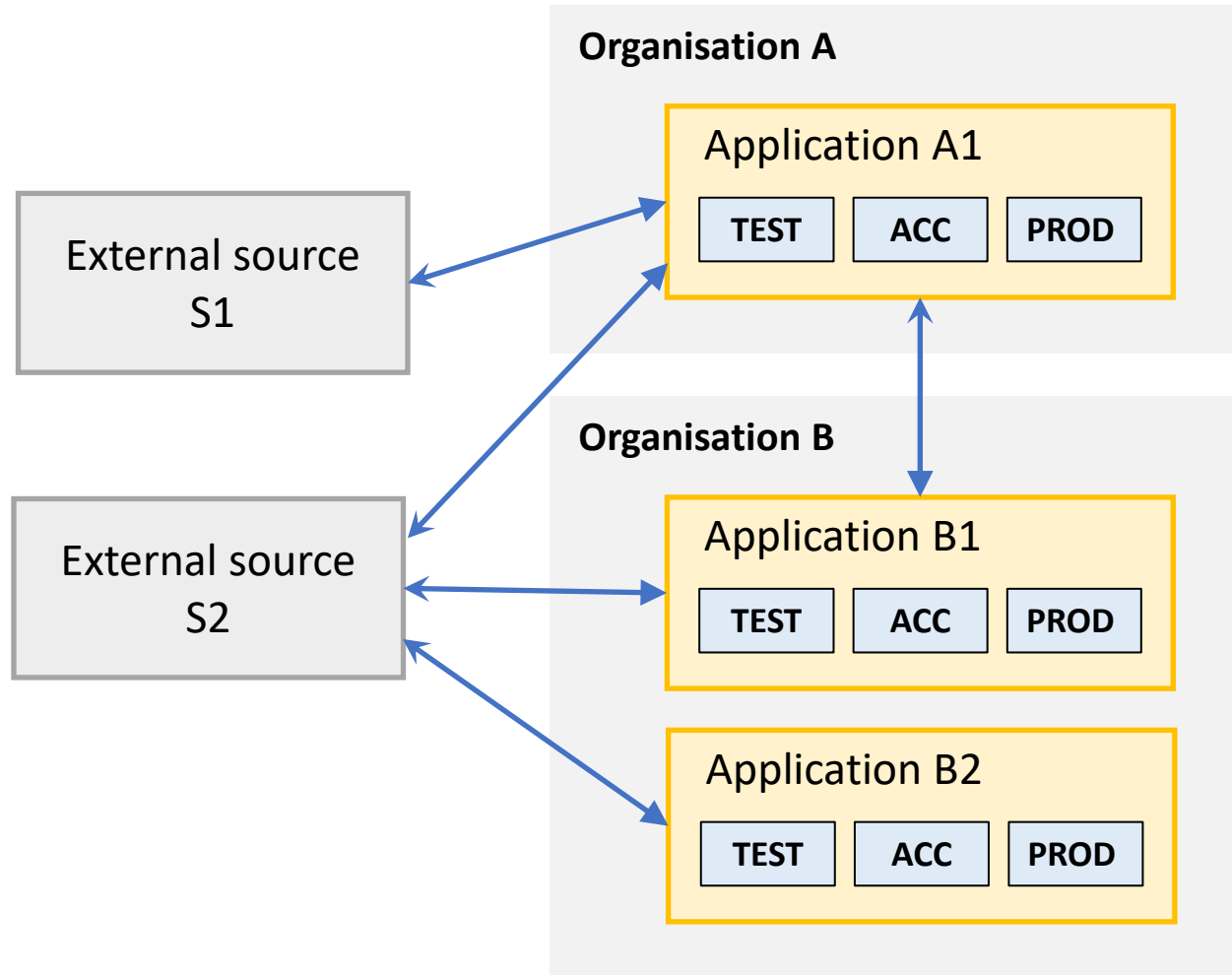
### **GDPR, Art 32.**

[...] *the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:*

- the **pseudonymisation** and encryption of personal data;*
- [...]



# Reality in public sector



Question customer

**How to improve privacy  
in TEST & ACC?**

Completely fictional data **not** an option

**Because**

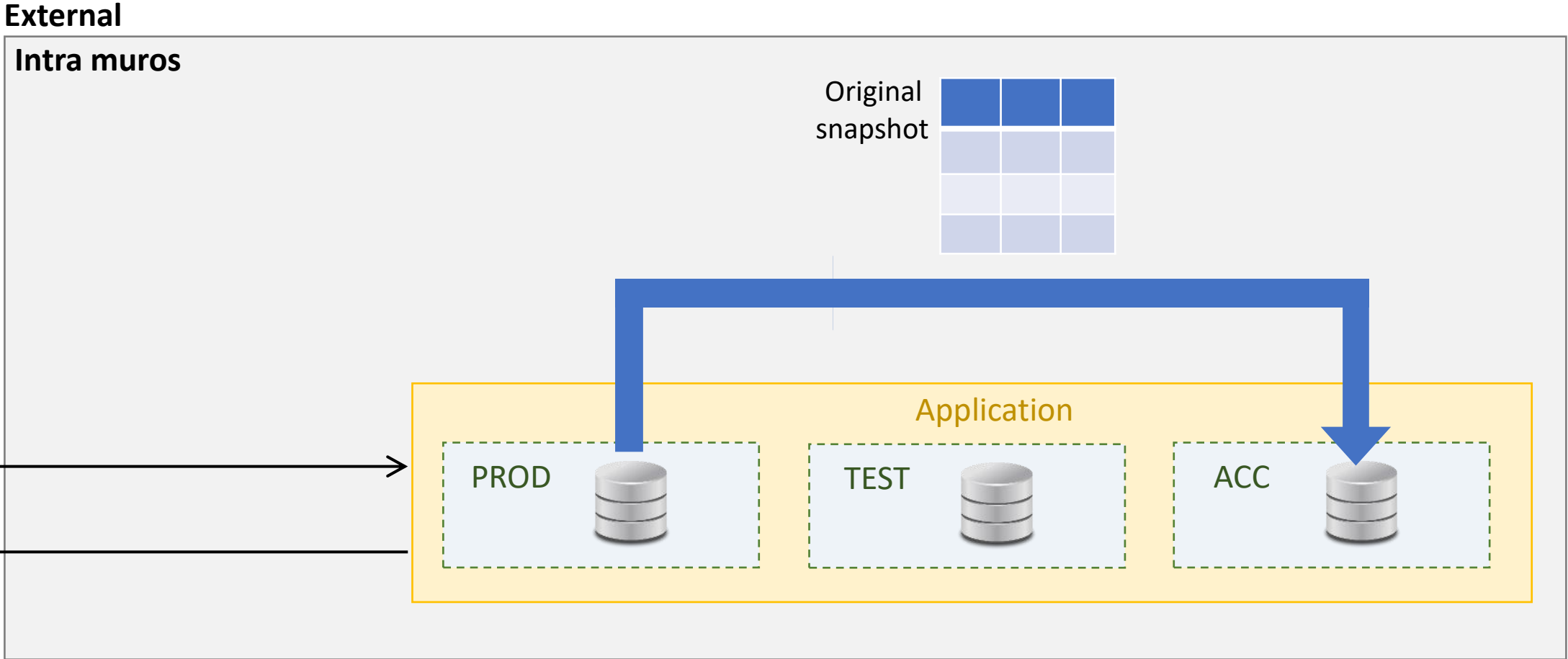
- Exchange personal data
- We would miss edge cases
- Complex data models
- Labor intensive = expensive

# Format-Preserving Pseudonymisation

- Problem statement
- **Concept**
- Experimental service
- Conclusion



# Current practice



External world

# Approach by member

Transforming batch of records with personal data copied to TEST or ACC

Structured identifiers	Unstructured identifiers		Domain-specific data		
Identifier	First name	Surname	...	...	...
18.32.08-903.41	Kasper	de Brouckère	A1	A2	A3
30.02.06-981.94	Melchior	Rogier	B1	B2	B3
72.43.27-109.21	Baltazar	Beernaert	C1	C2	C3

Pseudonymise      Shuffle      Shuffle

Identifier	First name	Surname
30.03.30-213.23	Melchior	Beernaert
66.08.15-286.27	Baltazar	de Brouckère
22.51.14-602.20	Kasper	Rogier

## 1. Pseudonymise

Replace structured identifier by format-preserving pseudonym

- Bidirectional
- By Smals Research

## 2. Shuffle

Column-wise permutation

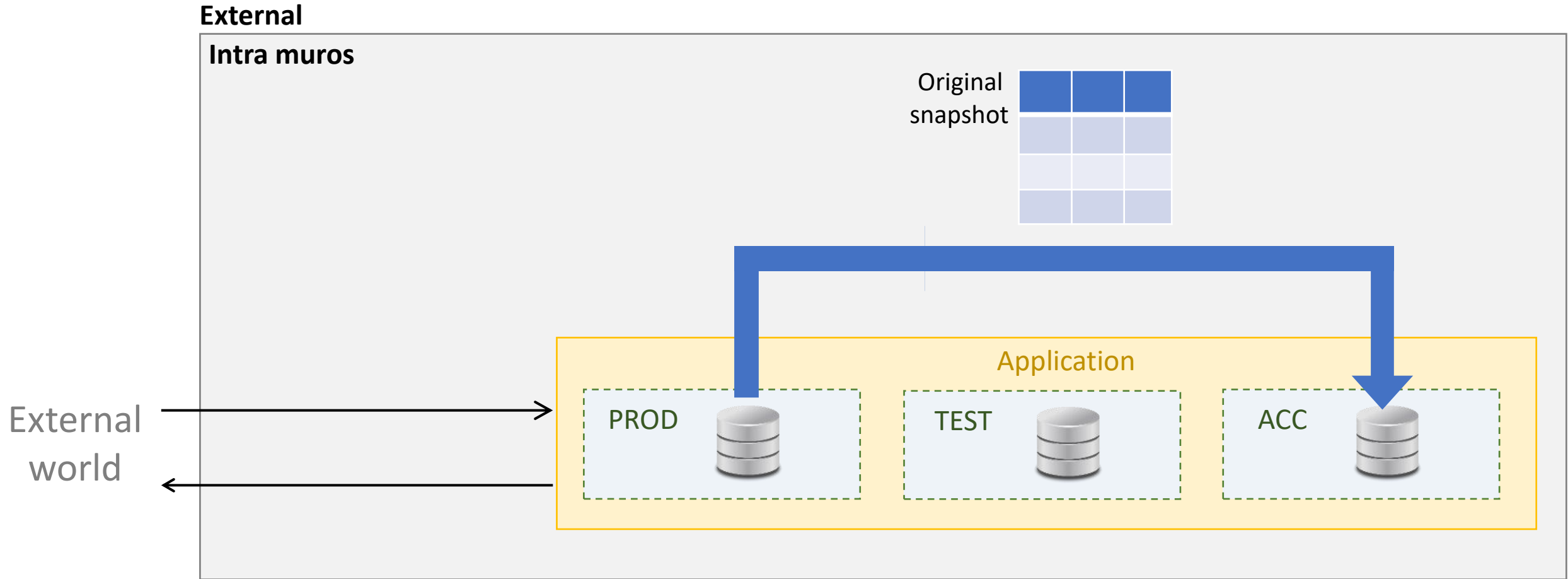
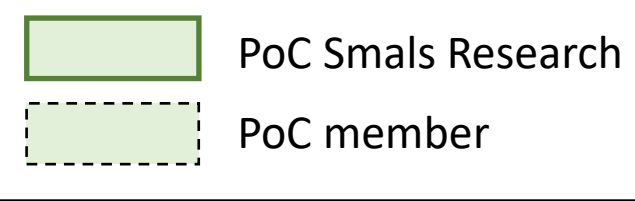
- Unidirectional
- By Customer

## Transformed snapshot

Identifier	First name	Surname	...	...	...
30.03.30-213.23	Melchior	Beernaert	A1	A2	A3
66.08.15-286.27	Baltazar	de Brouckère	B1	B2	B3
22.51.14-602.20	Kasper	Rogier	C1	C2	C3

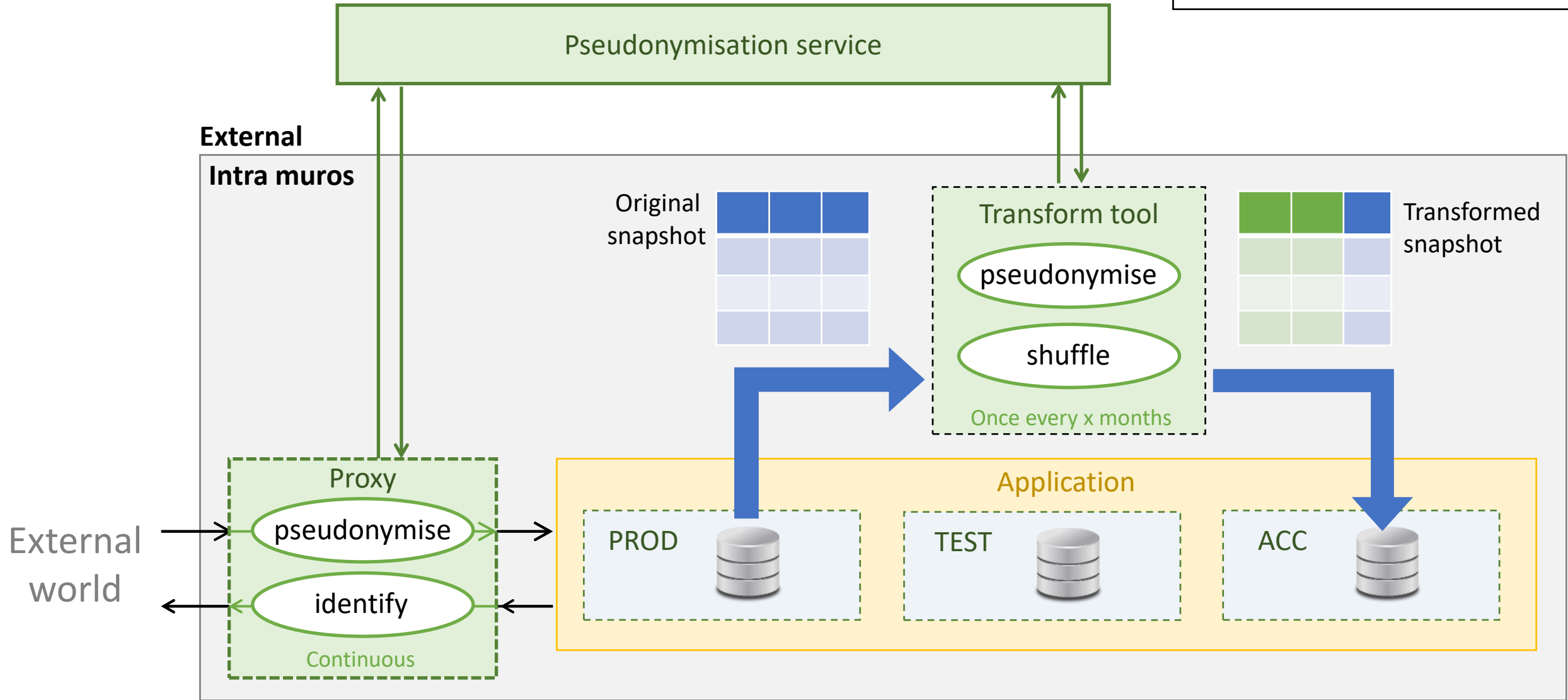
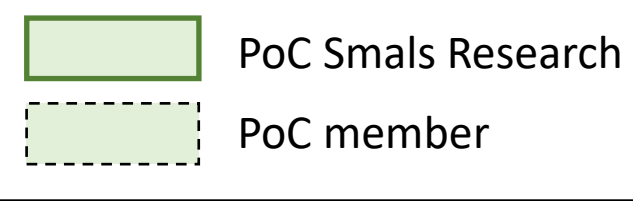
Records useful for TEST & ACC, while hard to identify!

# PoC in collaboration with customer

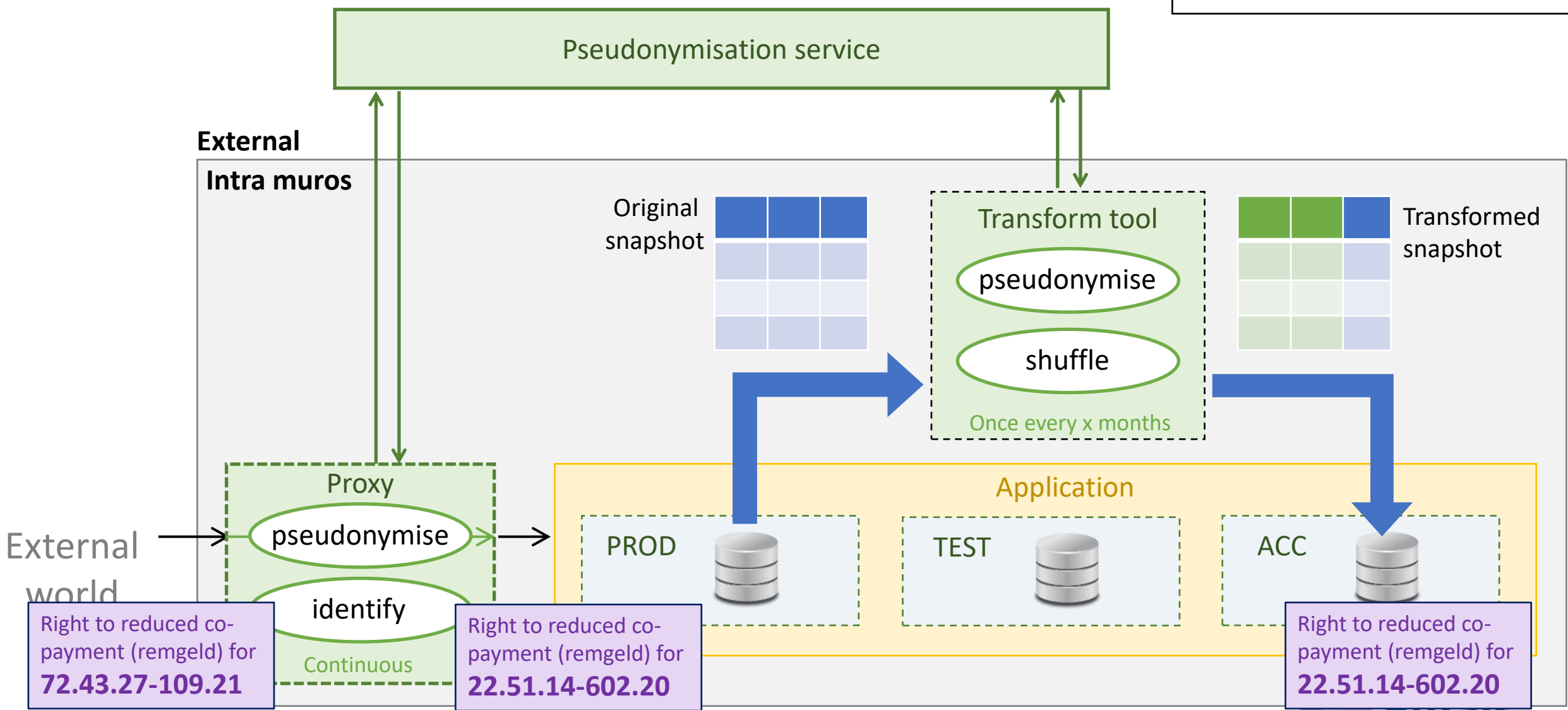
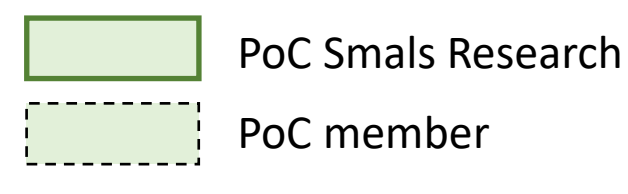




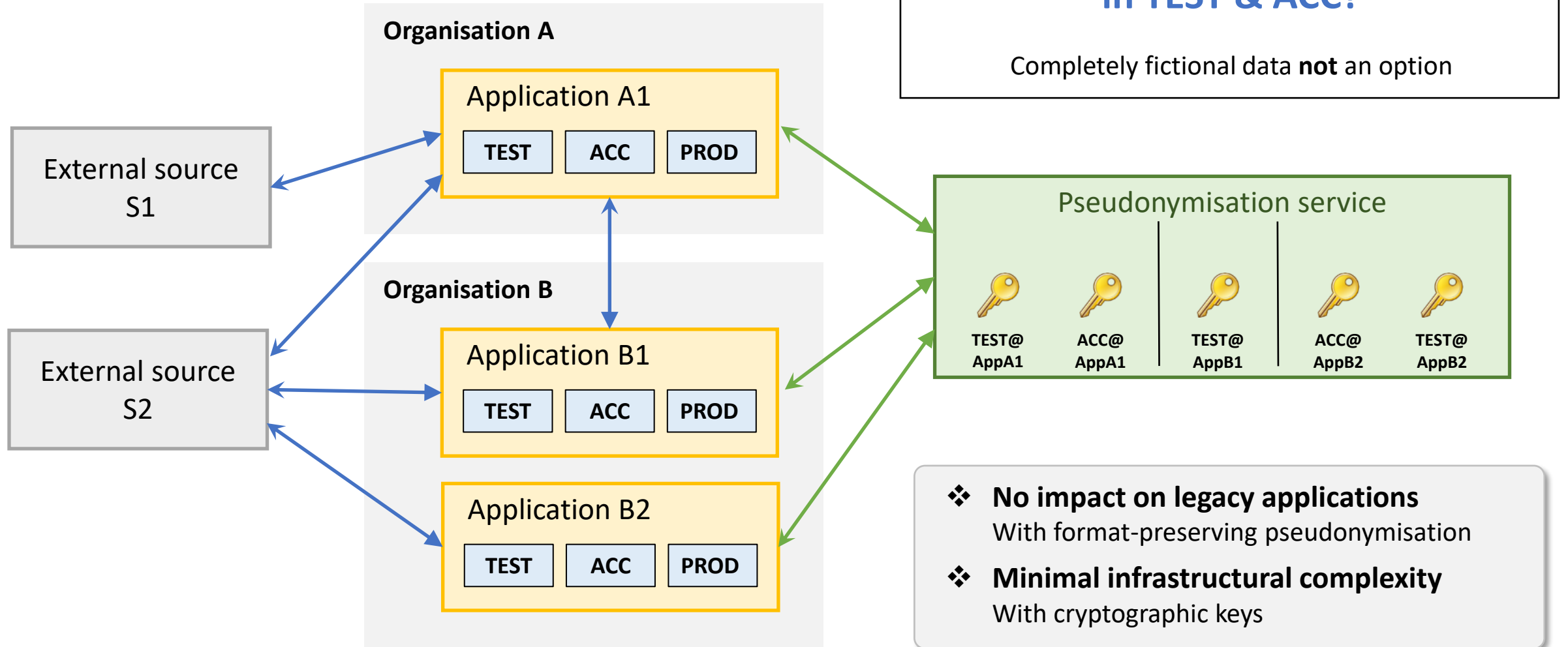
# PoC in collaboration with customer



# PoC in collaboration with customer

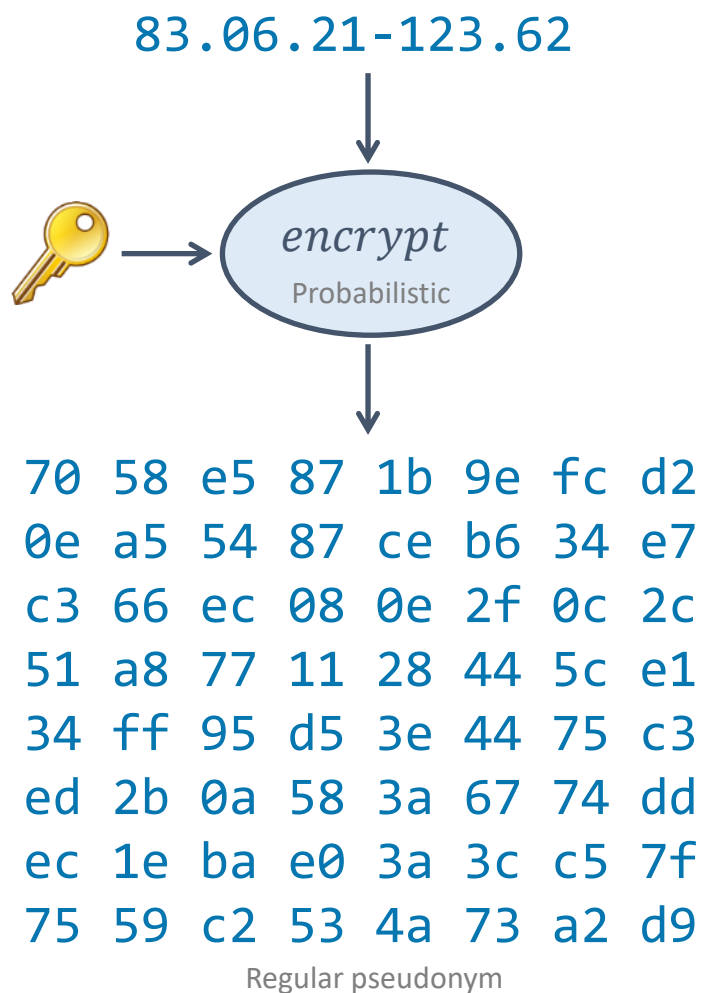


# Reality in public sector

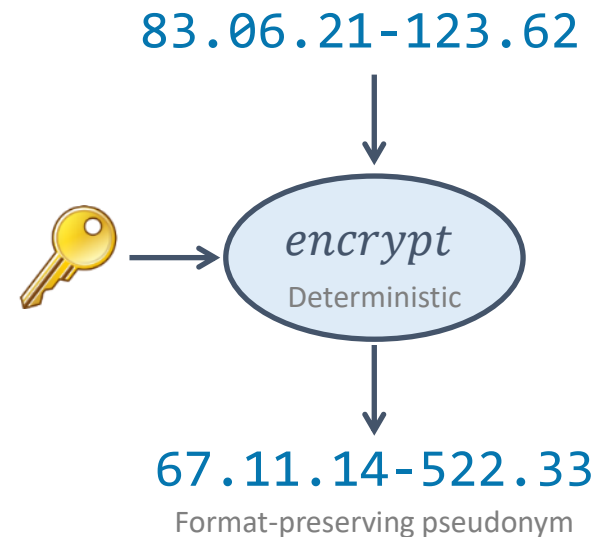


# Encryption

## TRADITIONAL ENCRYPTION



## FORMAT-PRESERVING ENCRYPTION



- ❖ Conversions happen on-the-fly
- ❖ Structure preserved, including valid checksum
- ❖ Described in NIST SP 800-38G Revision. 1 (2019)

# Format-Preserving Pseudonymisation

- Problem statement
- Concept
- **Experimental service**
- Conclusion





# Experimental REST service

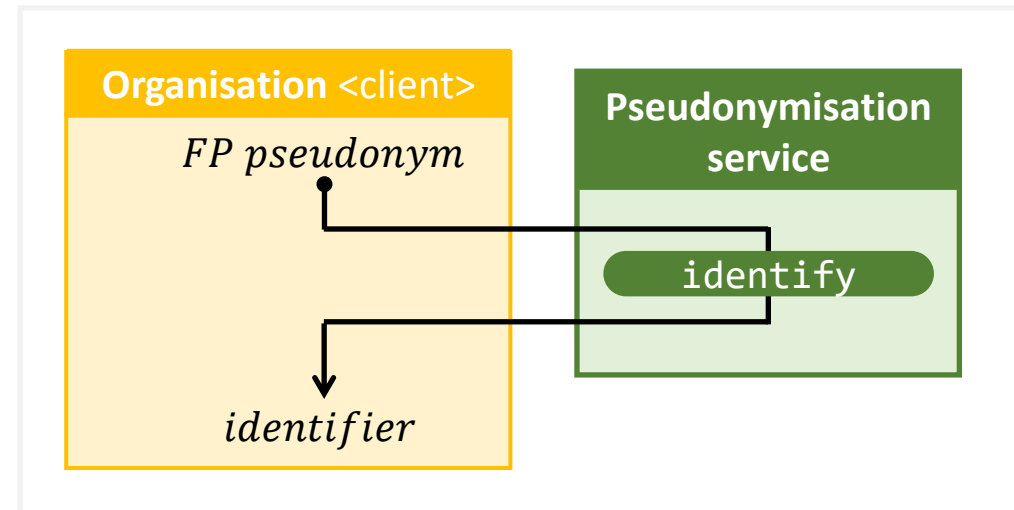
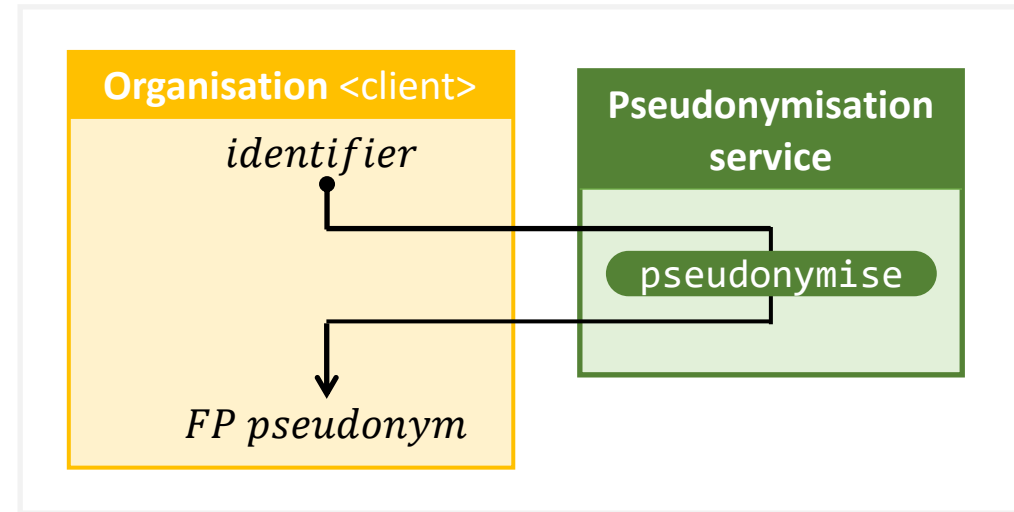
Built by Smals Research

## Rest API

- ✓ Pseudonymise & Identify
- ✓ GET and POST
- ✓ Also batch (POST only)

## Identifiers

- ✓ Support for Belgian social security numbers
- ✓ Extensible



# POST Request

```
1 {
2   "context": {
3     "security-group": "ehealth",
4     "application": "quatro",
5     "environment": "TEST"
6   },
7   "identifiers": [
8     "18.32.08-902.42",
9     "30.02.06-981.94",
10    "72.43.27-109.21",
11    "58.28.16-291.62",
12    "58.28.16-29X.61",
13    "58.28.16-291.90",
14    "79.27.28-621.96",
15    "30.43.04-205.53",
16    "93.26.17-802.47",
17    "33.24.16-568.07"
18  ]
19 }
```

- ✓ Easy to use
- ✓ Graceful error handling
- ✓ Efficient



# POST Response

```
1 {
2   "context": {
3     "security-group": "ehealth",
4     "application": "quatro",
5     "environment": "TEST"
6   },
7   "time": "2024-01-08T08:20:39.128207895Z",
8   "translation-info": {
9     "action": "pseudonymize",
10    "enabled": true
11  },
12  "translations": [
13    {
14      "identifier": "18.32.08-902.42",
15      "pseudonym": "30.43.30-213.41",
16      "valid": true
17    },
18    {
19      "identifier": "30.02.06-981.94",
20      "pseudonym": "66.08.15-286.27",
21      "valid": true
22    },
23    {
24      "identifier": "72.43.27-109.21",
25      "pseudonym": "22.51.14-602.20",
26      "valid": true
27    },
28    {
29      "identifier": "58.28.16-291.62",
30      "pseudonym": "null",
31      "valid": false,
32      "error": "checksum"
33    }
34  ]
35 }
```

# Format-Preserving Pseudonymisation

- Problem statement
- Concept
- Experimental service
- **Conclusion**



# Format-Preserving Pseudonymisation

## Building block

- To improve privacy in TEST and ACC environments
- Partial solution

## As a Service

- Simplifies logic organisation  
E.g. key management
- Stimulates reuse
- Separation of duties

## Status

Trying to go into project mode





# Innovation @ Smals Research

# Smart Pseudonymisation

Conversion from citizen identifiers to pseudonyms

## Format-Preserving Pseudonymisation

Retroactive protection of personal data in TEST & ACC of legacy applications



## eHealth Blind Pseudonymisation

Proactive protection of personal data in applications  
Privacy by Design



## Oblivious Join

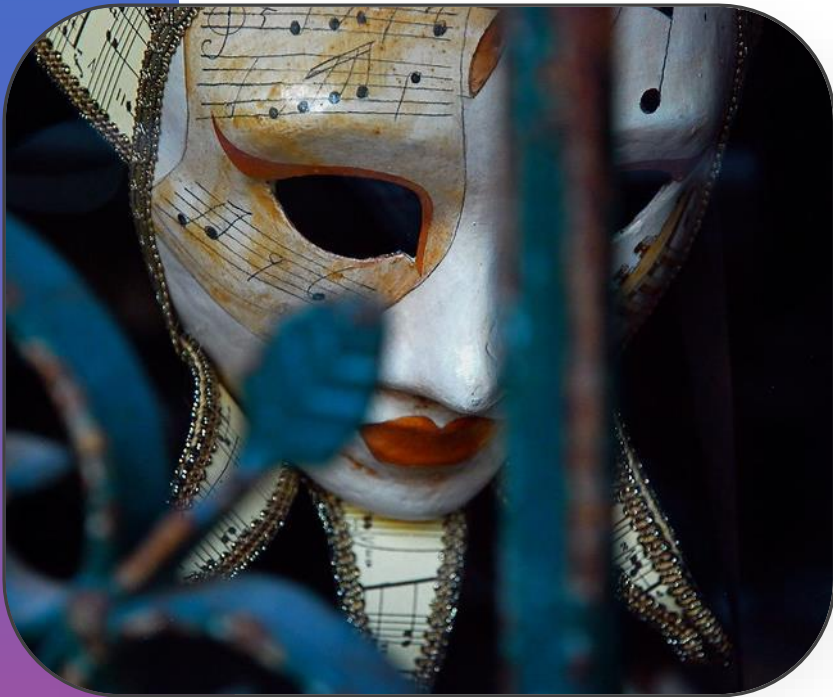
Non-trivial join & pseudonymise projects for research purposes  
Distributed & no integration





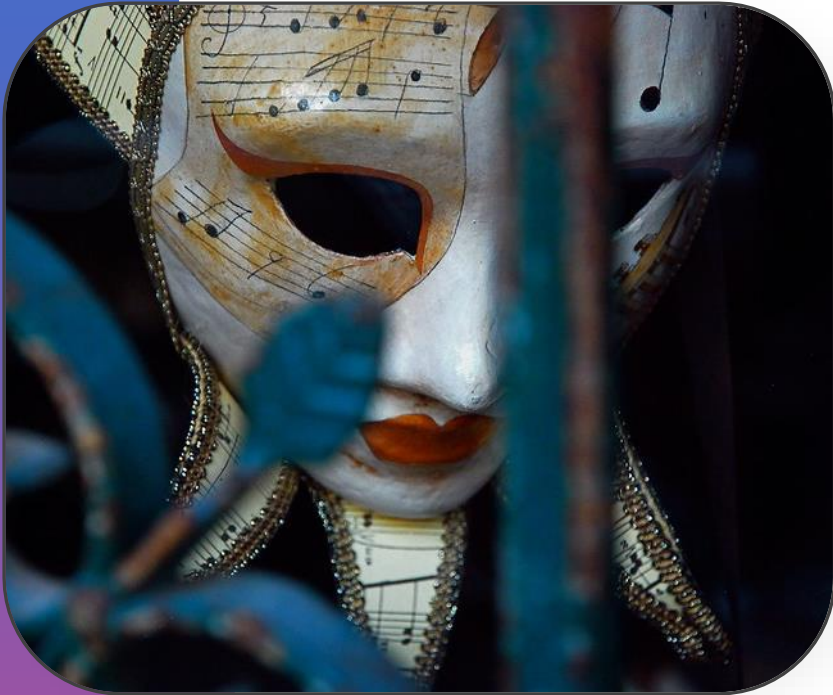
# eHealth Blind Pseudonymisation

- Problem statement
- Referral prescriptions
- Join & pseudonymise data for research
- Conclusion



# eHealth Blind Pseudonymisation

- **Problem statement**
- Referral prescriptions
- Join & pseudonymise data for research
- Conclusion





# Design principles



## Privacy by design

Privacy should be taken into account when designing and building products and services



## Separation of duties

Entity managing protection keys should not have access to protected data (and vice versa)



## Simplicity

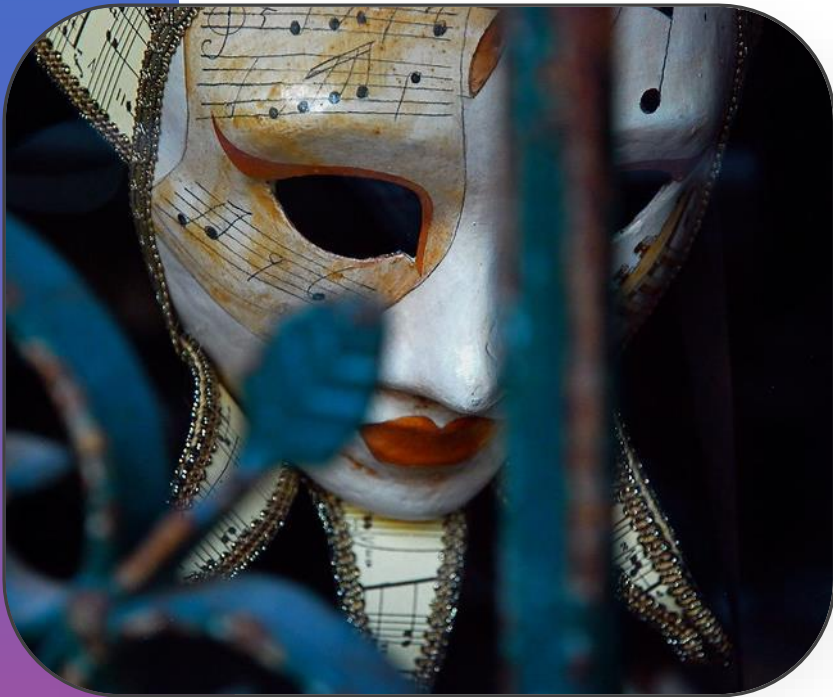
Complexity is the worst enemy of security





# eHealth Blind Pseudonymisation

- Problem statement
- **Referral prescriptions**
- Join & pseudonymise data for research
- Conclusion



## Use case 1 - Live

# Referral prescription = Verwijsvoorschrift / Prescription de renvoi

### What?

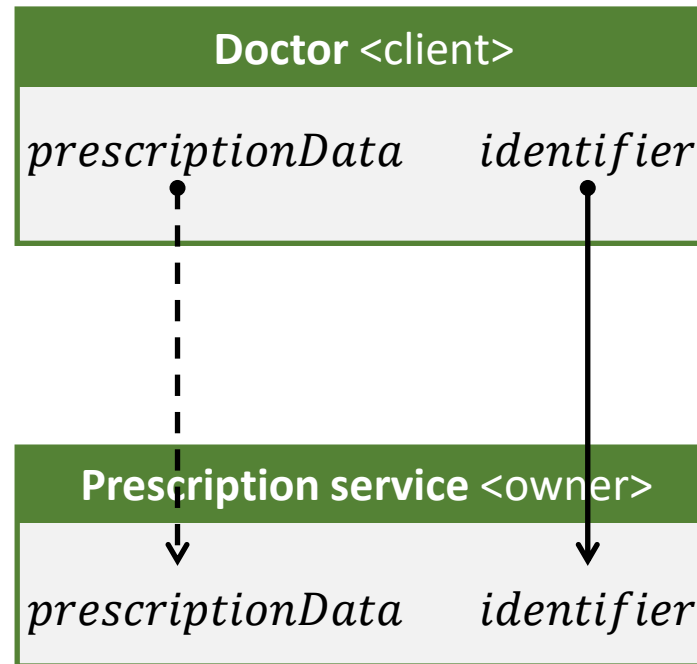
A certificate to start a certain treatment (e.g. physiotherapist, dieticians, speech therapists).

### Requirements

- ❖ **Pseudonymisation**  
Prescription service should never be able to link prescription data to a citizen
- ❖ **Partial encryption**  
The prescription service should not be able to access certain fields

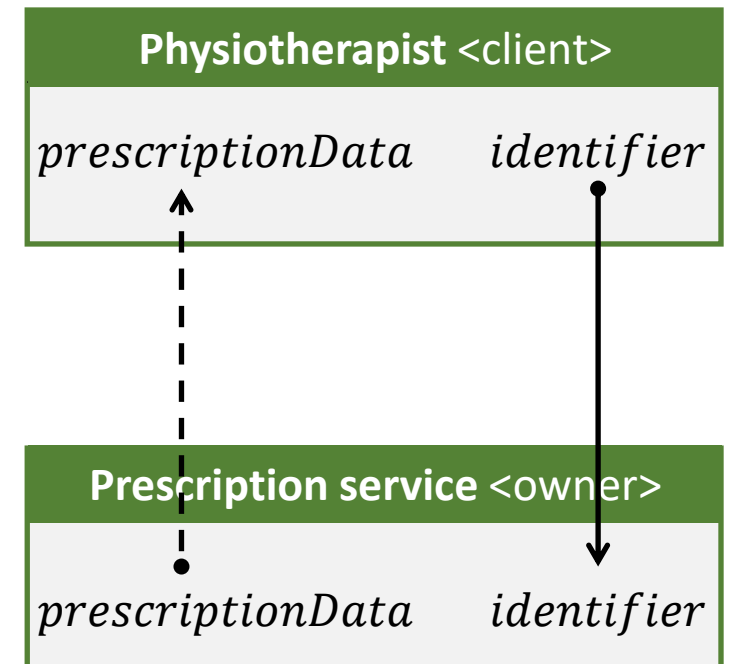
### Scenario 1

Doctor (client) requests Prescription service (owner) to register prescription



### Scenario 2

Physiotherapist (client) requests access to prescription for a specific citizen from Prescription service (owner)



# Blind Pseudo Service Pseudonymise

✓ **Each party only sees only what it needs to see**

- ❖ Client only sees identifiers
  - ❖ Owner only sees pseudonyms
  - ❖ Pseudon. service sees neither
- **Maximizes security & privacy**

✓ **Direct communication**

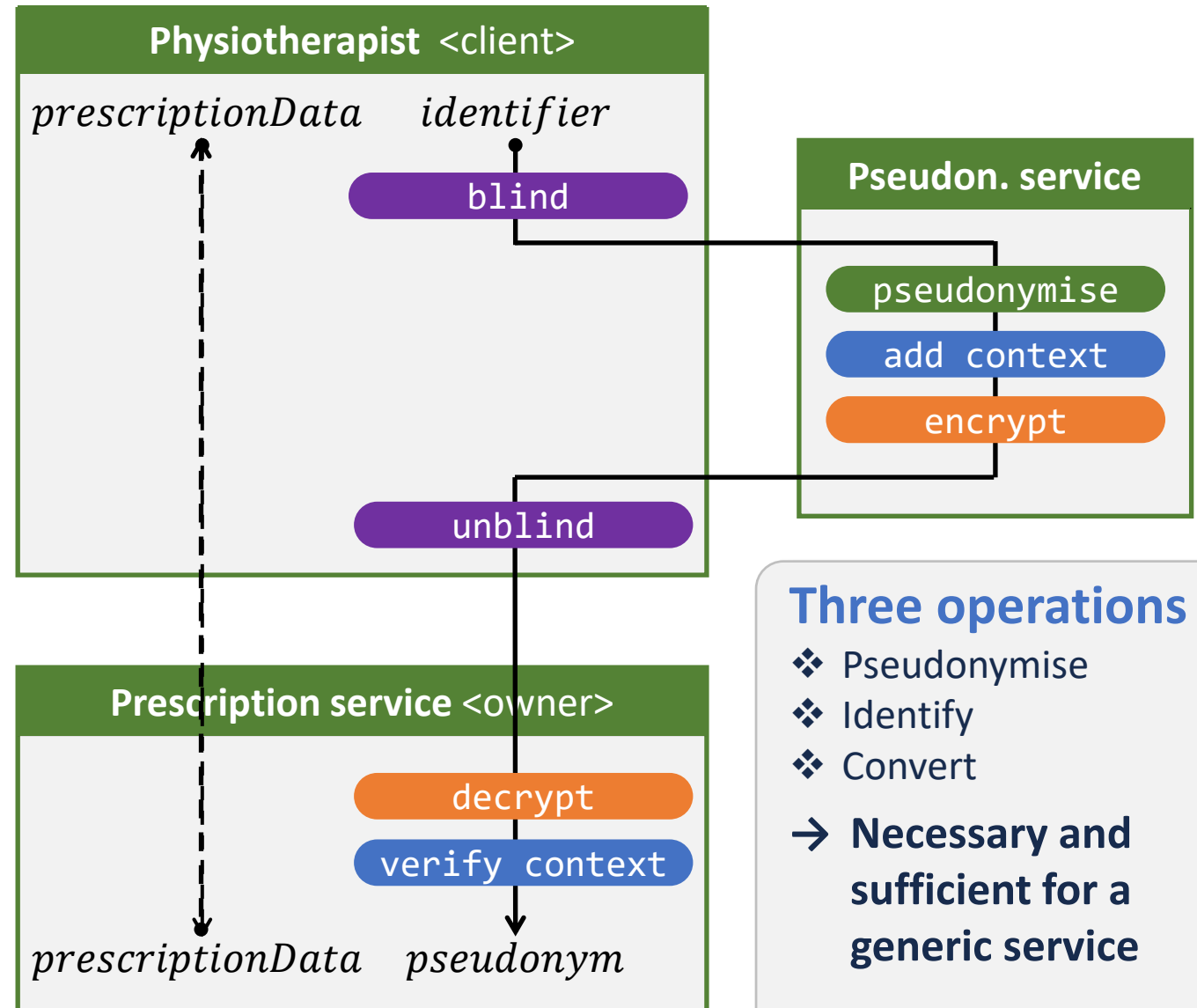
- ❖ Direct communication between healthcare professional and prescription service
- ❖ No in-between entity

✓ **Low-intrusive client-side**

- ❖ No extra keys required
- ❖ Relatively simple implementation

## Structure blinded identifier, blinded pseudonym and final pseudonym

(AV+VXF9H5LdTe4b1 SSC7bHjp6b2enJmf pIC6a3/jCR5fUHxX RSaRniYR8h7ugNqa lGvP49cZnv6lf9B7 2RUG0rA/, eSmII52CEtsZzSseU DY3YKltSgqhq1wLPm 9ncHBzGiv1wMlxmc1 jSmpW36GhTt/s1P5s hZGhG8ncoWKSgkJDy fw=)





## Use case 1 - Live

# Referral prescription = Verwijsvoorschrift / Prescription de renvoi

### What?

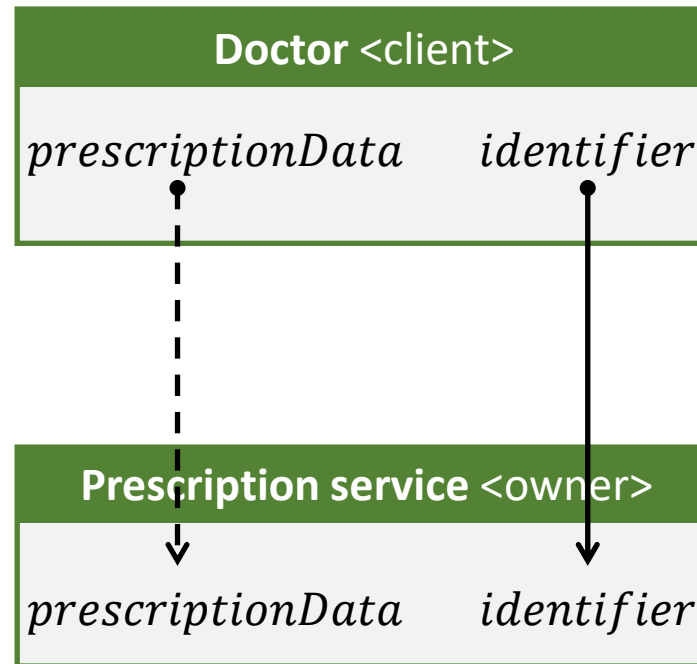
A certificate to start a certain treatment (e.g. physiotherapist, dieticians, speech therapists).

### Requirements

- ❖ **Pseudonymisation**  
Prescription service should never be able to link prescription data to a citizen
- ❖ **Partial encryption**  
The prescription service should not be able to access certain fields

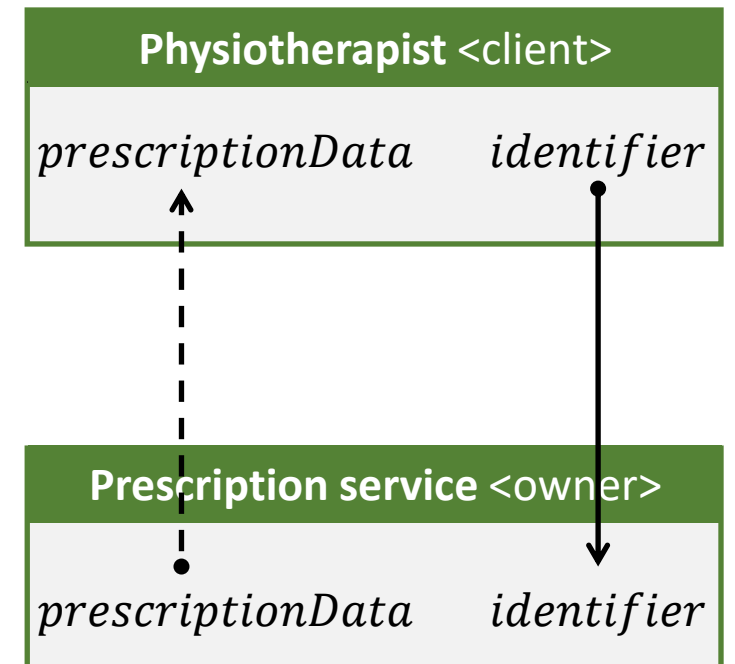
### Scenario 1

Doctor (client) requests Prescription service (owner) to register prescription



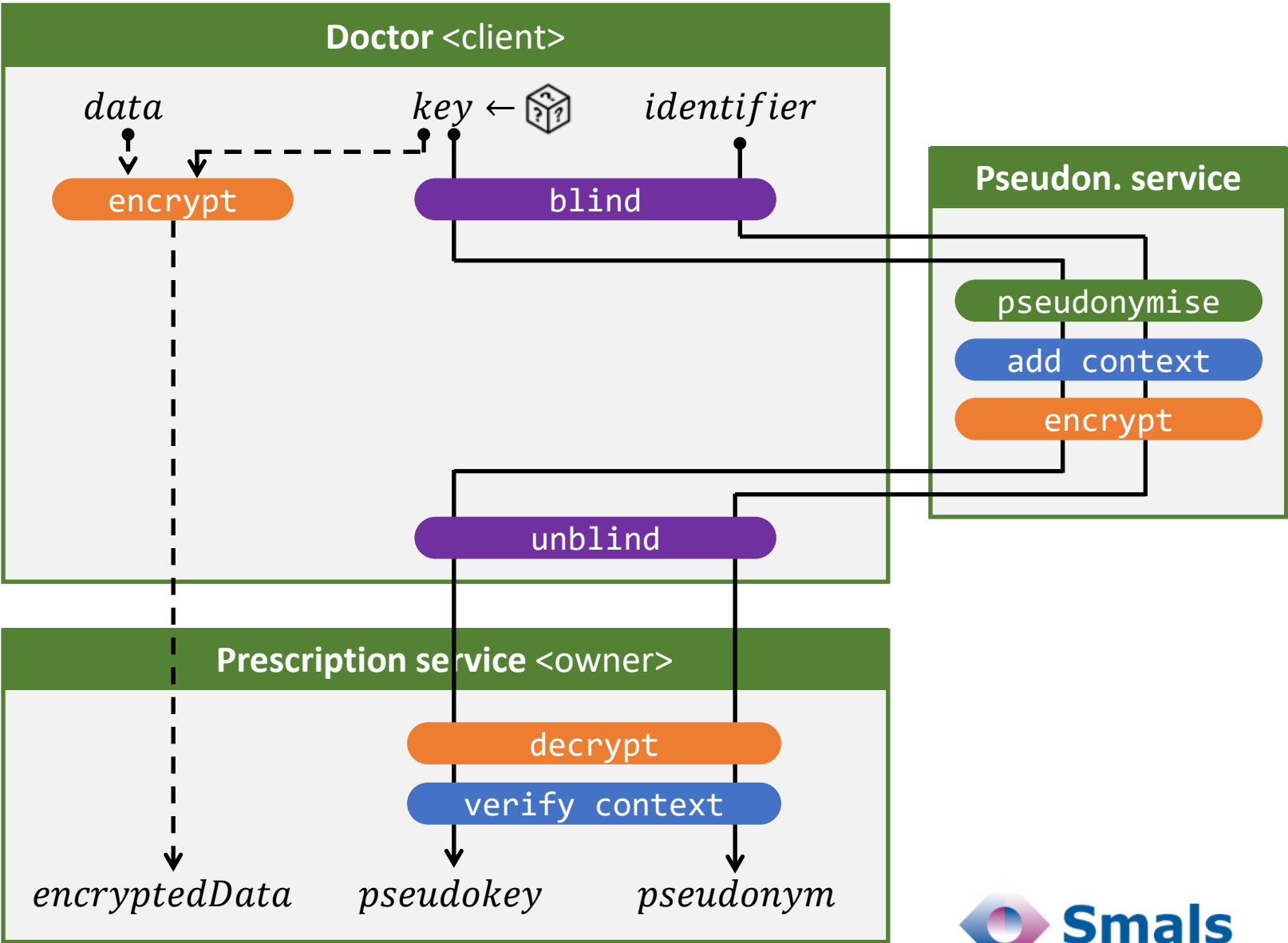
### Scenario 2

Physiotherapist (client) requests access to prescription for a specific citizen from Prescription service (owner)



# Blind Pseudonymisation Service

# Encrypt



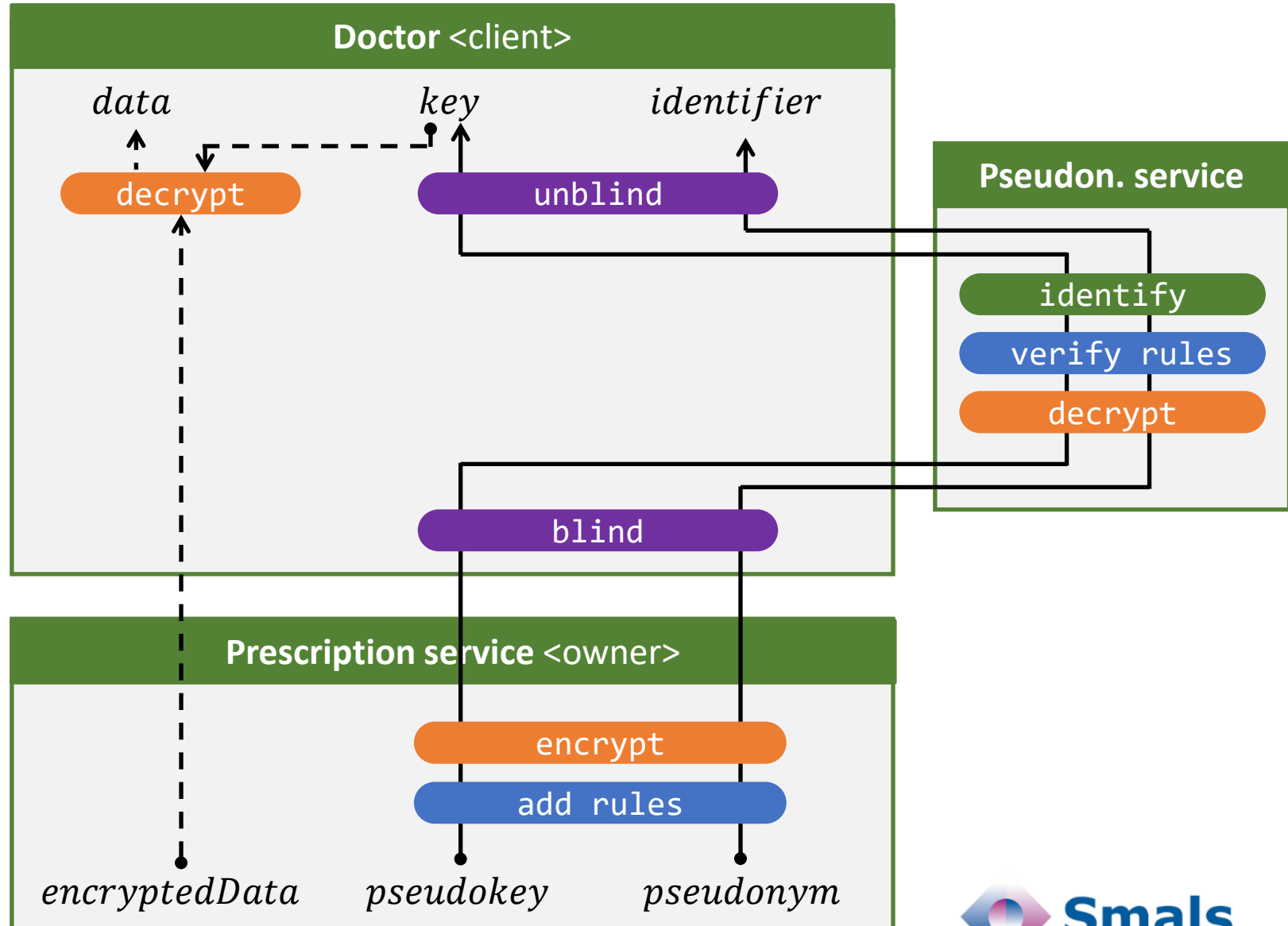
# Blind Pseudonymisation Service

## Decrypt

- ✓ Authorized healthcare professional can access data
- ✓ Prescription service cannot access data
- ✓ Pseudon. service cannot access key
- ✓ Quasi no new logic required

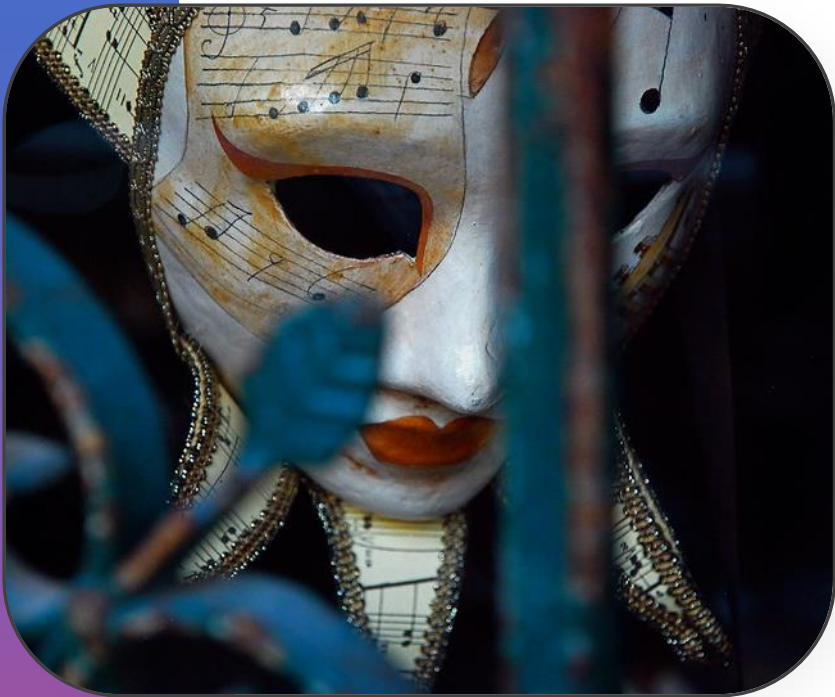
### Crucial that pseudon. service

- ❖ is independent
- ❖ is well secured
- ❖ has proper access control

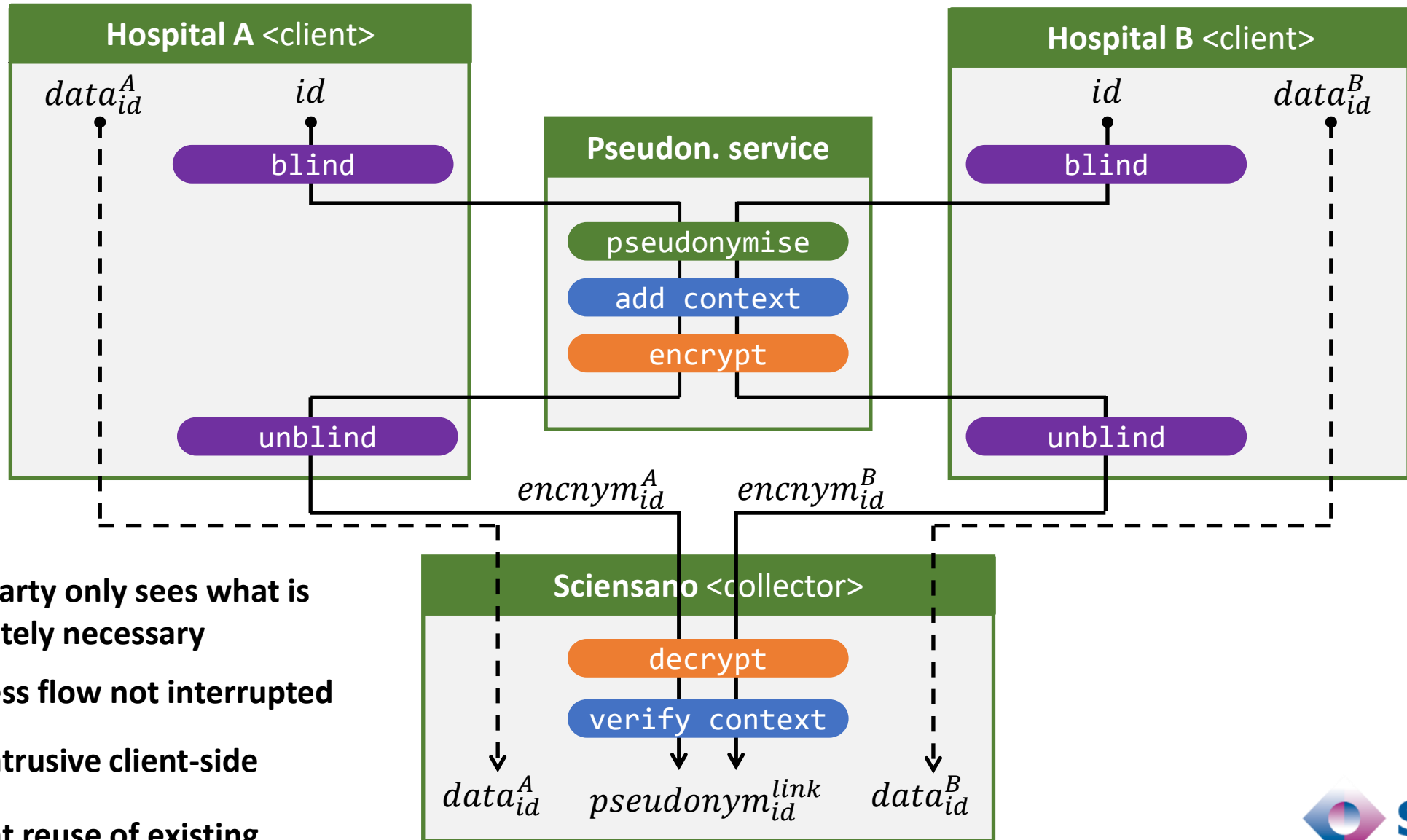


# eHealth Blind Pseudonymisation

- Problem statement
- Referral prescriptions
- **Join & pseudonymise data for research**
- Conclusion



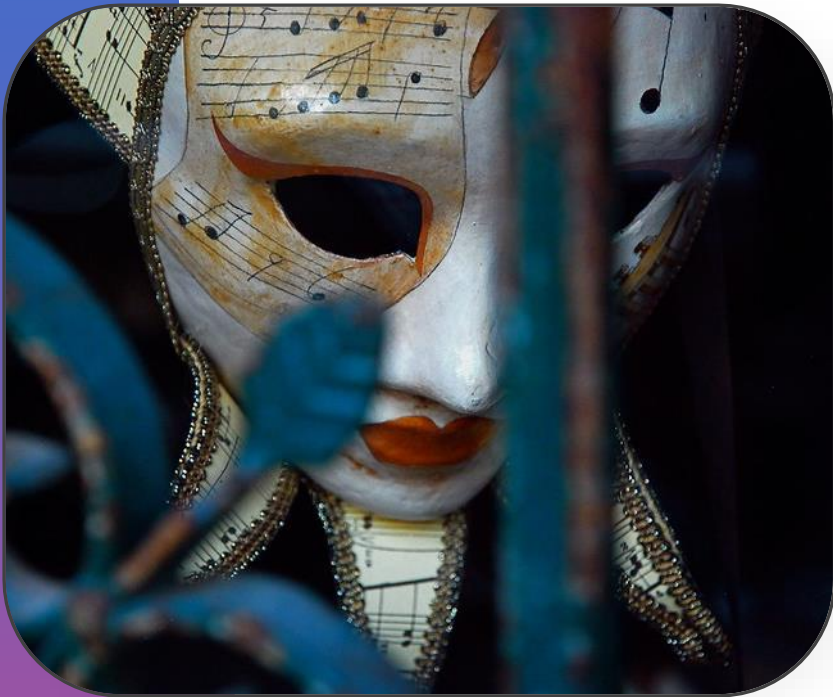
# Join & pseudonymise data for research



- ✓ Each party only sees what is absolutely necessary
- ✓ Business flow not interrupted
- ✓ Low-intrusive client-side
- ✓ Efficient reuse of existing infrastructure

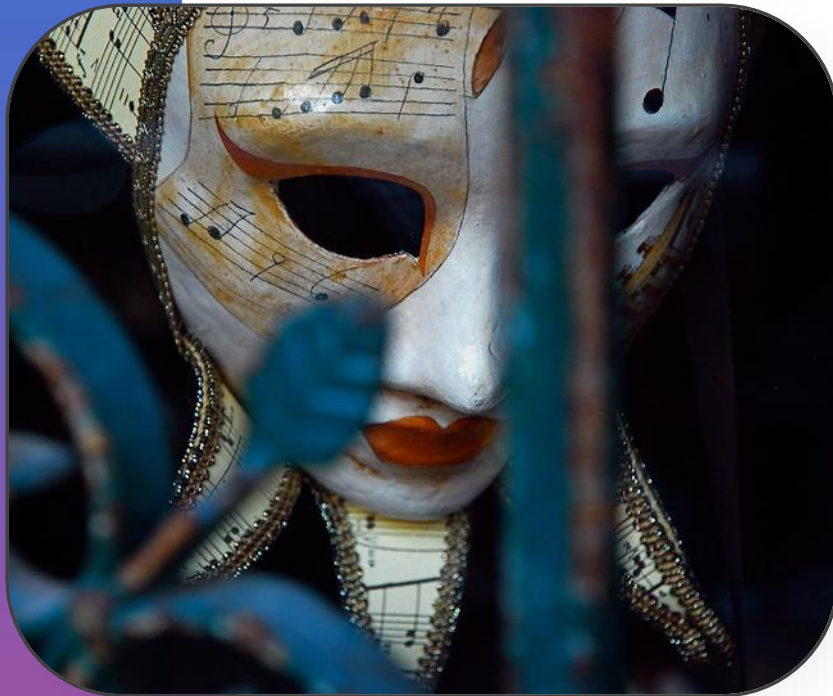
# eHealth Blind Pseudonymisation

- Problem statement
- Referral prescriptions
- Join & pseudonymise data for research
- **Conclusion**





# eHealth Blind Pseudonymisation



## Privacy by design

- ✓ Client only sees identifiers
- ✓ Service only sees pseudons.
- ✓ Pseudon. service sees neither



## Separation of duties

Pseudon. service and prescription service are separate entities



## Simplicity

- ✓ Versatility
- ✓ Low complexity client side



## Live with uptake

ePrescriptions, medication regimens, medical record summaries, vaccinations, allergies and intolerances, fertility.

# Innovation @ Smals Research

# Smart Pseudonymisation

Conversion from citizen identifiers to pseudonyms

## Format-Preserving Pseudonymisation

Retroactive protection of personal data in TEST & ACC of legacy applications



## eHealth Blind Pseudonymisation

Proactive protection of personal data in applications  
Privacy by Design



## Oblivious Join

Non-trivial join & pseudonymise projects for research purposes  
Distributed & no integration



# Oblivious Join

- Problem statement
- Concept
- In practice
- Conclusion





# Oblivious Join

- **Problem statement**
- Concept
- In practice
- Conclusion

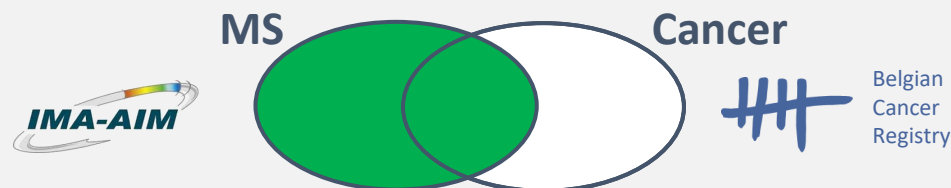


# Concrete case

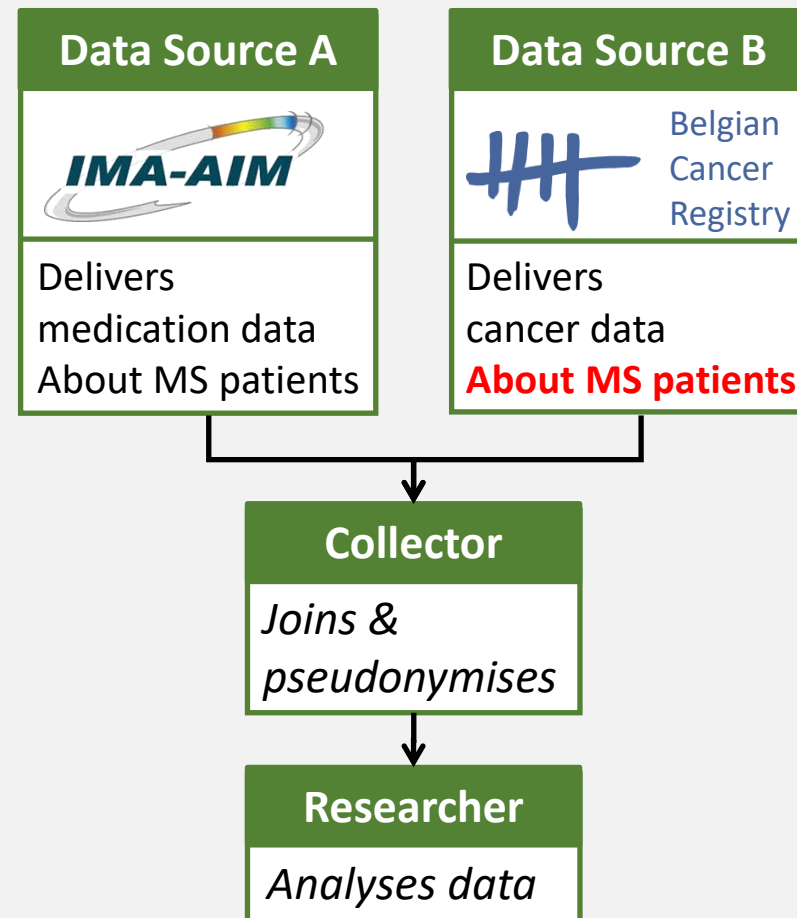
## Research question

Do MS patients who take medications with the molecule teriflunomide or alemtuzumab have an increased cancer risk compared to MS patients treated with other medications?

## Involved citizens



## Naive flow



How can BCR deliver only records about MS patients without learning who has MS?

# Current practice

## Observations

- ✗ Complex flow
- ✗ Expensive
- ✗ Bespoke
- ✗ Doesn't scale well
- ✗ Slow
- ✗ Security risk (data leakage)

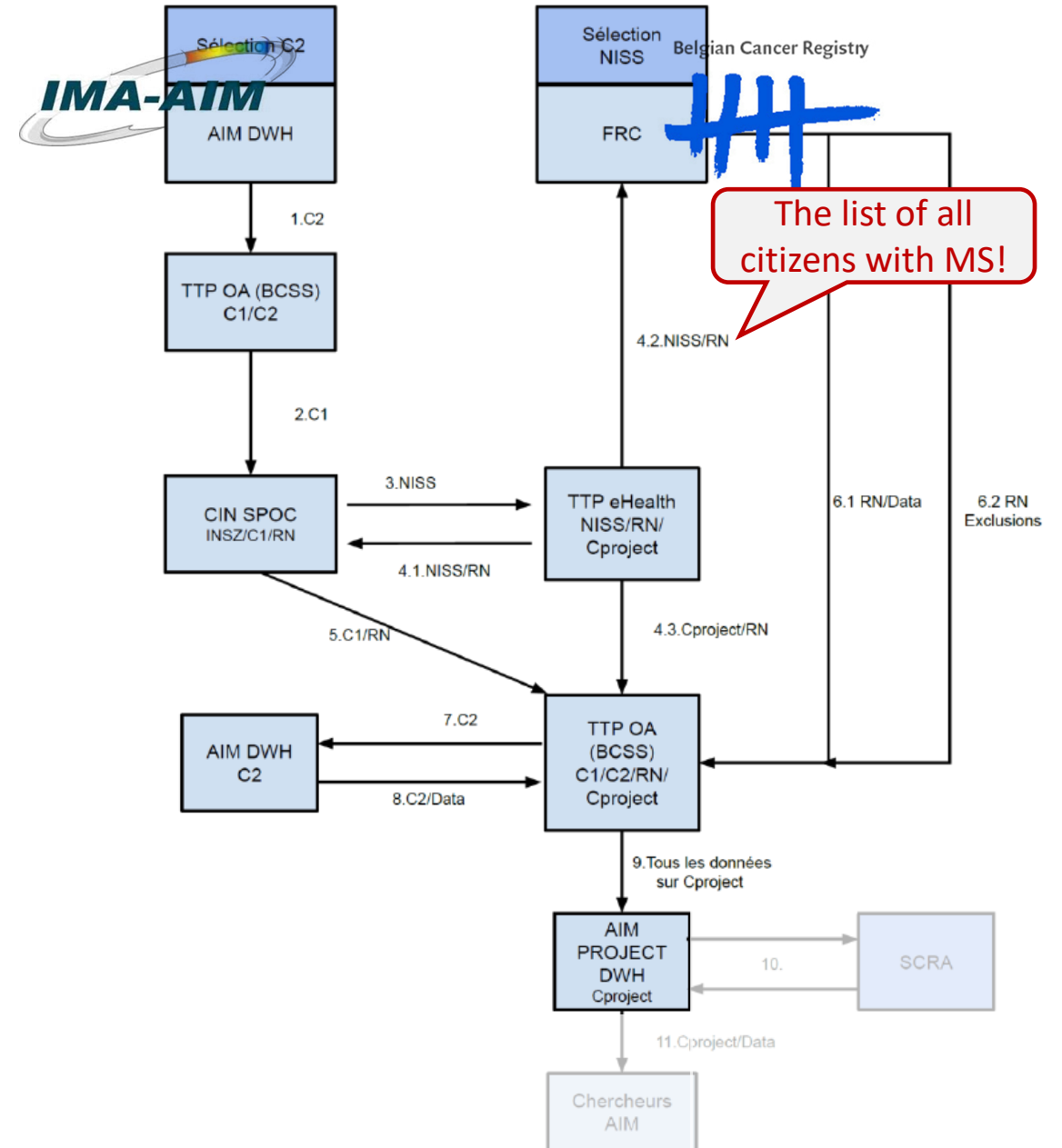
## Feedback

*"Lasts weeks, months, even years"*

*"Requires an exorbitant amount of resources"*

## Other countries

Heavy reliance on combination of trusted parties and strong legal regulations





# Challenge

Join and pseudonymise personal data originating from different sources

## Constraint

Not all data sources able to independently select relevant records

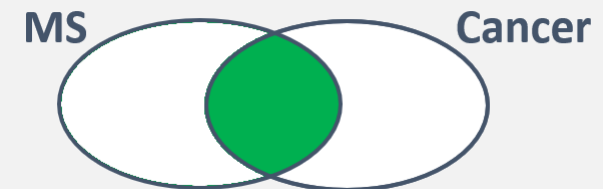
E.g., BCR unable to select records about citizens with MS

## Requirements

- ❖ **Privacy-friendly**  
Involved entities learns only the necessary
- ❖ **Uniform**  
Each research question is different, with different data and different data sources
- ❖ **No data aggregation**  
Researcher access to individual records
- ❖ **Easy to use**

## Focus: set intersection

Researcher wants pseudonymised data of citizens that have MS and cancer



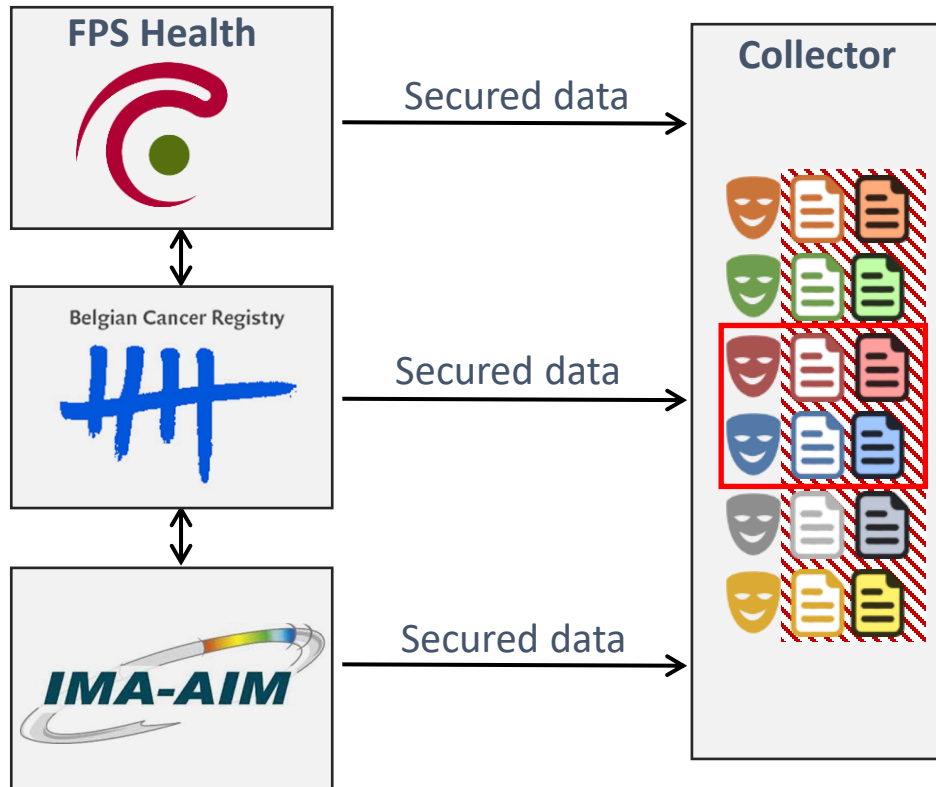
Extensible from there

# Oblivious Join

- Problem statement
- **Concept**
- In practice
- Conclusion



# Concept



## Data sources

- ❖ Do not learn any new personal or statistical data
- ❖ Only see identifiers of their data

## Collector

- ❖ Learns only minimum required pseudonymised personal data
- ❖ Learns high-level statistical data  
E.g. number of citizens with cancer diagnosis
- ❖ Only sees pseudonyms

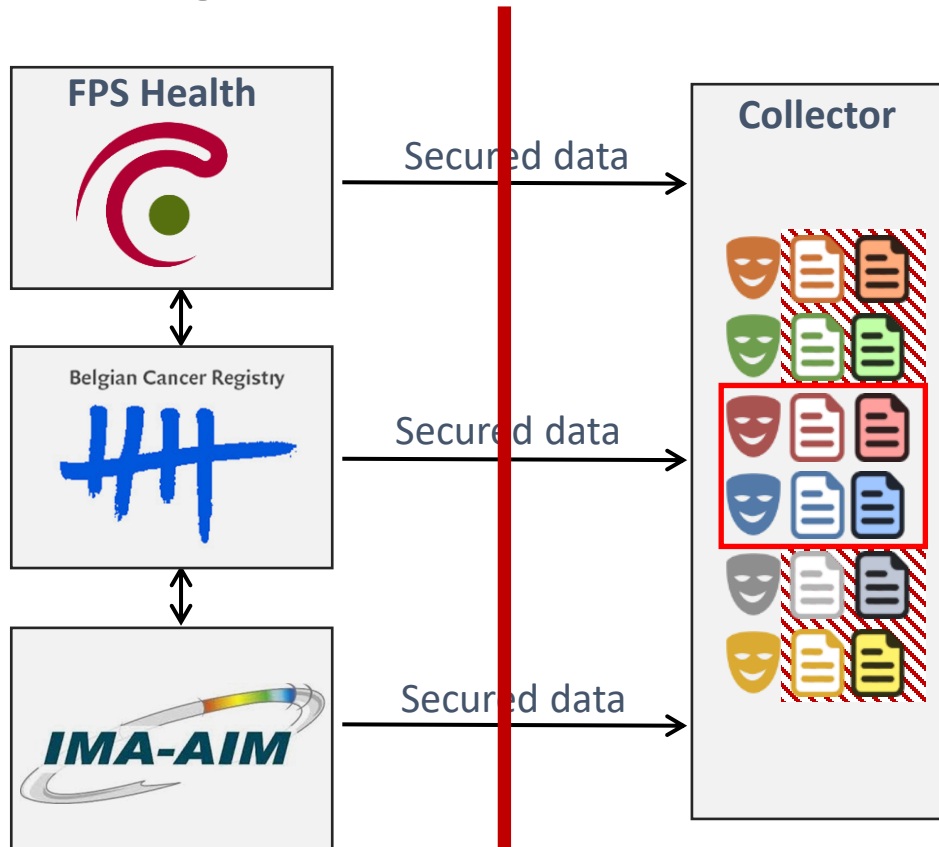
## Properties

- ✓ Privacy-friendly & secure
- ✓ Distributed: no pseudon. service
- ✓ Uniform & no integration
- ✓ Fast & cost-efficient

## 3 steps protocol

1. Fully automated agreements between data sources (no human intervention)
2. Each data source sends all potentially relevant data encrypted & pseudonymised to collector
3. Thanks previous agreements (step 1) collector can only decrypt & combine pertinent records

# Concept



**No collusion  
between data  
source and  
collector**

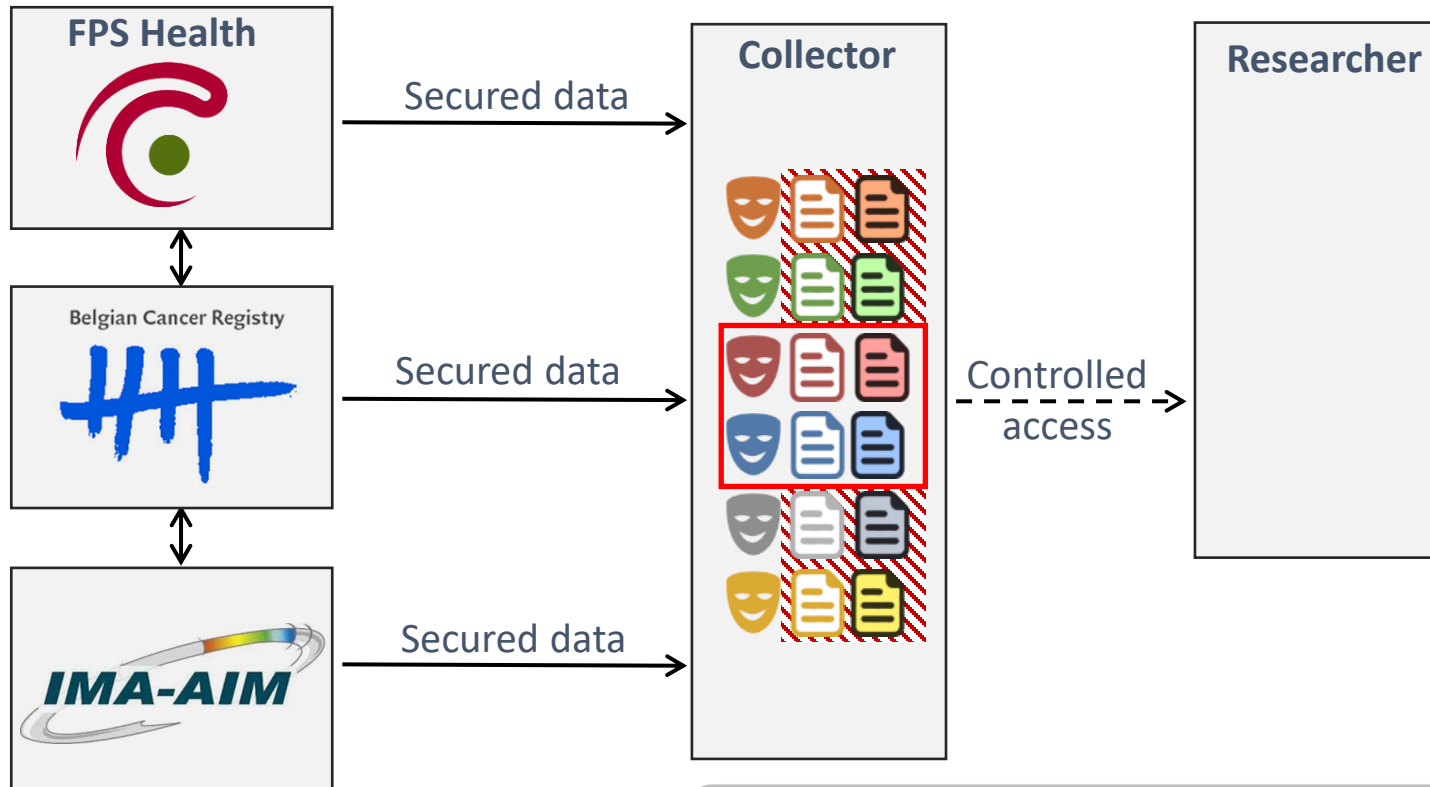
## Properties

- ✓ Privacy-friendly & secure
- ✓ Distributed: no pseudon. service
- ✓ Uniform & no integration
- ✓ Fast & cost-efficient

## 3 steps

1. Fully automated agreements between data sources (no human intervention)
2. Each data source sends all potentially relevant data encrypted & pseudonymised to collector
3. Thanks previous agreements (step 1) collector can only decrypt & combine pertinent records

# Concept



## Collector

### Independent and semi-trusted

1. Deletes asap irrelevant ciphertexts
2. Can do additional checks on the data
3. Controlled access to researcher

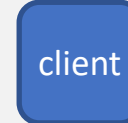
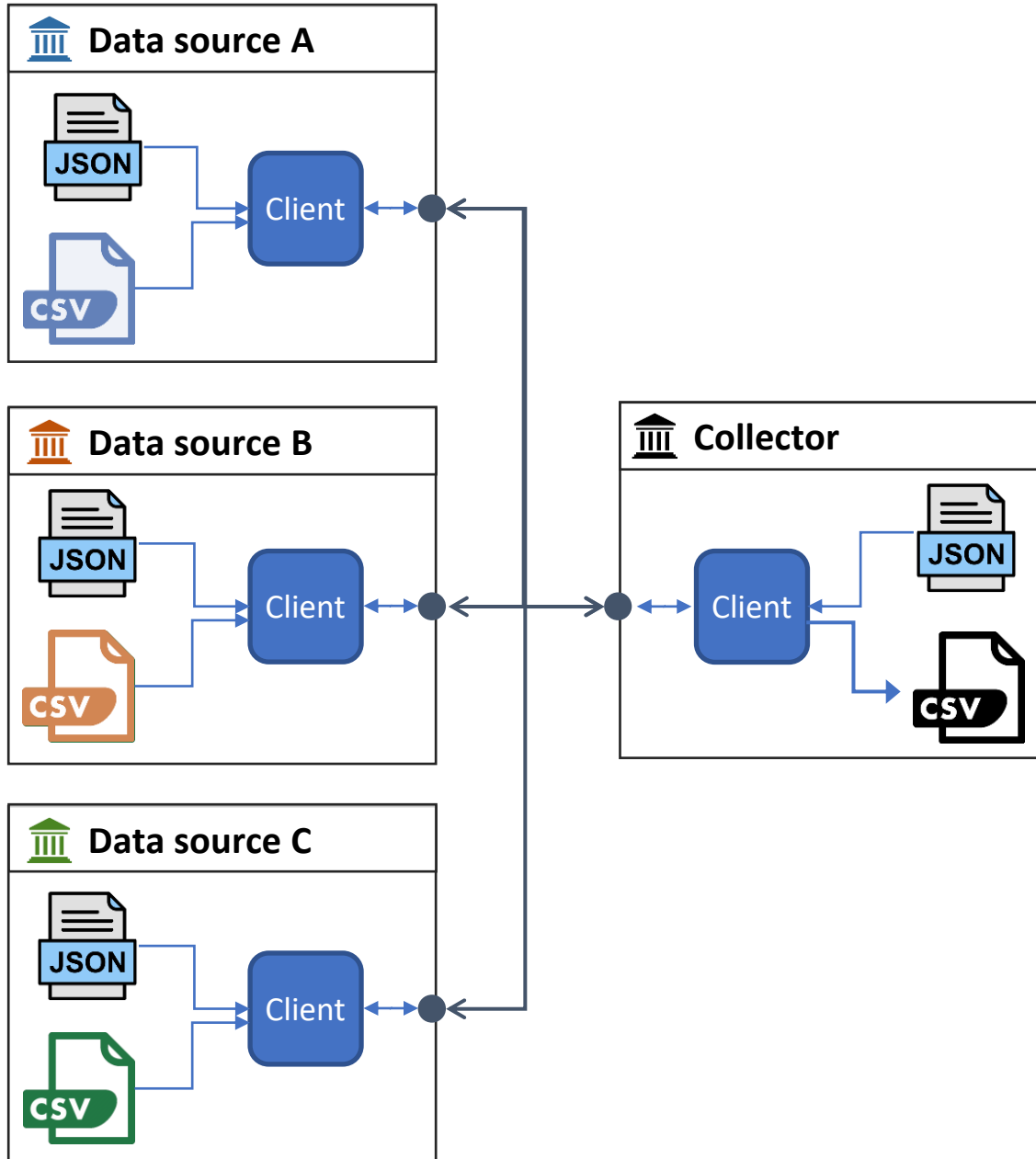
# Oblivious Join

- Problem statement
- Concept
- **In practice**
- Conclusion





# In practice



## Client

- Java jar
- No integration required → non-intrusive, flexible
- All parties use same client (software)
- Command-line interface



## Project description

- JSON file
- Created by coordinating party
- Contains all info required to execute protocol
- All parties use same project description



## Input files

- CSV file
- Created by individual data source (out of scope)
- Contains all, potentially relevant, identified personal data



## Output file

- CSV file
- Collector's output after protocol execution
- Contains minimal required joined & pseudonymised personal data

# Test with fictional data



Extract input CSV

## Data source 1 (IMA-AIM)

60.01.03-231.73	Teriflunomide
60.01.03-562.33	Alemtuzumab
60.01.03-697.92	Glatiramer acetate
60.01.04-606.56	Interferon beta
60.01.04-681.78	Dimethyl fumarate
60.01.05-045.05	Teriflunomide
60.01.05-186.58	Tysabri
60.01.05-617.15	Ocrelizumab
60.01.05-715.14	Alemtuzumab

**200 000 records**

E.g. Citizens with MS



Extract input CSV

## Data source 2 (BCR)

60.01.03-782.07	Melanoma	3	G1
60.01.04-124.53	Colorectal	1	G3
60.01.04-345.26	Prostate	2	G2
60.01.04-562.03	Breast	2	G1
60.01.05-045.05	Lung	1	G3
60.01.05-893.30	Pancreas	4	G2
60.01.06-401.07	Breast	3	G1
60.01.06-696.03	Stomach	2	G1
60.01.07-203.78	Thyroid	1	G3

**500 000 records**

E.g. Citizens with cancer



Extract input CSV

## Data source 3 (FPS Health)

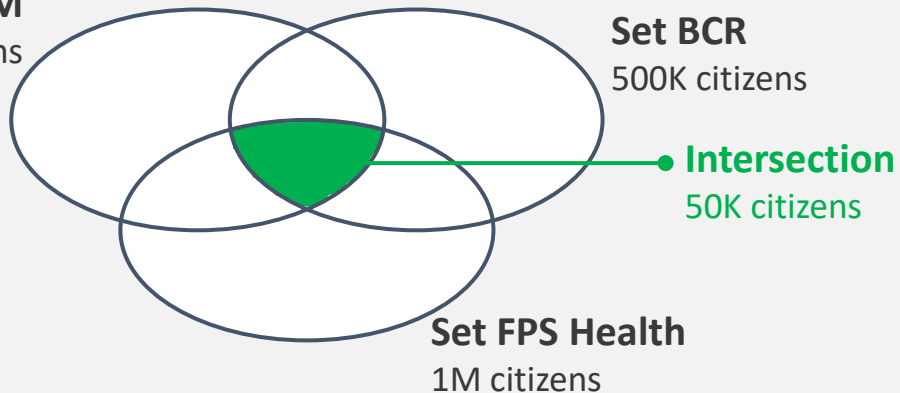
60.01.03-542.53	C
60.01.03-559.36	G
60.01.03-606.86	D
60.01.03-697.92	A
60.01.04-697.62	G
60.01.04-816.40	B
60.01.05-045.05	D
60.01.06-701.95	B
60.01.06-886.07	F

**1 000 000 records**

E.g. Citizens with high-risk profile

**Set IMA-AIM**

200K citizens



# Test with fictional data



## Extract input CSV

### Data source 1 (IMA-AIM)

60.01.03-231.73	Teriflunomide
60.01.03-562.33	Alemtuzumab
60.01.03-697.92	Glatiramer acetate
60.01.04-606.56	Interferon beta
60.01.04-681.78	Dimethyl fumarate
60.01.05-045.05	Teriflunomide
60.01.05-186.58	Tysabri
60.01.05-617.15	Ocrelizumab
60.01.05-715.14	Alemtuzumab

**200 000 records**

E.g. Citizens with MS



## Extract input CSV

### Data source 2 (BCR)

60.01.03-782.07	Melanoma	3	G1
60.01.04-124.53	Colorectal	1	G3
60.01.04-345.26	Prostate	2	G2
60.01.04-562.03	Breast	2	G1
60.01.05-045.05	Lung	1	G3
60.01.05-893.30	Pancreas	4	G2
60.01.06-401.07	Breast	3	G1
60.01.06-696.03	Stomach	2	G1
60.01.07-203.78	Thyroid	1	G3

**500 000 records**

E.g. Citizens with cancer



## Extract input CSV

### Data source 3 (FPS Health)

60.01.03-542.53	C
60.01.03-559.36	G
60.01.03-606.86	D
60.01.03-697.92	A
60.01.04-697.62	G
60.01.04-816.40	B
60.01.05-045.05	D
60.01.06-701.95	B
60.01.06-886.07	F

**1 000 000 records**

E.g. Citizens with high-risk profile

## Extract output CSV

### Collector (KSZ)

**50 000 records**



99338454821...	Teriflunomide	Lung	3	G1	F
12056965607...	Alemtuzumab	Cervix uteri	2	G2	B
15380767762...	Daclizumab	Pancreas	1	G2	A
15380767762...	Teriflunomide	Lung	1	G3	D
31309444464...	Ocrelizumab	Stomach	3	G1	C
99921347021...	Dimethyl fumarate	Breast	2	G2	H
69025938558...	Ofatumumab	Prostate	3	G3	A
38469942453...	Alemtuzumab	Melanoma	4	G1	E
18048091119...	Aubagio	Prostate	3	G3	D

## Who sees what?

- ❖ Data sources only see identifiers
- ❖ Collector only sees pseudonyms
- ❖ No pseudonymisation service

# Test with fictional data



## Extract input CSV

### Data source 1 (IMA-AIM)

60.01.03-231.73	Teriflunomide
60.01.03-562.33	Alemtuzumab
60.01.03-697.92	Glatiramer acetate
60.01.04-606.56	Interferon beta
60.01.04-681.78	Dimethyl fumarate
60.01.05-045.05	Teriflunomide
60.01.05-186.58	Tysabri
60.01.05-617.15	Ocrelizumab
60.01.05-715.14	Alemtuzumab

**200 000 records**

E.g. Citizens with MS



## Extract input CSV

### Data source 2 (BCR)

60.01.03-782.07	Melanoma	3	G1
60.01.04-124.53	Colorectal	1	G3
60.01.04-345.26	Prostate	2	G2
60.01.04-562.03	Breast	2	G1
60.01.05-045.05	Lung	1	G3
60.01.05-893.30	Pancreas	4	G2
60.01.06-401.07	Breast	3	G1
60.01.06-696.03	Stomach	2	G1
60.01.07-203.78	Thyroid	1	G3

**500 000 records**

E.g. Citizens with cancer



## Extract input CSV

### Data source 3 (FPS Health)

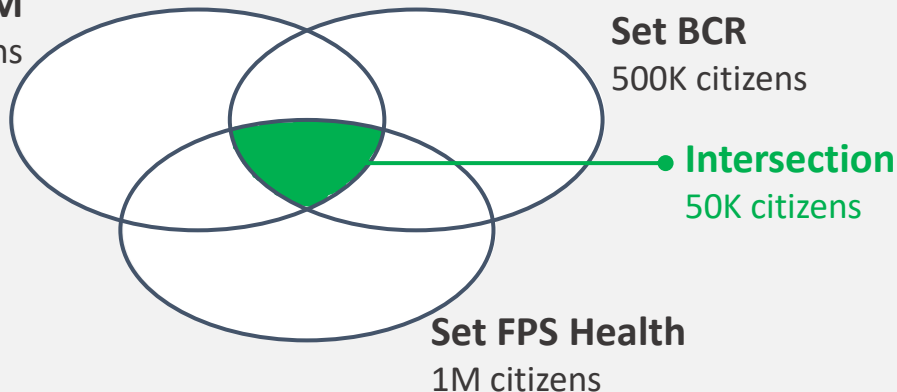
60.01.03-542.53	C
60.01.03-559.36	G
60.01.03-606.86	D
60.01.03-697.92	A
60.01.04-697.62	G
60.01.04-816.40	B
60.01.05-045.05	D
60.01.06-701.95	B
60.01.06-886.07	F

**1 000 000 records**

E.g. Citizens with high-risk profile

### Set IMA-AIM

200K citizens



## Performance test

### Parameters

- MinNbRecords: 10
- 128 bit security

### Infrastructure

- Data sources: 4 i9-7940x cores @ 3.10 GHz, 16GB RAM
- Collector: 2 i9-7940x cores @ 3.10 GHz, 16GB RAM

### Results

- **< 2 min calculations**
- Excl. a few hundred MBs data transfer

# Oblivious Join

- Problem statement
- Concept
- In practice
- **Conclusion**



# Collaboration universities

**Interdisciplinary paper (To be published in 2024)**

**Privacy-By-Design in the Belgian Public Sector**

Pseudonymising & Joining Personal Data Fragmented over Multiple Organisations



*In Public Governance and Emerging Technologies – Values, Trust, and Compliance by Design*



**Utrecht University**

**SPRINGER NATURE**

**Expert paper**

Ongoing  
<CONFIDENTIAL>



# Evaluation

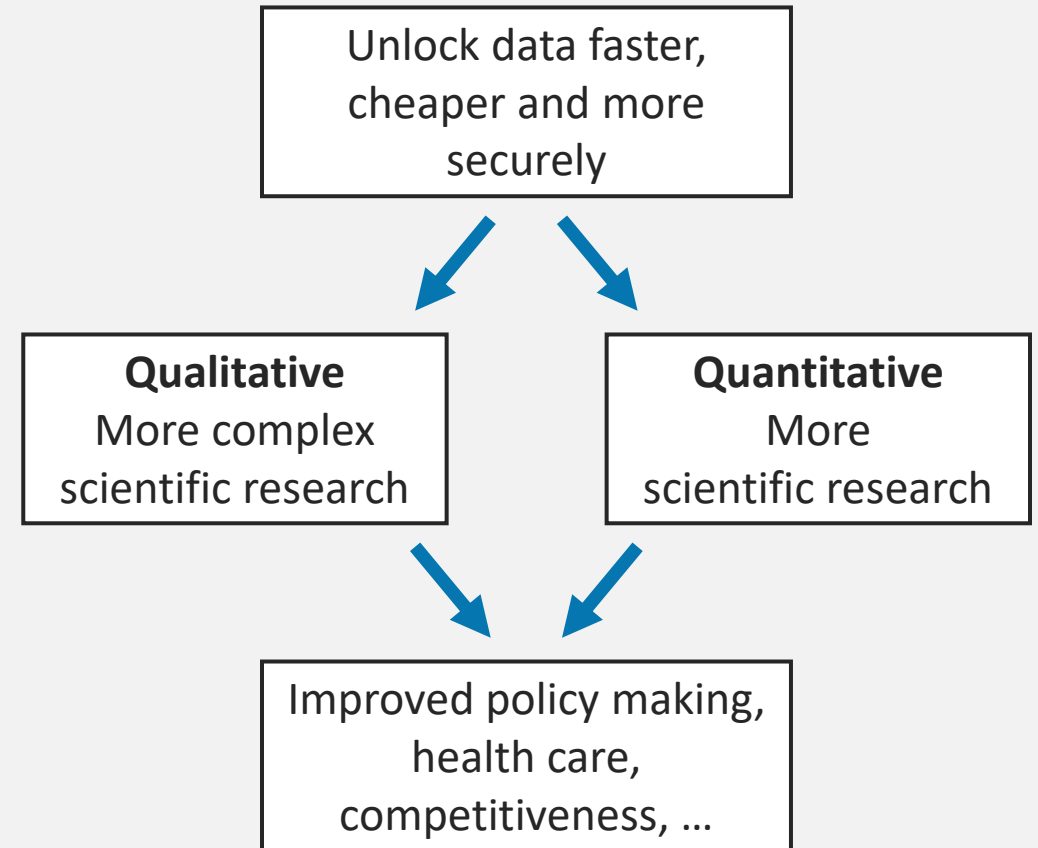
## Advantages

- ✓ Answer on business need
- ✓ Privacy-friendly & secure
- ✓ Distributed (no pseudonymisation service)
- ✓ Uniform & no integration
- ✓ Fast & cost-efficient
- ✓ Formal academic validation

## Challenges

- ⚠ Only passive interest
- ⚠ Still in research phase
- ⚠ Higher development complexity (but lower infra)
- ⚠ Extensions required

## Opportunities



# Wrapping up

# Innovation @ Smals Research

## Smart Pseudonymisation

Conversion from citizen identifiers to pseudonyms

### Format-Preserving Pseudonymisation

Retroactive protection of personal data in TEST & ACC of legacy applications



### eHealth Blind Pseudonymisation

Proactive protection of personal data in applications  
Privacy by Design



### Oblivious Join

Non-trivial join & pseudonymise projects for research purposes  
Distributed & no integration



Smart pseudonymisation can play a crucial role to protect personal data



Further reading  
[www.smalsresearch.be](http://www.smalsresearch.be)

# Thanks for your attention

If you have any questions, do not hesitate to contact me!  
See you at the Smals booth (Nb. 6)!

✉ [kristof.verslype@smals.be](mailto:kristof.verslype@smals.be)

☎ +32(0)2 7875376

in [linkedin.com/in/verslype](https://www.linkedin.com/in/verslype)

🌐 [www.smals.be](http://www.smals.be)  
[www.smalsresearch.be](http://www.smalsresearch.be)  
[www.cryptanium.eu](http://www.cryptanium.eu)



[www.smalsresearch.be/tag/pseudonymisation/](http://www.smalsresearch.be/tag/pseudonymisation/)

# Images



**Judy Dean**  
Creative Commons  
<https://flickr.com/photos/peterscherub/53152339550/>



**estorde**  
Creative commons  
<https://flickr.com/photos/estorde/4572006561>



**Pixabay**  
Pixabay License  
<https://pixabay.com/fr/photos/femme-les-yeux-masquer-carnaval-411494/>



**Aris Gionis**  
Creative Commons  
Flickr



**Daniel Bruce**  
Creative Commons  
<https://iconscout.com/free-icon/mask-126>



**Oscar Gende Villar**  
Creative Commons  
Flickr