Does Machine Learning Compromise the Privacy of Training Data?

Josep Domingo-Ferrer

josep.domingo@urv.cat

UNIVERSITAT ROVIRA i VIRGILI

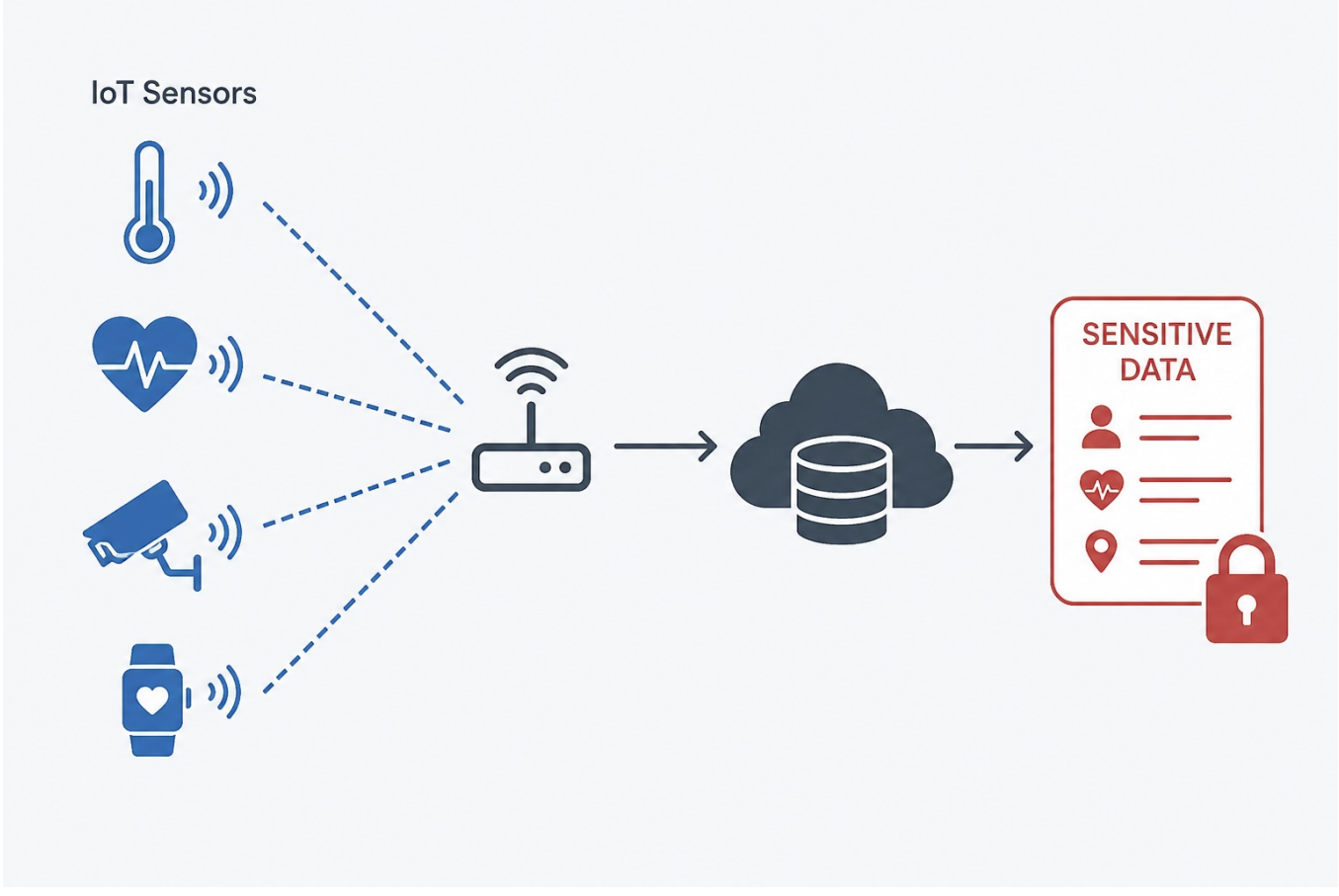
CYBER[SECURITY]CAT

ROMAE DIE XVII SEPTEMBRIS MMXXVI

Outline

Introduction

- Machine learning (ML) models are often trained on personal or otherwise sensitive data, like
  - Health data supplied by smartphones;
  - Sensitive data coming from IoT sensors.



AI Regulations

- Main regulations affecting AI in the EU: GDPR and AI Act.
- Chinese regulations on AI protect government more than citizens.
- U.S. regulations on AI are basically non-existing.

$\Longrightarrow$ The EU is the only large bloc committed to trustworthy AI, but the weakest one in AI technology

$\Longrightarrow$ The EU should avoid overcautious regulations that unnecessarily hamper its AI competitiveness.

Releasing a trained model $\neq$ releasing training data

- Releasing a model trained on personal data is not the same as directly released those training data.
- If only the model is released, a privacy attack is necessary to make inferences on the training data.
- I discuss how effective are the privacy attacks in the literature against predictive and generative AI in *real-world* conditions.
- I cover membership inference attacks (MIAs), property inference attacks, and reconstruction attacks.

Outline

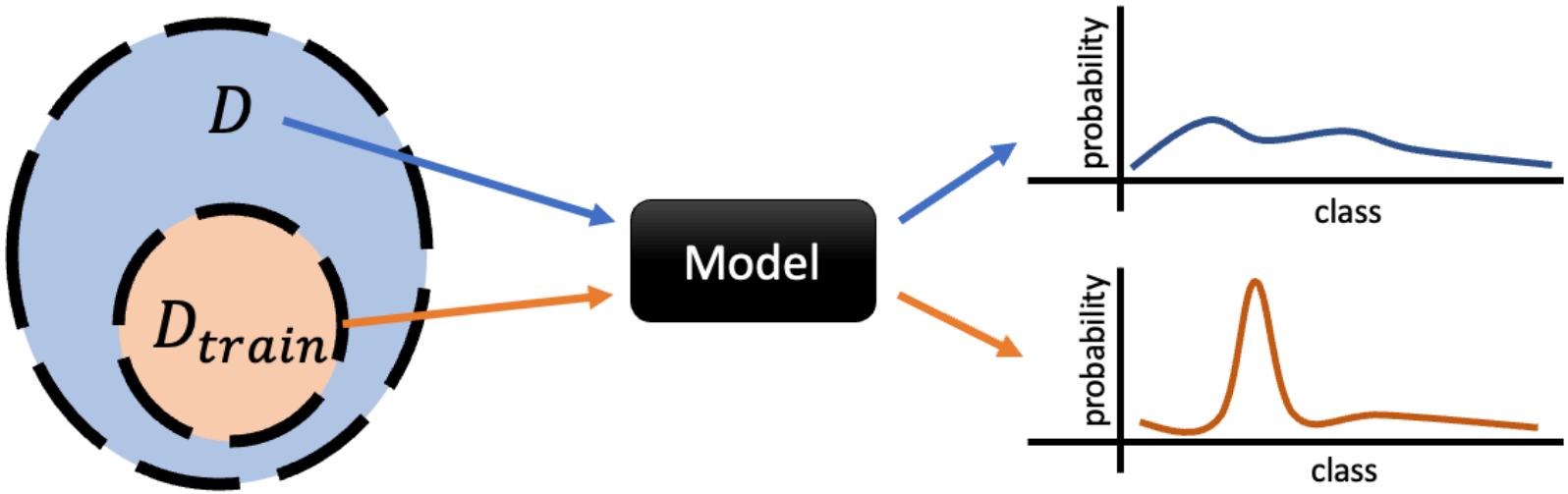
Background on privacy disclosure

Types of disclosure (that can occur independently):

- Identity disclosure. The attacker links some unidentified released piece of data to the subject to whom it corresponds.
- Attribute disclosure. The attacker determines the value of a confidential attribute (income, diagnosis. etc.) for a target subject after seeing the released data.
- Membership disclosure. The attacker determines whether a data point was included in the data used to train a released model.

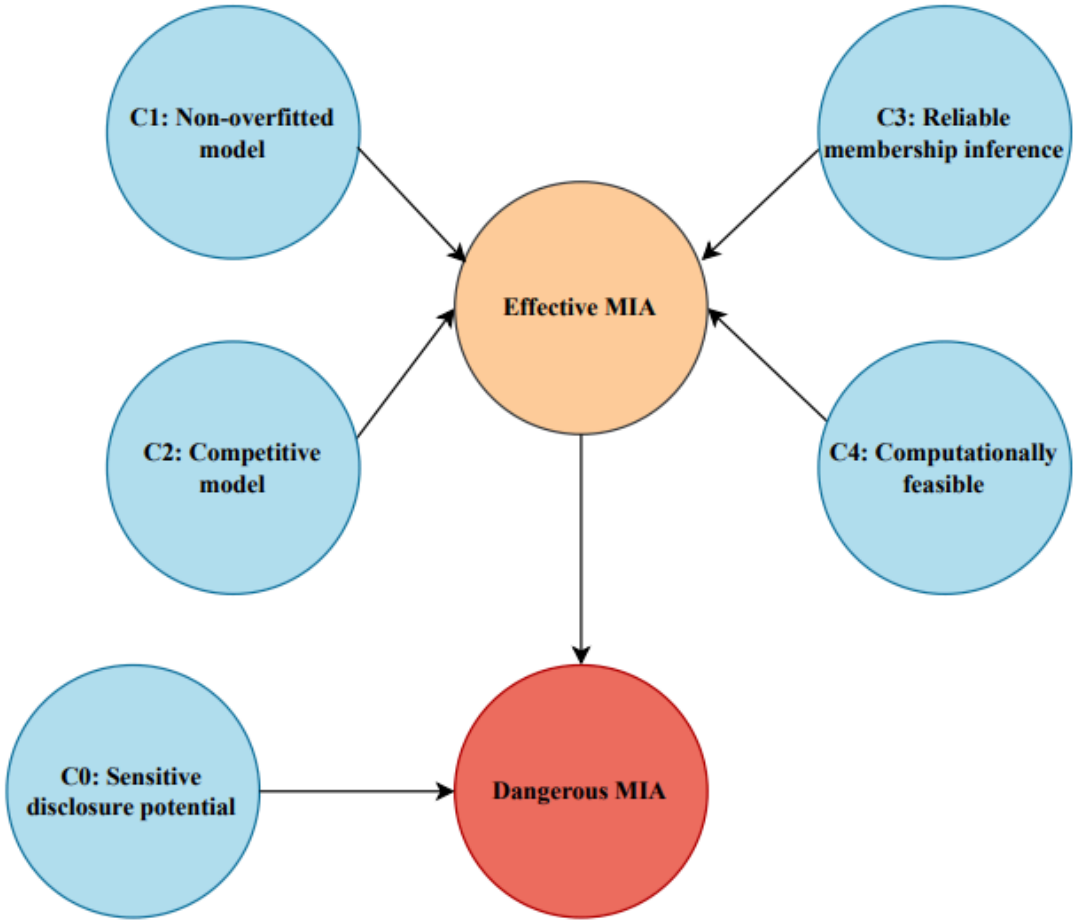
Outline

Membership inference attacks



Example: Attacker Alice wants to discover whether her neighbor Neil was a member of the training data for a model.

A framework for evaluating MIAs



N. Jebreel, D. Sánchez and J. Domingo-Ferrer, "A critical review on the effectiveness and privacy threats of membership inference attacks", in *ESORICS 2026*.

C0: Sensitive disclosure potential

C0 is the only condition depending on the training data, not the attack. Requirements for unequivocal inference:

- Exhaustivity. Otherwise, Alice might find a member with Neil's quasi-identifiers but who is not Neil.
- Non-diversity of unknown attributes. If Alice finds two members with Neil's quasi-identifiers but *different* incomes, Alice does not learn Neil's income.

MIA in the literature are not very dangerous because most training data sets are not exhaustive, and confidential attributes are naturally diverse.

C0: Sensitive disclosure potential

Related statistical disclosure control techniques:

- Non-exhaustivity inspires sampling. Protection inversely proportional to sampling factor and $\Pr(PU|SU)$.
- Non-diversity inspires *l*-diversity and *t*-closeness.

C1: Non-overfitted target model

- For an MIA to be effective, it must succeed against non-overfitted target models.
- We evaluated 61 pairs attack-data set in the literature.
- 24 lacked information on overfitting.
- 27 targeted overfitted models.
- Only 10 satisfied C1.

C2: Competitive target model

- The target model should have a utility (test accuracy) competitive with the SotA for its task. It should be a realistic model for deployment.
- Test accuracy was missing for 11 of the 61 attacks-data set pairs we assessed.
- Only 12 pairs use target models within 5% of the SotA.
- 38 pairs were vastly below the SotA.

C3: Reliable inference

- SotA focuses on TPR at extremely low FPR ($\leq 0.1\%$).
- However, low prior probability p of membership is usual and not accounted for.
- We supplement TPR at low FPR with

$$Prec = \frac{p \times TPR}{p \times TPR + (1 - p) \times FPR.}$$

- Only 16 out of 61 attack-data set pairs demonstrate reliable membership inference.

C4: Computational feasibility

- An MIA must be feasible with computational resources reasonably available to an attacker.
- An MIA whose cost exceeds that of training the target model is disproportionate.
- We assessed: (1) the number M of additional models needed; (2) the cost I of membership inference (low, moderate, high); and (3) number Q of model queries per target sample.
- 8 out of 13 attacks incur high computational costs.

C4: Computational feasibility

Attack	Resource requirements	Cost factors			Overall
		M	I	Q	cost
Shokri2017	10-100 additional models; K ML attack models; 1 query	High	Moderate	Low	High
Yeom2018	Simple thresholding rule on loss values; 1 query	Low	Low	Low	Low
Salem2019	Simple thresholding rule; 1K random data points per data set; 1 query	Low	Low	Low	Low
Sablayrolles2019	30 additional models; 1 query	High	Low	Low	High
Long2020	50 additional models; 1 query	High	Low	Low	High
Jayaraman2021	1 additional model; 100 queries	Moderate	Low	Low	Moderate
Song2021	1 additional model; 1 query	Moderate	Low	Low	Moderate
Liu2022	2 additional models; 1 ML attack model; multiple queries	High	Moderate	Moderate	High
Watson2022	1 additional model; 1 query	Moderate	Low	Low	Moderate
Ye2022	15-999 additional models; 1 query	High	Low	Low	High
Carlini2022 (LiRA)	32-256 additional models; under 100 queries	High	Low	Low	High
Bertran2023	Multiple quantile regression models; 1 query	Low	High	Low	High
Zarifzadeh2024	1-4 additional models; as per paper, 10% of training set size is sufficient	High	Low	High	High

Overall attack effectiveness

- No attack among those we reviewed simultaneously achieved C1, C2, C3, and C4.

MIAs against unlearning in generative ML

- It has been shown that MIAs on pre-trained LLMs are barely better than random guessing[1].
- Fine-tuned LLMs are much more vulnerable to MIAs on the data used for fine-tuning.

[1] M. Duan *et al.*, "Do membership inference attacks work on large language models?", arXiv:2402.07841, 2024.

Outline

Property inference attacks

- These attacks seek to infer a sensitive global property of a training data set.
- *E.g.*, "Google traffic was used in the training data", "images with noise were used for training", "mainly images of white males were used for training".
- Although discovering such properties is sensitive, it does not disclose private data on any individual.
- However, in federated learning, disclosing properties on a client's data set may be sensitive if all data relate to the same subject (*e.g.* fitness data in a smartphone).

Outline

Reconstruction attacks

- Before ML, Dinur and Nissim proposed a theory to reconstruct a database from query answers.
- In ML, reconstruction of training data is facilitated by model overfitting.
- Anti-overfitting techniques and differential privacy DP can be used to counter overfitting and hence reconstruction.

Original image

Reconstructed image

Complexity of reconstruction in ML

- For tabular training data, attribute ranges are finite (even for numerical attributes, due to limited precision).
- However, if the information content of an item X is $H(X)$, discovering X by exhaustive search amounts to discovering a random cryptographic key of $H(X)$ bits. IF $H(X) \geq 64$ bits, this is hard.

Total vs partial reconstruction

- *Total reconstruction*. If the attacker has unlimited resources, she can achieve total reconstruction with $\prod_{i=1}^{d} |A_i|$ MIAs assuming a tabular training data set with attributes $A_1, \ldots, A_d$.
- *Partial reconstruction*. A limited attacker focuses on MIAs for a subset of attribute combinations (guessing exercise). Betting on the most common combinations is likelier to succeed, but it is less privacy-sensitive.

Effectiveness of reconstruction

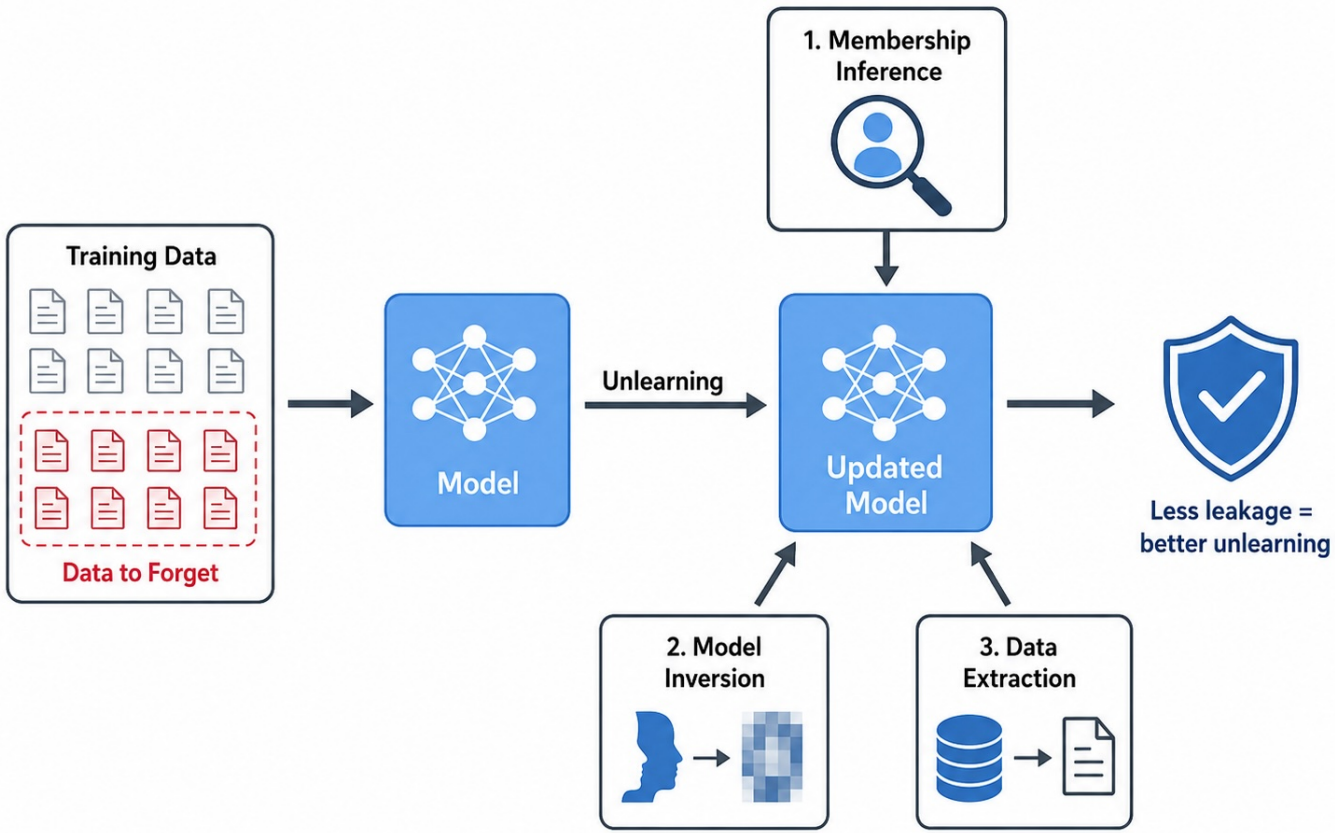
- State-of-the-art MIAs, such as LiRA, entail very substantial computational cost (shadow models).
- For non-tabular training data (unstructured text or multimedia), it is even more difficult to conduct an MIA for every potential image or unstructured text.
- It has been shown that MIAs on pre-trained LLMs are basically random guessing.
- However, fine-tuned LLMs are more vulnerable to MIAs on the fine-tuning data.

Effectiveness of reconstruction (II)

- If gradients are available to the attacker (as in federated learning), gradient inversion attacks can be used instead of MIAs.
- Gradient quantization or DP can be useful defenses against gradient inversion.
- Further, reconstruction without MIAs lacks a numerical success criterion when the attacker has no access to the training data.
- However, determining success may be easier when recontructing unlearned data (if those data are known to the attacker).

Outline

Privacy attacks and machine unlearning



Types of unlearning methods

- Exact unlearning, as if the data to be forgotten had never been seen by the model. High cost.
- Approximate unlearning. More efficient, but unlearning not guaranteed.
- Certified unlearning is a compromise, but requires assumptions on the ML models and loss functions, and guarantees not always clear.

$\Longrightarrow$ MIAs are used to check how well approximate unlearning works.

$\Longrightarrow$ They need to be effective and efficient.

Blockchain and machine (un)learning

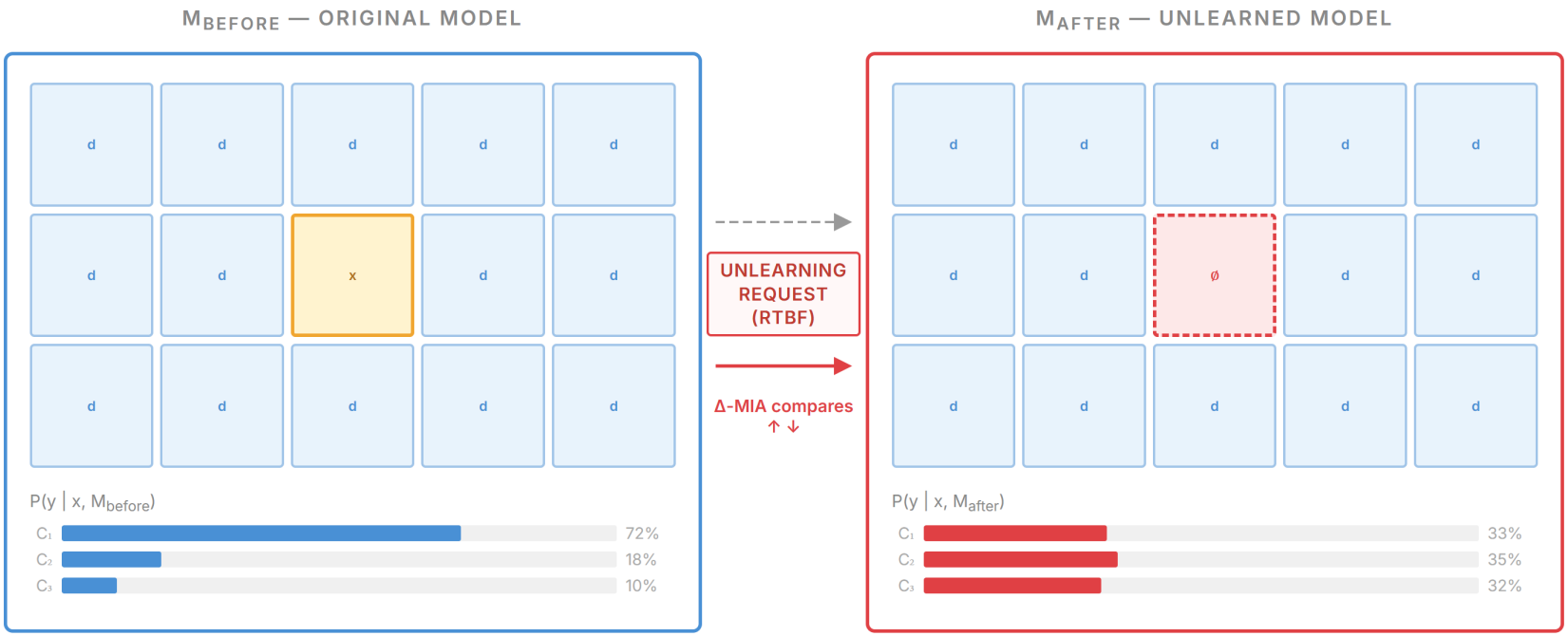
- It can make sense to record the successive versions of an ML in a blockchain.
- This helps provenance, accountability, regulatory auditability, or coordination across non-mutually trusted organizations (*e.g.*, in federated learning).
- In machine unlearning, blockchain can help prove which model versions existed before and after an unlearning request.

Clash between auditability and unlearning

- Implicit assumption in unlearning: the model before unlearning, M_{before}, becomes unavailable once it is replaced by the model after unlearning, M_{after}.
- If an attacker has access to both M_{before} and M_{after}, and the class probabilities output by both models for an input item x differ substantially, then x was in the training data of M_{before}[2].

[2] J. Domingo-Ferrer, N. Jebreel, and D. Sánchez, "Defenses against membership inference attacks on unlearned data", in *MDAI 2025*, LNAI 15957, pp. 145-149.

MIAs by comparing before and after unlearning



Possible mitigations

- Introduce guardrails that make class probabilities indistinguishable between M_{before} and M_{after} for unlearned items. $\Longrightarrow$ Problematic if models are inspectable.
- Do not record in blockchain all details/weights of successive models. $\Longrightarrow$ Not recording in blockchain is not a solution if successive models are downloadable by attackers.

Outline

Conclusions

- Our analysis casts doubts on the effectiveness of privacy attacks against ML in real-world conditions.
- MIAs are limited due to training data: non-exhaustivity and diversity of confidential attributes.
- They are also limited due to the nature of target models and attack methods.
- Property inference attacks are useful to audit general properties of the training data, but in general do not disclose data on any subject.
- Reconstruction attacks based on MIAs inherit MIA limitations and incur very significant cost: an MIA is needed for each potential data point.
- Reconstruction based on gradient inversion is used in generative ML, but it requires access to gradients and lacks a success criterion.

Conclusions (II)

- Since current privacy attacks are not that effective, utility-damaging defenses (such as DP) may often be unnecessary.
- Trustworthy ML regarding privacy may be easier to implement than assumed so far.
- This is good news for jurisdictions like the EU that are committed to reconciling strong AI regulation with competitiveness of their AI industry.

Conclusions (III)

- In unlearning, MIAs have a positive role to assess the effectiveness of approximate unlearning.
- However, auditability may facilitate MIA attacks based on comparing successive unlearned models (auditability-forgetting clash).
- Guardrails can be used to ensure indistinguishability of unlearned items, but they can be removed for inspectable models (inspectability-forgetting clash).

Further details

J. Domingo-Ferrer, "How worrying are privacy attacks against machine learning?", in *DPM-ESORICS 2025*, LNCS 16231, Springer, pp. 3-15.
J. Domingo-Ferrer, N. Jebreel, and D. Sánchez, "Defenses against membership inference attacks on unlearned data", in *MDAI 2025*, LNAI 15957, pp. 145-149.
N. Jebreel, M. Khalil, D. Sánchez, and J. Domingo-ferrer, "Revisiting the LiRA membership inference attack under realistic assumptions", in *PETS 2026*.
N. Jebreel, D. Sánchez and J. Domingo-Ferrer, "A critical review on the effectiveness and privacy threats of membership inference attacks", in *ESORICS 2026*.

Gratias vobis ago pro attentione vestra!

Thank you for your attention!

Gràcies per la vostra atenció!