Skövde 2017

Data privacy: an introduction (part II)

Vicenç Torra

February, 2017

School of Informatics, University of Skövde, Sweden

Outline

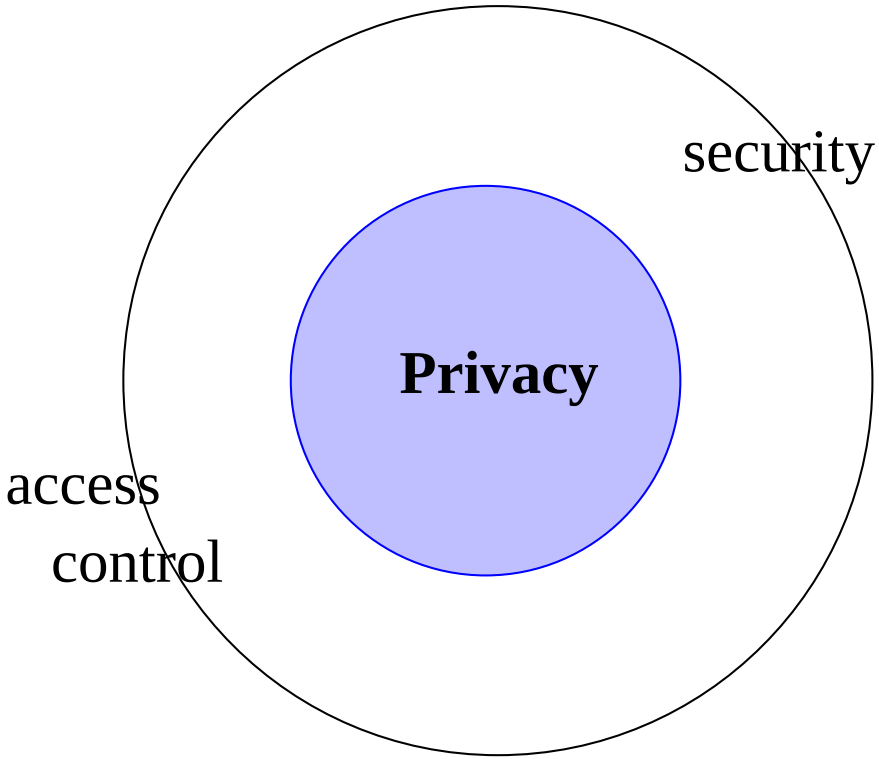
Basics

Introduction

- Data privacy (technological / computer science perspective)
  - Avoid the disclosure of sensitive information when processing data.

Introduction

- Data privacy: boundaries
  - Database in a computer or in a removable device
    $\Rightarrow$ access control to avoid unauthorized access
  - Data is transmitted
    $\Rightarrow$ security technology to avoid unauthorized access



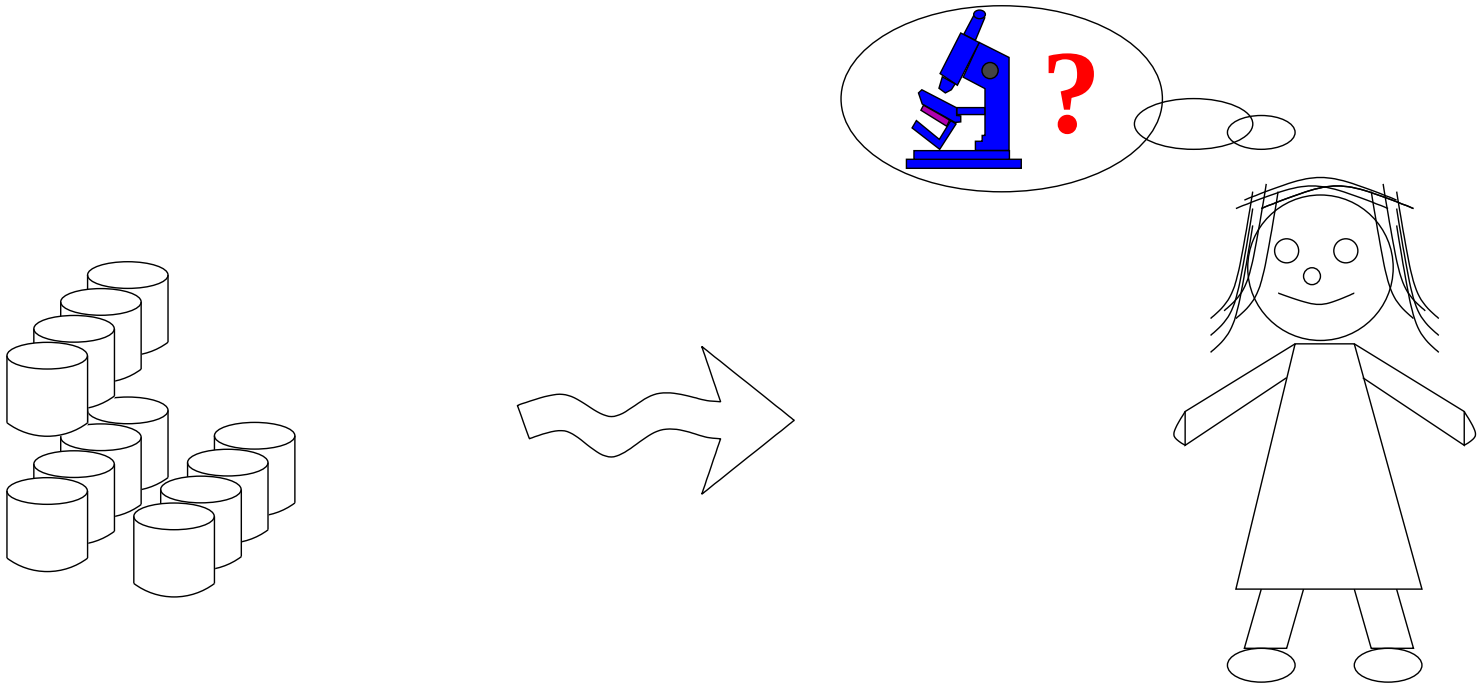
Introduction

- Data privacy: boundaries
  - Database in a computer or in a removable device
    $\Rightarrow$ access control to avoid unauthorized access
  - Data is transmitted
    $\Rightarrow$ security technology to avoid unauthorized access

- Data privacy: core
  - Data is/needs to be processed:
    $\Rightarrow$ statistics, data mining, machine learning
    $\Rightarrow$ compute indices, find patterns, build models
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.

Introduction

- Data privacy: core
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.



Difficulties

- Difficulties: Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston

  Names, Age, Sex, Occupation, Place of birth, Last place of residence, Yes/No, condition (healthy?)

LIST OF PASSENGERS.

REPORT AND LIST OF PASSENGERS taken on board the ... at London.

NAMES	AGE	SEX	OCCUPATION	PLACE OF BIRTH	Last Place of Residence	U.S.		CONDITION
[illegible]	18	Male	[illegible]	[illegible]	[illegible]	yes		[illegible]
[illegible]	20	-	[illegible]	[illegible]	[illegible]		no	-
[illegible]	24	-	Tailor	London	London		no	-
[illegible]	36	-	[illegible]	London	London		no	-
[illegible]	25	-	[illegible]	[illegible]	[illegible]	yes		-
[illegible]	19	-	-	[illegible]	[illegible]	yes		-
[illegible]	16	-	-	London (?)	London	yes		-
[illegible]	40	-	-	[illegible]	[illegible]	yes		-
[illegible]	25	-	-	[illegible]	[illegible]	yes		-
[illegible]	25	-	[illegible]	[illegible]	[illegible]	yes		-

Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - ⋆ 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.
  - Data from mobile devices:
    - ⋆ two positions can make you unique (home and working place)
  - AOL and Netflix cases (search logs and movie ratings)
  - Similar with credit card payments, shopping carts, search logs, ... (i.e., high dimensional data)

Difficulties

- Data privacy is "impossible", or not ?
  - Privacy vs. utility
  - Privacy vs. security
  - Computationally feasible

A classification – Dimensions

Dimensions: 1st

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

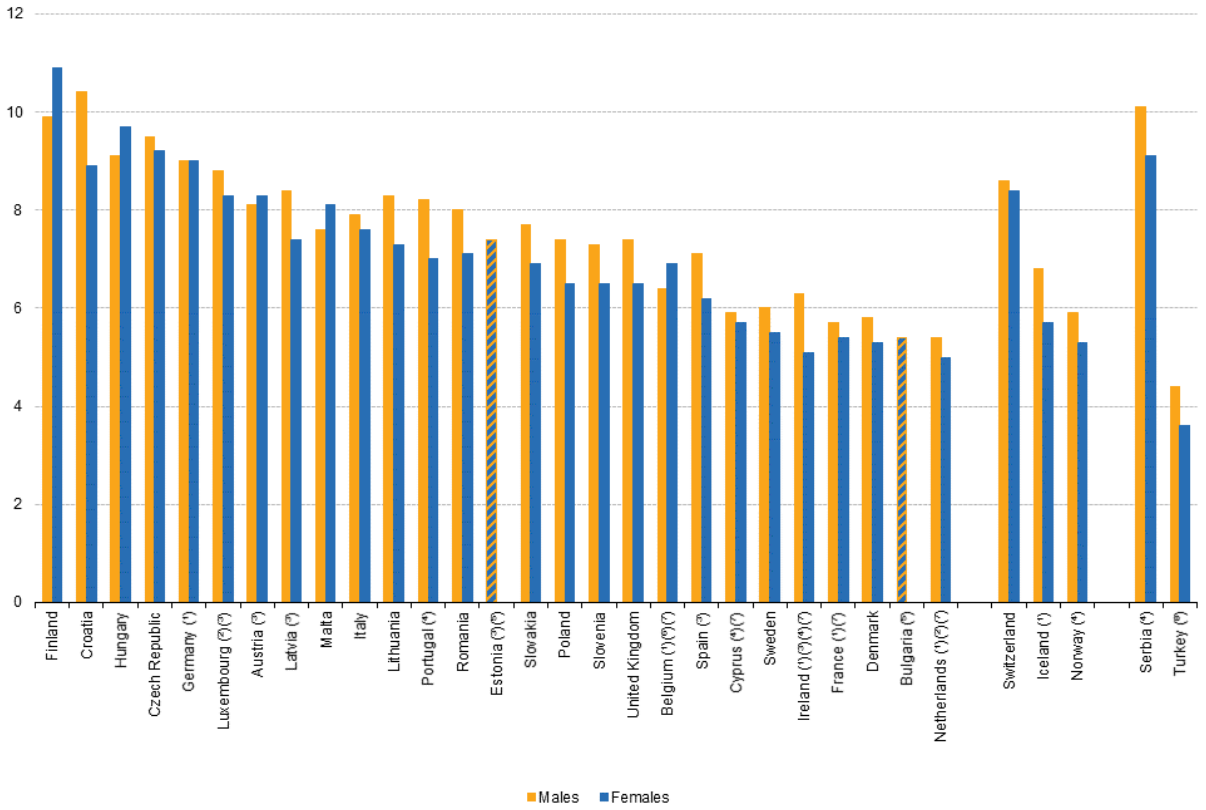
  $\Rightarrow$ Respondents' and holder's privacy implemented by holder.
  Different focus. Respondents are worried on their individual record, companies are worried on general inferences (e.g. to be used by competitors). E.g., protection of Ebenezer Scrooge's data

  (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

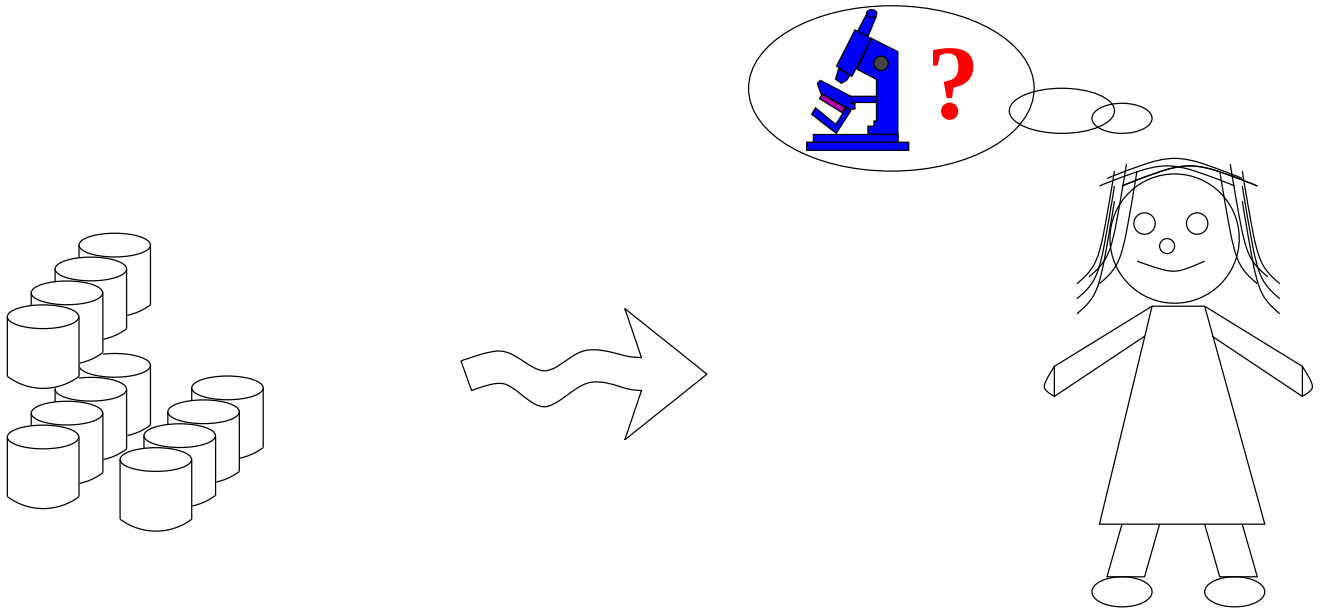
Dimensions: 2nd

- **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

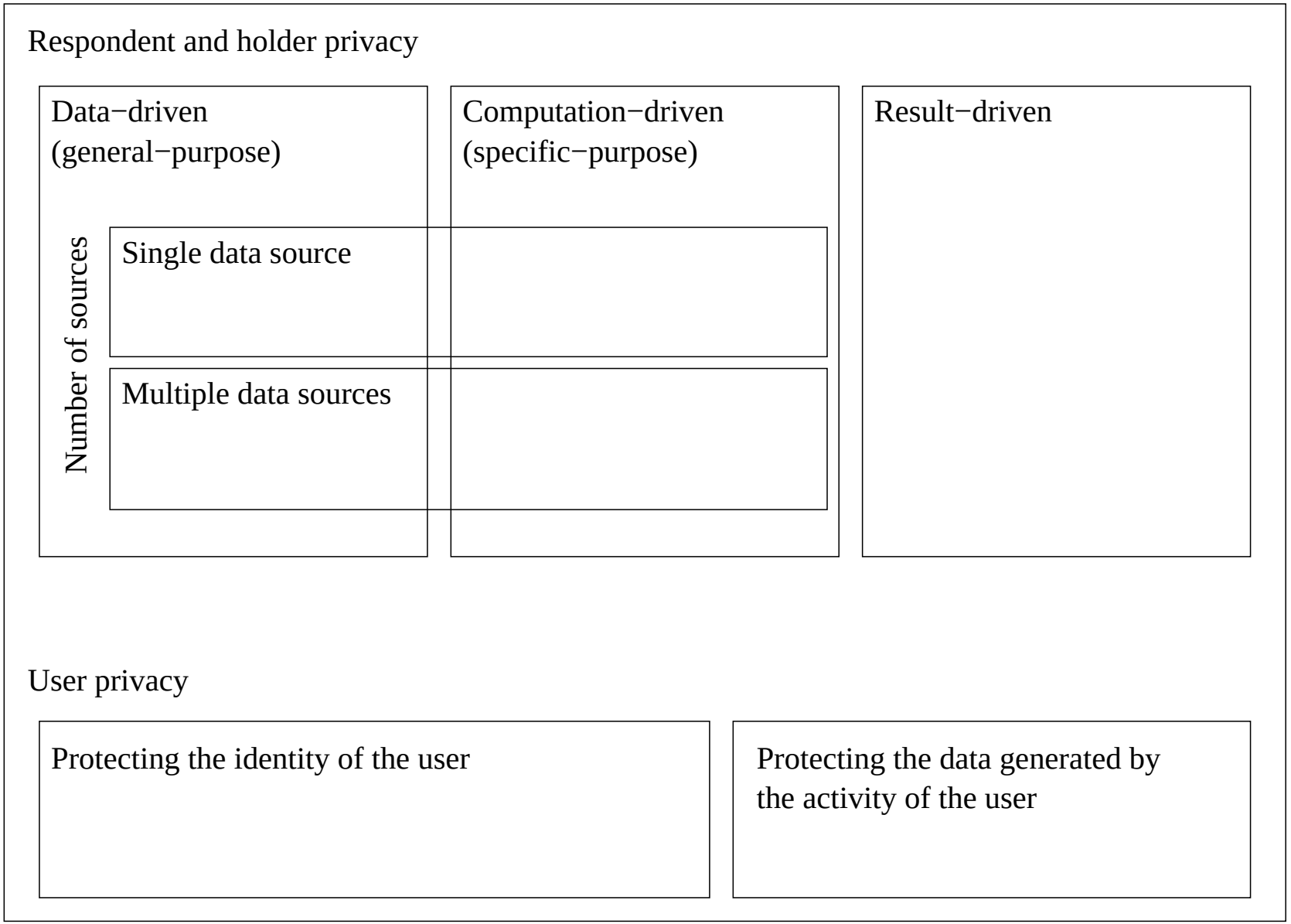
- **Dimension 2.** Knowledge on the analysis to be done
  - Data-driven or general purpose (*analysis not known*)
  - Computation-driven or specific purpose (*analysis known*)
  - Result-driven (*analysis known: protection of its results*)



Dimensions: 3rd

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

1st - 3rd Dimensions: Summary

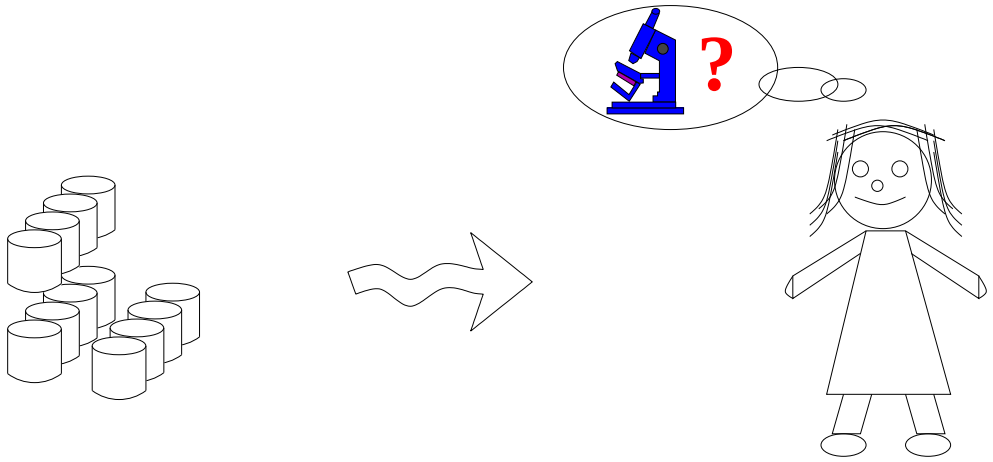


Masking methods

Masking methods

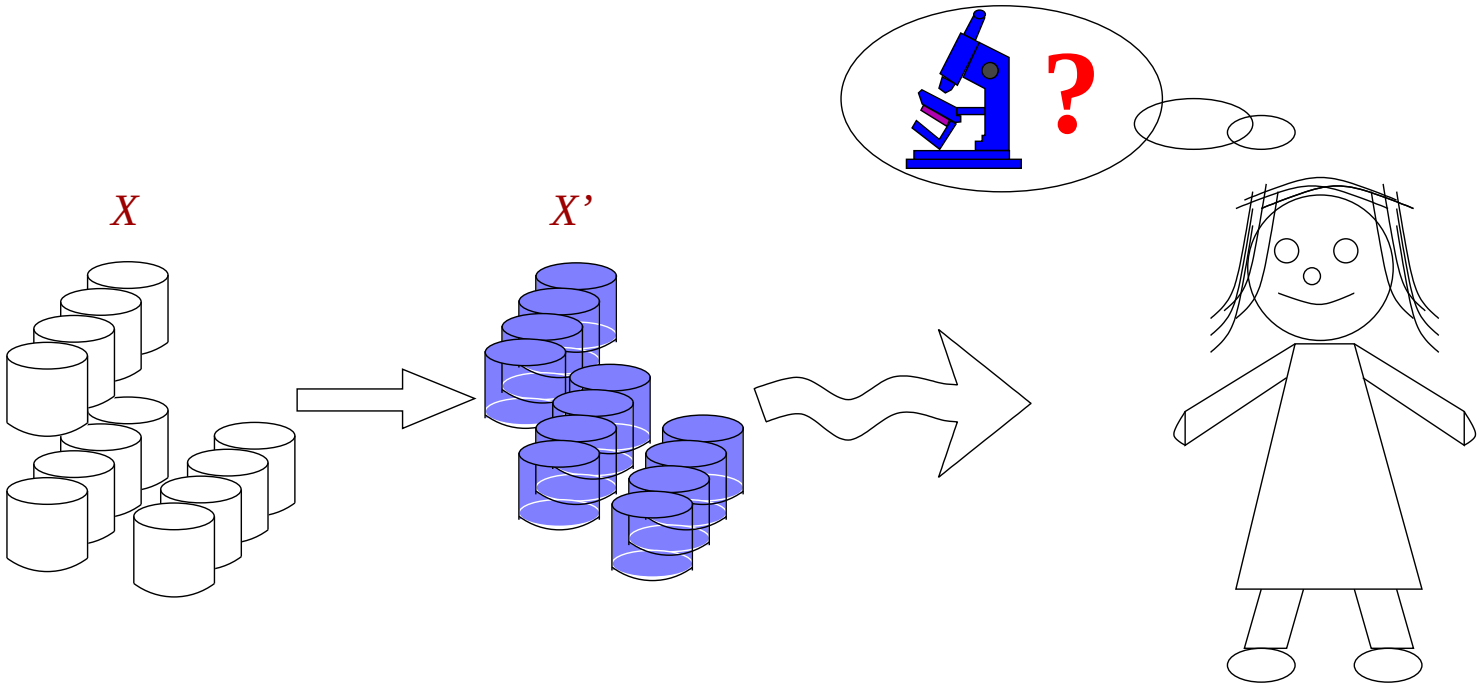
Respondent and holder privacy. Acc. to knowledge on the analysis

- Data-driven or general purpose (*analysis not known*)
  → masking methods / anonymization methods (one data source)
- Computation-driven or specific purpose (*analysis known*)
  → cryptographic protocols (multiple data sources)
  → masking methods (single data source, differential privacy)
- Result-driven (*analysis known: protection of its results*)
  → masking methods (one data source, Holder's privacy)



Masking methods

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



Masking methods

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

- Original X

Respondent	City	Age	Illness
ABD	Skövde	28	Cancer
COL	Mariestad	31	Cancer
GHE	Stockholm	62	AIDS
CIO	Stockholm	64	AIDS
HYU	Göteborg	58	Heart attack

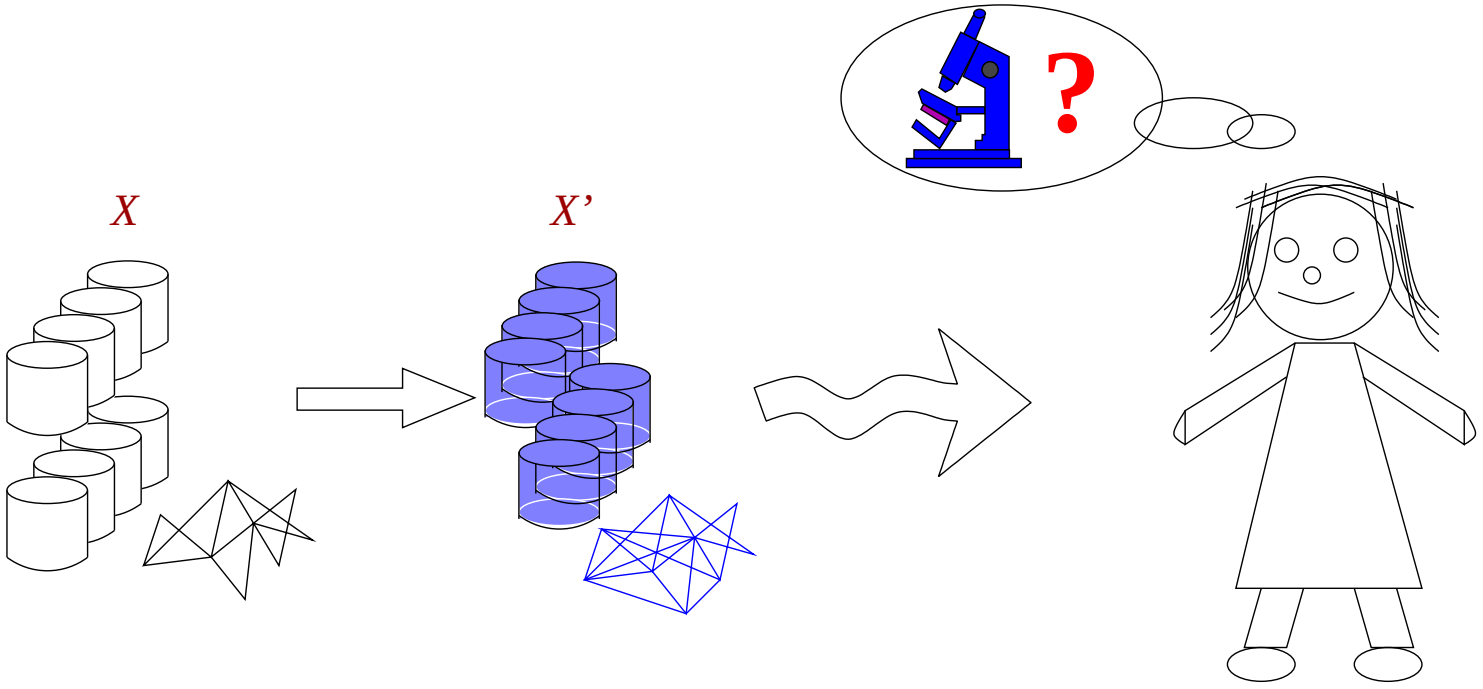
- Protected X'

Respondent	City	Age	Illness
ABD	Skövde or Mariestad	30	Cancer
COL	Skövde or Mariestad	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS

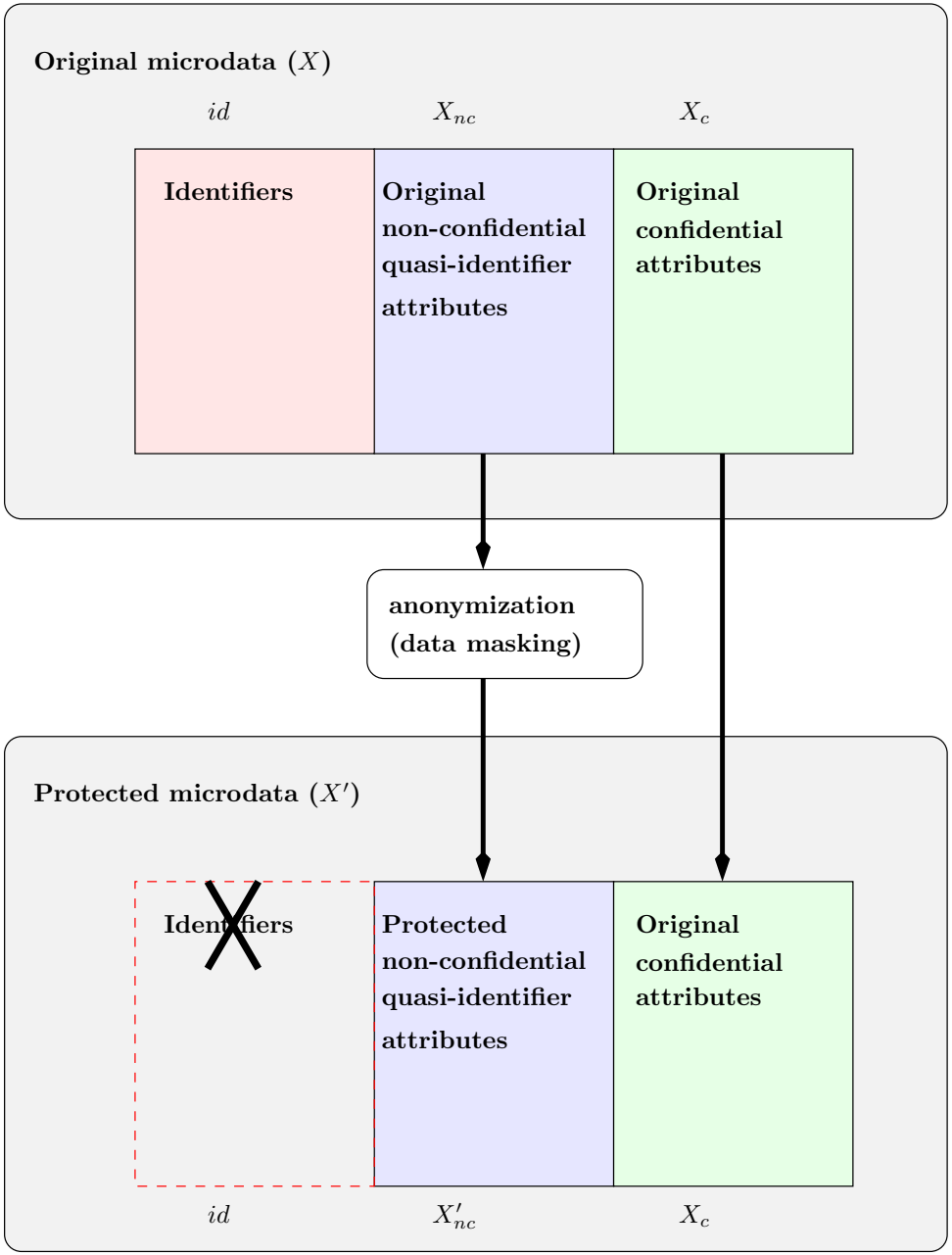
Masking methods

Approach valid for different types of data

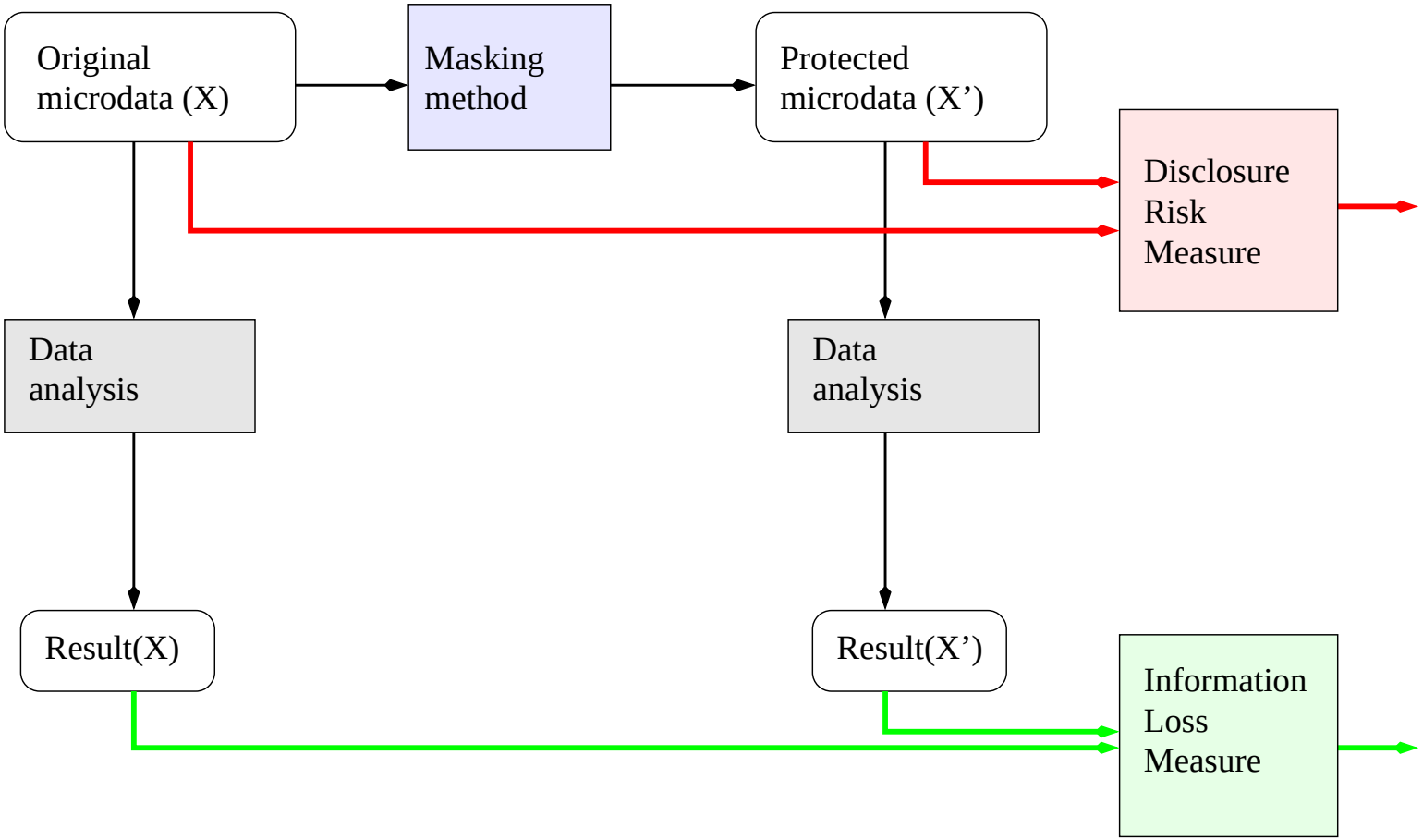
- Databases, documents, search logs, social networks, . . .
  (also masking taking into account semantics: wordnet, ODP)



Masking methods



Research questions



Masking methods

Masking methods. (anonymization methods)

Masking methods

Masking methods. (anonymization methods)

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping

Masking methods

Masking methods. (anonymization methods)

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression

Masking methods

Masking methods. (anonymization methods)

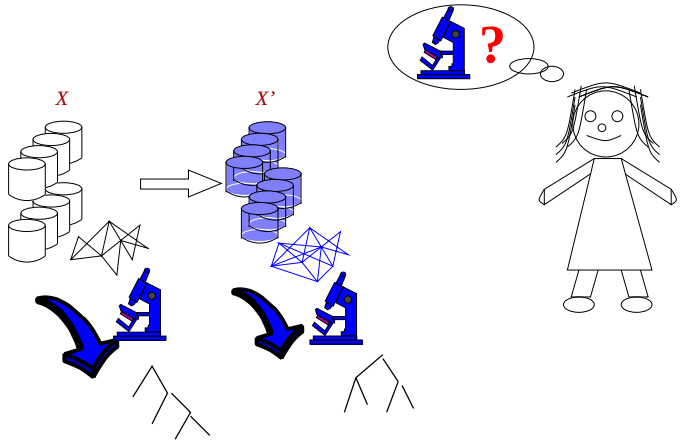
- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression
- Synthetic data generators. (less quality=not real data)
  E.g. (i) model from the data; (ii) generate data from model

Masking methods

Information loss measures. Compare X and X' w.r.t. analysis (f)

$$IL_f(X, X') = divergence(f(X), f(X'))$$

- f: generic vs. specific (data uses)
  - Statistics
  - Machine learning: Clustering and classification
    For example, classification using decision trees
  - ... specific measures for graphs



Privacy models and disclosure risk assessment

Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
  - Attribute disclosure: (e.g. learn about Alice's salary)
    - ⋆ Increase knowledge about an attribute of an individual
  - Identity disclosure: (e.g. find Alice in the database)
    - ⋆ Find/identify an individual in a masked file

Within machine learning, some attribute disclosure is expected.

Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures

Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  (minimize information loss vs. multiobjetive optimization)

Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  (minimize information loss vs. multiobjetive optimization)

Examples. Privacy models / disclosure risk measures

	Attribute disclosure	Identity disclosure
Boolean	Differential privacy	k−Anonymity
Quantitative	Interval disclosure	Re−identification (record linkage) Uniqueness

Thank you

References

Related references.

- Torra, V. (2017) Data privacy: Foundations, new developments, and the big data challenge, Springer, forthcomming.

- Aggarwal, C. C., Yu, P. S. (2008) (eds.) Privacy-Preserving Data Mining: Models and Algorithms, Springer.
- Duncan, G. T., Elliot, M., Salazar, J. J. (2011) Statistical confidentiality, Springer.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., de Wolf, P.-P. (2012) Statistical Disclosure Control, Wiley.
- Navarro-Arribas, G., Torra, V. (2015) (eds.) Advanced Research in Data Privacy, Springer.
- Vaidya, J., Clifton, C. W., Zhu, Y. M. (2006) Privacy Preserving Data Mining, Springer.