Publication de données respectueuse de la vie privée

Attaques VS Défense

Tristan Allard
Univ. Rennes, IRISA
tristan.allard@irisa.fr

La gouvernance des données massives critiques
Association Aristote
Le 11 Juin 2026
École polytechnique, Plateau de Saclay

Progress of the Talk

Ubiquitous personal data collection

Figure: Our lives: (1) stored in DBMSs for primary usages and (2) feeding analytics and/or machine learning algorithms (secondary usages).

Ubiquitous personal data collection

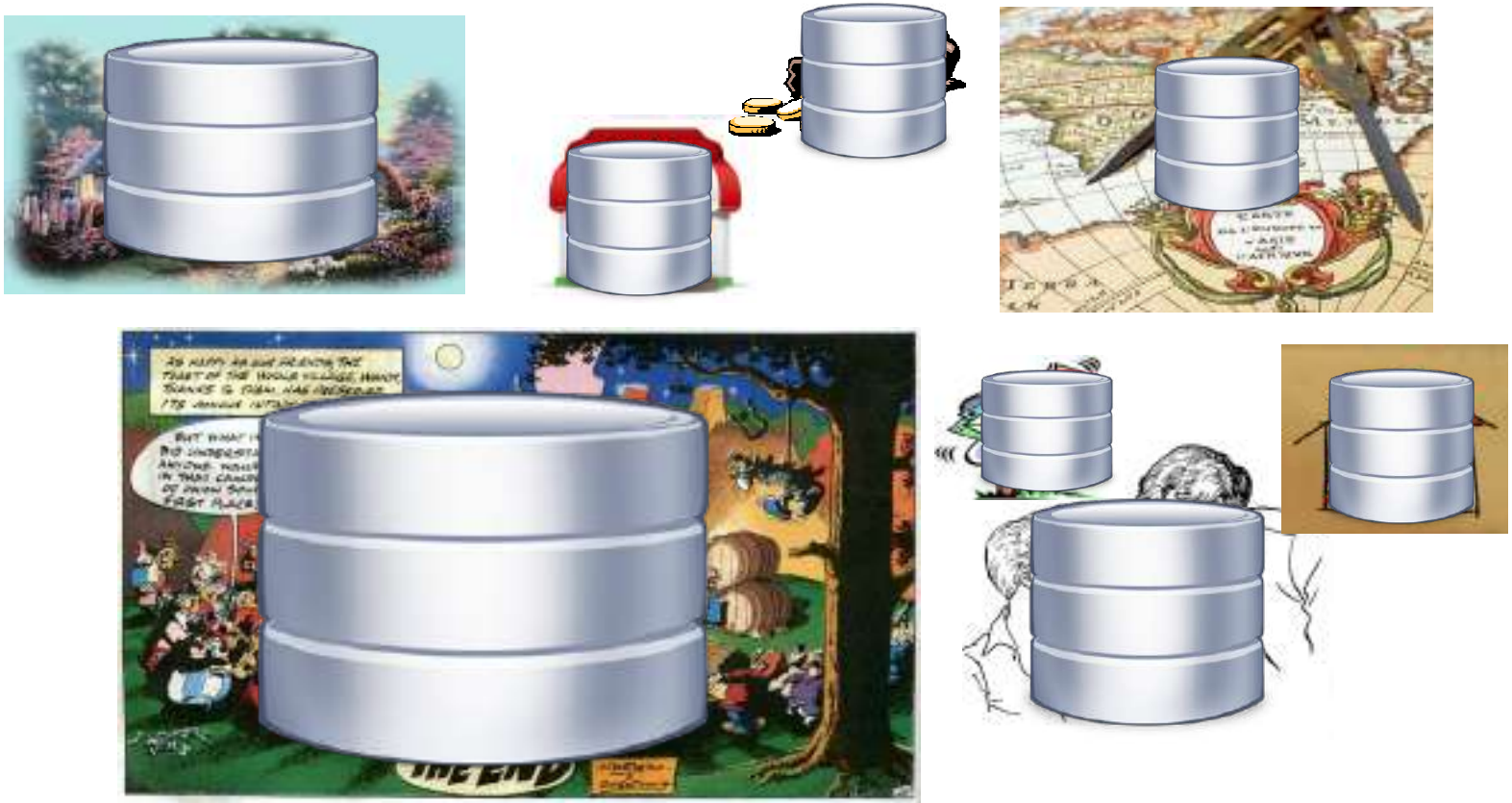


Figure: Our lives: (1) stored in DBMSs for primary usages and (2) feeding analytics and/or machine learning algorithms (secondary usages).

Ubiquitous personal data collection

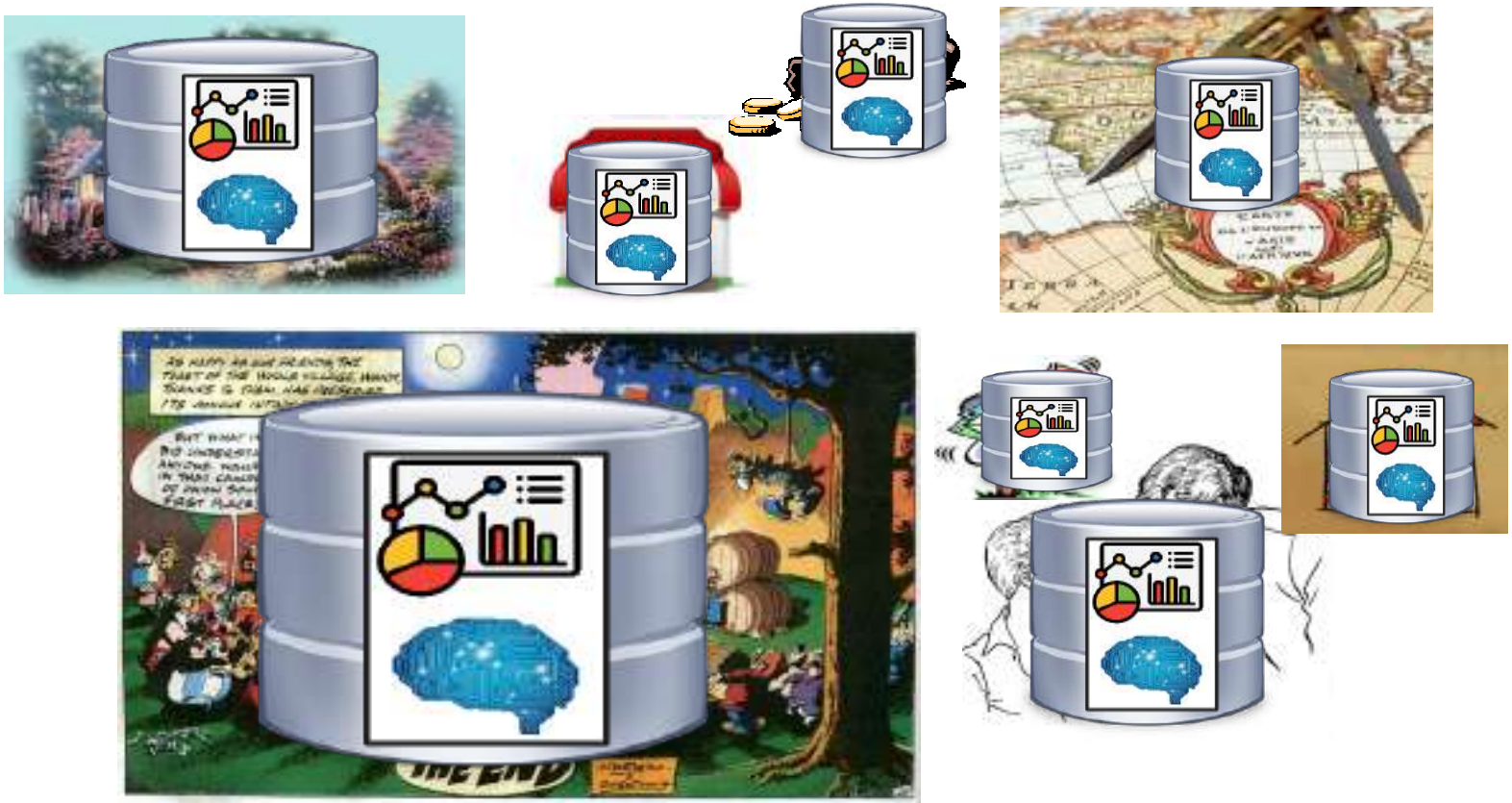


Figure: Our lives: (1) stored in DBMSs for primary usages and (2) feeding analytics and/or machine learning algorithms (secondary usages).

The New Oil

"Personal data is the new oil of the internet and the new currency of the digital world."
M. Kouneva, European Commissioner for Consumer Protection, March 2009

Privacy-Preserving Data Publishing

Privacy-Preserving Data Publishing (PPDP) :

- Publish *personal data* for analysis purposes (**accurate statistics**)...
- ...while preserving individuals' *privacy* (**uncertain point queries**)
- Also called:
  - *Sanitization* (in Computer Science)
  - *Anonymization* (in Laws, mainly)

Privacy-Preserving Data Publishing and the Law

Scope of the EU law:

- GDPR (General Data Protection Regulation) : the European regulation about the protection of personal data.
- Protects personal data.
- It does not apply to personal data "made anonymous".

$\Rightarrow$ Crucial to define "personal data" and "made anonymous" !

Record-level data *vs* Aggregate data

Privacy-preserving data publishing algorithms (for tabular data[1]):

- **Input:** tabular data, one row per individual.

Record-level data *vs* aggregate data: **imperfect but useful** taxonomy (see **[8]**).

[1] We focus on tabular data for the sake of clarity.

Record-level data *vs* Aggregate data

Privacy-preserving data publishing algorithms (for tabular data[1]):

- **Input:** tabular data, one row per individual.
- **Output:** either record-level data or aggregate data.

Record-level data *vs* aggregate data: **imperfect but useful** taxonomy (see **[8]**).

[1] We focus on tabular data for the sake of clarity.

Record-level data *vs* Aggregate data

Privacy-preserving data publishing algorithms (for tabular data[1]):

- **Input:** tabular data, one row per individual.
- **Output:** either record-level data or aggregate data.
  - **Record-level data:** results from *transforming* each input row into an output row (e.g., pseudonymization, *de-identification* like k-anonymity and followers).

Record-level data *vs* aggregate data: **imperfect but useful** taxonomy (see **[8]**).

[1] We focus on tabular data for the sake of clarity.

Record-level data *vs* Aggregate data

Privacy-preserving data publishing algorithms (for tabular data[1]):

- **Input:** tabular data, one row per individual.
- **Output:** either record-level data or aggregate data.
  - **Record-level data:** results from *transforming* each input row into an output row (e.g., pseudonymization, *de-identification* like k-anonymity and followers).
  - **Aggregate data:** results from *aggregating data across multiple input rows* (e.g., summary statistics, statistical queries, or even machine learning models).

Record-level data *vs* aggregate data: **imperfect but useful** taxonomy (see **[8]**).

[1] We focus on tabular data for the sake of clarity.

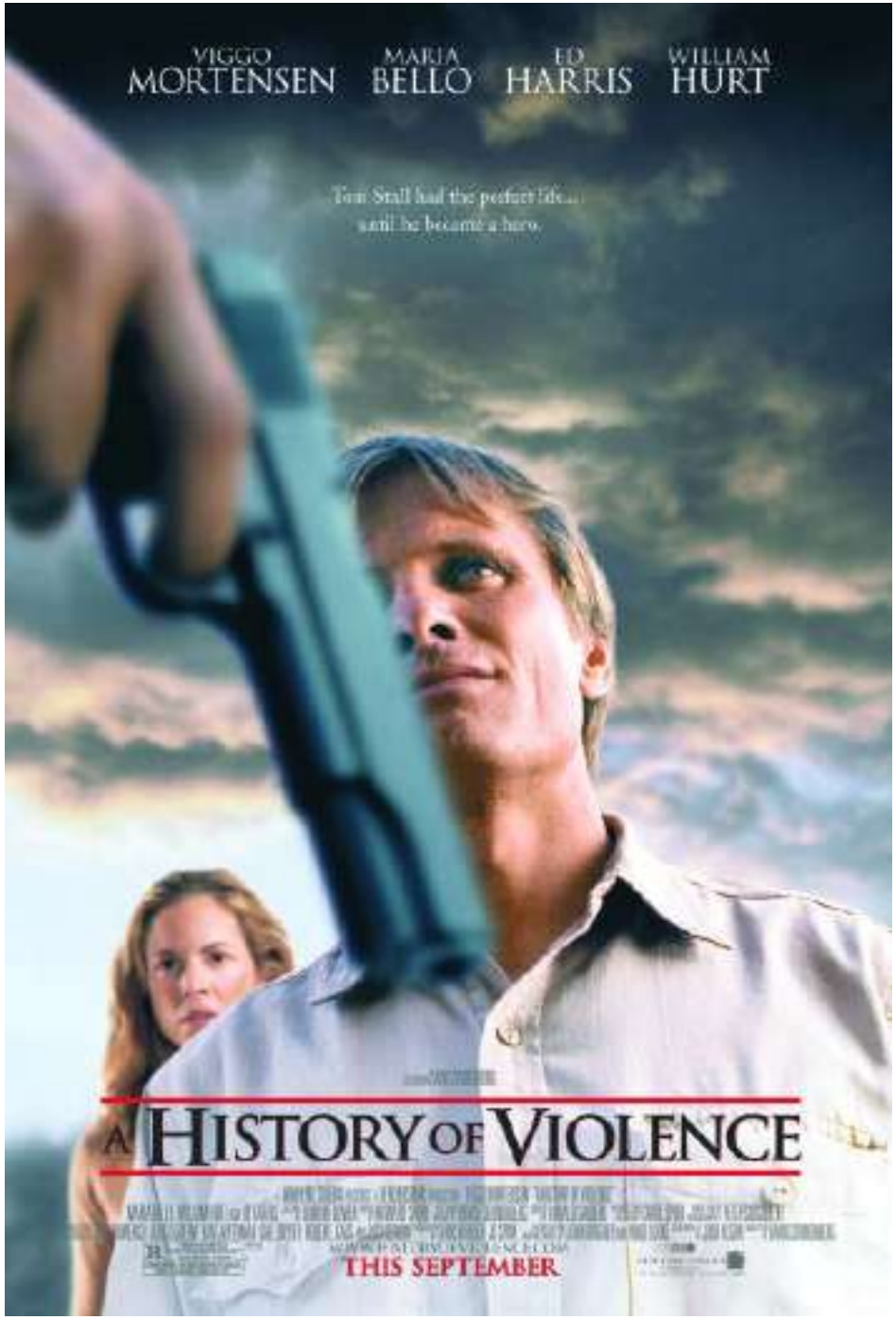
Menu

1. **Overview** of attacks on record-level data and on aggregate data.

Menu

1. **Overview** of attacks on record-level data and on aggregate data.
2. **Introduction** to sound protection approaches.

Progress of the Talk


VIGGO MORTENSEN
MARIA BELLO
ED HARRIS
WILLIAM HURT
Tom Stall had the perfect life...
until he became a hero.
A HISTORY OF VIOLENCE
THIS SEPTEMBER

Pseudonymization?

Record-level data.

Do you remember *what is pseudonymization*?

Pseudonymization?

Record-level data.

Do you remember *what is pseudonymization*?

Replace identifiers with a pseudonym, keep all the other columns.

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

$\Rightarrow$

UID	Zip	Age	Dis.
91263	75001	22	Cold
28781	75002	29	Flu
52689	75003	22	Cold
18429	75010	28	HIV

Table: **Left:** raw data (`Name` is an identifier) – **Right:** pseudonymized data (`UID` is the pseudonym, *one* UID per name, an arbitrary number)

Attacks over pseudonymization I

Record-level data.

Name	Zip	Age
Bob	75001	22

$\bowtie$

UID	Zip	Age	Dis.
91263	75001	22	Cold
28781	75002	29	Flu
52689	75003	22	Cold
18429	75010	28	HIV

Table: **Left:** background knowledge (`Name` is an identifier) – **Right:** pseudonymized data (`UID` is the pseudonym, *one* UID per name, an arbitrary number)

Attacks over pseudonymization II

Linkage attacks

1. **Exact matching attacks:** compute the exact match between pseudonymized records and background knowledge to perform the linkage (see, e.g., Governor Weld's re-identification [25, 23]).
   $\Rightarrow$ *A simple join.*
2. **Robust matching attacks:** compute an approximate match (see, e.g., netflix re-identification [19]).
   $\Rightarrow$ *Pairwise similarity between all possible pairs of (pseudonymous record, background knowledge) and optimal assignment.*

Attacks over pseudonymization III

Uniqueness study with ENEDIS on 2.5M electrical "consumption" records [27]: 1 power measurement (Watt) every 30 mins during 1 year.

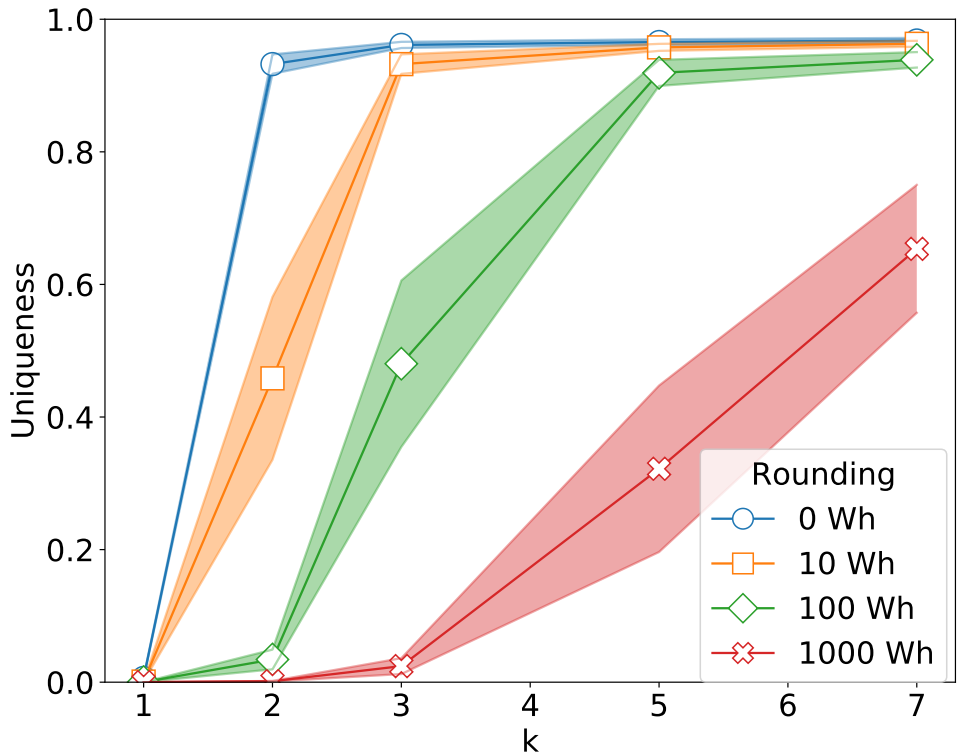


Figure: Average uniqueness wrt the **number of consecutive measures** considered (k) and to the **rounding** applied.

De-identification?

Record-level data.

De-Identification

- **Modify each input record** based on simple operations like, e.g., *generalization*, *swapping*, *naive perturbation*, *full deletion*.

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: **Left:** raw data – **Right:** a possible 2-anonymous release.

De-identification?

Record-level data.

De-Identification

- **Modify each input record** based on simple operations like, e.g., *generalization*, *swapping*, *naive perturbation*, *full deletion*.
- Two main approaches):

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: **Left:** raw data – **Right:** a possible 2-anonymous release.

De-identification?

Record-level data.

De-Identification

- **Modify each input record** based on simple operations like, e.g., *generalization*, *swapping*, *naive perturbation*, *full deletion*.
- Two main approaches):
  - **Based on privacy models:** k-anonymity, l-diversity, and followers (see below for privacy models).

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: **Left:** raw data – **Right:** a possible 2-anonymous release.

De-identification?

Record-level data.

De-Identification

- **Modify each input record** based on simple operations like, e.g., *generalization*, *swapping*, *naive perturbation*, *full deletion*.
- Two main approaches):
  - **Based on privacy models:** k-anonymity, l-diversity, and followers (see below for privacy models).
  - **Based on heuristics:** utility-driven (see [18] for a survey).

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: **Left:** raw data – **Right:** a possible 2-anonymous release.

Attacks over de-identification

Record-level data.

Any idea for attacking de-identified data?

Attacks over de-identification

Record-level data.

Any idea for attacking de-identified data?

- **Re-identification attack:** on attributes that are not part *degraded* (i.e., part of the quasi-identifier).

Attacks over de-identification

Record-level data.

Any idea for attacking de-identified data?

- **Re-identification attack:** on attributes that are not part *degraded* (i.e., part of the quasi-identifier).
- **Composition attacks:** non-composability [9].

Attacks over de-identification

Record-level data.

Any idea for attacking de-identified data?

- **Re-identification attack:** on attributes that are not part *degraded* (i.e., part of the quasi-identifier).
- **Composition attacks:** non-composability [9].
- **Minimality attacks:** undo generalization [6, 4].

Attacks over de-identification

Record-level data.

Any idea for attacking de-identified data?

- **Re-identification attack:** on attributes that are not part *degraded* (i.e., part of the quasi-identifier).
- **Composition attacks:** non-composability [9].
- **Minimality attacks:** undo generalization [6, 4].
- **DeFinetti attack:** exploit correlations between quasi-identifiers and sensitive data [16].

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

True, but. . . ?

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

True, but. . . ?

- **Differencing attacks** (see below)

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

True, but. . . ?

- **Differencing attacks** (see below)
- **Reconstruction attacks**: recover entire examples from the training set (e.g., [1, 2, 21, 32]).

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

True, but. . . ?

- **Differencing attacks** (see below)
- **Reconstruction attacks**: recover entire examples from the training set (e.g., [1, 2, 21, 32]).
- **Membership inference attacks**: predict the participation of an entity in the training set (e.g., [10, 11, 13, 17, 20, 24, 26, 30]).

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

True, but. . . ?

- **Differencing attacks** (see below)
- **Reconstruction attacks**: recover entire examples from the training set (e.g., [1, 2, 21, 32]).
- **Membership inference attacks**: predict the participation of an entity in the training set (e.g., [10, 11, 13, 17, 20, 24, 26, 30]).
- **Attribute inference attacks**: infer a sensitive attribute of a target record (e.g., [14, 29, 31, 32])

Aggregation, really protected?

Intuitively, aggregation **seems** to provide **some** protection:

- **No re-identification** attack
- **No minimality** attack
- **No DeFinetti** attack

True, but. . . ?

- **Differencing attacks** (see below)
- **Reconstruction attacks**: recover entire examples from the training set (e.g., [1, 2, 21, 32]).
- **Membership inference attacks**: predict the participation of an entity in the training set (e.g., [10, 11, 13, 17, 20, 24, 26, 30]).
- **Attribute inference attacks**: infer a sensitive attribute of a target record (e.g., [14, 29, 31, 32])
- *etc.*

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$

```
SELECT COUNT(*) FROM PATIENTS
WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
```

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$

```
SELECT COUNT(*) FROM PATIENTS
WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
```

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$

```
SELECT COUNT(*) FROM PATIENTS
WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
```

    Assume that the result is: $q_1(\mathcal{D}) = 1$

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_2(\mathcal{D}) = 100$

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_2(\mathcal{D}) = 100$
  - $q_3 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE NOT (AGE=26 AND ZIP=12345 AND GENDER='m')
    AND DIAG LIKE 'FLU'
    ```

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_2(\mathcal{D}) = 100$
  - $q_3 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE NOT (AGE=26 AND ZIP=12345 AND GENDER='m')
    AND DIAG LIKE 'FLU'
    ```

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_2(\mathcal{D}) = 100$
  - $q_3 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE NOT (AGE=26 AND ZIP=12345 AND GENDER='m')
    AND DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_3(\mathcal{D}) = 99$

Illustration: Differencing Attacks (the 1990s)

Aggregate data.

- Private table : `PATIENTS (AGE, ZIP, GENDER, DIAG)`
- Background knowledge : Bill is a man, 26 years old, zipcode 12345, present in `PATIENTS`.
- Consider the following queries :
  - $q_1 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE (AGE=26 AND ZIP=12345 AND GENDER='m')
    ```
    Assume that the result is: $q_1(\mathcal{D}) = 1$
  - $q_2 =$
    ```
    SELECT COUNT(*) FROM PATIENTS WHERE DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_2(\mathcal{D}) = 100$
  - $q_3 =$
    ```
    SELECT COUNT(*) FROM PATIENTS
    WHERE NOT (AGE=26 AND ZIP=12345 AND GENDER='m')
    AND DIAG LIKE 'FLU'
    ```
    Assume that the result is: $q_3(\mathcal{D}) = 99$
- Preventing queries that lead to differences with a low cardinality: in general **NP-Hard** [3], and leads to **leaks** [15].

Illustration: Linear Reconstruction Attacks (the 2000s)

Aggregate data.

In general, with a sufficient number of `COUNT` queries (**polynomial** in the DB size), **and even with naive noise addition**, it is possible to build and solve a system of linear equations for obtaining the target values.

[2] A real-life system `https://aircloak.com/` has been "recently" broken (2018) based on this kind of attack [5].

Illustration: Linear Reconstruction Attacks (the 2000s)

Aggregate data.

In general, with a sufficient number of `COUNT` queries (**polynomial** in the DB size), **and even with naive noise addition**, it is possible to build and solve a system of linear equations for obtaining the target values.

Real-life illustration[2]

Assume a `loan DB` with a `clientId` column, a secret bit `loanStatus`, and an *exact* `COUNT(*)` interface to the DB. Perform a sufficient number of queries as follows and solve the system.

```
SELECT COUNT(*) FROM loans WHERE clientId BETWEEN RANDOM1
and RANDOM2 AND loan = 1
```

[2] A real-life system `https://aircloak.com/` has been "recently" broken (2018) based on this kind of attack [5].

Attacks over machine learning models

Aggregate data.

In the context of machine learning models:

- **Differencing attacks:** NO differencing attacks (not relevant).

Attacks over machine learning models

Aggregate data.

In the context of machine learning models:

- **Differencing attacks:** NO differencing attacks (not relevant).
- **Reconstruction attacks:** VULN.

Attacks over machine learning models

Aggregate data.

In the context of machine learning models:

- **Differencing attacks:** NO differencing attacks (not relevant).
- **Reconstruction attacks:** VULN.
- **Membership inference attacks:** VULN.

Attacks over machine learning models

Aggregate data.

In the context of machine learning models:

- **Differencing attacks:** NO differencing attacks (not relevant).
- **Reconstruction attacks:** VULN.
- **Membership inference attacks:** VULN.
- **Attribute inference attacks:** VULN.

Attacks over machine learning models

Aggregate data.

In the context of machine learning models:

- **Differencing attacks:** NO differencing attacks (not relevant).
- **Reconstruction attacks:** VULN.
- **Membership inference attacks:** VULN.
- **Attribute inference attacks:** VULN.
- *etc.*

Illustration: Shadow Training for Membership Inferences

Shadow training is often used for MIA. Intuitively:

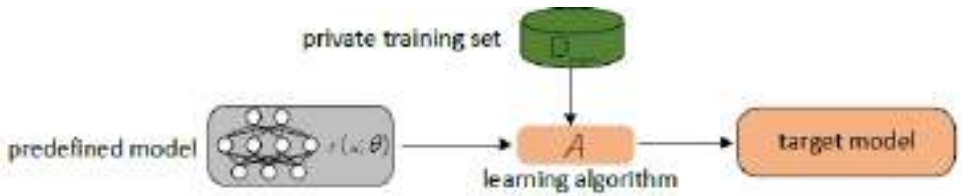


Figure: Intuitions on shadow training [12]

While **shadow training is popular**, the **attacks vary** on the details...

Illustration: Shadow Training for Membership Inferences

Shadow training is often used for MIA. Intuitively:

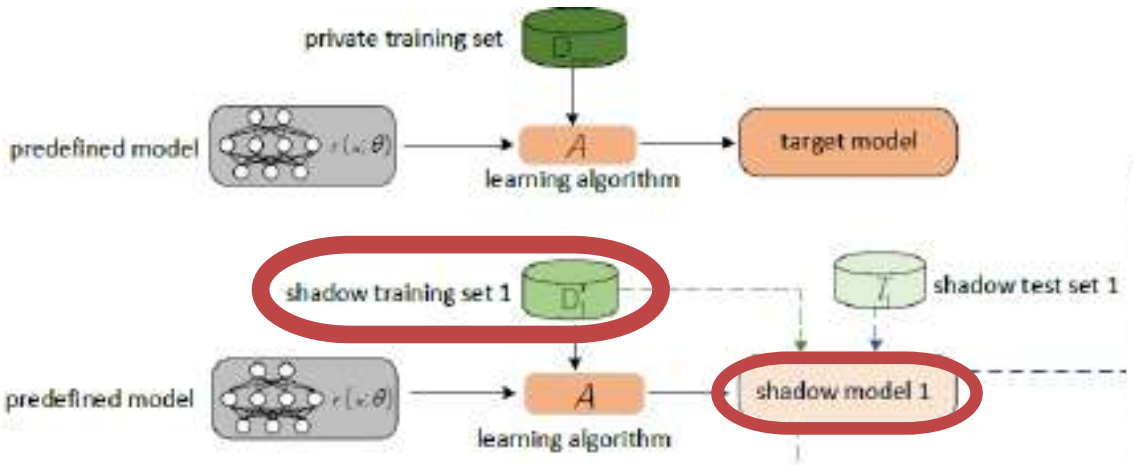


Figure: Intuitions on shadow training [12]

While **shadow training is popular**, the **attacks vary** on the details...

Illustration: Shadow Training for Membership Inferences

Shadow training is often used for MIA. Intuitively:

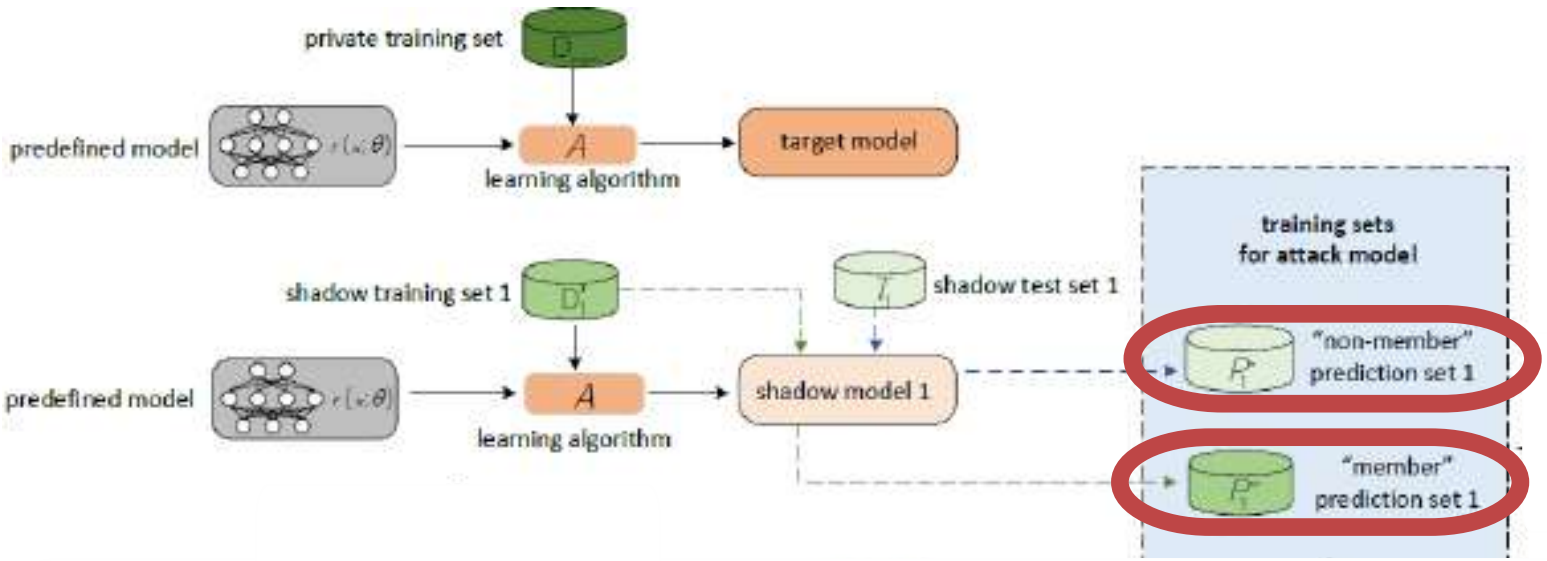


Figure: Intuitions on shadow training [12]

While **shadow training is popular**, the **attacks vary** on the details...

Illustration: Shadow Training for Membership Inferences

Shadow training is often used for MIA. Intuitively:

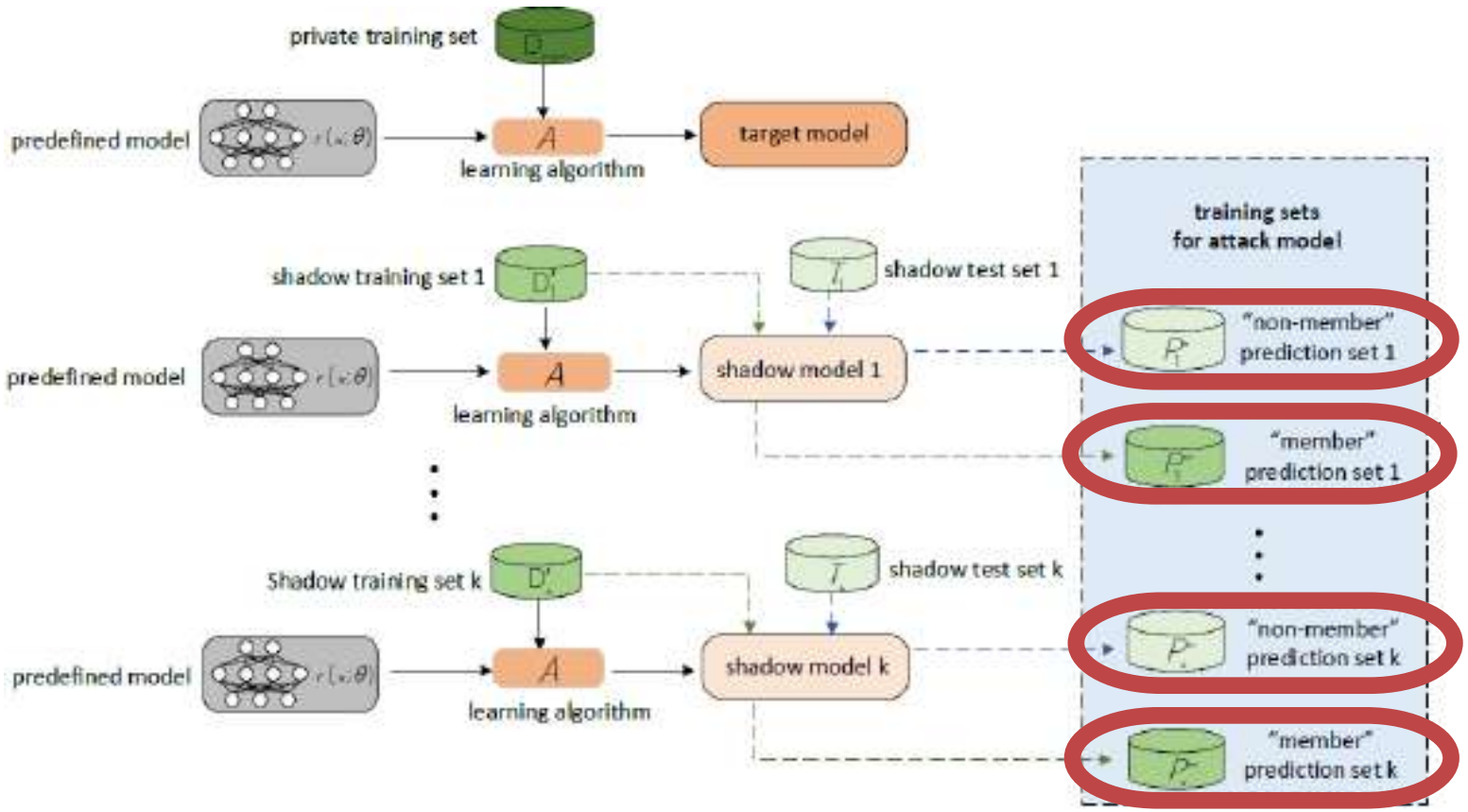


Figure: Intuitions on shadow training [12]

While **shadow training is popular**, the **attacks vary** on the details...

Illustration: Shadow Training for Membership Inferences

Shadow training is often used for MIA. Intuitively:

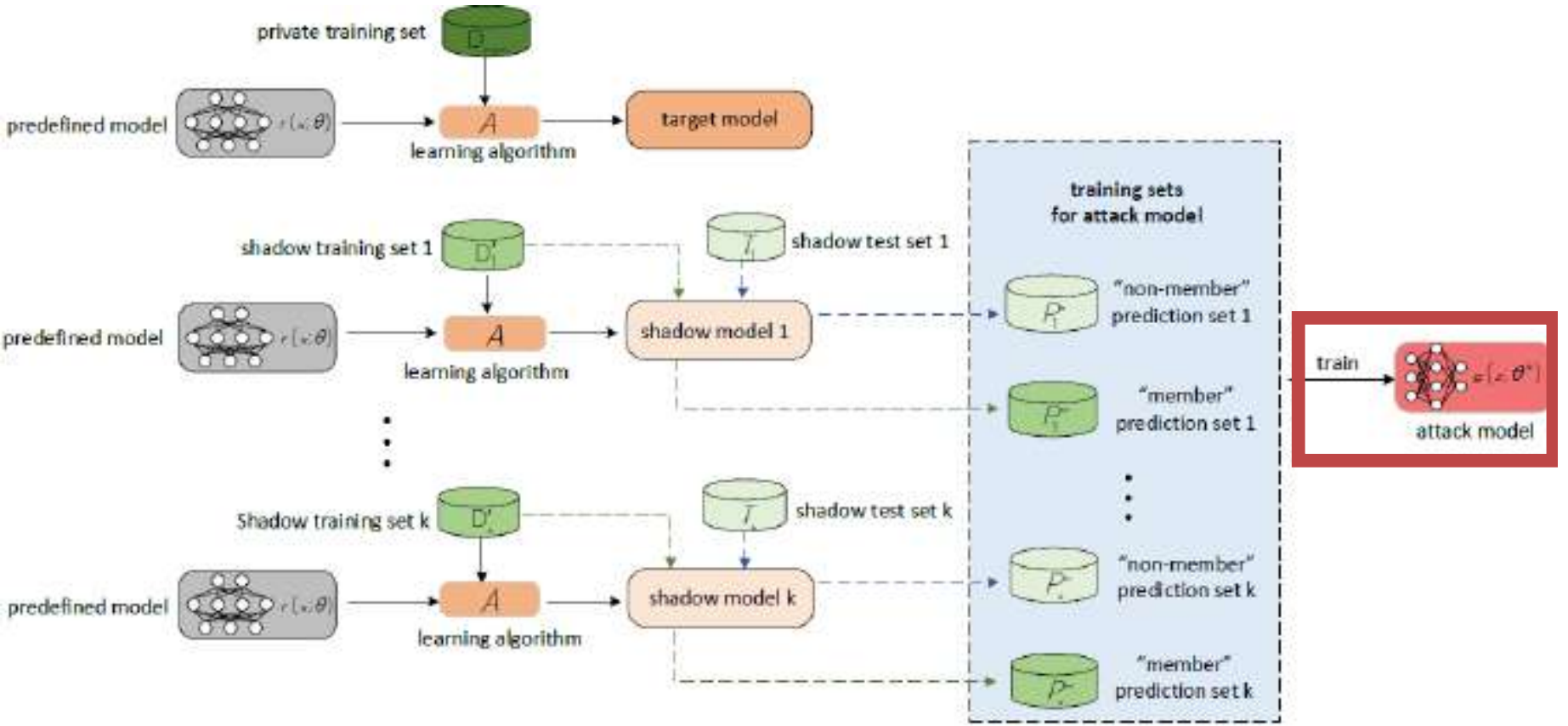


Figure: Intuitions on shadow training [12]

While **shadow training is popular**, the **attacks vary** on the details...

Overview of the LIRA Attack (the 2020's)

- **Goal:** MIA on *neural network* classifiers (n classes) $f_\theta : \mathcal{X} \to [0,1]^n$.
- **Background knowledge:** row of the target, distribution of the training data.
- **How:** exploit the *model classification confidence*.

Overview of the LIRA Attack (the 2020's)

- **Goal:** MIA on *neural network* classifiers (n classes) $f_\theta : \mathcal{X} \to [0,1]^n$.
- **Background knowledge:** row of the target, distribution of the training data.
- **How:** exploit the *model classification confidence*.
  - **Prepare:** *Shadow training* $\Rightarrow$ confidence distributions of the most-confident class when non-member VS when member.

Overview of the LIRA Attack (the 2020's)

- **Goal:** MIA on *neural network* classifiers (n classes) $f_\theta : \mathcal{X} \to [0,1]^n$.
- **Background knowledge:** row of the target, distribution of the training data.
- **How:** exploit the *model classification confidence*.
  - **Prepare:** *Shadow training* $\Rightarrow$ confidence distributions of the most-confident class when non-member VS when member.
  - **Exploit:** *Likelihood-ratio Test* given the classification confidence of the target and the two distributions.

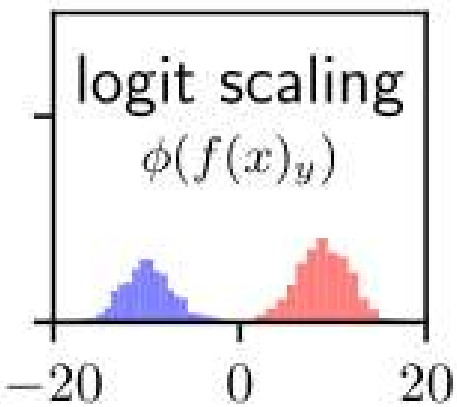


Figure: Illustration of the separability of the classification confidence distributions (confidence distribution in blue when the target is non-member, in red when member).

What is **unprotected** is. . .

What is **unprotected** is. . . **unprotected**!

"Natural protection" or *"empirical guarantees"* $\Rightarrow$ **Unprotected**.

What is **unprotected** is. . . **unprotected**!

"Natural protection" or *"empirical guarantees"* $\Rightarrow$ **Unprotected**.

Figure: Snake Challenge: `https://snake-challenge.github.io/`

Progress of the Talk

Components of a Privacy-preserving Data Publishing Solution

Three **needed** components:

1. **Privacy model (needed)**: What does it mean for the data released to be privacy-preserving?
   $\Rightarrow$ *The* **privacy guarantee** *that holds.*

In the following: illustration on **differential privacy**.

Components of a Privacy-preserving Data Publishing Solution

Three **needed** components:

1. **Privacy model (needed)**: What does it mean for the data released to be privacy-preserving?
   $\Rightarrow$ *The* **privacy guarantee** *that holds.*
2. **Privacy mechanism**: How to produce the privacy-preserving data to be released?
   $\Rightarrow$ *The* **algorithm** *used for providing the privacy guarantee.*

In the following: illustration on **differential privacy**.

Components of a Privacy-preserving Data Publishing Solution

Three **needed** components:

1. **Privacy model (needed)**: What does it mean for the data released to be privacy-preserving?
   $\Rightarrow$ *The* **privacy guarantee** *that holds.*
2. **Privacy mechanism**: How to produce the privacy-preserving data to be released?
   $\Rightarrow$ *The* **algorithm** *used for providing the privacy guarantee.*
3. **Utility metric**: How much useful is the released data?
   $\Rightarrow$ *Impact of the protection on data quality.*

In the following: illustration on **differential privacy**.

Let's start with the first component: the **model**.

What is the **guarantee** *that we* **target***?*

Let's start with the first component: the **model**.

What is the **guarantee** *that we* **target***?*

Warning: **necessary but often omitted** in real-life systems.

The Differential Privacy Paradigm

- **Global trends are not private** and must be learnt : there must be a knowledge gain !
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend.

The Differential Privacy Paradigm

- **Global trends are not private** and must be learnt : there must be a knowledge gain !
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend.

Differential Privacy Paradigm

A function f satisfies differential privacy iif: **the possible impact of any individual on the function's output is limited**.

The Differential Privacy Paradigm

- **Global trends are not private** and must be learnt : there must be a knowledge gain !
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend.

Differential Privacy Paradigm

A function f satisfies differential privacy iif: **the possible impact of any individual on the function's output is limited**.

Initial Model

ϵ-differential privacy (from [7])

A **random function** $\mathtt{f}$ satisfies ϵ-differential privacy iff:
For all $\mathcal{D}$ and $\mathcal{D}'$ **differing in at most one record**, and for **any** $\mathcal{S} \subseteq Range(\mathtt{f})$, then: $\mathtt{Pr}[\mathtt{f}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \times \mathtt{Pr}[\mathtt{f}(\mathcal{D}') \in \mathcal{S}]$

Remember: *think of f as a COUNT or a SUM perturbed by adding random noise to its output.*

Initial Model

ϵ-differential privacy (from [7])

A **random function** $\mathtt{f}$ satisfies ϵ-differential privacy iff:
For all $\mathcal{D}$ and $\mathcal{D}'$ **differing in at most one record**, and for **any** $\mathcal{S} \subseteq Range(\mathtt{f})$, then: $\mathtt{Pr}[\mathtt{f}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \times \mathtt{Pr}[\mathtt{f}(\mathcal{D}') \in \mathcal{S}]$

Remember: *think of f as a COUNT or a SUM perturbed by adding random noise to its output.*

Intuitively: **with or without the fox $\Rightarrow$ similar output probabilities**.

Intuitions - Mechanism

- Do not publish your dataset!

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...
- For ex : $q = \texttt{number of patients that have flu}$

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...
- For ex : $q = \texttt{number of patients that have flu}$
- How to **hide** the **impact** of any single individual participation to a **COUNT**, to a **SUM**?

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...
- For ex : $q = \texttt{number of patients that have flu}$
- How to **hide** the **impact** of any single individual participation to a **COUNT**, to a **SUM**?
  - **Add random noise** to the true result ! Answer $\mathbf{q}(\mathcal{D}) + \texttt{noise}$

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...
- For ex : $q = \texttt{number of patients that have flu}$
- How to **hide** the **impact** of any single individual participation to a **COUNT**, to a **SUM**?
  - **Add random noise** to the true result ! Answer $\mathbf{q}(\mathcal{D}) + \texttt{noise}$
  - Such that the noise is **proportional to the impact of a single individual**.

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...
- For ex : $q = \texttt{number of patients that have flu}$
- How to **hide** the **impact** of any single individual participation to a **COUNT**, to a **SUM**?
  - **Add random noise** to the true result ! Answer $\mathbf{q}(\mathcal{D}) + \texttt{noise}$
  - Such that the noise is **proportional to the impact of a single individual**.
  - For ex : $\texttt{noise}$ above should be proportionnal to the **impact** of one individual on q, *i.e.*,, proportionnal to 1 !

Intuitions - Mechanism

- Do not publish your dataset!
- Differential privacy originally considers **statistics** (counts, sums)...
- For ex : $q = \texttt{number of patients that have flu}$
- How to **hide** the **impact** of any single individual participation to a **COUNT**, to a **SUM**?
  - **Add random noise** to the true result ! Answer $\mathbf{q}(\mathcal{D}) + \texttt{noise}$
  - Such that the noise is **proportional to the impact of a single individual**.
  - For ex : $\texttt{noise}$ above should be proportionnal to the **impact** of one individual on q, *i.e.*,, proportionnal to 1 !
  - What if q had been a sum of salaries ?

What about attacks?

- Differential privacy directly opposes to membership inference attacks.
- But attack families are all related...

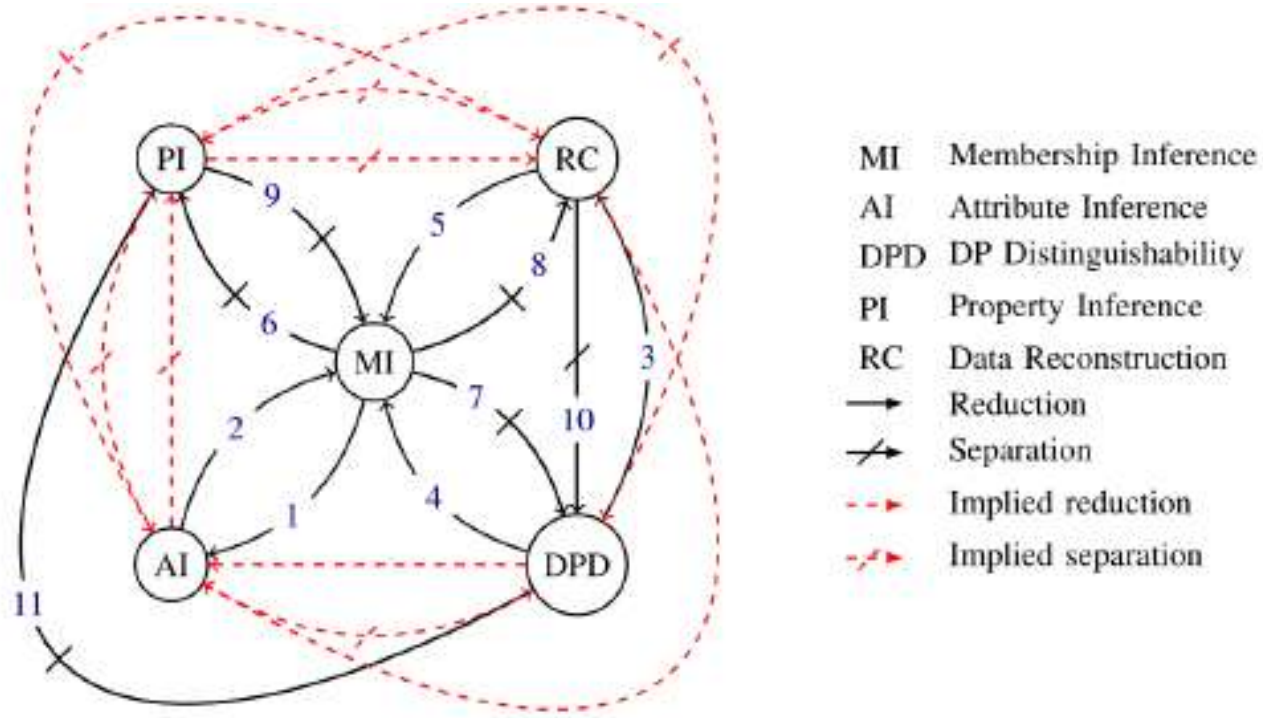


Fig. 1. Relations among adversary goals (under selected threat models). A solid arrow from node A to B means that security against A (i.e. a nontrivial advantage bound) implies security against B. A struck-through arrow from A to B means that security against A does not imply in general security against B; we show this separation with a construction that is secure against A but completely insecure against B. Dashed arrows are implied by solid arrows. Labels over solid arrows refer to the theorem showing the relationship. Some separations stem from differences in adversary capabilities, e.g. MI $\not\rightarrow$ RC.

Figure: "Relations among attack goals" [22] (obtained through reductions between games)

Differential Privacy Properties

- **Self-composability** : composing the outputs of two independant releases sanitized by differentially-private function(s) satisfies differential privacy :
  - Where $\epsilon_{final} = \sum \epsilon_i$ If input datasets are **not** disjoint
  - Or $\epsilon_{final} = \max \epsilon_i$ otherwise

Differential Privacy Properties

- **Self-composability** : composing the outputs of two independant releases sanitized by differentially-private function(s) satisfies differential privacy :
  - Where $\epsilon_{final} = \sum \epsilon_i$ If input datasets are **not** disjoint
  - Or $\epsilon_{final} = \max \epsilon_i$ otherwise
- **No breach from post-processing** :
  - (*Laplace mechanism is independent from data*)
  - Any function applied to a differentially-private input produces a differentially-private output

Inherent Limits

- Noise distribution centered on 0 ...
  $\Rightarrow$ Average of noises converges to 0 ...
  $\Rightarrow$ No unlimited number of queries !
- Composability properties $\Rightarrow$ the privacy parameter ϵ can be seen as a **budget** that must be distributed over the queries to execute ($\epsilon_{final} = \sum \epsilon_i$)

Progress of the Talk

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**
- Vulnerable approaches are **common**.

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**
- Vulnerable approaches are **common**.
- Differential privacy is the current *de facto* standard thanks to its **sound privacy guarantees** and **composability properties**.

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**
- Vulnerable approaches are **common**.
- Differential privacy is the current *de facto* standard thanks to its **sound privacy guarantees** and **composability properties**.
  - **Strong support in academia** (e.g., Gödel Prize in 2017)

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**
- Vulnerable approaches are **common**.
- Differential privacy is the current *de facto* standard thanks to its **sound privacy guarantees** and **composability properties**.
  - **Strong support in academia** (e.g., Gödel Prize in 2017)
  - **Major data-centric organisations** have switched to differential privacy (e.g., Google, the Census Bureau, LinkedIn, Facebook)

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**
- Vulnerable approaches are **common**.
- Differential privacy is the current *de facto* standard thanks to its **sound privacy guarantees** and **composability properties**.
  - **Strong support in academia** (e.g., Gödel Prize in 2017)
  - **Major data-centric organisations** have switched to differential privacy (e.g., Google, the Census Bureau, LinkedIn, Facebook)
  - **Drawbacks:** noise magnitude, limited privacy budget, not one-size-fits-all, need rare expertise.

Conclusion

- Privacy-preserving data publishing : a **deep and wide literature**
- Vulnerable approaches are **common**.
- Differential privacy is the current *de facto* standard thanks to its **sound privacy guarantees** and **composability properties**.
  - **Strong support in academia** (e.g., Gödel Prize in 2017)
  - **Major data-centric organisations** have switched to differential privacy (e.g., Google, the Census Bureau, LinkedIn, Facebook)
  - **Drawbacks:** noise magnitude, limited privacy budget, not one-size-fits-all, need rare expertise.
- **Active research:** synthetic data generation (with DP-like privacy guarantees), DP ML, **DP for protecting distributed algorithms** [28, 33], *etc.*

Progress of the Talk

[1] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. X. Song.
The secret sharer: Evaluating and testing unintended memorization in neural networks.
In *USENIX Security Symposium*, 2018.

[2] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. X. Song, Ú. Erlingsson, A. Oprea, and C. Raffel.
Extracting training data from large language models.
In *USENIX Security Symposium*, 2020.

[3] F. Y. L. Chin.
Security problems on inference control for sum, max, and min queries.
J. ACM, 33:451–464, 1986.

[4] A. Cohen.
Attacks on deidentification's defenses.
In *USENIX Security Symposium*, 2022.

[5] A. Cohen and K. Nissim.

Linear program reconstruction in practice.
CoRR, abs/1810.05692, 2018.

[6] G. Cormode, N. Li, T. Li, and D. Srivastava.
Minimizing minimality and maximizing utility.
Proceedings of the VLDB Endowment, 3:1045 – 1056, 2010.

[7] C. Dwork.
Differential privacy.
In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[8] A. Gadotti, L. Rocher, F. Houssiau, A.-M. Creţu, and Y.-A. de Montjoye.
Anonymization: The imperfect science of using data while preserving privacy.
Science Advances, 10(29):eadn7053, 2024.

[9] S. R. Ganta, S. P. Kasiviswanathan, and A. D. Smith.
Composition attacks and auxiliary information in data privacy.

In *Knowledge Discovery and Data Mining*, 2008.

[10] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro.
Logan: Membership inference attacks against generative models.
Proc. Priv. Enhancing Technol., 2019(1):133–152, 2019.

[11] B. Hilprecht, M. Härterich, and D. Bernau.
Monte carlo and reconstruction membership inference attacks against generative models.
Proceedings on Privacy Enhancing Technologies, 2019:232 – 249, 2019.

[12] H. Hu, Z. A. Salcic, G. Dobbie, and X. Zhang.
Membership inference attacks on machine learning: A survey.
ACM Computing Surveys (CSUR), 2022.

[13] J. H. Hyeong, J. Kim, N. Park, and S. Jajodia.
An empirical study on the membership inference attack against tabular data synthesis models.
In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022.

[14] B. Jayaraman and D. Evans.
Are attribute inference attacks just imputation?
Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, 2022.

[15] K. Kenthapadi, N. Mishra, and K. Nissim.
Simulatable auditing.
Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2005.

[16] D. Kifer.
Attacks on privacy and definetti's theorem.
Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009.

[17] K. Leino and M. Fredrikson.
Stolen memories: Leveraging model memorization for calibrated white-box membership inference.
In *USENIX Security Symposium*, 2019.

[18] G. J. Matthews and O. Harel.
Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy.
Statistics Surveys, 5:1–29, 2011.

[19] A. Narayanan and V. Shmatikov.
Robust de-anonymization of large sparse datasets.
In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.

[20] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou.
White-box vs black-box: Bayes optimal strategies for membership inference.
In *International Conference on Machine Learning*, 2019.

[21] A. Salem, A. Bhattacharyya, M. Backes, M. Fritz, and Y. Zhang.
Updates-leak: Data set inference and reconstruction attacks in online learning.

In *USENIX Security Symposium*, 2019.

[22] A. Salem, G. Cherubin, D. Evans, B. Kopf, A. J. Paverd, A. Suri, S. Tople, and S. Zanella-B'eguelin.
Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning.
2023 IEEE Symposium on Security and Privacy (SP), pages 327–345, 2023.

[23] P. Samarati.
Protecting respondents identities in microdata release.
IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.

[24] R. Shokri, M. Stronati, C. Song, and V. Shmatikov.
Membership inference attacks against machine learning models.
2017 IEEE Symposium on Security and Privacy (SP), pages 3–18, 2017.

[25] L. Sweeney.
k-anonymity: a model for protecting privacy.

Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.

[26] A. Voyez, T. Allard, G. Avoine, P. Cauchois, E. Fromont, and M. Simonin.
Membership Inference Attacks on Aggregated Time Series with Linear Programming.
In *Proceedings of the 19th International Conference on Security and Cryptography (SECRYPT '22)*, 2022.

[27] A. Voyez, T. Allard, G. Avoine, P. Cauchois, E. Fromont, and M. Simonin.
The privacy cost of fine-grained electrical consumption data.
Scientific Reports, 15, 2025.

[28] Q. Wu, A. C. Zhou, T. Allard, S. Ibrahim, Y. Feng, L. Li, and A. E. Abbadi.
Cfdgraph: Privacy-preserving graph processing for large-scale collaborative fraud detection.
In *Proceedings of the 42nd International Conference on Data Engineering (ICDE '26)*, 2026.

[29] X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton.
A methodology for formalizing model-inversion attacks.
2016 IEEE 29th Computer Security Foundations Symposium (CSF), pages 355–370, 2016.

[30] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha.
Privacy risk in machine learning: Analyzing the connection to overfitting.
In *The 31st IEEE Computer Security Foundations Symposium*, CSF, Oxford, UK, July 2018.

[31] S. Yeom, I. Giacomelli, A. Menaged, M. Fredrikson, and S. Jha.
Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning.
J. Comput. Secur., 28:35–70, 2020.

[32] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang.
Inference attacks against graph neural networks.
In *USENIX Security Symposium*, 2022.

[33] A. C. Zhou, R. Qiu, T. Lambert, T. Allard, S. Ibrahim, and A. E. Abbadi.
Pgpregel: an end-to-end system for privacy-preserving graph processing in geo-distributed data centers.
In *Proceedings of the 13th Symposium on Cloud Computing*, 2022.

Progress of the Talk

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*
  - **a name**,

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*
  - **a name**,
  - **an identification number**,

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*
  - **a name**,
  - **an identification number**,
  - **location data**,

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*
  - **a name**,
  - **an identification number**,
  - **location data**,
  - **an online identifier**

[3] https://gdpr-info.eu/art-4-gdpr/

An Encompassive Definition of Personal Data

GDPR Article 4[3] (1) : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*
  - **a name**,
  - **an identification number**,
  - **location data**,
  - **an online identifier**
  - *or to one or more factors specific to the* **physical**, **physiological**, **genetic**, **mental**, **economic**, **cultural** *or* **social identity** *of that natural person;*"

[3] https://gdpr-info.eu/art-4-gdpr/

But a Fuzzy Definition of Privacy-Preserving Data Publishing Techniques

GDPR Recital 26 [4] :

- "**The principles of data protection should therefore not apply to anonymous information**,

⇒ Crucial to design **strong** privacy-preserving data publishing techniques!

[4] https://gdpr-info.eu/recitals/no-26/

But a Fuzzy Definition of Privacy-Preserving Data Publishing Techniques

GDPR Recital 26 [4] :

- "**The principles of data protection should therefore not apply to anonymous information**,
- *namely information which does not relate to an identified or identifiable natural person*

⇒ Crucial to design **strong** privacy-preserving data publishing techniques!

[4] `https://gdpr-info.eu/recitals/no-26/`

But a Fuzzy Definition of Privacy-Preserving Data Publishing Techniques

GDPR Recital 26 [4] :

- "**The principles of data protection should therefore not apply to anonymous information**,
- *namely information which does not relate to an identified or identifiable natural person*
- *or to* **personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable**.

$\Rightarrow$ Crucial to design **strong** privacy-preserving data publishing techniques!

[4] https://gdpr-info.eu/recitals/no-26/

But a Fuzzy Definition of Privacy-Preserving Data Publishing Techniques

GDPR Recital 26 [4] :

- "**The principles of data protection should therefore not apply to anonymous information**,
- *namely information which does not relate to an identified or identifiable natural person*
- *or to* **personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable**.
- *This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*"

⇒ Crucial to design **strong** privacy-preserving data publishing techniques!

[4] `https://gdpr-info.eu/recitals/no-26/`

LIRA: Questions and Answers

- **Model's confidence score?**

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.
  - Used because the logit of the confidence scores is well approximated by a normal distribution (see below).

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.
  - Used because the logit of the confidence scores is well approximated by a normal distribution (see below).
- **Parametric modeling?**

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.
  - Used because the logit of the confidence scores is well approximated by a normal distribution (see below).
- **Parametric modeling?**
  - Do not estimate the distributions IN *VS* OUT, but assume they are normal distributions (and estimate the mean and standard deviations only – much simpler).

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.
  - Used because the logit of the confidence scores is well approximated by a normal distribution (see below).
- **Parametric modeling?**
  - Do not estimate the distributions IN *VS* OUT, but assume they are normal distributions (and estimate the mean and standard deviations only – much simpler).
  - There exist:
    In-distribution: $Q_{in}(x, y) = \{f \leftarrow \mathcal{T}(D \cup \{(x, y)\}) | D \sim \mathcal{D}\}$
    Out-distribution: $Q_{out}(x, y) = \{f \leftarrow \mathcal{T}(D \setminus \{(x, y)\}) | D \sim \mathcal{D}\}$

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.
  - Used because the logit of the confidence scores is well approximated by a normal distribution (see below).
- **Parametric modeling?**
  - Do not estimate the distributions IN *VS* OUT, but assume they are normal distributions (and estimate the mean and standard deviations only – much simpler).
  - There exist:
    In-distribution: $Q_{in}(x, y) = \{f \leftarrow \mathcal{T}(D \cup \{(x, y)\}) | D \sim \mathcal{D}\}$
    Out-distribution: $Q_{out}(x, y) = \{f \leftarrow \mathcal{T}(D \setminus \{(x, y)\}) | D \sim \mathcal{D}\}$
  - We estimate them by Gaussians: $\hat{Q_{in}} \sim \mathcal{N}(\mu_{in}, \sigma^2_{in})$ and $\hat{Q_{out}} \sim \mathcal{N}(\mu_{out}, \sigma^2_{out})$

LIRA: Questions and Answers

- **Model's confidence score?**
  - Real-valued score in $[0, 1]$, a probability (often resulting from a $\texttt{softmax}(\cdot)$ on the outputs of the classifier.
  - Denoted $A_S(x)_y$, where x is a set of features and y is a label (and A_S is the trained model, like above).
- **Logit scaling?**
  - $\phi(p) = \log \frac{p}{1-p}$ where p is the score defined above.
  - Used because the logit of the confidence scores is well approximated by a normal distribution (see below).
- **Parametric modeling?**
  - Do not estimate the distributions IN *VS* OUT, but assume they are normal distributions (and estimate the mean and standard deviations only – much simpler).
  - There exist:
    In-distribution: $Q_{in}(x, y) = \{f \leftarrow \mathcal{T}(D \cup \{(x, y)\}) | D \sim \mathcal{D}\}$
    Out-distribution: $Q_{out}(x, y) = \{f \leftarrow \mathcal{T}(D \setminus \{(x, y)\}) | D \sim \mathcal{D}\}$
  - We estimate them by Gaussians: $\hat{Q_{in}} \sim \mathcal{N}(\mu_{in}, \sigma^2_{in})$ and $\hat{Q_{out}} \sim \mathcal{N}(\mu_{out}, \sigma^2_{out})$
- **Likelihood-ratio Test**: is $\mathbb{P}(\texttt{confobs} | \mathcal{N}(\mu_{in}, \sigma^2_{in}))$ higher than $\mathbb{P}(\texttt{confobs} | \mathcal{N}(\mu_{out}, \sigma^2_{out}))$?