Research directions on data privacy

Vicenç Torra

November 2020

Dept. CS, Umeå University, Sweden

Background

- A kind of justification ...
  - Started in this field in 2000 (before the data privacy hype).
  - Background on AI and data aggregation,
  - *Statistics perspective:*
    - ⋆ Data uses should go beyond statistics & regression (now clear)
  - *Machine learning/data mining:*
    - ⋆ Sensitive data is an issue, and it is pervasive.
      Its 'smell' is infiltrating the ML models.
    - ⋆ Trade-off privacy & utility for ML uses.

Background

- Research topics:
  - Privacy from a computational point of view
  - Privacy-aware for machine learning and statistics

Outline

Two motivating examples

Privacy models

Data-driven and general purpose: masking databases

Computation-driven or specific purpose

Two motivation examples

Two motivating examples

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

Two motivating examples

- Anonymization is more difficult than it seems

Two motivating examples

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

Two motivating examples

- Case #1. A database with people.
  - Solution. Remove names and identity card/passport numbers
  - This does not work ......!!

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

Image from wikipedia

Two motivating examples

- Difficulties: Naive anonymization does not work
  - (Sweeney, 1997; 2000[1]) on USA population
    - ⋆ 87.1% (216 /248 million) is likely to be uniquely identified by 5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% (9.1 /248 million) is likely to be uniquely identified by 5-digit ZIP, gender, Month and year of birth
- Difficulties: highly identifiable data
  - AOL and Netflix cases (reidentification: search logs/movie ratings)
  - Similar with credit card payments, shopping carts ...
    $\Rightarrow$ high dimensional data: unique people: reidentification
  - Data from mobile devices: (two variables)
    - ⋆ two positions can make you unique (home and working place)

[1] L. Sweeney, Simple Demographics Often Identify People Uniquely, CMU 2000

Privacy models

- Difficulties: highly identifiable data.
  - University: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)

Privacy models

- Difficulties: highly identifiable data.
  - University: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No “personal data”,

    Umeå, CS, No

    Umeå, CS, No

    Umeå, CS, Yes

    Lycksele, CS, No

    Umeå, BA MEDIA STUDIES, No

    Umeå, BA MEDIA STUDIES, Yes

  is this ok ?

Privacy models

- Difficulties: highly identifiable data.
  - University: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",

    Umeå, CS, No

    Umeå, CS, No

    Umeå, CS, Yes

    Lycksele, CS, No

    Umeå, BA MEDIA STUDIES, No

    Umeå, BA MEDIA STUDIES, Yes

    is this ok ?

    NO!!:
  - E.g., there is only one student of anthropology living in Täfteå.

    Täfteå, Anthropology, Yes

    Only one black mask in the death star

Two motivating examples

- Case #2. Mean salary
  - Solution. Mean salary is an aggregate, not personal data. Compute $\sum_{i=1}^{n} x_i / n$

Two motivating examples

- Case #2. Mean salary
  - 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, $\Rightarrow$ disclosure of a rich person

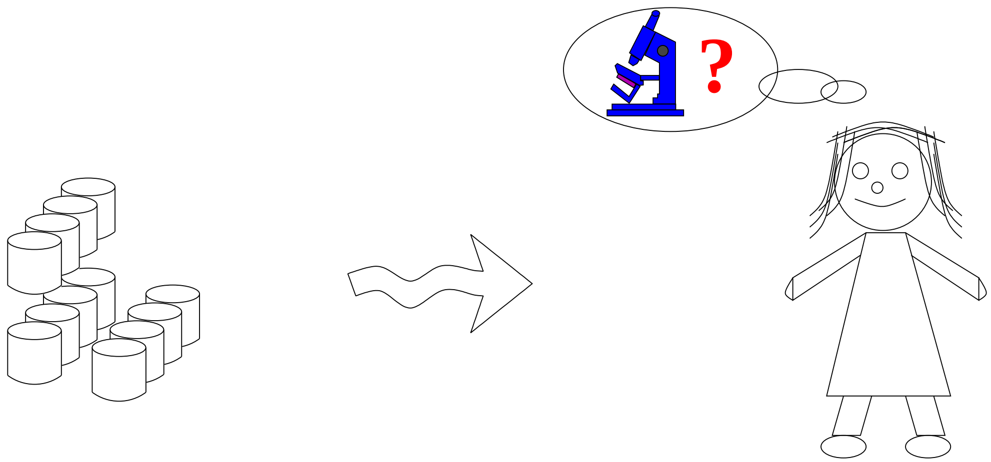
Two motivating examples

- Case #2. Mean salary
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Mean income is not "personal data", is this ok ? NO!!:
  - Example[2]: 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean $= 3300$
  - Adding Ms. Rich's salary 100,000 Eur/month: 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

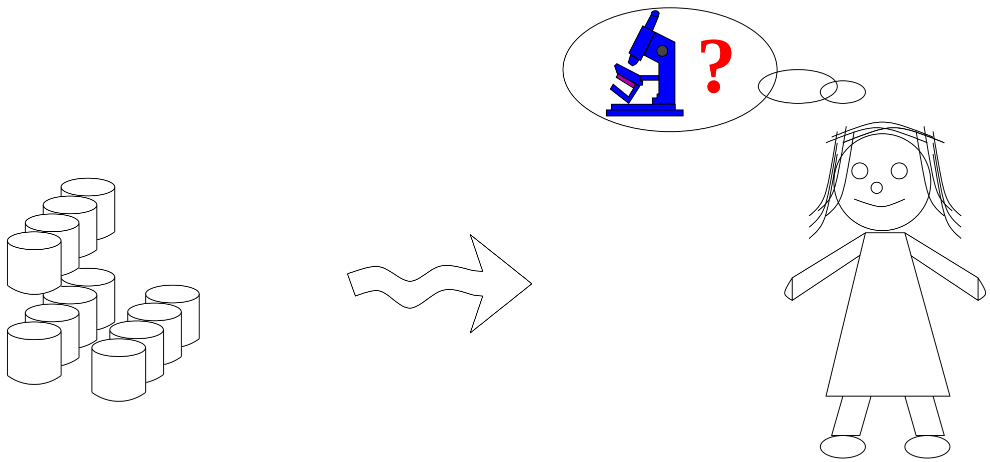
[2] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur (accessed nov. 2020):
https://www.frsrecruitment.com/articles/market-insights/average-wage-in-ireland

Privacy models



Privacy models

Privacy model. A computational definition for privacy.



Privacy models

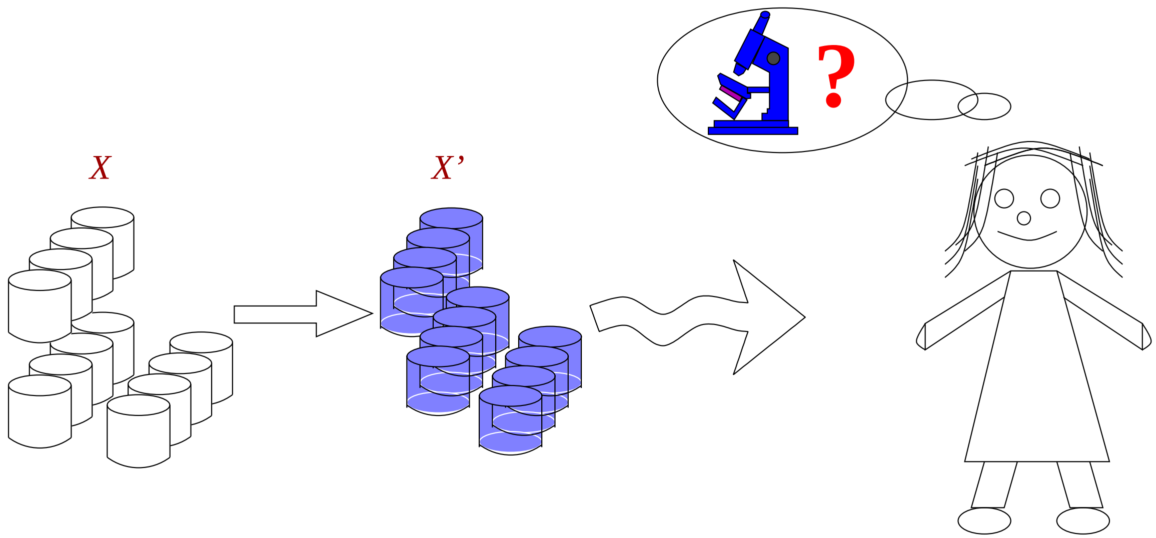
(Some) Privacy models. Computational definitions for privacy.

- **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.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.

Privacy models

Privacy models. A computational definition for privacy.

- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Result privacy.** Avoid results when an algorithm is applied to DB X

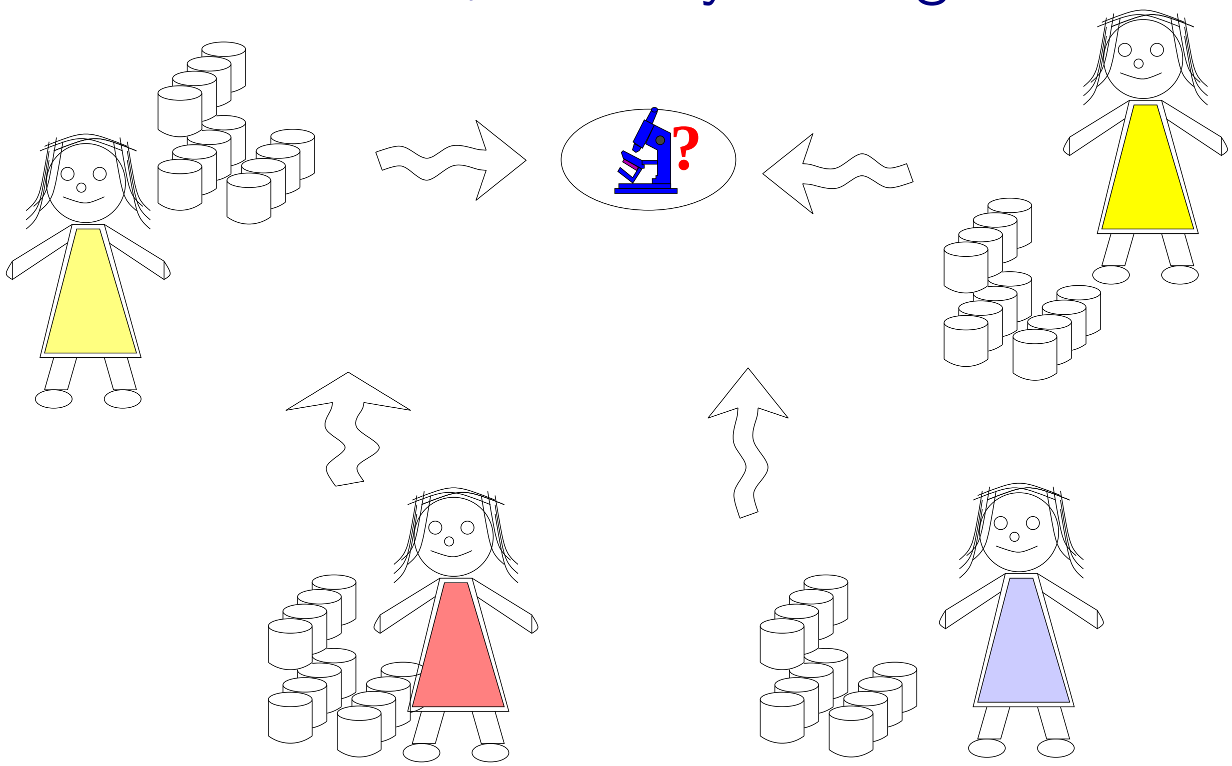


can we find  ? we don't want this possible ...

Privacy models

Privacy models. A computational definition for privacy. Examples.

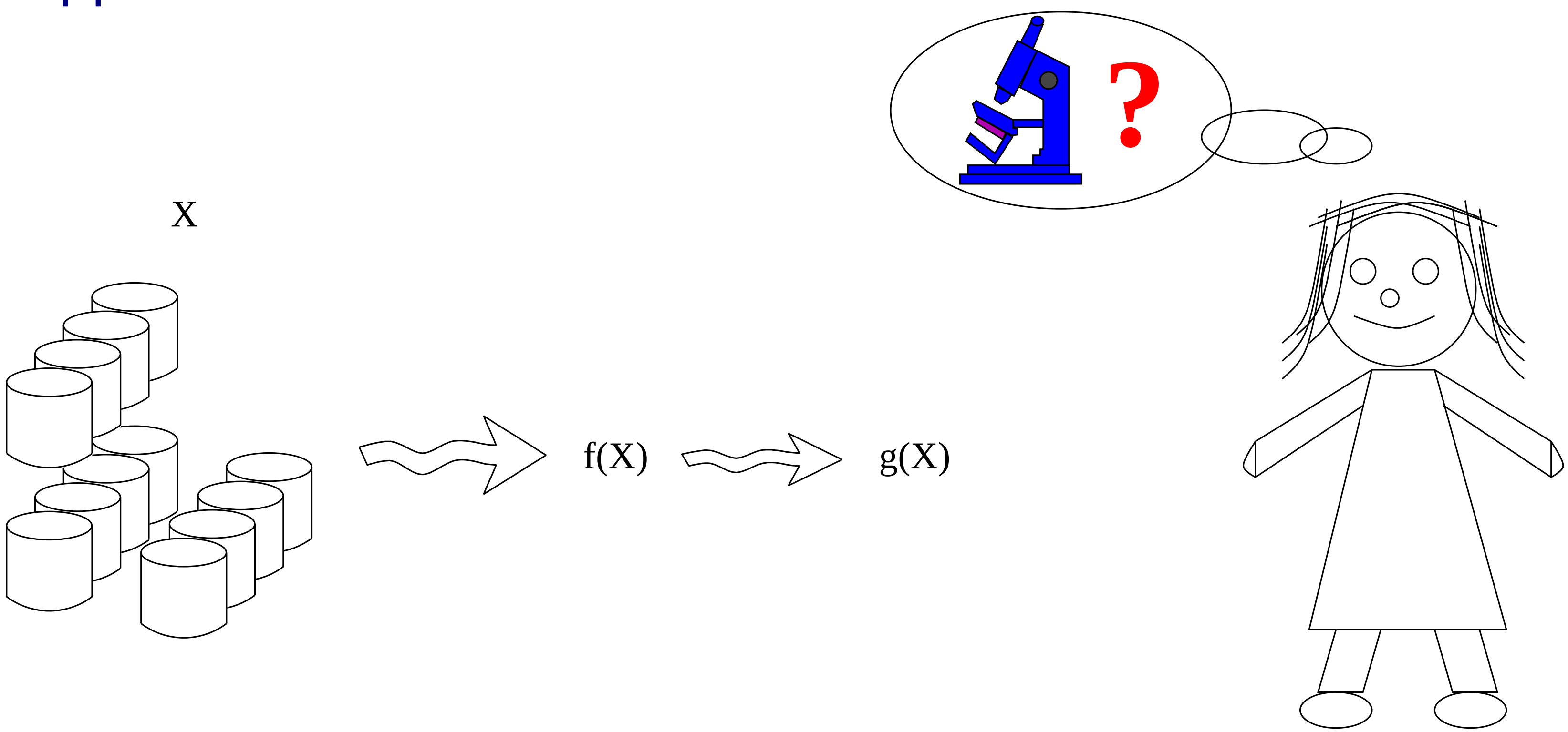
- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.



Privacy models

Privacy models. A computational definition for privacy. Examples.

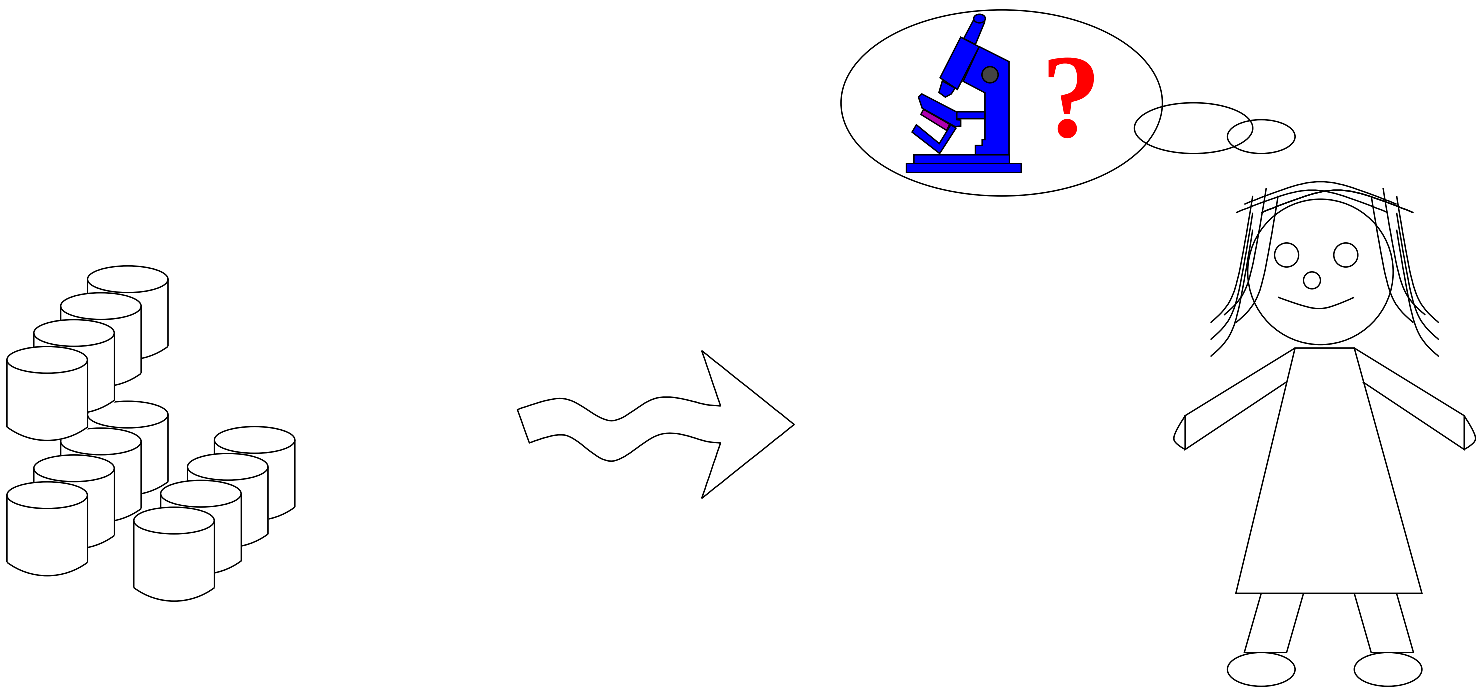
- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.



Privacy models

Our research

- privacy models: k-anonymity, differential privacy, integral privacy
- disclosure risk measures: reidentification (modeling attacks)
- data protection mechanisms: microaggregation, and others

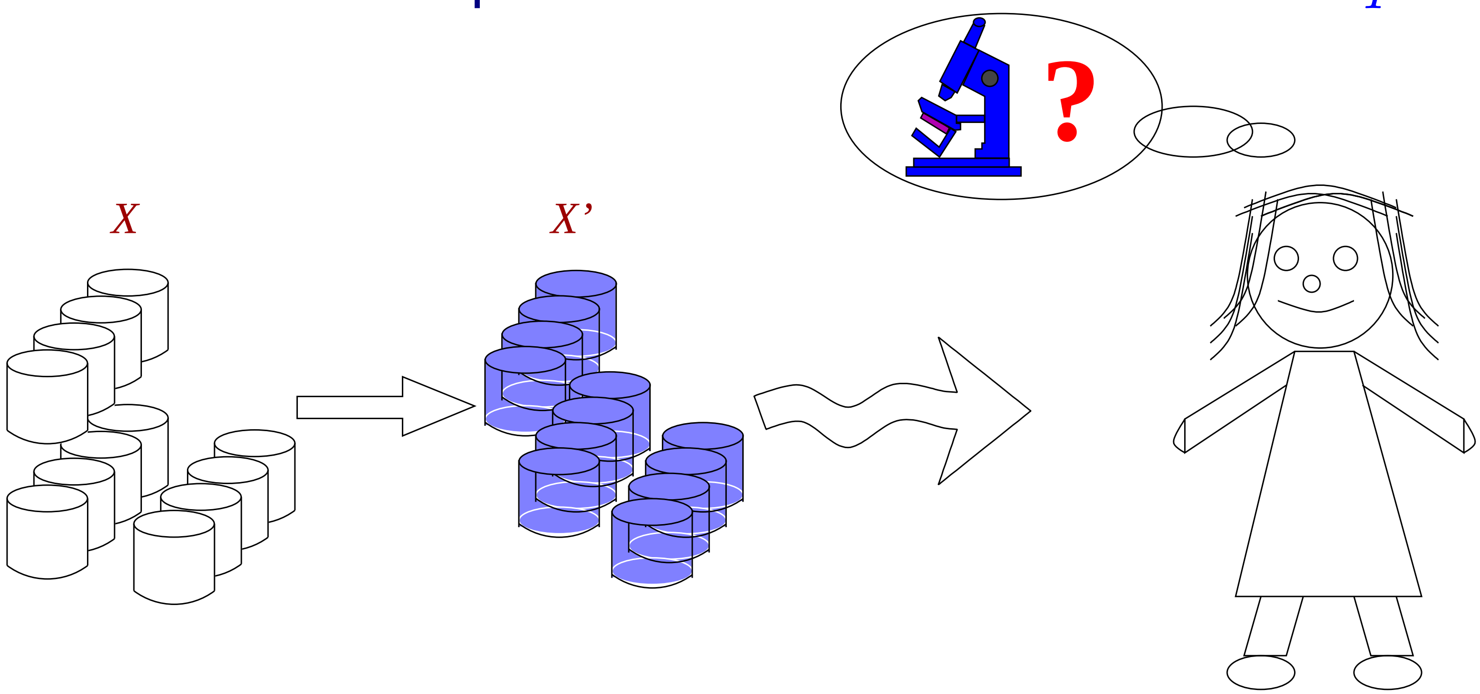


Data-driven and general purpose masking databases

Masking methods

Data-driven or general purpose (*analysis not known*)

- Privacy model: Reidentification / k-anonymity.
- Privacy mechanisms: Anonymization / masking methods: Given a data file X compute a file X' with data of *less quality*.

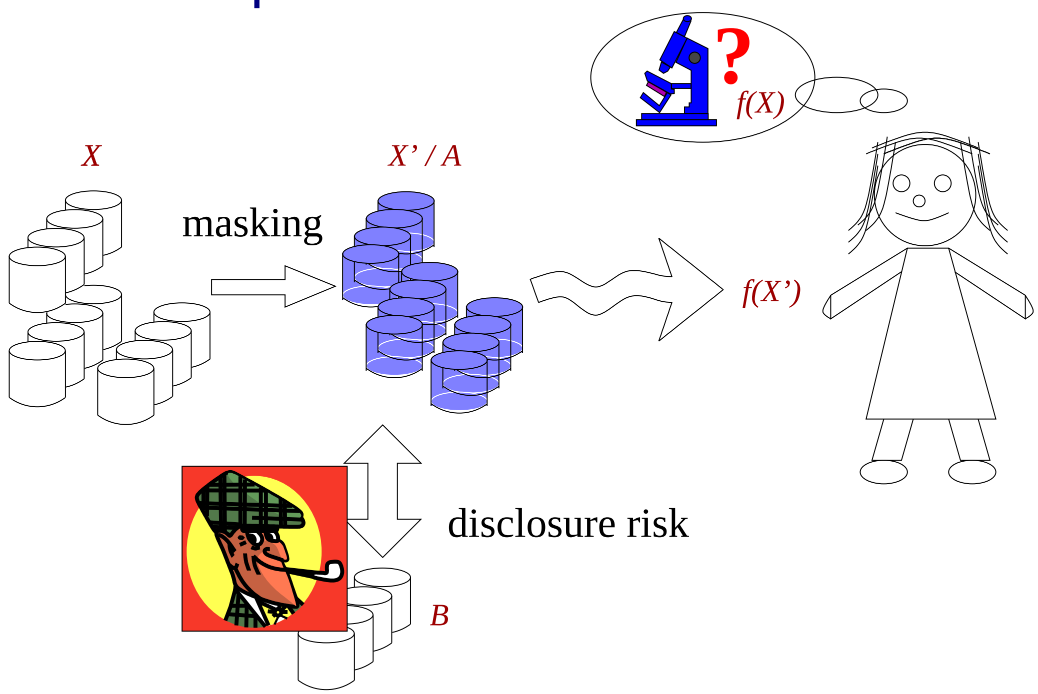


Questions: masking, less quality=information loss

Masking methods

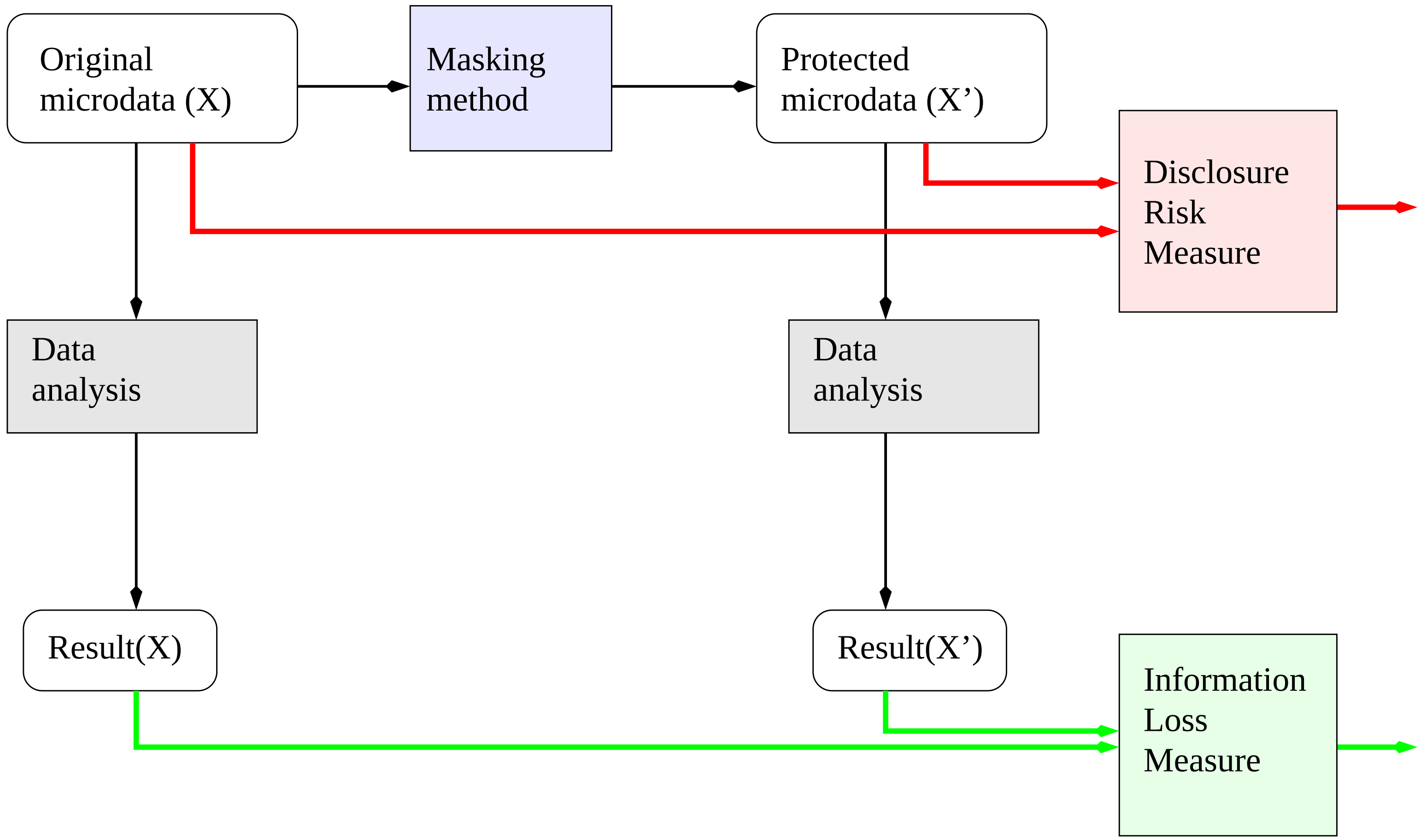
Data-driven or general purpose (*analysis not known*)

- Privacy model: Reidentification / k-anonymity.
- Privacy mechanisms: Anonymization / masking methods:
  Given a data file X compute a file X' with data of *less quality*.



Questions: masking, less quality=information loss, disclosure risk

Masking methods



Questions: masking, less quality=information loss, disclosure risk

Research questions: (i) masking methods

Masking methods. (anonymization methods) $X' = \rho(X)$

- **Privacy models**
  - **k-anonymity.** Single-objective optimization: utility
  - **Privacy from re-identification.** Multi-objective: trade-off U/Risk
- **Families of masking methods**
  - Perturbative. (less quality=erroneous data)

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

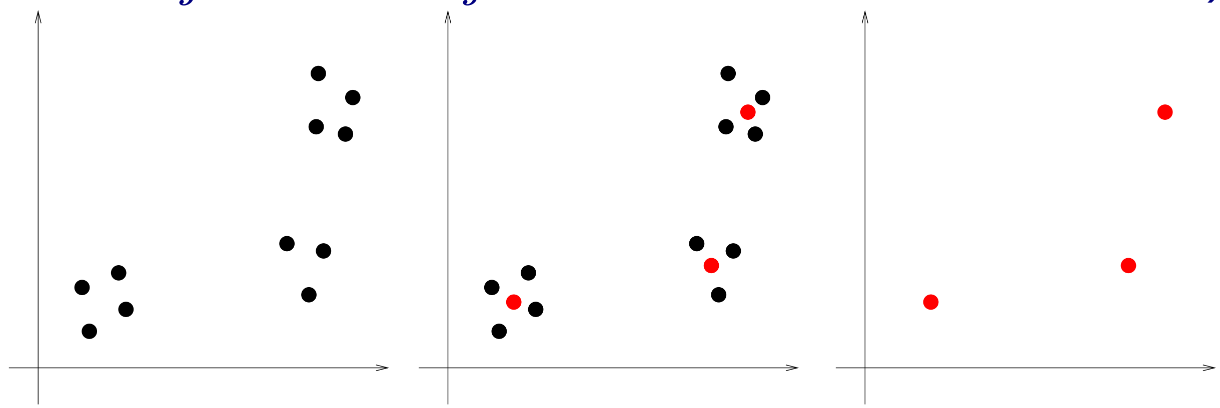
    E.g. generalization, suppression
  - Synthetic data generators. (less quality=not real data)

    E.g. (i) model from the data; (ii) generate data from model

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. Microaggregation (k records clusters)

- **Privacy models.** k-Anonymity and privacy from re-identification
- **Formalization.** $u_{ij} = 1$ iff x_j in ith cluster; v_i centroid)



Data: (age, salary)
Original cluster: {(20,1000), (21,1100), (23, 1020), (24, 1080)}
Protected one: {(22, 1050), (22, 1050), (22, 1050), (22, 1050)}

Minimize $SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2$
Subject to $\sum_{i=1}^{g} u_{ij} = 1$ for all $j = 1, \dots, n$
$2k \geq \sum_{j=1}^{n} u_{ij} \geq k$ for all $i = 1, \dots, g$
$u_{ij} \in \{0, 1\}$

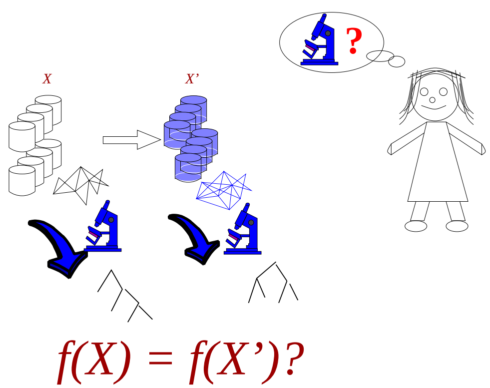
Research questions: (ii) information loss/data utility

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

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

- f: depends on X; generic vs. specific data uses.
  - Statistics, ML: clustering & classification, centrality-graphs, ...
  - For classification using decision trees $f = DT$:

$$accuracy(DT(X)) \text{ vs. } accuracy(DT(X'))$$



Research questions: (ii) information loss/data utility

- Typical comparison of methods w.r.t. IL/utility and Risk

	PIL	DR	Accuracy, ACC DT	NB	k-NN	SVM	Area Under Curve, AUC DT	NB	k-NN	SVM
Original	0.00%	100.00%	54.22%	54.78%	53.93%	54.56%	71.60%	73.30%	71.60%	70.30%
Noise, $\alpha = 3$	7.90%	74.56%	54.39%	51.81%	53.36%	54.49%	73.09%	73.41%	71.48%	70.50%
Noise, $\alpha = 10$	24.65%	38.95%	53.67%	51.88%	51.62%	54.37%	73.24%	73.42%	70.55%	70.49%
Noise, $\alpha = 100$	73.94%	4.10%	51.04%	52.21%	48.17%	53.20%	72.06%	73.98%	66.47%	69.50%
MultNoise, $\alpha = 5$	13.50%	50.81%	54.44%	51.90%	52.36%	54.39%	73.51%	73.42%	71.22%	70.50%
MultNoise, $\alpha = 10$	24.81%	24.75%	54.20%	51.76%	54.20%	54.32%	73.15%	73.42%	72.67%	70.41%
MultNoise, $\alpha = 100$	74.29%	0.00%	50.73%	52.12%	50.90%	53.27%	71.00%	73.90%	68.10%	69.52%
RS p-dist, $p = 2$	22.12%	51.12%	53.19%	51.23%	53.99%	54.37%	70.95%	73.24%	74.15%	70.57%
RS p-dist, $p = 10$	29.00%	23.49%	53.55%	51.85%	54.35%	54.18%	71.84%	73.52%	73.17%	70.40%
RS p-dist, $p = 50$	39.96%	7.80%	40.63%	50.56%	37.32%	53.20%	59.24%	73.17%	57.75%	69.50%
CBFS, $k = 5$	39.05%	13.73%	54.56%	51.64%	54.01%	54.54%	74.10%	73.29%	73.26%	70.62%
CBFS, $k = 25$	58.08%	6.65%	53.31%	51.95%	53.05%	54.01%	73.48%	73.10%	74.22%	70.23%
CBFS, $k = 100$	63.55%	4.32%	51.30%	51.59%	53.53%	54.10%	71.16%	73.24%	74.56%	70.31%
CBFS 2-sen, $k = 25$	58.08%	0.55%	53.31%	52.00%	53.05%	54.13%	73.44%	73.10%	74.22%	70.30%
CBFS 3-sen, $k = 25$	73.00%	0.00%	45.00%	42.00%	43.00%	41.00%	62.00%	61.00%	63.00%	60.00%
CBFS 2-div, $k = 25$	61.55%	0.40%	52.72%	51.57%	52.84%	54.37%	72.13%	73.24%	73.09%	70.36%
CBFS 3-div, $k = 25$	86.00%	0.00%	38.00%	39.00%	38.00%	40.00%	60.00%	61.00%	62.00%	63.00%
IPSO $g = 2$	65.09%	1.66%	52.81%	51.52%	50.11%	53.39%	72.36%	73.61%	68.06%	69.66%
IPSO $g = 3$	58.93%	4.93%	51.45%	51.09%	49.87%	52.41%	69.58%	73.22%	68.24%	68.81%
IPSO $g = 4$	58.56%	1.81%	52.05%	51.23%	50.68%	52.52%	70.41%	73.22%	68.52%	69.00%

Abalone (4177 records, 9 attr, 3 classes) w/ different SDC perturbation methods[3].

[3] Herranz, Matwin, Nin, Torra (2010) Classifying data from protected statistical datasets. C&S.

Research questions: (ii) information loss/data utility

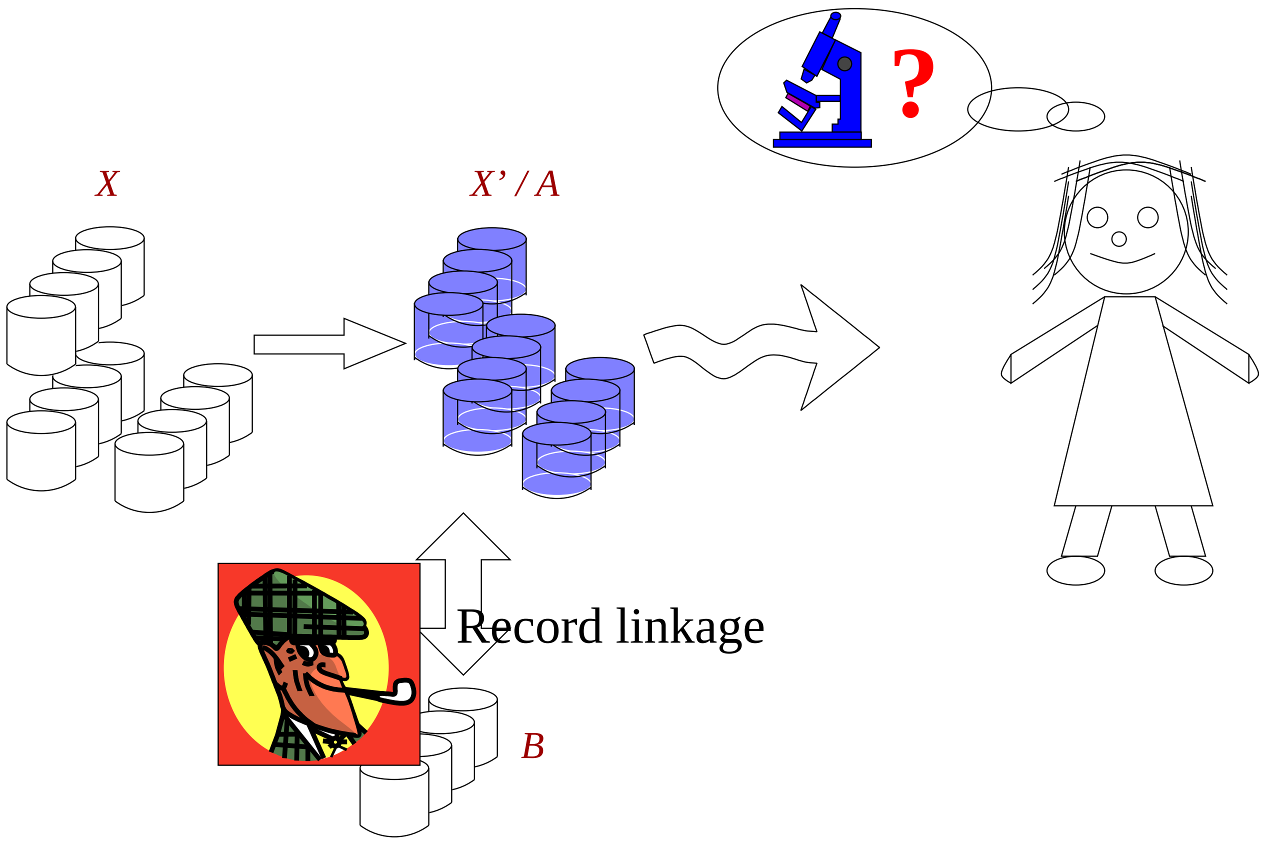
Goal of masking methods:
good trade-off information loss - disclosure risk

ML models, accuracy and masking methods

- Masking methods: not always equivalent to a loss of accuracy
  There are cases in which the performance is even improved. Aggarwal and Yu (2004) report that 'in many cases, the classification accuracy improves because of the noise reduction effects of the condensation process'. The same was concluded in [Sakuma and Osame, 2017] for recommender systems: 'we observe that the prediction accuracy of recommendations based on anonymized ratings can be better than those based on non-anonymized ratings in some settings'. [Torra, 2017]

Research questions: (iii) disclosure risk assessment

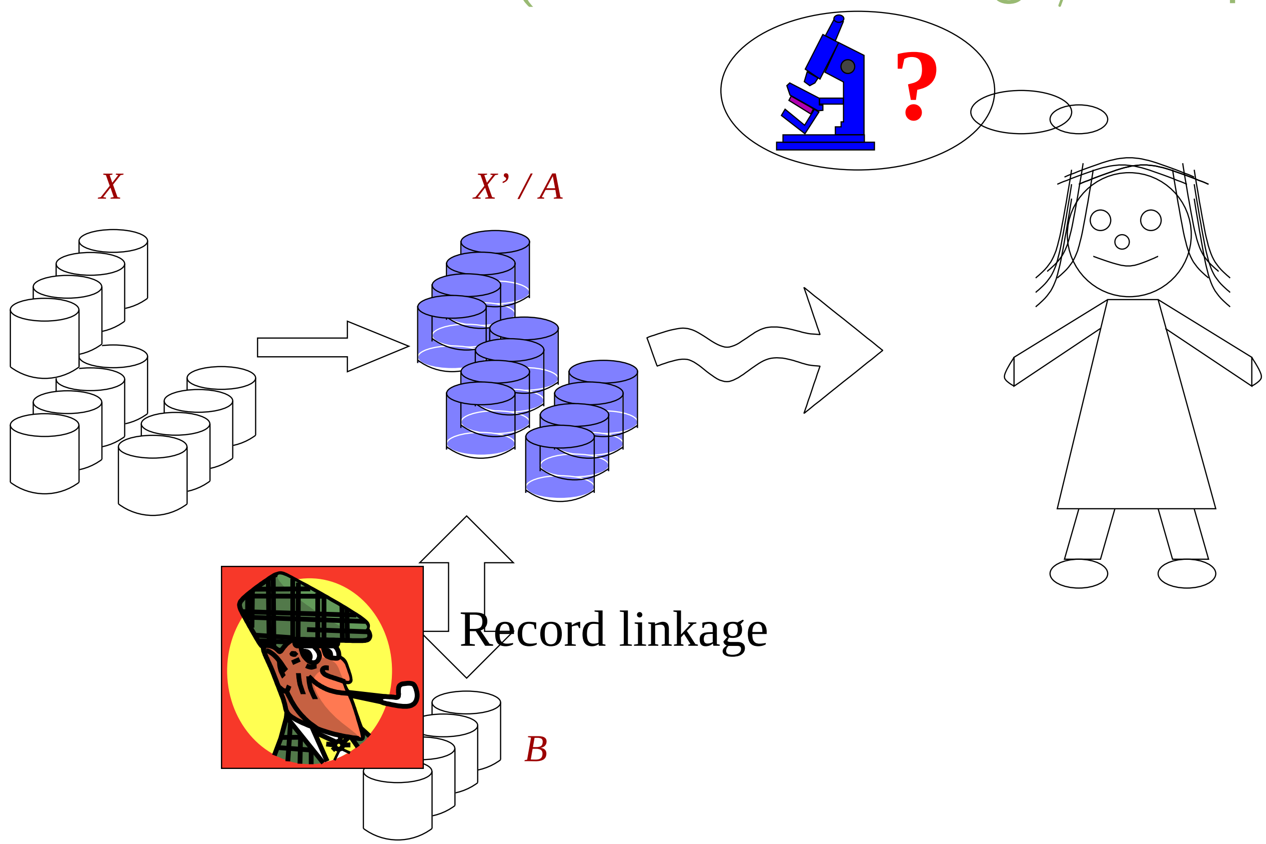
- **Privacy from re-identification**. Identity disclosure[4]. Scenario:
  - A: File with the protected data set
  - B: File with the data from the intruder (subset of original X)



[4] Identity disclosure vs. attribute disclosure: Finding Alice in DB vs. Δ knowledge on Alice's salary

Research questions: (iii) disclosure risk assessment

- **Privacy from re-identification**. Worst-case scenario (maximum knowledge) to give upper bounds of risk:
  - transparency attacks (information on how data has been protected)
  - largest data set (original data)
  - best re-identification method (best record linkage/best parameters)



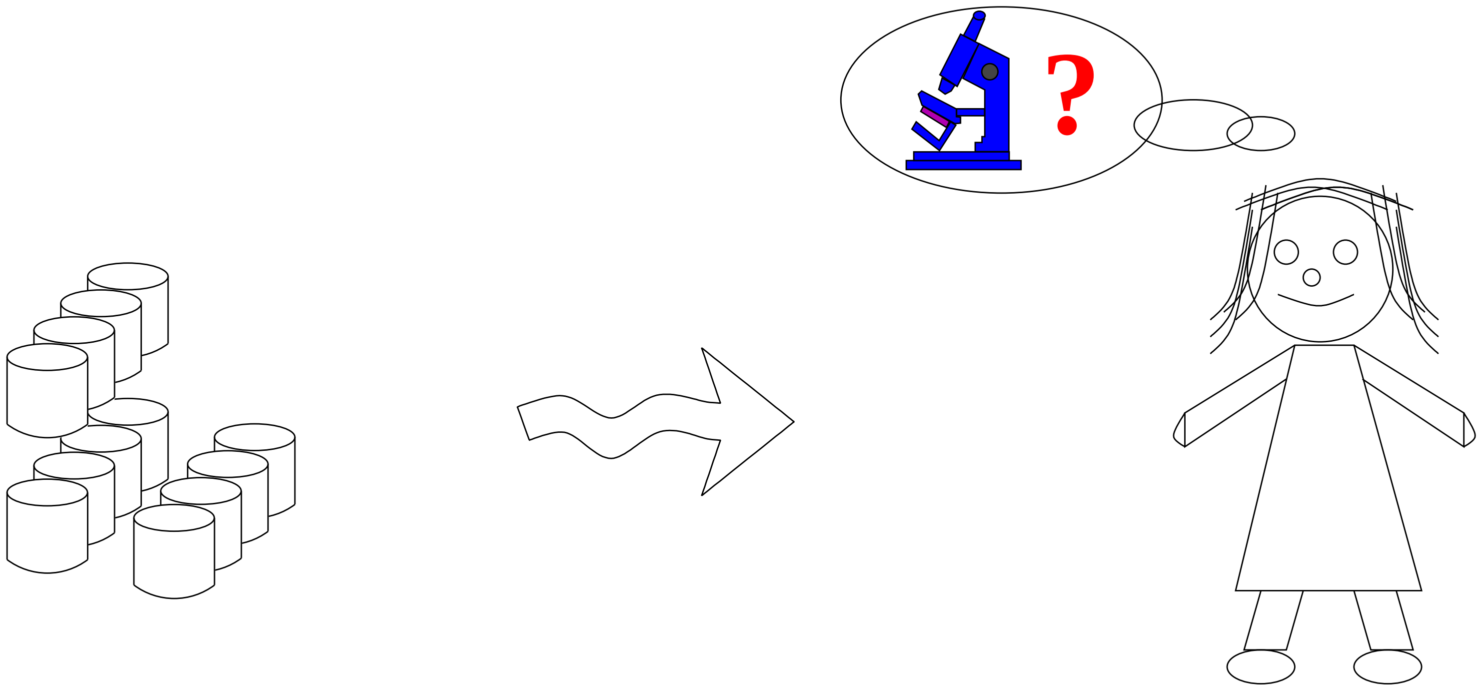
Computation-driven or specific purpose

perturb output

Integral privacy

Computation-driven or specific purpose (*analysis known*)

- Privacy model: differential, integral privacy
- Privacy mechanism: for algorithm A?



Integral privacy

Integral privacy, and differential privacy

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

Integral privacy

Integral privacy, and differential privacy

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

- Simple integrally private function:
  A an algorithm that is 1 if the number of records in D is even, and 0 if the number of records in D is odd.
  That is, $f(D) = 1$ if and only if $|D|$ is even.

Model selection in machine learning

Finding. Recurrent models[5] appear also in machine learning

- If we sample a database and build ML models (e.g., decision trees), some models appear more frequently, recurrent models

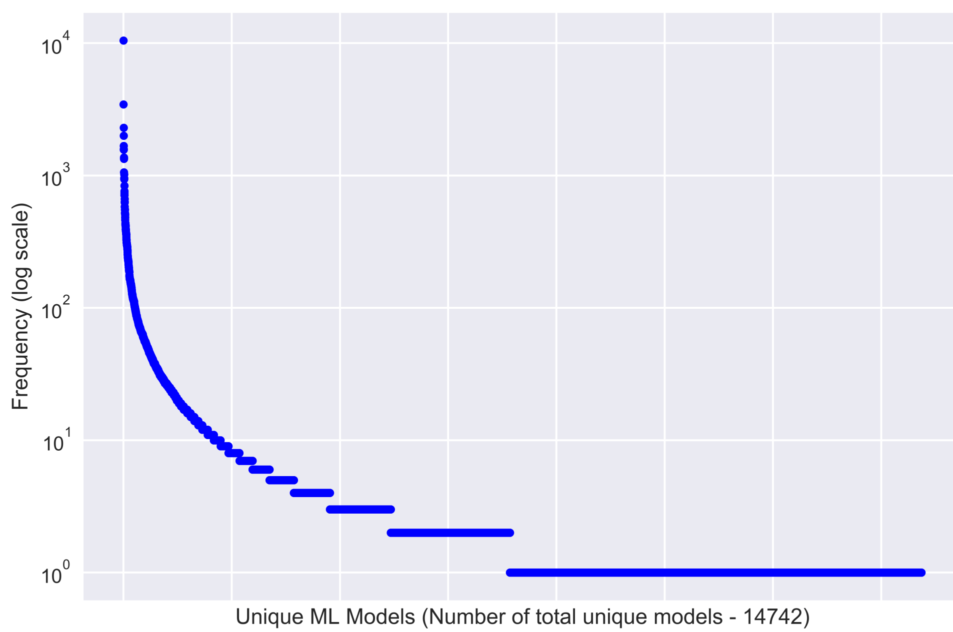
[5] Senavirathne & Torra (2019) Integrally private model selection for decision trees, Computers and Security 83 167-181

Model selection in machine learning

Finding. Recurrent models appear also in machine learning

- **Recurrent models?** Large set of generators
- **Generators?** DB generator of m_1 if $f(DB) = m_1$

Decision trees with Iris dataset. Models/freq.

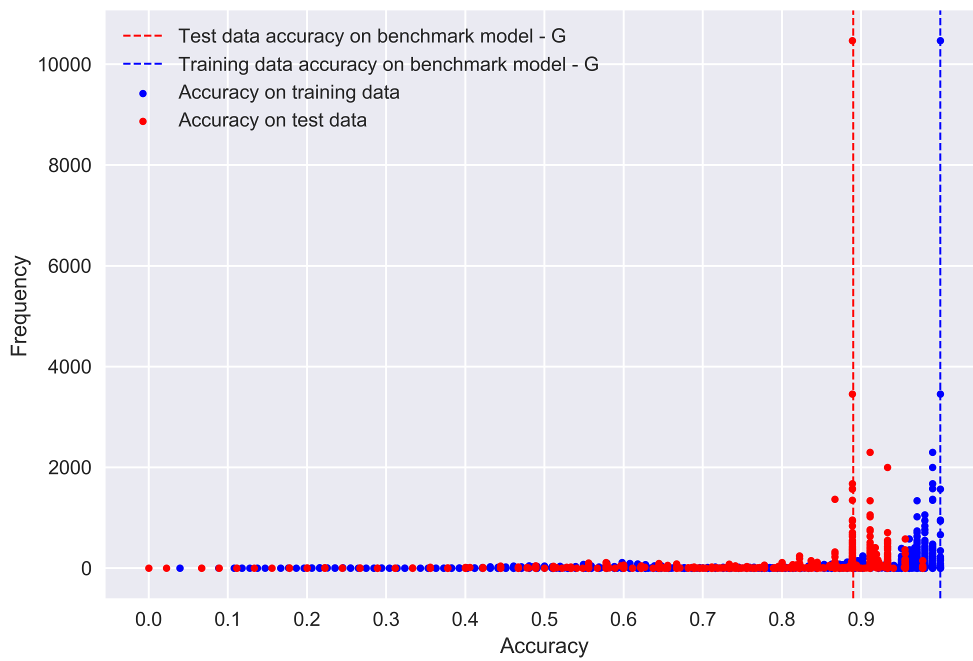


Model selection in machine learning

Finding N. 1. Recurrent models appear also in machine learning

Finding N. 2. Recurrent models may have good accuracy

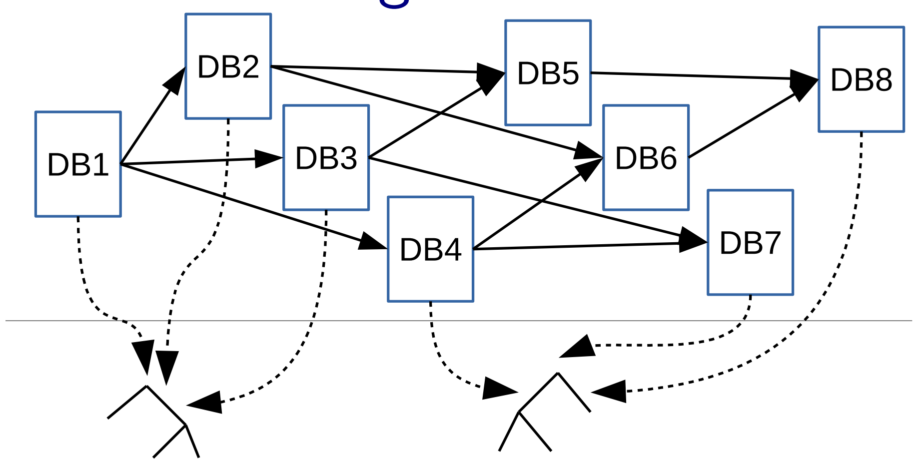
- accuracy + frequency. DT with Iris. Acc./freq.



Integral privacy

Integral privacy. (*analysis known*)

- (Original) motivation: modifications to a database (right to rectification, right to erasure)
- Goal: protect the DB and changes in the DB.



Integral privacy

- **Integral privacy** for a single database when applying an algorithm A.
  - Consider inferences on the database from the output (model).
  - Let $G \in \mathcal{G}$, A an algorithm, $S^* \subseteq P$ some background knowledge on the data set used to compute G. Integral privacy is when the set $Gen^*(G, S^*)$ is *large* and

$$\cap_{m \in Gen^*(G,S^*)} m = \emptyset.$$

- Integral privacy, and plausible deniability
  - IP satisfies plausible deniability if for any record r in P such that $r \notin S^*$, there is a set/database $\sigma \in Gen^*(G, S^*)$ such that $r \notin \sigma$.
- Our definition satisfies plausible deniability

Summary

Summary

- Data privacy
  - Naive anonymization does not work
  - Data-driven / *masking* databases
  - Computation-driven / *masking* output

Thank you

References

Related references.

- V. Torra (2017) Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer.
- http://ppdm.cat/dp/