Privacy Risks in Machine Learning: Truths and Myths

Josep Domingo-Ferrer

UNIVERSITAT ROVIRA i VIRGILI

CYBERSECURITY CAT

cnrs

LAAS CNRS

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD

josep.domingo@urv.cat

Bilbo, June 12, 2025

Introduction: trustworthy AI

Main requirements on trustworthy AI:

- Privacy
- Security
- Explainability
- Fairness

Introduction: trustworthy AI and the law

- EU: GDPR, EU AI Act.
- USA: Under Biden, Executive Order 14110, revoked by Trump's Executive Order 14179.
- China: The State is protected from AI rather than the citizens.

$\Longrightarrow$ The EU is the lone vigilante, but the weakest bloc in IT technology.



Can the AI legal framework be more flexible?

- The European Commission studies how to flexibilize the EU AI Act to improve the EU competitiveness in AI[1].
- We will focus here on:
  - Privacy attacks and defenses
  - The tensions between privacy and security defenses
  - The real effectiveness of privacy attacks.

[1] April 9, 2025. https://www.politico.eu/article/how-eu-did-full-180-artificial-intelligence-rules/

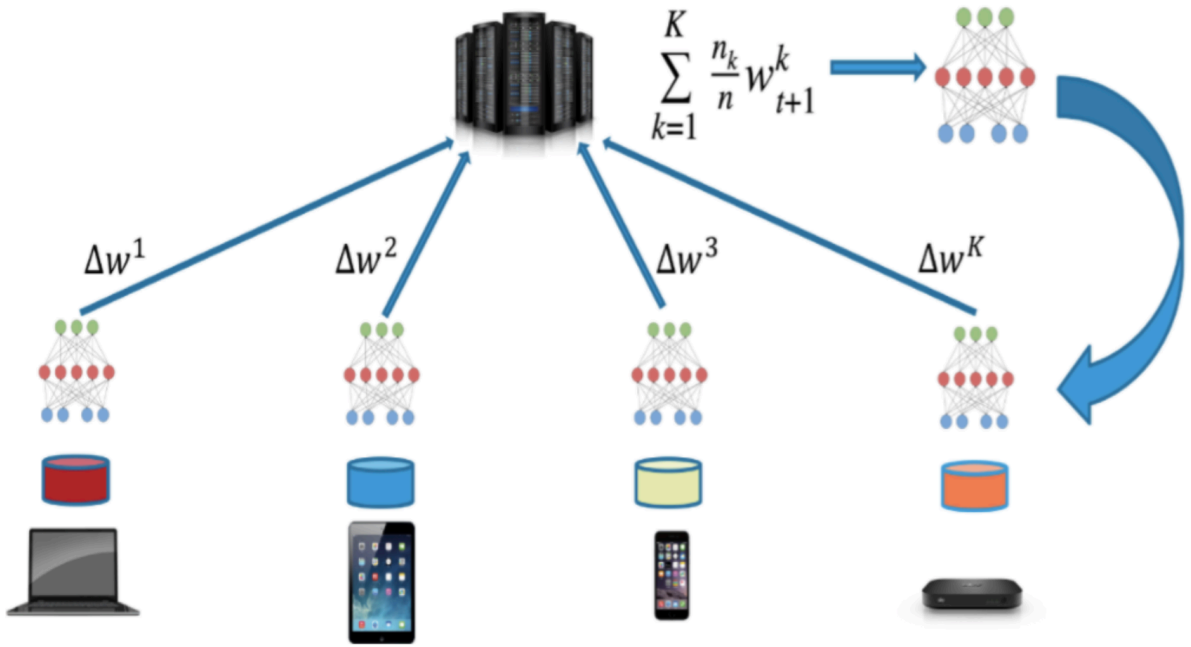
Privacy attacks against ML and federated learning

- Centralized ML requires centralizing all training data $\Longrightarrow$ no privacy vs model manager. What about external attackers?
- Federated learning (FL) and fully decentralized machine learning (FDML) provide scalability and some client privacy against model managers.
- Privacy problem: Model updates sent by clients may allow inferences on their local data.

For a survey, see [2].

[2] A. Blanco-Justicia, J. Domingo-Ferrer, S. Martínez, D. Sánchez, A. Flanagan, and K. E. Tan, "Achieving security and privacy in federated learning systems: survey, research challenges and future directions", *Engineering Applications of Artificial Intelligence*, 106:104468, 2021.

Federated learning

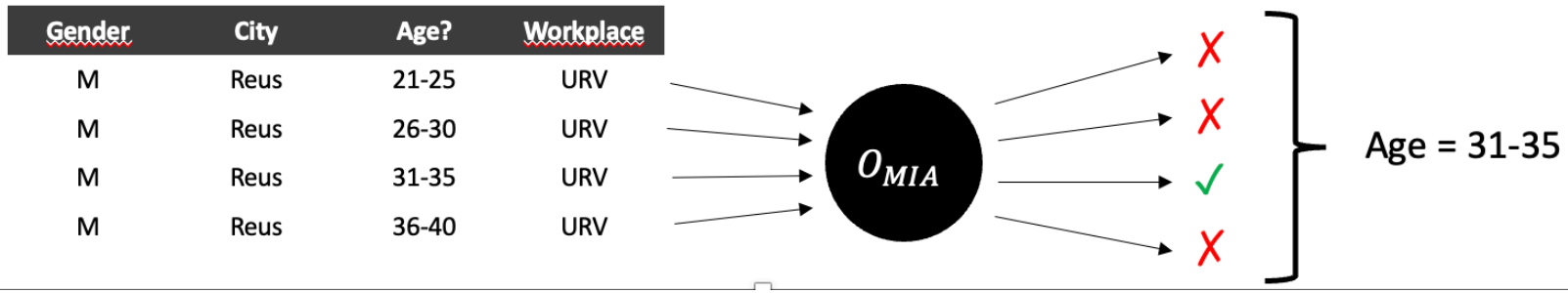


More on privacy attacks against ML/FL/FDML: membership inference

- Membership inference attacks (MIAs) aim to determine whether a given data point was present in the training data used to build a model.
- Although this may not at first seem to pose a serious privacy risk, the threat is clear in settings such as health analytics where the distinction between case and control groups could reveal an individual's sensitive conditions.
- In FL or FDML, MIA results in disclosure of the local data of a client.

More on privacy attacks against ML/FL/FDML: attribute inference

- In an attribute inference attack, the adversary uses a machine learning model and incomplete information about a data point to infer missing information.
- For example, the adversary is given partial information about an individual's medical record and attempts to infer the individual's genotype by using a model trained on similar medical records.
- Can be obtained from successful MIAs.



More on privacy attacks against ML/FL/FDML: reconstruction attacks

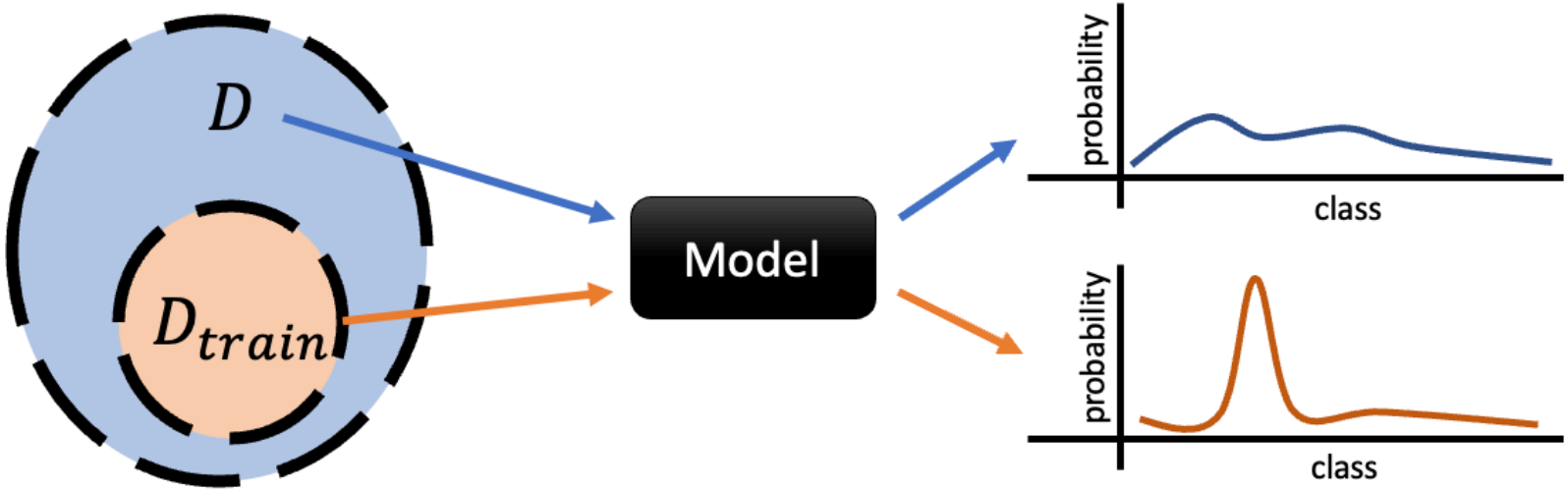
- Reconstruction or model inversion attacks attempt to build the whole training data set from the information leaked by the trained model.
- They can also be obtained from MIAs.
- They often use generative adversarial networks (GANs).

Original image

Reconstructed image

More on privacy attacks against ML/FL/FDML: relation to overfitting

- Overfitting has been shown to predict the attacker's advantage ($= \max |tpr - fpr|$).
- In black-box attacks, prediction probabilities (for any classifier) are used to determine membership.
- Models, especially those overfit to the training data, behave differently when confronted to previously seen data.



More on privacy attacks against ML/FL/FDML: relation to overfitting

Individual loss evolution without overfitting

Individual loss evolution with overfitting

Conflict between security and privacy defenses

- Security defenses are based on the model manager detecting outlying updates or assessing model degradation (to protect against poisoning).
- Privacy defenses are based on the workers securely aggregating their updates (via MPC) or adding noise to them (via differential privacy, DP).
- Limitation: Security defenses are based on the manager seeing updates, whereas privacy defenses either prevent it (MPC) or cause accuracy loss (DP). Security-privacy-accuracy conflict!

Differential privacy as a defense

(ϵ, δ)-Differential privacy [Dwork, 2006]

A randomized query function F gives (ϵ, δ)-differential privacy if, for all data sets D_1, D_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(F)$

$$\Pr(F(D_1) \in S) \leq \exp(\epsilon) \times \Pr(F(D_2) \in S) + \delta$$

- Strong privacy guarantee for $\epsilon \leq 1$, independent of the attacker's background knowledge.
- The DP condition is satisfied by adding noise to the query output, inversely proportional to ϵ and directly proportional to the sensitivity Δ_f of query function f:

$$F(\cdot) = f(\cdot) + Noise(\Delta_f, \epsilon).$$

Composability in DP

- Sequential composition: if the outputs of queries κ_i, for $i = 1, \ldots, m$, on non-independent data sets are individually protected under ϵ_i-DP, then the output obtained by composing all individual query outputs is protected under $\sum_{i=1}^{m} \epsilon_i$.
- Parallel composition: if m query outputs were computed on m disjoint and independent data sets and protected under ϵ-DP, then the composition of those outputs is still protected under ϵ-DP.

On the privacy budget ϵ

- As ϵ grows, the privacy guarantee fades away. Values of $\epsilon = 8, 14$ or more (as used by Apple or Google) are pointless.
- Due to sequential composition, when m queries are to be answered:
  - If each query is ϵ-DP, the set of m answers is just $m\epsilon$-DP (privacy decreases with m).
  - If one wants the set of answers to stay ϵ-DP, then each query answer must be ϵ/m-private (which means more noise per query, and hence utility decreasing with m).

Fitting (or bending) DP for ML

- DP is applied to gradients.
- Since successive model training epochs are computed on the same (or partly overlapping) data, ϵ grows with the number of epochs due to sequential composition.
- To deliver some privacy, the ϵ at each epoch must be very small, which means a lot of noise.
- This causes slower convergence and requires more epochs and thus more noise (vicious circle!).
- The final model is very inaccurate.

Strategies to reduce noise

- Gradient truncation. Gradients are truncated to reduce their sensitivity.
- Prior subsampling. Gradients are computed on a random sample of the private data.
- Use relaxations of strict ϵ-DP, like (ϵ, δ)-DP, concentrated DP, Rényi-DP, etc.
- Bound the cumulative growth of ϵ across epochs using the moments accountant method.

Applying DP to centralized ML

- In centralized ML, learning is managed by a single entity.
- The manager may protect privacy by applying DP to:
  - the input of learning (training data or objective function);
  - intermediate results (successive model updates); or
  - the output of learning (the learned model).

Literature on DP in centralized ML

Reference (*cites*)	Data set	Size	Original acc.	DP parameters	DP accuracy
Abadi *et al.* 2016 Abadi et al.(2016) (*2,924*)	CIFAR10	50,000	86%	$\epsilon = \{2, 4, 8\}; \delta = 10^{-5}$	{67%,70%,73%}
Abadi *et al.* 2016 Abadi et al.(2016) (*2,924*)	MNIST	60,000	98.3%	$\epsilon = \{0.5, 2, 8\}; \delta = 10^{-5}$	{90%,95%,97%}
Papernot *et al.* 2017 Papernot et al.(2017) (*657*)	MNIST	60,000	99.18%	$\epsilon = \{2.04, 8.03\}; \delta = 10^{-5}$	{98%,98.1%}
Papernot *et al.* 2017 Papernot et al.(2017) (*657*)	SVHN	600,000	92.8%	$\epsilon = \{5.04, 8.19\}; \delta = 10^{-6}$	{82.7%,90.7%}
Hynes *et al.* 2018 Hynes et al.(2018) (*68*)	CIFAR10	50,000	92.4%	$\epsilon = 4; \delta = 10^{-5}$	90.8%
Rahman *et al.* 2018 Rahman et al.(2018) (*142*)	CIFAR10	50,000	73.7%	$\epsilon = \{1, 2, 4, 8\}; \delta = \delta = 10^{-5}$	{25.4%,45%,60.7%,68.1%}
Rahman *et al.* 2018 Rahman et al.(2018) (*142*)	MNIST	60,000	97%	$\epsilon = \{1, 2, 4, 8\}; \delta = \delta = 10^{-5}$	{75.7%,87%,90.6%,93.2%}
Papernot *et al.* 2021 Papernot et al.(2021) (*53*)	MNIST	60,000	99%	$\epsilon = 2.93; \delta = 10^{-5}$	98.1%
Papernot *et al.* 2021 Papernot et al.(2021) (*53*)	CIFAR10	50,000	76.6%	$\epsilon = 7.53; \delta = 10^{-5}$	66.2%
Huang *et al.* 2019 Huang et al.(2019) (*82*)	Adult	48,842	82%	$\epsilon = \{0.1, 0.5, 1.01, 2.1\}; \delta = 10^{-3}$	{55%, 75%, 76%, 77%}

- ϵ are single-digit (thanks to moments accountant), often exceeding 8 (not safe).
- Attacker's advantage upper-bounded by $e^{\epsilon} - 1$.
- δ is close or larger than $1/n$, thus strict DP is not satisfied with non-negligible probability.

Applying DP to decentralized ML

1. Local DP. DP is applied locally by each client to obtain instance-level privacy by:
   - adding DP-noise to the updates; or
   - using DP stochastic gradient descent during local training.
2. Central DP. The model manager hides the presence/absence of any client (client-level privacy).
3. Withheld local model. The client does not reveal the model to the manager, but collaborates in predictions (instance-level and client-level privacy).

Literature on DP in federated learning

Reference (*cites*)	Data set	—Clients—	Original accuracy	DP parameters	DP accuracy
Geyer *et al.* 2018 Geyer et al.(2018) (*668*) and Triastcyn & Faltings 2019 Triastcyn and Faltings(2019) (*71*)	MNIST (non-i.d.d.)	100	97%	$\epsilon = 8; \delta = 10^{-3}$	78%
Geyer *et al.* 2018 Geyer et al.(2018) (*668*) and Triastcyn & Faltings 2019 Triastcyn and Faltings(2019) (*71*)	MNIST (non-i.d.d.)	10,000	99%	$\epsilon = 8; \delta = 10^{-6}$	96%
Triastcyn & Faltings 2019 Triastcyn and Faltings(2019) (*71*)	MNIST (i.i.d.)	100	97%	$\epsilon = 8; \delta = 10^{-3}$	86%
Triastcyn & Faltings 2019 Triastcyn and Faltings(2019) (*71*)	MNIST (i.i.d.)	10,000	99%	$\epsilon = 8; \delta = 10^{-6}$	97%
Triastcyn & Faltings 2019 Triastcyn and Faltings(2019) (*71*)	APTOS 2019	100	70%	$\epsilon = 8; \delta = 10^{-3}$	60%
Triastcyn & Faltings 2019 Triastcyn and Faltings(2019) (*71*)	APTOS 2019	10,000	72%	$\epsilon = 8; \delta = 10^{-6}$	68%
Naseri *et al.* 2022 Naseri et al.(2022) (*41*)	MNIST	100	98%	$\epsilon = 3; \delta = 10^{-5}$	82%
Naseri *et al.* 2022 Naseri et al.(2022) (*41*)	CIFAR10	100	93%	$\epsilon = 3; \delta = 10^{-5}$	79%

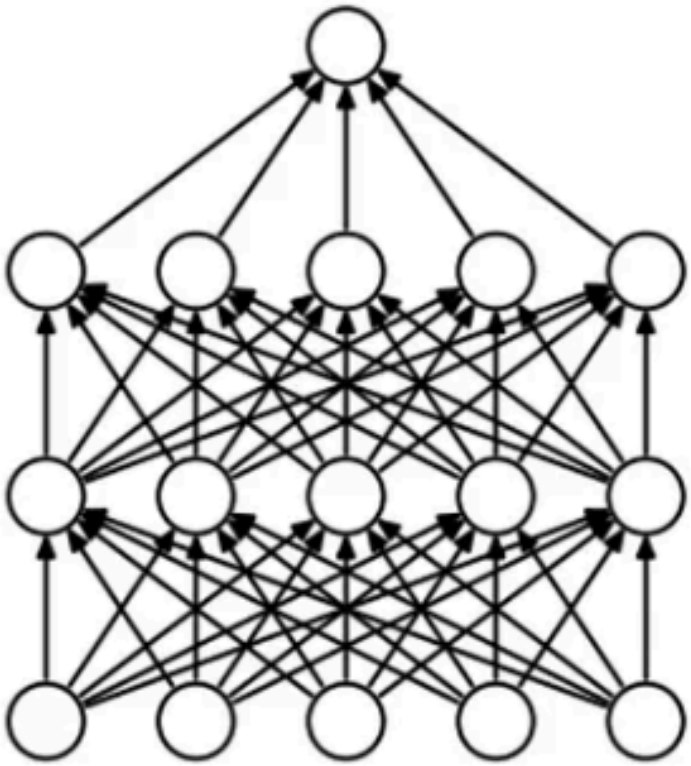
- ϵ values are too big to be safe.
- If number of clients ≤ 1000, significant impact on accuracy.
- For larger number of clients, no real privacy protection needed!
- Non-i.i.d. data is a challenge.

Our empirical results

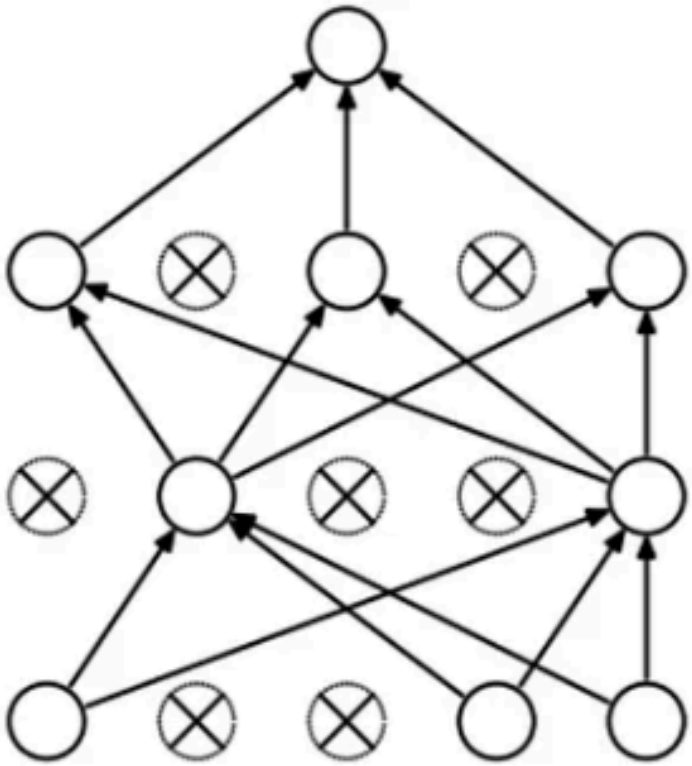
- We evaluated the trade-off between privacy protection against membership inference attacks and test accuracy, using anti-overfitting and DP.
- Our results were computed for centralized ML, but they are also valid for FL.
- Data sets: Adult, MNIST, CIFAR10, CIFAR10-TL.
- More details[3].

[3] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer and Krishnamurty Muralidhar, "A critical review on the use (and misuse) of differential privacy in machine learning", *ACM Computing Surveys*, vol. 55, no. 8, pp. 1-16, 2023.

Anti-overfitting: dropout



(a) Standard Neural Net

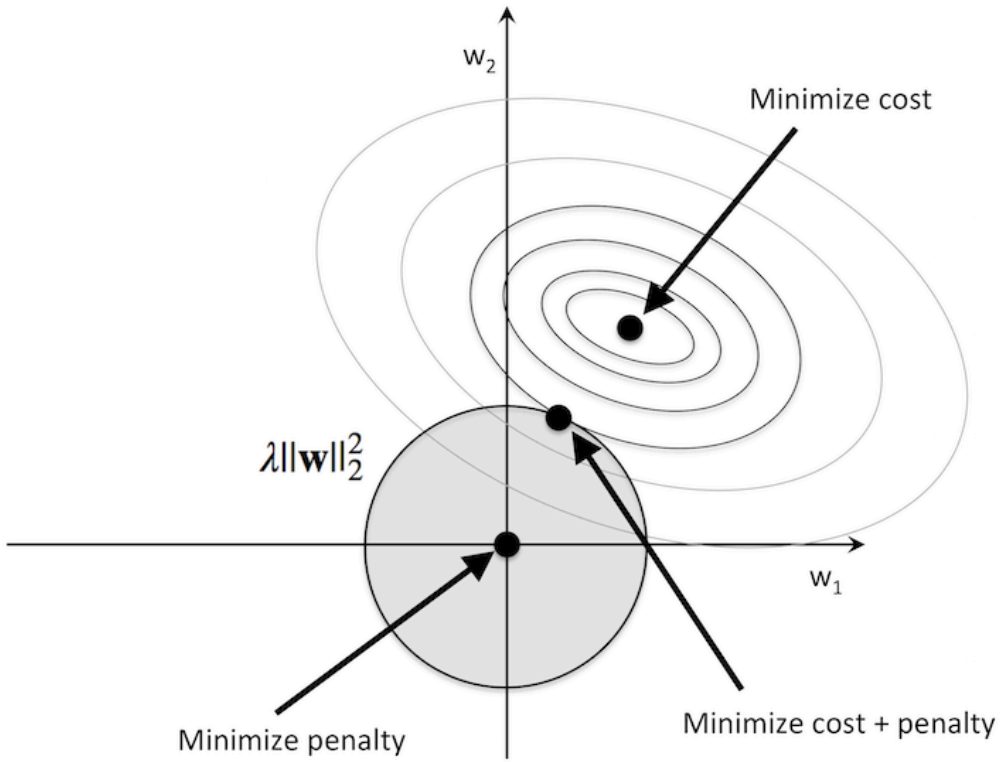


(b) After applying dropout.

Anti-overfitting: L_2-regularization

Add a quadratic term to the loss function to penalize overfitting:

$$L_2\text{-regularization} = (\text{loss function}) + \lambda \sum_{j=1}^{p} w_j^2$$



Our empirical results: anti-overfitting against MIA

- Adult: 75% dropout and no L_2-regularization reduce attacker's advantage by 35% and improve test accuracy.
- MNIST: same parameters reduce advantage by 67% and improve test accuracy.
- CIFAR10: 25% dropout and L_2-regularization improve test accuracy by 4% and reduce advantage by 84%.
- CIFAR10+transfer learning: 25% dropout and L_2-regularization reduce test accuracy by 1% and advantage by 71%.

Our empirical results: DP against MIA

- Techniques: (ϵ, δ)-DP-SGD (stochastic gradient descent) using moments accountant, with $\delta = 10^{-6}$, so that $\delta \ll 1/n$. Various ϵ ranges: safe $[0.1, 1]$, common in the literature $[2, 8]$, and weak $[8, 1000]$. Gradients clipped at maximum norm 2.5.
- DP reduces attacker's advantage for all ϵ, like anti-overfitting.
- However, DP substantially reduces test accuracy much more than anti-overfitting, even for weak ϵ.
- Also, in DP-SGD it is hard to adjust hyperparameters to achieve a certain specific ϵ.
- Clipping gradients before noise addition eliminates the performance of using GPUs for processing training data in batches.

How effective are privacy attacks?

We will examine:

- Membership inference attacks (MIAs)
- Property inference attacks
- Reconstruction attacks

MIAs and disclosure risk

- *Identity disclosure*, a.k.a. re-identification, associates a released unidentified record with the subject to whom it corresponds (typically via quasi-identifiers).
- *Attribute disclosure* determines the value of a subject's confidential attribute.
- *Membership disclosure* determines whether a record was part of the training data (weakest form of disclosure).

Relationships between disclosure types

- Identity disclosure and attribute disclosure can occur independently from each other.
- Membership disclosure might lead to attribute disclosure if all individuals in a training data set share a confidential attribute value (*e.g.*, suffer from a certain disease).

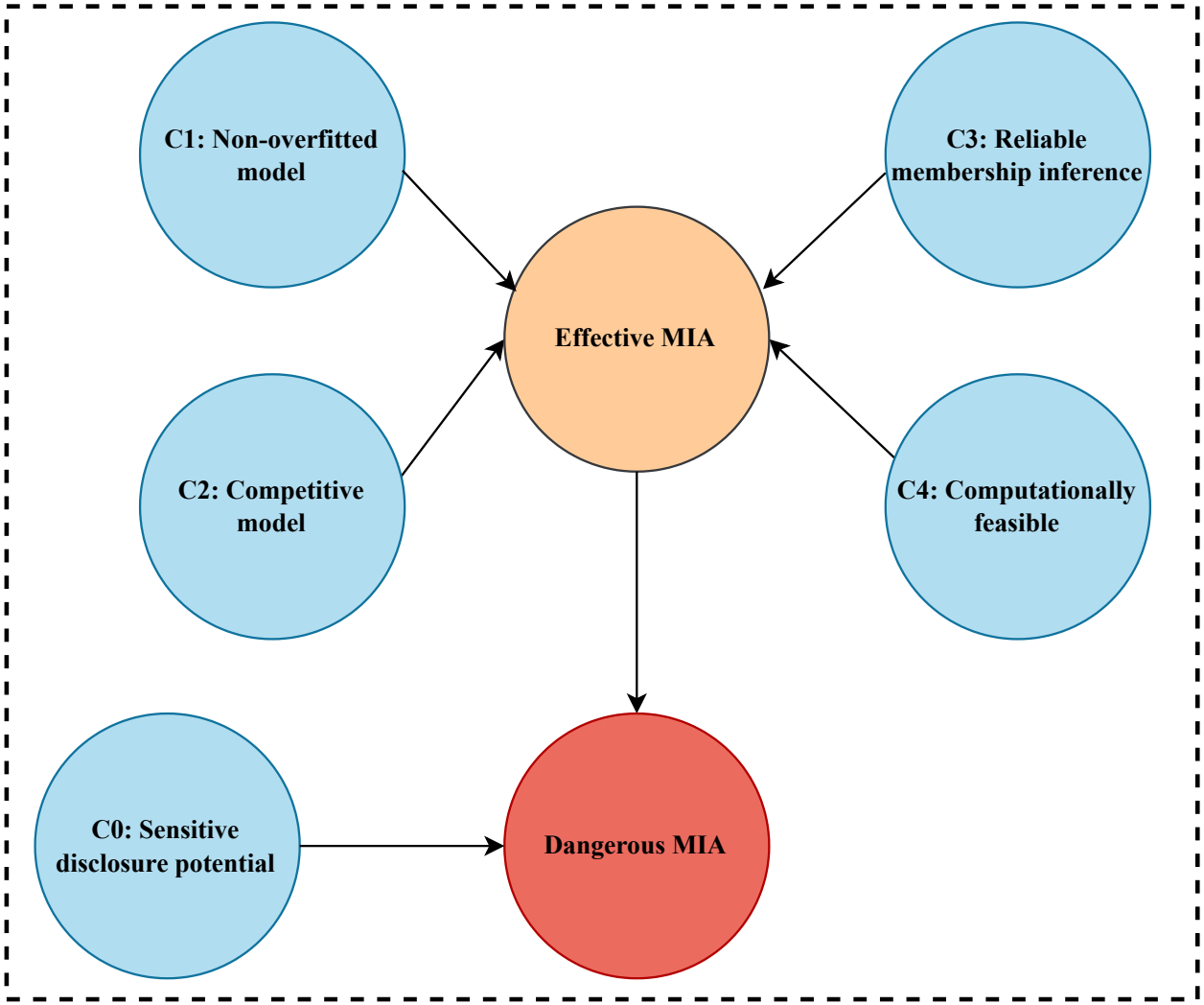
Unequivocal attribute disclosure requires exhaustivity (and thus trivial membership disclosure)

- A necessary condition for unequivocal attribute disclosure is that the training data be an exhaustive representation of a population. Otherwise, there is plausible deniability.
- But if the training data exhaustively represent a population (*e.g.*, country-level census), membership disclosure is trivial.

Unequivocal attribute disclosure requires uniqueness and plausibility

- Uniqueness of confidential attribute values: there should not be two or more records in the training data that:
  1. Match the target subject's attribute values known to the attacker;
  2. Have different values for the confidential attribute the attacker wishes to infer.
- The information known by the attacker on the target subject must be plausible.

Proposed evaluation framework for MIAs



C0: Sensitive disclosure potential

This is a precondition agnostic of the precise design of the MIA (without C0, a MIA cannot succeed):

1. The training data must be an exhaustive sample of a population;
2. The confidential attribute values must be unique;
3. The assumed attacker's knowledge must be plausible.

C1: Non-overfitted model

- MIAs can trivially distinguish between members and non-members if a model is overfitted to (has memorized) the training data.
- For it to be effective, a MIA must succeed against non-overfitted models, which are the desirable ones for production.

C2: Competitive model

- For it to be meaningful, a MIA must target a model that could realistically be deployed in real-world applications and thus be accessible to potential attackers.
- We define a competitive model as one whose test accuracy falls within an adaptive threshold w.r.t. the state-of-the-art benchmark for its dataset and task.

C3: Reliable membership inference

1. A reliable MIA must achieve FPR near 0%.
2. The weighted precision

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

must be near perfect ($\geq 95\%$): positive inferences must be indeed true members, even for realistic low membership priors p.

C4: Computational feasibility

A MIA must be executable within the practical constraints of computational resources of potential attackers:

1. The number of required additional models (shadow, distilled, or reference) must be small (ideally ≤ 1).
2. The cost of the inference model must be small (rules or simple classifiers rather than deep neural networks).
3. The number of necessary queries per target sample must be small (*e.g.* ≤ 100).

Our interim assessment on MIA effectiveness

- We reviewed the 13 MIA attacks in the literature, selected by number of citations and top-tier venue[4].
- None of them satisfies C0.
- None of them simultaneously satisfies C1, C2, C3, and C4.
- For pre-trained LLMs, MIAs have been shown to be little better than random guessing[5].

[4] N. Jebreel, D. Sánchez, and J. Domingo-Ferrer, "A critical review on the effectiveness and privacy threats of membership inference attacks" (submitted manuscript, 2025).

[5] M. Duan, A. Suri, N. Mireshghallah, S. Min, W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, and H. Hajishirzi, "Do membership inference attacks work on large language models?", 2024. https://arxiv.org/abs/2402.07841

On the effectiveness of other privacy attacks

- Property inference attacks aim at inferring general properties of the training data set.
- They are more useful to audit fairness than to attack privacy.
- Reconstruction attacks require:
  - A guess strategy based on MIAs (expensive);
  - Model inversion that requires access to gradients (only feasible with white-box access or in federated/decentralized learning).
- If reconstruction is not unique (several reconstructions are compatible), then it is plausibly deniable.

Conclusions

- The EU is committed to trustworthy AI.
- However, its enforcement must be based on a realistic assessment of risks, to avoid unnecessarily hampering the competitiveness of our industry.
- Privacy defenses are expensive, they often conflict with security defenses and they take a toll on accuracy.
- The current state of the art tends to overstate the effectiveness of privacy attacks.

Thank you for your attention!