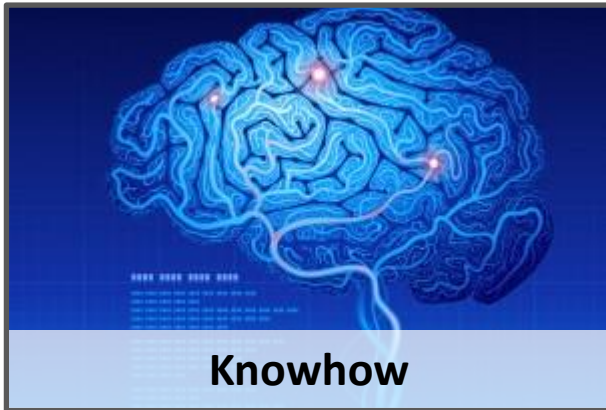


Linking Together Personal Data in the Era of Big Data & GDPR



Smals

ICT for society



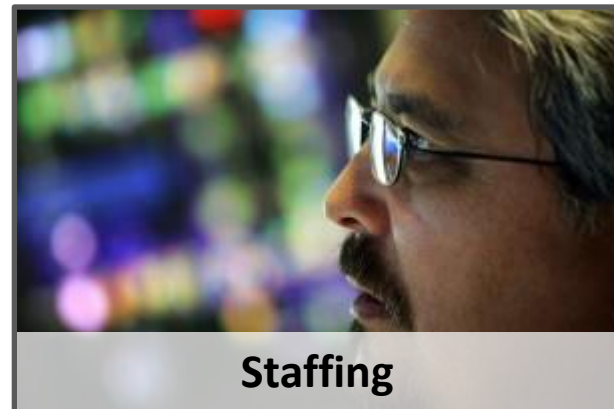
Knowhow



Development



Infrastructure



Staffing



E-gov award
AGORIA



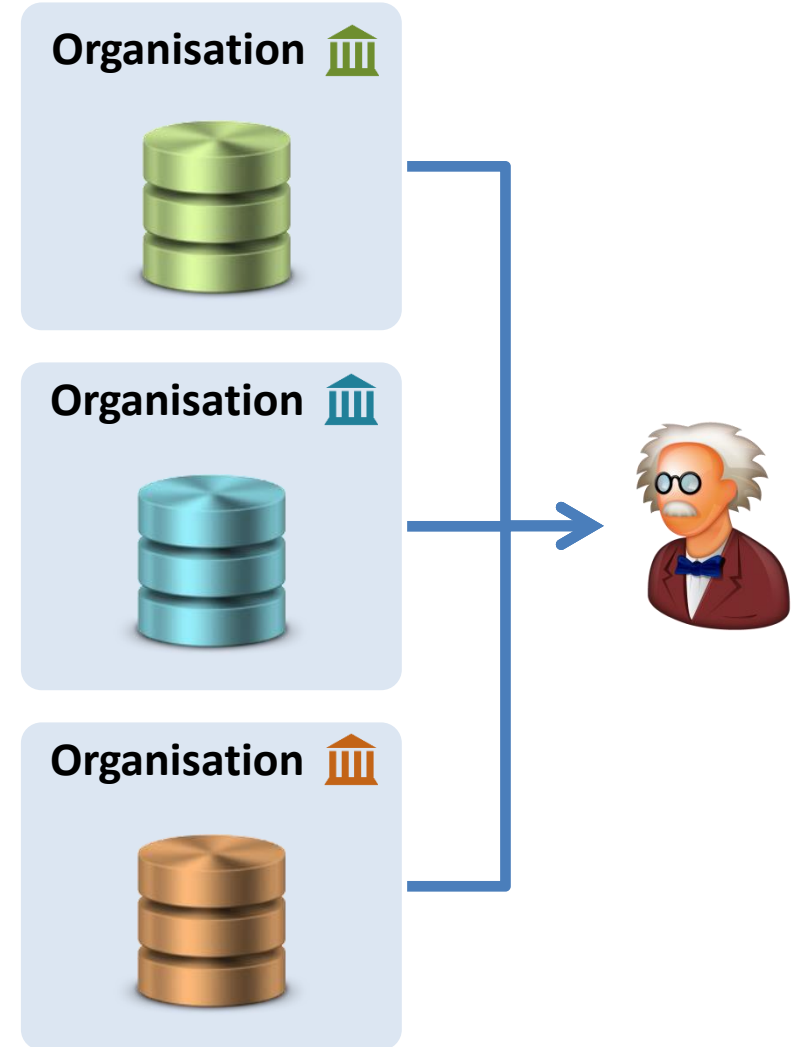
CERTIFIED EXCELLENCE IN EMPLOYEE CONDITIONS

Linking Together Personal Data

A fictional example

A research team wants to analyse medical, financial and demographic data from all citizens born in or after 1990 with a wage of at least € 50 000 per year who are self-employed as secondary activity.

However, these data are maintained by separate governmental organisations and, hence, need to be linked together.



GDPR

Anonymous Data

Unlinkable
to natural person

GDPR not
applicable

Complex (big) data cannot
be anonymised without
rendering the data useless

Pseudonymous Data (NEW)

Linkable with
additional info
to natural person

GDPR applicable
some provisions relaxed
Encouraged by GDPR

Easier to use for
secondary purposes and
for scientific, historical or
statistical research

Identified Data

Linkable without
additional info
to natural person

GDPR fully
applicable

Only for specific, explicit
and legitimate purposes
(purpose limitation)



Reidentification risk (more left is better)

Key Pseudonyms

68.08.05-078.47	↔	1F1tAaz5x1HU	Man	Hypertension
83.01.25-123.77	↔	3BcMuv1VJqmw	Woman	Schizophrenia
76.04.18-042.23	↔	1Nf311Qb8rLD	Man	Pneumonia

 ,  and  are pseudonyms

Attribute Pseudonyms (Indirect Identifiers)



4710 Man 05/08/1968 Hypertension



8434 Woman 25/01/1983 Schizophrenia



1050 Man 18/04/1976 Pneumonia

Extra information

id	ZIP	Sex	DoB
	4710	M	05/08/1968
	8434	F	25/01/1983
	1050	M	18/04/1976

**ZIP+Sex+DoB
is a pseudonym**

Linking Together Personal Data

A fictional example

*A **research team** wants to analyse medical, financial and demographic data from all **citizens** born in or after 1990 with a wage of at least € 50 000 per year who are self-employed as secondary activity.*

*However, these data are maintained by separate **governmental organisations** and, hence, need to be linked together.*

Scientists
Analyse data sets

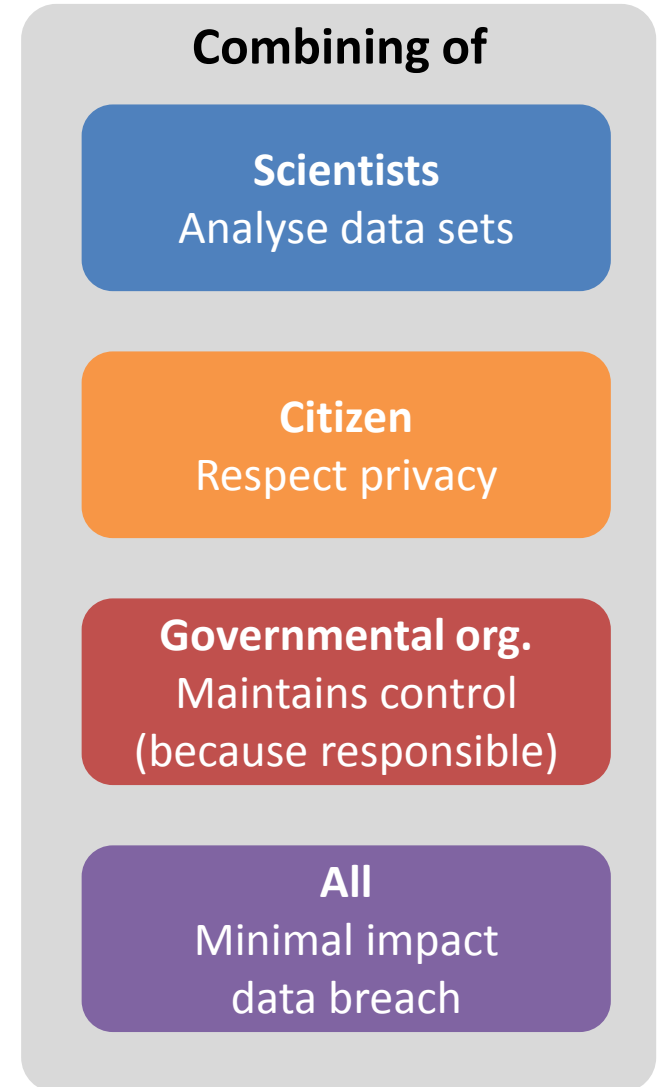
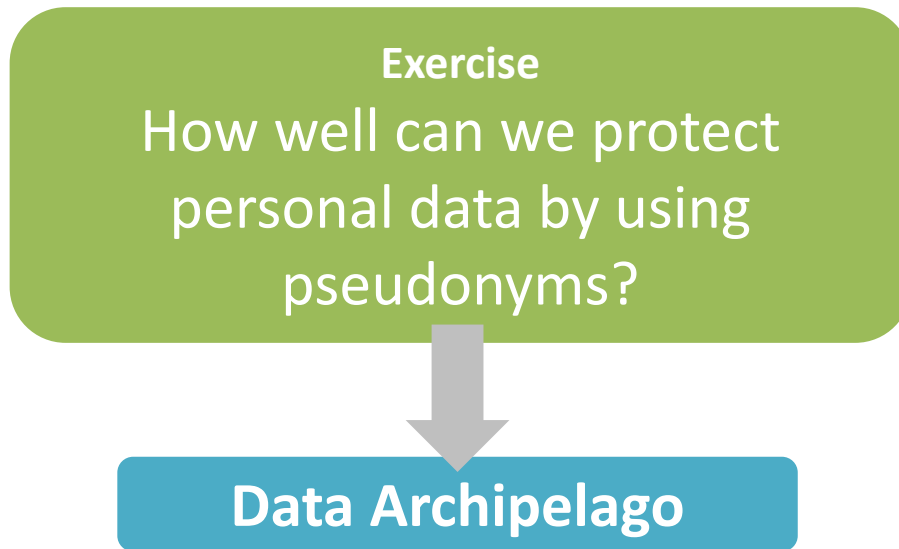
Citizen
Respect privacy

Governmental org.
Maintains control
(because responsible)

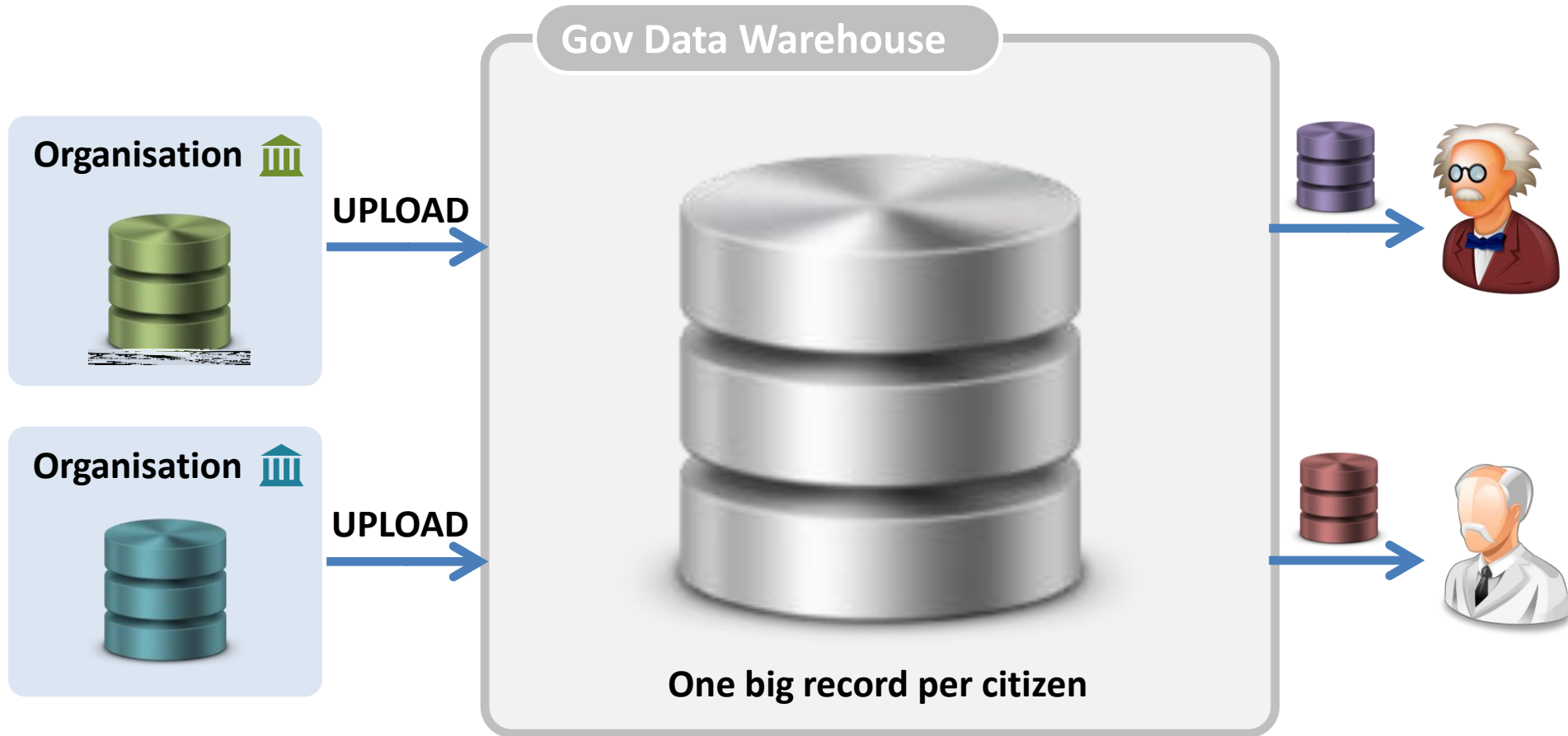
All
Minimal impact
data breach

Linking Together Personal Data

A fictional example



The Naive Approach



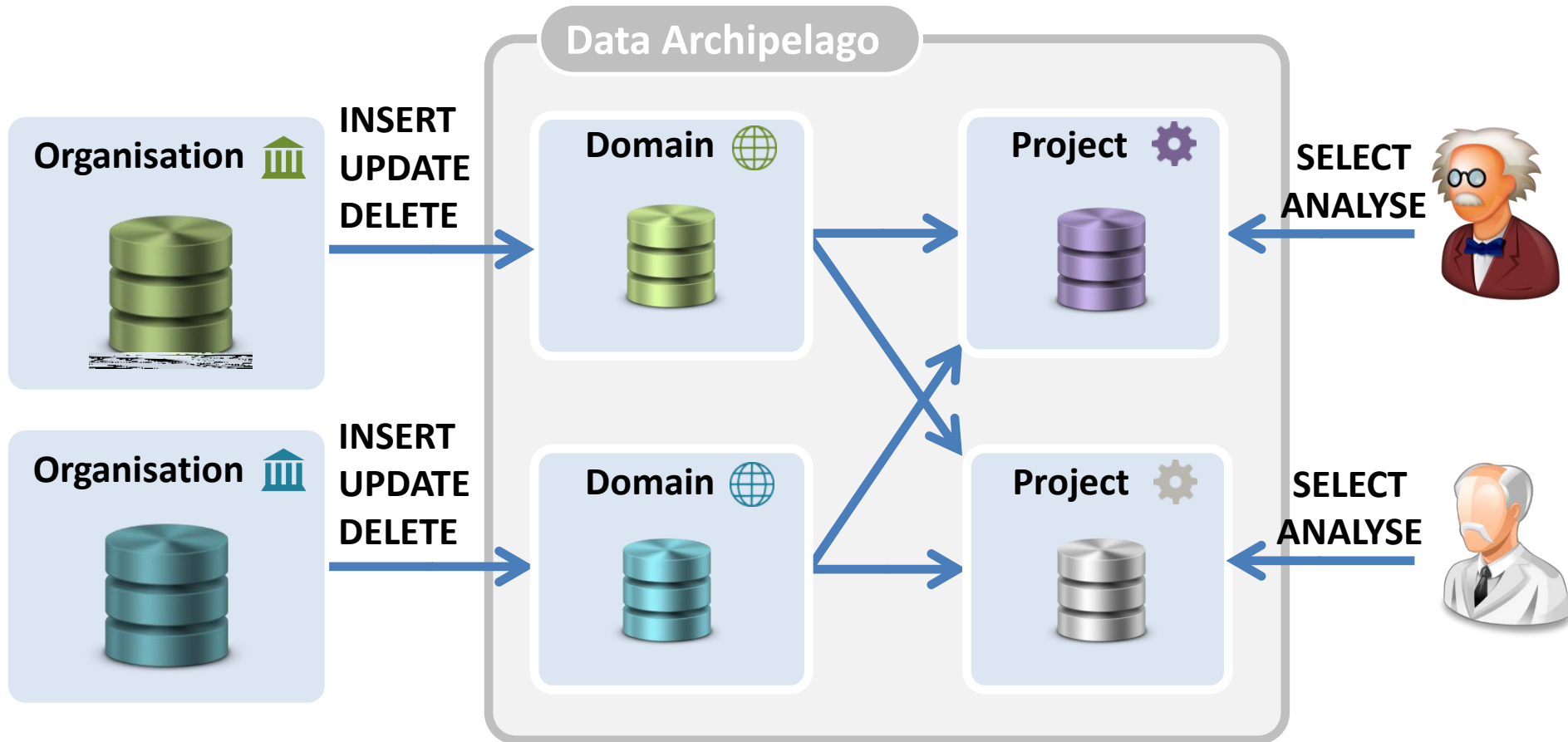
Governmental
organisation no control

Data breach
dramatic

Privacy
risks

Easy linking
together data

Concept



Domain

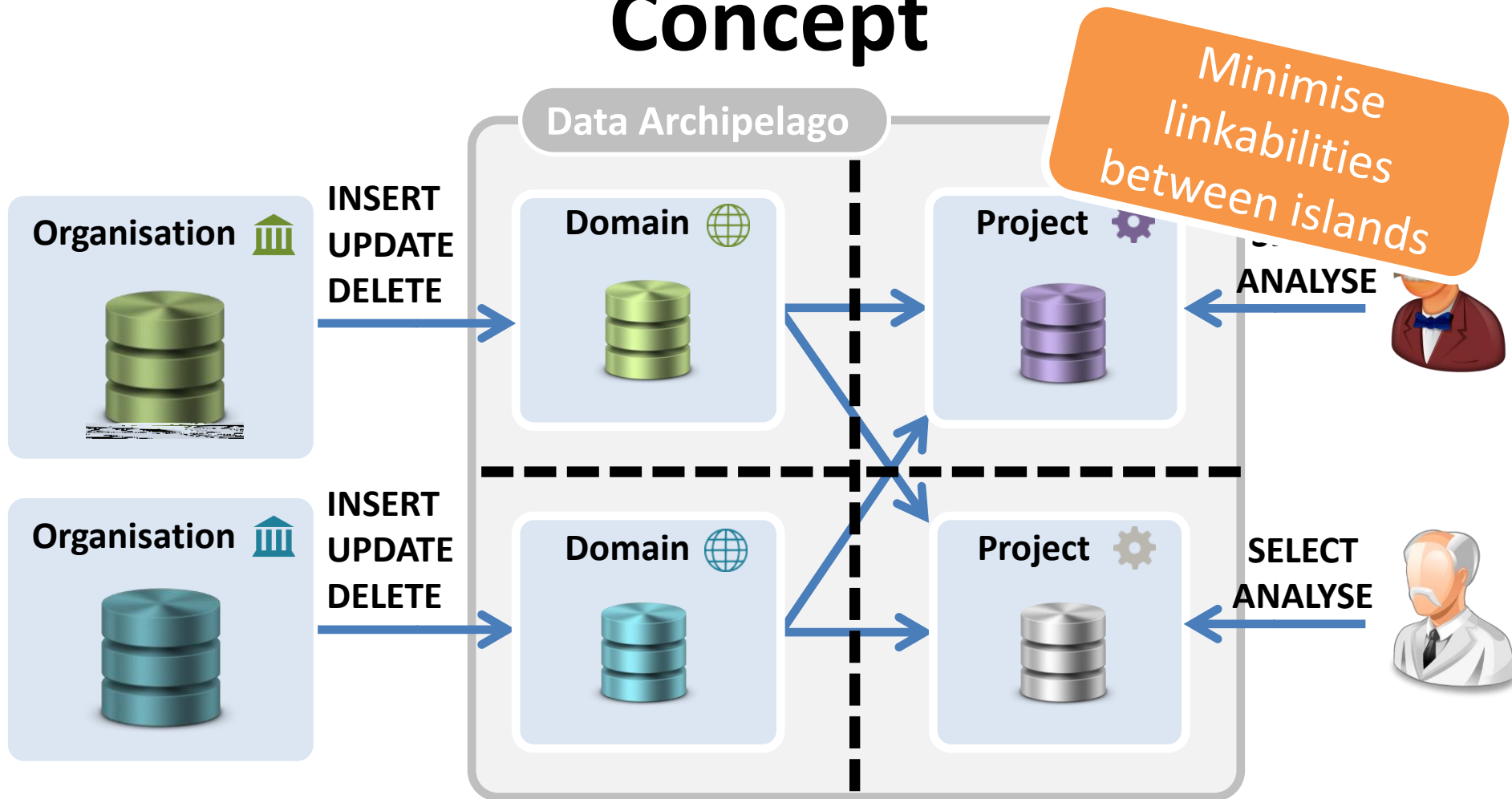
- Managed & controlled by one organisation
- Permanent
- low performance requirements



Project

- Receives minimal required data
- Access control & monitoring
- Temporal
- High performance

Concept



Maximal control
gov. organisation

Smaller impact
if data breach

Better
privacy

Easy linking
together data

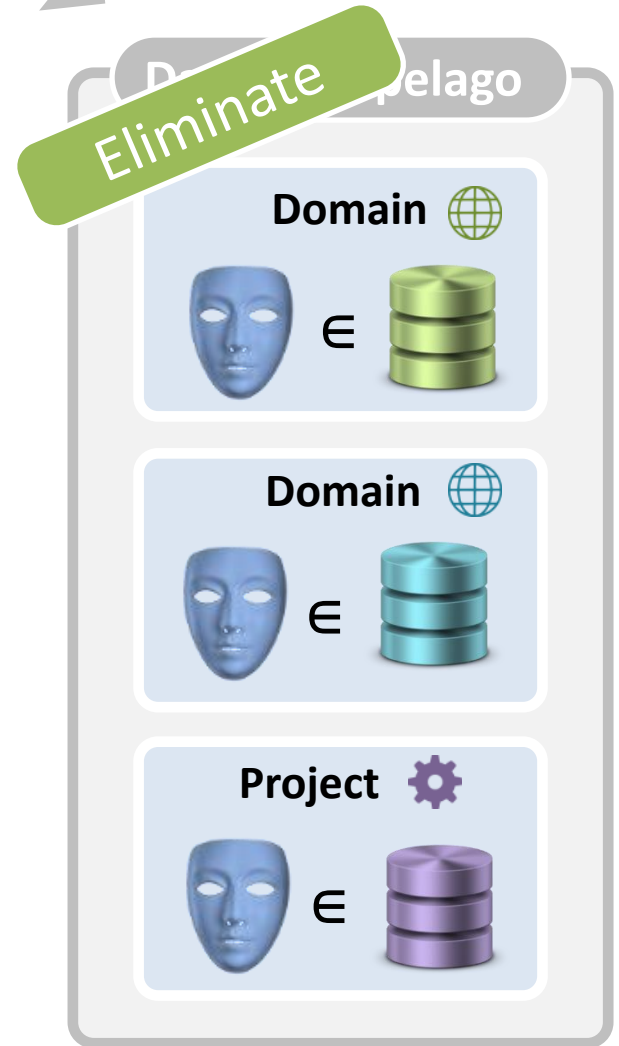
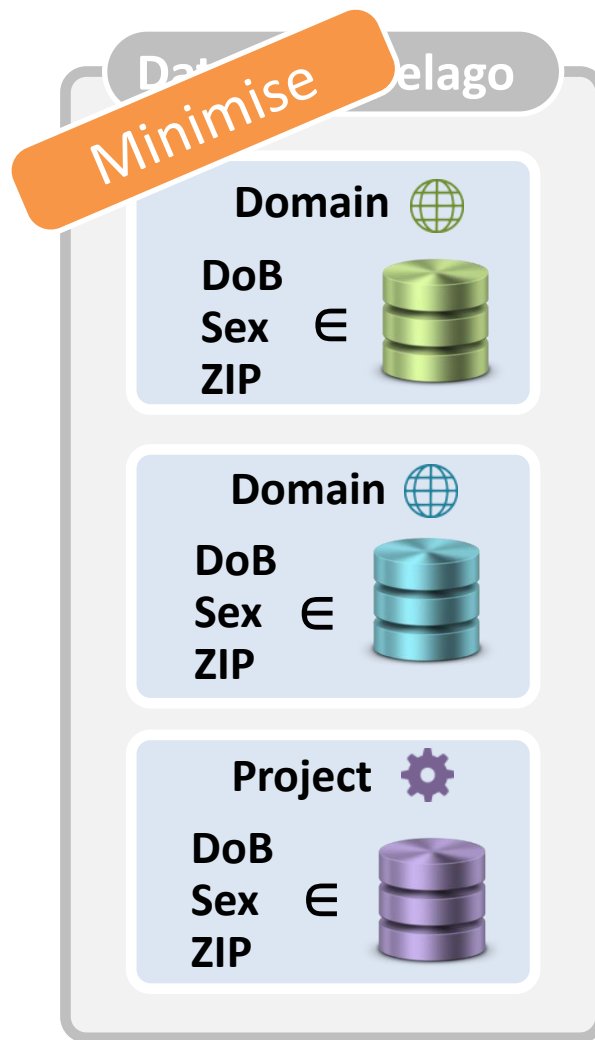
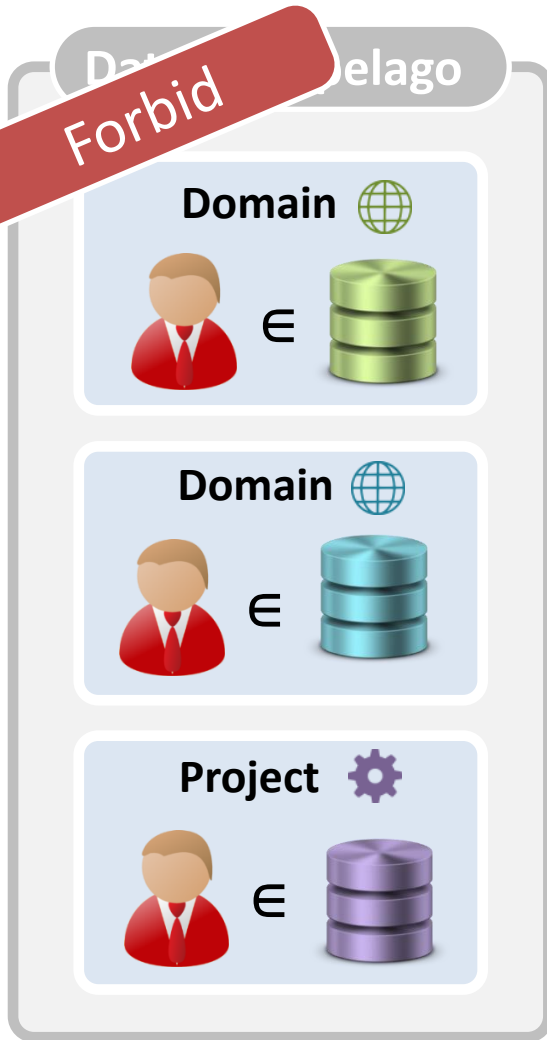
Linkability



Forbid

Minimise

Eliminate



Linkability



Forbid

Minimise

Eliminate

Domain 



€



Domain 



€



Project 



€



Domain 

DoB

Sex

ZIP

€



Domain 

DoB

Sex

ZIP

€



Project 

DoB

Sex

ZIP

€



Domain 



€



Domain 



€



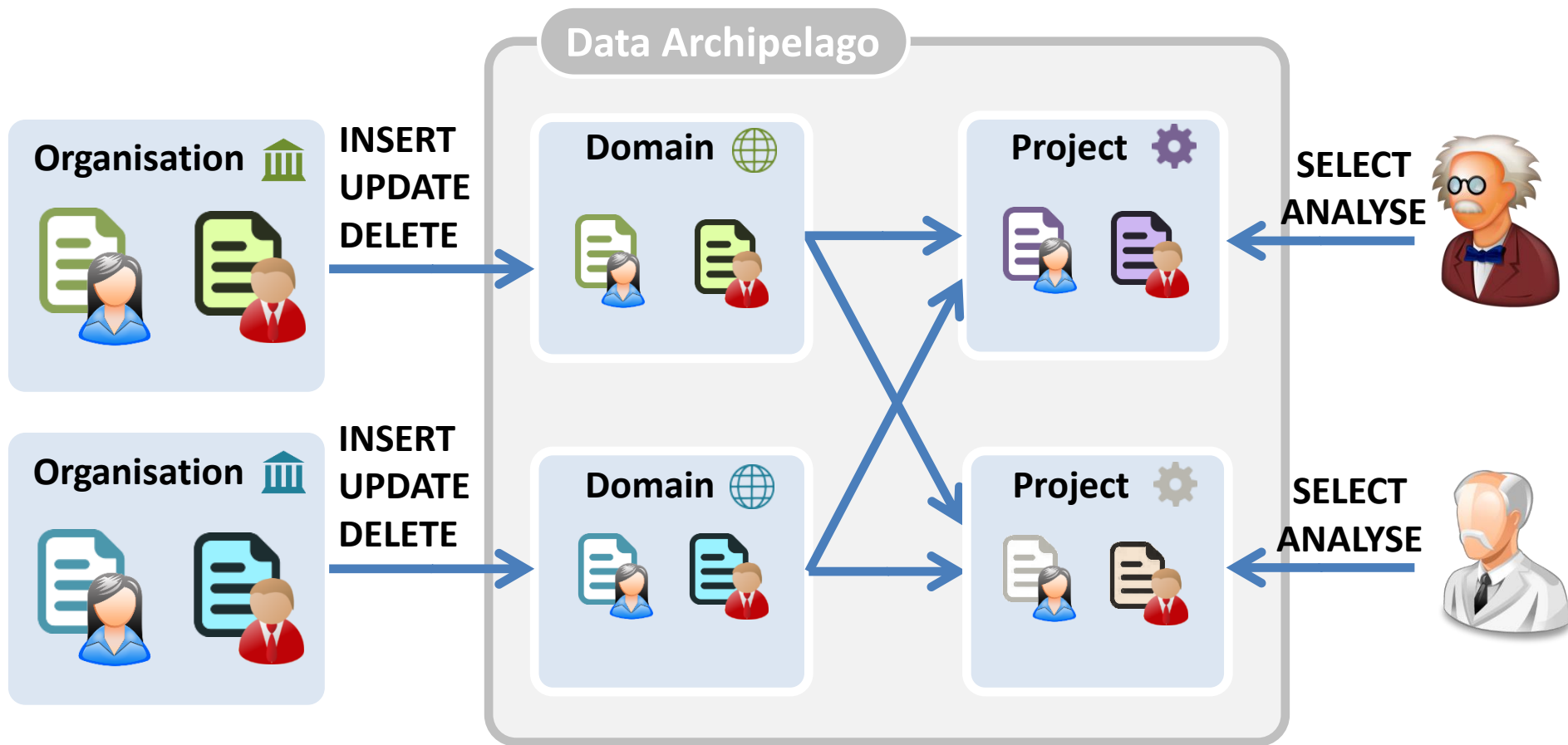
Project 



€



Linkabilities with Identifiers



Every island knows the identifier of citizen

GDPR fully applicable → Not a good idea

Linkability



Forbid

Minimise

Eliminate

Domain 



€



Domain 



€



Project 



€



Domain 

DoB
Sex
ZIP

€



Domain 

DoB
Sex
ZIP

€



Project 

DoB
Sex
ZIP

€



Domain 



€



Domain 



€



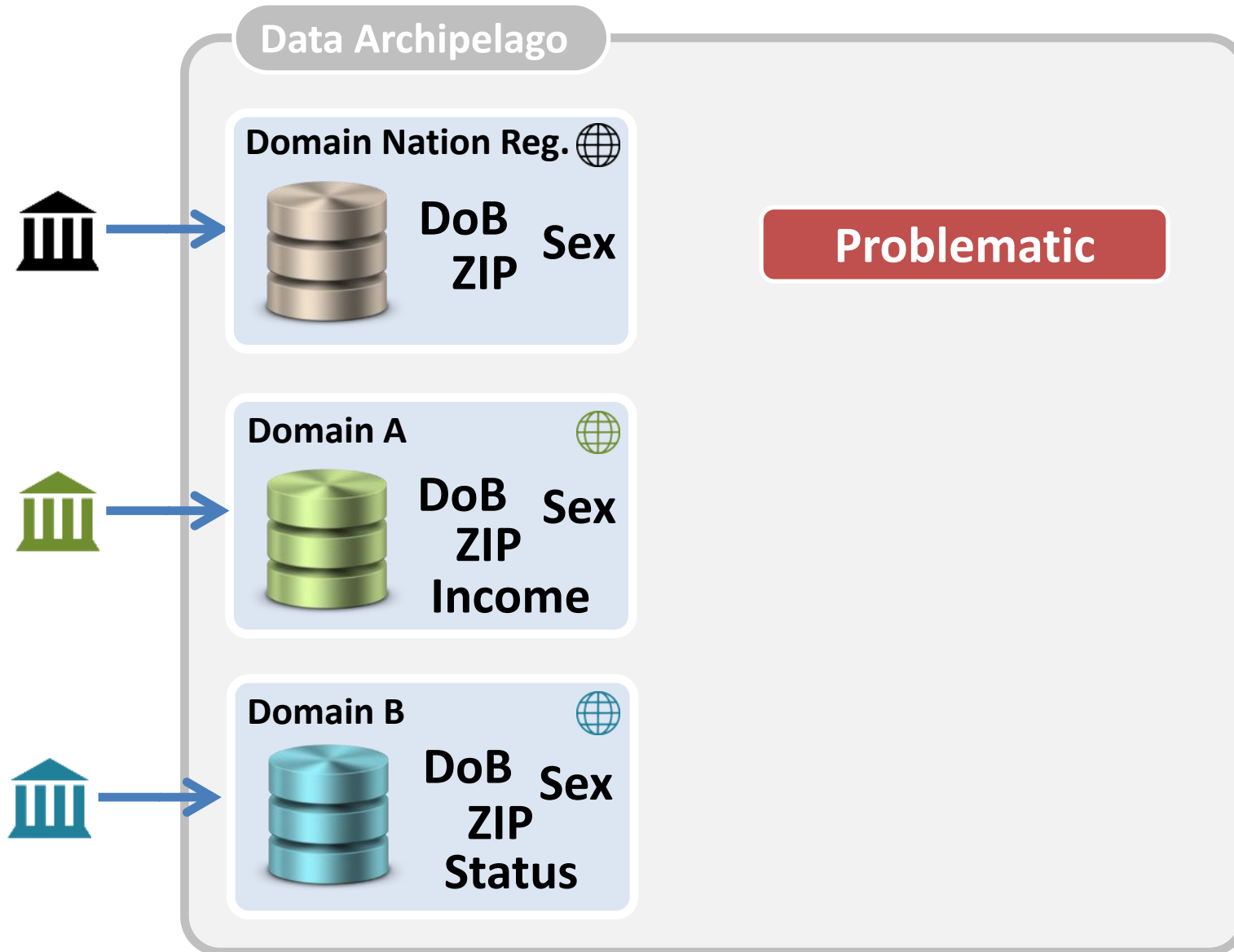
Project 



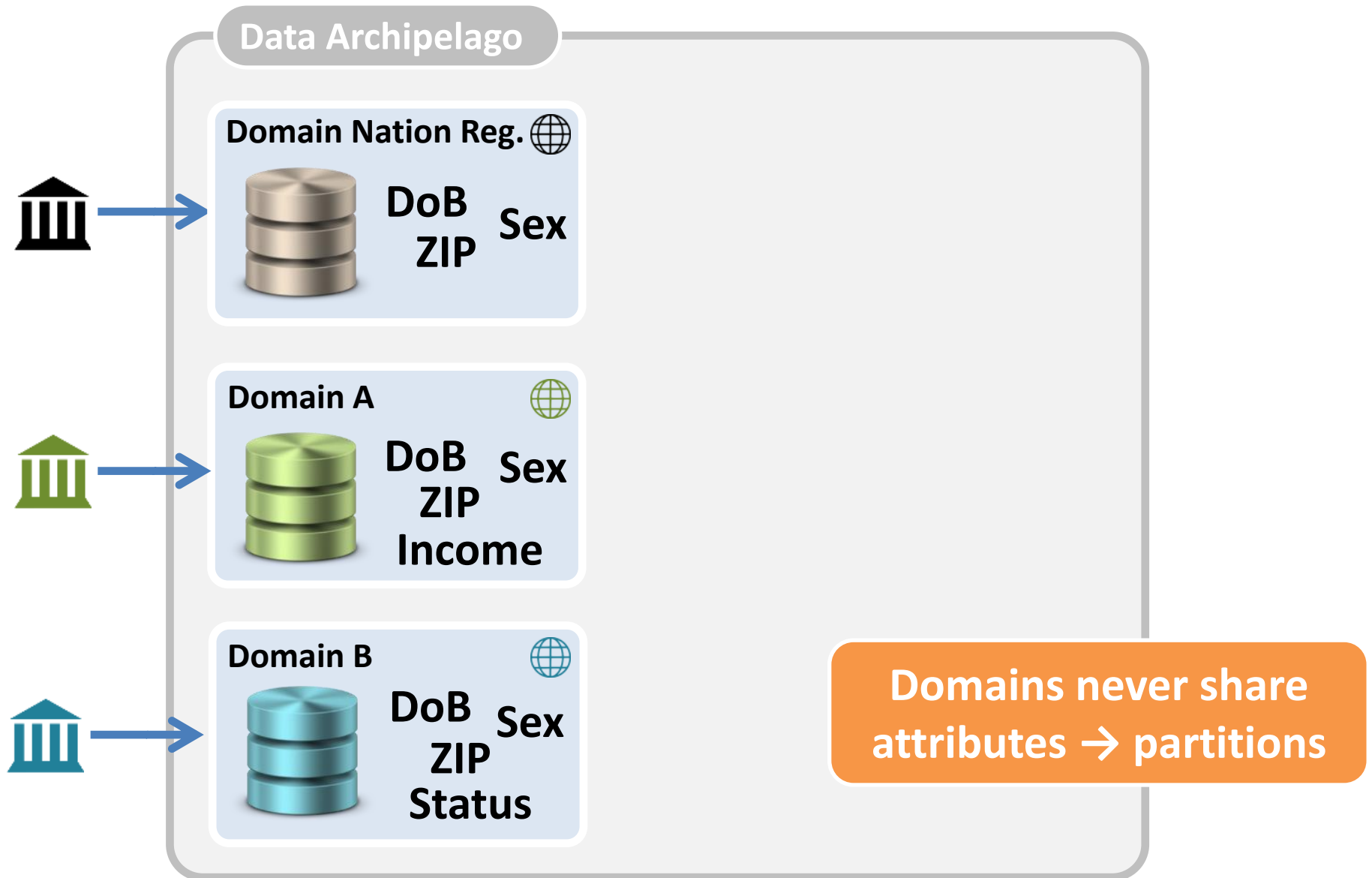
€



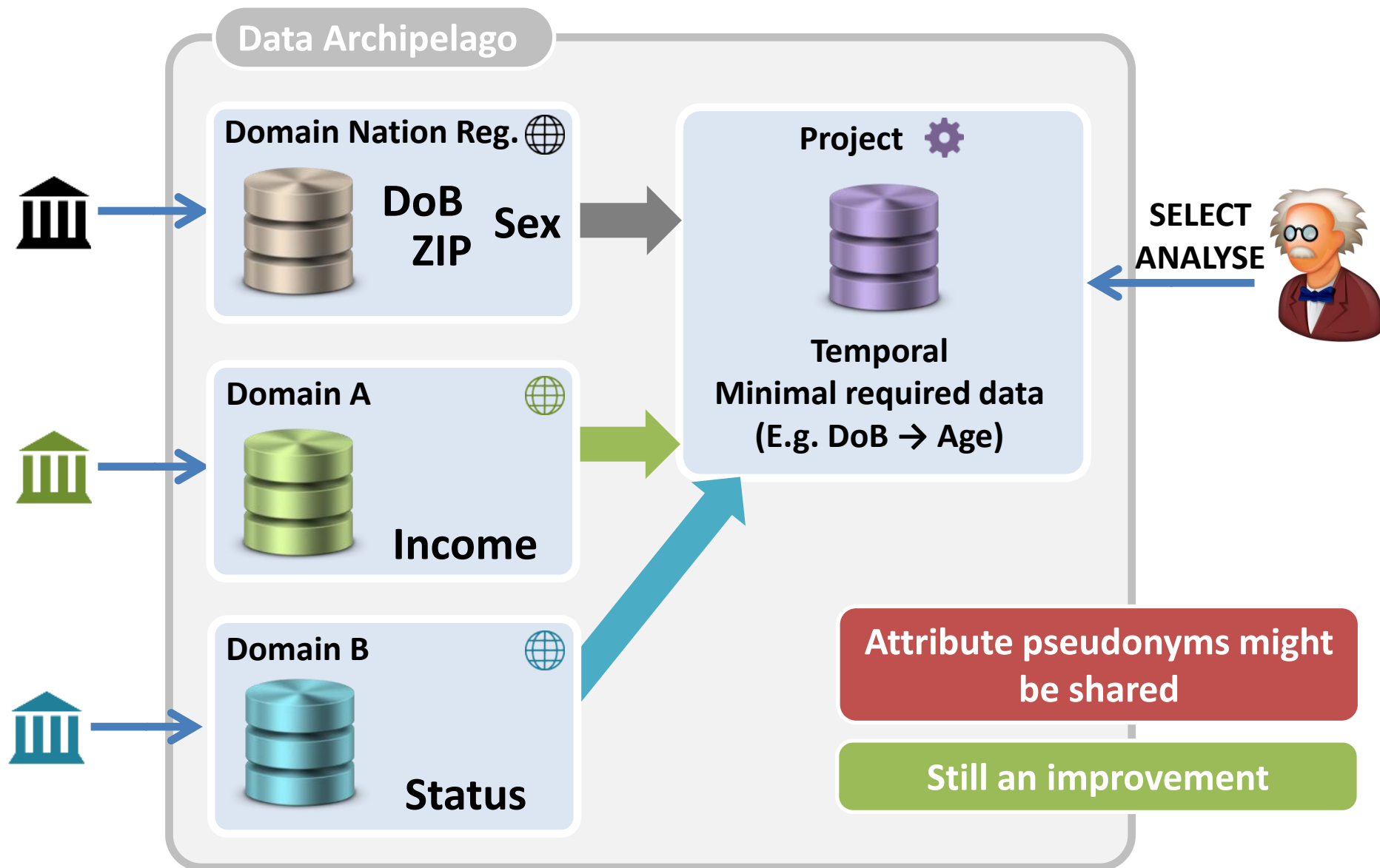
Attribute Linkability



Attribute Linkability



Attribute Linkability



Linkability



Forbid

Minimise

Eliminate

Domain 



€



Domain 



€



Project 



€



Domain 

DoB

Sex

ZIP

€



Domain 

DoB

Sex

ZIP

€



Project 

DoB

Sex

ZIP

€



Domain 



€



Domain 



€



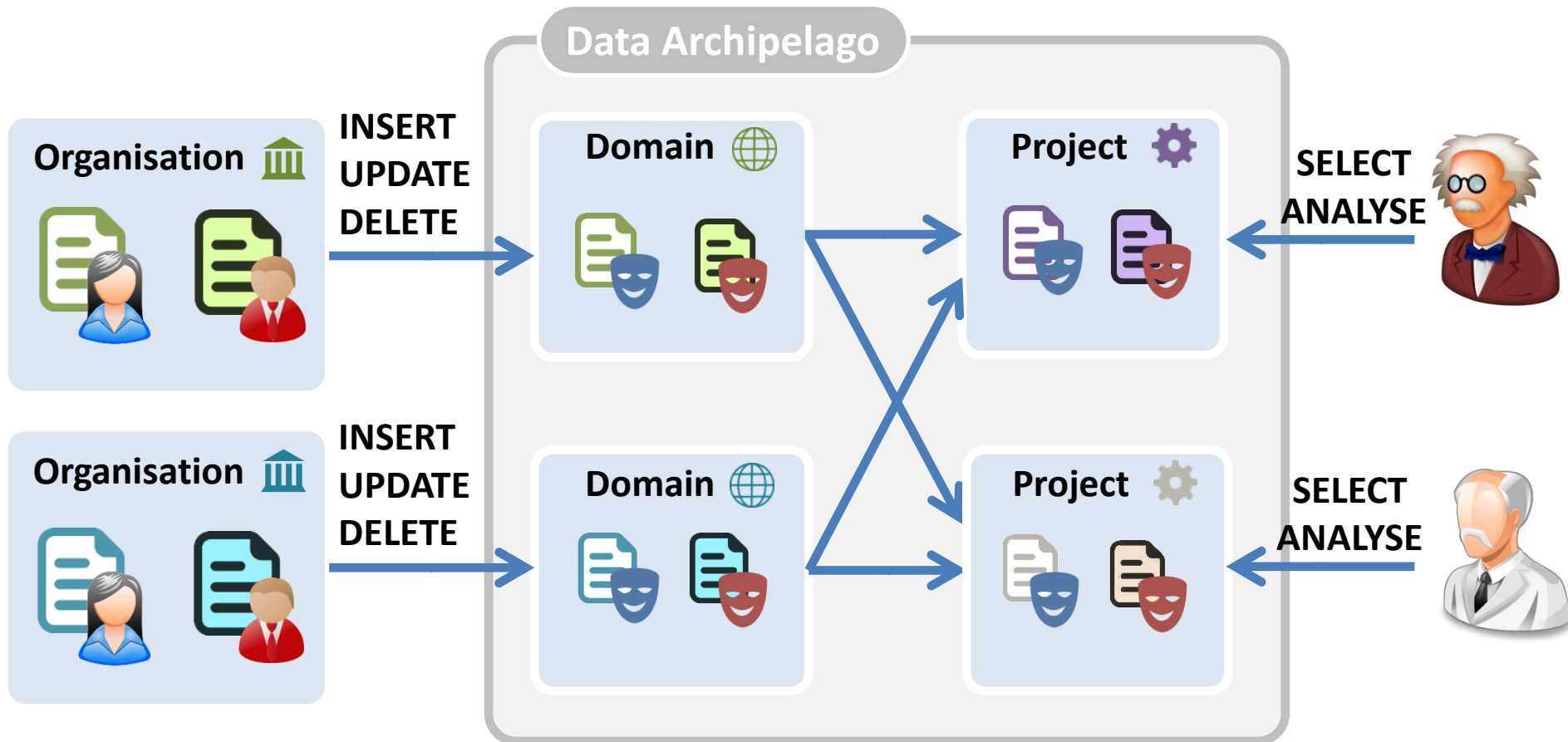
Project 



€

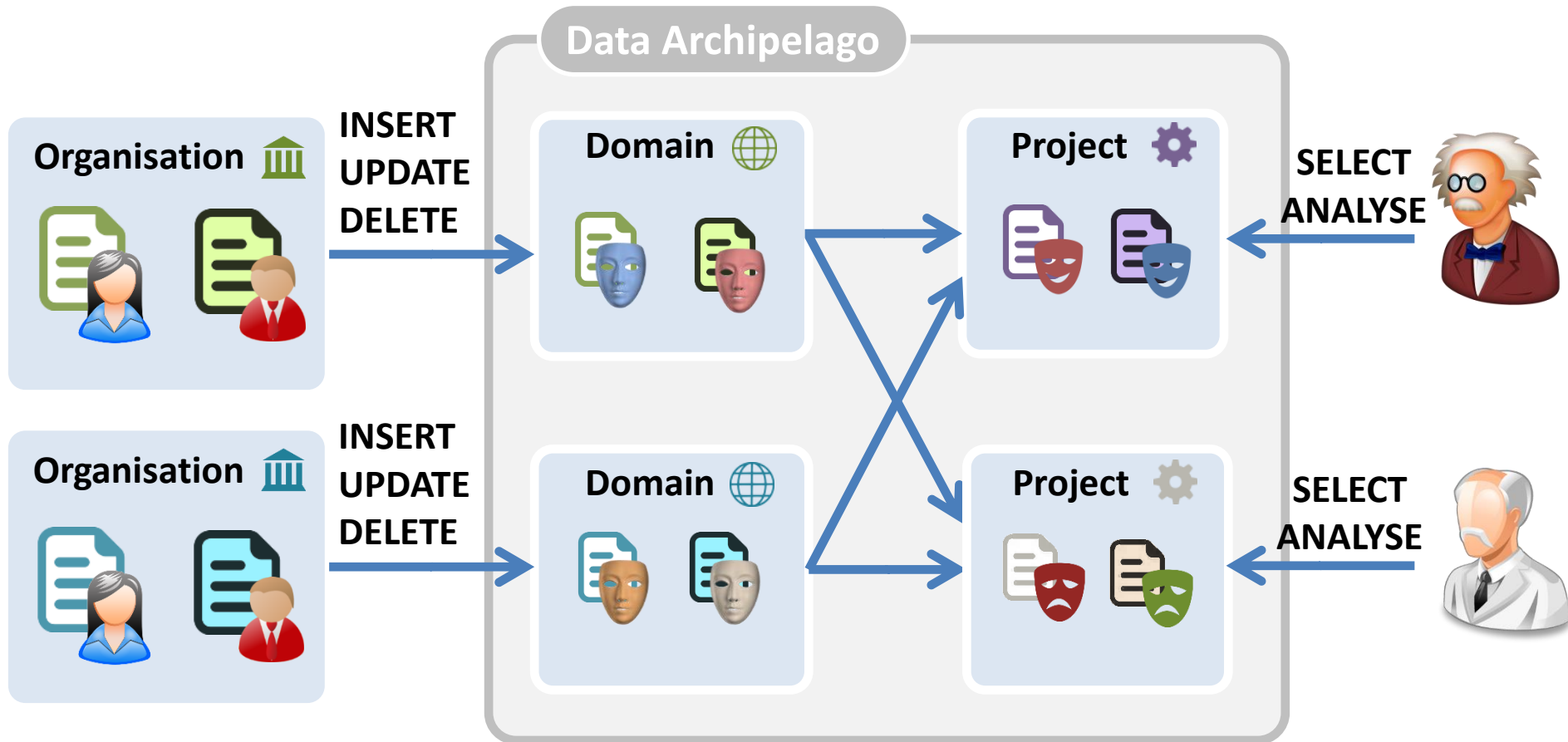


Linkabilities with Pseudonyms



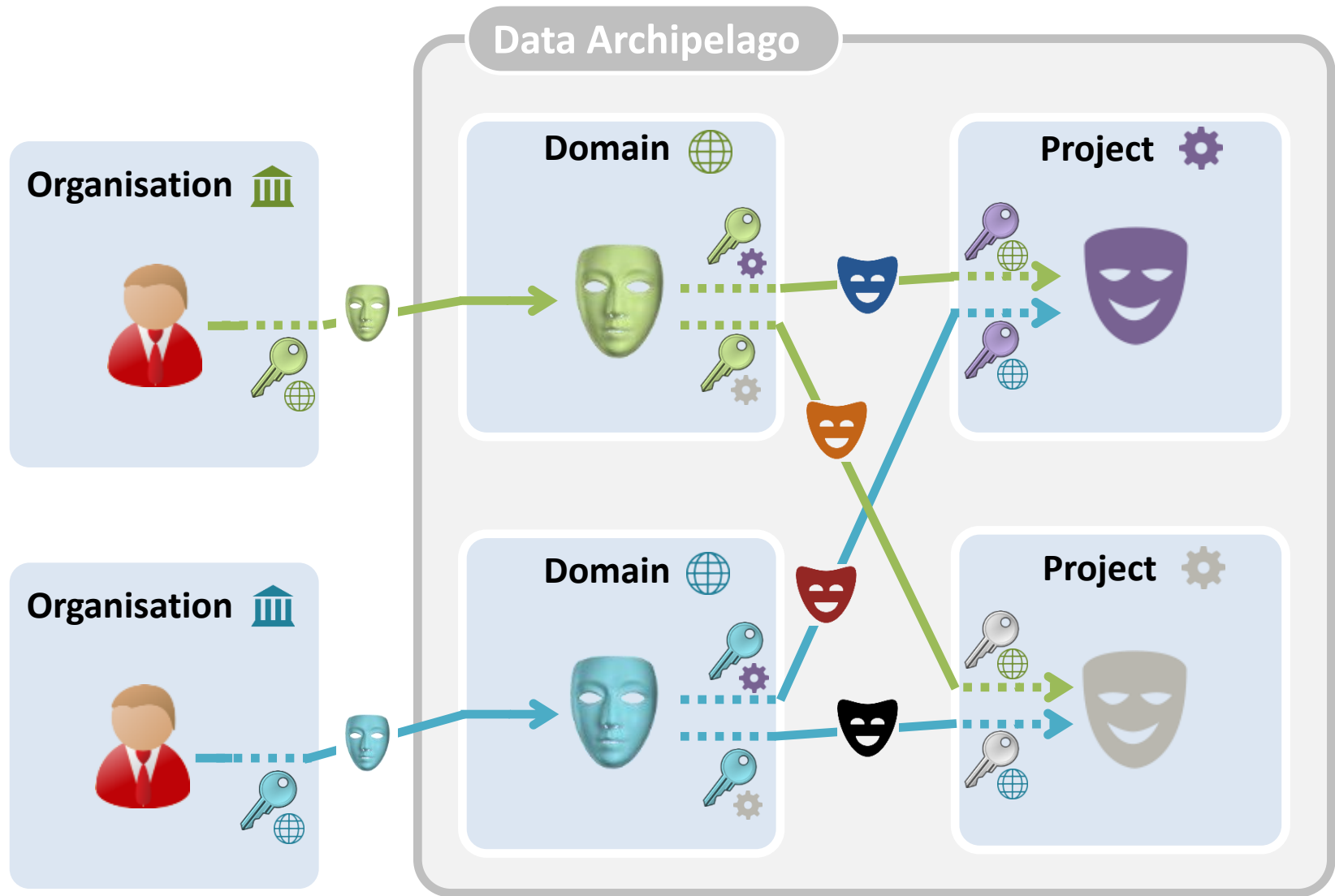
Citizen is known on each island under the same pseudonym

In a Perfect World...



Every citizen has a separate pseudonym for each island
Pseudonyms unlinkable to each other and to identifier

Identifiers & Pseudonyms



New link requires at least two parties => isolation

A Project Example

Project needs data about

Domain National Register 
Year of Birth $\geq$ 1990

Domain A 
self-employed
secondary activity

Domain B 
wage >
50.000€/year

Steps

1. Project approval (machtigingsaanvraag)



2. Generation & distribution of keys

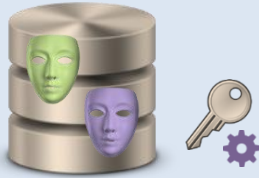


3. Data collected by project

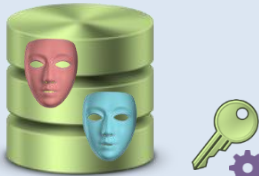
Identifiers & Pseudonyms

Data Archipelago

Domain NR 



Domain A 



Domain B 



Project 



Identifiers & Pseudonyms

Data Archipelago

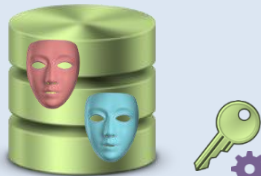
Domain NR 



“YoB $\geq$ 1990”



Domain A 



“self-employed
secondary activity”



Domain B 



“Wage > 50 000€”



Project 



Identifiers & Pseudonyms

Data Archipelago

Domain NR 



Domain A 



Domain B 



Project 



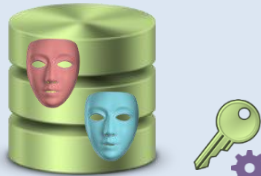
Identifiers & Pseudonyms

Data Archipelago

Domain NR 



Domain A 



Domain B 



Project 



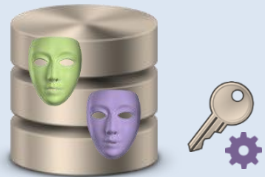
$\cap$



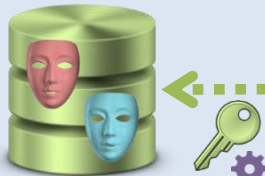
Identifiers & Pseudonyms

Data Archipelago

Domain NR 



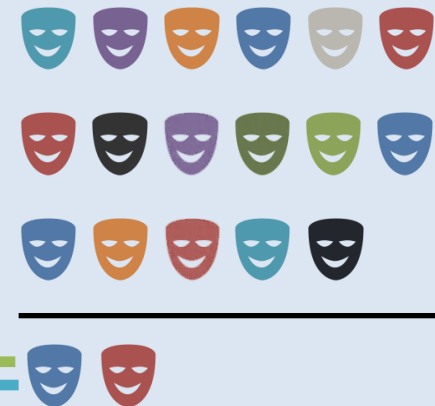
Domain A 



Domain B 



Project 



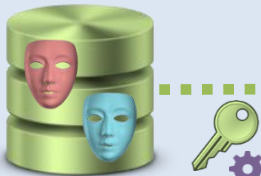
Identifiers & Pseudonyms

Data Archipelago

Domain NR 



Domain A 



Domain B 



Project 

Maximal control by domains
(organisations)

Project receives only minimal
required data



Trust / Abuse

No one learns that I ask data about too many pseudonyms



Data Archipelago

Domain NR 



I learn about  and  :
 $YoB \geq 1990$ and $wage > 50.000/year$ 

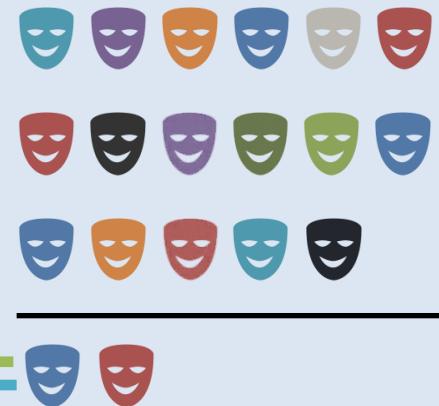
Domain A 



Domain B 



Project 



Report in progress

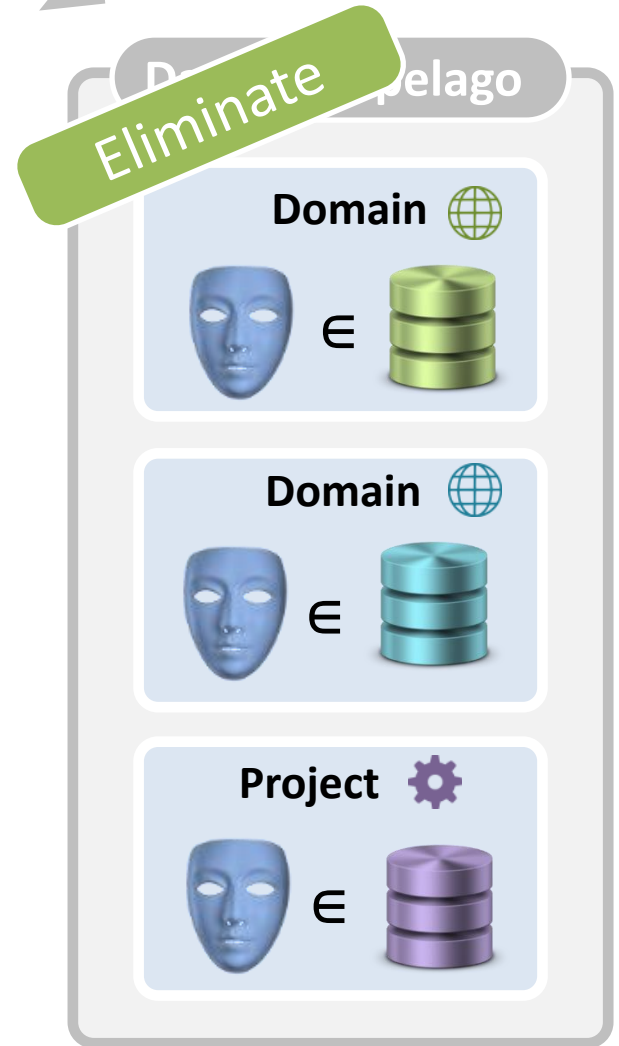
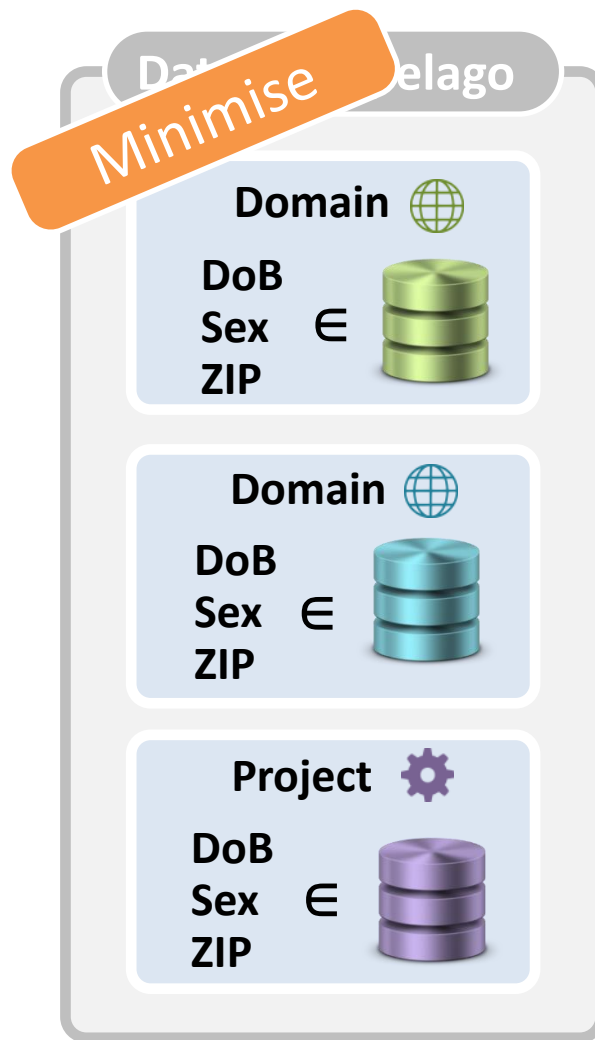
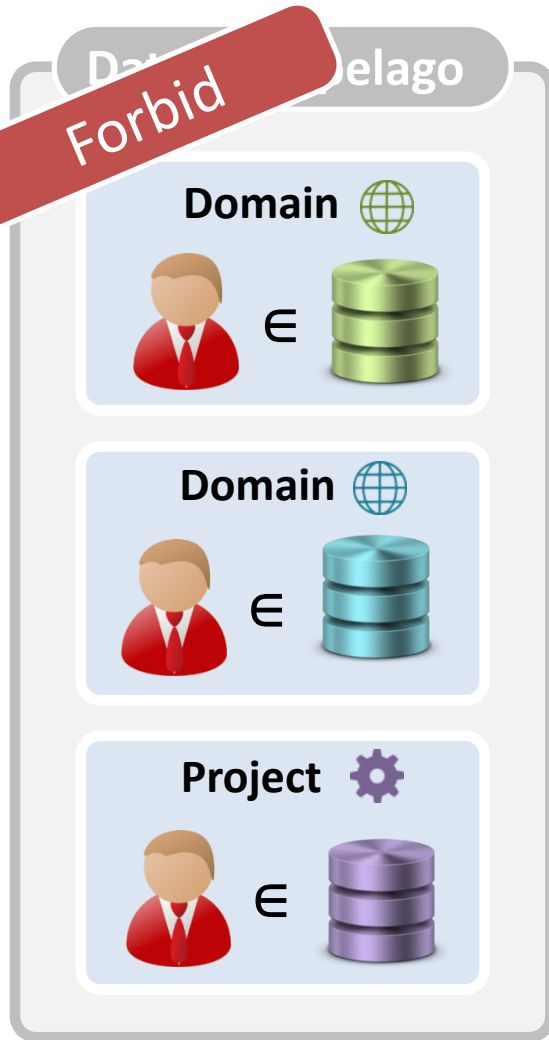
Linkability



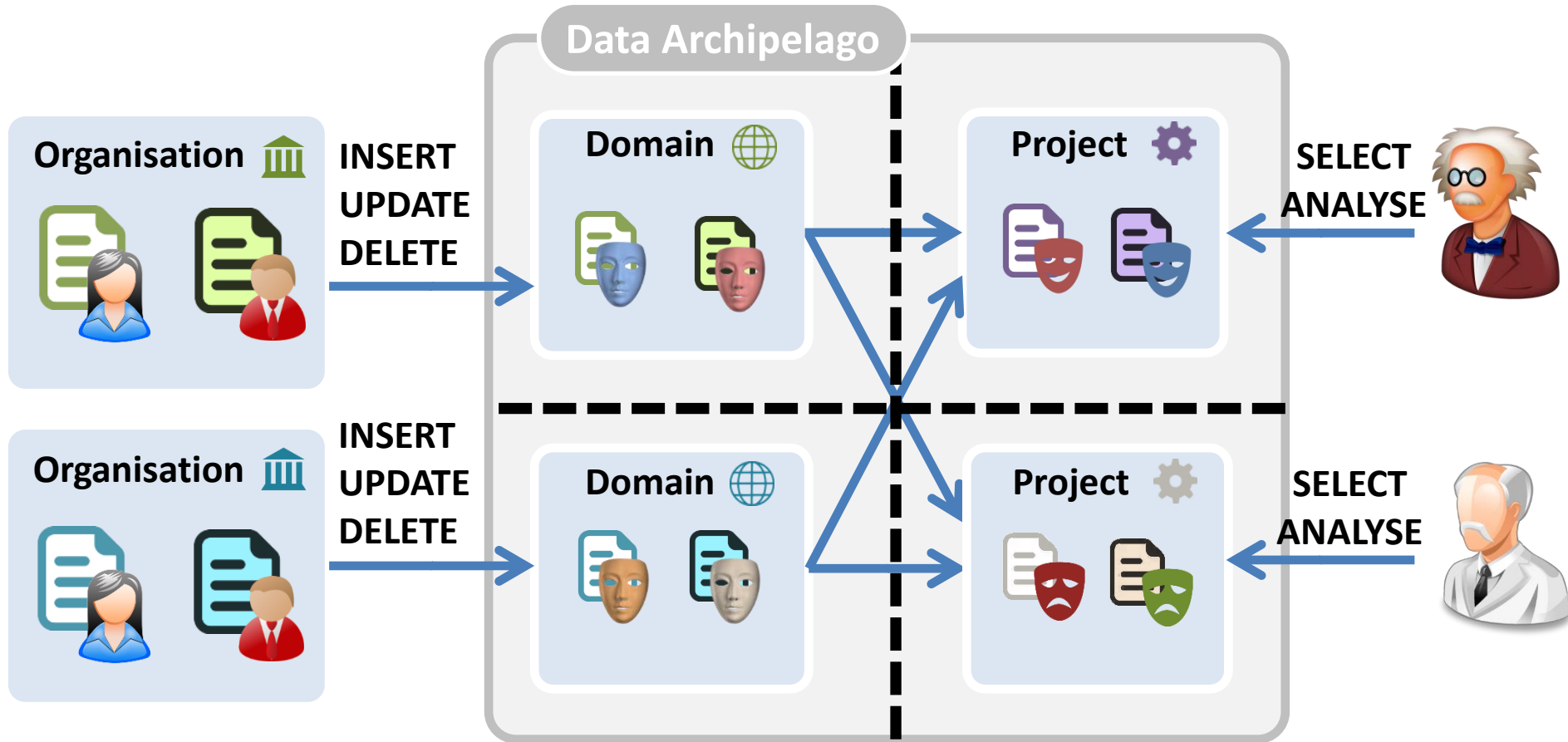
Forbid

Minimise

Eliminate



Maximal isolation



In Summary

Efficient linking together of data

Organisation more control over data

- Decides what data to domain
- Cooperation required to link data in project

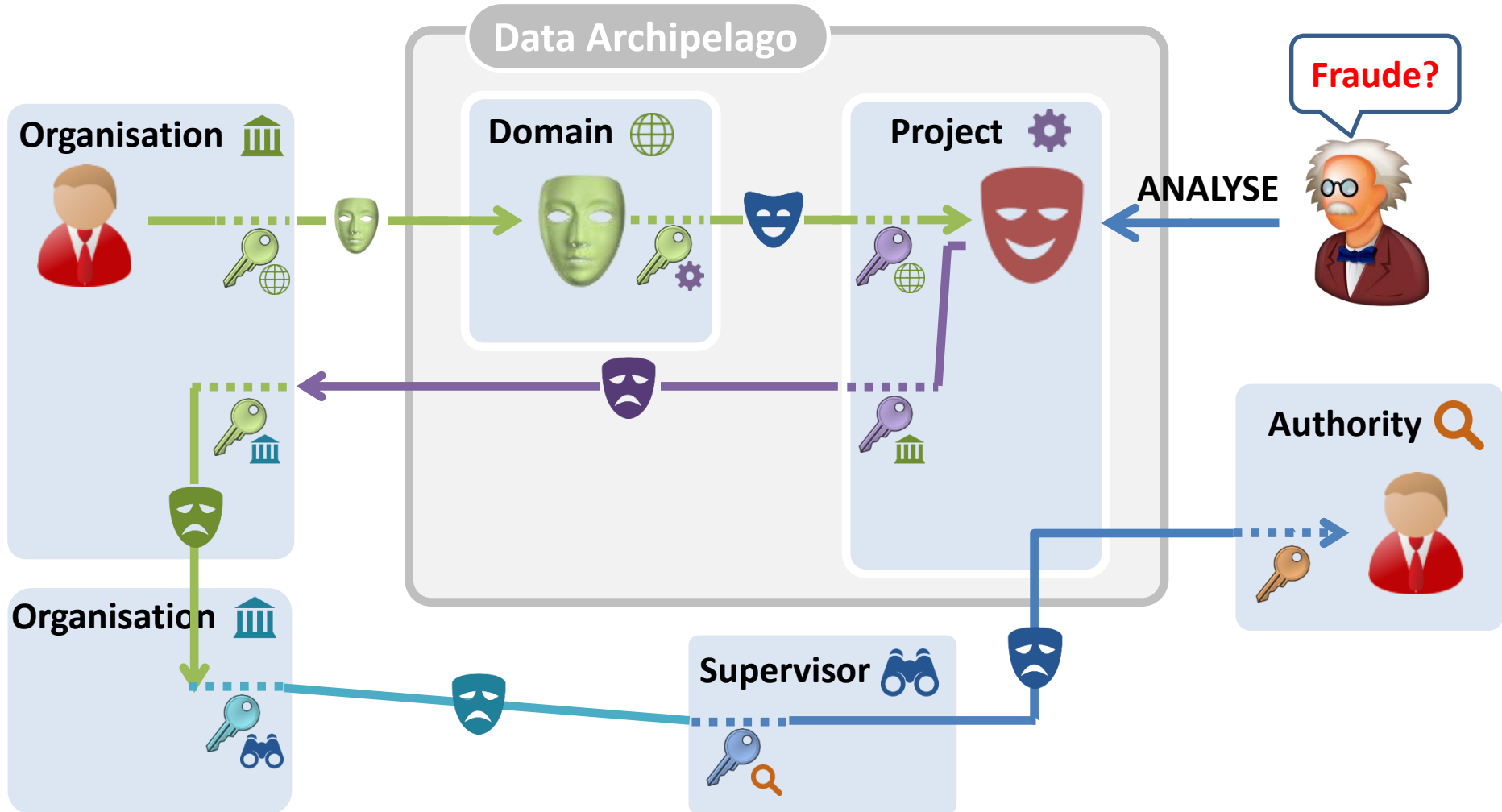
Minimal impact in case of data breach

- By minimising linkabilities
- More probable that use for secondary purposes and research allowed by GDPR

Flexible, case-by-case deanonymisation

- E.g. for fraud detection
- E.g. Approval Privacy Commission required for each deanonymisation
- Privacy Commission does not learn identity suspect

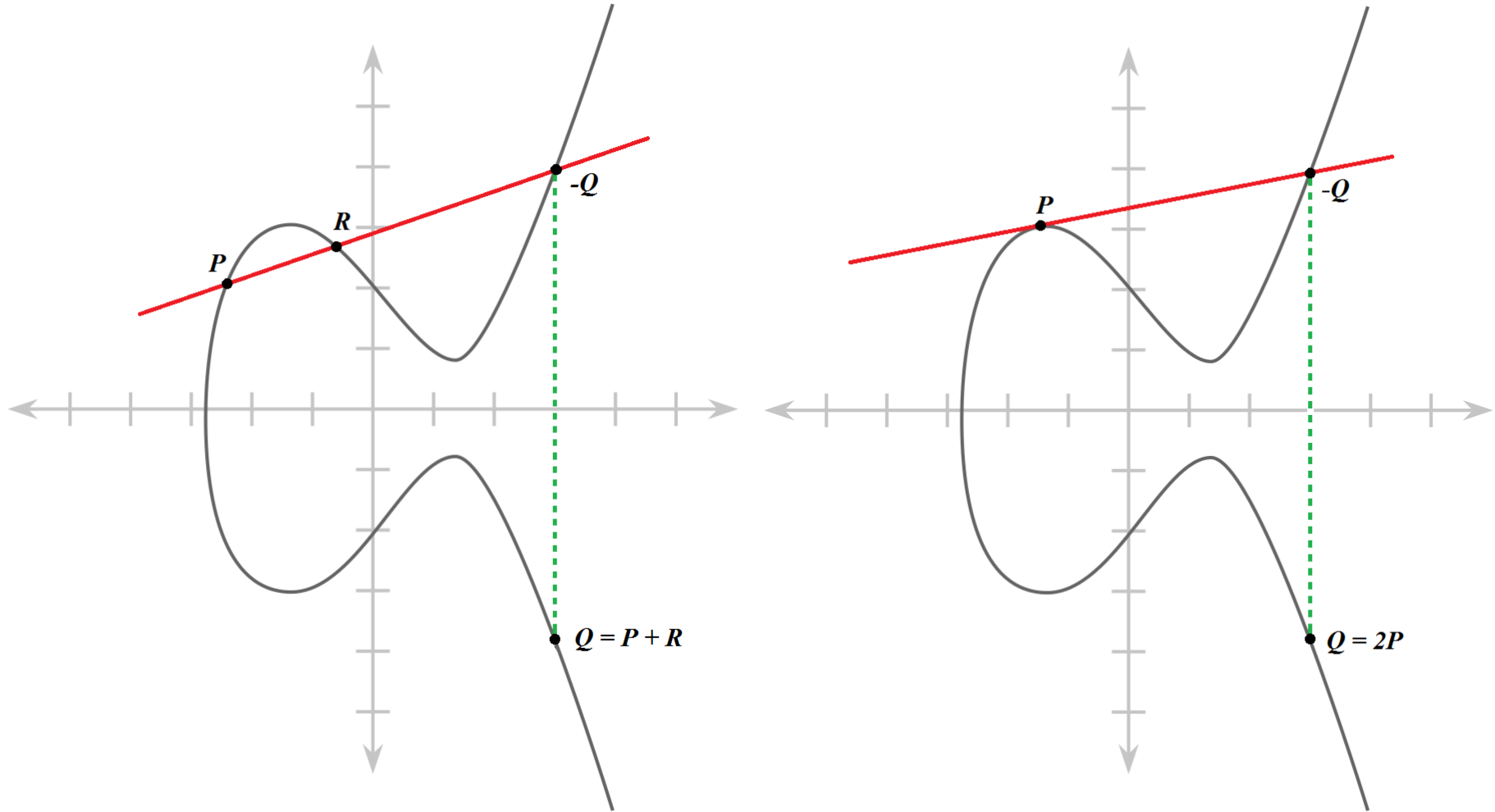
Deanononymisation



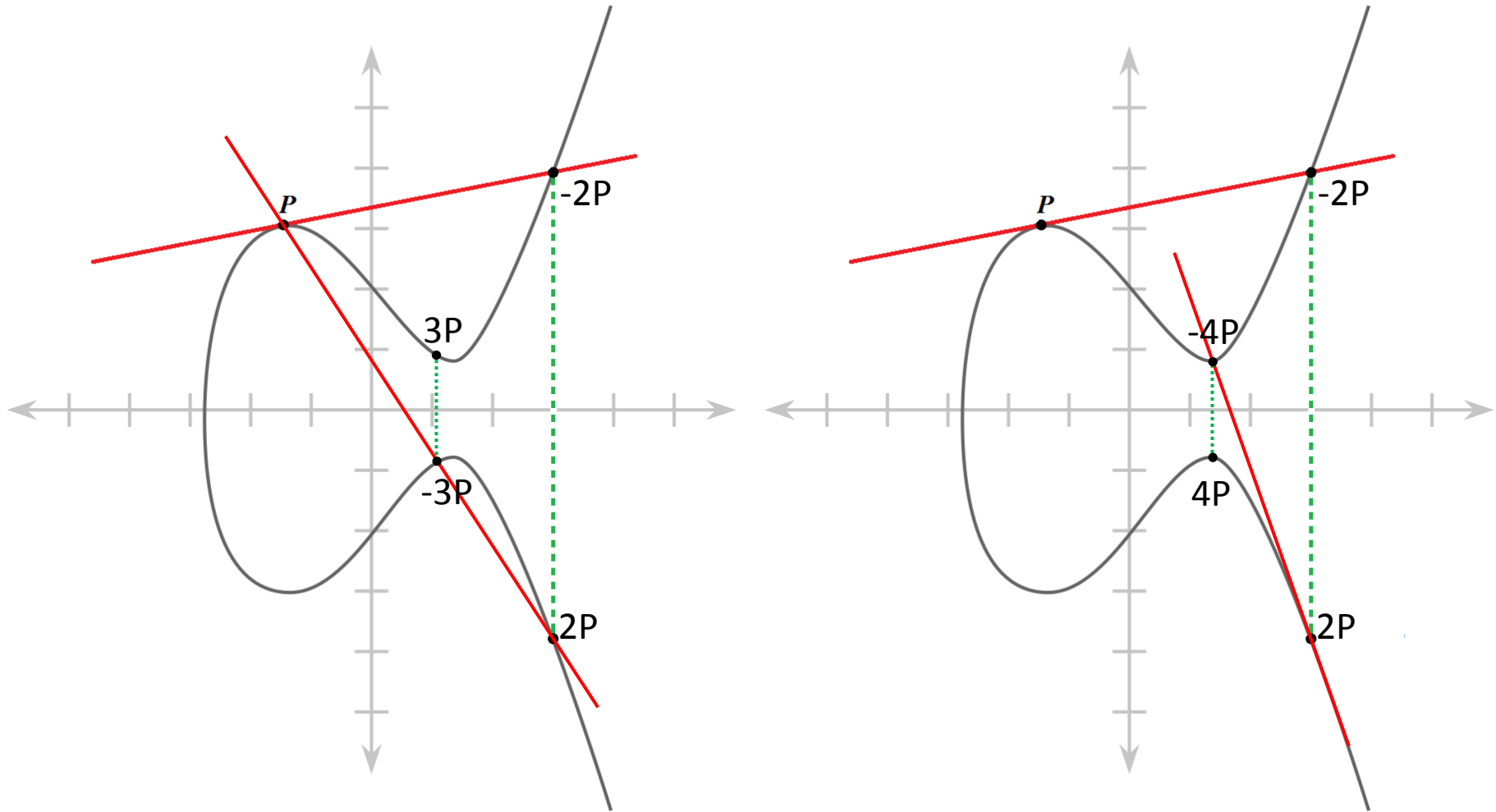
Under the hood



Elliptic Curves (EC)



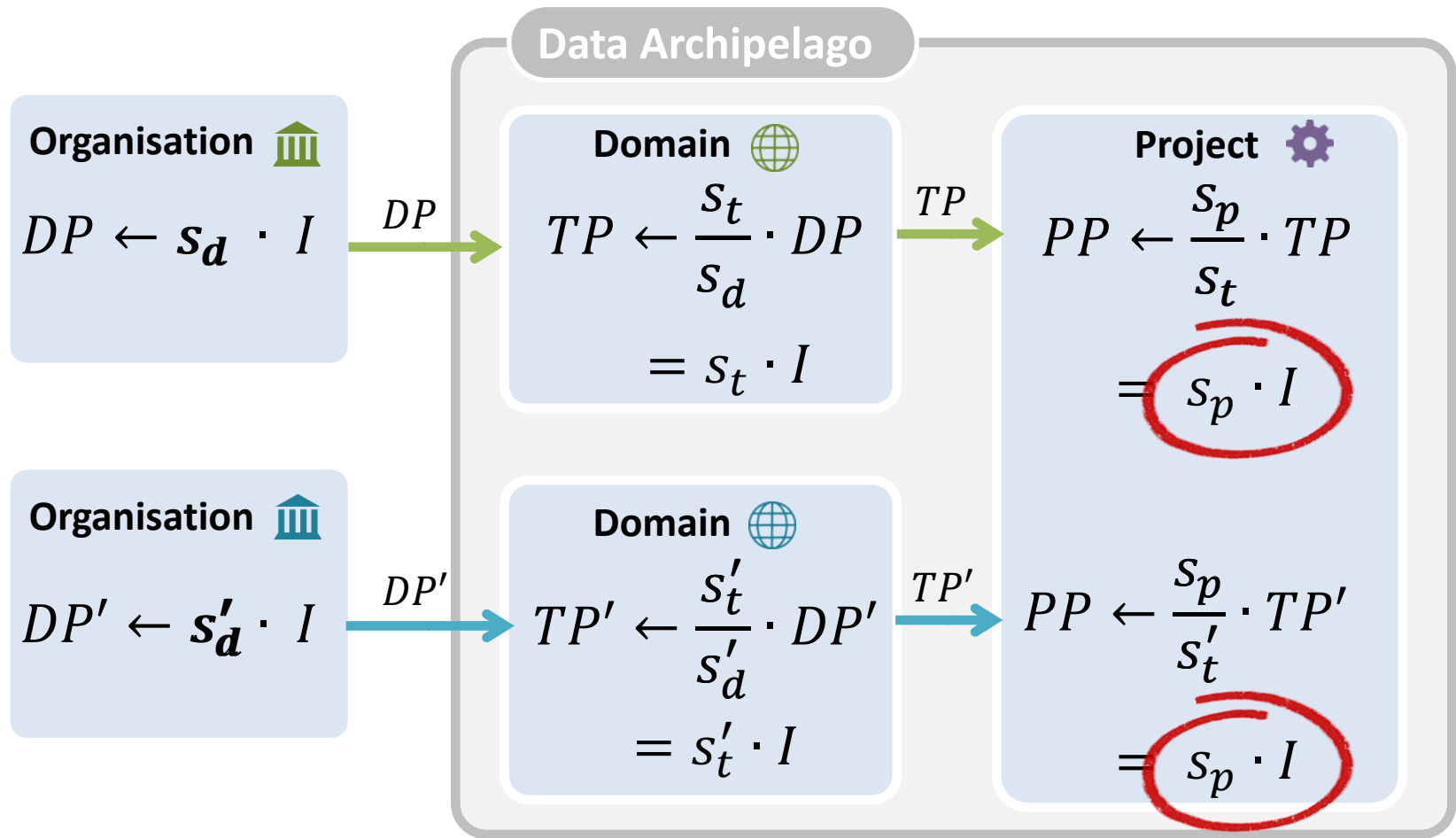
Elliptic Curves (EC)



Easy: $Q \leftarrow n \cdot P = \underbrace{P + P + \dots + P}_{n \text{ times}} \quad n \in \mathbb{Z}_q$

Hard: $n \leftarrow P, Q$

Central Idea



Easy: $Q \leftarrow n \cdot P = \underbrace{P + P + \dots + P}_{n \text{ times}} \quad n \in \mathbb{Z}_q$

Hard: $n \leftarrow P, Q$

Proof-of-Concept

Theoretical model also works in practice

Performance pseudonym conversion

PC Windows 7 Enterprise (64bit) on a single 2,66Ghz Intel i5 core

RSA			EC		
Key size	One operation	Ops / hour	Key size	One operation	Ops / hour
1536 bit	58ms	62070	192 bit	0,4ms	9 million
2048 bit	135ms	26700	224 bit	0,6ms	6 million
3072 bit	440ms	8180	256 bit	0,7-0,8ms	4-5 million

1 million pseudonym conversions => 12,5 minutes

Linking Together Personal Data

A fictional example

A research team wants to analyse medical, financial and demographic data from all citizens born in or after 1990 with a wage of at least € 50 000 per year who are self-employed as secondary activity.

However, these data are maintained by separate governmental organisations and, hence, need to be linked together.

A step forward

Scientists

Analyse data sets

Citizen

Respect privacy

Governmental org.

Maintains control
(because responsible)

All

Minimal impact
data breach

Kristof Verslype



02 787 53 76



Kristof.verslype@smals.be



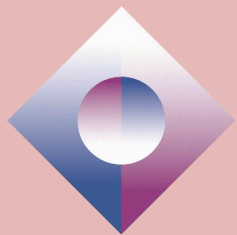
@KristofVerslype



be.linkedin.com/in/verslype



Smals



www.smals.be



@Smals_ICT



www.smalsresearch.be



@SmalsResearch

