UF Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

CIS6930: Trustworthy Machine Learning

# Can machine unlearning preserve membership privacy?

Manas Gupta
Malvika Jadhav
Swarnabha Roy

POWERING THE NEW ENGINEER TO TRANSFORM THE FUTURE

Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

# Problem Statement

To study of Membership Inference Attacks on Unlearnt Classification models

# Objectives

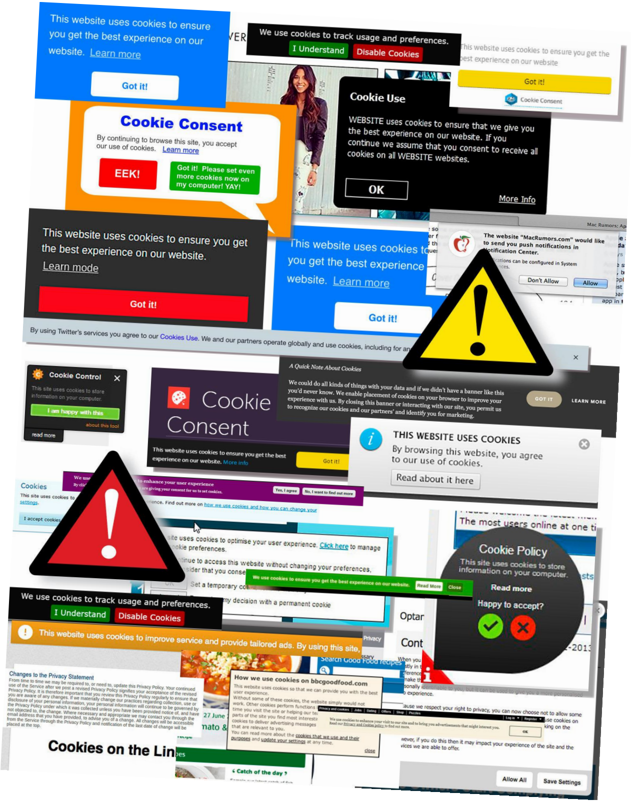
1. To implement machine unlearning on a classification model trained on the CIFAR-10 dataset
2. To implement and test a Membership Inference Attack on the unlearnt model
3. To investigate whether machine unlearning plays a role in preserving membership privacy

UF Herbert Wertheim College of Engineering
UNIVERSITY *of* FLORIDA

# Motivation

## Sharing personal data improves our experience online but at what cost?

## Can machines forget what they have learnt just like we do?

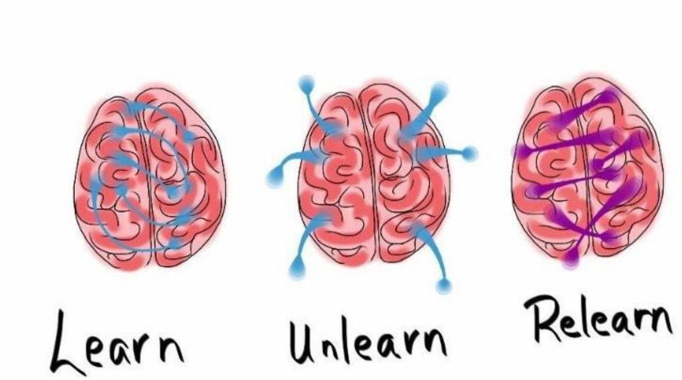


Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

# ML Techniques/ Background

# Machine Unlearning

- What is machine unlearning?
- Ways to perform machine unlearning.
  - SISA
  - Logit based filtration




*[Image credit: Flickr user: Giulia Forsythe]*

# SISA Training Approach

**Paper**: *Bourtoule et al.,Machine unlearning, 2021 IEEE Symposium on Security and Privacy (SP)*

**Goal**: To reduce the influence of individual data points on the trained model

**Steps**:

- **Sharding**
- **Isolation**
- **Slicing**
- **Aggregation**

# Membership Inference Attack

**Paper:** ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models

- Posterior Attack

UF Herbert Wertheim College of Engineering
UNIVERSITY *of* FLORIDA

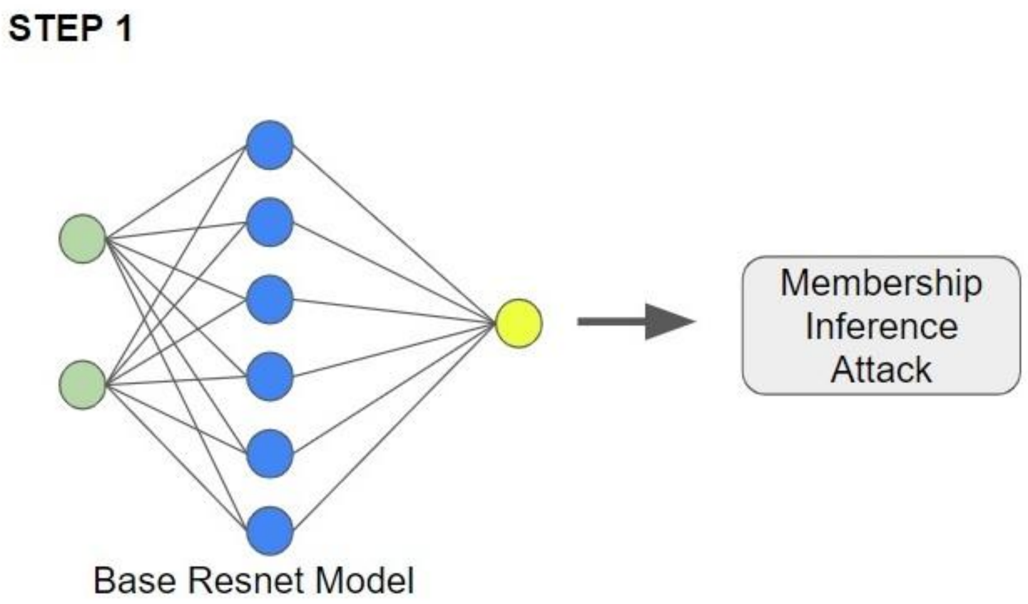
# Experimental Methodology

# Experimental Methodology

Step 1: Posterior attack on base model

Step 2: Machine Unlearning using SISA training

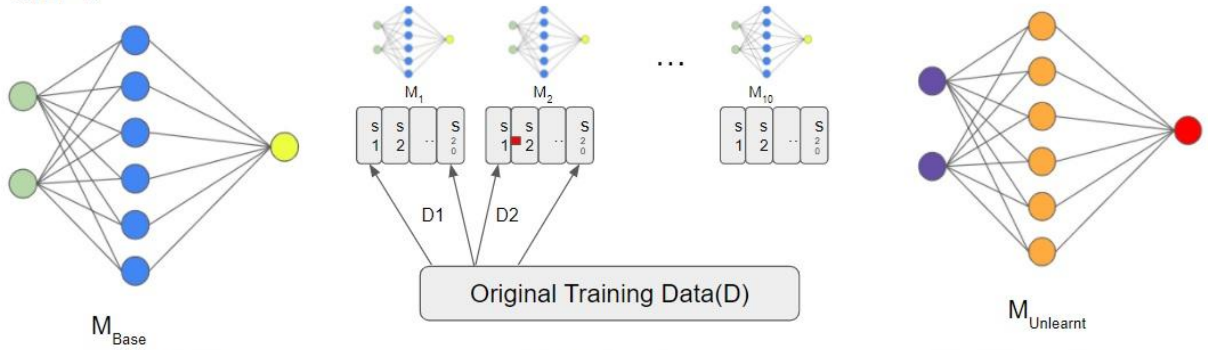
Step 3: Posterior attack on unlearnt model




Step 1: Posterior attack on base model

Dn : n[th] data shard from training data n = 1..10
Sm : n[th] data slice from data shard m = 1..20
Mn : Model corresponding to n[th] data shard
: Data slice to unlearn

**STEP 2**



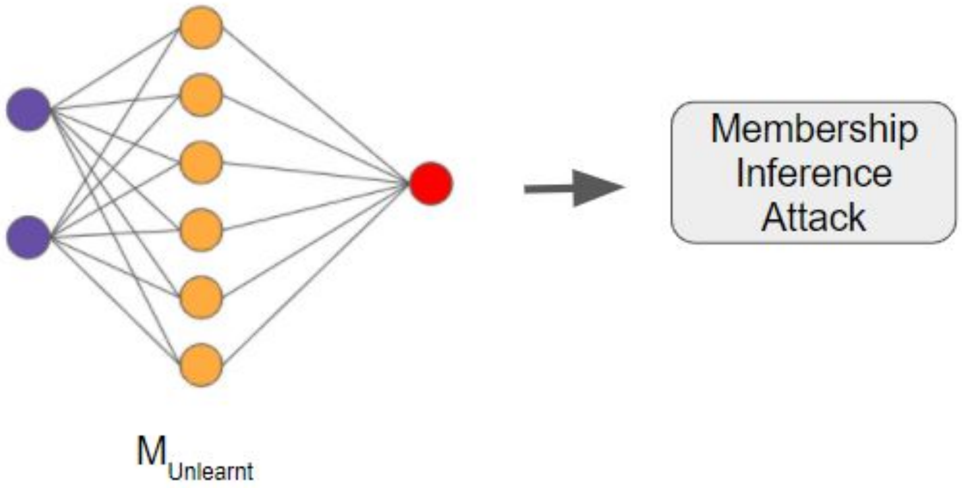

Base Resnet Model

SISA Training

Model after Unlearning

Step 2: Machine Unlearning using SISA training

STEP 3




Step 3: Posterior attack on unlearnt model

UF Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

# Tools and infrastructure

# Tools

- Python : Tensorflow and Keras libraries
- Tesla V100 GPU (Google Vertex AI Cloud)

Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

# Dataset

# CIFAR - 10

- 60000 - 32 x 32 colored images
- 10 classes
- Train and Test data sets

Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

# Results

# Base Architecture

- Training size - 50k, testing size-10k
- 10 classes
- Handout dataset accuracy - 89%, loss - 0.57.
- Posterior attack:
  - Sample size: 2000, balanced in and out classes
  - Attack accuracy: 85%, advantage: 0.17

# Machine Unlearning - SISA

- Training size - 45k, testing size-10k
- Size (Slices) of data removed randomly - 5k
- Constituent model: training accuracy of 82% and testing accuracy of 68%
- Posterior attack:
  - Accuracy: 52.6%
  - Advantage: 0.05

# Conclusion

# Conclusion

- During our implementation we also modified the aggregation step in SISA training to generate output in terms of confidence values opposed to a target label.

- Experimental results show that unlearning techniques can bring the membership inference attack accuracy very close to random guessing.

- Another observation is that using a training method that works with small sized data shards can result in compromising accuracy of overall predictions especially if performed on overparameterized models.

# Thank You.

Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

# References

# References

- Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers,Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In 2021 IEEE Symposium on Security and Privacy (SP), pages 141–159. IEEE, 2021.

- Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models.arXiv preprint arXiv:1806.01246, 2018.