# Pseudonymization as a Service:

## Compartmentalizing and Controlling Data Processing in Evolving Systems with Micropseudonymization

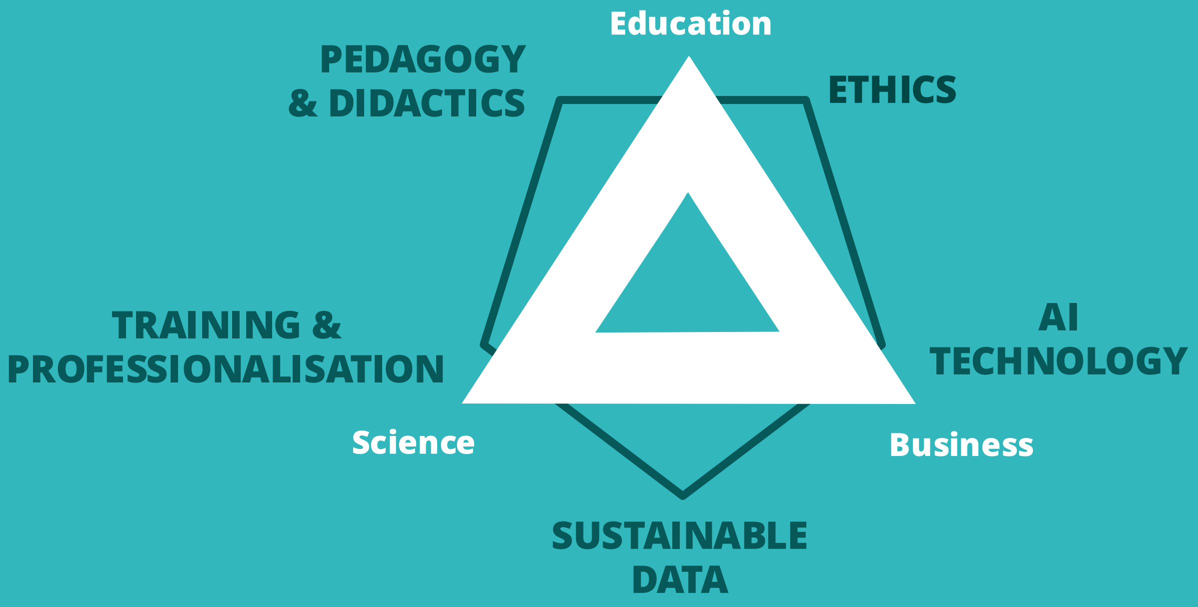
17 November 2025

**Job Doesburg**, Bernard van Gastel, Erik Poll
Radboud University
— iCIS | Software Science department
— NOLAI | Sustainable Data focus area

NOLAI

# NOLAI

NATIONAL
EDUCATION LAB AI



**Facts & Figures:**

- responsible AI for primary & secondary education
- €143M funding for 10 years
- 64 partners (2025):
  schools, businesses, research institutes
- ~77 co-creation projects in 10 years

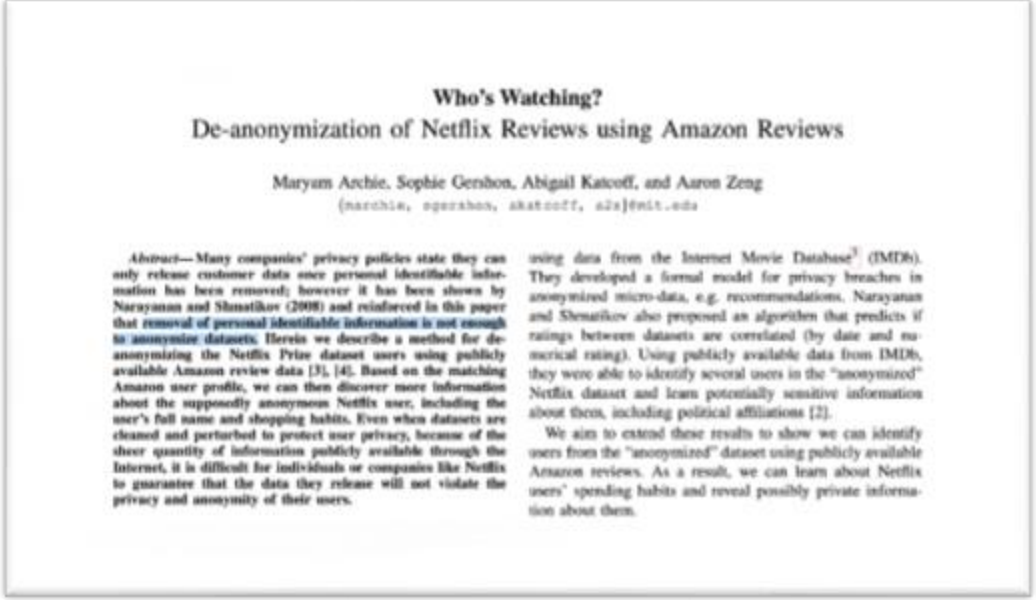
NOLAI
NATIONAL
EDUCATION LAB AI

## NOLAI Research Data Platform:

- Consent registration
- Surveys
- File uploads
- Voice & video recordings
- Application usage logs
- AI model training
- …?

## The challenges we face:

- Complex & evolving infrastructure
- Rapid development speed
- Privacy!



Who's Watching?

De-anonymization of Netflix Reviews using Amazon Reviews

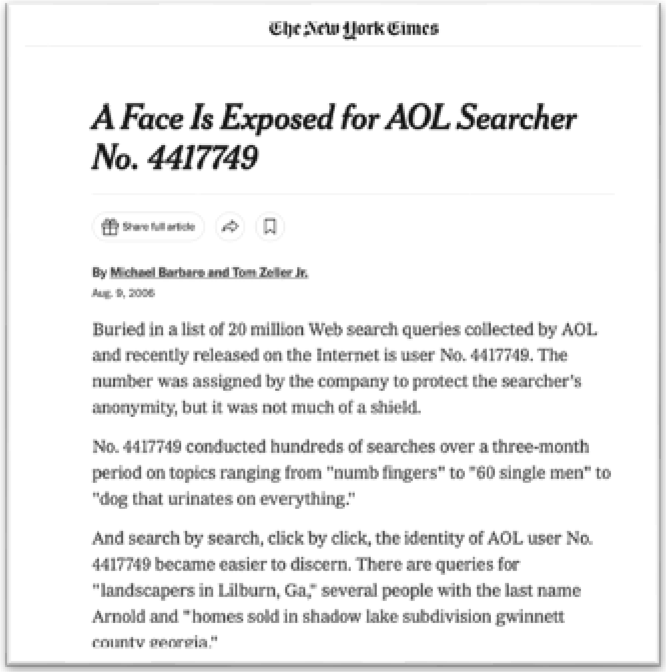
Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aaron Zeng

{marchie, sgershon, akatcoff, a2z}@mit.edu

*Abstract*—Many companies' privacy policies state they can only release customer data once personal identifiable information has been removed; however it has been shown by Narayanan and Shmatikov (2008) and reinforced in this paper that removal of personal identifiable information is not enough to anonymize datasets. Herein we describe a method for de-anonymizing the Netflix Prize dataset users using publicly available Amazon review data [3], [4]. Based on the matching Amazon user profile, we can then discover more information about the supposedly anonymous Netflix user, including the user's full name and shopping habits. Even when datasets are cleaned and perturbed to protect user privacy, because of the sheer quantity of information publicly available through the Internet, it is difficult for individuals or companies like Netflix to guarantee that the data they release will not violate the privacy and anonymity of their users.

using data from the Internet Movie Database[3] (IMDb). They developed a formal model for privacy breaches in anonymized micro-data, e.g. recommendations. Narayanan and Shmatikov also proposed an algorithm that predicts if ratings between datasets are correlated (by date and numerical rating). Using publicly available data from IMDb, they were able to identify several users in the "anonymized" Netflix dataset and learn potentially sensitive information about them, including political affiliations [2].

We aim to extend these results to show we can identify users from the "anonymized" dataset using publicly available Amazon reviews. As a result, we can learn about Netflix users' spending habits and reveal possibly private information about them.



The New York Times

*A Face Is Exposed for AOL Searcher No. 4417749*

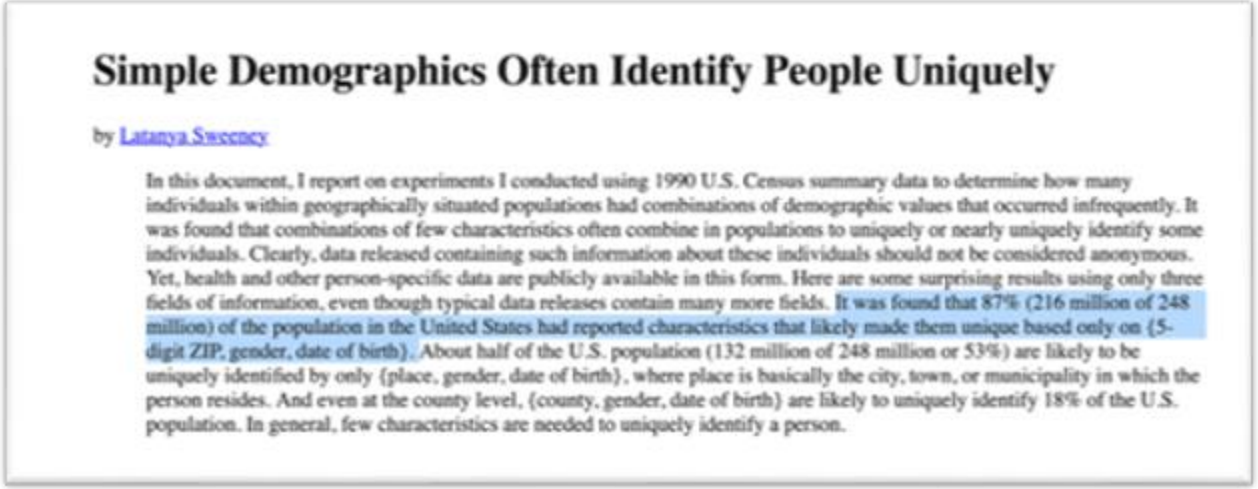
By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."



# Simple Demographics Often Identify People Uniquely

by Latanya Sweeney

In this document, I report on experiments I conducted using 1990 U.S. Census summary data to determine how many individuals within geographically situated populations had combinations of demographic values that occurred infrequently. It was found that combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals. Clearly, data released containing such information about these individuals should not be considered anonymous. Yet, health and other person-specific data are publicly available in this form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

# Data processing systems are growing increasingly complex

NOLAI

# This complexity increases the chance of data breaches

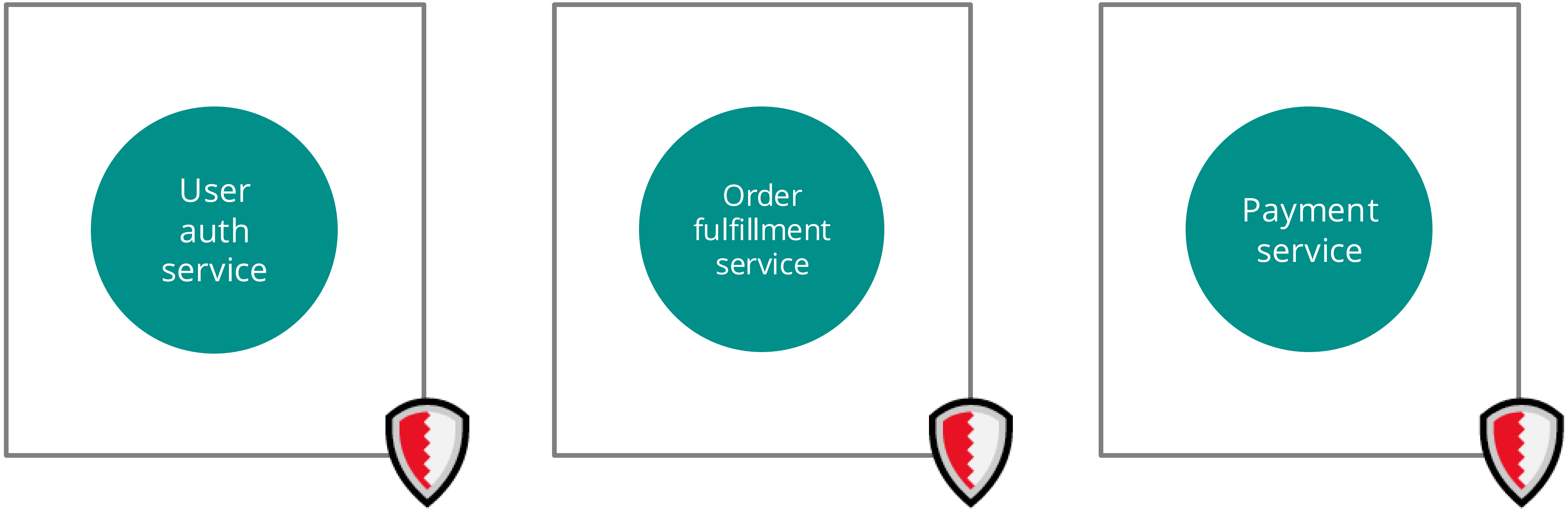
# Micropseudonymization

- In many real-world systems, **anonymisation** or **deleting** data is **not** an option
- The **impact** of data breaches depends on the linkability/identifiability of data
- Minimize data linkability with **micropseudonymization**:
  use different pseudonyms for data in different (sub)systems

# COMPARTMENTALIZATION

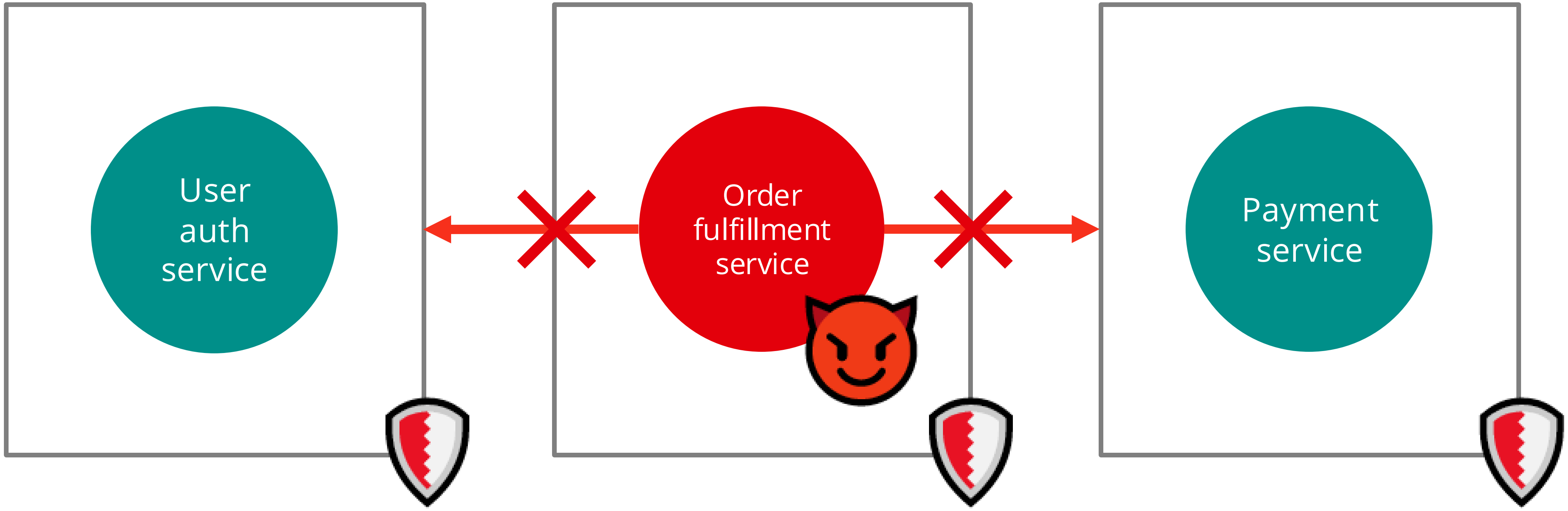
PSEUDONYMIZATION AS A SERVICE

# Functional compartmentalization



PSEUDONYMIZATION AS A SERVICE

# Functional compartmentalization



PSEUDONYMIZATION AS A SERVICE

# Data compartmentalization

User auth service

| email | passwd |
|---|---|
| bob@example.net | ******* |
| alice@fake.com | ******* |
| charlie@fakemail.nl | ******* |
| devin@example.co | ******* |
| … | … |

Order fulfillment service

| email | order id | parcelnr |
|---|---|---|
| bob@example.net | 202502 | 3S12345 |
| alice@fake.com | 202503 | 3S93752 |
| charlie@fakemail.nl | 202504 | C890252 |
| devin@example.co | 202505 | C892734 |
| … | … | |

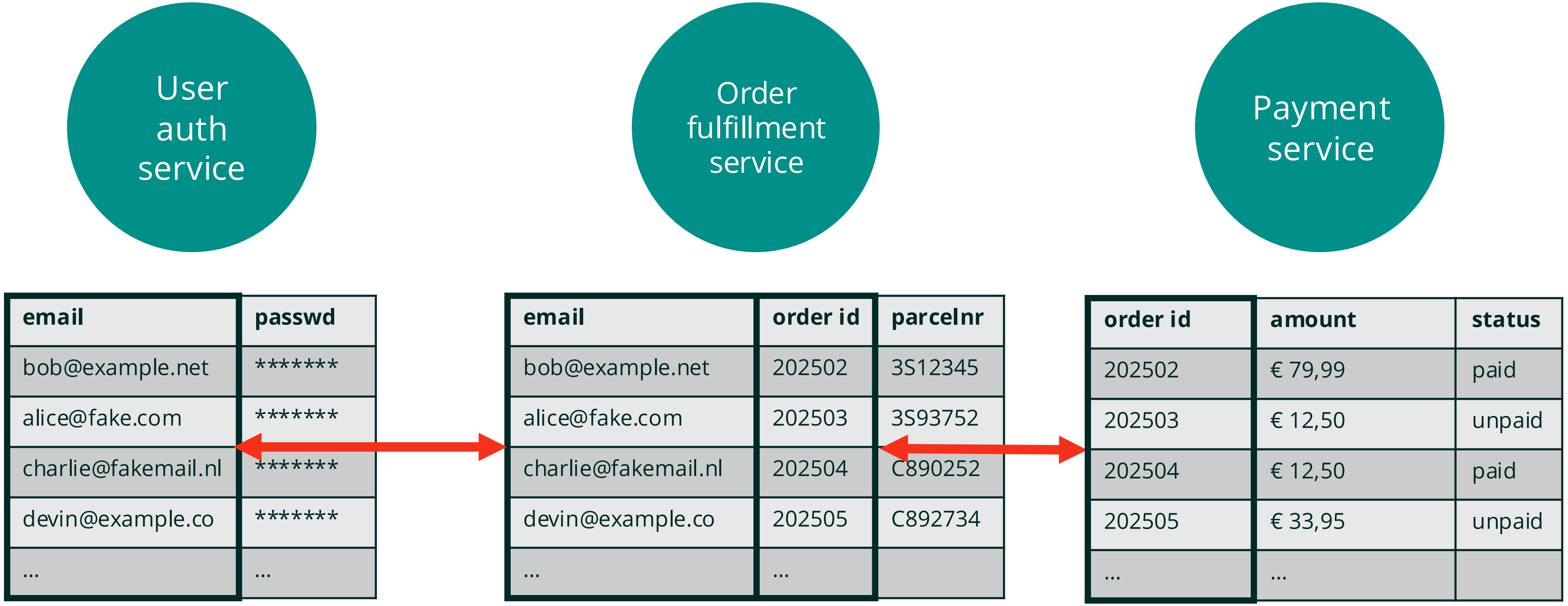
Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

PSEUDONYMIZATION AS A SERVICE

# Data compartmentalization

DATA FORMS A COMPARTMENT WHEN IT CANNOT BE LINKED TO OTHER DATA



| email | passwd |
|---|---|
| bob@example.net | ******* |
| alice@fake.com | ******* |
| charlie@fakemail.nl | ******* |
| devin@example.co | ******* |
| … | … |

| email | order id | parcelnr |
|---|---|---|
| bob@example.net | 202502 | 3S12345 |
| alice@fake.com | 202503 | 3S93752 |
| charlie@fakemail.nl | 202504 | C890252 |
| devin@example.co | 202505 | C892734 |
| … | … | |

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

PSEUDONYMIZATION AS A SERVICE

# Data compartmentalization

BECAUSE EMAILS ARE PUBLIC,
DATA IS NOT ONLY LINKABLE,
BUT EVEN IDENTIFIABLE!

User auth service

| email | passwd |
|---|---|
| bob@example.net | ******* |
| alice@fake.com | ******* |
| charlie@fakemail.nl | ******* |
| devin@example.co | ******* |
| … | … |

Order fulfillment service 😈

| email | order id | parcelnr |
|---|---|---|
| bob@example.net | 202502 | 3S12345 |
| alice@fake.com | 202503 | 3S93752 |
| charlie@fakemail.nl | 202504 | C890252 |
| devin@example.co | 202505 | C892734 |
| … | … | |

⚠

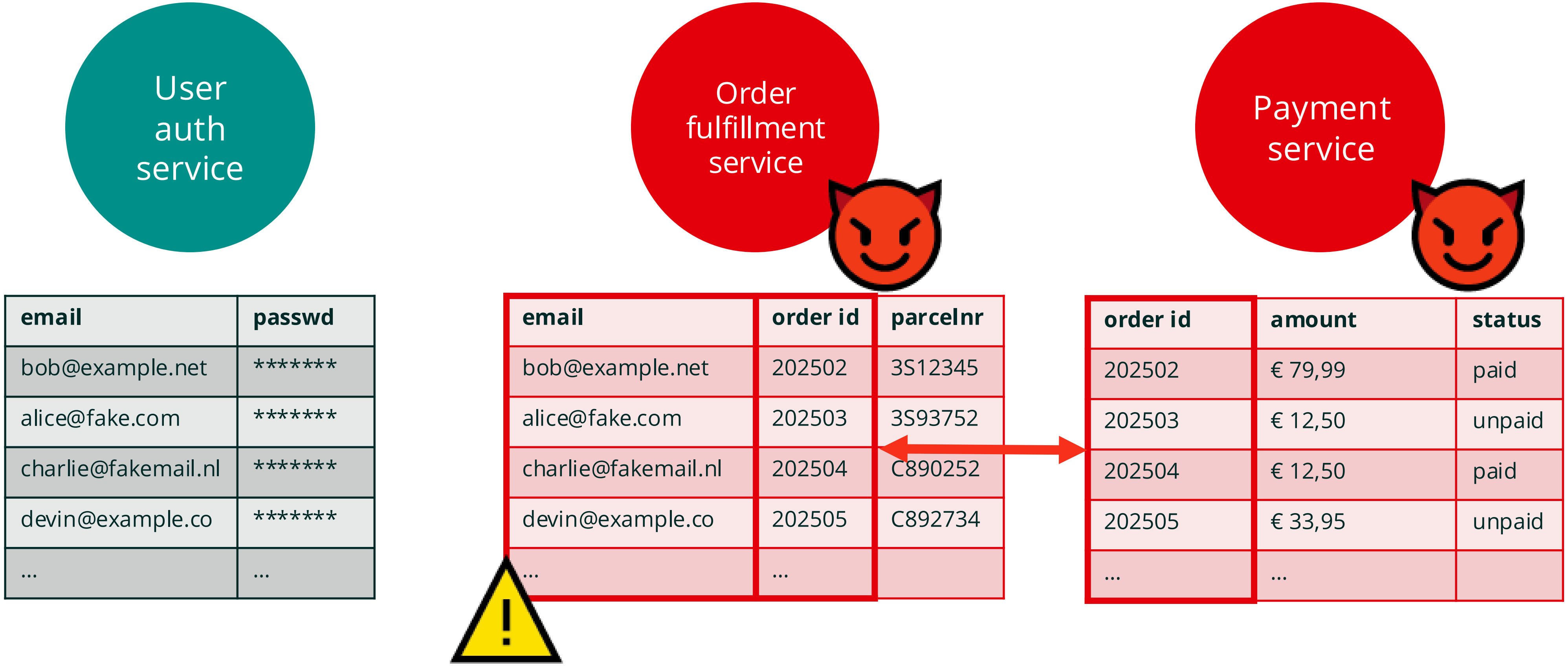
Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

PSEUDONYMIZATION AS A SERVICE

# Data compartmentalization

BECAUSE EMAILS ARE PUBLIC,
DATA IS NOT ONLY LINKABLE,
BUT EVEN IDENTIFIABLE!



User auth service

| email | passwd |
|---|---|
| bob@example.net | ******* |
| alice@fake.com | ******* |
| charlie@fakemail.nl | ******* |
| devin@example.co | ******* |
| … | … |

Order fulfillment service

| email | order id | parcelnr |
|---|---|---|
| bob@example.net | 202502 | 3S12345 |
| alice@fake.com | 202503 | 3S93752 |
| charlie@fakemail.nl | 202504 | C890252 |
| devin@example.co | 202505 | C892734 |
| … | … | |

Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

# Data compartmentalization

User auth service

| uid | email | passwd |
|---|---|---|
| 23 | bob@example.net | ******* |
| 24 | alice@fake.com | ******* |
| 25 | charlie@fakemail.nl | ******* |
| 26 | devin@example.co | ******* |
| … | … | … |

Order fulfillment service

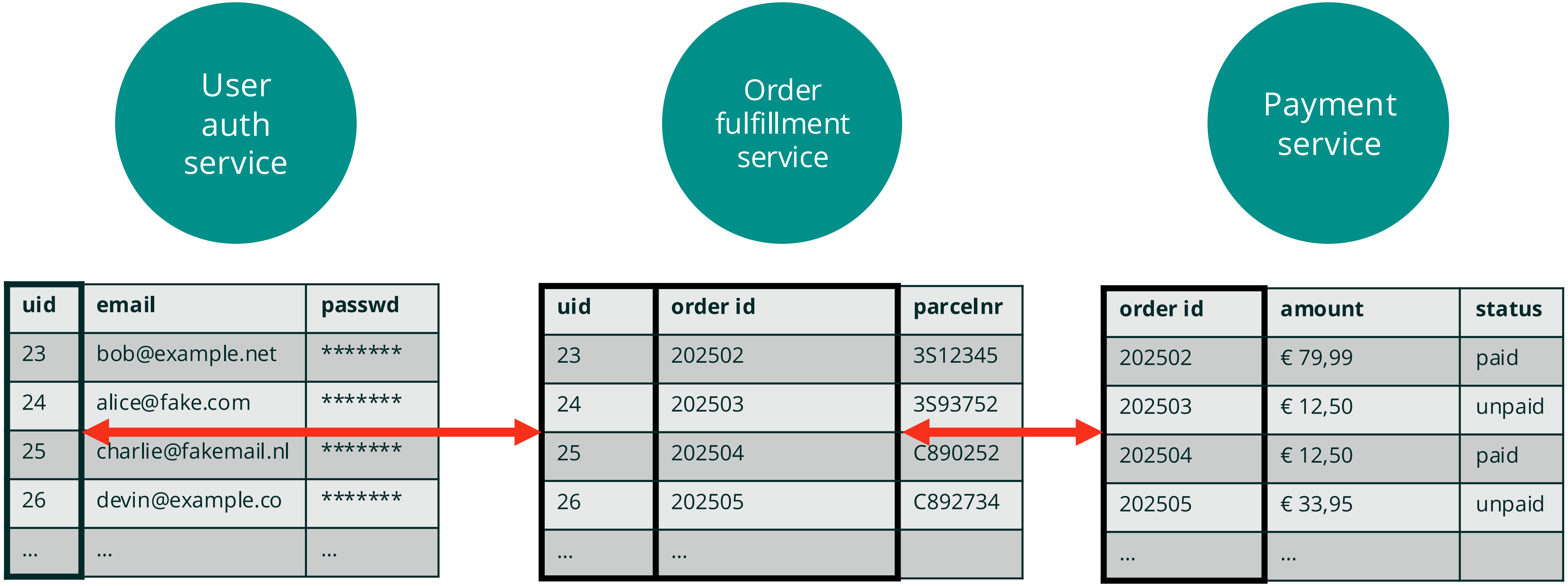
| uid | order id | parcelnr |
|---|---|---|
| 23 | 202502 | 3S12345 |
| 24 | 202503 | 3S93752 |
| 25 | 202504 | C890252 |
| 26 | 202505 | C892734 |
| … | … | |

Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

PSEUDONYMIZATION AS A SERVICE

# Data compartmentalization



User auth service

| uid | email | passwd |
|---|---|---|
| 23 | bob@example.net | ******* |
| 24 | alice@fake.com | ******* |
| 25 | charlie@fakemail.nl | ******* |
| 26 | devin@example.co | ******* |
| … | … | … |

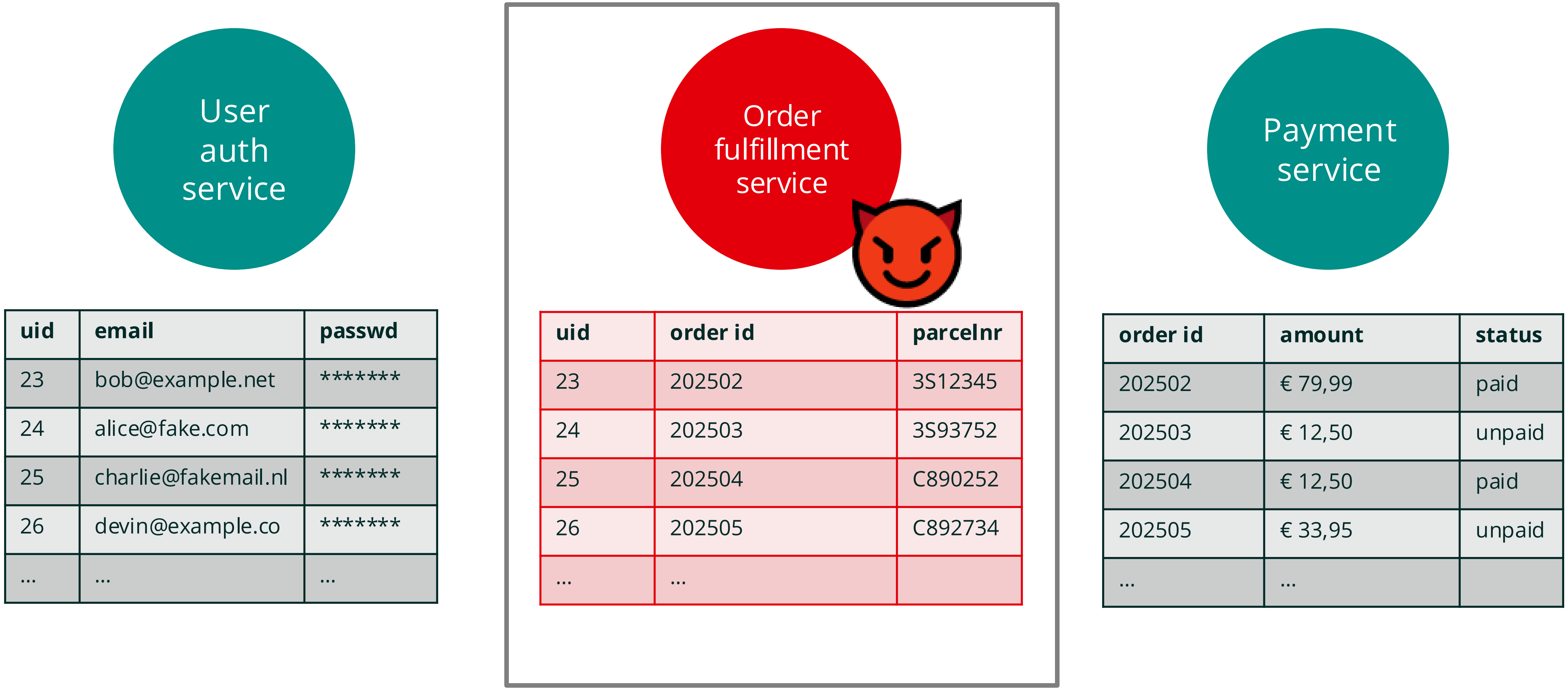
Order fulfillment service

| uid | order id | parcelnr |
|---|---|---|
| 23 | 202502 | 3S12345 |
| 24 | 202503 | 3S93752 |
| 25 | 202504 | C890252 |
| 26 | 202505 | C892734 |
| … | … | |

Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

# Data compartmentalization



User auth service

| uid | email | passwd |
|---|---|---|
| 23 | bob@example.net | ******* |
| 24 | alice@fake.com | ******* |
| 25 | charlie@fakemail.nl | ******* |
| 26 | devin@example.co | ******* |
| … | … | … |

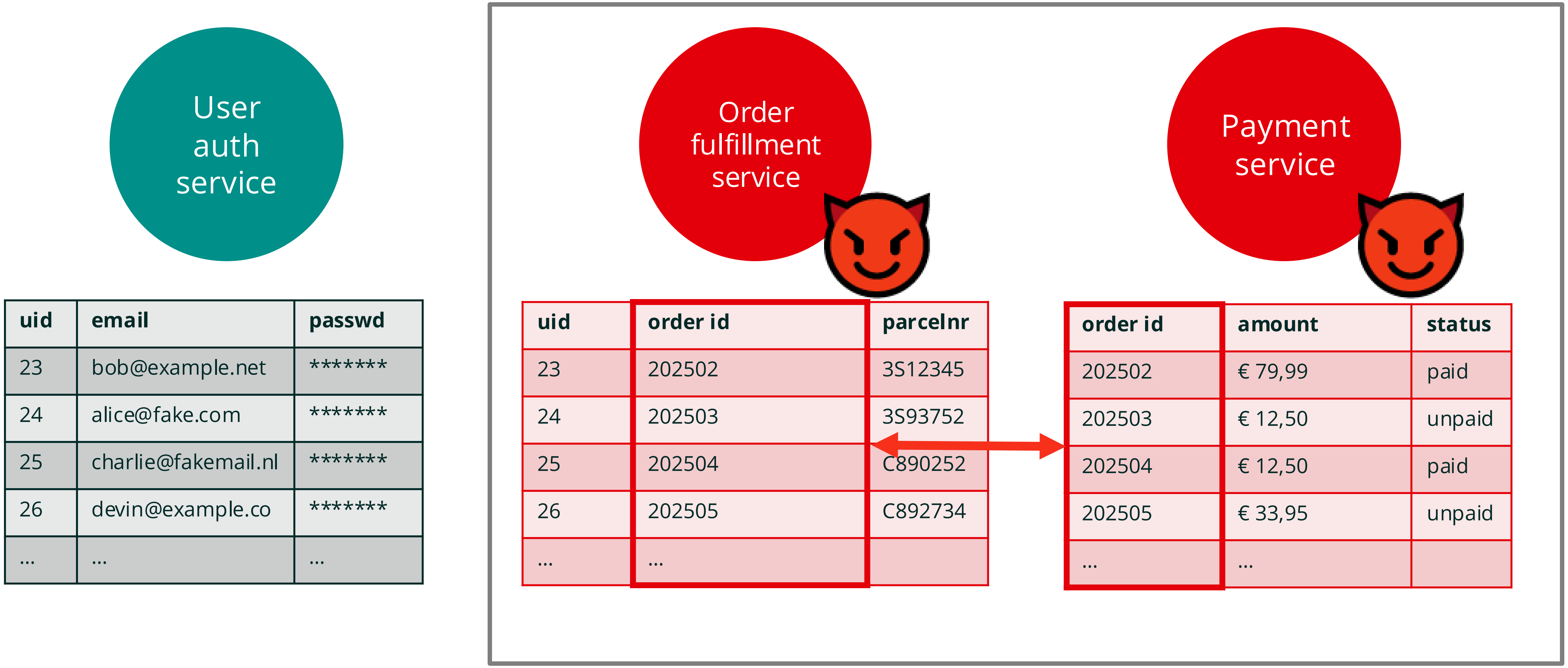
Order fulfillment service

| uid | order id | parcelnr |
|---|---|---|
| 23 | 202502 | 3S12345 |
| 24 | 202503 | 3S93752 |
| 25 | 202504 | C890252 |
| 26 | 202505 | C892734 |
| … | … | |

Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

# Data compartmentalization



User auth service

| uid | email | passwd |
|---|---|---|
| 23 | bob@example.net | ******* |
| 24 | alice@fake.com | ******* |
| 25 | charlie@fakemail.nl | ******* |
| 26 | devin@example.co | ******* |
| … | … | … |

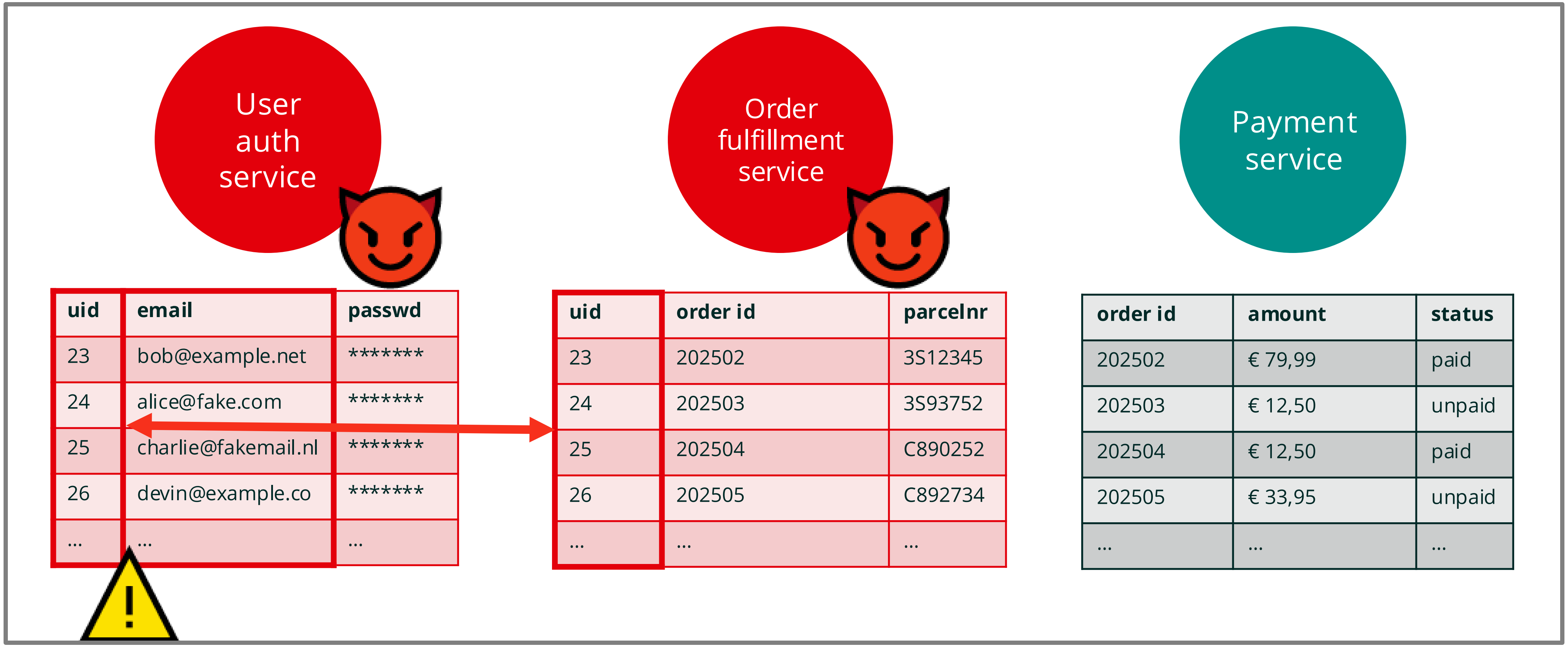
Order fulfillment service

| uid | order id | parcelnr |
|---|---|---|
| 23 | 202502 | 3S12345 |
| 24 | 202503 | 3S93752 |
| 25 | 202504 | C890252 |
| 26 | 202505 | C892734 |
| … | … | |

Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | |

# Data compartmentalization



User auth service

| uid | email | passwd |
|---|---|---|
| 23 | bob@example.net | ******* |
| 24 | alice@fake.com | ******* |
| 25 | charlie@fakemail.nl | ******* |
| 26 | devin@example.co | ******* |
| … | … | … |

Order fulfillment service

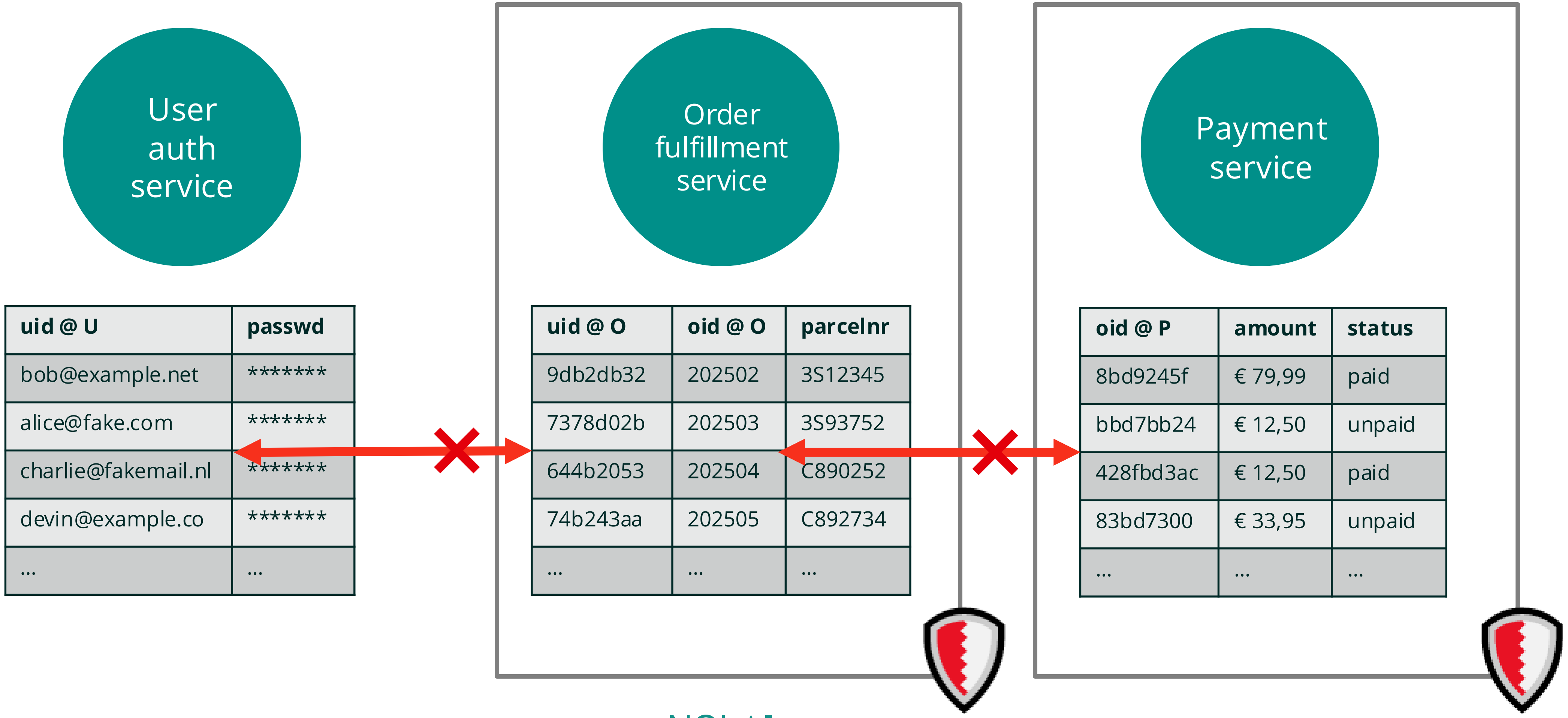
| uid | order id | parcelnr |
|---|---|---|
| 23 | 202502 | 3S12345 |
| 24 | 202503 | 3S93752 |
| 25 | 202504 | C890252 |
| 26 | 202505 | C892734 |
| … | … | … |

Payment service

| order id | amount | status |
|---|---|---|
| 202502 | € 79,99 | paid |
| 202503 | € 12,50 | unpaid |
| 202504 | € 12,50 | paid |
| 202505 | € 33,95 | unpaid |
| … | … | … |

PSEUDONYMIZATION AS A SERVICE

# Micropseudonymization



User auth service

| uid @ U | passwd |
|---|---|
| bob@example.net | ******* |
| alice@fake.com | ******* |
| charlie@fakemail.nl | ******* |
| devin@example.co | ******* |
| … | … |

Order fulfillment service

| uid @ O | oid @ O | parcelnr |
|---|---|---|
| 9db2db32 | 202502 | 3S12345 |
| 7378d02b | 202503 | 3S93752 |
| 644b2053 | 202504 | C890252 |
| 74b243aa | 202505 | C892734 |
| … | … | … |

Payment service

| oid @ P | amount | status |
|---|---|---|
| 8bd9245f | € 79,99 | paid |
| bbd7bb24 | € 12,50 | unpaid |
| 428fbd3ac | € 12,50 | paid |
| 83bd7300 | € 33,95 | unpaid |
| … | … | … |

# Micropseudonymization

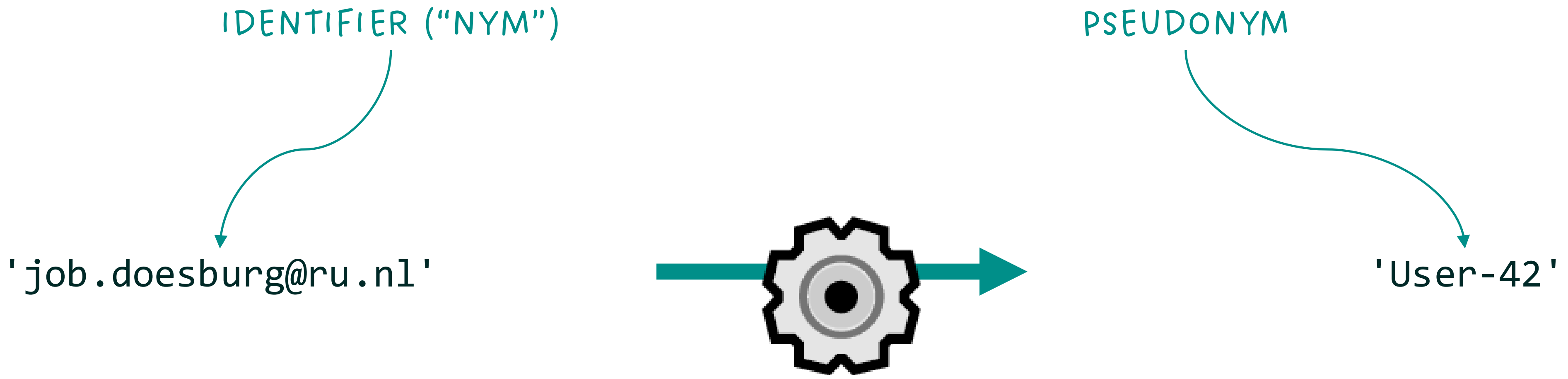
- Each (sub)system uses its own pseudonyms, instead of shared identifiers
  - Functional compartments aligned with data compartments

- Pseudonymization *by default* prevents data linkage (= *Privacy by Design*)
  - Assuming no quasi-identifiers in the data
  - But doing this very granularly actually eliminates quasi-identifiers!

- **When system gets compromised, its data remains unlinkable**
  **⇒ data compartmentalization**

# PSEUDONYMIZATION

# Pseudonymization



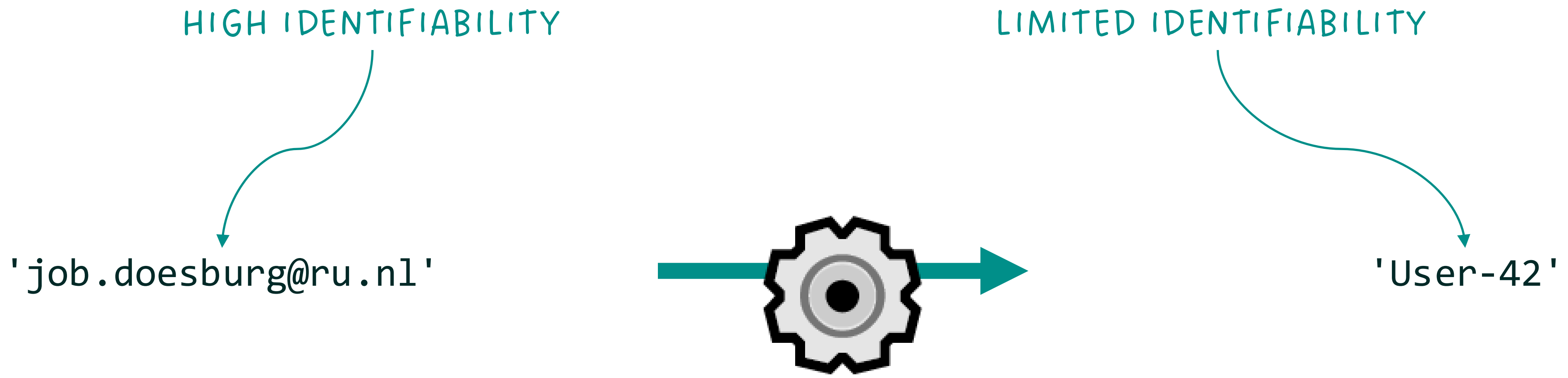
IDENTIFIER ("NYM")

'job.doesburg@ru.nl'

PSEUDONYM

'User-42'

# Pseudonymization



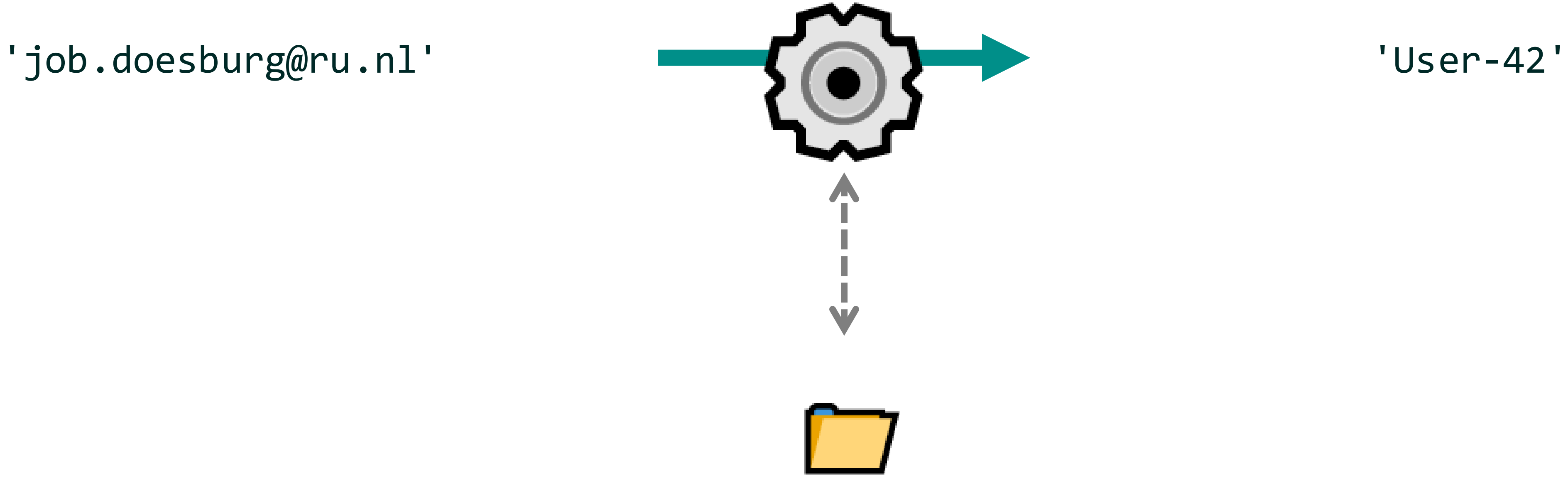
HIGH IDENTIFIABILITY

'job.doesburg@ru.nl'

LIMITED IDENTIFIABILITY

'User-42'

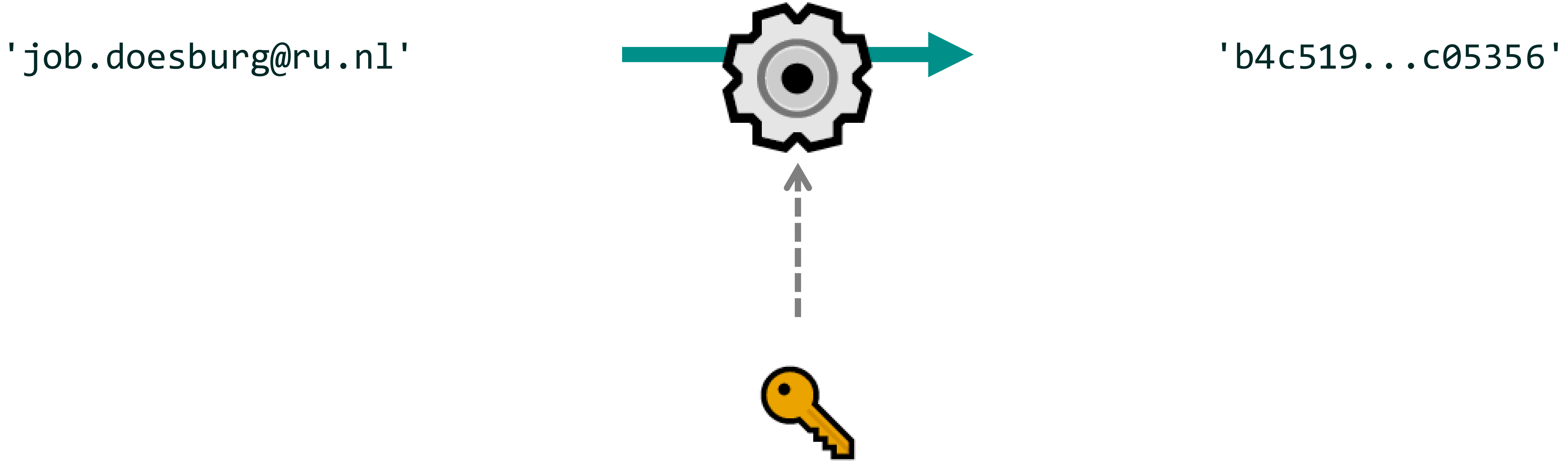
# Mapping-table pseudonymization



'job.doesburg@ru.nl'

'User-42'

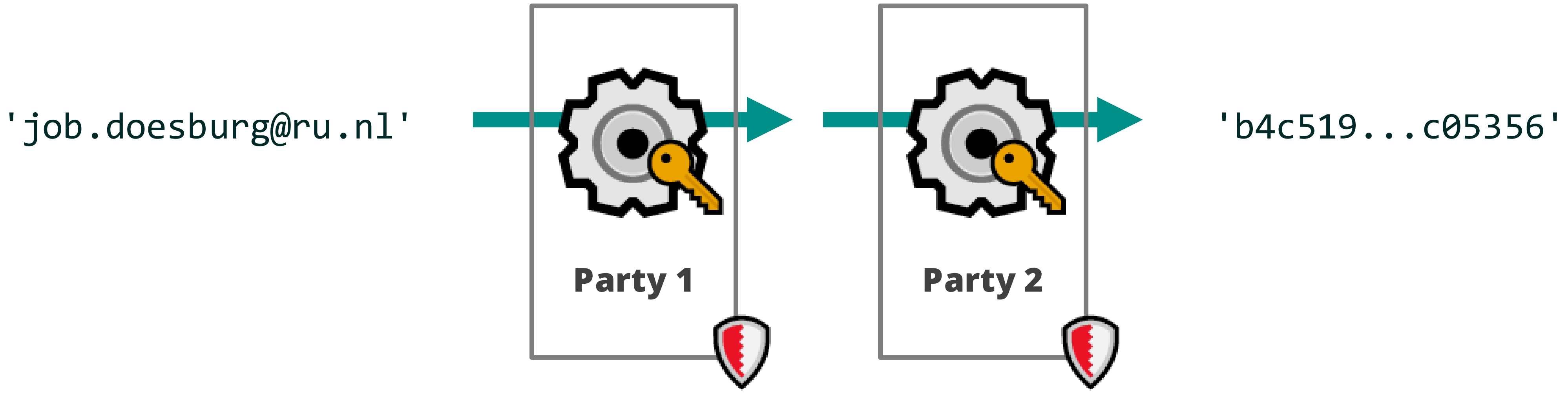
# Random (cryptographic) pseudonymization



'job.doesburg@ru.nl'

'b4c519...c05356'

# Distributed pseudonymization



'job.doesburg@ru.nl'

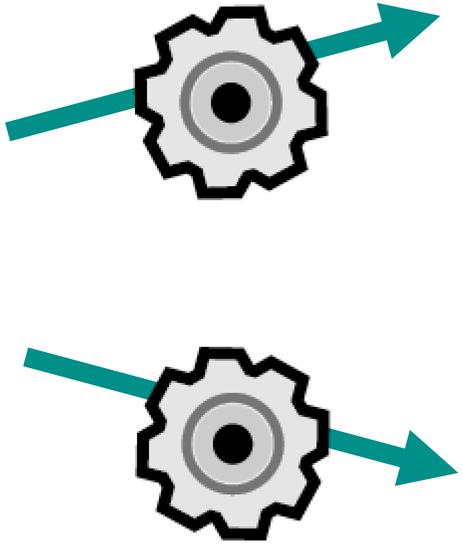
Party 1

Party 2

'b4c519...c05356'

# Pseudonymization to multiple domains

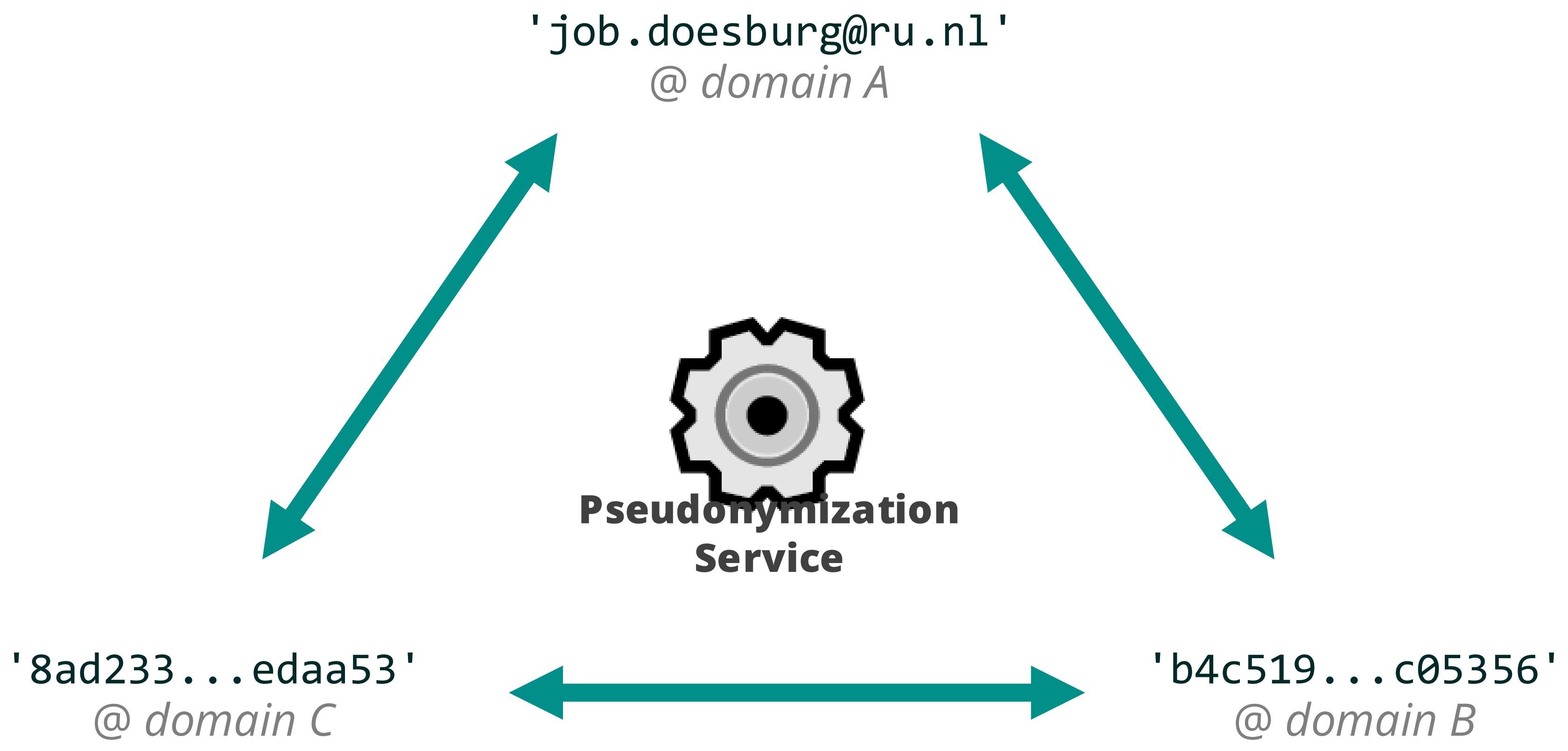
'job.doesburg@ru.nl'
*@ domain A*



'b4c519...c05356'
*@ domain B*

'8ad233...edaa53'
*@ domain C*

# Transitive pseudonymization



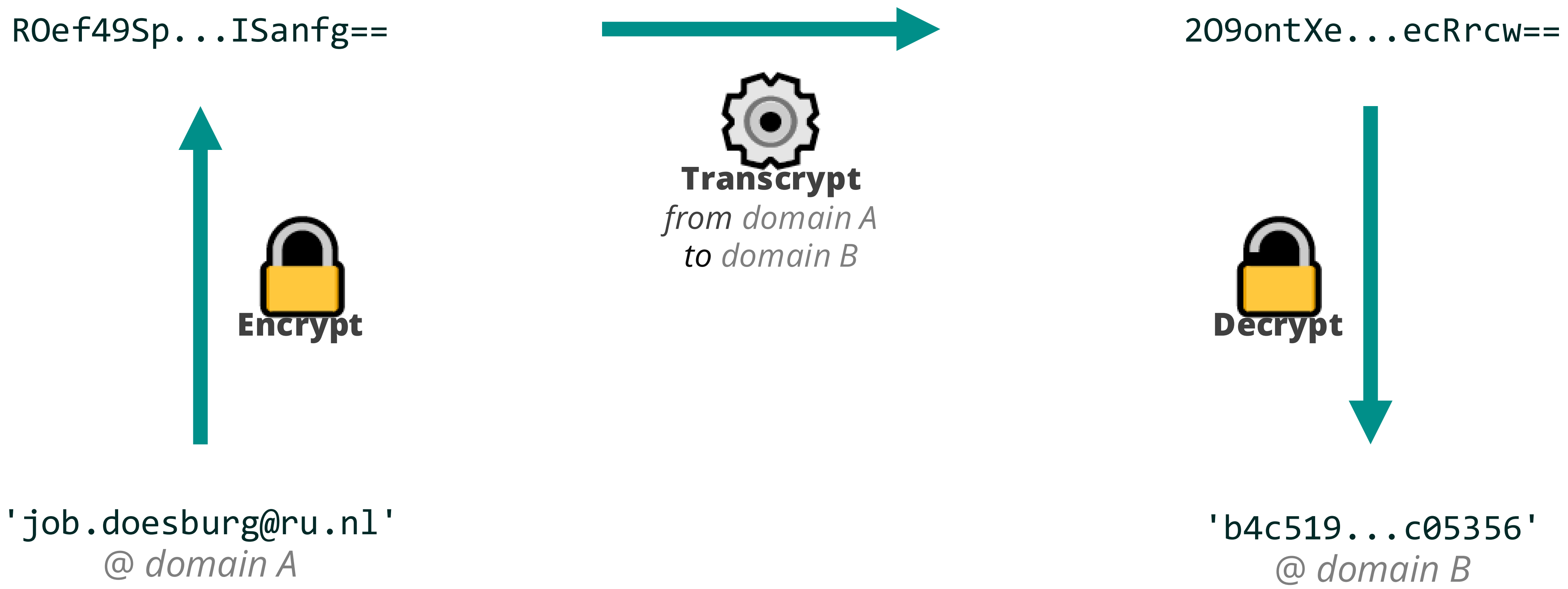
'job.doesburg@ru.nl'
*@ domain A*

**Pseudonymization Service**

'8ad233...edaa53'
*@ domain C*

'b4c519...c05356'
*@ domain B*

# Blind pseudonymization



'job.doesburg@ru.nl'
*@ domain A*

**Encrypt**

`ROef49Sp...ISanfg==`

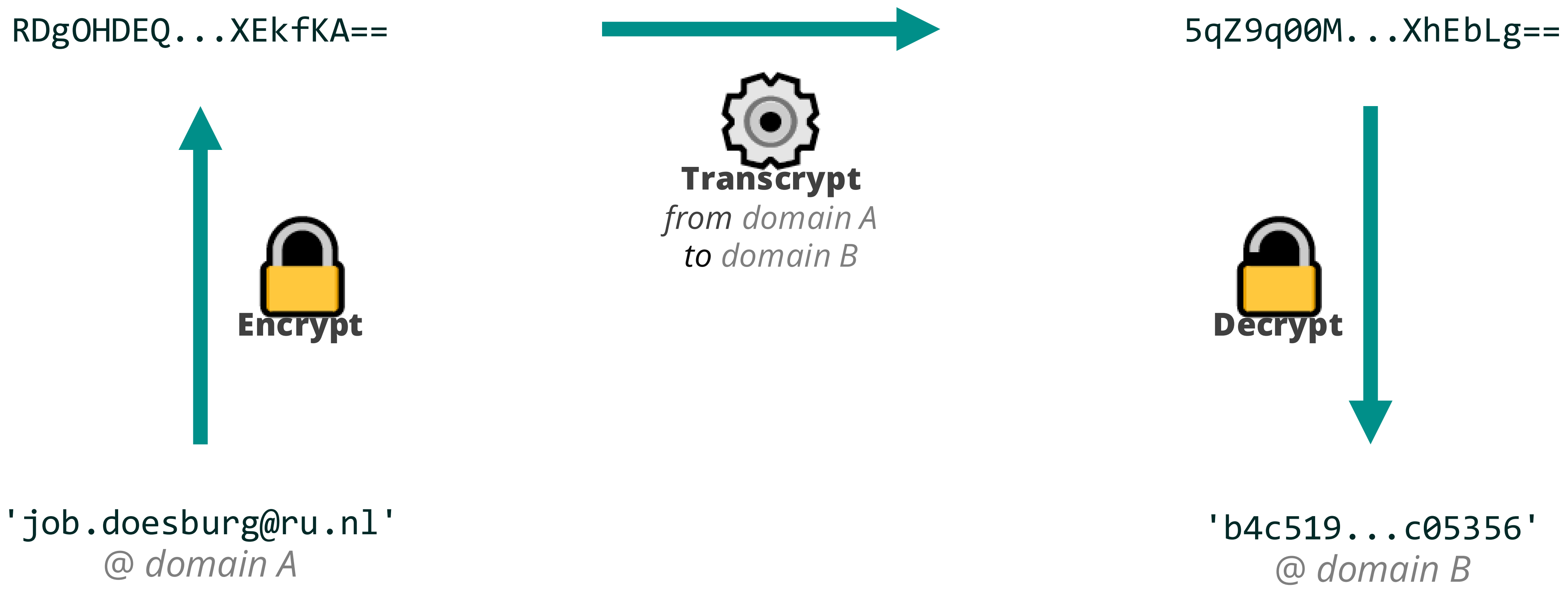
**Transcrypt**
*from domain A*
*to domain B*

`2O9ontXe...ecRrcw==`

**Decrypt**

'b4c519...c05356'
*@ domain B*

# Polymorphic pseudonymization



RDgOHDEQ...XEkfKA==

5qZ9q00M...XhEbLg==

**Encrypt**

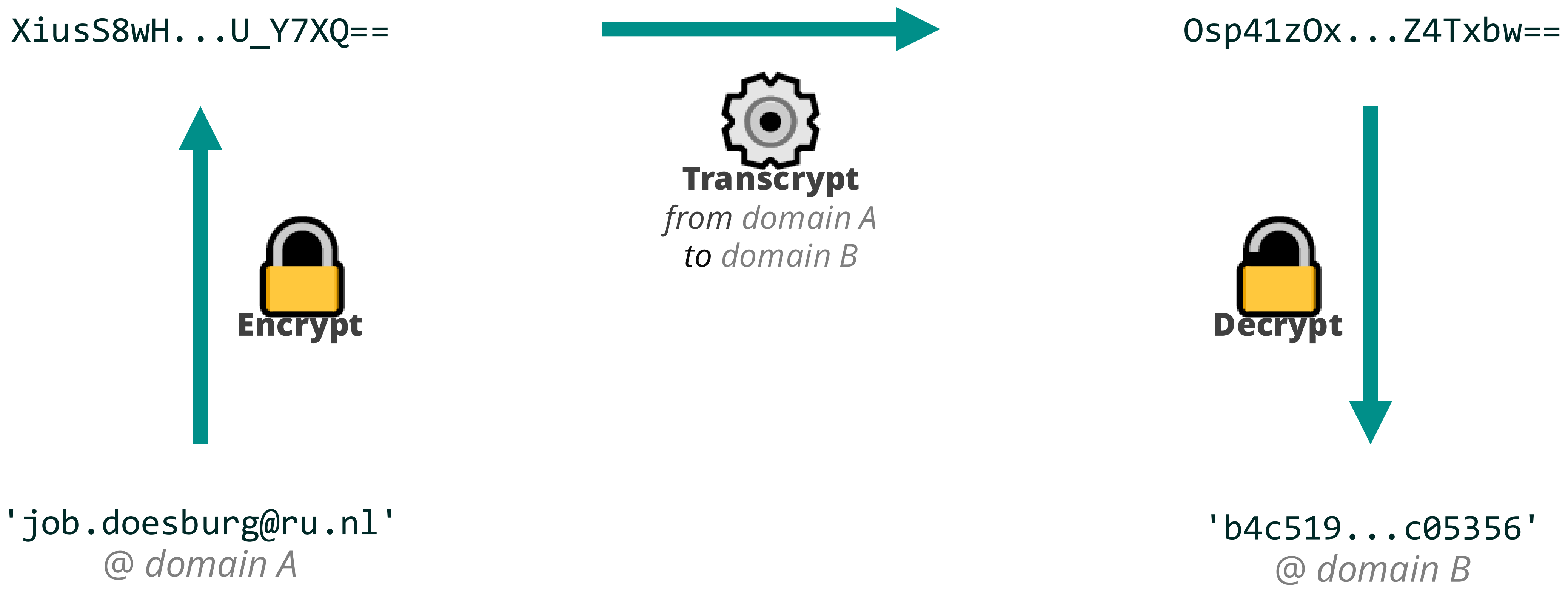
**Transcrypt**
*from domain A*
*to domain B*

**Decrypt**

'job.doesburg@ru.nl'
*@ domain A*

'b4c519...c05356'
*@ domain B*

# Polymorphic pseudonymization



XiusS8wH...U_Y7XQ==

Encrypt

'job.doesburg@ru.nl'
*@ domain A*

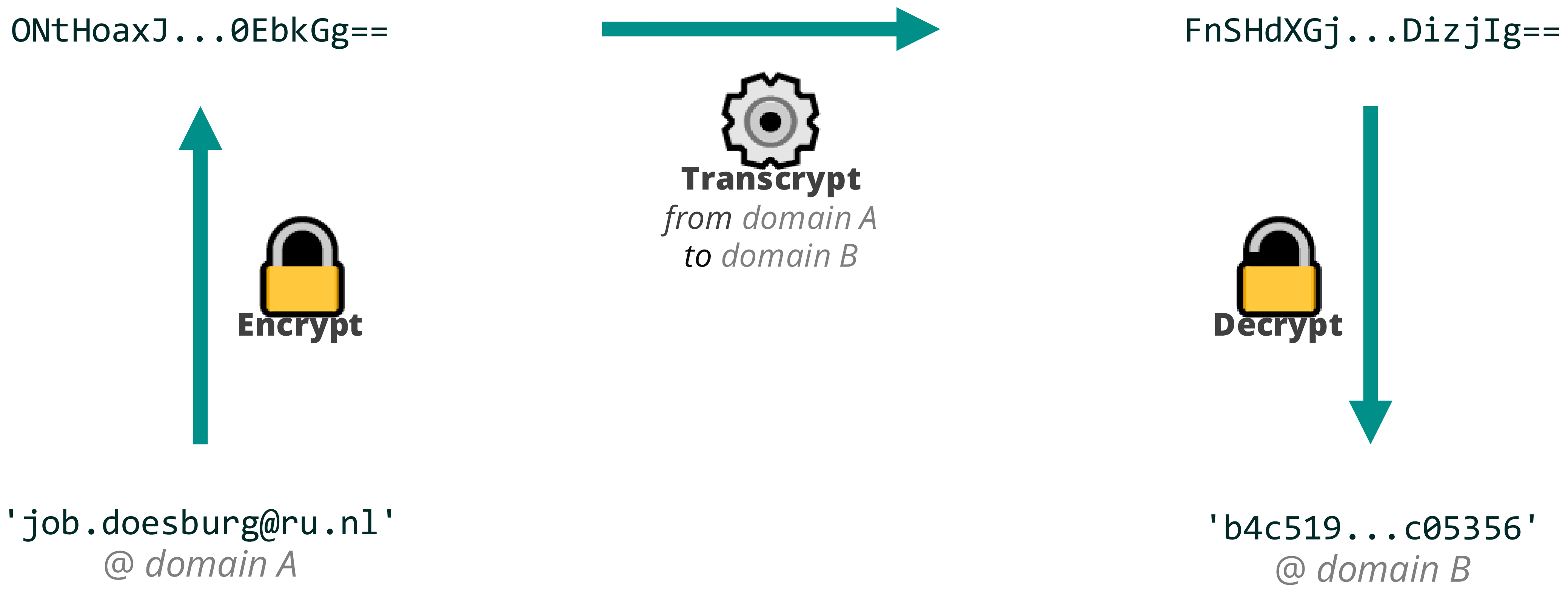
**Transcrypt**
*from domain A*
*to domain B*

Osp41zOx...Z4Txbw==

Decrypt

'b4c519...c05356'
*@ domain B*

# Polymorphic pseudonymization



ONtHoaxJ...0EbkGg==

FnSHdXGj...DizjIg==

**Encrypt**

**Transcrypt**
*from domain A*
*to domain B*

**Decrypt**

'job.doesburg@ru.nl'
*@ domain A*

'b4c519...c05356'
*@ domain B*

# Transitive + blind + polymorphic + distributed = 'secure' pseudonymization

Can pseudonymize in any direction without linking

Doesn't see what values it's pseudonymizing

Doesn't see if it's pseudonymizing the same thing twice

No single party owns the full pseudonymization key

PSEUDONYMIZATION AS A SERVICE

# The PEP framework

- Polymorphic encryption and pseudonymization (Verheul & Jacobs, 2017)

- Current applications in:
  - DigiD Hoog (the Dutch national eID system)
  - PEP Responsible Data Sharing Repository (medical research)
  - PubHubs (pseudonymous social networking)

BSNk PP

DigiD

Responsible Data Sharing Repository

# PSEUDONYMIZATION AS A SERVICE

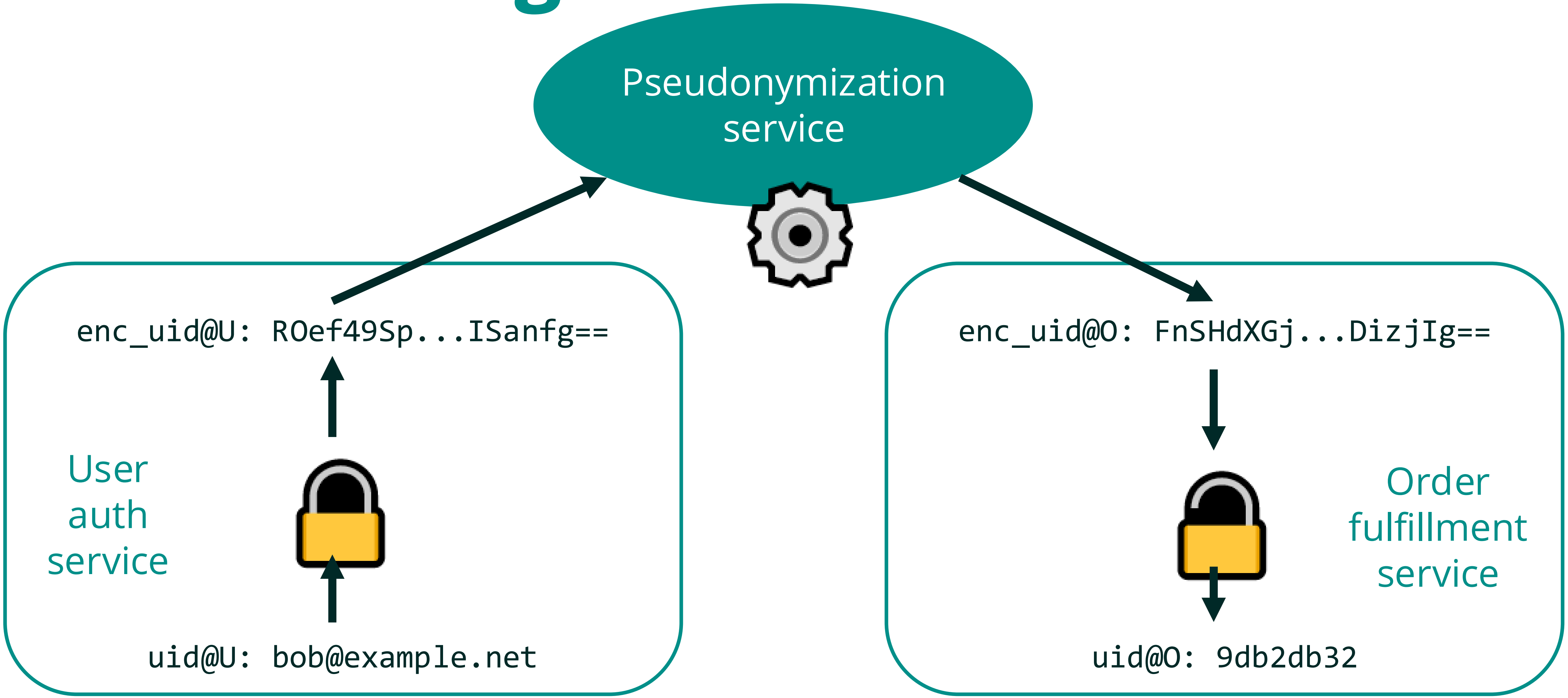
# Pseudonymization *as a Service*

- Central pseudonymization service can blindly *monitor* and *control* data linkage
- Integrating a new subsystem only requires updating access rules
- **Keep grip on data in an increasingly complex architecture**
  **⇒ enable privacy-preserving system evolution**
- **Retrofitting is possible**: start with migrating one subsystem to use pseudonyms, slowly compartmentalize more systems

PSEUDONYMIZATION AS A SERVICE

# Data exchange

**Event: user starts authenticated session**

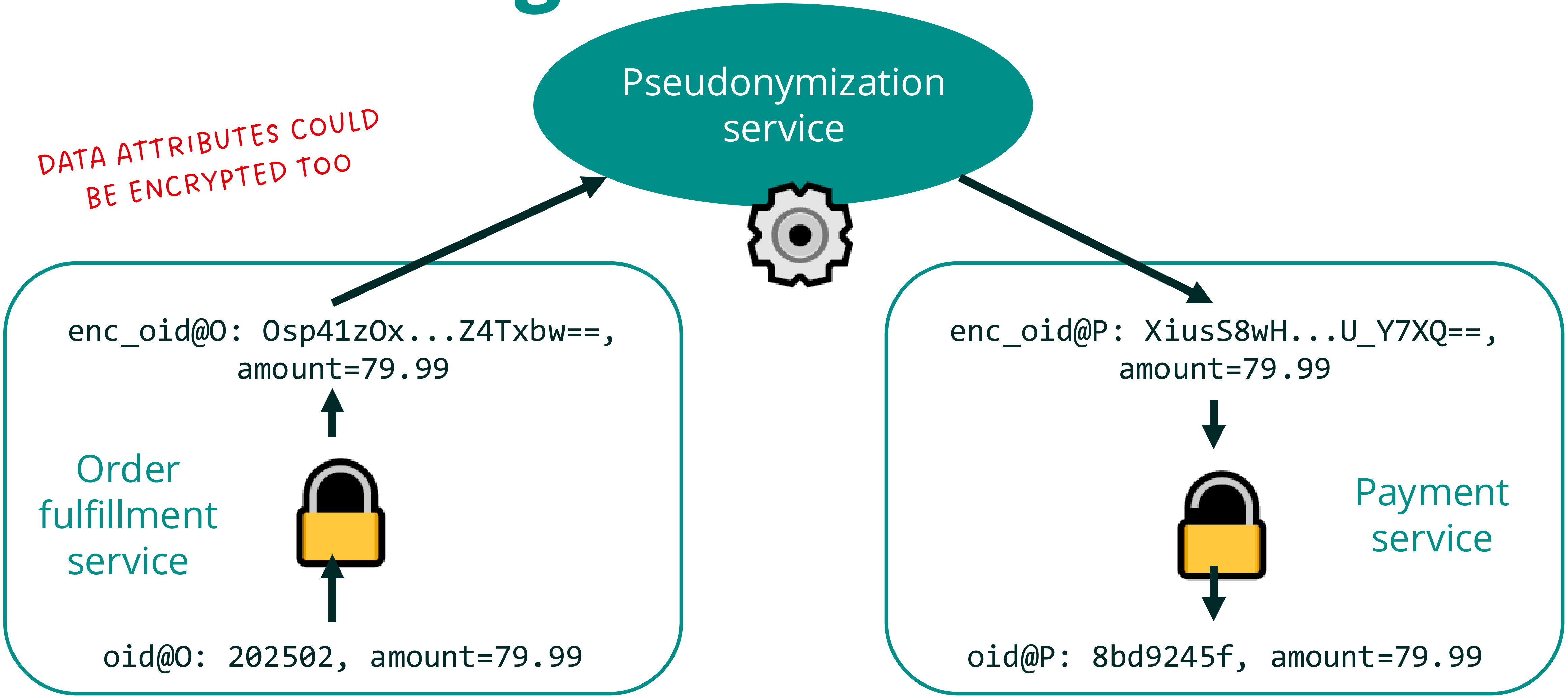


Pseudonymization service
enc_uid@U: ROef49Sp...ISanfg==
User auth service
uid@U: bob@example.net
enc_uid@O: FnSHdXGj...DizjIg==
Order fulfillment service
uid@O: 9db2db32

PSEUDONYMIZATION AS A SERVICE

# Data exchange

**Event: user places order**



DATA ATTRIBUTES COULD BE ENCRYPTED TOO

```
enc_oid@O: Osp41zOx...Z4Txbw==,
amount=79.99

oid@O: 202502, amount=79.99
```

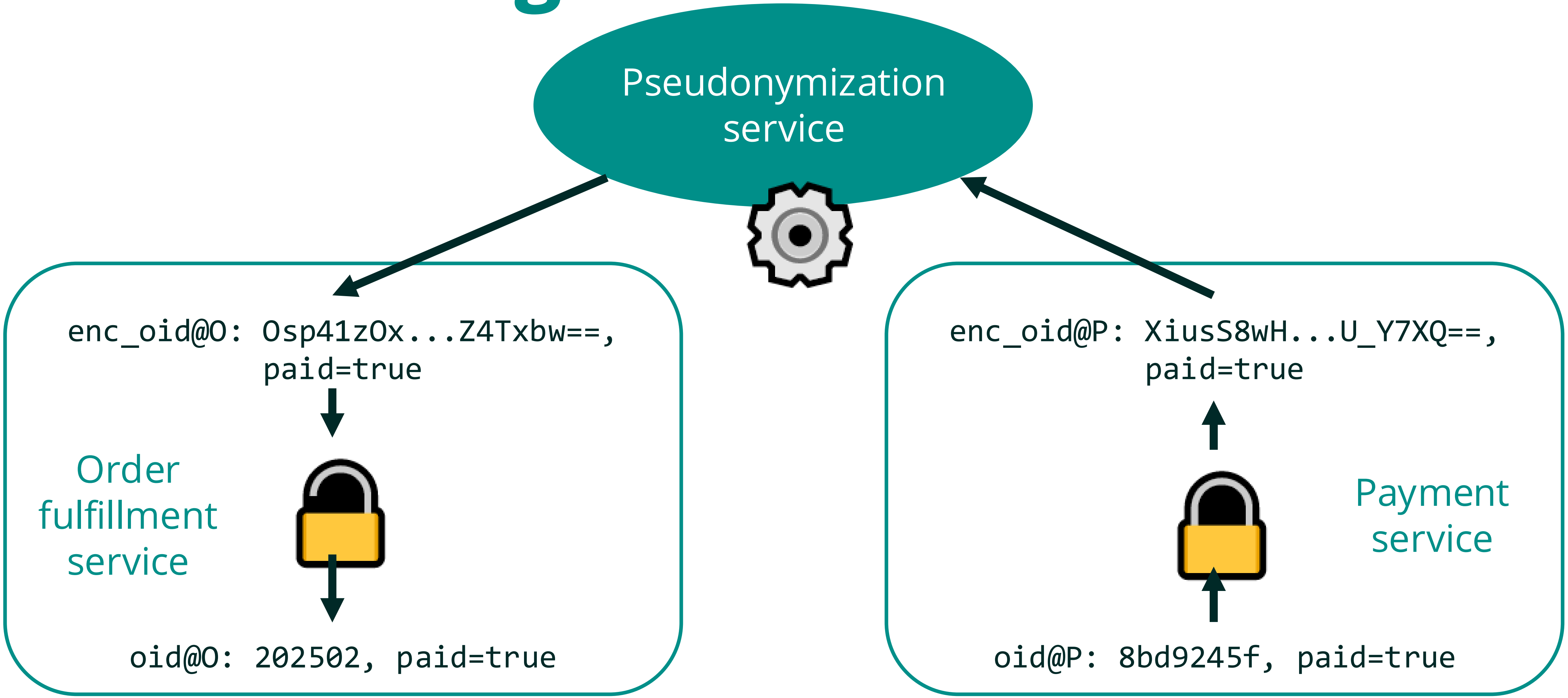
```
enc_oid@P: XiusS8wH...U_Y7XQ==,
amount=79.99

oid@P: 8bd9245f, amount=79.99
```

PSEUDONYMIZATION AS A SERVICE

# Data exchange

**Event: payment received**



Pseudonymization service

Order fulfillment service

```
enc_oid@O: Osp41zOx...Z4Txbw==,
           paid=true
```

```
oid@O: 202502, paid=true
```

Payment service

```
enc_oid@P: XiusS8wH...U_Y7XQ==,
           paid=true
```

```
oid@P: 8bd9245f, paid=true
```

PSEUDONYMIZATION AS A SERVICE

# NOLAI research data platform

- **Evolving platform**: new data sources are connected frequently
  - Many partners, many systems, rapid development: high chance of compromise
  - Micropseudonymization limits impact of potential data breaches
- **Integrating new/existing systems is relatively easy**
  e.g. via API wrapper that converts identifiers (such as existing LimeSurvey service)
- **Centralized data governance** despite decentralized system architecture

PSEUDONYMIZATION AS A SERVICE

# Take-aways

**Open-source development ongoing on GitHub**
Server in Rust, clients in Rust, JS/TS & Python

https://github.com/NOLAI/paas-server

- The **impact** of data breaches depends on the linkability/identifiability of data
- Minimize data linkability with **micropseudonymization**:
  use different pseudonyms for data in different subsystems
- A blind, transitive, **central pseudonymization service** can securely convert pseudonyms between subsystems, while monitoring and controlling data linkage
- ***Pseudonymization by default*** **enables privacy-preserving system evolution**

# Future work

- **Pseudonymization service integrity** using zero-knowledge proofs
- **Data subject authentication** to allow users to prove their pseudonymous identity
- **Distributed tracing** with pseudonymization of trace IDs