The University of Osaka

Data privacy:
From sensitive data to
privacy preserving machine learning models

Vicenç Torra

December 2025

Umeå University, Sweden

Outline

1. Introduction
   - A context
   - Privacy for machine learning and statistics
   - Two motivating examples
2. Privacy for computations
   - Differential privacy
   - Integral privacy
   - First results: decision trees
   - Integrally private means
   - Integrally private deep learning
   - Integrally private clustering

Introduction

A context:

Data-driven machine learning/statistical models

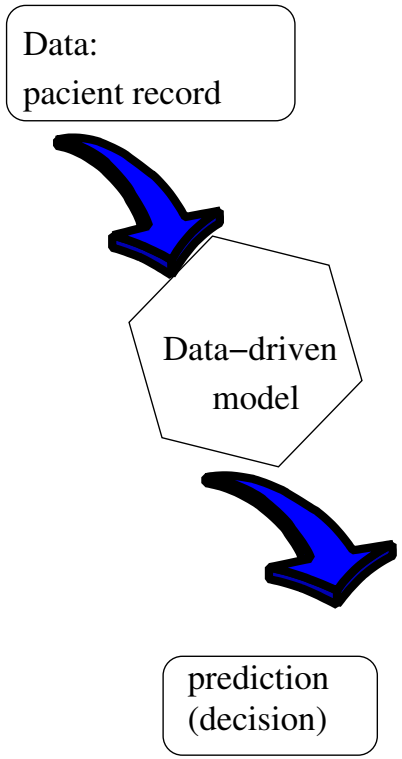
Prediction using (machine learning/statistical) models

- Data is collected to be used
  (otherwise, better not to collect them[1])

[1] Concept: Data minimization

Prediction using (machine learning/statistical) models

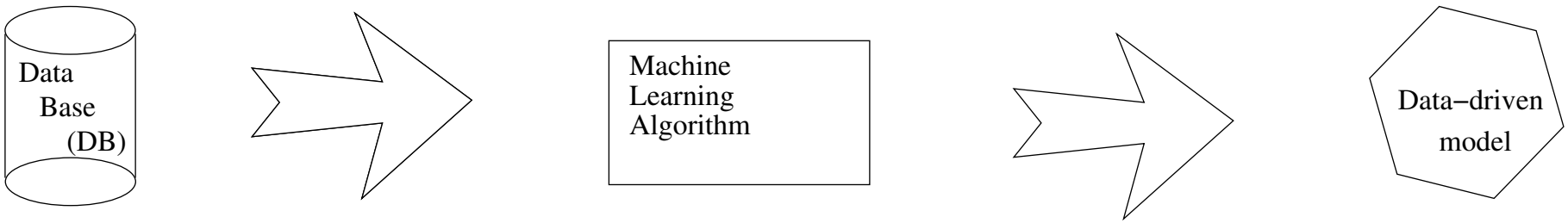
- Application of a model for decision making
  data ⇒ prediction/decision



- Example: predict the length-of-stay at admission

Data-driven machine learning/statistical models

- From (huge) databases, build the "decision maker"
  - Use (logistic) regression, deep lerning, neural networks, . . . classification algorithms, decision trees, . . .



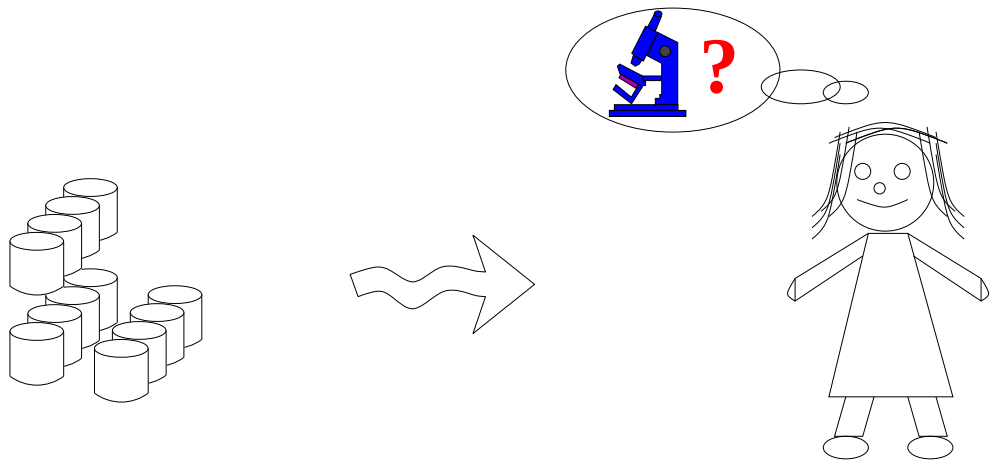
- Example: build a predictor from hospital historical data about length-of-stay at admission

Privacy for machine learning and statistics:

Data-driven machine learning/statistical models

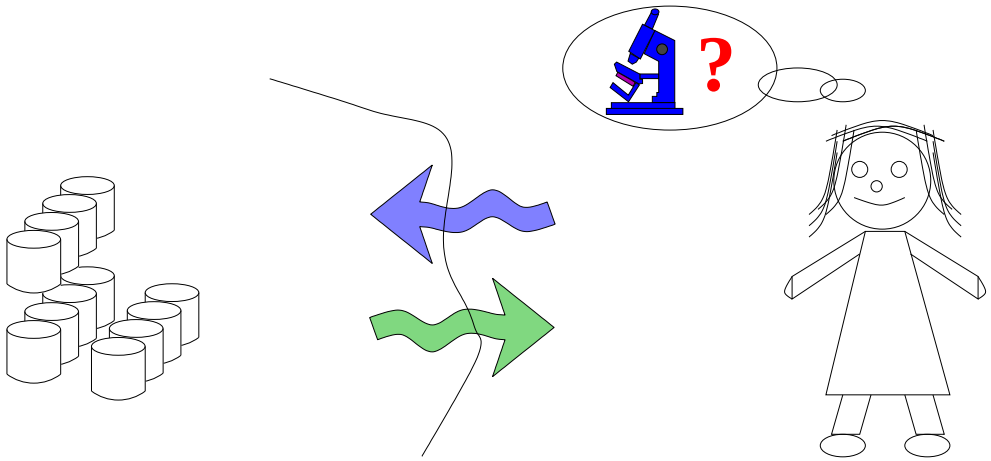
Data is sensitive

- Who/how is going to create this model (this "decision maker")?
- Case #1. Sharing (part of the data)



Data is sensitive

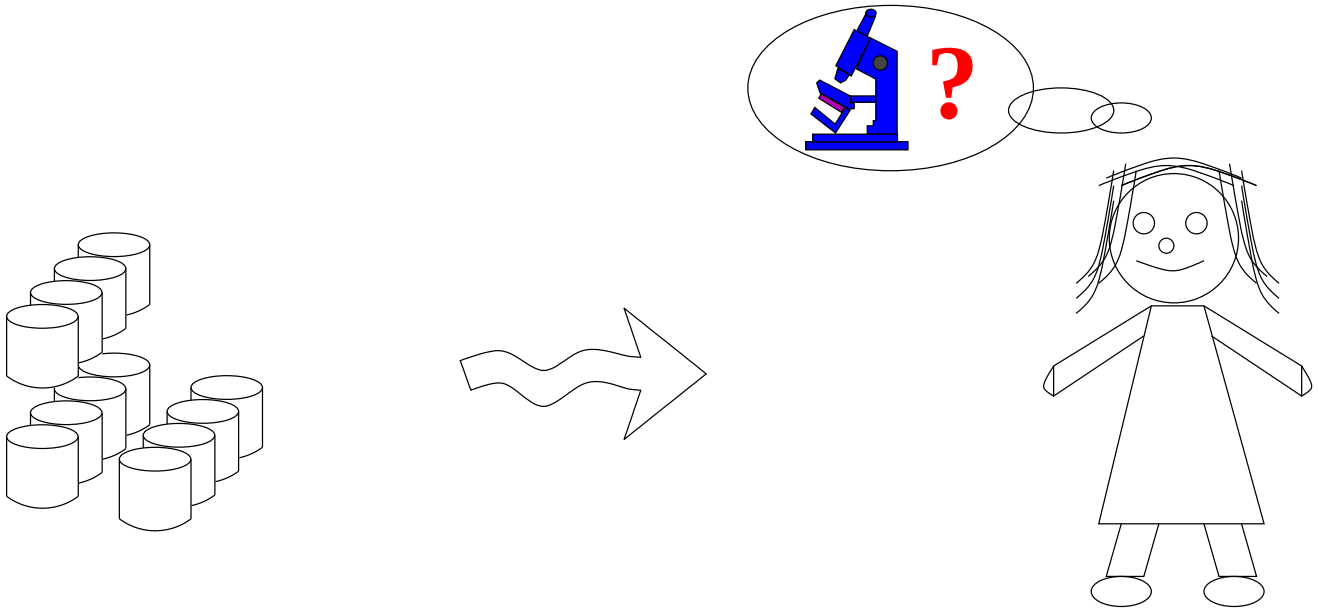
- Who/how is going to create this model (this "decision maker")?
- Case #2. Not sharing data, only querying data



Two motivating examples

Data is sensitive

- 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.

Data is sensitive

- Case #1. Sharing (part of the data)
- Q: How different children ages and diagnoses affect this length of stay? Average length of stay is decreasing in the last years due to new hospital policies?
- Data: Existing database with previous admissions (2010–2019).
  To avoid disclosure a view of the DB restricting records to children born before 2019 and only providing for these records year of birth, town, year of admission, illness, and length of stay.

Data is sensitive

- Case #1. Sharing (part of the data)

Year birth	Year Admission	Town	Illness	Length stay (days)
2017	2019	Umeå	a	3
2015	2020	Umeå	b	2
2011	2020	Luleå	c	5
2017	2019	Luleå	a	2
2016	2020	Dorotea	b	4
2016	2020	Holmöns	d	2
2015	2019	Täfteå	e	4
2015	2019	Täfteå	e	4
2015	2018	Täfteå	e	4
2015	2018	Täfteå	e	4

- Is this data safe?

Data is sensitive

- Case #1. Sharing (part of the data)

Year birth	Year Admission	Town	Illness	Length stay (days)
2017	2019	Umeå	a	3
2015	2020	Umeå	b	2
2011	2020	Luleå	c	5
2017	2019	Luleå	a	2
2016	2020	Dorotea	b	4
2016	2020	Holmöns	d	2
2015	2019	Täfteå	e	4
2015	2019	Täfteå	e	4
2015	2018	Täfteå	e	4
2015	2018	Täfteå	e	4

- Is this data safe?

Holmöns 63, Täfteå 1383, Luleå 49123, Umeå 83249, Dorotea 2366

Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?

Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?
  - Mean income is not "personal data", is this ok ? NO!!:

Data is sensitive: computation leads to disclosure

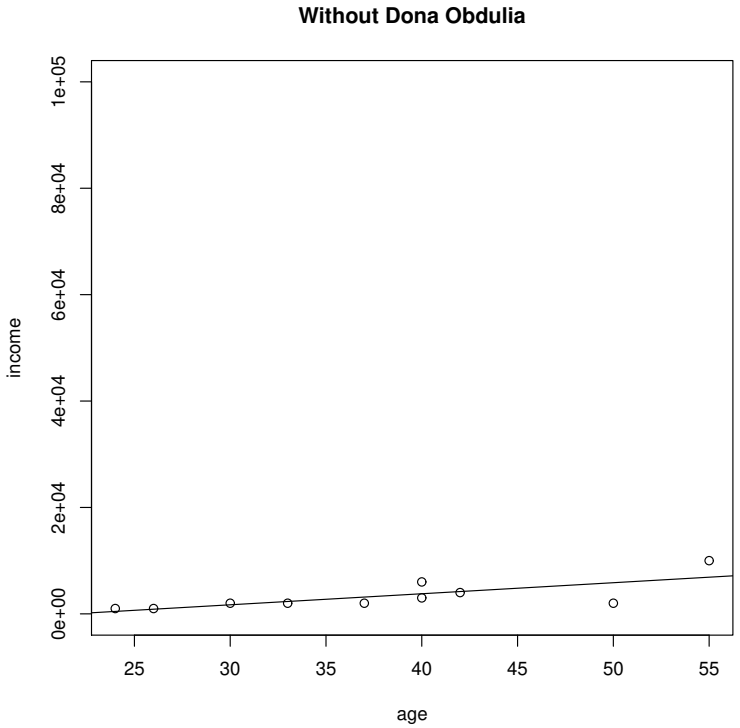
- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?
  - Mean income is not "personal data", is this ok ? NO!!:
  - Example 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300

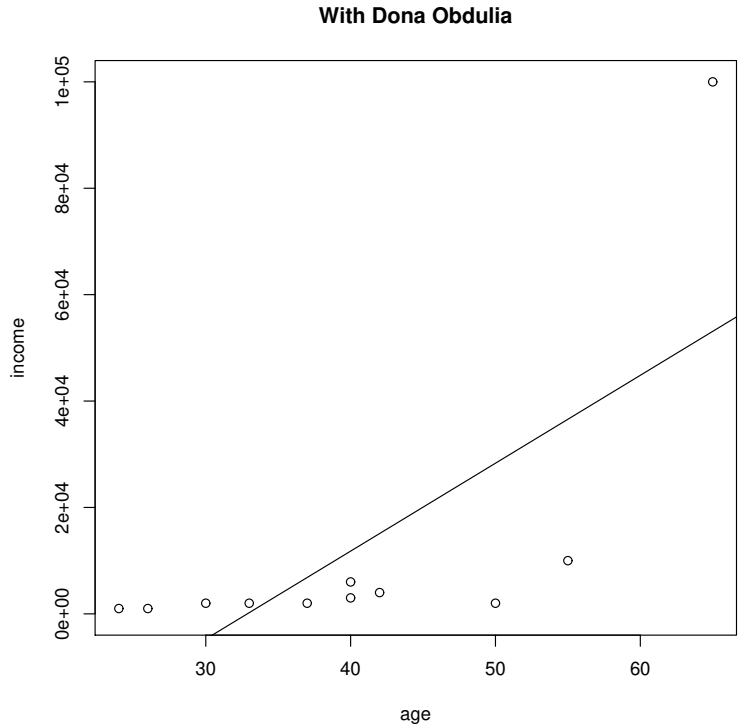
Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation.
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town (Bunyola)?
  - Mean income is not "personal data", is this ok ? NO!!:
  - Example 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

Data is sensitive: computation leads to disclosure

- Case #2. Sharing a computation. Example 2





- Regression of income with respect to age with (right) and without (left) the record of Dona Obdúlia
  - income = -4524.2 + 207.5 age (without Ms. Rich = Dona Obdúlia)
  - income = -54307 + 1652 age (with Ms. Rich = Dona Obdúlia)

Differential privacy

Differential privacy

- Differential privacy (Dwork, 2006).
  - Motivation:
    - ▻ the result of a query should not depend on the presence (or absence) of a particular individual
    - ▻ the impact of any individual in the output of the query is limited

    differential privacy ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis (Dwork, 2006)

Differential privacy

- Mathematical definition of differential privacy
  (in terms of a probability distribution on the range of the function/query)
  - A function K_q for a query q gives ϵ-differential privacy if for all data sets D_1 and D_2 differing in at most one element, and all $S \subseteq Range(K_q)$,

$$\frac{Pr[K_q(D_1) \in S]}{Pr[K_q(D_2) \in S]} \leqslant e^{\epsilon}.$$

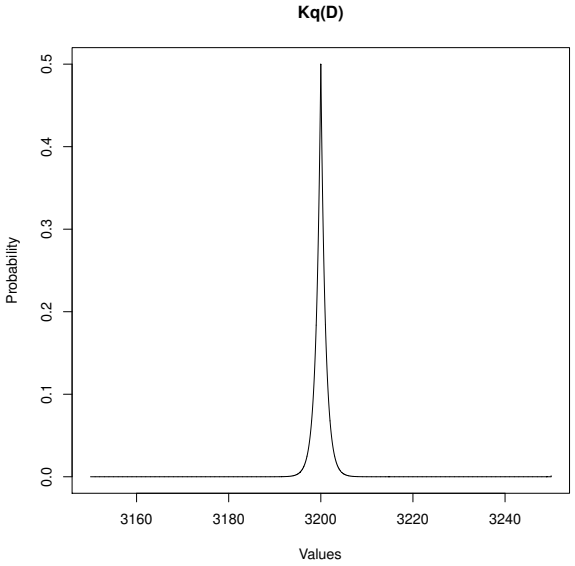
  (with 0/0=1) or, equivalently,

$$Pr[K_q(D_1) \in S] \leqslant e^{\epsilon} Pr[K_q(D_2) \in S].$$

- ϵ is the level of privacy required (privacy budget). The smaller the ϵ, the greater the privacy we have.

Differential privacy

- Differential privacy
  - A function K_q for a query q gives ϵ-differential privacy if . . .
    - ▻ $K_q(D)$ ~~is a constant. E.g.,~~
      $$~~K_q(D) = 0~~$$
    - ▻ $K_q(D)$ is a randomized version of $q(D)$:
      $$K_q(D) = q(D) + \textit{and some appropriate noise}$$



Differential privacy

- Implementation
  - For numerical data, usual to add Laplacian noise.
  - The larger the sensitivity of the function, the larger the noise
  - The larger the privacy required, the larger the noise.
- Properties
  - Plausible deniability: to an extend, in terms of ϵ

Integral privacy

Privacy models

Privacy models: for computations

- Option 1:
  - Privacy for re-identification (to data) + computation
  - k-Anonymity (to data) + computation
- Option 2:
  - Differential privacy directly to the computation
- We introduced for Option 2:
  - Integral privacy

Privacy models

Integral privacy: for a computation or algorithm f

- $f(X)$ is private if there are different ways to reach $f(X)$, i.e., different databases X which are different enough.

Integral privacy

Integral privacy: for computation f.

Some preliminaries ...

- P the population, f be a function or algorithm that given a data set $S \subseteq P$ computes an output $f(S)$ that belongs to another domain $\mathcal{G}$.
- Given G in $\mathcal{G}$, previous knowledge S^* with $S^* \subset P$, the set of possible generators of G is:

$$Gen(G, S^*) = \{S' | S^* \subseteq S' \subseteq P, f(S') = G\}.$$

We use $Gen^*(G, S^*) = \{S' \backslash S^* | S^* \subseteq S' \subseteq P, f(S') = G\}$
(when no information is known on S^*, we use $S^* = \varnothing$

Integral privacy

Integral privacy: for function f, definition:

- P data, $f : S \to \mathcal{G}$, S^* background knowledge, $Gen(G, S^*)$ databases that generate G and are consistent with background knowledge S^*. Then, integral privacy is satisfied when $Gen(G, S^*)$ is large and diverse.

Integral privacy

Integral privacy: for function f, definition:

- P data, $f : S \to \mathcal{G}$, S^* background knowledge, $Gen(G, S^*)$ databases that generate G and are consistent with background knowledge S^*. Then, integral privacy is satisfied when $Gen(G, S^*)$ is large=at least k databases and diverse:

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

Requirements: why? / what?

- Empty intersection to avoid all generators sharing a record (e.g., avoiding membership inference attacks)
- $Gen(G, S^*)$ large. large = k-flavor.

Integral privacy vs Differential privacy

Integral privacy, and differential privacy

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

Integral privacy vs Differential privacy

Integral privacy, and differential privacy

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

- Simple integrally private function:
  f 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.

Integral privacy vs Differential privacy

Pros and cons:

- Cons:
  - If S^* is all population P but a record. Not "strong" guarantees.
- Pros:
  - 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

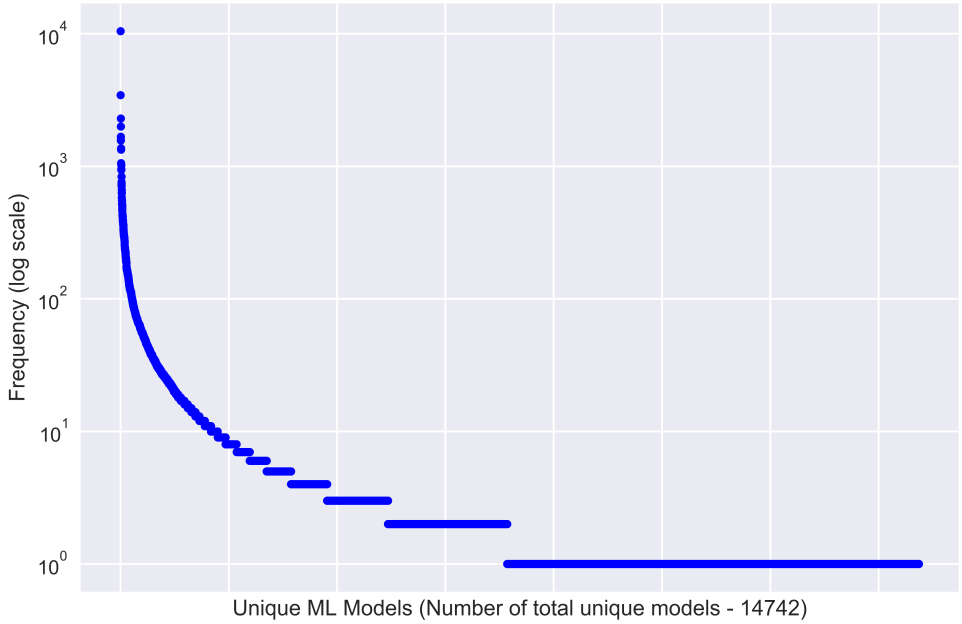
First results: decision trees

Integral privacy

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.

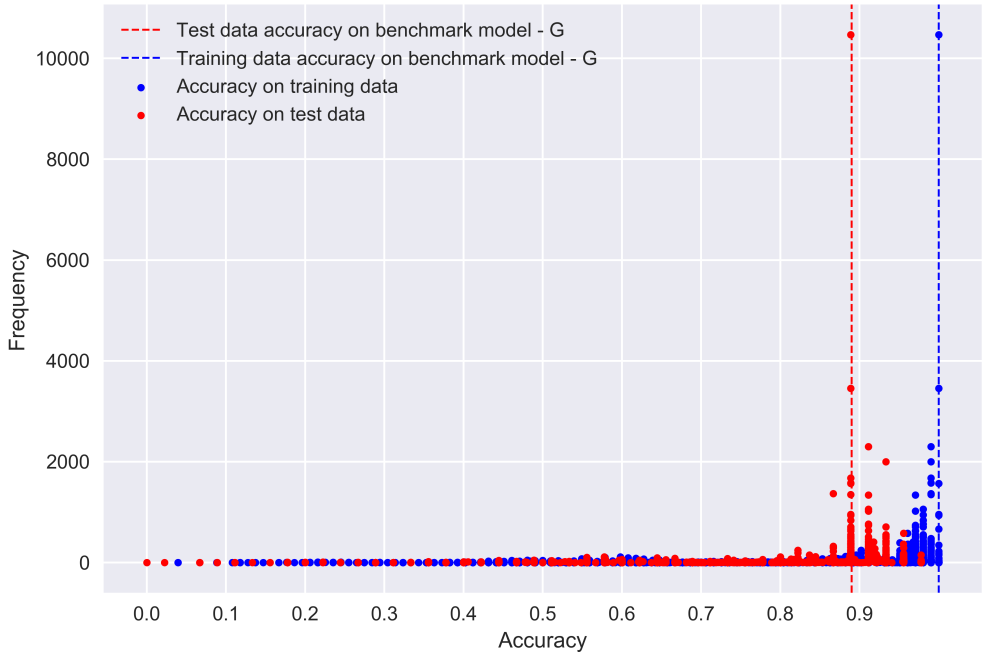


Integral privacy

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

How to implement ML models

- Sampling the database (DT)
  - Create databases from the original database
  - Create models m for each database db
    (db = generator of m)
  - Compare models and generators
- Partition the database (SVM, DL)
  - Create a database from each part
  - Create models
  - If the models are the same, by construction they satisfy the privacy constraints
    (or models similar enough)

Integrally private means

Integral privacy

How to implement IP mean (numerical database)

- Round numbers in the database
  - All number multiples of r
- Sample the database and build subsets
- Compute means of subsets
- Take a frequent mean such that satisfies the privacy constraints
  E.g., at least k generators with empty intersection

Integral privacy

How to implement IP mean (numerical database)

- k is a privacy requirement, and relates to distortion
  - larger k, larger distortion
- Larger r in rounding, larger distortion
- Amount of distortion also depends on the query
  - See mean vs. maximum / minimum
    (to produce the same $maximum$ we will need larger rounding)

Integrally private deep learning

Integrally private deep learning

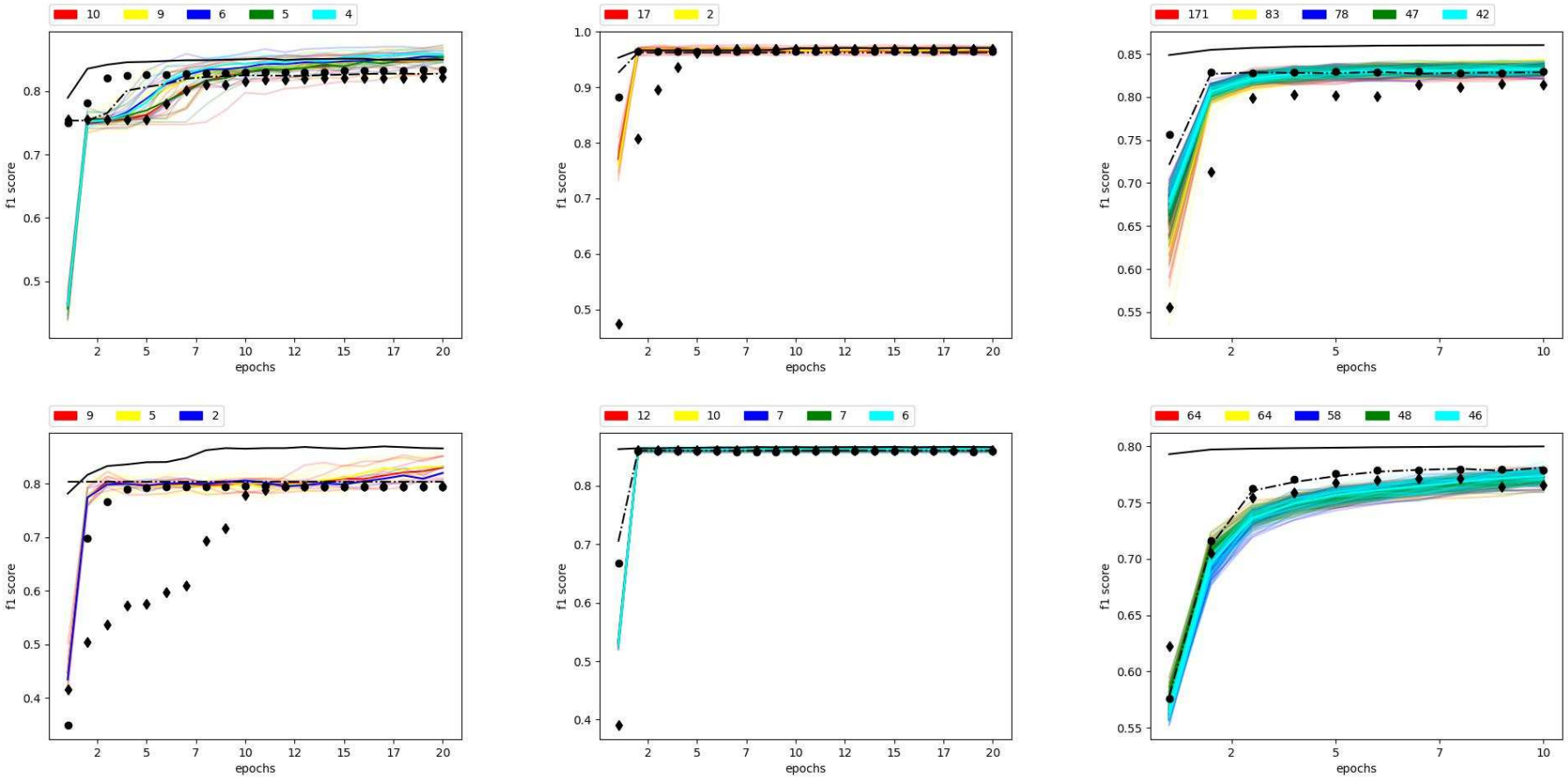
- Approach
  - Generate n subsamples each of size N having the same class-distribution as the original.
  - Compute models and cluster them so that each cluster has models that are utmost ϵ different from each other.
  - Choose a cluster of models with recurrent models and high utility.
  - Return the mean of these models.

Integrally private deep learning

- Experiments
  - Data sets: Adult, Susy, ai4i, HepMass, Churn_Modelling, Diabetes. (Churn Modelling and Adult: categorical data; Diabetes is multi-class; both small ($\sim$ 10-50K instances), medium ($\sim$ 250K instances) and large datasets ($\sim$ 5-7 million instances)
  - Subsamples between 500 and 10000.
  - An architecture of 5-layered DNN with 3-hidden layers with 5-10-5 neurons.
    (considered other architectures as well)
  - Taken $\epsilon = 0.05$ for all datasets

F1 score of top 5 recurring models (training data)

- Adult, ai4i, HepMass, Churn Modelling, Diabetes, Susy Datasets
  - high ($\epsilon \sim 0.1$, ◆), moderate ($\epsilon \sim 0.5$, · –, low privacy ($\epsilon \sim 1.0$, •)



Applications of integrally private neural networkse

- Results for concept drift detection

- Machine unlearning
  These results are based on the fact that integrally private solutions provide plausible deniability

Integrally private clustering:

κ-centroid c-means

Formalization of the problem

- Same underlying idea, but for clustering
- Different approach:
  - Integrally private solution by construction

Formalization of the problem

Informal description:

- Database X
- Macro-clusters: c
- Micro-clusters: κ

So, $c \times \kappa$ disjoint groups or parts

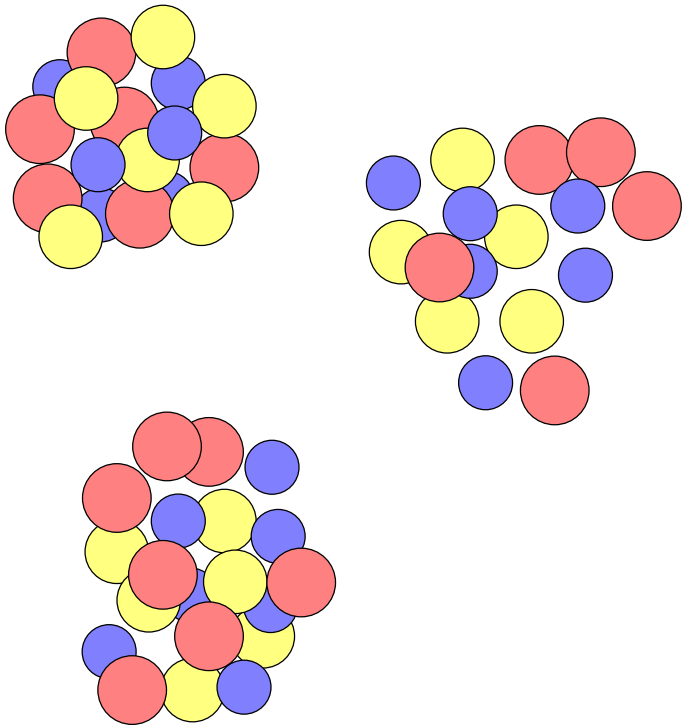
- Macro-clusters are distinct and distant
- Micro-clusters of a macro-cluster are similar and overlapping in the data space

Formalization of the problem

Data and parameters:

- Database X
- Macro-clusters: c
- Micro-clusters: κ

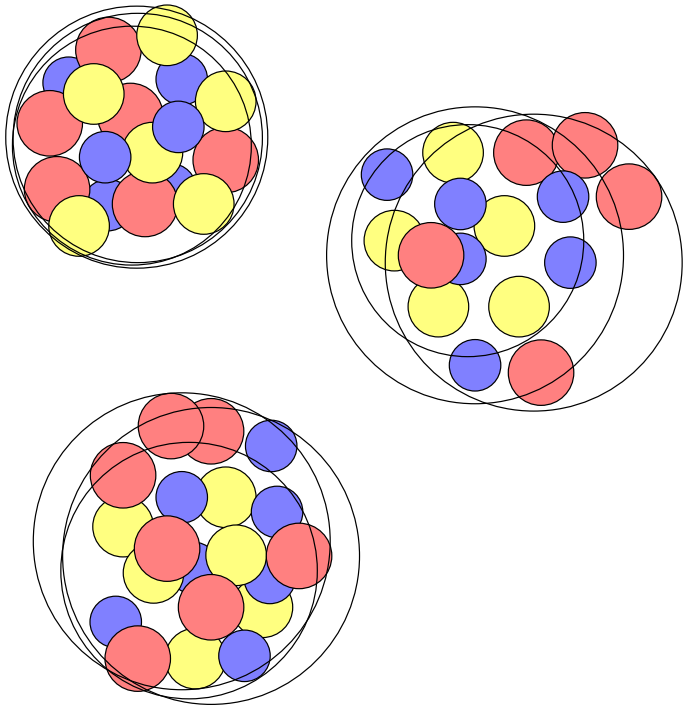
Data points and clusters:



Formalization of the problem

Notation

- centroids: v_{jk} for $j = 1, \ldots, c$ and $k = 1, \ldots, \kappa$ be the centroid of kth micro-centroid of the jth macro-cluster.
- assignment: $\mu_{jk}(x)$ represent the membership of x to the kth micro-centroid of the jth macro-cluster. We assume $\mu_{jk} \in \{0, 1\}$.



Formalization of the problem

Parameters: X,
A (difference on number of records), δ (distance for centroids)

$$\min J(\mu, v) = \sum_{j=1}^{c} \sum_{k=1}^{\kappa} \sum_{x \in X} \mu_{jk}(x) ||x - v_{jk}||^2$$

subject to $\sum_{j=1}^{c} \sum_{k=1}^{\kappa} \mu_{jk}(x) = 1$ for all $x \in X$

$$|\sum_{x \in X} \mu_{jk_1}(x) - \sum_{x \in X} \mu_{jk_2}(x)| \leqslant A$$
for all $j \in \{1, \ldots, c\}$, $k_1 \neq k_2 \in \{1, \ldots, \kappa\}$

$$||v_{jk_1} - v_{jk_2}||^2 \leqslant \delta$$
for all $j \in \{1, \ldots, c\}$, $k_1 \neq k_2 \in \{1, \ldots, \kappa\}$

$$\mu_{jk}(x) \in \{0, 1\}$$
for all $j \in \{1, \ldots, c\}$, $k \in \{1, \ldots, \kappa\}$, and $x \in X$

Experiments

Implementation:

- (Clustering +) Genetic algorithms
- MDAV to produce k-size clusters
  so all clusters have the same number of records,
  better partition of macro-clusters into micro-clusters
  (better approximation of δ)

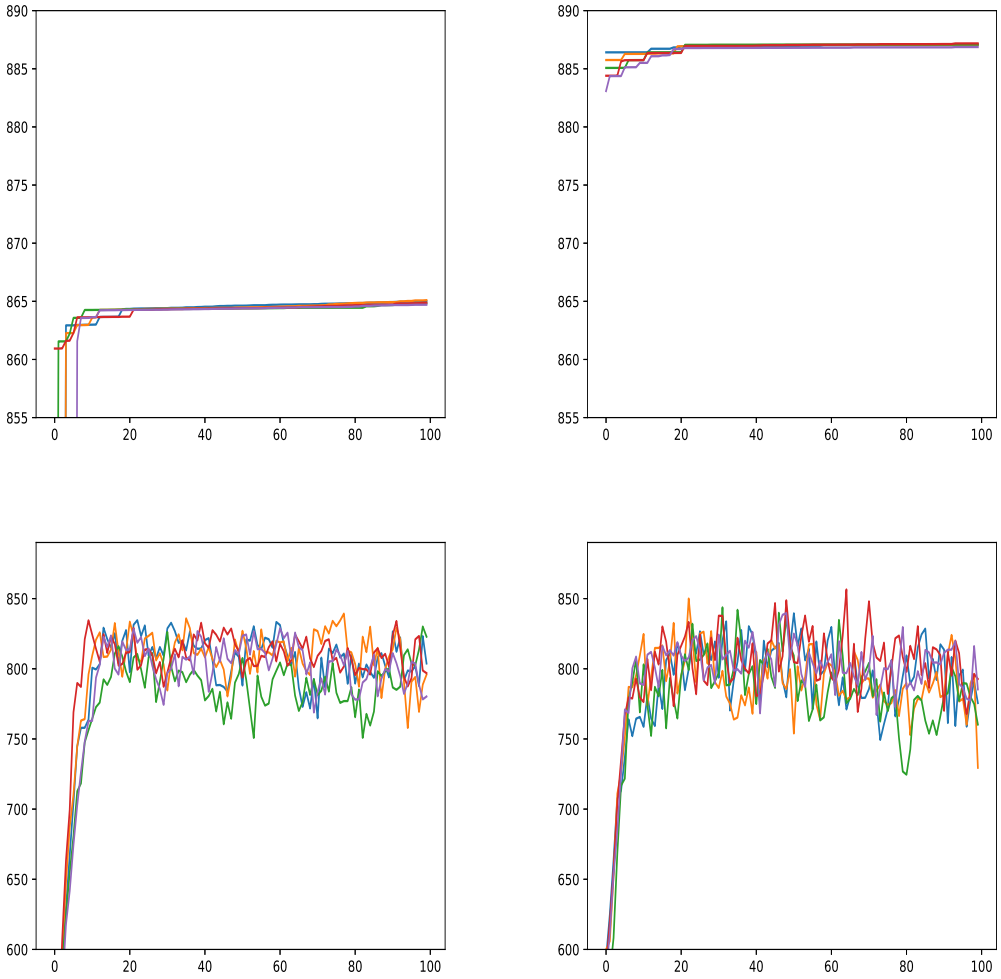
Parameters: $\delta = 0.0005$, $A = 5$
(5 runs, 100 epochs; $c = 2, \kappa = 3$ also $c = 4, \kappa = 10$)

Dataset: Concrete and CASC

Experiments

Example: Concrete, $c = 2, \kappa = 3$

(best top,mean bottom; random (left) and MDAV (right)



Discussion

Discussion:

- Solution satisfies integral privacy constraints (κ parts with empty intersection); but,
- the optimization with $\kappa \neq 1$ and the full dataset X, and a reduced problem (say X_k) with one of the subsets, may lead to different results; but,
  - separated enough clusters will produce same results for X and X_k,
  - clustering algorithms lead to local optimal,
- So, maybe good enough?

Discussion

Discussion:

- Database changes. We want models that do not change.

Thank you

References

- A. K. Varshney, V. Torra: Efficient federated unlearning under plausible deniability. Mach. Learn. 114(1): 25 (2025)
- Y. Kanzawa, V. Torra: Integrally private model construction via optimization: a case in linear regression, USB Proceedings MDAI 2025.
- Y. Kanzawa, V. Torra: On Some Variants for Anticlustering. SCIS/ISIS 2024: 1-6
- A. K. Varshney, V. Torra: Integrally Private Model Selection for Deep Neural Networks. DEXA (2) 2023: 408-422
- A. K. Varshney, V. Torra: Concept Drift Detection Using Ensemble of Integrally Private Models. PKDD/ECML Workshops (5) 2023: 290-304
- V. Torra, G. Navarro-Arribas, E. Galván: Explaining Recurrent Machine Learning Models: Integral Privacy Revisited. PSD 2020: 62-73
- N. Senavirathne, V. Torra: Integrally private model selection for decision trees. Comput. Secur. 83: 167-181 (2019)
- N. Senavirathne, V. Torra: Integral Privacy Compliant Statistics Computation. DPM/CBT@ESORICS 2019: 22-38
- N. Senavirathne, V. Torra: Approximating Robust Linear Regression With An Integral Privacy Guarantee. PST 2018: 1-10
- V. Torra, G. Navarro-Arribas: Integral Privacy. CANS 2016: 661-669