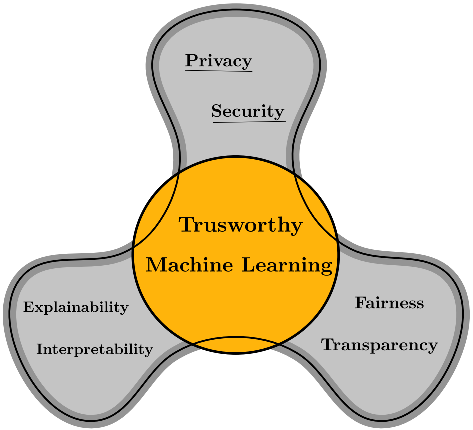
Projet IA
Database Privacy in Modern Computer Science
A gentle introduction to k-anonymity and differential privacy

Alexandre VÉRINE - Blaise DELATTRE

Université Paris Dauphine - PSL

December 11, 2025

Trustworthy Machine Learning



- Massive use of machine learning algorithms raises major issues.
- Industries and governments **have to** treat this issues (GDPR 2018).
- This course focuses on the notions of **Security** and **Privacy**.

Main questions today: How modern computer science can be a treat for individuals privacy? How can practitioners address this issue?

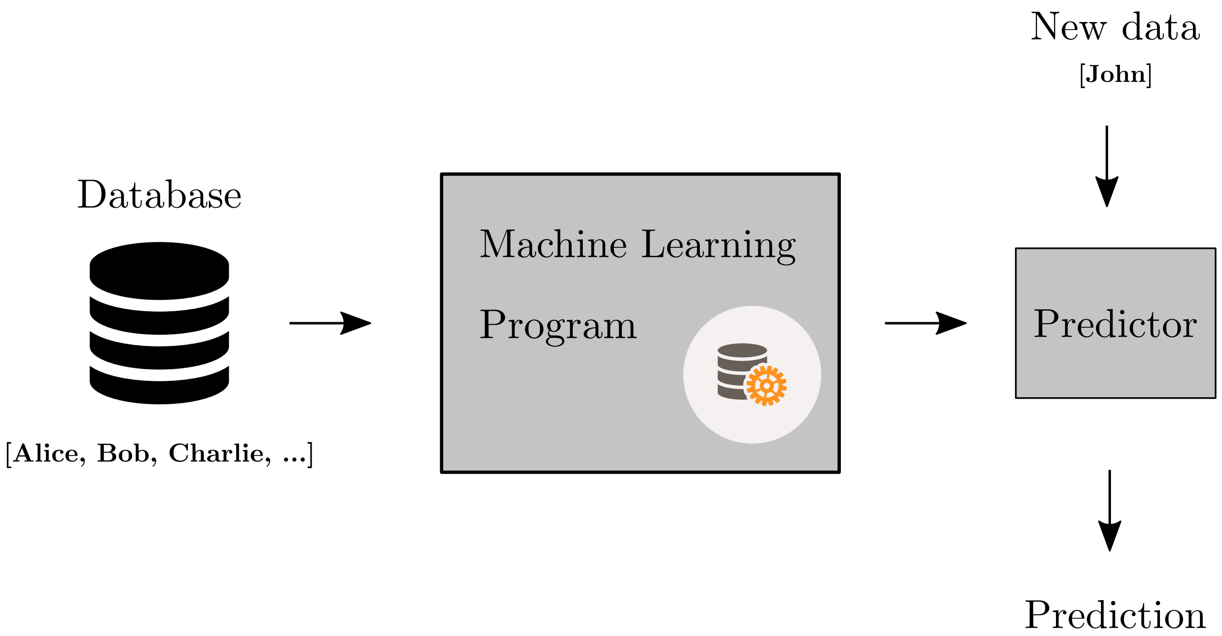
Table of contents

1. Modern computer science: a threat for individual privacy? 4

2. k-anonymity / l-diversity / t-closeness 10

3. Differential privacy: a principled privacy preserving theory 17

Modern computer science is data driven



Sensitive databases

- Databases are massively used in many sensitive domains e.g. **cyber-security, banking, healthcare.**
- **Healthcare:** One want to know that smoking causes cancer, but not that Alice smokes/has a cancer.

Question: Is machine learning compatible with privacy?

Classical data anonymization does not preserve privacy

- Some algorithms (e.g. KNN or SVM) release directly the database including sensitive characteristics. This is obviously not private!
- To protect privacy, a workaround is to remove from the database any information which trivially identifies an individual such as “name” and “social security number” fields, etc. **(not always a good idea)**
- **Sweeney**: *"87% of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}"*

Advanced data anonymization does not work either

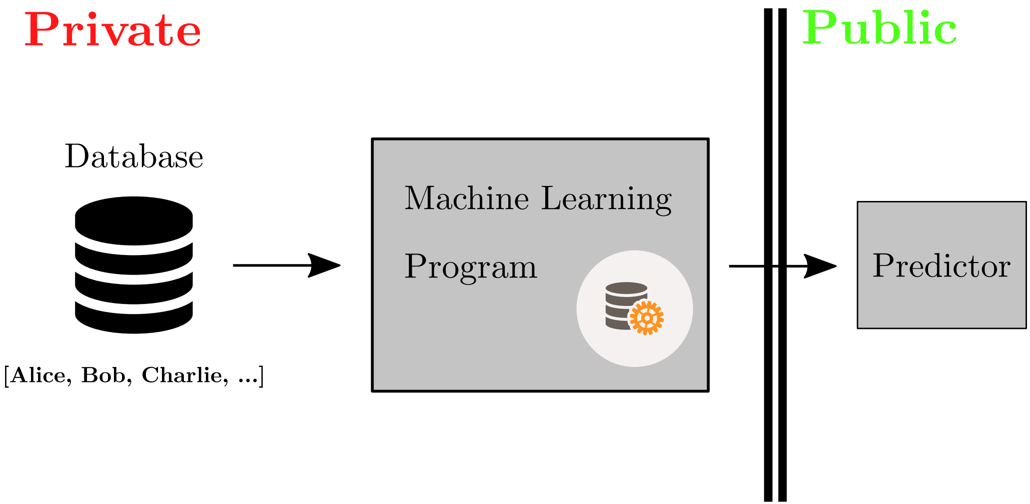
So called advanced anonymization:

- Unique identifiers removed
- Shuffle of some characteristics
- Modification of some ratings.

Narayanan et al. showed that few movie ratings suffice to uniquely identify anonymized users from Netflix prize dataset by linking with IMDB publicly available database. This was a subject of lawsuits and had a major impact on Netflix's privacy policy.

No access to the database is not sufficient

If there is no anonymization possible, users should not have access to the database (the database is **private**), but only to a **public** predictor.



This defense has also been broken by **membership inference attacks.**

How to protect the database ?

Let us suppose that the Adversary can recover the database. We want to pre-process it such that still keeps individual privacy.

Name	Age	Sex	Smoker ?
Alice	20	F	1
Bob	22	M	1
Charlie	21	F	0
Dona	21	F	0
Ernest	50	M	0
Fred	57	M	0
Grace	55	F	0
Henry	62	M	0

- Solution 1: k-anonymity $\rightarrow$ l-diversity $\rightarrow$ t-closeness
- Solution 2: Differential privacy

Definition of k-anonymity

A database is said k-anonymous each row is identical with at least k-1 other rows (excluding the sensitive columns).

To do so we suppress or **generalize** rows until the condition is met:

1. Select some feature(s) of interest, e.g. age
2. Group the rows according to the selected feature(s)
3. Unify the attributes within each group. For example:
   - Replace the numerical attributes by the median within the group
   - Replace the categorical attribute by the set of existing categories within the group

Getting k-anonymity with k=2

Name	Age	Sex	Smoker ?
Alice	20	F	1
Bob	22	M	1
Charlie	21	F	0
Dona	21	F	0
Ernest	50	M	0
Fred	57	M	0
Grace	55	F	0
Henry	62	M	0

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	1
Charlie	21	F	0
Dona	21	F	0
Ernest	52.5	{F,M}	0
Fred	59.5	M	0
Grace	52.5	{F,M}	0
Henry	59.5	M	0

Name	Age	Sex	Smoker ?
Alice	20	F	1
Bob	22	M	1
Charlie	21	F	0
Dona	21	F	0
Ernest	50	M	0
Fred	57	M	0
Grace	55	F	0
Henry	62	M	0

Getting k-anonymity with k=4

Name	Age	Sex	Smoker ?
Alice	20	F	1
Bob	22	M	1
Charlie	21	F	0
Dona	21	F	0
Ernest	50	M	0
Fred	57	M	0
Grace	55	F	0
Henry	62	M	0

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	1
Charlie	21	{F,M}	0
Dona	21	{F,M}	0
Ernest	56	{F,M}	0
Fred	56	{F,M}	0
Grace	56	{F,M}	0
Henry	56	{F,M}	0

Name	Age	Sex	Smoker ?
Alice	20	F	1
Bob	22	M	1
Charlie	21	F	0
Dona	21	F	0
Ernest	50	M	0
Fred	57	M	0
Grace	55	F	0
Henry	62	M	0

From k-anonymity to l-diversity

l-diversity is an **extension** of the k-anonymity that handles the sensitive column(s). A database (that is already k-anonymous) is said to be l-diverse if each group has l distinct values for the sensitive field.

- Here we only have 2 possible values hence we can only have l=1 or 2
- l diversity can be implemented two ways:
  - Use l-diversity as another constraint for k-anonymity (notebook)
  - Replace arbitrarily some values to force the condition

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	1
Charlie	21	{F,M}	0
Dona	21	{F,M}	0
Ernest	56	{F,M}	1
Fred	56	{F,M}	0
Grace	56	{F,M}	0
Henry	56	{F,M}	0

From l-diversity to t-closeness

t-closeness is another **extension** of the k-anonymity that also handles the sensitive column(s). A database is said to be t-close if the **distance** between the distribution of the sensitive attribute within any group and the distribution of the attribute in the whole table is no more than a threshold t.

Example of distance Let us for example use the Total variation distance:

$$TV(P_{overall}, P_{group}) = 1/2\left(|P_{overall}(0) - P_{group}(0)| + |P_{overall}(1) - P_{group}(1)|\right)$$

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	1
Charlie	21	{F,M}	0
Dona	21	{F,M}	0
Ernest	52.5	{F,M}	0
Fred	59.5	{F,M}	0
Grace	52.5	{F,M}	0
Henry	59.5	{F,M}	0

- $P_{overall}(0) = 3/4,\ P_{overall}(1) = 1/4$
- $P_{group1}(0) = 1/2,\ P_{group1}(1) = 1/2$
- $TV(P_{overall}, P_{group1}) = 1/4$

T-CLOSENESS, WHAT T TO CHOOSE ?

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	1
Charlie	21	{F,M}	0
Dona	21	{F,M}	0
Ernest	52.5	{F,M}	0
Fred	59.5	{F,M}	0
Grace	52.5	{F,M}	0
Henry	59.5	{F,M}	0

$t \geq 1/4$

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	0
Charlie	21	{F,M}	0
Dona	21	{F,M}	0
Ernest	56	{F,M}	1
Fred	56	{F,M}	0
Grace	56	{F,M}	0
Henry	56	{F,M}	0

t small.

Conclusion on k-anonymity/l-diversity/t-closeness

- Simple concepts that can be easy to apply
- How to choose k/l/t? Not easy to trade-off privacy and accuracy
- No formal guarrantees
- To go further: Differential Privacy.

Informal definition of differential privacy

- The adversary can recover the database.
- To protect the individuals, the database should be "the same" when Alice is in the database and when she is not.
- Differential Privacy formalize this notion

"The outcome of any analysis is essentially equally likely, independent of whether any individuals joins, or refrains from joining the database". Cynthia Dwork, 2006.

Randomized algorithm

A **randomized algorithm** $\mathcal{A}$ is an algorithm that outputs a random variable instead of deterministic values.

Private database d

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	1
Bob	21	{F,M}	1
Charlie	21	{F,M}	0
Dona	21	{F,M}	0
Ernest	52.5	{F,M}	0
Fred	59.5	{F,M}	0
Grace	52.5	{F,M}	0
Henry	59.5	{F,M}	0

$\mathcal{A}$

Public resulting database $\mathcal{A}(d)$

Name	Age	Sex	Smoker ?
Alice	21	{F,M}	X_1
Bob	21	{F,M}	X_2
Charlie	21	{F,M}	X_3
Dona	21	{F,M}	X_4
Ernest	56	{F,M}	X_5
Fred	56	{F,M}	X_6
Grace	56	{F,M}	X_7
Henry	56	{F,M}	X_8

Where $X_1, \dots, X_n$ are Bernoulli random variable

Formal definition of differential privacy

A randomized algorithm $\mathcal{A}$ is called ϵ-**differentially private** if for any $S \subset Range(\mathcal{A})$ and for all pair of database $d \sim d'$:

$$\mathbb{P}\left[\mathcal{A}(d) \in S\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}(d') \in S\right]$$

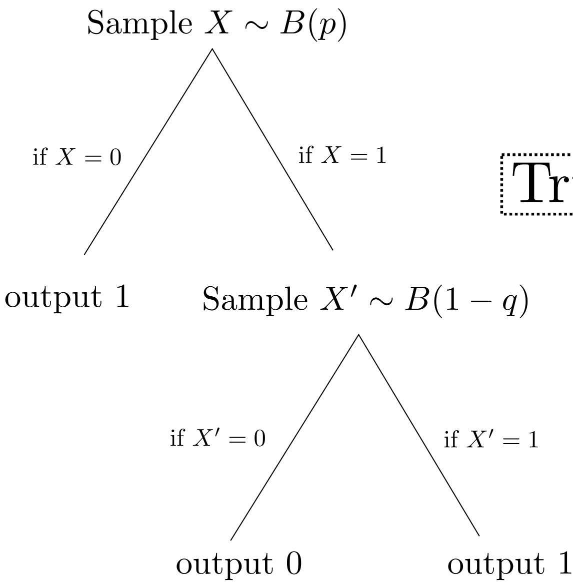
$d \sim d'$ means that d and d' differ at most from one individual.

- What value of ϵ is good? Typically ≤ 1 is good (but be careful).
- How can we craft $\mathcal{A}$ to have differential privacy? Randomized response.

Randomized response

For each row:

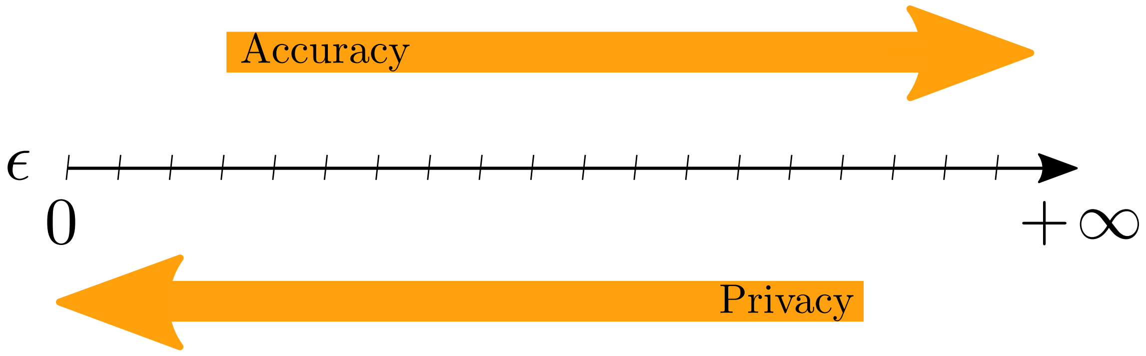
- With probability 1-p, we leave the true sensitive value.
- With probability p, we change the value and set it to 1 with probability q and 0 with probability 1−q.



$\boxed{\text{True value for Alice: 1}}$

No free lunch: trade-off accuracy/privacy

- Randomized response is ϵ-differential privacy, with $\epsilon = -ln(pq)$.
- Other noise injection trick exist (e.g. Laplace mechanism)
- Privacy is not free, but one can certify some level of privacy/accuracy.



It simply states that accuracy and privacy are not trivially combined.

Take-home message

- **Trustworthy Machine Learning**, is rapidly gaining in interest.
- It is **hard** but **not impossible** to release private databases.
- k-anonymity and its extensions are a good start but do not provide formal guarrantees.
- **Differential privacy** is a theoretically grounded framework for privacy preserving data management.
- It preserves privacy at the **expense of some controlled loss of accuracy.**