Privacy models for machine learning and statistics

Vicenç Torra

May 2021

Dept. CS, Umeå University, Sweden

Background

- My background ...
  - Started in this field in 2000 (before the data privacy hype).
  - How to make data useful and private for statistics and ML
  - Research topics:
    - ▷ Privacy from a computational point of view
    - ▷ Privacy-aware for machine learning and statistics

Outline

A context

- Data analysis and data-driven models
- Data analysis and data-driven models + privacy

Privacy models

- Two motivating examples
- Privacy models
- Privacy models: Avoiding reidentification
- Privacy models: Avoiding inference from calculations

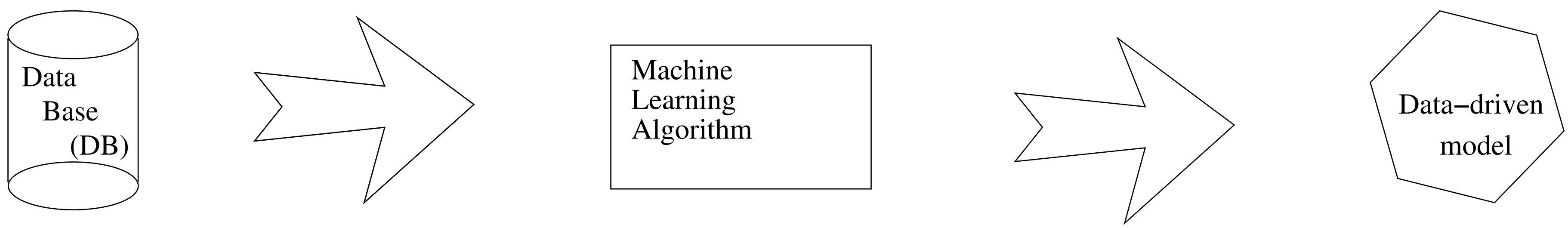
A context:

Machine learning and statistics

Data analysis and data-driven models

Data-driven models

- Data-driven model
  (regression, logistic regression, neural networks, etc.) for prediction, image processing, decision support systems, etc.



Data-driven models

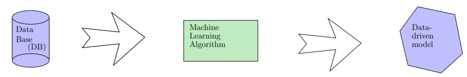
- Machine learning (usage, informal)
  - Data access (relevant and irrelevant data)
  - Exploratory data analysis
  - Model building (several models)
    (different types of models, different (hyper-)parameters)
  - Select a good model
    (whatever good means)

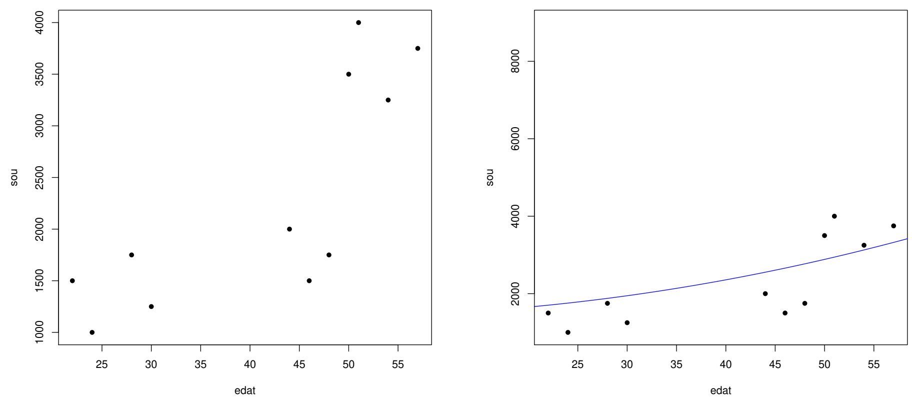
- Example
  - Hospital length stay at time of admission[1]

[1] https://www.nature.com/articles/sdata201635

Data-driven models

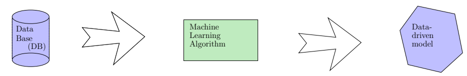
- Build a data-driven model: age → income

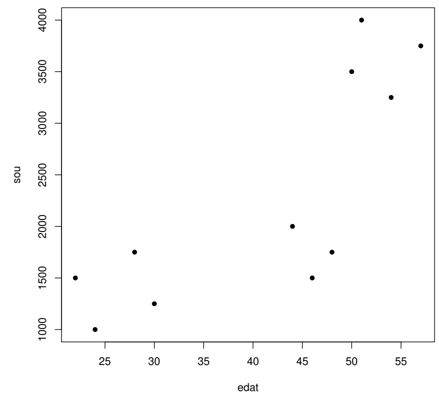


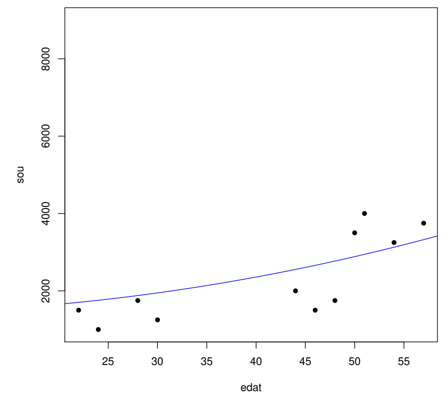


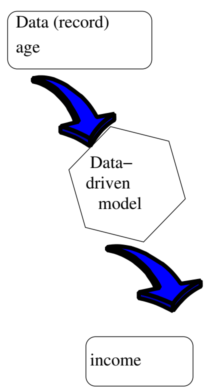
Data-driven models

- Build a data-driven model: age → income









$$\text{income} = 1418.63 + 0.5864 * \text{age}^2$$

Income of Aina (age=25, income=?)

Data analysis and data-driven models

and privacy

Data-driven models and privacy

- Relevant questions for privacy
  - Who has data access? ⇒ access control
    - ▷ Different actors have different roles/permissions (data access):
      Admissions, pharmacy technician, clinical laboratory, physician, etc.
    - ▷ But also
      Health information technician, and data scientists

Data-driven models and privacy

- Relevant questions for privacy
  - Who has data access? ⇒ access control
    - ▷ Different actors have different roles/permissions (data access): Admissions, pharmacy technician, clinical laboratory, physician, etc.
    - ▷ But also Health information technician, and data scientists
- Access control is not enough
  - Access seems ok but inferences may imply disclosure

Data-driven models and privacy

- Relevant questions for privacy
  - From what you are allowed to access,
    can you infer something you shouldn't learn? E.g.,
    - Can you find someone you know from the information you are allowed to access?
    - Can you learn sensitive information from anonymized / view of database?
    - Can you learn sensitive information from aggregated data?
    - Can you learn sensitive information from a model?
  - If so, what should we do instead? $\Rightarrow$ Data privacy

Data-driven models and privacy

- Relevant questions for privacy
  - From what you are allowed to access,
    can you infer something you shouldn't learn? E.g.,
    - Can you find someone you know from the information you are allowed to access?
    - Can you learn sensitive information from anonymized / view of database?
    - Can you learn sensitive information from aggregated data?
    - Can you learn sensitive information from a model?
  - If so, what should we do instead? $\Rightarrow$ Data privacy

Data privacy is (not only) about data leakages
(privacy vs. security and access control)

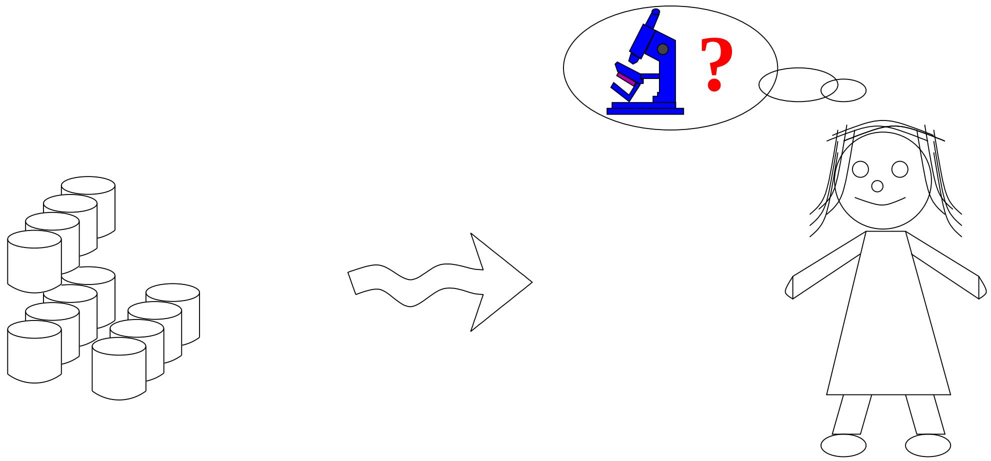
Data privacy

Two motivating examples

Anonymization is more difficult than it seems

Two motivating examples

- Case #1. A database with people. Hospital data.
  - Solution. Remove names and identity card/passport numbers



Two motivating examples

- Case #1. A database with people. Hospital data.
  - Solution. Remove names and identity card/passport numbers
  - Naive anonymization does not work ......!!
    Sensitive information can still be inferred.
    Other attributes can be used to find a record

~~Darth Vader~~, Washington National Cathedral, Northwest, Washington D.C.

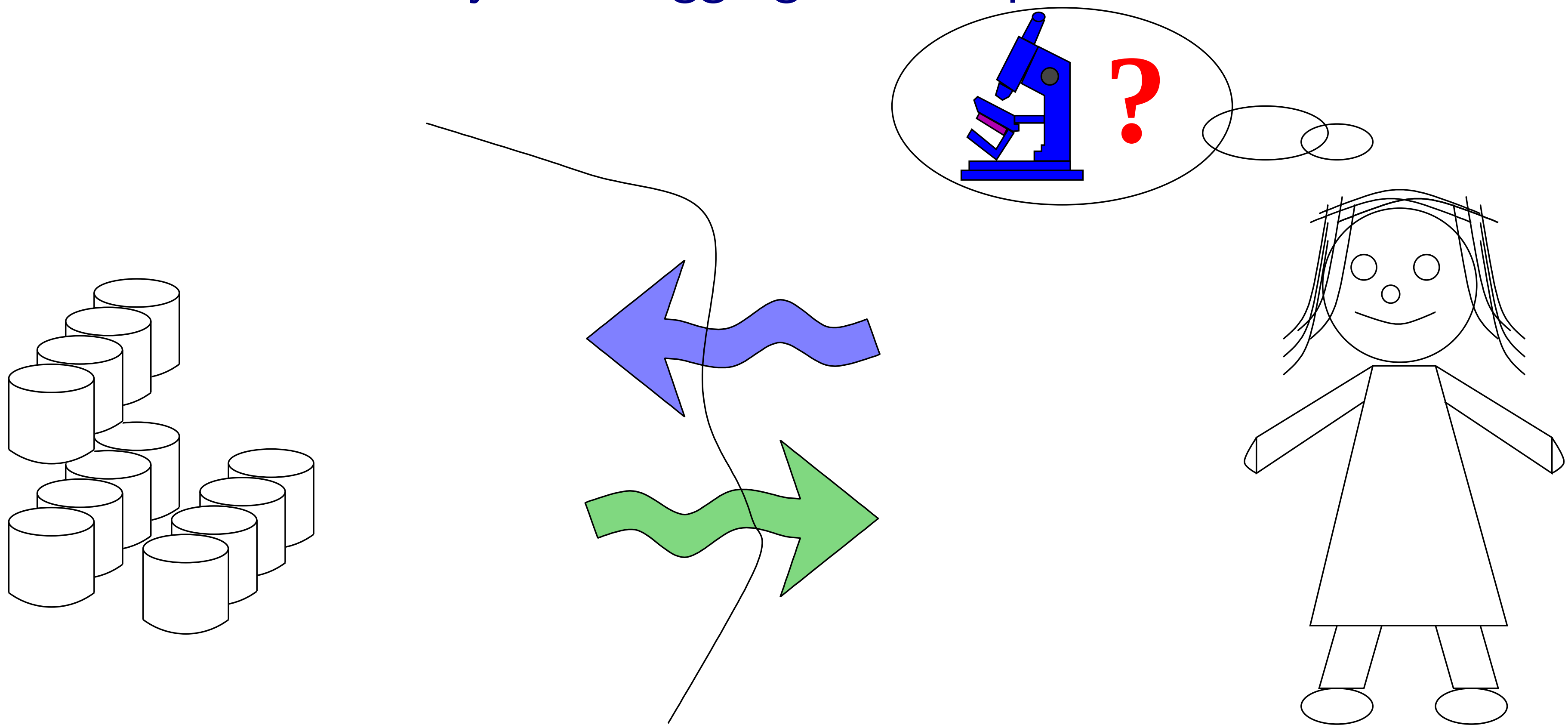
Image from wikipedia

Two motivating examples

- Difficulties: Naive anonymization does not work
  - Cases about disclosure from incorrect anonymization
    - ▷ AOL, Netflix (search logs, film ratings)
    - ▷ 3.7% (9.1 /248 million) likely to be uniquely identified by 5-digit ZIP, gender, Month and year of birth
  - Similarly
    - ▷ Mobile positions (two positions identify)
    - ▷ fidelity cards, credit card payments, shopping carts ...
  - High dimensional data + highly identifiable data

Two motivating examples

- Case #2. Mean salary (or, in general, any other computation – ML)
  - Solution. Mean salary is an aggregate, not personal data.



Two motivating examples

- Case #2. Mean salary (or, in general, any other computation – ML)
  - Solution. Mean salary is an aggregate, not personal data.
    Compute $\sum_{i=1}^{n} x_i/n$

Two motivating examples

- Case #2. Mean salary (or, in general, any other computation – ML)
  - Solution. Mean salary is an aggregate, not personal data.
    Compute $\sum_{i=1}^{n} x_i/n$
  - This does not work ......!!

    'I sense something. A presence I have not felt since . . . '

    (Darth Vader, Star Wars IV: A new hope)

  - A simple function can give information on who is in the database
    - ▷ Mean salary of psychiatric unit by town
      For a given town, ⇒ disclosure of a rich person

Two motivating examples

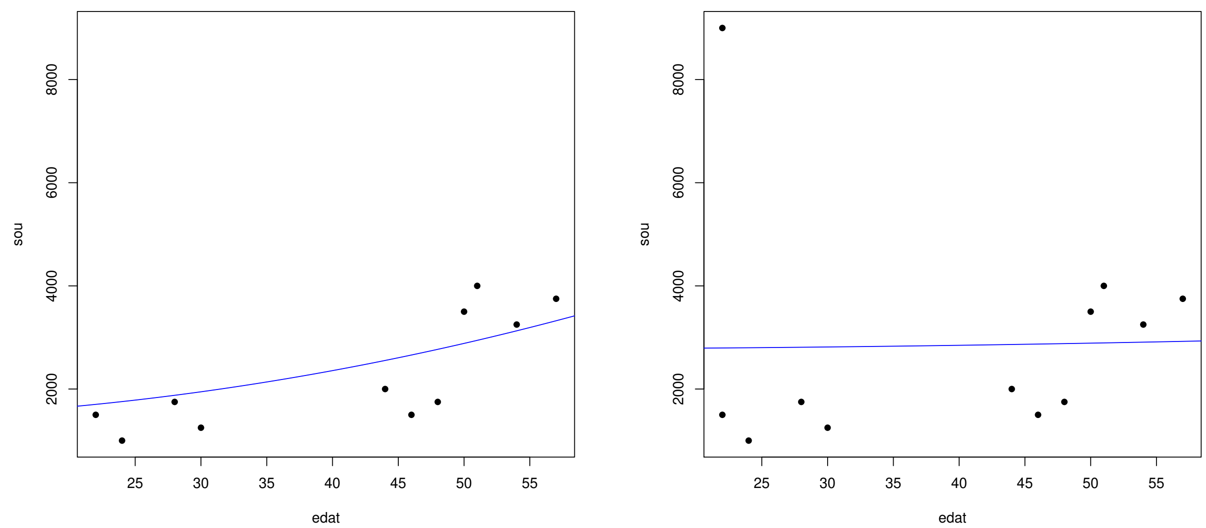
- Case #2. Mean salary
  - Q: Mean income of admitted (unit, town) – psychiatric unit
    (similar problem: mean salary by town)
  - Mean income is not "personal data", is this ok ?
    - ▷ Example:
      1000 2000 3000 2000 1000 6000 2000 10000 2000 4000
      $\Rightarrow$ mean = 3300
    - ▷ Adding Ms. Rich's salary 100,000 Eur/month:
      $\Rightarrow$ mean = 12090,90 !
      (a extremely high salary changes the mean significantly)
      $\Rightarrow$ We infer Ms. Rich from Town was attending the unit

  Obi-Wan Kenobi is in the Death Star

Two motivating examples

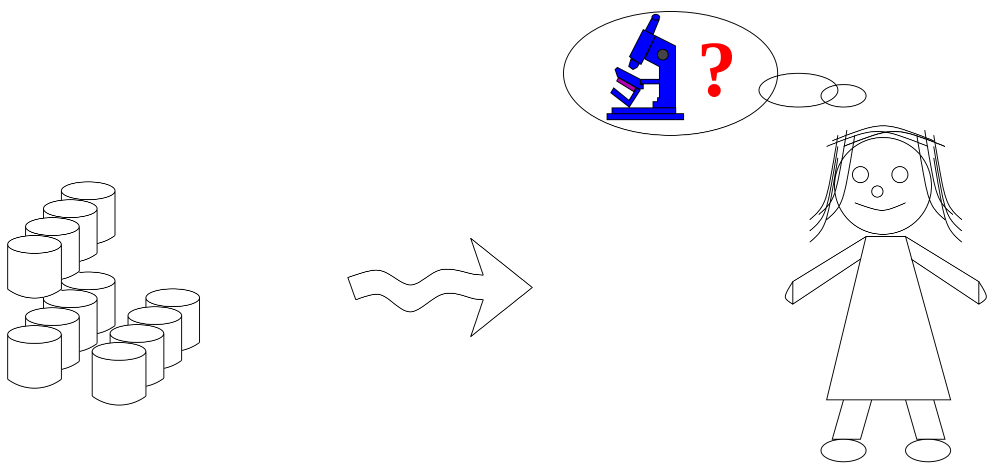
Example #2. Another computation

- Q: Regressions (and other ML models)
  membership attacks (Ms Rich data as has been used?)



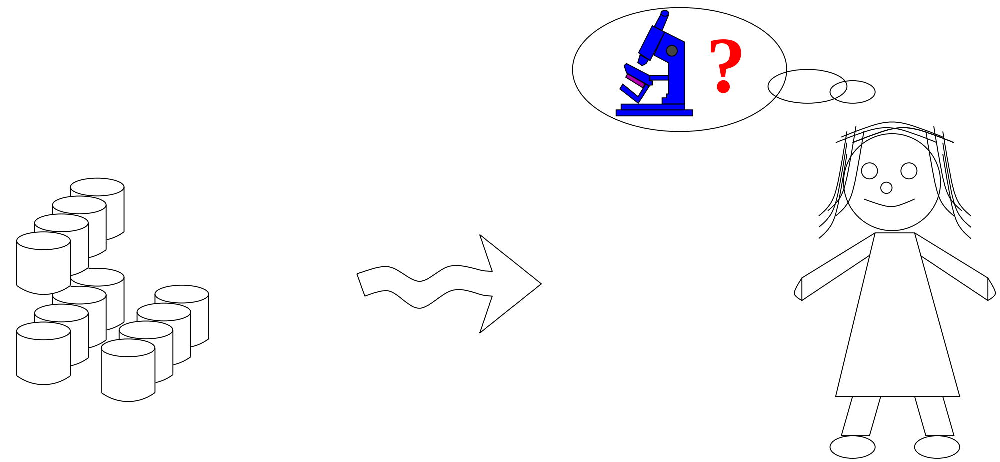
income $=$ 1418.63 $+$ 0.5864 * age^2 vs. income $=$ 2774 $+$ 0.04639 * age^2

Privacy models



Privacy models

Privacy model. A computational definition for privacy.



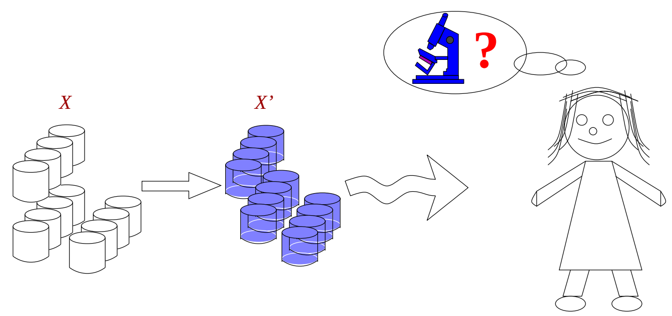
Privacy models

(Some) Privacy models. Computational definitions for privacy.

- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.
- **Result privacy.** We want to avoid some results when an algorithm is applied to a database.
- **Differential privacy.** Output of a query does not change when a record is added/removed from a DB.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.

Privacy models:

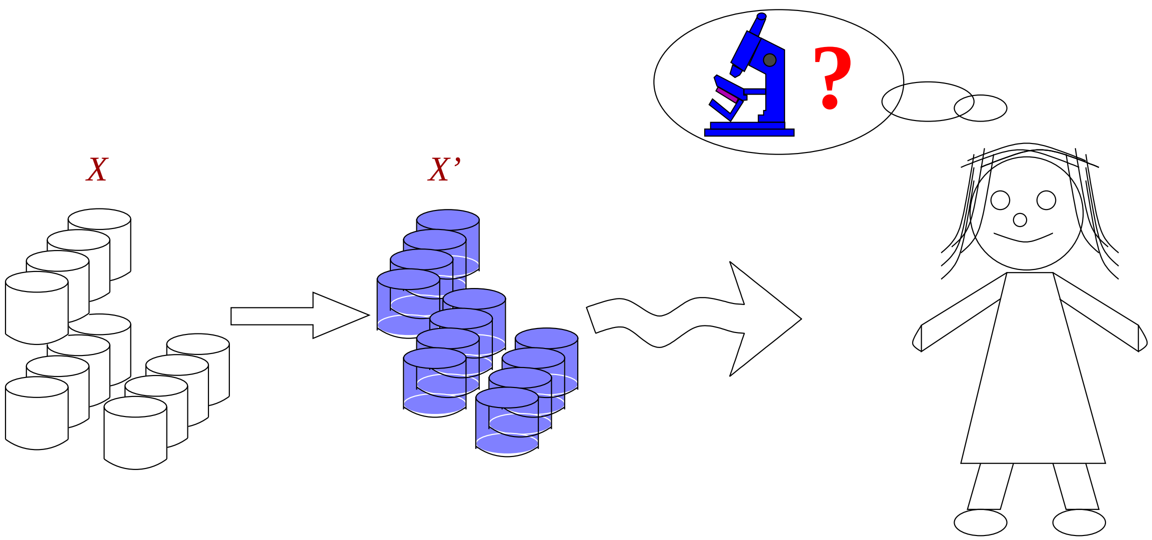
Avoiding reidentification



Privacy models

Privacy models. A computational definition for privacy.

- **Reidentification privacy.** Avoid finding a record
- **k-Anonymity.** k indistinguishable records



can we find  ? we don't want this possible ...

Privacy models

Privacy models. A computational definition for privacy.

- **Reidentification privacy.** Avoid finding a record
- **k-Anonymity.** k indistinguishable records

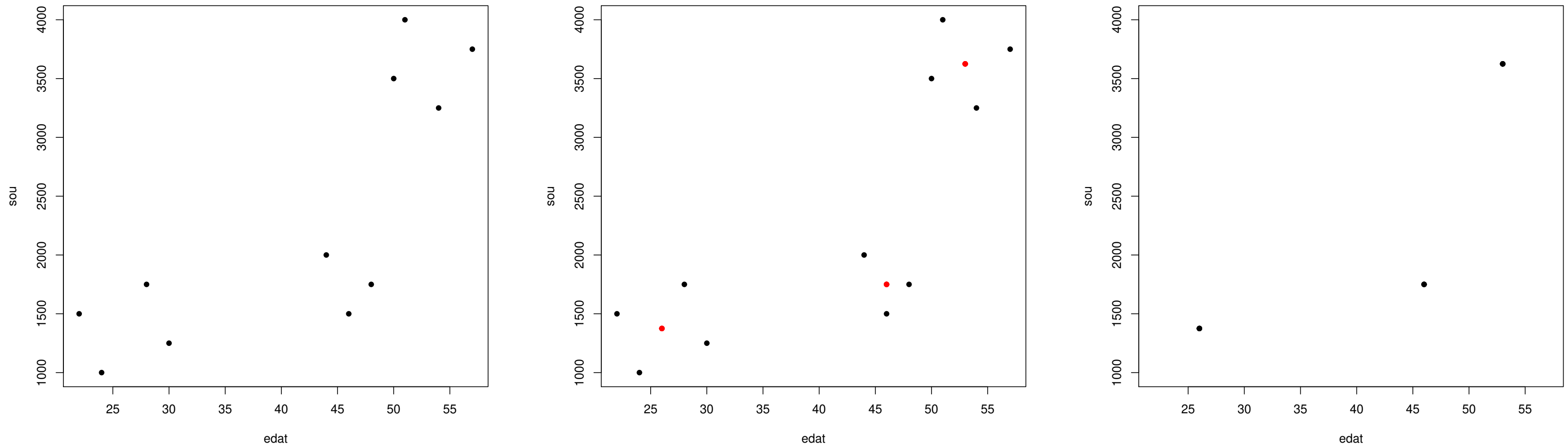
How? Change the level of detail or add noise to the data

- Additive noise:
  $x' = x + r$ with $r \sim N(0, b)$: 2019 $\rightarrow$ 2018
- Generalization: $x' = county(town(x))$:
  Maynooth $\rightarrow$ Kildare (Ireland)
- Microaggregation:
  We build clusters with a minimum size and publish means
- ...

Microaggregation

Data protection. Microaggregation. Clusters: at least k records

- **Privacy model.** k-Anonymity ($k = 3$)



Database: (age, income)

- Original cluster: {(22,1500), (24,1000), (28, 1750), (30, 1250)}
- Protected cluster: {(26, 1375),(26, 1375),(26, 1375),(26, 1375)}

- **Formalization.** $u_{ij} = 1$ iff x_j is in the i-th cluster; v_i centroide

  Minimize $\quad SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2$

  Subject to $\sum_{i=1}^{g} u_{ij} = 1$ for all $j = 1, \ldots, n$

  $2k \geq \sum_{j=1}^{n} u_{ij} \geq k$ for all $i = 1, \ldots, g$

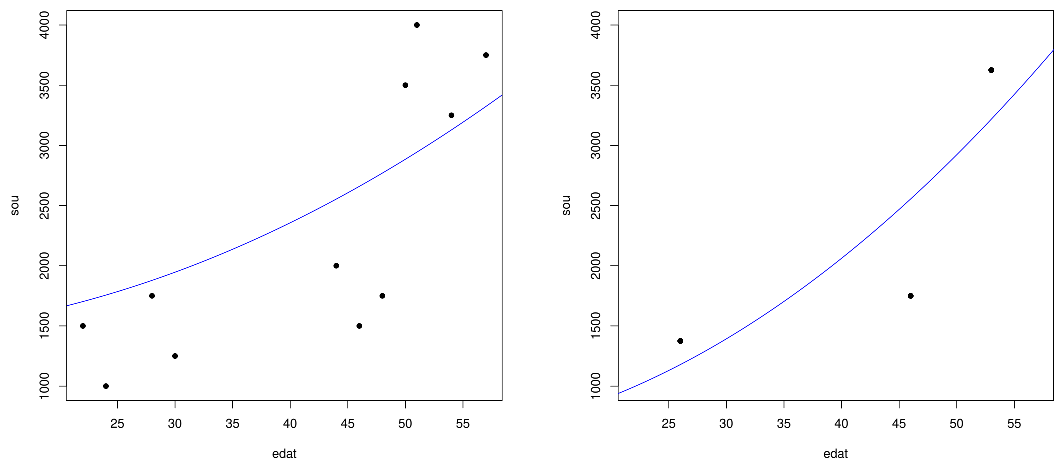
  $u_{ij} \in \{0, 1\}$

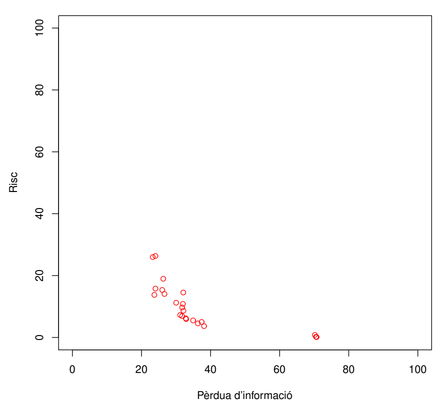
Microaggregation

Data protection. Microaggregation. Clusters: at least k records

- Clusters ensure anonymity, but
  we also want to preserve utility

  Can we infer Aina's salary? (age=25, income=?)

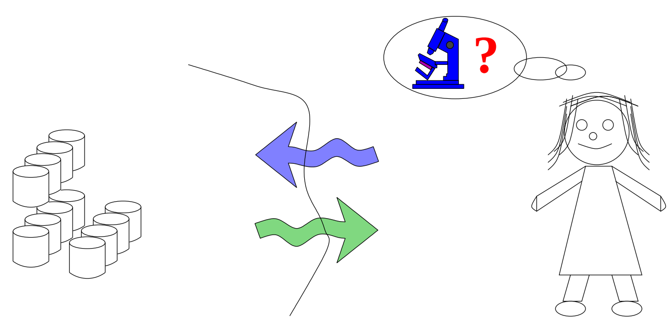




Fuzzy microaggregation. The boundaries of clusters are not crisp, we can assign a record to several clusters, and reduce influence of outliers (income of Ms Rich)

Privacy models:

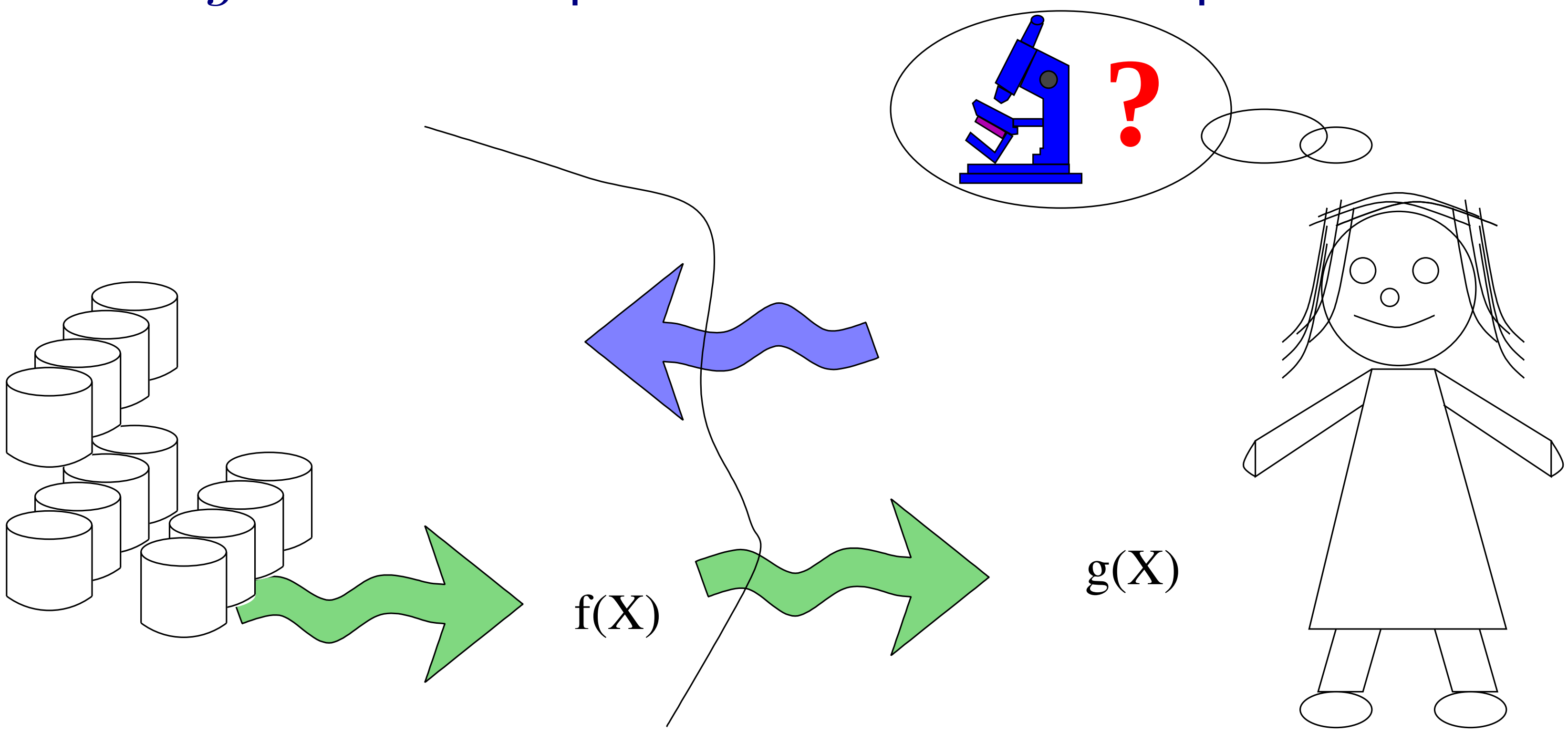
Avoiding inference from calculations



Privacy models

Privacy models. Avoiding inferences from computations

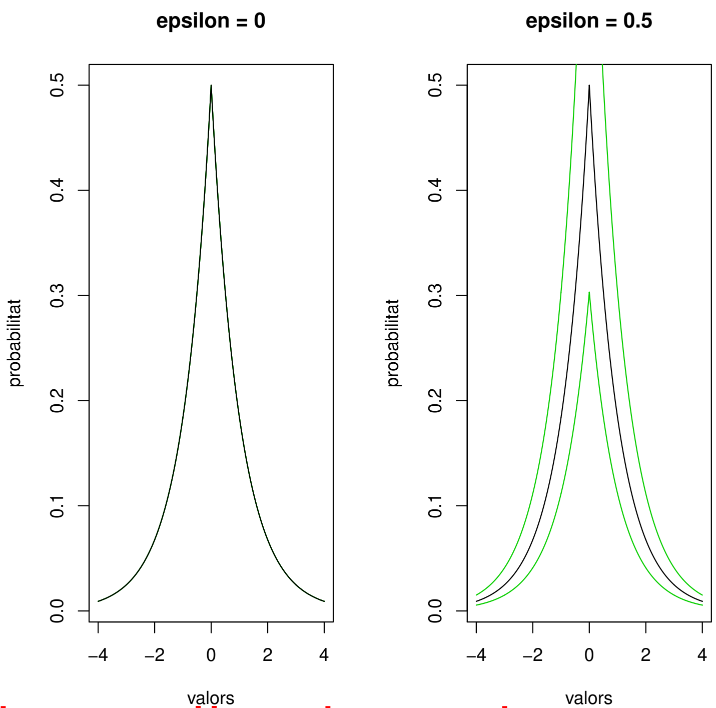
- **Differential privacy.**
  The outcome does not depend (much) on the presence (absence) of a record
- Implementation: instead of $f(X)$ compute $g(X)$,
  and so that g does not depend so much on the input add noise



Privacy models

Privacy models. Avoiding inferences from computations

- **Differential privacy.**
  Result does not depend (much) on the presence (absence) of a record
- Implementation: instead of $f(X)$ compute $g(X)$, typically $g(X) = f(X) + r$ with $r \sim L(0, b)$ (Laplace distribution)



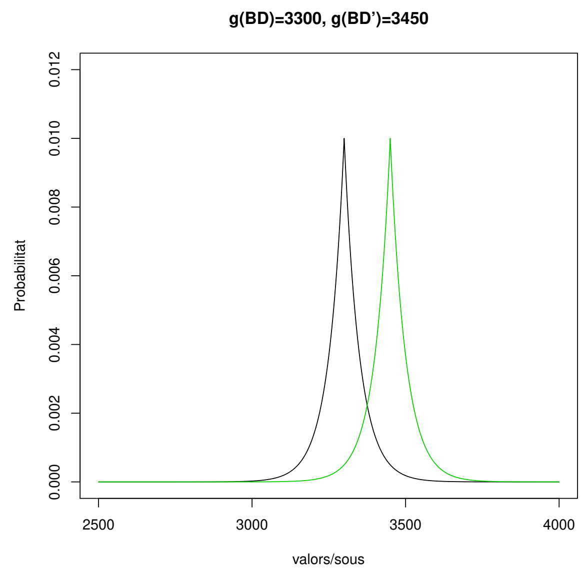
Definition. The result $g(D)$ satisfies differential privacy in degree ϵ if for all BD_1 and BD_2 it holds for all $S \subseteq Range(K_q)$, $Pr[K_q(BD_1) \in S] \leq e^{\epsilon} Pr[K_q(BD_2) \in S]$

- The smaller the ϵ, the more similar the two distributions

Privacy models

Privacy models. Avoiding inferences from computations

- **Differential privacy.** Implementation
  - Define $g(X) = f(X) + r$ with $r \sim L(0, b)$ (Laplace distribution)
  - Example $f(BD) = 3300$ and $f(BD') = 3450$, with Laplace distribution $L(0, 50)$



- The value b in $L(0, b)$ depends on ϵ (the privacy level) and the sensitivity of the function f to the possible DBs

Privacy models

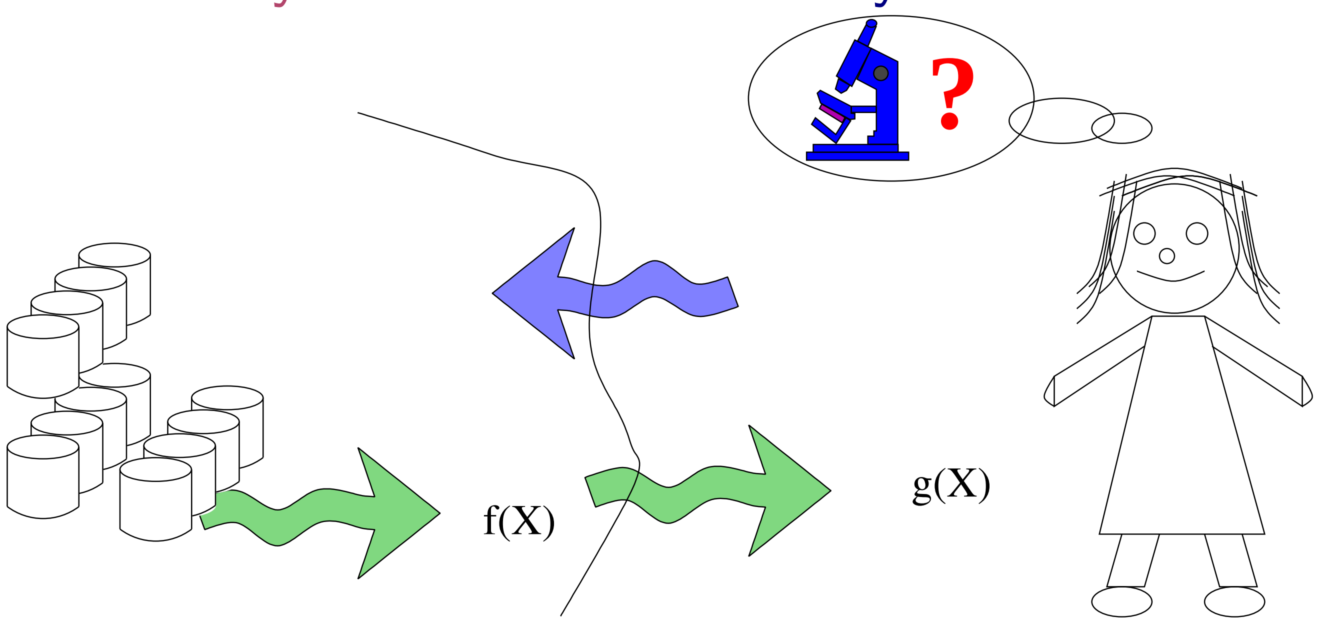
Privacy models. Avoiding inferences from computations

- **Differential privacy.**
  Other mechanisms for non-numerical functions
  and, for example, for neural networks/deep learning, decision trees
- Solutions are robust to membership attacks
  (recall Ms. Rich!)

Privacy models

Privacy models. Avoiding inferences from computations

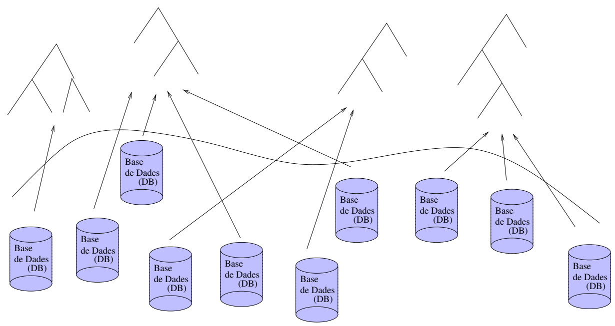
- **Integral privacy.**
  The outcome is a recurrent result
  - Several databases can provide the same result
- Privacy:
  - k databases generate the same result (k-anonymity)
  - plausible deniability: I wasn't there – Says Ms. Rich



Privacy models

Privacy models. Avoiding inferences from computations

- **Integral privacy.**
  The outcome is a recurrent result



Definition. The result $G = g(D)$ satisfies integral privacy given background knowledge S^* if $Gen(G, S^*)$ is large (k BDs) and $\bigcap_{g \in Gen^*(G,S^*)} g = \emptyset$.
where $Gen(G, S^*) = \{S' | S^* \subseteq S' \subseteq P, A(S') = G\}$
$Gen^*(G, S^*) = \{S' \setminus S^* | S^* \subseteq S' \subseteq P, A(S') = G\}$

- k different databases,
  not sharing records (and different enough)
  to avoid membership attacks

Privacy models

Privacy models. Avoiding inferences from computations

- **Integral privacy.**

  The outcome is a recurrent result

  - Example

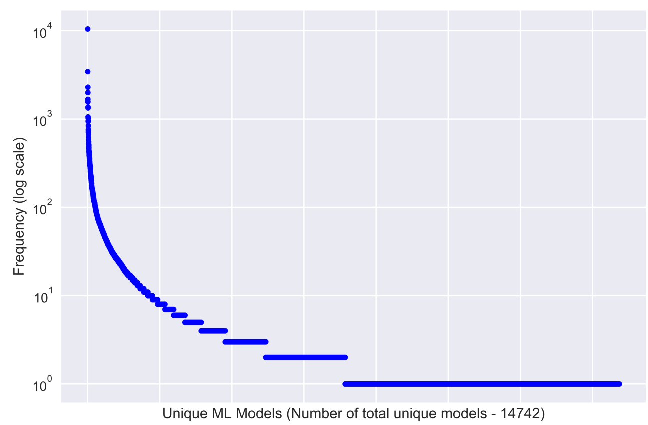
$$\{1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000\} \cup \{100000\}$$

  - Several subsets return the same output: mean equal 3000
    - ▷ {3000}
    - ▷ {2000, 4000}
    - ▷ {6000, 2000, 1000}
    - ▷ {10000, 1000, 1000, 3000}
    - ▷ {6000, 4000, 1000, 2000, 3000, 2000}

Privacy models

Privacy models. Avoiding inferences from computations

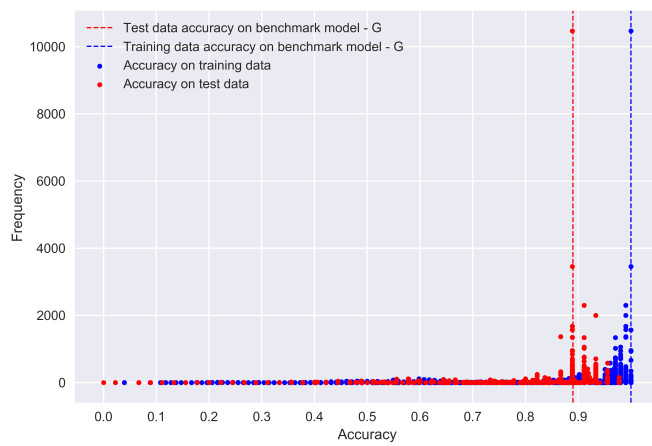
- **Integral privacy.**
  The outcome is a recurrent result
- Recurrent models also appear in machine learning
- Decision trees built from a database (Iris dataset). Models/freq.



Privacy models

Privacy models. Avoiding inferences from computations

- **Integral privacy.**
  The outcome is a recurrent result
- Recurrent models can also have good accuracy
- Decision trees built from a database (Iris dataset). Accuracy/freq.



Privacy models

Privacy models. Avoiding inferences from computations

- Differential privacy, smooth function
$$f(D) \sim f(D \oplus x)$$
where $D \oplus x$ represents adding a record x to a database D
- Integral privacy, recurrent function
If $f^{-1}(G)$ is the set of all (real) databases that can generate G, we require $f^{-1}(G)$ to be a large and diverse set.

Privacy models

Privacy models. Avoiding inferences from computations

- Differential privacy, smooth function
$$f(D) \sim f(D \oplus x)$$
where $D \oplus x$ represents adding a record x to a database D
- Integral privacy, recurrent function
If $f^{-1}(G)$ is the set of all (real) databases that can generate G, we require $f^{-1}(G)$ to be a large and diverse set.

- An example of a simple function that satisfies integral privacy is:
A an algorithm that returns 1 if the number of records of D is even and 0 if it is odd
That is, $f(D) = 1$ if and only if $|D|$ is even.

Summary

Summary

- Achieve a good anonymization is challenging
  (if we want data to be useful, of course!)

- It is possible to obtain data and models protected enough to be useful and with certain privacy levels.

Summary

- My research ...
  - Data masking methods for SQL/noSQL (microaggregation, rank swapping)
  - Disclosure risk assessment for masked data. Worst-case scenario: transparency attacks + machine learning to identify best parameters
  - Differential privacy + Integral privacy
  - Federated learning

Thank you

References

Related references.

- V. Torra, Fuzzy microaggregation for the transparency principle. J. Appl. Log. 23 (2017) 70-80.
- D. Abril, G. Navarro-Arribas, V. Torra, Supervised Learning Using a Symmetric Bilinear Form for Record Linkage, Information Fusion 26 (2015) 144-153.
- N. Senavirathne, V. Torra, Integrally private model selection for decision trees, Comput. Secur. 83 (2019) 167-181.
- V. Torra, G. Navarro-Arribas, E. Galván, Explaining Recurrent Machine Learning Models: Integral Privacy Revisited. Proc. PSD 2020 62-73.
- V. Torra (2017) Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer.
- http://ppdm.cat/dp/