On Handling Data Minimisation for Workflows: *Preliminary Approach*

Saliha _Irem_ BESIK

besiksal@informatik.hu-berlin.de
Supervisor: Prof. Johann-Christoph Freytag, Ph.D.

dbis
INSTITUT FÜR INFORMATIK
HUMBOLDT-UNIVERSITÄT ZU BERLIN

HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

▸*Slides at https://irem.dev*   @irembesik

GENERAL DATA PROTECTION REGULATION

GDPR Article 5 - *Principles relating to processing of personal data*

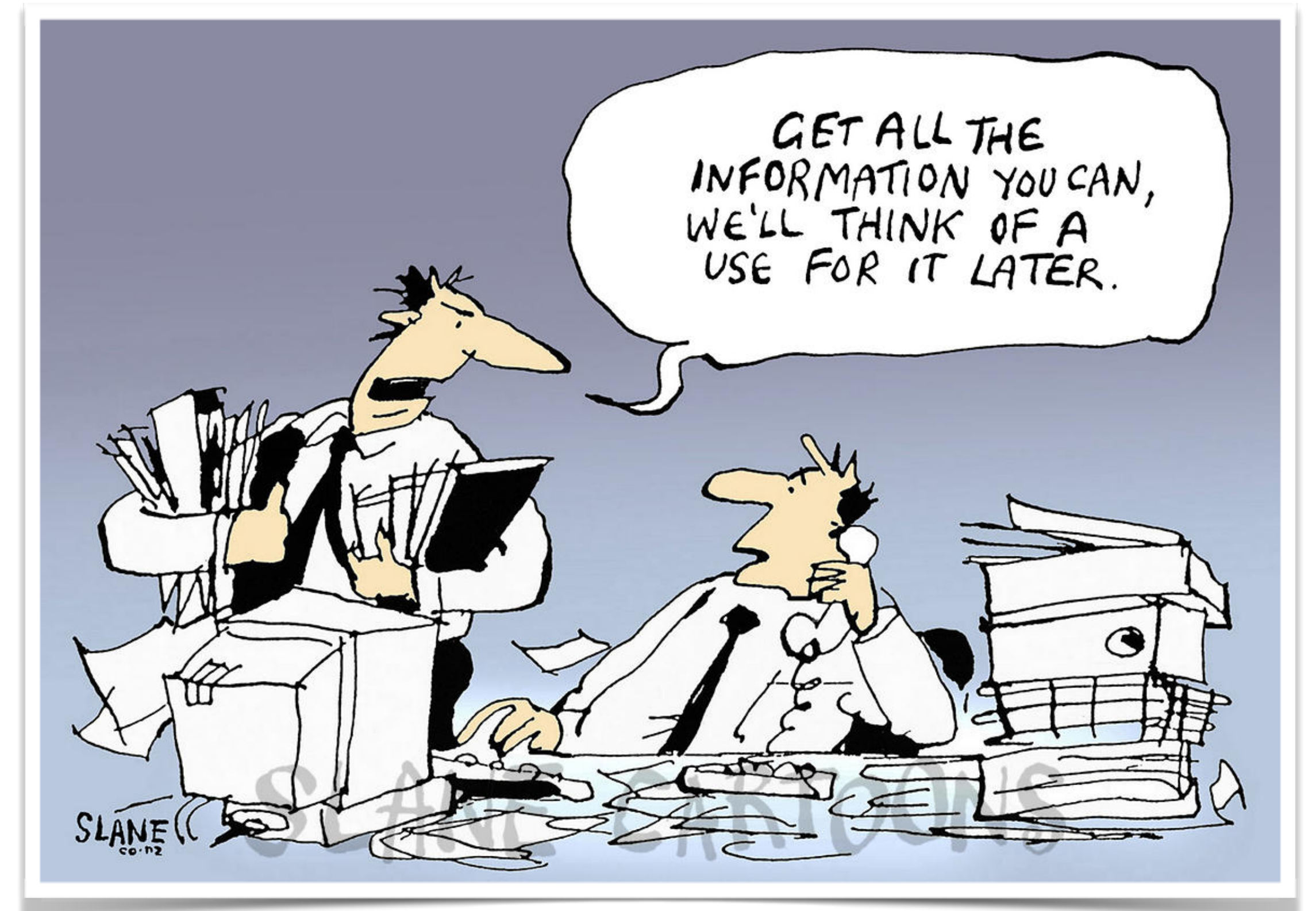| Lawful processing	Data Minimisation	Storage Limitation
Purpose Limitation	Accuracy	Integrity & Confidentiality

GENERAL DATA PROTECTION REGULATION

GDPR Article 5 - *Principles relating to processing of personal data*
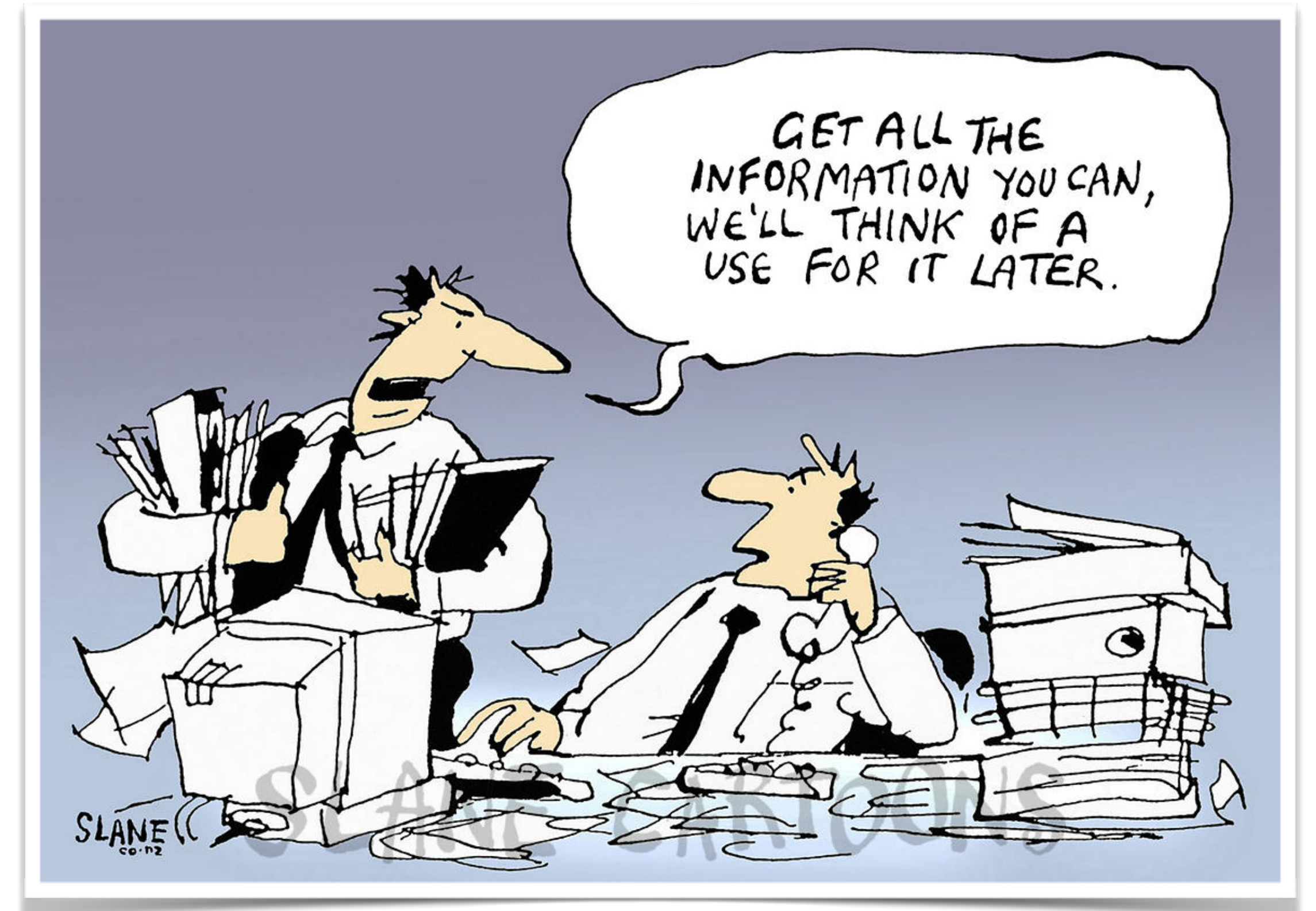
> Data Minimisation

WHAT IS *DATA MINIMISATION?*

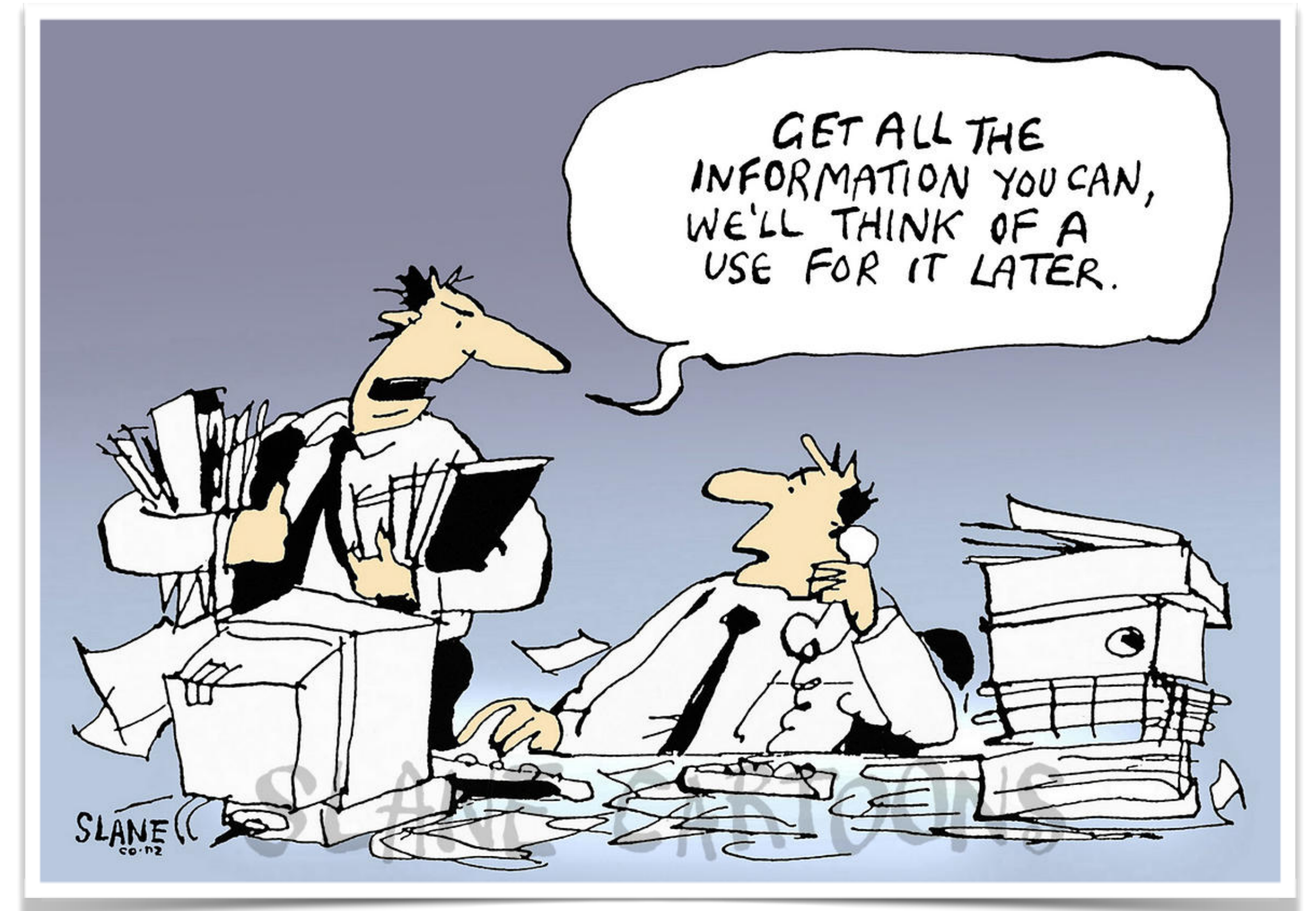# WHAT IS *DATA MINIMISATION?*

Personal data should be

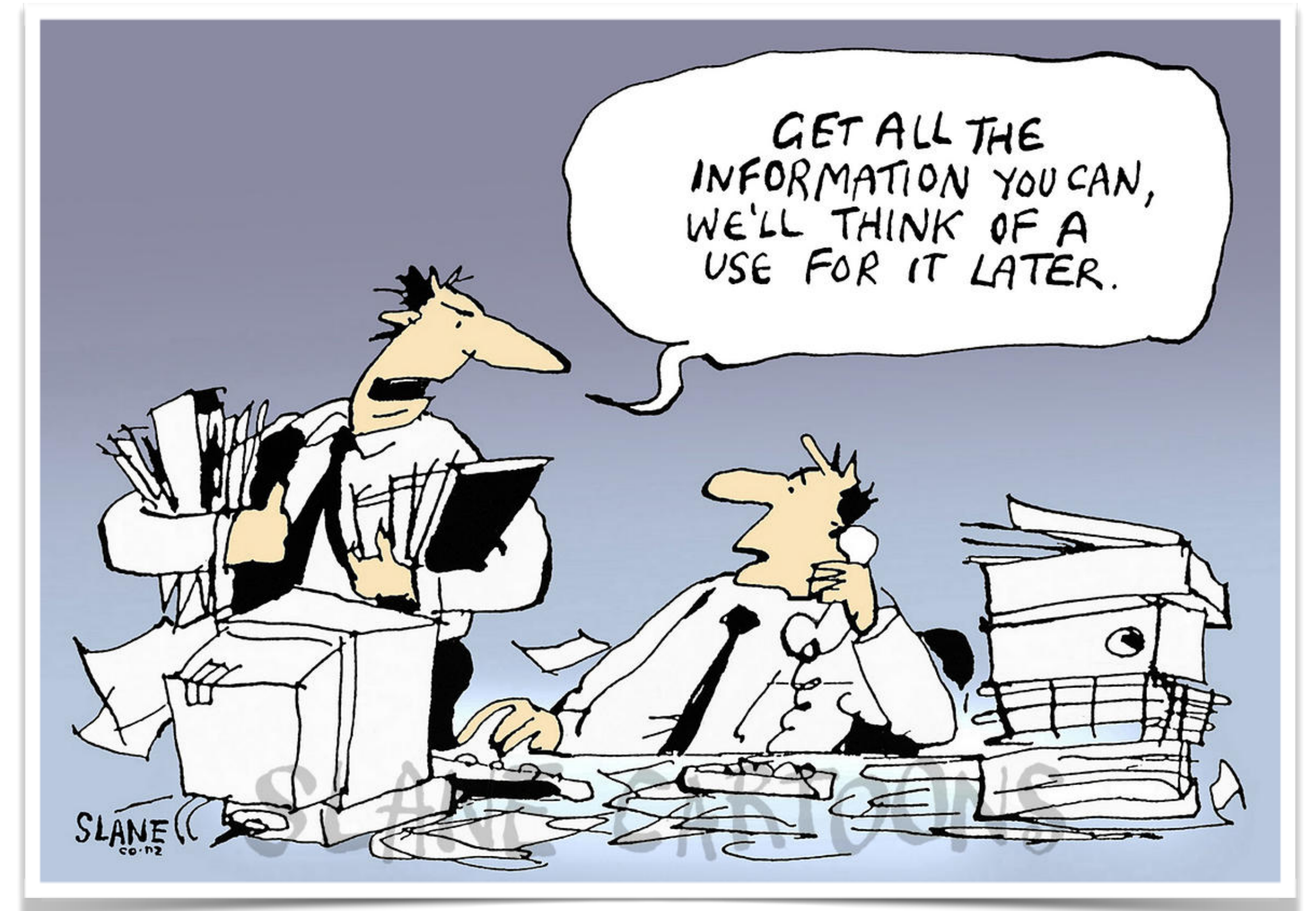# WHAT IS *DATA MINIMISATION?*

Personal data should be

- **adequate**

WHAT IS *DATA MINIMISATION?*
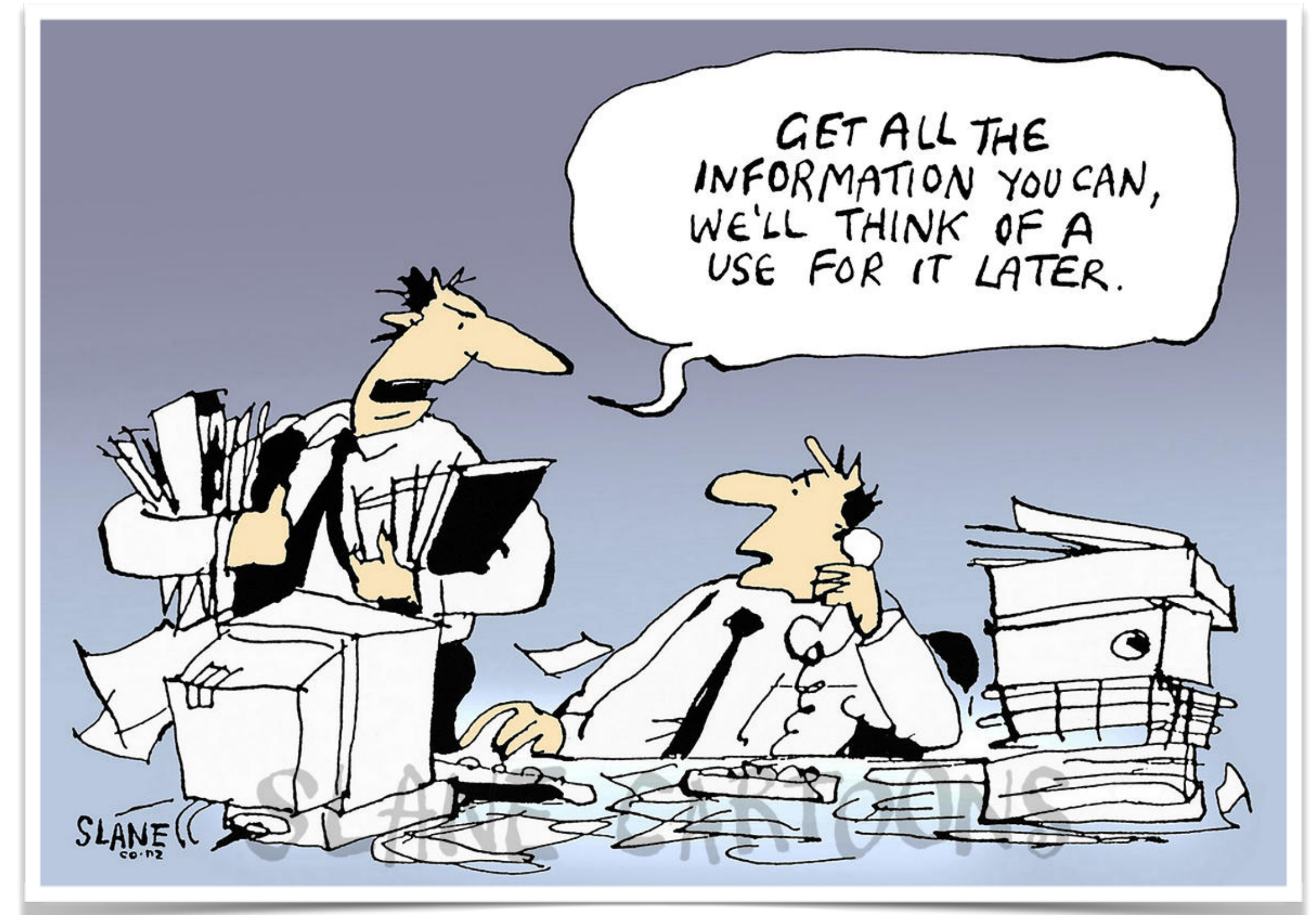
Personal data should be

- **adequate**

- **relevant** and

WHAT IS *DATA MINIMISATION?*

Personal data should be

- **adequate**

- **relevant** and

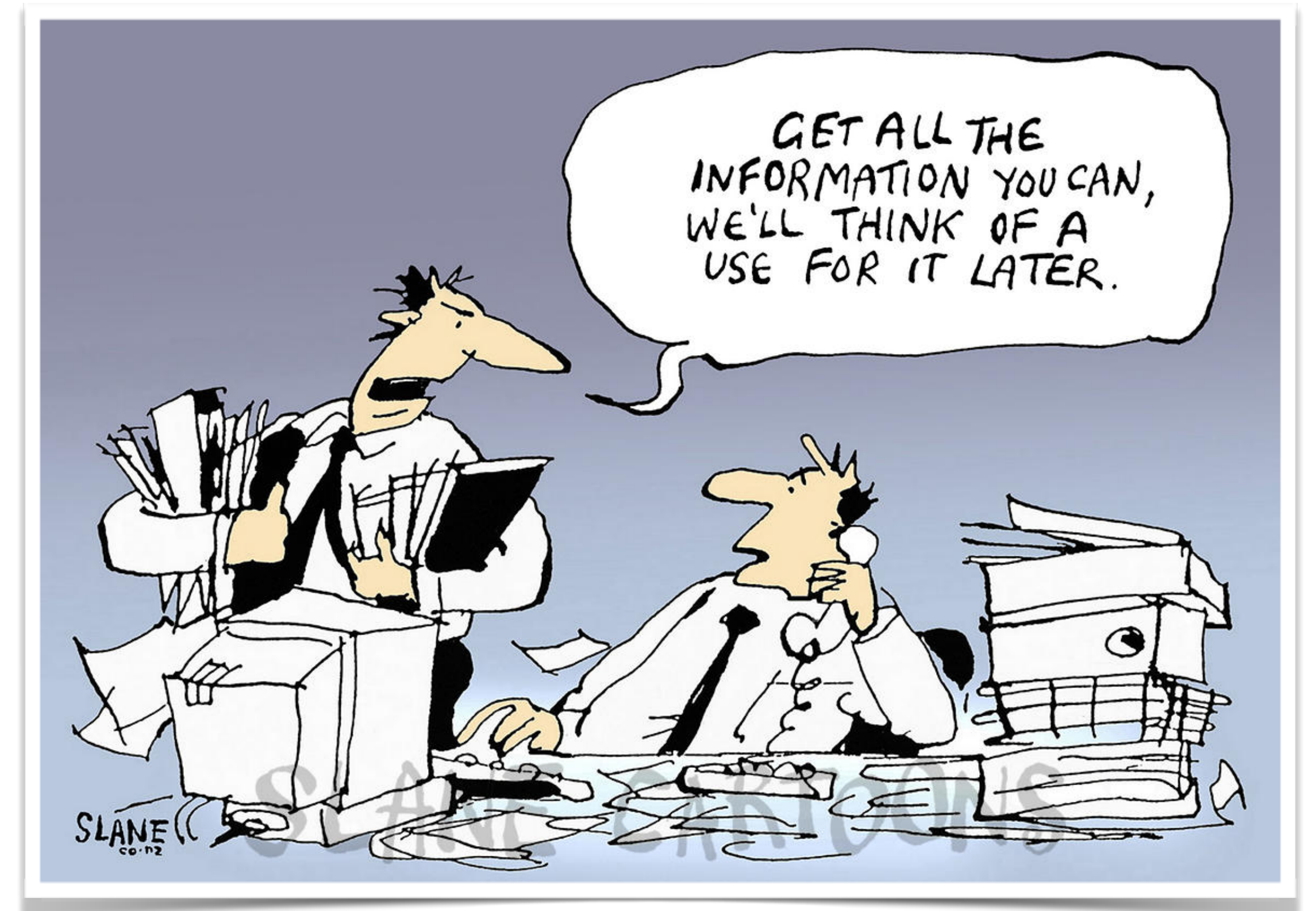- **limited to what is necessary**

WHAT IS *DATA MINIMISATION?*

Personal data should be

- **adequate**

- **relevant** and

- **limited to what is necessary**

in relation to the **purposes** for which they are processed

WHAT IS *DATA MINIMISATION?*

Personal data should be

- **adequate**

- **relevant** and

- **limited to what is necessary**

in relation to the **purposes** for which they are processed

OUTLINE

▷ Motivation

▷ Foundation

▷ Research Problem

▷ Approach

▷ Summary § Outlook

OUTLINE

▷ **Motivation**

    ▷ **Privacy by Design via Workflows**

▷ Foundation

▷ Research Problem

▷ Approach

▷ Summary § Outlook

MOTIVATING EXAMPLE: *NEWBORN SCREENING*

demographic data

sensitive blood data

Desk

Lab

Pediatrician

medical data

Dieses Feld mit den **Daten der Mutter** ausfüllen:

Krankenkasse bzw. Kostenträger

Name, Vorname des Versicherten

geb. am

Kassen-Nr.   Versicherten-Nr.   Status

Betriebsstätten-Nr.   Arzt-Nr.   Datum

Leerkarten-Grund: ☐ verstorben
(bei Einsendung ohne ☐ Verlegung
Material ankreuzen) ☐ Entl. < 36 h

Screening-ID

Daten des Kindes:

Nachname

Geburtsdatum

Tag   Monat   Jahr   Std.   Min.

Geschlecht   Gestationswoche   ☐ Mehrling

☐ M   ☐ W   +   lfd. Nummer

Labor-Nr.

Telefonnummer der **Mutter** mit Vorwahl

Einsender

Telefonnummer des **Einsenders** mit Vorwahl

Abrechnung: ☐ **Privat**

Privatversicherte bitte hier unterschreiben

Besonderes:

☐ Transfusion

am:

☐ **weiteres:**

SN B00362101   SN B00362101

Bitte vollständig durchtränken

Vorname

Datum/Uhrzeit der Abnahme: Geburtsgewicht   Geburtenbuch-Nr.

Tag   Monat   Jahr   Std.   Min.   g.

Hörscreening: unauffällig   auffällig

Wiederholungs-untersuchung   ☐ nicht durchgeführt   TEOAE: **bds.** ☐   ☐ R ☐ L

AABR: **bds.** ☐   ☐ R ☐ L
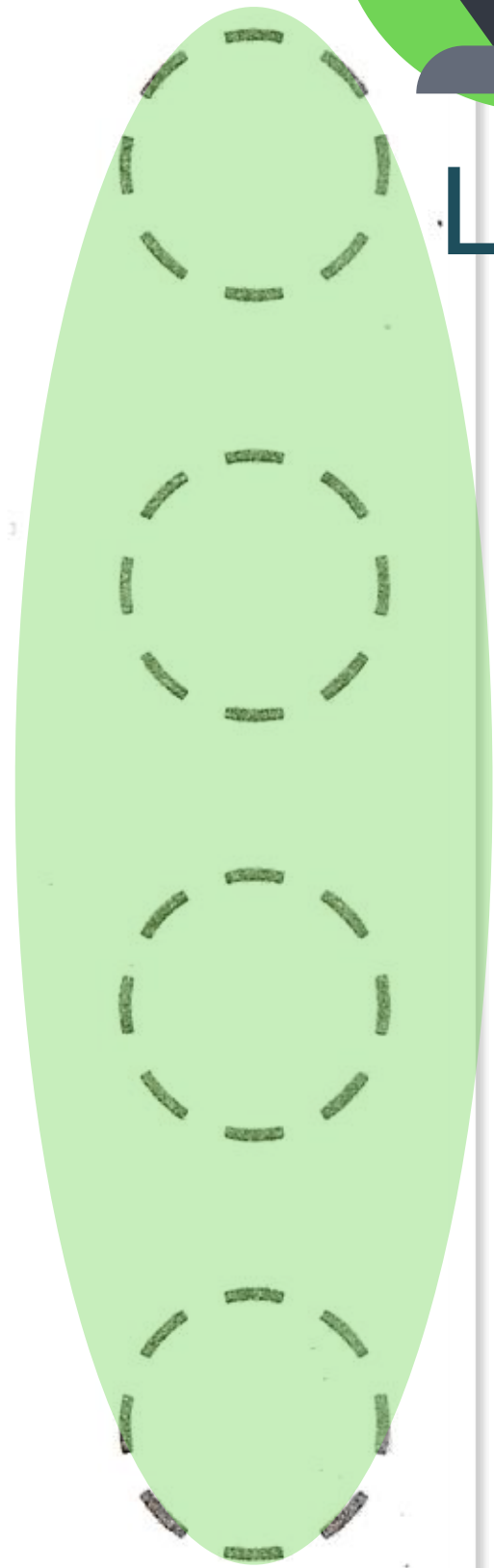
MOTIVATION: PRIVACY BY DESIGN

GDPR says: Consider privacy at design phase…

MOTIVATION: PRIVACY BY DESIGN

GDPR says: Consider privacy at design phase…

Good News: Workflows might help!

MOTIVATION: PRIVACY BY DESIGN

GDPR says: Consider privacy at design phase…

Good News: Workflows might help!
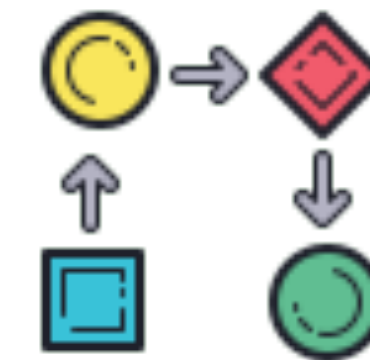
A Workflow includes a series of tasks *to achieve a goal*

MOTIVATION: PRIVACY BY DESIGN

GDPR says: Consider privacy at design phase…

Good News: Workflows might help!
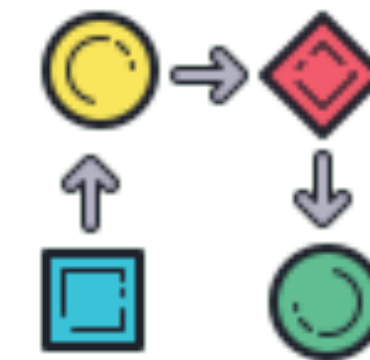
A Workflow includes a series of tasks *to achieve a goal*

▸ also how tasks are performed, in what order, and by whom

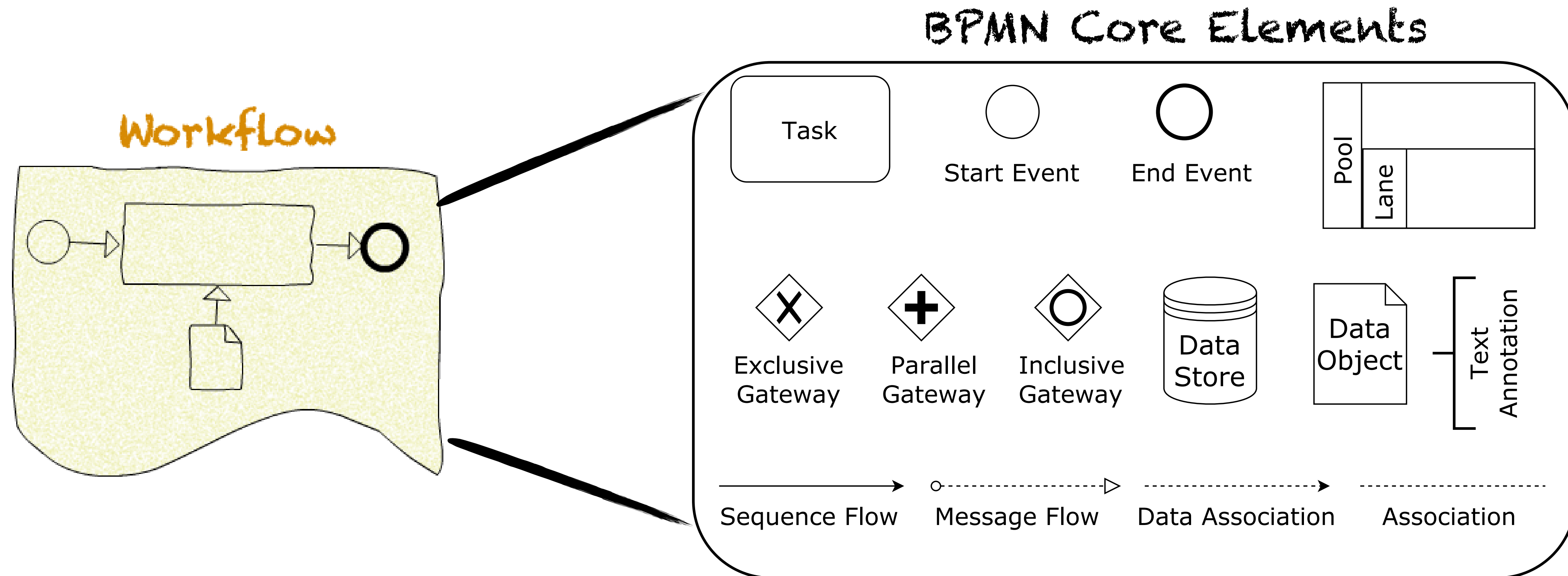# PRIVACY BY DESIGN VIA WORKFLOWS

PRIVACY BY DESIGN VIA WORKFLOWS

Workflow (Model) ≈ *Business Process Modeling Notation (BPMN) Model*

PRIVACY BY DESIGN VIA WORKFLOWS

Workflow (Model) ≈ Business Process Modeling Notation (BPMN) Model

BPMN Core Elements

Task — Start Event — End Event — Pool / Lane

Exclusive Gateway — Parallel Gateway — Inclusive Gateway — Data Store — Data Object — Text Annotation

Sequence Flow — Message Flow — Data Association — Association
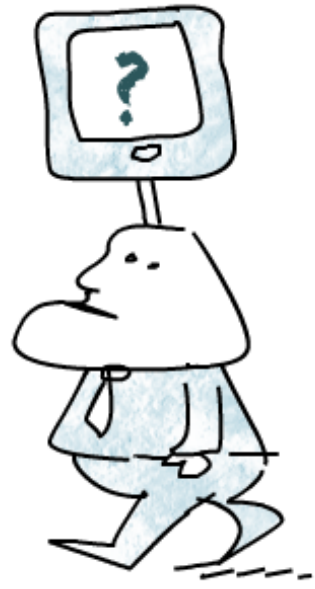
OUTLINE
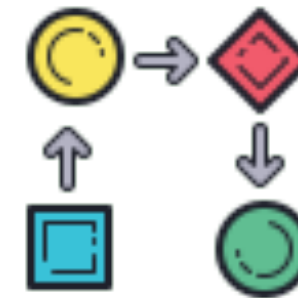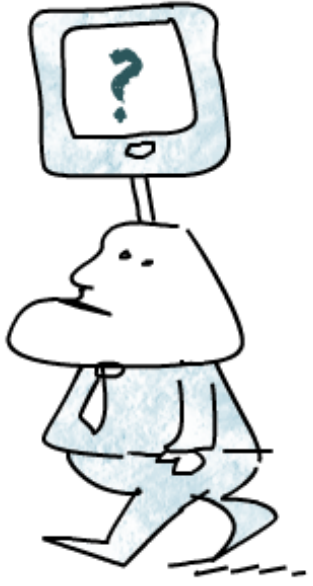
FOUNDATION

Which sources needed to handle data minimisation?
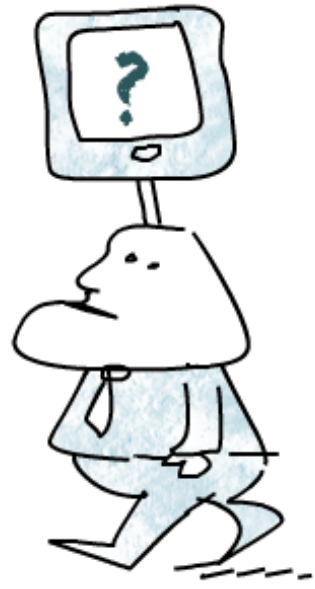
FOUNDATION

Which sources needed to handle data minimisation?

1- *Data-Aware* Workflow

FOUNDATION

Which sources needed to handle data minimisation?

1- *Data-Aware* Workflow

Which *data attributes* are (potentially) used for which *purpose* in the Workflow

FOUNDATION

Which sources needed to handle data minimisation?

1- *Data-Aware* Workflow

Which *data attributes* are (potentially) used for which *purpose* in the Workflow
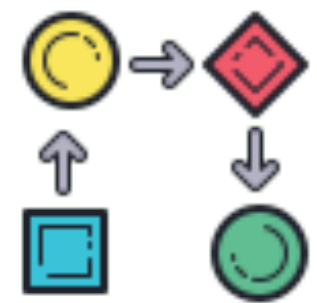
2- **Privacy Policy**
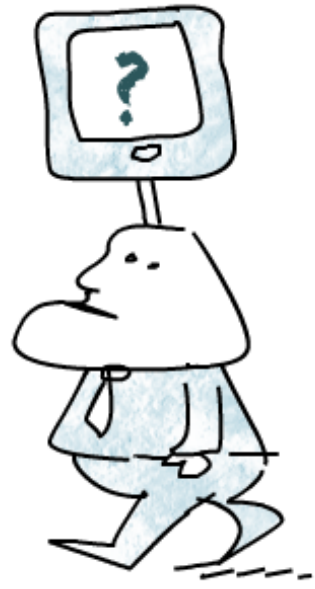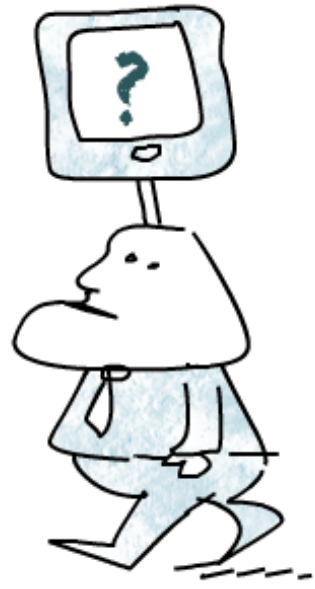
FOUNDATION

Which sources needed to handle data minimisation?

1- *Data-Aware* Workflow ← Which *data attributes* are (potentially) used for which *purpose* in the Workflow
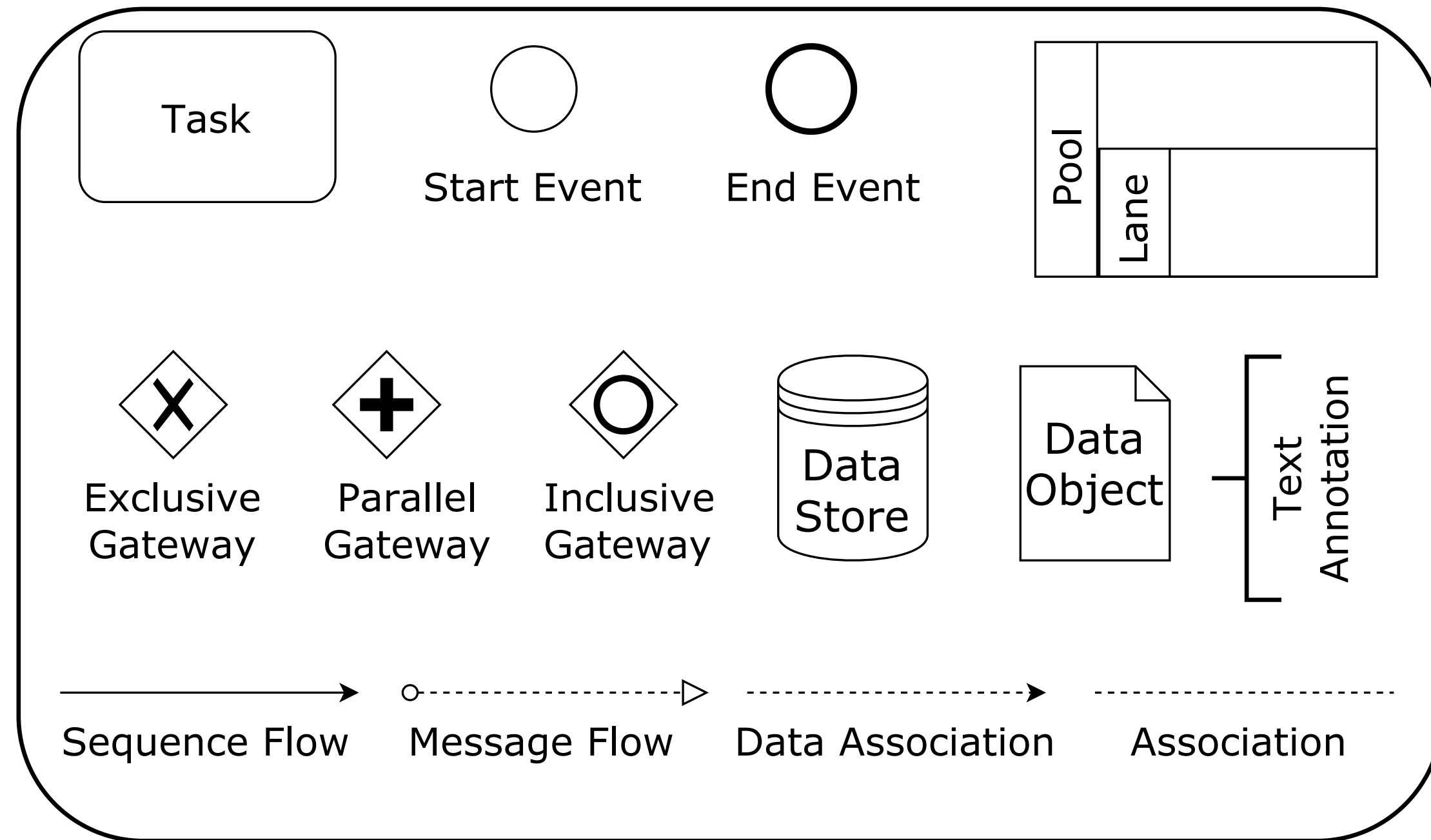
2- **Privacy Policy** ← Which *data attributes* are required to accomplish a certain *purpose*

Workflow

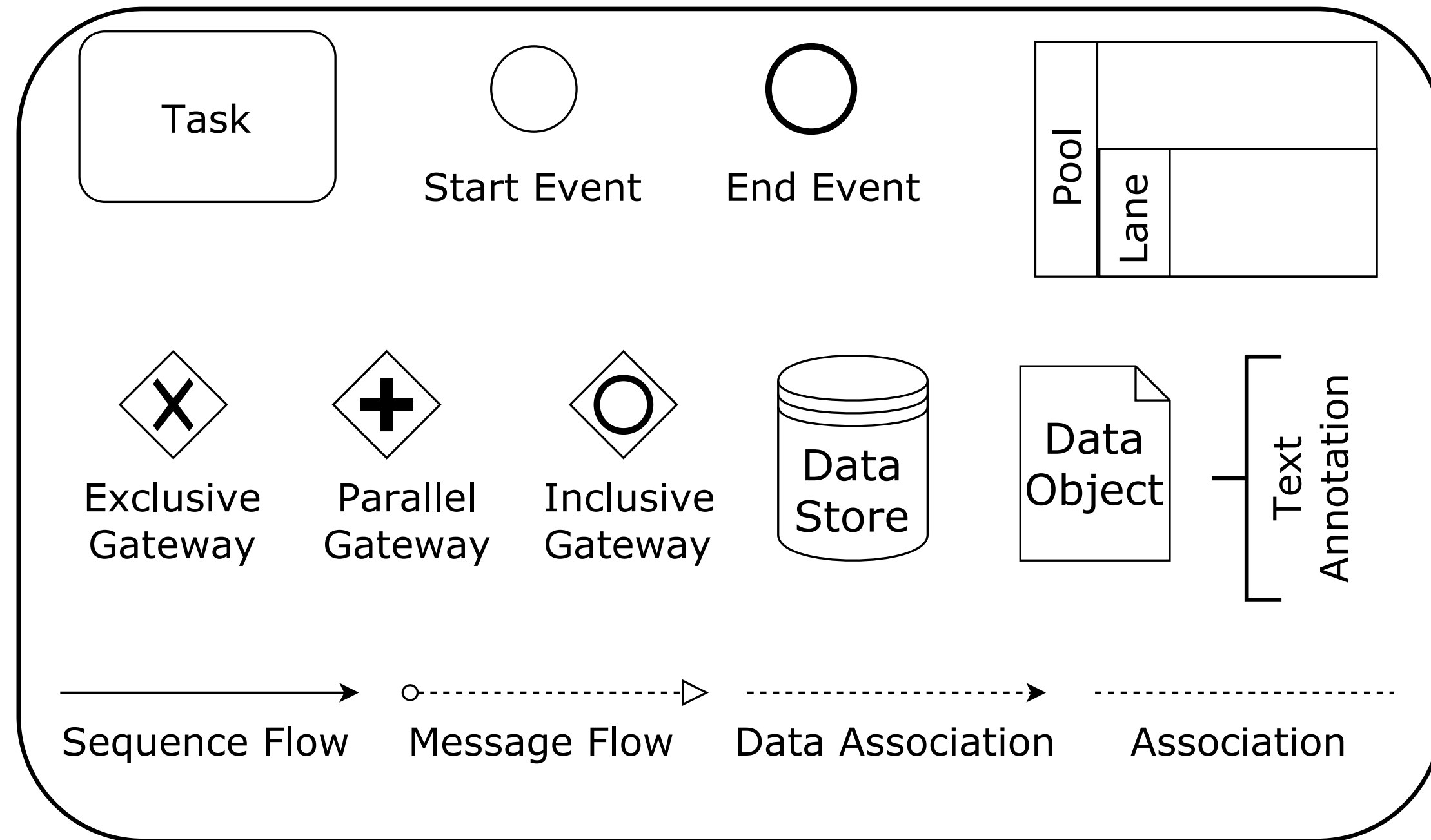|			
Task	Start Event	End Event	Pool / Lane
Exclusive Gateway	Parallel Gateway	Inclusive Gateway	Data Store
Data Object	Text Annotation		
Sequence Flow	Message Flow	Data Association	Association

BPMN Core Elements

Workflow ⟶ Data-Aware Workflow

BPMN Core Elements:
- Task
- Start Event
- End Event
- Pool / Lane
- Exclusive Gateway
- Parallel Gateway
- Inclusive Gateway
- Data Store
- Data Object
- Text Annotation
- Sequence Flow
- Message Flow
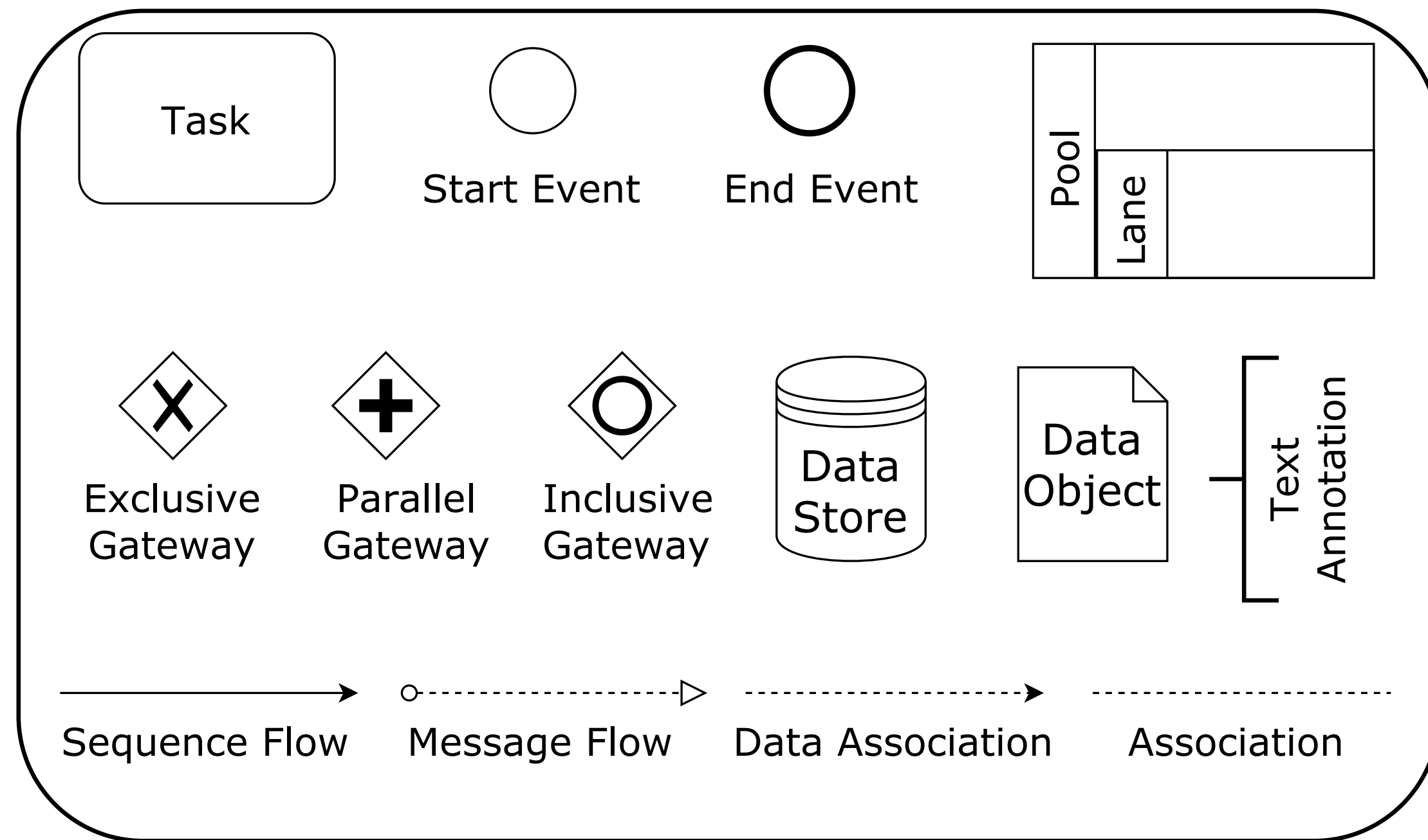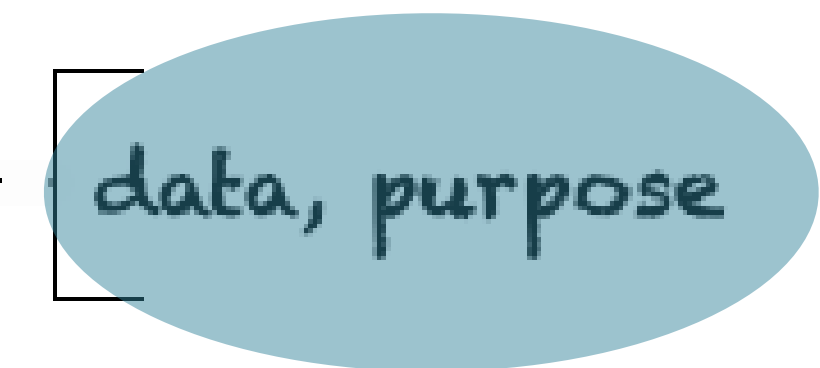- Data Association
- Association

Data Operation Task + data, purpose → Data

Workflow → Data-Aware Workflow

BPMN Core Elements:
- Task
- Start Event
- End Event
- Pool / Lane
- Exclusive Gateway
- Parallel Gateway
- Inclusive Gateway
- Data Store
- Data Object
- Text Annotation
- Sequence Flow
- Message Flow
- Data Association
- Association

+
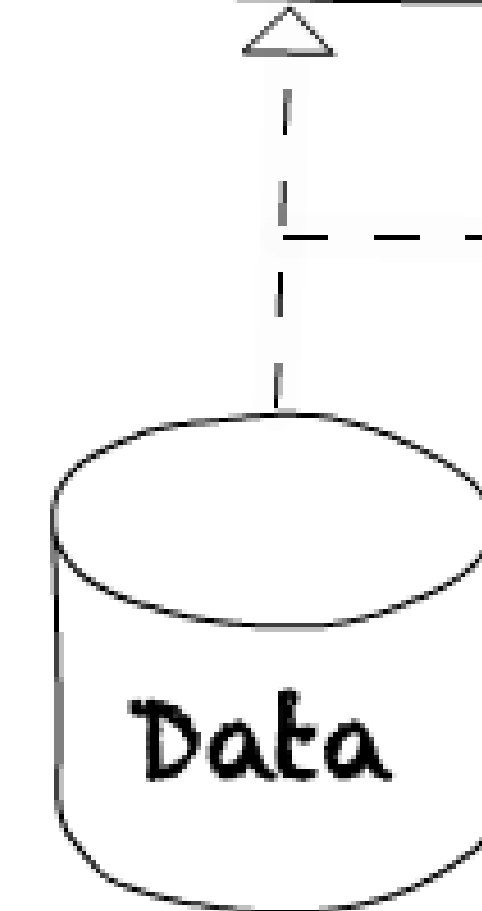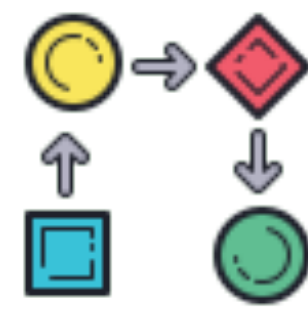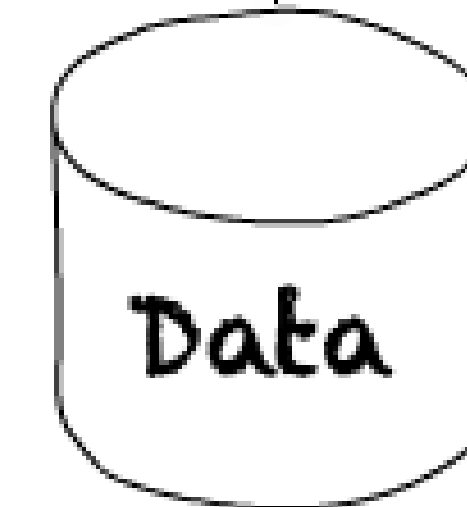
Data Operation Task
data, purpose
Data
Data Annotation

Workflow → Data-Aware Workflow

BPMN Core Elements

Task

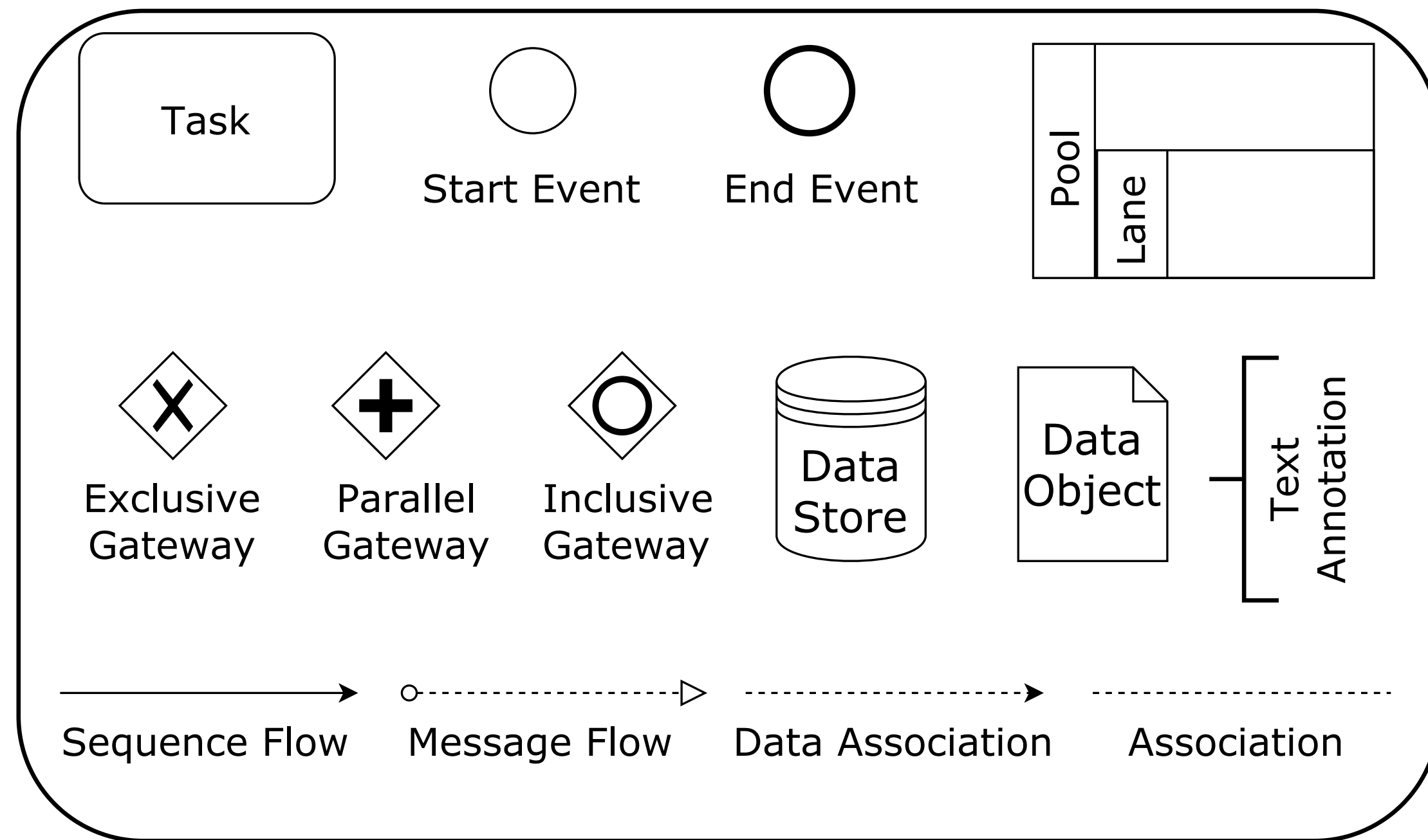Start Event   End Event

Pool / Lane

Exclusive Gateway   Parallel Gateway   Inclusive Gateway   Data Store   Data Object   Text Annotation

Sequence Flow   Message Flow   Data Association   Association

+

Data Operation Task

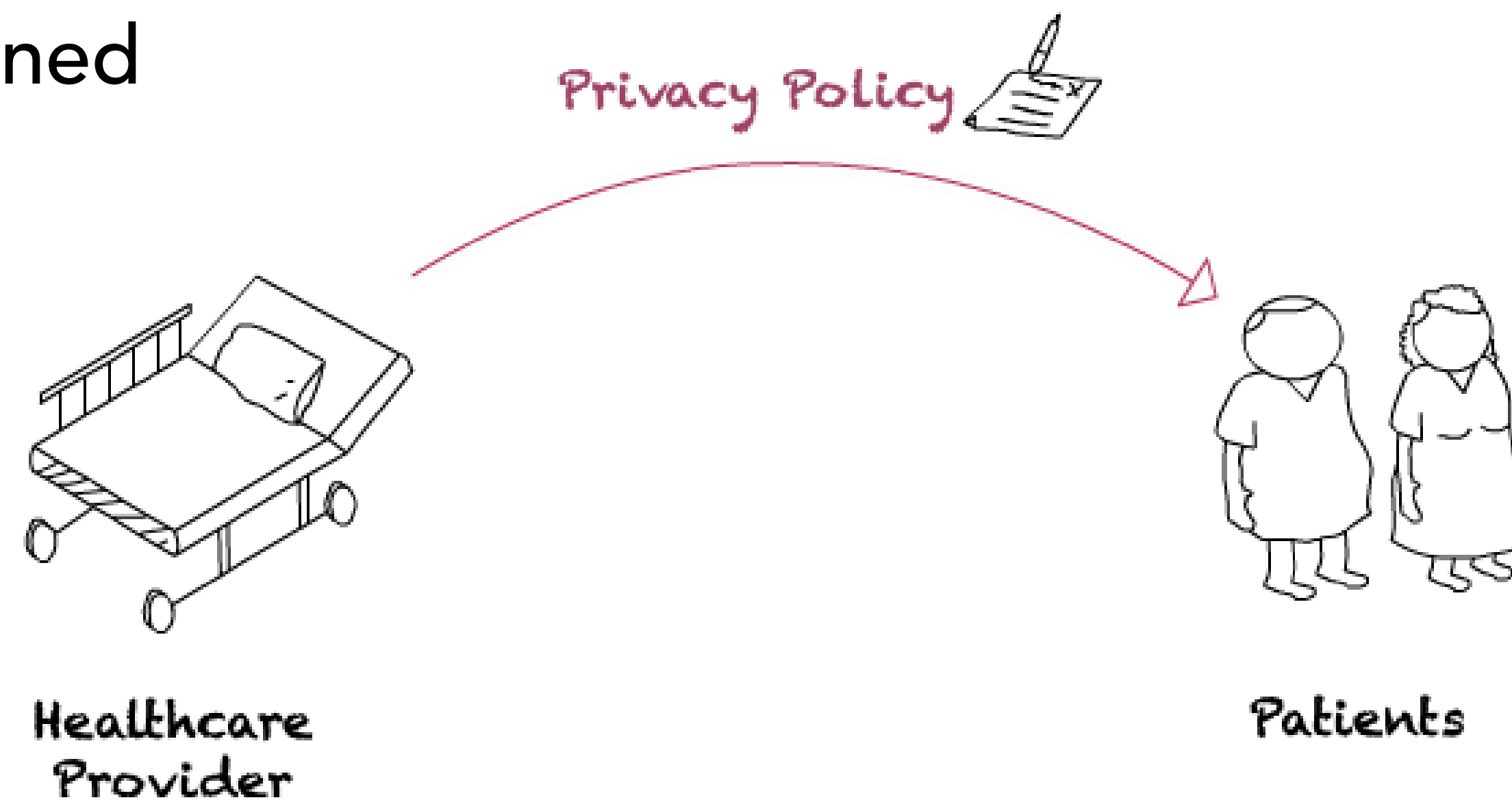data, purpose
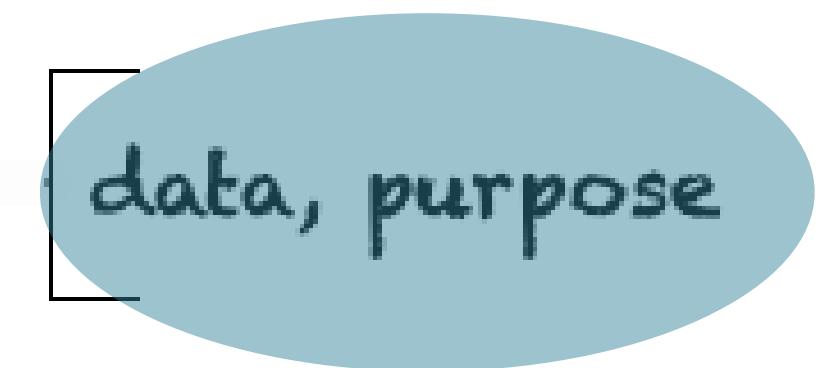
Data

Data Annotation

* Different types of Data Handling in BPMN are stated in [1]

[1] Besik, Saliha Irem, and Johann-Christoph Freytag. "Ontology-Based Privacy Compliance Checking for Clinical Workflows."
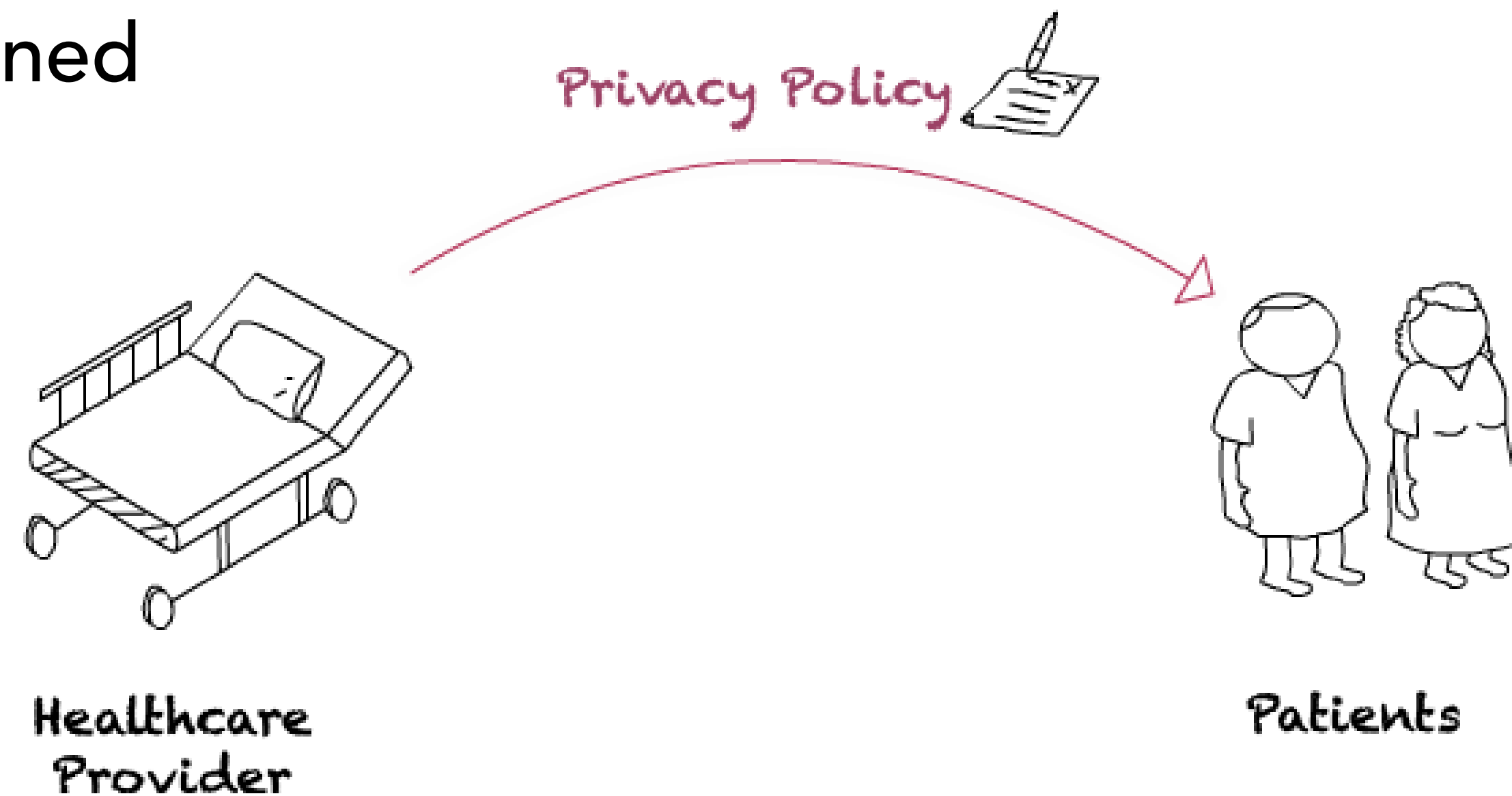
PRIVACY POLICY

- what data is collected for what purposes

- the modality of data processing, whether it is obligatory or voluntary

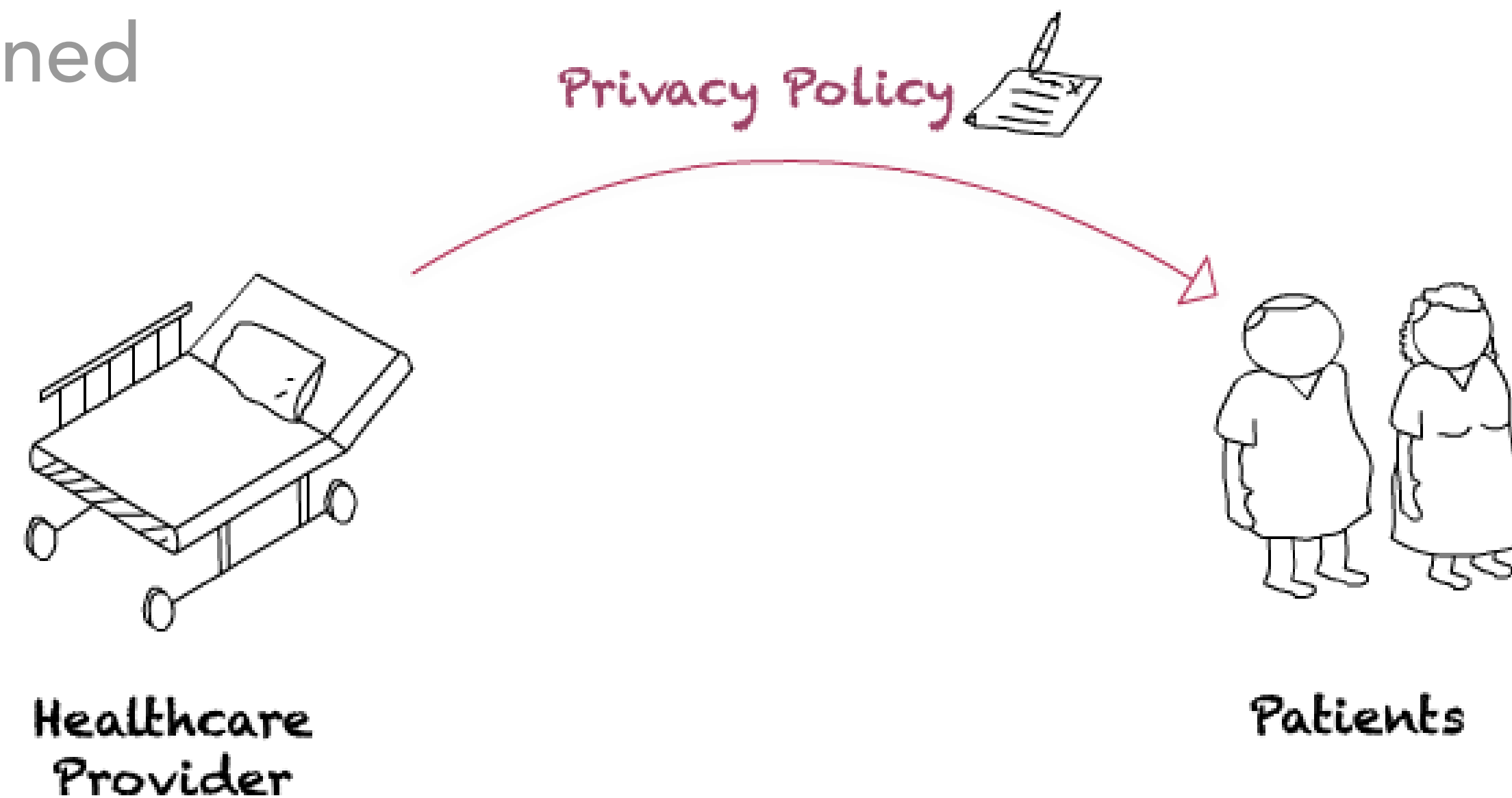- how long it is retained

PRIVACY POLICY

- what data is collected for what purposes ——————— *Data Minimization*

- the modality of data processing, whether it is obligatory or voluntary

- how long it is retained

Privacy Policy

Healthcare Provider

Patients

PRIVACY POLICY

- **what data is collected for what purposes** ———————— *Data Minimization*

- the modality of data processing, whether it is obligatory or voluntary

- how long it is retained

Privacy Policy

Healthcare
Provider

Patients

PRIVACY POLICY

- **what data is collected for what purposes** ———————— *Data Minimization*

- the modality of data processing, whether it is obligatory or voluntary

- how long it is retained

Privacy Policy

Healthcare
Provider

Patients
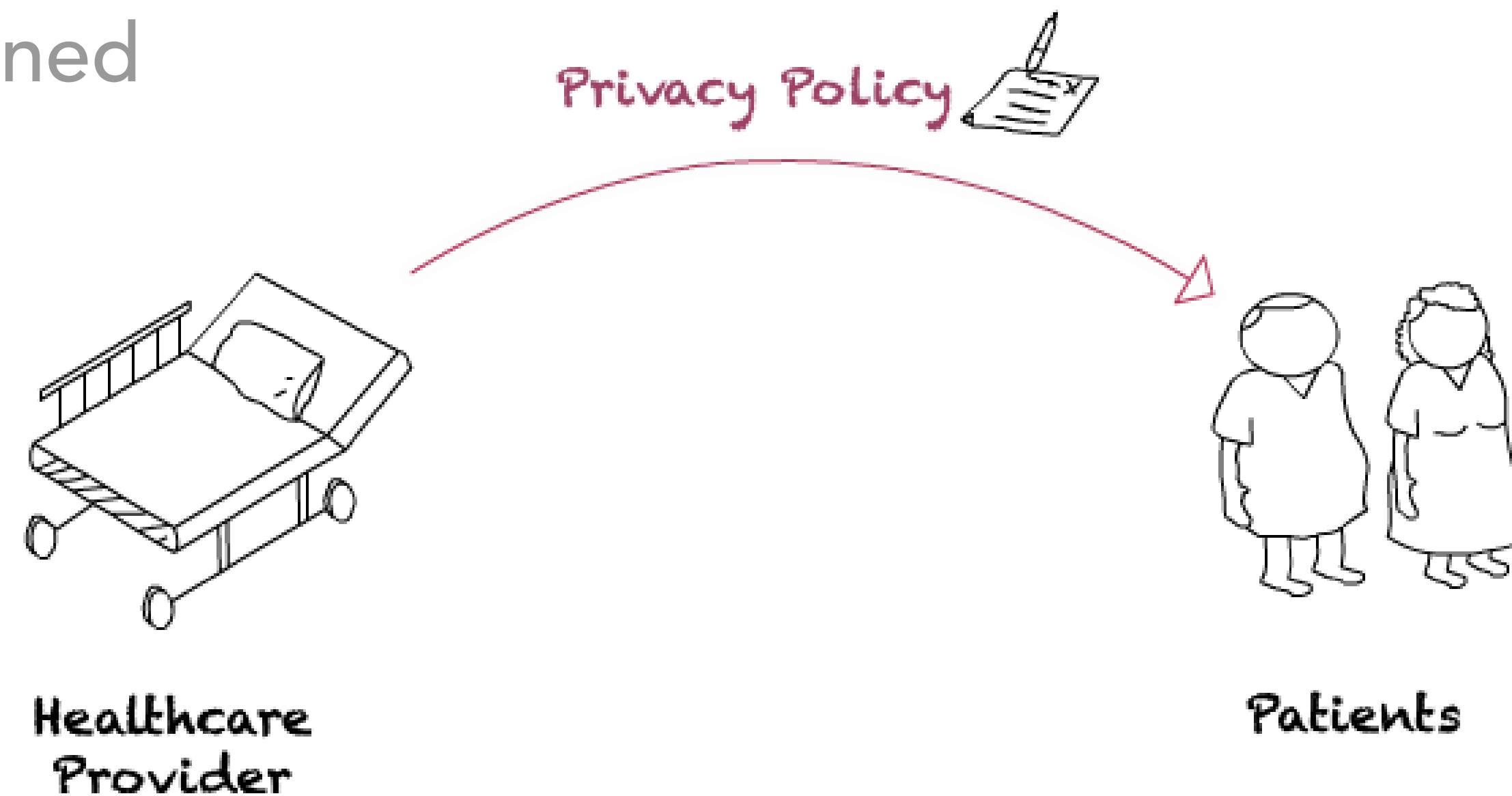
Privacy Metadata

[2] Besik, Saliha Irem, and Johann-Christoph Freytag. "A formal approach to build privacy-awareness into clinical workflows." SICS (2019): 1-12.

[3] Shastri, Supreeth, et al. "Understanding and Benchmarking the Impact of GDPR on Database Systems." appears in VLDB`2020.

OUTLINE

▷ Motivation

▷ Foundation

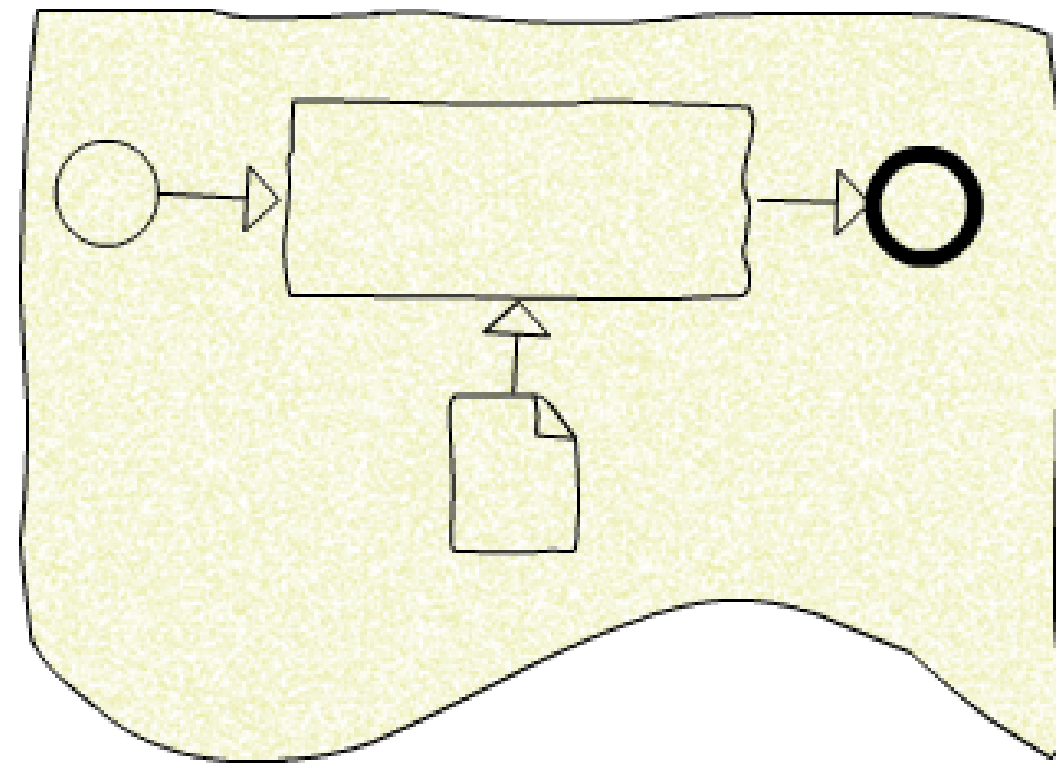▶ **Research Problem**

▷ Approach

▷ Summary § Outlook
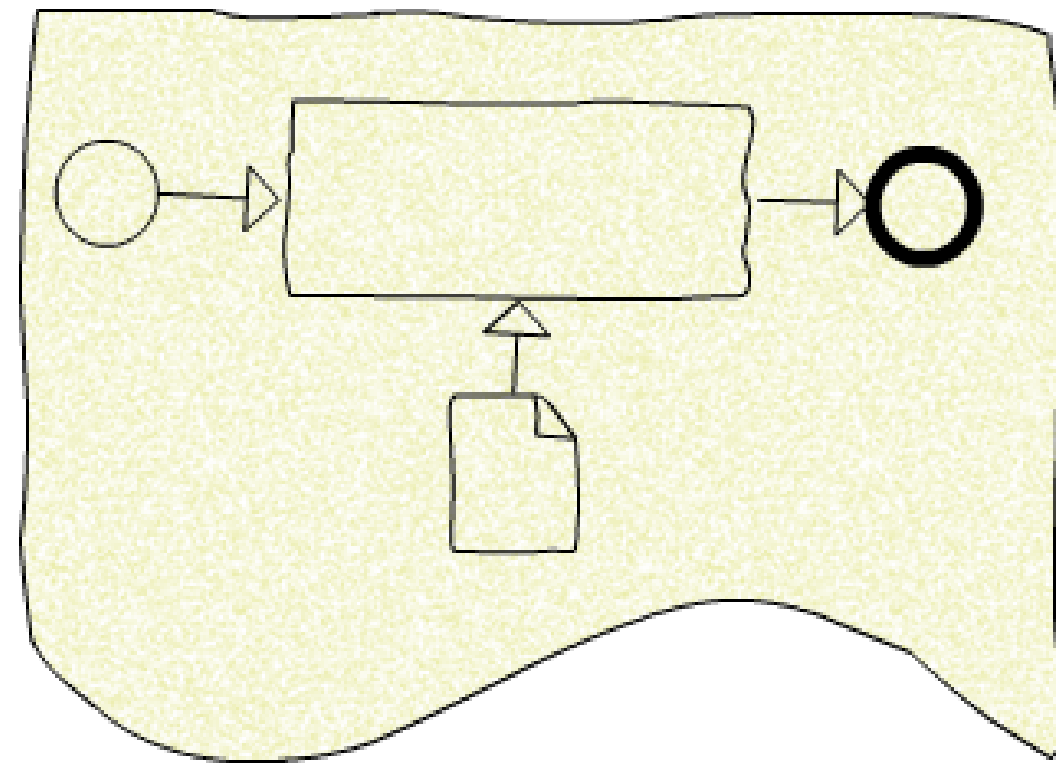
RESEARCH PROBLEM

RESEARCH PROBLEM
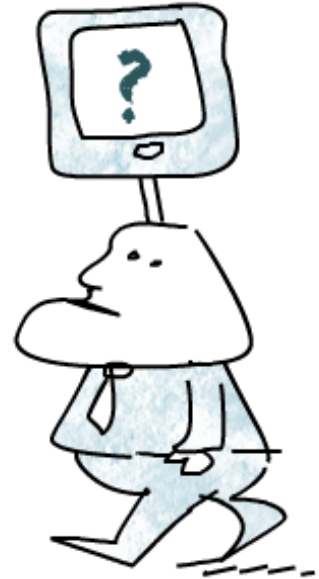

Workflow

might violate data minimisation
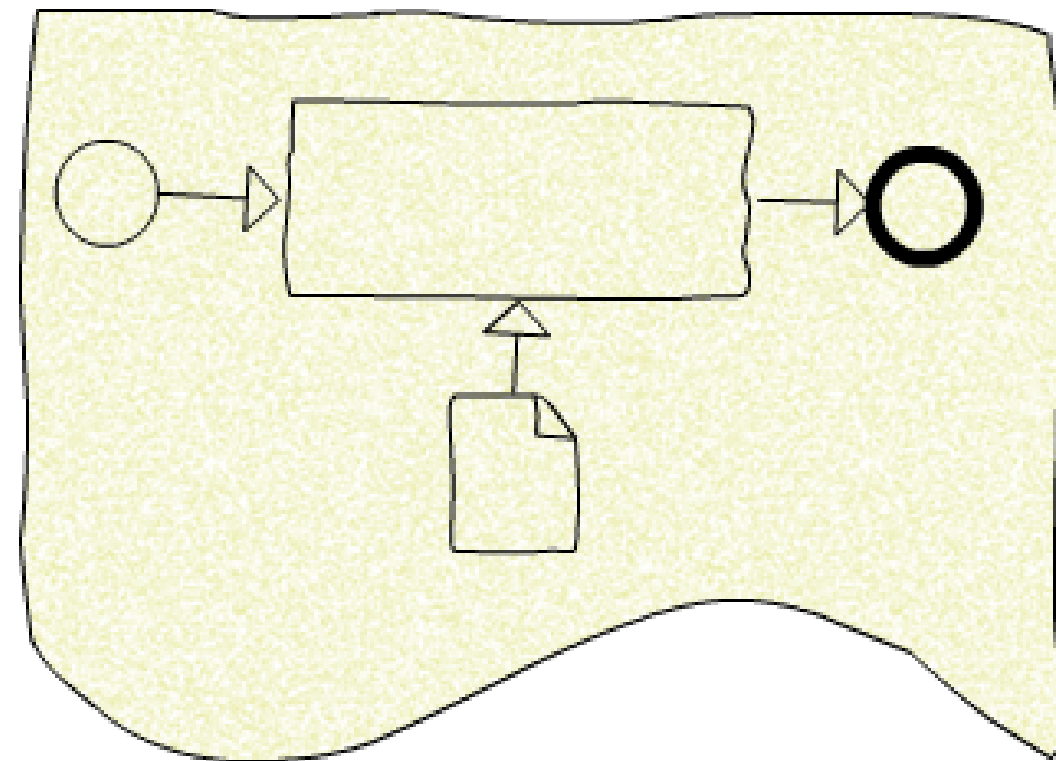
RESEARCH PROBLEM
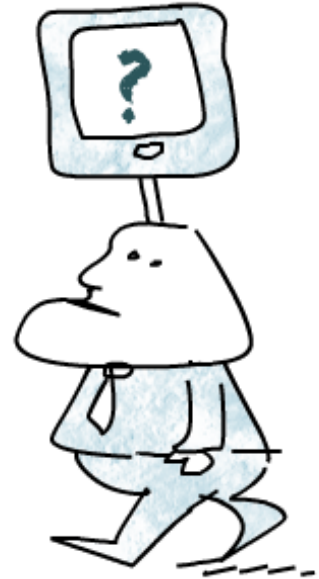

might violate data minimisation

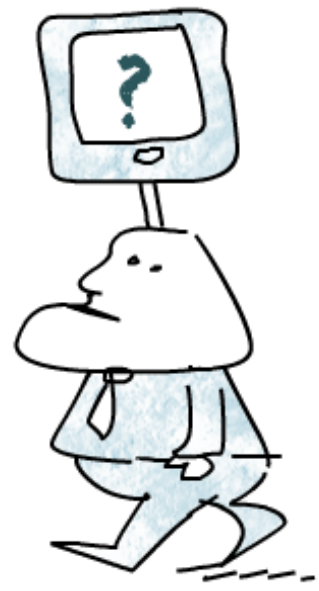How to **detect** data minimisation violations?

RESEARCH PROBLEM

Workflow

might violate data minimisation

How to **detect** data minimisation violations?

How to **recover** WFs from data minimisation violations?

When a Workflow violate data minimisation?

Personal data should be

- **adequate**

- **relevant** and

- **limited to what is necessary**

in relation to the purposes for which they are processed

When a Workflow violate data minimisation?

Personal data should be

- **adequate**

- **relevant** and

- **limited to what is necessary**

*in relation to the **purposes** for which they are processed*

Privacy violation occurs when there is

When a Workflow violate data minimisation?

Personal data should be

- **adequate**

- **relevant** and

- **limited to what is necessary**

*in relation to the **purposes** for which they are processed*

Privacy violation occurs when there is

1. **Missing Data**

When a Workflow violate data minimisation?

Personal data should be

Privacy violation occurs when there is

- **adequate** ⟶ 1. **Missing Data**

- **relevant** and ⟶ 2. **Irrelevant Data** or

- **limited to what is necessary**

*in relation to the **purposes** for which they are processed*

When a Workflow violate data minimisation?

Personal data should be

Privacy violation occurs when there is

- **adequate** ⟶ 1. **Missing Data**

- **relevant** and ⟶ 2. **Irrelevant Data** or

- **limited to what is necessary** ⟶ 3. **Redundant Data**

*in relation to the **purposes** for which they are processed*

OUTLINE

▷ Motivation

▷ Foundation

▷ Research Problem

▷ **Approach**

       ▷ **Detection**   ▷ **Missing Data**

                          ▷ **Irrelevant Data**

                          ▷ **Redundant Data**

       ▷ **Recovery**   ▷ **Irrelevant Data**

▷ Summary § Outlook

MISSING DATA - *DETECTION*

1- When a data object is to be read without having been written by any preceding task or event
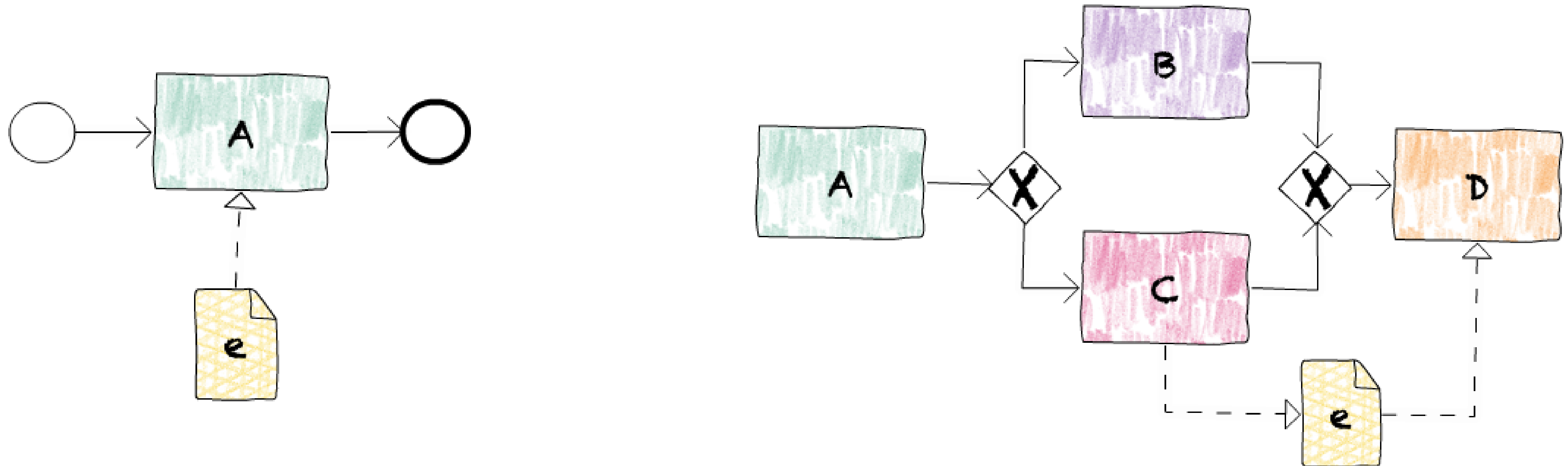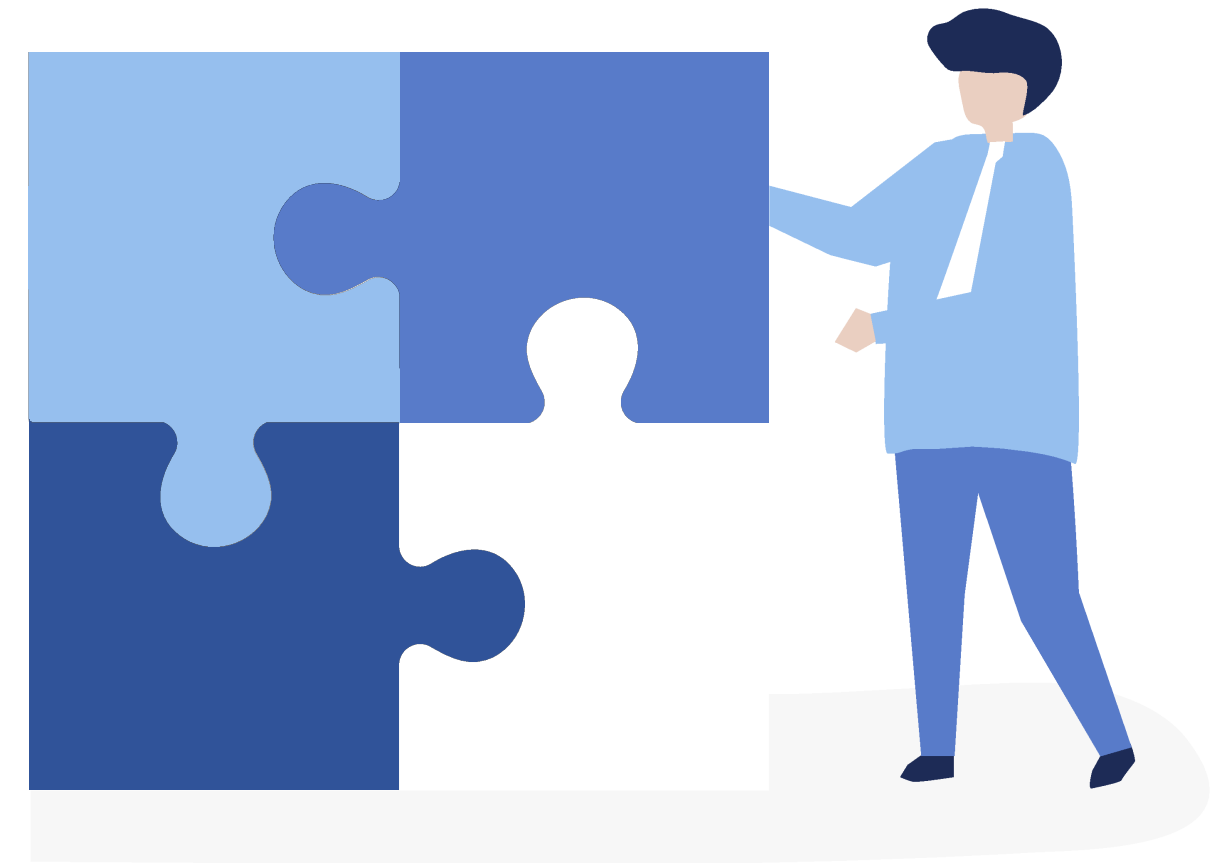
MISSING DATA - *DETECTION*

1- When a data object is to be read without having been written by any preceding task or event
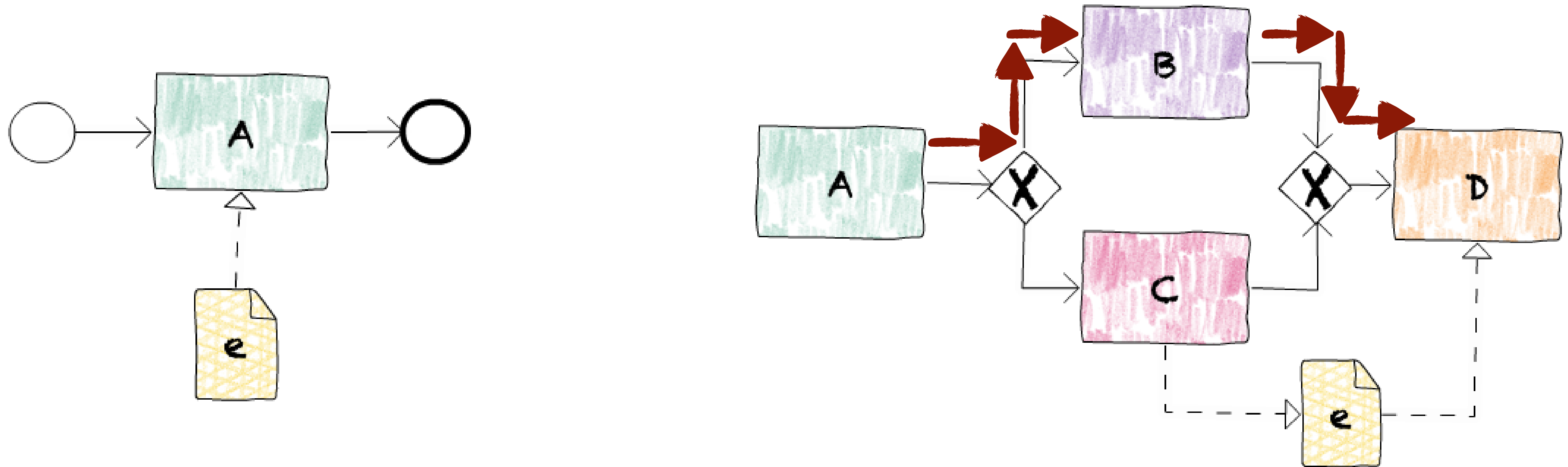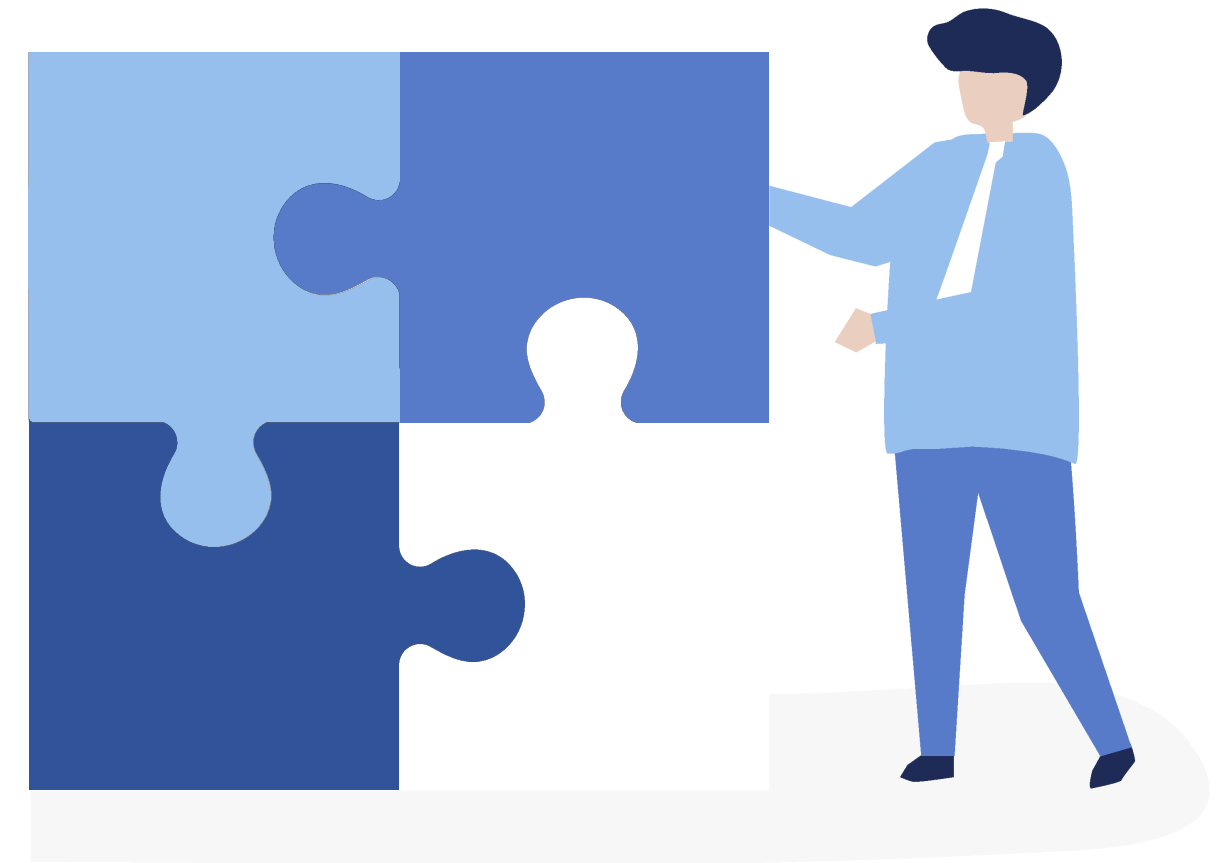
MISSING DATA - *DETECTION*

1- When a data object is to be read without having been written by any preceding task or event
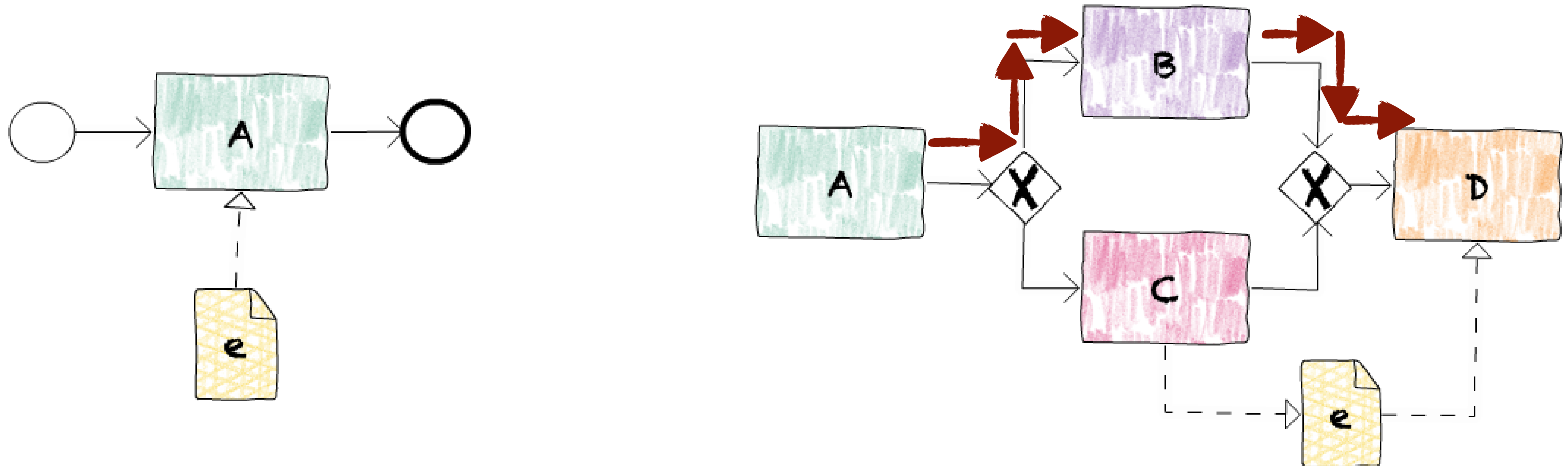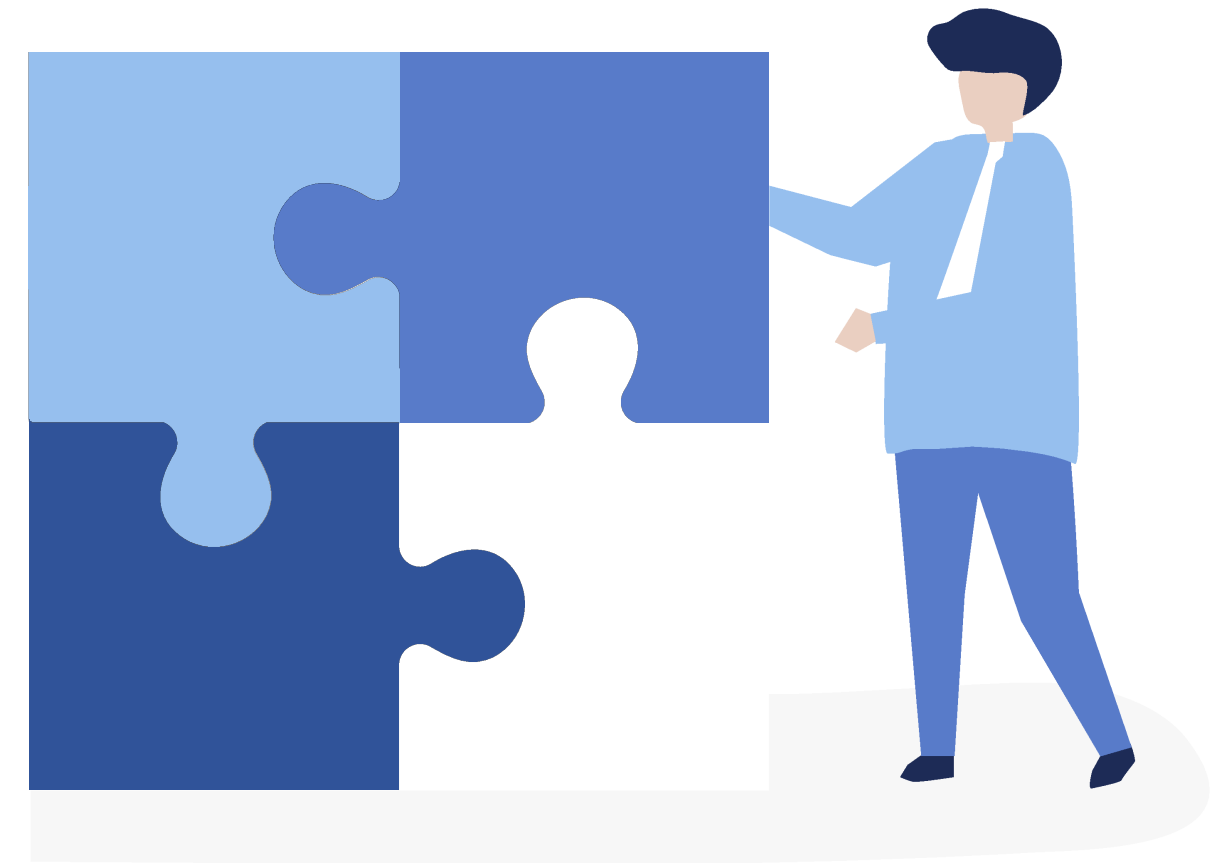
"Potential" Violator & Close World Assumption

MISSING DATA - *DETECTION*
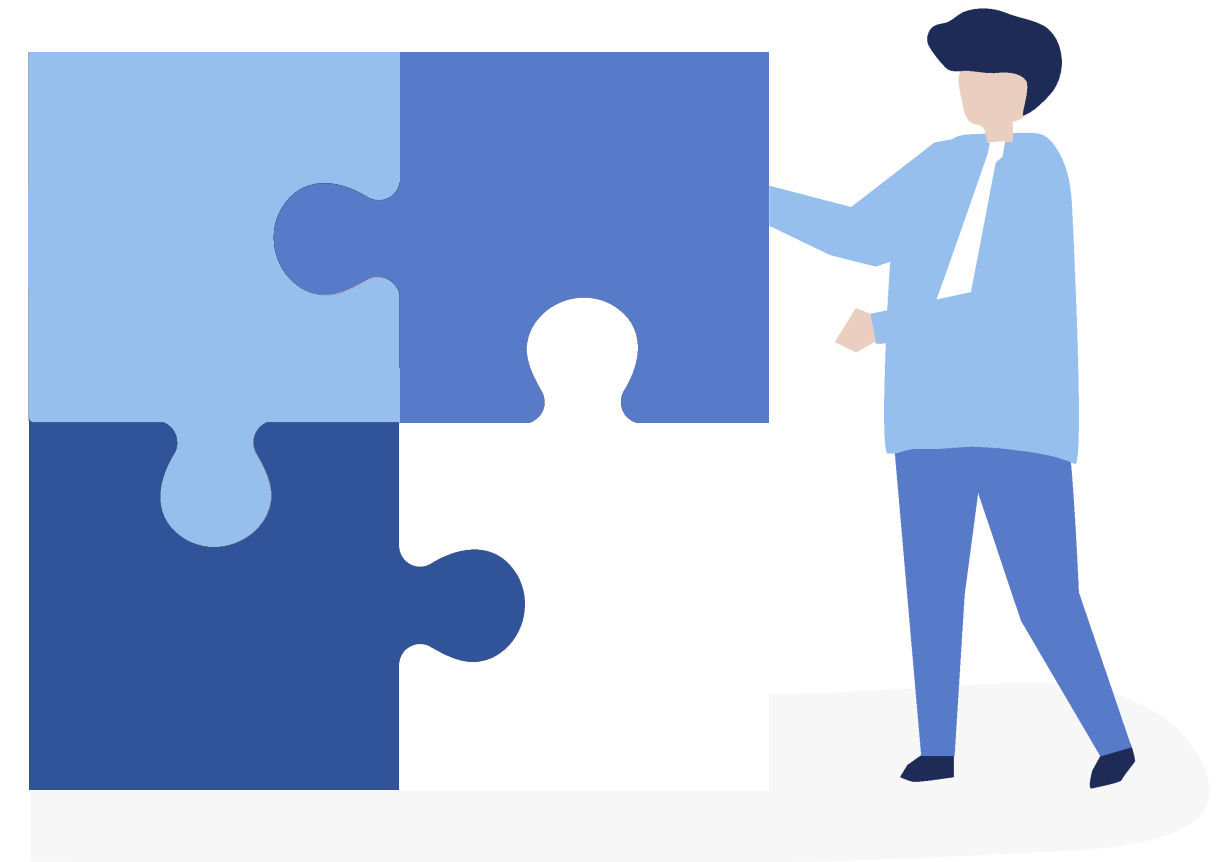
2- When some data attributes are not adequate to accomplish the stated purpose

d: age, p: treatment

d: name, p: treatment
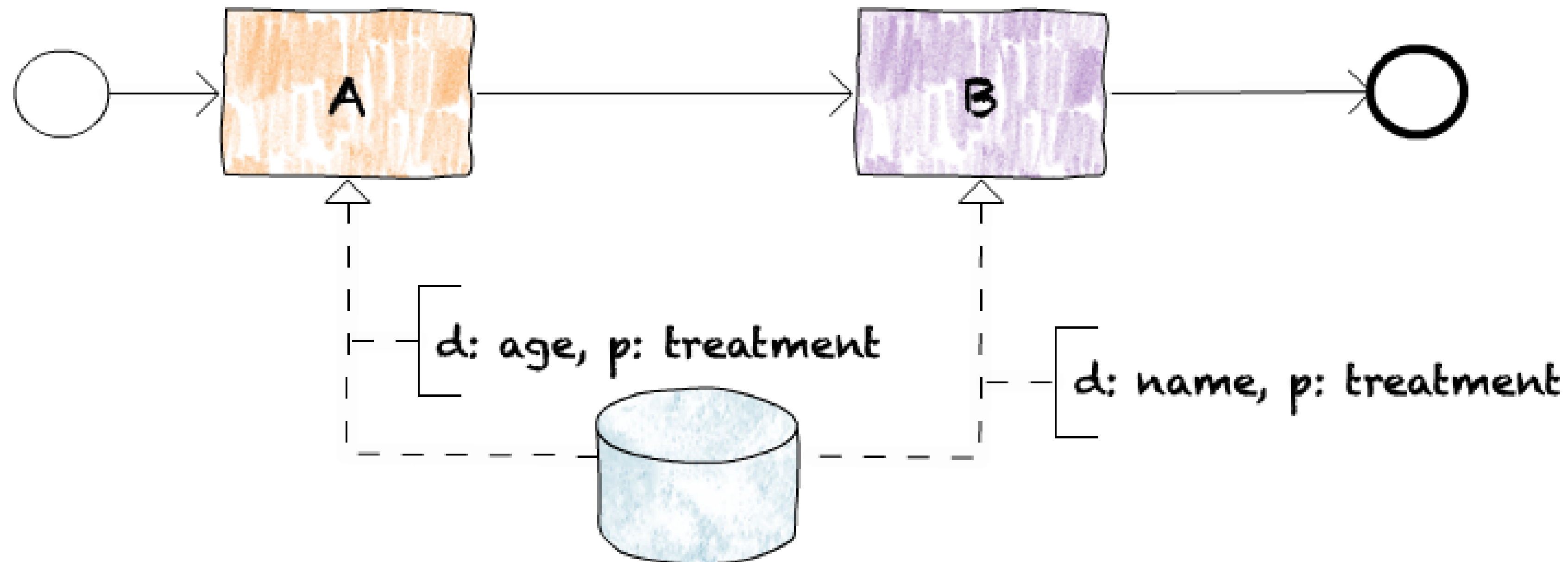
Policy: treatment requires name, age and blood-type

MISSING DATA - *DETECTION*
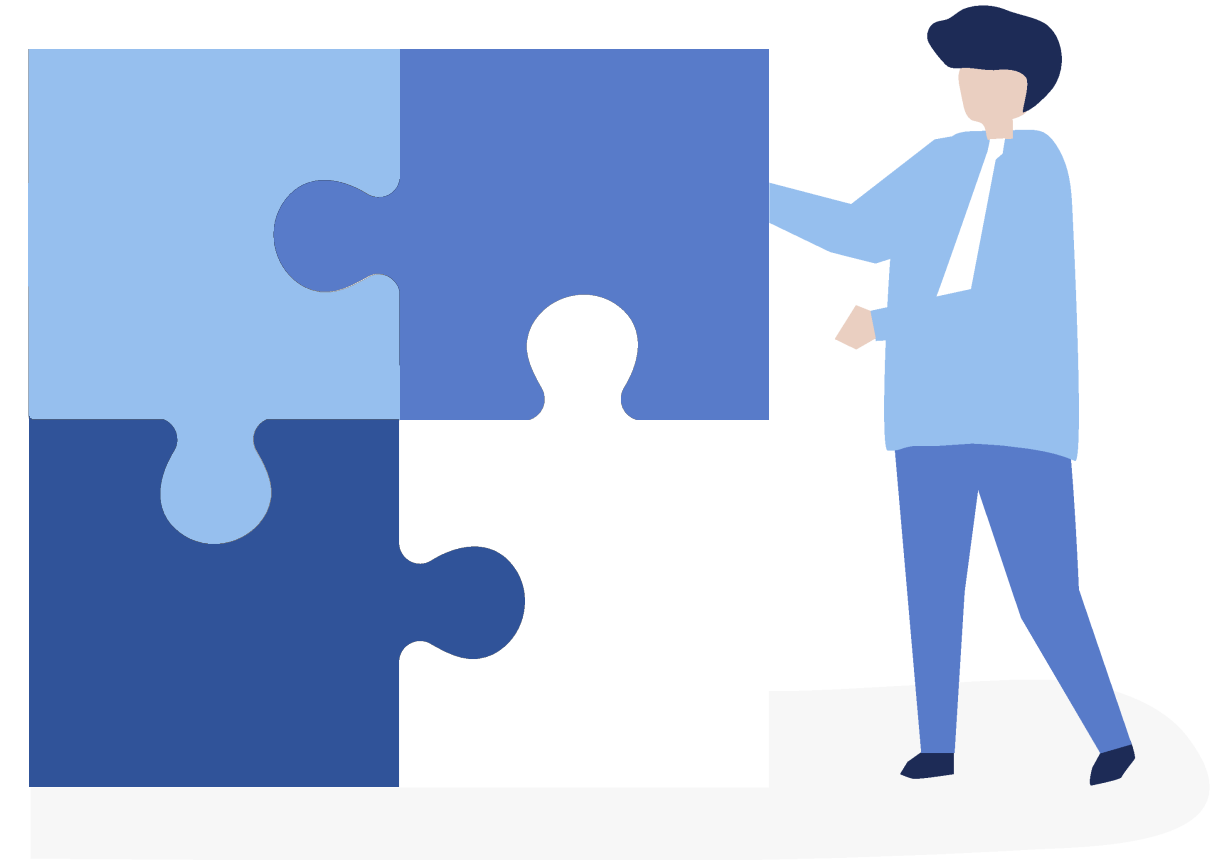
2- When some data attributes are not adequate to accomplish the stated purpose
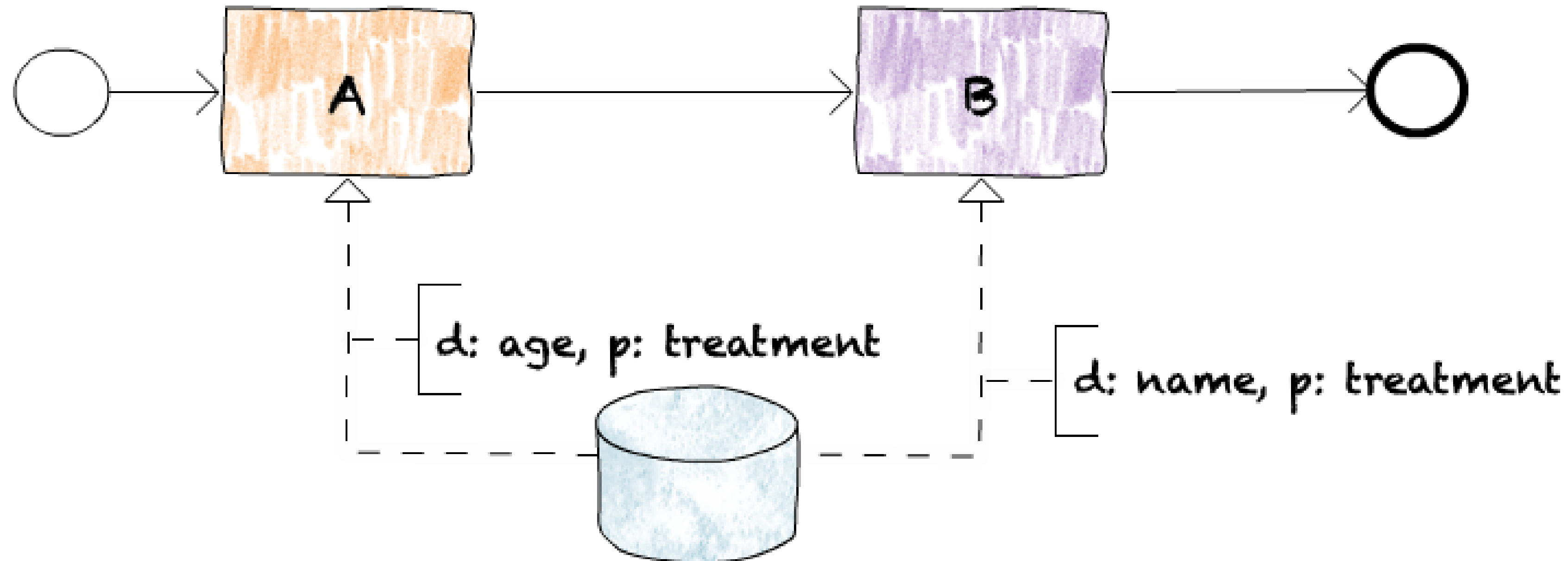
Policy: treatment requires name, age and blood-type

IRRELEVANT DATA - *DETECTION*
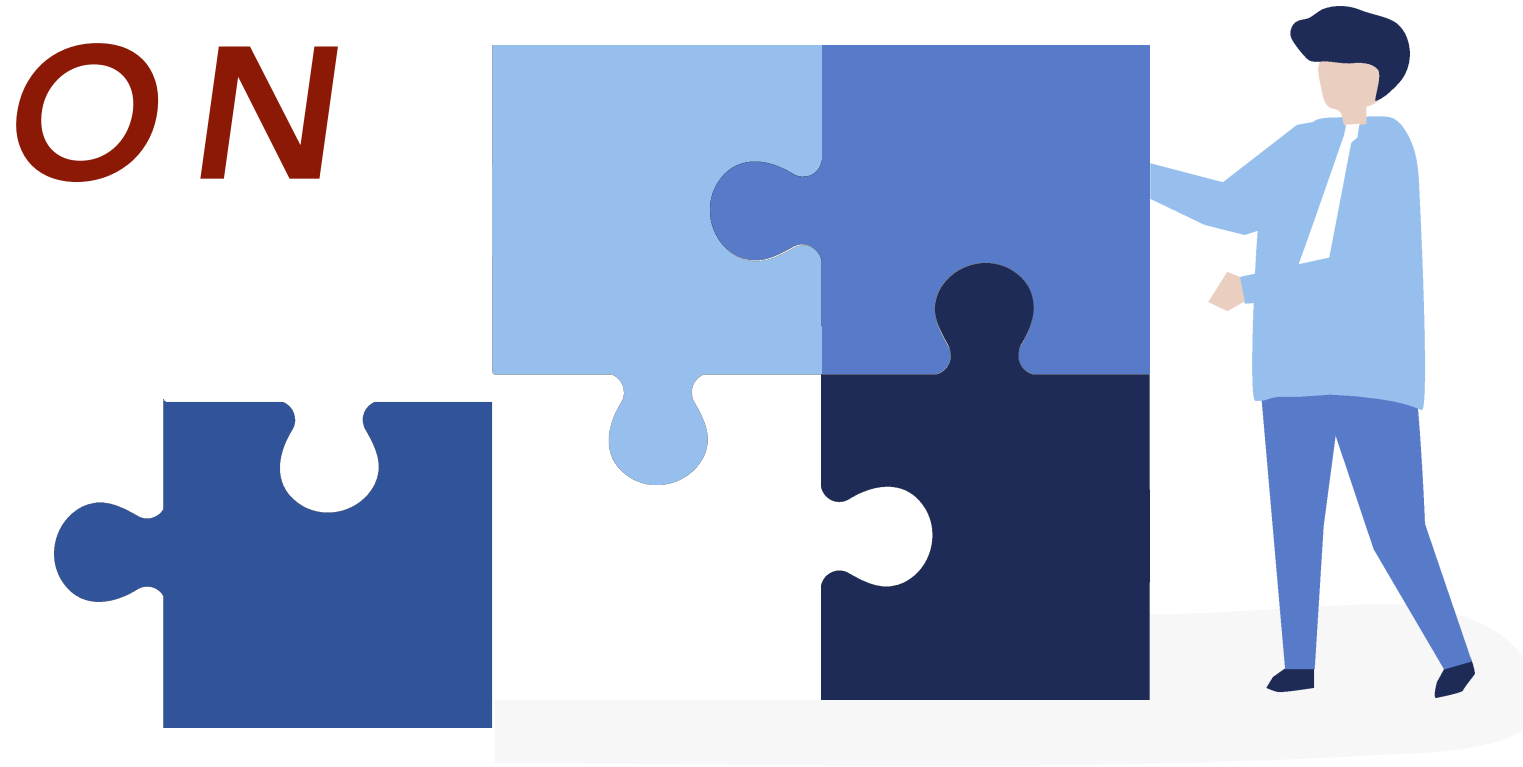
When some data attributes in the WF are not relevant

to the stated purpose
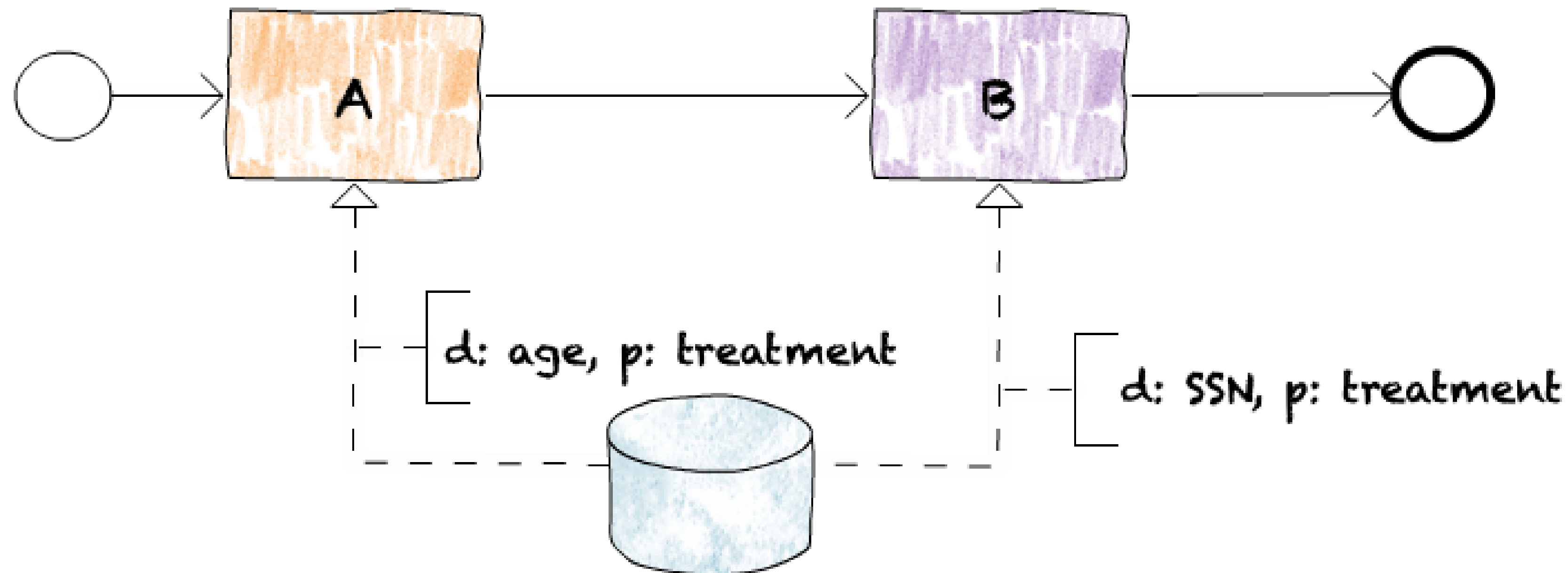
**Policy:** treatment requires name, age and blood-type

IRRELEVANT DATA - *DETECTION*
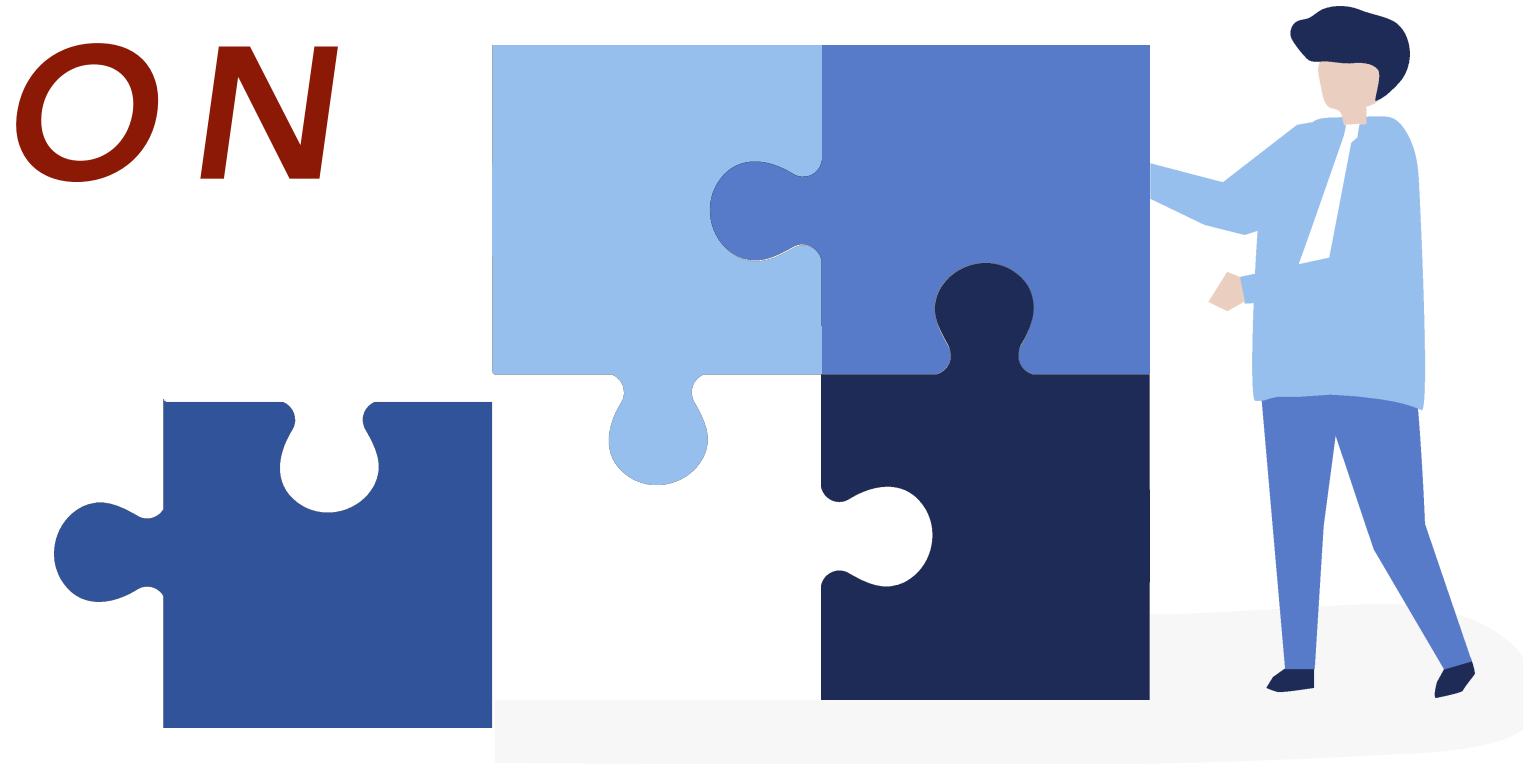
When some data attributes in the WF are not relevant

to the stated purpose
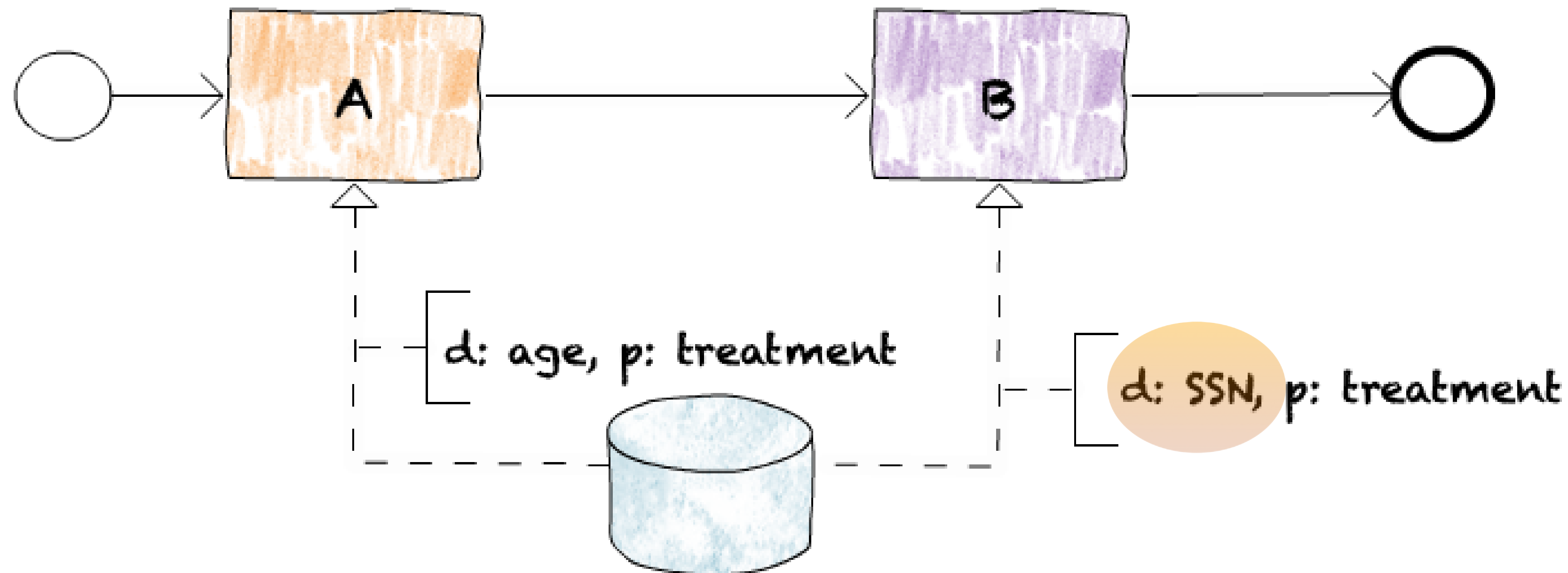
Policy: treatment requires name, age and blood-type

REDUNDANT DATA - *DETECTION*

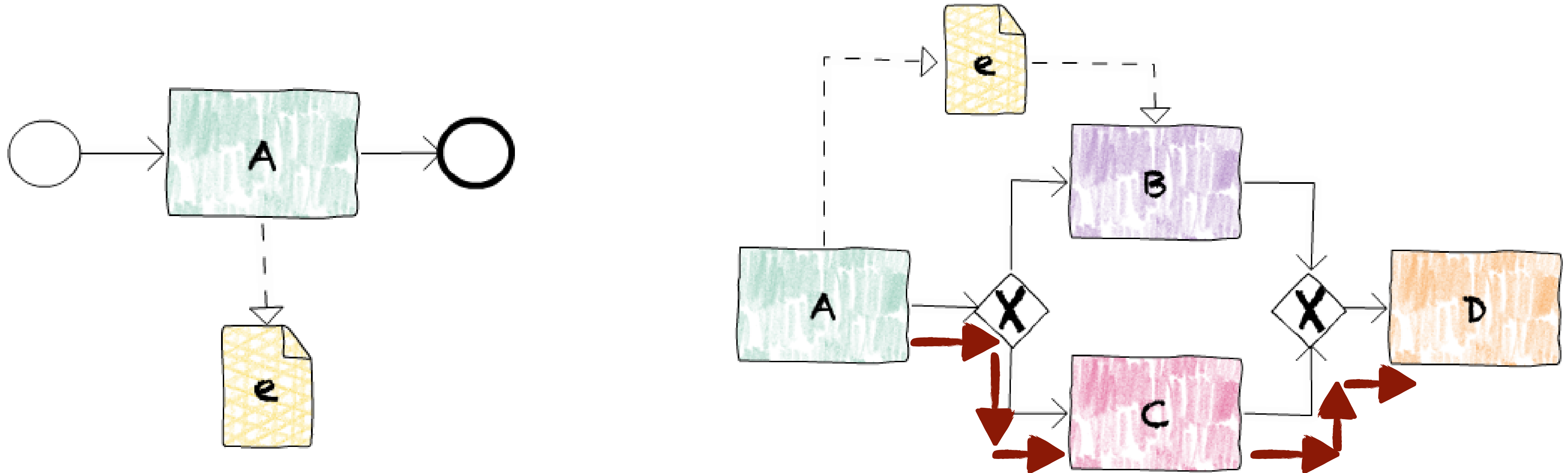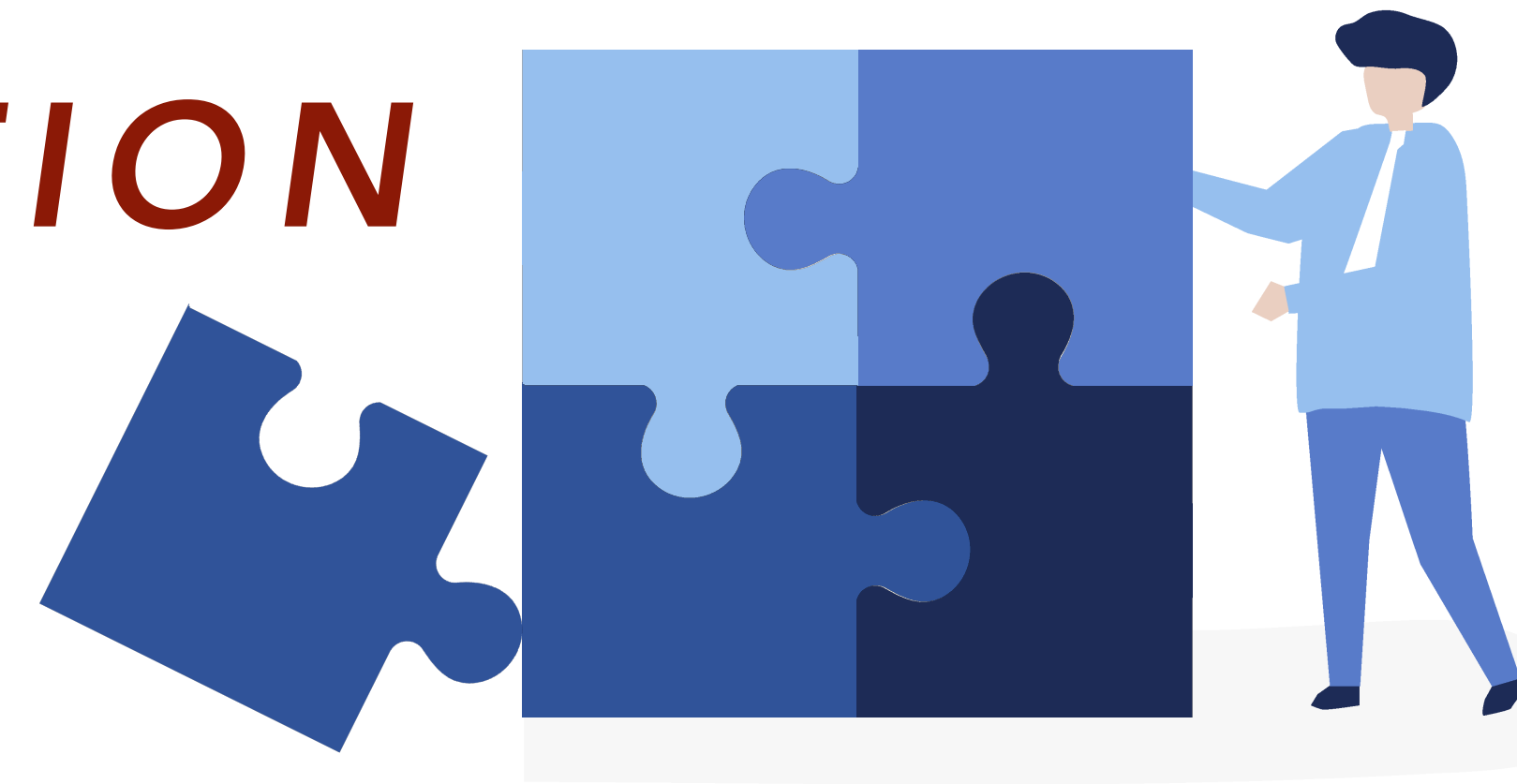1- When a data object written by a task if it is neither read by any subsequent task nor passed to outside by an event
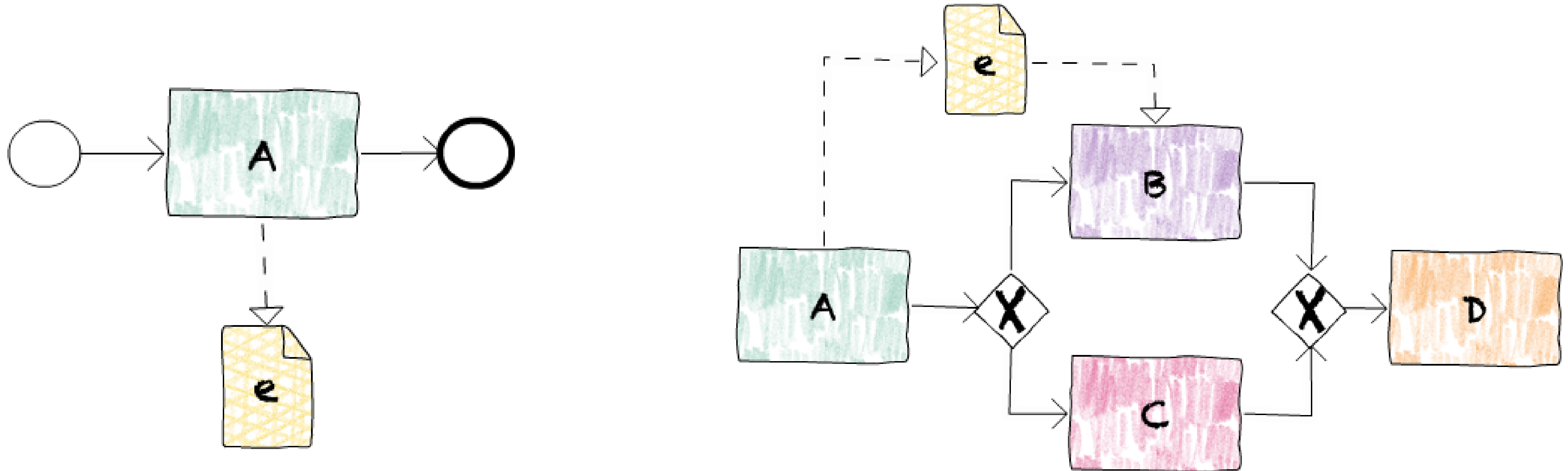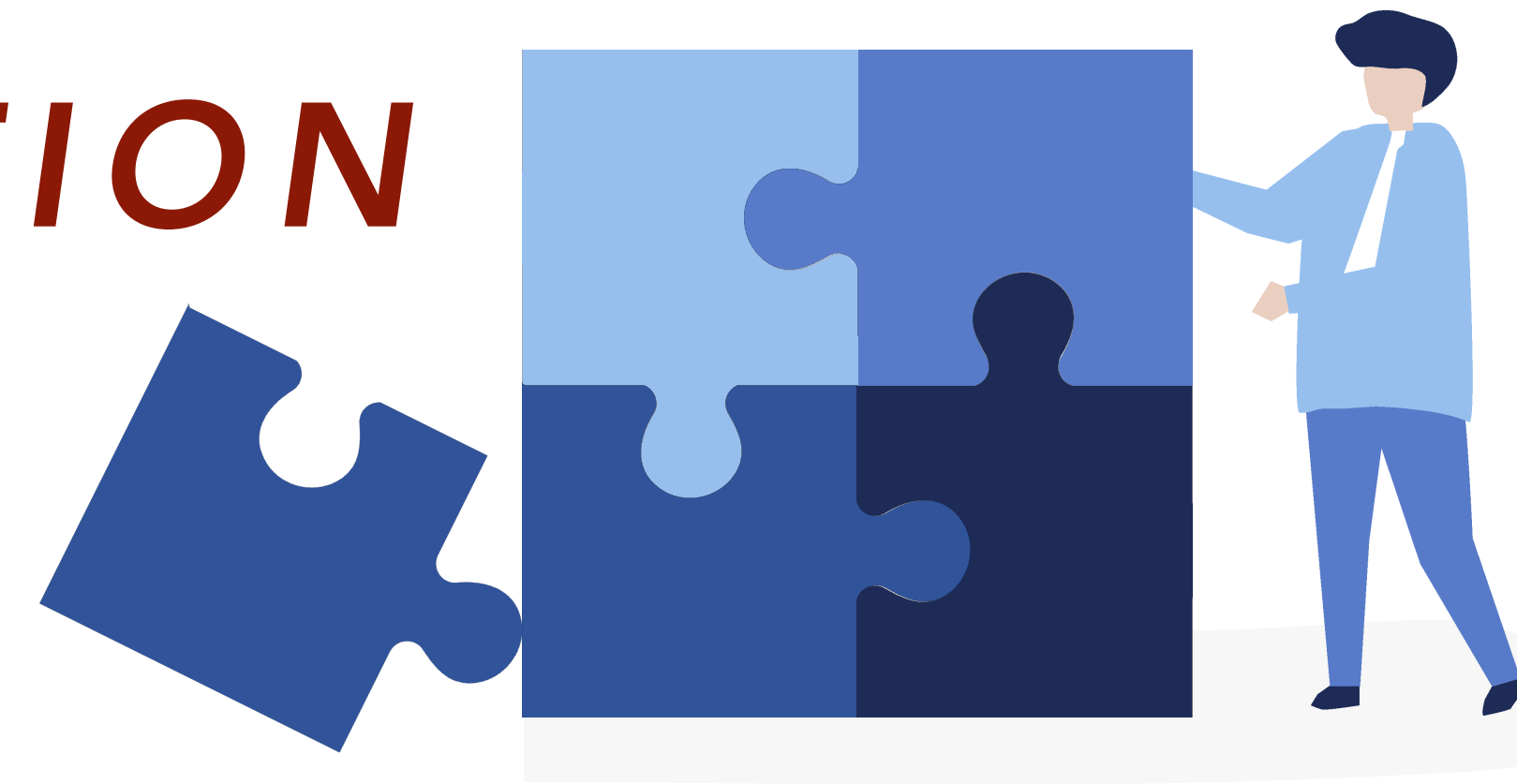
REDUNDANT DATA - *DETECTION*

1- When a data object written by a task if it is neither read by any subsequent task nor passed to outside by an event
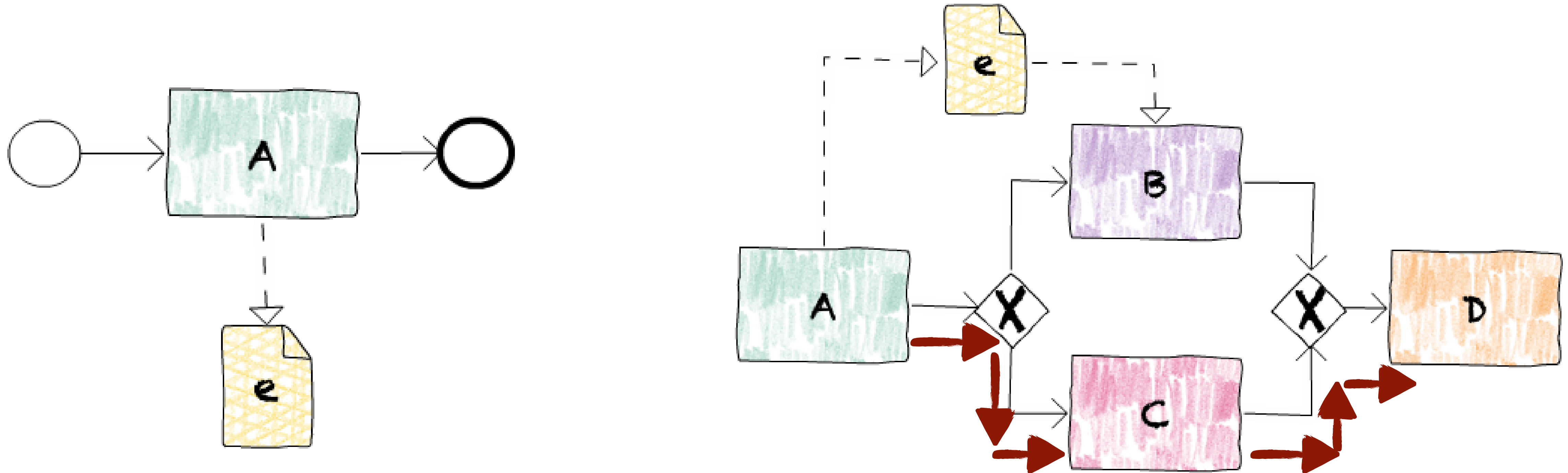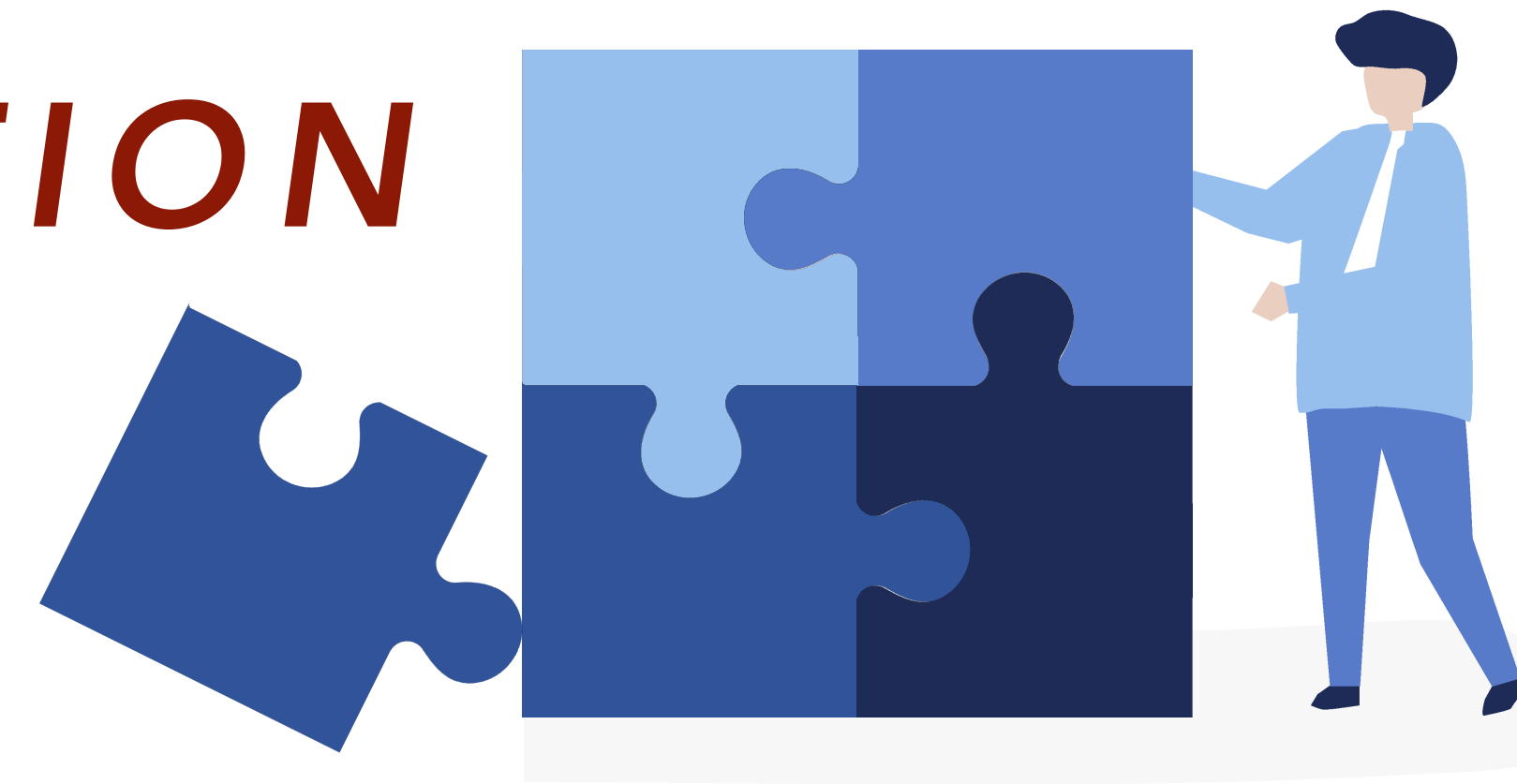
REDUNDANT DATA - *DETECTION*

1- When a data object written by a task if it is neither read by any subsequent task nor passed to outside by an event
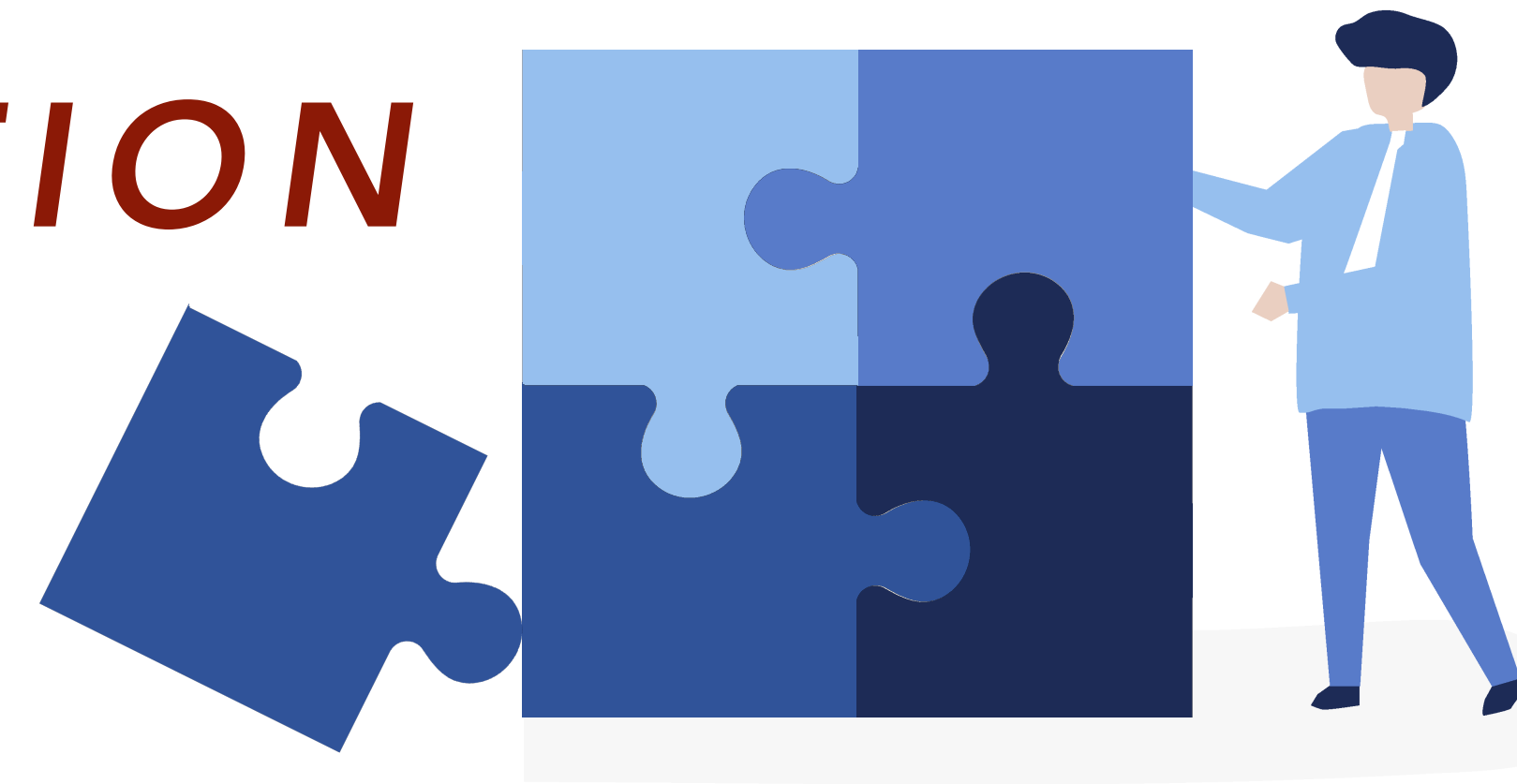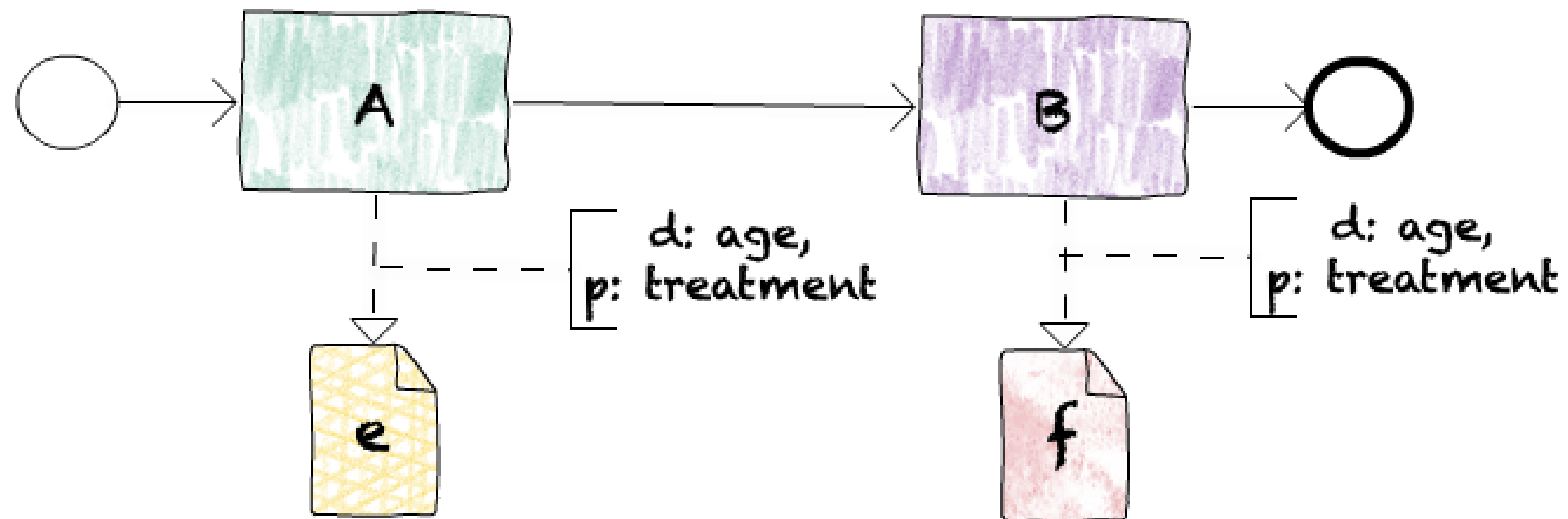
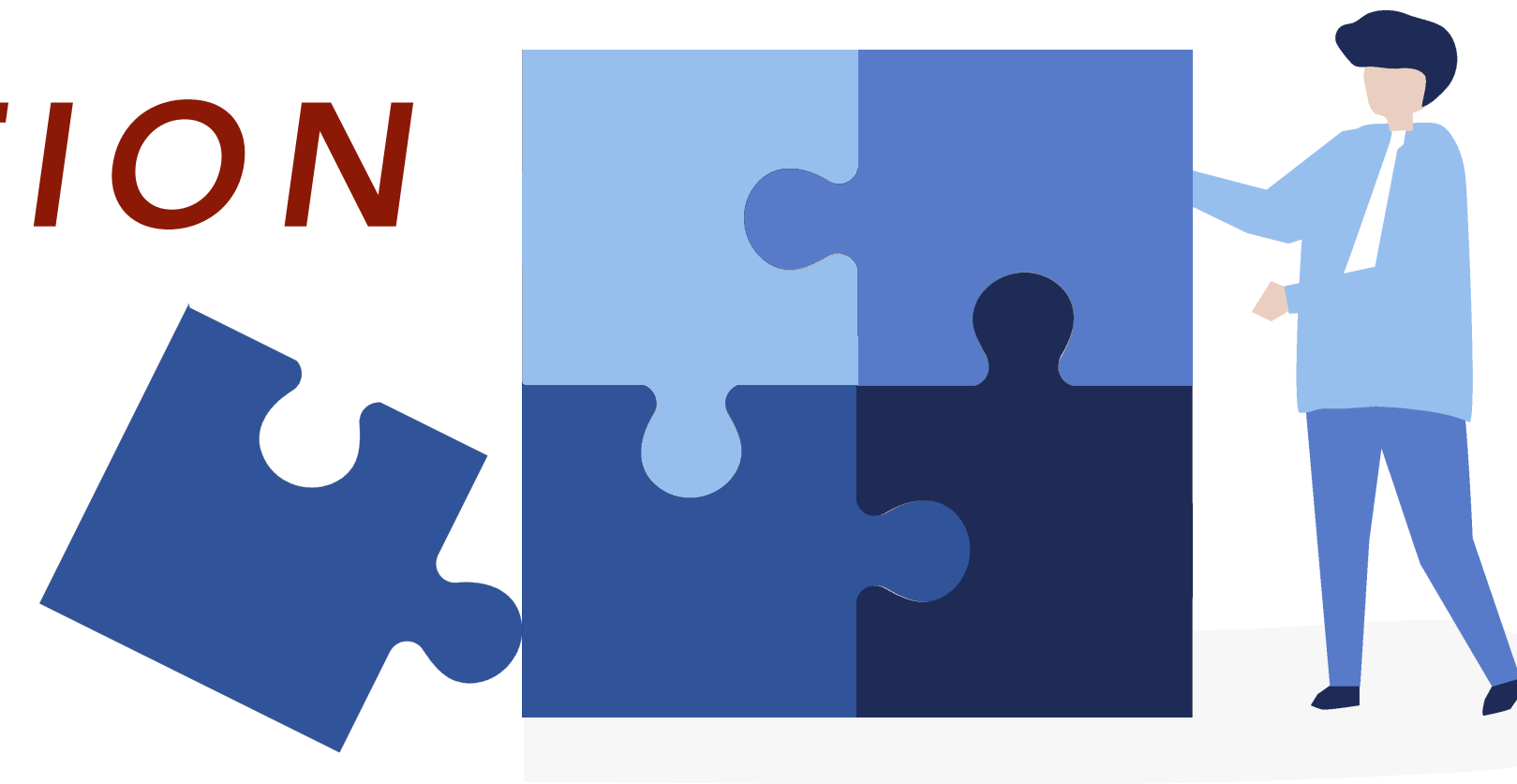*"Potential" Violator & Close World Assumption*

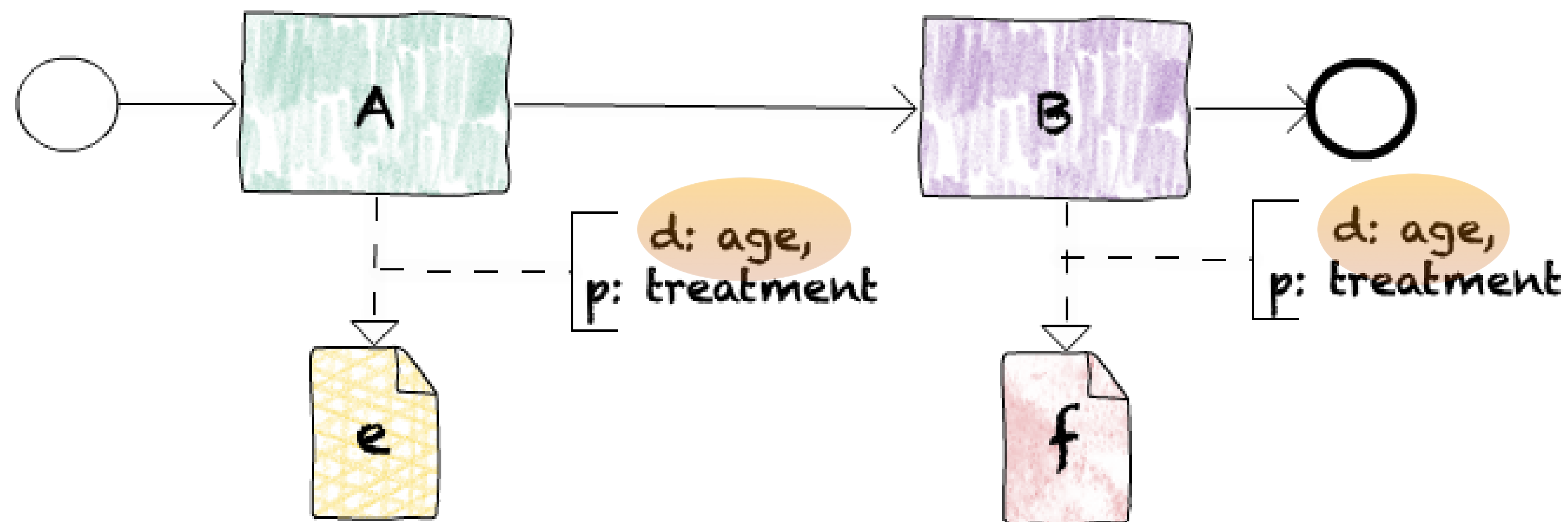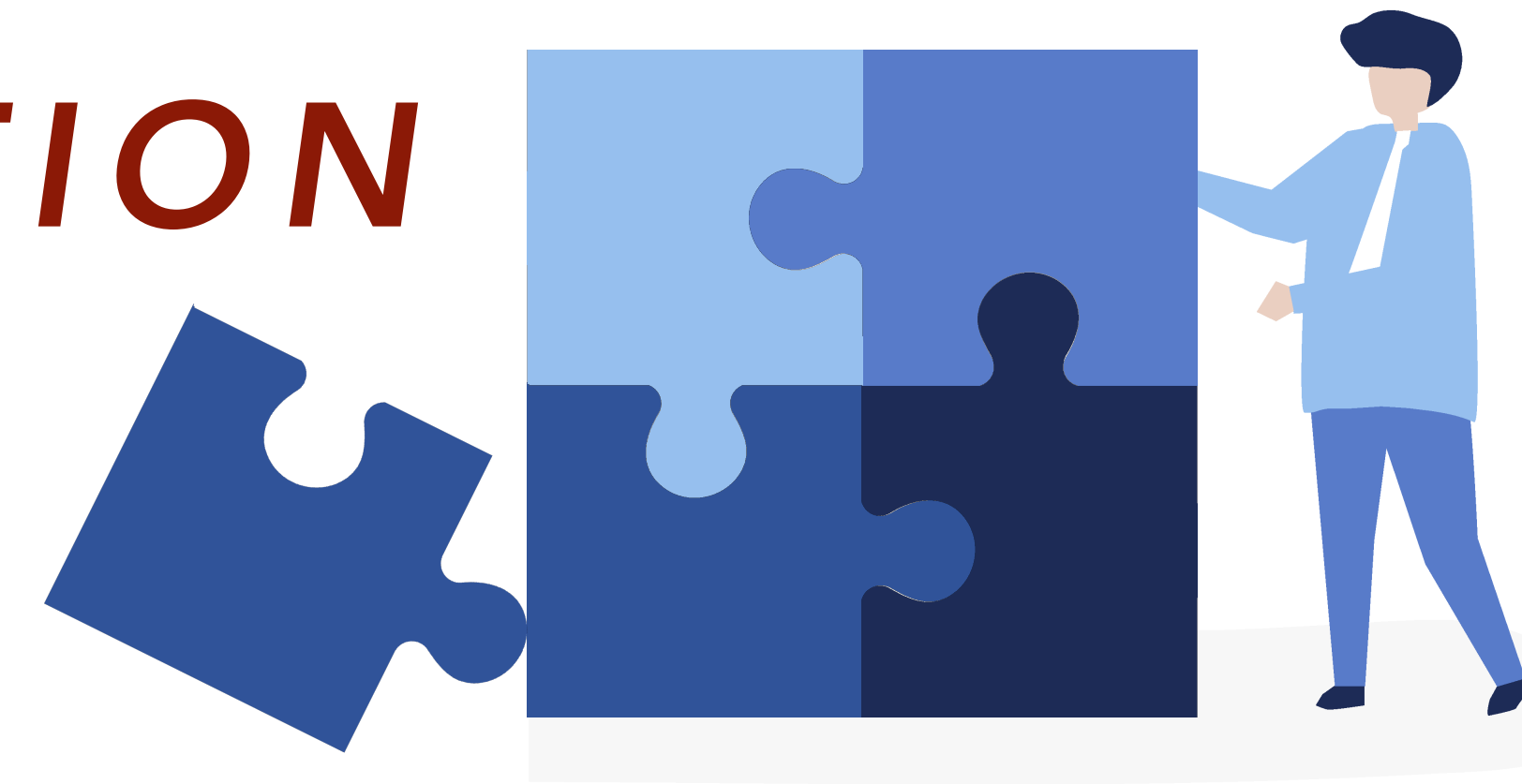2- When the same piece of data is stored in different files or in different tables within a single database

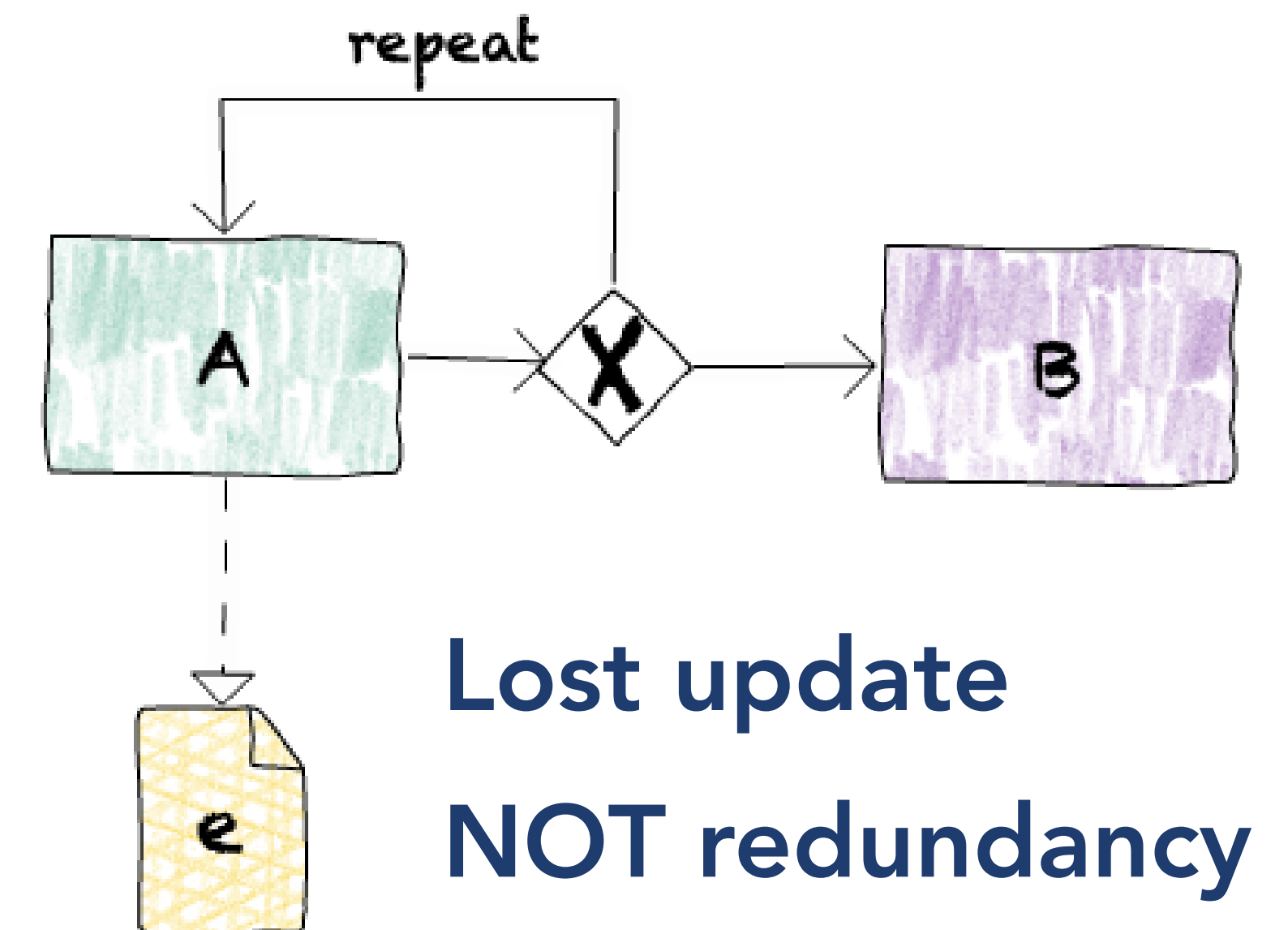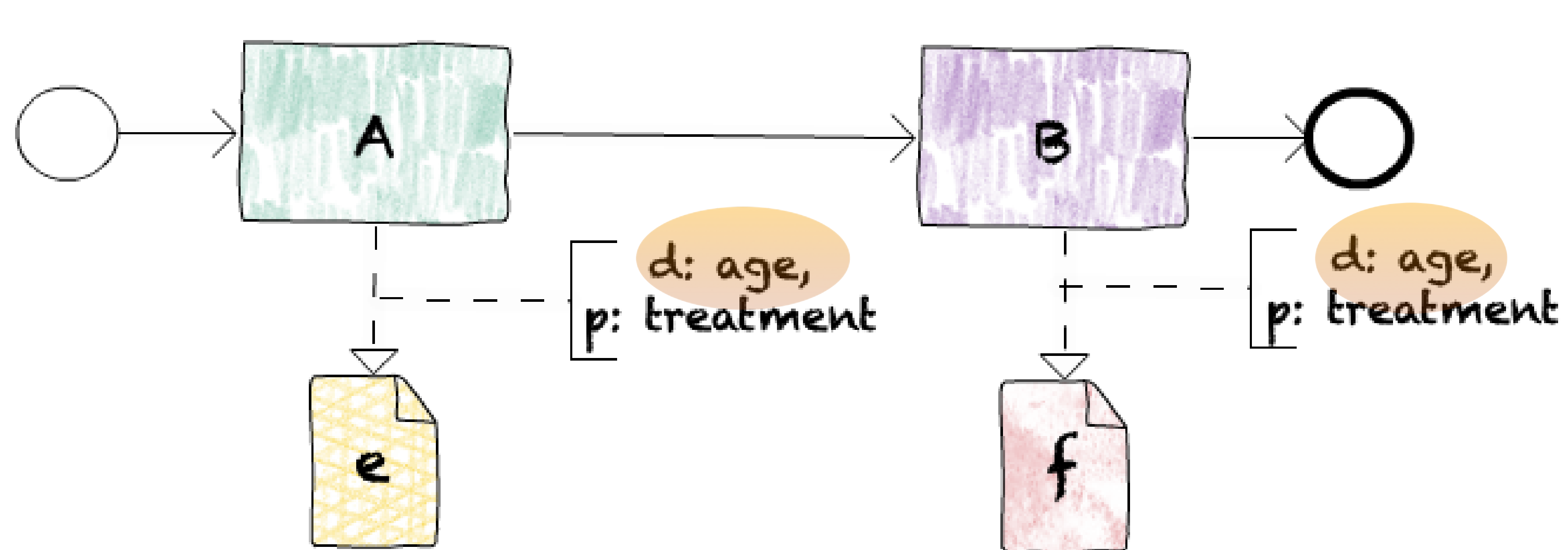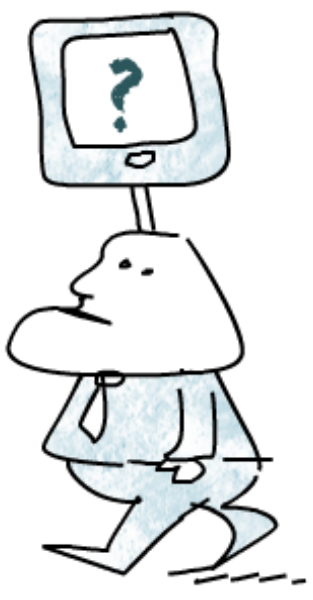2- When the same piece of data is stored in different files or in different tables within a single database
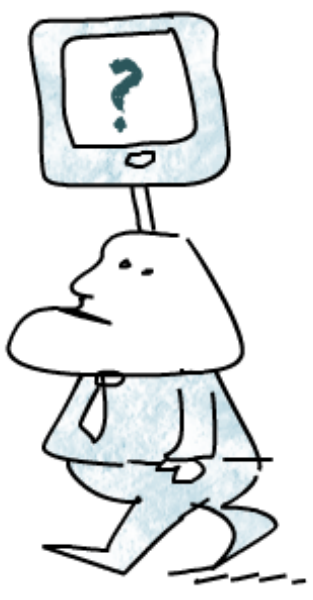
2- When the same piece of data is stored in different files or in different tables within a single database
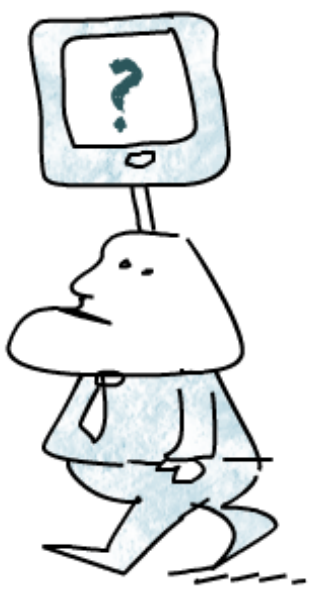

Lost update
NOT redundancy

How to **recover** WFs from data minimisation violations?

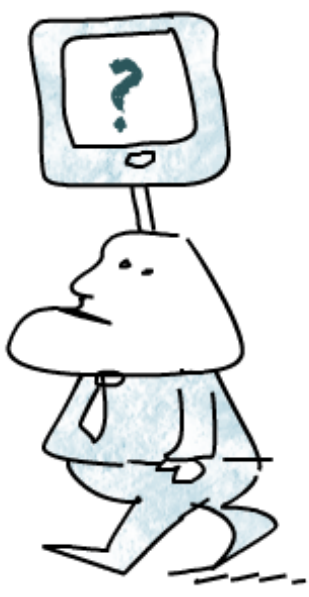1- Business Layer Recovery Strategies

How to **recover** WFs from data minimisation violations?
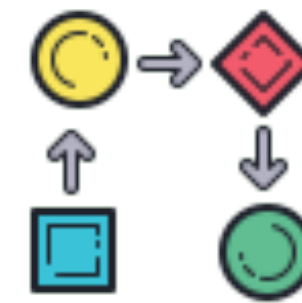
1- Business Layer Recovery Strategies → *Exception Handler: Trying Alternatives, Inserting / Cancelling Behavior…

*[4] Reichert, Manfred, and Barbara Weber. "*Enabling flexibility in process-aware information systems: challenges, methods, technologies.*"

How to **recover** WFs from data minimisation violations?

1- Business Layer Recovery Strategies → *Exception Handler: Trying Alternatives, Inserting / Cancelling Behavior…
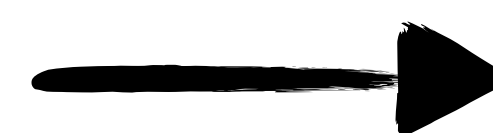
2- Data Layer Recovery Strategies

*[4] Reichert, Manfred, and Barbara Weber. "*Enabling flexibility in process-aware information systems: challenges, methods, technologies.*"
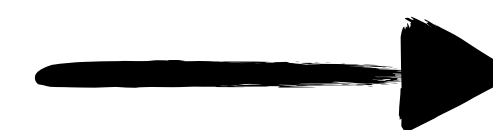
How to **recover** WFs from data minimisation violations?

1- Business Layer Recovery Strategies ➡️ *Exception Handler: Trying Alternatives, Inserting / Cancelling Behavior…
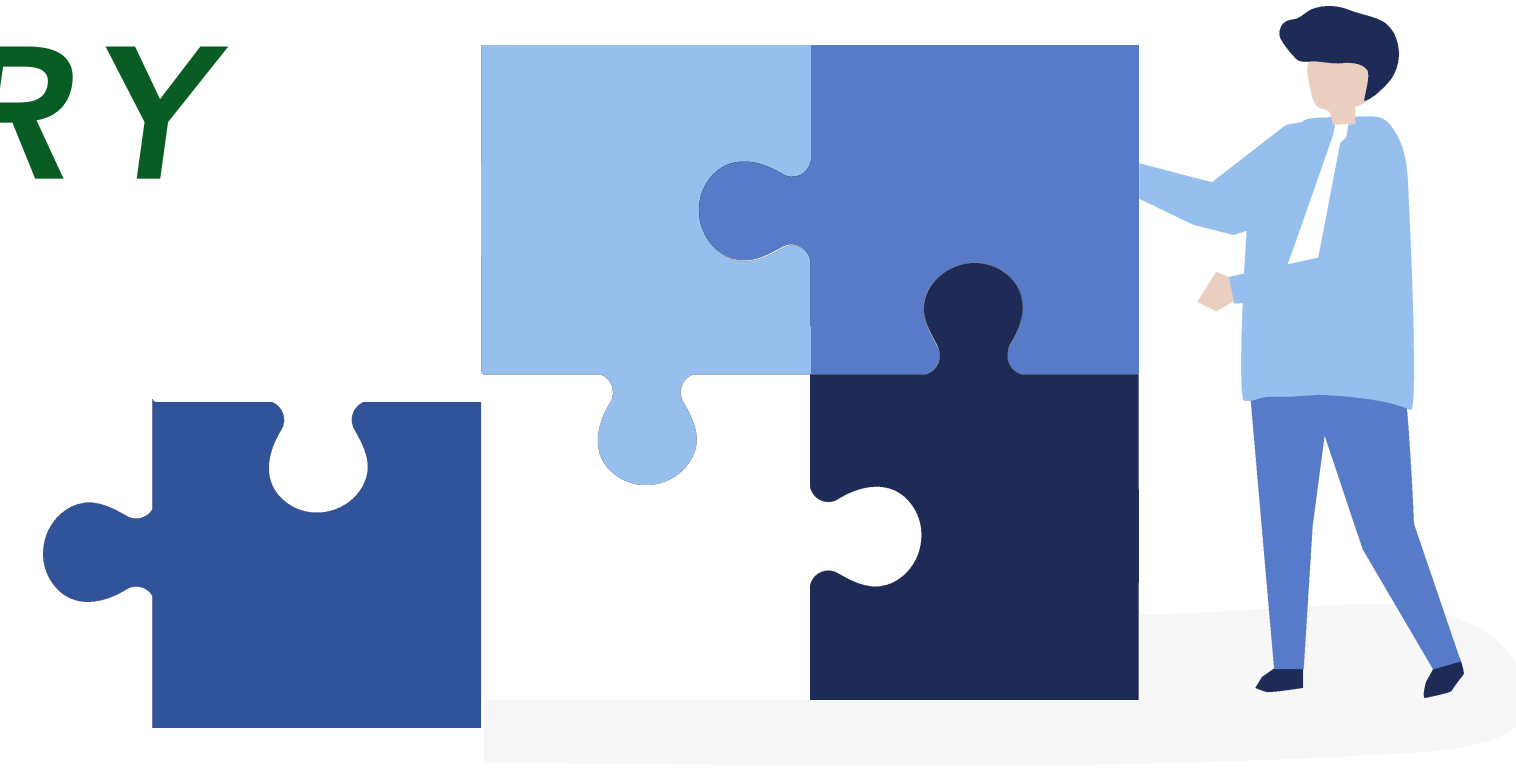
2- Data Layer Recovery Strategies ➡️ Data anonymization, undetectability, unobservability…

*[4] Reichert, Manfred, and Barbara Weber. "*Enabling flexibility in process-aware information systems: challenges, methods, technologies.*"
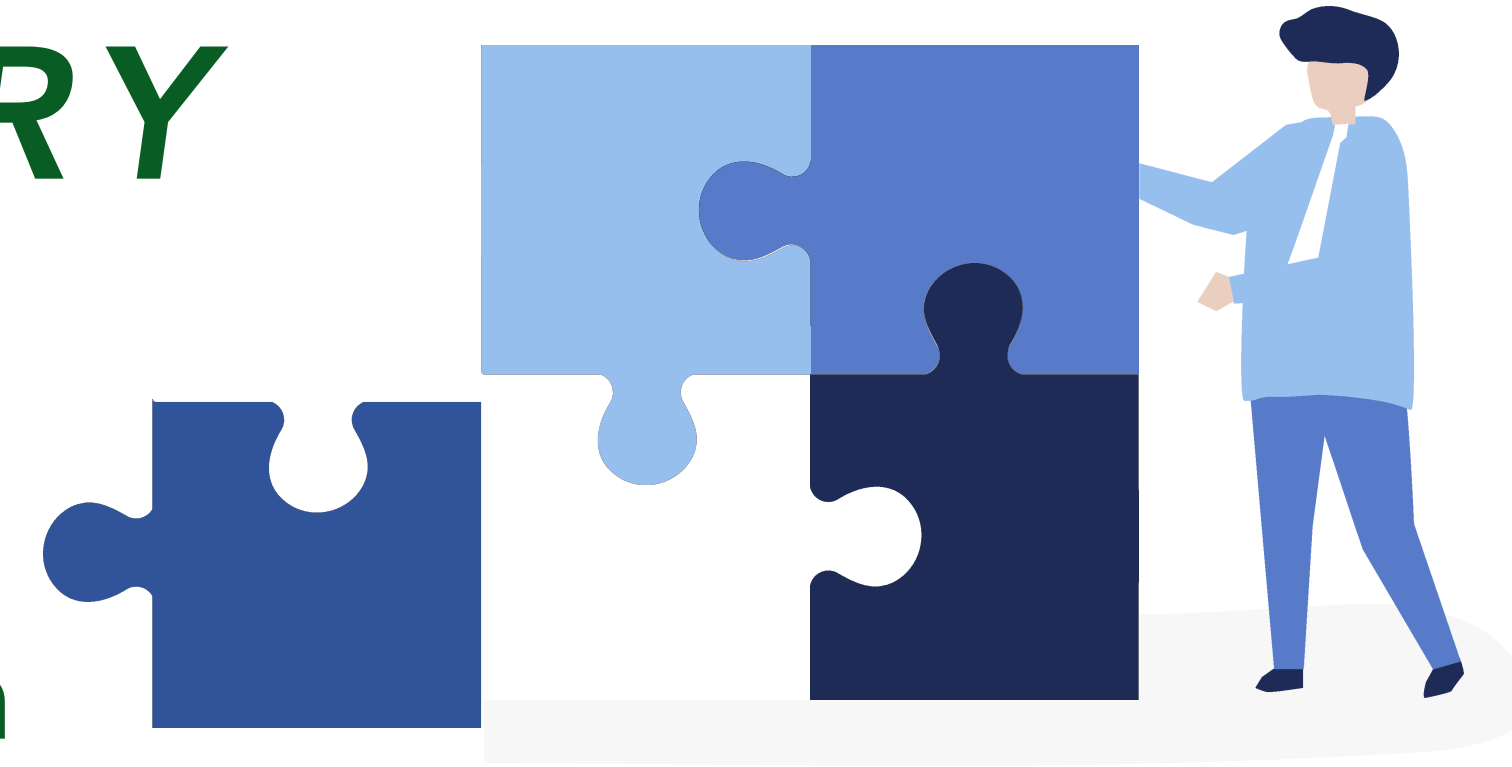
IRRELEVANT DATA - *RECOVERY*

1- If there is no data dependency

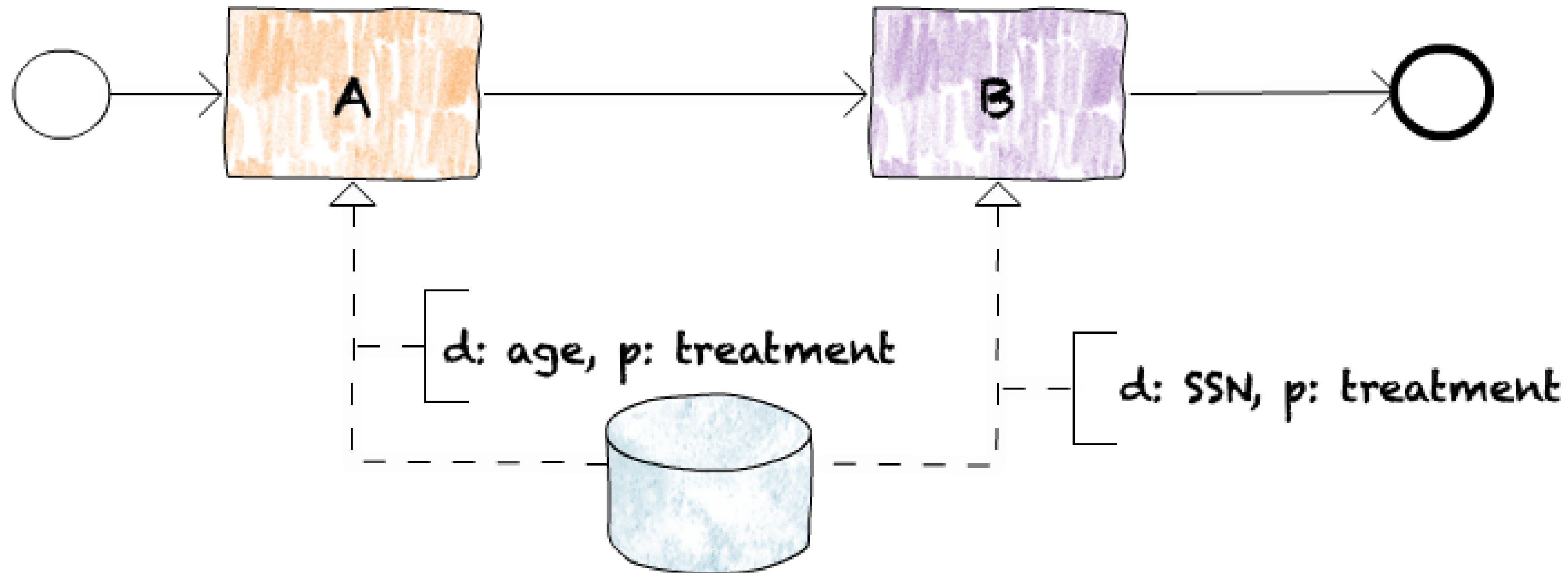# IRRELEVANT DATA - *RECOVERY*
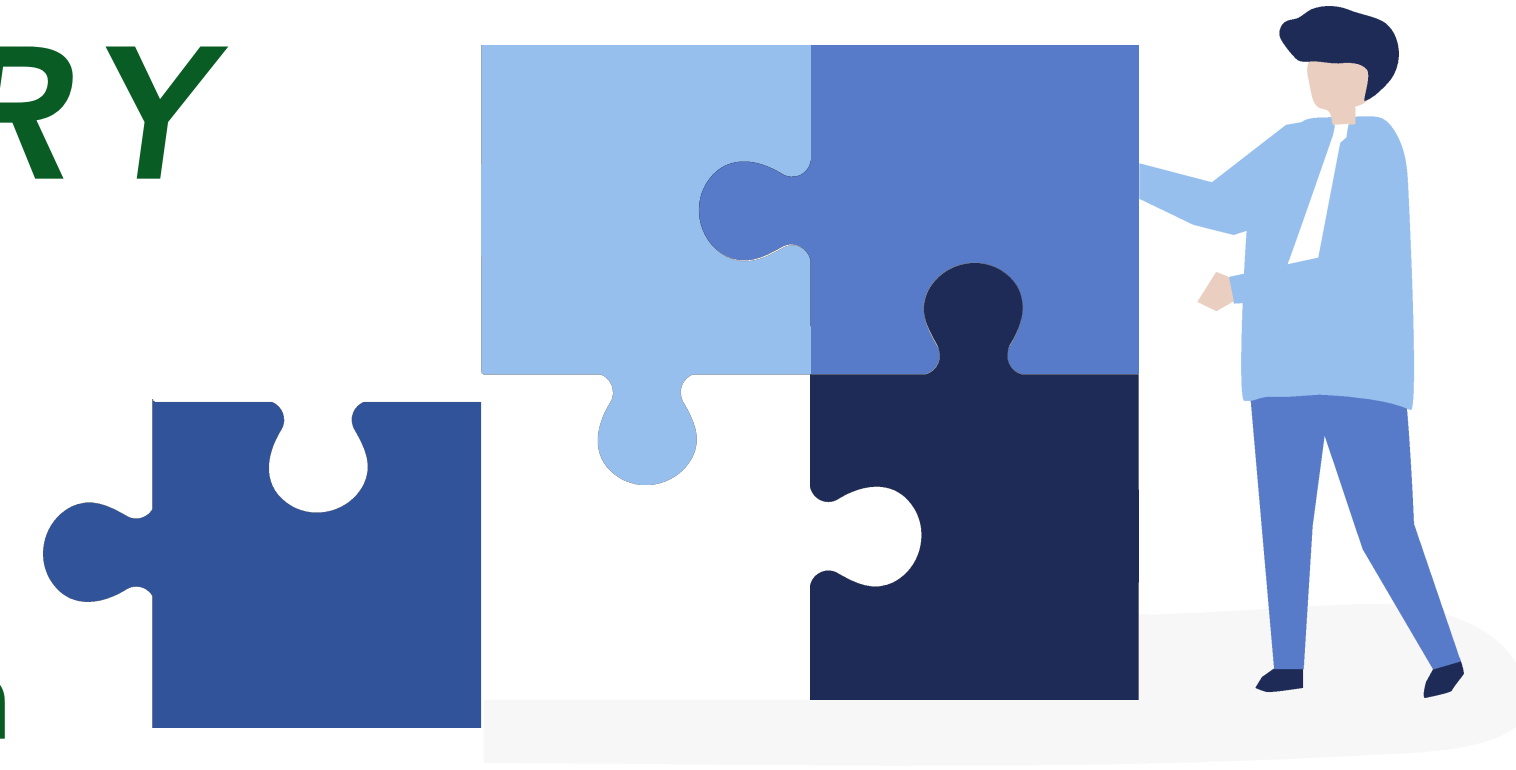
1- If there is no data dependency

Strategy: canceling behaviour by deleting irrelevant data

IRRELEVANT DATA - *RECOVERY*

1- If there is no data dependency

Strategy: canceling behaviour by deleting irrelevant data

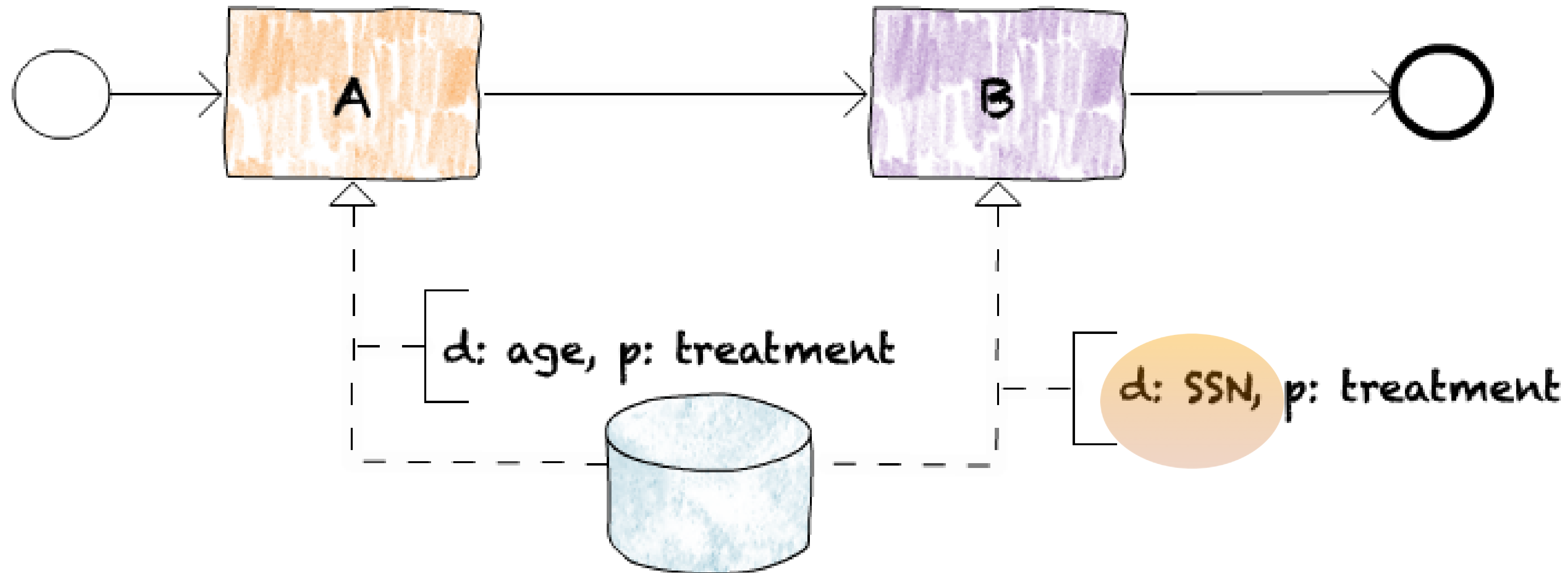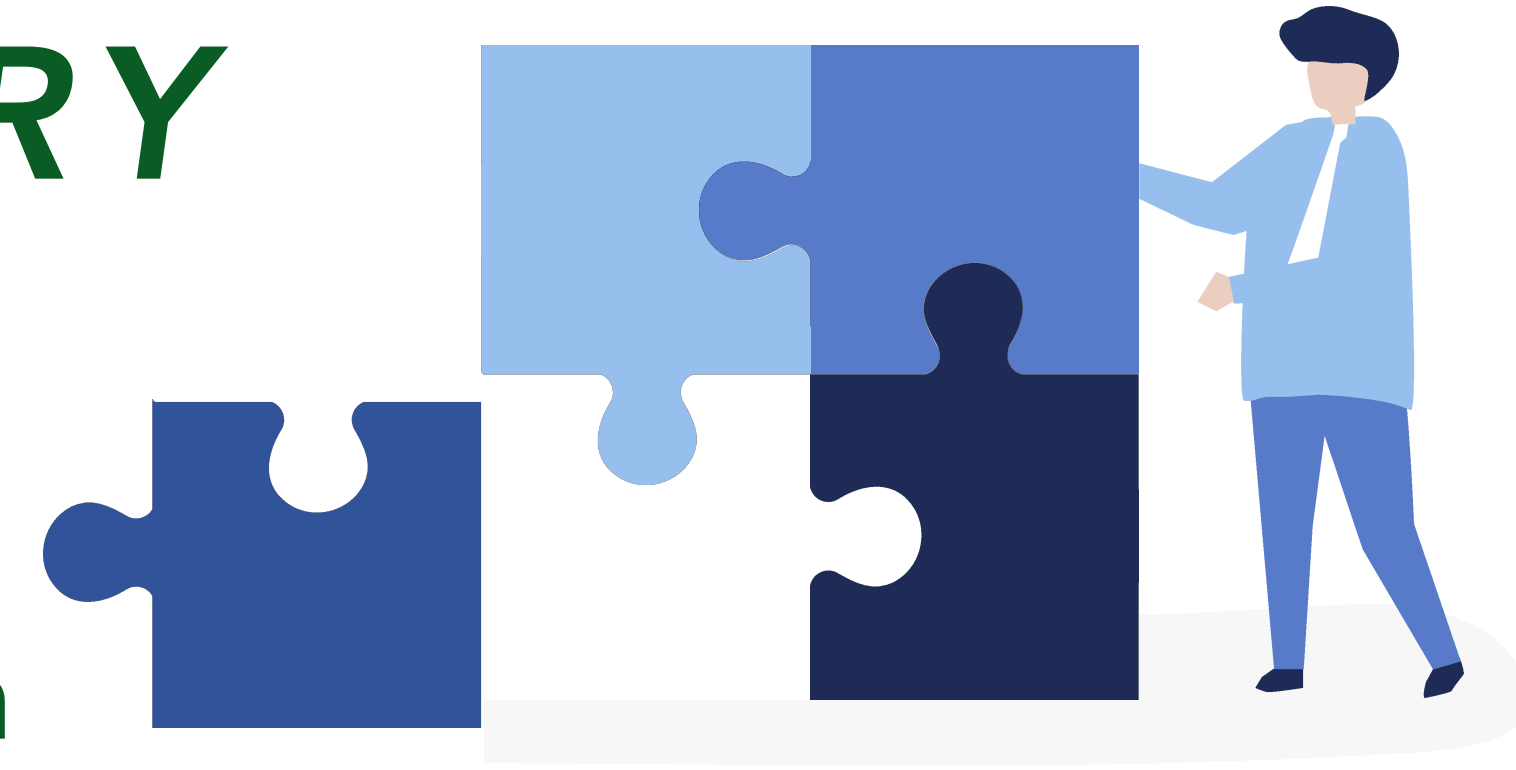d: age, p: treatment

d: SSN, p: treatment

Policy: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*

1- If there is no data dependency

Strategy: canceling behaviour by deleting irrelevant data

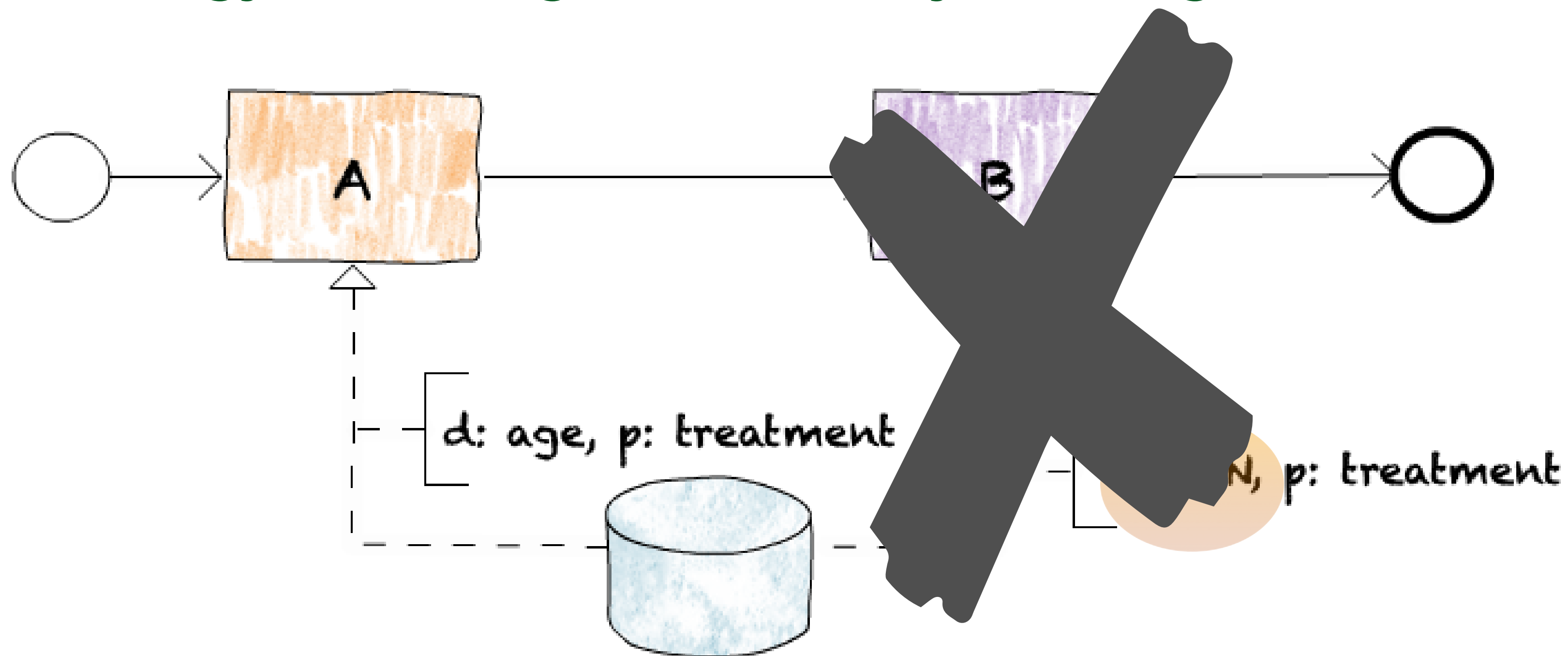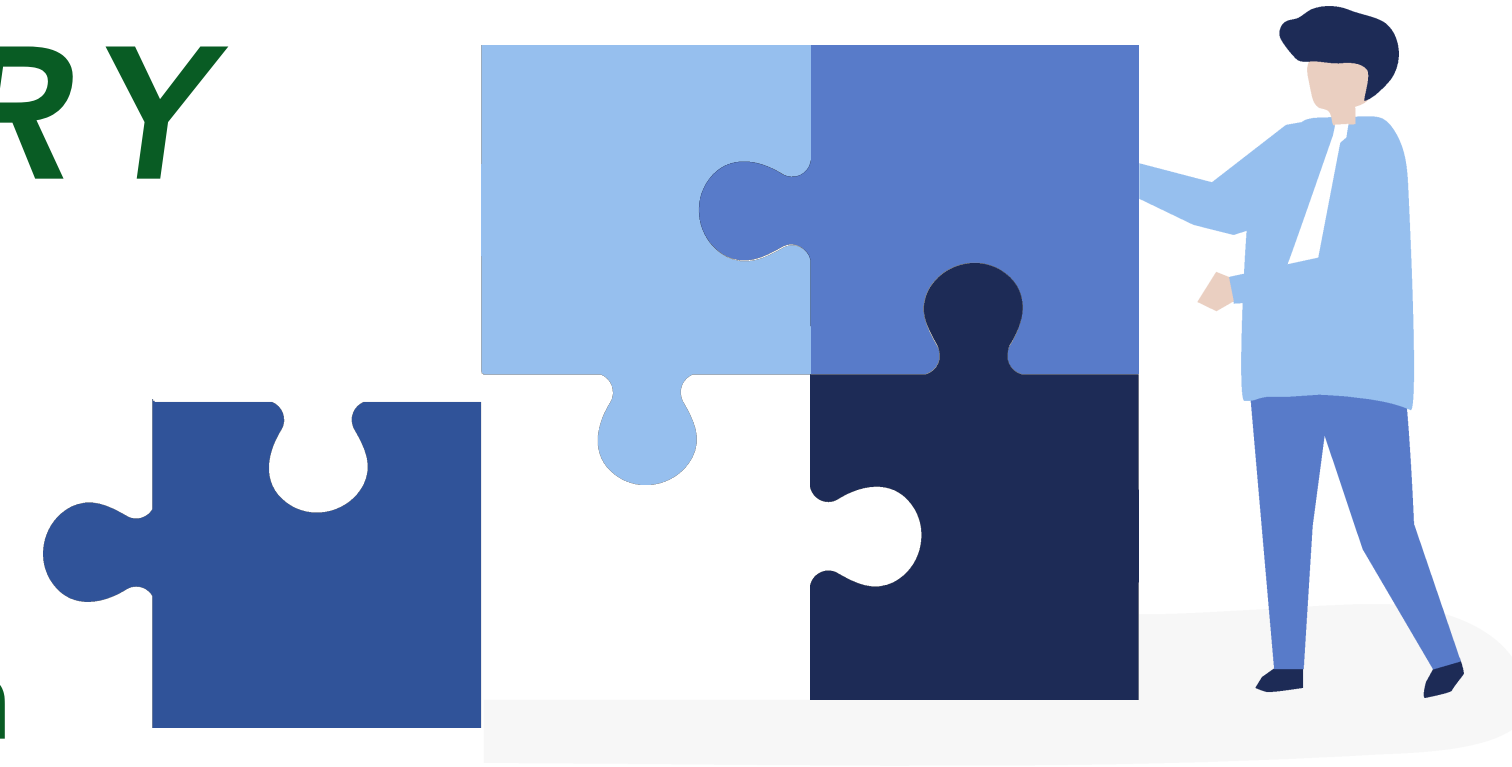d: age, p: treatment

d: SSN, p: treatment

Policy: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*

1- If there is no data dependency

Strategy: canceling behaviour by deleting irrelevant data

Policy: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*

1- If there is no data dependency

Strategy: canceling behaviour by deleting irrelevant data
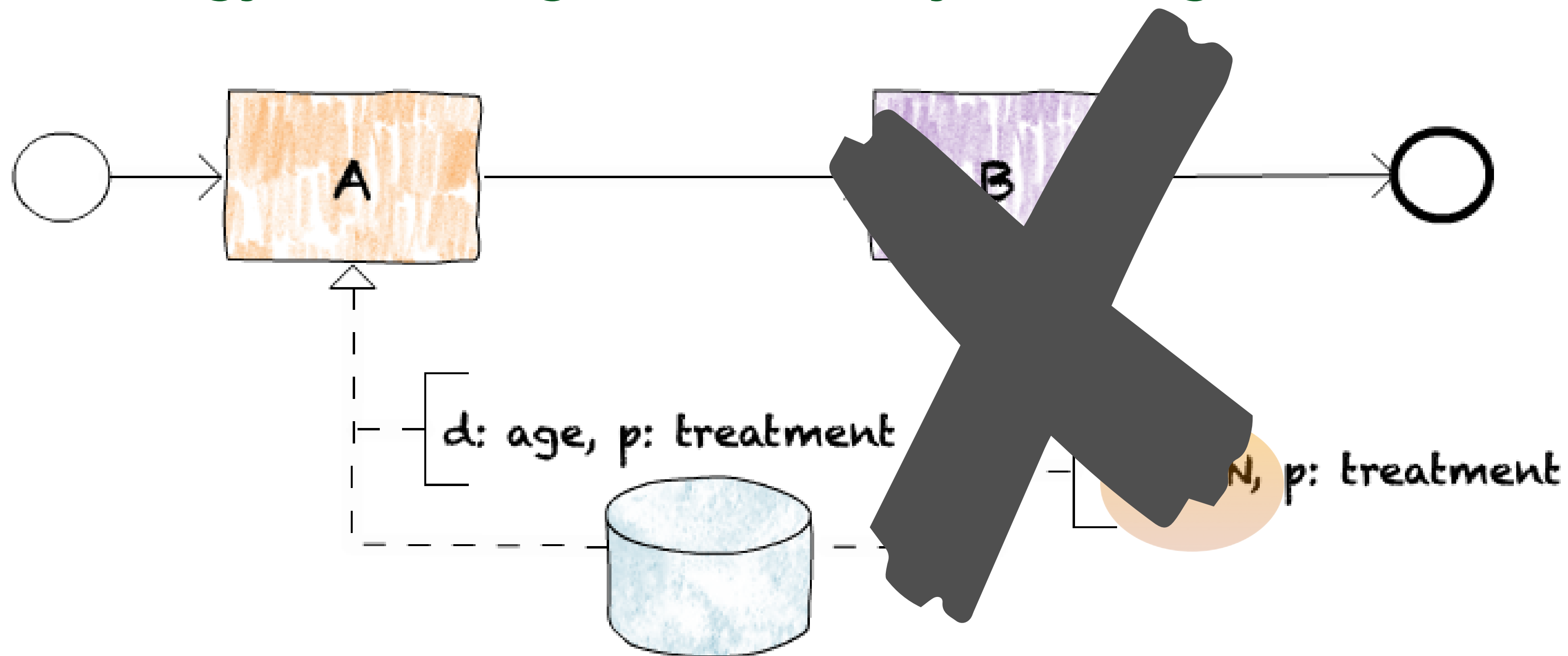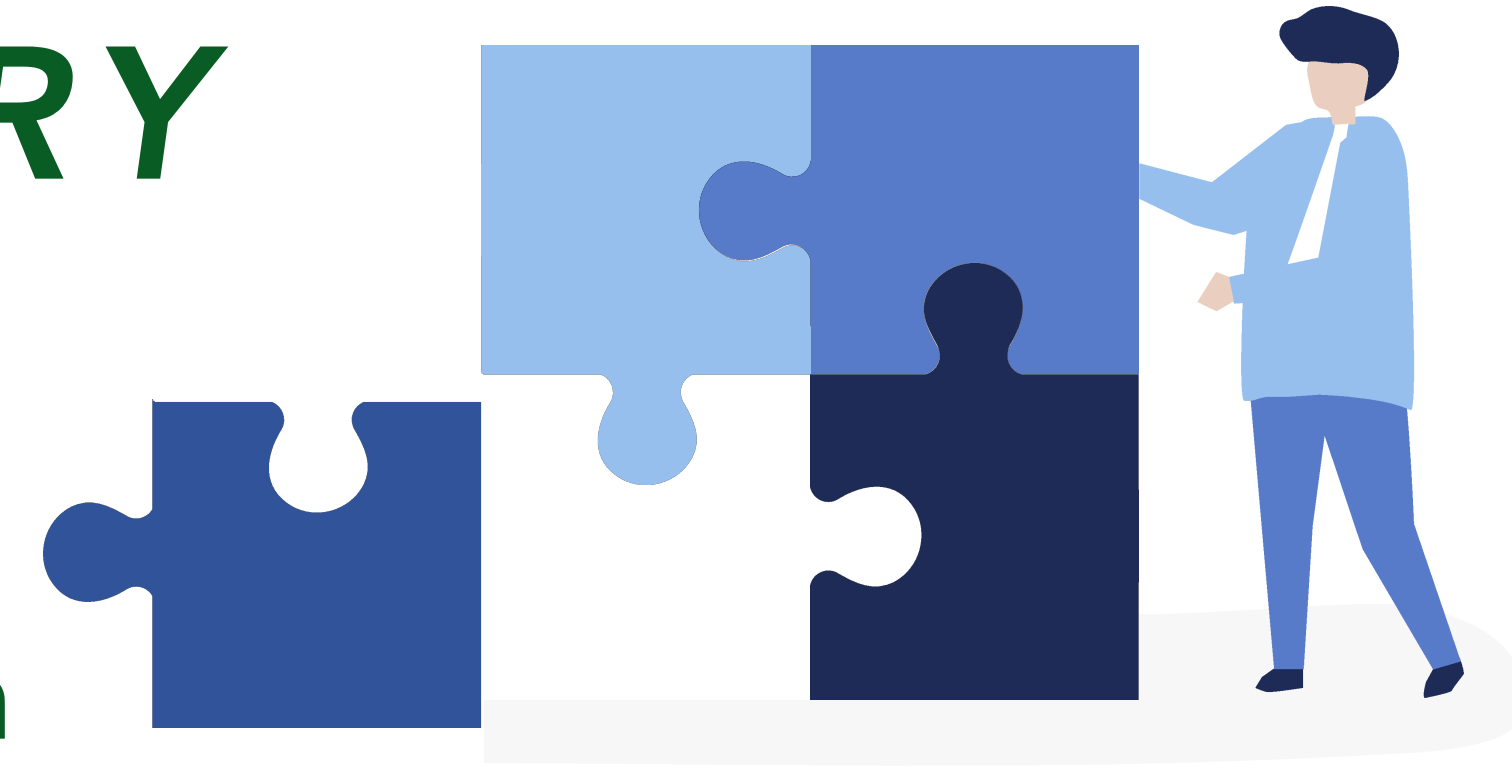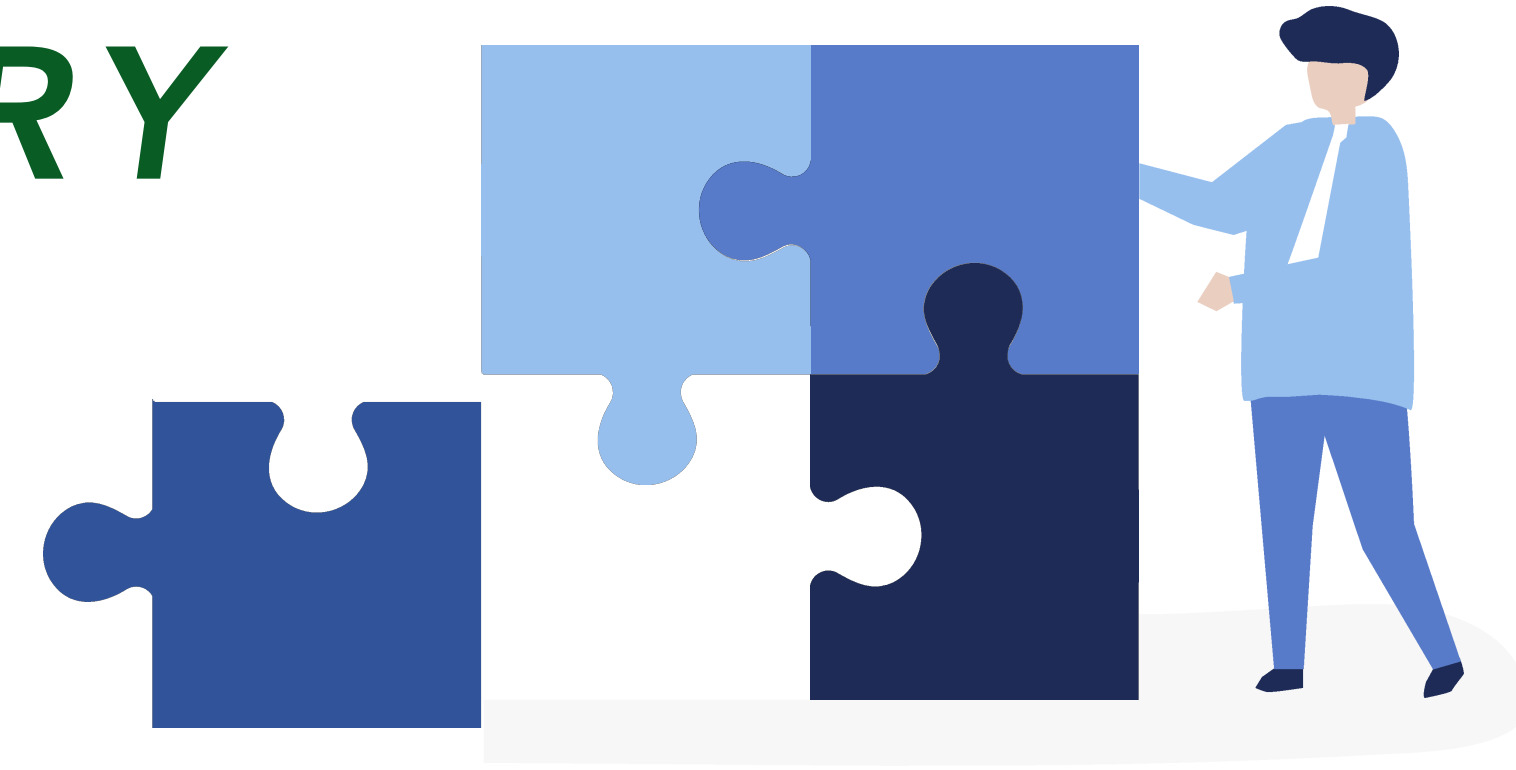

Policy: treatment requires name, age and blood-type

Problem: deleting the data / task might violate the temporal dependency,

also result in information loss
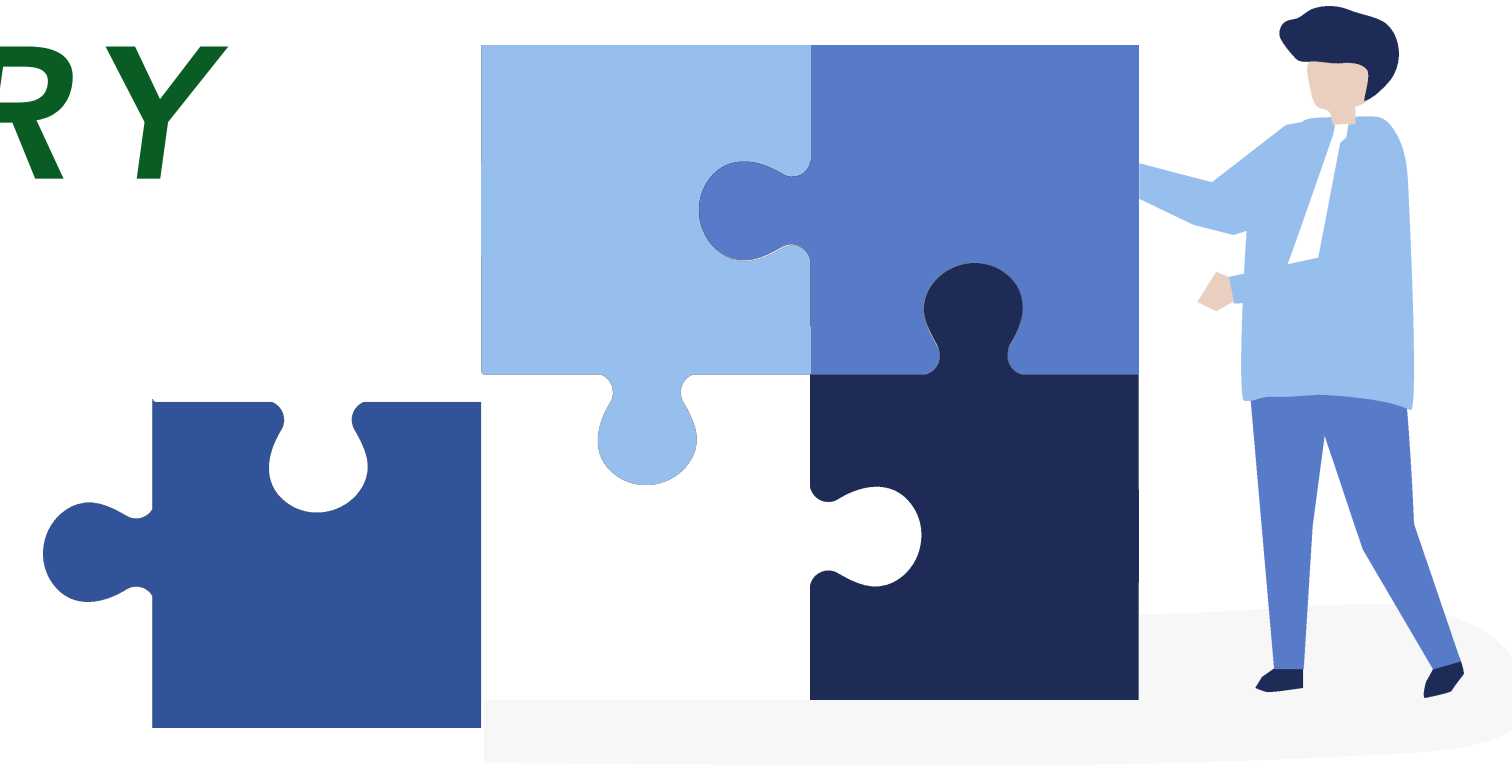
IRRELEVANT DATA - *RECOVERY*

2- If it is writing operation & no data dependency

IRRELEVANT DATA - *RECOVERY*
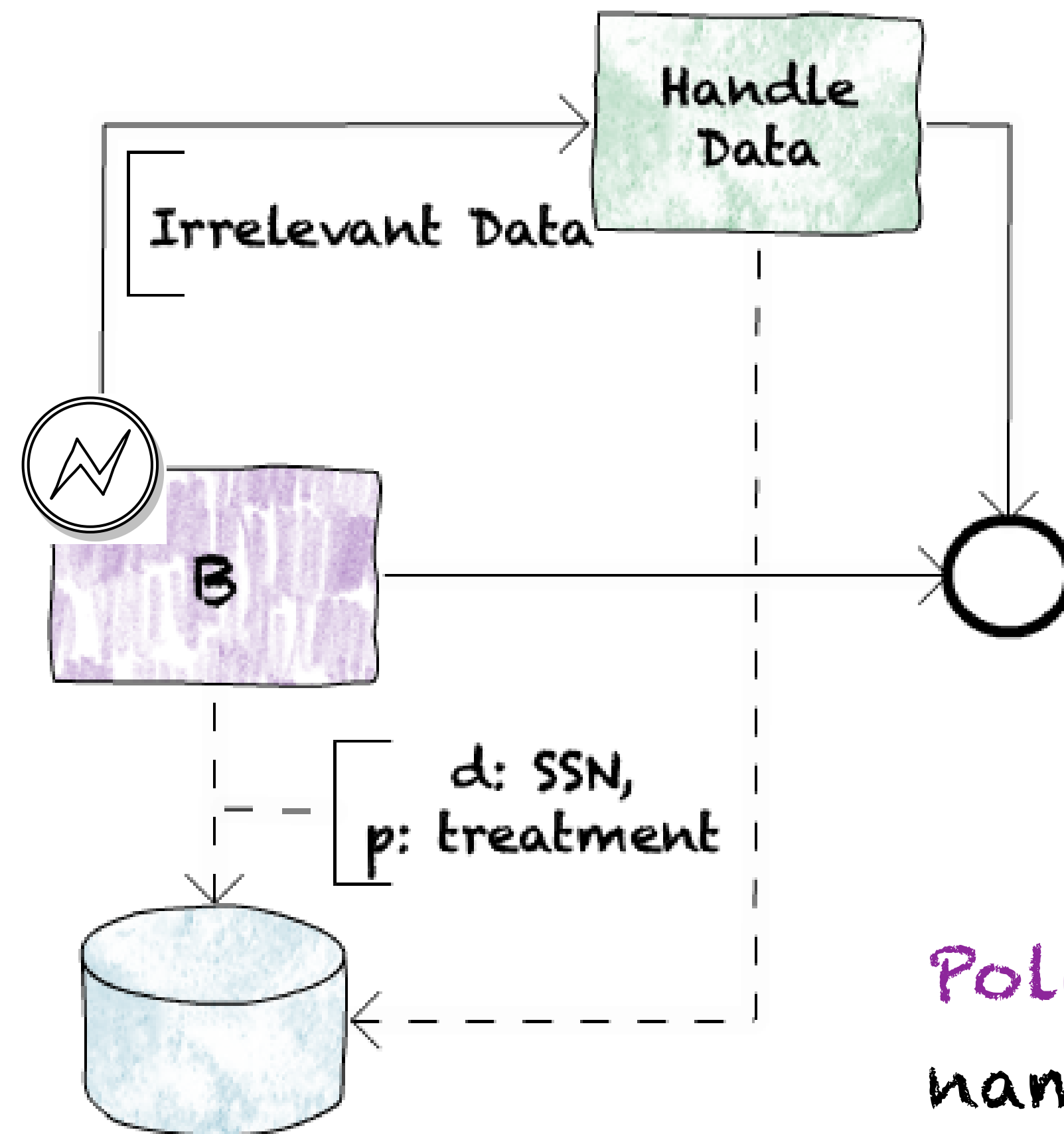
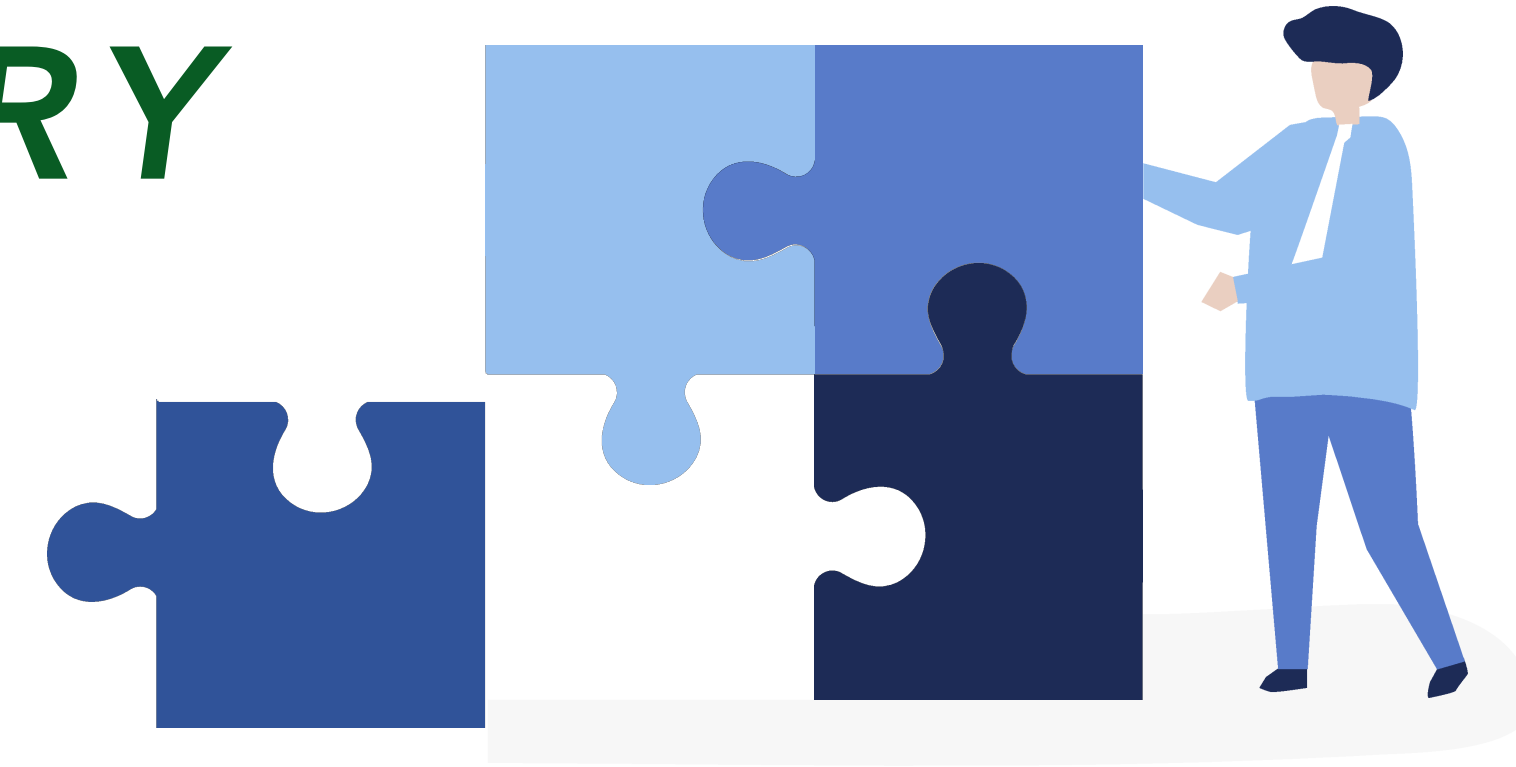2- If it is writing operation & no data dependency

Strategy: Inserting behaviour - Handle Data

IRRELEVANT DATA - *RECOVERY*

2- If it is writing operation & no data dependency

Strategy: Inserting behaviour - Handle Data

IRRELEVANT DATA - *RECOVERY*

2- If it is writing operation & no data dependency

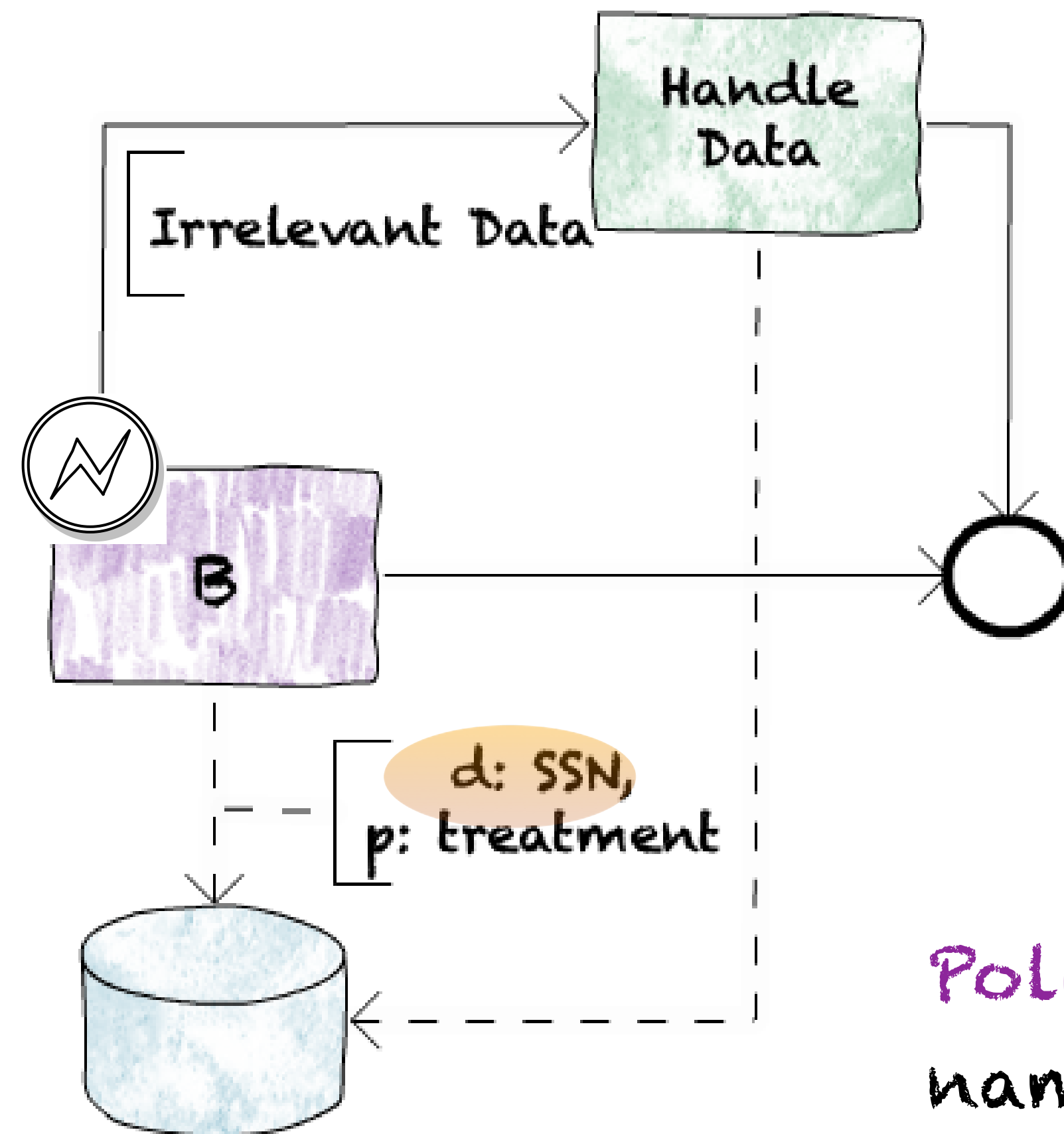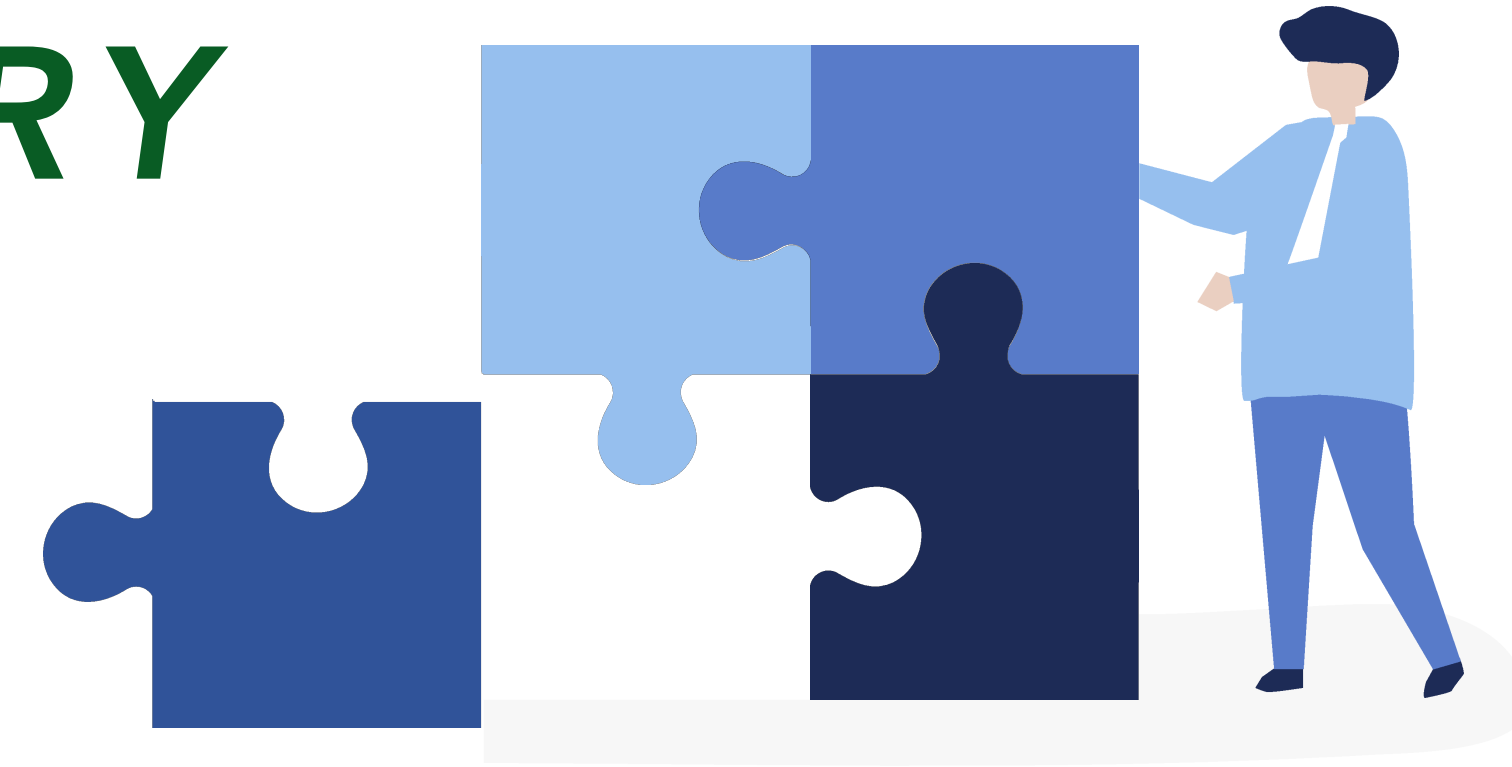**Strategy: Inserting behaviour - Handle Data**

Policy: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*

2- If it is writing operation & no data dependency

Strategy: Inserting behaviour - Handle Data

Handle Data: Make the irrelevant data anonymized, unobservable etc.

Irrelevant Data

Handle Data

B

d: SSN,
p: treatment

Policy: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*

2- If it is writing operation & no data dependency

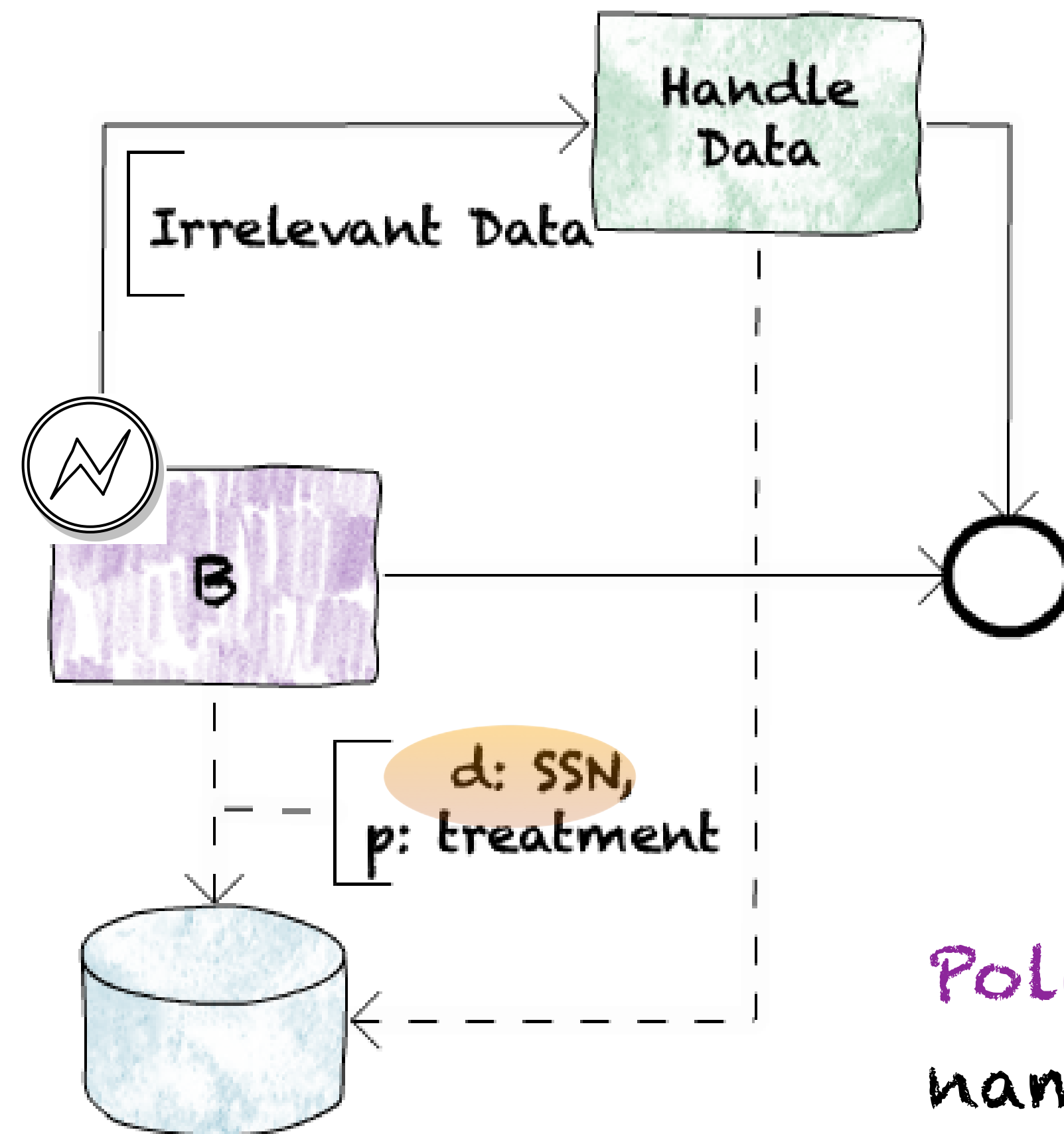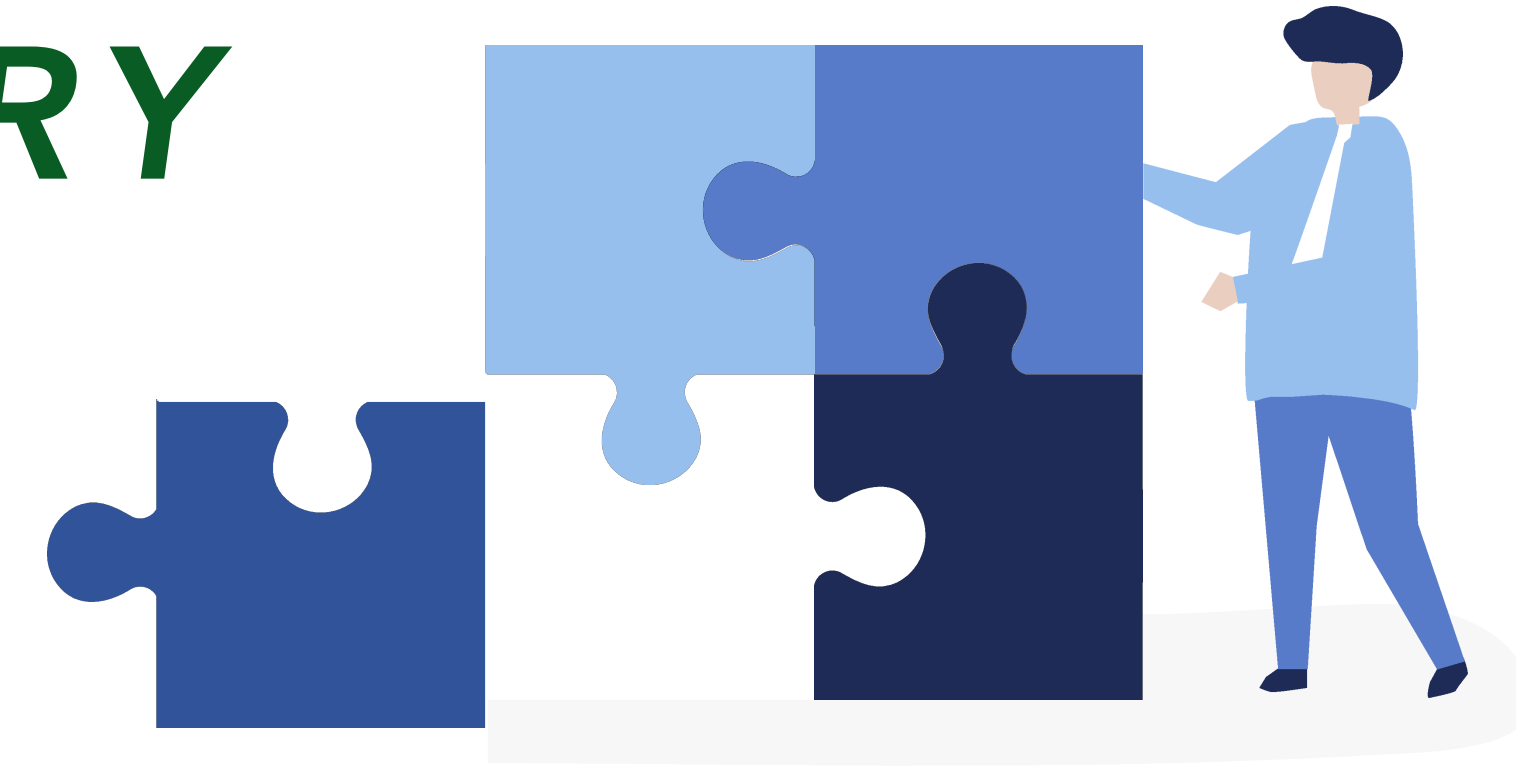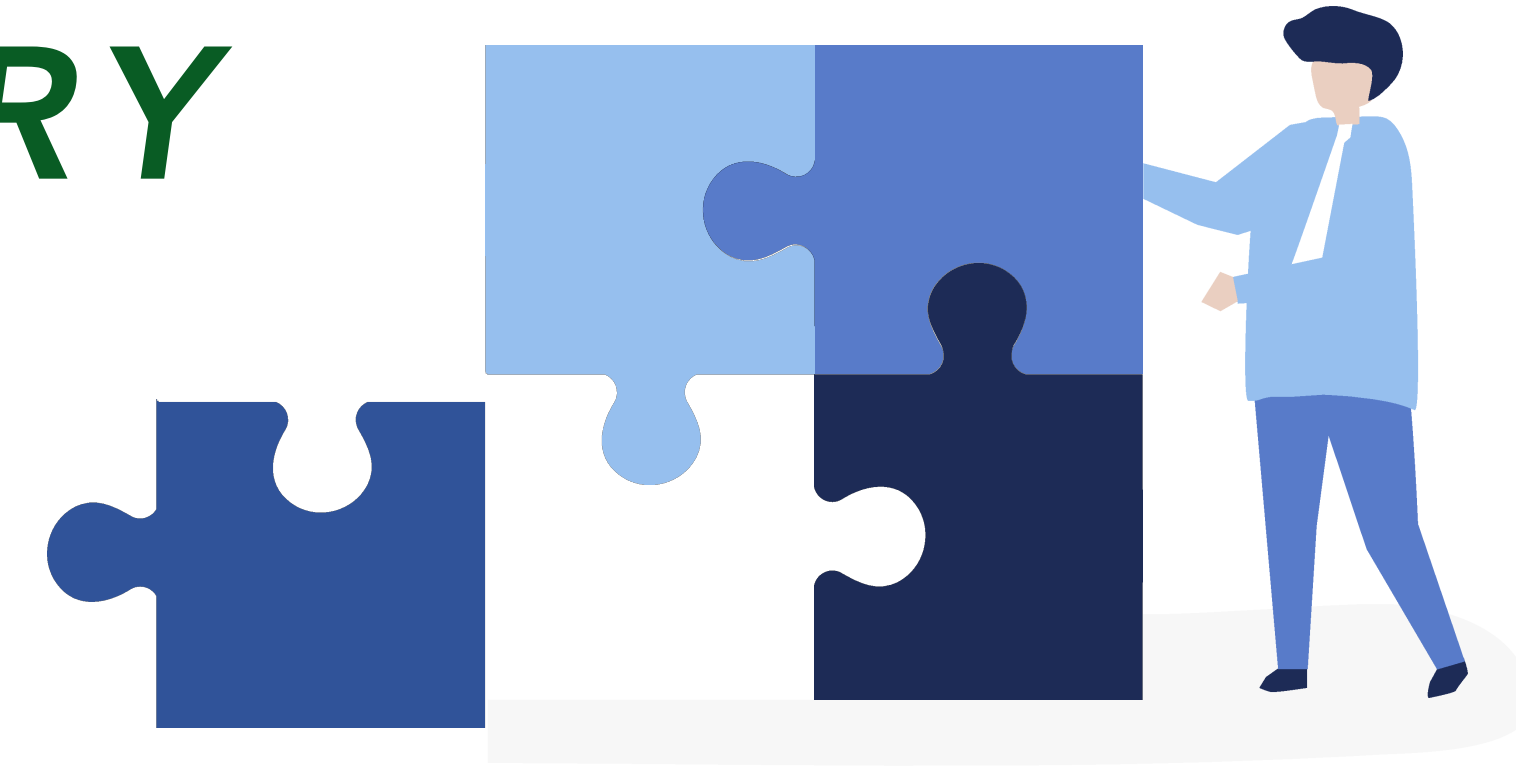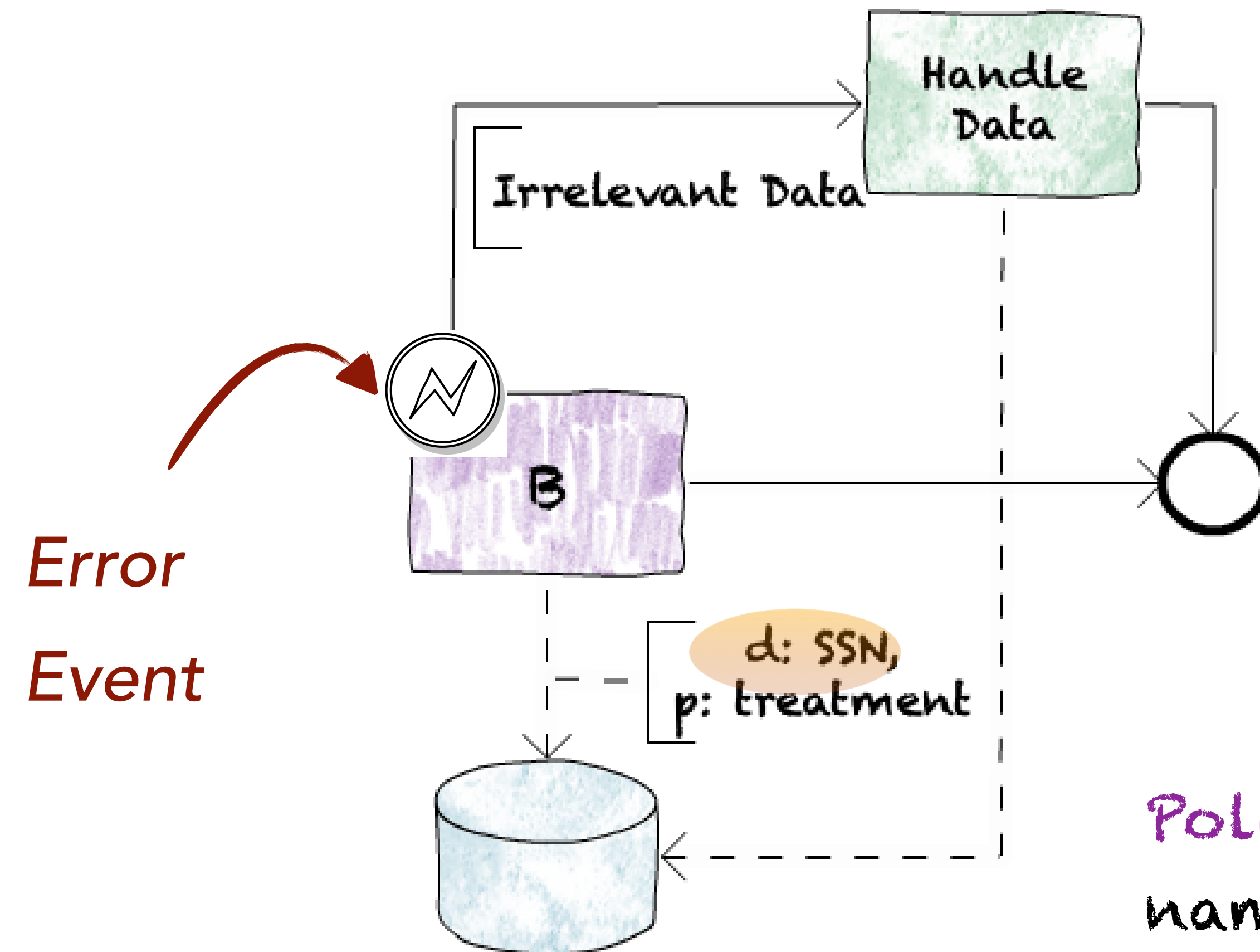**Strategy: Inserting behaviour - Handle Data**

Handle Data: Make the irrelevant data anonymized, unobservable etc.
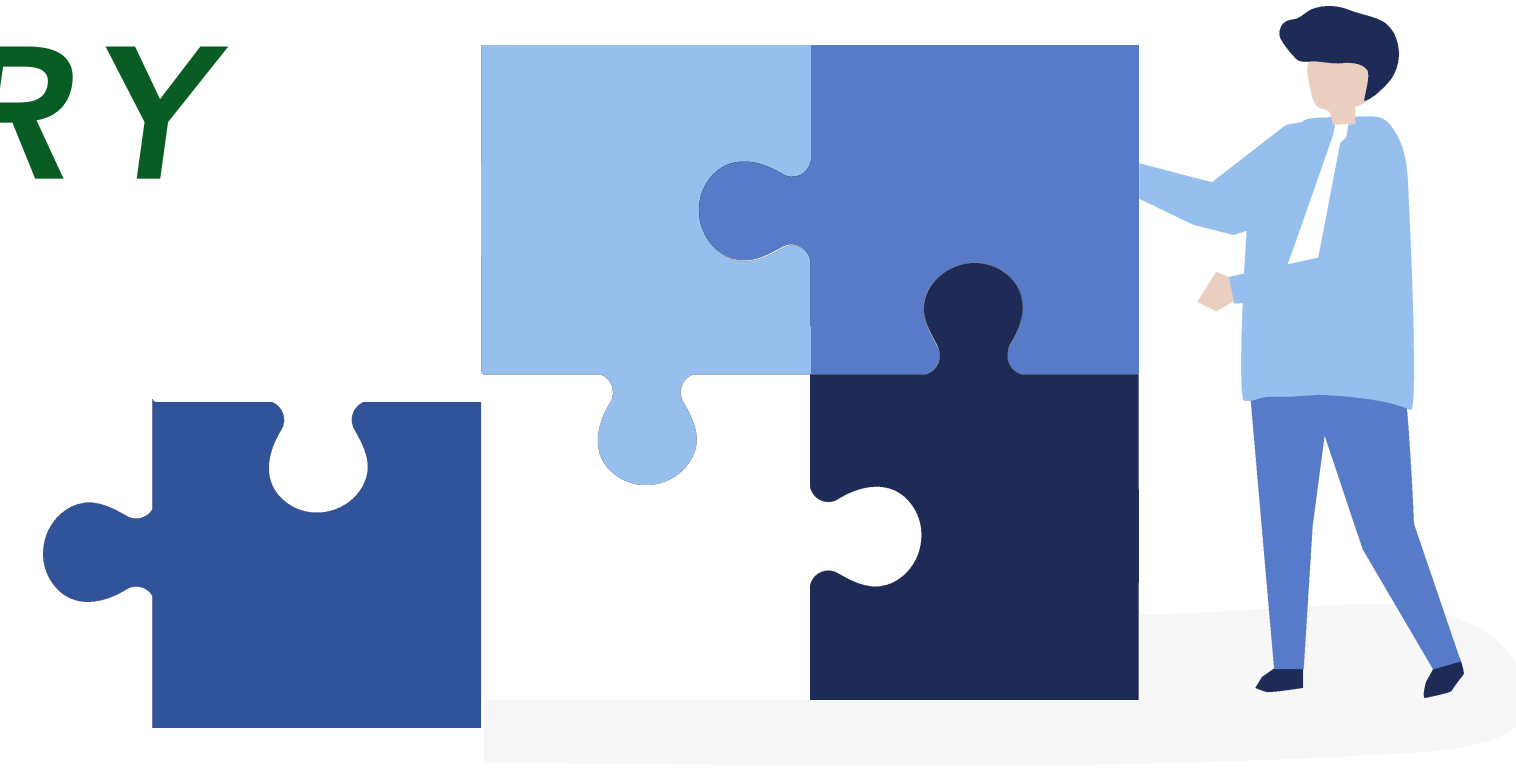
Error Event

d: SSN,
p: treatment

Policy: treatment requires name, age and blood-type
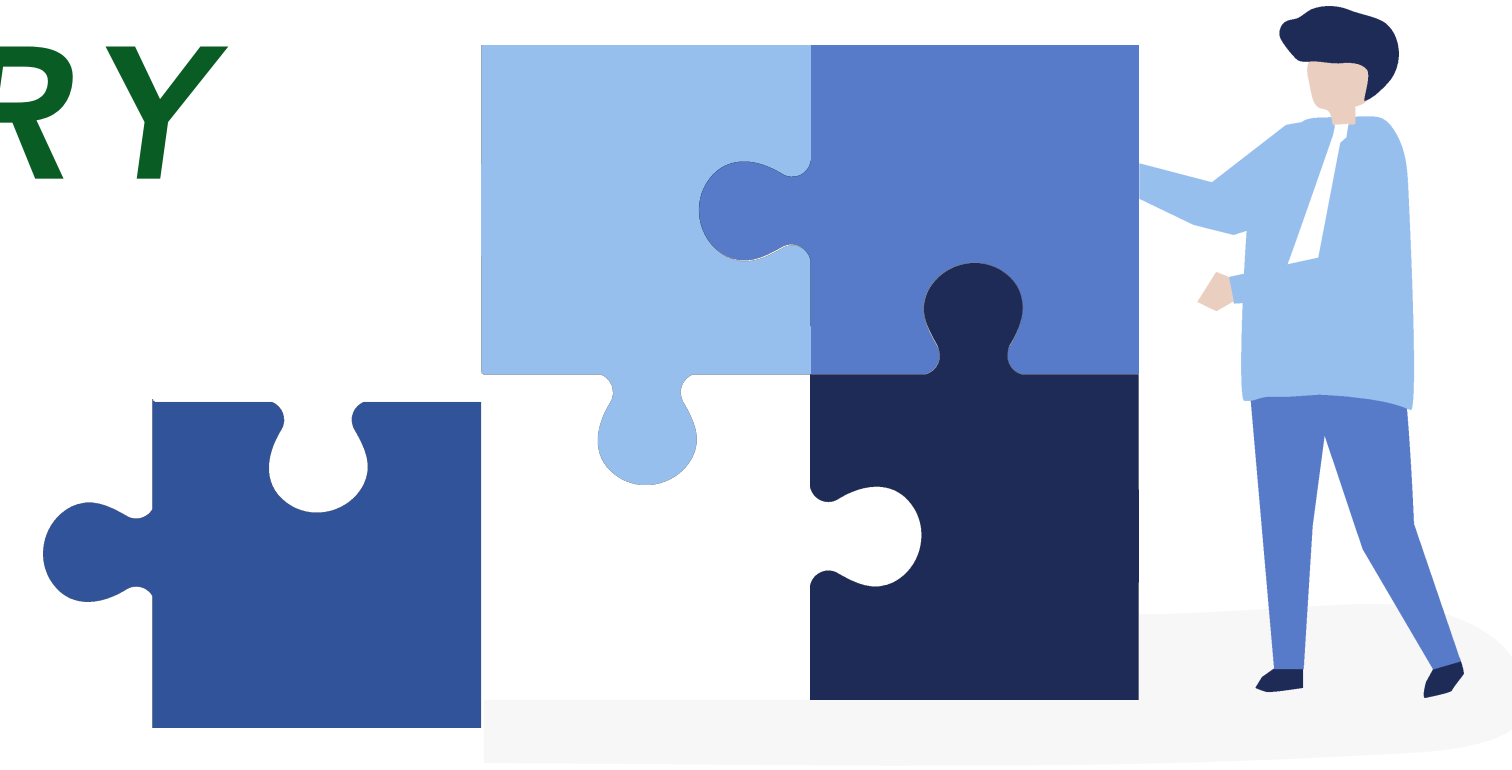
IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

Both Strategies Canceling & Inserting behaviour
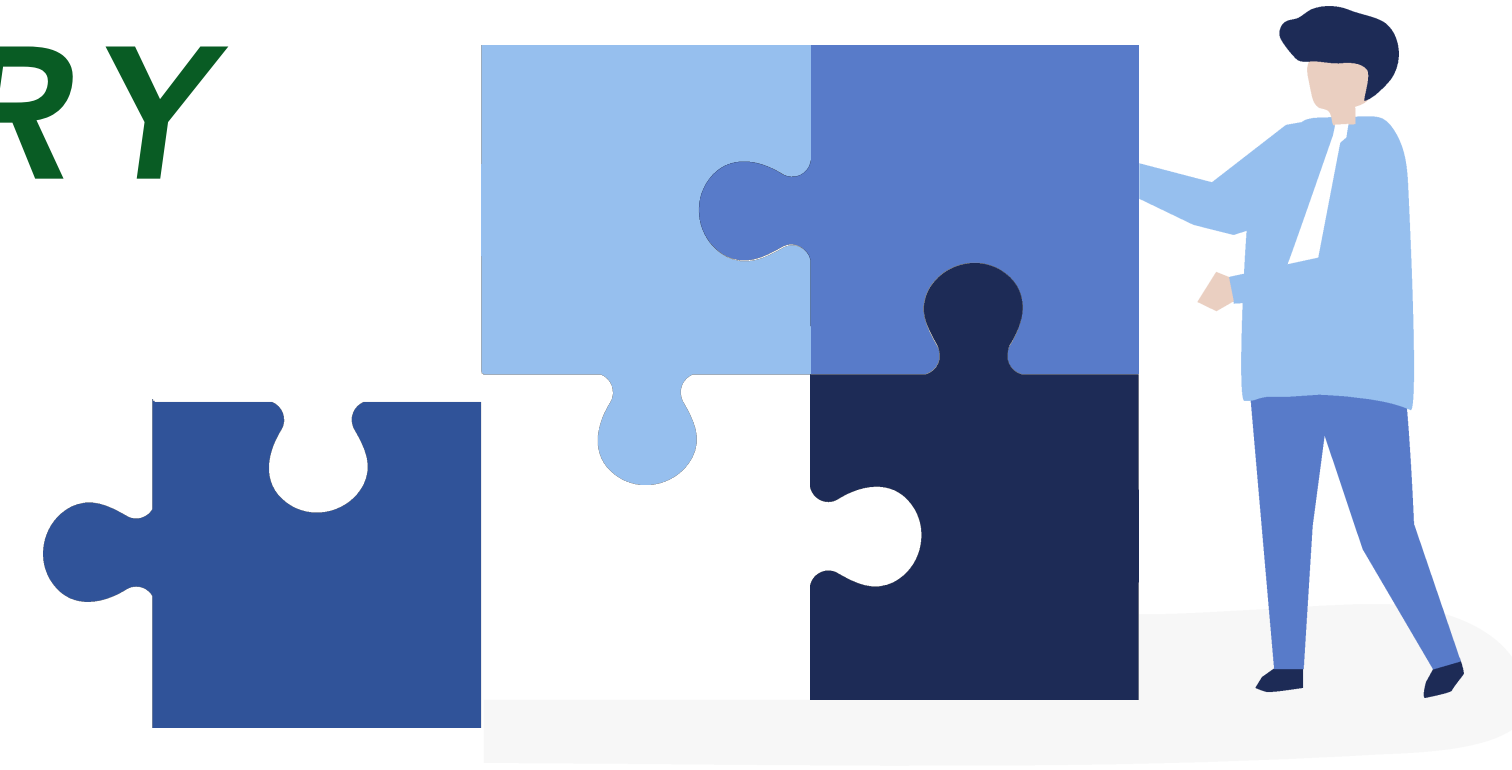
can be applied

IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

Both Strategies Canceling & Inserting behaviour

can be applied

Problem: might trigger new violations!

IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

Both Strategies Canceling & Inserting behaviour

can be applied

Problem: might trigger new violations!

Policy: payment
requires SSN

IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

Both Strategies Canceling & Inserting behaviour

can be applied

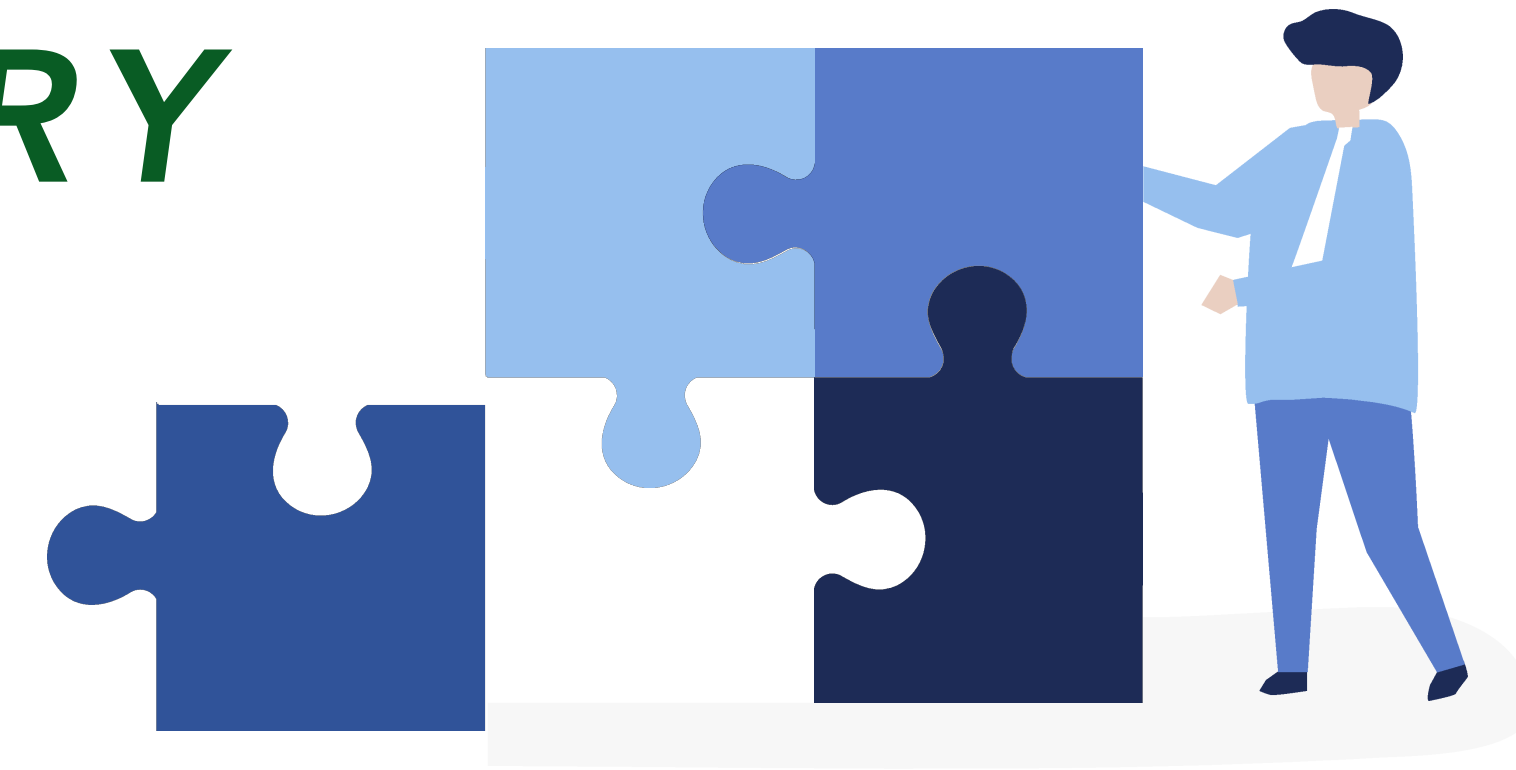*Problem: might trigger new violations!*
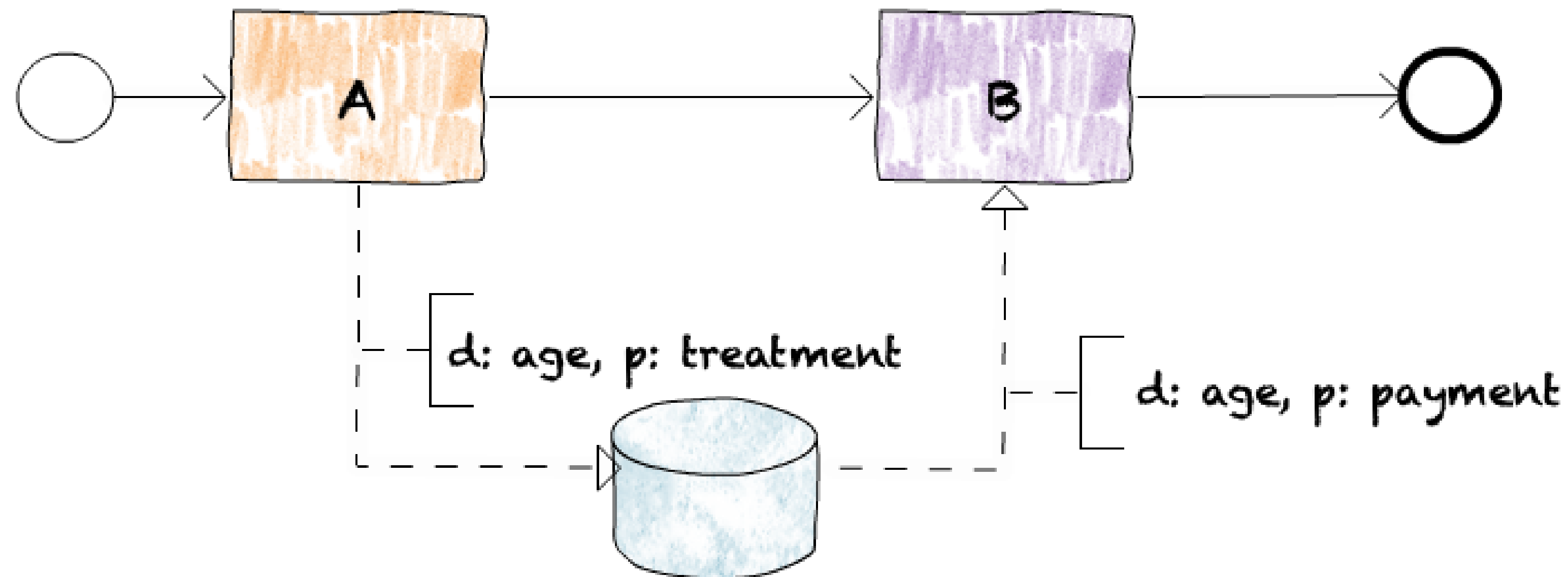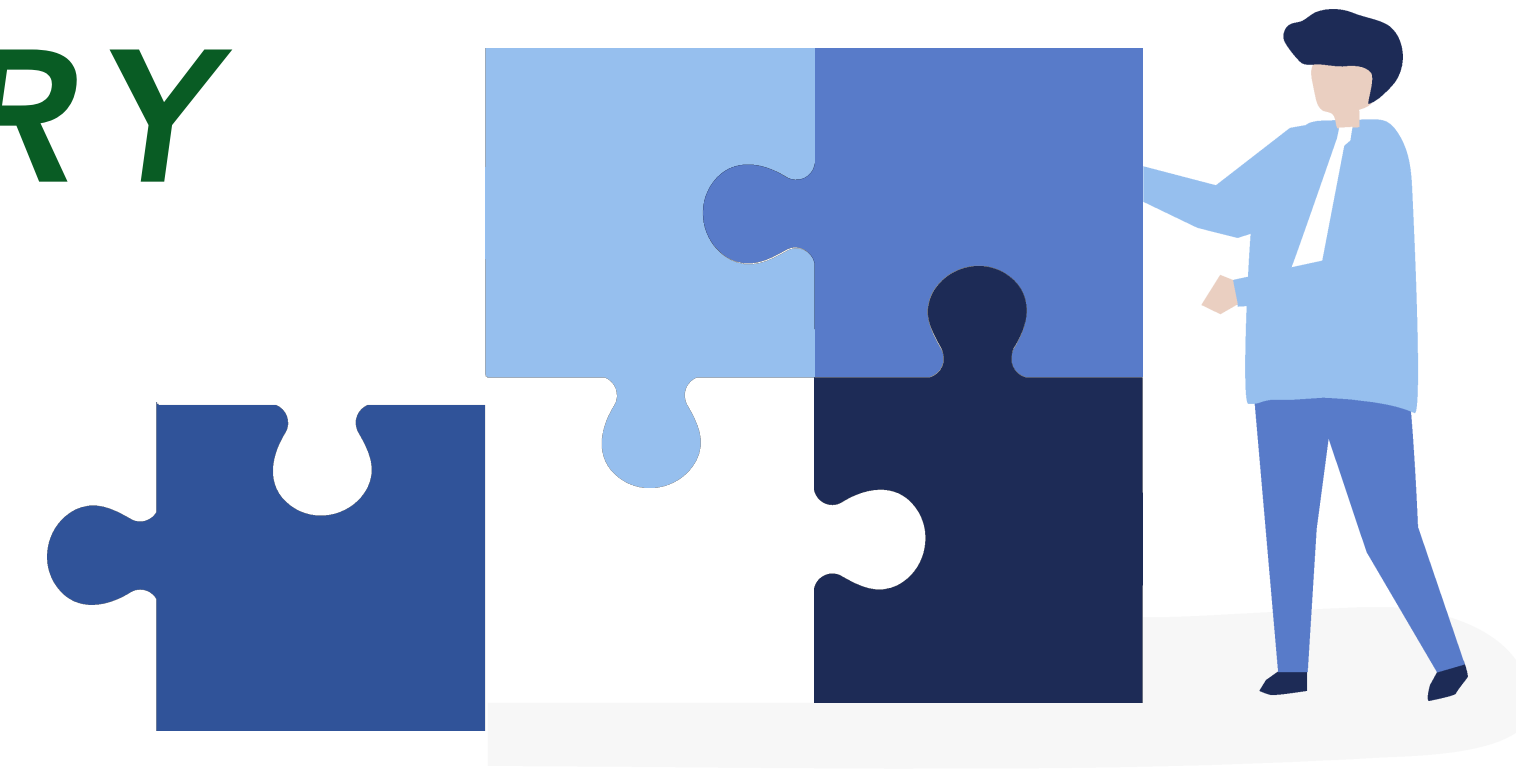
Policy: payment requires SSN
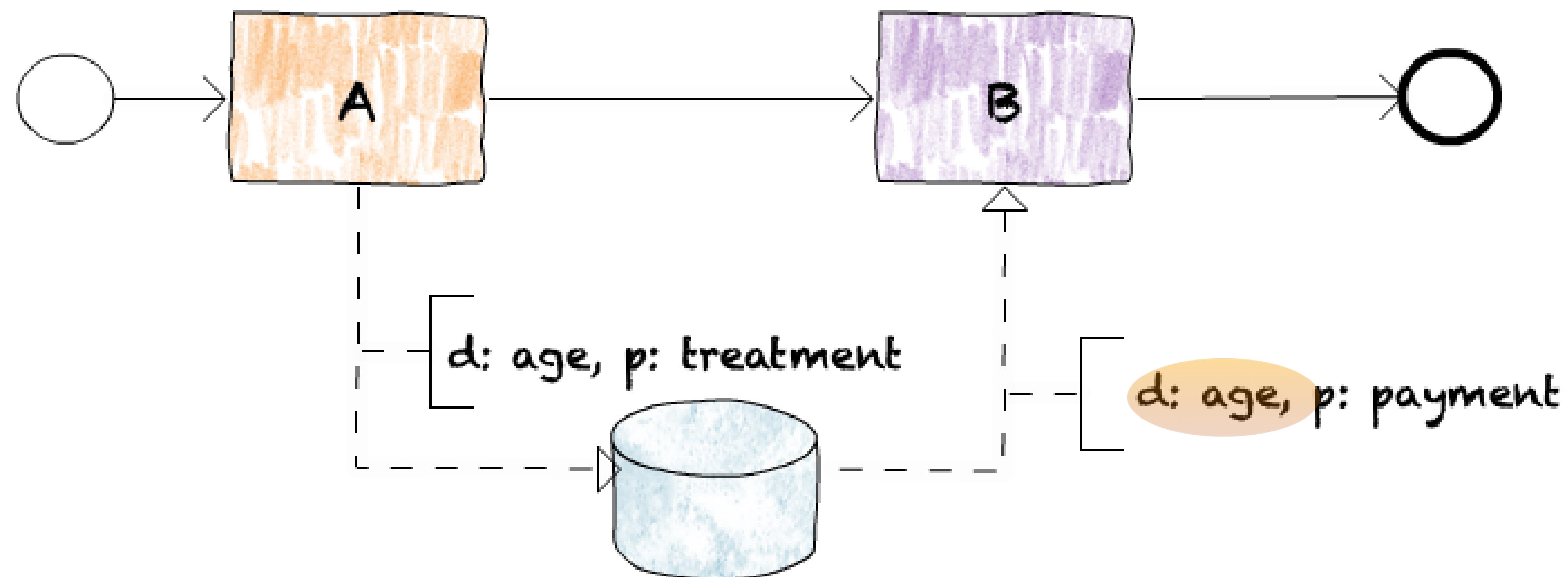
IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

Both Strategies Canceling & Inserting behaviour

can be applied

Problem: might trigger new violations!

Policy: payment requires SSN
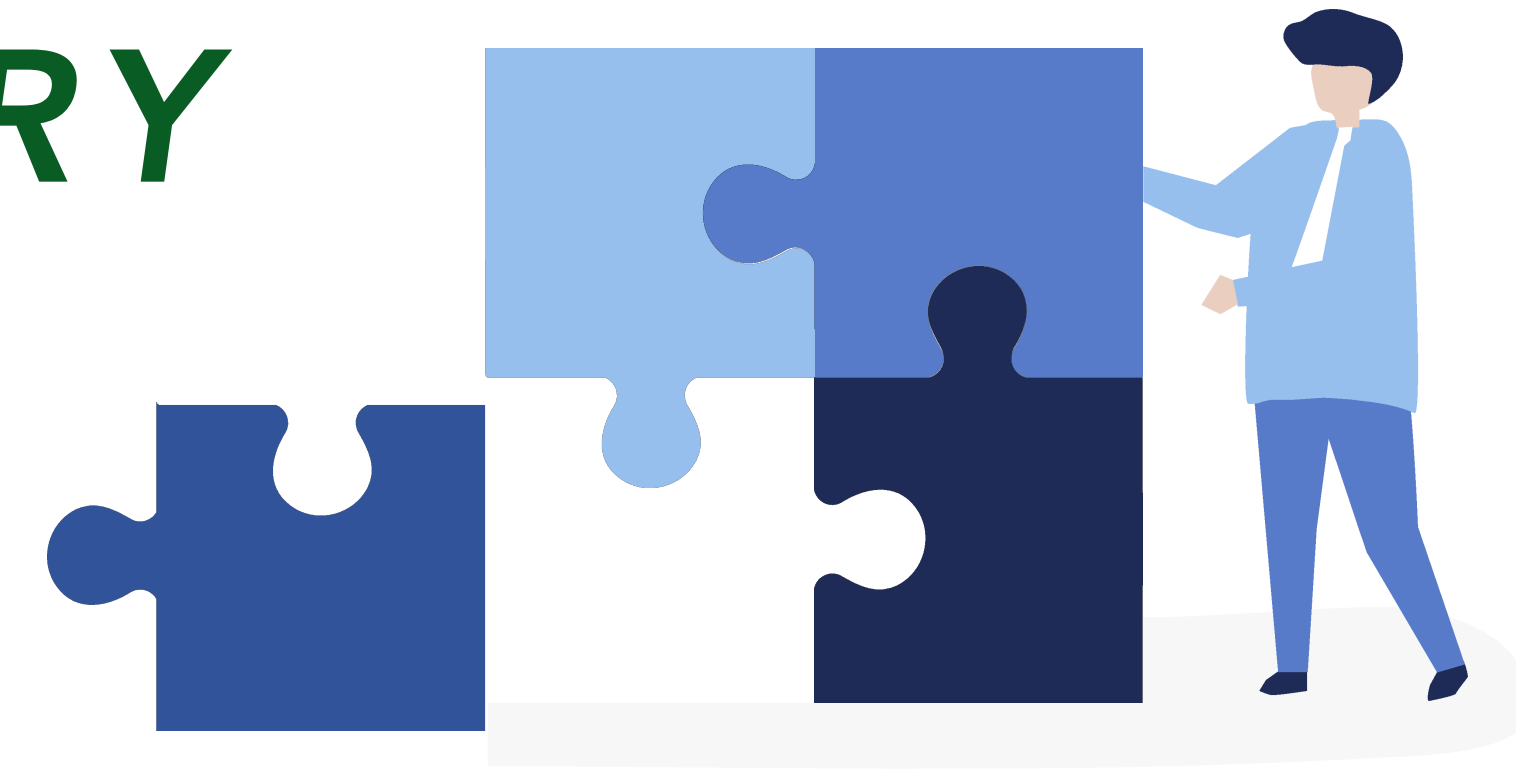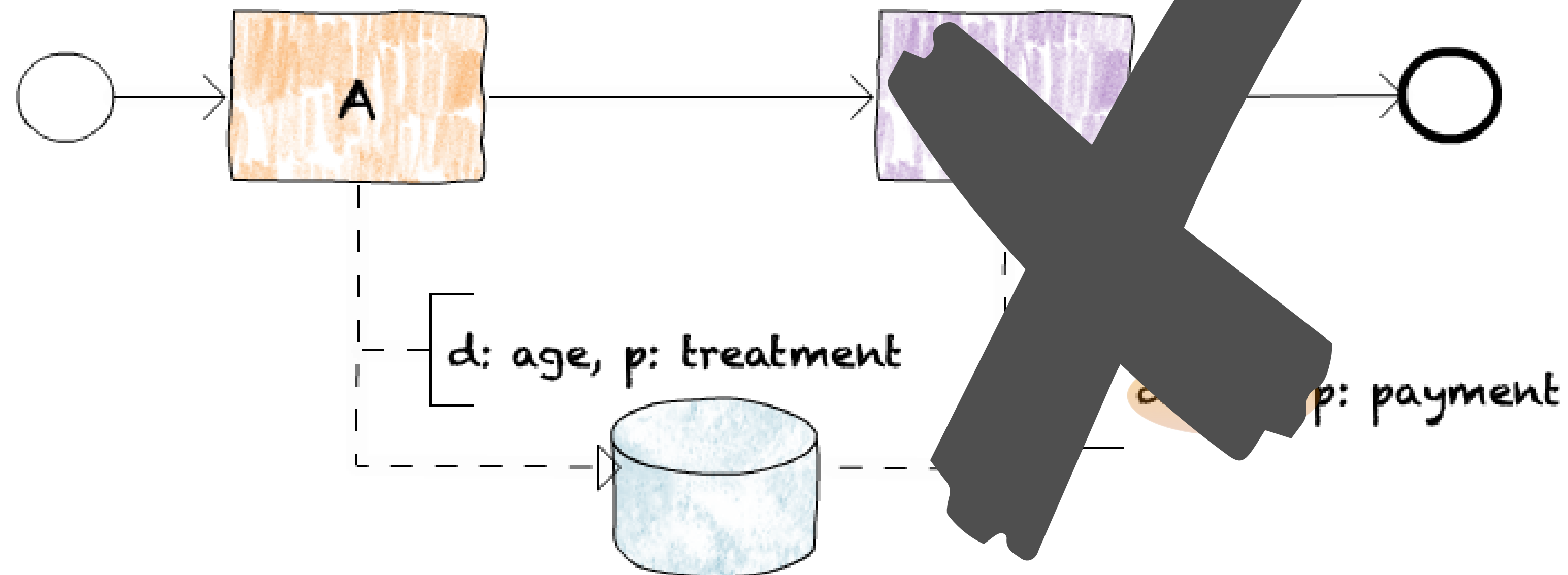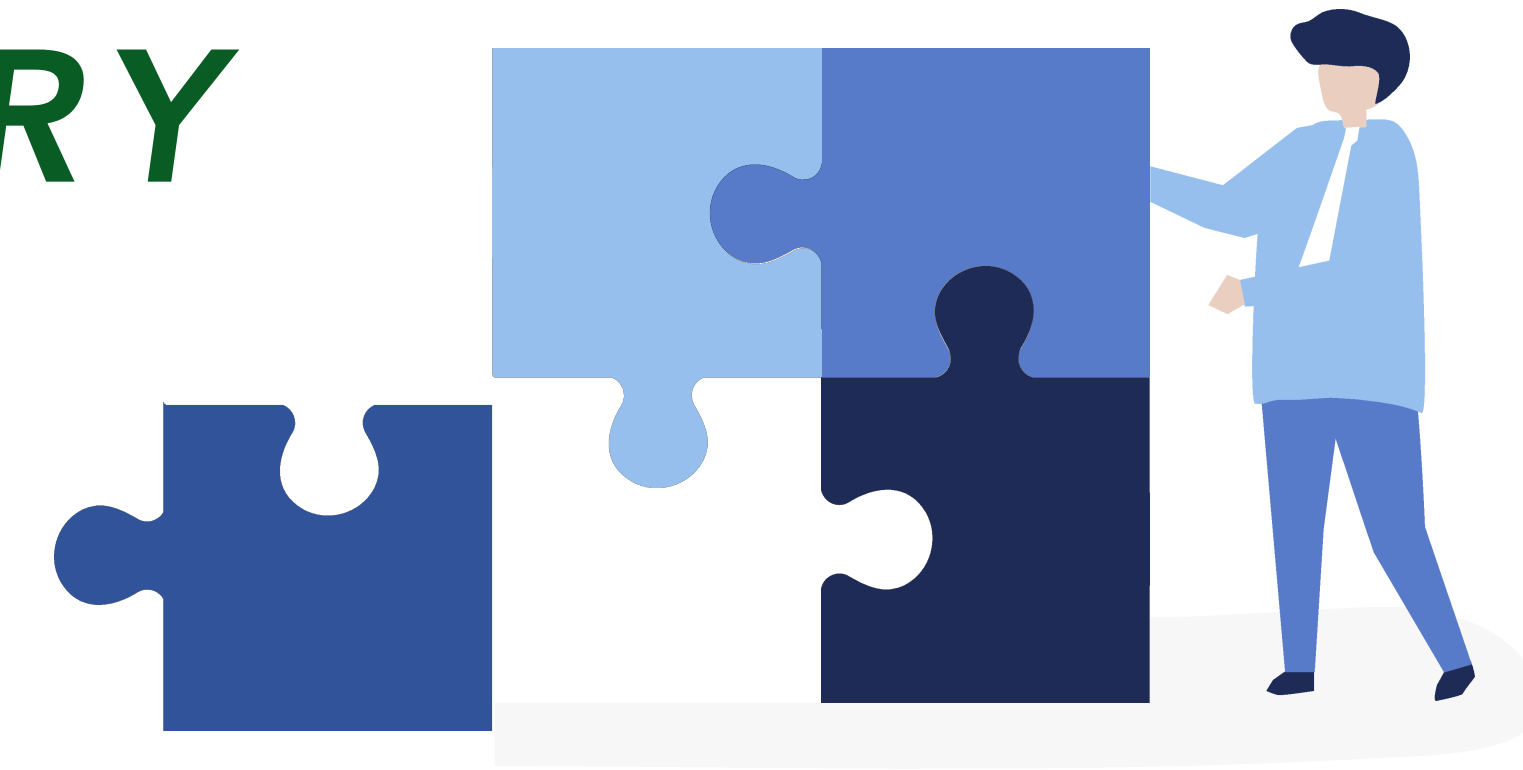
A

d: age, p: treatment

p: payment
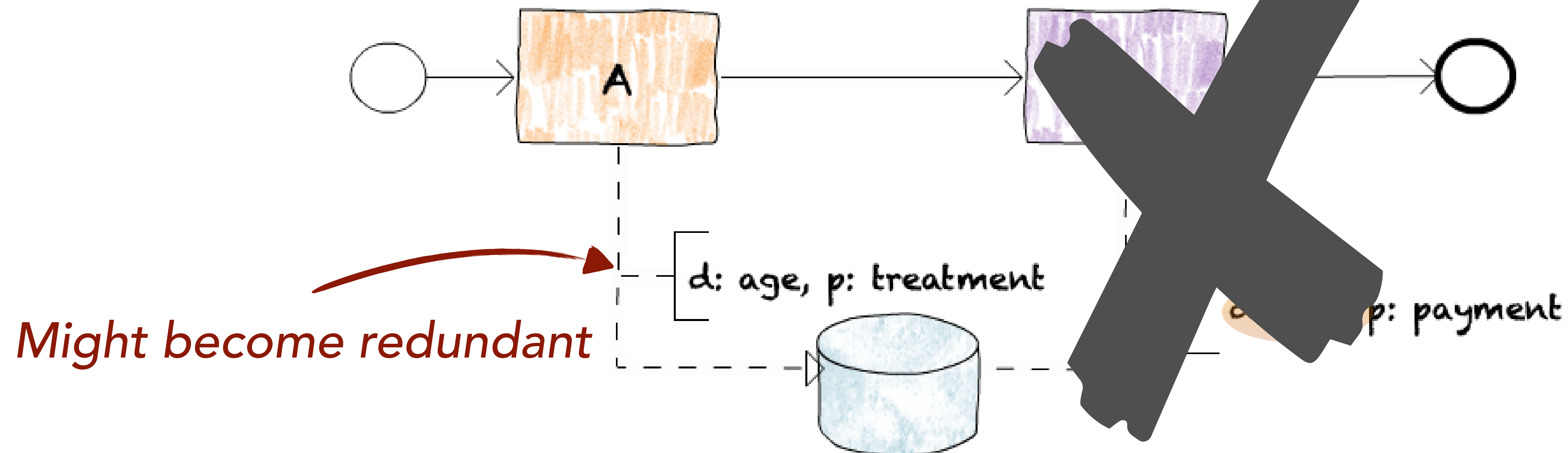
IRRELEVANT DATA - *RECOVERY*

3- If there is data dependency

Both Strategies Canceling & Inserting behaviour

can be applied

Problem: might trigger new violations!

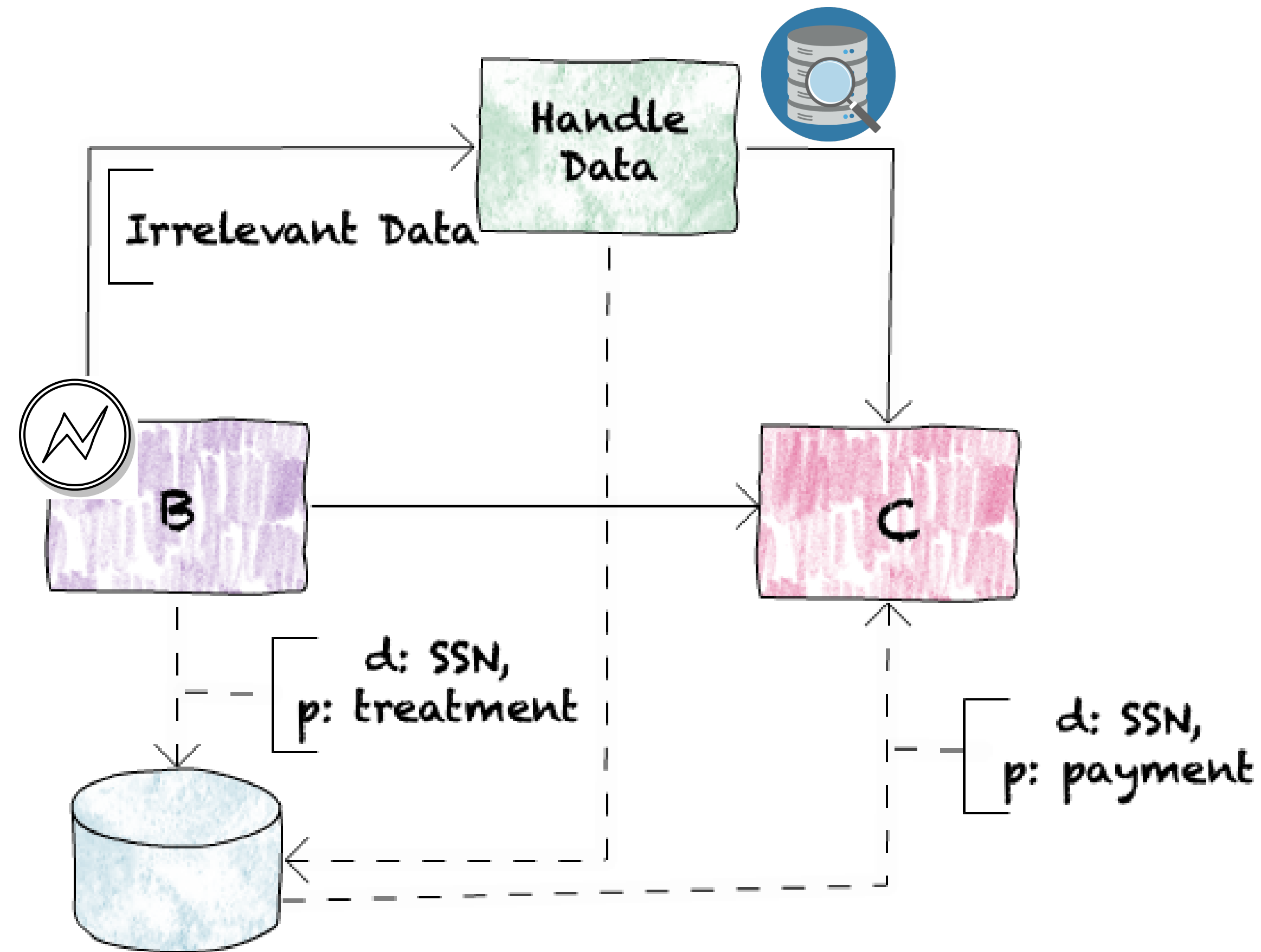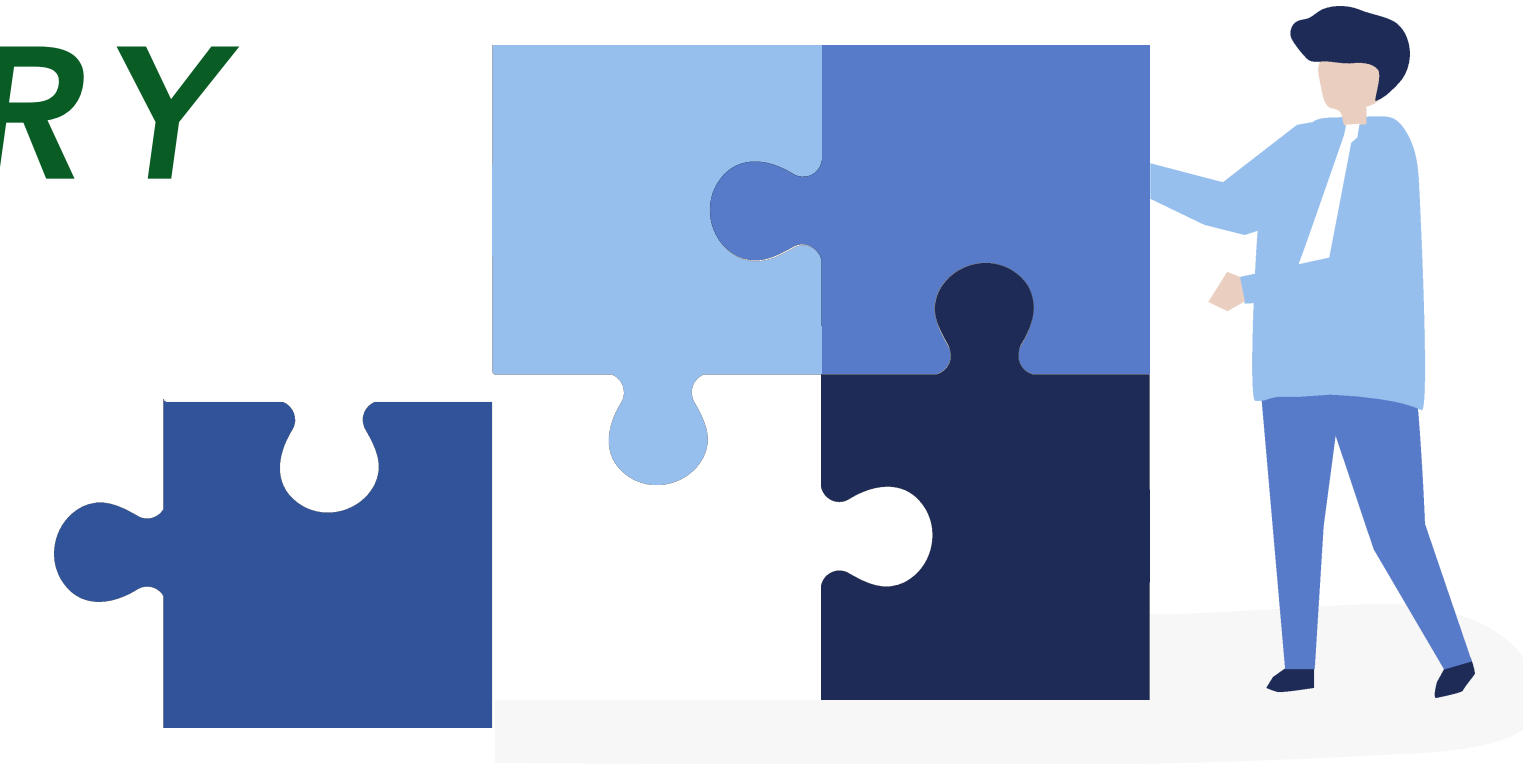**Policy:** payment requires SSN

A

d: age, p: treatment

Might become redundant

p: payment

IRRELEVANT DATA - *RECOVERY*

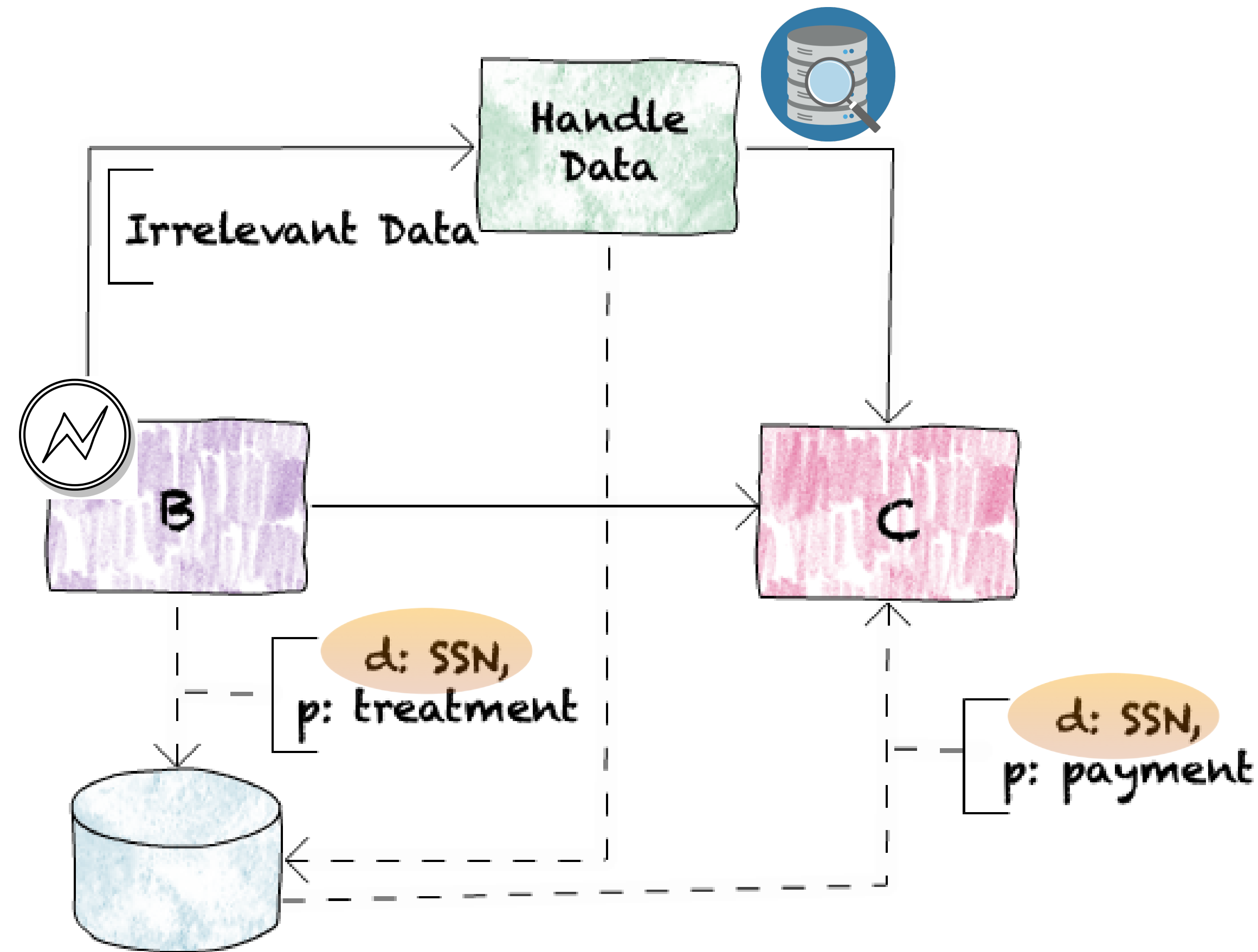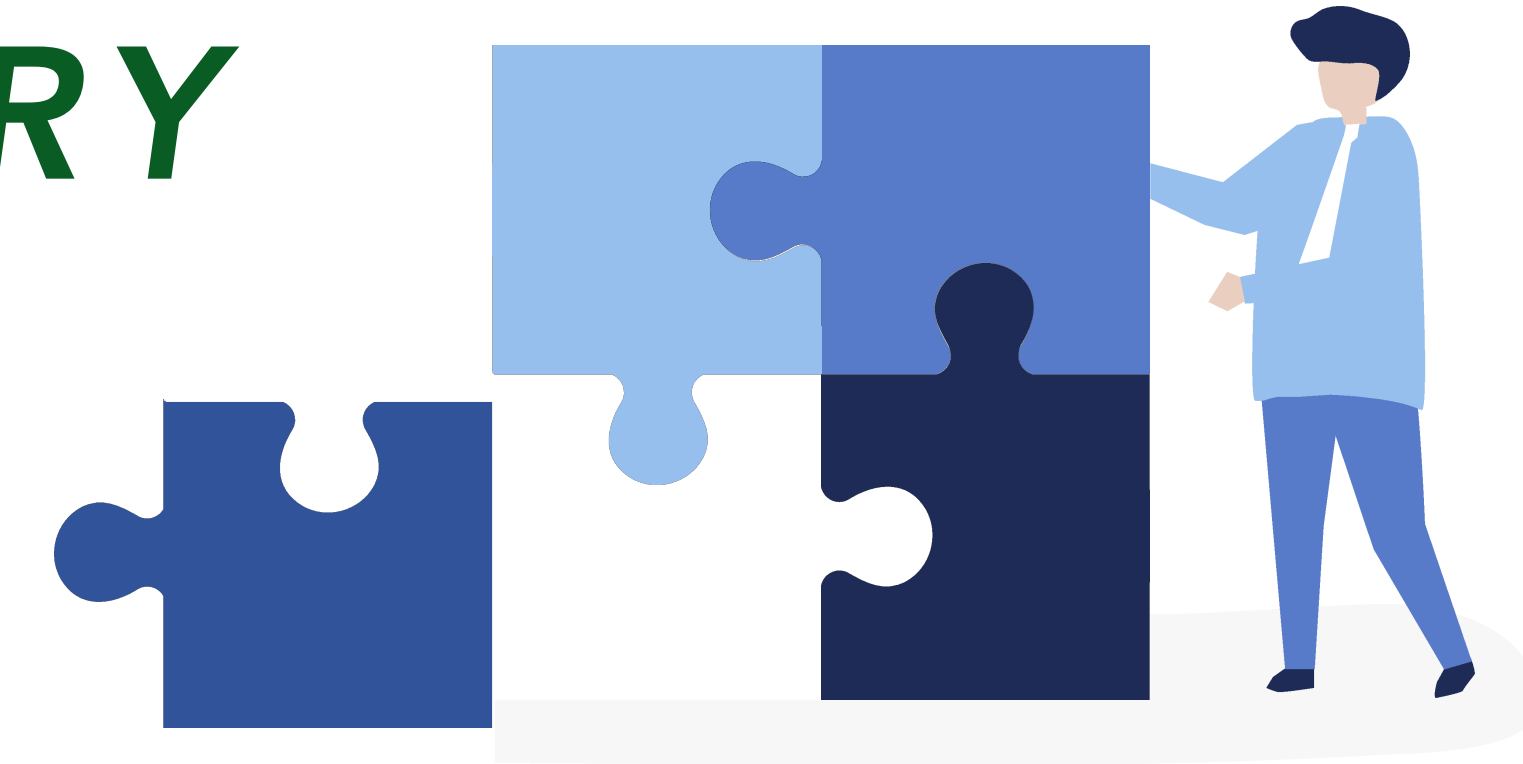*Problem: might trigger new violations!*

Policy#1: payment requires SSN

Policy#2: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*

Problem: might trigger new violations!

Policy#1: payment requires SSN

Policy#2: treatment requires name, age and blood-type

IRRELEVANT DATA - *RECOVERY*
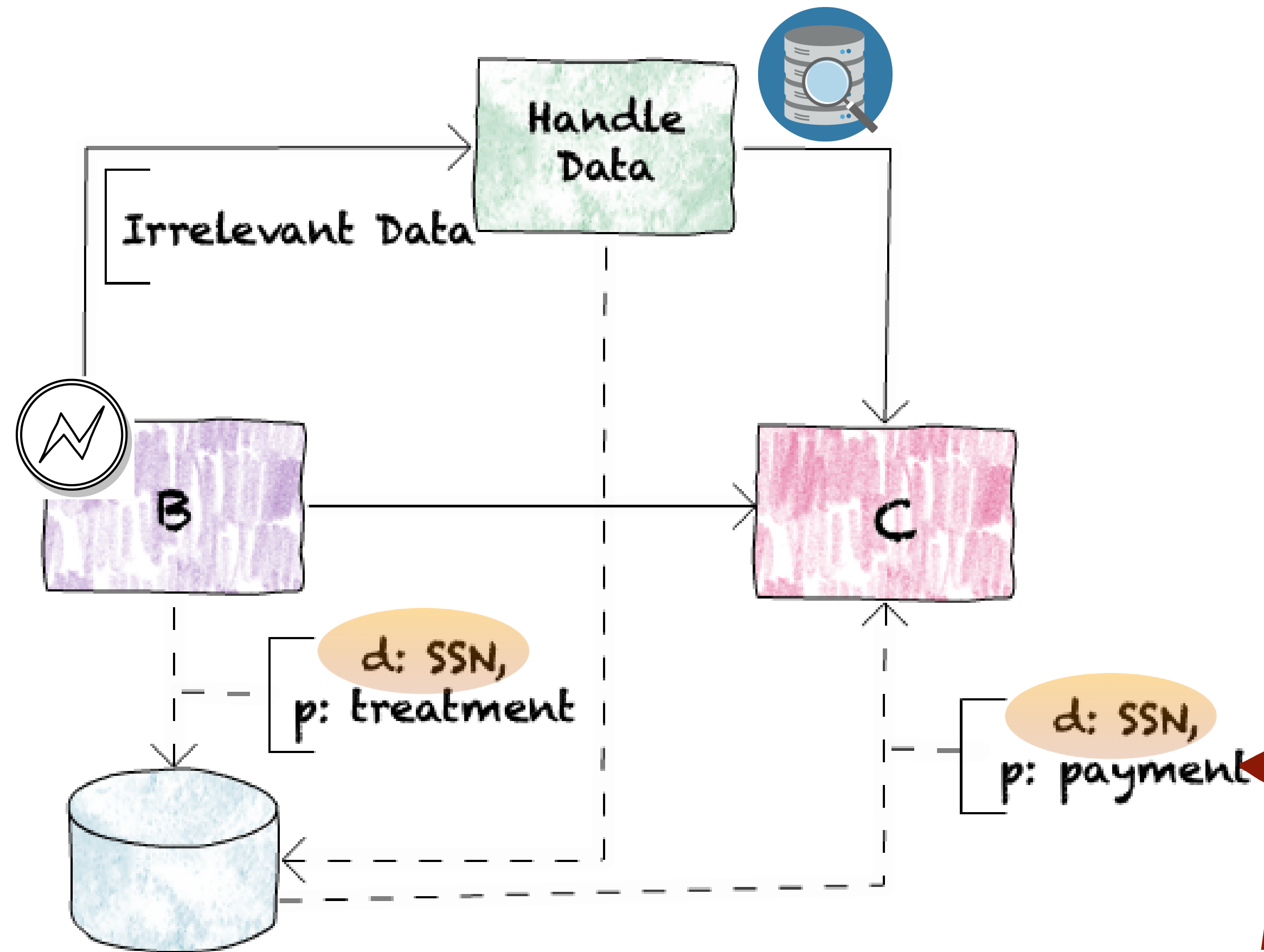
Problem: might trigger new violations!

Policy#1: payment requires SSN
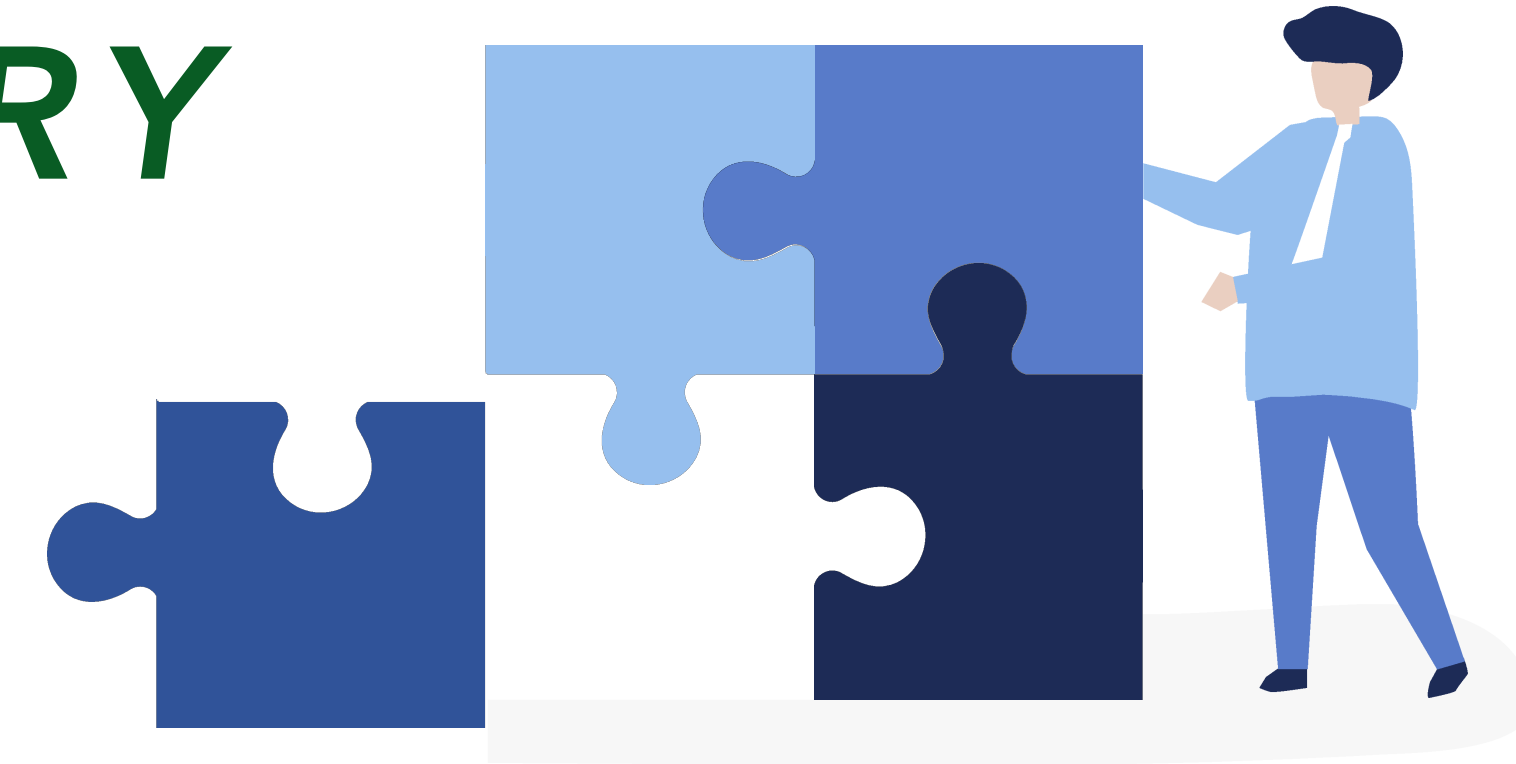
Policy#2: treatment requires name, age and blood-type

Might trigger missing data issue

SUMMARY

SUMMARY

‣ What is *Data Minimisation*?

    ‣ Personal data must be *adequate*, *relevant* and *limited to what is necessary*

SUMMARY
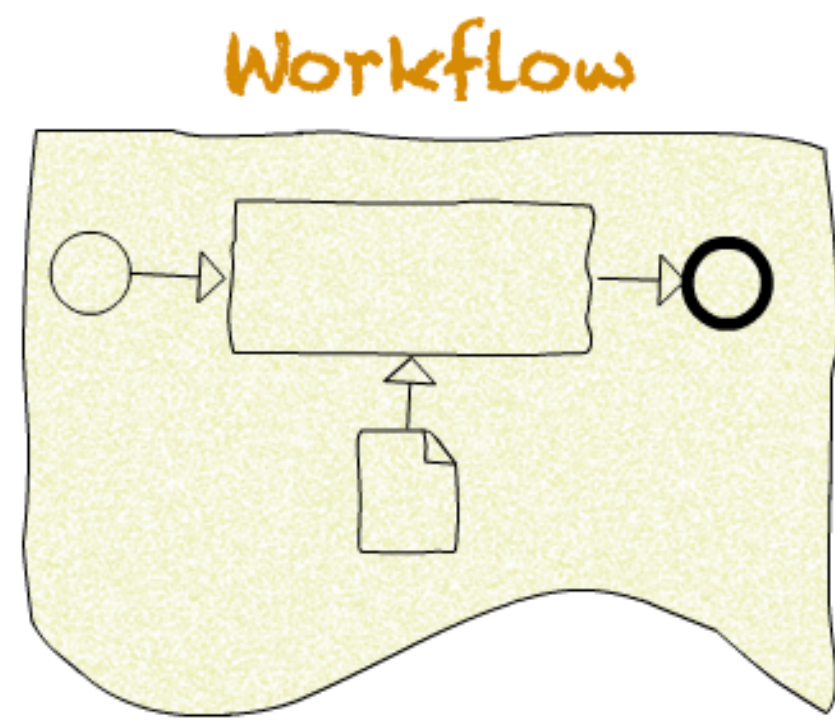
▸ What is *Data Minimisation*?

  ▸ Personal data must be *adequate*, *relevant* and *limited to what is necessary*

▸ What are needed to handle data minimisation for workflows?

  ▸ **Data-Aware Workflow**
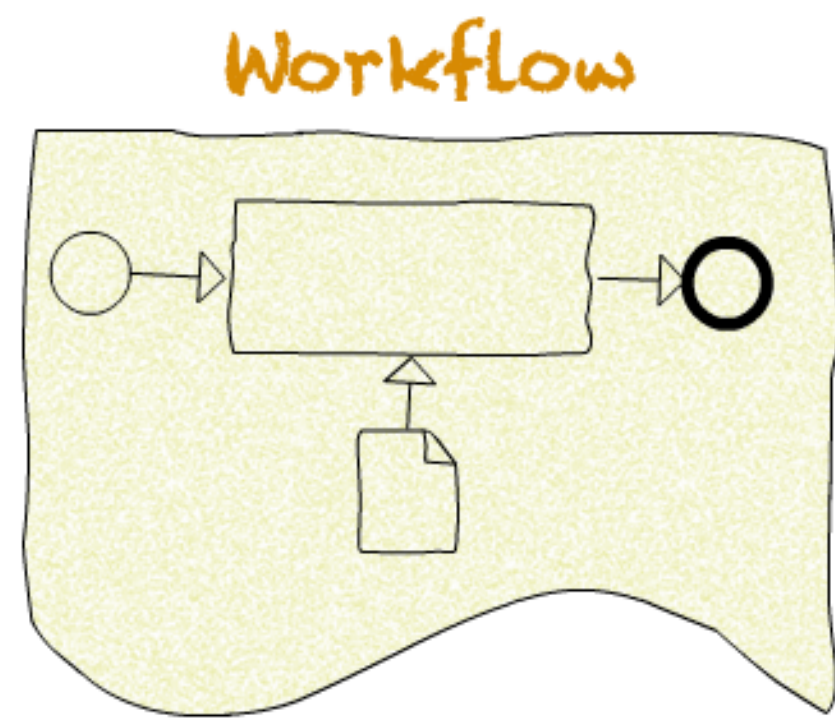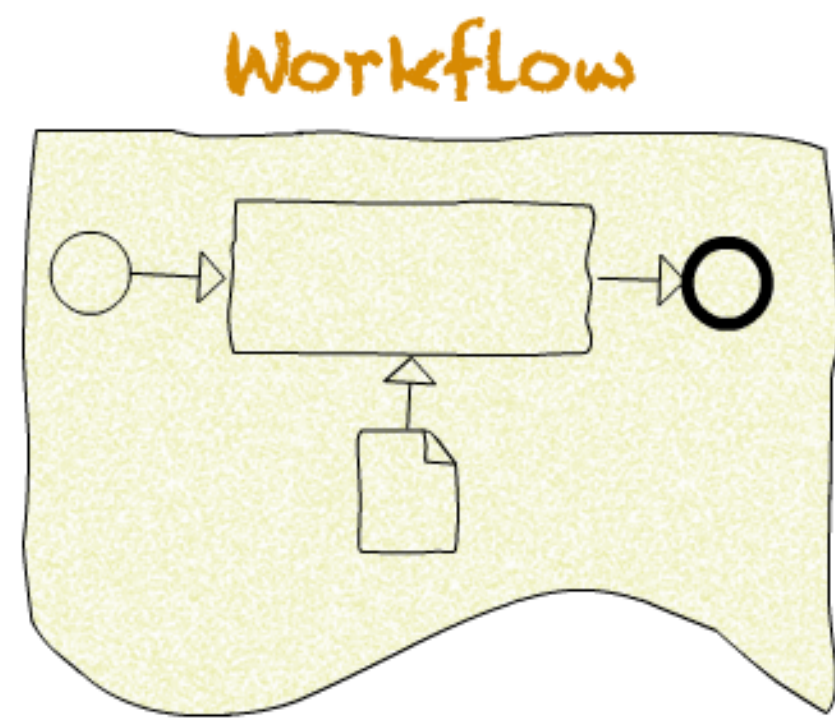
  ▸ **Privacy Policy**

Workflow

might violate data minimisation

might violate data minimisation

‣ How to **detect** data minimisation violations?

‣ How to **recover** WFs from data minimisation violations?

Workflow

→ *might violate data minimisation*

▸ How to **detect** data minimisation violations?

▸ How to **recover** WFs from data minimisation violations?

OUTLOOK

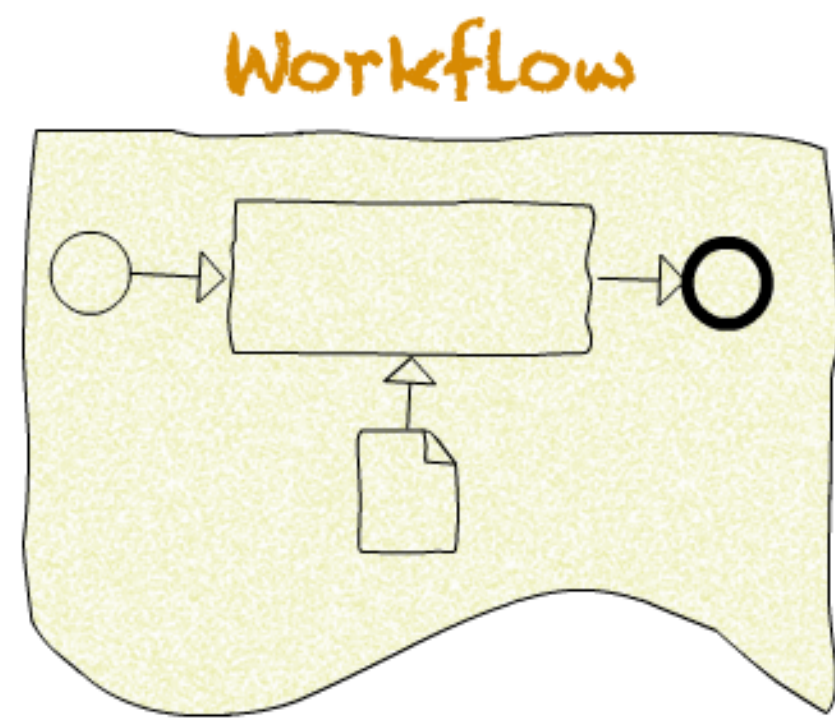◉ Analysis of the **optimality** of the recovery strategies

might violate data minimisation

‣ How to **detect** data minimisation violations?

‣ How to **recover** WFs from data minimisation violations?

OUTLOOK

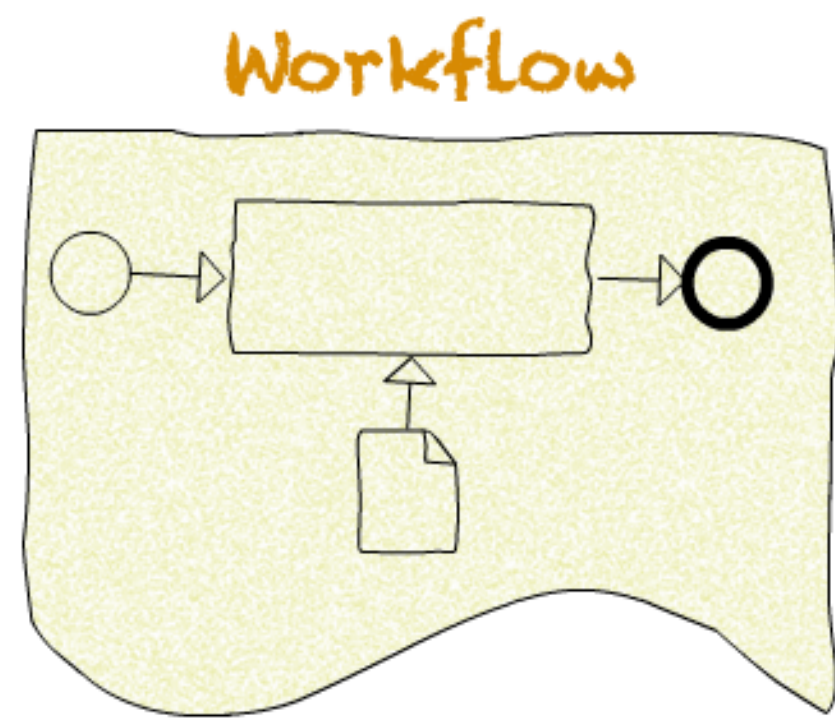◉ Analysis of the **optimality** of the recovery strategies

◉ Conducting a **use case study** to show applicability

Workflow

→ *might violate data minimisation*

▸ How to **det...**

▸ How to **rec...**

Thank you!!!

OUTLOOK

◉ Analysis of the **optimality** of the recovery strategies

◉ Conducting a **use case study** to show applicability

Workflow

→ *might violate data minimisation*

▸ How to **detect** data minimisation violations?

▸ How to **recover** WFs from data minimisation violations?

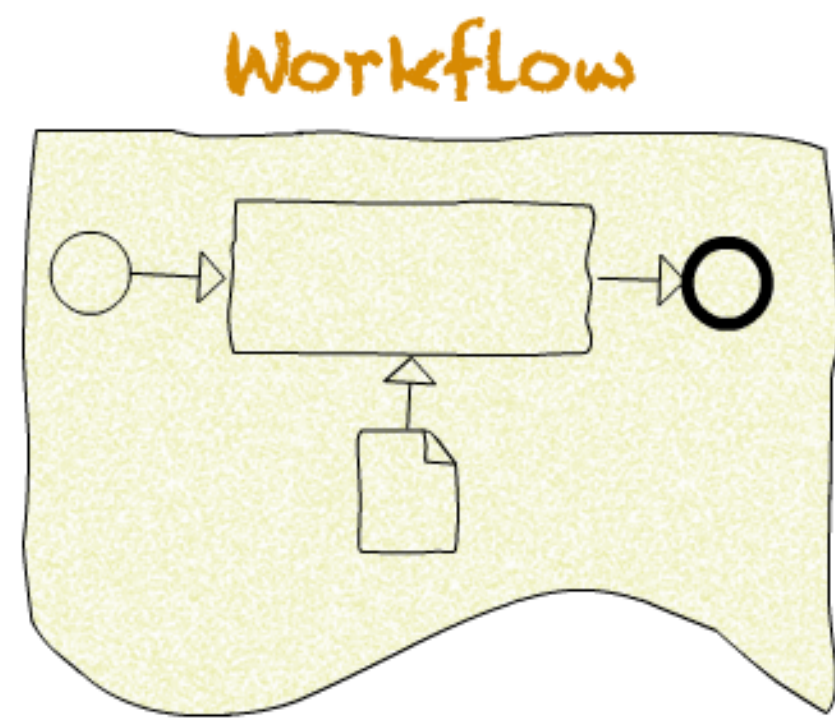## OUTLOOK

◉ Analysis of the **optimality** of the recovery strategies

◉ Conducting a **use case study** to show applicability