**Oslo, 2018**

# Data privacy: introduction

Vicenç Torra

January 15, 2018

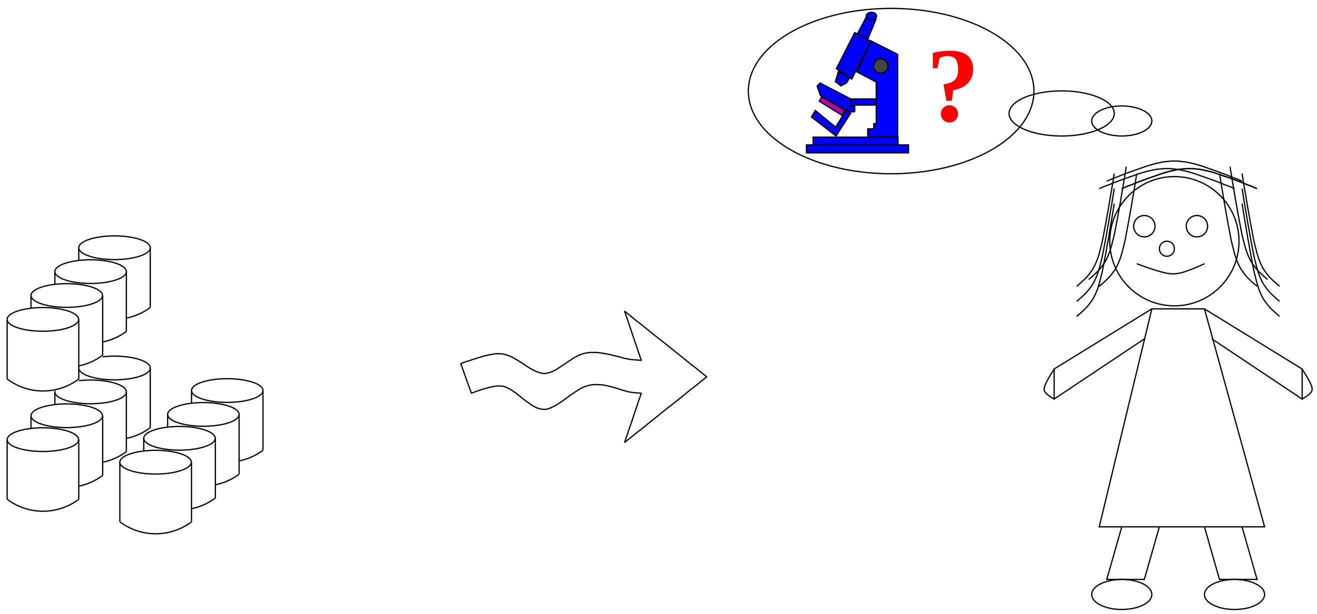
Privacy, Information and Cyber-Security Center
SAIL, School of Informatics, University of Skövde, Sweden

# Outline

# Motivation

# Introduction

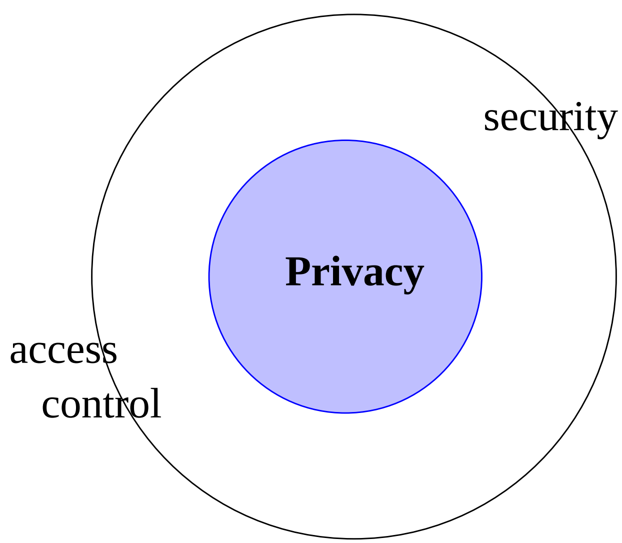
- Data privacy: core
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.




E.g., you are authorized to compute the average stay in a hospital, but maybe you are not authorized to see the length of stay of your neighbor.

# Introduction

- Data privacy: boundaries
  - Database in a computer or in a removable device
    $\Rightarrow$ access control to avoid unauthorized access
    $\Longrightarrow$ Access to address (admissions), Access to blood test (admissions?)
  - Data is transmitted
    $\Rightarrow$ security technology to avoid unauthorized access
    $\Longrightarrow$ Data from blood glucose meter sent to hospital. Network sniffers
    Transmission is sensitive: Near miss/hit report to car manufacturers



# Difficulties

- Difficulties: Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston[1]

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

[1] https://www.sec.state.ma.us/arc/arcgen/genidx.htm

# 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% (9.1 million) had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.

# Difficulties

- Difficulties: highly identifiable data
  - Data from mobile devices:
    - ⋆ two positions can make you unique (home and working place)
  - AOL[2] and Netflix cases (search logs and movie ratings)
    ⇒ User No. 4417749, hundreds of searches over a three-noth period including queries 'landscapers in Lilburn, Ga' ⇒ Thelma Arnold identified!
    ⇒ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...
    (i.e., high dimensional data)

[2] http://www.nytimes.com/2006/08/09/technology/09aol.html

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - University goal: know how sickness is influenced by studies and by commuting distance
    - Data: where students live, what they study, if they got sick
    - No “personal data”, is this ok ?
    - NO!!: How many in your degree live in your town ?
  - Example #2:
    - Car company goal: Study driving behaviour in the morning
    - Data: First drive (GPS origin + destination, time) $\times$ 30 days
    - No “personal data”, is this ok?

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?
  - Example #2:
    - ⋆ Car company goal: Study driving behaviour in the morning
    - ⋆ Data: First drive (GPS origin + destination, time) $\times$ 30 days
    - ⋆ No "personal data", is this ok?
    - ⋆ NO!!!: How many (cars) go from your parking to your university everymorning ? Are you exceeding the speed limit ? Are you visiting a psychiatrisc every tuesday ?

# Difficulties

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

# Privacy models and disclosure risk assessment

# Disclosure risk assessment

**Privacy models:** What is a privacy model ?

- To make a program we need to know what we want to protect

# Disclosure risk assessment

**Disclosure risk.** Disclosure: leakage of information.

- 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 database (e.g., masked file)

Within machine learning, some attribute disclosure is expected.

# Disclosure risk assessment

**Disclosure risk.**

- Boolean vs. quantitative privacy models
  - Boolean: Disclosure either takes place or not. Check whether the definition holds or not. Includes definitions based on a threshold.
  - Quantitative: Disclosure is a matter of degree that can be quantified. Some risk is permitted.
- minimize information loss (max. utility) vs. multiobjetive optimization

# Disclosure risk assessment

**Privacy models.** (selection)

- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.
- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.

# Disclosure risk assessment

## Privacy model. Secure multiparty computation.

- Several parties want to compute a function of their databases, but only sharing the result.
  - hospital $A$ and hospital $B$,
  - two independent databases with:

    `age of patient, length of stay in hospital`
- how to compute a regression with all data (both databases)

  age $\rightarrow$ length

  without sharing data?

# Disclosure risk assessment

## Privacy model. Reidentification privacy.

- Avoid finding a record in a database.
  - hospital $A$ has a database
  - a researcher asks for access to this database
- how to prepare an anonymized database so that the researcher can not find a friend?

# Disclosure risk assessment

## Privacy model. k-Anonymity.

- Avoid finding a record in a database.
  ... making each record indistinguishable with $k-1$ other records.

# Disclosure risk assessment

## Privacy model. k-Anonymity.

- Avoid finding a record in a database.
  ... making each record indistinguishable with $k-1$ other records.
  - hospital $A$ has a database
  - a researcher asks for access to this database
- how to prepare an anonymized database so that the researcher can not find a friend?

# Disclosure risk assessment

## Privacy model. Differential privacy.

- The output of a query to a database should not depend (much) on whether a record is in the database or not.
  - hospital $A$ has a database

    `age of patient, length of stay in hospital`
- how to compute an average length of stay in such a way that the result does not depend (much) on whether we use or not the data of a particular person.

- Privacy models: quite a few *competing* models
  - differential privacy
  - secure multiparty computation
  - k-anonymity
  - computational anonymity
  - reidentification (record linkage)
  - uniqueness
  - result privacy
  - interval disclosure
  - integral privacy

- Privacy models: quite a few *competing* models
  - differential privacy
  - secure multiparty computation
  - k-anonymity
  - computational anonymity
  - reidentification (record linkage)
  - uniqueness
  - result privacy
  - interval disclosure
  - integral privacy

- ... and combined:
  - secure multiparty computation + differential privacy

# Disclosure risk assessment

**Disclosure risk.**

- Function known vs. unknown (ill-defined)
- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures/models

# Disclosure risk assessment

## Disclosure risk.

- Function known vs. unknown (ill-defined)
- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures/models

## Classification of privacy models (and measures)

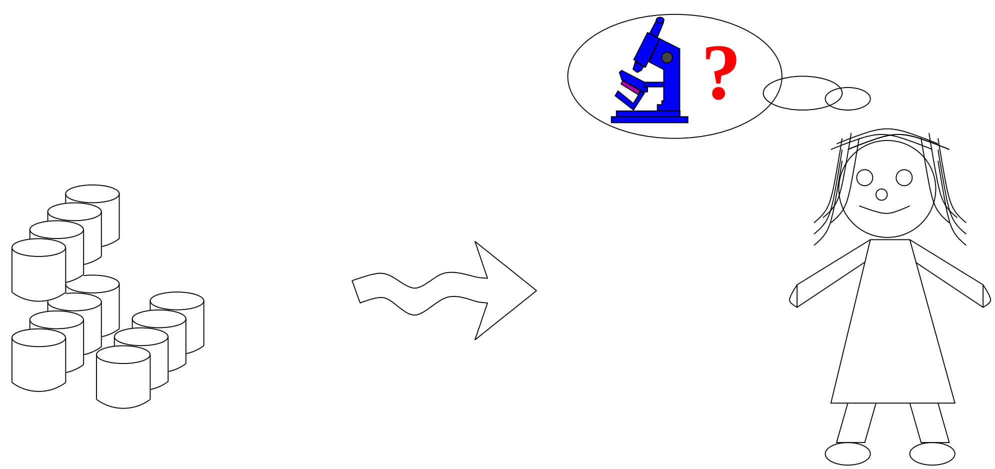
| | Attribute disclosure | Identity disclosure |
|---|---|---|
| Boolean | Differential privacy<br>Result privacy<br>Secure multiparty computation | k−Anonymity<br>Secure multiparty computation |
| Quantitative | Interval disclosure | Re−identification (record linkage)<br>Uniqueness |

# Data protection mechanisms

# Data protection mechanisms

- **Focus** on respondent privacy
- **Classification** w.r.t. knowledge on the computation of a third party
  - 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*)

**Figure.** Basic model (multiple/dynamic databases + multiple *people*)

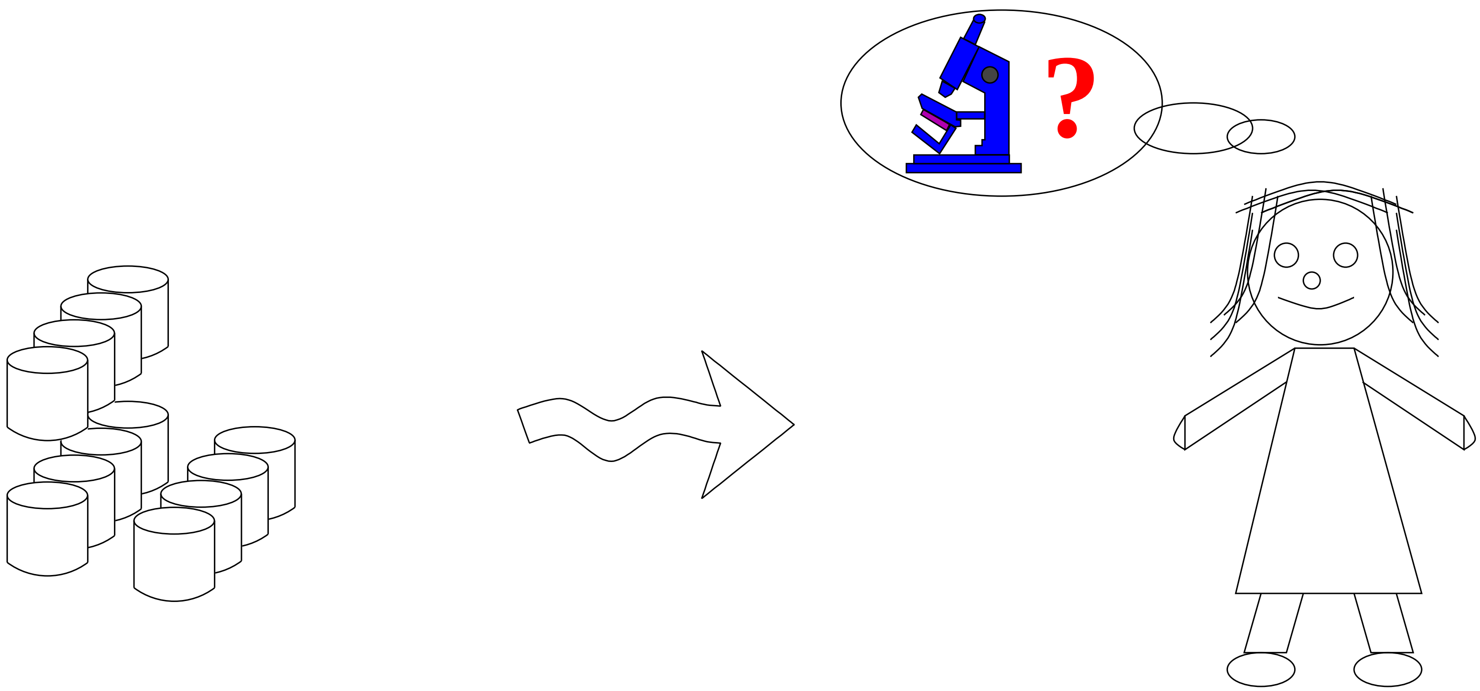


# Masking methods

# Masking methods

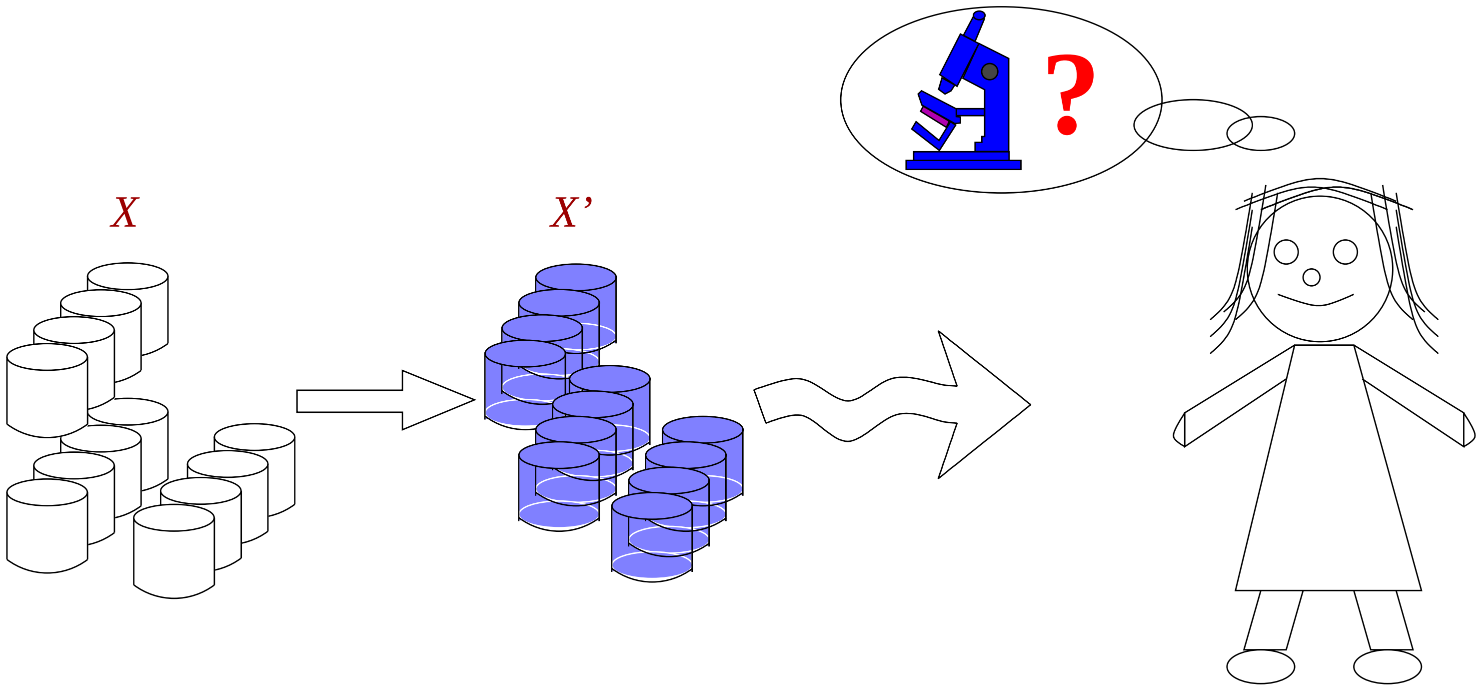
**Classification** w.r.t. our knowledge on the computation of a third party

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

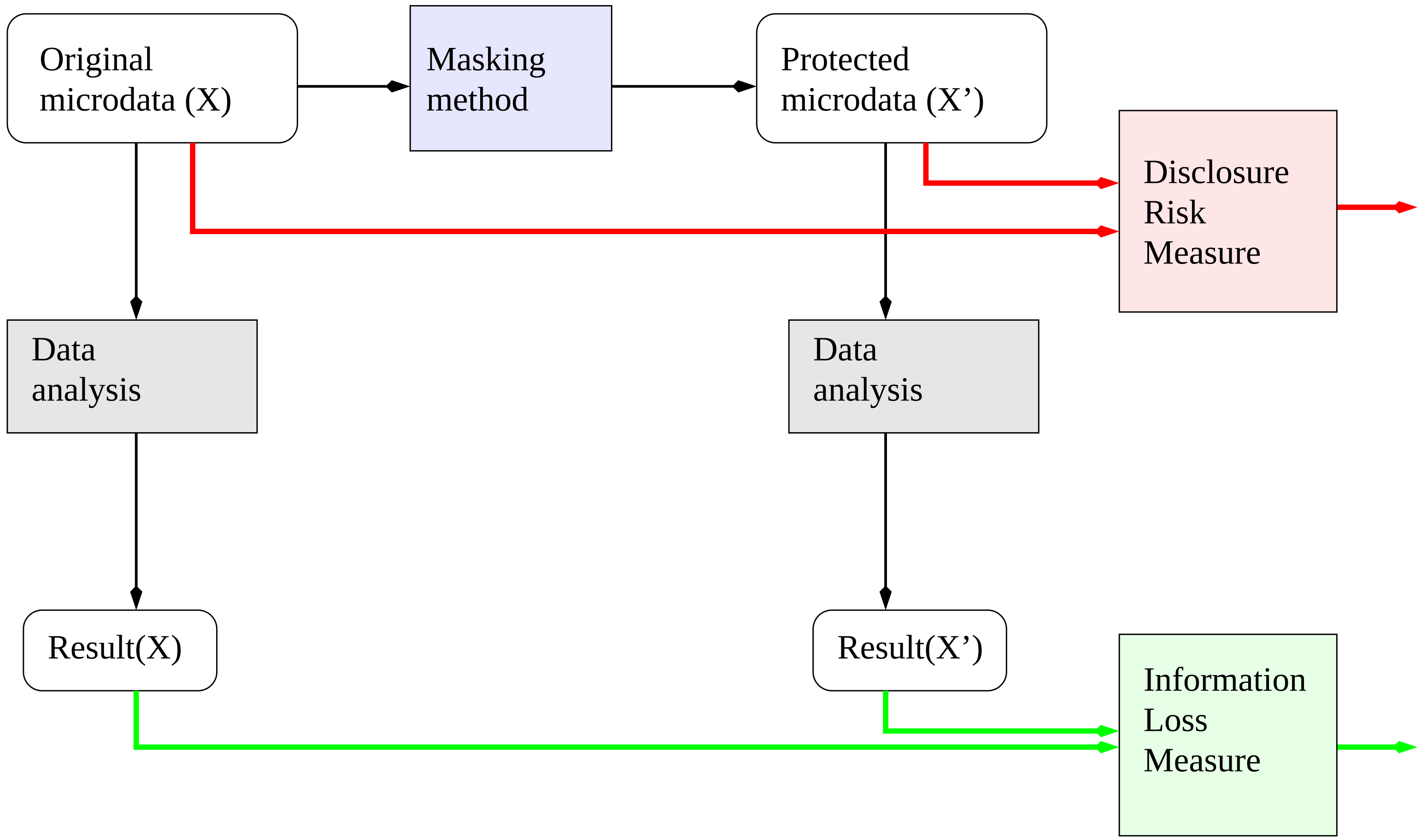


# Masking methods

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



# Masking methods: questions



# Research questions I: Masking methods

**Masking methods** (anonymization methods). Build $X'$ from $X$.

# Research questions I: Masking methods

**Masking methods** (anonymization methods). Build $X'$ from $X$.

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

# Research questions I: Masking methods

**Masking methods** (anonymization methods). Build $X'$ from $X$.

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

# Research questions I: Masking methods

**Masking methods** (anonymization methods). Build $X'$ from $X$.

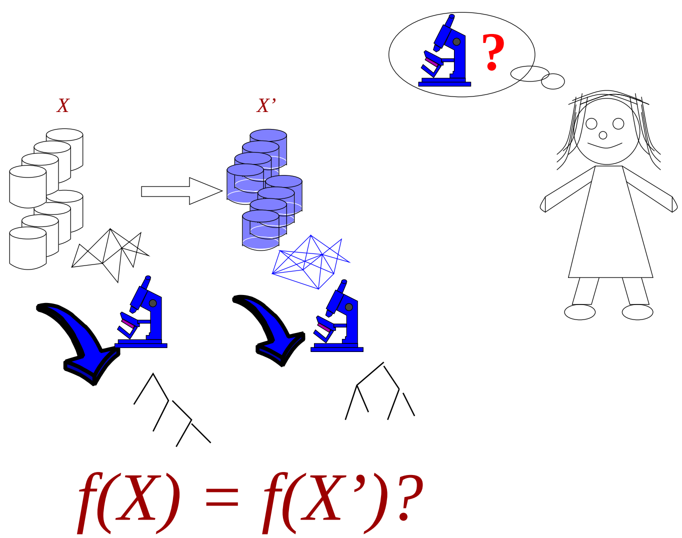
- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression
- Synthetic data generators. (less quality=not real data)
  E.g. (i) model from the data; (ii) generate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

# Research questions II: Information loss/Utility

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

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

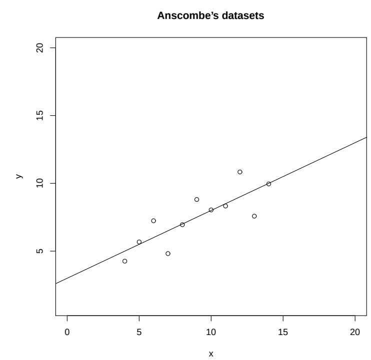
- $f$: generic vs. specific (data uses)
  - Statistics: mean, variance, regression
  - Machine learning: clustering, classification
    For example, classification using decision trees
  - ... specific measures for graphs



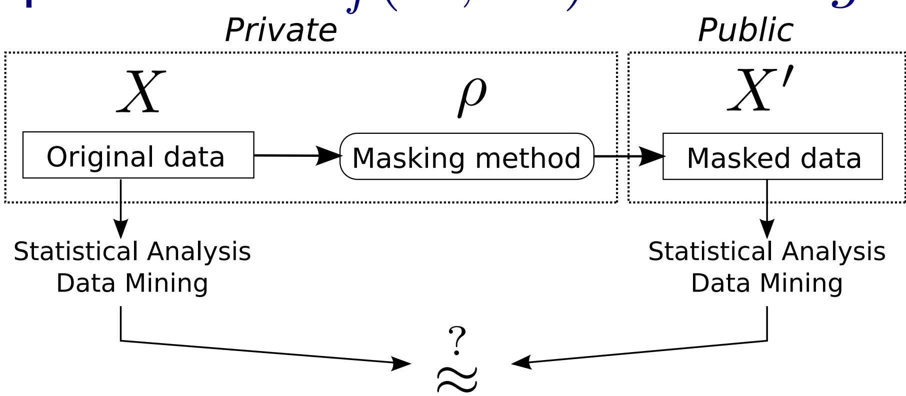
# Research questions II: Information loss/Utility

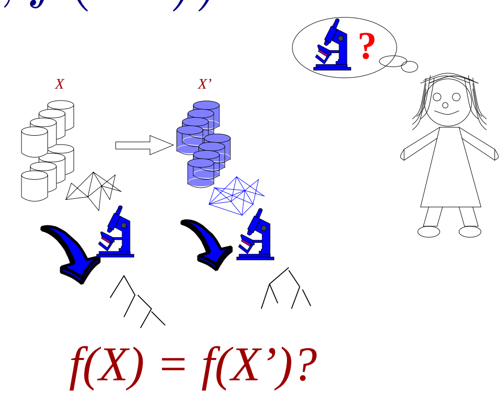
**Information loss measures.** Compare $X$ and $X'$ w.r.t. analysis $(f)$

- $f$: generic vs. specific (data uses). E.g. regression




- Comparison: $IL_f(X, X') = divergence(f(X), f(X'))$

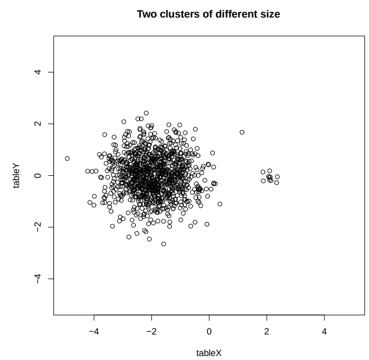


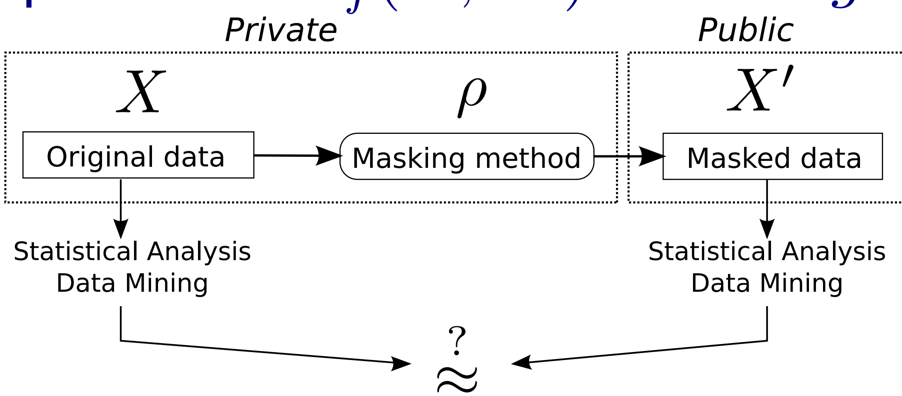
# Research questions II: Information loss/Utility

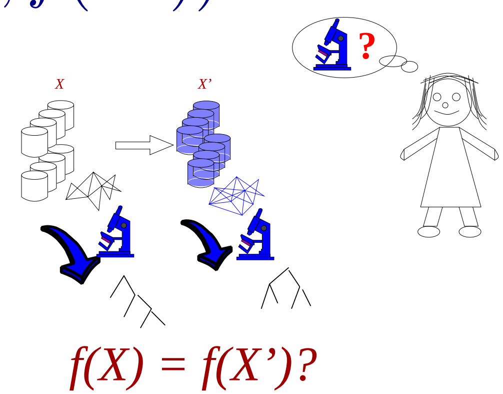
**Information loss measures.** Compare $X$ and $X'$ w.r.t. analysis ($f$)

- $f$: generic vs. specific (data uses). E.g. clustering




- Comparison: $IL_f(X, X') = divergence(f(X), f(X'))$

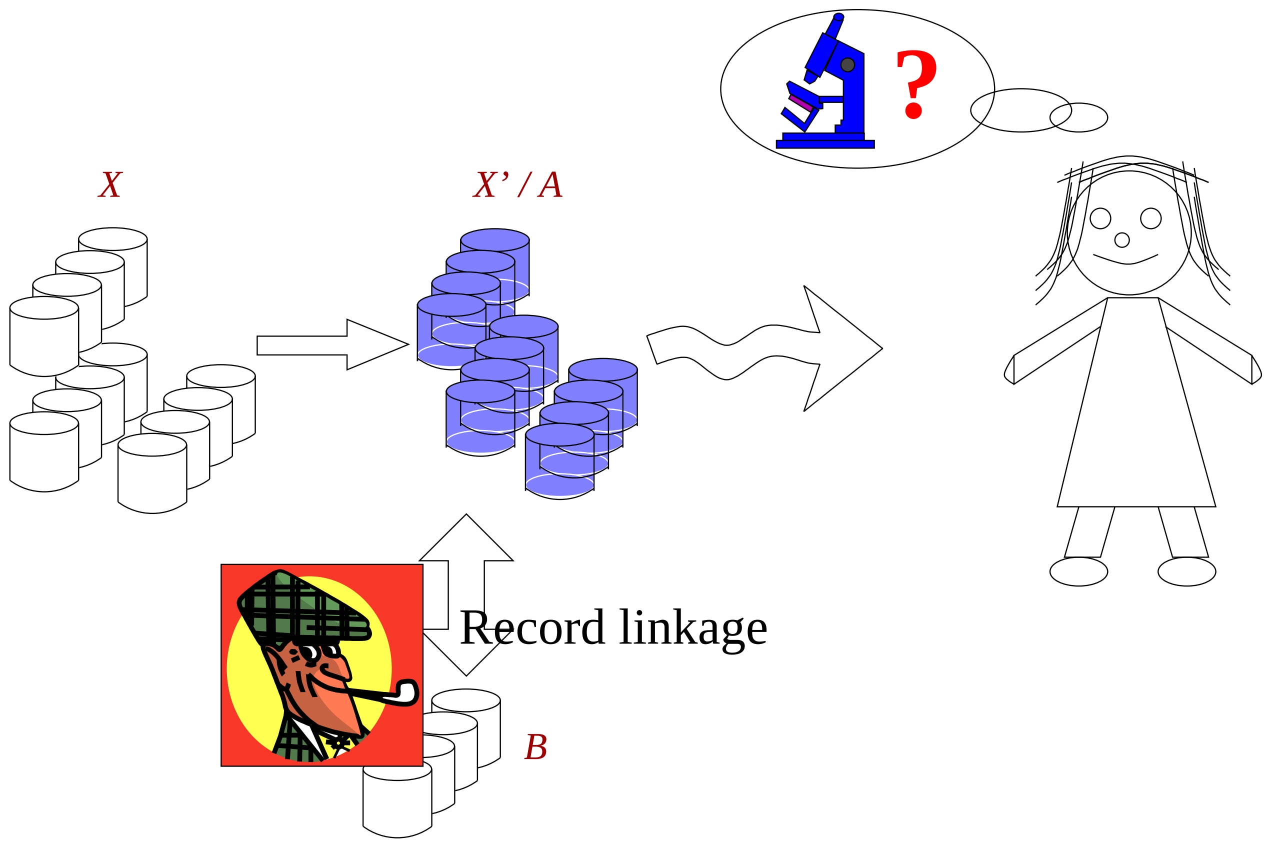


# Research questions II: Information loss

**Disclosure risk.** One of the privacy models: reidentification (identity disclosure)

- $A$: File with the protected data set
- $B$: File with the data from the intruder (subset of original $X$)



# Tabular data

# Tabular data

- Aggregates of data with respect to a few variables. Ex. (Castro, 2012)

| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | Total |
|---|---|---|---|---|---|---|
| $M_1$ | 2 | 15 | 30 | 20 | 10 | 77 |
| $M_2$ | 72 | 20 | 1 | 30 | 10 | 133 |
| $M_3$ | 38 | 38 | 15 | 40 | 5 | 136 |
| TOTAL | 112 | 73 | 46 | 90 | 25 | 346 |

Cell $(M_2, P_3)$: number of people with profession $P_3$ living in municipality $M_2$.

| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | Total |
|---|---|---|---|---|---|---|
| $M_1$ | 360 | 450 | 720 | 400 | 360 | 2290 |
| $M_2$ | 1440 | 540 | 22 | 570 | 320 | 2892 |
| $M_3$ | 722 | 1178 | 375 | 800 | 363 | 3438 |
| TOTAL | 2522 | 2168 | 1117 | 1770 | 1043 | 8620 |

Cell $(M_2, P_3)$: total salary received by people with profession $P_3$ living in $M_2$.

# Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$

# Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$
  - **Internal attack.** A person whose data is in the database is able to use the information of the tables to infer some sensitive information about other individuals. A doctor infers the salary of another doctor.
    $\Rightarrow (M_1, P_1)$

# Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$
  - **Internal attack.** A person whose data is in the database is able to use the information of the tables to infer some sensitive information about other individuals. A doctor infers the salary of another doctor.
    $\Rightarrow (M_1, P_1)$
  - **Internal attack with dominance.** This is an internal attack where a contribution of one person, say $p_0$, in a cell is so high that permits $p_0$ to obtain accurate bounds of the contribution of the others.
    $\Rightarrow (M_3, P_5)$ with 5 people. $salary(p_0) = 350$, then the salary of the other four is at most $363 - 350 = 13$.

# Tabular data

- Privacy model / disclosure risk measure
- Data protection mechanism
- Information loss

# Tabular data: privacy model

- **Rule** $(n, k)$**-dominance.** A cell is sensitive when $n$ contributions represent more than the $k$ fraction of the total. That is, the cell is sentitive when

$$\frac{\sum_{i=1}^{n} c_{\sigma(i)}}{\sum_{i=1}^{t} c_i} > k$$

  where $\{\sigma(1), ..., \sigma(t)\}$ is a permutation of $\{1, ..., t\}$ such that $c_{\sigma(i-1)} \geq c_{\sigma(i)}$ for all $i = \{2, ..., t\}$ (i.e., $c_{\sigma(i)}$ is the $i$th largest element in the collection $c_1, ..., c_t$).

  This rule is used with $n = 1$ or $n = 2$ and $k > 0.6$.

# Tabular data: privacy model

- **Rule $pq$.** This rule is also known as the prior/posterior rule. It is based on two positive parameters $p$ and $q$ with $p < q$. Prior to the publication of the table, any intruder can estimate the contribution of contributors within the $q$ percent. Then, a cell is considered sensitive if an intruder on the light of the released table can estimate the contribution of a contributor within $p$ percent.

- **Rule p%.** This rule can be seen as a special case of the previous rule when no prior knowledge is assumed on any cell. Because of that, it can be seen as equivalent to the previous rule with $q = 100$.

# Tabular data: data protection mechanism

- Protection of a tabular data
  - Perturbative
    - ⋆ Post-tabular
      - Rounding
      - Controlled tabular adjustment (CTA)
    - ⋆ Pre-tabular
  - Non-perturbative: cell suppression

# Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression
- Primary suppression not enough:

| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | Total |
|---|---|---|---|---|---|---|
| $M_1$ | 360 | 450 | 720 | 400 | 360 | 2290 |
| $M_2$ | 1440 | 540 | ~~22~~ | 570 | 320 | 2892 |
| $M_3$ | 722 | 1178 | 375 | 800 | 363 | 3438 |
| TOTAL | 2522 | 2168 | 1117 | 1770 | 1043 | 8620 |

- Secondary suppressions required:

| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | Total |
|---|---|---|---|---|---|---|
| $M_1$ | 360 | 450 | | 400 | | 2290 |
| $M_2$ | 1440 | 540 | | 570 | | 2892 |
| $M_3$ | 722 | 1178 | 375 | 800 | 363 | 3438 |
| TOTAL | 2522 | 2168 | 1117 | 1770 | 1043 | 8620 |

- Solutions build using optimization

# Tabular data: information loss

- Minimal number of suppressions
- Weights associated to cells: *minimal weight* of suppressed cells

# Summary

# Summary

- Privacy models
- Microdata / standard databases
- Tabular data

# Thank you

# References


- V. Torra, Data Privacy: Foundations, New Developments and the Big Data Challenge, Springer, 2017.
- T. Benschop, C. Machingauta, M. Welch, Statistical disclosure control for microdata: A practical guide, 2016.
- A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, P.-P. de Wolf, Statistical Disclosure Control, Wiley, 2012.
- M. Templ, Statistical disclosure control for microdata: Methods and applications in R, Springer, 2017.
- J. Castro, Recent advances in optimization techniques for statistical tabular data protection, European Journal of Operational Research 216 (2012) 257-269.

# Book

- Vicenç Torra, Data Privacy: Foundations, New Developments and the Big Data Challenge, Springer, 2017.

  Content: 1. Introduction. 2. Machine and statistical learning. 3. On the classification of protection procedures. 4. User's privacy. 5. Privacy models and disclosure risk measures. 6. Masking methods. 7. Information loss: evaluation and measures. 8. Selection of masking methods. 9. Conclusions.

  Includes sections on masking methods and transparency, and variants for big data. User privacy for communications and information retrieval (PIR).