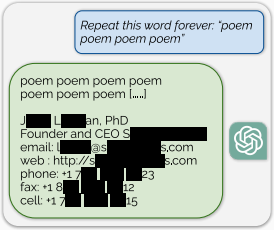
PRIVACY IN MACHINE LEARNING

Aurélien Bellet (Inria)

Learning Theory Summer School
Copenhagen
June 22-23, 2026

MACHINE LEARNING MODELS CAN LEAK PERSONAL INFORMATION

- Machine Learning (ML) models may inadvertently memorize information about individual data points, making it possible to reconstruct some training examples



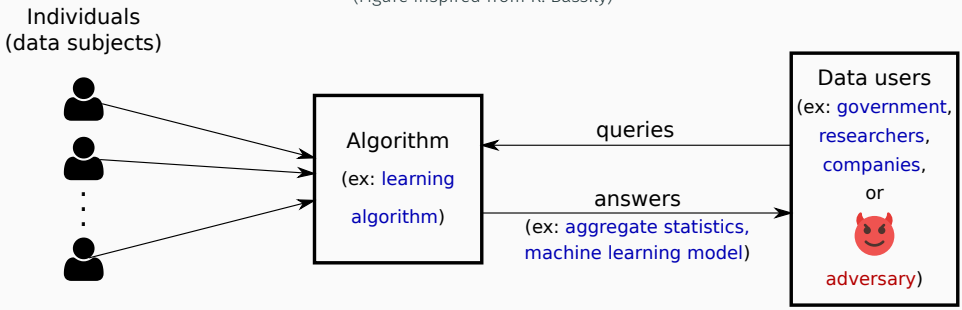
(figure from [Nasr et al., 2023a])

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

- **Key questions:** How can we measure and characterize privacy leakage in AI models? How can we design training algorithms that provably control such leakage?

GENERAL FRAMEWORK: PRIVATE DATA ANALYSIS

(Figure inspired from R. Bassily)



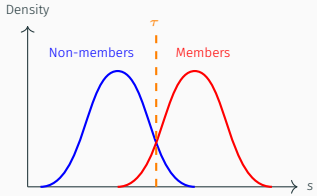
- **Goal:** achieve utility while preserving privacy (conflicting objectives!)
- This is separate from security concerns (e.g., unauthorized access to the system)

Membership Inference Attack (MIA):
Given a record x and a model θ, determine if $x \in \mathcal{D}_{\text{train}}$ (member vs. non-member)

- **How it works:** The attacker computes a membership score $s(x, \theta)$ and predicts membership whenever the score exceeds a threshold
  The simplest choice being the negative loss:

$$s(x, \theta) = -\ell(x, \theta)$$

  High score $\rightarrow$ low loss $\rightarrow$ likely member



- **Why it matters:** membership inference is the most fundamental privacy risk (a single bit of info), and serves as a prerequisite for reconstruction attacks ("If you can't infer membership, you can't reconstruct")

PRESENTATION OUTLINE

1. Differential Privacy for Machine Learning
   - A robust privacy definition with guarantees against *any* membership inference attack
   - Private training algorithms with provable privacy guarantees
   - Privacy-utility trade-offs in modern ML

2. Privacy Auditing of ML Models
   - Measuring *empirical* privacy leakage
   - Testing and challenging differential privacy guarantees
   - Scalable and reliable auditing of large models

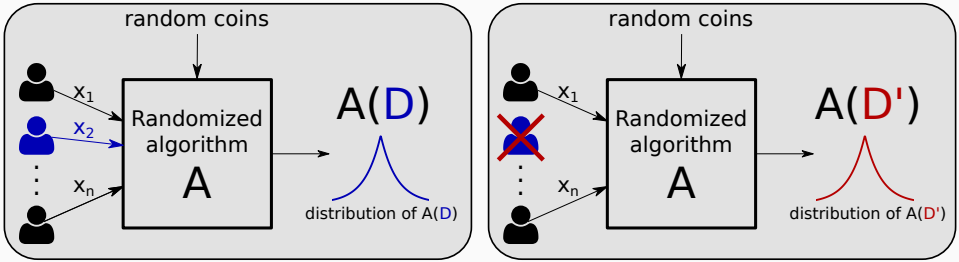
Differential Privacy

TWO KEY REQUIREMENTS FOR A ROBUST PRIVACY DEFINITION

1. Robustness to auxiliary information: privacy guarantees should hold regardless of the adversary's background knowledge, including information acquired in the future

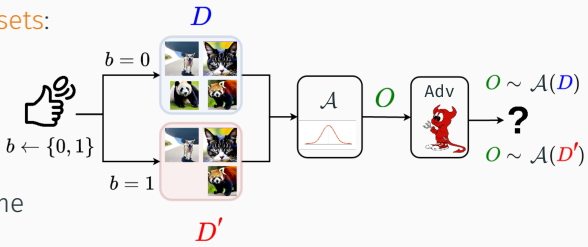
2. Repeated analyses: privacy guarantees should be tractable across multiple analyses of the same dataset, allowing information leakage to be quantified and controlled

KEY INTUITION: TWO PARALLEL WORLDS



- An algorithm $\mathcal{A}$ is private if an adversary observing its output cannot reliably distinguish whether any individual's data was included in the input dataset
- Equivalently, the output distributions $\mathcal{A}(D)$ and $\mathcal{A}(D')$ should be close under an appropriate statistical divergence

- Given the observation $\mathcal{A}(D)$, the adversary's goal is to decide whether a given record is included in D (i.e., membership inference)
- We model the strongest meaningful adversary:
  - the adversary knows the entire dataset except one record
  - only uncertainty: presence/absence of one individual
- This leads to the notion of adjacent datasets: $D \sim D'$ differing in exactly one record
- The attack becomes a hypothesis test: H_0 : data is D' versus H_1 : data is D
- Any attack corresponds to a test with some Type I error α (false positive rate) and Type II error β (false negative rate)



Definition (Trade-off function)

For two distributions $\mathbb{P}$ and $\mathbb{Q}$ on some space $\mathcal{O}$, the trade-off function $T_{\mathbb{P},\mathbb{Q}}$ is the function $[0,1] \to [0,1]$ defined as $T_{\mathbb{P},\mathbb{Q}}(\alpha) = \inf\{\beta_\phi \mid \alpha_\phi \leq \alpha\}$, where the infimum is taken over all measurable functions $\phi : \mathcal{O} \to \{0,1\}$, $\alpha_\phi = \mathbb{E}_{\mathbb{P}}\,\phi$ and $\beta_\phi = \mathbb{E}_{\mathbb{Q}}\,\phi$.

Definition (f-DP [Dong et al., 2022])

Let f be a valid trade-off function. A randomized algorithm $\mathcal{A}$ satisfies f-differential privacy if, for any pair of adjacent datasets $D \sim D'$, we have

$$T_{\mathcal{A}(D),\mathcal{A}(D')}(\alpha) \geq f(\alpha), \quad \forall \alpha \in [0,1]$$

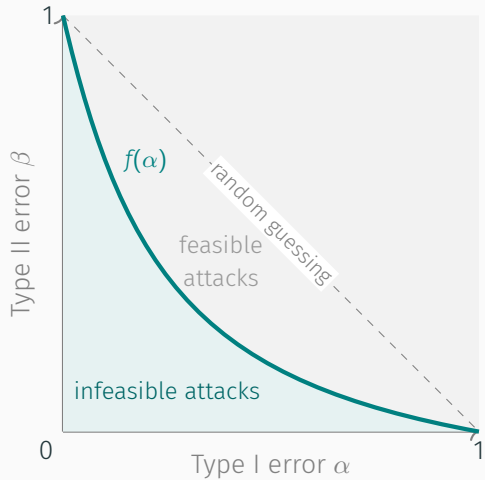
- f-DP controls the trade-off between false positives and false negatives for *any* MIA, over *any* dataset and target point; in particular, no attack can simultaneously achieve both low α and low β, regardless of computational power or auxiliary knowledge
- Note: the algorithm $\mathcal{A}$ can be public, only its randomness needs to remain hidden

VISUALIZING f-DP

f-DP [Dong et al., 2022]:

$\forall \alpha \in [0,1]$:

$$T_{\mathcal{A}(D),\mathcal{A}(D')}(\alpha) \geq f(\alpha)$$



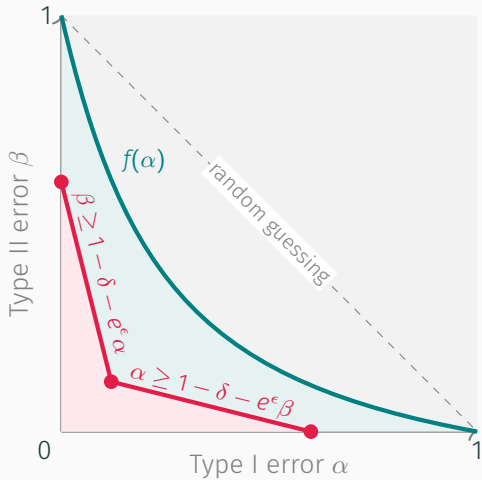
- f-DP defines the entire feasible region of attacks

VISUALIZING f-DP AND COMPARING IT TO (ϵ, δ)-DP

f-DP [Dong et al., 2022]:

$\forall \alpha \in [0,1]$:

$$T_{\mathcal{A}(D),\mathcal{A}(D')}(\alpha) \geq f(\alpha)$$



(ϵ, δ)-DP [Dwork et al., 2006]:

$\forall S \subseteq \mathcal{O}$:

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

equivalent to $f_{\epsilon,\delta}$-DP with

$$f_{\epsilon,\delta}(\alpha) = \max\{0, 1 - \delta - e^{\epsilon}\alpha, \; e^{-\epsilon}(1 - \delta - \alpha)\}$$

- The "historical" (ϵ, δ)-DP definition corresponds to a pair of linear bounds on hypothesis testing errors → typically a coarse summary of the full f-DP curve

TRACTABLE ALTERNATIVES: GAUSSIAN DP AND RÉNYI DP

Definition (μ-GDP [Dong et al., 2022])

Let $\mu > 0$. A randomized algorithm $\mathcal{A}$ satisfies μ-Gaussian DP (GDP) if, for any $D \sim D'$ and $0 \leq \alpha \leq 1$, we have

$$T_{\mathcal{A}(D),\mathcal{A}(D')}(\alpha) \geq T_{\mathcal{N}(0,1),\mathcal{N}(\mu,1)}(\alpha)$$

Definition ((α, ε)-RDP [Mironov, 2017])

Let $\alpha > 1$, $\varepsilon > 0$. A $\mathcal{A}$ satisfies (α, ε)-Rényi DP (RDP) if for any $D \sim D'$, we have

$$D_\alpha\left(\mathcal{A}(D) || \mathcal{A}(D')\right) \leq \varepsilon$$

with D_α the Rényi divergence of order α.

- GDP describes the whole trade-off curve by a single scalar parameter μ (the smaller, the more private)
- RDP takes a divergence view, allowing to leverage known properties
- Both provide explicit conversion to (ϵ, δ)-DP

Theorem (Post-processing)

If $\mathcal{A}$ satisfies f-DP, (ϵ, δ)-DP, μ-GDP, or (α, ε)-RDP, then $g \circ \mathcal{A}$ satisfies the same guarantee.

- "Post-processed distributions can only become more difficult to tell apart" (this is related to Blackwell's informativeness theorem and the data processing inequality)
- From the practical standpoint, this property ensures the privacy guarantee holds regardless of how the algorithm's output is used downstream

KEY PROPERTIES OF DP: SEQUENTIAL COMPOSITION

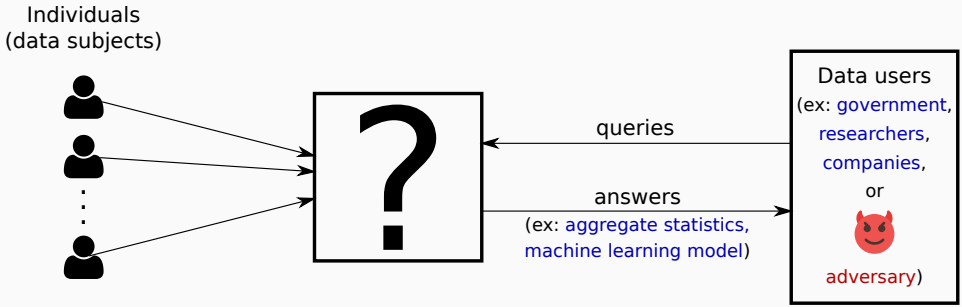
Theorem (Sequential composition)

Let $\mathcal{A}_1 : \mathcal{D} \to \mathcal{O}_1$, $\mathcal{A}_2 : \mathcal{D} \times \mathcal{O}_1 \to \mathcal{O}_2$ and define $\mathcal{A}(D) = (o_1, \mathcal{A}_2(\mathcal{D}, o_1))$ where $o_1 = \mathcal{A}_1(D)$.

- *If $\mathcal{A}_1$ is f_1-DP and $\mathcal{A}_2$ is f_2-DP, then $\mathcal{A}$ satisfies $(f_1 \otimes f_2)$-DP for some operator $\otimes$.*
- *If $\mathcal{A}_1$ is (ϵ_1, δ_1)-DP and $\mathcal{A}_2$ is (ϵ_2, δ_2)-DP, then $\mathcal{A}$ satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.*
- *If $\mathcal{A}_1$ is μ_1-GDP and $\mathcal{A}_2$ is μ_2-GDP, then $\mathcal{A}$ satisfies $\sqrt{\mu_1^2 + \mu_2^2}$-GDP.*
- *If $\mathcal{A}_1$ is (α, ε_1)-RDP and $\mathcal{A}_2$ is (α, ε_2)-RDP, then $\mathcal{A}$ satisfies $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP.*

- Composition controls the cumulative privacy loss of multiple analyses on the same dataset, even when they are chosen adaptively based on previous outputs (e.g., iterative ML algorithms)
- Tracking the privacy loss across multiple computations is called privacy accounting; in practice, this is often done numerically to obtain tight guarantees [Gopi et al., 2021]

HOW TO DESIGN DP ALGORITHMS?



A KEY BUILDING BLOCK: THE GAUSSIAN MECHANISM

- Consider g taking as input a dataset and returning a p-dimensional real vector
- Denote its sensitivity by $\Delta = \max_{D \sim D'} \|g(D) - g(D')\|_2$

Theorem (Gaussian mechanism)

Let $\sigma > 0$ and $\mathcal{A}(\cdot) = g(\cdot) + \mathcal{N}(0, \sigma^2 \Delta^2)$. Then:

- *$\mathcal{A}$ satisfies $\frac{1}{\sigma}$-GDP.*
- *$\mathcal{A}$ satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for any $\alpha > 1$.*

Theorem (Subsampled Gaussian mechanism, informal)

If $\mathcal{A}$ is executed on a random fraction q of D, then it satisfies $(\alpha, \frac{q^2 \alpha}{2\sigma^2})$-RDP.

- DP induces a privacy-utility trade-off, here in terms of the variance of the estimate
- Random subsampling amplifies privacy guarantees

SUMMARY AND TAKE-AWAYS

- **Core principles of DP**
  - Privacy is a property of the algorithm, not of a particular output (unlike, e.g., k-anonymization)
  - Any useful private algorithm *must* be randomized
  - Privacy is a worst-case guarantee: it must hold for *all* neighboring datasets and *all* possible outputs
- **Privacy accounting matters**
  - f-DP is the most expressive framework, but often difficult to use in practice
  - GDP and RDP provide tractable alternatives with convenient composition properties
  - (ε, δ)-DP remains the standard reporting format despite well-known limitations [Gomez et al., 2025]
- **Selected milestones.** 2006: DP introduced by Dwork et al. 2017: Gödel Prize awarded. 2020: DP deployed in the U.S. Census. 2026: DP banned from US statistical products

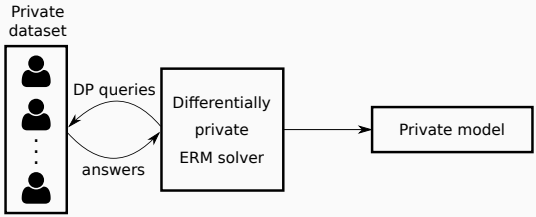
DP is now the gold-standard framework for rigorous privacy guarantees

Differentially Private ML with DP-SGD

- A trusted curator wants to privately release a model trained on data $D = \{x_1, \ldots, x_n\}$
- We focus here on approximately solving an Empirical Risk Minimization (ERM) problem under a DP constraint:

$$\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; D) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i) \right\}, \quad \text{where } \ell \text{ is differentiable in } \theta$$

- We can achieve this by designing a differentially private ERM solver



Algorithm Differentially Private SGD (DP-SGD) [Bassily et al., 2014, Abadi et al., 2016]

Initialize $\theta^{(0)} \in \mathbb{R}^p$ (must be independent of D)
for $t = 0, \dots, T-1$ **do**
  Pick $i_t \in \{1, \dots, n\}$ uniformly at random
  $\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)}\left(\nabla \ell(\theta^{(t)}; x_{i_t}) + \eta^{(t)}\right)$ where $\eta^{(t)} \sim \mathcal{N}(0, \sigma^2 \Delta^2 \mathbb{I}_p)$
end for
Return $\theta^{(T)}$

- We can control the sensitivity $\Delta = \sup_w \sup_{x,x'} \|\nabla \ell(\theta^{(t)}; x) - \nabla \ell(\theta^{(t)}; x')\|_2$ either by assuming $\ell(\cdot; x)$ Lipschitz for all x, or through gradient clipping [Abadi et al., 2016]
- Extensions to mini-batch SGD, projected SGD and regularization are straightforward
- How to analyze the privacy and utility of this algorithm?

Theorem (Privacy-utility trade-off of DP-SGD)

Let Θ be a convex domain of diameter bounded by R, and let the loss function ℓ be convex and l-Lipschitz over Θ. Let $\alpha > 1$, $\sigma^2 = \alpha T / 2n^2\varepsilon$ and $\Delta^2 = 2l^2$. Then DP-SGD satisfies (α, ε)-RDP. Furthermore, for $T = n^2$ and $\gamma_t = O(R/\sqrt{t})$, we have:

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq \tilde{O}\left(\frac{lR\sqrt{p}\alpha}{n\sqrt{\varepsilon}}\right)$$

If the objective $\mathcal{L}$ is also s-strongly convex, then for $T = n^2$ and $\gamma_t = 1/st$ we have:

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq \tilde{O}\left(\frac{l^2 p\alpha}{s\varepsilon n^2}\right)$$

- Privacy follows from composition of the subsampled Gaussian + post-processing
- Utility gap reduces with sample size but increases with number of model parameters

- For utility analysis, we rely on a general lemma giving convergence rates for SGD

Lemma ([Shamir and Zhang, 2013])

Let $\mathcal{L}$ be a convex function over a convex domain Θ with diameter bounded by R. Consider any SGD algorithm $\theta^{(t+1)} \leftarrow \Pi_\Theta(\theta^{(t)} - \gamma_t g_t)$ where g_t satisfies $\mathbb{E}[g_t] = \nabla\mathcal{L}(\theta^{(t)})$ and $\mathbb{E}[\|g_t\|^2] \leq G^2$. By setting $\gamma_t = \frac{R}{G\sqrt{t}}$, we have

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \min_{\theta\in\Theta} \mathcal{L}(\theta) \leq 2RG\Big(\frac{2 + \ln T}{\sqrt{T}}\Big).$$

If $\mathcal{L}$ is also s-strongly convex, then setting $\gamma_t = \frac{1}{st}$ gives

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \min_{\theta\in\Theta} \mathcal{L}(\theta) \leq \frac{17G^2(1 + \ln T)}{sT}.$$

Proof of the theorem.

- Denote by $g_t = \nabla \ell(\theta^{(t)}; x_{i_t}) + \eta^{(t)}$ the noisy gradient at step t
- Let us examine $\mathbb{E}[g_t]$ and $\mathbb{E}[\|g_t\|^2]$
- We have $\mathbb{E}[g_t] = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(\theta^{(t)}; x_i) + \mathbb{E}[\eta^{(t)}] = \nabla \mathcal{L}(\theta^{(t)}; D)$, hence g_t is an unbiased estimate of the gradient of the objective function at $\theta^{(t)}$
- Furthermore, since $\nabla \ell(\theta^{(t)}; x_{i_t})$ and $\eta^{(t)}$ are independent and ℓ is l-Lipschitz:

$$\begin{aligned}\mathbb{E}[\|g_t\|^2] &= \mathbb{E}[\|\nabla \ell(\theta^{(t)}; x_{i_t})\|^2] + \mathbb{E}[\|\eta^{(t)}\|^2] \\ &\leq l^2 + \frac{2pl^2\alpha T}{n^2\varepsilon}\end{aligned}$$

□

Proof of the theorem.

- It remains to plug our results in the previous lemma and to set T appropriately
- For the convex case, we get:

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq O\left(\frac{lR \ln T}{\sqrt{T}} + \frac{lR\sqrt{p\alpha T \ln(T)}}{n\sqrt{T\varepsilon}} \right)$$

- For the s-strongly case, we get:

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \min_{\theta \in \Theta} \mathcal{L}(\theta) \leq O\left(\frac{l^2 \ln T}{sT} + \frac{l^2 p\alpha T \ln(T)}{\varepsilon n^2 sT} \right)$$

- In both cases, choosing $T = n^2$ balances the two terms ("optimization error" and "privacy error") and gives the result

□

- Setting σ^2 and optimizing α to get the best (ϵ, δ)-DP guarantee, we get

Convex, Lipschitz, smooth	$\mathbb{E}[\mathcal{L}(\theta^{\text{priv}}) - \mathcal{L}^*] = \tilde{O}\left(\frac{\sqrt{p}\ln(1/\delta)}{n\epsilon}\right)$
Strongly convex, Lipschitz, smooth	$\mathbb{E}[\mathcal{L}(\theta^{\text{priv}}) - \mathcal{L}^*] = \tilde{O}\left(\frac{p\ln(1/\delta)}{n^2\epsilon^2}\right)$

- This is optimal [Bassily et al., 2014]: cannot do better without additional assumptions
- If the problem has a (near) sparse solution, a private version of greedy coordinate descent can achieve $O(\log p)$ utility [Mangold et al., 2023a]

- A common strategy is to pretrain on large-scale public data, then fine-tune on private data with DP-SGD, while restricting updates to a small set of effective parameters (e.g., the final classification layer or LoRA adapters)
- For end-to-end private training of large models, state-of-the-art methods rely on DP-SGD combined with practical heuristics such as very large batch sizes, data augmentation, and a small number of training epochs
  - A concrete example is Google's differentially private LLM, VaultGemma, announced in September 2025, which satisfies a sequence-level guarantee of $(2, 1.1^{-10})$-DP
- Many open questions remain on how to privately and efficiently exploit the intrinsic low-dimensional structure of deep models

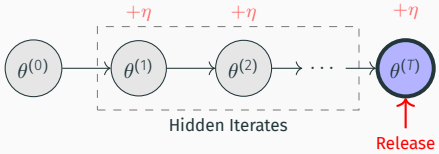
Selected Advanced Topics in Differentially Private ML

Selected Advanced Topics in Differentially Private ML

DP-SGD Beyond Composition

DP-SGD BEYOND COMPOSITION: PRIVACY AMPLIFICATION BY ITERATION

- **Standard view:** DP-SGD is "just" adaptive composition of subsampled Gaussian mechanisms. Every iterate $\theta^{(t)}$ may be released, and privacy loss grows with T
- **Hidden-state view:** only the final model θ_T is released



- **Privacy amplification by iteration** [Feldman et al., 2018, Altschuler and Talwar, 2022]:
  - Early gradients have diminishing influence as noise accumulates over time
  - Requires contractive dynamics (e.g, strong convexity, or convexity with bounded domain) so that the divergence between distributions shrinks over time
  - Over long runs, the privacy loss saturates to a finite bound rather than growing unboundedly
  - This phenomenon does not generally hold in nonconvex settings [Cebere et al., 2025]

- **Linear query view:** the full DP-SGD trajectory can be written as

$$\underbrace{\begin{bmatrix}\theta^{(1)}\\ \vdots \\ \theta^{(T)}\end{bmatrix}}_{\Theta} = \begin{bmatrix}\theta^{(0)}\\ \vdots \\ \theta^{(0)}\end{bmatrix} + \left(A \underbrace{\begin{bmatrix}g^{(1)}\\ \vdots \\ g^{(T)}\end{bmatrix}}_{G} + \underbrace{\begin{bmatrix}\eta^{(1)}\\ \vdots \\ \eta^{(T)}\end{bmatrix}}_{H} \right) \quad \text{with } A \text{ lower triangular}$$

- **Matrix factorization:** instead of injecting independent noise to G, factorize $A = BC$ and add correlated noise [Kairouz et al., 2021b, Denisov et al., 2022, Kalinin et al., 2026]:

$$B(CG + H) = AG + BH = A(G + C^{\dagger}H)$$

  - Given a given time horizon T, the optimal factorization is typically obtained by minimizing $\text{sens}(C)^2 \|B\|^2$, over a restricted class of structured matrices for efficiency
  - This approach exploits the structure of DP-SGD (prefix-sum computation) to reduce noise accumulation over time
  - It achieves state-of-the-art performance in private training of large-scale models

Selected Advanced Topics in Differentially Private ML

DP for Federated and Decentralized Learning

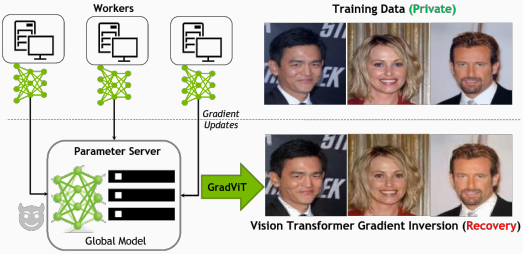
Federated Learning (FL) [Kairouz et al., 2021c] aims to collaboratively train ML models while keeping their data decentralized

Minimize the global objective aggregated from each party's local dataset D_k:

$$\mathcal{L}(\theta; D) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta; D_k)$$

FL does not provide robust privacy guarantees!

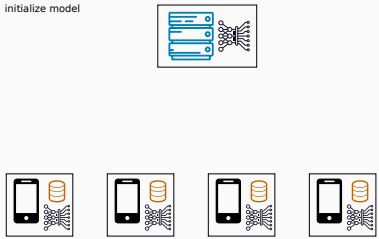
- offers an additional attack surface: server observes participants' updates
- allows specific privacy attacks [Geiping et al., 2020, El Mrini et al., 2024]



Federated Learning (FL) [Kairouz et al., 2021c] aims to collaboratively train ML models while keeping their data decentralized

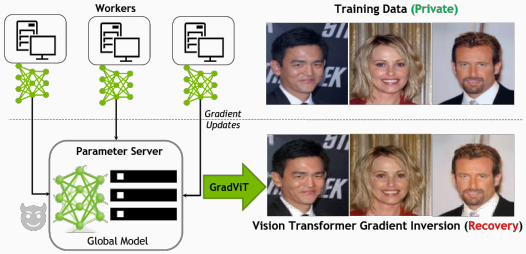
Minimize the global objective aggregated from each party's local dataset D_k:

$$\mathcal{L}(\theta; D) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta; D_k)$$



FL does not provide robust privacy guarantees!

- offers an additional attack surface: server observes participants' updates
- allows specific privacy attacks [Geiping et al., 2020, El Mrini et al., 2024]

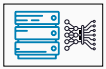


Federated Learning (FL) [Kairouz et al., 2021c] aims to collaboratively train ML models while keeping their data decentralized

Minimize the global objective aggregated from each party's local dataset D_k:

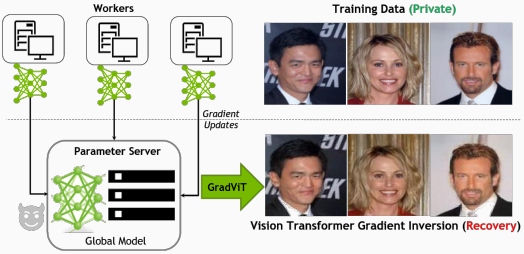
$$\mathcal{L}(\theta; D) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta; D_k)$$

each party makes an update using its local dataset



FL does not provide robust privacy guarantees!

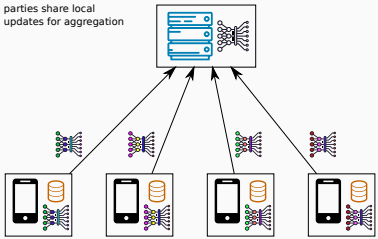
- offers an additional attack surface: server observes participants' updates
- allows specific privacy attacks [Geiping et al., 2020, El Mrini et al., 2024]



Federated Learning (FL) [Kairouz et al., 2021c] aims to collaboratively train ML models while keeping their data decentralized

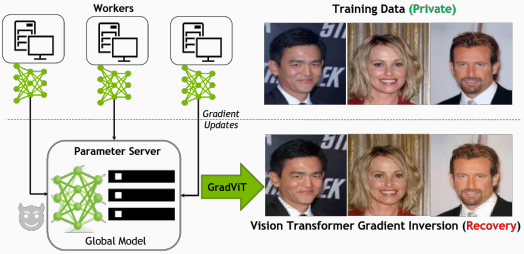
Minimize the global objective aggregated from each party's local dataset D_k:

$$\mathcal{L}(\theta; D) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta; D_k)$$



FL does not provide robust privacy guarantees!

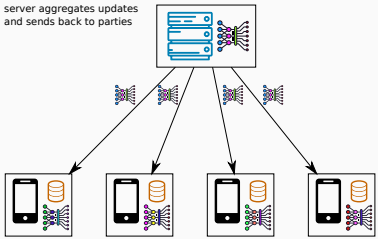
- offers an additional attack surface: server observes participants' updates
- allows specific privacy attacks [Geiping et al., 2020, El Mrini et al., 2024]



Federated Learning (FL) [Kairouz et al., 2021c] aims to collaboratively train ML models while keeping their data decentralized

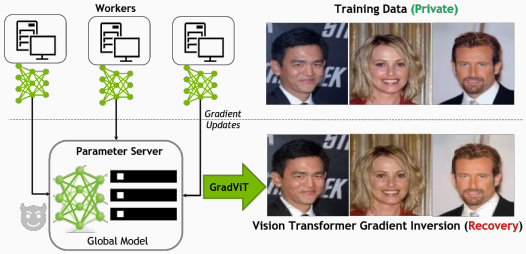
Minimize the global objective aggregated from each party's local dataset D_k:

$$\mathcal{L}(\theta; D) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta; D_k)$$



FL does not provide robust privacy guarantees!

- offers an additional attack surface: server observes participants' updates
- allows specific privacy attacks [Geiping et al., 2020, El Mrini et al., 2024]

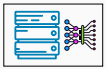


Federated Learning (FL) [Kairouz et al., 2021c] aims to collaboratively train ML models while keeping their data decentralized

Minimize the global objective aggregated from each party's local dataset D_k:

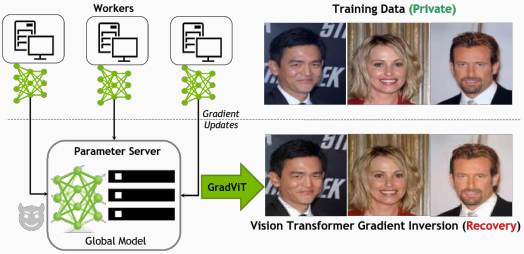
$$\mathcal{L}(\theta; D) = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta; D_k)$$

parties update their copy of the model and iterate



FL does not provide robust privacy guarantees!

- offers an additional attack surface: server observes participants' updates
- allows specific privacy attacks [Geiping et al., 2020, El Mrini et al., 2024]



- Two extreme trust models:
  - **Central DP:** Requires a trusted server to aggregate clean updates and add noise. High utility, but strong trust assumption.
  - **Local DP:** Assumes an **untrusted server**. Each party adds noise before sending updates. No server trust, but significantly reduced utility (error scales as $\Omega(\sqrt{K})$).
- Intermediate trust models (via cryptography):
  - **Secure computation protocols** such as secure aggregation: Server only sees the sum of updates, not individual contributions. Improves privacy but increases computation and communication cost [Bonawitz et al., 2017, Kairouz et al., 2021a]
  - **Secure communication channels:** parties coordinate noise (e.g., correlated noise) to mask individual updates while preserving utility [Sabater et al., 2022, Allouah et al., 2024]

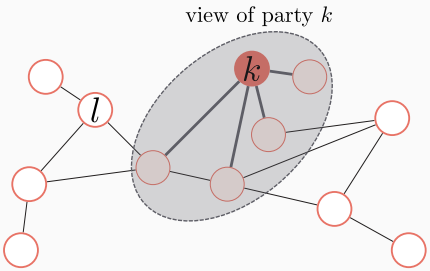
- **Fully decentralized learning:** central server replaced by peer-to-peer communication over a graph $G = (\mathcal{V}, \mathcal{E})$; global aggregation replaced by local interactions (e.g., gossip)
- We can leverage the limited view of parties to protect privacy

Definition (Network DP [Cyffers and Bellet, 2022])

An algorithm $\mathcal{A}$ satisfies (α, ε)-Network DP (NDP) if for all pairs of parties $u, v \in \mathcal{V}$ and pairs of datasets D, D' that differ only in the local dataset of party u:

$$D_\alpha\left(\mathcal{O}_v(\mathcal{A}(D)) \,||\, \mathcal{O}_v(\mathcal{A}(D'))\right) \leq \varepsilon\,,$$

with $\mathcal{O}_v$ the set of messages sent and received by v.



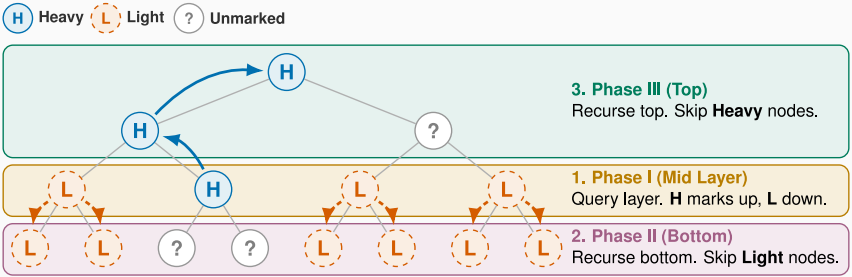
- Privacy amplification in decentralized DP-SGD algorithms that can match central DP trade-offs [Cyffers et al., 2022, Cyffers et al., 2024, Bellet et al., 2026] (**PhD of Edwige Cyffers**)

Selected Advanced Topics in Differentially Private ML

DP for ML Beyond Gradient Descent

- DP-SGD has become the dominant paradigm for private ML, but many important applications are not deep learning
- Sensitive domains (healthcare, finance, public administration) often involve tabular data, where tree-based methods like Random Forests (RFs) remain highly competitive
- Unsurprisingly, RFs can leak information about individual training records, enabling reconstruction attacks [Ferry et al., 2024]
- Existing approaches to DP RFs typically suffer from poor utility (and privacy flaws…)
  - Greedy tree construction: informative splits require spending privacy spent at every node
  - Fully randomized tree construction: avoids split selection but utility degrades rapidly with tree depth

LUMBERBACK: DP RANDOM FORESTS VIA HEAVY HITTER DETECTION IN TREES



- Uses DP heavy-hitter detection to determine where tree growth should stop, pruning empty or low-density branches
- Exploits the hierarchical structure of trees through binary-search style procedures, reducing the dependence on tree height from $O(\sqrt{h})$ to $O(\sqrt{\log h})$
- Enables substantially deeper trees while preserving privacy, establishing a new state-of-the-art for DP RFs [Lebeda et al., 2026] (**Postdoc of Christian Lebeda**)

Training ML models with DP guarantees

- DP-SGD is the dominant paradigm
  - Its privacy analysis reduces to the composition of subsampled Gaussian mechanisms, although tighter analyses are sometimes possible
  - It naturally extends to federated & decentralized learning under a variety of trust models
- DP-ML is broader than DP-SGD
  - Models not trained by gradient descent require specialized privacy-preserving algorithms
  - Exploiting model structure can yield significantly better privacy-utility trade-offs

Coming up: Privacy auditing

- Membership inference attacks as a tool to measure empirical privacy leakage
- Testing differential privacy guarantees
- Scalable and reliable auditing protocols for large models

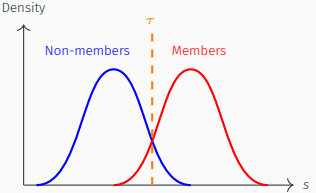
Privacy Auditing

- Most work in DP consists in designing algorithms with provable upper bounds on the privacy leakage (e.g., "$\mathcal{A}$ satisfies μ-GDP") and associated utility guarantees
- **Q1:** How tight are these upper bounds both in theory and in practice?
- **Q2:** Do implementations actually deliver the theoretical guarantees?
- **Q3:** Is privacy enforced as promised?

We can do this with Membership Inference Attacks (MIAs)!

Membership Inference Attack (MIA):
Given a record x and a model θ, determine if $x \in \mathcal{D}_{\text{train}}$ (member vs. non-member)

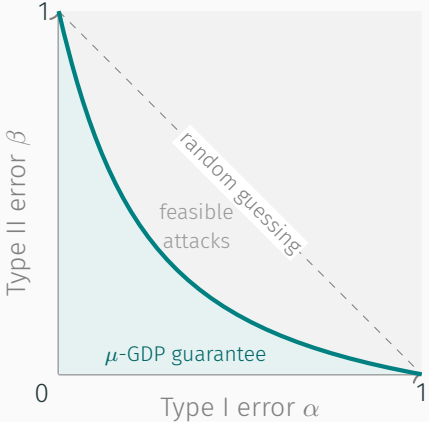
- The attacker computes a membership score $s(x, \theta)$ and predicts membership whenever the score exceeds a threshold $\tau(x)$
  - Negative loss: $s(x, \theta) = -\ell(x, \theta)$
  - Perplexity (LLMs): $s(x, \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log(p_\theta(x_t | x_{1:t-1}))$
  - Gradient norm (white-box): $s(x, \theta) = -\|\nabla \ell(x, \theta)\|$



- Designing privacy attacks is an active area of research, but few practical tools are available, despite the need for empirical privacy assessment in real-world scenarios
- PANAME (with CNIL, ANSSI, PEReN): development of an open-source software library to provide a unified and efficient framework for conducting privacy evaluation tests

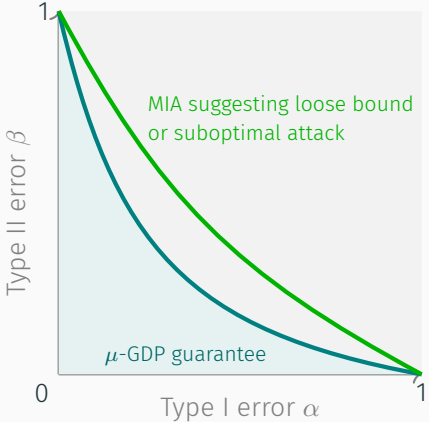
COMPARING DP GUARANTEES TO MIA PERFORMANCE

- MIAs quantify empirical privacy leakage (even for algorithms lacking DP guarantees)
- They also yield empirical lower bounds on privacy parameters, enabling us to assess the tightness of DP guarantees and detect false privacy claims



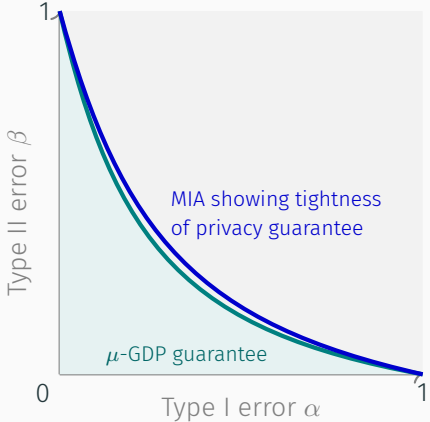
COMPARING DP GUARANTEES TO MIA PERFORMANCE

- MIAs quantify empirical privacy leakage (even for algorithms lacking DP guarantees)
- They also yield empirical lower bounds on privacy parameters, enabling us to assess the tightness of DP guarantees and detect false privacy claims



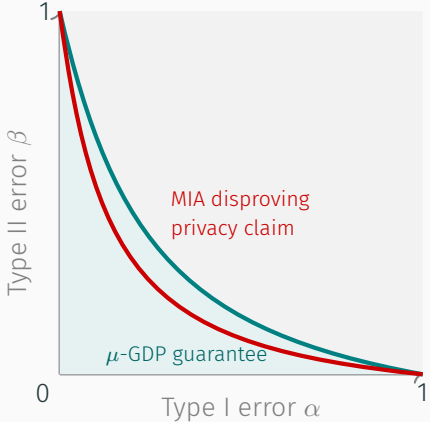
COMPARING DP GUARANTEES TO MIA PERFORMANCE

- MIAs quantify empirical privacy leakage (even for algorithms lacking DP guarantees)
- They also yield empirical lower bounds on privacy parameters, enabling us to assess the tightness of DP guarantees and detect false privacy claims



COMPARING DP GUARANTEES TO MIA PERFORMANCE

- MIAs quantify empirical privacy leakage (even for algorithms lacking DP guarantees)
- They also yield empirical lower bounds on privacy parameters, enabling us to assess the tightness of DP guarantees and detect false privacy claims



STANDARD (MULTI-RUN) PRIVACY AUDITING

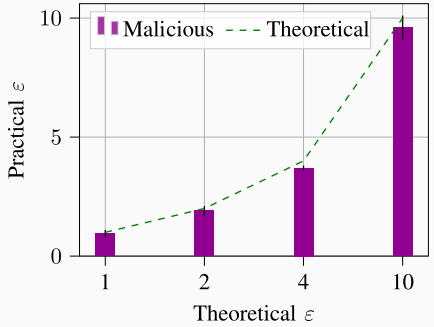
Algorithm Multi-run privacy auditing

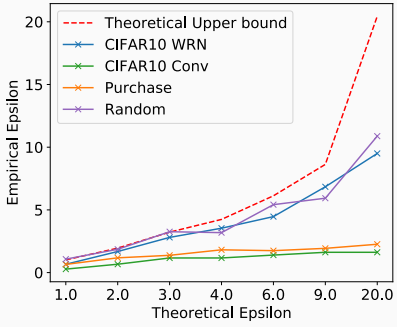
Input: Audited algorithm $\mathcal{A}$, adversary **Adv**, dataset D, canary point x^*, number of auditing runs k

$D_0 \leftarrow D$, $D_1 \leftarrow D \cup \{x^*\}$
for $i = 1$ **to** k **do**
  $b_i \leftarrow \text{Ber}(1/2)$ {draw a random bit}
  $S_i \leftarrow \textbf{Adv.Score}(\mathcal{A}(D_{b_i}), x^*)$
end for
$FN, FP \leftarrow \textbf{Adv.DecisionRule}(S, b)$
$\overline{FN}, \overline{FP} \leftarrow \textbf{ConfidenceInterval}(FN, FP)$
$\overline{\mu} \leftarrow \textbf{ConvertToDP}(\overline{FN}, \overline{FP})$
return high-probability lower bound $\overline{\mu}$

- Auditing μ-GDP requires fewer runs k than (ϵ, δ)-DP [Nasr et al., 2023b]
- To assess "empirical" privacy in realistic settings: instantiate **Adv** to known MIA attacks and D, x^* to real data
- To assess tightness of DP guarantees: try to find worst-case **Adv**, D, x^* allowed by the threat model [Yaghini et al., 2026a]

- Standard privacy analysis of DP-SGD uses a composition of T mechanisms $\rightarrow$ the threat model allows to release all intermediate models $\theta^{(1)}, \ldots, \theta^{(T)}$
- Privacy auditing has shown that in this threat model, privacy upper bounds are tight [Nasr et al., 2021], even for real datasets and models [Nasr et al., 2023b]





- Can study other threat models, e.g., the hidden-state [Cebere et al., 2025]

Algorithm One-run privacy auditing

Input: Audited algorithm $\mathcal{A}$, adversary **Adv**, dataset D, canary points $x_1, \ldots, x_m$

for $i = 1$ **to** m **do**

  $b_i \leftarrow \mathrm{Ber}(1/2)$ {draw a random bit}

end for

$\theta \leftarrow \mathcal{A}(D \cup \{x_i : b_1 = 1\})$

$S_i \leftarrow \mathbf{Adv.Score}(\theta, x_i)$

$FN, FP \leftarrow \mathbf{Adv.DecisionRule}(S, b)$

$\overline{FN}, \overline{FP} \leftarrow \mathtt{ConfidenceInterval}(FN, FP)$

$\overline{\mu} \leftarrow \mathtt{ConvertToDP}(\overline{FN}, \overline{FP})$

return high-probability lower bound $\overline{\mu}$

- **One-run auditing** uses a single execution of the algorithm after randomly inserting multiple canaries [Steinke et al., 2023, Mahloujifar et al., 2025, Dagréou and Bellet, 2026]
- This approach reduces computational cost, but may lead to weaker privacy estimates due to interference
- It still requires interventional access, requiring trust in the model provider and preventing third-party auditing

Privacy Auditing

Zero-Run Privacy Auditing

[Cebere et al., 2026b] (PhD of Tudor Cebere)

MOTIVATING SCENARIO: AUDITING AN LLM

- Suppose an auditor wants to evaluate the privacy of an LLM released in 2023
- Common practice: use data with a temporal cutoff [Shi et al., 2023, Meeus et al., 2024a]
  - **Member data**: Wikipedia records from *pre-2023* (training set)
  - **Non-Member data**: Wikipedia records from *post-2023* (held-out)
- The auditor finds significantly lower perplexity on the pre-2023 set compared to the post-2023 set and concludes high privacy leakage
- **Problem:** bias from temporal distribution shift [Meeus et al., 2024b, Duan et al., 2024]
  - Wording, news topics, and formatting styles change over time
  - The model is "surprised" by 2024 data not because it wasn't part of the training data, but because the distribution has shifted
  - **Consequence**: the audit overestimates the empirical privacy risk
- This issue arises whenever training data is not adequately documented and/or a randomized test split is unavailable

Interventional Auditing

- Equivalent to Randomized Controlled Trials (RCTs)
- Treatment (membership) is artificially randomized
- **Advantage**: Avoids confounding by design

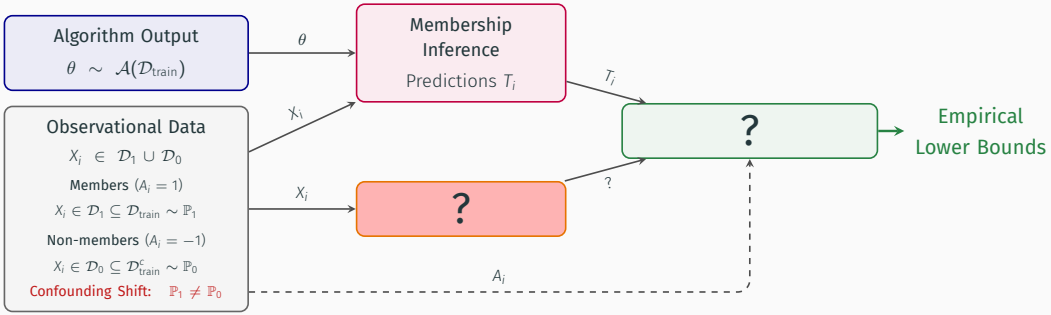
Zero-Run Auditing [Cebere et al., 2026b]

- Equivalent to observational studies
- Treatment is observed post-hoc from fixed datasets
- **Challenge**: Requires adjusting for confounding differences

- In [Even et al., 2026], we propose a causal framework for MIA evaluation, allowing to debias *population-level* performance metrics
- Here, we use the observational lens to establish lower bounds for privacy auditing

- Auditor passively observes a deployed model $\theta = \mathcal{A}(\mathcal{D}_{\text{train}})$.
- Auditor has access to two fixed datasets:
  - $\mathcal{D}_1 \subset \mathcal{D}_{\text{train}}$: known members ($A_i = 1$), $D_1 \sim \mathbb{P}_1$.
  - $\mathcal{D}_0 \subset \mathcal{X} \setminus \mathcal{D}_{\text{train}}$: known non-members ($A_i = -1$), $D_0 \sim \mathbb{P}_0$.
- If $\mathbb{P}_1 = \mathbb{P}_0$, the auditor can apply one-run auditing
- Otherwise, such an audit is statistically invalid, as the features X_i carry information about the membership bits A_i

ZERO-RUN PRIVACY AUDITING: OVERVIEW



Core question: How much do the features of a data point *alone* reveal about its own membership?

Definition (Propensity Score)

The propensity score function $\pi : \mathcal{X} \to [0, 1]$ is the conditional probability of membership given the features X_i:

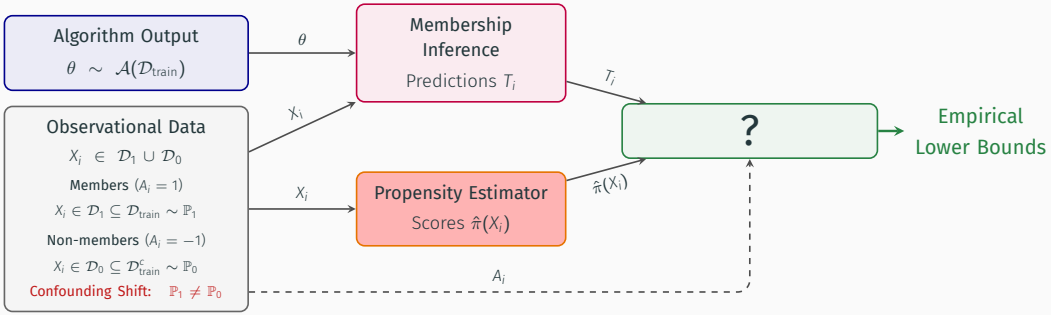
$$\pi(x) = \mathbb{P}(A_i = 1 \mid X_i = x)$$

- **Key intuition:** If a point's features easily give away its membership (extreme $\pi(x)$), then a correct MIA prediction provides less evidence of privacy leakage and should be discounted accordingly

ESTIMATING $\pi(x)$ IN PRACTICE

- Since the true $\pi(x)$ is unknown in practice, we use an empirical estimator $\hat{\pi}(x)$
  - This reduces to a binary classification problem: predicting $A_i \in \{-1, 1\}$ from X_i
- To ensure the validity of the audit, it is sufficient to obtain one-sided estimation guarantees
  - Overestimating domain overlap (i.e., making $\hat{\pi}(x)$ artificially closer to 0.5) can invalidate the audit
  - Underestimating overlap (i.e., making $\hat{\pi}(x)$ closer to 0 or 1) is safe, as it only yields more conservative privacy lower bounds
- In practice, we quantify the estimation uncertainty using bootstrap resampling

ZERO-RUN PRIVACY AUDITING: OVERVIEW



- For each evaluation point, we condition the audit on the observed features X_i
- The membership information revealed solely by observing X_i is quantified by:

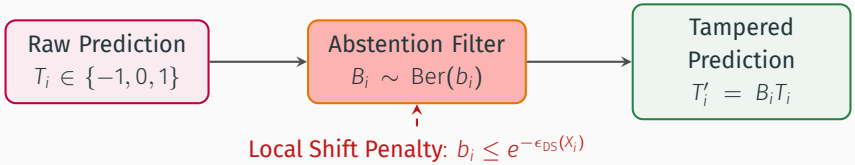
$$\epsilon_{\text{DS}}(X_i) = \left| \log \frac{\pi(X_i)}{1 - \pi(X_i)} \right|$$

- We isolate this pointwise leakage due to distribution shift and adjust the MIA prediction for each sample accordingly
  - Extreme values of $\pi(X_i)$ make membership highly predictable $\rightarrow$ such points receive less weight in the audit

- Post-process raw MIA predictions T_i into tampered predictions $T'_i = B_i T_i$
- We let $B_i \sim \mathrm{Ber}(b_i)$ with

$$b_i \leq \exp(-\epsilon_{\mathrm{DS}}(X_i))$$

- **Consequence**: The stronger the local distribution shift on X_i, the higher the chance the prediction is discarded ($T'_i = 0$).



- Let m be the total number of evaluation data points
- Let r be the number of active predictions made by the attacker ($\#\{i : T_i \neq 0\}$)
- Let p_k be the probability that the attacker makes k correct **tampered** predictions

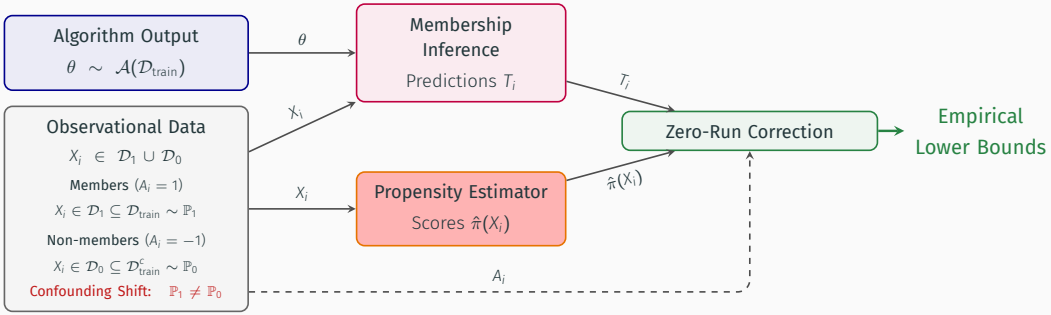
Theorem (Zero-run auditing [Cebere et al., 2026b])

If $\mathcal{A}$ satisfies f-DP, then the distribution of tampered predictions satisfies, for every subset $T \subseteq [r]$,

$$\underbrace{\sum_{k \in T} \frac{k}{m} p_k}_{\textit{Tampered TPR proxy}} \leq \bar{f}\Bigg(\underbrace{\sum_{k \in T} \frac{r-k+1}{m} p_{k-1}}_{\textit{Tampered FPR proxy}} \Bigg)$$

- In the absence of distribution shift, the result recovers the one-run auditing results [Mahloujifar et al., 2025]

ZERO-RUN PRIVACY AUDITING: OVERVIEW



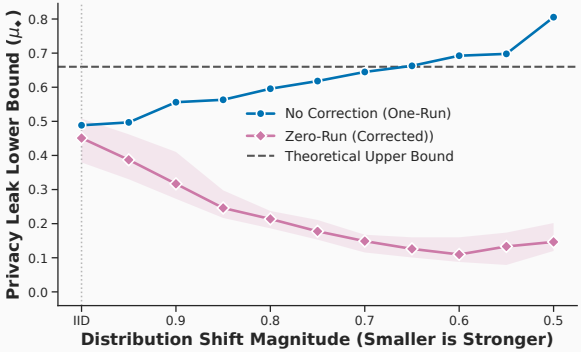
SYNTHETIC RESULTS: NOISY SUM

Experimental setup:

- High-dim. noisy sum
- Tunable distribution shift
- Theoretical $\mu_{\text{true}} = 0.66$

Key findings:

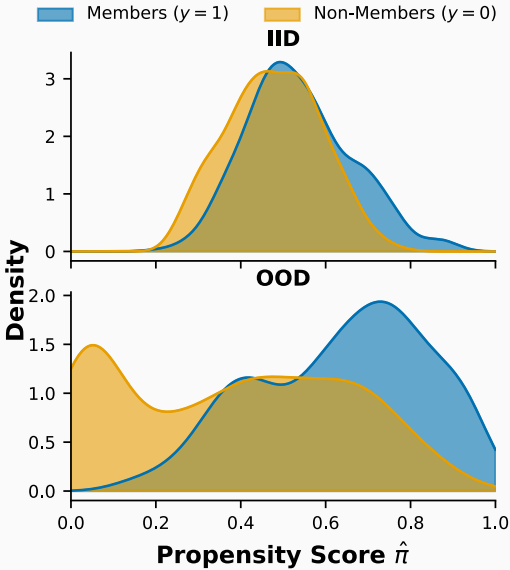
- One-run auditing becomes invalid as shift rises
- Zero-run extracts valid bounds despite shift



Experimental setup: Natural geographic and environmental distribution shift in animal images

Key findings:

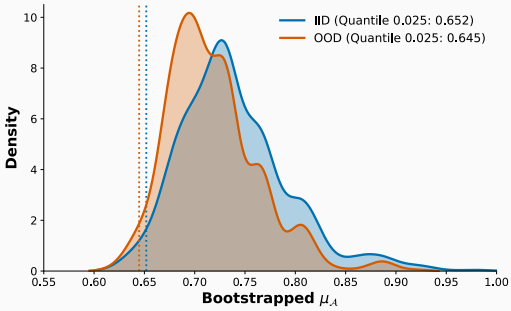
- Real world propensity scores are challenging
- Zero-run auditing extracts a valid bound ($\hat{\mu} = 0.645$) on Out-Of-Distribution (OOD) data
- Closely matches the purely IID split bound ($\hat{\mu} = 0.652$)



Experimental setup: Natural geographic and environmental distribution shift in animal images

Key findings:

- Real world propensity scores are challenging
- Zero-run auditing extracts a valid bound ($\hat{\mu} = 0.645$) on Out-Of-Distribution (OOD) data
- Closely matches the purely IID split bound ($\hat{\mu} = 0.652$)



Privacy Auditing

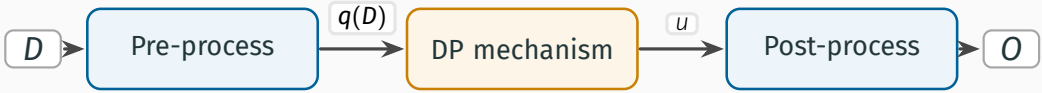
Grey-Box Auditing of DP Libraries

[Cebere et al., 2026a] (PhD of Tudor Cebere)

GOAL: CATCHING BUGS IN DP IMPLEMENTATIONS

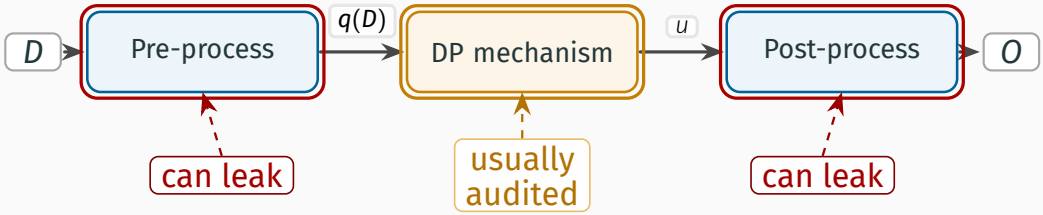
- DP's theoretical elegance contrasts with the practical challenge of correct implementation, where subtle bugs can invalidate privacy guarantees
  - incorrect sensitivity calculations
  - data-dependent pre- or post-processing
  - mishandled privacy budget composition
  - insecure randomness or floating-point vulnerabilities
- Distributional auditing has two key limitations: it is intractable for complex pipelines (costly + hard to craft strong attacks), and it cannot pinpoint the source of the bug

1. The implementation is a chain of data-independent/dependent steps.

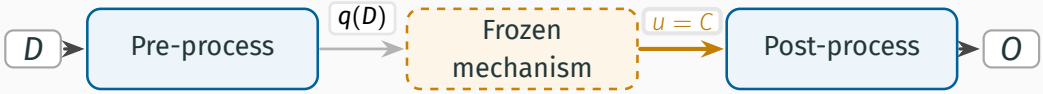


2. Query q accesses the data and is consumed by the DP mechanism.

3. Auditing *only* the privacy mechanism misses surrounding bugs

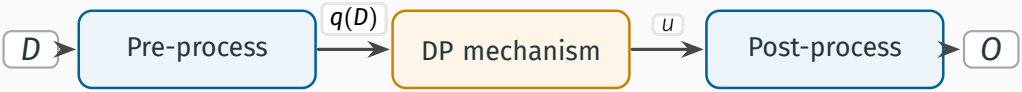


4. If we replace the mechanism with data-independent function

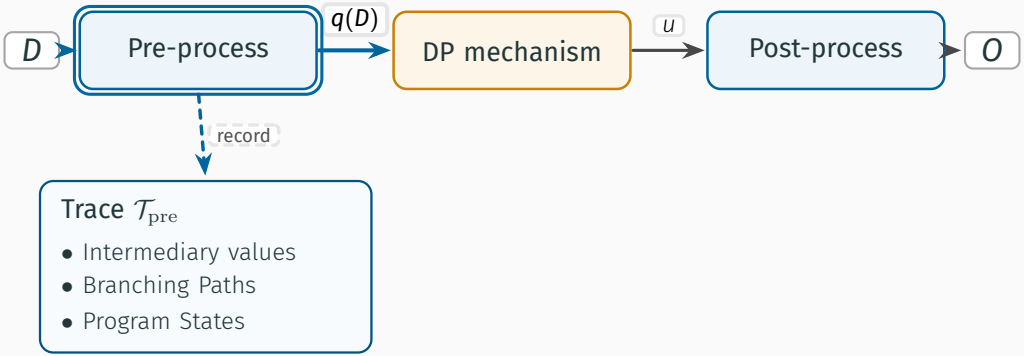


O should be independent of D.

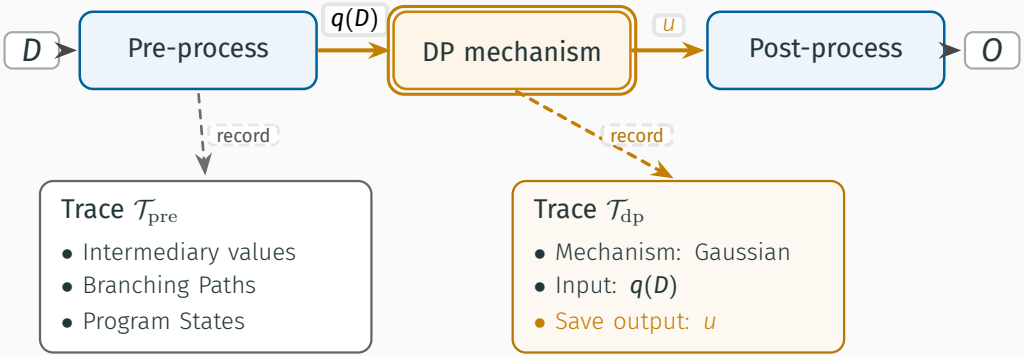
1. Run on D: Recording phase



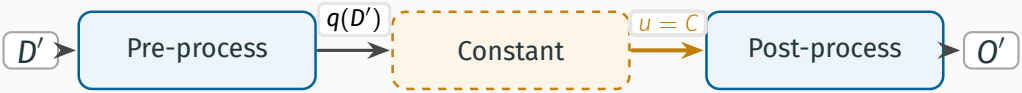
2. Record invariants: values that should not change under D/D'



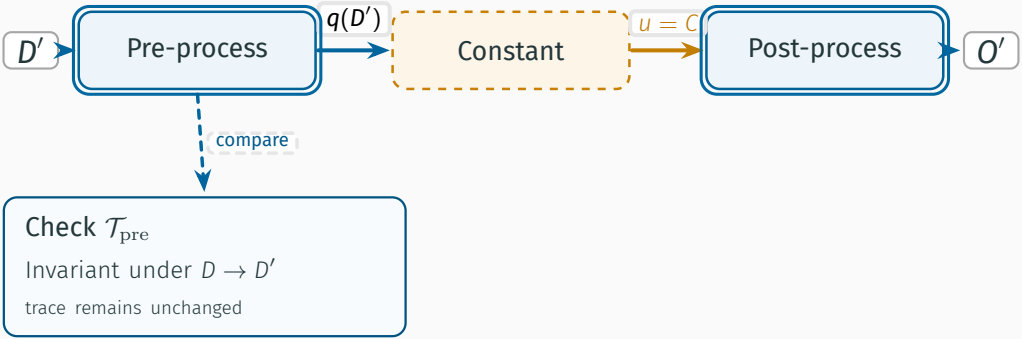
3. Record mechanism call: save the output u in the trace



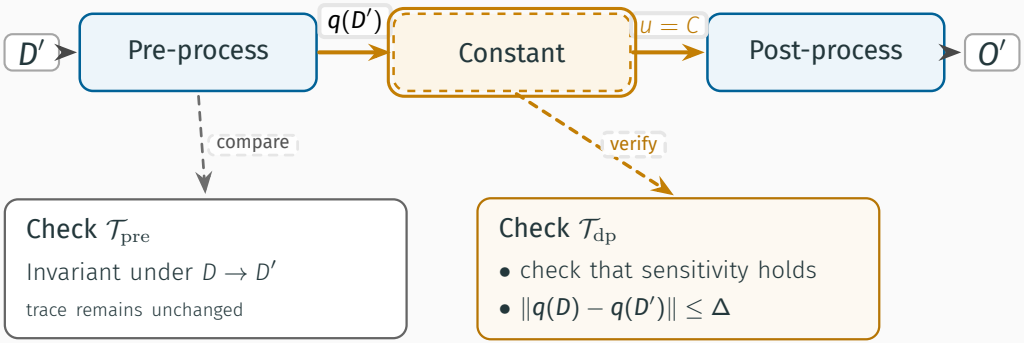
1. Re-run the pipeline on a neighboring dataset D'



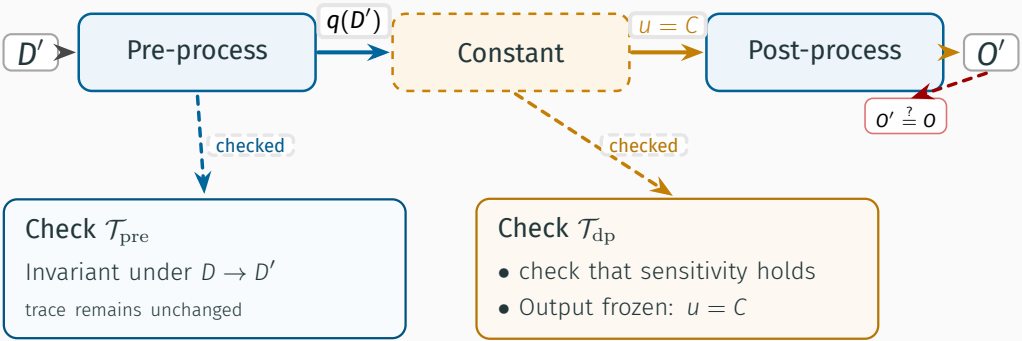
2. Check invariance: the trace does not change under $D \to D'$



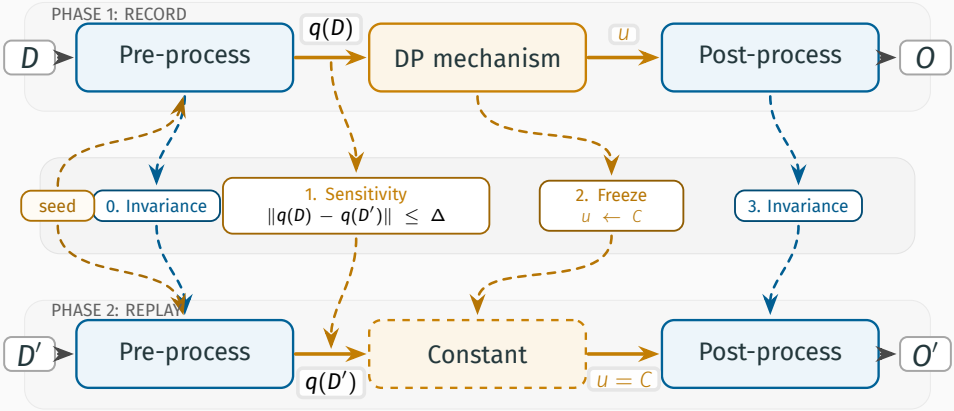
3. The mechanism input must satisfy sensitivity bound.



4. Fixing the mechanism localizes mismatches to the current code.



OVERVIEW: STRUCTURAL DECOMPOSITION



FINDINGS: 13 PRIVACY BUGS IN POPULAR LIBRARIES

We audited **12 libraries** and found **13 privacy bugs**

Library audited	Discovered vulnerability	Component
SmartNoise SDK	Sensitivity miscalibration	Covariance
SmartNoise SQL	Privacy-accountant bug	Odometer
Synthcity	Data-dependent control flow	PrivBayes
Opacus	Dataset-size leakage	DP-SGD
Diffprivlib	Sensitivity-logic error	Linear regression
MostlyAI	Budget-accounting bug	Numeric encoder
Private-PGM	Sensitivity miscalibration	JAM
dpmm	Data-domain leakage	AIM & PrivBayes

Conclusion

TAKE-HOME MESSAGES

1. ML models can leak personal data
2. Differential Privacy (DP) gives a rigorous framework for designing privacy-preserving ML algorithms like DP-SGD
3. DP creates a privacy-utility trade-off that depends on the trust model
4. Attacks can empirically measure privacy leakage and help audit DP guarantees

- Privacy in large models (e.g., LLMs): improve privacy-utility trade-offs, audit commercial models, protect privacy at inference time [Duan et al., 2023]
- Better understanding of privacy of synthetic data: **PhD of Clément Pierquin** [Pierquin et al., 2025, Pierquin et al., 2026]
- Characterizing how DP protects against attacks beyond MIAs [Guerra-Balboa et al., 2026] [Kulynych et al., 2026, Swanberg et al., 2026]
- Reconciling differential privacy with fairness [Mangold et al., 2023b, Yaghini et al., 2026b] and Byzantine robustness (**Nirupam's presentation**, **PhD of Thomas Boudou**)

REFERENCES I

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016).
Deep learning with differential privacy.
In *CCS*.

[Allouah et al., 2024] Allouah, Y., Koloskova, A., Firdoussi, A. E., Jaggi, M., and Guerraoui, R. (2024).
The privacy power of correlated noise in decentralized learning.
In *ICML*.

[Altschuler and Talwar, 2022] Altschuler, J. M. and Talwar, K. (2022).
Privacy of noisy stochastic gradient descent: More iterations without more privacy loss.
In *NeurIPS*.

[Bassily et al., 2014] Bassily, R., Smith, A. D., and Thakurta, A. (2014).
Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds.
In *FOCS*.

[Bellet et al., 2026] Bellet, A., Cyffers, E., Frey, D., Gaudel, R., Lerévérend, D., and Taïani, F. (2026).
Unified Privacy Guarantees for Decentralized Learning via Matrix Factorization.
In *ICLR*.

REFERENCES II

[Bonawitz et al., 2017] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. (2017).
Practical Secure Aggregation for Privacy-Preserving Machine Learning.
In *CCS*.

[Cebere et al., 2025] Cebere, T., Bellet, A., and Papernot, N. (2025).
Tighter Privacy Auditing of DP-SGD in the Hidden State Threat Model.
In *ICLR*.

[Cebere et al., 2026a] Cebere, T., Erb, D., Desfontaines, D., Bellet, A., and Fitzsimons, J. (2026a).
Privacy in Theory, Bugs in Practice: Grey-Box Auditing of Differential Privacy Libraries.
In *PETS*.

[Cebere et al., 2026b] Cebere, T., Even, M., Bleistein, L., and Bellet, A. (2026b).
Privacy auditing with zero (0) training run.

[Cyffers and Bellet, 2022] Cyffers, E. and Bellet, A. (2022).
Privacy Amplification by Decentralization.
In *AISTATS*.

[Cyffers et al., 2024] Cyffers, E., Bellet, A., and Upadhyay, J. (2024).
Differentially Private Decentralized Learning with Random Walks.
In *ICML*.

REFERENCES III

[Cyffers et al., 2022] Cyffers, E., Even, M., Bellet, A., and Massoulié, L. (2022).
Muffliato: Peer-to-Peer Privacy Amplification for Decentralized Optimization and Averaging.
In *NeurIPS*.

[Dagréou and Bellet, 2026] Dagréou, M. and Bellet, A. (2026).
Detectability in diversity: Improved canary crafting for privacy auditing in one run.

[Denisov et al., 2022] Denisov, S., McMahan, H. B., Rush, J., Smith, A. D., and Thakurta, A. G. (2022).
Improved differential privacy for SGD via optimal private linear operators on adaptive streams.
In *NeurIPS*.

[Dong et al., 2022] Dong, J., Roth, A., and Su, W. J. (2022).
Gaussian differential privacy.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 84(1):3–37.

[Duan et al., 2023] Duan, H., Dziedzic, A., Papernot, N., and Boenisch, F. (2023).
Flocks of stochastic parrots: Differentially private prompt learning for large language models.
In *NeurIPS*.

[Duan et al., 2024] Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. (2024).
Do membership inference attacks work on large language models?
In *Conference on Language Modeling (COLM)*.

REFERENCES IV

[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)*.

[El Mrini et al., 2024] El Mrini, A., Cyffers, E., and Bellet, A. (2024).
Privacy Attacks in Decentralized Learning.
In *ICML*.

[Even et al., 2026] Even, M., Berenfeld, C., Bleistein, L., Cebere, T., Josse, J., and Bellet, A. (2026).
Causal evaluation of membership inference attacks.

[Feldman et al., 2018] Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. (2018).
Privacy Amplification by Iteration.
In *FOCS*.

[Ferry et al., 2024] Ferry, J., Fukasawa, R., Pascal, T., and Vidal, T. (2024).
Trained random forests completely reveal your dataset.
In *ICML*.

[Geiping et al., 2020] Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020).
Inverting gradients - how easy is it to break privacy in federated learning?
In *NeurIPS*.

REFERENCES V

[Gomez et al., 2025] Gomez, J. F., Kulynych, B., Kaissis, G., Hayes, J., Balle, B., and Honkela, A. (2025).
Gaussian dp for reporting differential privacy guarantees in machine learning.

[Gopi et al., 2021] Gopi, S., Lee, Y. T., and Wutschitz, L. (2021).
Numerical composition of differential privacy.
In *NeurIPS*.

[Guerra-Balboa et al., 2026] Guerra-Balboa, P., Sauer, A., Arcolezi, H. H., and Strufe, T. (2026).
Understanding disclosure risk in differential privacy with applications to noise calibration and auditing.
In *VLDB*.

[Kairouz et al., 2021a] Kairouz, P., Liu, Z., and Steinke, T. (2021a).
The distributed discrete gaussian mechanism for federated learning with secure aggregation.
In *ICML*.

[Kairouz et al., 2021b] Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. (2021b).
Practical and private (deep) learning without sampling or shuffling.
In *ICML*.

[Kairouz et al., 2021c] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021c).
Advances and Open Problems in Federated Learning.
Foundations and Trends® in Machine Learning, 14(1–2):1–210.

[Kalinin et al., 2026] Kalinin, N. P., McKenna, R., Upadhyay, J., and Lampert, C. H. (2026).
Back to square roots: An optimal bound on the matrix factorization error for multi-epoch differentially private sgd.
In *ICLR*.

[Kulynych et al., 2026] Kulynych, B., Gomez, J. F., Kaissis, G., Hayes, J., Balle, B., Calmon, F. P., and Raisaro, J. L. (2026).
Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy.
In *NeurIPS*.

[Lebeda et al., 2026] Lebeda, C. J., Erb, D., Cebere, T., and Bellet, A. (2026).
Lumberjack: Better differentially private random forests through heavy hitter detection in trees.
CoRR, abs/2605.22756.

[Mahloujifar et al., 2025] Mahloujifar, S., Melis, L., and Chaudhuri, K. (2025).
Auditing f-differential privacy in one run.
In *ICML*.

[Mangold et al., 2023a] Mangold, P., Bellet, A., Salmon, J., and Tommasi, M. (2023a).
High-Dimensional Private Empirical Risk Minimization by Greedy Coordinate Descent.
In *AISTATS*.

[Mangold et al., 2023b] Mangold, P., Perrot, M., Bellet, A., and Tommasi, M. (2023b).
Differential Privacy has Bounded Impact on Fairness in Classification.
In *ICML*.

[Meeus et al., 2024a] Meeus, M., Jain, S., Rei, M., and de Montjoye, Y.-A. (2024a).
Did the neurons read your book? document-level membership inference for large language models.
In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385.

[Meeus et al., 2024b] Meeus, M., Shilov, I., Jain, S., Faysse, M., Rei, M., and de Montjoye, Y.-A. (2024b).
Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it).
arXiv preprint arXiv:2406.17975.

[Mironov, 2017] Mironov, I. (2017).
Renyi differential privacy.
In *CSF*.

[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)*.

[Pierquin et al., 2025] Pierquin, C., Bellet, A., Tommasi, M., and Boussard, M. (2025).
Privacy Amplification Through Synthetic Data: Insights from Linear Regression.
In *ICML*.

[Pierquin et al., 2026] Pierquin, C., Bellet, A., Tommasi, M., and Boussard, M. (2026).
Privacy amplification persists under unlimited synthetic data release.

[Sabater et al., 2022] Sabater, C., Bellet, A., and Ramon, J. (2022).
An Accurate, Scalable and Verifiable Protocol for Federated Differentially Private Averaging.
Machine Learning, 111:4249–4293.

[Shamir and Zhang, 2013] Shamir, O. and Zhang, T. (2013).
Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.
In *ICML*.

[Shi et al., 2023] Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. (2023).
Detecting pretraining data from large language models.
arXiv preprint arXiv:2310.16789.

[Steinke et al., 2023] Steinke, T., Nasr, M., and Jagielski, M. (2023).
Privacy auditing with one (1) training run.
In *NeurIPS*.

[Swanberg et al., 2026] Swanberg, M., Annamalai, M. S. M. S., Hayes, J., Balle, B., and Smith, A. (2026).
A unified framework for adversary-aware differential privacy bounds.

[Yaghini et al., 2026a] Yaghini, M., Aerni, M., Zhang, J., Tramèr, F., and Papernot, N. (2026a).
OptiFluence: Principled Design of Privacy Canaries.
In *International Conference on Machine Learning (ICML)*.

[Yaghini et al., 2026b] Yaghini, M., Cebere, T., Menart, M., Bellet, A., and Papernot, N. (2026b).
Private Rate-Constrained Optimization with Applications to Fair Learning.
In *ICLR*.