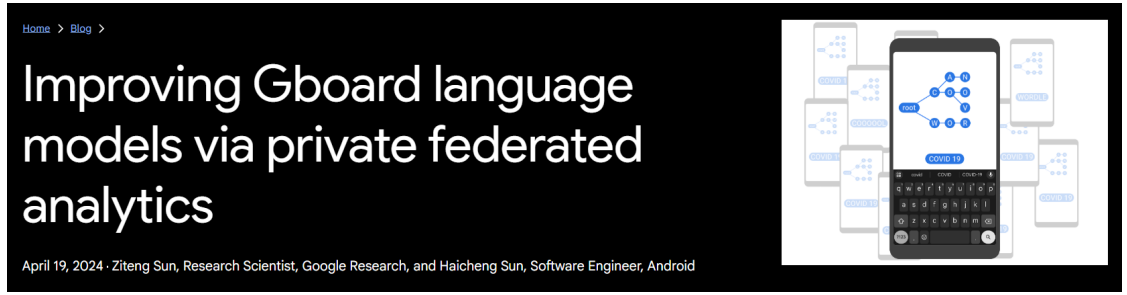
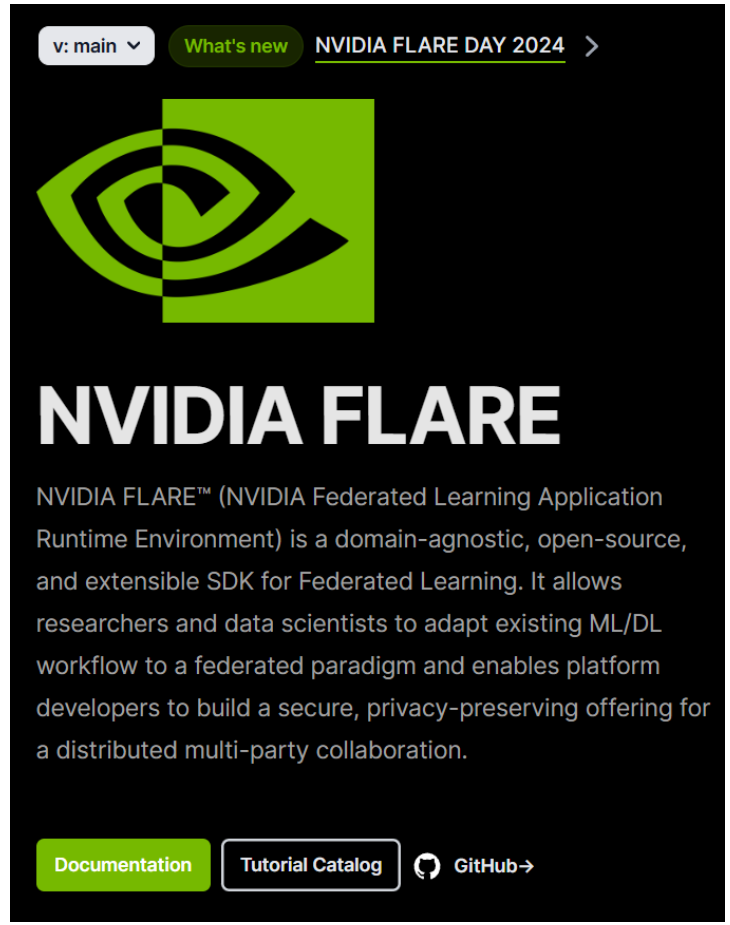
MedLeak: Multimodal Medical Data Leakage in Secure Federated Learning with Crafted Models

Shanghao Shi, Md Shahedul Haque, Abhijeet Parida,
Chaoyu Zhang, Marius George Linguraru, Y. Thomas Hou,
Syed Muhammad Anwar, and Wenjing Lou

Federated Learning Applications



Gboard



Mobile Apps
Computing Platforms
Healthcare
…

Rhino Health Platform Powers Hospital-Based Federated Learning Consortium

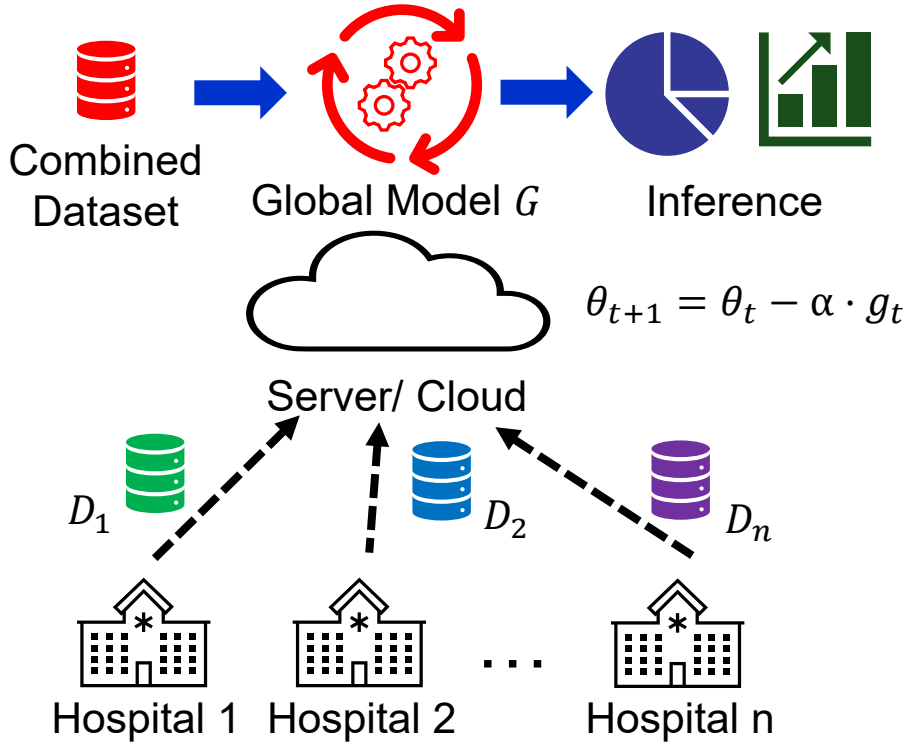
Healthcare Institutions Around the Globe Collaborate with Disparate Data Securely to Transform Healthcare AI Development and Clinical Translation

May 05, 2022 09:00 ET | Source: Rhino Health

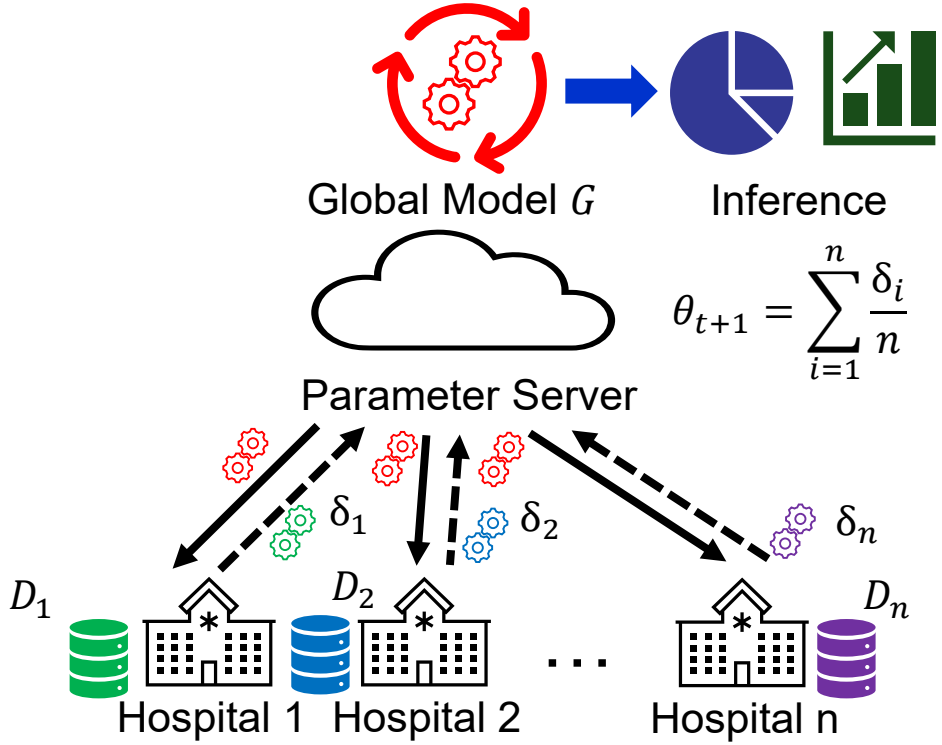
Follow

VIRGINIA TECH

Centralized v.s. Federated Learning

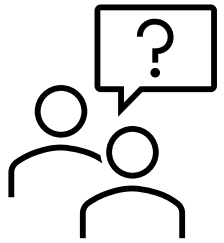


- Centralized Learning
  - Participants ***share*** data with the server.



- Federated Learning
  - Participants collaboratively train models.
  - ***Participants' data remains local***.
  - ***Only model updates are shared.***

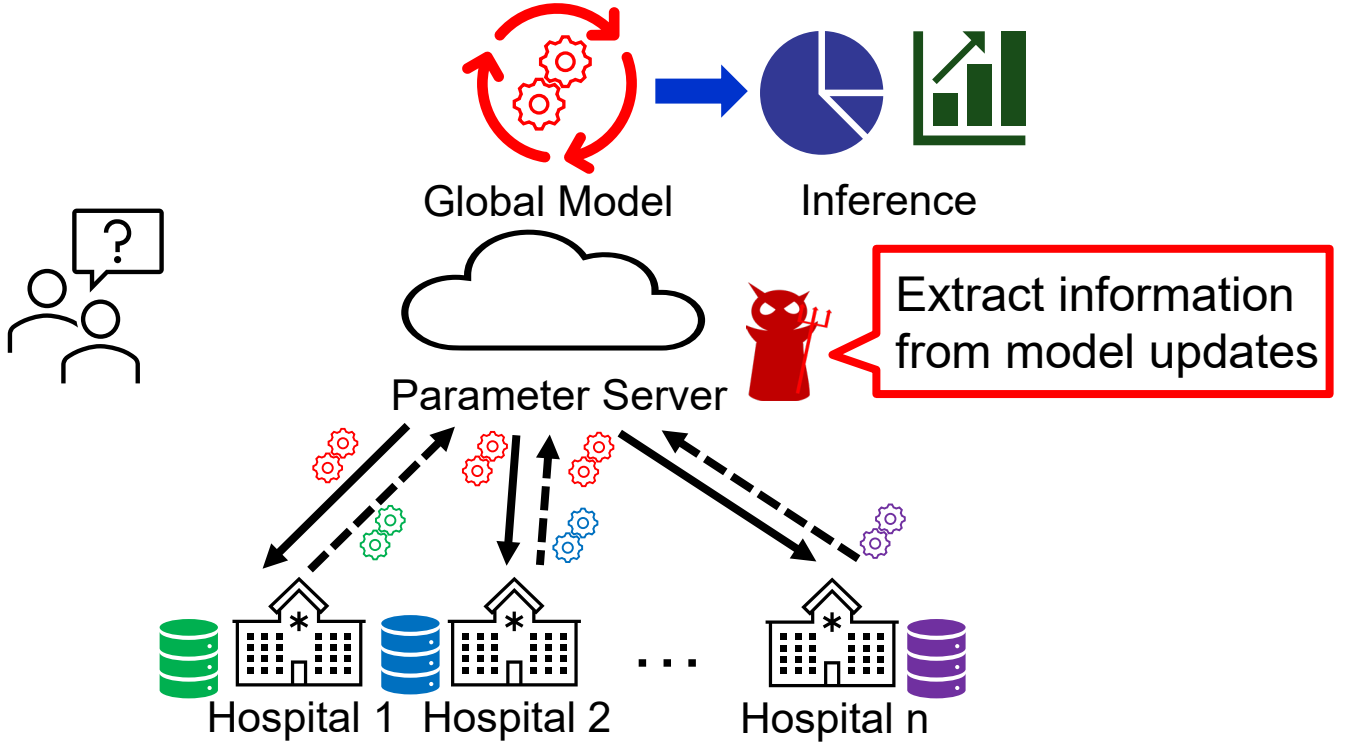
Privacy Concern for Federated Learning



Can federated learning actually preserve data privacy?

Privacy Attacks in Federated Learning

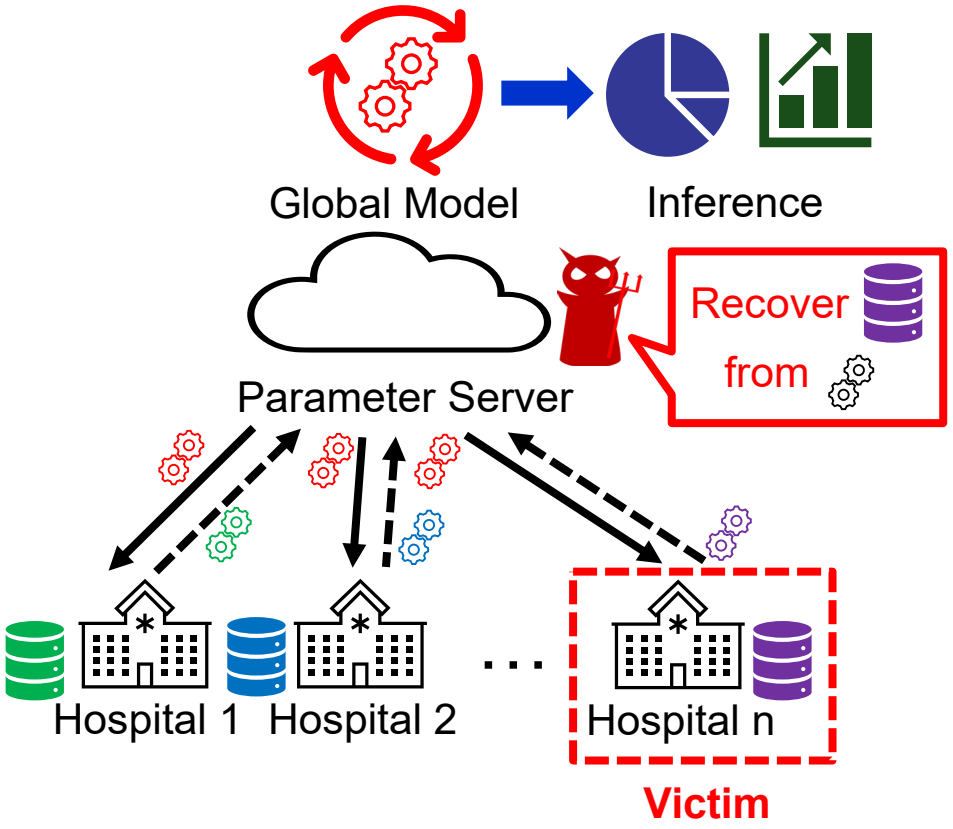
- Privacy Leakage Possibilities:
  - The **global model** G and **individual model updates** δ_i are exposed.
  - Can the attacker retrieve any meaningful information from them?
- State-of-the-art Privacy Attacks:
  - Membership Inference Attack
  - **Model Inversion Attack**
- State-of-the-art Defenses:
  - Differential Privacy
  - **Secure Aggregation**



- Federated Learning
  - Participants collaboratively train models.
  - Participant data remains ***local***.
  - Only model updates are shared.

Model Inversion Attack

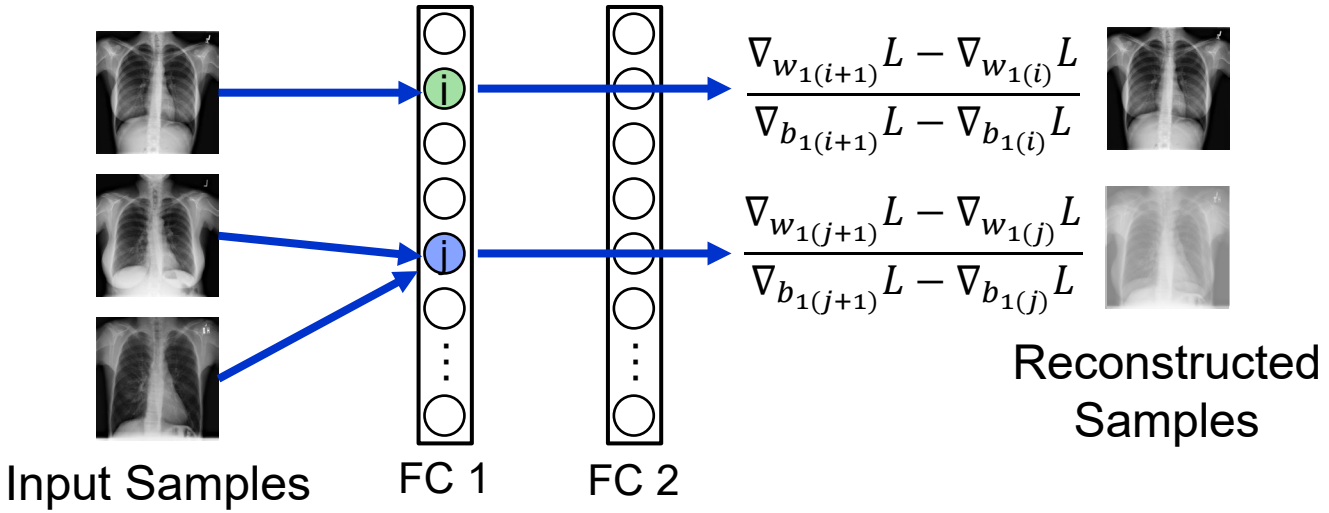
- Launched by the **parameter server.**
- Protected by **Secure Aggregation**:
  - The individual model update δ_i is hidden.
- Attacker's Knowledge:
  - Global model G
  - **Aggregated** local model $\sum_{i=1}^{n} \delta_i$
- Goal: Reverse aggregated model update $\sum_{i=1}^{n} \delta_i$ back to local samples D_i.
  - $D_i = Reverse(\sum_{i=1}^{n} \delta_i)$



Linear Leakage

- "Linear leakage" can reconstruct its inputs from gradients [5].
  - Need knowledge about the target's data distribution φ.
    - Can be estimated by an auxiliary dataset D_{aux}.
  - Craft a two-layer attack module according to φ.
  - Each row vector (neuron) can reconstruct one sample.

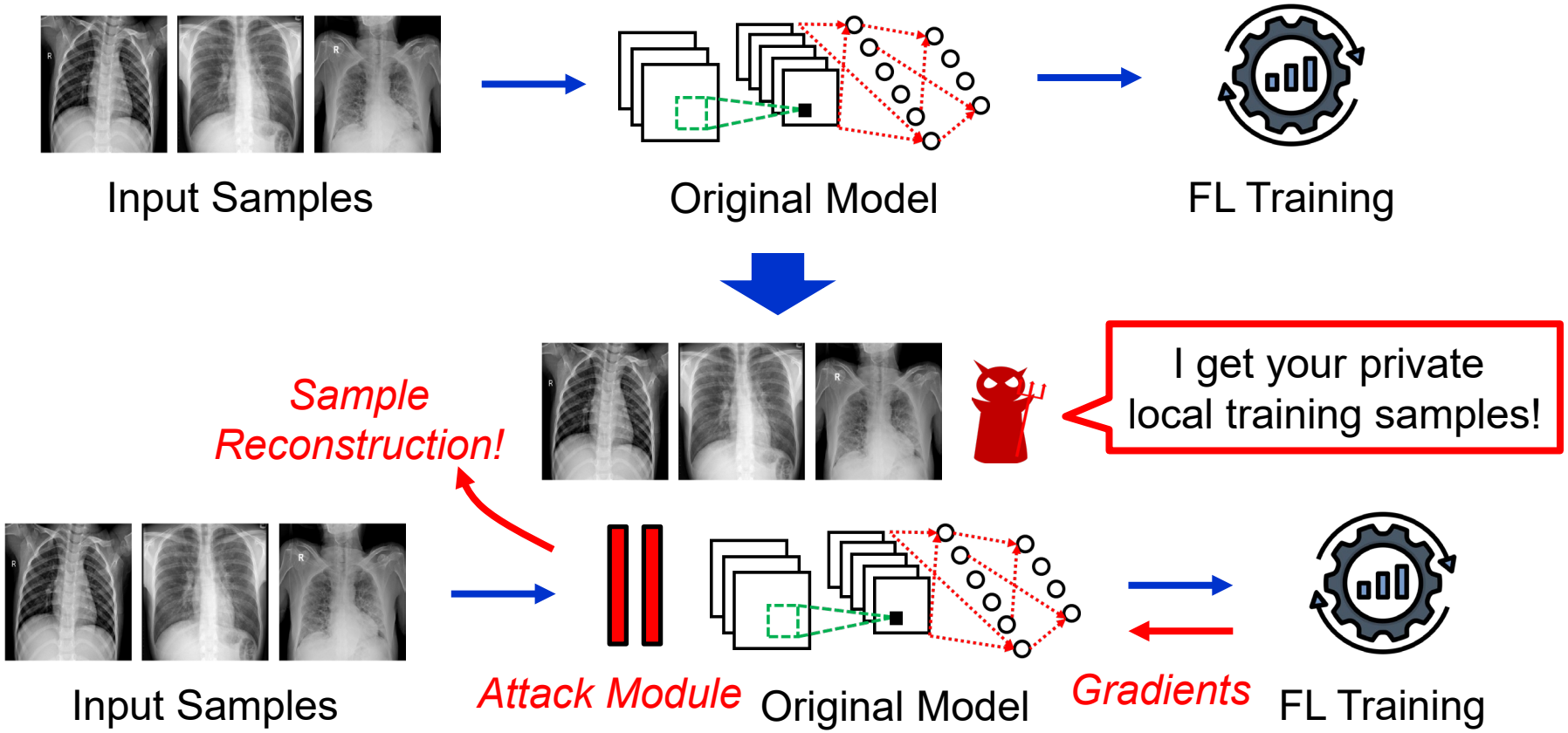
One neuron
One reconstruction!



[1] Fowl, Liam, Jonas Geiping, Wojtek Czaja, Micah Goldblum, and Tom Goldstein. "Robbing the fed: Directly obtaining private data in federated learning with modified models." arXiv preprint arXiv:2110.13057 (2021).

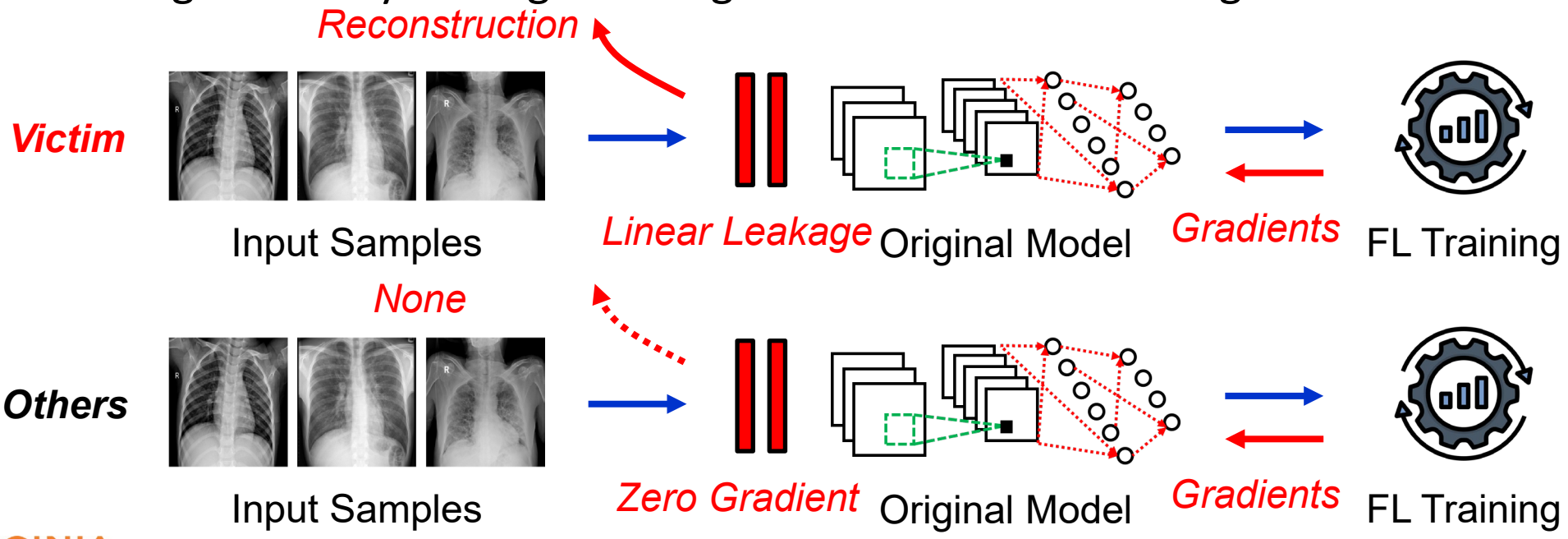
Attack via Model Crafting

- Craft a two-linear layer attack module.
- Insert the attack module in front of the original architecture.



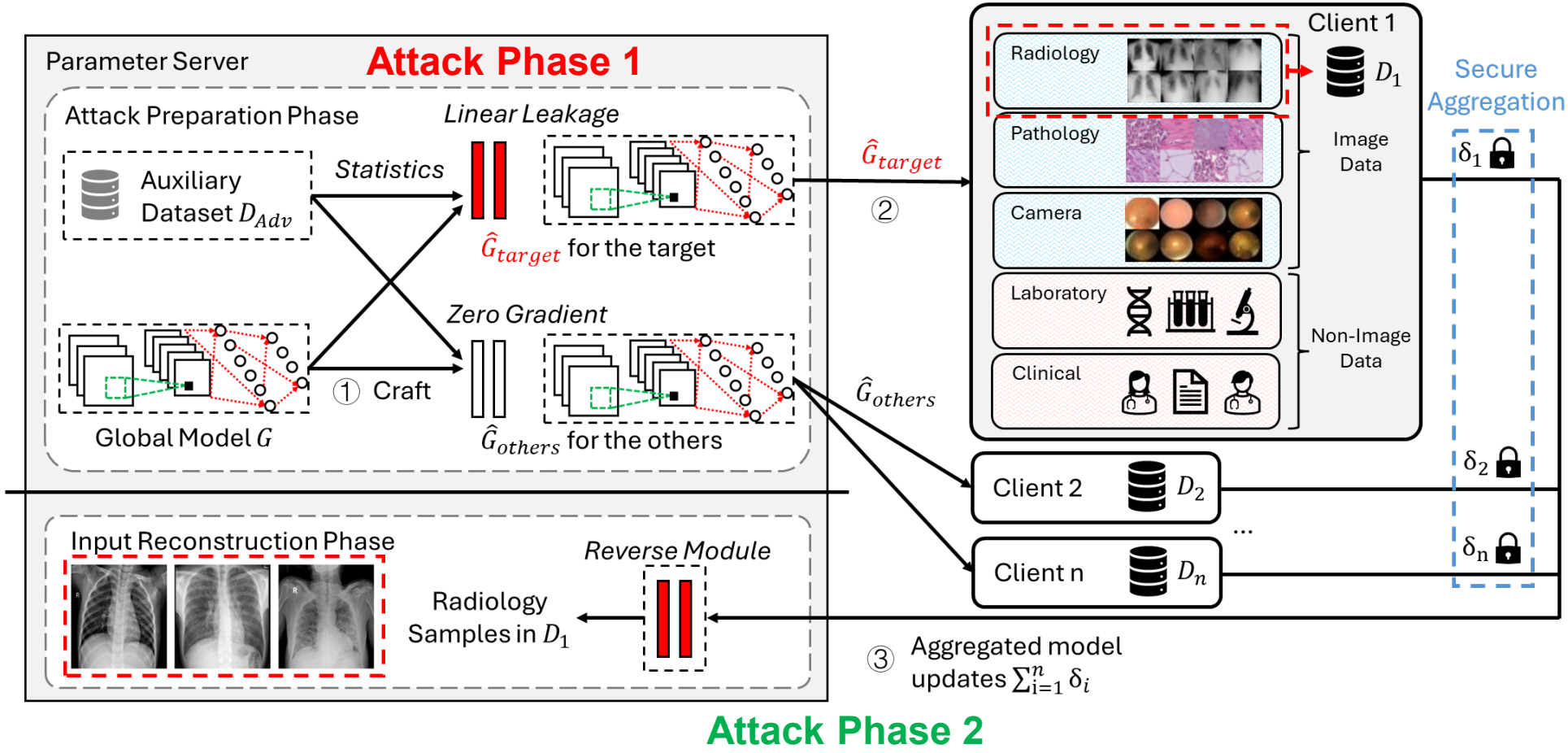
Separate the Victim from Others

- Can we achieve targeted attack? (**Bypass secure aggregation**)
- We use **zero gradient** module to separate the victim from others.
  - $ReLU(x) = \begin{cases} x, when\ x \geq 0 \\ 0, when\ x < 0 \end{cases}$, and its gradient satifies $ReLU'(x) = \begin{cases} 1, when\ x \geq 0 \\ 0, when\ x < 0 \end{cases}$.
  - Zero out gradient by forcing the weight W and bias b to be negative.



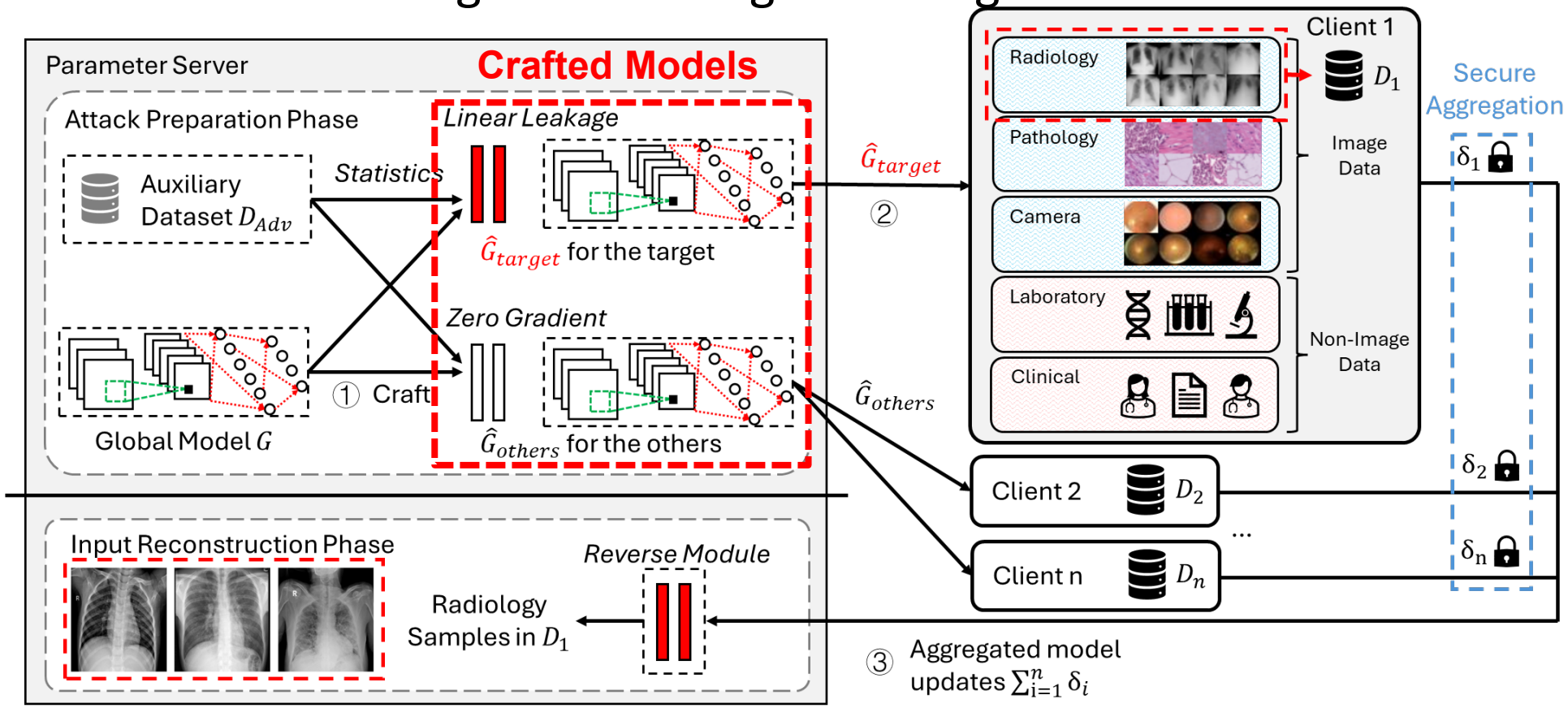
Attack Flow

- MedLeak is a **two-phase** attack: including the attack preparation and input reconstruction.



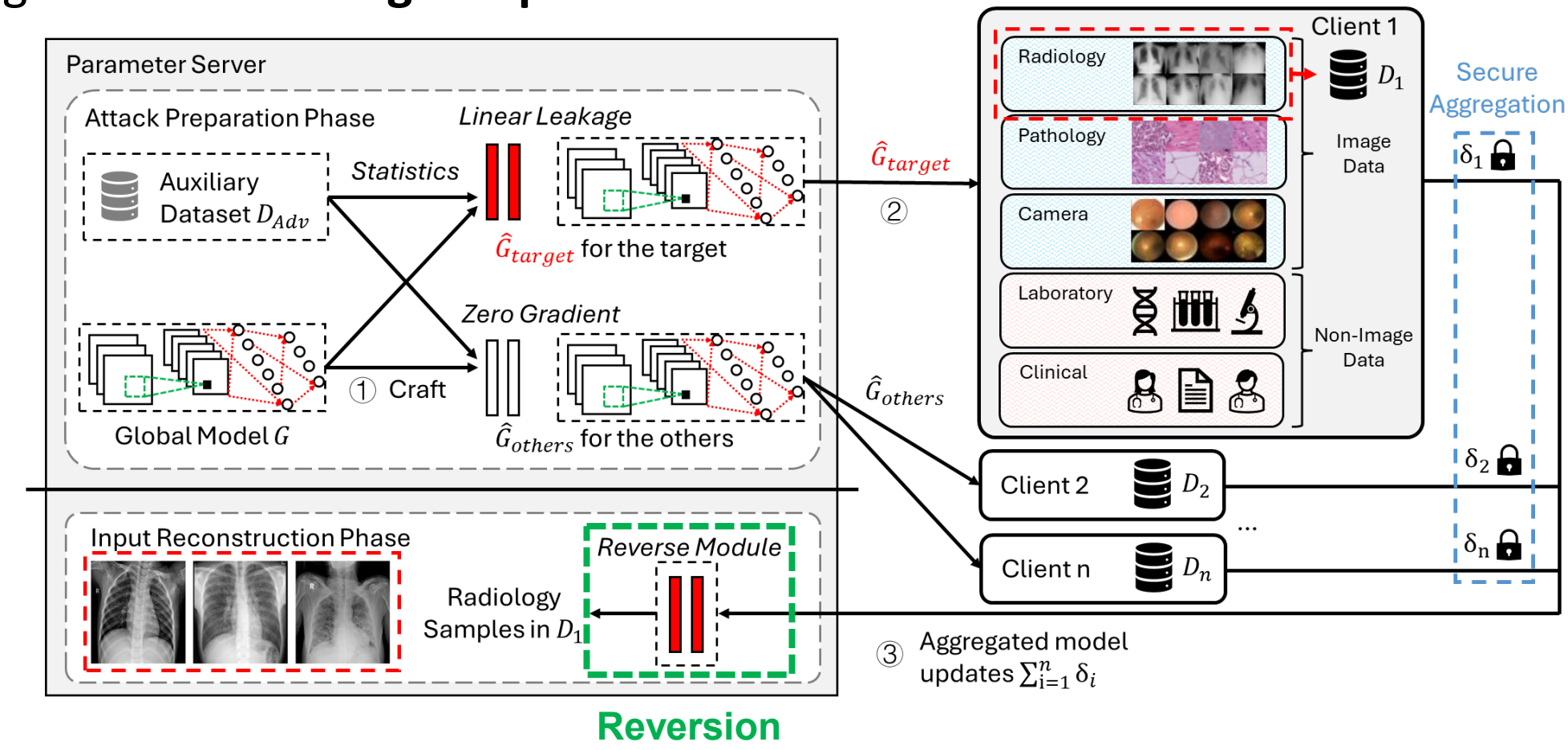
Attack Preparation

- The first attack phase is conducted offline by the server to **craft the adversarial attack modules** including linear leakage & zero gradient.



Input Reconstruction

- In the second phase, the server reverse the **aggregated model updates** back the target's **local training samples**.



Implementation

- Experiment Settings
  - **Medical Image**: COVIDx CXR-4, Kaggle Brain Tumor MRI, and MedMNIST datasets.
  - **Medical Text**: MedAbstract dataset.
  - FL client number: **5 to 30**.
  - Local training: **1 to 5** rounds.
- Evaluation metrics
  - Peak signal-to-noise ratio (**PSNR**) score:
    - $PSNR = 20log_{10}(\frac{Max_I}{\sqrt{MSE}})$
  - Structural similarity index measure (**SSIM**) score:
    - $SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_1)}$
  - Reconstruction successful **rate** and attack **time**.

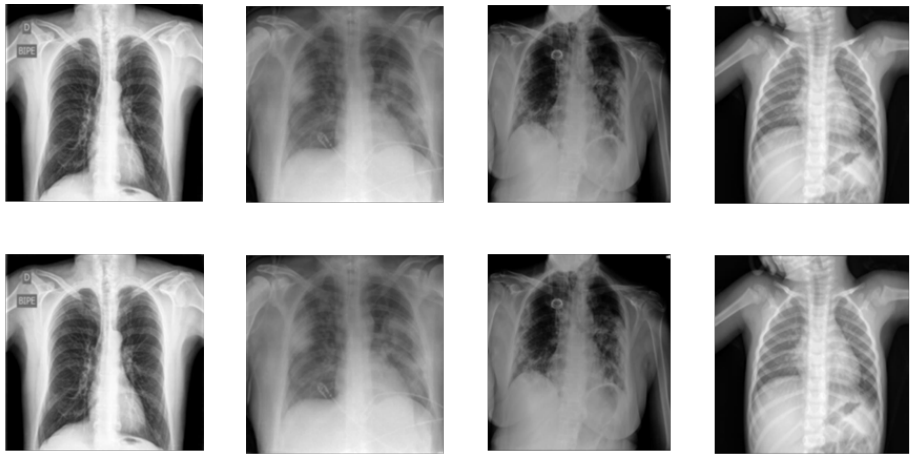
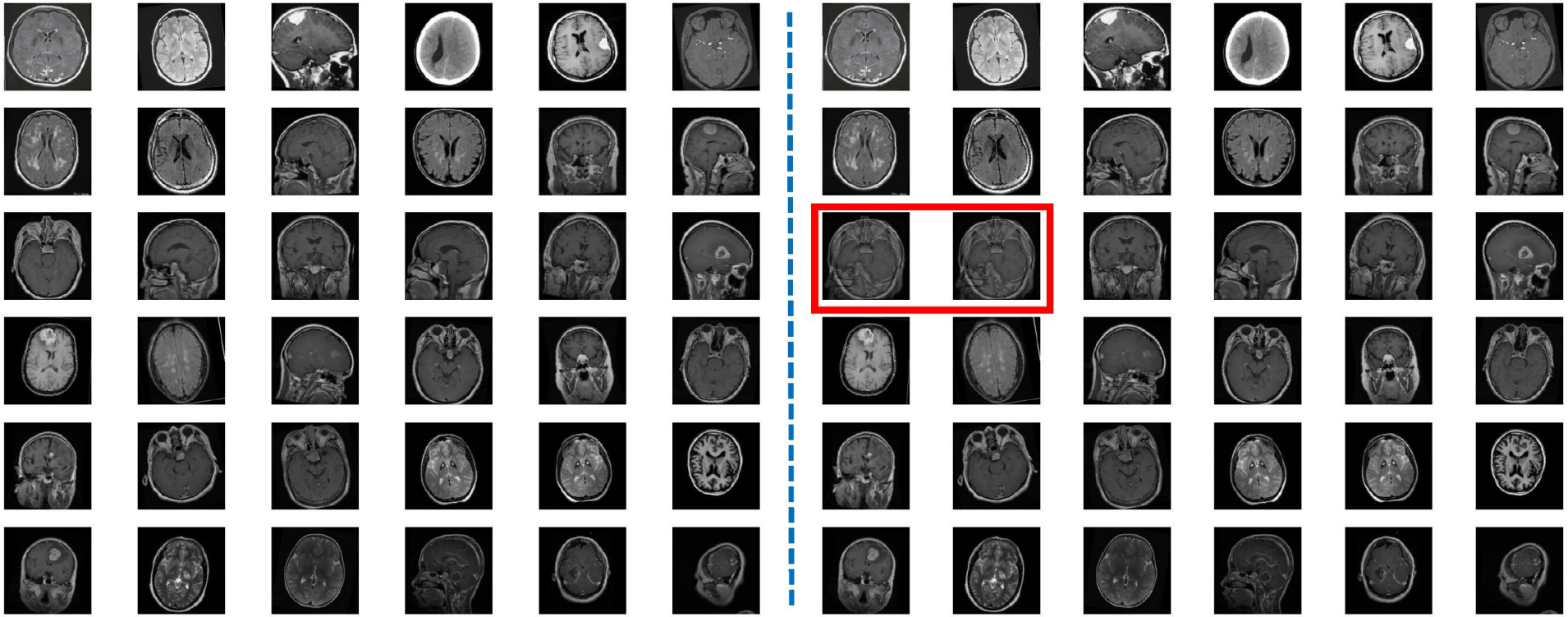


Image Reconstruction Samples

Reconstruction Examples

- We select one batch of 36 images from the Kaggle Brain Tumor MRI dataset.
- **Original images are on the left**, and **reconstructed ones are on the right**.
  - 34 out of 36 images were successfully reconstructed!



Reconstruction Results

- Reconstruction performance on the COVIDx CXR-4 dataset over **different** reconstruction **batch sizes**.
  - Our attack can reconstruct **hundreds of samples** simultaneously with decent reconstruction rates and quantitative scores.
  - Our attack can be accomplished within **a few seconds**.

Batch Size	Dataset	Pixel Size	Rate	PSNR	SSIM	Time (in sec)
100	COVIDx CXR-4	224x224	0.95	121.75	0.96	6.022
200	COVIDx CXR-4	224x224	0.89	112.66	0.99	7.003
300	COVIDx CXR-4	224x224	0.88	105.12	0.96	8.121
400	COVIDx CXR-4	224x224	0.86	97.30	0.99	8.762
500	COVIDx CXR-4	224x224	**0.81**	**95.86**	**0.99**	**9.763**

Medical Downstream Tasks

- We conducted a binary classification task (**COVID-19 detection**) on the **actual** and **recovered images** with a pre-trained ViT-S model.
  - We evaluated the classification performance on common ML evaluation **metrics**.
  - Recovered images achieve nearly **the same performance** as original ones.

Original Images | Reconstructed Images | Binary COVID Classifier

Model	Image	AUPR	TNR	TPR	ACC	AUC
ViT-S (SSL)	original	0.9375	0.80	0.857	0.829	0.905
	recovered	0.9207	0.90	0.719	0.805	0.919
ViT-S (Finetuned)	original	0.9745	0.97	0.931	0.953	0.969
	recovered	0.9653	0.886	0.938	0.912	0.966

Medical Text Recovery

- Medical text data contains a huge amount of **private personal records**.
  - We targeted the **Med Abstract** dataset [6], which consists of 14438 medical abstracts (each has **a few hundred** words) describing the patients' health conditions in **five different classes**.
- Evaluation metrics
  - Word error rate (**WER**)
  - Reconstruction **rate** and attack **time.**

Batch Size	Max Len	Rate	WER	Time (in sec)
100	200	0.7550	0.0047	1.058
	300	0.7585	0.0052	1.514

Infection during chronic epidural catheterization: diagnosis and treatment. A potentially serious complication of long-term epidural catheterization in cancer patients is infection. The early signs of infection were studied in 350 patients in whom long-term epidural catheters were inserted…

Infection during chronic epidural catheterization: diagnosis and treatment. A potentially serious complication of long-term epidural catheterization in cancer patients is infection. The early signs of infection were studied in 350 patients in whom long-term epidural catheters were inserted…

WER:0.0052

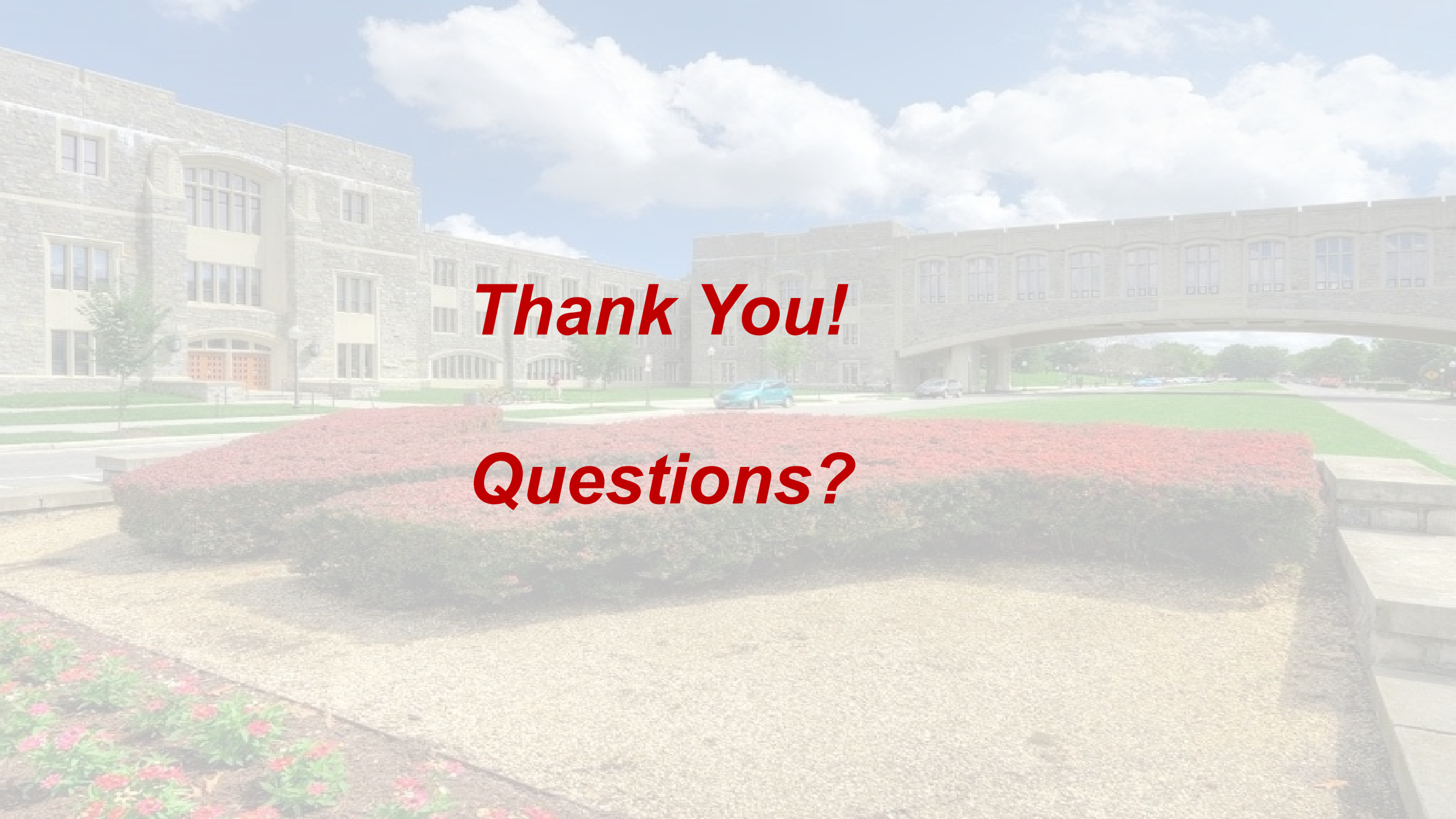
Text Reconstruction Sample

[2] Tim Schopf, Daniel Braun, and Florian Matthes. 2023. Evaluating Unsupervised Text Classification: Zero-Shot and Similarity-Based Approaches. In Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval (Bangkok, Thailand) (NLPIR '22). Association for Computing Machinery, New York, NY, USA, 6–15.

Summary

- MedLeak is a novel model inversion attacks (MIA) that challenge the fundamental **privacy-preserving** property of the FL systems.
  - The attack can **efficiently and accurately** reconstruct site-specific medical images and text records.
  - The existing **secure aggregation** mechanism is **ineffective** against this advanced MIA.

Paper Link


Thank You!
Questions?