LoTUS: Large-Scale Machine Unlearning with a Taste of Uncertainty

Christoforos N. Spartalis[1,2] Theodoros Semertzidis[2] Efstratios Gavves[1,3] Petros Daras[2]

Universiteit van Amsterdam

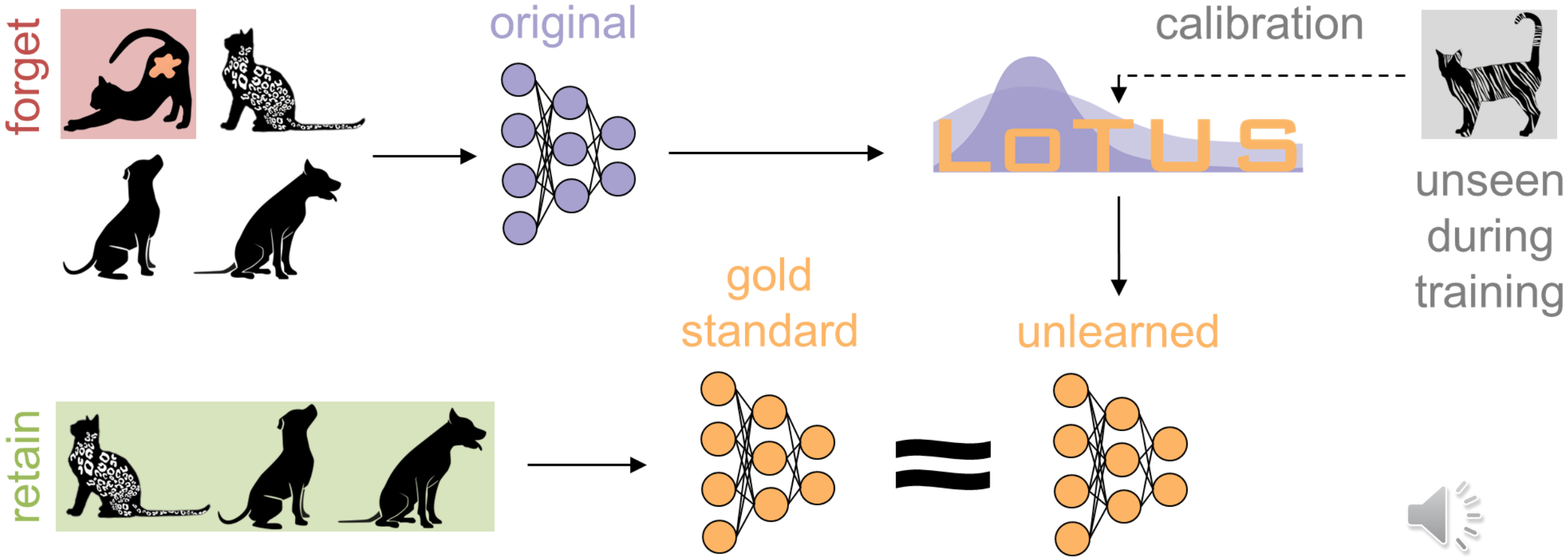
CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

ARCHIMEDES

CVPR Nashville JUNE 11-15, 2025

ATLANTIS

Machine Unlearning



Entropy-based Unlearning

Why Unlearning?

- DNNs memorize sample-specific information
- Privacy leakage in overconfident predictions
- Unlearning by increasing model's uncertainty

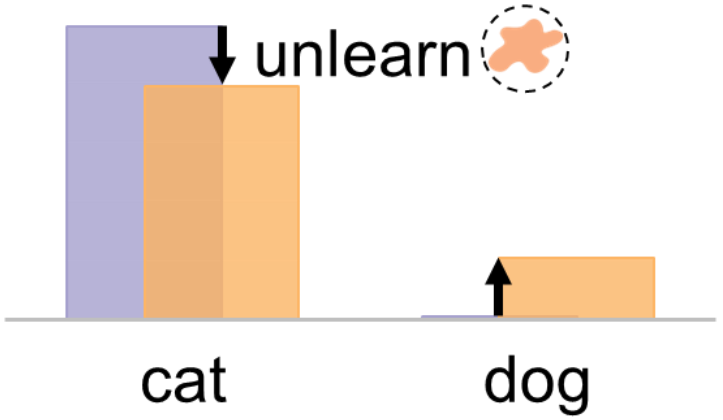
How much Uncertainty?

- LoTUS: 1st method to answer that
- Better balance between forgetting-retention
- Information-Theoretic Framework

X_S: forget set

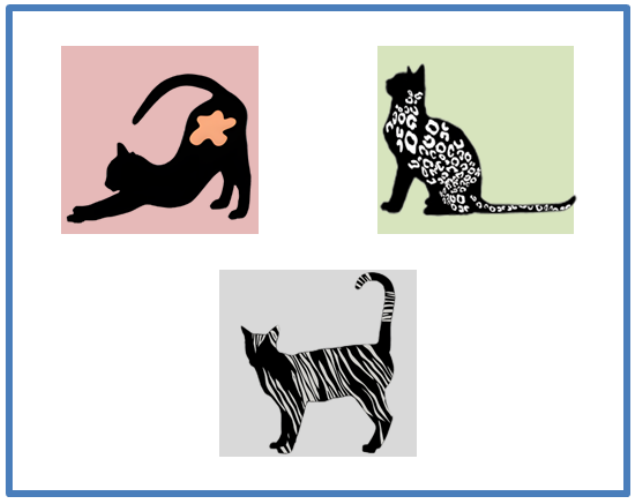
$f_{or}(X_S)$ & $f_{un}(X_S)$

model outputs



Information-Theoretic Framework

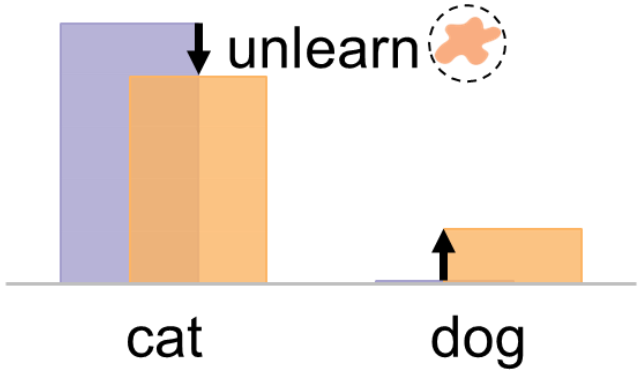
X: general population



X_S: forget set

$f_{or}(X_S)$ & $f_{un}(X_S)$

model outputs



$$I(\, f_{or}(X_S); X_S \,) = I(\, f_{or}(X_S); X \,) + I(\, f_{or}(X_S); X_S \mid X \,)$$

total info captured by the model for the forget set

global info from general features in the training set (e.g., body shape of cats)

additional subset-specific info from unique features memorized by the model

Objective

$$I(\, f_{or}(X_S); X_S \,) \quad = \quad I(\, f_{or}(X_S); X \,) \quad + \quad I(\, f_{or}(X_S); X_S \mid X \,)$$

total info captured by the model for the forget set

global info from general features in the training set (e.g., body shape of cats)

additional subset-specific info from unique features  memorized by the model

Objective 

Forget: $I(\, f_{un}(X_S); X_S \mid X \,) \triangleq 0$

Retain: $I(\, f_{un}(X_S); X \,) \triangleq I(\, f_{or}(X_S); X \,)$

$$\Big\} \; I(\, f_{un}(X_S); X_S \,) \triangleq I(\, f_{or}(X_S); X \,)$$

Unseen set as a "perfectly unlearned" set : $I(\, f_{or}(X_S); X_S \,) = I(\, f_{or}(X_S); X \,)$

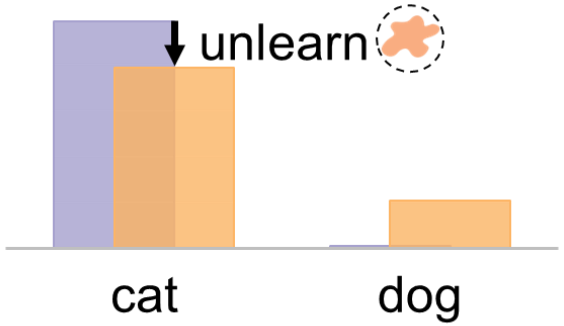
$I(f_{un}($  $);$  $) = I(f_{or}($  $);$  $) \Rightarrow H($  $\mid f_{un}($  $)) = H($  $\mid f_{or}($  $))$

$\text{Accuracy}(f_{un}($  $)) = \text{Accuracy}(f_{or}($  $))$

Logits Tempering Unlearning Strategy

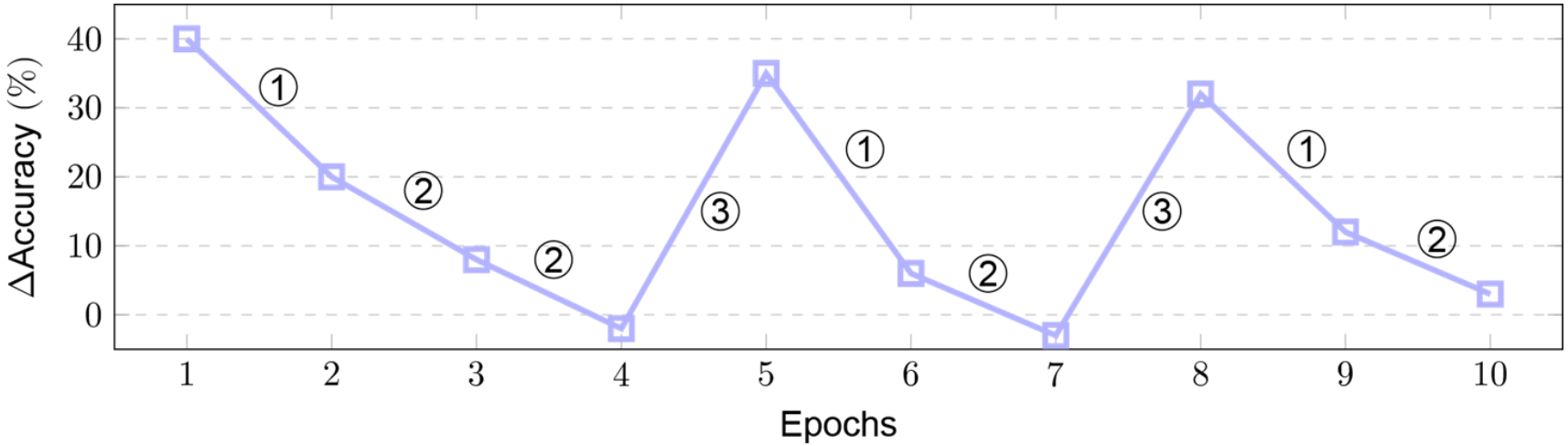
Forget loss: $gs(f_{or}(\text{  }), \tau_d) \odot \log s(f_{un}(\text{  }))$

Retain loss: $gs(f_{or}(\text{  }), \tau \to 0^+) \odot \log s(f_{un}(\text{  }))$



$$\tau_d = \exp\left(\alpha\left(\text{Accuracy}(f_{un}(\text{  })) - \text{Accuracy}(f_{or}(\text{  }))\right)\right)$$

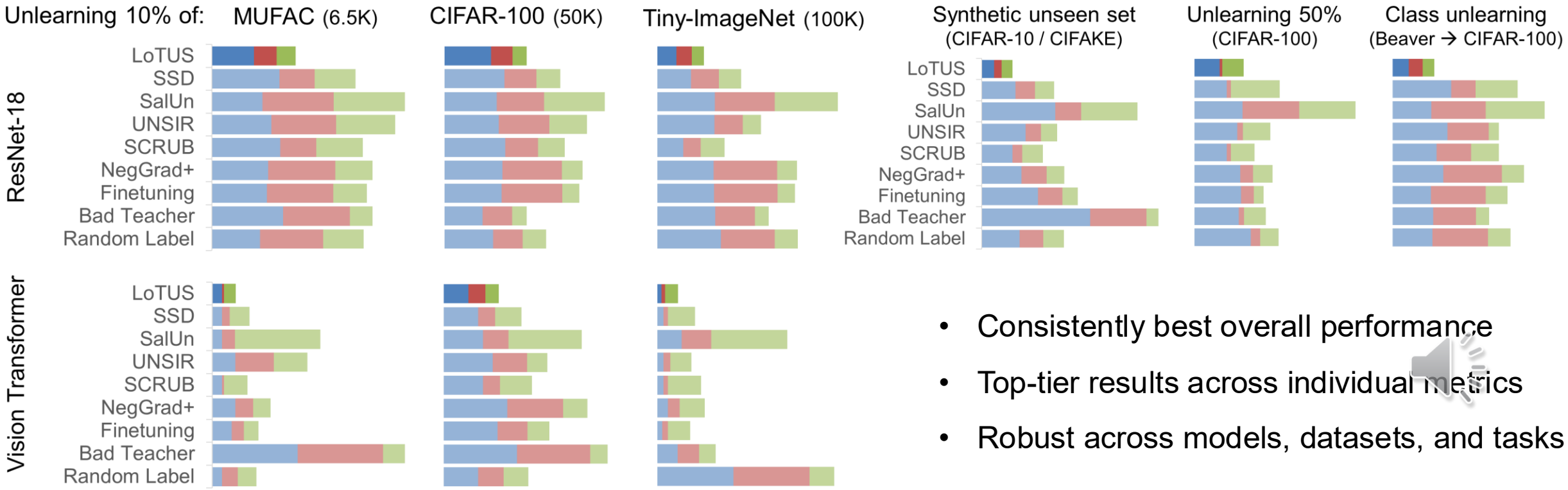
Dynamic Control over Uncertainty



- ΔAcc > 0 → τ > 1
- ΔAcc ⬇ → τ ⬇
- ΔAcc = 0 → τ = 1
- ΔAcc < 0 → τ < 1

Results

- **Average Gap** (↓): Balance between forgetting and retention
- **JSD** (↓): Unlearning efficacy & Resilience to the Streisand Effect
- **Runtime Estimation** (↓): Unlearning efficiency



- Consistently best overall performance
- Top-tier results across individual metrics
- Robust across models, datasets, and tasks

Large-Scale Real-World Benchmarking

The gold standard model is **not available**

- Pre-trained ViT
- ImageNet-1K (1.2M)
- Limited data access

Method	RF-JSD $\times 1e4$ ($\downarrow$)	Time ($\downarrow$)	Retain Acc.	MIA Acc.
Original	$1.22_{\pm 0.01}$	(pre-trained)	$0.94_{\pm 0.00}$	$0.71_{\pm 0.00}$
Finetuning	$2.22_{\pm 0.02}$	$16.24_{\pm 0.03}$	$0.97_{\pm 0.00}$	$0.78_{\pm 0.00}$
NegGrad+	$2.17_{\pm 0.02}$	$18.10_{\pm 0.03}$	$0.97_{\pm 0.00}$	$0.80_{\pm 0.00}$
Rnd Labeling	$1.80_{\pm 0.09}$	$19.37_{\pm 0.03}$	$0.95_{\pm 0.01}$	$0.74_{\pm 0.01}$
Bad Teacher	$3.16_{\pm 3.25}$	$11.66_{\pm 0.03}$	$0.77_{\pm 0.21}$	$0.52_{\pm 0.18}$
SCRUB	$1.24_{\pm 0.01}$	$24.49_{\pm 0.03}$	$0.94_{\pm 0.00}$	$0.71_{\pm 0.00}$
SSD	$1.23_{\pm 0.01}$	$22.61_{\pm 0.10}$	$0.94_{\pm 0.00}$	$0.71_{\pm 0.00}$
UNSIR	$2.54_{\pm 0.03}$	$33.12_{\pm 0.03}$	$0.99_{\pm 0.00}$	$0.77_{\pm 0.01}$
SalUn	$1.83_{\pm 0.03}$	$59.27_{\pm 0.37}$	$0.95_{\pm 0.00}$	$0.74_{\pm 0.01}$
LoTUS	$\mathbf{1.11}_{\pm \mathbf{0.01}}$	$\mathbf{10.72}_{\pm \mathbf{0.01}}$	$0.94_{\pm 0.00}$	$0.61_{\pm 0.01}$

Novel Metric: Retrain-Free Jensen-Shannon Divergence

$JSD\left(f_{un}(\ ) \,||\, f_{gold}(\ )\right)$

$RF\text{-}JSD\left(f_{un}(\ ) \,||\, f_{or}(\ )\right)$

Highly correlated

Contributions

- ✓ **Information-Theoretic Framework** for formalizing Machine Unlearning
- ✓ **LoTUS**: Scalable and effective entropy-based unlearning strategy
- ✓ **RF-JSD**: Evaluation metric for large-scale and real-world benchmarking

Thank you for your attention and interest!

Scan for: code, paper, video, blog, and slides

Highly modular code for benchmarking machine unlearning in classification tasks:
github.com/cspartalis/LoTUS