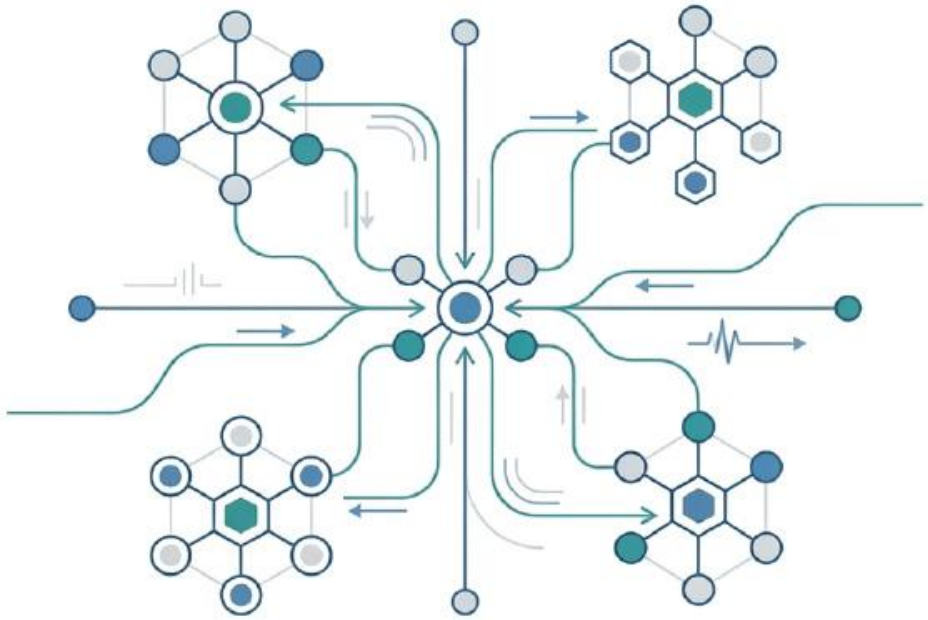
ΑΡΧΗ ΠΡΟΣΤΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ



Επιστημονική Ημερίδα της Αρχής:
2η Ημέρα Διαλόγου με την Ερευνητική Κοινότητα

22 Οκτωβρίου 2026

Privacy in Machine Learning: From k-Anonymity and Differential Privacy to Machine Unlearning and the Right to Erasure (Art. 17 GDPR)

Vassilios S. Verykios, Vice Rector, Professor · **Georgios Feretzakis**, PhD, Researcher & Adjunct Teaching Staff
Hellenic Open University — Big Data Analytics and Anonymization Laboratory

EDPB Opinion 28/2024: a high bar for “anonymous” AI models

A model trained on personal data is anonymous only if both tests hold:

Test 1 · Extraction

The likelihood of directly — or probabilistically — extracting personal data from the model is insignificant.

Test 2 · Querying

The likelihood of obtaining such data through queries, intentionally or not, is insignificant.

If unlawful processing is established

supervisory authorities may consider proportionate corrective measures:

- Fines
- Temporary limitation of processing
- Erasure of part or all of the dataset
- Erasure of the model itself
- **Retraining — as a remedial action**

Assessed case by case — evidence supervisory authorities may weigh:

Source selection | Data minimisation | Privacy-preserving training (e.g., DP) | Output filtering | Resistance to attacks | Documentation

Source: EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models (adopted 17 Dec 2024).

Privacy risk follows personal data through the ML lifecycle



Data	Training	Trained model	Deployment	Erasure request
RISK Re-identification & linkage	RISK Gradients leak individual records	RISK Unintended memorization	RISK Membership inference & extraction	RISK Influence persists in weights
SAFEGUARD k-anonymity and its extensions	SAFEGUARD Differential privacy in training	SAFEGUARD Dedup, memorization audits	SAFEGUARD Attack testing, output filters	SAFEGUARD Machine unlearning (Art. 17)

Today's path: data-side protections → training-time guarantees → memorization and extraction → unlearning as a route to erasure.

k-Anonymity hides each record in a crowd of at least k

Each combination of quasi-identifiers **(age, ZIP, sex)** appears at least k times — via generalisation and suppression.

Age	ZIP	Sex	Diagnosis
30–39	151**	*	Diabetes
30–39	151**	*	Asthma
30–39	151**	*	Hypertension
40–49	145**	*	**Cancer**
40–49	145**	*	**Cancer**
40–49	145**	*	**Cancer**

Illustrative 3-anonymous table. The second group shares one diagnosis — anyone known to be in it is exposed (homogeneity attack).

87% vs. 61–63%

unique by ZIP code, full birth date and sex. Sweeney (2000): 87% for 1990. Golle (2006) re-estimated ~61% (1990) and ~63% (2000): lower, still significant.

Known limits

1. **Homogeneity & background knowledge**
   → ℓ-diversity, t-closeness
2. **Curse of dimensionality**
   rich, high-dimensional data lose most utility
3. **Syntactic, not semantic**
   protection depends on assumed attacker knowledge

Differential privacy bounds any one record’s influence

DEFINITION · (ε, δ)-DP

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S] + \delta$$

for all adjacent datasets D, D’ differing in one protected unit — a record by default; an individual under person-level DP. Smaller ε → stronger protection.

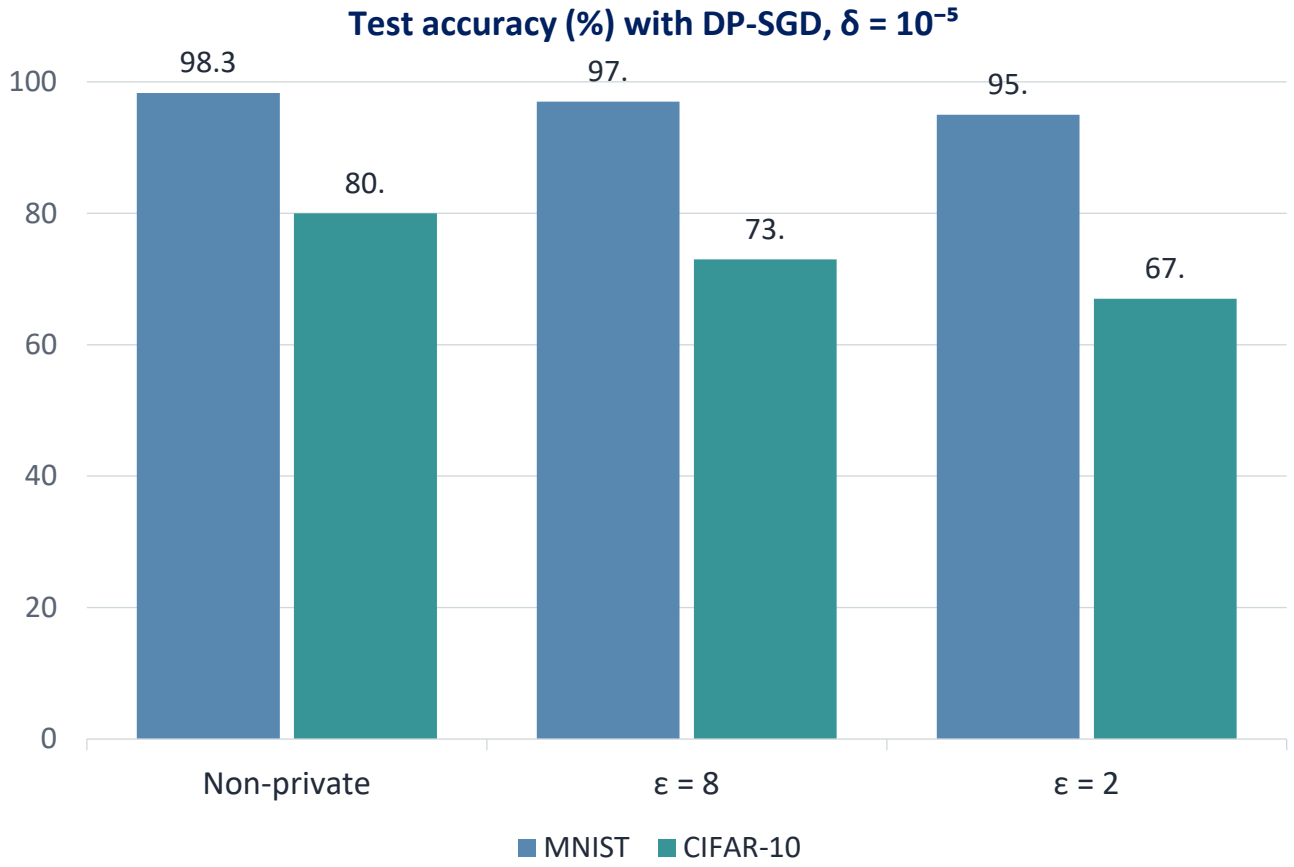
- Robust to auxiliary information
- Composes across releases
- Immune to post-processing

DP-SGD: private training in three steps

1. **Clip**
   each example’s gradient to norm C — bounds the contribution of a single training example
2. **Add noise**
   calibrated Gaussian noise to the aggregated gradient
3. **Account**
   track cumulative privacy loss (ε, δ) across all training steps

Dwork et al. (2006); Abadi et al., “Deep Learning with Differential Privacy”, ACM CCS 2016.

Privacy has a price: accuracy falls as ε tightens



–13 pts

CIFAR-10 accuracy at ε = 2 vs. the non-private baseline

On sensitive data the cost is uneven

- Small clinical cohorts: the same noise drowns more of the signal
- Clipping and noise can disproportionately reduce accuracy for under-represented or hard-to-learn subgroups
- Report subgroup performance alongside ε — not only average accuracy

Illustrative results from Abadi et al., ACM CCS 2016 (MNIST: 90% at ε = 0.5); CIFAR-10 experiments used pretrained convolutional features. Bagdasaryan et al., NeurIPS 2019.

Models memorize — and memorized data can be extracted

>10,000

unique verbatim training examples recovered from ChatGPT (GPT-3.5, 2023)

$200

query budget sufficient for the attack

150×

more training data emitted under a simple “divergence” prompt attack

Nasr, Carlini et al., “Scalable Extraction of Training Data from (Production) Language Models”, 2023.

Membership inference

Was this person’s record in the training set?

Shokri et al., 2017

Training-data extraction

Verbatim names, emails, phone numbers recovered by prompting

Carlini et al., 2021

Attribute inference

Infer a sensitive attribute from partial knowledge of a person

Fredrikson et al., 2015

Memorization grows with **model size**, **data duplication** and **prompt context length** *(Carlini et al., ICLR 2023)*

ΑΡΧΗ ΠΡΟΣΤΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ

Machine unlearning: removing one person's influence

← Stronger guarantee · higher cost

Weaker guarantee · lower cost →

Retrain from scratch | **SISA sharded retraining** | **Certified removal** | **Heuristic fine-tuning**

Exact unlearning

The model is identical in distribution to one never trained on the data.

- **Full retraining:** the reference standard; prohibitive for large models
- **SISA:** retrain only the shard that held the data; exact relative to the same sharded procedure, not a monolithic model (Bourtoule et al., IEEE S&P 2021)

Approximate unlearning

Weights approximate the retrained model; guarantees range from formal to empirical.

- **Certified removal:** provable guarantee under stated assumptions, e.g. L2-regularised linear models (Guo et al., ICML 2020)
- **LLM methods:** gradient ascent / fine-tuning on a "forget set"; evaluated with benchmarks such as TOFU (Maini et al., 2024)

Article 17 GDPR meets the model: what does “erasure” mean?

Whether Art. 17 reaches a model depends on the ground, whether the model processes personal data and the exceptions; **controllers must select effective measures, demonstrate compliance and respect any supervisory orders.**

Response to an erasure request	What it actually does	Technical strength	Cost
Delete from the training corpus	Stops future use; the trained model is unchanged	Weak if memorized	Low
Suppress in outputs (filters)	Attempts to suppress targeted data in outputs; model weights remain unchanged	Masks, does not erase	Low
Approximate unlearning	Reduces the data’s influence in the weights	Partial; hard to verify	Medium
Exact unlearning (SISA / retrain)	Model as if never trained on the data	Strong — yet not compliance by itself	High

Art. 17(1) erasure “without undue delay” where a ground applies (e.g., consent withdrawn, unlawful processing)

Art. 17(3)(d) research is no automatic exception: Art. 89(1) safeguards must be met and erasure must be likely to render impossible or seriously impair the research

Art. 19 notify recipients to whom the data were disclosed, unless impossible or disproportionate effort

Where the technical guarantees stop

Verification

Black-box tests cannot conclusively establish forgetting. Exact and certified methods give formal guarantees under stated assumptions; heuristic LLM unlearning stays hard to verify.

Relearning & leakage

Approximately unlearned knowledge can resurface after light fine-tuning or rephrased prompts.

What ε really means

Protection depends on ε, δ and the unit of privacy — record or person. A large ε protects little.

Copies & derivatives

Distilled models, fine-tunes, embeddings and RAG indices carry the data beyond the base model.

Technical guarantees are evidence for accountability — **not a substitute for it.**

For controllers: design for erasure from day one

1. Map and minimise

Training-data lineage, deduplication, PII filtering before training

2. Test the anonymity claim

Membership-inference and extraction tests with measurable results kept as evidence

3. Train privately where feasible

DP-SGD with documented ε, δ and unit of privacy

4. Architect for unlearning

Sharded training (SISA), adapters, versioned checkpoints

5. Layer deployment controls

Output filtering, rate limits and monitoring — complements, not substitutes

6. Operationalise Art. 17

Covers data, models, derivatives, RAG and backups; DPIA reviewed where required

EDPB CEF report on erasure (Feb 2026; 32 SAs, 764 controllers): inadequate procedures, difficulties erasing backup data and reliance on ineffective anonymisation.

Key takeaways

1. **Anonymised data ≠ anonymous model.** The EDPB test asks what can be extracted from the model, not what was intended.

2. **Data-side techniques help, at a price.** k-anonymity is fragile; DP gives a formal bound but may reduce utility, particularly for small or under-represented cohorts.

3. **Unlearning can support Art. 17 — guarantees depend on the method.** Exact and certified approaches are possible; heuristic LLM unlearning remains hard to verify.

FURTHER READING

Feretzakis, G.; Verykios, V.S. **Trustworthy AI: Securing Sensitive Data in Large Language Models.** *AI* 2024, 5, 2773–2800. https://doi.org/10.3390/ai5040134

EDPB (2026). **CEF 2025 report: implementation of the right to erasure by controllers.**



Thank you

Questions & discussion

Vassilios S. Verykios · Georgios Feretzakis
Hellenic Open University — Big Data Analytics and Anonymization Laboratory