

Classification of procedures

Vicenç Torra

January, 2018

SAIL + PICS, School of Informatics, University of Skövde, Sweden

Outline

1. Dimensions
2. 1st dimension
3. 2nd dimension
4. 3rd dimension
5. Other dimensions
6. Roadmap of data protection methods

Dimensions

Data Privacy: Dimensions

Data Privacy

- 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)
- Holder's privacy (or owner's)
- User's privacy (*active*)

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.
- **Actors:** Database of patients
 - 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.
- **Actors:** Database of patients
 - Holder: the hospital
 - Respondents: the patients
- **Actors:** Database of queries
 - 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.
- **Actors:**
 - Holder: The insurance company
 - Respondent: Customers

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.
- **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

⇒ 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.

⇒ User's privacy implemented by the user

Data Privacy

Classification 1: On whose privacy is being sought

- Respondents' privacy (*passive* data supplier)
 - (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 owner's)
 - (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*)
 - (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.

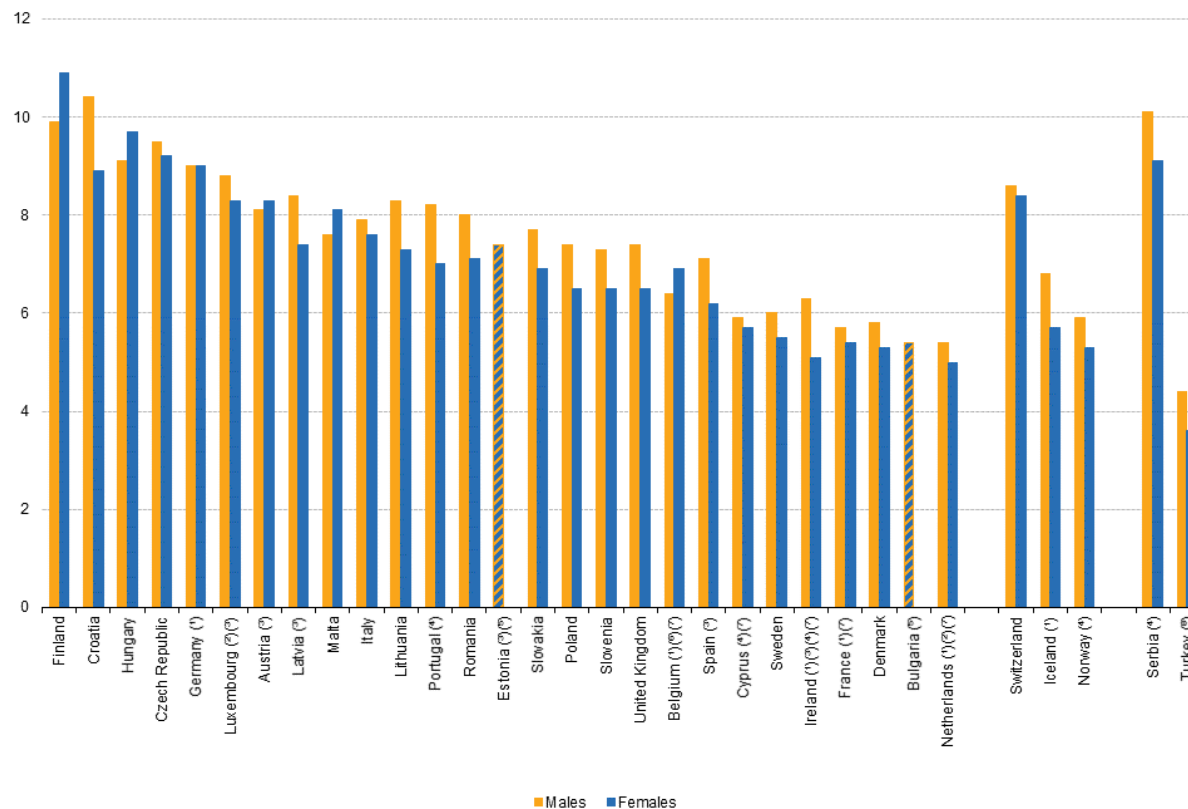
Dimensions: 2nd

- **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. The retailer is very much concerned about alcohol consumption and wants to avoid the data miner inferring rules about baby diapers and beers¹

¹A 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]).

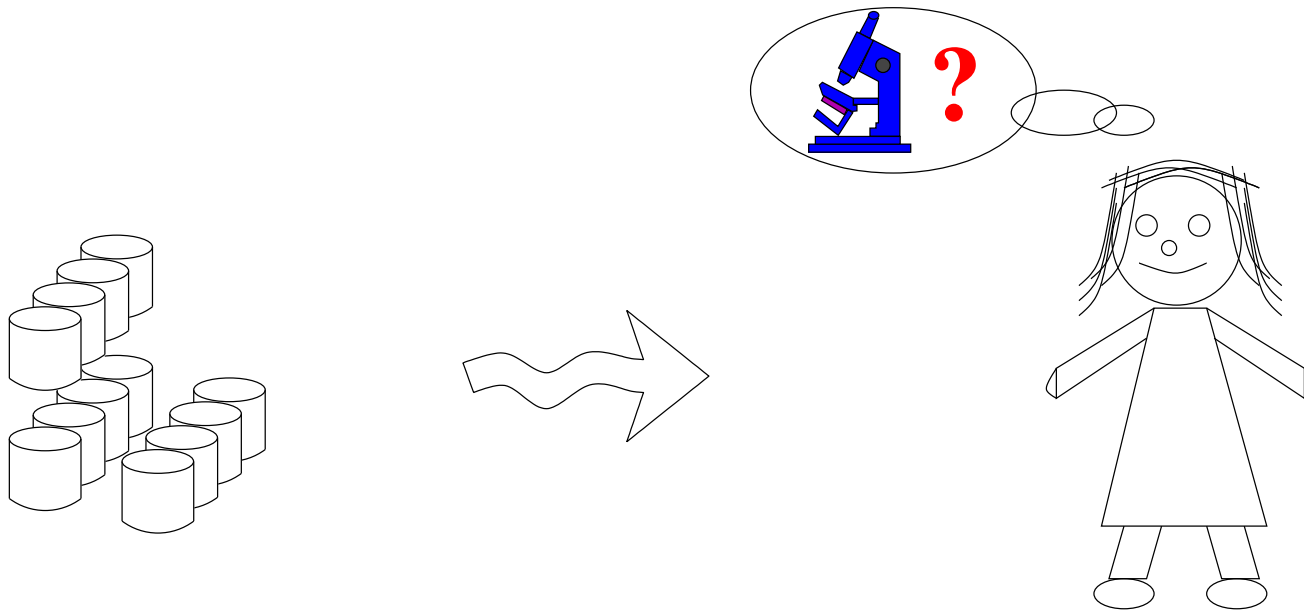
Dimensions: 2nd

- **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*)
 - anonymization methods / masking methods
 - Computation-driven or specific purpose (*analysis known*)
 - cryptographic protocols, differential privacy
 - Result-driven (*analysis known: protection of its results*)



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)
 - Semantic of terms
 - Metadata and constrained data
 - Knowledge rich disclosure risk assessment

Dimensions: Other

- Cryptographic vs. masking methods
 - **Cryptographic approach.** 100% privacy, 100% accuracy, no flexibility and high computational cost.
 - **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		
	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

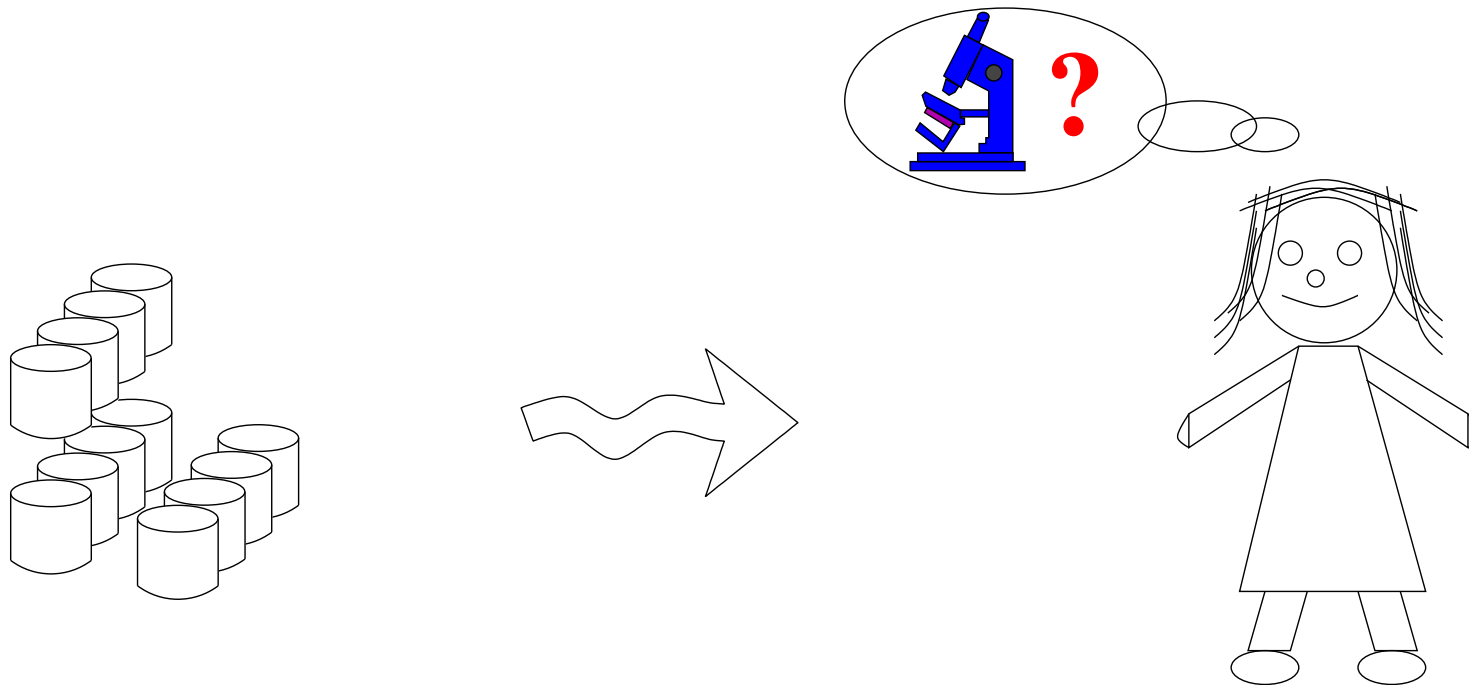
Dimensions: Other

- 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

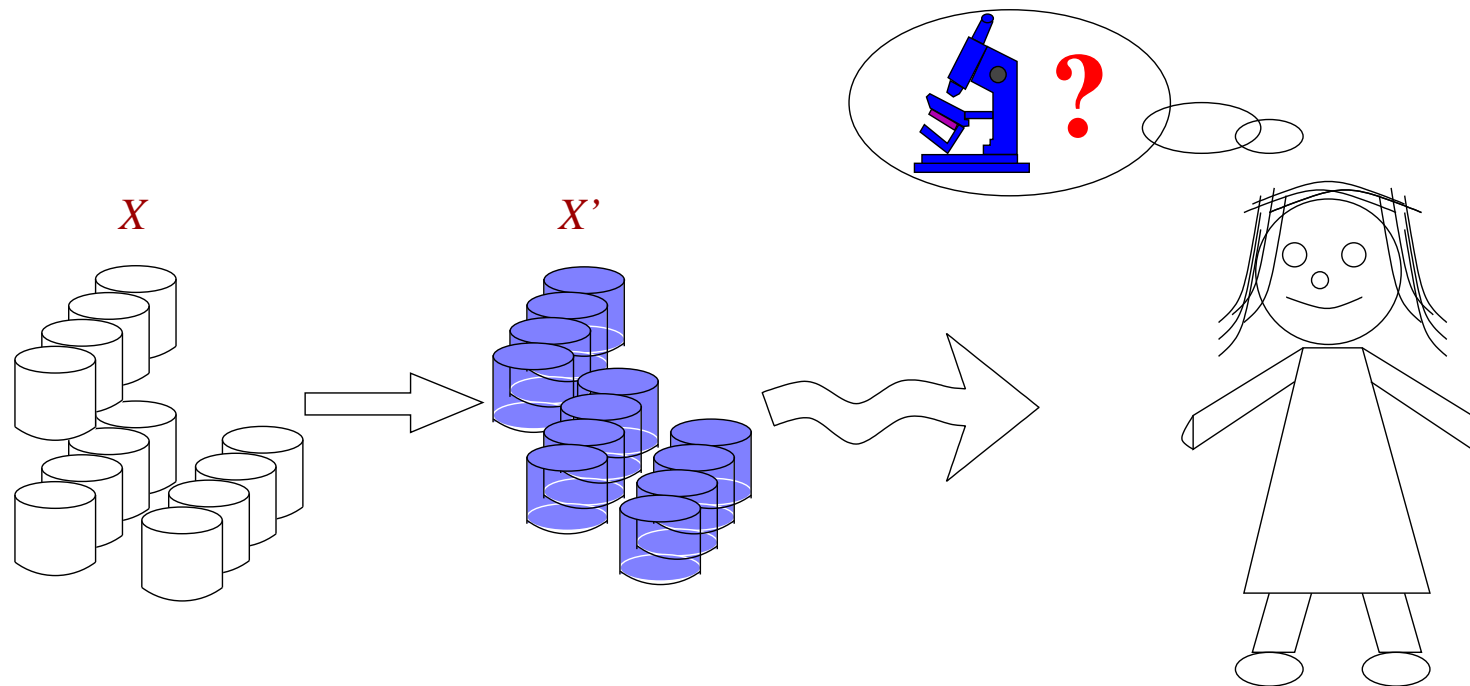
Classification w.r.t. our knowledge on the computation of a third party

- Data-driven or general purpose (*analysis not known*)
→ anonymization methods / masking methods \times s



Masking methods

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



Masking methods: questions

