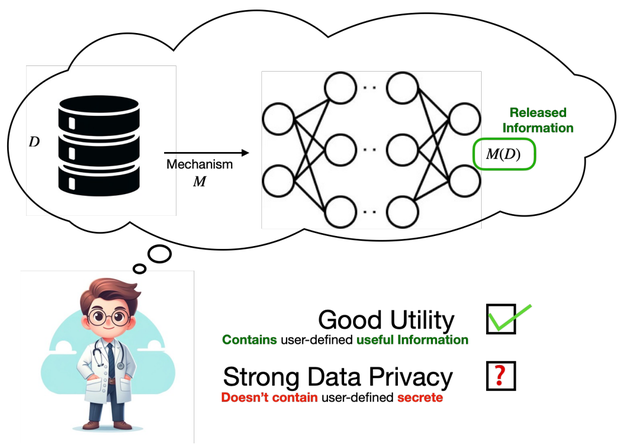
Black-Box Privacy Auditing for Deployed ML: Detecting Violations and Calibrating Noise to Upgrade Privacy

Problem Statement and Goals

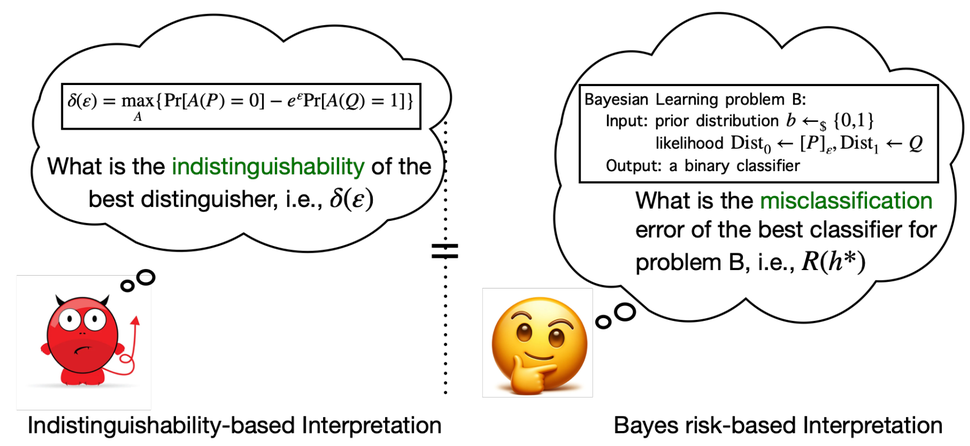


Problem: Deployed Machine Learning (ML) should protect users' data privacy. However, auditing the guarantee (with API-only access) is hard.

Viewing data privacy as indistinguishability between a real world (with a user's data) and an ideal world (without it), our **Goals** are:

- **Efficient black-box privacy estimation**
- **Detecting violations** when deployed mechanisms do not meet claimed privacy guarantees.
- **Measure widely used randomized ML models** without known privacy analysis
- **Upgrade privacy** by calibrating the ML algorithm's noise using the estimated privacy parameters

Approach



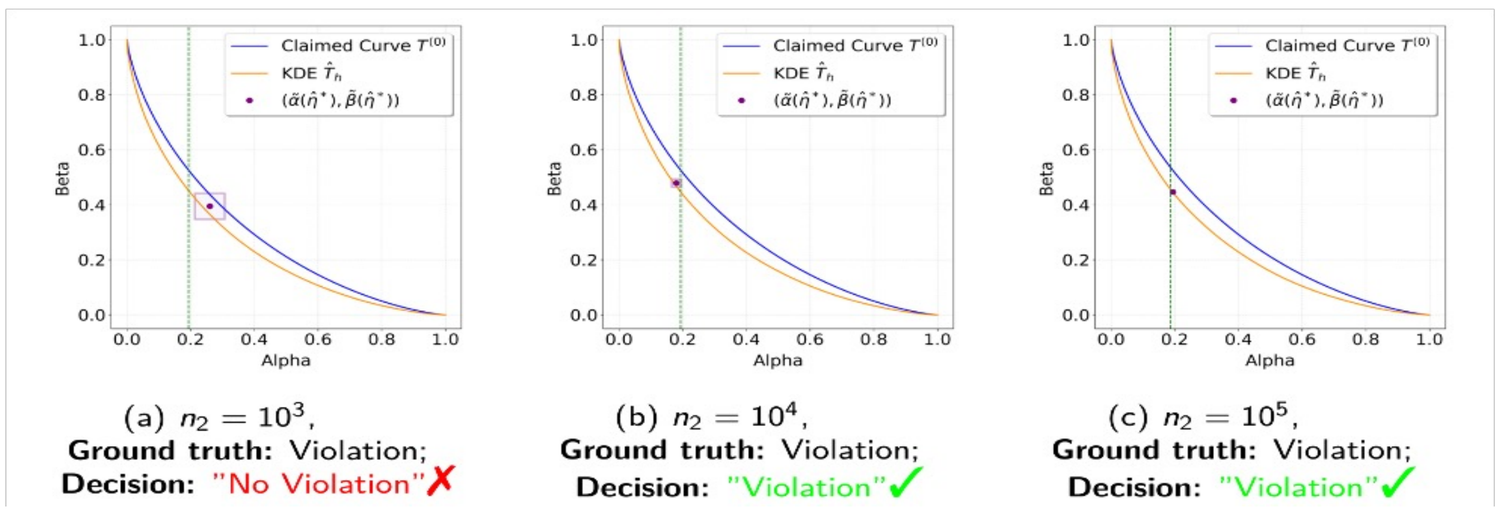
Indistinguishability-based Interpretation — Bayes risk-based Interpretation

Our **core ideas** are:

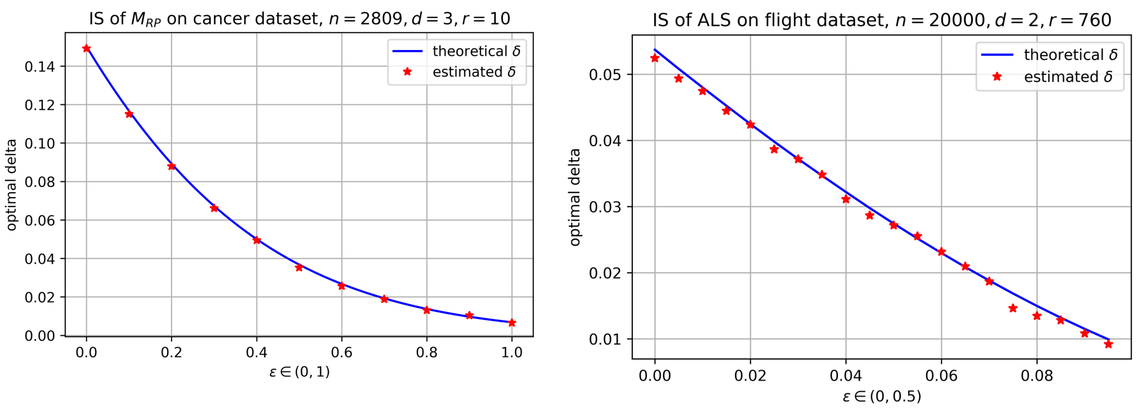
- **Converting privacy quantity** (e.g, DP, f-DP) **estimation into a learning/testing problem,** making privacy estimable from black-box samples.
- **Instantiating estimators** using nonparametric classifiers/tests (e.g., kNN, KDE)
- **From estimation to action:** Use the resulting estimates to (i) audit for violations, (ii) characterize privacy of algorithms, and (iii) guide mechanism upgrades by reducing output distinguishability (e.g., regularization + calibrated noise).

Results

- **Violation detection:** flags faulty mechanisms when sampled outputs contradict a claimed DP / f-DP curve (example: faulty Gaussian).



- **Privacy Auditing beyond standard private algorithms:** empirical privacy estimates for random projection and approximate least squares, matching theory when available.



- **Privacy upgrades:** Differential Privacy mechanism for random projection and least square with improved utility

Database	Method	ε	δ	PDR	DPR
Cancer n=2809	$\mathcal{M}_{RP}$	1	$1/n$	18.942 ± 0.036	0.399 ± 0.005
	Sheffet(2019)	1	$1/n$	56.116 ± 0.130	0.379 ± 0.005
Song n=327346	$\mathcal{M}_{RP}$	1	$1/n$	2.146 ± 0.001	0.787 ± 0.002
	Sheffet(2019)	1	$1/n$	7.373 ± 0.004	0.445 ± 0.001
Flight n=515345	$\mathcal{M}_{RP}$	1	$1/n$	1.763 ± 0.005	1.064 ± 0.004
	Sheffet(2019)	1	$1/n$	3.989 ± 0.013	0.868 ± 0.004

- **Ongoing and Future works:**
  - Sequential Auditing framework
  - End-to-end implementation for complex pipelines: private learning/training -> release & serving -> online privacy monitoring -> automatic noise recalibration

Yu Wei
Georgia Institute of Technology

Black-Box Privacy Auditing for Deployed ML: Detecting Violations and Calibrating Noise to Upgrade Privacy

Abstract

Organizations increasingly deploy machine learning on sensitive data and rely on differential privacy (DP) to limit information leakage. In practice, verifying that an implementation achieved the desired level of privacy is challenging: deployed ML algorithms are complex, may deviate from standard analyses, and are often accessible only through an API.

We present a unified toolkit for black-box estimation and auditing of privacy in machine learning. Under the view that privacy as indistinguishability between output distributions produced on neighboring datasets, we show that key DP quantities --- minimal δ(ε) for (ε,δ)-DP and the trade-off function for f-DP --- can be reduced to the Bayes risk of an optimal classifier distinguishing outputs from neighboring inputs. This connection lets us estimate privacy by training nonparametric classifiers (e.g., k-nearest neighbors) on samples obtained by querying the target mechanism, yielding finite-sample confidence regions and certified empirical lower bounds on privacy loss.

We demonstrate three outcomes: (1) auditing can detect privacy violations in deployed mechanisms, including faulty noise calibration; (2) estimation can characterize the privacy of randomized algorithms without a standard privacy analysis such as random projection and approximate least squares, with empirical estimates closely tracking theoretical curves when available; and (3) estimation combined with analytic tools can ‘upgrade’ an algorithm into a DP-compliant variant with minimal modification, illustrated by a differentially private random projection mechanism and a least square mechanism that improve utility compared to prior baselines on multiple datasets.

Selected publications:

USENIX Security ’25 (General-Purpose f-DP Estimation and Auditing in a Black-Box Setting);
IEEE S&P ’24 (Eureka: A general framework for black-box differential privacy estimators);

arXiv ’23 (The Normal Distributions Indistinguishability Spectrum and its Application to Privacy-Preserving Machine Learning).

Yu Wei
Georgia Institute of Technology