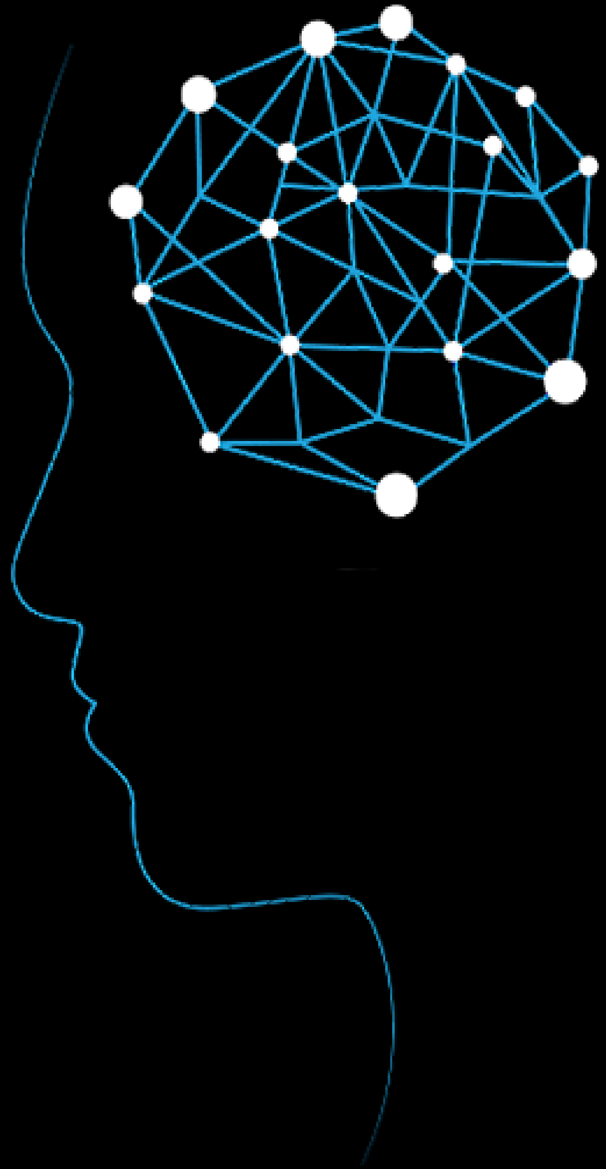
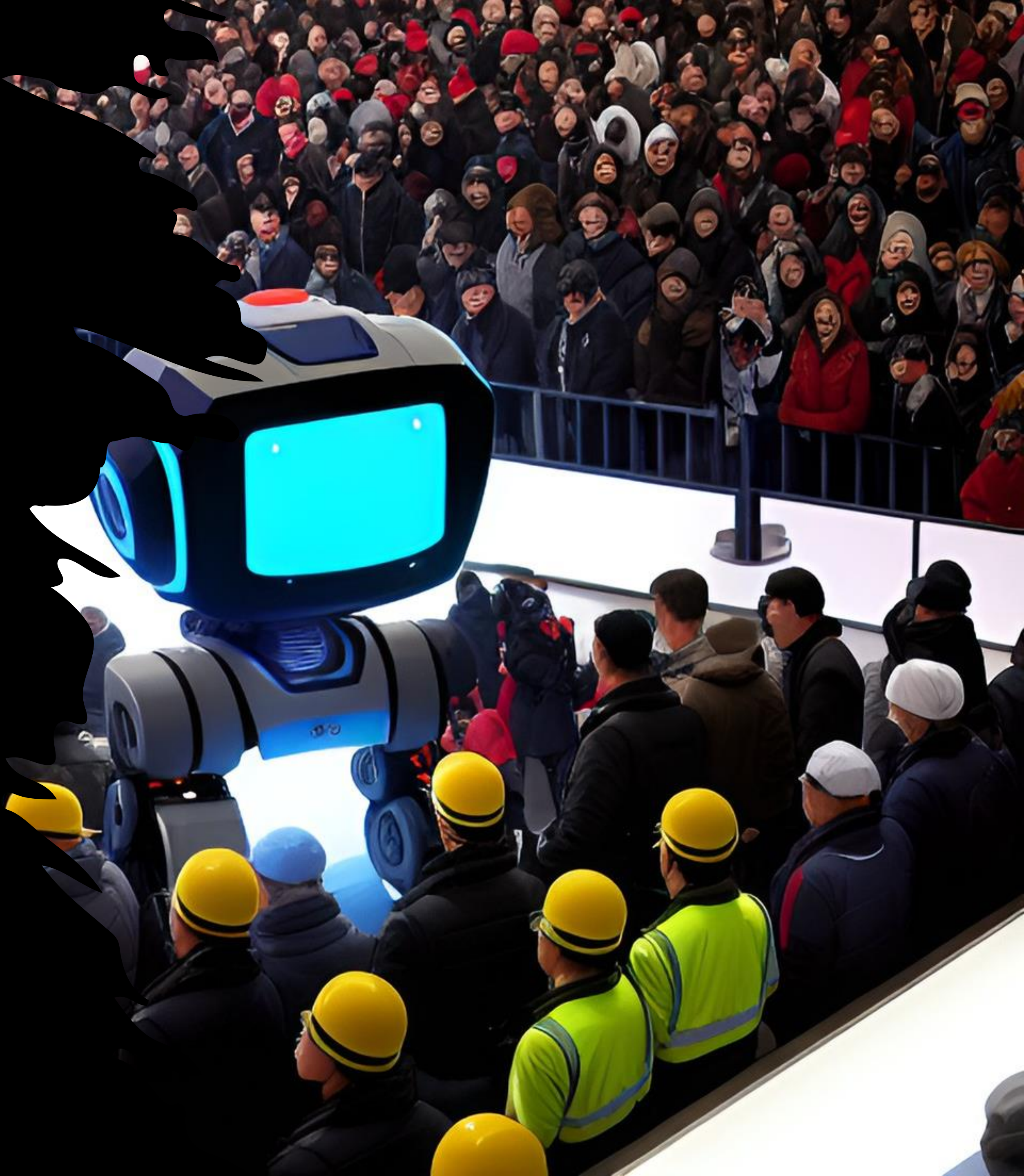
Secure Software Systems

From Advantages to Adversaries: Safeguarding Security in Federated Machine Learning

Alexandra Dmitrienko,
Julius Maximilians Universität Würzburg



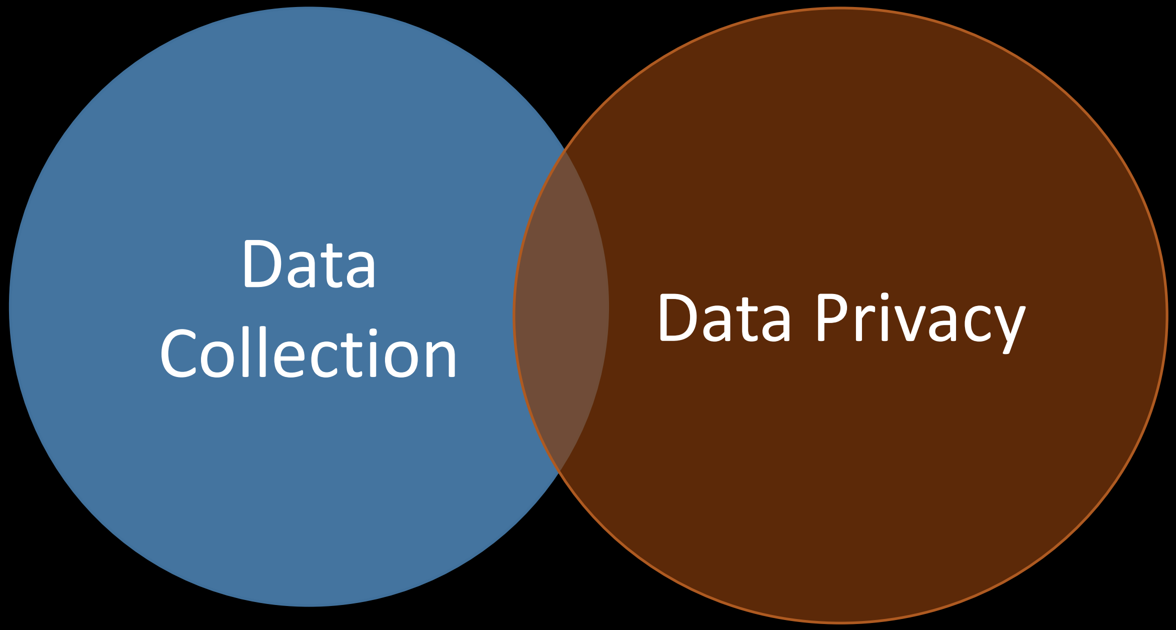
The AI Pandemic



Privacy Challenge of AI

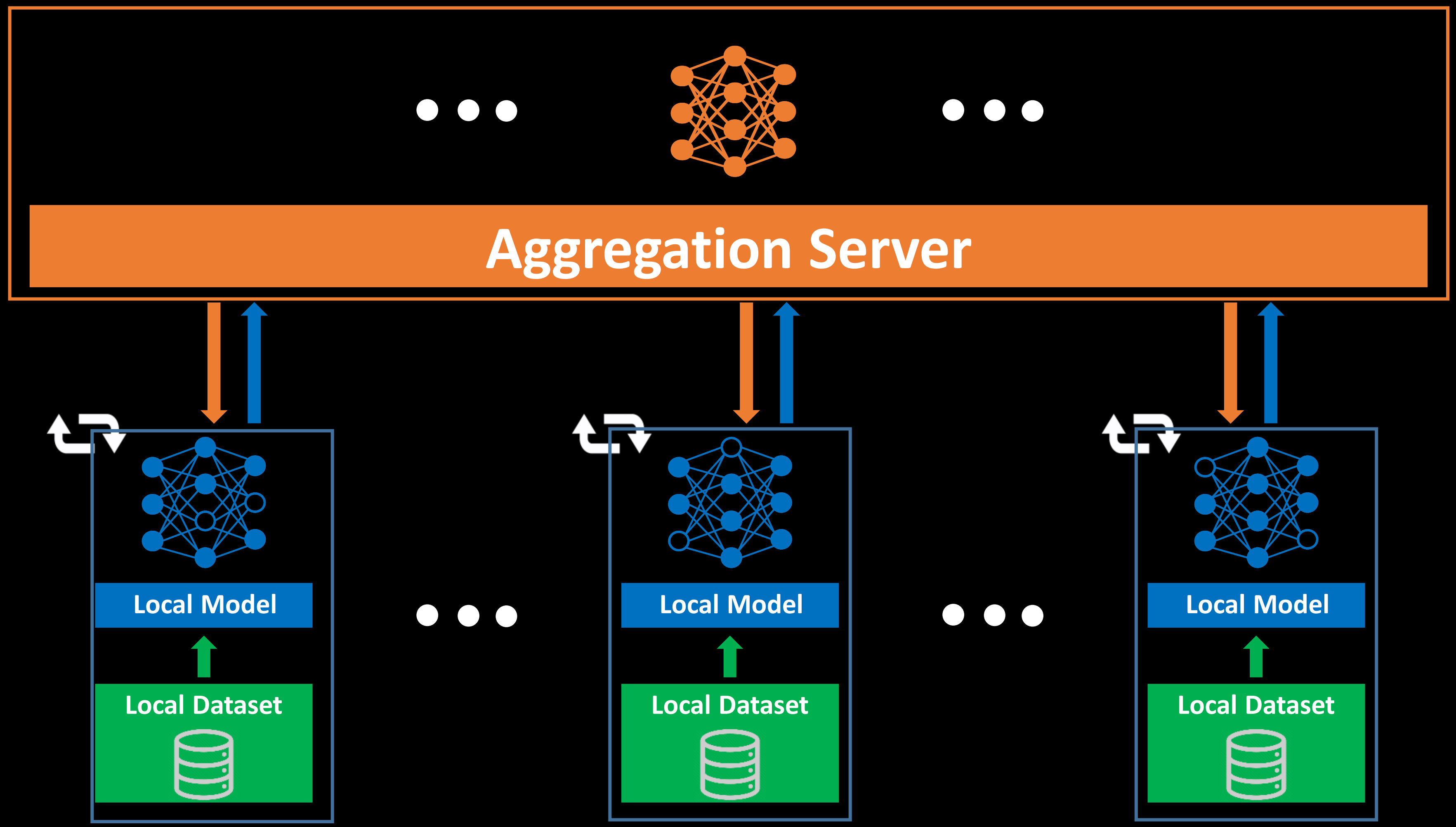
Data-hungry AI

Requirement on large-scale data collection contradicts privacy requirements



Federated Learning can help!

Federated Learning Training

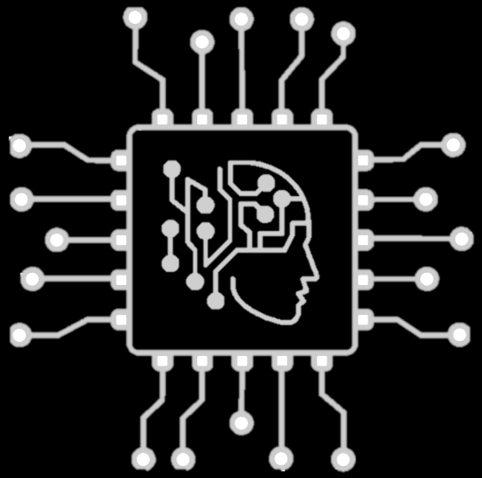


Promised Benefits of Federated Learning

User Privacy

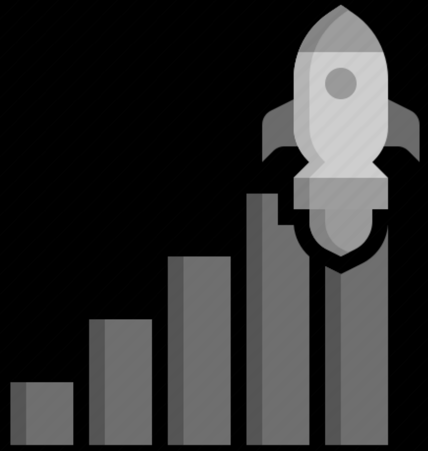
[McMahan et. al PMLR 2017]

Hardware Efficiency



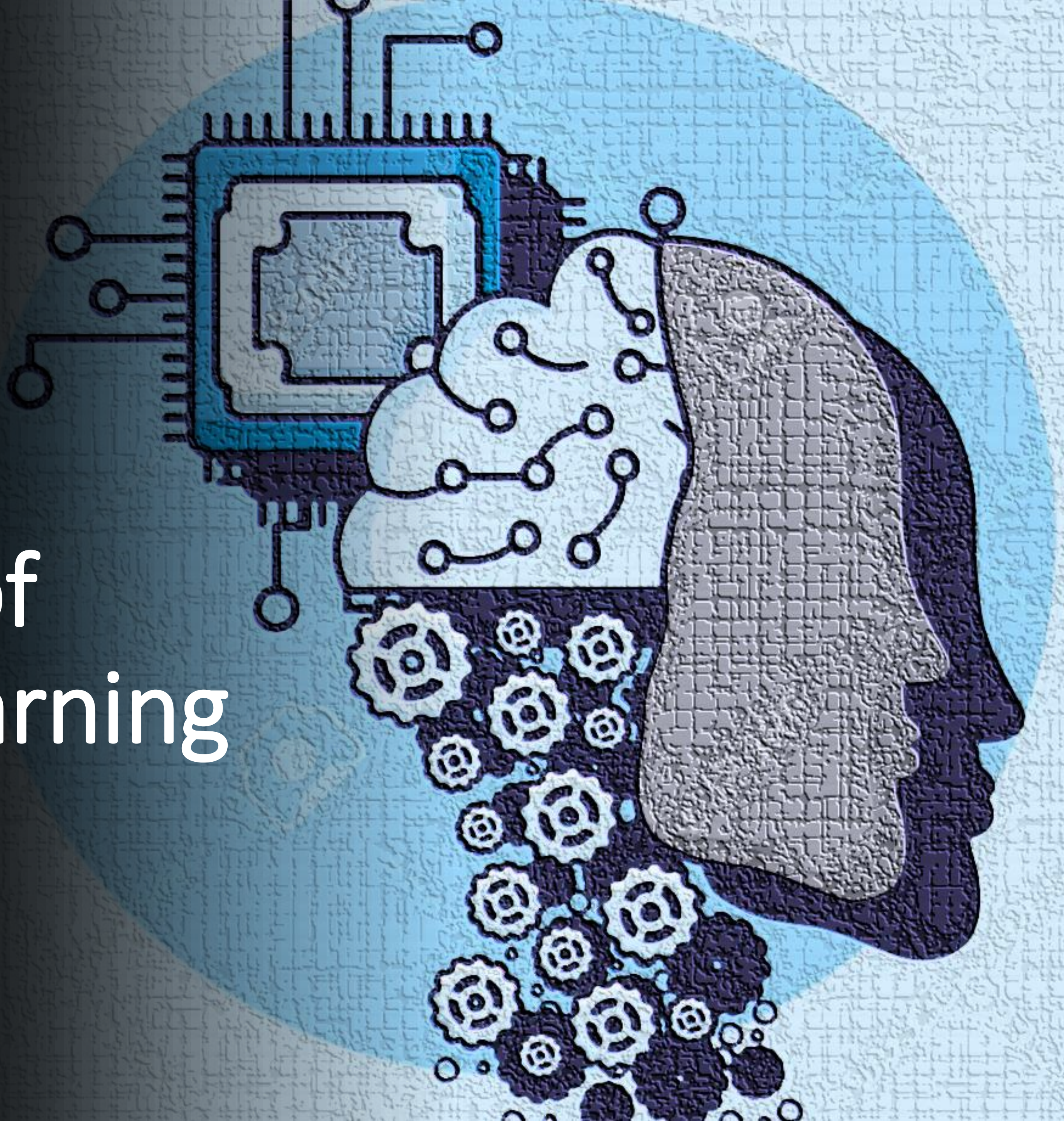
[Kairouz et. al arXiv 2019]

Performance Boosting



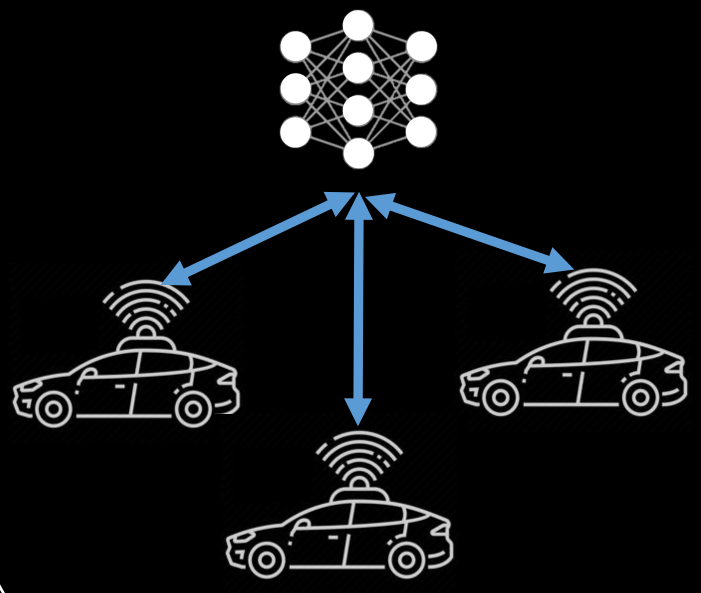
[Fereidooni et. al NDSS 2022]

Applications of Federated Learning



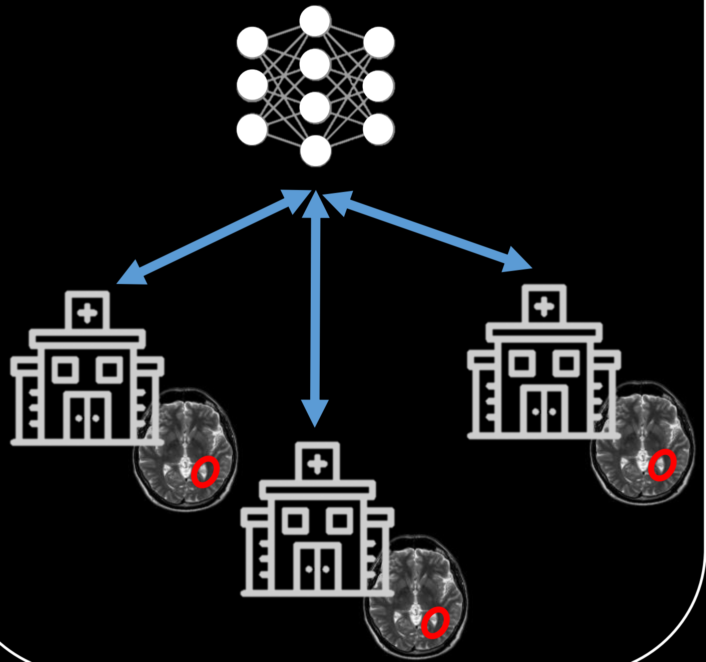
Examples of Federated Learning Applications

Autonomous Driving
Improve object recognition



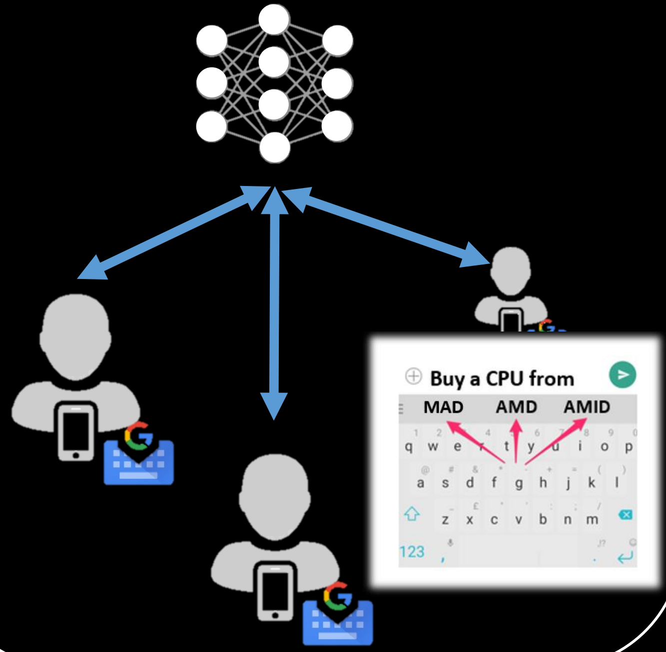
[Jallepalli et. al IEEE BigDataService 2021]

Medical Assistance
Collaboratively learn Brain Tumor Segmentation



[Intel & Pennsylvania[1]]

Word Suggestion
Train word suggestion



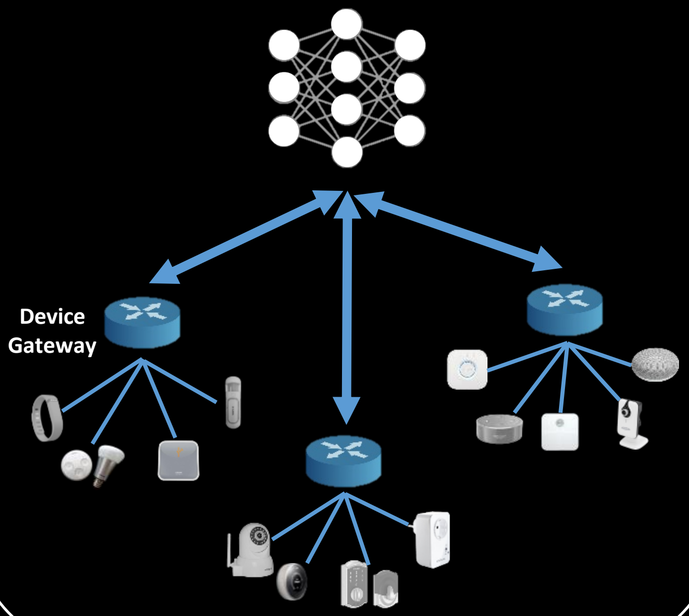
[Hard et. al arXiv 2018]

---

[1] https://newsroom.intel.com/news/intel-works-university-pennsylvania-using-privacy-preserving-ai-identify-brain-tumors

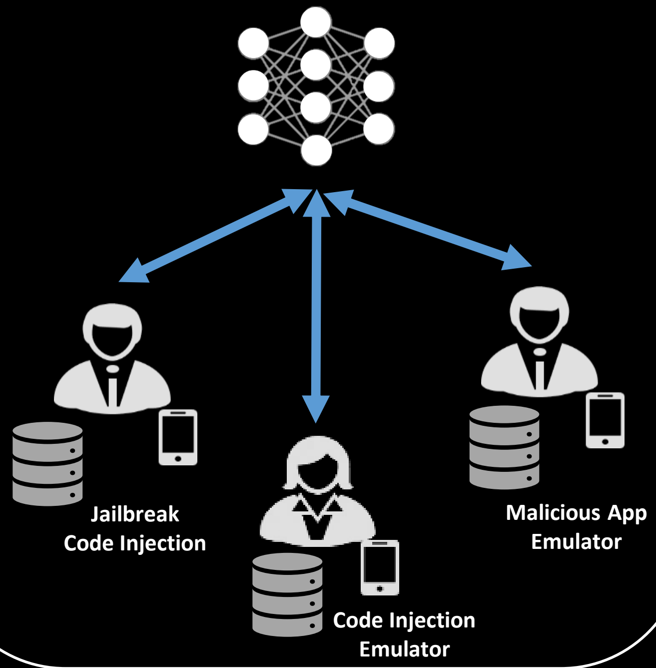
Examples of Federated Learning Applications

IoT Anomaly Detection System
Improve detecting of compromised IoT devices



[Nguyen et. al ICDCS 2019]

Our work:
Sharing Cyber-Risk Intelligence
Improve risk detection & management



[Fereidooni et. al NDSS 2022]

Sharing Cyber-Risk Intelligence

FedCRI: Federated Mobile Cyber-Risk Intelligence

Hossein Fereidooni[1], Alexandra Dmitrienko[2], Phillip Rieger[1], Markus Miettinen[1], Ahmad-Reza Sadeghi[1], and Felix Madlener[3]

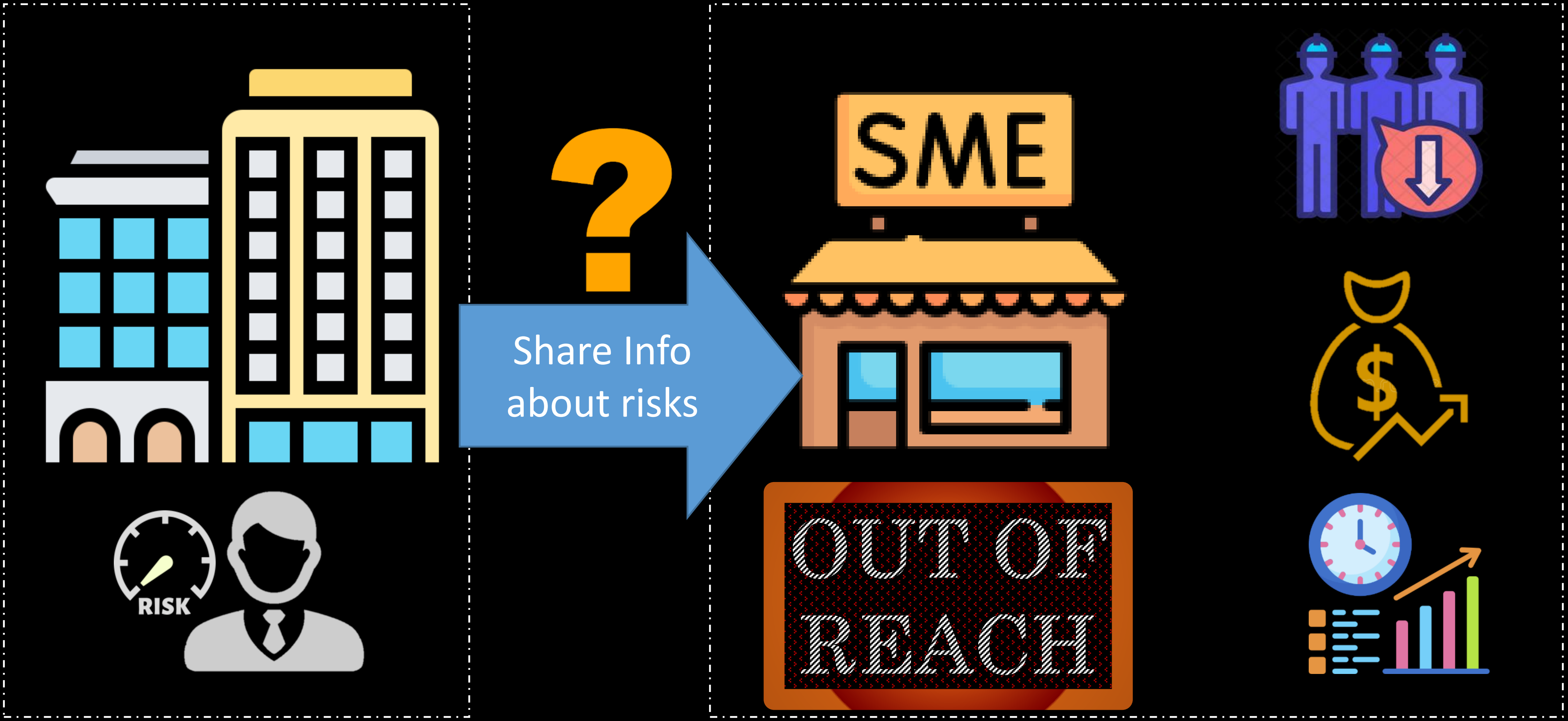
[1]TU Darmstadt, [2]Uni Wuerzburg, [3]KOBIL GmbH

Network and Distributed Security Symposium (NDSS), 2022

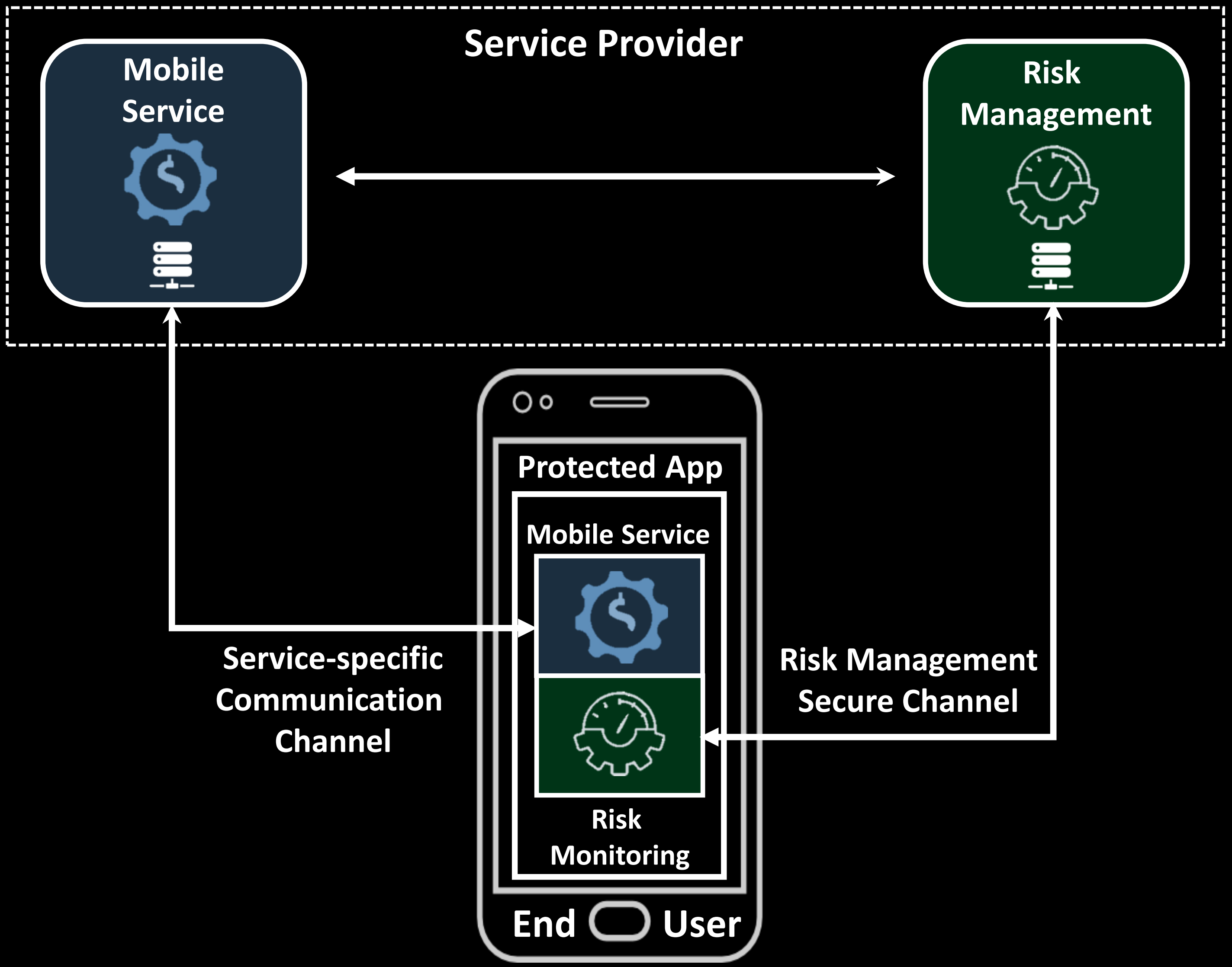
Rapid Growth of Mobile Services

Rapid Growth of Mobile Services

Problem Statement



State-of-the-art: Risk Analysis Frameworks



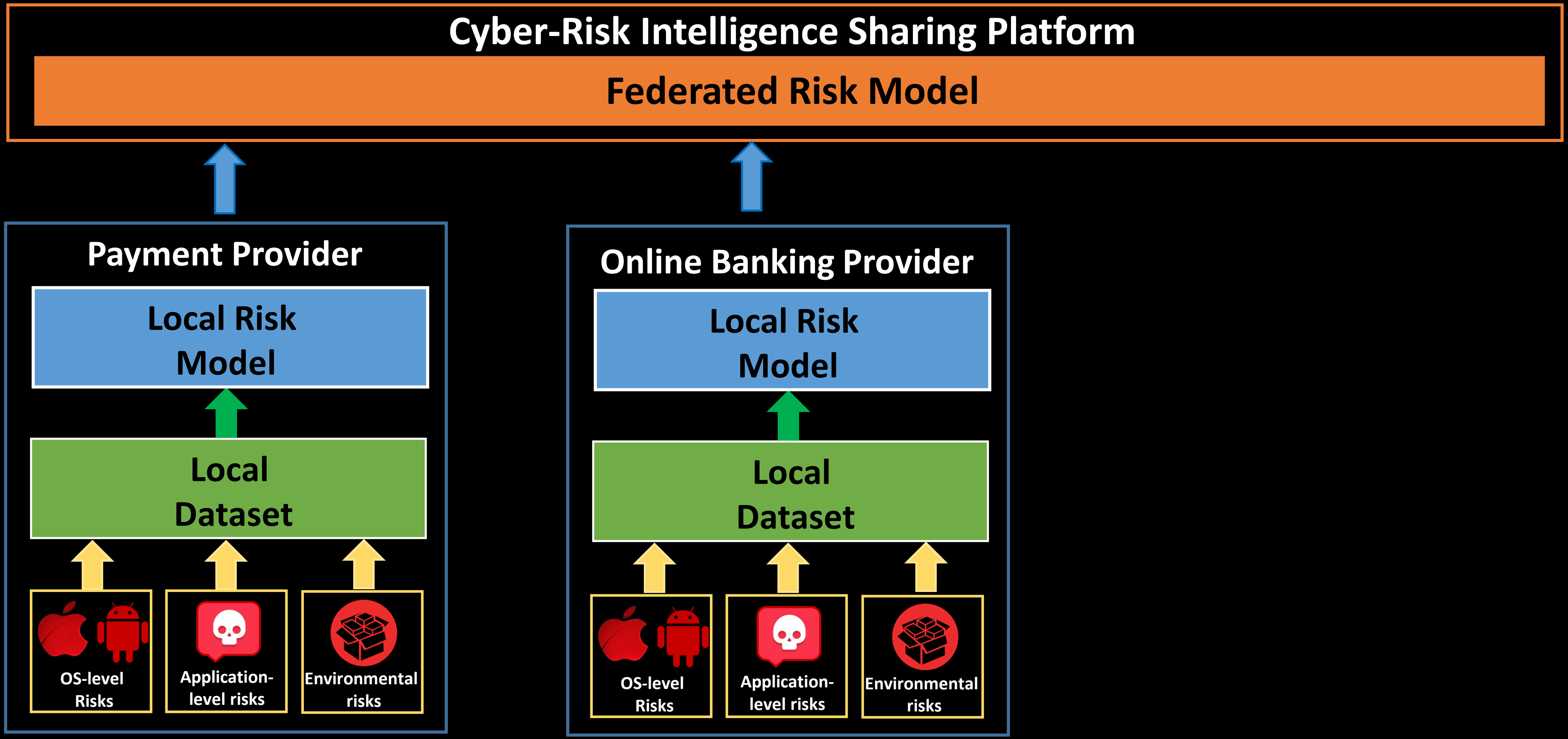
Risk Categories

OS-level Risks
(Jailbreak/Rooted)
(Code Injection)

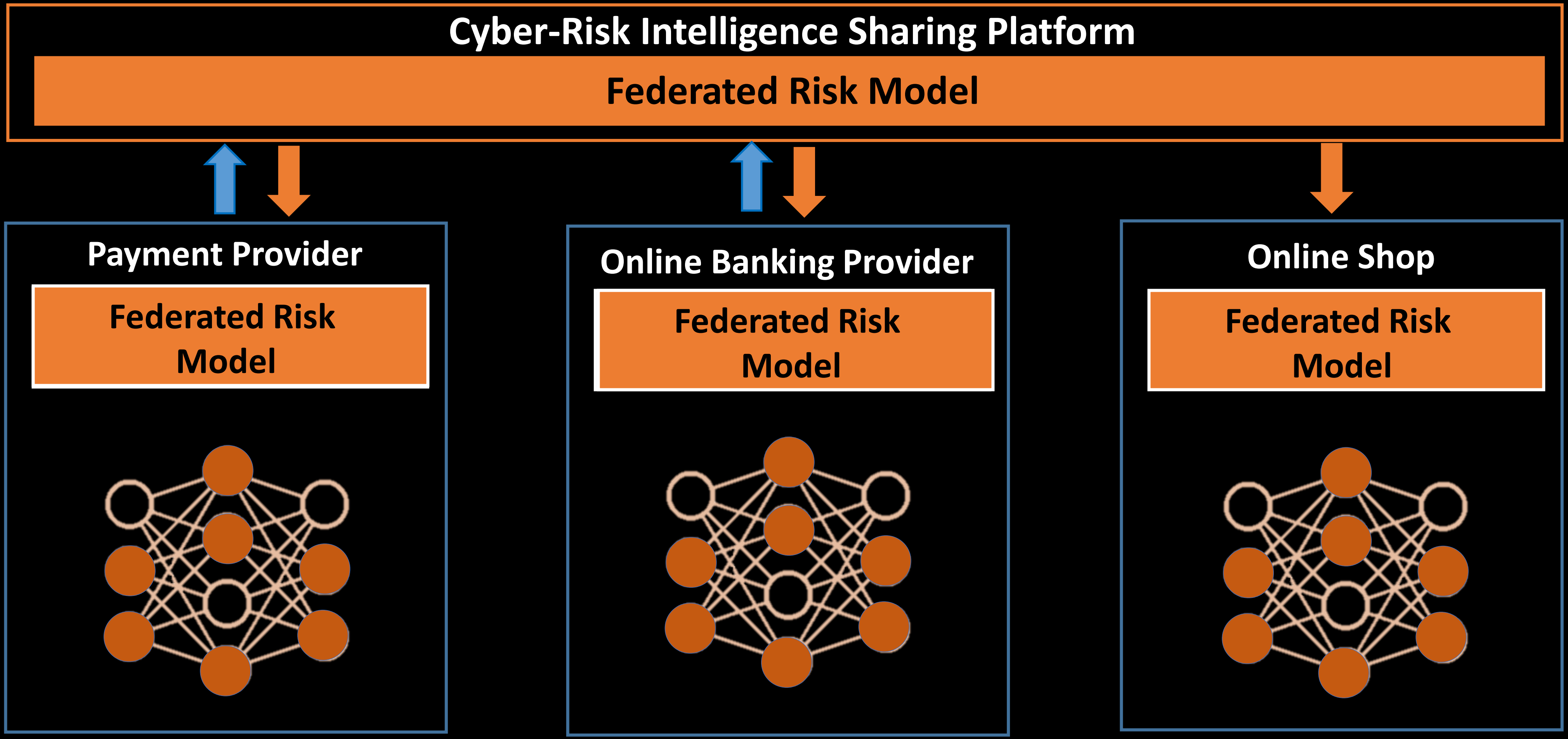
Application-level risks
(app permissions)

Environmental risks
(Emulator/VM)

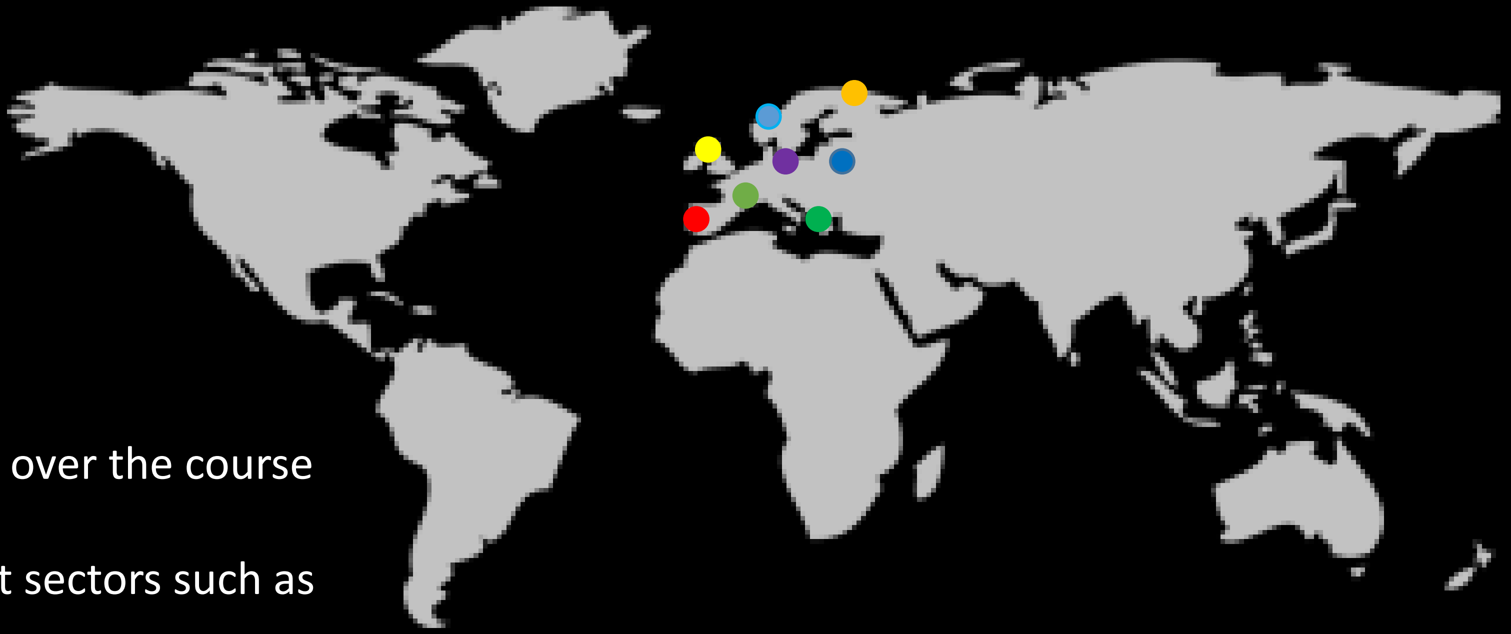
Federated Cyber-Risk Intelligence (FedCRI) Platform



Federated Cyber-Risk Intelligence (FedCRI) Platform



Dataset



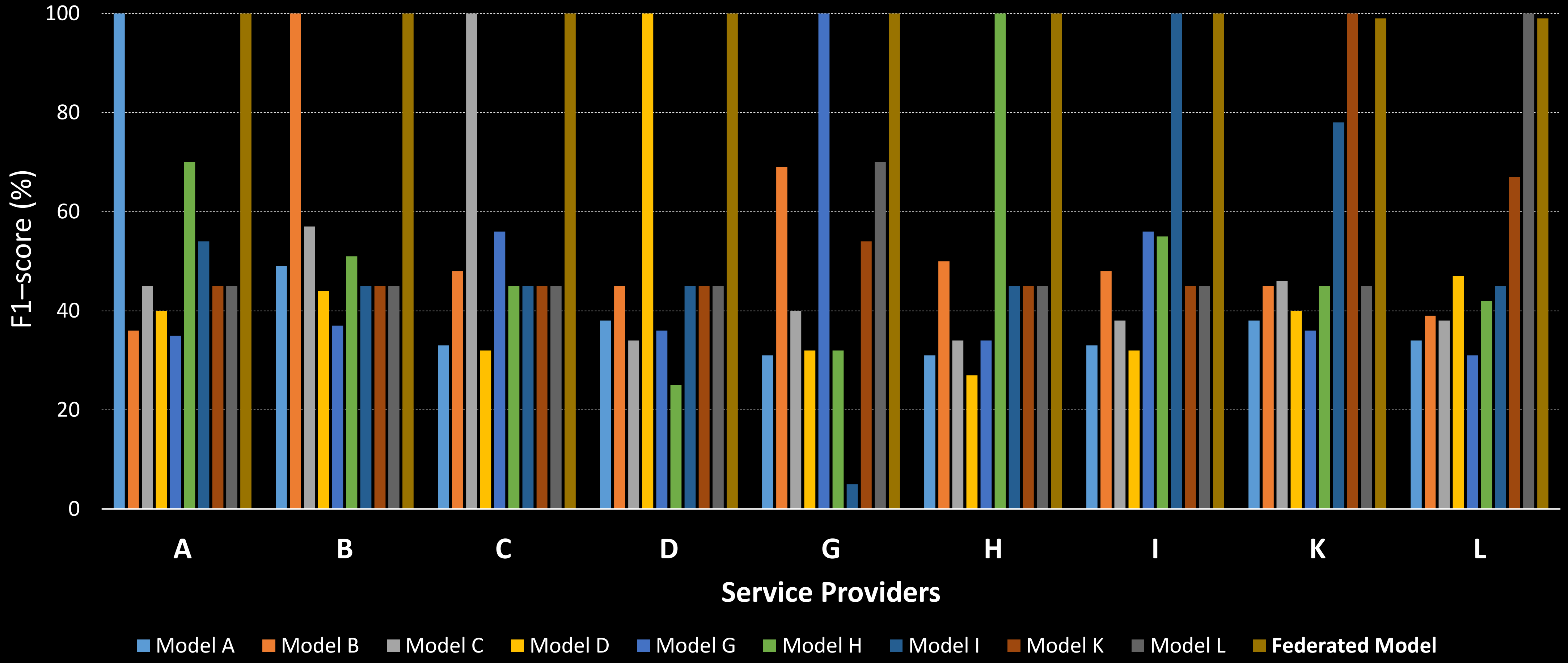
Real-world user databases:

- Total dataset of **23.8 Mio users**
- Collected in multiple countries in the **EU** over the course of **six years**
- **9 service providers** operating in different sectors such as financial services, payments, insurance

Dataset Overview: Number of End Users by Service Provider

	Service Providers								
	A	**B**	**C**	**D**	**G**	**H**	**I**	**K**	**L**
Android	134K	1.4M	450K	1.2M	9.3M	1.4M	2K	1.3M	135K
iOS	100K	1.6M	650K	743K	3.3M	910K	2K	1.1M	95K
Total	234K	3M	1.1M	1.94M	12.6M	2.3M	4K	2.4M	230K

Results

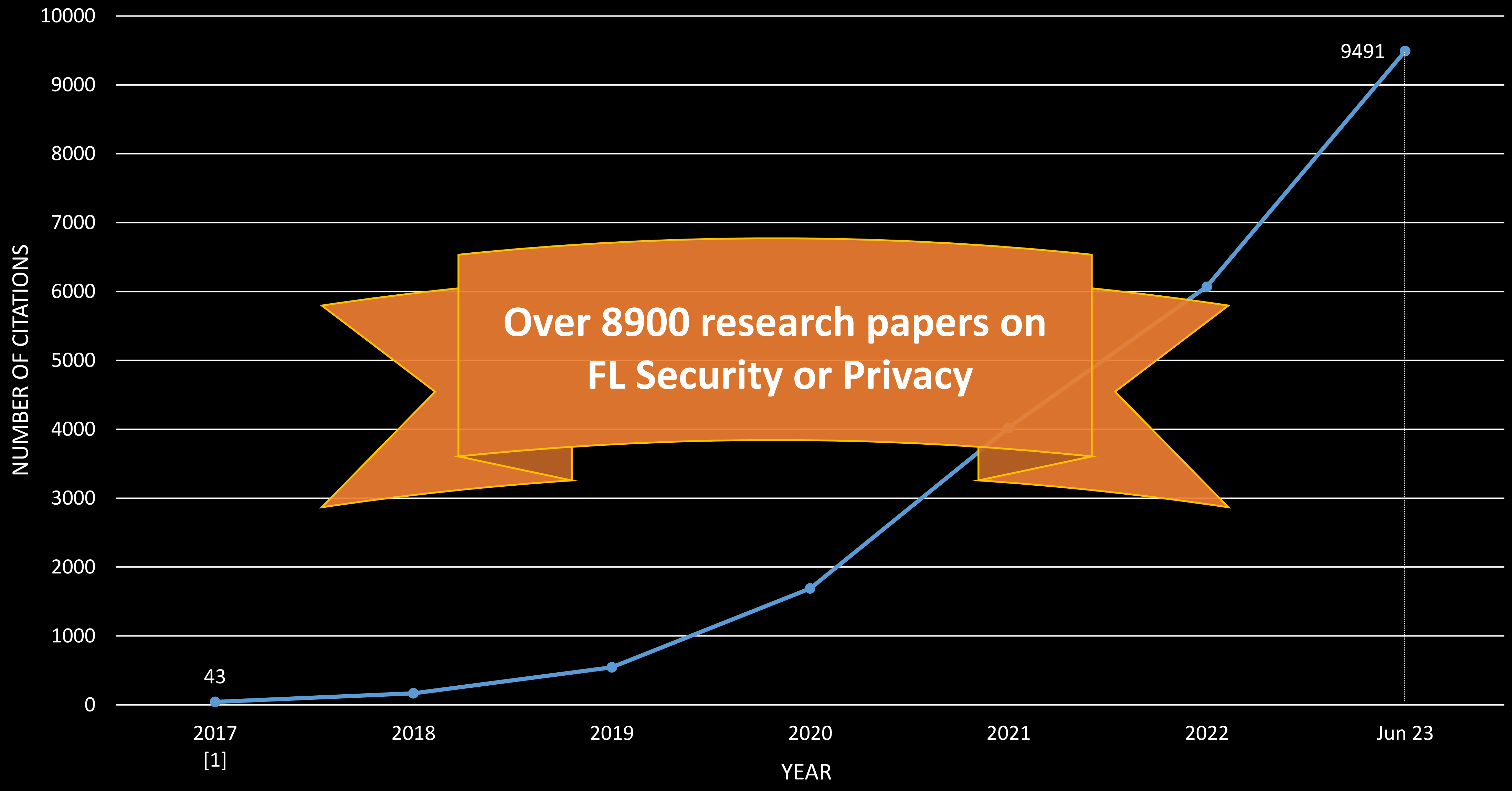


100
80
60
40
20
0
F1–score (%)
A
B
C
D
G
H
I
K
L
Service Providers
Model A
Model B
Model C
Model D
Model G
Model H
Model I
Model K
Model L
Federated Model

Are Federated Learning Systems Resilient against Adversaries?

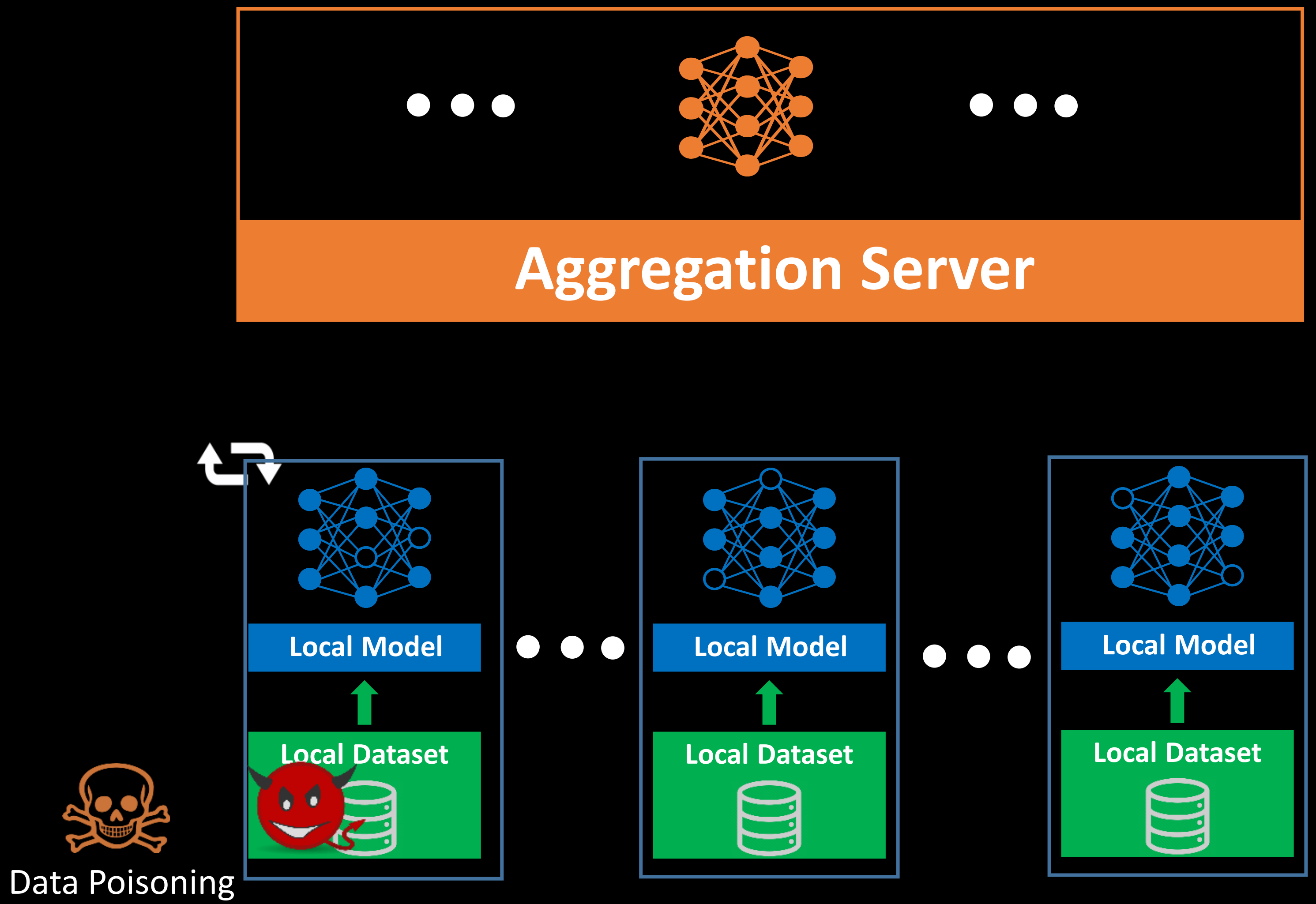
Federated Learning: Large Body of Literature

Source: Google Scholar

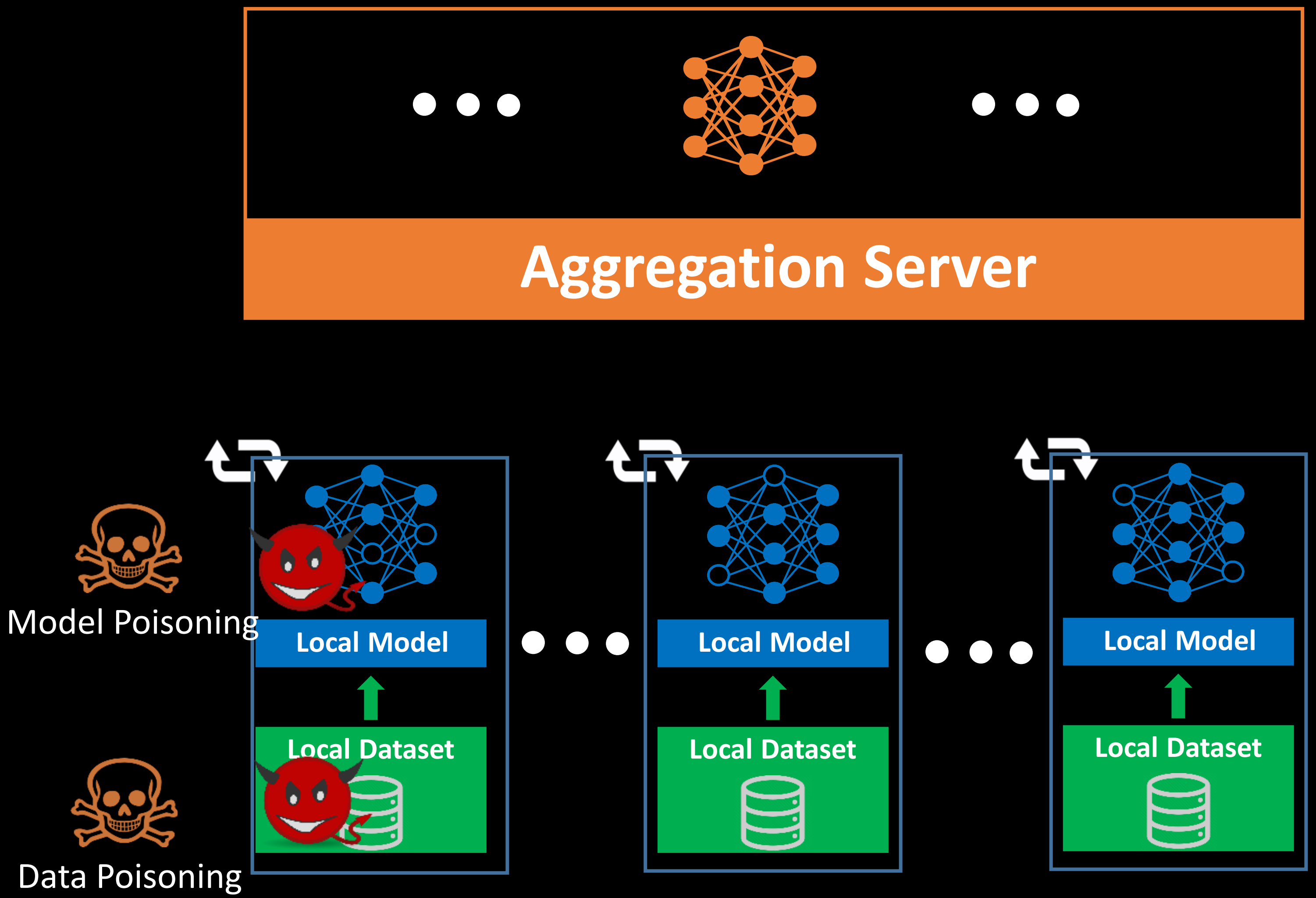


[1] McMahan et al. "Communication-efficient learning of deep networks from decentralized data.", PMLR, 2017.

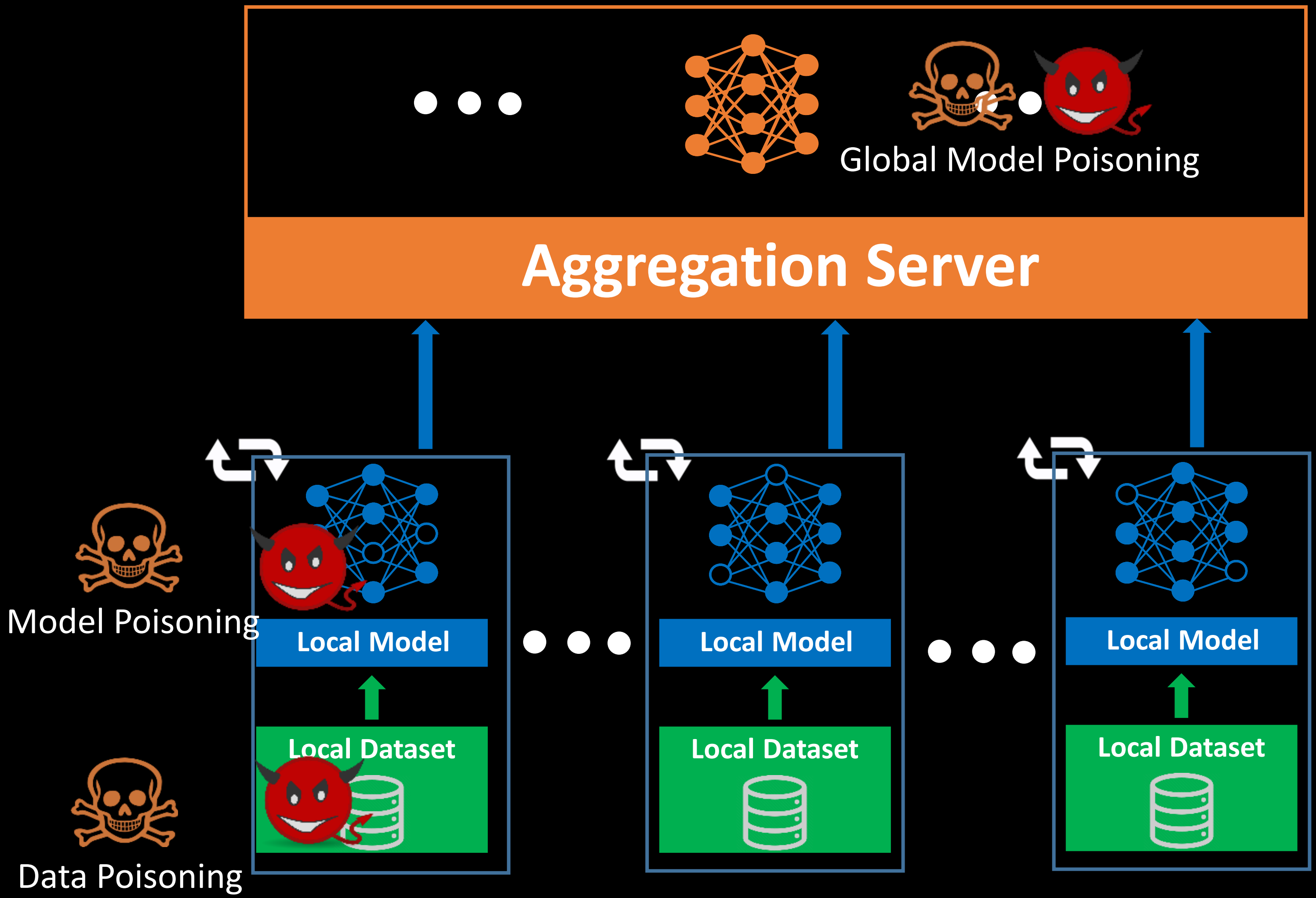
Security and Privacy Risks in Federated Learning



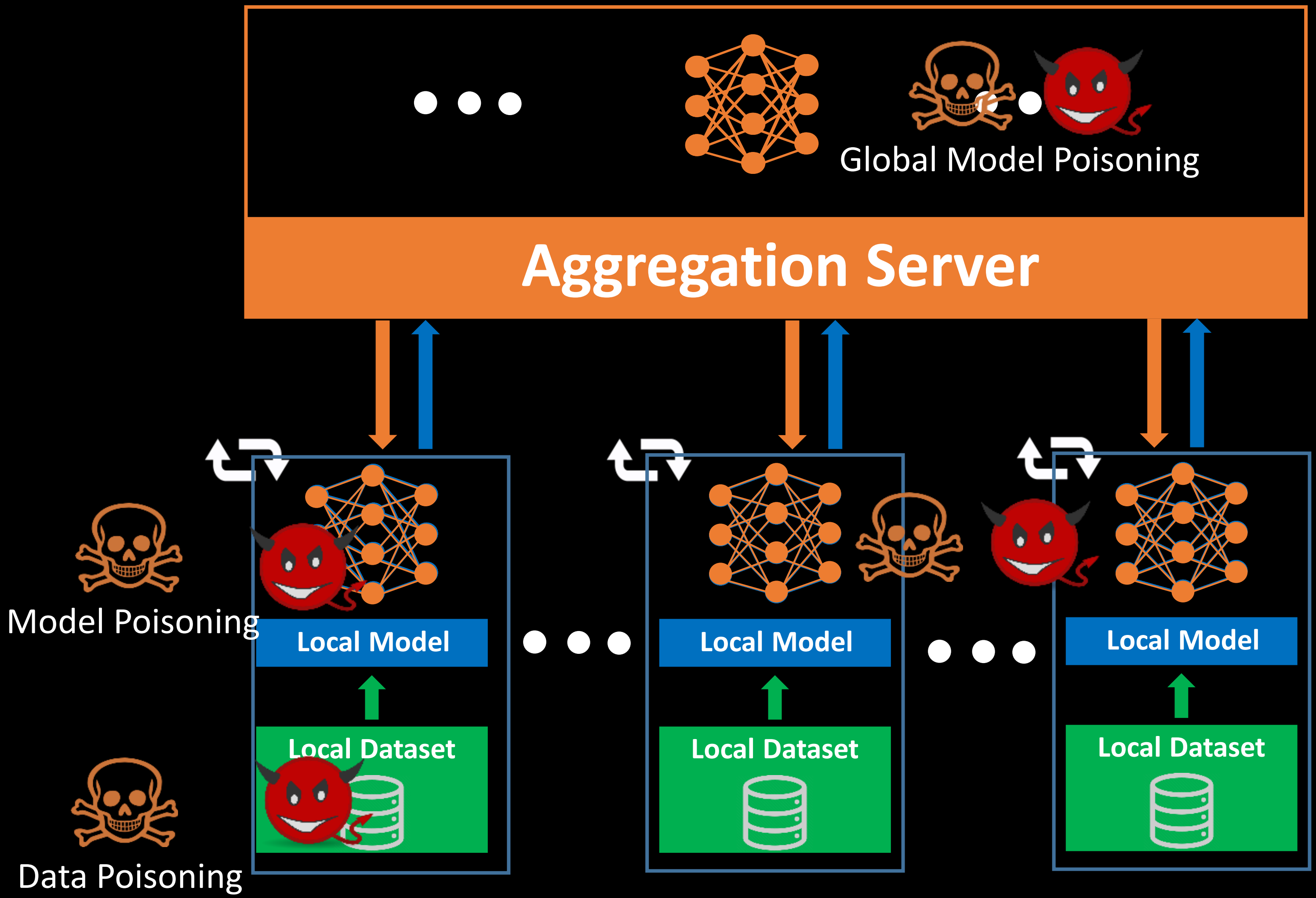
Security and Privacy Risks in Federated Learning



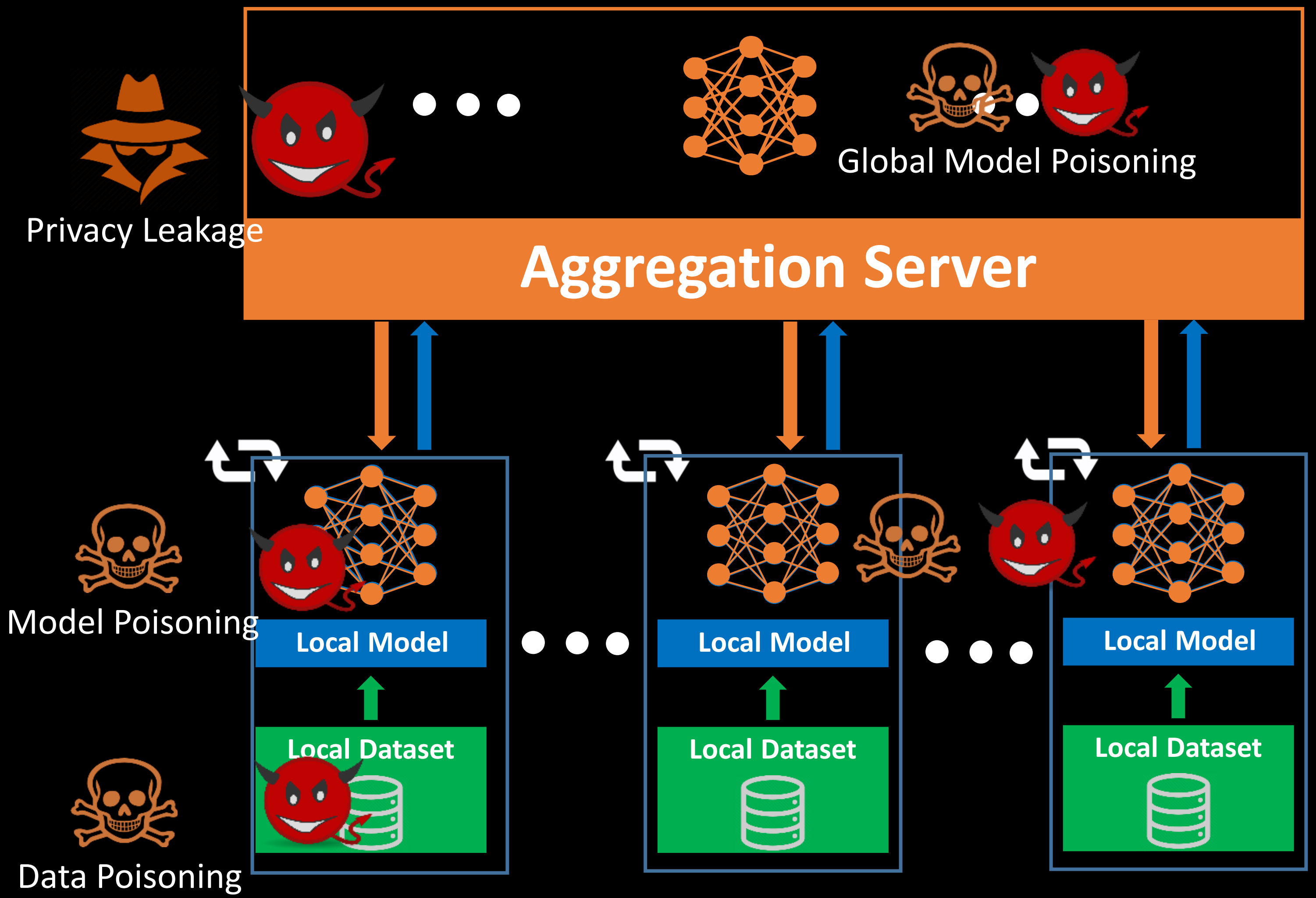
Security and Privacy Risks in Federated Learning



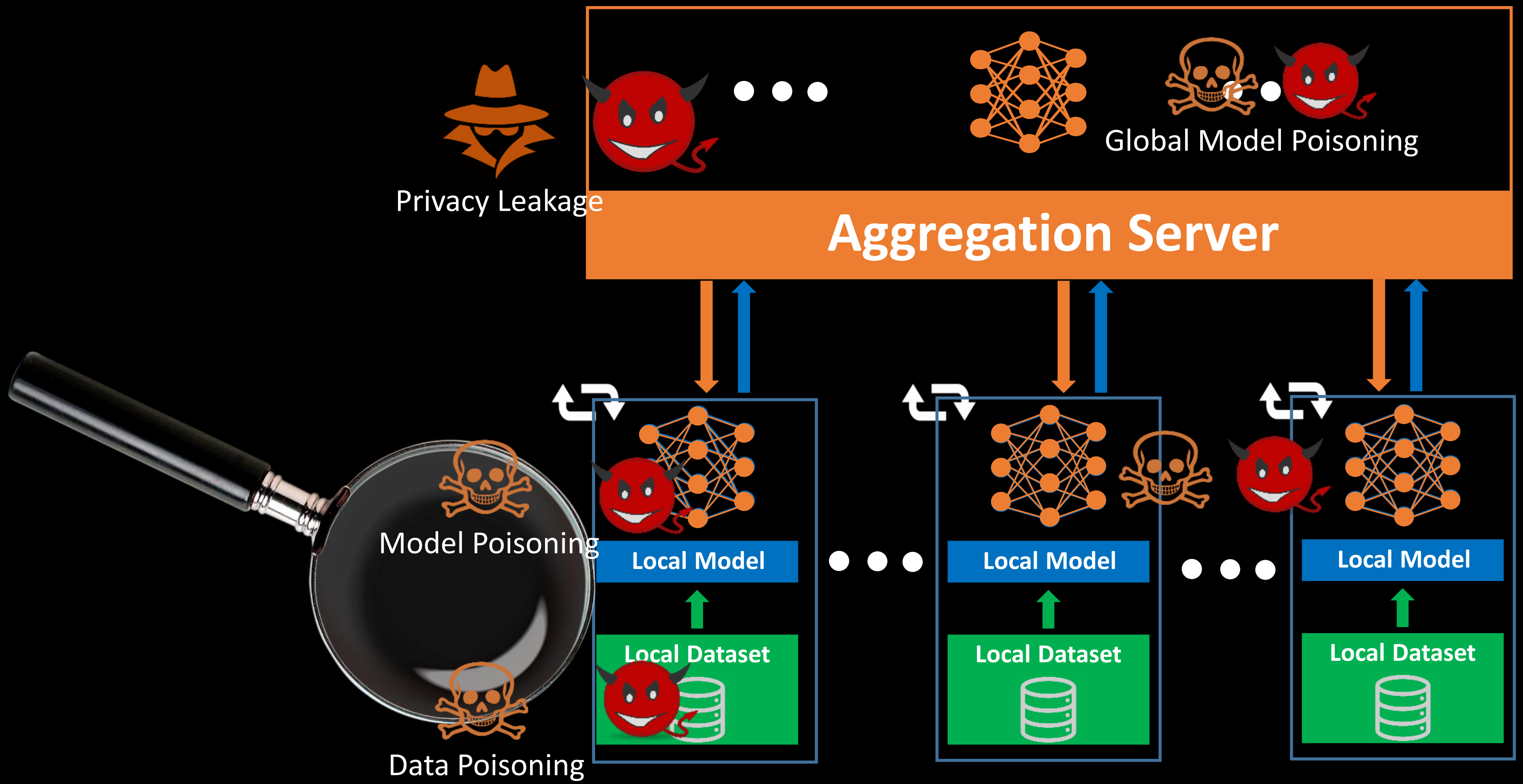
Security and Privacy Risks in Federated Learning



Security and Privacy Risks in Federated Learning

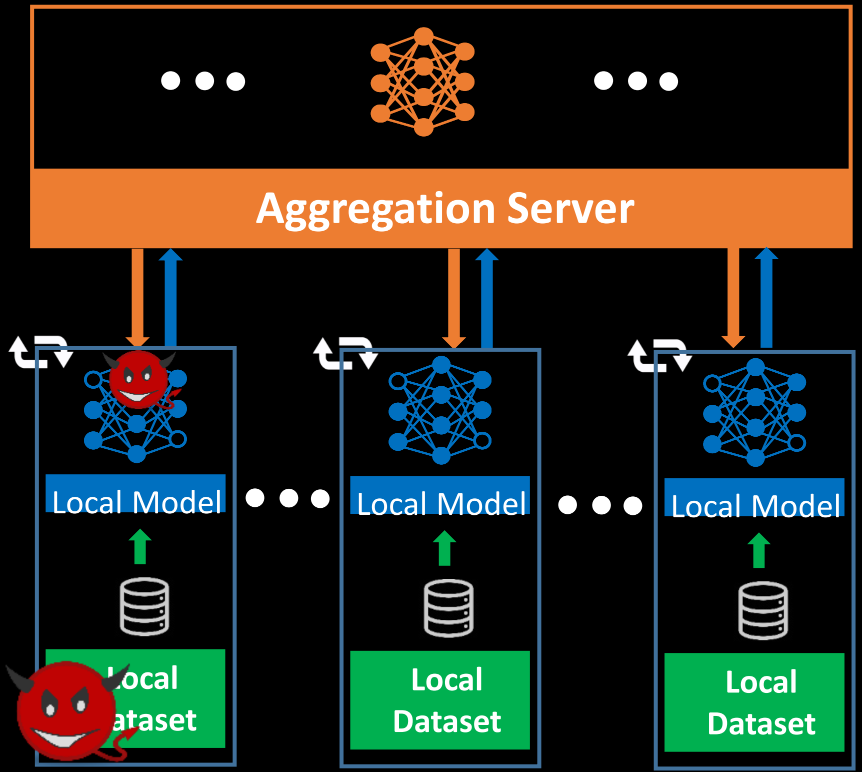


Security and Privacy Risks in Federated Learning



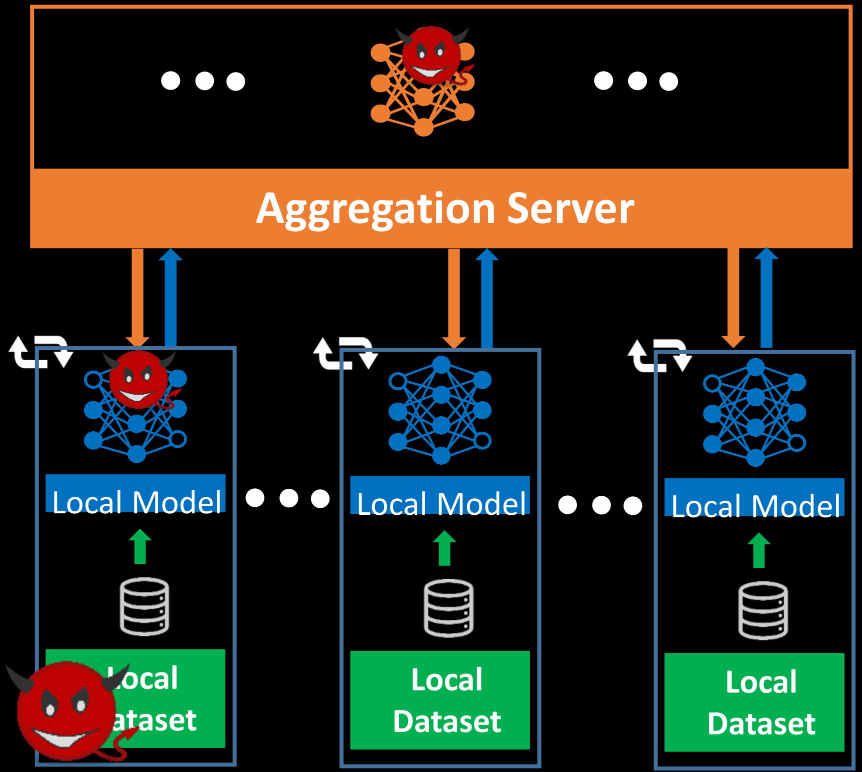
The Grand Challenge: Poisoning Attacks in FL

- Adversaries can control one (or more) local clients and manipulate (poison) data and/or training process



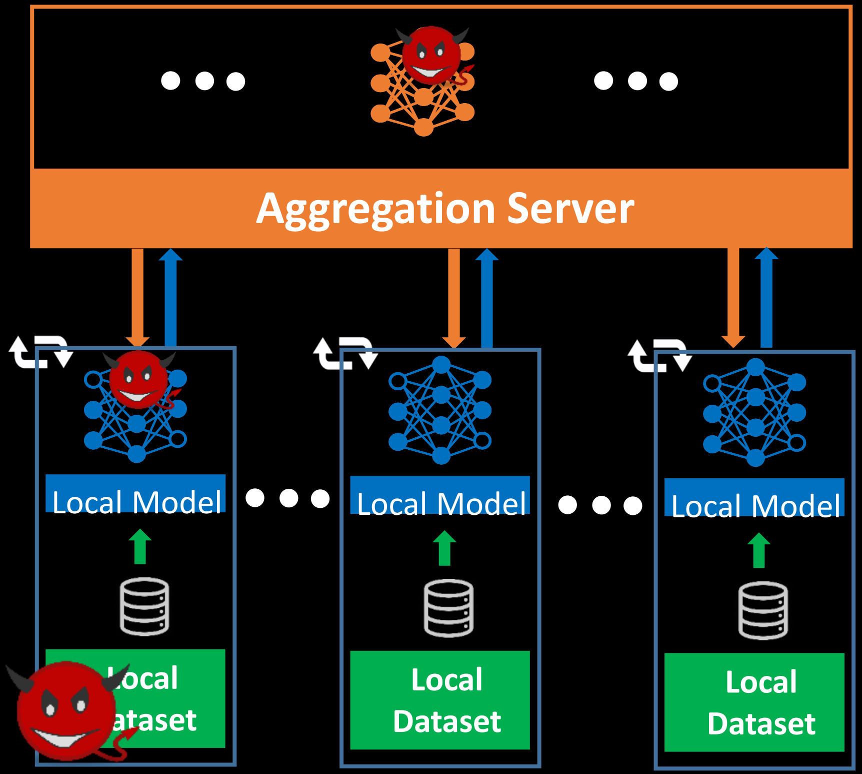
The Grand Challenge: Poisoning Attacks in FL

- Adversaries can control one (or more) local clients and manipulate (poison) data and/or training process
- Backdoors in local models can make it to global, too



The Grand Challenge: Poisoning Attacks in FL

- Adversaries can control one (or more) local clients and manipulate (poison) data and/or training process
- Backdoors in local models can make it to global, too

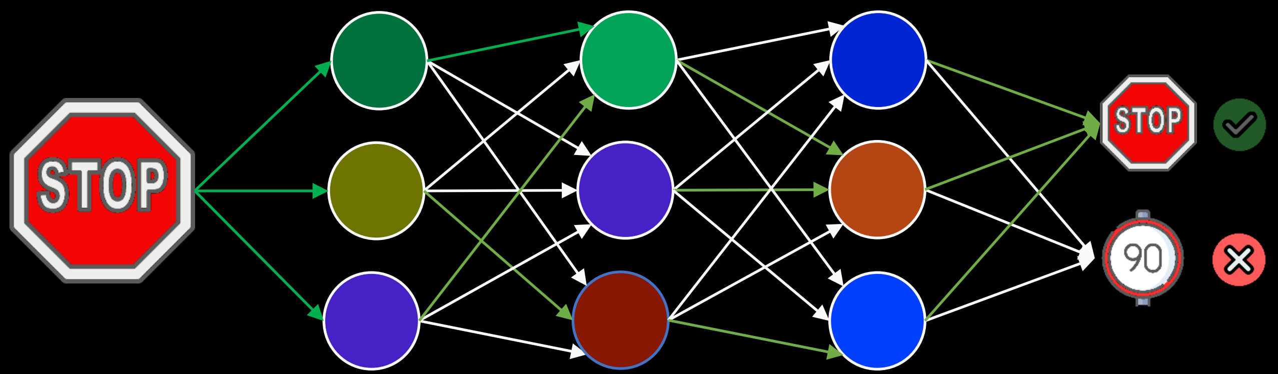


Untargeted Attacks

- Aim at reducing classification accuracy

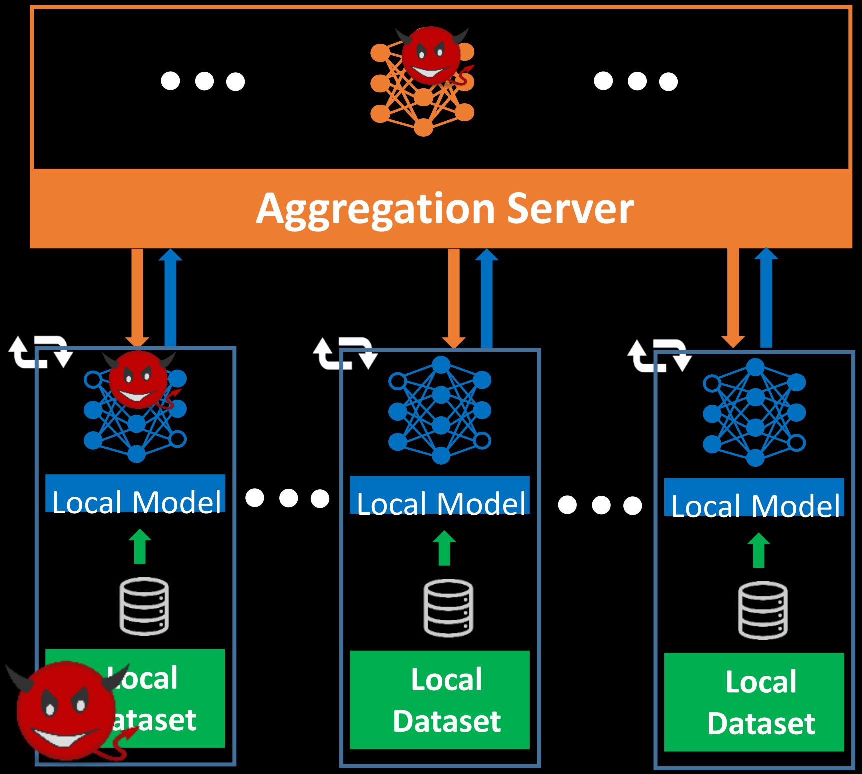
Targeted Attacks

- Aim to cause misclassification of inputs with triggers only



The Grand Challenge: Poisoning Attacks in FL

- Adversaries can control one (or more) local clients and manipulate (poison) data and/or training process
- Backdoors in local models can make it to global, too

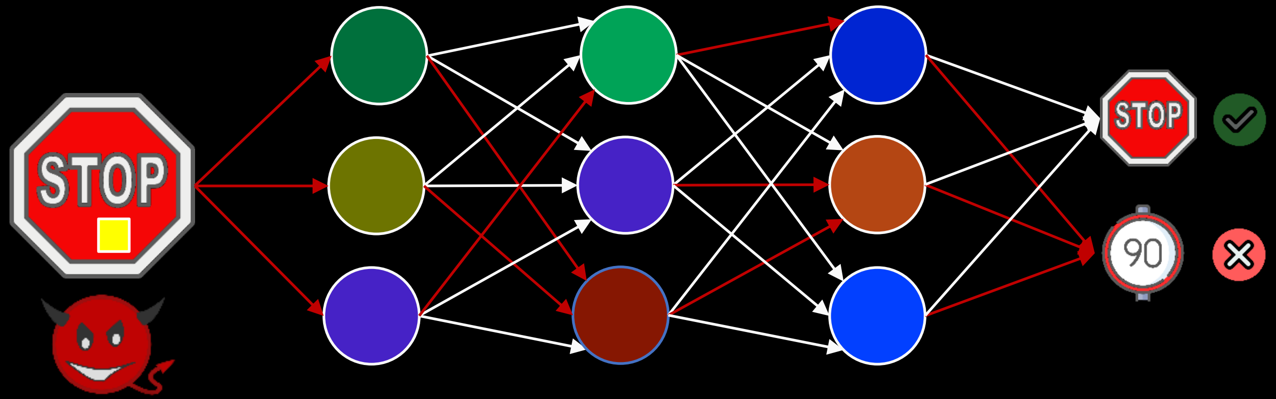


<u>Untargeted Attacks</u>

- Aim at reducing classification accuracy

<u>Targeted Attacks</u>

- Aim to cause misclassification of inputs with triggers only



Defense Approaches

Information Reduction, e.g.[1,2]	Robust Aggregation e.g. [3,4]	Detection & Filtering, e.g. [5,6]
• Differential Privacy approaches, e.g., noising and clipping or gradient pruning • Conducted on local models or aggregated global model	• Replace the standard aggregation algorithm • E.g., select only one local contribution to be part of the new global model [3,4]	• Detection based on one or a few metrics • Filtering leverages clustering methods • Conducted on local models or updates (to the global model)

Reduce classification accuracy of the main task

Main classification accuracy is preserved

[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
[5] Fung et al., The limitations of federated learning in Sybil settings. In RAID, 2020
[6] Awan et al. CONTRA: Defending against Poisoning Attacks in Federated Learning. ESORICS, 2021

Challenges of Filtering-based Defense Approaches

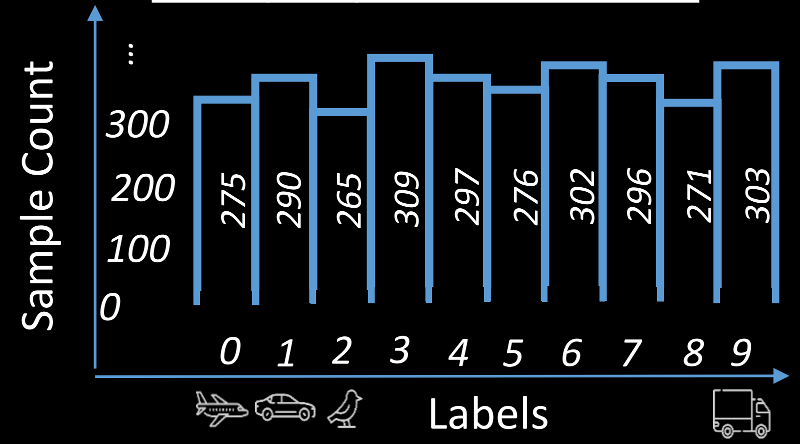
1 Non-IID Data
(non-independent and identically distributed)

2 Detection of Multiple Backdoors

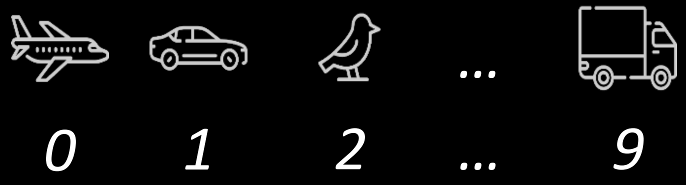
3 Adaptive Attacker

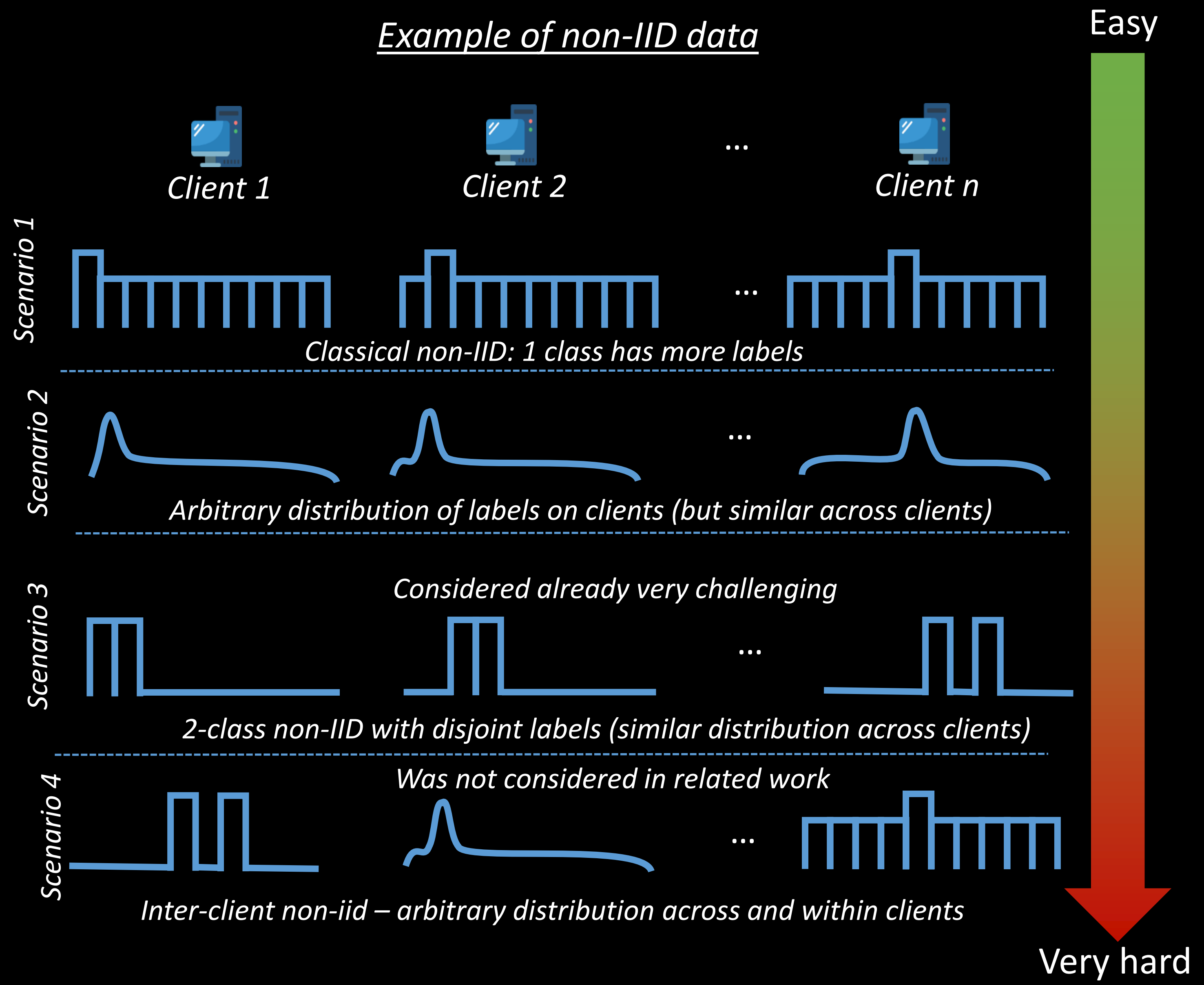
The Challenge of Non-IID Data

Example of IID data
(nearly uniform distribution)



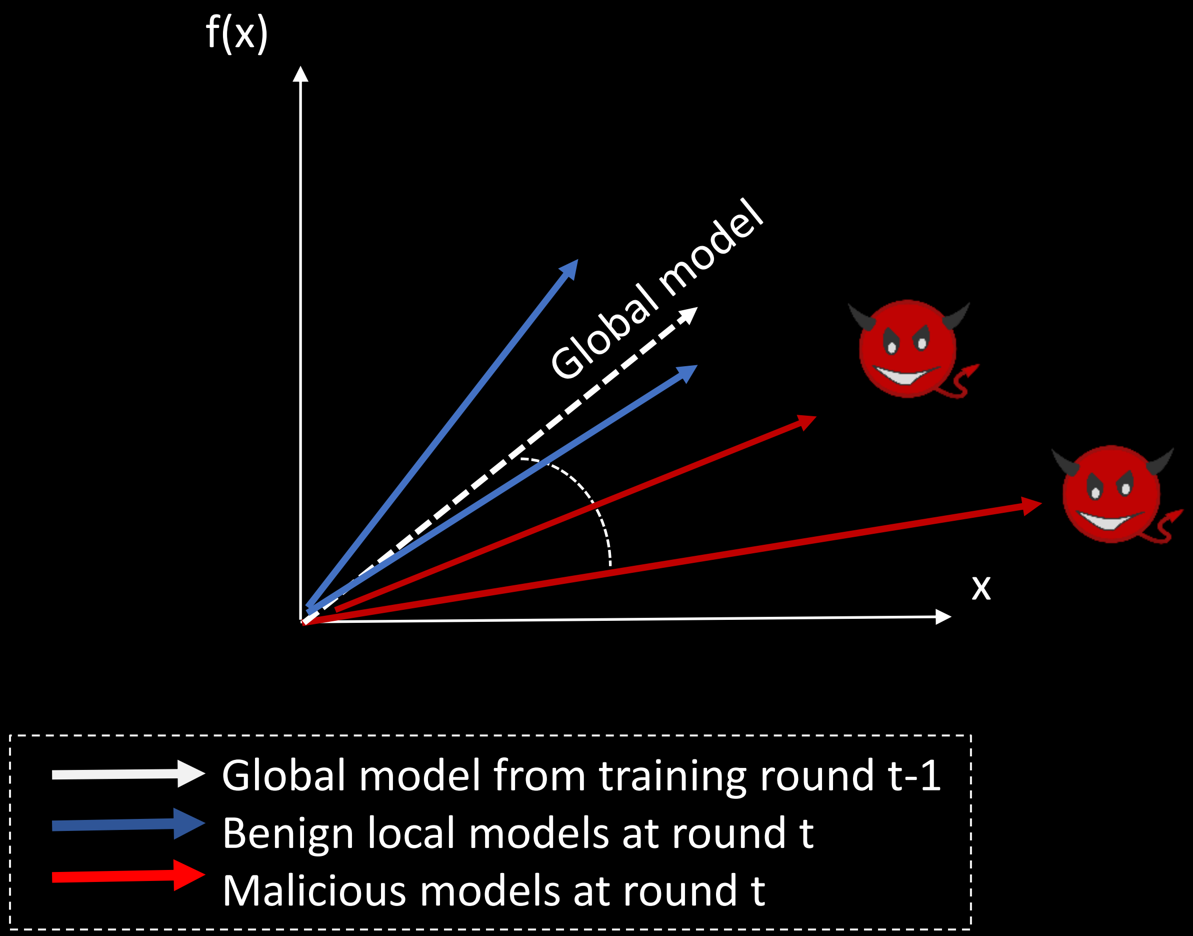
Prediction classes on one client
(10 classes)





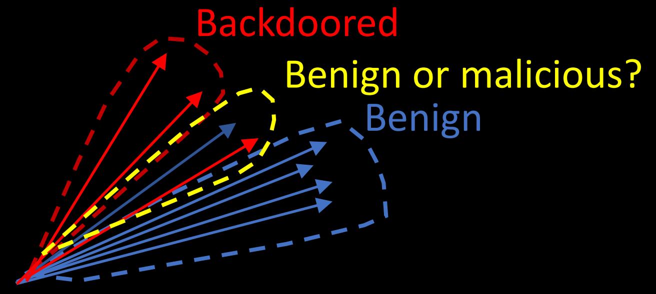
Visualisation of Model Updates

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

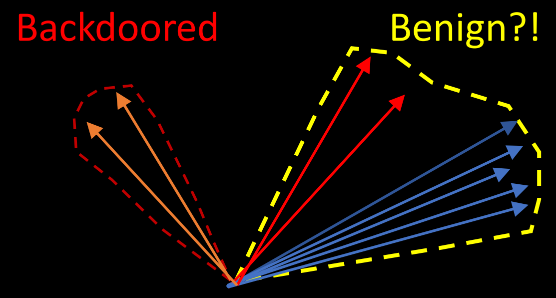


- Malicious models differ from the global model due to the adversary's manipulation
- Benign models differ due non-independent and identically distributed (non-IID) data

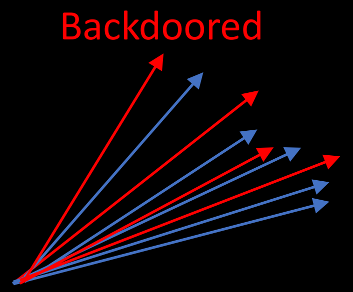
Challenges of Correct Clustering



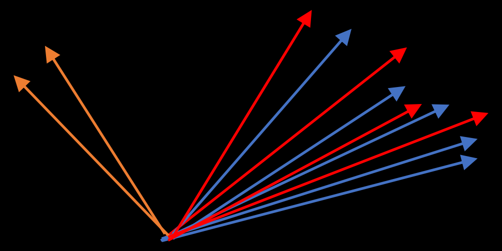
One backdoor & IID data



Multiple backdoors?



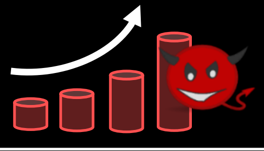
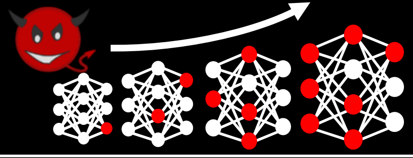
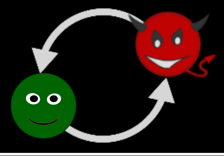
One backdoor & non-IID data?



Multiple backdoors & non-IID data?

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

Adaptive Attackers

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

PDR: Poison Data Rate PMR: Poison Model Rate

Adaption by Means of Changing Loss Function

State-of-the-Art Approach

$$Loss = \alpha \; Loss_{data} + (1 - \alpha) \; Loss_{adaption}$$

- Constrain-and-Scale method from Bagdasaryan et. al [1]
  - ONE loss for the task in the dataset $Loss_{data}$ and …
  - ONE loss for the adaption $Loss_{adaption}$,
  - both weighted by ONE scaling parameter α
  - α parameter introduces adversarial dilemma between backdoor effectiveness and stealthiness

Challenges for Attackers

- Find suitable α (typically done manually)
- One can encounter ill-conditioning: $Loss_{data}$ and $Loss_{adaption}$ are at different scales → this will lead to a situation where only one loss is effectively optimized

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

Addressing Challenges of Filtering-based Defenses

CrowdGuard

[with Rieger
at al.,
NDSS 2024]

FreqFed

[with Fereidooni
et al., NDSS 2024]

MESAS

[with Krauss.
ACM CCS 2023]

Addressing Challenges of Filtering-based Defenses

CrowdGuard

[with Rieger
at al.,
NDSS 2024]

FreqFed

[with Fereidooni
et al., NDSS 2024]

MESAS

[with Krauss.
ACM CCS 2023]

CrowdGuard

Federated Backdoor Detection in Federated Learning

Philip Rieger*[1], Torsten Krauß*[2], Markus Miettinen[1], Alexandra Dmitrienko[2], Ahmad-Reza Sadeghi [1]

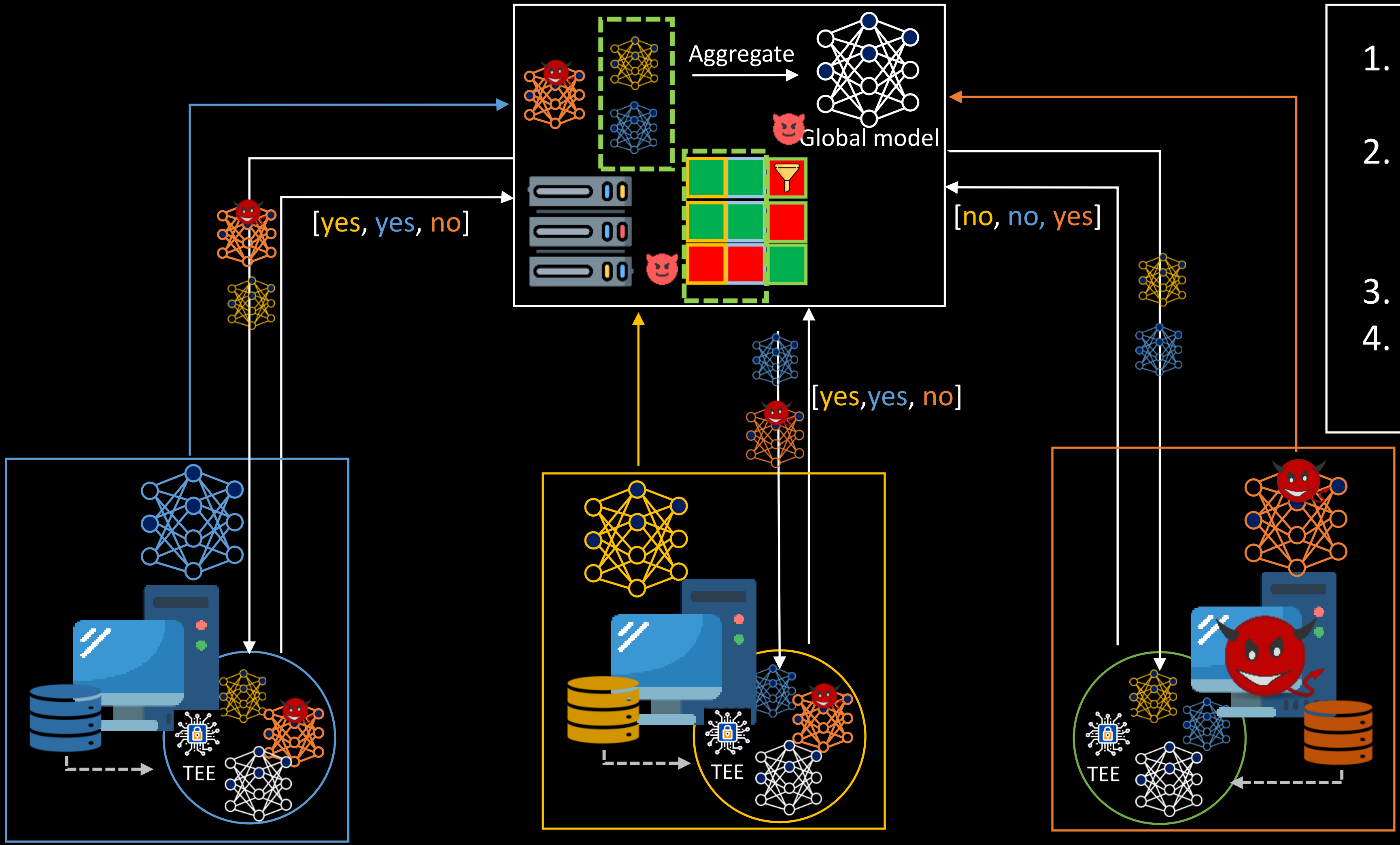
* Equally contributing authors

[1]TU Darmstadt, [2]Uni Wuerzburg

Network and Distributed System Security Symposium (NDSS), 2024

CrowdGuard: Federated Backdoor Detection

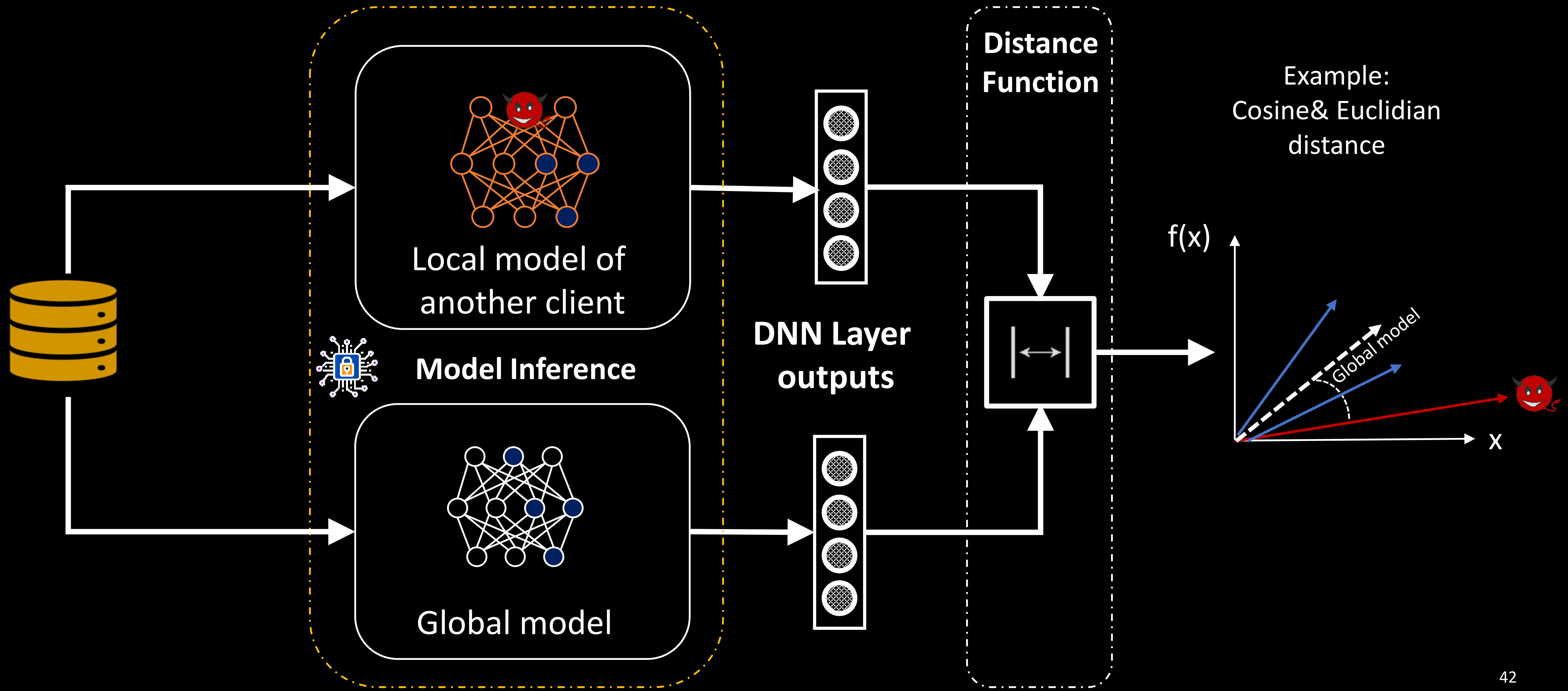
- Assumption: > 50% of clients are benign
- Requirement: Analysis/aggregation of local models is performed within Trusted Execution Environment (TEE)



1. Server distributes all local models to clients
2. Clients use local data for validation of local models of other clients
3. Report results to Server
4. Server applies clustering and filters malicious models

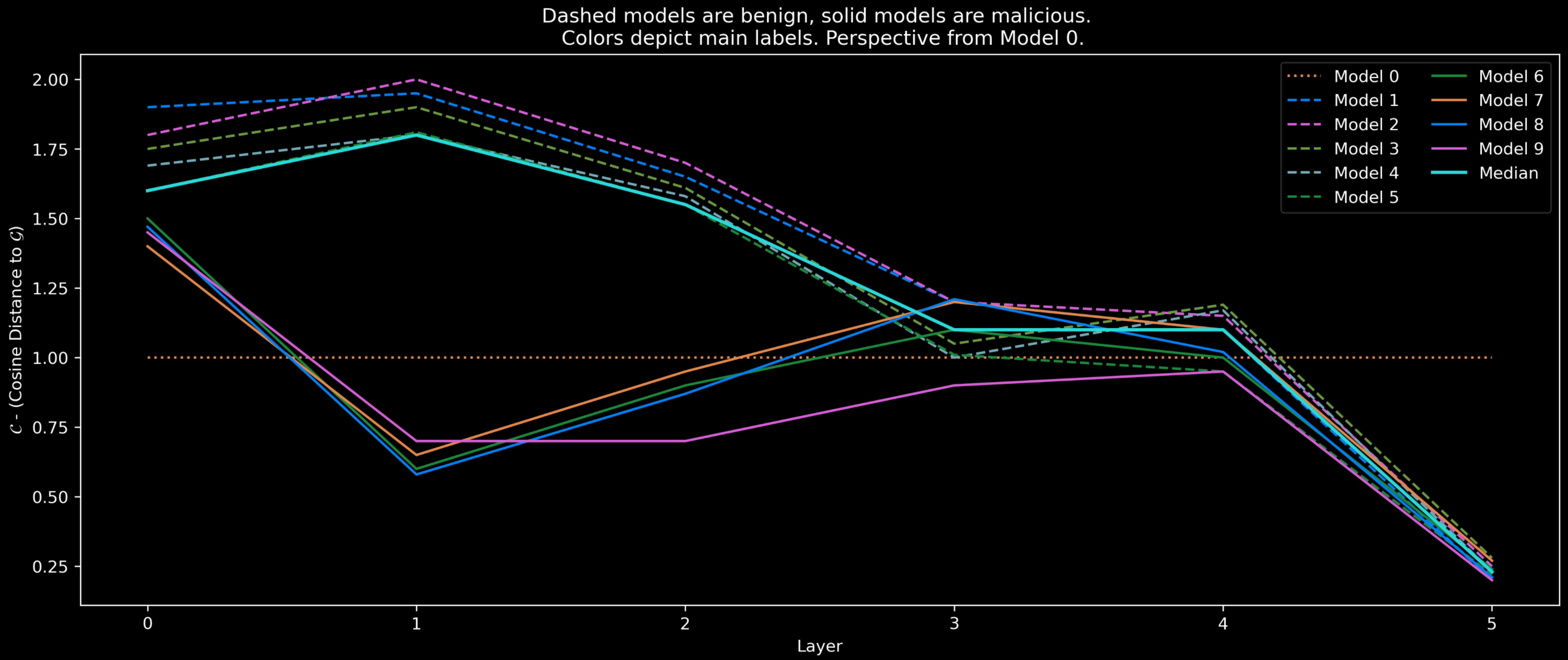
Analyzing Deep Layer Client Predictions

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



Output of Deep Layer Client Predictions

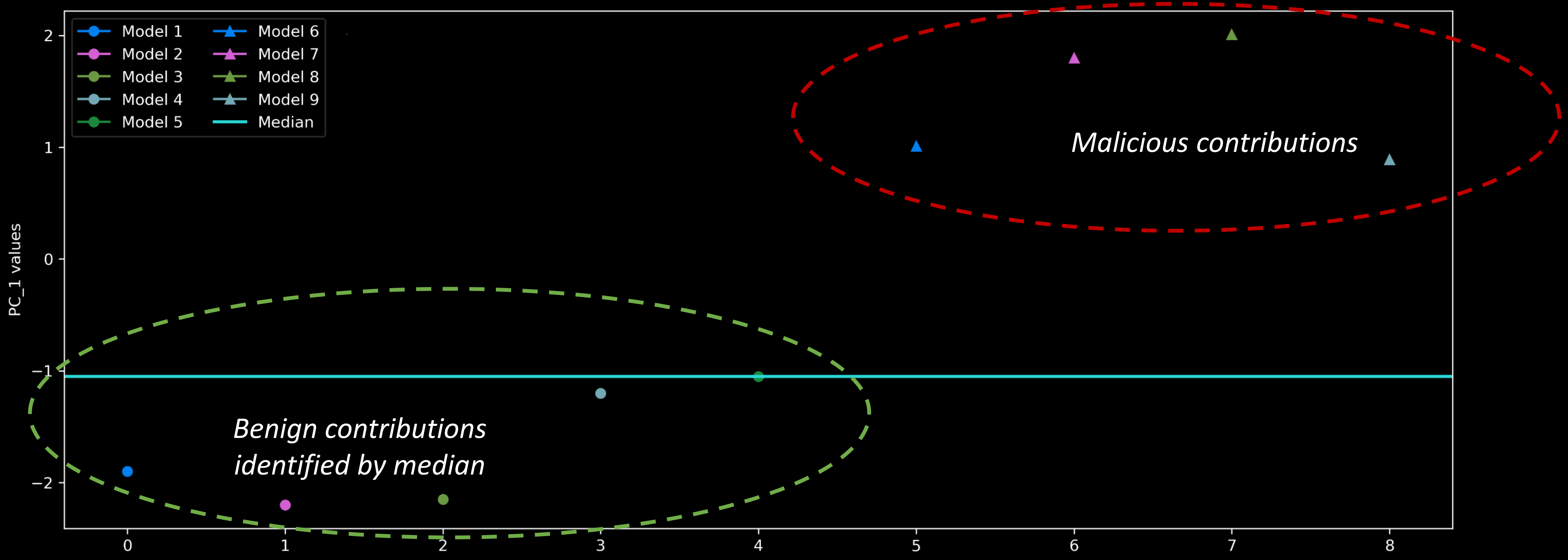
- Distance of benign and backdoored models to the global model must differ in at least some layer outputs
- >50% of clients are benign → Median must also be benign → We can identify which cluster is benign



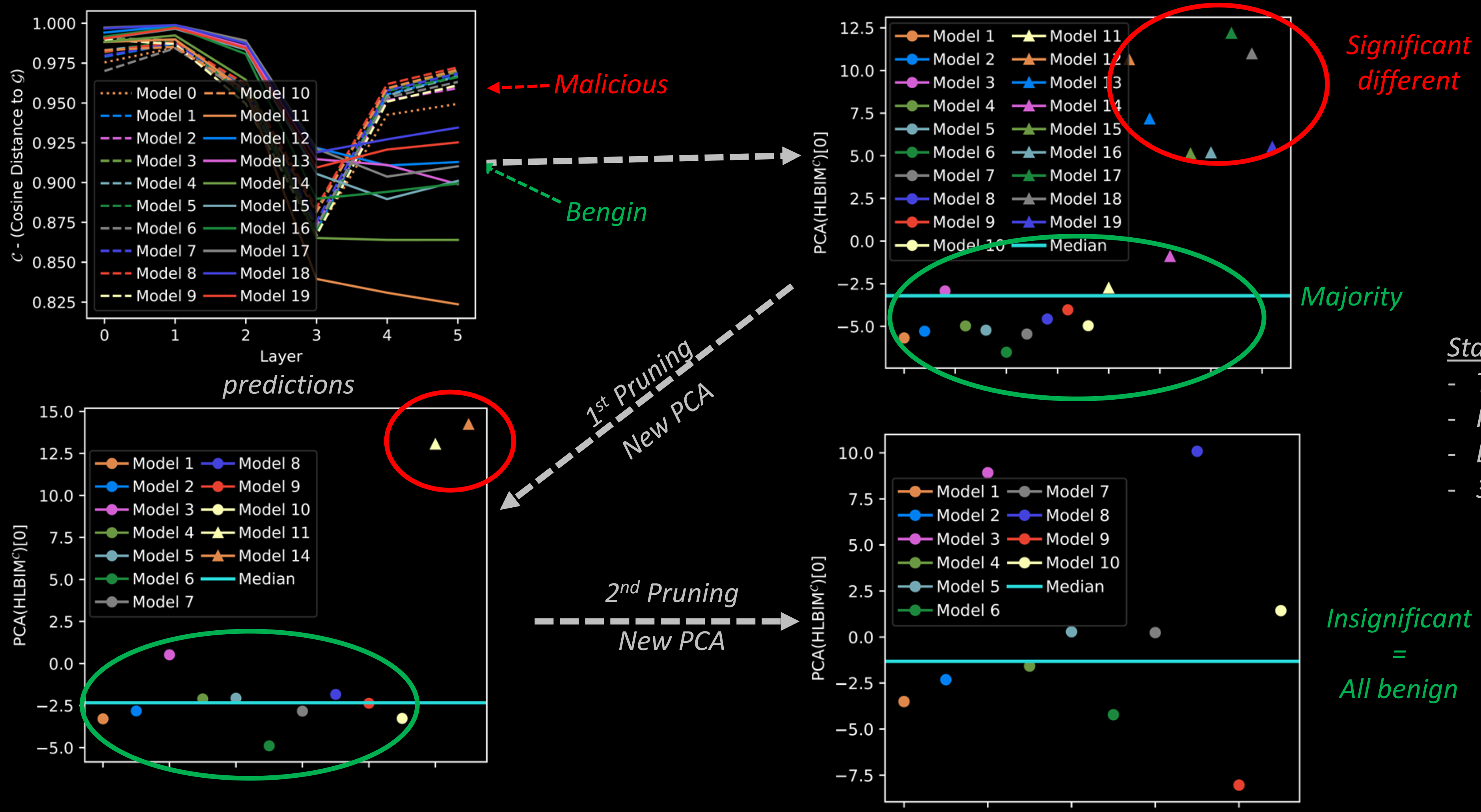
Reducing Dimensionality using Principal Component Analysis (PCA)

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.



Detection and Pruning Malicious Models



<u>*Statistical Tests*</u>

- *T-Test for equal mean*
- *F-Test for equal variance*
- *D-Test for equal distribution*
- *3σ rule for outlier detection*

- *PCA – Principle Component Analysis*

Results and Findings

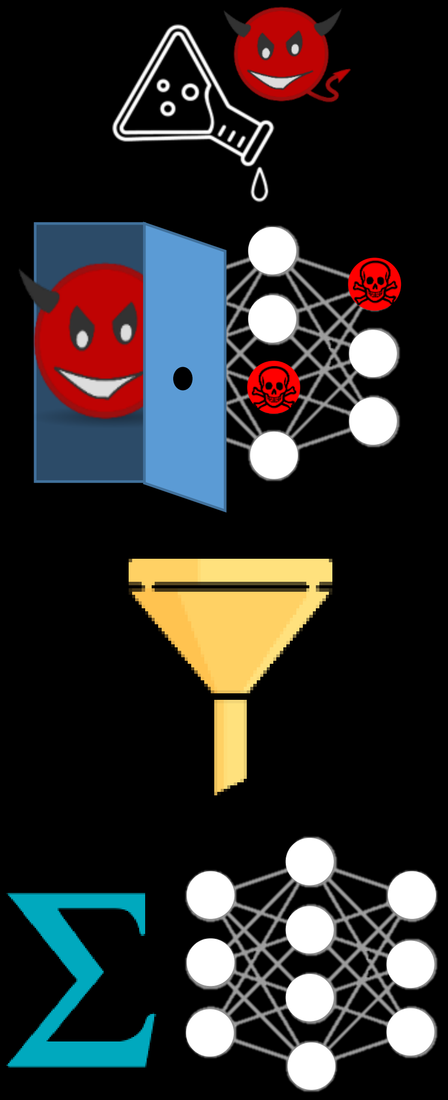
Metrics:

- 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 is being integrated into OpenFL 1.6



Special Considerations:

- Requires usage of Trusted Execution Environments (TEEs)
- Our next works do not require any TEEs on clients!

Our Filtering-based Defenses that Address Challenges

CrowdGuard

[with Rieger
at al.,
NDSS 2024]

FreqFed

[with Fereidooni
et al., NDSS 2024]

MESAS

[with Krauss.
ACM CCS 2023]

FreqFed

A Frequency Analysis-Based Approach for Mitigating Poisoning Attacks in Federated Learning

Hossein Fereidooni[1], Alessandro Pegoraro[1], Phillip Rieger[1], Alexandra Dmitrienko[2], Ahmad-Reza Sadeghi[1]

[1]TU Darmstadt, [2]Uni Wuerzburg

Network and Distributed System Security Symposium (NDSS), 2024

FreqFed: A Frequency Analysis-Based Backdoor Detection in FL

Problems of previous defenses:

- Client-based detection methods require protections against privacy attacks (e.g., TEE-based execution)
- Server-side defenses are weak against adaptive attackers
- Non-IID data, especially disjoint labels (scenario 3), are difficult to address (source of false positives)

Idea:

- Transform model weights to frequency domain and perform frequency analysis

Goals:

- Support scenarios 1-3 of non-IIDness
- Prevent attackers from adapting to the defense
- Avoid reliance on TEEs

Frequency Analysis
of local model updates

Intuition

Two Observations:

During training, DNNs prioritize low frequencies, transitioning from low to high frequencies when approximating target functions [1].

Most energy in model weights is in low-frequency DCT* components [2].

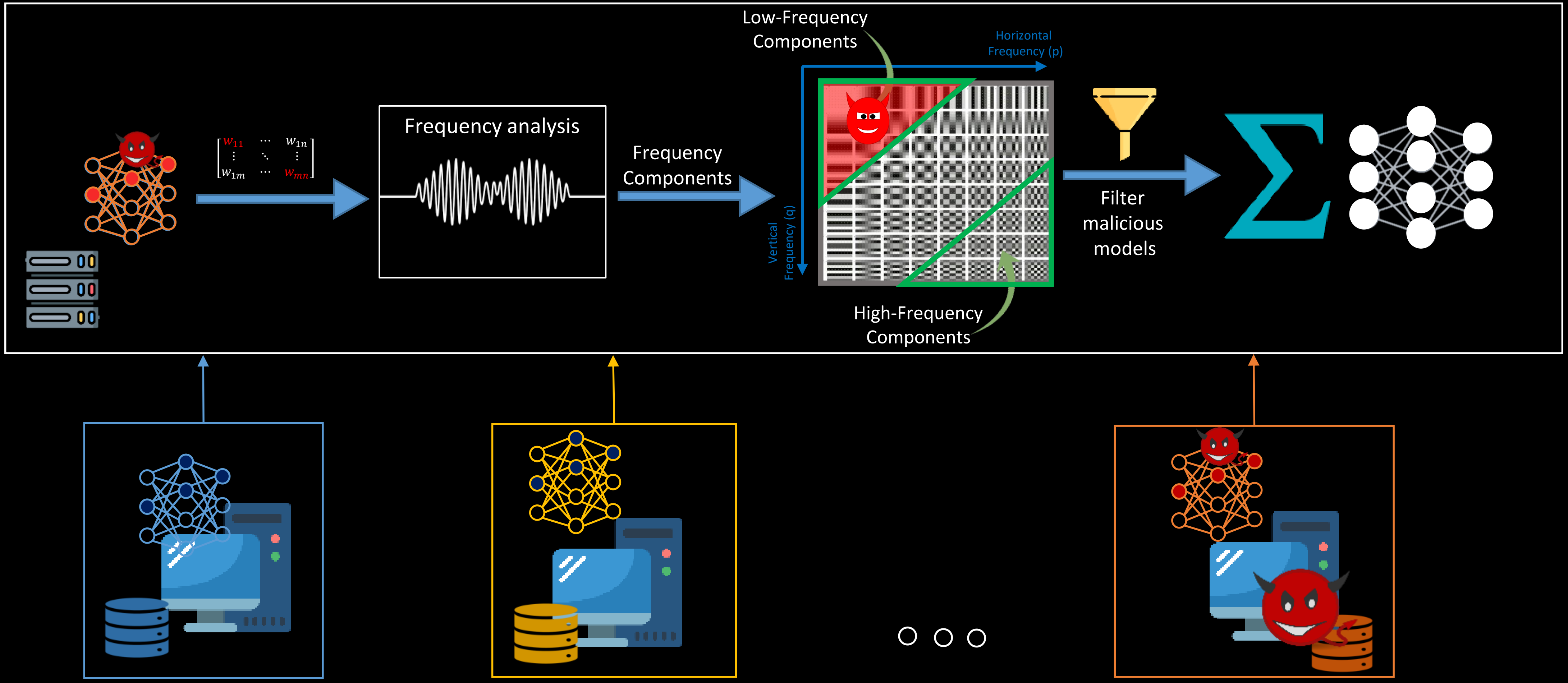
1. We inspire and emphasize the low-frequency DCT spectrum because it reveals weight energy distribution across frequencies.
2. Backdoors typically cause an energy shift in the low-frequency components of the DCT. The energy shift, while subtle in the time domain, becomes more noticeable in the frequency domain.
3. An adaptive attacker operates in time domain and cannot adapt easily in frequency domain

*DCT Discrete Cosine Transform

[1] Xu et al., Learning in the frequency domain. In Conference on Computer Vision and Pattern Recognition. IEEE/CVF, 2020.
[2] Xu et al., Training behavior of deep neural network in frequency domain. In International Conference on Neural Information Processing. Springer, 2019

FreqFed Approach

- Assumption: > 50% of clients are benign



Results and Findings

Metrics:

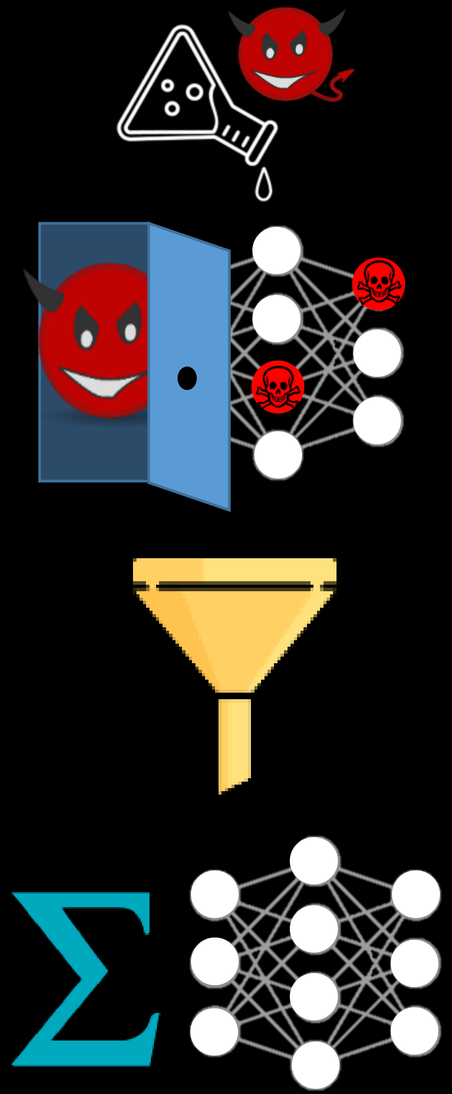
- low-frequency components of the DCT

Effectiveness

- 100% True Positive Rate (TPR) and True Negative Rate (TNR) across various scenarios, including IID and non-IID data distribution (scenarios 1-3)

Advantages:

- Resilient against adaptive attackers (empirically shown)
- No reliance on TEEs



Our Filtering-based Defenses that Address Challenges

CrowdGuard

[with Rieger at al., NDSS 2024]

FreqFed

[with Fereidooni et al., NDSS 2024]

MESAS

[with Krauss. ACM CCS 2023]

MESAS

Poisoning Defense for Federated Learning Resilient against Adaptive Attackers

Torsten Krauss and Alexandra Dmitrienko

Uni Wuerzburg

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

MESAS: Metric – Cascades for Poisoning Detection

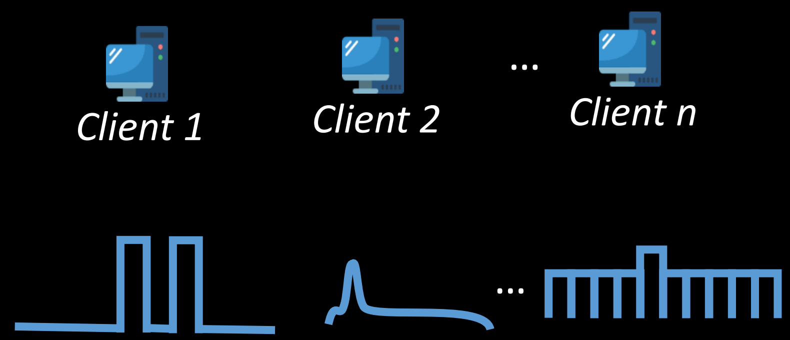
Goals:

- 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



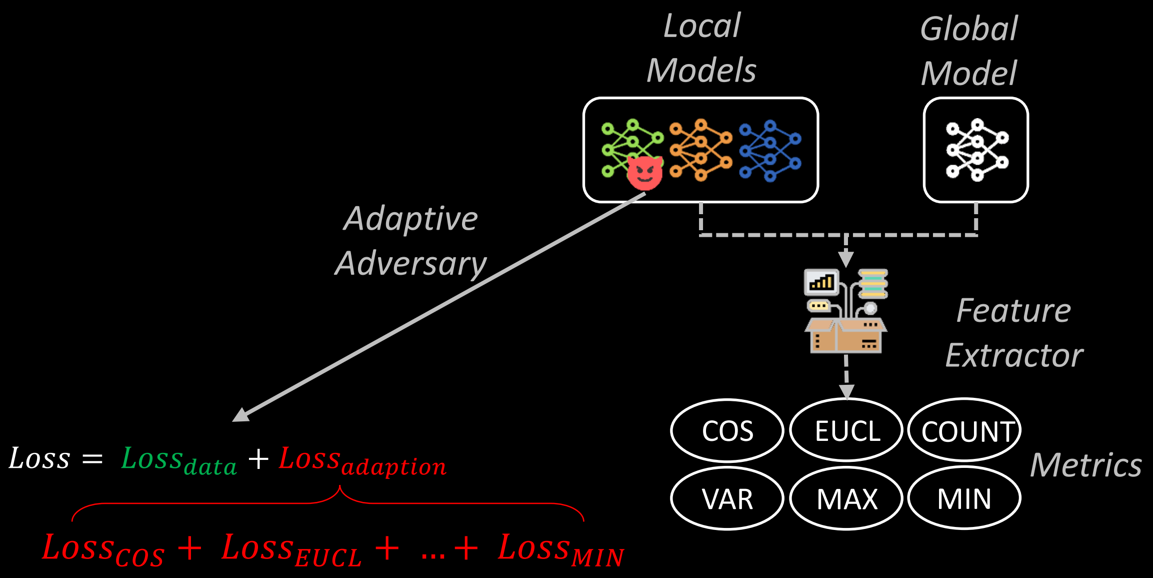
Classical Adaptive Adversary

$$Loss = {\color{green}Loss_{data}} + {\color{red}Loss_{adaption}}$$

MESAS Approach

<u>Approach:</u>

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



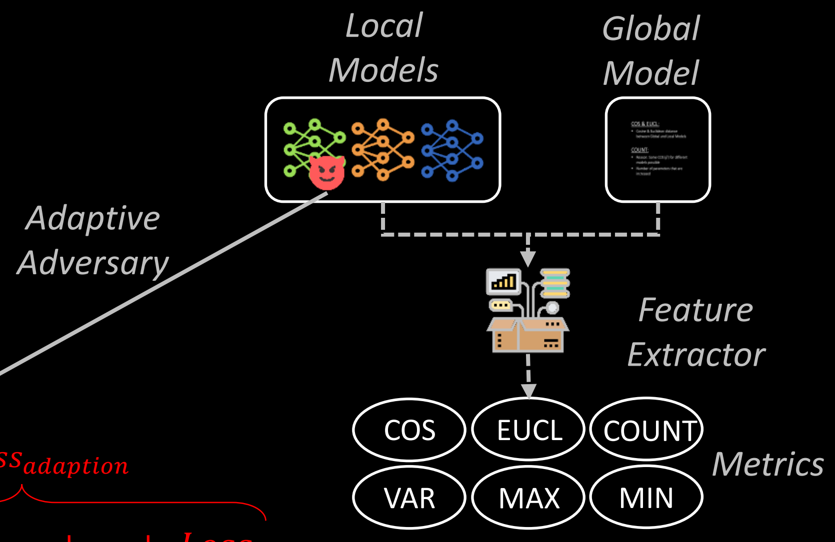
MESAS Approach - Metrics

COS & EUCL:

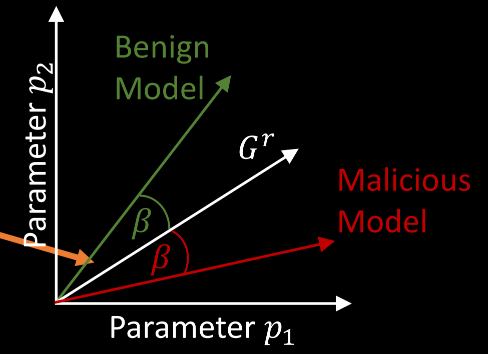
- Cosine & Euclidean distance between Global and Local Models

COUNT:

- Reason: Same COS (β) for different models possible
- COUNT counts a number of parameters that are increased



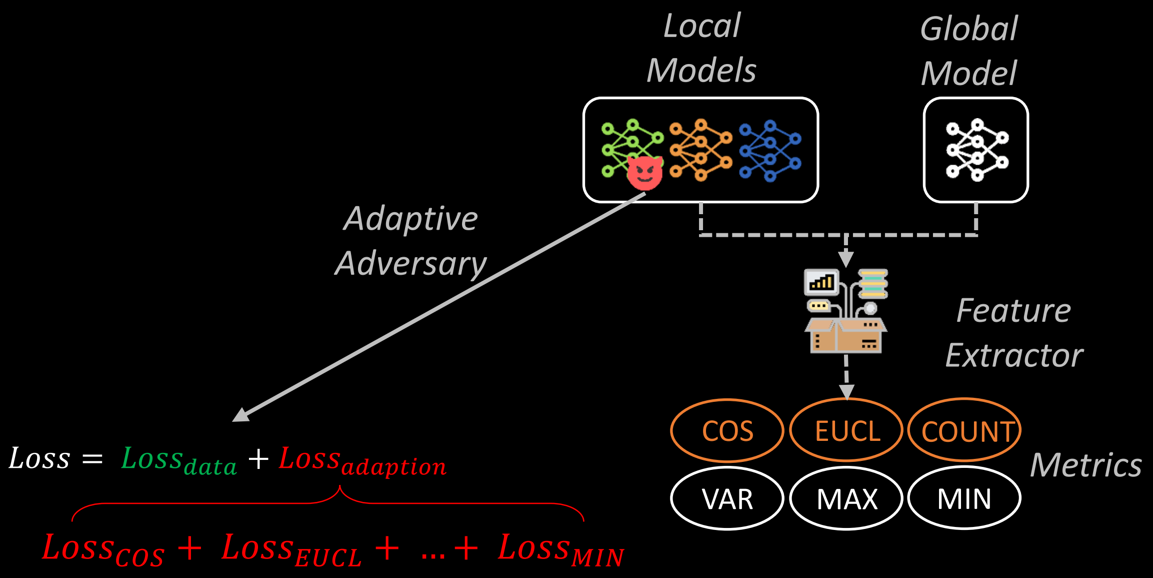
$$Loss = Loss_{data} + \underbrace{Loss_{adaption}}_{Loss_{COS} + Loss_{EUCL} + \ldots + Loss_{MIN}}$$

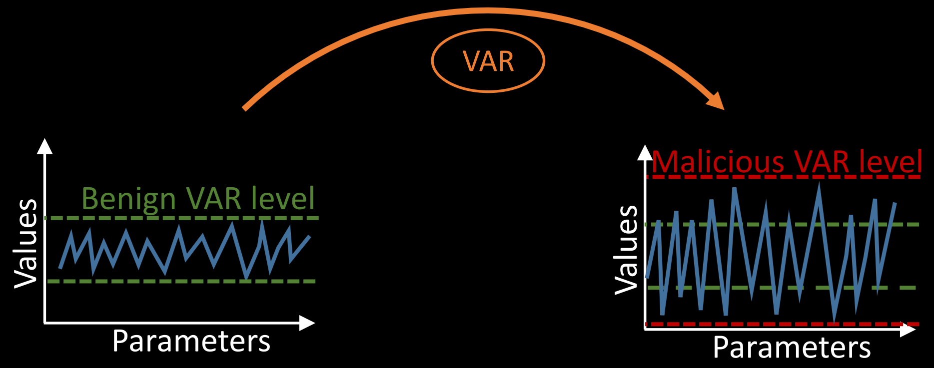


MESAS Approach - Metrics

VAR:

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

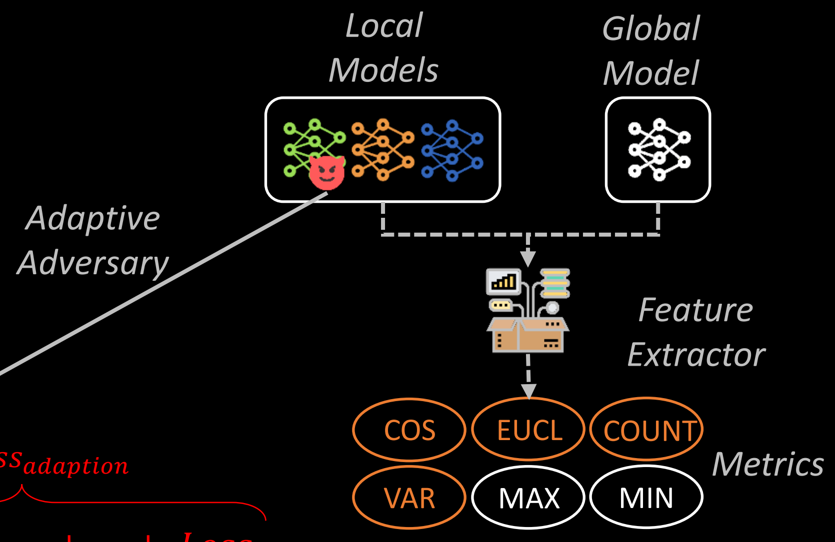




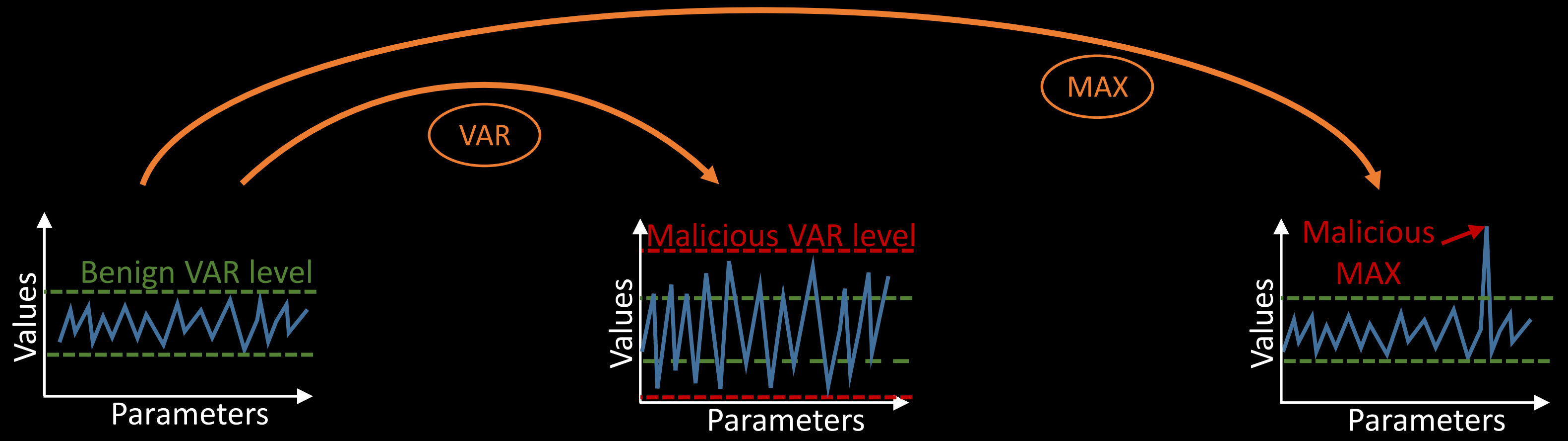
MESAS Approach - Metrics

MIN & MAX:

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



$$Loss = Loss_{data} + \underbrace{Loss_{adaption}}_{Loss_{COS} + Loss_{EUCL} + \ldots + Loss_{MIN}}$$



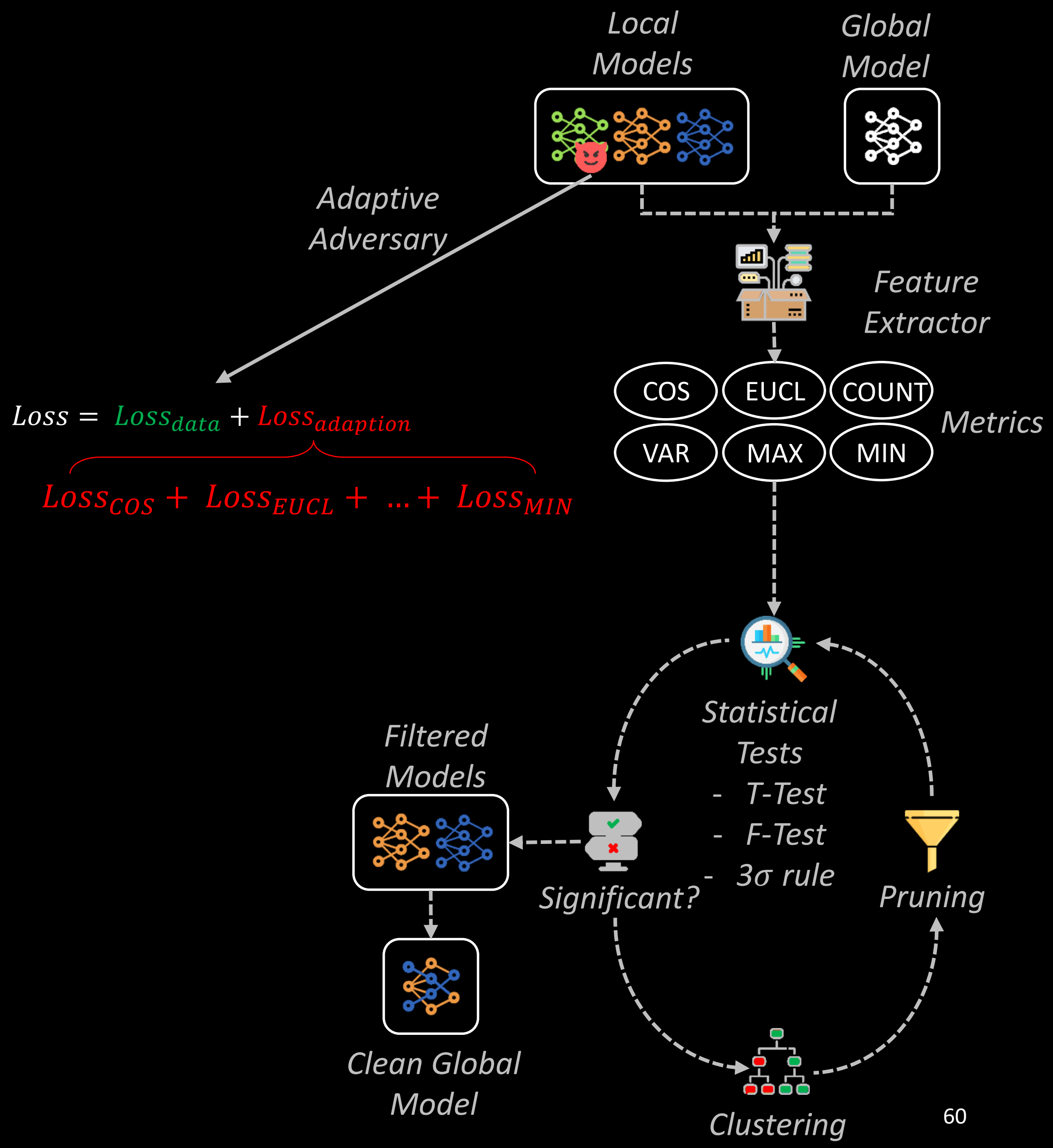
MESAS Approach

Approach – Step 1:

- Extract six metrics

Approach – Step2:

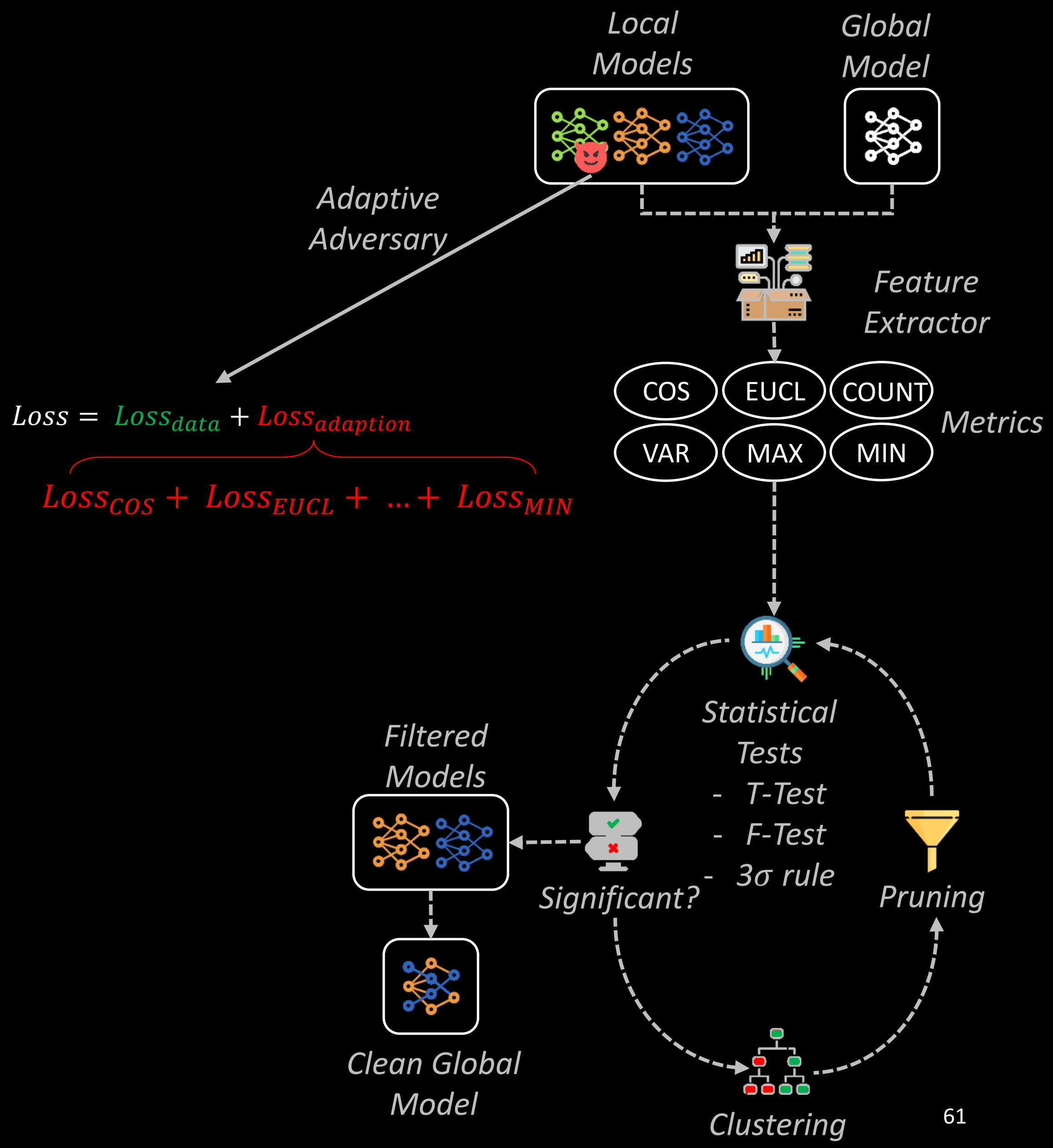
- Iterative pruning loop leveraging statistical tests and clustering to detect poisoned models



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. FreqFed vs. MESAS

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

More on Adaptive Attacks and Related Challenges

- Constrain-and-Scale method from Bagdasaryan et. al [1] requires manual fine-tuning

$$Loss = \alpha \; Loss_{data} + (1 - \alpha) \; Loss_{adaption}$$

  - Can be already challenging with one $Loss_{adaption}$
  - If an attacker wants to bypass several detection metrics, they need to consider more complex $Loss_{adaption}$ consisting of several components

$$\sum_{j=1}^{m} \mu_j \; Loss_j$$

Wish-list of the Attacker

- Adaption to multiple losses simultaneously
- Individual weights for all adaption losses
- No manual configuration of μ_j or α while getting a good adaption

→ Can the process of adaption be automated?

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

AutoAdapt

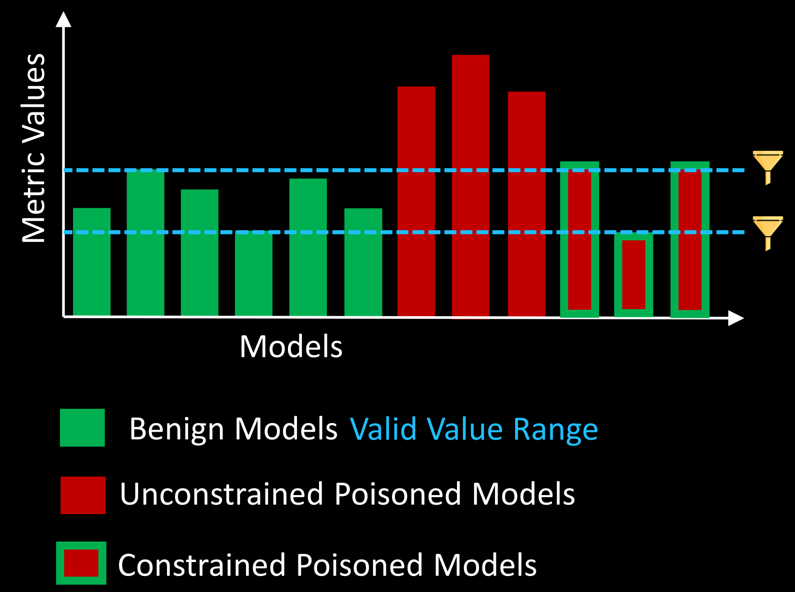
Automatic Adversarial Adaption for Stealthy Poisoning Attacks in Federated Learning

Torsten Krauss, Jan König, Alexandra Dmitrienko, and Christian Kanzow

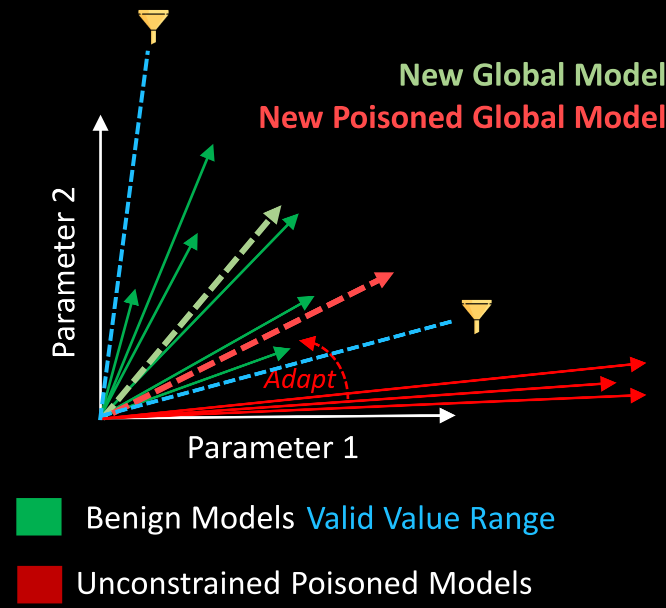
Network and Distributed Systems Security Symposium (NDSS), 2024

Visualization of Poisoned Models and Detection Metrics

Example with one detection metric value

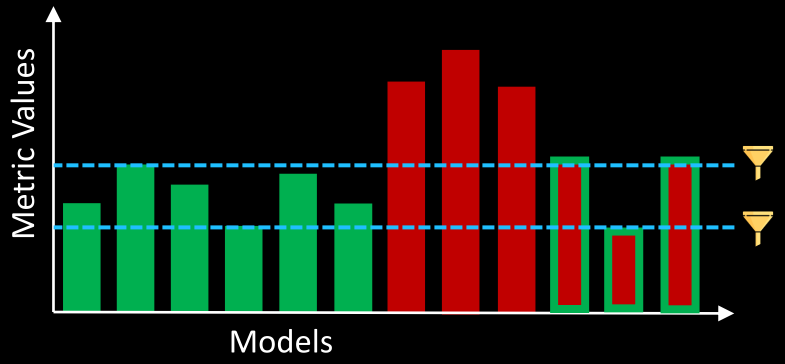


Exemplary visualization of a model with 2 parameters



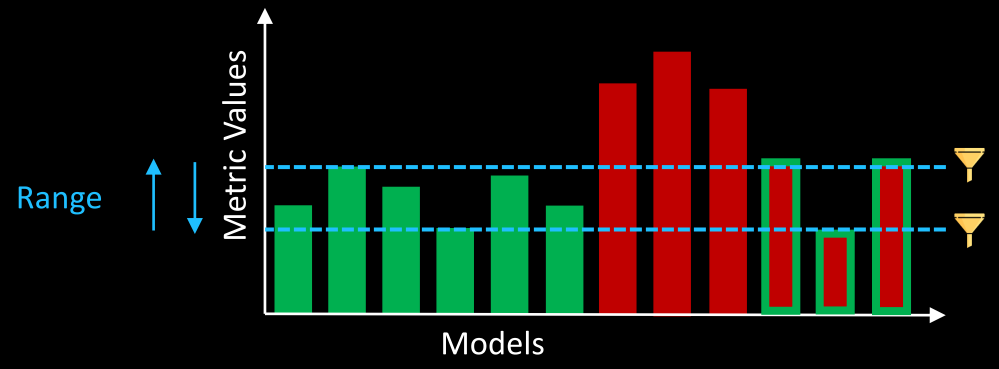
$$Loss = Loss_{data} + Loss_{adaption}$$

AutoAdapt: Automatic Adversarial Adaption



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

AutoAdapt: Automatic Adversarial Adaption



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

$$\Big\downarrow \textit{Trade-Off}$$

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

$$\Big\downarrow \textit{AutoAdapt}$$

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

$$Loss_{AutoAdapt} = \frac{1}{2\alpha_{AutoAdapt}} \sum_{j=1}^{m} \left(\left| \max(0, \mu_j + \alpha_{AutoAdapt}\, Loss_j) \right|^2 - \mu_j^2 \right)$$

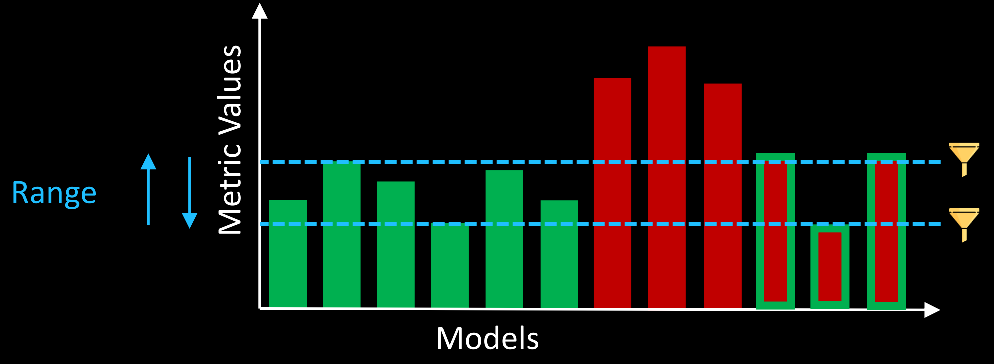
$$\mu_j = \begin{cases} \mu_j + \alpha_{AutoAdapt}\, Loss_j, & if\, Loss_j \geq 0 \\ 0, & if\, Loss_j < 0 \end{cases}$$

Solution

- Replace α with Augmented Lagrangian (AL)* Method
- Extend AL for multiple range constraints (if we want to detect in several metrics)
- No manual hyperparameters
  → $\alpha_{AutoAdapt}$ is insensitive
- Automatic switching off of the $Loss_{AutoAdapt}$ for constraints that are already fulfilled

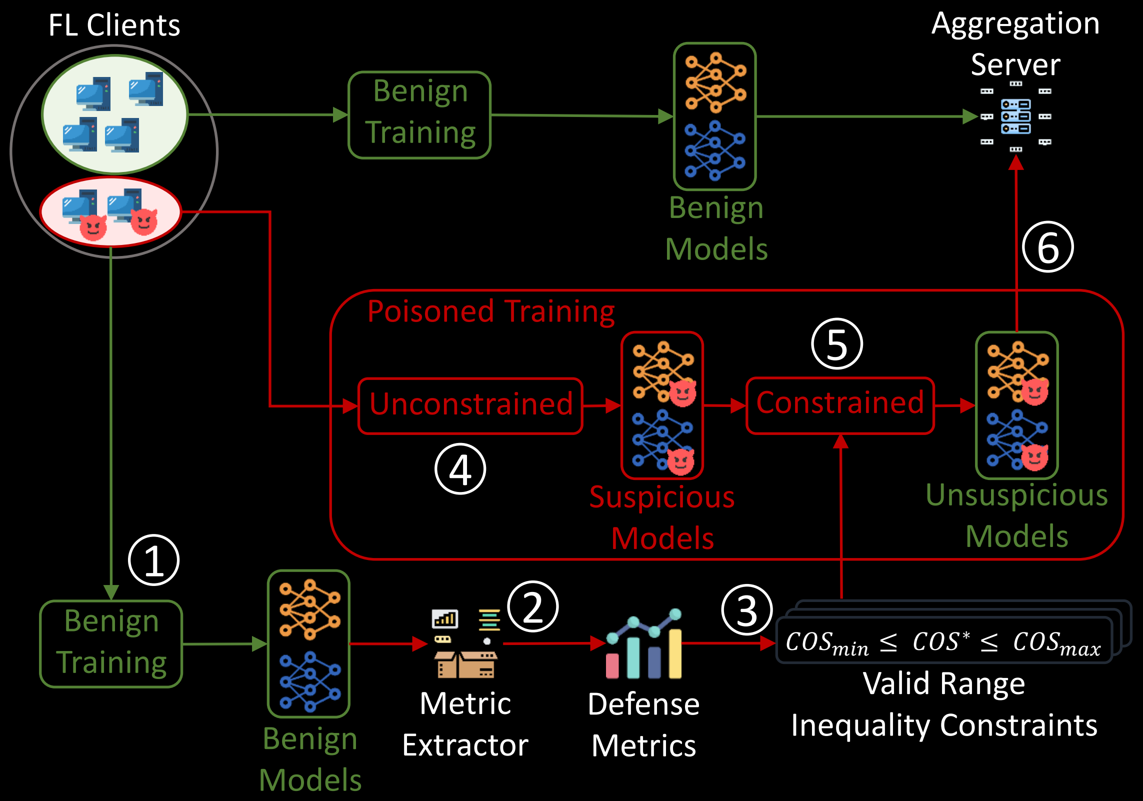
***Augmented Lagrangian methods** are a certain class of algorithms for solving constrained optimization problems

AutoAdapt: Automatic Adversarial Adaption

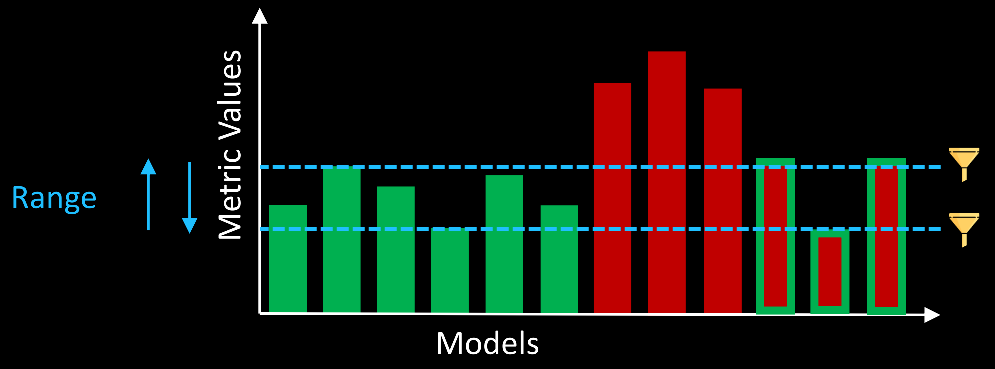


$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

Workflow



AutoAdapt: Automatic Adversarial Adaption



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

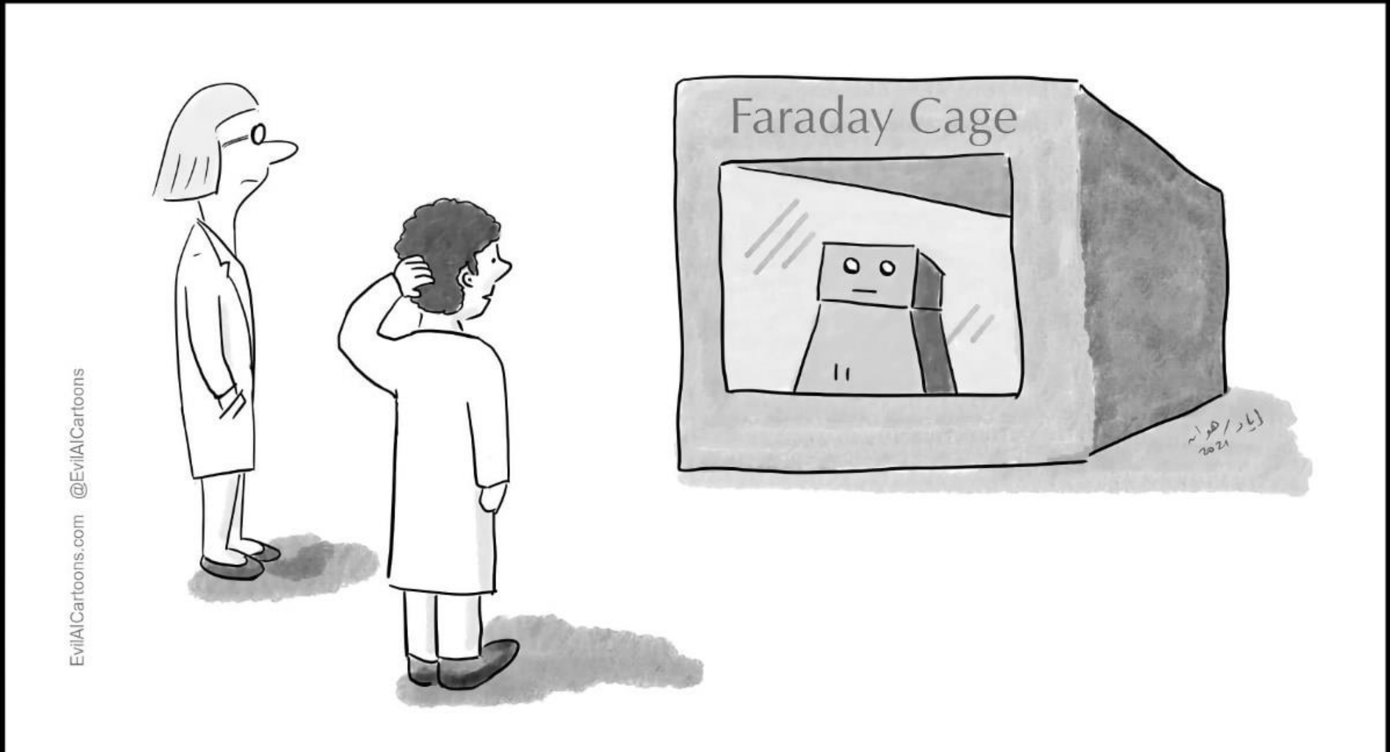
Results

- Successfull adaption to multiple range constraints simultaneously
- Adaption on a model-wise and layer-wise level
- Showcased circumvention of five state-of-the-art defenses
- Quick adaption (mostly within 1-3 training epochs)

→ We propose to use AutoAdapt as a new baseline for evaluation of new FL poisoning defenses

Conclusion

- Federated Learning helps solving high data demand vs. privacy dilemma
- Similar to centralized ML, FL is also prone to untargeted and targeted **poisoning attacks**
- An arm raise between attacks and defenses is going on and will continue



*"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