Basics of Personal Data Minimization in ARX

Afshin Amighi, Ahmad Omar

Hogeschool Rotterdam: CMI-INF , Creating 010

June 26, 2024

SURF Security- en Privacyconferentie

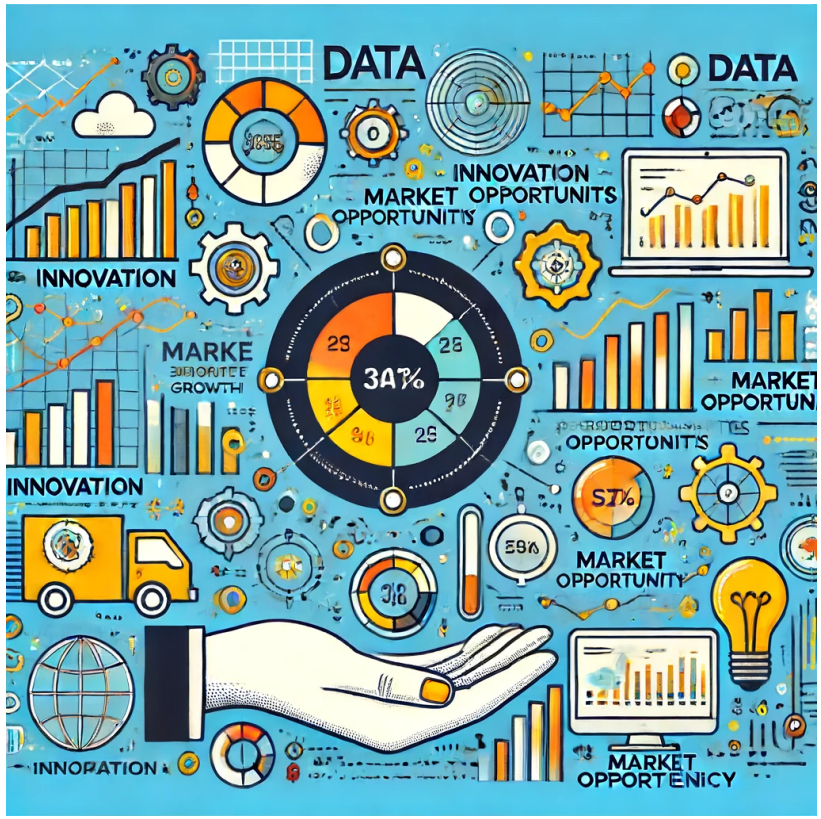
Agenda

- Motivation
- SDC Technologies
- k-Anonymity
- ARX
- Discussion

Motivation

Public Data and Data Sharing

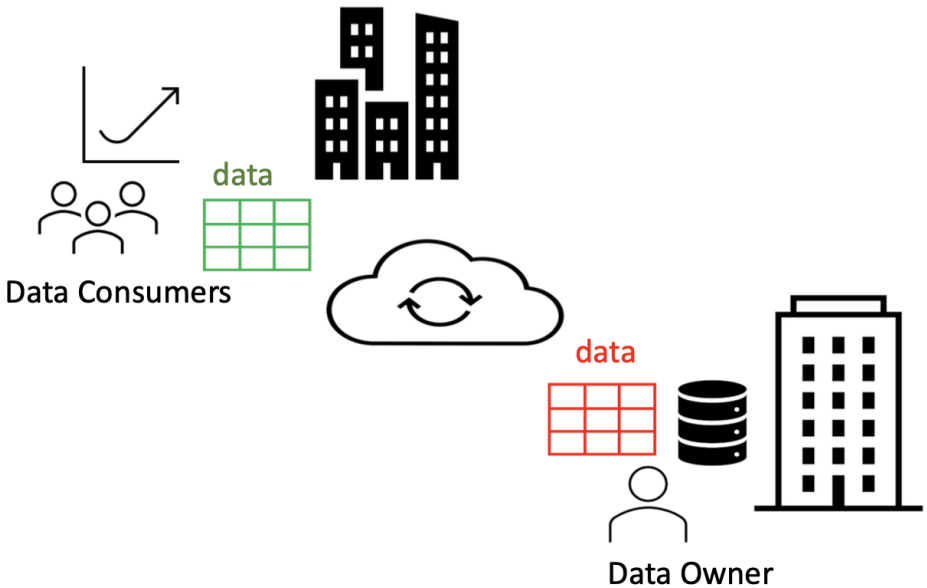
- **Data as an Economic Driver:**
  - Innovation and Research
  - Efficiency and Productivity
- **Economic Benefits:**
  - Market Opportunities
  - Transparency and Accountability



Motivation

Data Protection

- **GDPR:**
  - Key Principles: Lawfulness, fairness, transparency, etc.
- **Rights and Obligations:**
  - Processors (Consumers)
  - Controllers
  - Owners



SDC Technologies

Basics

Statistical Disclosure Control (SDC): statistical methods used to protect the confidentiality of individuals in datasets while maintaining the utility of the data for analysis.

- Suppression
- Generalization
- Masking
- Aggregation

k-Anonymity

Basics

Definition

Ensuring each record is indistinguishable from **at least** k-1 other records with respect to Quasi Identifiers (QIDs).

Attributes (columns): Explicit identifiers, Quasi identifiers, Sensitive attributes, non-sensitive attributes.

- **Implementation:** Suppression and Generalization
- **Benefits:** Reduces risk of re-identification
- **Limitations:** May reduce data utility

3-Anonymized:

Age	Gender	ZIPCode	Occupation
25-30	*	1234*	Engineer
25-30	*	1234*	Doctor
25-30	*	1234*	Teacher
30-35	*	1234*	Nurse
30-35	*	1234*	Lawyer
30-35	*	1234*	Scientist

k-Anonymity

Example

Question

What is the value of k?

Hint: Highlight rows with identical generalised values (quasi identifiers). Each group with a different color (or number).

Age	Gender	ZIP Code	Occupation
20-25	*	123**	Engineer
25-30	*	143**	Doctor
20-25	*	123**	Teacher
30-35	*	127**	Nurse
25-30	*	123**	Lawyer
20-25	*	123**	Scientist
25-30	*	143**	Engineer
25-30	*	123**	Doctor
30-35	*	127**	Teacher

k-Anonymity

Example

Question

What is the value of k?

Hint: Highlight rows with identical generalised values (quasi identifiers). Each group with a different color.

Age	Gender	ZIP Code	Occupation
20-25	*	123**	Engineer
25-30	*	143**	Doctor
20-25	*	123**	Teacher
30-35	*	127**	Nurse
25-30	*	123**	Lawyer
20-25	*	123**	Scientist
25-30	*	143**	Engineer
25-30	*	123**	Doctor
30-35	*	127**	Teacher

ARX

Introduction

An open-source software for anonymizing personal data

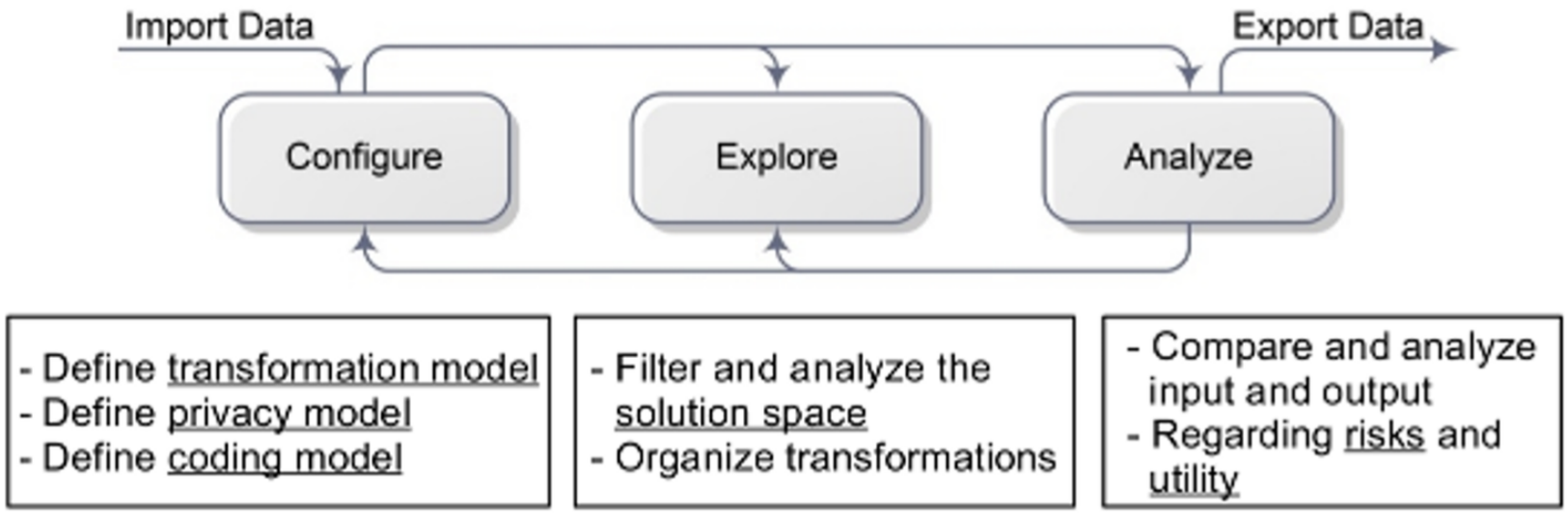
- **Usage:** Import, anonymize, evaluate, and export datasets
- **Source:** https://arx.deidentifier.org/

Features:

- Data Transformation
- Privacy and Risk Models
- Data Utility Metrics

ARX

Features



ARX

Demo

Practice

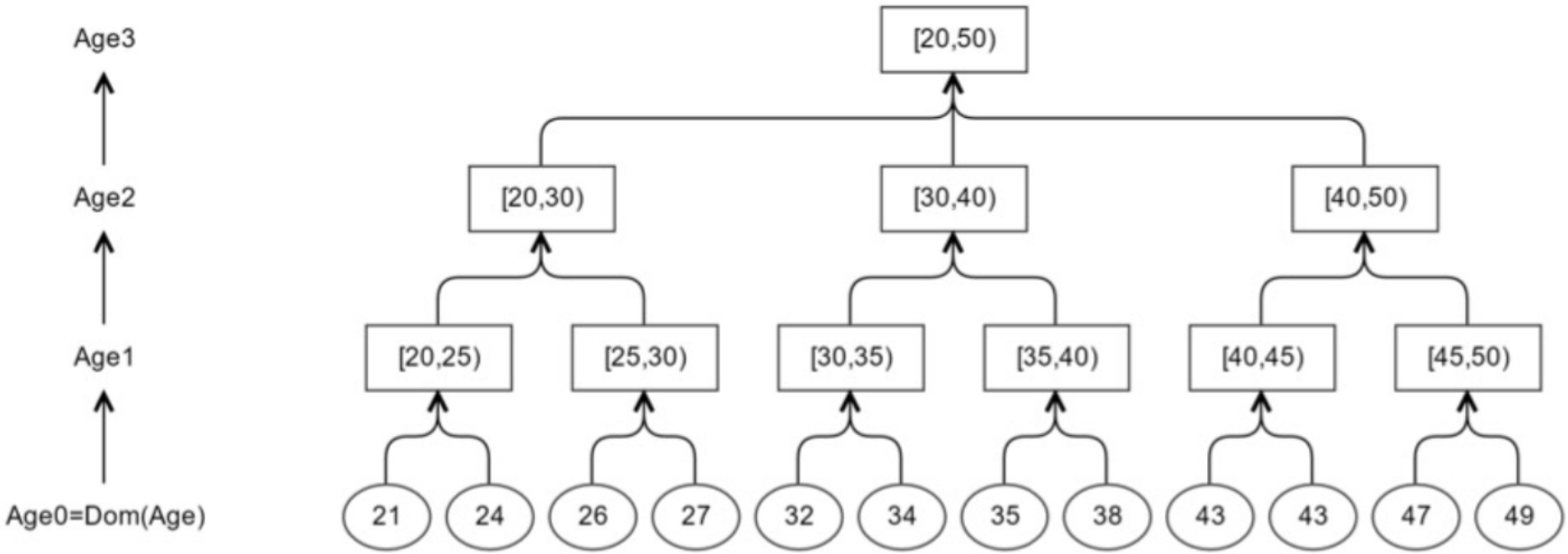
Anonymizing a small dataset in ARX ...

k-Anonymity

taxonomy tree

Definition

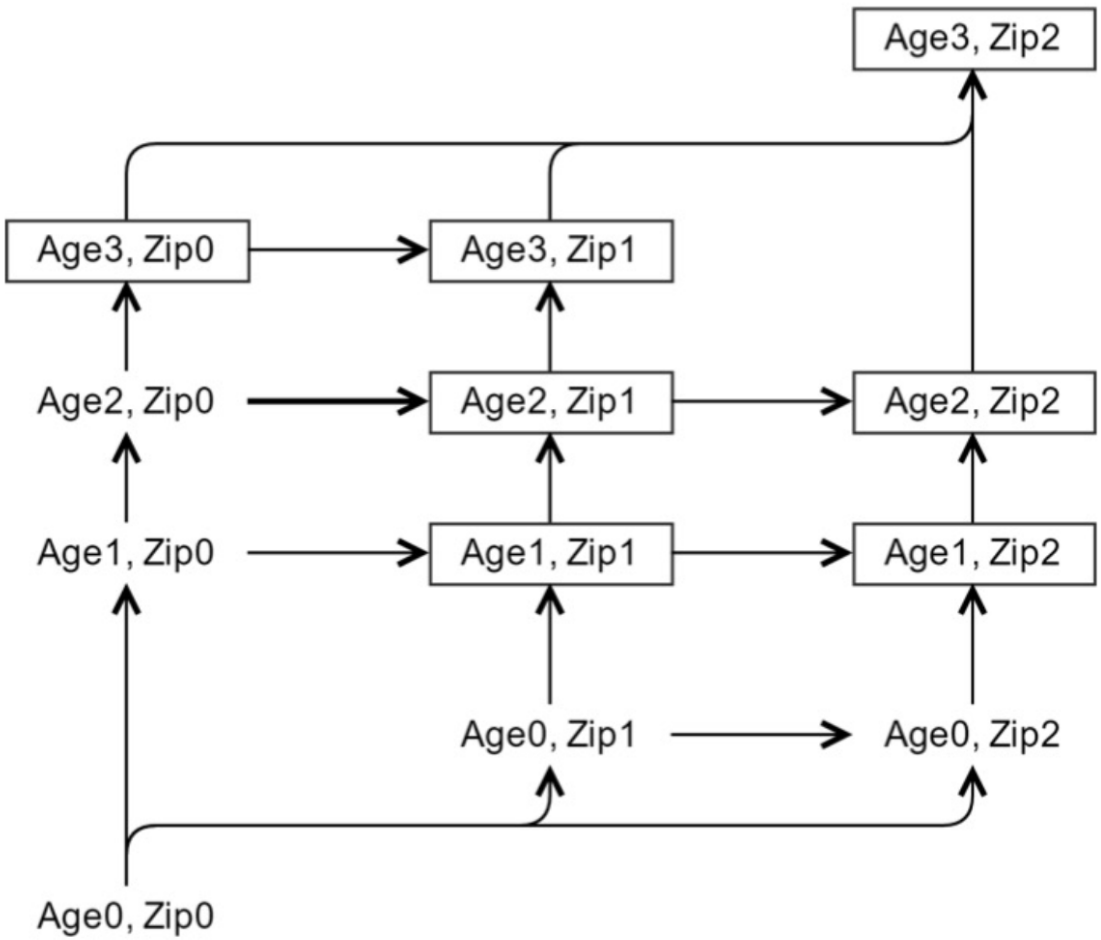
Mapping the values of an attribute to a generalised value.



k-Anonymity

Algorithm Search

The k-anonimity algorithm searches the space of the combined taxonomy trees.



Discussions

Challenges:

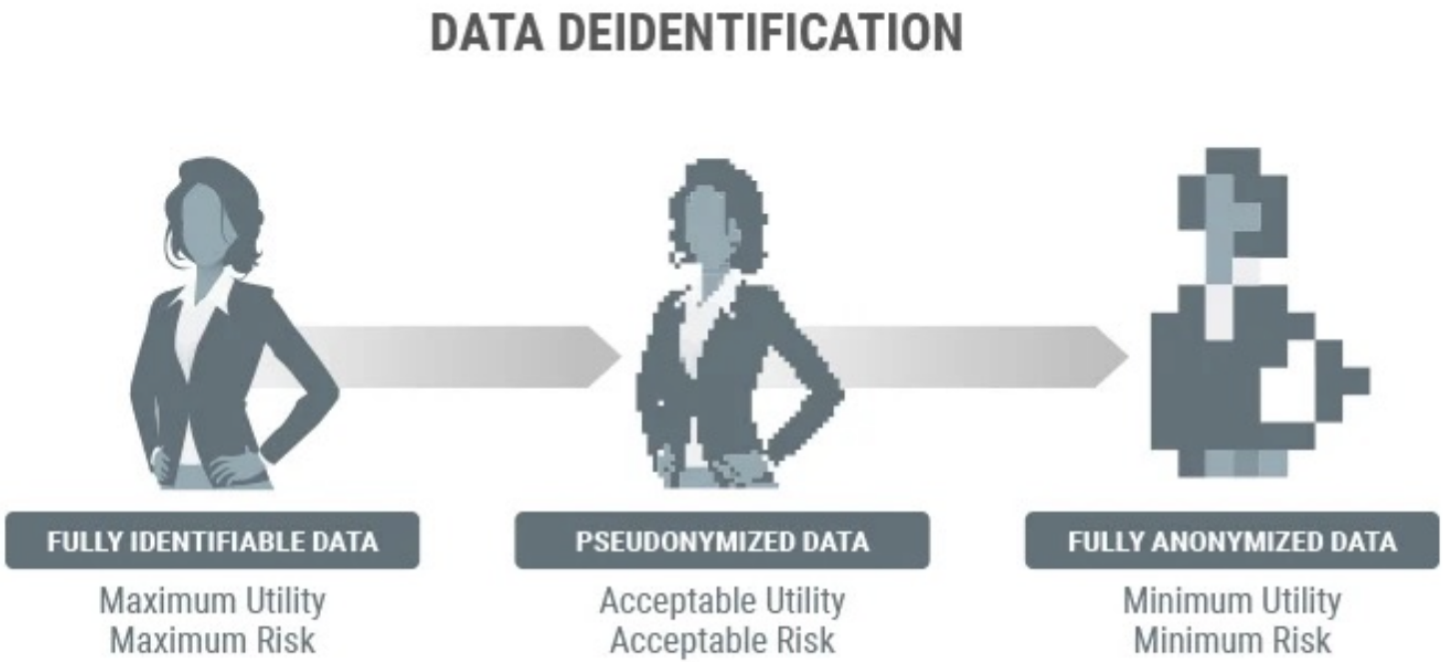
- Balance between data utility and data privacy
- Adapting SDC technologies within organisations

Published research papers:

- ICISS'21: WiP: A Distributed Approach for Statistical Disclosure Control Technologies
- DG.O'24: Directions for Enhancing the Use of Personal Data Minimization Technology in Public Organizations

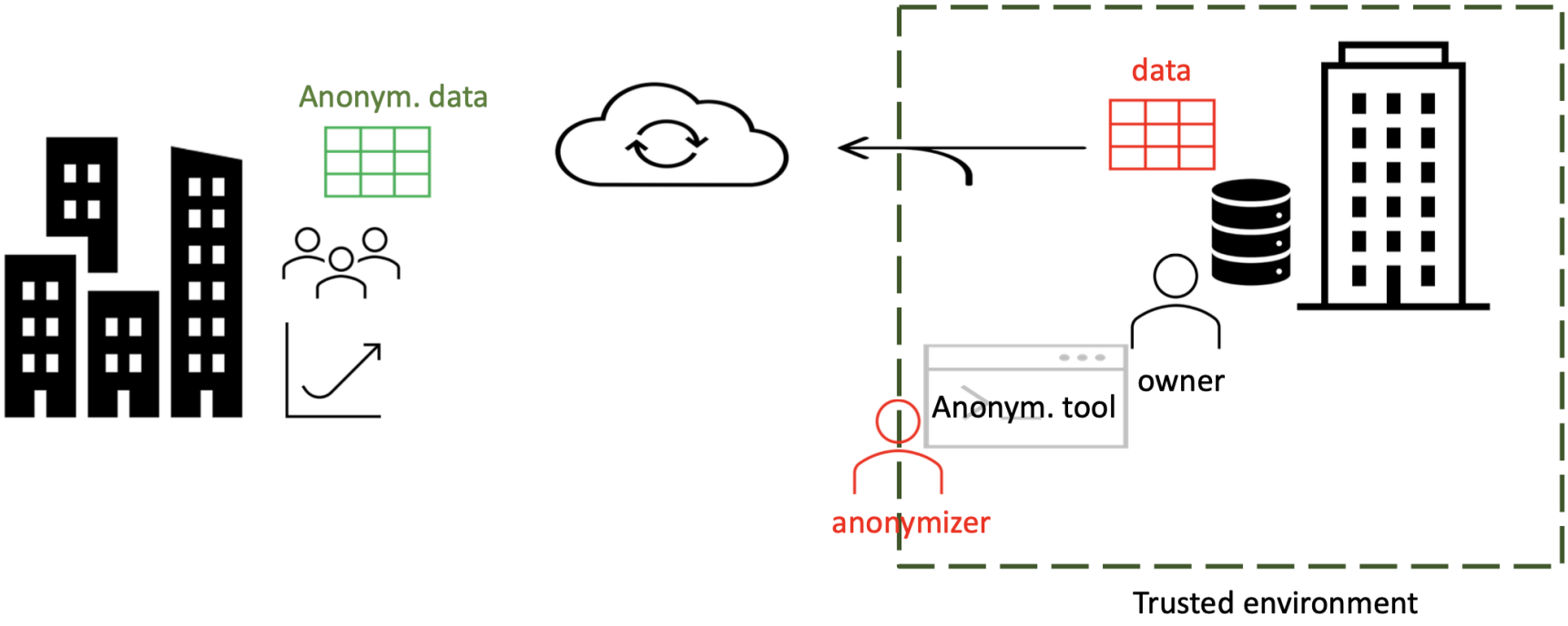
Discussions

Risk vs. Utility



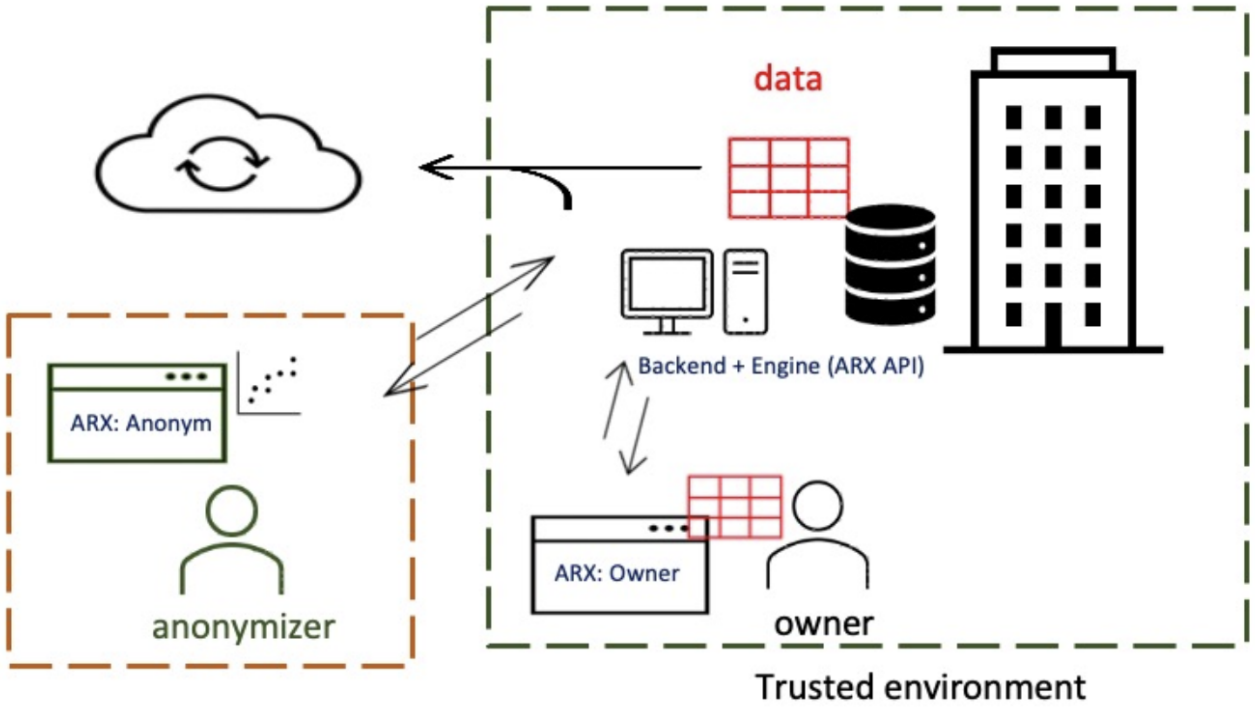
Discussions

Adapting SDC Technologies



Discussions

Our proposal: ICISS'21



Data stays within the trusted environment (secure)

The anonymization: can be outsourced (organisationally scalable)

The anonymization: experts in both domain and SDC (usable)

Conclusion

- **SDC Technologies:** Well developed algorithms, tools, techniques.
- **Challenges:** Usability, adapting technologies and trusted tools.

Thanks

Feel free to reach us ...

Afshin Amighi
a.amighi@hr

Ahmad Omar
a.omar@hr