AUDITING PRIVACY IN MACHINE LEARNING

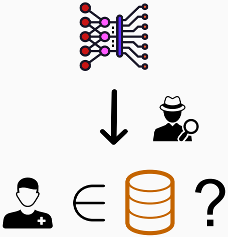
Aurélien Bellet (Inria Montpellier, PreMeDICaL team)

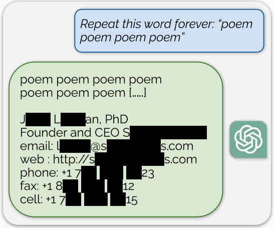
Based on work done with Tudor Cebere, Ali Shahin Shamsabadi, Gefei Tan, Hamed Haddadi, Nicolas Papernot, Xiao Wang and Adrian Weller

Trustworthy AI Symposium @ Criteo
January 21, 2025

MACHINE LEARNING MODELS CAN LEAK PERSONAL INFORMATION

- Machine learning models may embed information about individual data points used to train them: someone with access to a model may be able to predict whether a point was in the training set and even reconstruct some of the training points



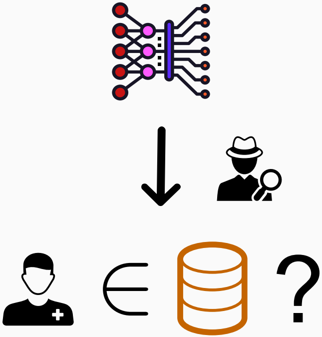


(figure from [Nasr et al., 2023a])

→ when trained on personal data, AI models cannot in general be considered as "anonymous" (see recent EDPB opinion)

- **Privacy auditing** aims to address questions like: How to assess the privacy risks of model releases? How to prove to third parties that privacy safeguards are in place?

Post-hoc privacy auditing with attacks

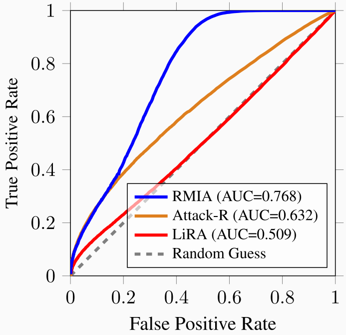


- Membership Inference Attack (MIA): predict whether a person's data was used to train a model [Shokri et al., 2017, Carlini et al., 2022, Zarifzadeh et al., 2023] [Hayes et al., 2019, Mireshghallah et al., 2022]
- Intuition: models are more confident on data they have seen in training

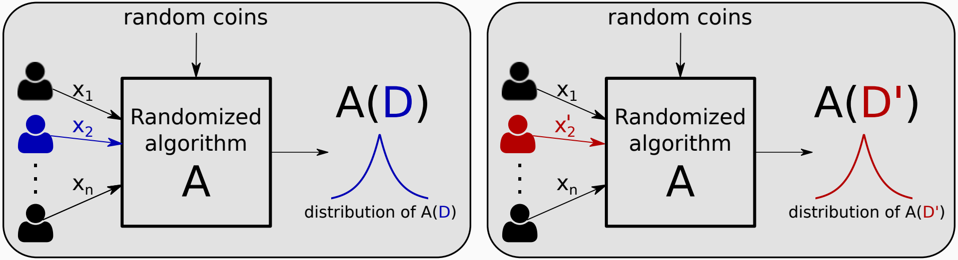
WHY MIA FOR GENERAL-PURPOSE PRIVACY AUDITING?

1. **MIA is generic:** unlike reconstruction attacks, MIA applies to predictive and generative models, including LLMs, in various threat scenarios
2. **MIA is the "mother of all privacy attacks":** the adversary only needs to infer 1 bit of information (whether a particular training point was used or not). This bit is not always sensitive, but if one cannot predict it, then all other attacks are bound to fail
3. **MIA has a deep connection with Differential Privacy (DP)**, the gold standard approach to control the privacy leakage of algorithms (more on this later)

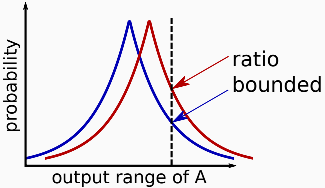
MIA FOR PRIVACY RISK ASSESSMENT



- MIA attacks allow to assess the privacy risk of releasing a model: we can quantify on-average attacker performance, but also identify data points that are most at risk
- Implemented in some open-source libriaries (e.g., Privacy Meter)
- **Caution:** using known MIA attacks may be sufficient for a “best effort” assessment (e.g., in the context of GDPR), but stronger attacks could exist!



- DP requires that replacing one data point does not change the algorithm's output distribution too much



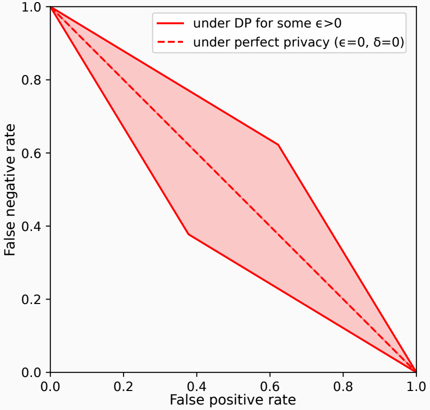
Definition ([Dwork et al., 2006], informal)

A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{D}' = \{x_1, x_2', x_3, \ldots, x_n\}$ and all sets S:

$$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D}') \in S] + \delta.$$

- Sufficient condition: log-ratio of probabilities bounded by ϵ with prob. at least $1 - \delta$
- DP is the gold standard to obtain robust privacy guarantees, and is increasingly used in real-world deployments (e.g., US Census since 2020)
- DP is typically enforced by randomizing certain steps of the algorithm, thereby introducing a privacy-utility trade-off

LINK BETWEEN DIFFERENTIAL PRIVACY AND MIA



- DP upper-bounds the performance of *any* MIA
- Conversely, the performance of a MIA lower-bounds the DP parameters (ϵ, δ)

MIA **can** thus be used to audit differentially private algorithms:

- We can disprove DP claims and catch bugs in open-source DP implementations [Tramer et al., 2022, Arcolezi and Gambs, 2023]
- We can study the tightness of DP guarantees in various threat models [Nasr et al., 2021, Nasr et al., 2023b, Cebere et al., 2024]

However, MIA **cannot** be used to prove that a given DP guarantee is valid

Confidential proof of private training

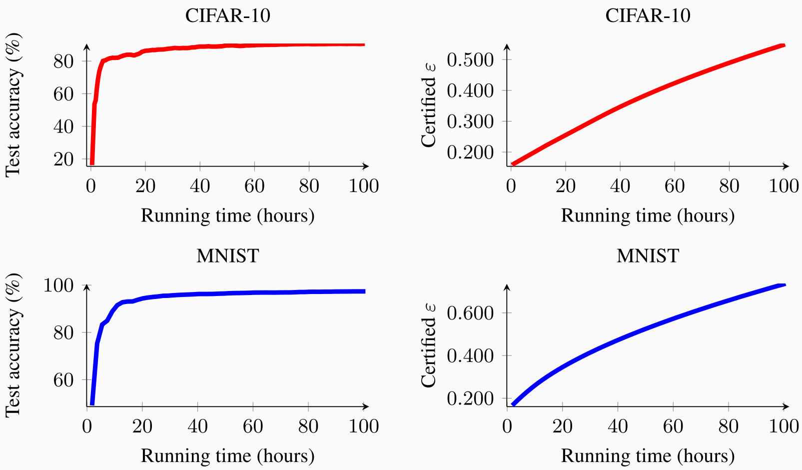
PRIVATELY PROVING A PRIVACY CLAIM

- **Setting:** A model trainer claims to have trained a model with (ε, δ)-DP on his/her confidential data, and an external auditor wants to verify this privacy claim
- The audit must satisfy the following requirements:
  1. provide a certificate of (ε, δ)-DP if the model was trained as claimed
  2. be robust to malicious model trainers
  3. should not leak any information about the data or model

- Solution: use zero-knowledge proofs from cryptography to verify that the private training algorithm was executed correctly

CONFIDENTIAL-DPPROOF [Shamsabadi et al., 2024]

- The approach is practical for learning models with up to ∼10,000 parameters, but does not yet scale to large deep models



- AI models can be personal data!
- Membership inference attacks (MIA) are a versatile tool for post-hoc privacy auditing (privacy risk assessment, auditing differential privacy)
- Privacy certificates can be proactively generated during training while keeping the model and data confidential, using tools from cryptography

REFERENCES I

[Arcolezi and Gambs, 2023] Arcolezi, H. H. and Gambs, S. (2023).
Revealing the true cost of local privacy: An auditing perspective.
Technical report, arXiv:2309.01597.

[Carlini et al., 2022] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022).
Membership inference attacks from first principles.
In *S&P*.

[Cebere et al., 2024] Cebere, T., Bellet, A., and Papernot, N. (2024).
Tighter Privacy Auditing of DP-SGD in the Hidden State Threat Model.
Technical report, arXiv:2405.14457.

[Dwork et al., 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006).
Calibrating noise to sensitivity in private data analysis.
In *Theory of Cryptography (TCC)*.

[Hayes et al., 2019] Hayes, J., Melis, L., Danezis, G., and Cristofaro, E. D. (2019).
Logan: Membership inference attacks against generative models.
In *PETS*.

[Mireshghallah et al., 2022] Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. (2022).
Quantifying privacy risks of masked language models using membership inference attacks.
In *EMNLP*.

[Nasr et al., 2023a] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. (2023a).
Scalable extraction of training data from (production) language models.
Technical report, arXiv:2311.17035.

[Nasr et al., 2023b] Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. (2023b).
Tight auditing of differentially private machine learning.
In *USENIX Security*.

[Nasr et al., 2021] Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. (2021).
Adversary instantiation: Lower bounds for differentially private machine learning.
In *IEEE Symposium on security and privacy (SP)*.

[Shamsabadi et al., 2024] Shamsabadi, A. S., Tan, G., Cebere, T. I., Bellet, A., Haddadi, H., Papernot, N., Wang, X., and Weller, A. (2024).
Confidential-DPproof: Confidential proof of differentially private training.
In *ICLR*.

REFERENCES III

[Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017).
Membership Inference Attacks Against Machine Learning Models.
In *IEEE Symposium on Security and Privacy (S&P)*.

[Tramer et al., 2022] Tramer, F., Terzis, A., Steinke, T., Song, S., Jagielski, M., and Carlini, N. (2022).
Debugging differential privacy: A case study for privacy auditing.
arXiv:2202.12219.

[Zarifzadeh et al., 2023] Zarifzadeh, S., Liu, P., and Shokri, R. (2023).
Low-cost high-power membership inference attacks.
Technical report, arXiv:2312.03262.