Research on Practical Privacy-Preserving Machine Learning

Privacy-enhancing technologies let us extract value from sensitive data without exposing individuals. We present practical advances in federated learning and noise injection under secure aggregation — and how combining them yields better privacy–utility trade-offs. With tightening regulation and growing stakes around personal data, demonstrable privacy protection is no longer optional.

Henrik Forsgren and Rickard Brännvall
RISE – Research Institutes of Sweden

Are you managing the privacy risk of your AI?

Health records, financial data, and personal communications hold enormous value for research and innovation — but organisations that use them face growing regulatory requirements and reputational risk. The EU's AI Act, EHDS, and GDPR increasingly demand demonstrable evidence that individuals are protected, not just policy commitments. Privacy-enhancing technologies (PETs) provide the technical means to meet these demands.

Instance specific noise injection

Protect Privacy Where It Matters

Problem: Membership inference attacks expose whether someone's data was in training. Differential privacy adds noise for protection — but hurts accuracy and is expensive.

- Gradient clipping is computationally expensive
- DP-SGD often require model changes

Our approach: Forgotten-by-Design (FbD)
Instead of uniform global noise, we downweight the most vulnerable samples using per-instance vulnerability scores. Trades formal guaranties for better accuracy at empirical privacy.

Why is Forgotten-by-Design practical?

- No modifications to model architecture required
- Much better accuracy–privacy trade-off than DP-SGD
- Regulators ask for demonstrable privacy protection

Federated Learning with Heterogeneous Data

Keep data at the source — bring the computation to the data.

Problem: Data heterogeneity across sites (hospitals, regions, devices) degrades federated models. Existing approaches either ignore client differences (FedAvg), require costly cluster discovery (IFCA, DAC), or maintain per-client models (Ditto).

Our approach: Each client condition their training of the global model on training data statistics computed locally at each client. Zero extra communication.

Experiments E1–E4:

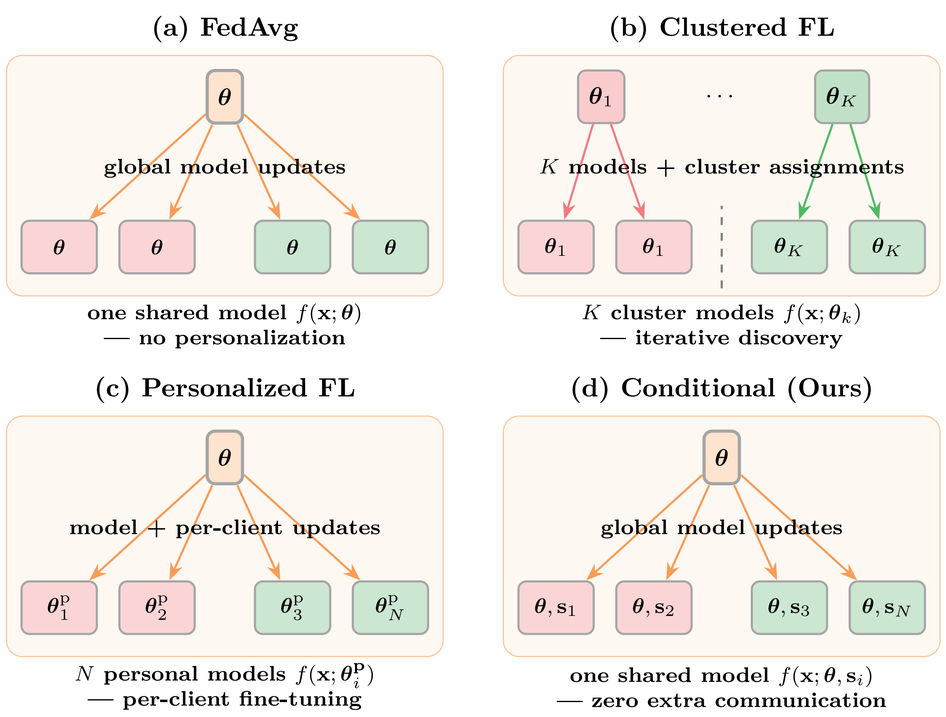
- E1 Label shift — clients see different subsets of classes
- E2 Covariate shift — different input distributions
- E3 Concept shift — same labels map different concepts
- E4 Combined heterogeneity — multiple shifts

Evaluated across 97 configurations and sparsity levels.

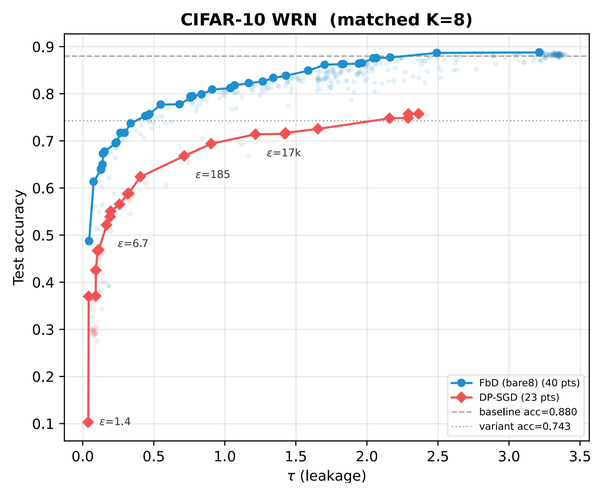
Exp.	Config	FedAvg	Oracle	Cond	Best Other
E1	CIFAR-10 K=5	.325	.927	.929	.906 (IFCA)
E2a	CIFAR-10 K=3	.943	.979	.981	.972 (DAC)
E2b	FMNIST rot. K=4	.789	.900	.895	.868 (DAC)
E3a	FMNIST sem. Rich	.671	.950	.950	.943 (IFCA)
E3b	CIFAR-10 perm. K=2	.408	.795	.808	.773 (DAC)
E4a	MNIST+FM Rich	.939	.957	.955	.953 (IFCA)
E4b	CIFAR-10 C=3	.743	.914	.935	.880 (Ditto)

Why is Client Conditional FL practical?

- No additional communication or disclosure beyond FedAvg
- Only adds <1% model parameters
- Fully compatible with secure aggregation and DP-SGD



Left figure is comparing our Client Conditional method with the conventional Persolized and Clustered methods for heterogeneous federated learning. Illustrated for N clients partioned over K (unknown) clusters. Our method doesn't require peer identification or cluster discovery, and adds no communication overhead or private data disclosure compared to the vanilla Federated Averaging.
The table shows that it ourperforms standard heterogeneous federated learning benchmark methods.



Forgotten by Design (FbD) instance specific noise-injection outperforms conventional Differential Privacy (DP-SGD). Figure shows utility-privacy trade-off, that is, accuracy versus traceability τ.

Other Privacy Enhancing Technologies at RISE

Fully homomorphic encryption enables end-to-end computation on encrypted data without decryption.

- Not practical for training
- Encrypted inference remains challenging.
- Co-design of Transformers for faster inference.

Synthetic data mimics statistical patterns in source data

- Useful for testing pipelines and interoperability
- High fidelity for ML remains difficult.
- Synthetic ≠ anonymised unless proven.

Privacy testing for validation of PET efficiency.

- Membership inference attacks, data reconstruction, ...

Combining PETs. No single PET covers all attack surfaces.

- FHE based protocols for federated aggregation are practical
- Secure aggregation combined with DP for better utility.

HOW WE CAN HELP

PET advisory. Matching privacy technologies to your data flows, threat models, and regulatory requirements (GDPR, EHDS, AI Act).

Privacy testing. Auditing models and synthetic datasets for information leakage – prepare evidence for regulatory reporting (DPIA).

Explore opportunities through the **TEFs** (Test & Experiment Facilities) and **AI Factory MIMER.**

RISE Research Institutes of Sweden
www.ri.se / info@ri.se

CONTACT
Rickard Brännvall, rickard.brannvall@ri.se