THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Opening Minds • Shaping the Future • 啟迪思維 • 成就未來

# Privacy Protection in Federated Learning


JUNXIAO WANG

PolyU Edge Intelligence Lab

DEPARTMENT OF COMPUTING
電子計算學系

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

# Introduction to PEIL

The PolyU Edge Intelligence Laboratory (PEIL) [1]

Directed by **Prof. Dr. Song Guo**

Team members:

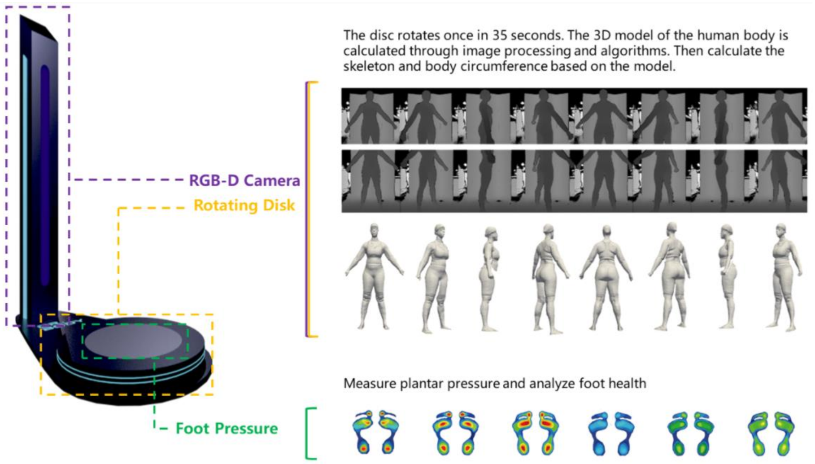
- 1 Research Assistant Professor
- 4 Postdoctoral Fellow
- 17 PhD Students

PEIL's Research Interests are as Follows:

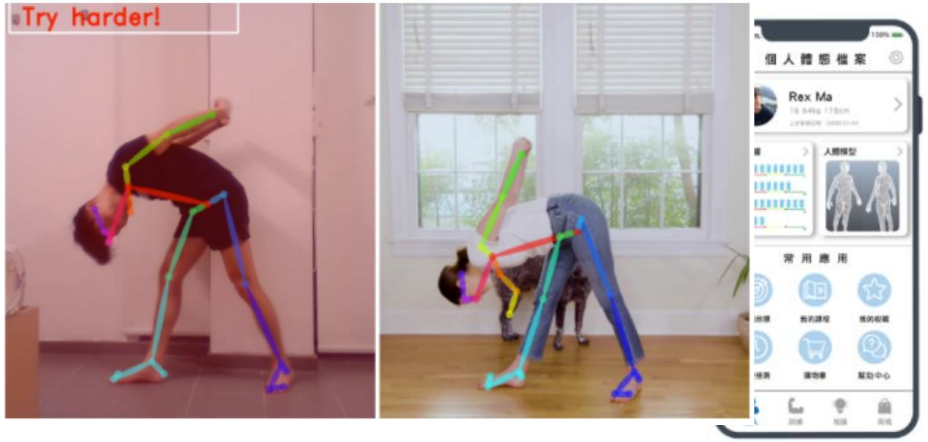
- Edge AI and Federated Learning
- AI Empowered Internet-of-Things
- Edge Computing Driven Ubiquitous Blockchain

[1] https://peilab.comp.polyu.edu.hk/

## Pathways to Impact (Smart Healthcare)




(a) Scoliosis Diagnosis




(b) Scoliosis Recovery

# Topics of This Talk

**1.** Privacy Protection and Federated Learning

See what's the privacy protection trend and how it performs in federated learning

**2.** Gradient Leakage in Federated Learning

Identify the threats of gradient leakage attack and how we can defend it

**3.** Machine Unlearning in Federated Learning

Identify what's the federated unlearning how we can achieve that efficiently

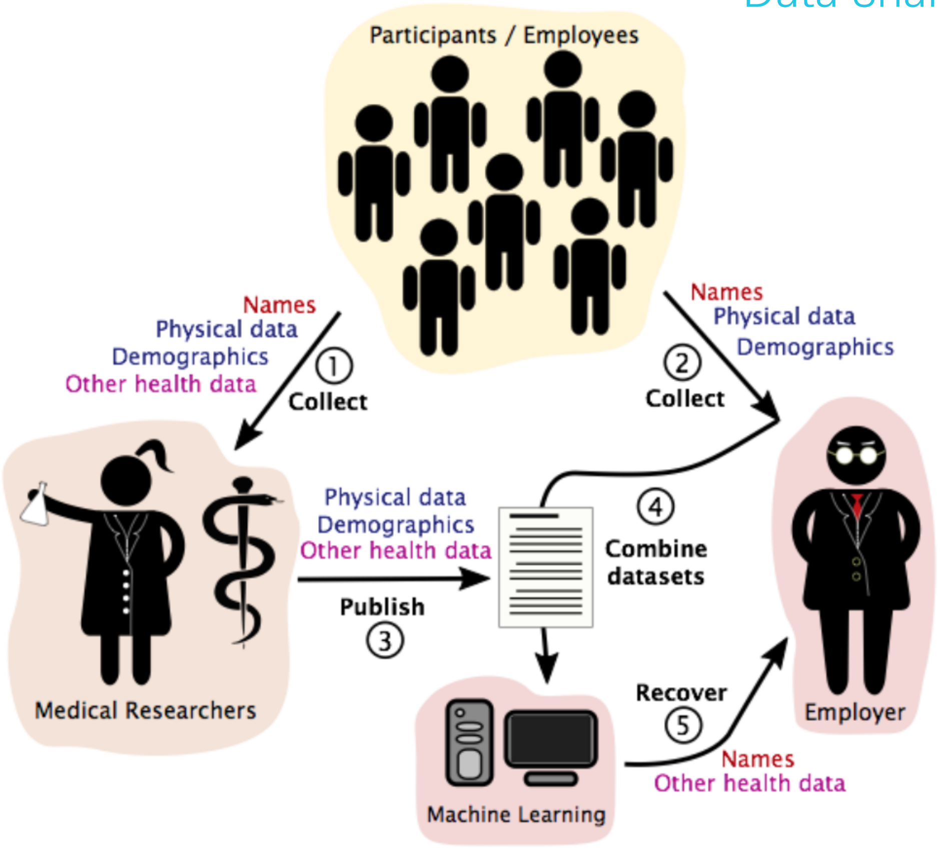
# Part 1.

# Privacy Protection and Federated Learning

**See what's the privacy protection trend and how it performs in federated learning**

# Introduction to Privacy Protection Trend

Data Sharing *vs.* Privacy Protection




(a) Identifying People via their Health Data


(b) PCPD of Hong Kong

# Introduction to Privacy Protection Trend

New Privacy Legislation:

- Calls for Transparency and Clarity of Data
- Empowers Users to Remove their Data

## No one's ready for GDPR

*'Very few companies are going to be 100 percent compliant on May 25th'*

By Sarah Jeong | @sarahjeong | May 22, 2018, 3:28pm EDT

## Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control

**Authors:** Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, Igor Santos Authors Info & Affiliations

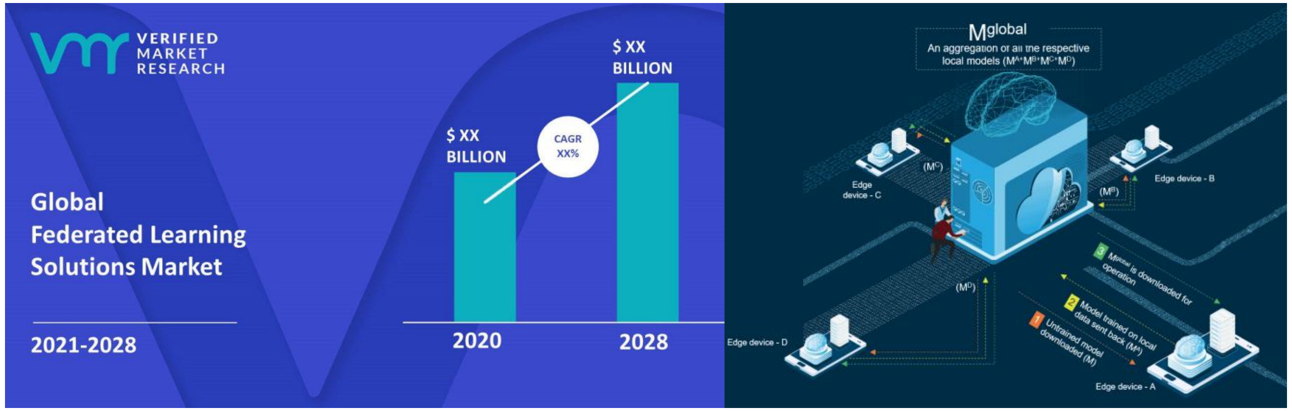
## A Study on Subject Data Access in Online Advertising After the GDPR

Authors | Authors and affiliations

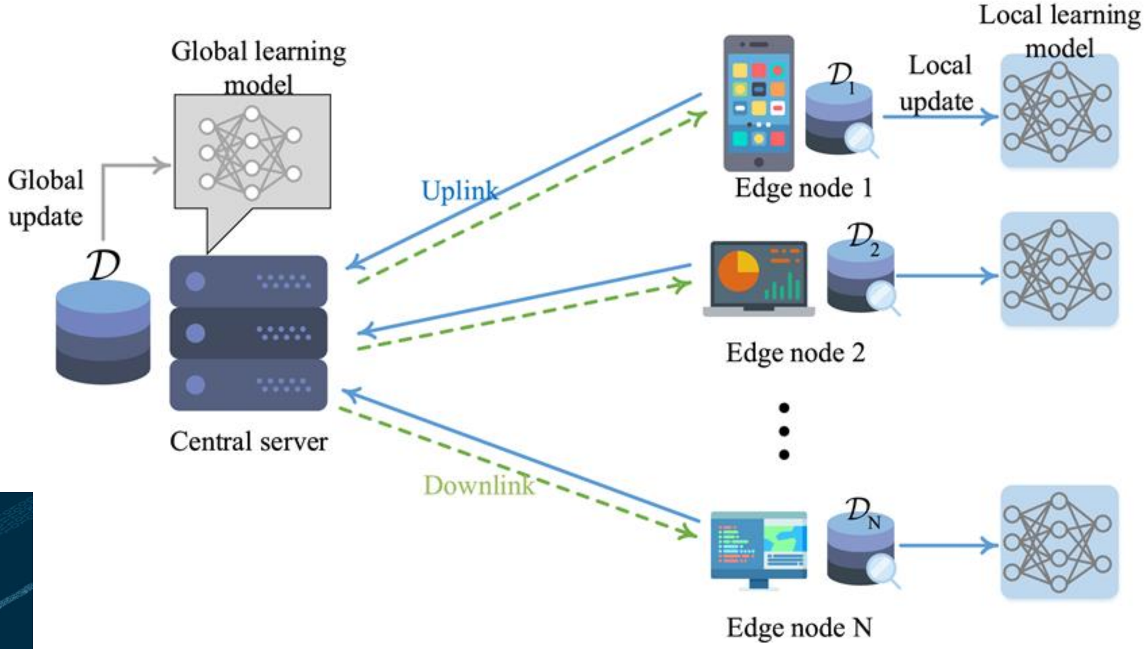
Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, Norbert Pohlmann

# Introduction to Federated Learning

(a) TensorFlow Federated (TFF): **a framework for implementing Federated Learning**




(b) Market Statistics and Application of FL




(c) FL workflow: How Federated Learning performs

[1]https://www.tensorflow.org/federated/
[2]https://www.everestgrp.com/
[3]https://www.verifiedmarketresearch.com/

# Part 2.

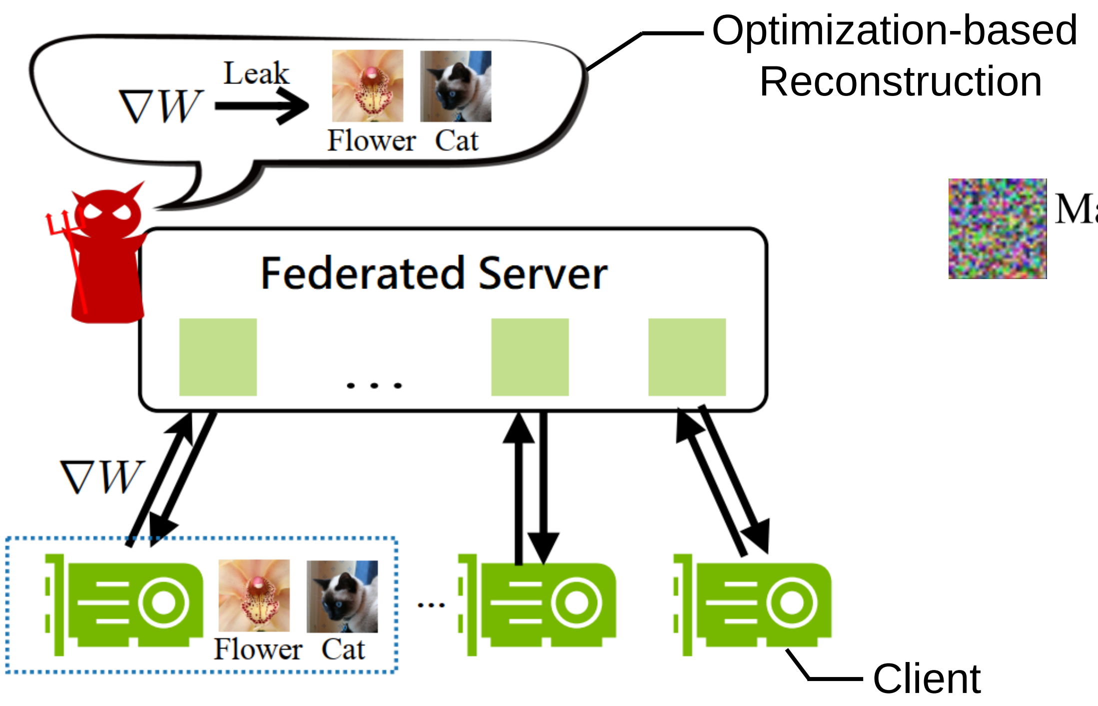
## Gradient Leakage in Federated Learning

**Identify the threats of gradient leakage attack and how we can defend it**

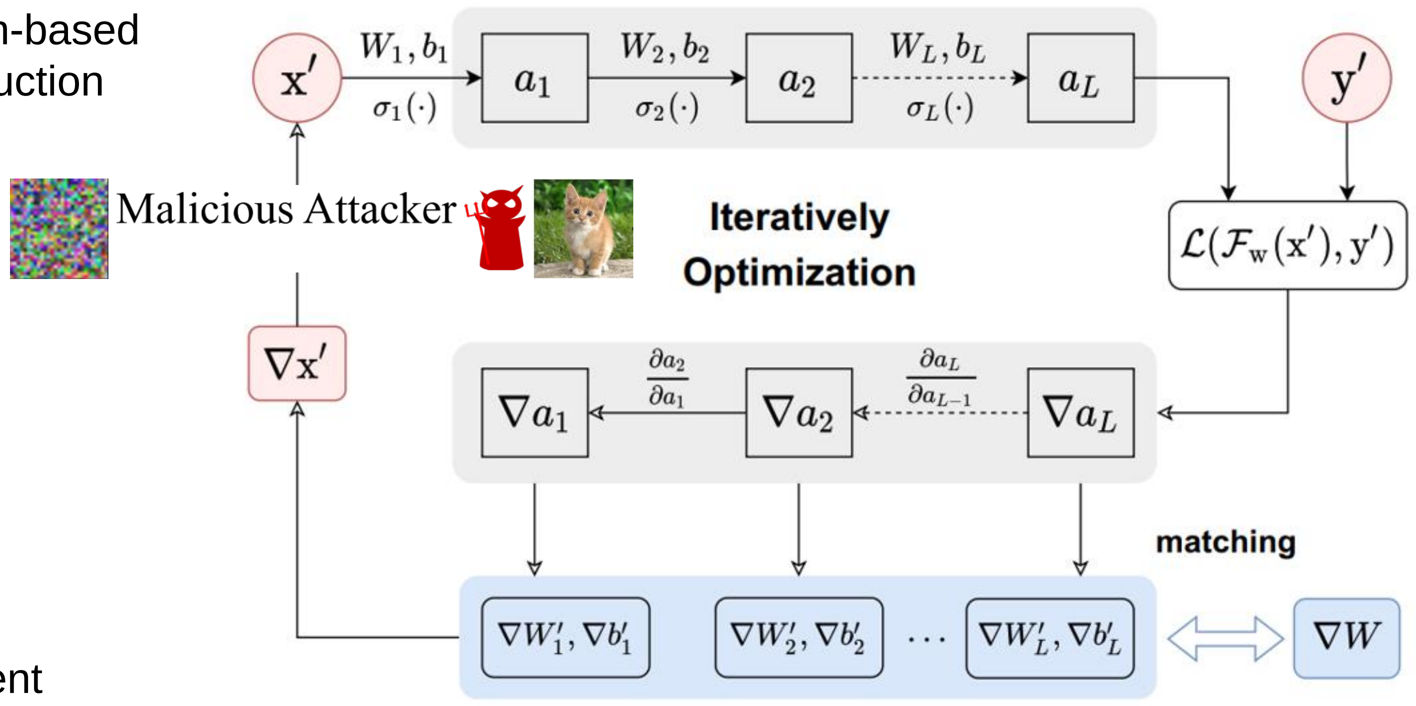
# Gradient Leakage Attack: Deep Leakage from Gradients

MIT, NeurIPS 2019 [1]

- Background: An ***honest-but-curious*** **attacker**, who can be the **federated server**. The attacker can observe **gradients of a victim** and he attempts to **recover data from gradients**.




(a) Threat Model [1]




(b) Workflow of the Optimization-based Reconstruction

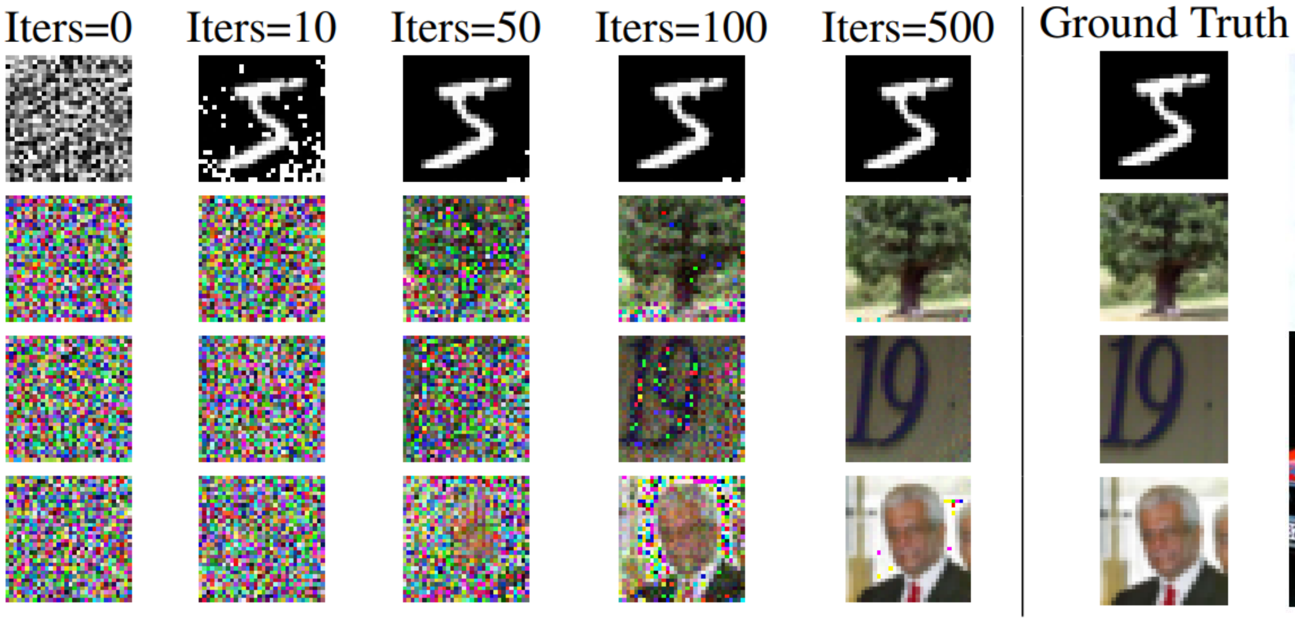
# **Gradient Leakage Attack** pixel-wise level for images

**Deep Leakage from Gradients**

MIT, NeurIPS 2019 [1]

**Inverting Gradients**

Siegen, NeurIPS 2020 [2]




(a) Deep Leakage on Images from MNIST, CIFAR-100, SVHN and LFW [1]

(b) Additional Positive Cases for a Trained ResNet-18 on ImageNet [2]

**Question: How to Protect Privacy from Gradients? Cryptographic Methods?**

# **Existing Defenses against Gradient Leakage** pros and cons

- **General Privacy Protection Methods**

  - Homomorphic Encryption (HE)

    - Advantages: Gradient Aggregation is Performed on Ciphertexts.

  - Multi-Party Computation (MPC)

    - Advantages: Zero-Knowledge of Gradient Aggregation's Input/Output.

    - **Limitations: High Computation and Communication Overhead**

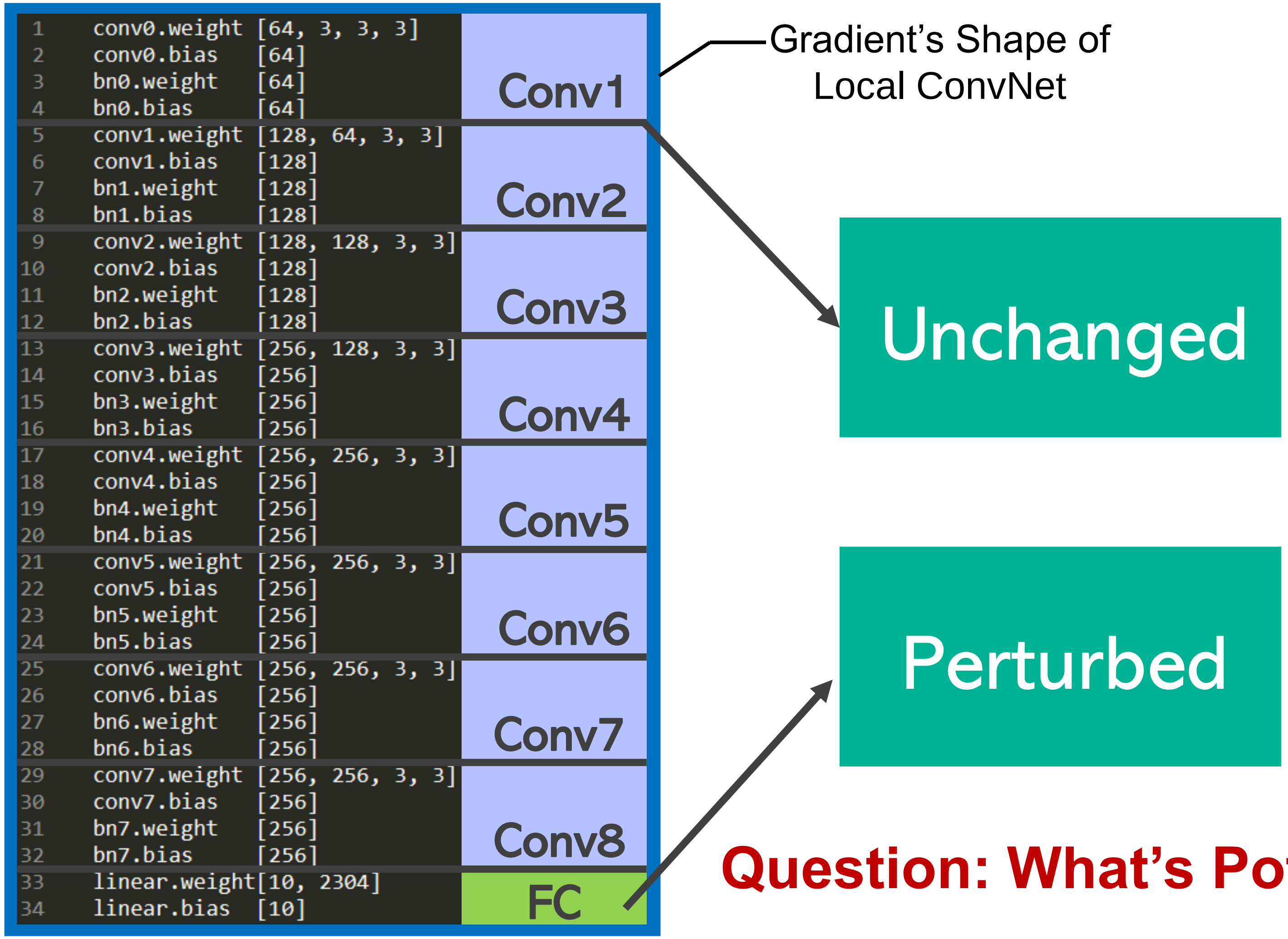
  - Local Differential Privacy (LDP)

    - Advantages: Identify Samples from Gradients within Theoretical Bound.

    - **Limitations: High Convergence Accuracy Loss**

## Defense Specific to Gradient Leakage Attack

"Provable Defense against Privacy Leakage in Federated Learning", Duke, CVPR 2021



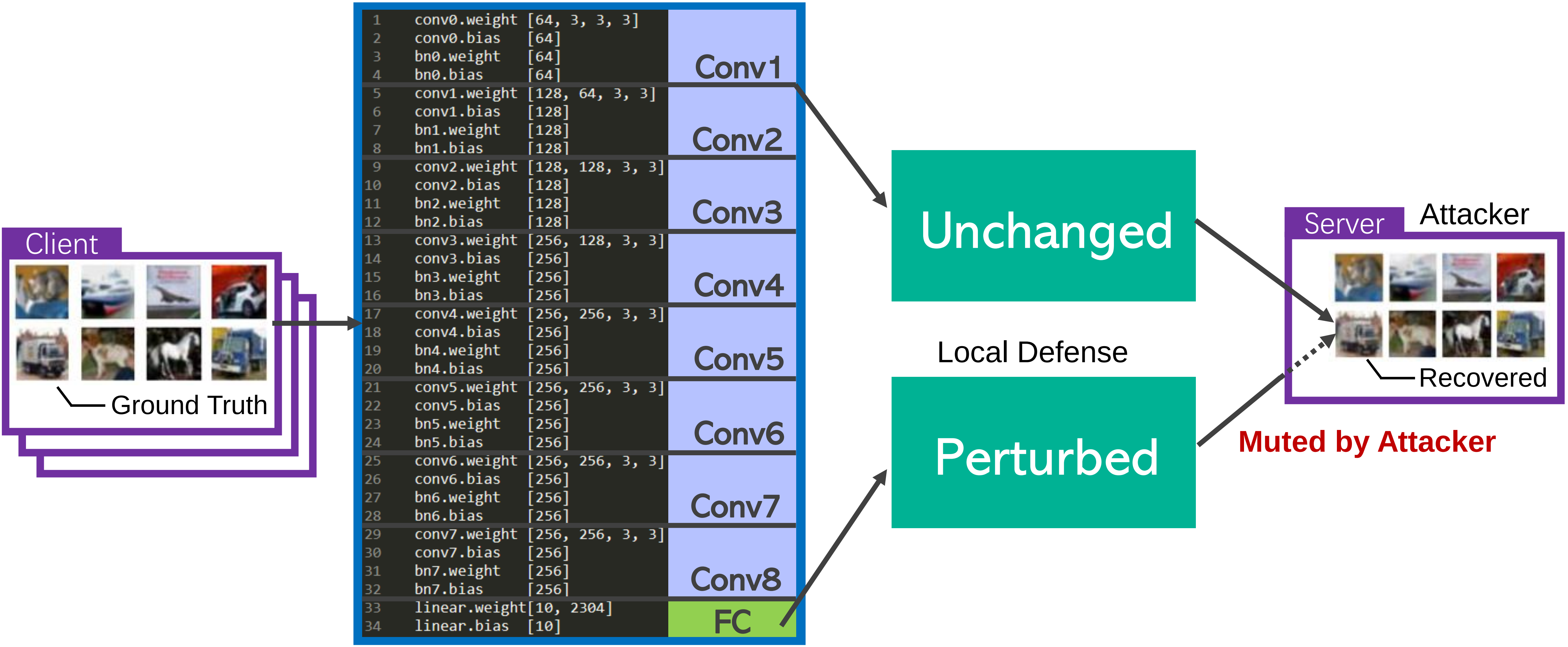

- Advantages: It only Perturbs a Certain Single Layer of Local Gradients (e.g., FC Layer).

In order to Lower Perturbation Footprints and Accuracy Loss.

**Question: What's Potential Risk of this Rigid Pattern?**

## Defense Specific to Gradient Leakage Attack

- Limitations: Rigid Pattern is easily broken down once the Perturbed Layer is Muted by the Attacker.



# Targets of Defense against Gradient Leakage

- **Lightweight, Accuracy-Guaranteed, Privacy-Adequate Defense**

  - Lightweight in Overhead (Computation, Storage, Communication)

    - **Cryptographic Methods e.g., HE**, **MPC** are with significant Overhead.

  - Guaranteed in Convergence Accuracy Loss

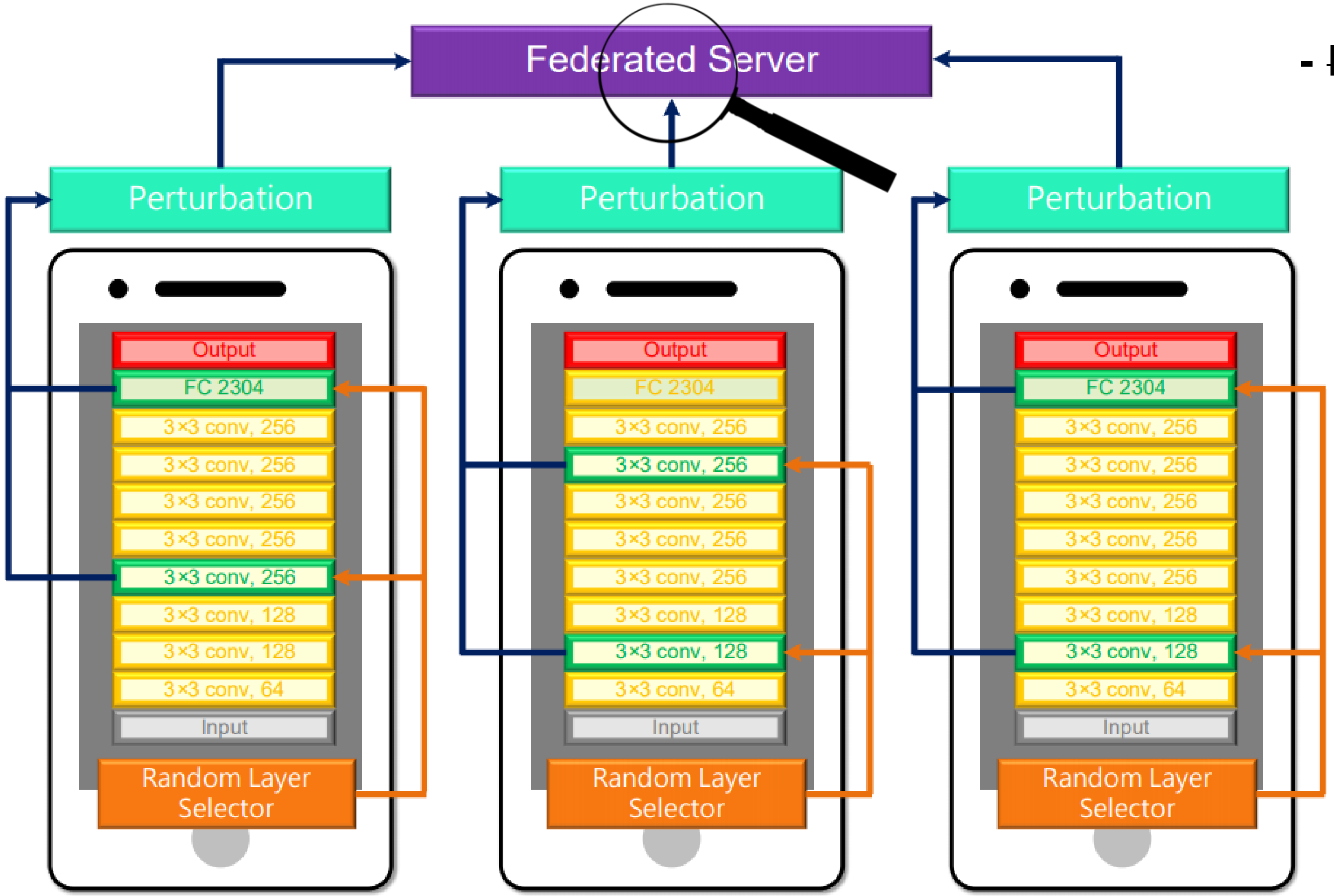
    - **Methods like LDP** are with significant Accuracy Loss.

  - Adequate in Privacy Protection and Hard to Break Down

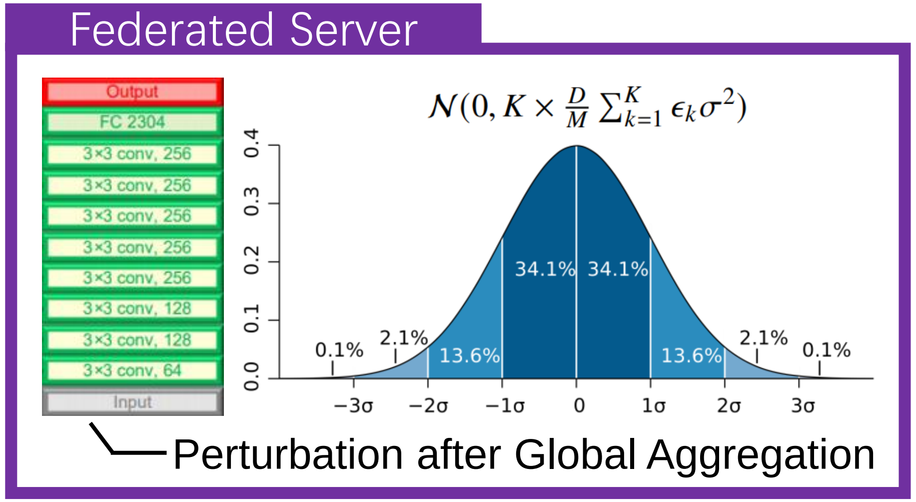
    - **Methods with Rigid Pattern** are easily Inferred and Broken Down.

# **Defense against Gradient Leakage** basic idea

- Inspiration: Each Client Randomly Selects Part of Local Gradients to Perturb



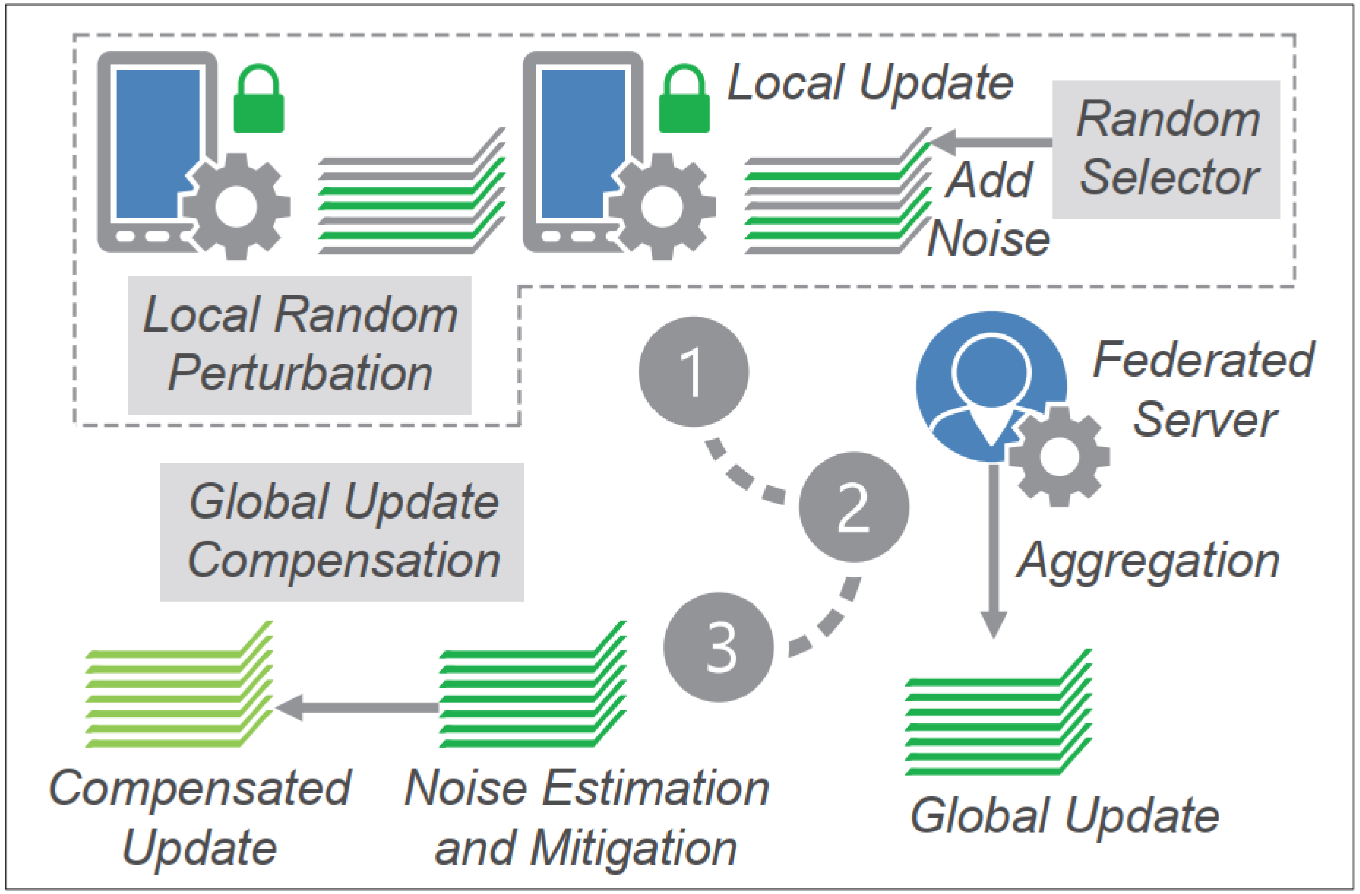

- ~~Rigid Pattern~~ **Random Pattern**
  - **Defense Becomes Hard to Break Down.**
  - **No Significant Overhead.**
  - **Perturbation Can be Compensated.**




J. WANG, S. GUO, et al. “Protect Privacy from Gradient Leakage Attack in Federated Learning,” INFOCOM 2022.

# **Defense against Gradient Leakage** workflow

- The workflow consists of two stages: Local Random Perturbation and Global Update Compensation.




- **Local Random Perturbation**

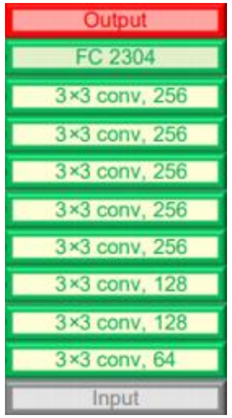
  - Randomly select a certain part of slices from local gradients and add artificial noise to these selected slices.

- **Global Update Compensation**

  - Derive from the perturbed gradients, more accurate information about the original gradients as a compensation for the global update.

# **Defense against Gradient Leakage** more considerations

- Privacy Leakage Risk Evaluation and Gradient Slicing




- Cons: <u>Different layers have different risks of privacy leakage.</u>

Each Slice of Gradients has Balanced Privacy Protection

(a) Random Perturbation is based on Gradient's Logical Layers <u>e.g., Convolutional Layer (Conv) or Fully-Connected Layer (FC).</u>

(b) Random Perturbation is based on Gradient's Slices <u>where Each Slice has Equivalent Defense.</u>

- Prevent Global Compensation from Being Abused by Attacker
  - [**Optional**]: <u>Local Clipping Operation</u>
    (Clipping Selected Gradients and Scaling them to similar range corresponding to the Scale of Perturbation)

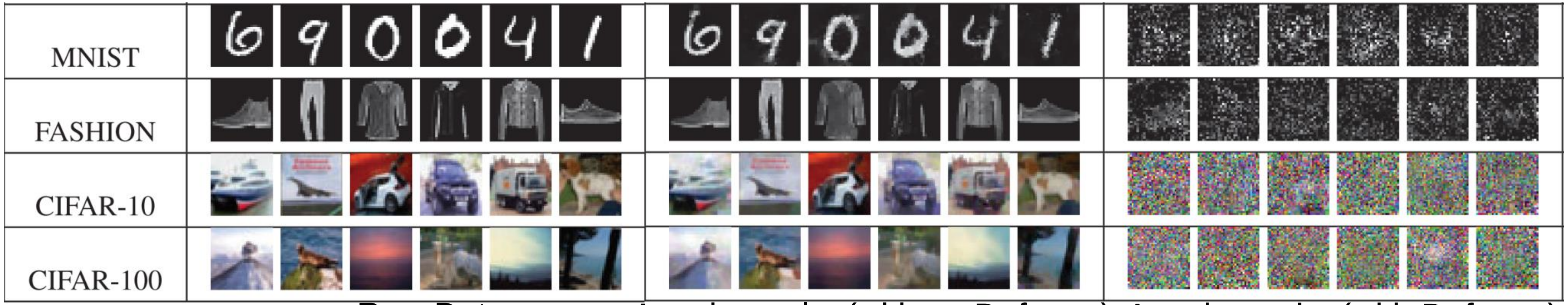
    \- <u>Global Compensation is still Valid</u>.

# Experimental Settings

- Attack Methods
  - [1] DGA, Deep Leakage from Gradients, NeurIPS2019.
  - [2] GIA, Inverting Gradients, NeurIPS2020.
- Baseline Defense Methods
  - [1] GC, Gradient Compression.
  - [2] DP, Differential Privacy, DP-Gaussian and DP-Laplacian.
  - [3] PLD, Provable Defense against Privacy Leakage in Federated Learning, CVPR2021.
- Cared Metrics
  - [1] Attack Reconstruction Quality (Image Similarities).
    - Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM).
  - [2] Accuracy (ACC) of Global Model on the Testing Set.
  - [3] Average Round Time (ART) of Training.
- Datasets and Model
  - MNIST, Fashion-MNIST, CIFAR, Convolutional Networks (LeNet)

# Experimental Results

- Privacy Protection Perspective




(a) Visualization of Privacy Protection Results.

[A] Measure on Different Defenses against the DGA.

| | MNIST - ACC 91.69% without defenses | | | | Fashion-MNIST - ACC 91.80% without defenses | | | | CIFAR-10 - ACC 54.15% without defenses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] |
| PSNR | **9.41** | 9.52 | 9.36[9.39] | 9.57[18.49] | **9.66** | 9.83 | 9.57[9.62] | 9.89[19.78] | **9.61** | 9.79 | 9.55[9.52] | 9.88[24.48] |
| SSIM | **4.6E-2** | 5.1E-2 | 4.1E-2[4.3E-2] | 5.3E-2[6.4E-1] | **7.3E-2** | 7.7E-2 | 7.1E-2[6.5E-2] | 8.2E-2[8.4E-1] | **2.5E-2** | 2.6E-2 | 2.3E-2[2.4E-2] | 2.9E-2[8.8E-1] |

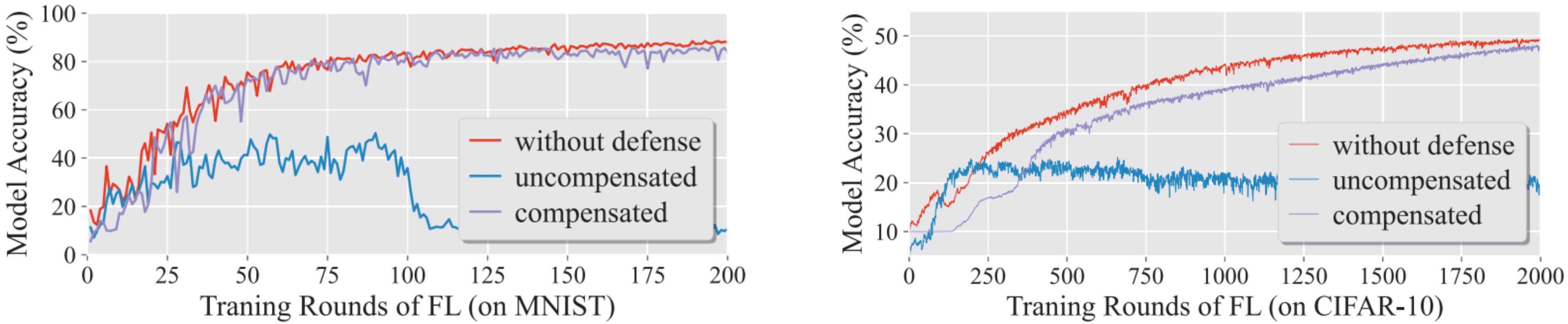
[B] Measure on Different Defenses against the GIA.

| | MNIST - ACC 88.14% without defenses | | | | Fashion-MNIST - ACC 86.57% without defenses | | | | CIFAR-10 - ACC 49.31% without defenses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] |
| PSNR | **9.83** | 10.01 | 9.66[9.59] | 10.43[19.61] | **9.91** | 9.98 | 9.74[9.80] | 10.14[21.23] | **10.11** | 10.32 | 9.95[9.86] | 10.79[27.04] |
| SSIM | **4.9E-2** | 5.1E-2 | 4.4E-2[4.6E-2] | 5.7E-2[7.3E-1] | **7.5E-2** | 8.3E-2 | 6.8E-2[6.7E-2] | 8.9E-2[9.5E-1] | **4.1E-2** | 4.2E-2 | 3.0E-2[3.4E-2] | 4.4E-2[9.3E-1] |

(b) Numerical Results of Privacy Protection (PSNR, SSIM).

# Experimental Results

- Convergence Accuracy Perspective




(a) Visualization of Convergence Accuracy Results.

- Overhead Perspective

[A] Measure on Different Defenses against the DGA.

| | MNIST - ACC 91.69% without defenses | | | | Fashion-MNIST - ACC 91.80% without defenses | | | | CIFAR-10 - ACC 54.15% without defenses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] |
| ACC | **90.43%** | 36.52% | 10.37%[10.21%] | 87.77%[-] | **89.29%** | 33.11% | 10.10%[9.98%] | 86.35%[-] | **52.47%** | 29.84% | 10.19%[10.00%] | 49.91%[-] |
| ART | **+8.45%** | +4.63% | +3.91%[3.74%] | +14.52%[-] | **+8.11%** | +3.75% | +3.89%[4.04%] | +13.20%[-] | **+8.97%** | +3.58% | +4.03%[4.31%] | +14.09%[-] |

[B] Measure on Different Defenses against the GIA.

| | MNIST - ACC 88.14% without defenses | | | | Fashion-MNIST - ACC 86.57% without defenses | | | | CIFAR-10 - ACC 49.31% without defenses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] | **Ours** | GC | DP-G[-L] | PLD[-muted] |
| ACC | **86.87%** | 32.29% | 10.46%[9.85%] | 84.09%[-] | **84.65%** | 30.38% | 9.86%[9.77%] | 81.10%[-] | **47.73%** | 23.35% | 10.01%[10.16%] | 45.16%[-] |
| ART | **+9.07%** | +4.90% | +3.84%[3.66%] | +16.12%[-] | **+8.62%** | +4.23% | +4.14%[3.99%] | +15.86%[-] | **+9.33%** | +4.08% | +4.15%[4.02%] | +16.43%[-] |

(b) Numerical Results of Accuracy (ACC) and Average Round Time (ART).

# Part 3.

# Machine Unlearning in Federated Learning

**Identify what's the federated unlearning how we can achieve that efficiently**

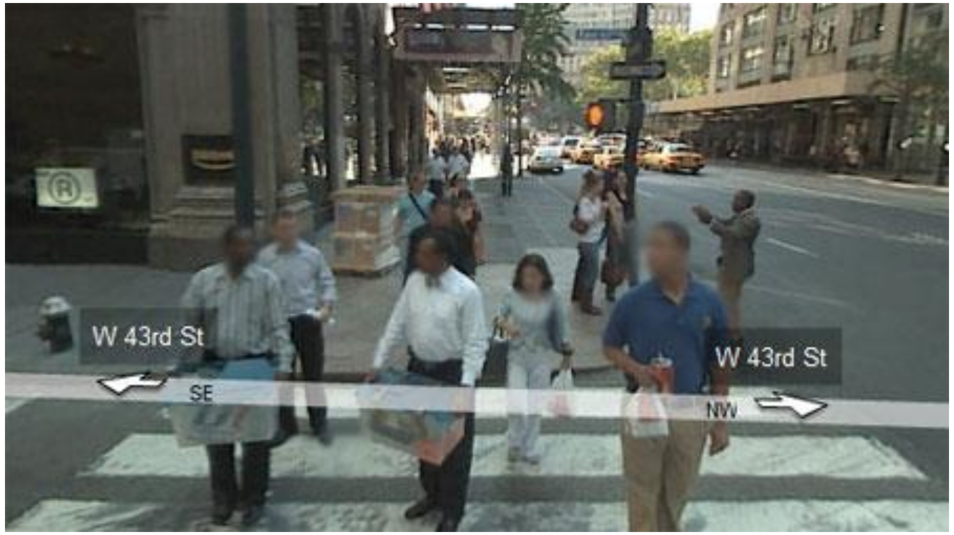
# What's Machine Unlearning


- **Users have the right to Unlearn Sensitive Data from Trained ML Models.**

➢ What's the Class Unlearning ?

↓

**A specific class of data** needs to be removed from Trained ML Model.




Street View Images with Facial

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

# Class Unlearning and Approximate: Challenges in FL

## ➢ General Machine Unlearning – Retraining from Scratch

**Advantage:** Determined to be effective and convincing.
**Disadvantage:** Computational and time overhead associated with fully retraining models affected by training data erasure can be prohibitively expensive.

## ➢ Existing Approximate Machine Unlearning

They require global access to training data.

- The number of participant devices is usually much smaller than the total devices.
- The non-IID training data across different participants

↓

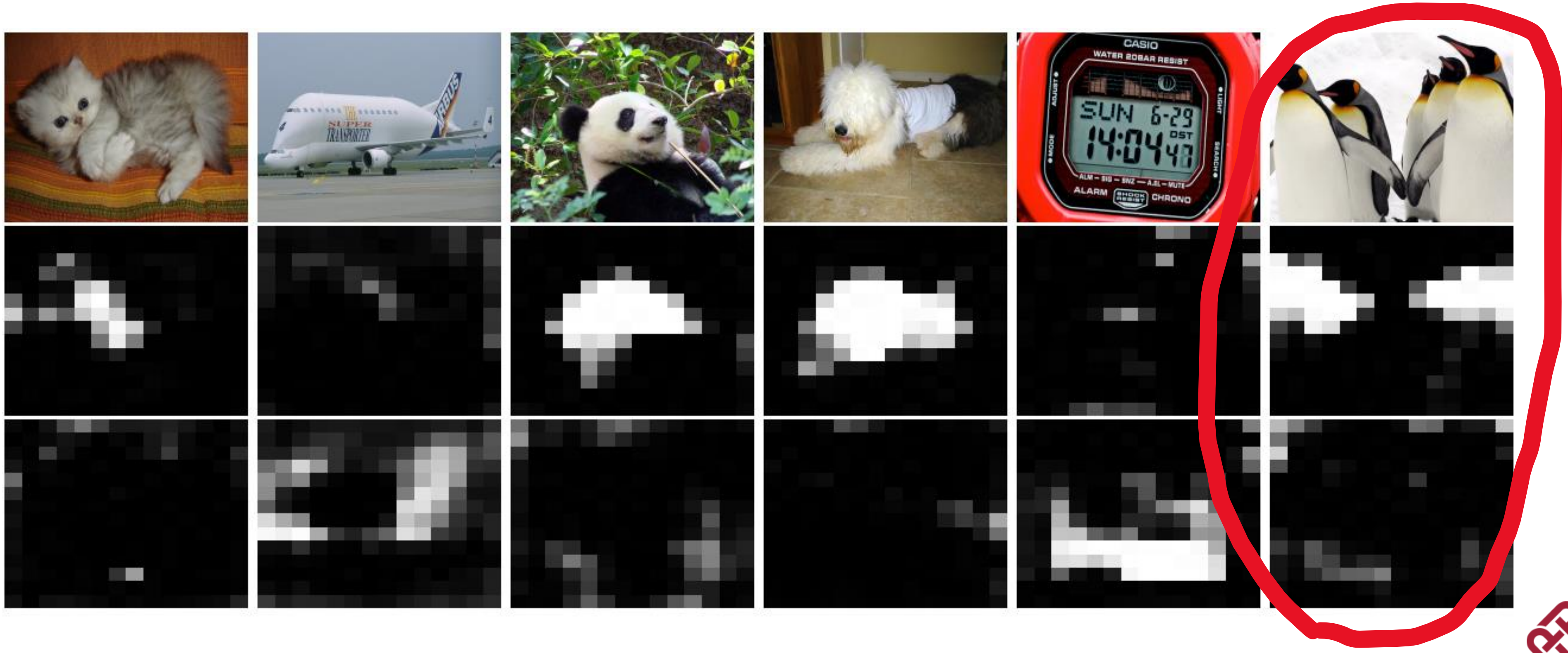
Incomplete and severely biased local training data, the existing approximate unlearning method can only offer inaccurate model approximations.

# Class Machine Unlearning

➢ Feature Maps vs. Raw Data (in term of Class Discrimination)

- High-level feature maps contain <u>more information about the class</u>.
- Incomplete and biased training data in the same class <u>share similar high-level features</u>.

Class discrimination of feature maps can be learned through a small set of collaboration.



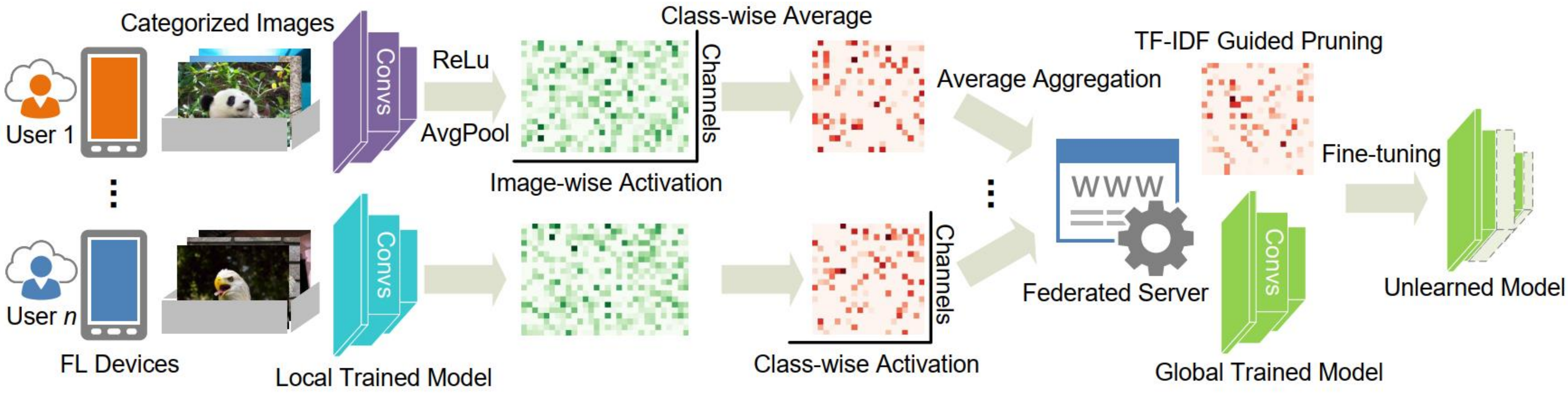

Class unlearning
-> <u>Channel pruning</u>

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

# Class Machine Unlearning

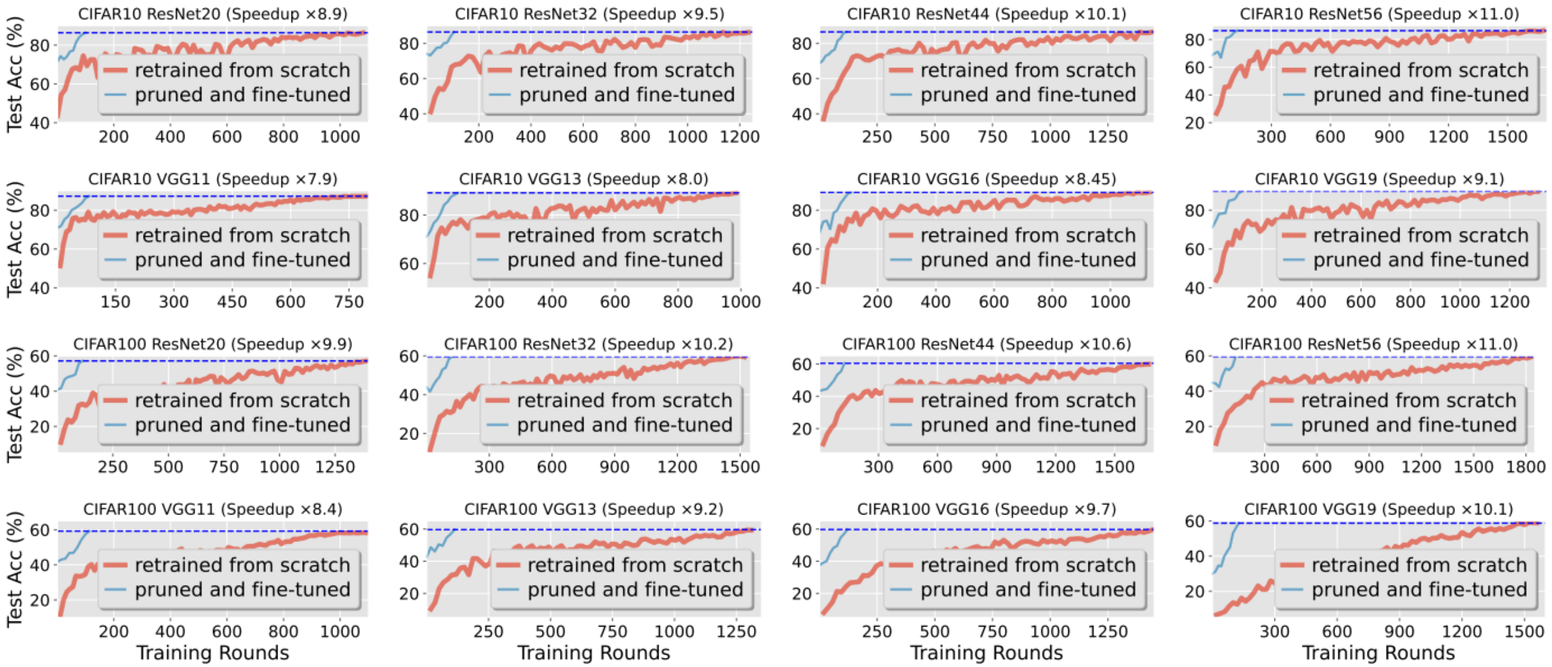
- Workflow of Class Unlearning in FL

- Participant clients transform their private images to generate local representations.
- Server builds a pruner based on the relationship between the target class and channels.
- TF-IDF (widely used in NLP) now is used by to compute the relationship between the target class and channels. Straightforward but efficient.




J. WANG, S. GUO, et al. "Federated Unlearning via Class-Discriminative Pruning," WWW 2022.

# Experimental Results




- <u>Speedup</u> is significant compared to retraining.
- <u>Information erasure</u> is the same to retraining.

| | CIFAR10 | | | | | | | | CIFAR100 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Raw model | | After-pruned | | Fine-tuned | | Re-trained | | Raw model | | After-pruned | | Fine-tuned | | Re-trained | |
| Accuracy | U-set | R-set | U-set | R-set | U-set | R-set | U-set | R-set | U-set | R-set | U-set | R-set | U-set | R-set | U-set | R-set |
| ResNet20 | 91.50% | 83.33% | 00.00% | 20.79% | 00.00% | 86.40% | 00.00% | 86.33% | 54.00% | 50.01% | 00.00% | 05.38% | 00.00% | 57.17% | 00.00% | 57.11% |
| ResNet32 | 94.20% | 83.71% | 00.00% | 11.58% | 00.00% | 86.40% | 00.00% | 86.14% | 52.00% | 51.67% | 00.00% | 01.06% | 00.00% | 59.62% | 00.00% | 59.42% |
| ResNet44 | 89.90% | 83.94% | 00.00% | 22.19% | 00.00% | 86.48% | 00.00% | 86.34% | 48.00% | 53.25% | 00.00% | 01.22% | 00.00% | 60.41% | 00.00% | 59.85% |
| ResNet56 | 93.10% | 84.02% | 00.00% | 11.11% | 00.00% | 86.42% | 00.00% | 86.38% | 44.00% | 52.91% | 00.00% | 01.32% | 00.00% | 59.60% | 00.00% | 59.28% |
| VGG11 | 88.20% | 84.72% | 00.00% | 18.29% | 00.00% | 87.24% | 00.00% | 87.13% | 50.00% | 53.55% | 00.00% | 01.28% | 00.00% | 59.25% | 00.00% | 58.20% |
| VGG13 | 91.50% | 84.19% | 00.00% | 15.17% | 00.00% | 89.18% | 00.00% | 89.09% | 60.00% | 51.88% | 00.00% | 03.82% | 00.00% | 59.65% | 00.00% | 59.27% |
| VGG16 | 91.60% | 84.38% | 00.00% | 17.79% | 00.00% | 89.20% | 00.00% | 89.30% | 44.00% | 50.34% | 00.00% | 01.46% | 00.00% | 59.72% | 00.00% | 59.57% |
| VGG19 | 88.80% | 83.53% | 00.00% | 11.11% | 00.00% | 89.72% | 00.00% | 89.62% | 52.00% | 52.15% | 00.00% | 01.02% | 00.00% | 58.78% | 00.00% | 58.96% |

# ***Thank you!***

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學