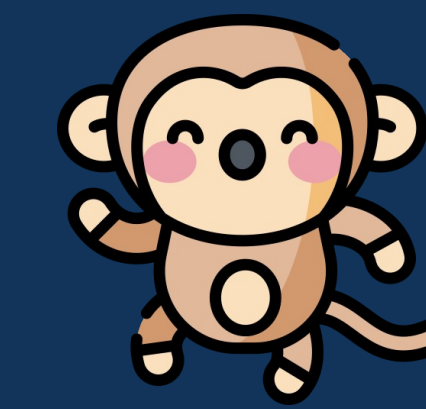


Models Designed to Forget: Machine Unlearning via Key Deletion

Sonia Laguna, Jorge da Silva Gonçalves, Moritz Vandenhirtz, Alain Ryser, Irene Cannistraci[†], Julia E. Vogt[†]

Department of Computer Science, ETH Zurich, Switzerland

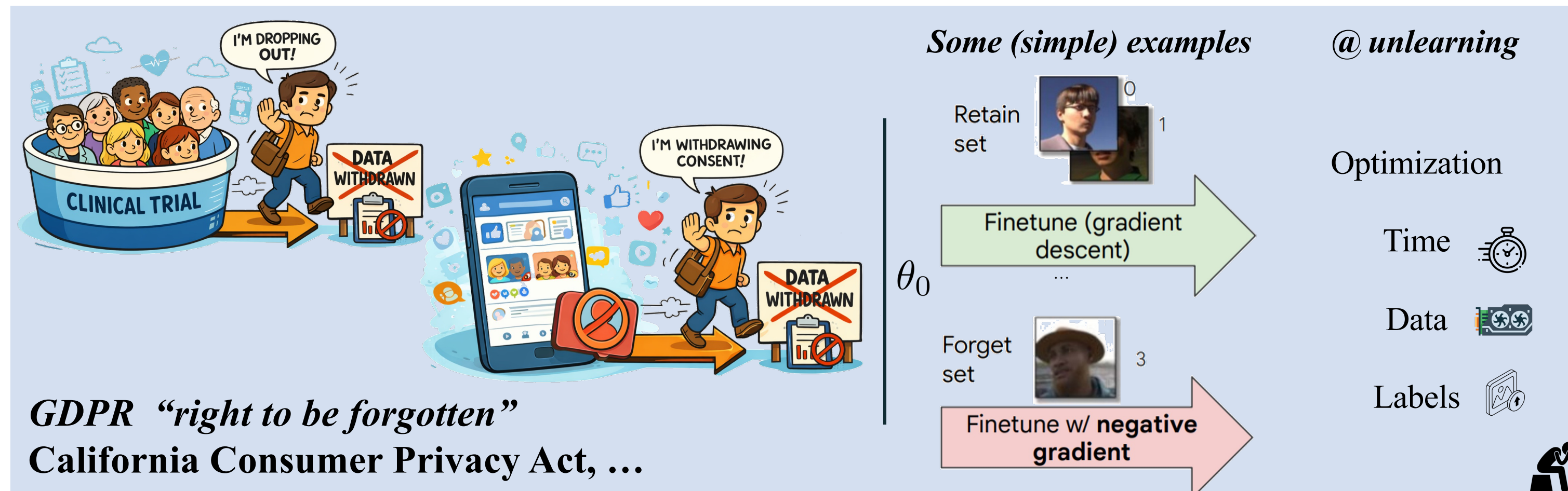


✉ slaguna@inf.ethz.ch

[†] Shared senior authorship



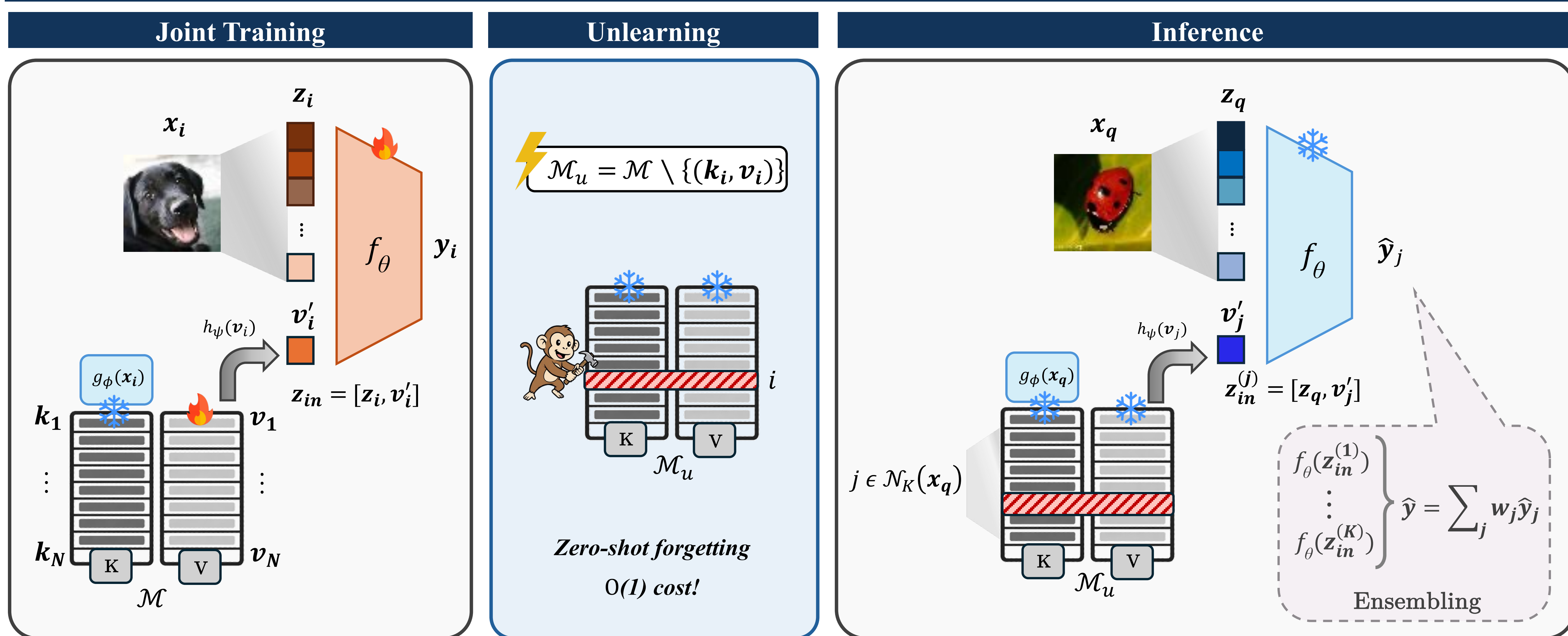
Approximate Unlearning



Rethinking Machine Unlearning

- **Change of paradigm:** *Unlearning by design*
- Real world: Requests can be **anticipated**
- **Trained to support forgetting**
- **MUNKEY:** Machine UNlearning via **KEY** Deletion

MUNKEY: Machine UNlearning via KEY deletion



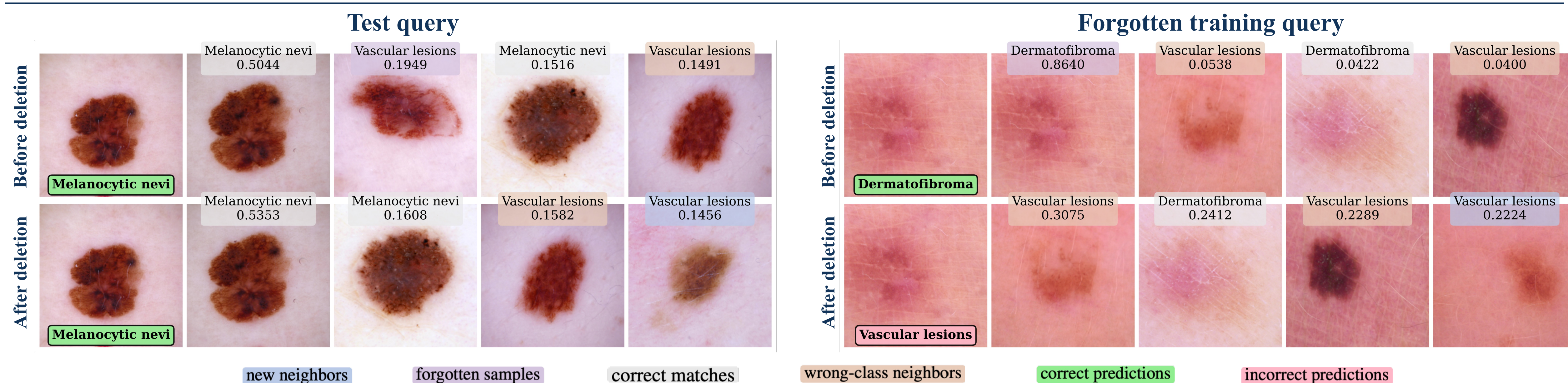
Memory: frozen keys, learnable exemplars

$$\mathcal{M} = \{(\mathbf{k}_i, \mathbf{v}_i)\}_{i=1}^N, \quad \mathbf{k}_i = g_\phi(\mathbf{x}_i)$$

Inference: similarity-weighted ensemble

$$w_j = \frac{\exp(\text{sim}(\mathbf{k}_q, \mathbf{k}_j)/\tau)}{\sum_{i \in \mathcal{N}_K(\mathbf{x}_q)} \exp(\text{sim}(\mathbf{k}_q, \mathbf{k}_i)/\tau)}$$
$$\hat{\mathbf{y}} = \sum_{j \in \mathcal{N}_K(\mathbf{x}_q)} w_j f_\theta([\mathbf{z}_q, h_\psi(\mathbf{v}_j)])$$

Effect of retrieval



Forgetting and utility

Random deletion of 10% instances; ViT-Tiny
TA / RA / FA: test / retain / forget accuracy.

Extended datasets and ablations in the paper

Method	DermaMNIST · 10% deletion					CIFAR-10 · 10% deletion				
	TA ↑	RA ↑	FA	MIA	Gap ↓	TA ↑	RA ↑	FA	MIA	Gap ↓
Retrain (T)	76.46 ±0.22	100.0 ±0.00	74.86 ±1.77	48.88 ±0.69	—	78.13 ±0.64	91.62 ±0.55	77.57 ±0.83	49.77 ±0.32	—
Retrain (M)	81.25 ±0.66	99.70 ±0.08	80.29 ±1.41	50.04 ±1.05	—	92.38 ±0.06	99.93 ±0.03	92.05 ±0.23	49.90 ±0.34	—
KNN	74.60 ±0.24	85.54 ±0.14	74.81 ±1.23	49.94 ±0.90	—	91.06 ±0.05	94.56 ±0.06	91.02 ±0.13	50.16 ±0.10	—
CF	75.28 ±0.98	97.50 ±1.55	94.48 ±2.23	64.41 ±1.96	9.71 ±0.99	77.66 ±0.50	91.74 ±0.48	87.25 ±1.04	55.86 ±0.16	4.09 ±0.44
CF-k	76.94 ±0.42	99.86 ±0.13	99.90 ±0.13	67.10 ±3.73	10.97 ±1.05	79.60 ±0.63	94.30 ±0.55	94.16 ±0.61	58.93 ±0.40	7.48 ±0.41
EU-k	76.89 ±0.58	99.60 ±0.23	99.52 ±0.18	69.27 ±0.23	11.47 ±0.51	79.44 ±0.72	94.22 ±0.55	94.00 ±0.65	58.87 ±0.30	7.36 ±0.42
L1-sparse	75.81 ±0.45	92.36 ±1.80	86.81 ±1.17	57.69 ±1.13	7.26 ±0.78	52.80 ±0.55	53.47 ±0.75	52.97 ±0.86	50.79 ±0.79	22.28 ±0.48
NegGrad+	76.26 ±0.37	96.31 ±1.03	83.48 ±1.25	55.83 ±1.14	4.87 ±0.69	77.29 ±0.27	91.08 ±0.42	83.13 ±0.89	53.65 ±0.32	2.71 ±0.41
Amnesiac	75.56 ±0.23	98.30 ±0.85	82.00 ±2.38	47.01 ±1.11	2.93 ±0.73	78.31 ±0.35	92.17 ±0.75	85.67 ±1.05	52.51 ±0.21	2.89 ±0.46
SCRUB	73.97 ±1.81	76.02 ±2.48	74.68 ±1.23	49.62 ±0.25	6.85 ±0.96	76.51 ±0.78	85.31 ±0.71	83.72 ±0.94	54.18 ±0.47	4.62 ±0.48
SSD	75.07 ±2.08	92.07 ±10.2	89.73 ±12.9	63.29 ±8.28	9.65 ±4.67	78.54 ±1.08	93.29 ±0.71	92.84 ±0.96	58.91 ±0.43	6.62 ±0.52
AMUN	76.46 ±0.69	98.33 ±0.94	95.05 ±0.64	63.70 ±0.27	9.17 ±0.59	77.71 ±0.46	91.25 ±0.56	86.85 ±3.46	55.90 ±1.94	4.05 ±1.05
CoUn	79.50 ±0.35	99.80 ±0.05	87.24 ±1.33	54.92 ±1.29	5.42 ±0.67	80.68 ±0.44	95.72 ±0.01	91.32 ±0.54	55.75 ±0.34	6.60 ±0.36
MUNKEY	81.06 ±0.44	99.71 ±0.08	84.00 ±1.22	51.19 ±1.19	1.27 ±0.64	92.63 ±0.20	99.94 ±0.01	92.53 ±0.52	51.40 ±0.44	0.56 ±0.21

Utility

Forgetting

Trade-off

Utility

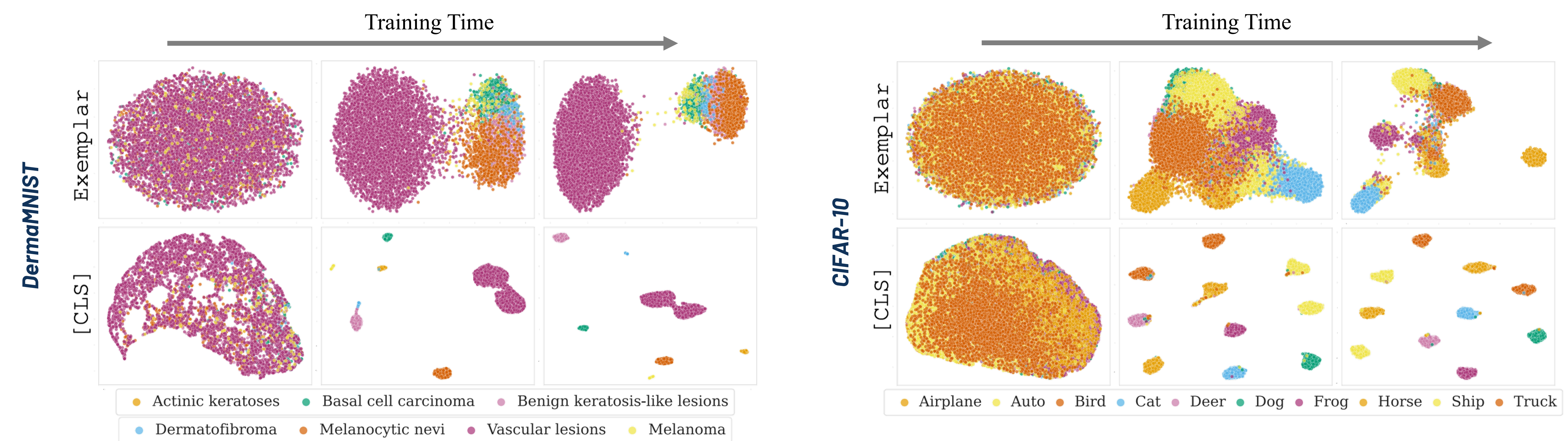
Forgetting

Trade-off

Minimal overhead, fast deletion

Method	Cost	DermaMNIST	CIFAR-10
Standard ViT	Train / epoch	24.7 s	174.6 s
MUNKEY	Train / epoch	25.9 s	178.5 s
Fastest baseline	Unlearn	23.1 s	105.8 s
Slowest baseline	Unlearn	313.4 s	1940.5 s
MUNKEY	Unlearn	≈ 0 s	≈ 0 s

Information encoded in the memory



Exemplar tokens retain finer structure; [CLS] representations become class-clustered

Summary

- **The Problem:** Post-hoc unlearning is expensive.
- **The Paradigm:** *Unlearning by design*: train models to support forgetting.
- **The Solution:** **MUNKEY** enables near-instantaneous key deletion from a memory bank with high utility.



References Baselines

Kurmanji et al. Towards Unbounded Machine Unlearning. NeurIPS, 2023.
Khalil, Y. H., et al. Coun: Empowering machine unlearning via contrastive learning. NeurIPS 2025
Ebrahimpour-Borojeny, A., et al. Not all Wrong is Bad: Using Adversarial Examples for Unlearning. ICLR 2025

Di et al. Adversarial Machine Unlearning. ICLR, 2025.

Foster et al. Fast Machine Unlearning without Retraining through Selective Synaptic Dampening. AAAI, 2024.