Classification of procedures

Vicenç Torra

November, 2022

Umeå University

Outline

Dimensions

Data Privacy: Dimensions

Classification of data protection procedures

- Alternative dimensions for classification
  - Classification 1:
    - On whose privacy is being sought
  - Classification 2:
    - On the computations to be done
  - Classification 3:
    - On the number of data sources

Dimensions. 1st classification

On whose privacy is being sought

Data Privacy

Dimension 1: On whose privacy is being sought

Data Privacy

Dimension 1: On whose privacy is being sought

Subjects involved: Respondent, owner and user

Data Privacy

Dimension 1: On whose privacy is being sought

Subjects involved: Respondent, owner and user

- Respondents' privacy (*passive* data supplier, data subject)
- Holder's privacy (or owner's, controller's)
- User's privacy (*active*)

GDPR: (Article 4)

- Data subject: (Undefined): 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person
- Data controller: the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data
- Data processor: a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller
- Third party: a natural or legal person, public authority, agency or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data

Dimensions: 1st

- **Ex. 3.1.** A hospital collects data from patients and prepares a server to be used by researchers to explore the data.

Dimensions: 1st

- **Ex. 3.1.** A hospital collects data from patients and prepares a server to be used by researchers to explore the data.
  - **Case 1. Database of patients.**
    Actors:
    - Holder: the hospital
    - Respondents: the patients

Dimensions: 1st

- **Ex. 3.1.** A hospital collects data from patients and prepares a server to be used by researchers to explore the data.
  - **Case 1. Database of patients.**
    Actors:
    - ▷ Holder: the hospital
    - ▷ Respondents: the patients
  - **Case 2. Database of queries.**
    Actors:
    - ▷ Holder: the hospital
    - ▷ Respondents: researchers
    - ▷ User's: researchers if they want to protect the queries

Dimensions: 1st

- **Ex. 3.2.** An insurance company collects data from customers for internal use. A software company develops new software. A fraction of the database is transferred to the software company for software testing.
  - **Database transferred to a software company.**
    Actors:
    - ▷ Holder: The insurance company
    - ▷ Respondent: Customers
    - ▷ The software company is neither data processor nor third party if they do not process personal data but pseudonymized

Dimensions: 1st

- **Ex. 3.4.** Two supermarkets with fidelity cards record all transactions of customers. The two directors will mine relevant association rules from their databases. In the extent possible, each director do not want the other to access to own records.
  - **Two supermarkets and two DBs to mine**
    Actors:
    - ▷ Holder: Supermarkets
    - ▷ Respondent: Customers

Dimensions: 1st

- **Dimension 1.** Whose privacy is being sought REVISITED
  - Respondents' privacy (*passive* data supplier)
  - Holder's (or owner's) privacy
  - User's (*active*) privacy

  $\Rightarrow$ Respondents' and holder's privacy implemented by holder.
  Different focus. Respondents are worried on their individual record, companies are worried on general inferences (e.g. to be used by competitors). E.g., protection of Ebenezer Scrooge's data

  (E. Scrooge | misanthropic, tightfisted, money addict)

  The hospital may be interested on hiding the number of addiction relapses.
  $\Rightarrow$ User's privacy implemented by the user

Data Privacy

Classification 1: On whose privacy is being sought

- Respondents' privacy (*passive* data subject)
  - (Ex. 3.1) Researcher cannot find an individual in the hospital data, cannot learn about an illness of a friend.
  - (Ex. 3.2) Employees in the software company don't learn anything from the dataset used for testing.
- Holder's privacy (or controller)
  - (Ex. 3.4) One supermarket cannot link a record with another one in the other database that *belongs* to the same customer. One supermarket cannot infer information for its economical advantage.
- User privacy (*active* data subject)
  - (Ex. 3.1) The hospital cannot learn that a researcher is studying the number of *failures* of Doctor Hide.

Dimensions. 2nd classification

On the computations to be done

Dimensions: 2nd

- **Ex. 3.6.** Aitana, the director of hospital A, contacts Beatriu, the director of hospital B. She proposes to compute a linear regression model to estimate the number of days patients stay in hospital using their databases.

Dimensions: 2nd

- **Ex. 3.7.** Elia, a researcher on epidemiology, has contacted Aitana the director of a hospital chain. She wants to access the database because she studies flu and she wants to compare how the illness spreads every year in Chicago and in Miami.

Dimensions: 2nd

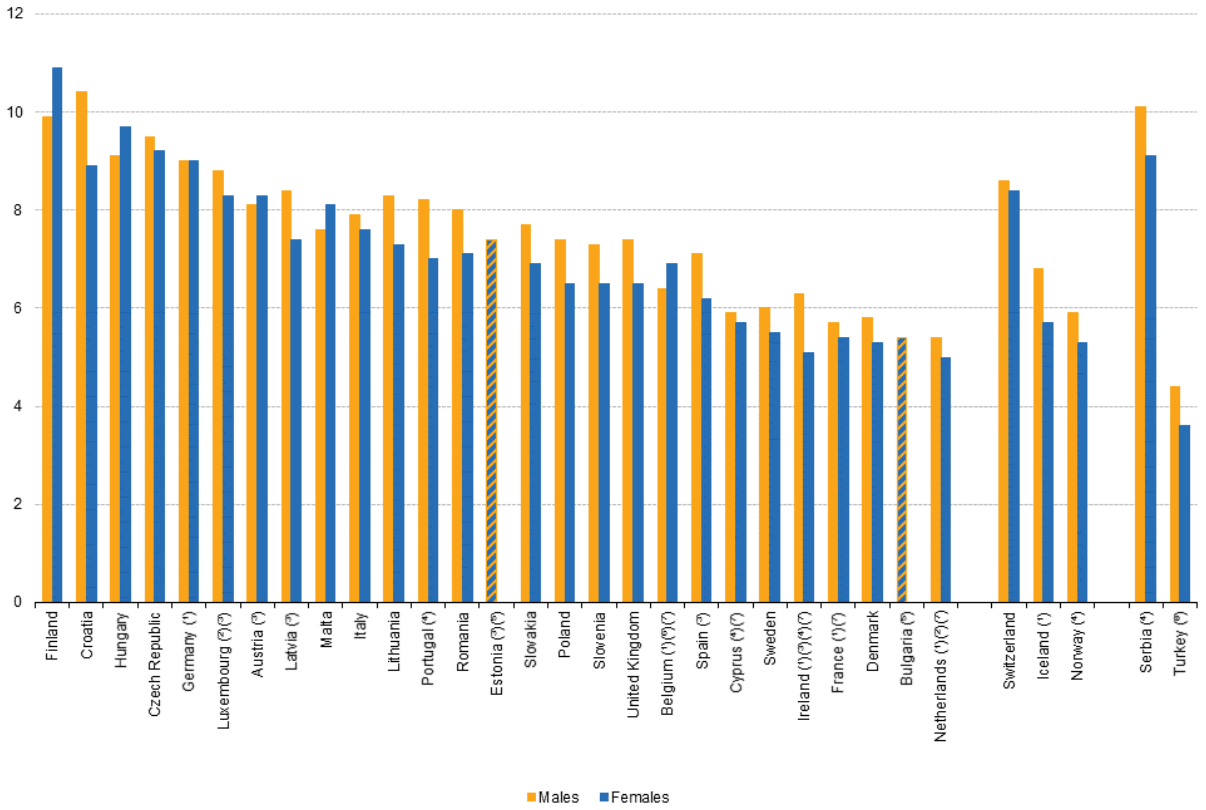
- **Ex. 3.8.** A retailer specialized in baby goods publishes a database with the information gathered from customers with their fidelity card. This database is to be used by a data miner to extract some association rules[1]. The retailer is very much concerned about alcohol consumption and wants to avoid the data miner inferring rules about baby diapers and beers[2]

[1] Association rules. Rules of the form, if someone buys A, B, C also buys D, E, F

[2] A classic example in the literature of association rule mining is about the discovery of a rule stating that men that buy diapers also buy beers (see e.g. [6]).

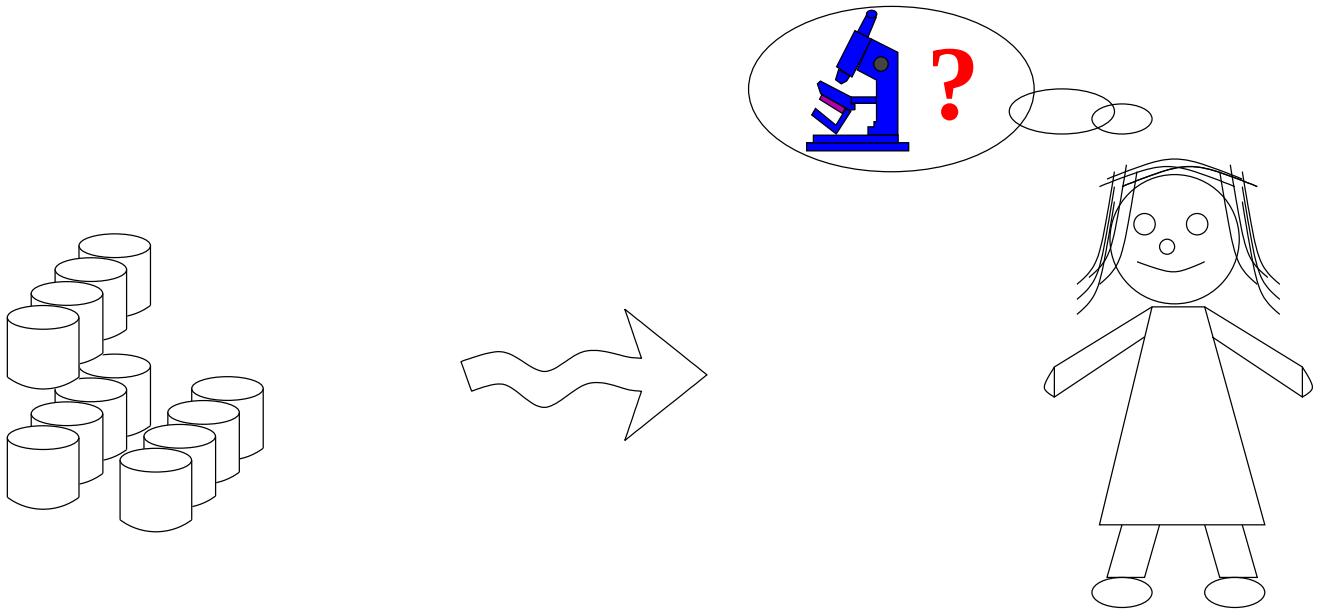
Dimensions: 2nd

- **Dimension 2.** Knowledge on the analysis to be done
  - Full knowledge. Average length of stay for hospital in-patient
  - Partial or null knowledge. A model for mortgage risk prediction (but we do not know what kind of model will be used)



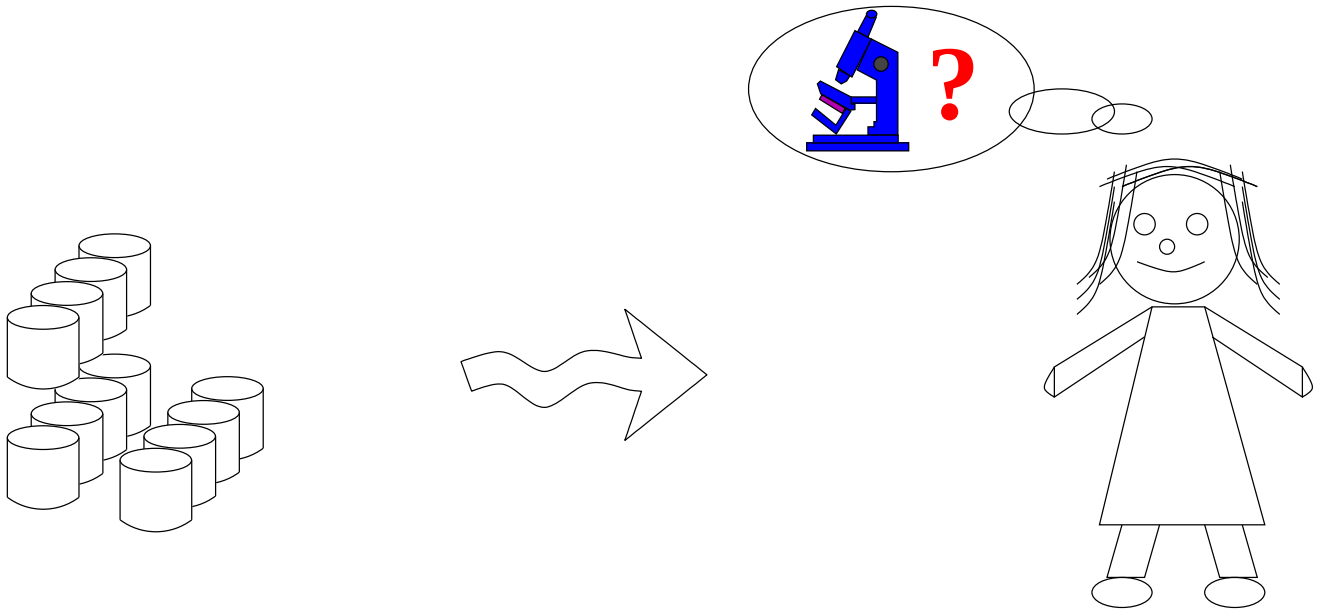
Dimensions: 2nd

- **Dimension 2.** Knowledge on the analysis to be done
  - Data-driven or general purpose (*analysis not known*)
    → Model for mortgage risk prediction, Ex.3.7. Illness spreads,
  - Computation-driven or specific purpose (*analysis known*)
    → Mean length stay, Ex.3.6. Linear regression
  - Result-driven (*analysis known: protection of its results*)
    → Ex.3.8. No rules: baby diapers ⇒ beers



Dimensions: 2nd

- **Dimension 2.** Knowledge on the analysis to be done
  - Data-driven or general purpose (*analysis not known*)
    $\rightarrow$ anonymization methods / masking methods
  - Computation-driven or specific purpose (*analysis known*)
    $\rightarrow$ cryptographic protocols, differential privacy
  - Result-driven (*analysis known: protection of its results*)
    $\rightarrow$ result-driven approaches (tailored masking methods)



Dimensions. 3rd classification

On the number of data sources

Dimensions: 3rd

- **Dimension 3.** Number of data sources
  - Single data source. (single owner)
  - Multiple data sources. (multiple owners)

Other dimensions / classifications

Dimensions: Other

- **Knowledge intensive data privacy** (Ch. 3.1.4)
  (non-numerical data, categories and terms with semantics)
  - Semantic of terms.
    - Recoding cities by counties, recoding of occupations[3], ontologies and diccionaries[4].
    - Use in data protection, information loss, and risk assessment
  - Metadata and constrained data.
    - Relationships between variables (e.g. constraints on variables) and/or records (members in the same household, queries from the same person or computer).
  - Knowledge rich disclosure risk assessment.
    - NoSQL and free text in blogs and online social networks may be used for intruders
    - Models (machine and statistical learning models) can be learnt from data and used to infer sensitive attributes of a respondent and/or to infer attributes that can be used to reidentify a respondent.

[3] https://en.wikipedia.org/wiki/International_Standard_Classification_of_Occupations
[4] https://wordnet.princeton.edu/

Dimensions: Other

- **Cryptographic vs. masking methods**
  - **Cryptographic approach.**
    - ▷ Secure multiparty computation. 100% privacy, 100% accuracy, no flexibility and high computational cost.
    - ▷ Homomorphic encryption. High computational cost.
    - ▷ Using a certain privacy model (e.g., homomorphic encryption) does not imply that the result is *safe* (does not lead to disclosure): combination of several models.
  - **Masking methods.** Trade-off between privacy and accuracy, low computational cost.

Dimensions: Other

- **Semantic and syntactic methods**
  - The literature on security distinguishes between perfect secrecy and semantic security. Perfect secrecy ensures that a ciphertext does not provide any information without knowing the key. In semantic security it is possible in theory to get information from the ciphertext, but it is not computationally feasible to get this information.
  - k-Anonymity is considered a syntactic method, differential privacy is considered an algorithmic and semantic method. Computational anonymity (Sec. 5.8.3) also a semantic method, although based on k-anonymity.
  - **Note:** Semantic not as in ontologies.

Data Privacy

Respondent and owner privacy

	Data–driven (general–purpose)	Computation–driven (specific–purpose)	Result–driven
Number of sources: Single data source			
Number of sources: Multiple data sources			

User privacy

Protecting the identity of the user	Protecting the data generated by the activity of the user

Roadmap of data protection methods

Dimensions: Other

- Respondent and holder privacy
  - **Data-driven methods from a single or from multiple databases**.
    - ▷ Masking methods (Ch. 3.3 & Ch. 6)
    - ▷ Information loss (Ch. 7) and disclosure risk measures (Ch. 5)
  - **Computation-driven methods with several data sources**.
    - ▷ Typically holder privacy. Cryptographic approaches (Ch. 3.4.2)
  - **Computation-driven methods for a single database release**.
    - ▷ Query a database. Differential privacy (Sec. 3.4.1).
    - ▷ Ill-defined case: Masking methods (Ch. 3.3 & Ch. 6)
  - **Result-driven methods**.
    - ▷ Holder privacy. Also to avoid discriminatory knowledge inferred from databases. (Sec. 3.5).

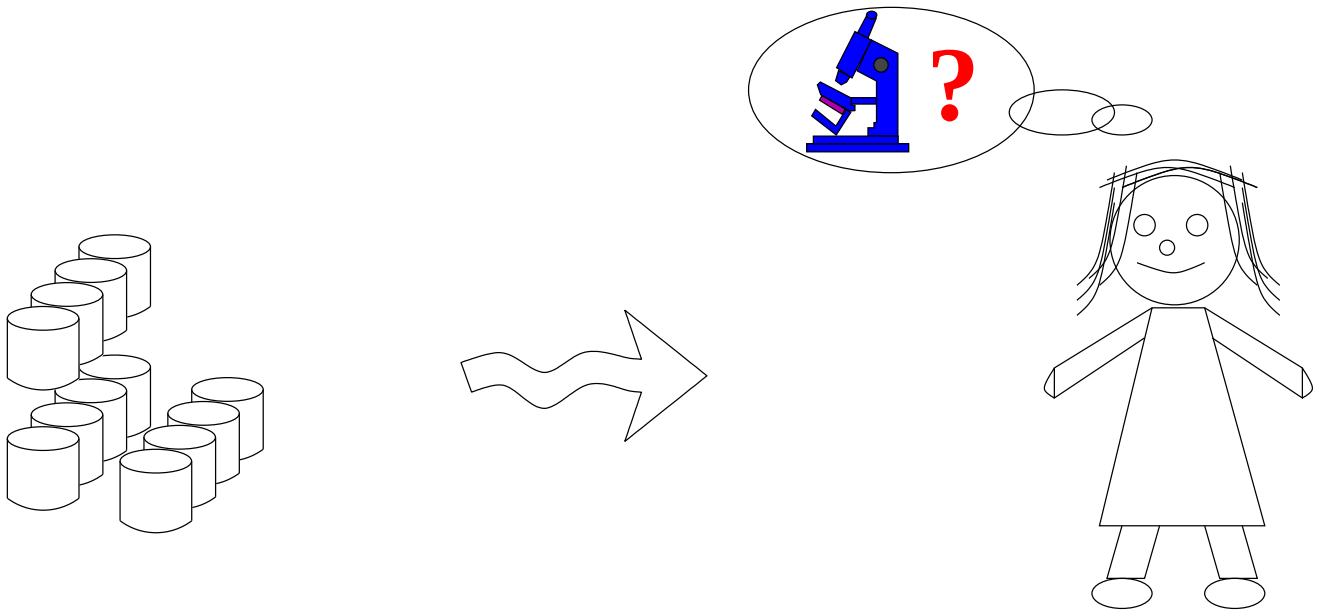
Masking methods

Classification w.r.t. our knowledge on the computation of a third party

- Data-driven or general purpose (*analysis not known*)
  $\rightarrow$ anonymization methods / masking methods

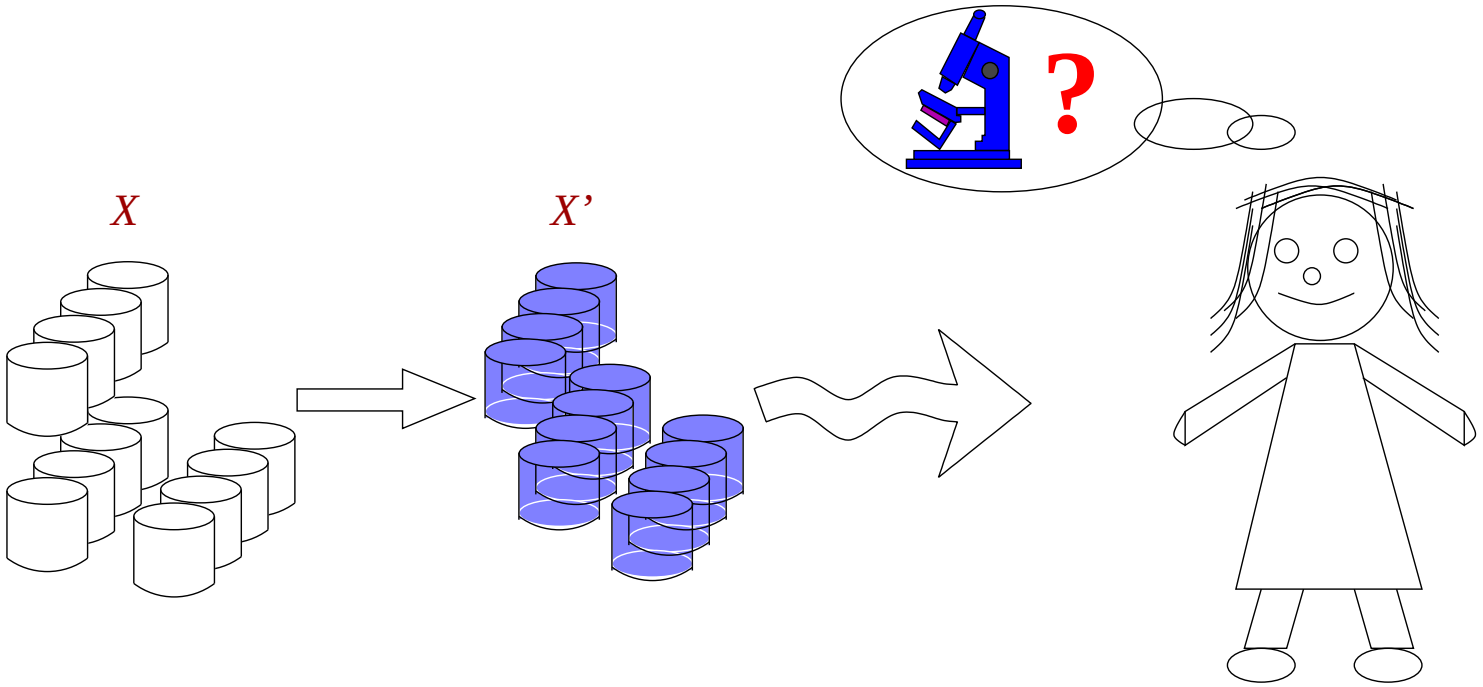
GDPR: (Article 4)

- Pseudonymisation: the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;



Masking methods

Anonymization/masking method: Given a data file X compute a file X' with data of *less quality*.



Masking methods: questions