Machine Unlearning

Lucas Bourtoule*, Varun Chandrasekaran*, Christopher A. Choquette-Choo*, Hengrui Jia*,
Adelin Travers*, Baiwu Zhang*, David Lie, Nicolas Papernot

*Joint Lead Authors

Motivation

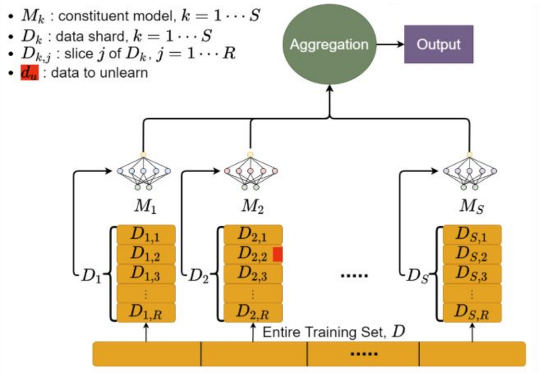
1. Synergy **missing between legal and tech. experts**
2. Complex interplay **between data and parameters**

Concrete Problem: *Unlearn* data from trained ML models (e.g., DNNs) such that removal guarantee is comprehensible

Prior Approaches

Differentially Private Learning [Abadi et al., 2016]	**Statistical Query Learning [Cao et al., 2015]**
1. Requires ε=0 for compliance 2. Strongly influences accuracy 3. Guarantee is probabilistic ☹	1. Applicable for simple models 2. Can make limited number of queries 3. No known algorithm for DL models ☹

SISA Training

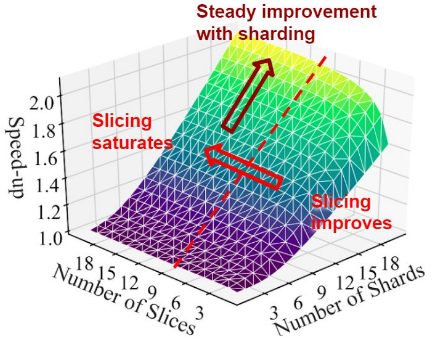


- M_k : constituent model, $k = 1 \cdots S$
- D_k : data shard, $k = 1 \cdots S$
- $D_{k,j}$: slice j of D_k, $j = 1 \cdots R$
- d_u : data to unlearn

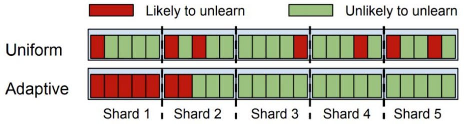
Aggregation
Output
M_1
M_2
M_S
D_1
D_2
D_S
$D_{1,1}$ $D_{1,2}$ $D_{1,3}$ $D_{1,R}$
$D_{2,1}$ $D_{2,2}$ $D_{2,3}$ $D_{2,R}$
$D_{S,1}$ $D_{S,2}$ $D_{S,3}$ $D_{S,R}$
.....
Entire Training Set, D

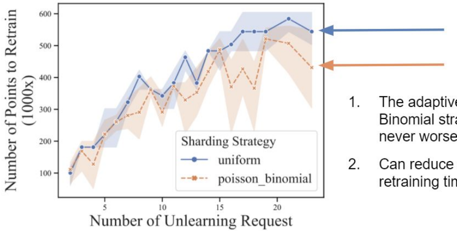
Tuneable Knob	Retraining speed-up	Storage Cost	Accuracy
Sharding	⬆		⬇
Slicing	⬆	⬇	
Aggregation Strategy			⬆

Advantage of Sharding & Slicing



Distribution Aware Sharding





1. The adaptive Poisson Binomial strategy is never worse
2. Can reduce analytical retraining time.