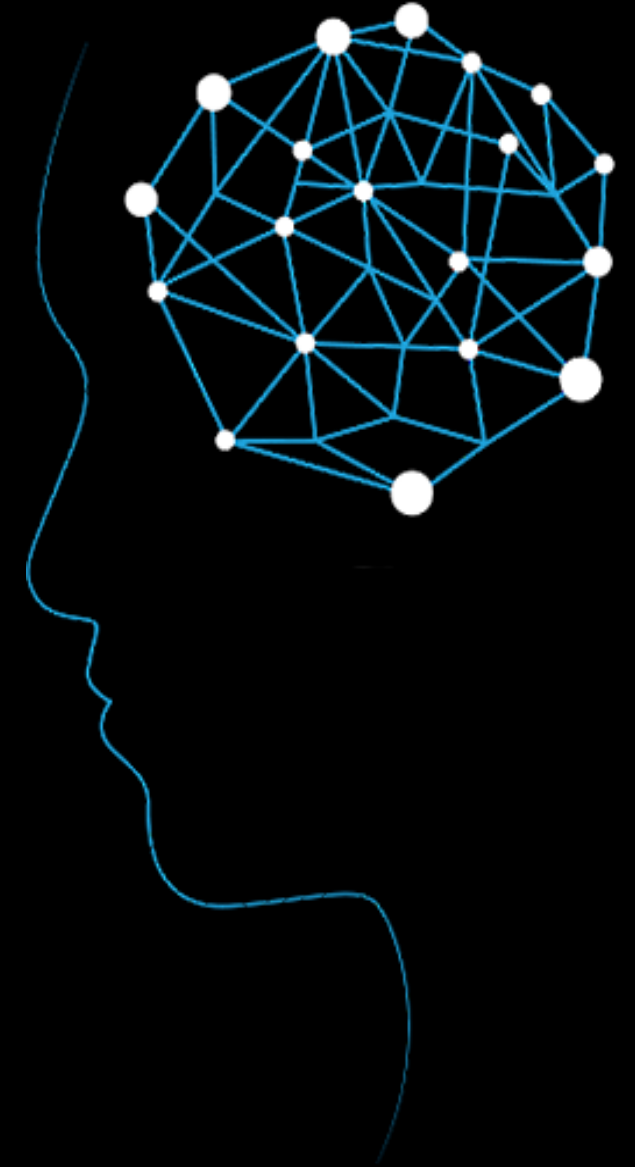


Confronting Adaptive Attackers in Federated Learning: Challenges and Countermeasures

Alexandra Dmitrienko,
Julius Maximilians Universität Würzburg

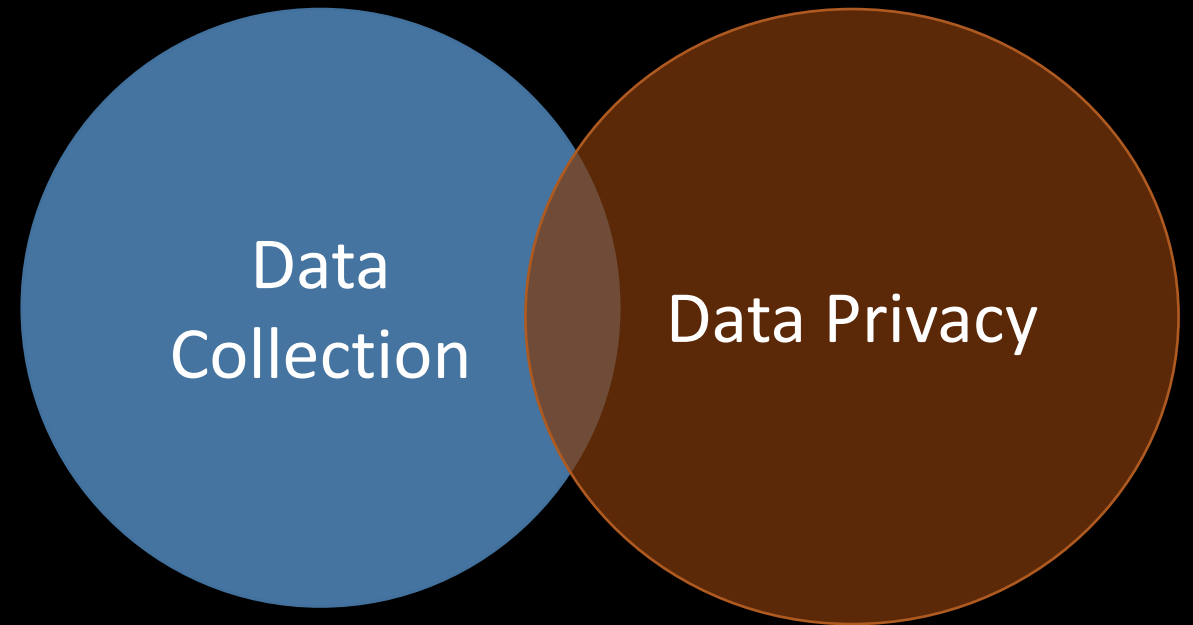


Privacy Challenge of AI

Data-hungry AI



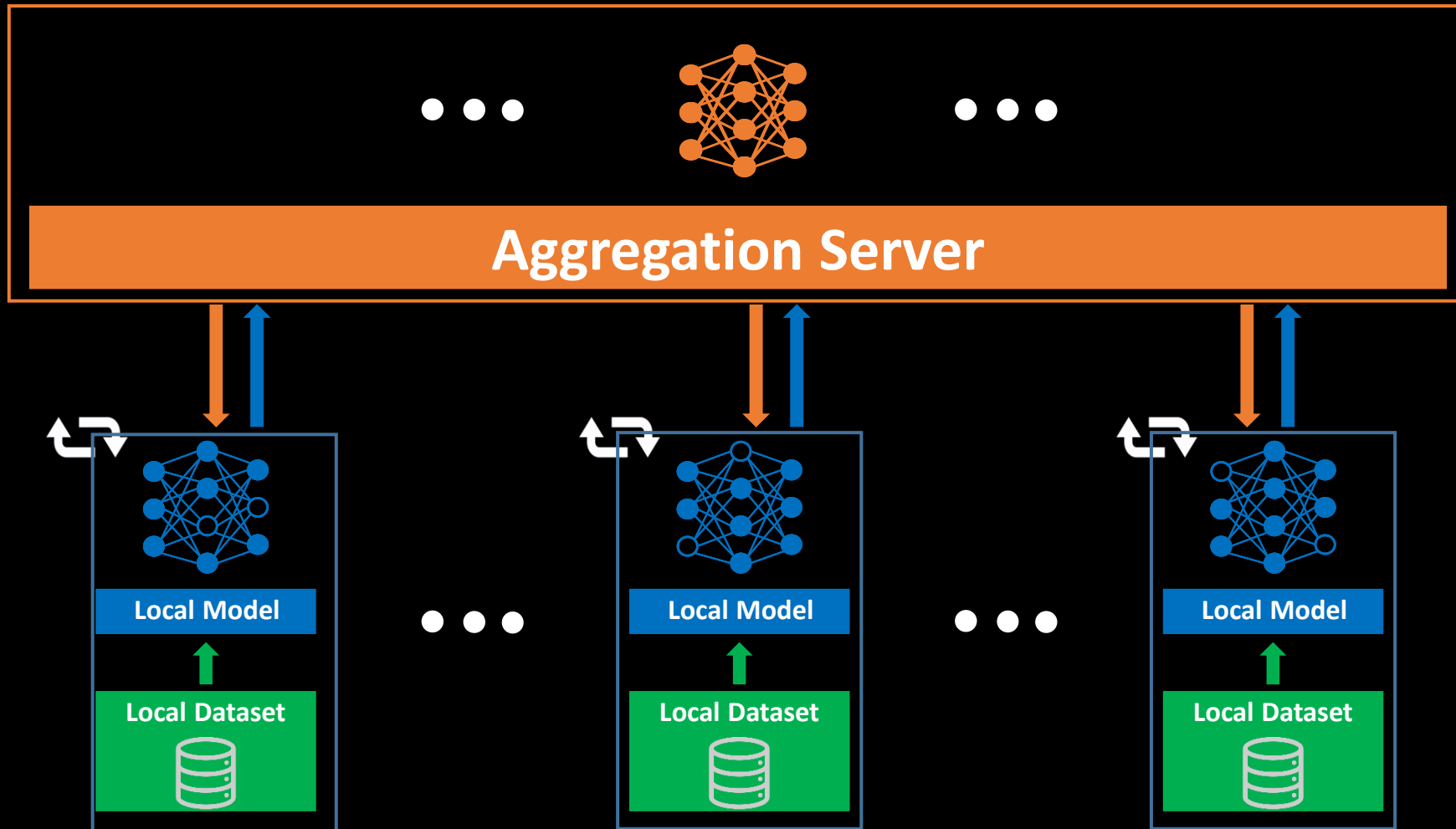
Requirement on large-scale data collection
contradicts privacy requirements





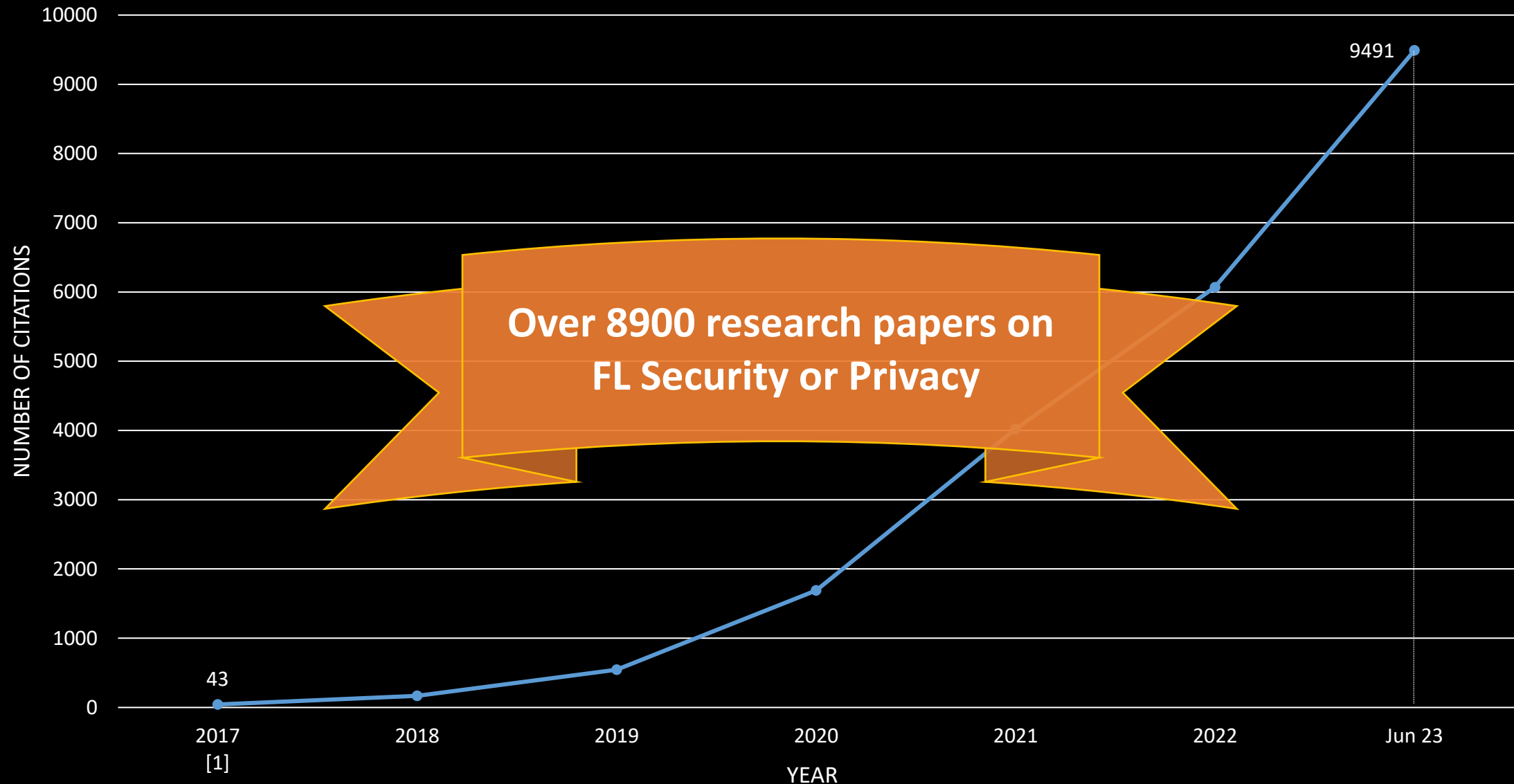
Federated Learning can help!

Federated Learning Training



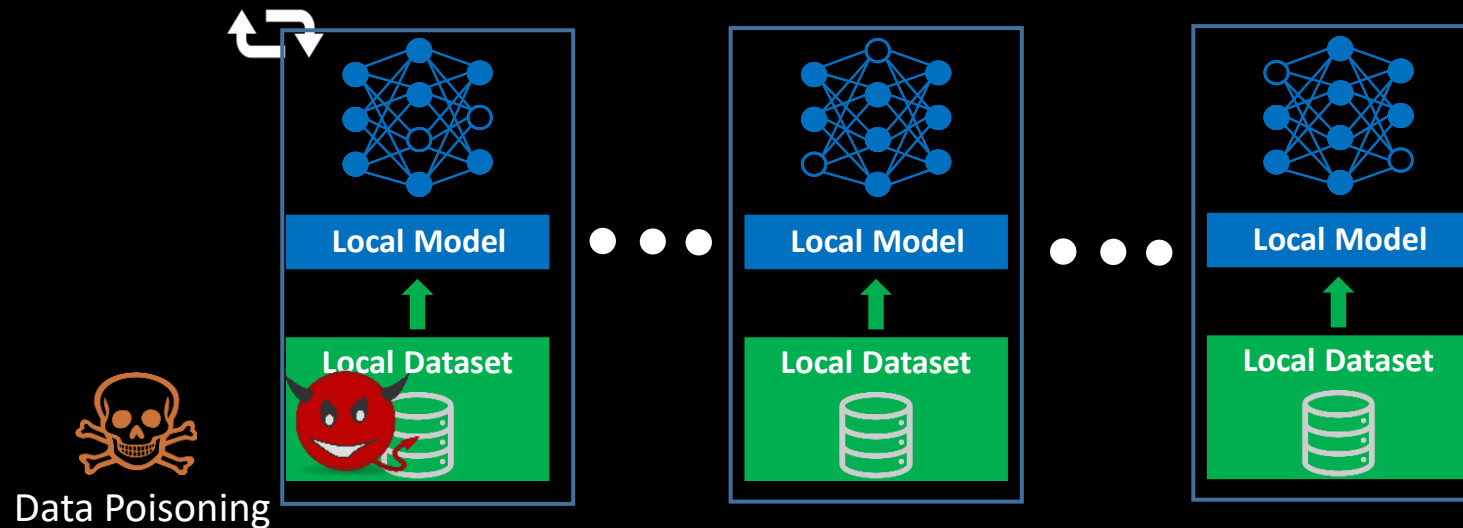
Federated Learning: Large Body of Literature

Source: Google Scholar

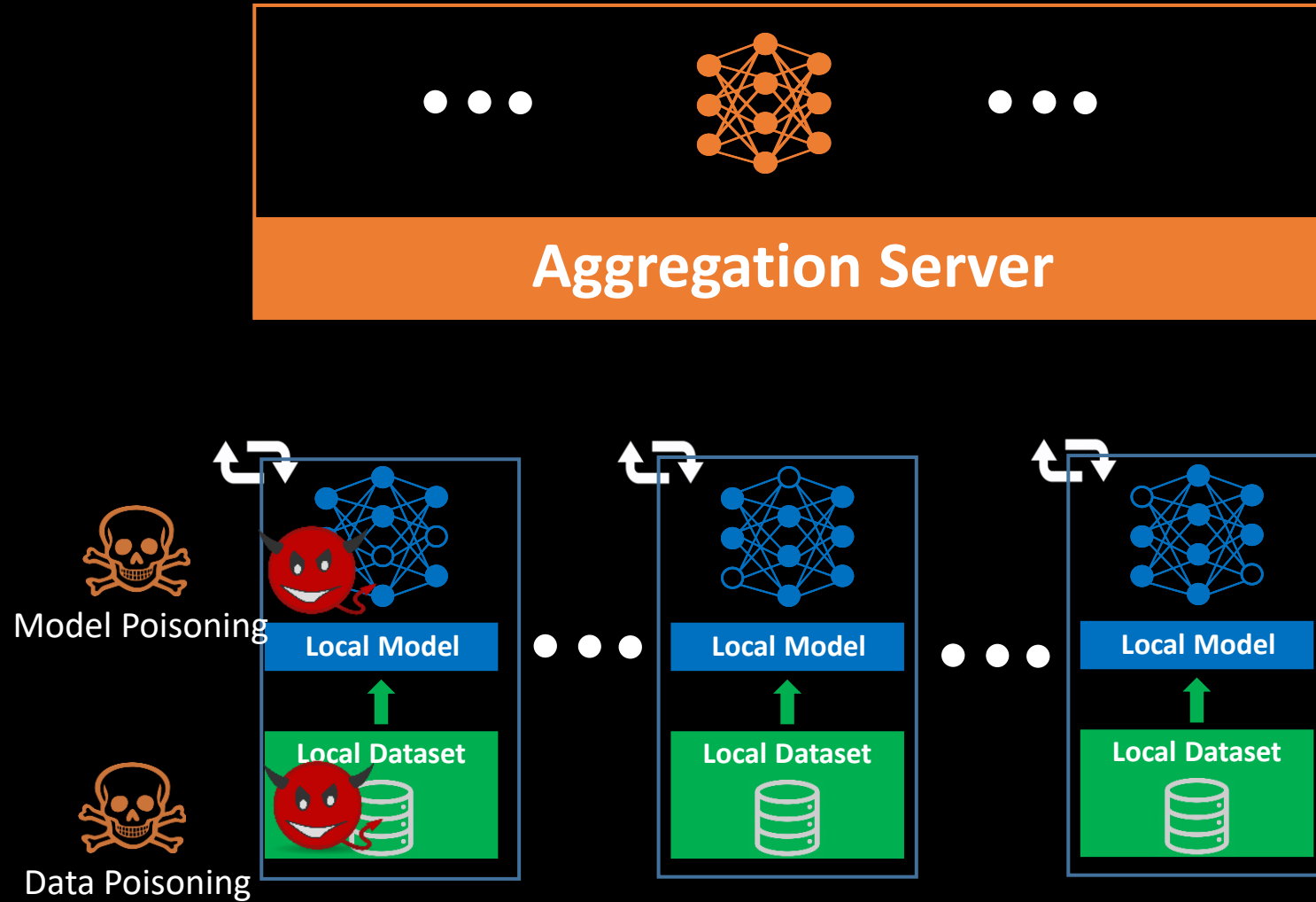


[1] McMahan et al. "Communication-efficient learning of deep networks from decentralized data.", PMLR, 2017.

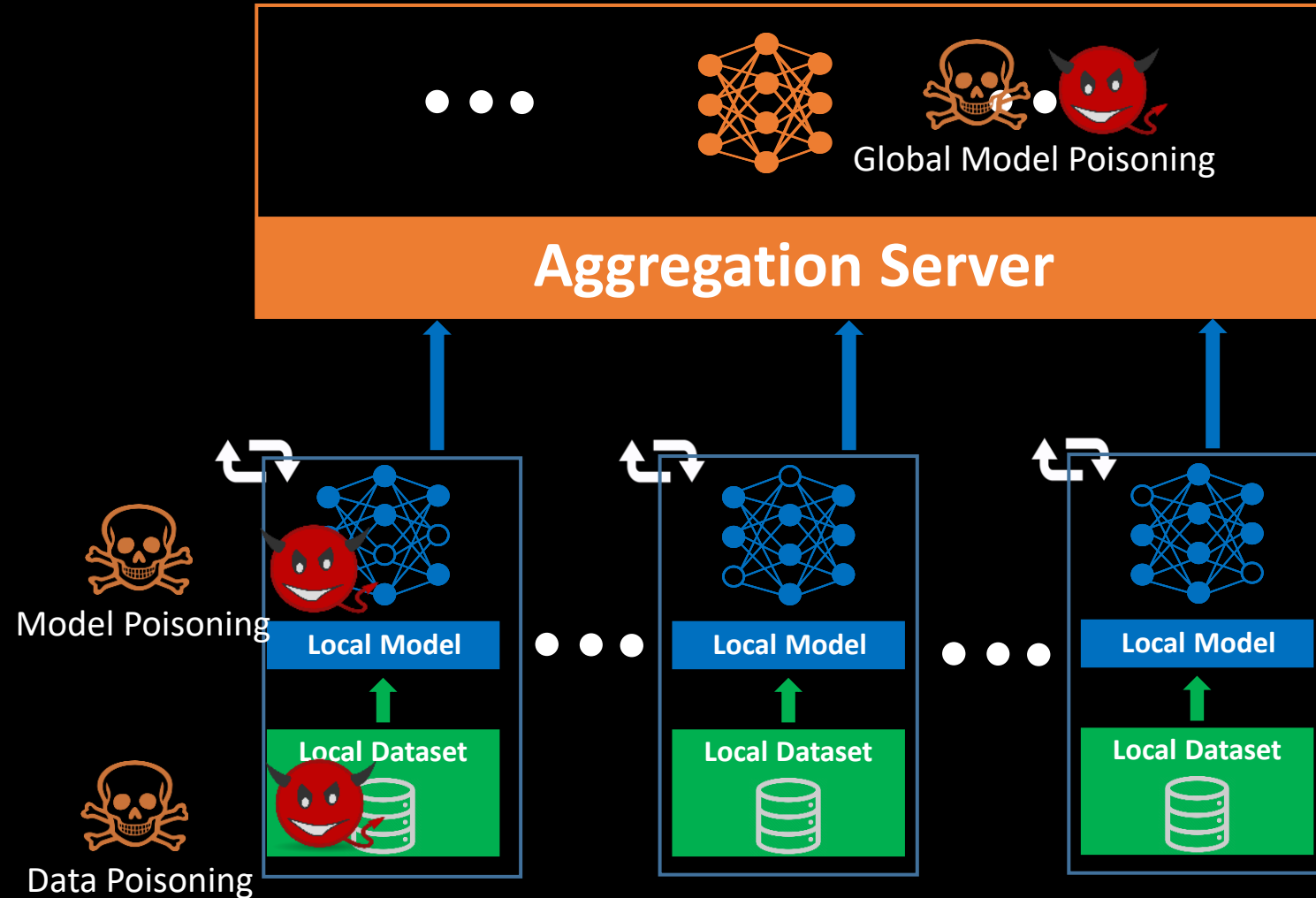
Security and Privacy Risks in Federated Learning



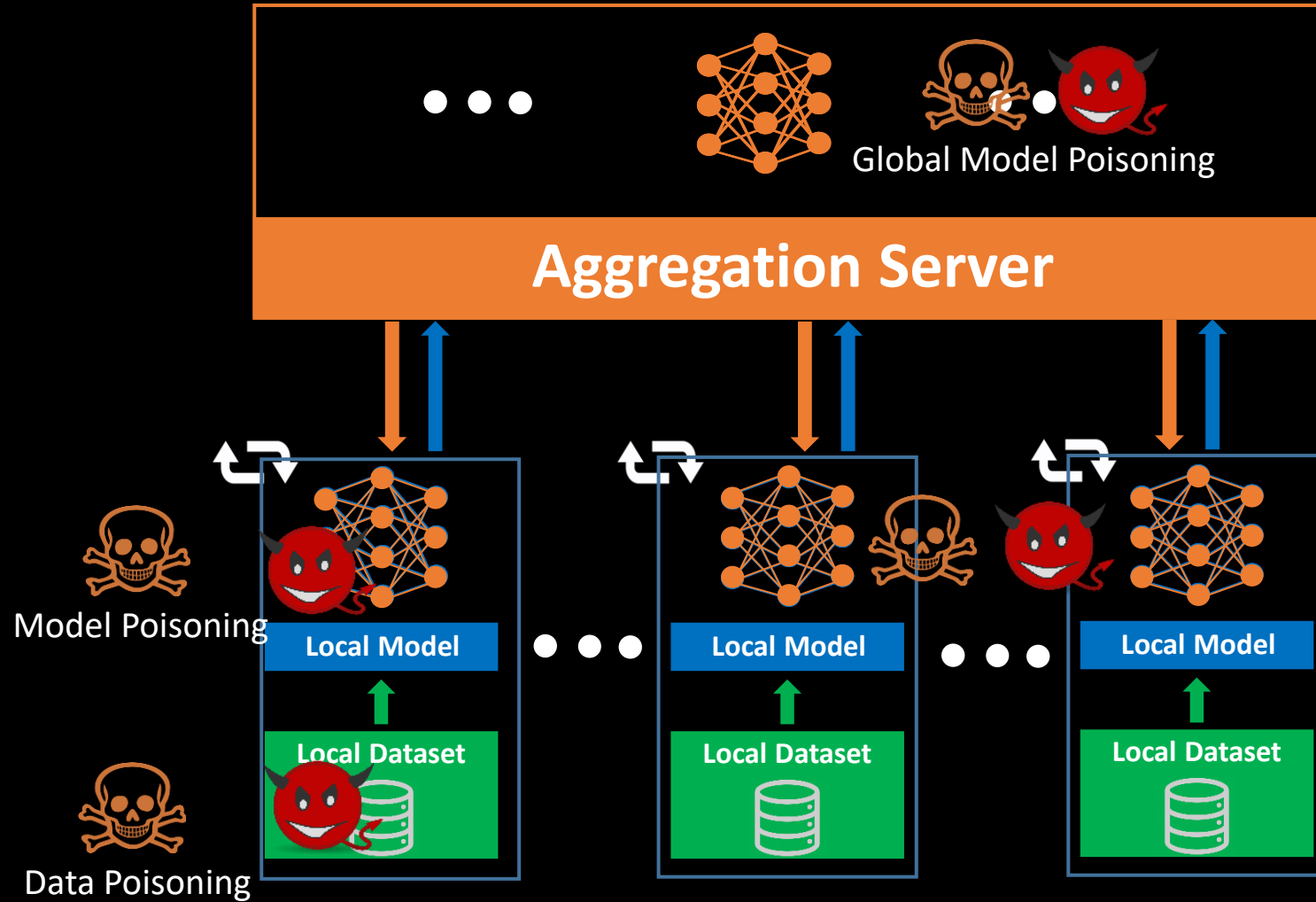
Security and Privacy Risks in Federated Learning



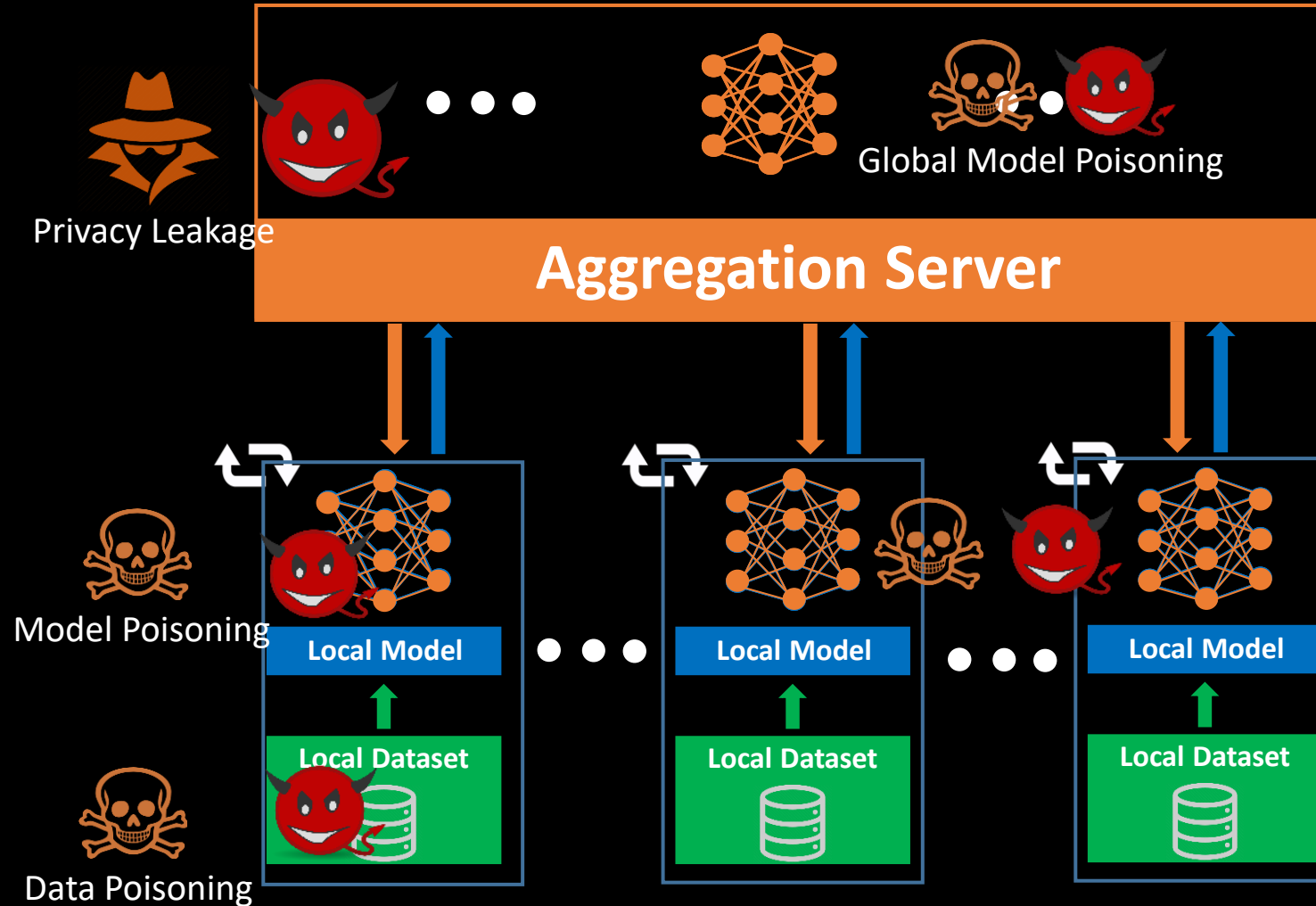
Security and Privacy Risks in Federated Learning



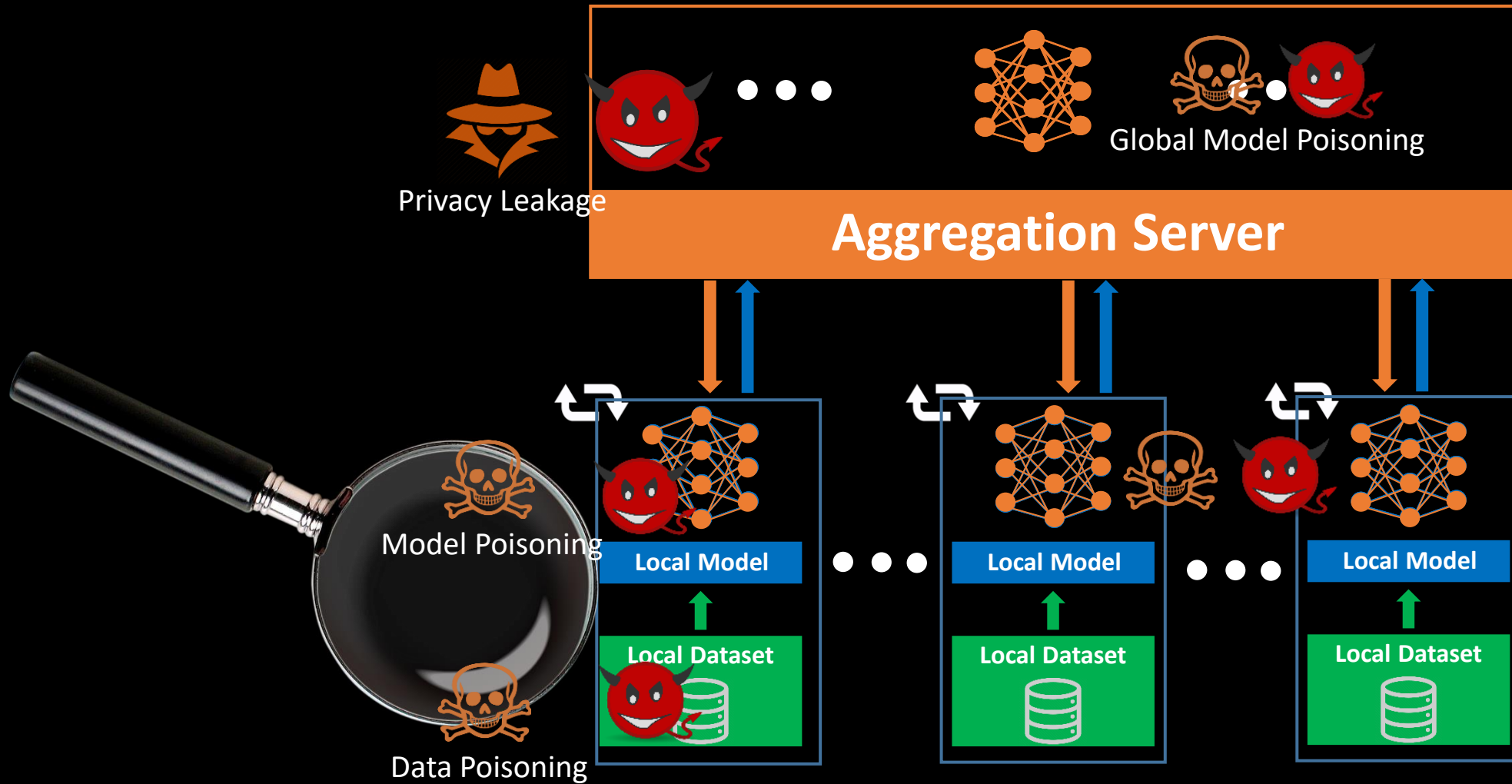
Security and Privacy Risks in Federated Learning



Security and Privacy Risks in Federated Learning



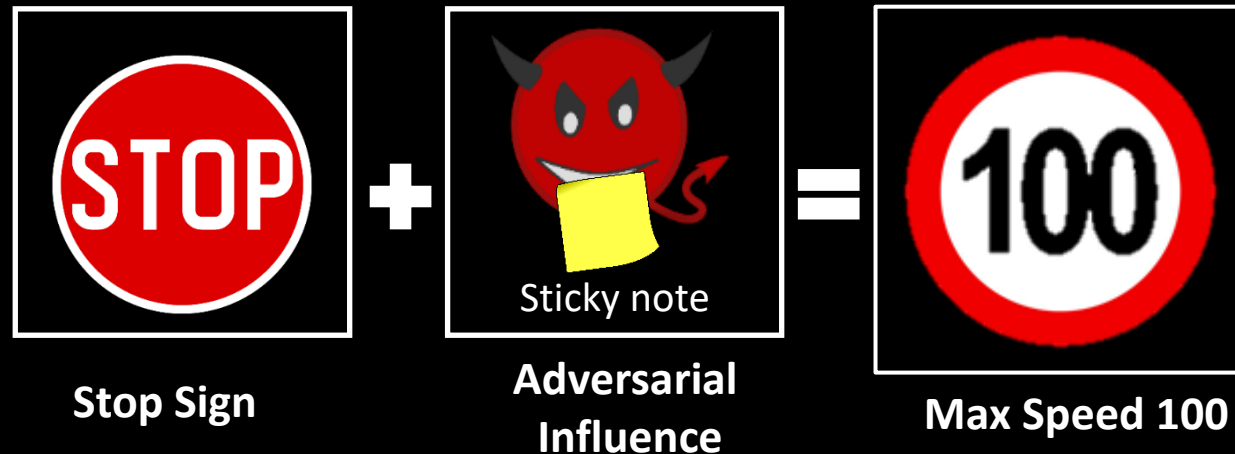
Security and Privacy Risks in Federated Learning



Security Risks: Poisoning Attacks

- Errors in classification by ML models can have devastating effects
- Security weaknesses are especially concerning if ML models are deployed in security or safety-critical applications
- Untargeted poisoning:
 - Models can be trained on poor quality data, thus lowering classification accuracy
- Targeted poisoning, or backdoors:
 - Attackers can induce (attacker-chosen) errors only on specific inputs, and without lowering accuracy on main classification task

Hypothetical attack on Self-Driving Cars



Defense Approaches

Information Reduction

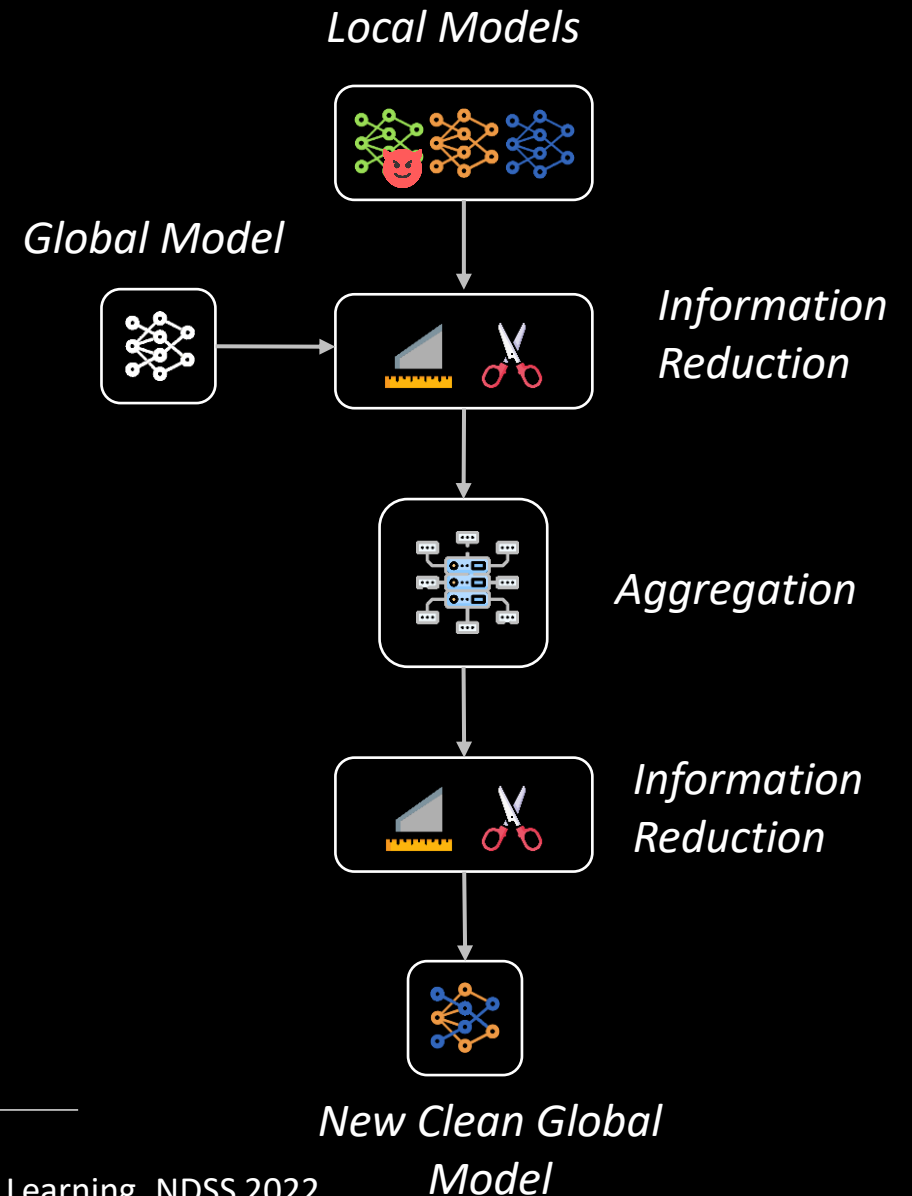
- Differential Privacy approaches, e.g., noising and clipping [1,2], gradient pruning [2]
- Conducted on local models or aggregated global model

Robust Aggregation

- Replace the standard aggregation algorithm
- E.g., select only one local contribution to be part of the new global model
 - Either a complete local model, based on update density [3]
 - Or parameter-wise, based on the mean/median of each parameter [4]

Limitations

- Reduces classification accuracy on the main task



[1] E. Bagdasaryan et al., How To Backdoor Federated Learning. *AISTATS*, 2020

[2] Naseri et al., Local and Central Differential Privacy for Robustness and Privacy in Federated Learning, NDSS 2022

[3] Blanchard, et al, Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. NIPS, 2017

[4] Yin, et al, Byzantine-robust distributed learning: Towards optimal statistical rate. PMLR, 2018

Defense Approaches

Detection & Filtering [1,2,3,4]

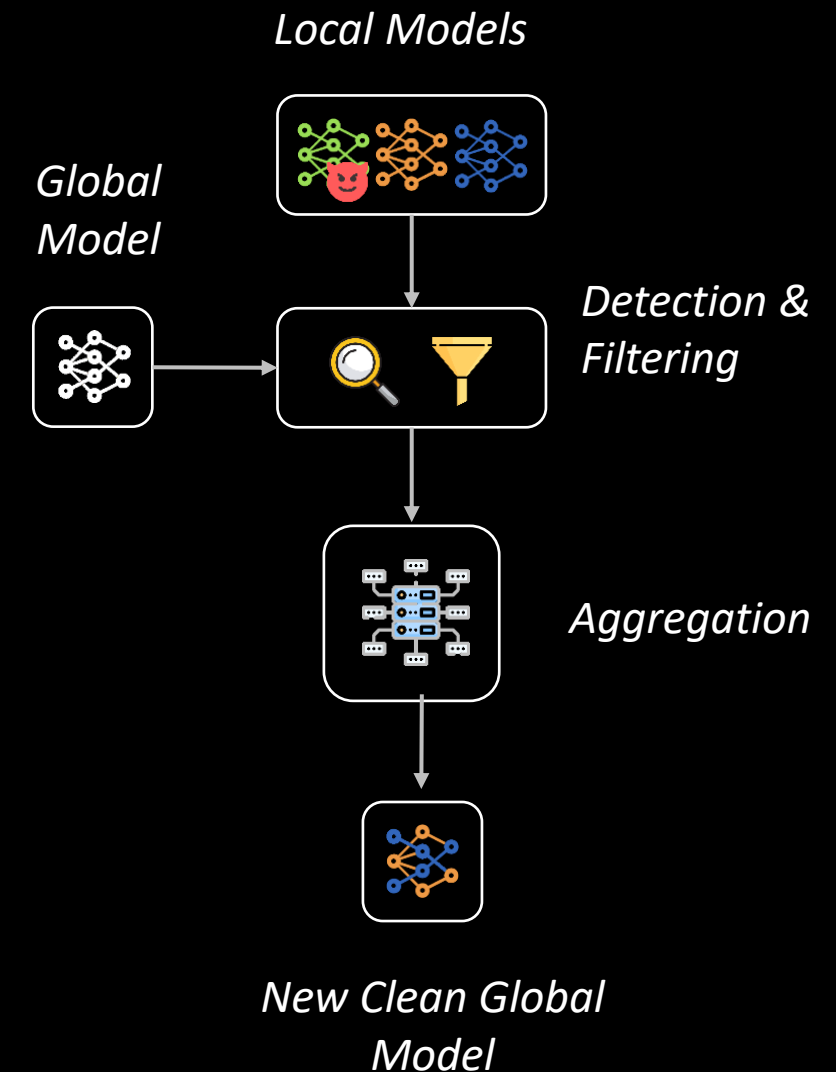
- Conducted on local models or updates (to the global model)
- Detection based on one or a few metrics
- Filtering leverages clustering methods

Advantages

- Classification accuracy on the main task is not reduced

Challenges

- Accurate distinguishing of malicious model updates vs. benign updates from clients with unusual data distributions (non-IID data)
- Detection of multiple backdoors
- Adaptive adversary



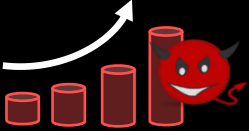
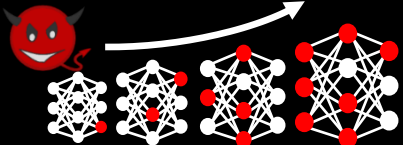
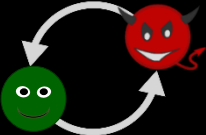

[1] Fung et al., The limitations of federated learning in Sybil settings. In RAID, 2020 (

[2] Awan et al. CONTRA: Defending against Poisoning Attacks in Federated Learning. ESORICS, 2021

[3] Shen et al., Auror: Defending Against Poisoning Attacks in Collaborative Deep Learning Systems. ACSAC, 2016

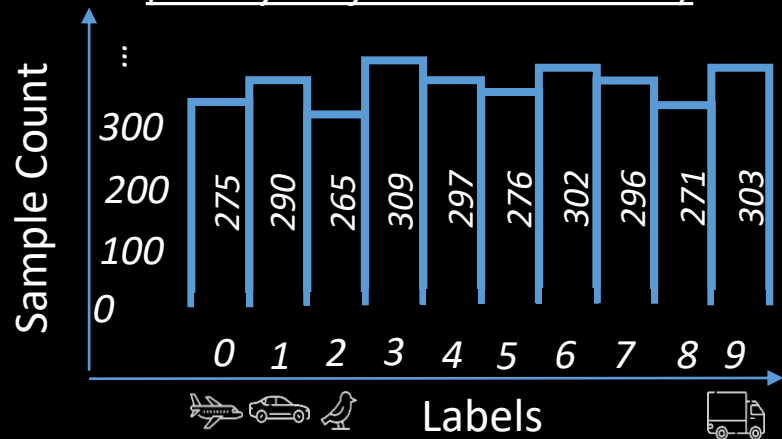
[4] Muñoz-González et al., Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging. In arXiv preprint:1909.05125, 2019

Adaptive Attack Strategies

<p>Changing PDR</p> 	<p>Adapt number of samples for backdoor behavior in training data</p>
<p>Changing PMR</p> 	<p>Adapt number of malicious clients that inject the backdoor</p>
<p>Changing Behaviour</p> 	<p>Behave benign or malicious in different training rounds</p>
<p>Changing Loss Function</p>  $Loss_{train} = Loss_{benign} + Loss_{adv}$	<p>Adding an additional adaptation loss to constrain weights</p> $Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$

The Challenge of Non-IID Data

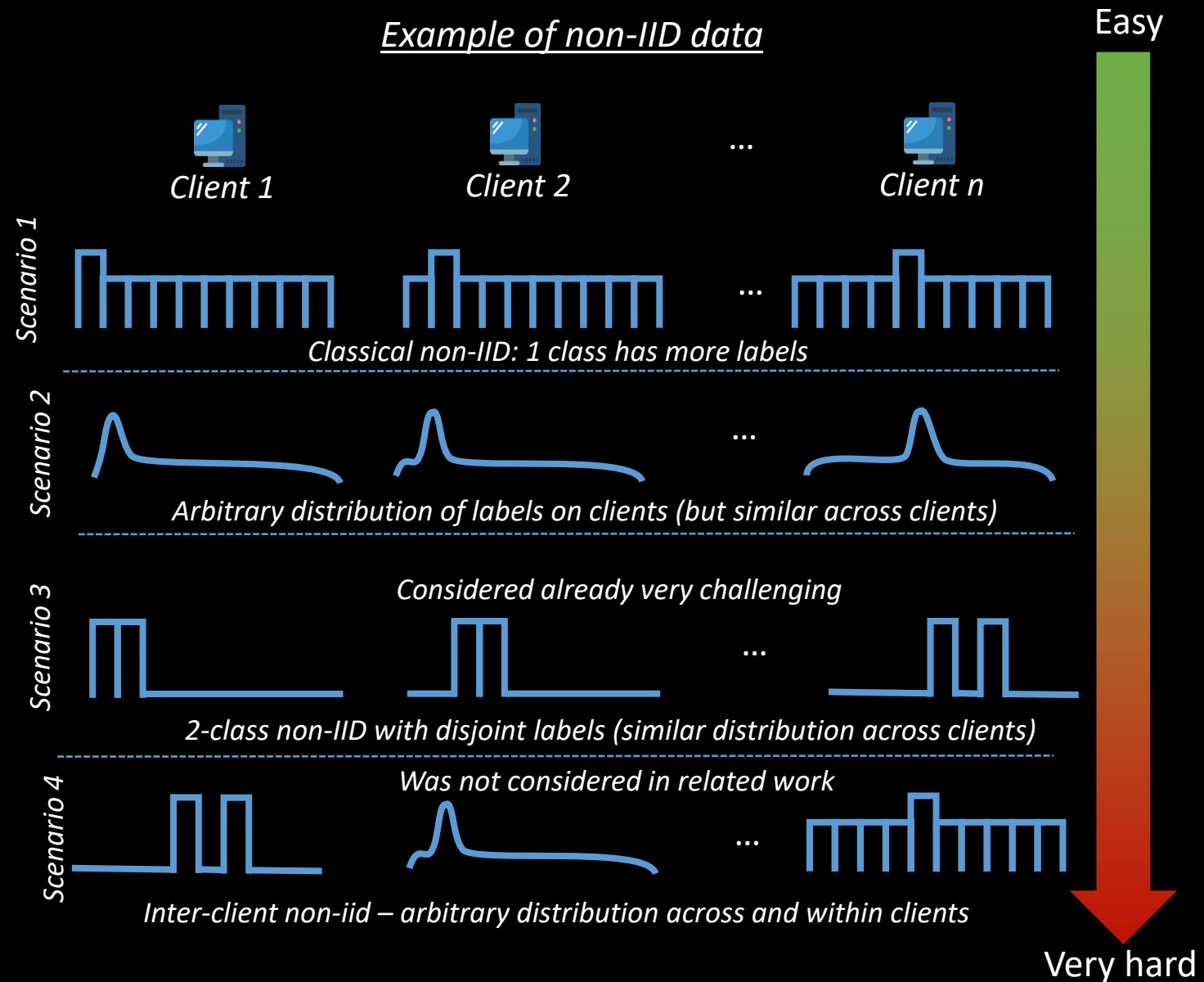
Example of IID data
(nearly uniform distribution)



Prediction classes on one client
(10 classes)

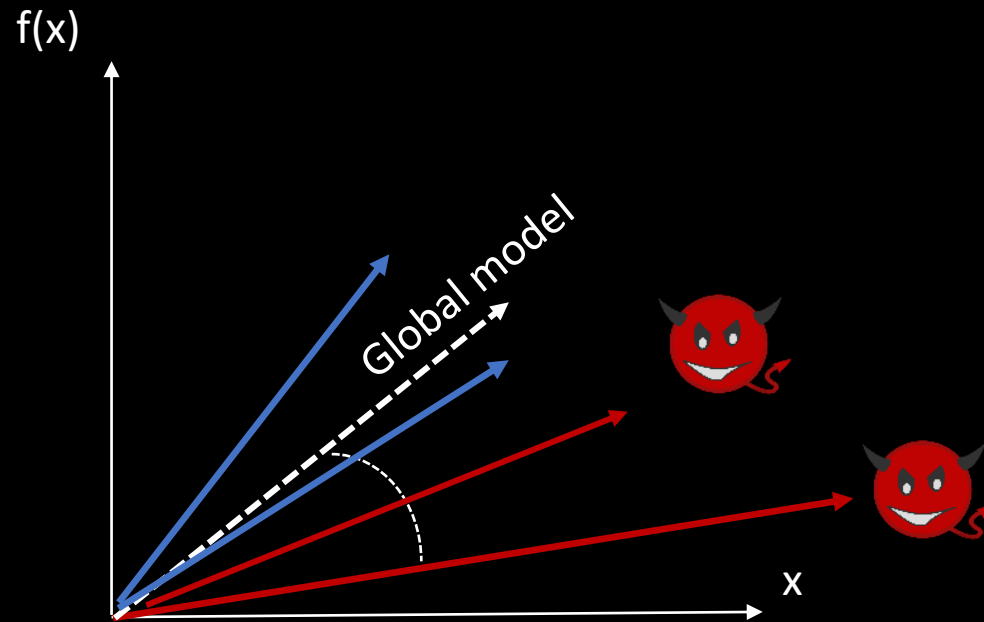


Example of non-IID data



Visualisation of Model Updates

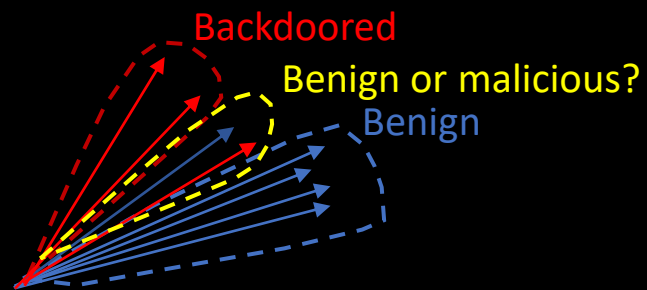
- Let's imagine that the model is a simple linear function $f(x) = ax+b$, where a and b are model parameters



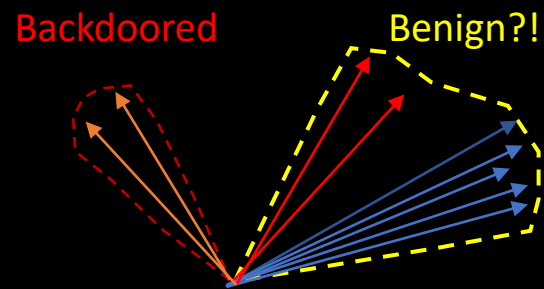
- Malicious models differ from the global model due to the adversary's manipulation
- Benign models differ due non-independent and identically distributed (non-IID) data

- Global model from training round $t-1$
- Benign local models at round t
- Malicious models at round t

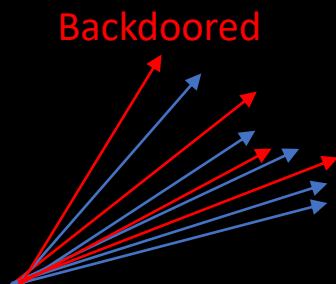
Challenges of Correct Clustering



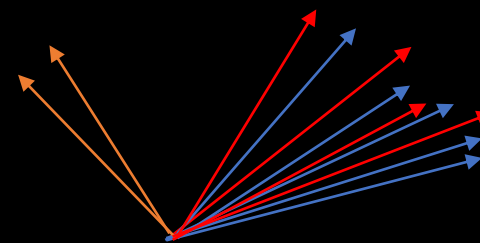
One backdoor & IID data



Multiple backdoors?



One backdoor & non-IID data?



Multiple backdoors & non-IID data?

- Global model from training round $t-1$
- Benign models at round t
- Malicious models at round t

Two Solutions



CrowdGuard

[with Rieger
at al.,
NDSS 2024]



MESAS

[with Krauss.
ACM CCS 2023]



CrowdGuard

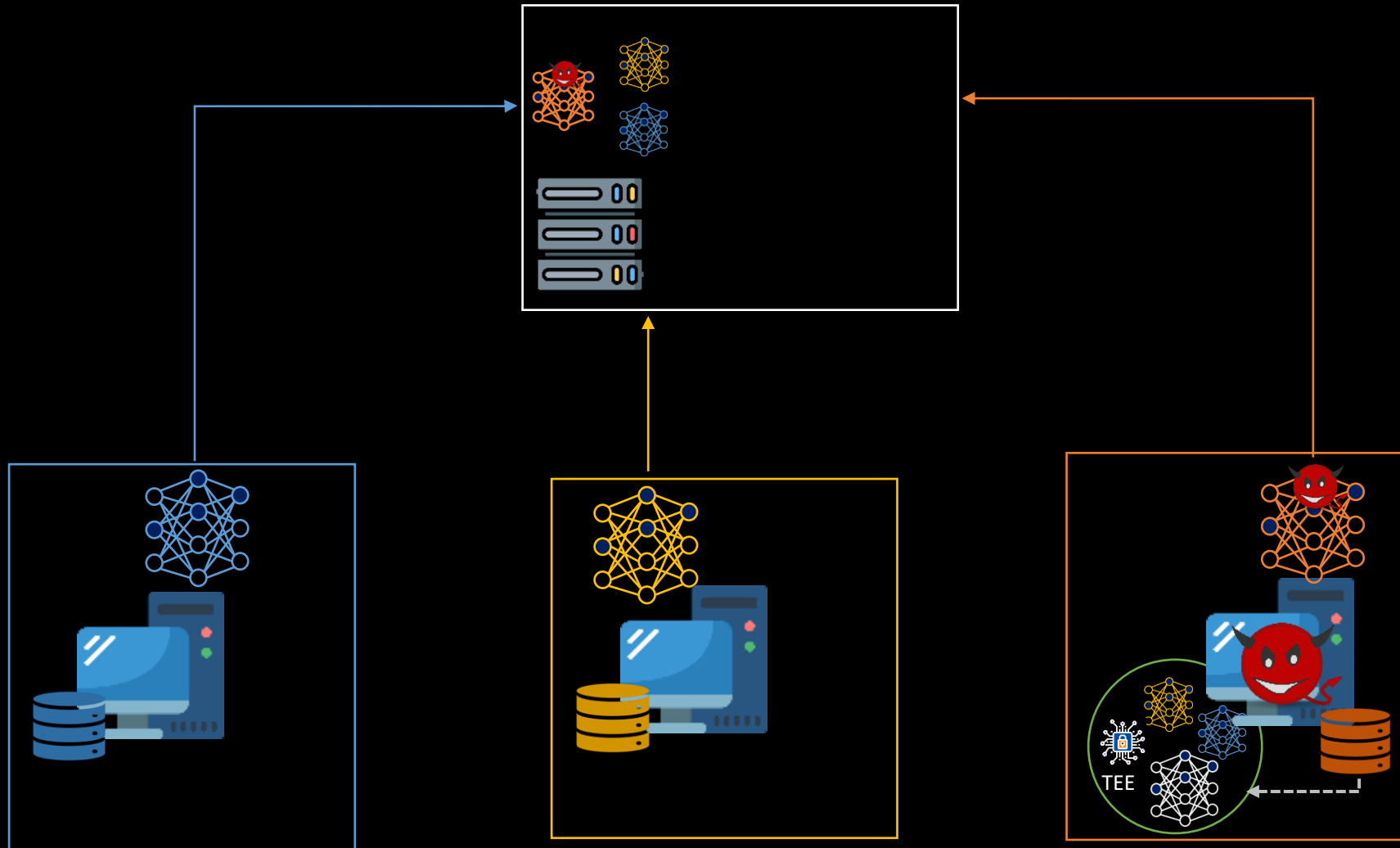
Federated Backdoor Detection in Federated Learning

Philip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko, Ahmad-Reza Sadeghi

Network and Distributed System Security Symposium (NDSS), 2024

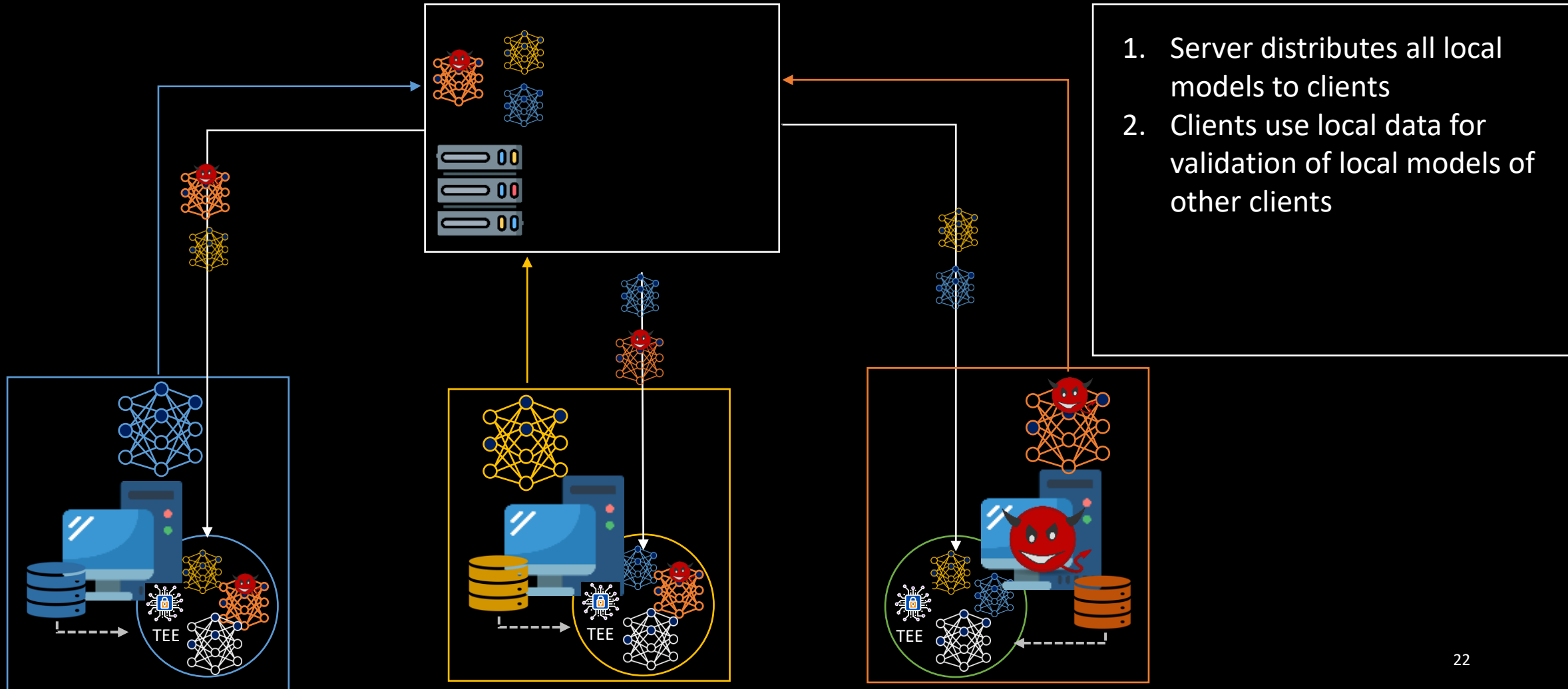
CrowdGuard: Federated Backdoor Detection

- Assumption: $> 50\%$ of clients are benign



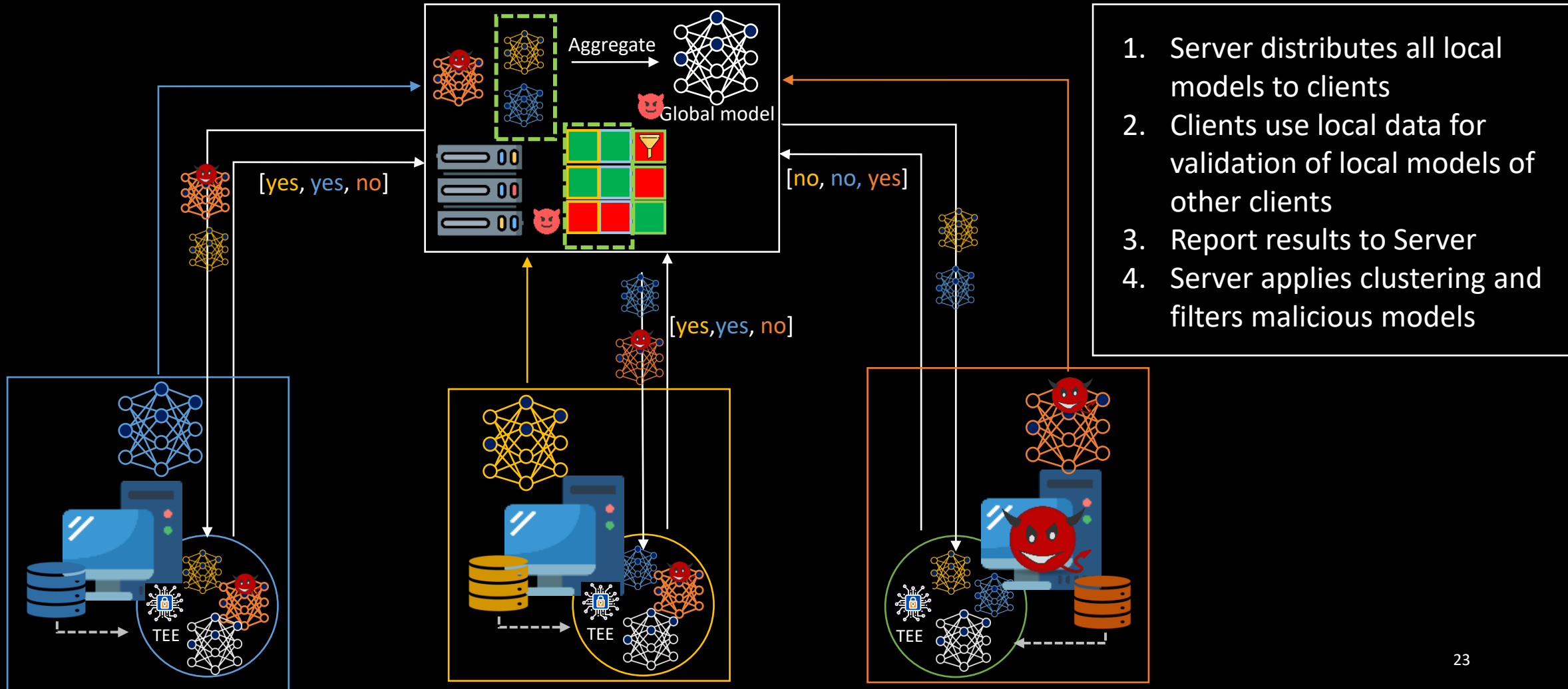
CrowdGuard: Federated Backdoor Detection

- Assumption: > 50% of clients are benign
- Requirement: Analysis/aggregation of local models is performed within Trusted Execution Environment (TEE)



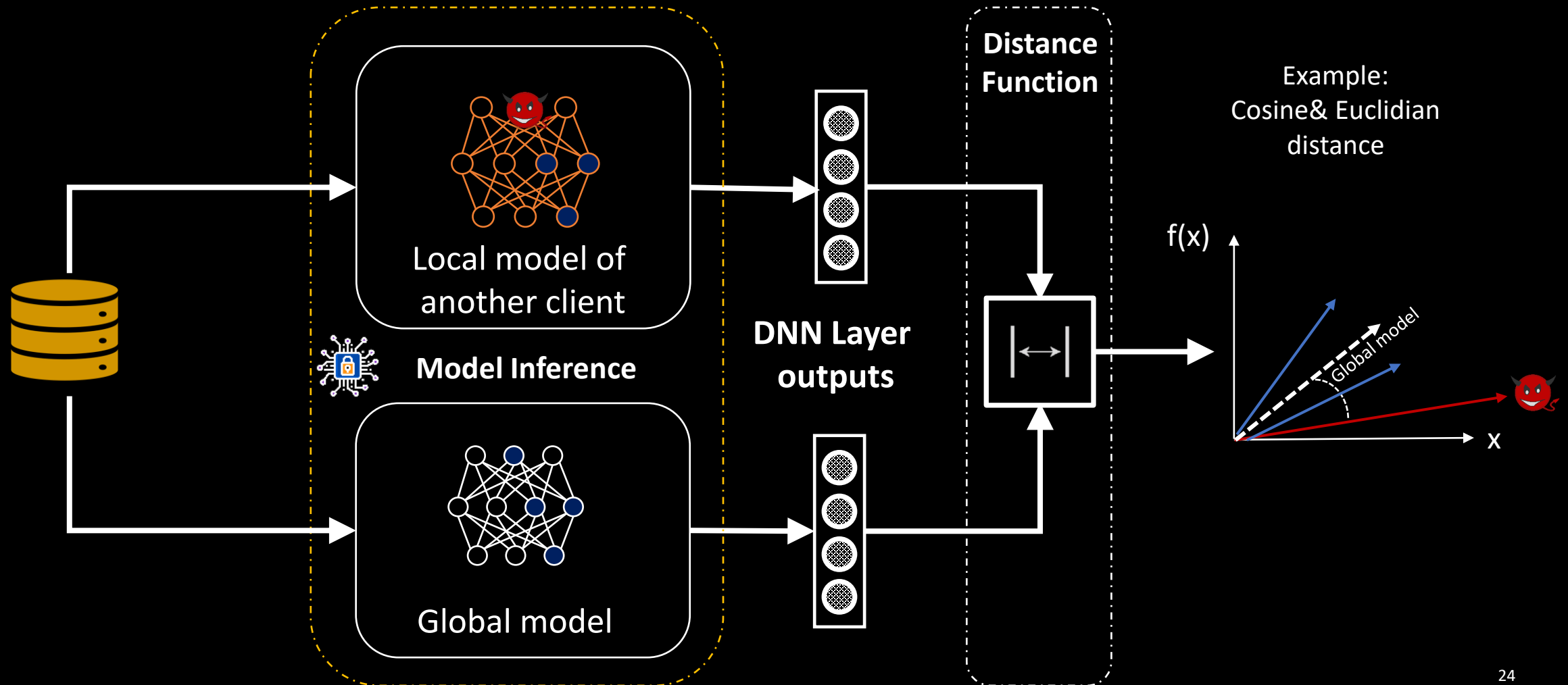
CrowdGuard: Federated Backdoor Detection

- Assumption: > 50% of clients are benign
- Requirement: Analysis/aggregation of local models is performed within Trusted Execution Environment (TEE)



Analyzing Deep Layer Client Predictions

- Repeat for every sample of every label and average results within the label



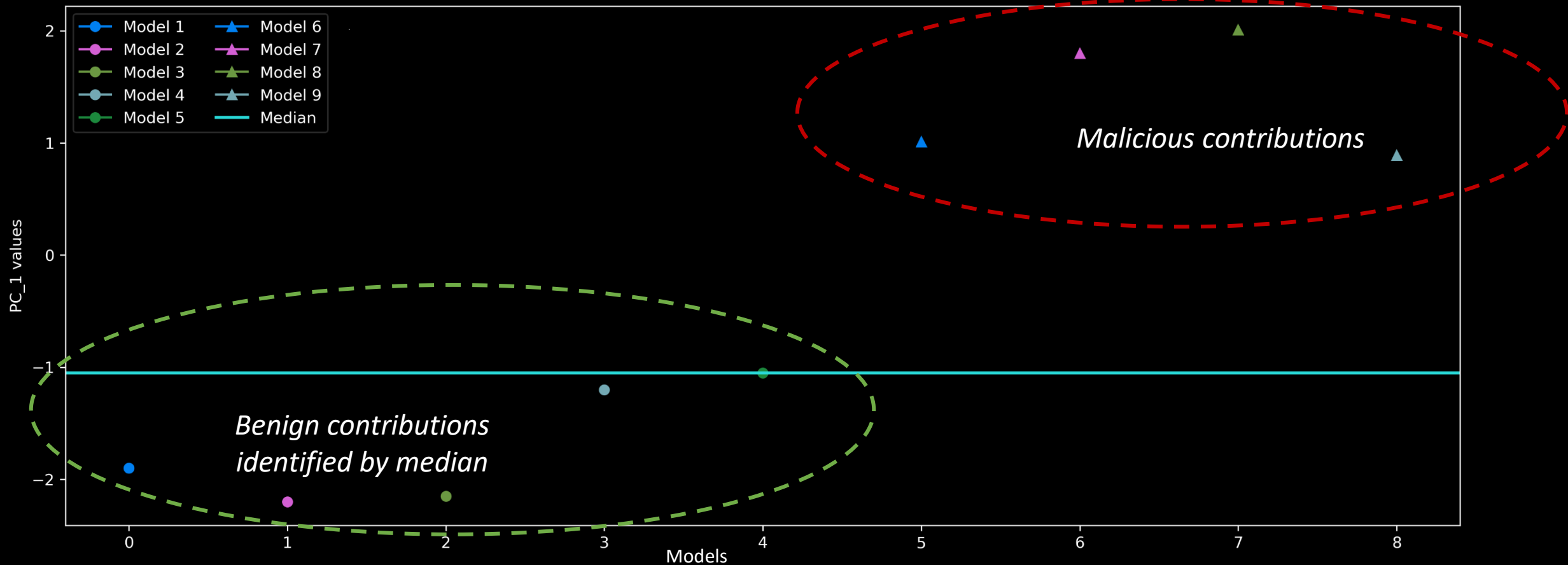
Reducing Dimensionality using Principal Component Analysis (PCA)

Setup: 10 clients (11 benign & 9 malicious) – Analysis on client 0

Values: Principal component 1 values

Metric: Cosine and Euclidian distance of the prediction to the prediction of the Global Model

Benign models are circles, malicious models are triangles. Colors depict main labels.



Results and Findings

Metrics:

- Cosine and Euclidian distance of local model to global model layer outputs
- PCA is effective for dimensionality reduction
- We additionally derive so-called HLBIM metric which helps to separate benign and malicious models more effectively

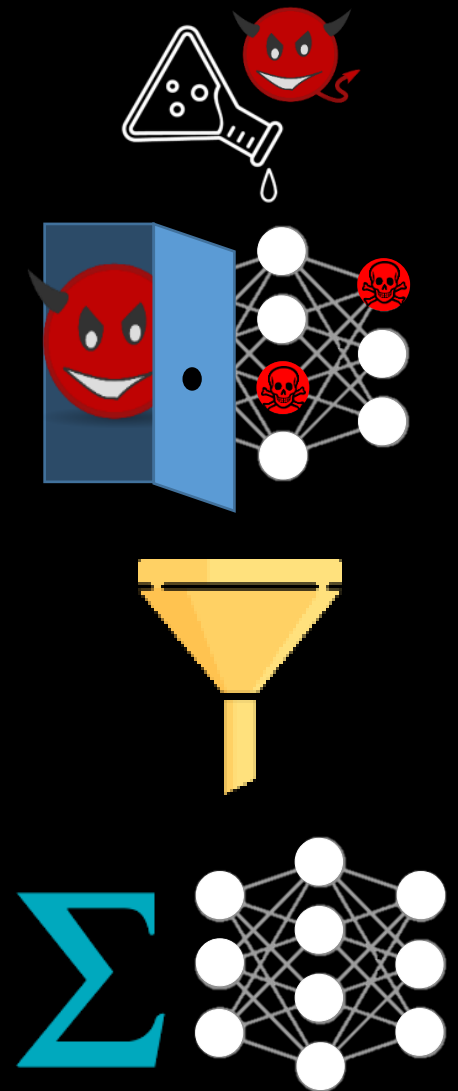
Effectiveness and Advantages:

- 100% True Positive Rate (TPR) and True Negative Rate (TNR) across various scenarios, including IID and non-IID data distribution (scenarios 1-3)
- Per design resilient against adaptive attackers

→ CrowdGuard will be integrated into OpenFL 1.6

Special Considerations:

- Requires usage of Trusted Execution Environments (TEEs)
- Our next work, MESAS, does not require any TEEs on clients!



MESAS

Poisoning Defense for Federated Learning Resilient
against Adaptive Attackers

Torsten Krauss and Alexandra Dmitrienko

ACM Conference on Computer and Communications Security (CCS), 2023

MESAS: Metric – Cascades for Poisoning Detection

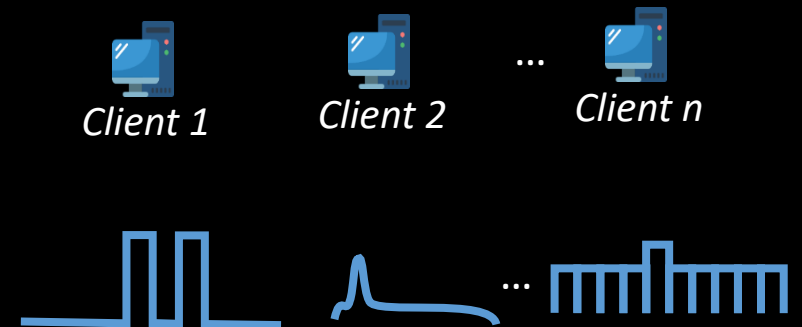
Goals:

- Support arbitrary non-IID client datasets (including scenario 4)
- Prevent attackers from adapting to the defense without relying on TEEs

Idea:

- Use many metrics for detection of poisoned models at the same time
- Intuition: For an adaptive attacker, it should be harder (if at all possible?) to adapt to many metrics

*The most challenging non-IID scenario:
Arbitrary distribution between and across clients*



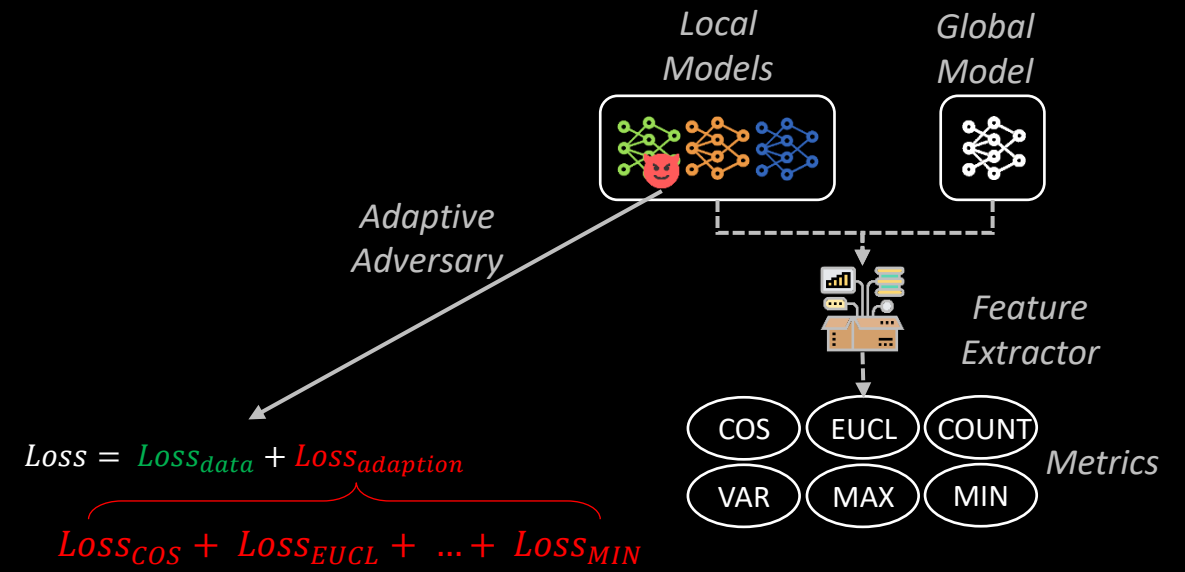
Classical Adaptive Adversary

$$Loss = Loss_{data} + Loss_{adaption}$$

MESAS Approach

Approach:

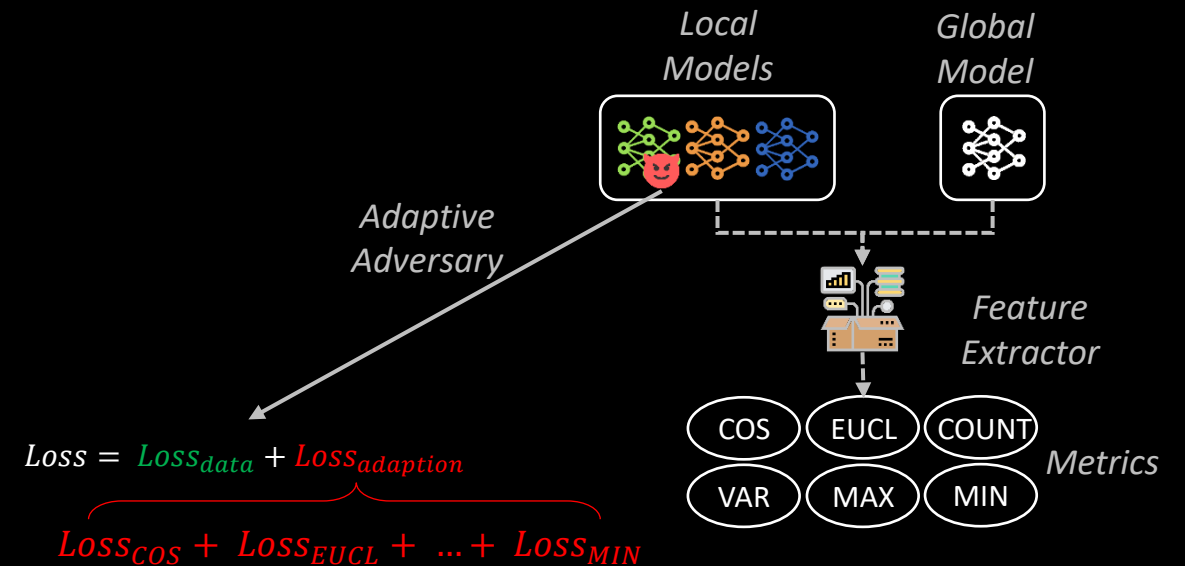
- Detection and pruning based on six well-chosen metrics
- Force the attacker into a heavy multi-objective optimization problem
 - Hardening the adversarial dilemma between backdoor effectiveness and stealthiness



MESAS Approach - Metrics

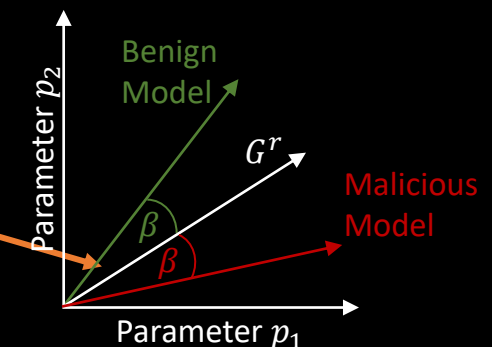
COS & EUCL:

- Cosine & Euclidean distance between Global and Local Models



COUNT:

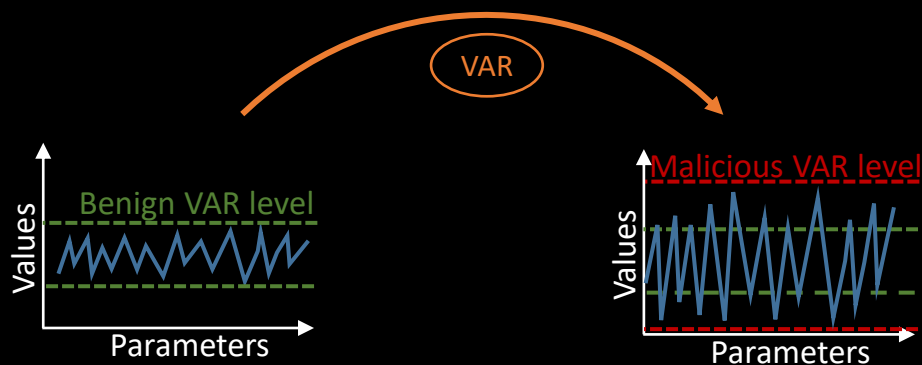
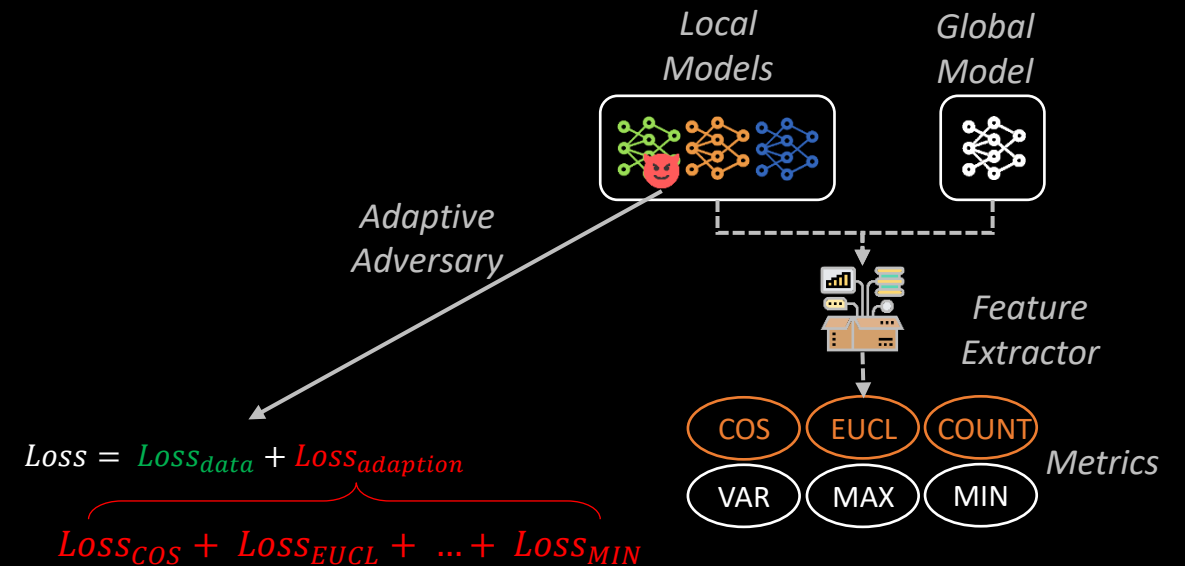
- Reason: Same COS (β) for different models possible
- Number of parameters that are increased



MESAS Approach - Metrics

VAR:

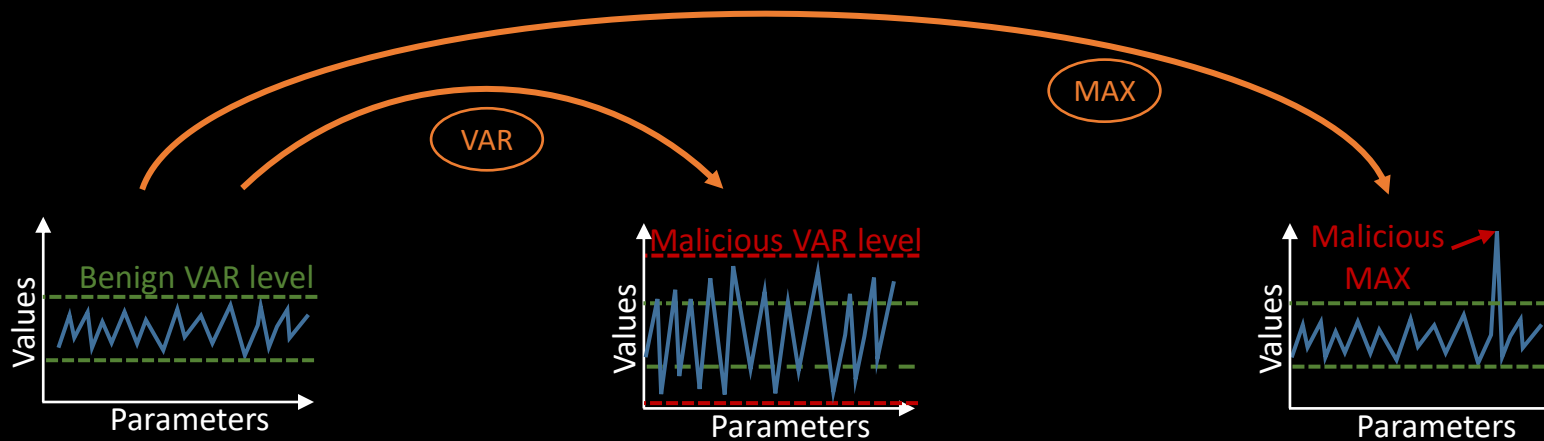
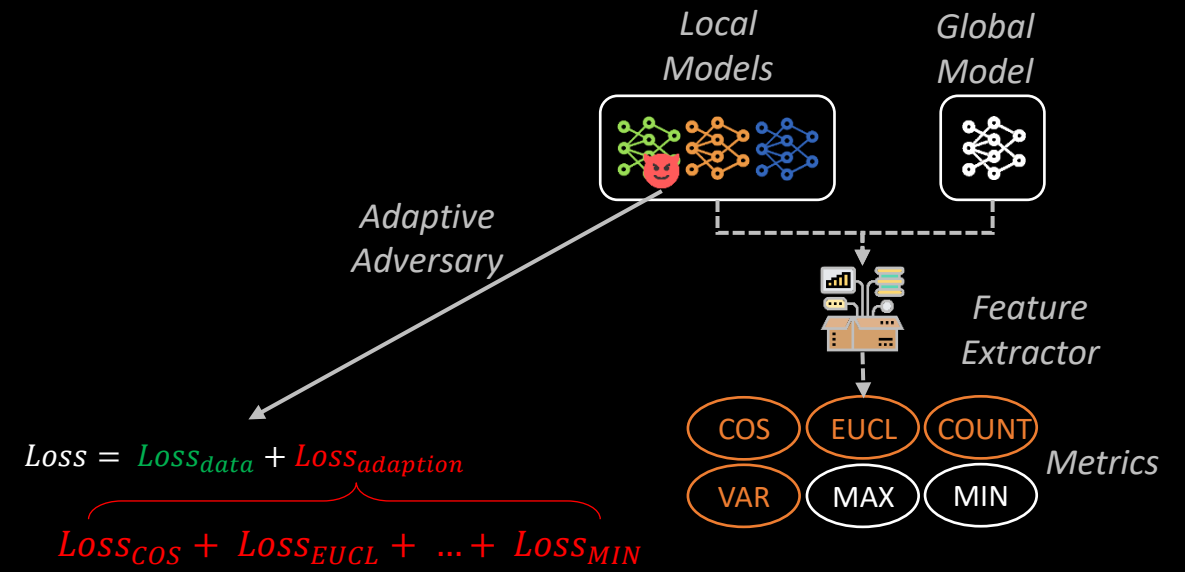
- COS, EUCL, and COUNT can look benign, but still a backdoor can be embedded
- Adversary could increase the variance of updates



MESAS Approach - Metrics

MIN & MAX:

- Variances in general are not heavily influenced by extreme outliers
- An adversary could embed a backdoor into outliers



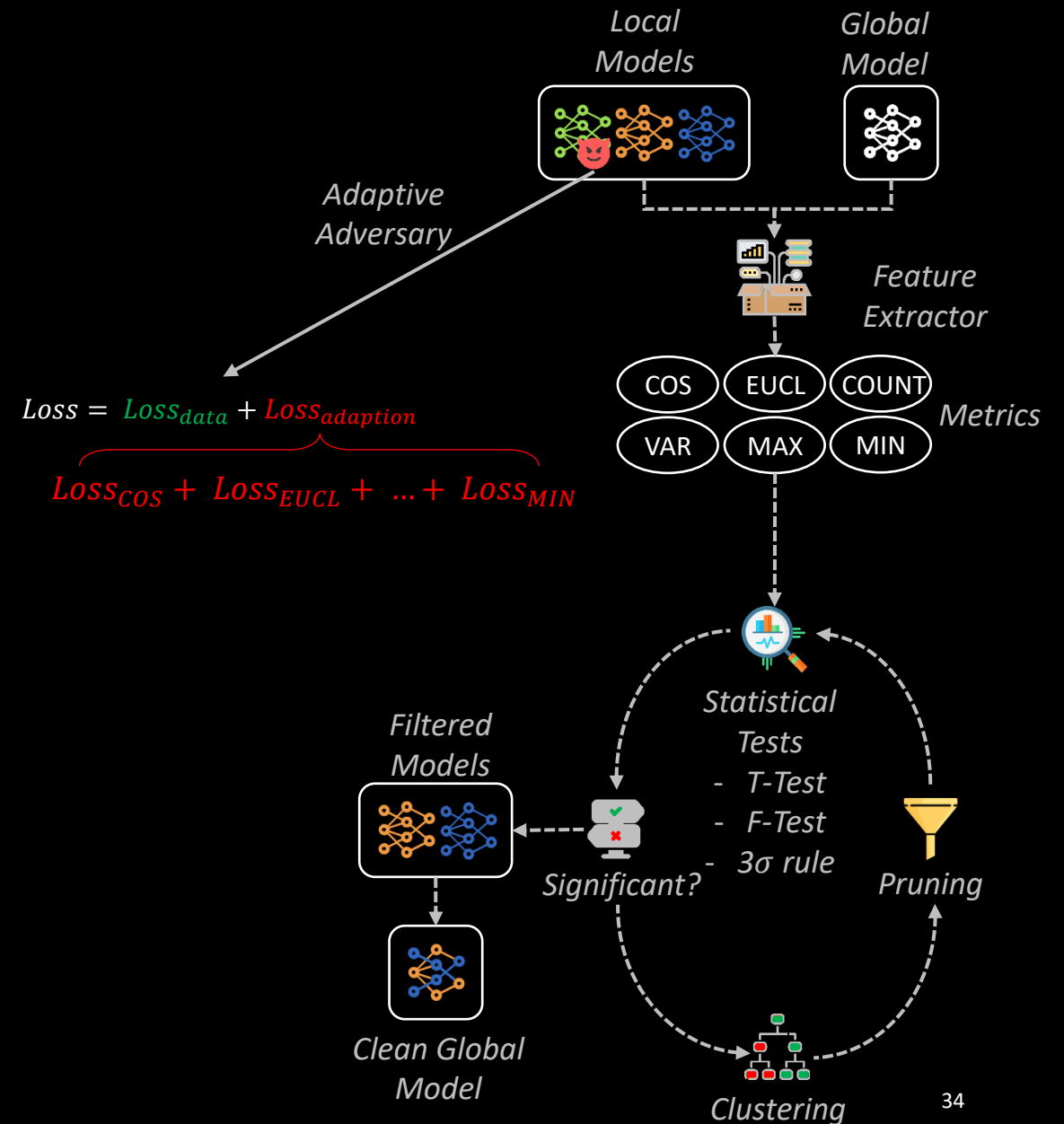
MESAS Approach

Approach – Step 1:

- Extract six metrics

Approach – Step2:

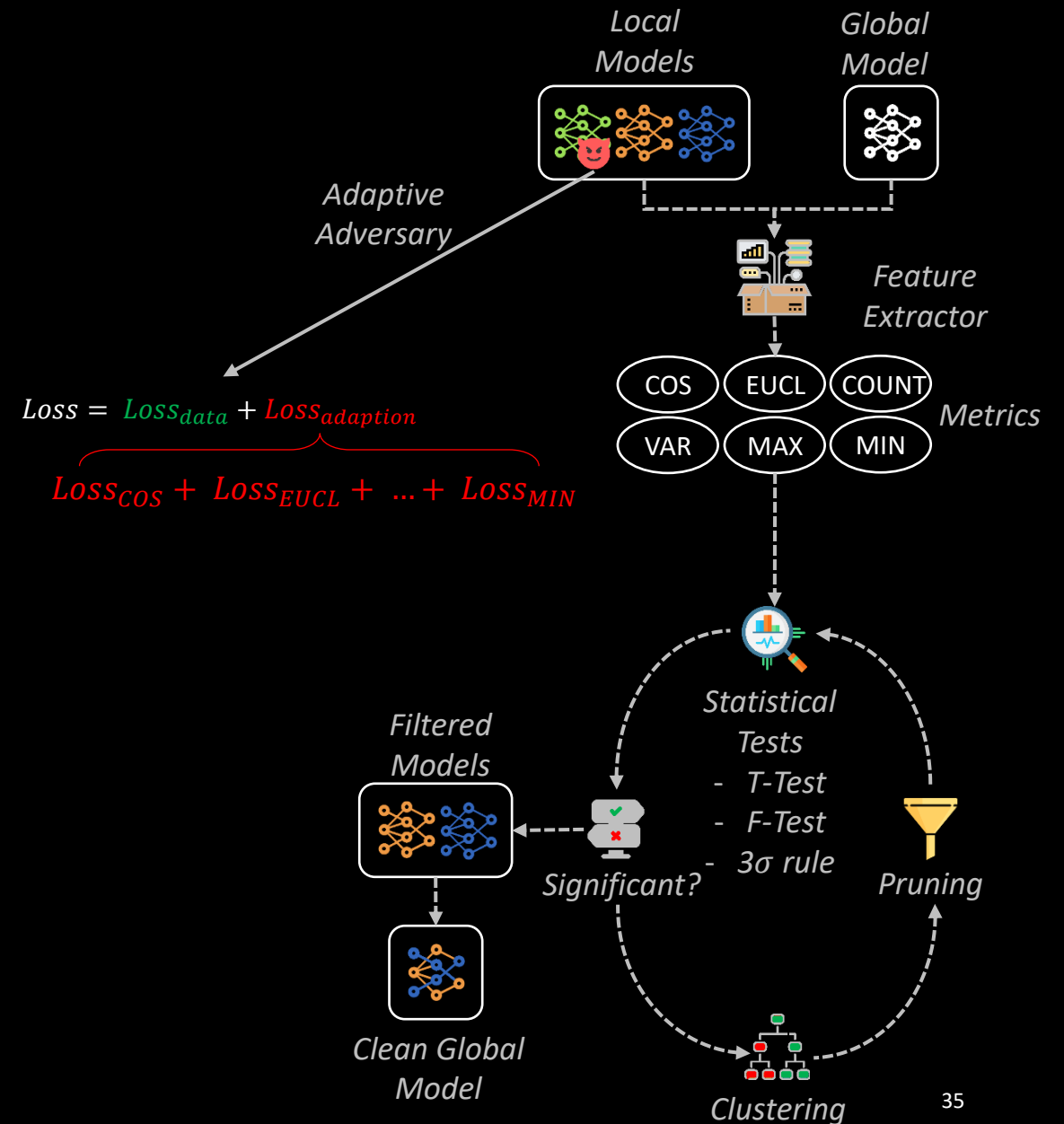
- Iterative pruning loop leveraging statistical tests and clustering to detect poisoned models



MESAS Results

Evaluation:

- Metrics have mutual effects during adaptation
- We demonstrate empirically that an attacker cannot adapt to all of them at the same time
- It works even for the most challenging non-IID scenario with arbitrary distribution across clients!



CrowdGuard vs. MESAS Comparison

	CrowdGuard	MESAS
What is analyzed?	Prediction layer outputs	Local models
Where the analysis is performed?	Clients	Server
Utilized metrics	Cosine & Euclidian distances between global and local models	Six metrics: Cosine & Euclidian distances, COUNT, Variance, Outliers (MIN & MAX)
Resilience against adaptive attacker	Resilient per design	Demonstrated empirically
Non-IIDness	Scenarios 1-3	Scenarios 1-4
Additional requirements	TEE on clients	-

Conclusion & Further Research



- AI-based algorithms find wide adaptation in many areas, including **security-critical applications**

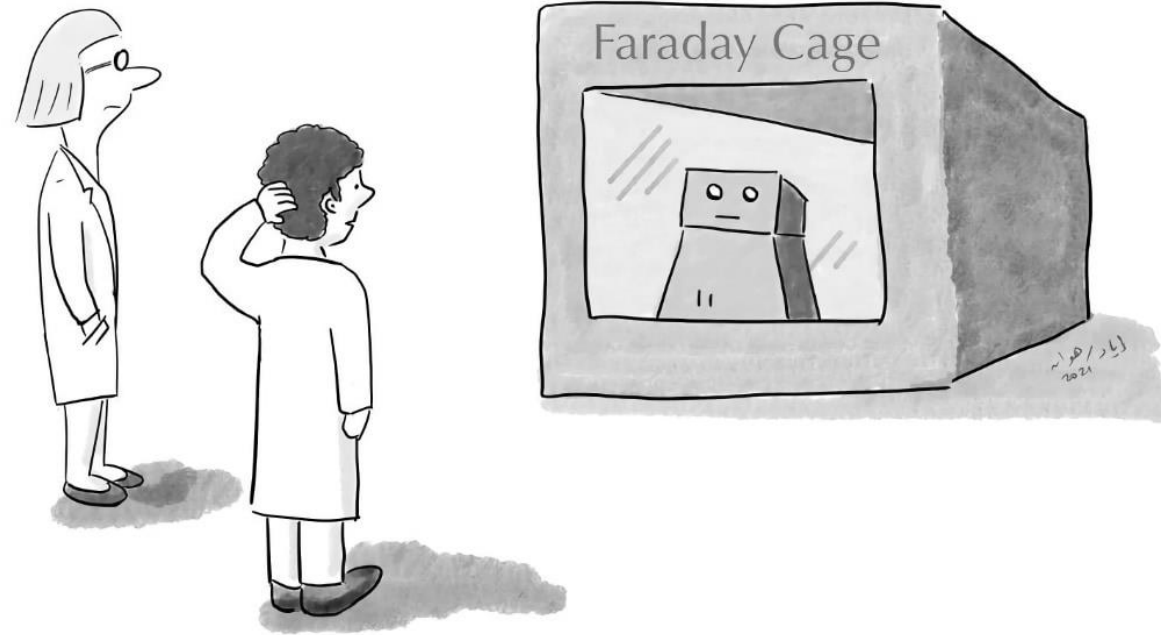


- AI algorithms are prone to untargeted and targeted **poisoning attacks**



- MESAS and CrowdGuard significantly advance the state of the art
- Yet each of them has own limitations that could be overcome in future works

EvilAI Cartoons.com @EvilAICartoons



*“If we let it out, there’s an 85% chance it would cure cancer.
But there’s also a 0.01% chance it takes over the world!”*

<https://www.evilaicartoons.com/archive>