Comprehensive Privacy Analysis of Deep Learning:

Passive and Active White-box Inference Attacks against Centralized and Federated Learning

Milad Nasr[1], Reza Shokri[2], Amir Houmansadr[1]

[1]University of Massachusetts Amherst, [2]National University of Singapore

Deep learning Tasks

Medical

Location

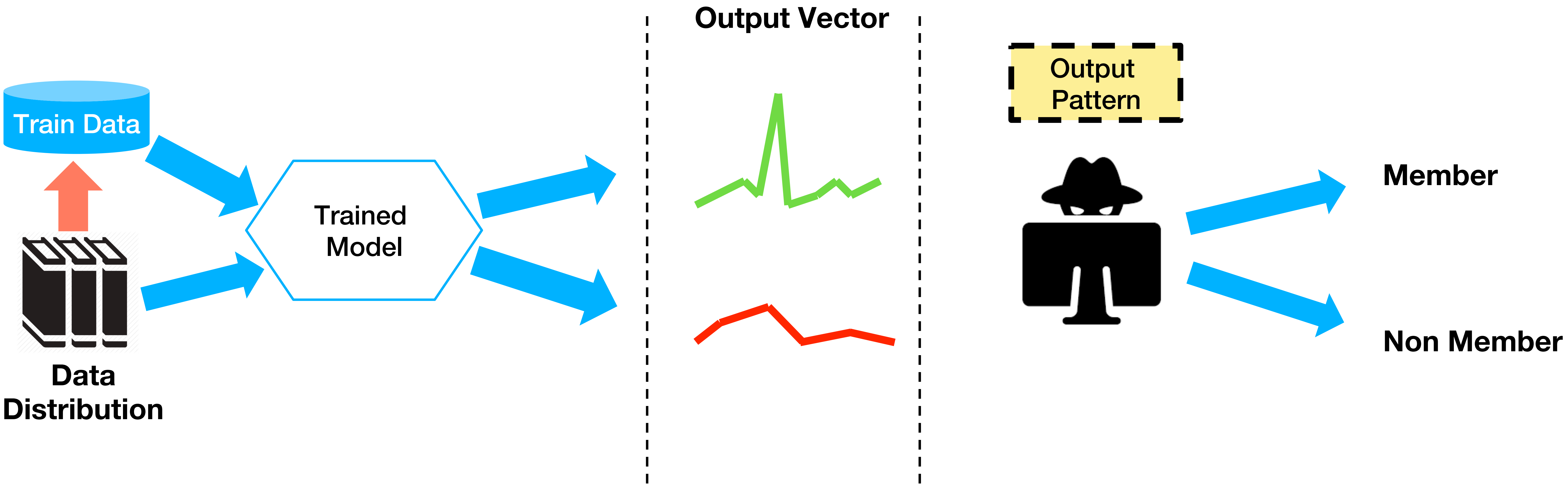
Financial

Personal History

Privacy Threats

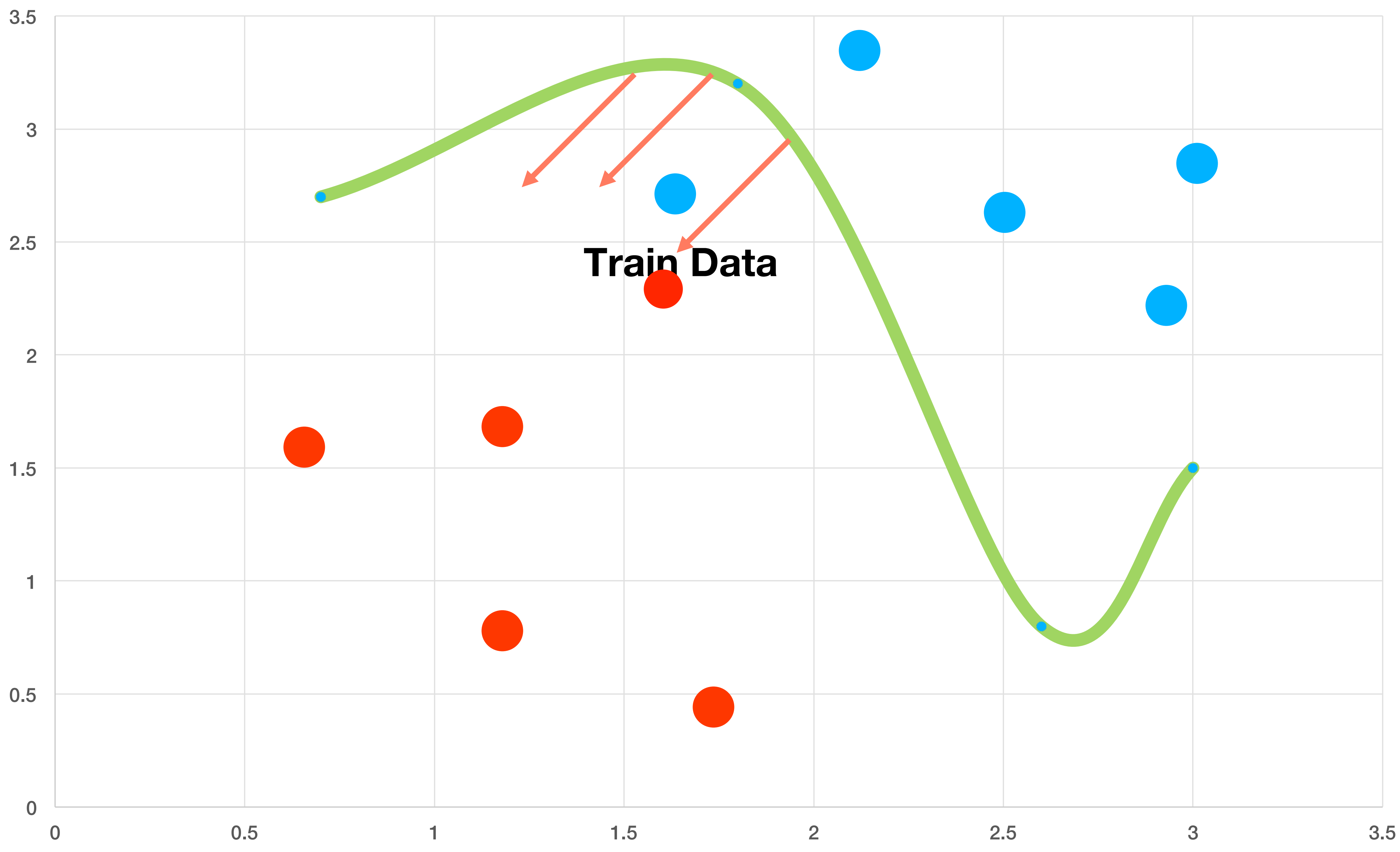
- We provide a comprehensive privacy analysis of deep learning algorithms.
    - Our objective is to measure information leakage of deep learning models about their training data
    - In particular we emphasize on membership inference attacks
    - Can an adversary infer whether or not a particular data record was part of the training set?

Membership Inference



What is the cause of this behavior ?

Training a Model



SGD:

W **Model parameters**

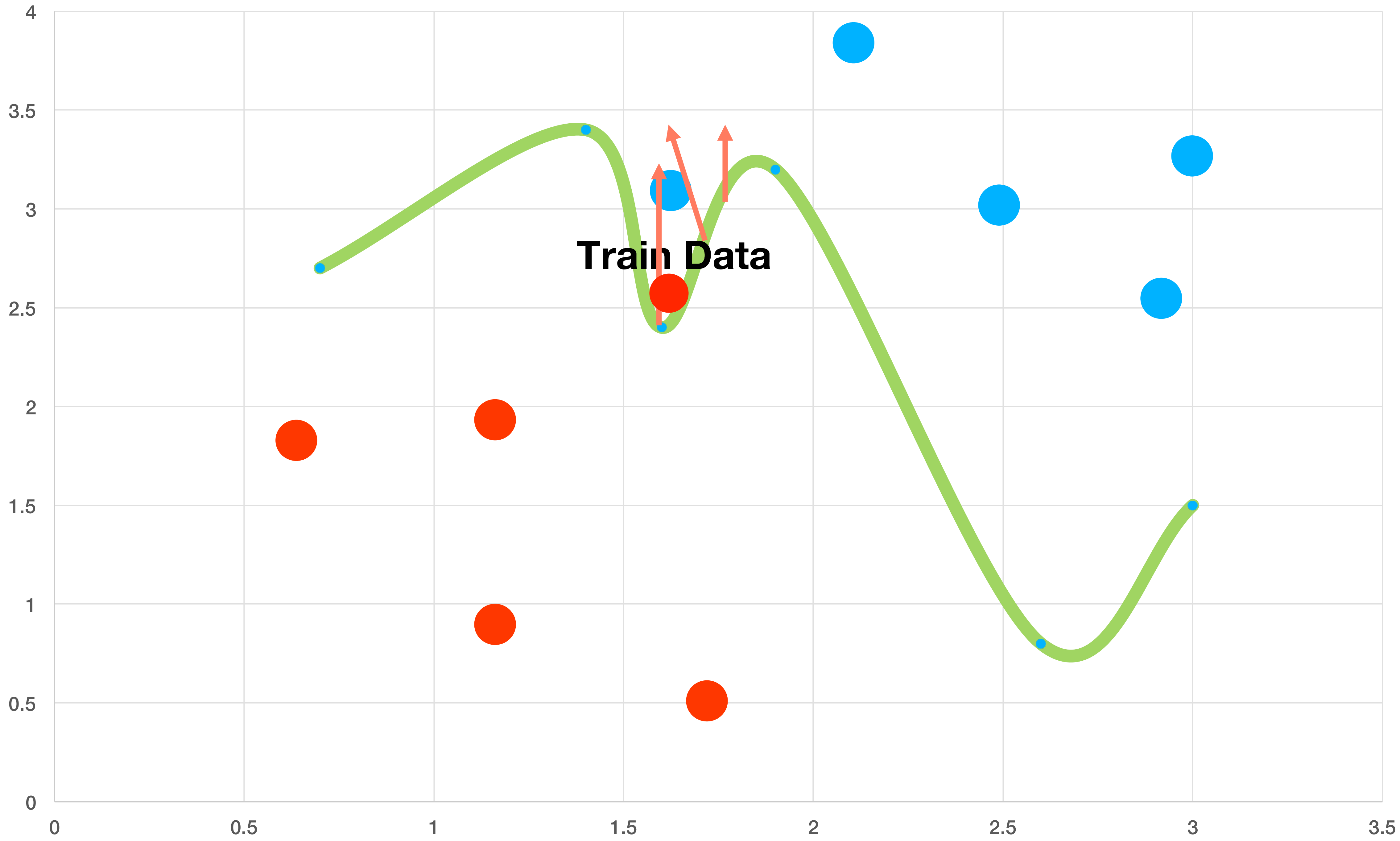
$\mathbf{L}$ **Loss**

$\nabla \mathbf{L}\downarrow \mathbf{w}$ **Loss gradient w.r.t parameters**

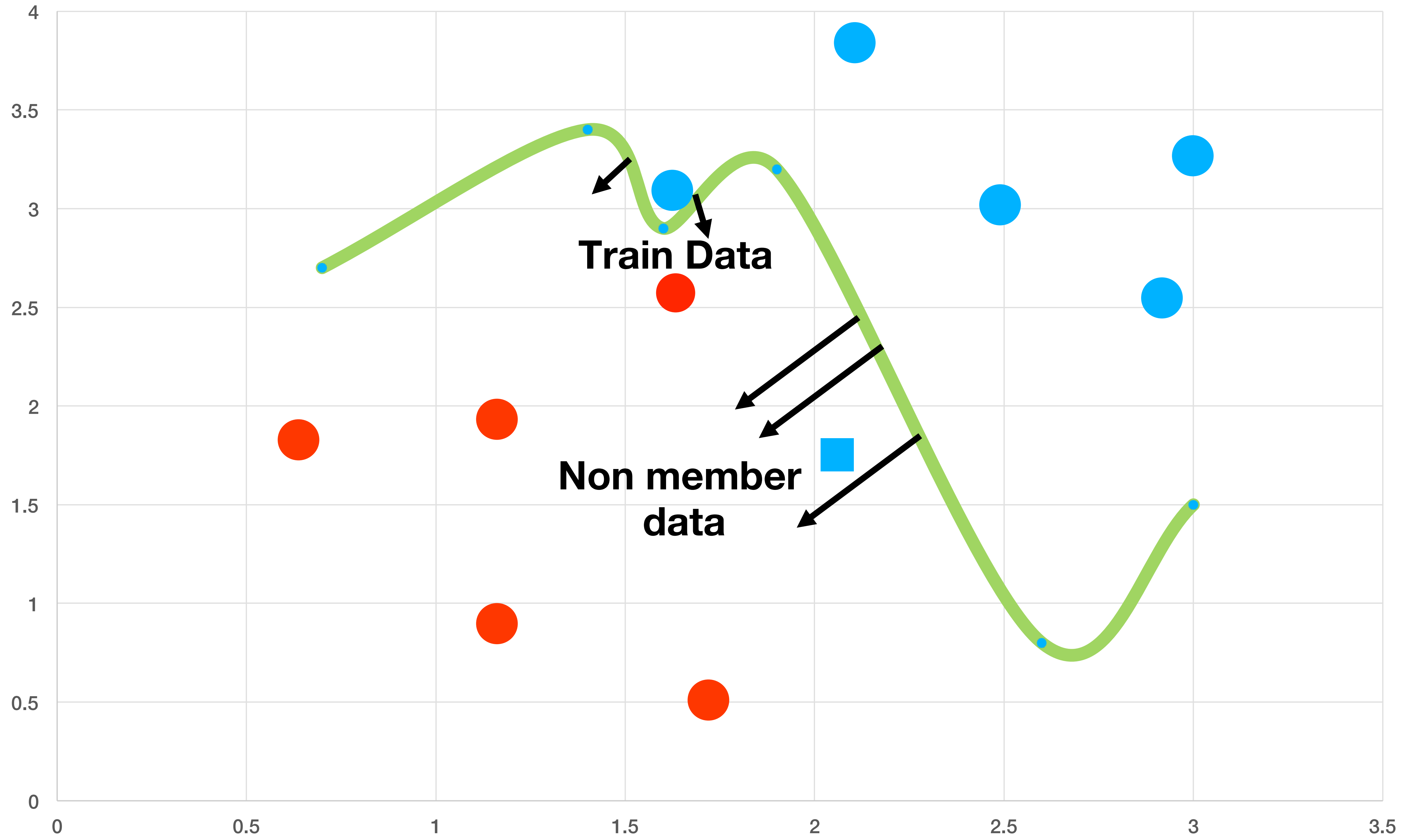
$$W = W - \alpha \nabla \mathbf{L}\downarrow \mathbf{w}$$

Model parameters change in the opposite direction of each training data point's gradient

Training a Model

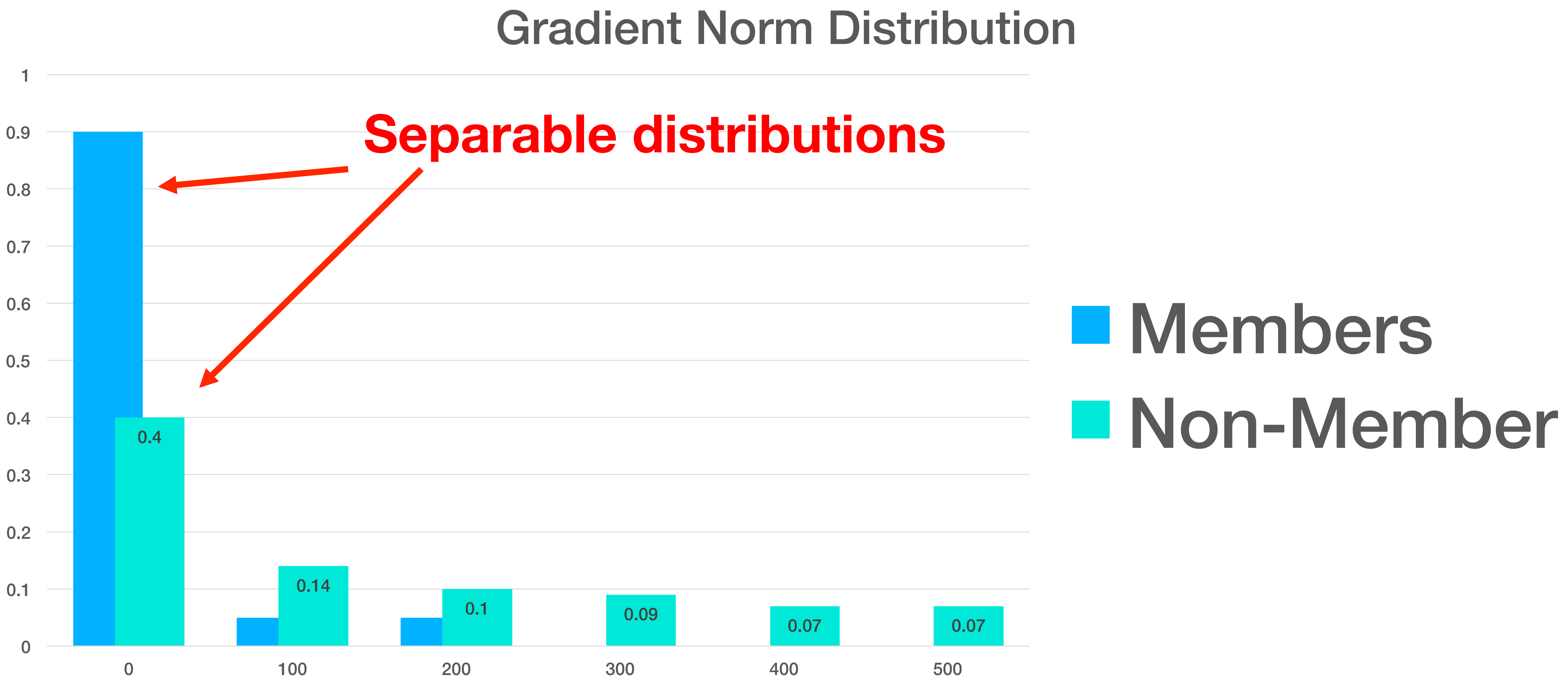


Training a Model



Gradients leak information by behaving differently for non-member data vs. member data.

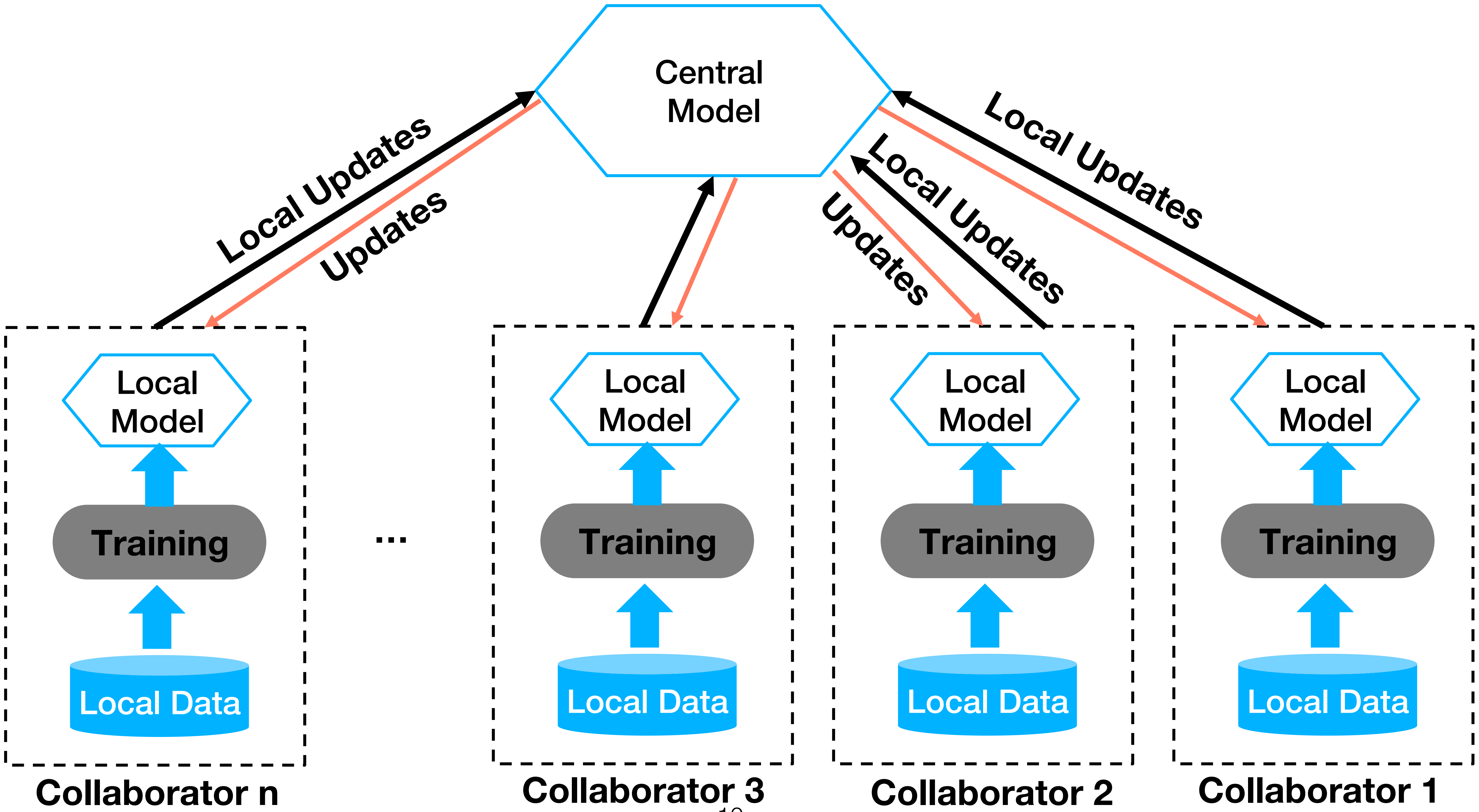
Gradients Leak Information



Different Learning/Attack Settings

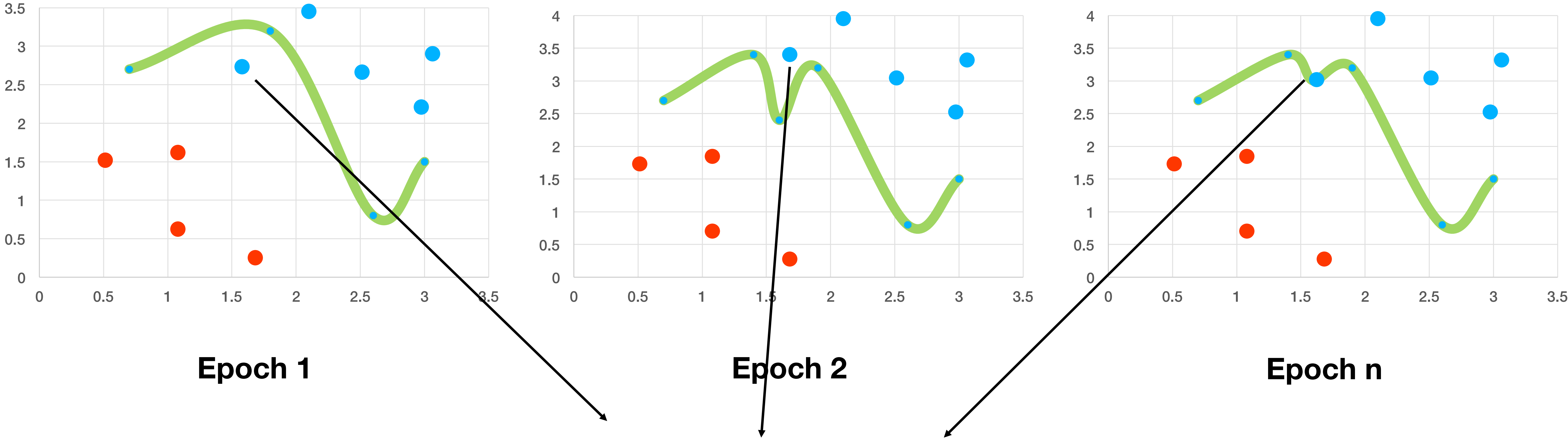
- Fully trained
  - Black/ White box
- Fine-tuning
- Federated learning
  - Central/ local Attacker
  - Passive/ Active

Federated Model



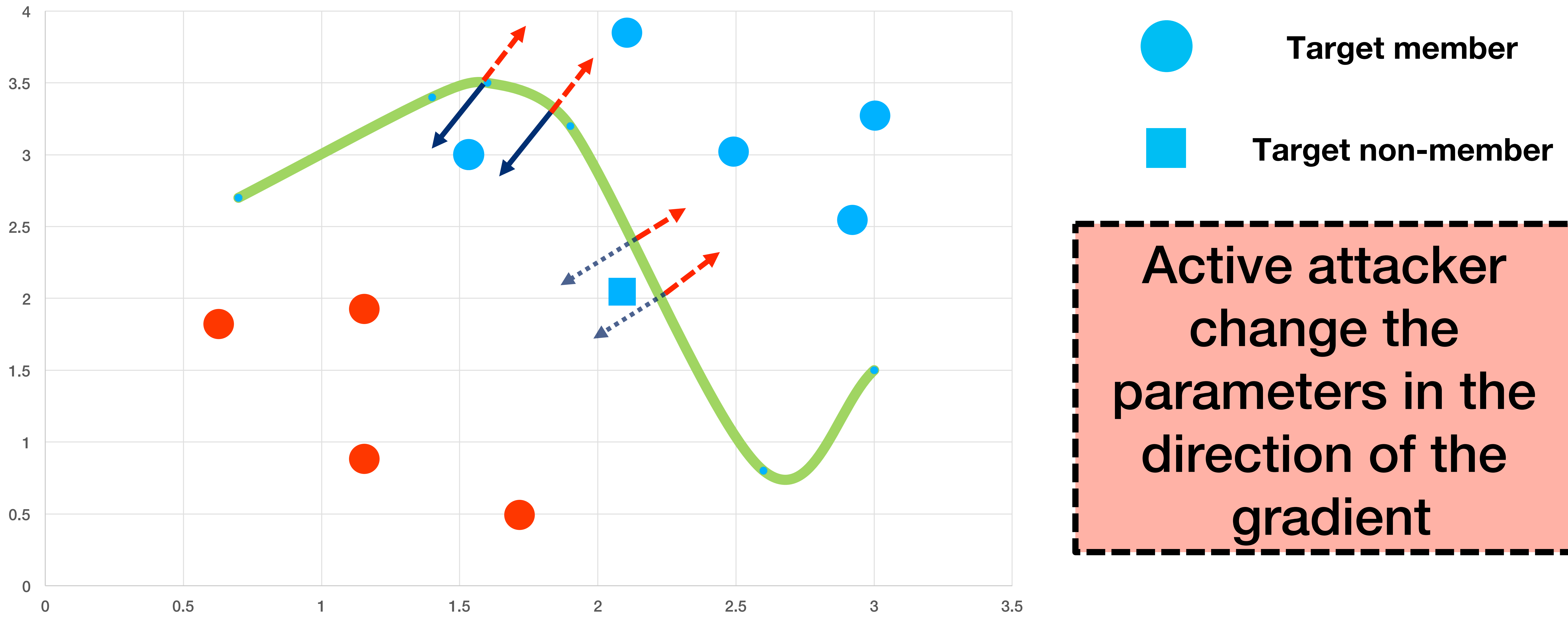
Federated Learning

Multiple observations:



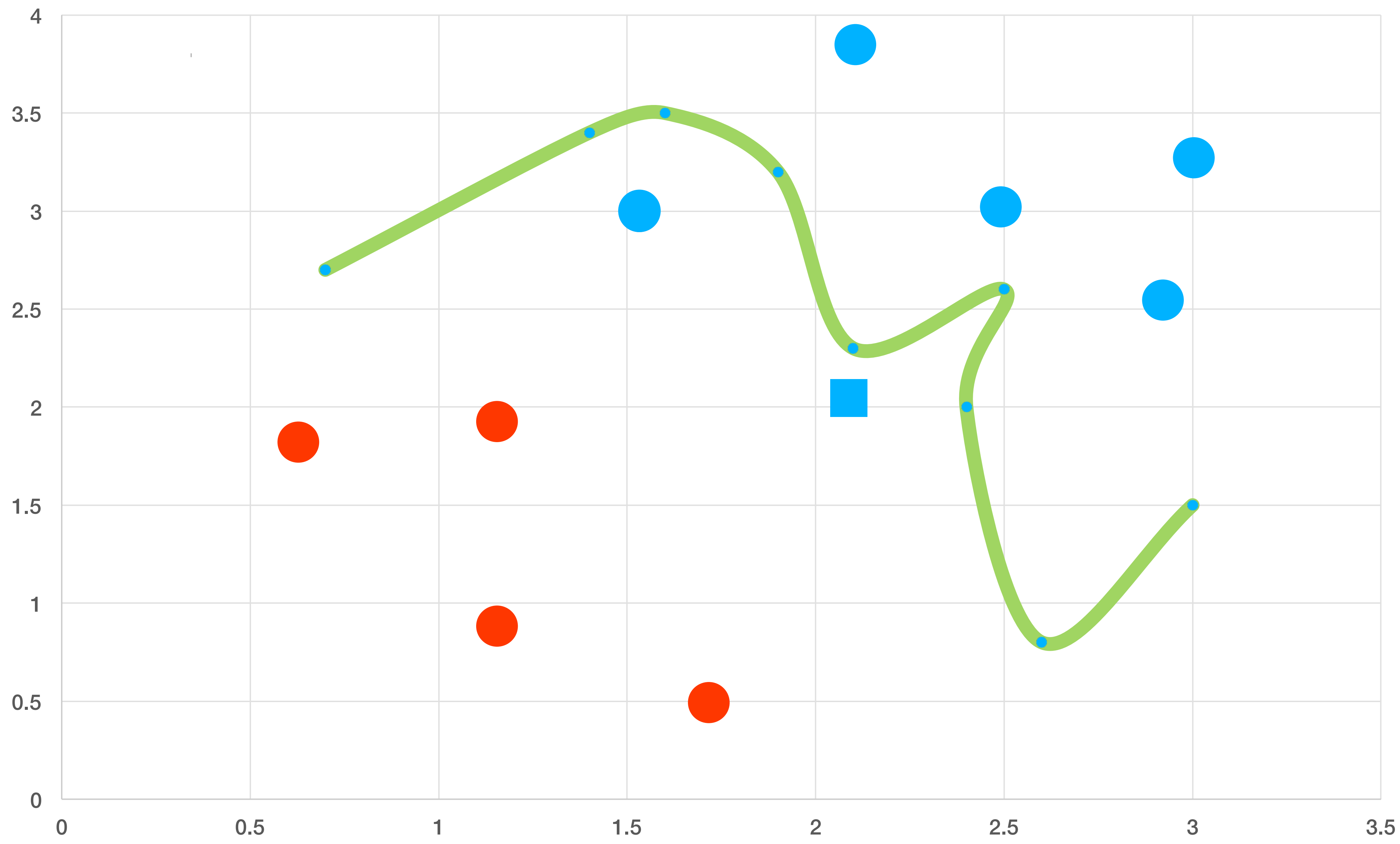
Every point leave traces on the target function

Active Attack on Federated Learning



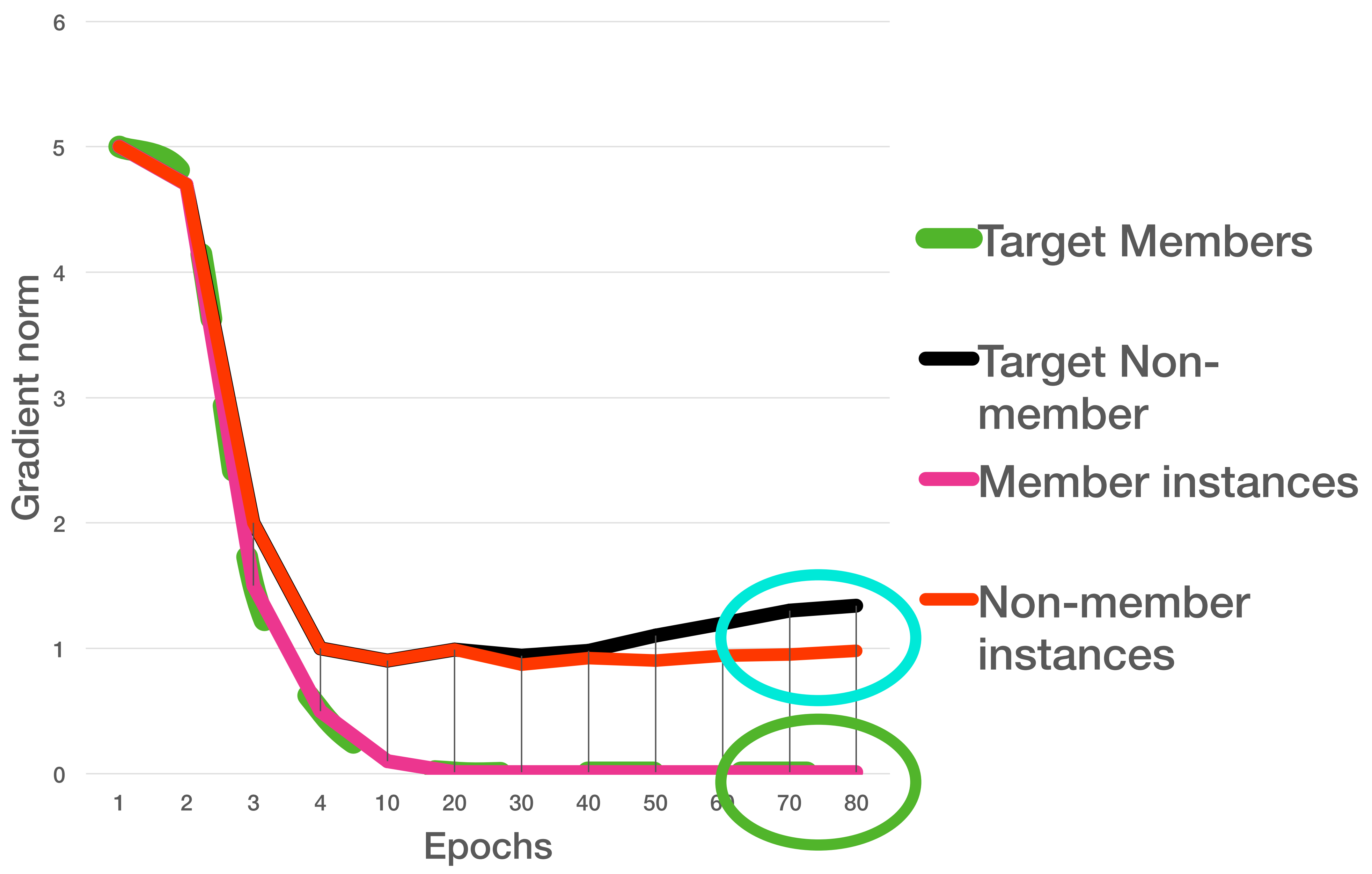
Active attacker change the parameters in the direction of the gradient

Active Attack on Federated Learning

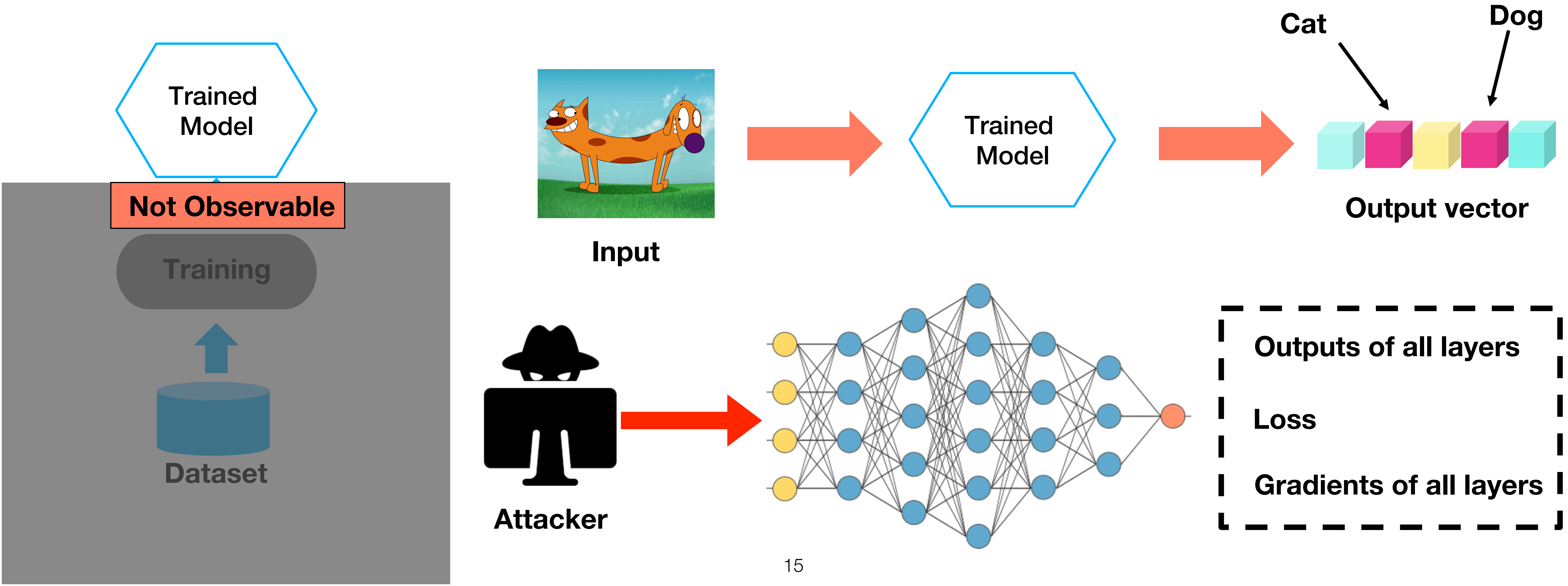


For the data points that are in the training dataset, local training will compensate for the active attacker

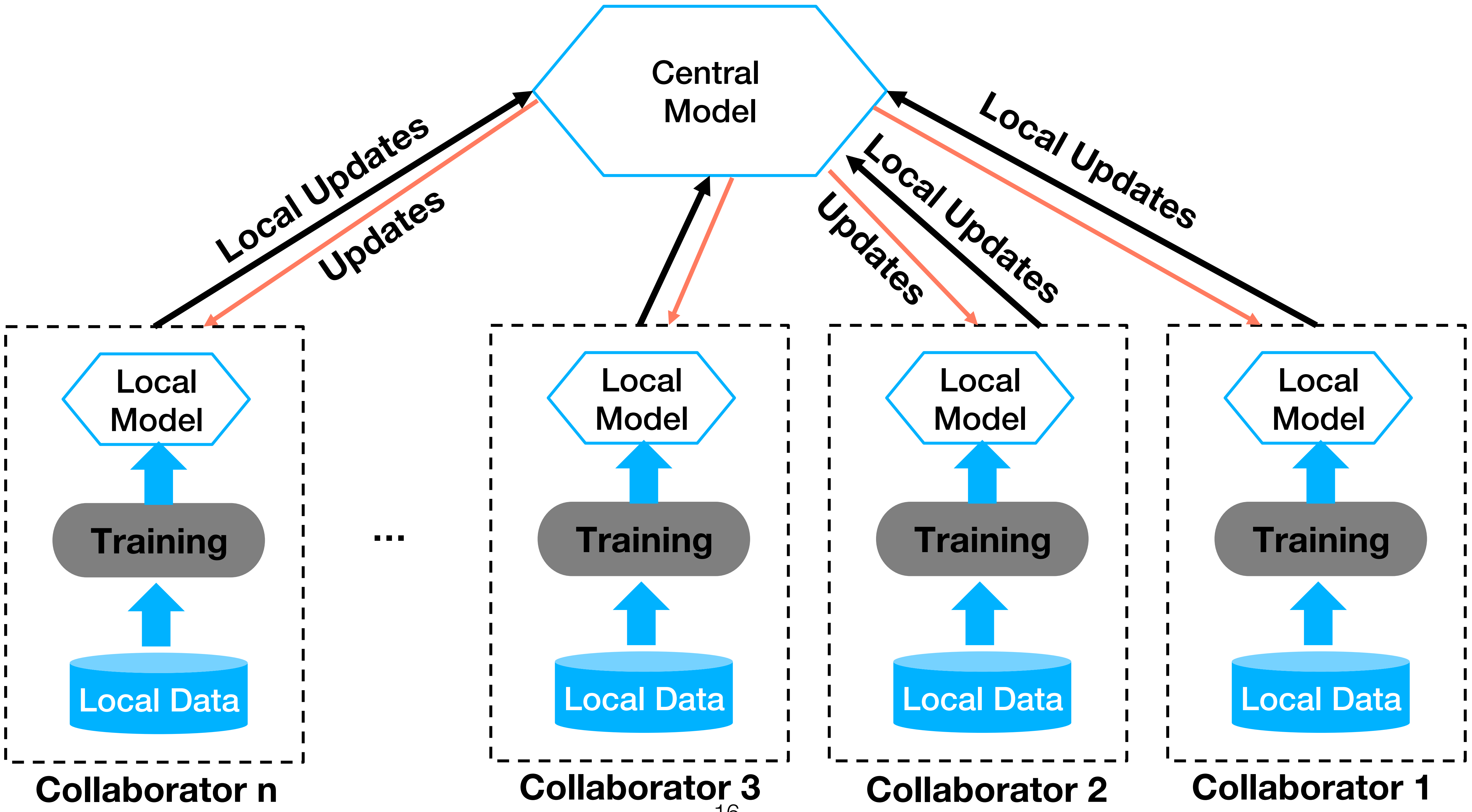
Active Attacks in Federated Model



Scenario 1: Fully Trained Model



Scenario 2: Central Attacker in Federated Model

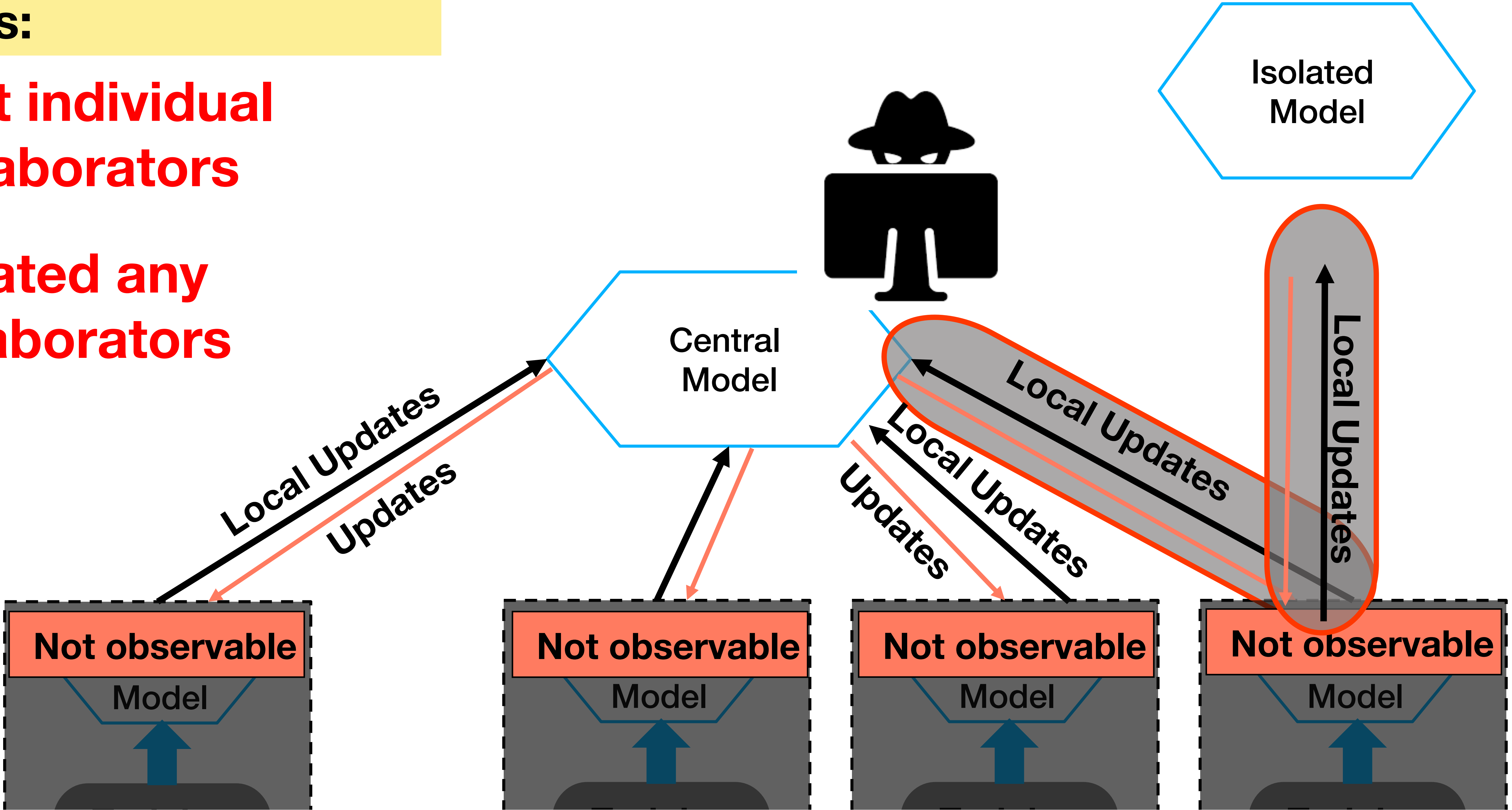


Scenario 2: Central Attacker in Federated Model

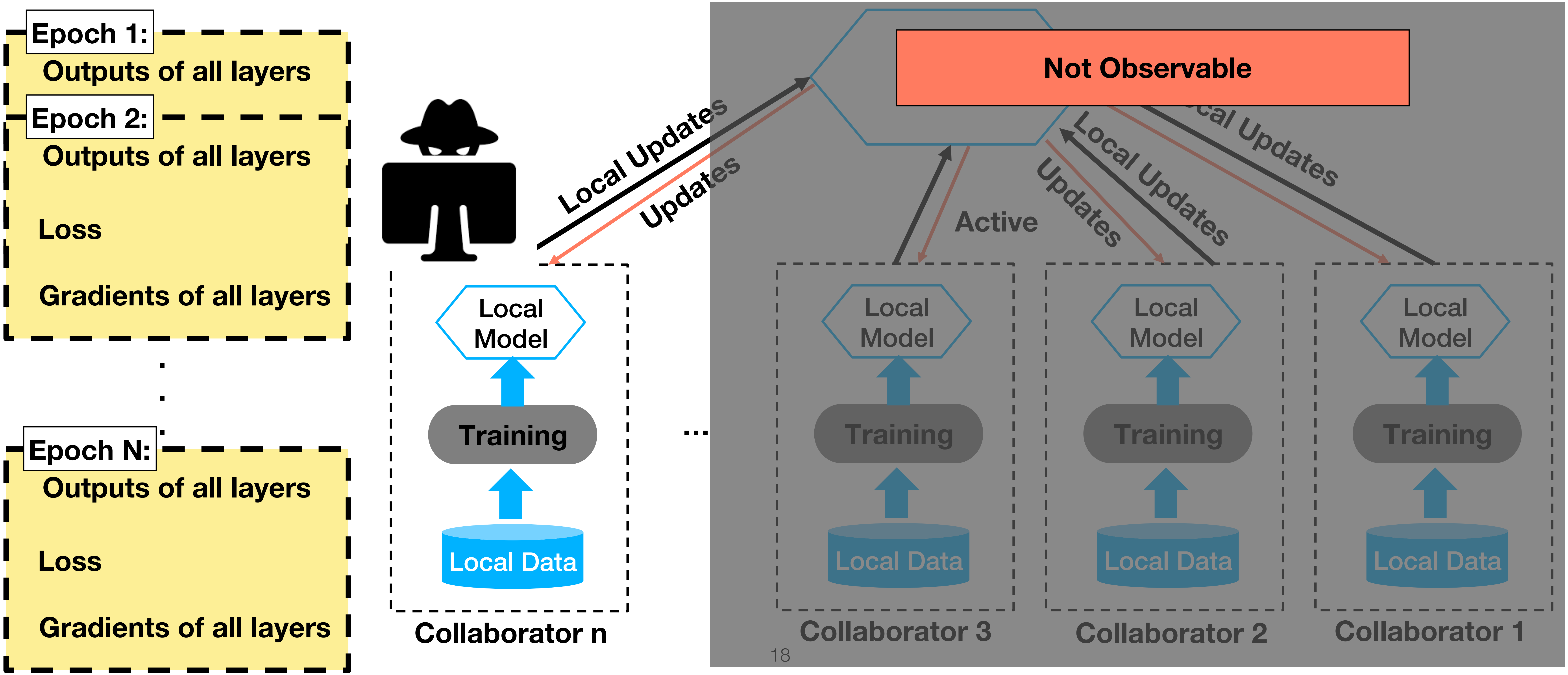
In addition to the local attacker observations:

Target individual collaborators

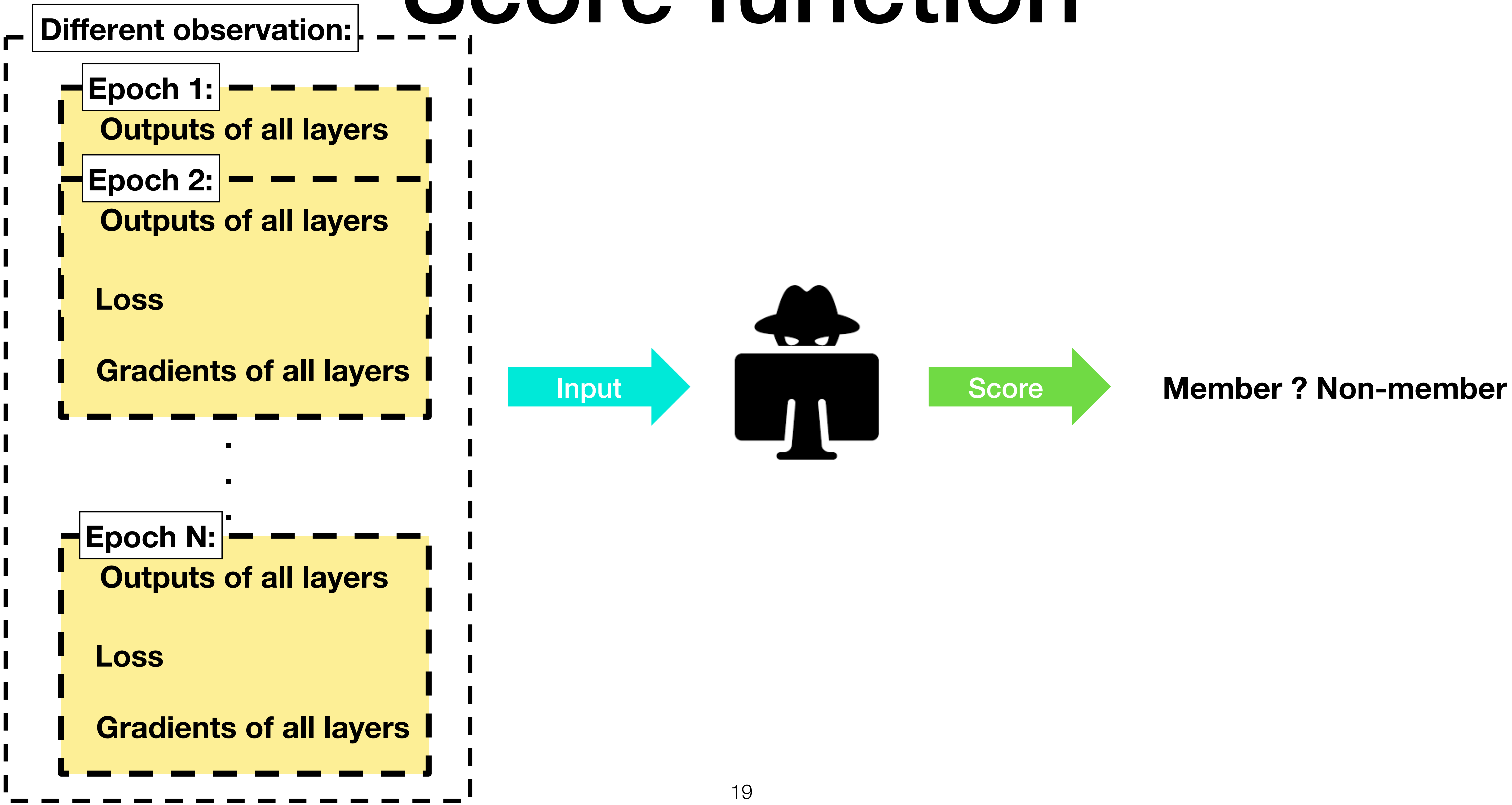
Isolated any collaborators



Scenario 3: Local Attacker in Federated Learning



Score function



Experimental Setup

- Unlike previous works, we used publicly available **pretrained models**
- We used all common regularization techniques
- We implemented our attacks in **PyTorch**
- We used following datasets:
  - *CIFAR100*
  - *Purchase100*
  - *Texas100*

Results

Pretrained Models Attacks

Gradients leak
significant information

Target Model				Attack Accuracy		
Dataset	Architecture	Train Accuracy	Test Accuracy	Black-box	White-box (Outputs)	White-box (Gradients)
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	[redacted]
CIFAR100	ResNet	89%	73%	62.2%	62.2%	[redacted]
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	[redacted]
Texas100	Fully Connected	81.6%	52%	63.0%	63.3%	[redacted]
Purchase100	Fully Connected	100%	80%	67.6%	67.6%	[redacted]

Last layer contains
the most information

Federated Attacks

Target Model		Global Attacker (the parameter aggregator)				Local Attacker (a participant)	
		Passive	Active			Passive	Active
Dataset	Architecture		Gradient Ascent	Isolating	Isolating Gradient Ascent		Gradient Ascent
CIFAR100	Alexnet	85.1%	88.2%	89.0%	92.1%	73.1%	76.3%
CIFAR100	DenseNet	79.2%	82.1%	84.3%	87.3%	72.2%	76.7%
Texas100	Fully Connected	66.4%	69.5%	69.3%	71.7%	62.4%	66.4%
Purchase100	Fully Connected	72.4%	75.4%	75.3%	82.5%	65.8%	69.8%

Global attack is more powerful than the local attacker

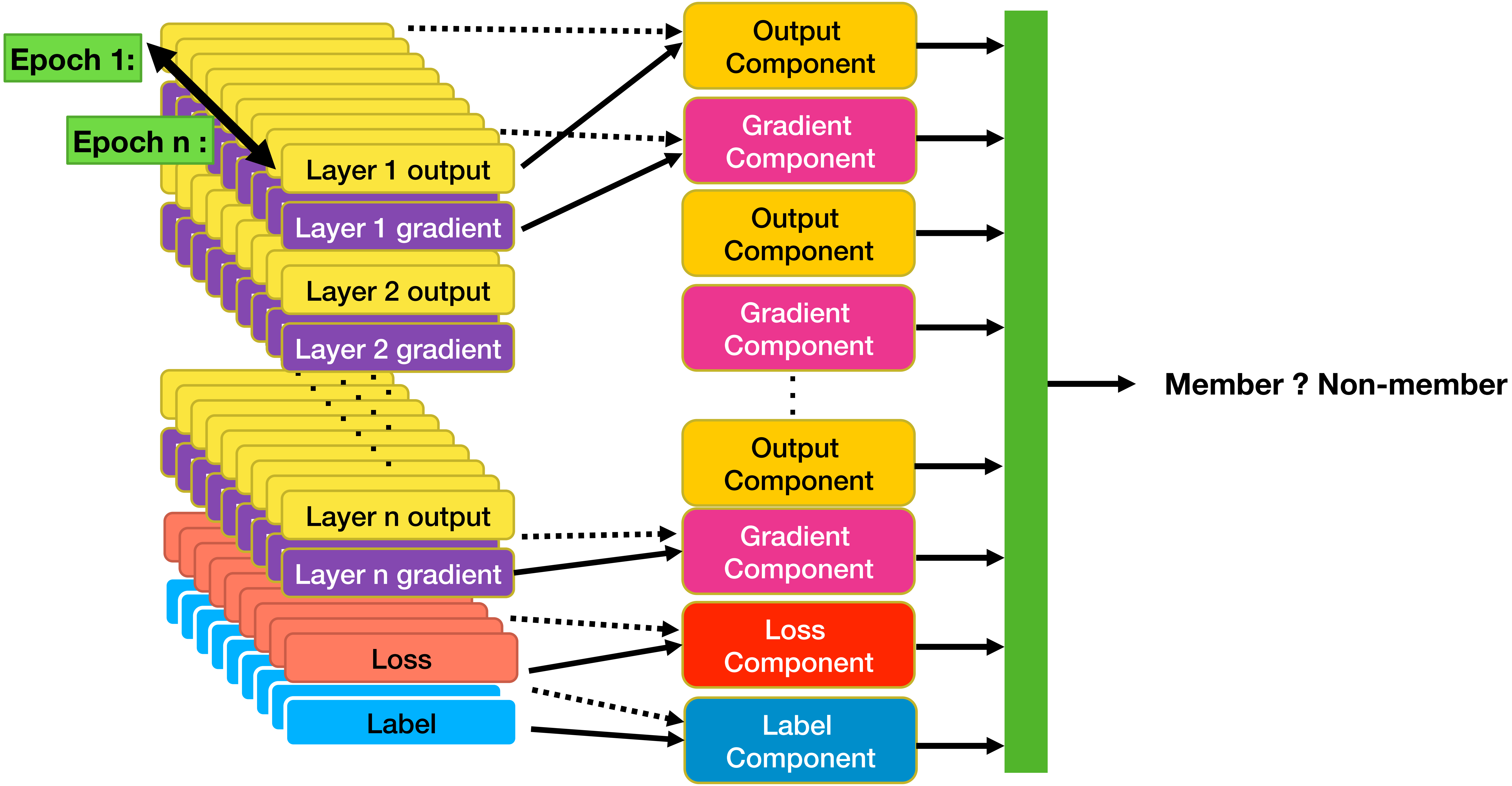
An active attacker can force SGD to leak more information

Conclusions

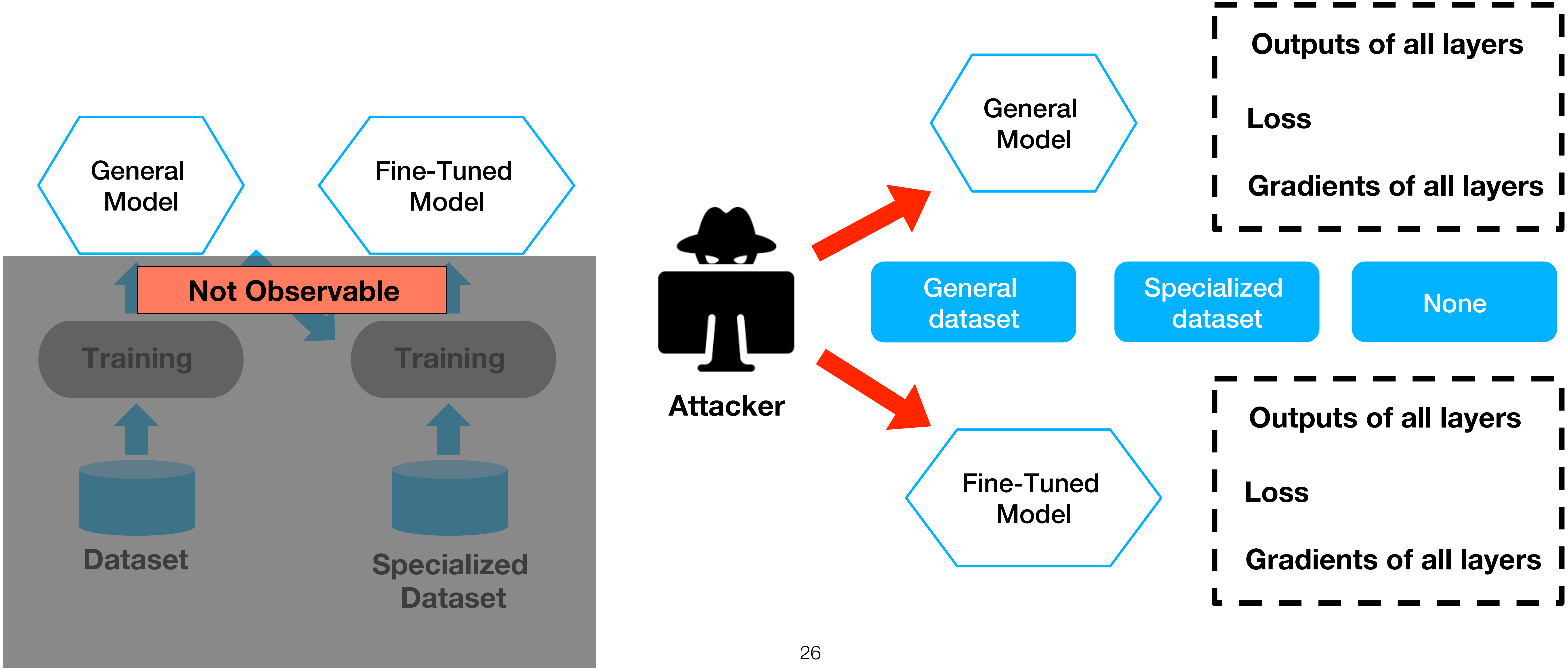
- We go beyond black-box scenario and try to understand **why a deep learning model leak information**
- **Gradients leak information** about the training dataset
- Attacker in the federated learning can take the advantage of **multiple observations** to leak more information
- In the federated setting, an attacker can **actively force SGD** to leak information

Questions ?

Overall Attack Model



Scenario 4: Fine-Tuning Model



Fine-tuning Attacks

Dataset	Arch	Distinguishing specialized/general datasets	Distinguishing general / non-member datasets	Distinguishing Specialized / non-member datasets
CIFAR100	Alexnet	62.1%	75.4%	71.3%
CIFAR100	DenseNet	63.3%	74.6%	71.5%
Texas100	Fully Connected	58.4%	68.4%	67.2%
Purchase100	Fully Connected	64.4%	73.8%	71.2%

Both specialized and general datasets are vulnerable to the membership attacks

Federated Attacks

Observed Epochs	Attack Accuracy
$5, 10, 15, 20, 25$	57.4%
$10, 20, 30, 40, 50$	76.5%
$50, 100, 150, 200, 250$	79.5%
$100, 150, 200, 250, 300$	85.1%

Number of Participants	Attack Accuracy
2	89.0%
3	78.1%
4	76.7%
5	67.2%

Fine-Tuning Model Leakage