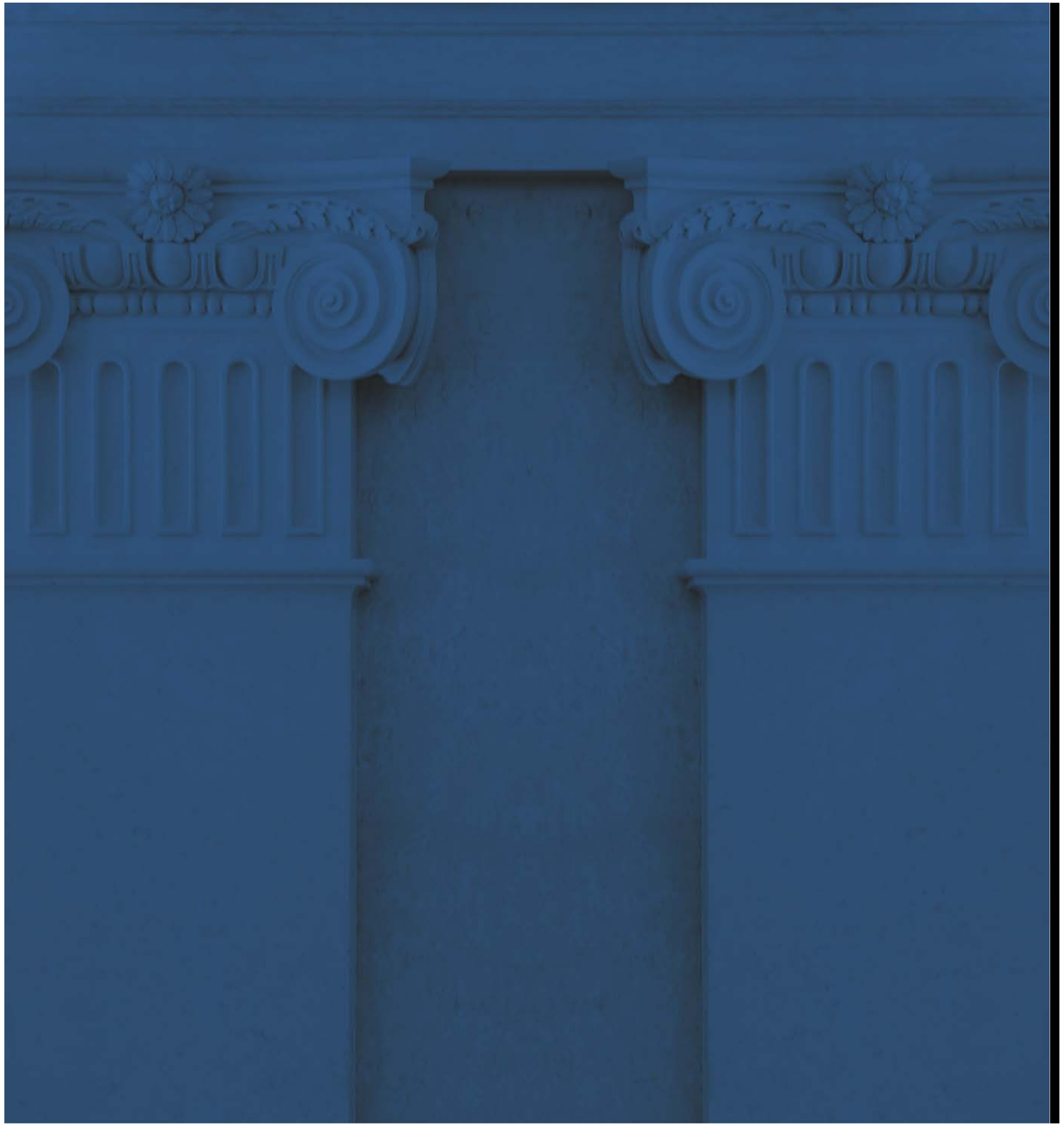


DATA ANONYMIZATION PRINCIPLES AT BANCO DE PORTUGAL

ANA F. CARVALHO
FRANCISCO FONSECA
MÁRIO LOURENÇO
RICARDO MARQUES

4TH IFC-BDI WORKSHOP ON
"DATA SCIENCE IN CENTRAL BANKING“

ROME, FEBRUARY 2025

BANCO DE
PORTUGAL
EUROSISTEMA

AGENDA

01 | MOTIVATION

02 | ANONYMIZATION PRINCIPLES

03 | DEFINING THE PATH TOWARDS IMPLEMENTATION

04 | CLOSING REMARKS

MOTIVATION

01

- How do we balance the need to **protect our data** while keeping it available to be easily **used for different purposes**?

MOTIVATION
GRANULAR DATA AND DATA PRIVACY

Granular data containing information on natural persons, which is increasingly available...

... Has great potential

- Self-service exploration for analytical purposes
- Elimination of redundancies which may lead to a reduction of the reporting burden of institutions
- Making full use of increasing computational capabilities and **ML/AI** algorithms

... But must be handled carefully

- Wrongful disclosure of data represents a very significant reputational and legal risk
- Data protection regulations are very demanding (GDPR)
- Transition to a cloud-based infrastructure poses further potential security risks

MOTIVATION

RECENT DEVELOPMENTS OF BANCO DE PORTUGAL'S DATA INFRASTRUCTURE

- Integrated Data Management Program
- Developing data culture increases the interest in exploring our data
- Streamlined data processing procedures with centralized data repositories
- BdP DataHub – an effort towards the integration of all data reported to the Bank

- Increasingly sensitive data
(CCR, Banking Deposits Database, Household's Income, …)
- Clear access policy that ensures that people only access sensitive data on a **need-to-know basis…**
- … while making sure that self-service access to data is still possible when appropriate.

MOTIVATION

THE ROLE OF ANONYMIZATION

- **Anonymization techniques** play a vital role in **increasing data protection**, complementing infrastructure security

- Anonymization models are **very diverse**, and their impact on **data integrity** varies

- Building upon previous experiences, BdP has defined a model that **minimizes loss of data integrity** but meaningfully increases **data security**

ANONYMIZATION PRINCIPLES

02

ANONYMIZATION PRINCIPLES

MAIN GOALS

Define

Ensure that core concepts – **personal data, anonymization, pseudonymization** – have **a common definition** at the institutional level

Prototype

Idealize and implement a working prototype for **an internal pseudonymization algorithm**

Idealize

Consider the **key issues** that must be tackled when defining a data privacy model for the **department**, and eventually for **the bank**

Implement

Objectively define the steps that must be taken **to implement the pseudonymization model** and establish the corresponding **governance model**

ANONYMIZATION PRINCIPLES

SOME DEFINITIONS

Personal data

Variables or sets of variables that:

- Individually or when combined;
- Directly or indirectly;

Relate to an individual and allow us to identify them with high confidence.

These variables can be divided into:

- Identifiers
- Quasi-identifiers

Pseudonymization

Process that transforms personal data such that the risk of identification is significantly reduced. Methods include:

- Eliminating high risk variables;
- Replacing identifiers with pseudonyms;
- Lowering the level of detail in the data.

There is generally **a strong positive relationship** between the **reduction in risk of identification** and the implicit **loss of information**.

ANONYMIZATION PRINCIPLES

MAIN PIECES

Main information pieces of any data anonymization/pseudonymization model

ORIGINAL (IDENTIFIED) DATA

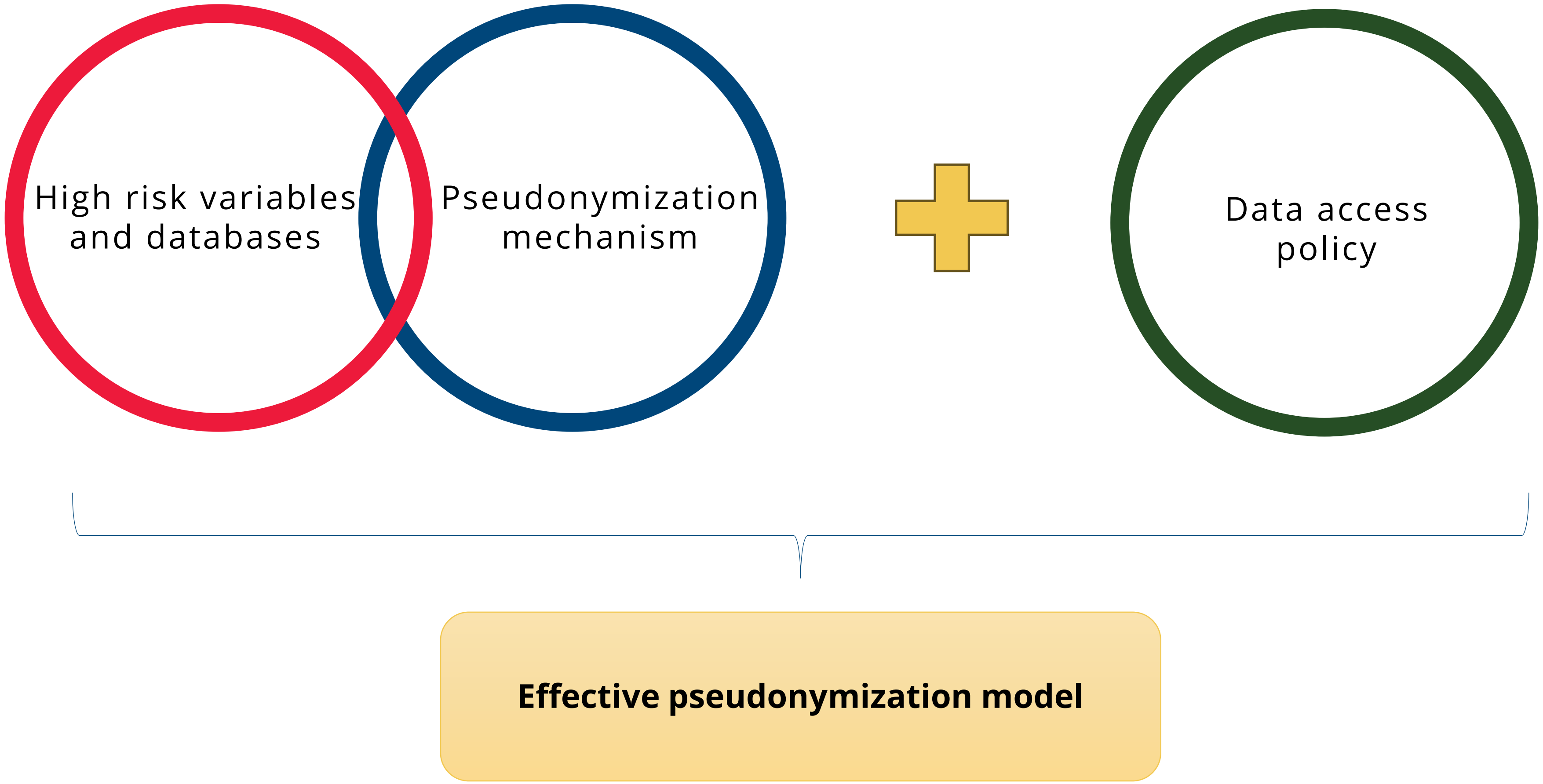
PSEUDONYMIZATION MECHANISM

PSEUDONYMIZED DATA

If a user knows two out of three of these pieces, he can reverse the pseudonymization process and subvert the model

ANONYMIZATION PRINCIPLES

WHAT WE MUST TACKLE



High risk variables and databases

Pseudonymization mechanism

Data access policy

Effective pseudonymization model

DEFINING THE PATH TOWARDS IMPLEMENTATION

03

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – MAIN TOPICS

1

Critical databases and variables

Evaluate **all databases** where the unit is an individual, and define a list of quasi-identifiers that pose a significant risk of identification

2

Define an access policy

Proposed policy comprised of **3 types of profile**, and focused on **democratizing the access to all pseudonymized data**

Pseudonymization algorithm

Define and implement an **internal pseudonymization algorithm** that can be applied to **all relevant databases** regardless of the types of identifier/quasi-identifiers

3

Data flow

Implement a **data flow** policy that **must be followed** when receiving, using, and sharing **any data where the unit is an individual**

4

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – CRITICAL DATABASES AND VARIABLES

1

We propose:

Evaluate all the databases within the Bank where the unit is an individual (including corporations, when relevant)

Define, at the institutional level, a list of quasi-identifiers that, by themselves or when combined, pose a significant risk of identification

When pseudonymizing a given database, we should refer to this list of variables to determine if any of them should be suppressed

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – ACCESS POLICY

2

We propose 3 types of profile:

ORIGINAL DATA (TYPE A)

Can access the **minimal necessary set** of **identified data** to carry out his work

Is ignorant of the pseudonymization mechanism and does **not have access to any pseudonymized data**

PSEUDONYMIZATION MECHANISM (TYPE B)

Has access to **all data necessary to conduct the pseudonymization process**

Responsible for implementing and applying the **algorithm**

Restricted to the smallest possible number of people

PSEUDONYMIZED DATA (TYPE C)

Has access to **all pseudonymized databases**

Is ignorant of the pseudonymization mechanism and does **not have access to any identified data**

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – ACCESS POLICY

2

Additionally, we propose:

Type B access should be attributed either to the **IT Department,** in the case of a Bank-wide model, or to **the Information Management Division,** in case of a Department-wide model

Type C access should be given to **every person** that does not have any **Type A** access

Across the Department and the Bank, teams should be set up in such a way as to **isolate all functions that require access to identified data**

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – PSEUDONYMIZATION ALGORITHM

3

We propose a pseudonymization algorithm:

- Recommended and considered sufficient within the context of the GDPR
- Focuses on identifiers, rather than quasi-identifiers
- Preserves the possibility of joining different databases with common identifiers
- Developing an “in-house” model is relatively simple, increasing trust and security
- Ensures reversibility, although the process should only be reversed if strictly necessary

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – PSEUDONYMIZATION ALGORITHM

3

Concerning the algorithm, we propose:

The pseudonymization algorithm **should be defined internally**, should ensure **unique pseudonyms,** and include at least one step that depends on **a key**

The algorithm should be sufficiently **flexible** to be applied to different types of identifiers (numerical, alphanumerical, etc.)

The algorithm should be applied **after defining** the variables that should be made available in the **anonymized database**

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – DATA FLOW

4

Concerning the data flow, we propose:

Any data sharing, **formal or ad hoc**, must follow the defined **data flows**

The model should be **agile and flexible,** so that people are not tempted to avoid following the data flow to share the data faster

CLOSING REMARKS

04

CLOSING REMARKS

WAY FORWARD

- Thorough **discussion on the viability of our recommendations**

  (particularly regarding the segmentation of different roles and the allocation of the responsibilities defined under our model)

- **Define who's who**

  (assign roles and responsibilities, evaluate the possibility of having a segmentation of roles between people who need to access identified data and those who don't)

- **Experimentation through different use cases**

  (apply pseudonymization techniques to different databases, making them available to specific sets of users)

QUESTIONS

FFONSECA@BPORTUGAL.PT
MFLLOURENCO@BPORTUGAL.PT