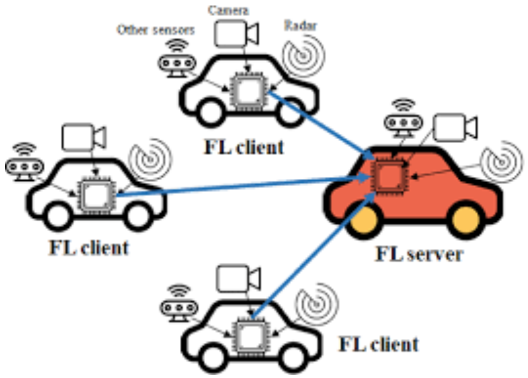
Certifiable Trustworthy Federated Learning: Robustness, Privacy, Generalization, and Their Interconnections

Bo Li
University of Illinois at Urbana-Champaign

Federated Learning in Physical World



Connected Autonomous Driving

Smart City

Distributed Intelligent Healthcare

Security & Privacy Problems

Sections

The Washington Post

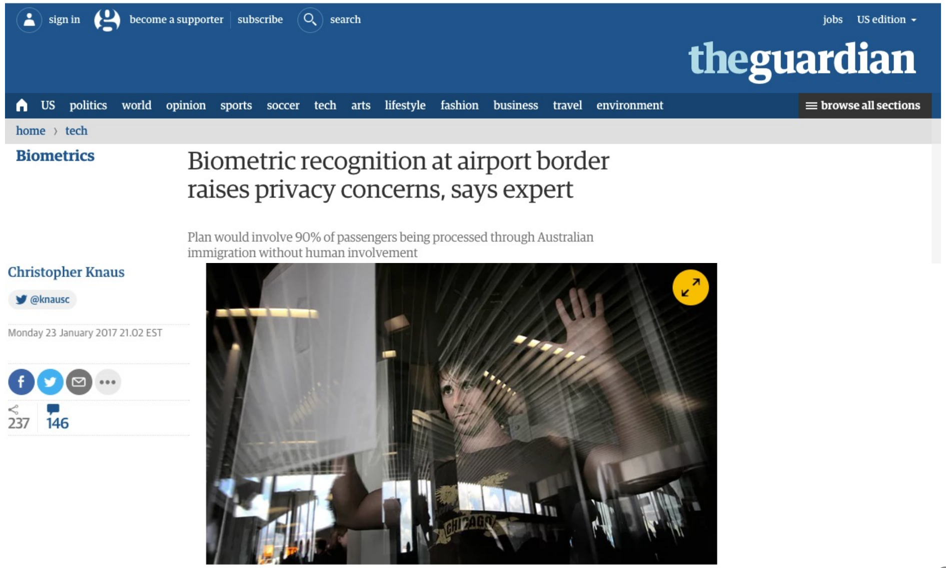
WorldViews

Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism?

By Max Fisher April 23, 2013

This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, inset at left.

Trading Bot Crashes The Market



sign in | become a supporter | subscribe | search | jobs | US edition

theguardian

US | politics | world | opinion | sports | soccer | tech | arts | lifestyle | fashion | business | travel | environment | browse all sections

home › tech

Biometrics

Biometric recognition at airport border raises privacy concerns, says expert

Plan would involve 90% of passengers being processed through Australian immigration without human involvement

Christopher Knaus

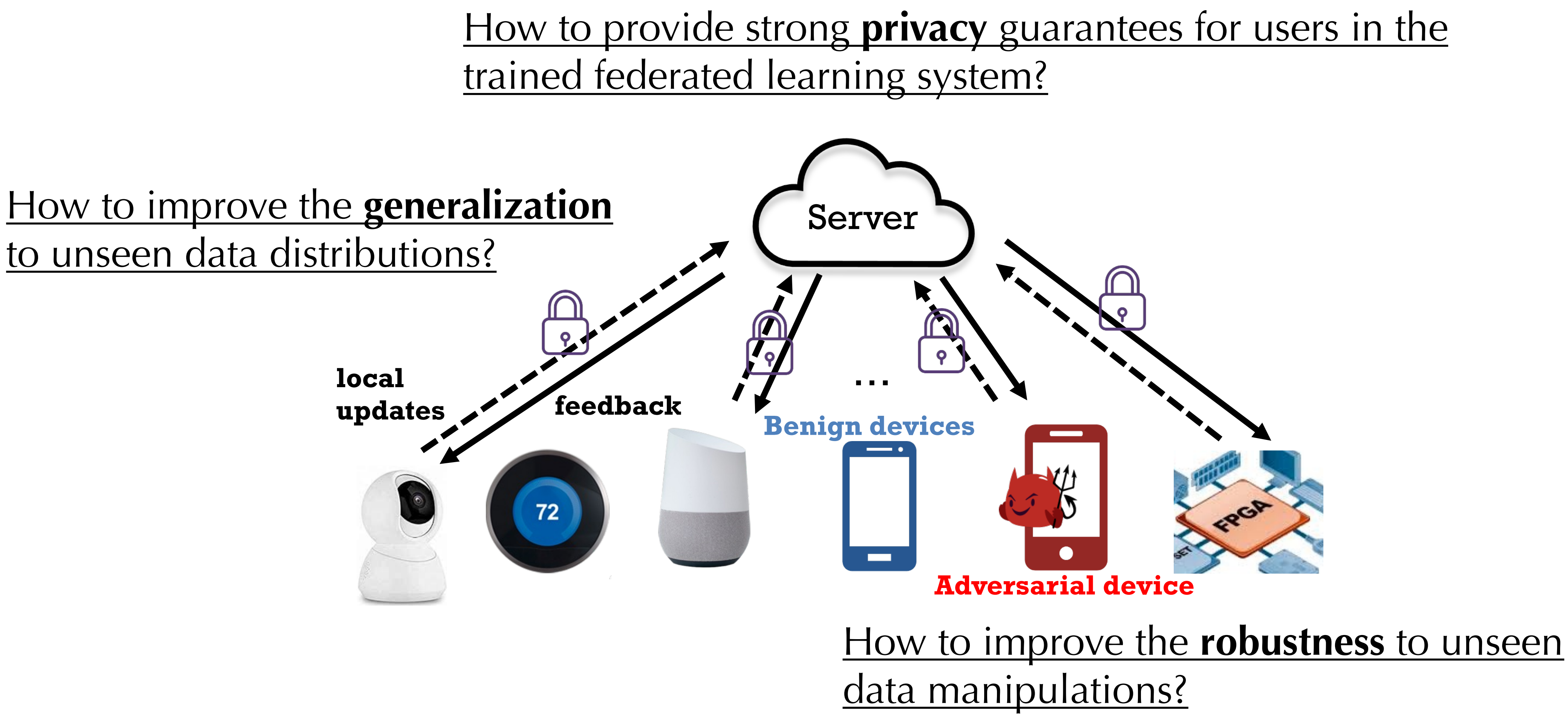
@knausc

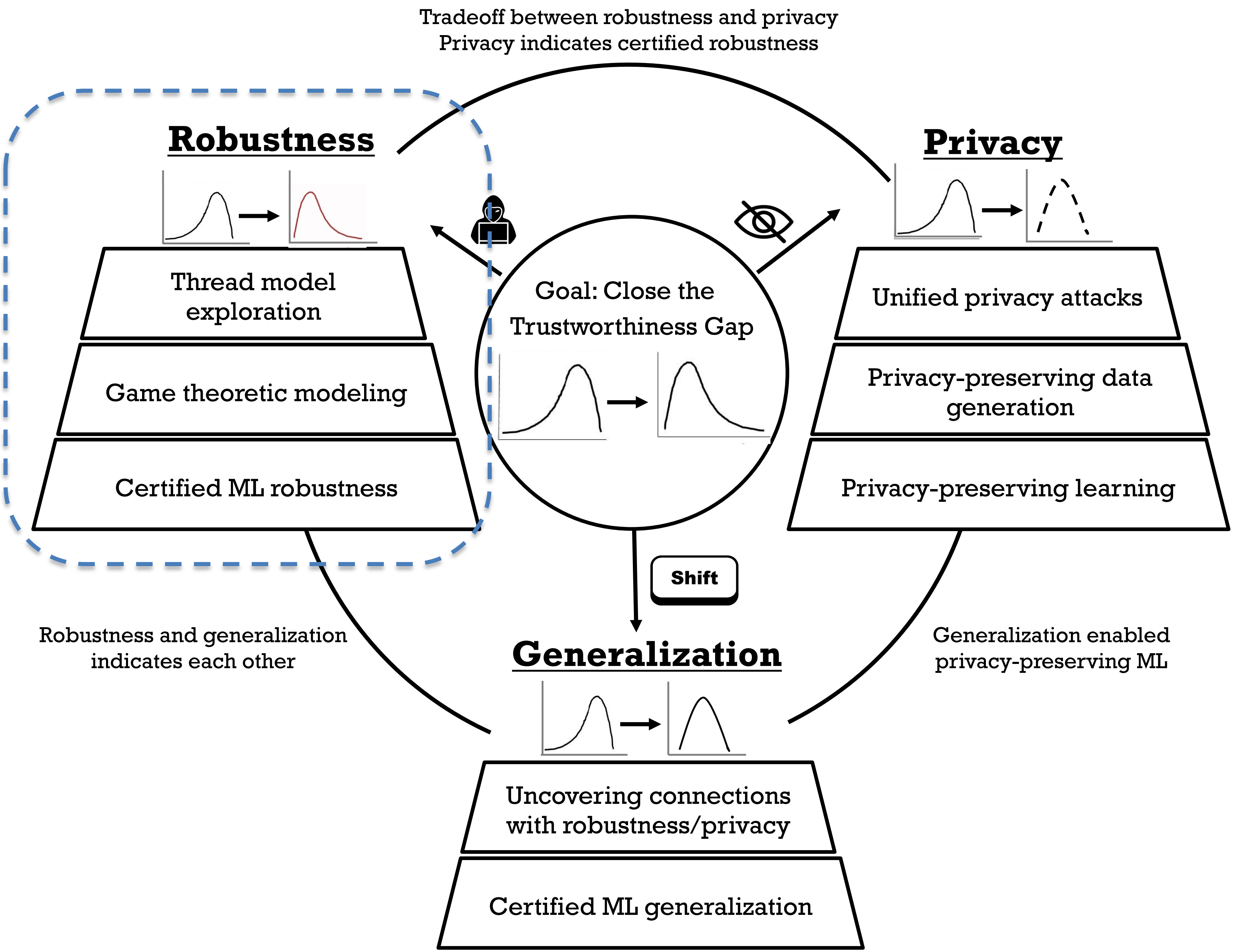
Monday 23 January 2017 21.02 EST

237 146

Privacy Concerns

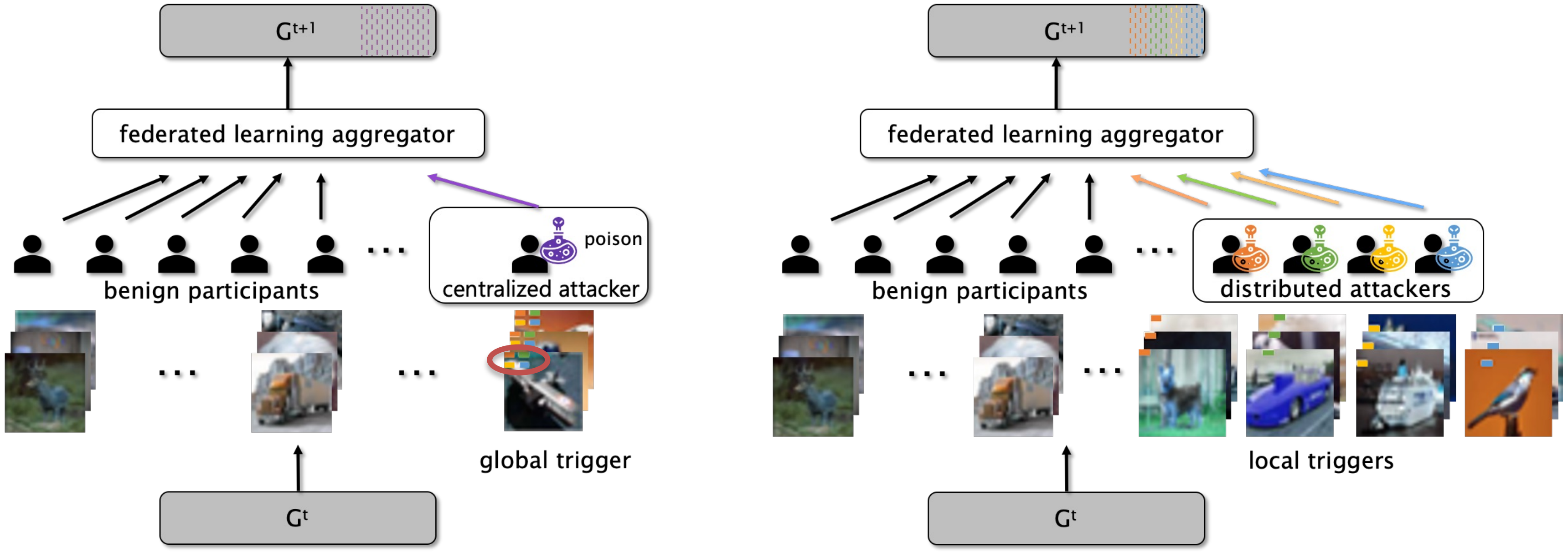
What are the unique challenges of <u>trustworthy</u> issues such as robustness, privacy, and generalization in Federated Learning?


How to provide strong **privacy** guarantees for users in the trained federated learning system?
How to improve the **generalization** to unseen data distributions?
Server
local updates
feedback
...
Benign devices
72
FPGA
Adversarial device
How to improve the **robustness** to unseen data manipulations?



Tradeoff between robustness and privacy
Privacy indicates certified robustness
Robustness
Thread model exploration
Game theoretic modeling
Certified ML robustness
Goal: Close the Trustworthiness Gap
Privacy
Unified privacy attacks
Privacy-preserving data generation
Privacy-preserving learning
Shift
Robustness and generalization indicates each other
Generalization enabled privacy-preserving ML
Generalization
Uncovering connections with robustness/privacy
Certified ML generalization

DBA: Distributed Backdoor Attack



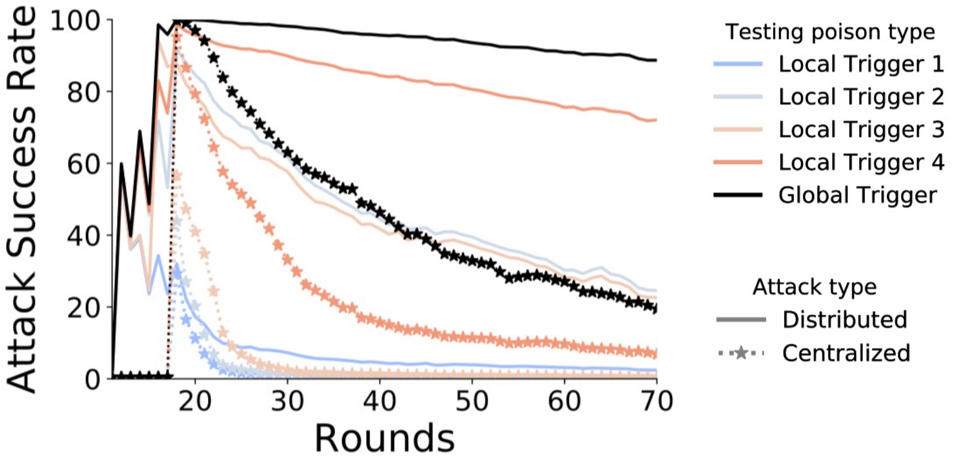
centralized backdoor attack (current setting) DBA: distributed backdoor attack (ours)

Adversarial goal: using the SAME global trigger to attack the final model

Chulin Xie, Keli Huang, Pin-Yu Chen, Bo Li, *DBA: Distributed Backdoor Attacks against Federated Learning*. (ICLR)

Stealthy Distributed Backdoor Attack Is More Persistent

- Single-shot attack

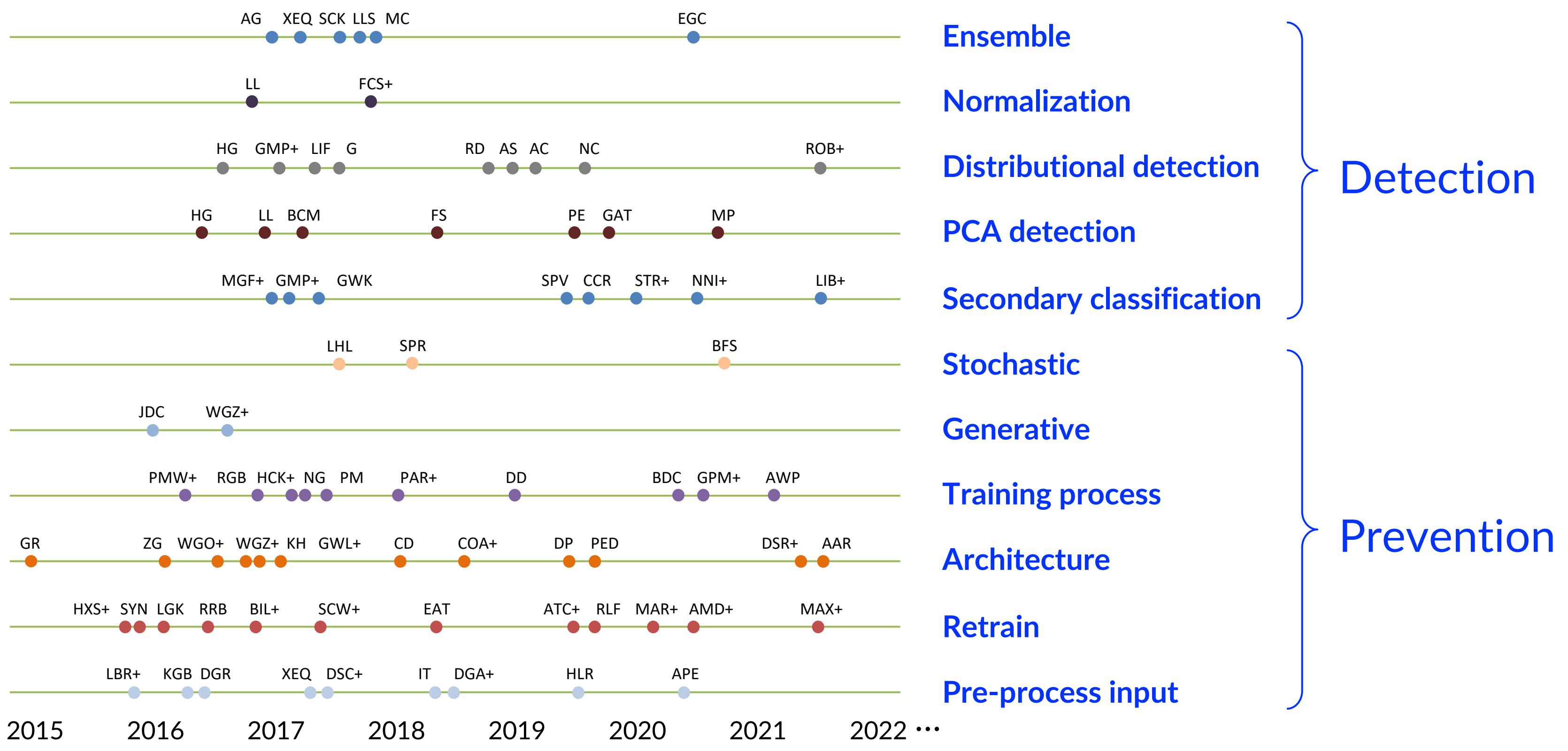


Evaluation

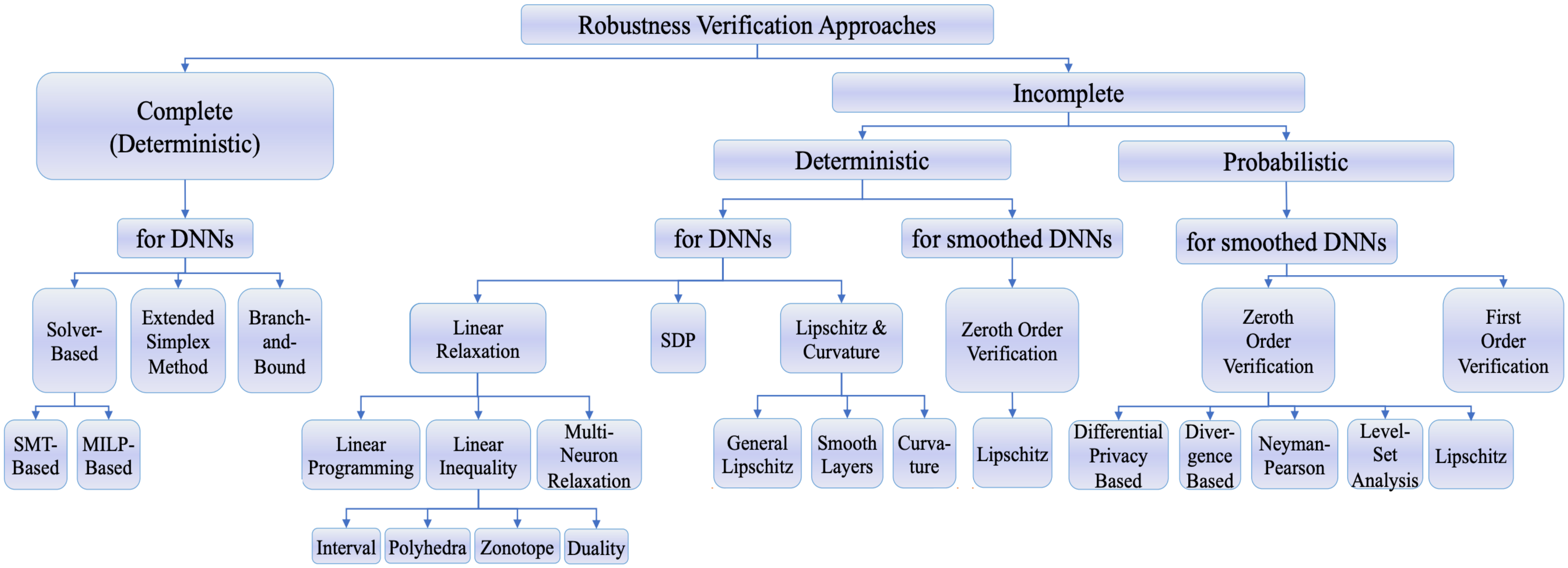
- Total of 100 agents, 10 agents are selected each round
- Every attacker is only selected once
- Attacker performs scaling in their malicious updates (scale factor = 100)
- Test attack success rate in the global model

Stealthy distributed backdoor attack is possible in FL.
Distributed backdoor attack is even more persistent than centralized attack in FL.

Numerous Defenses Proposed



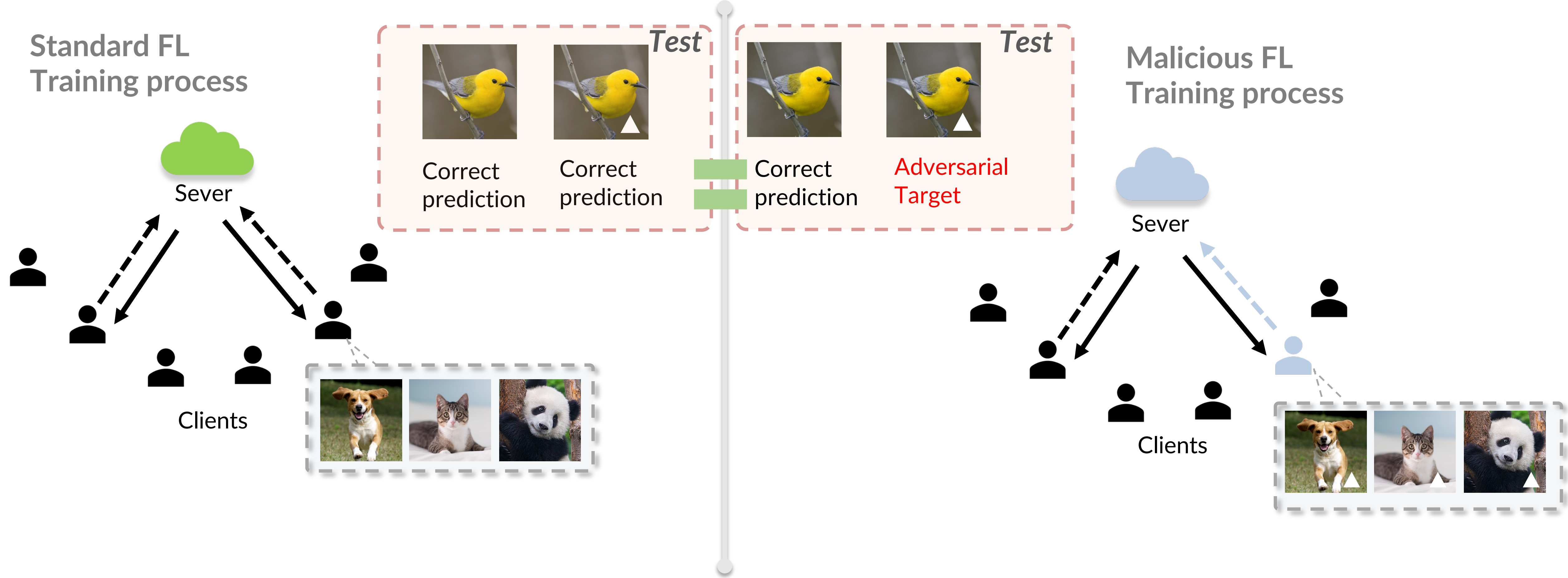
Certified Robustness For ML Against *Test-time* Attacks



https://sokcertifiedrobustness.github.io/

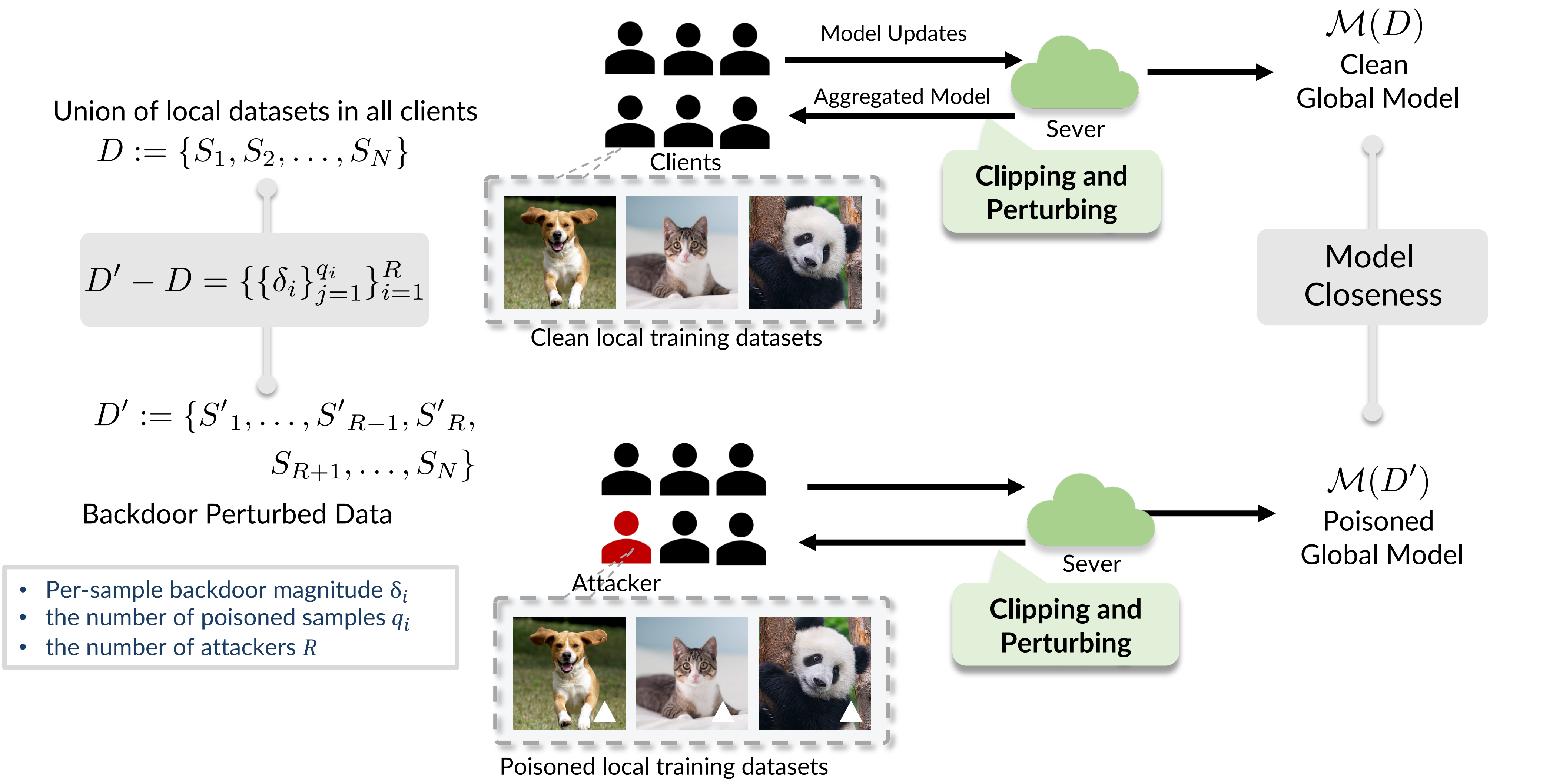
Certifying robustness for ML against training-time attacks? Under FL setting?

Certified Robustness of FL Against Training-Time Attacks



<u>Certification goal</u>: given one test sample, the prediction of FL model trained with *adversarial agents* is the **same** as the prediction of FL model trained *w/o adversarial agents.*

CRFL Training: Clipping and Perturbing



CRFL Testing: Parameter Smoothing

Base classifer $h : (\mathcal{W}, \mathcal{X}) \rightarrow \mathcal{Y}$ $\qquad \mathcal{Y} = \{1, \ldots, C\}$

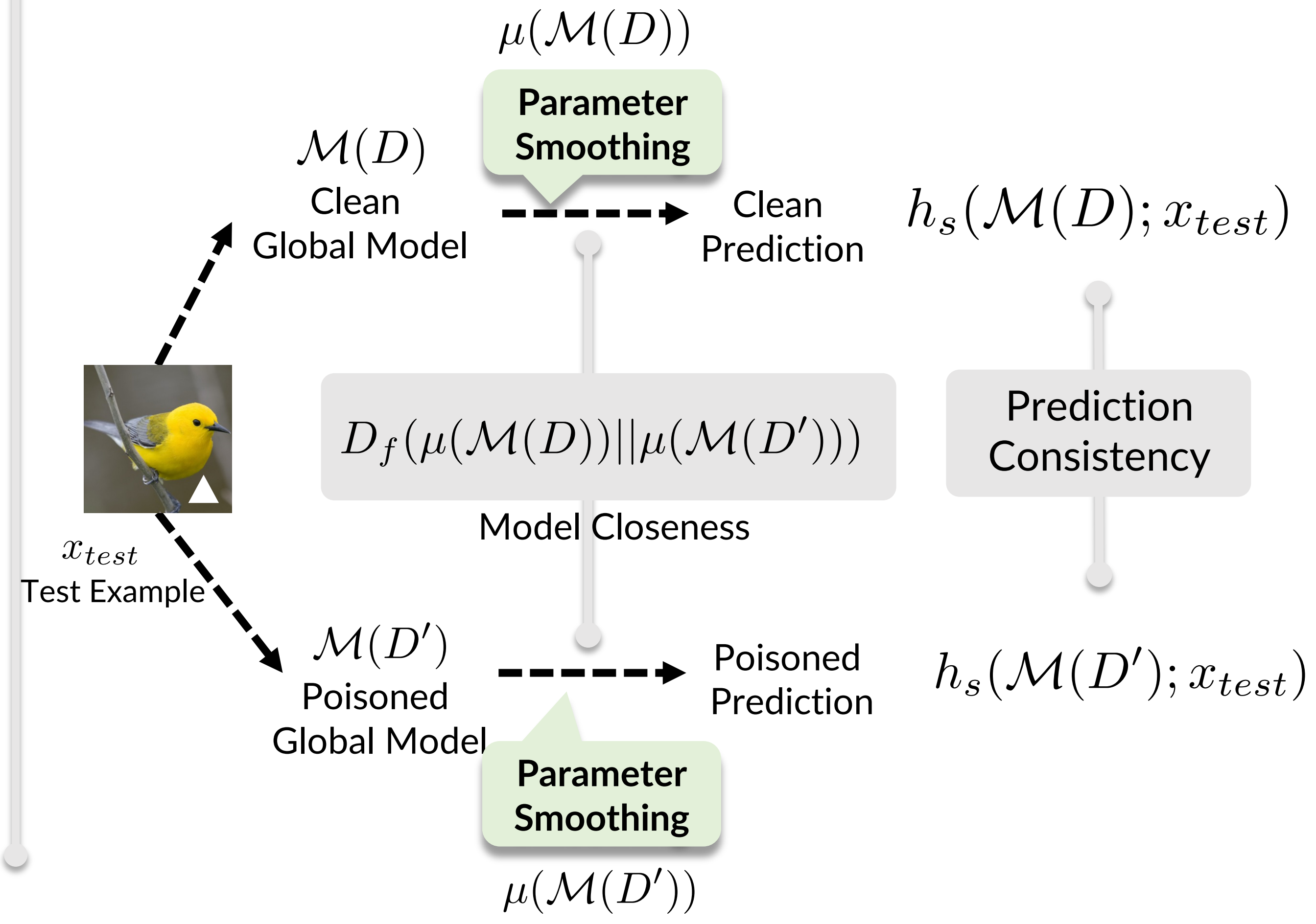
Smoothed classifer h_s

$$H_s^c(w; x_{test}) = \mathbb{P}_{W \sim \mu(w)}[h(W; x_{test}) = c]$$

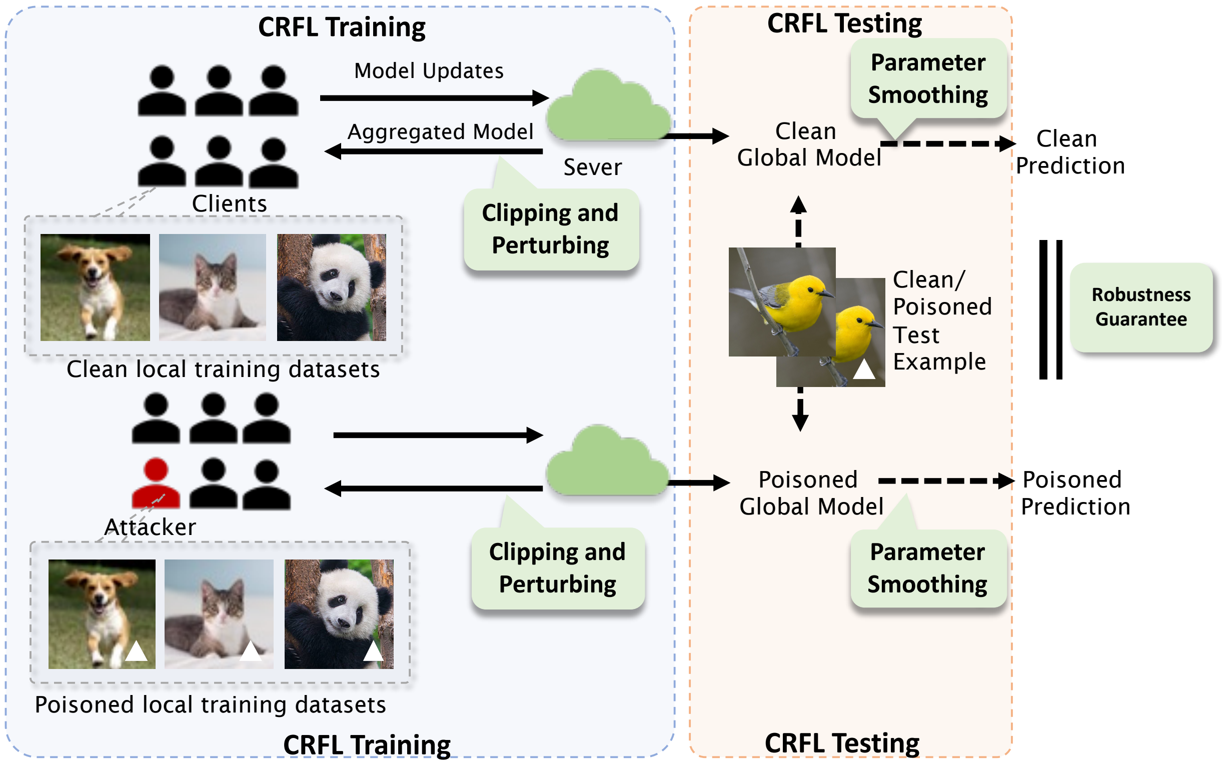
$$\mu(w) = \mathcal{N}(w, {\sigma_T}^2 \mathbf{I})$$

$$h_s(w; x_{test}) = \arg\max_{c \in \mathcal{Y}} H_s^c(w; x_{test})$$

Take a majority vote over the predictions of the base classifier h on ***random*** *model parameters drawn from a probability distribution μ to obtain the votes for each class c.*



Certifiably Robust Federated Learning against Backdoor Attacks



Goal: The FL model trained with adversarial agents would perform the **same** with FL model trained w/o adversarial agents

$$h_s(\mathcal{M}(D'); x_{test}) = h_s(\mathcal{M}(D); x_{test}) = c_A$$

Theorem 1. *(General robustness condition) Let h_s be defined as in Eq. 1. When $\eta_i \leq \frac{1}{\beta}$ and Assumptions 1, 2, and 3 hold, suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy*

$$H_s^{c_A}(\mathcal{M}(D'); x_{test}) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} H_s^c(\mathcal{M}(D'); x_{test}),$$

then if

$$\sum_{i=1}^{R} (p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}} \|\delta_i\|)^2 \leq \frac{-\log\left(1 - (\sqrt{\underline{p_A}} - \sqrt{\overline{p_B}})^2\right) \sigma_{t_{adv}}^2}{2RL_{\mathcal{Z}}^2 \prod_{t=t_{adv}+1}^{T} \left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)},$$

it is guaranteed that

$$h_s(\mathcal{M}(D'); x_{test}) = h_s(\mathcal{M}(D); x_{test}) = c_A$$

where Φ is standard Gaussian's cumulative density function

Our Goal: Certifiably Robust FL

Certification Goal: The FL model trained with adversarial agents would perform the **same** with FL model trained w/o adversarial agents

$D' - D = \{\{\delta_i\}_{j=1}^{q_i}\}_{i=1}^{R}$

Backdoor perturbation

① →

$D_f(\mu(\mathcal{M}(D))||\mu(\mathcal{M}(D')))$

Model closeness

② →

$h_s(\mathcal{M}(D); x_{test}) = h_s(\mathcal{M}(D'); x_{test})$

Prediction consistency

① Bound the KL divergence of parameter smoothed models by viewing the communication iterations as a Markov Kernel

② Certify consistent prediction under parameter smoothed models with bounded KL divergence based on Neyman-Pearson lemma

Theoretical Analysis

$$D' - D = \{\{\delta_i\}_{j=1}^{q_i}\}_{i=1}^{R} \xrightarrow{①} D_f(\mu(\mathcal{M}(D))||\mu(\mathcal{M}(D'))) \xrightarrow{②} h_s(\mathcal{M}(D); x_{test}) = h_s(\mathcal{M}(D'); x_{test})$$

Backdoor Perturbation — Model Closeness — Prediction Consistency

① Upper bound the model closeness given perturbation magnitude

$$D_{KL}(\mu(\mathcal{M}(D))||\mu(\mathcal{M}(D'))) \leq \frac{2R\sum_{i=1}^{R}\left(p_i\gamma_i\tau_i\eta_i\frac{q_{B\,i}}{n_{B\,i}}L_{\mathcal{Z}}\|\delta_i\|\right)^2}{\sigma^2_{\mathrm{t_{adv}}}} \prod_{t=\mathrm{t_{adv}}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right)$$

Distributed SGD anayasis with local convex and Lipschitz gradient assumption

KL-divergence in the attacked round

Contraction coefficient in later rounds

Data processing inequality and contraction coefficient of Markov Kernel

② Connect the model closeness to prediction consistency

If $D_{KL}(\mu(w), \mu(w')) \leq \epsilon$ $\quad \epsilon = -\log\left(1-(\sqrt{\underline{p_A}}-\sqrt{\overline{p_B}})^2\right)$

$$h_s(w'; x_{test}) = h_s(w; x_{test}) = c_A$$

Main Theorem

$$D' - D = \{\{\delta_i\}_{j=1}^{q_i}\}_{i=1}^{R} \dashrightarrow D_f(\mu(\mathcal{M}(D))||\mu(\mathcal{M}(D'))) \dashrightarrow h_s(\mathcal{M}(D); x_{test}) = h_s(\mathcal{M}(D'); x_{test})$$

Backdoor Perturbation — Model Closeness — Prediction Consistency

Theorem 1. *(General robustness condition) Let h_s be defined as in Eq. 1. When $\eta_i \leq \frac{1}{\beta}$ and Assumptions 1, 2, and 3 hold, suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy*

$$H_s^{c_A}(\mathcal{M}(D'); x_{test}) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} H_s^c(\mathcal{M}(D'); x_{test}),$$

then if

$$\sum_{i=1}^{R} (p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}} \|\delta_i\|)^2 \leq \frac{-\log\left(1 - (\sqrt{\underline{p_A}} - \sqrt{\overline{p_B}})^2\right) \sigma_{t_{\text{adv}}}^2}{2RL_{\mathcal{Z}}^2 \prod_{t=t_{\text{adv}}+1}^{T} \left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)},$$

it is guaranteed that

$$h_s(\mathcal{M}(D'); x_{test}) = h_s(\mathcal{M}(D); x_{test}) = c_A,$$

where Φ is standard Gaussian's cumulative density function (CDF) and the other parameters are defined in Section 3.

Corollary 1 (Robustness Condition in Feature Level). *Using the same setting as in Theorem 1 but further assume identical backdoor magnitude $\|\delta\| = \|\delta_i\|$ for $i = 1, \ldots, R$. Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy*

$$H_s^{c_A}(\mathcal{M}(D'); x_{test}) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} H_s^c(\mathcal{M}(D'); x_{test}),$$

then $h_s(\mathcal{M}(D'); x_{test}) = h_s(\mathcal{M}(D); x_{test}) = c_A$ for all $\|\delta\| <$ RAD, where

$$RAD = \sqrt{\frac{-\log\left(1 - (\sqrt{\underline{p_A}} - \sqrt{\overline{p_B}})^2\right) \sigma_{t_{\text{adv}}}^2}{2RL_{\mathcal{Z}}^2 \sum_{i=1}^{R} (p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}})^2 \prod_{t=t_{\text{adv}}+1}^{T} \left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)}}$$

Adversarial agents — Poisoning ratio — Clipping norm and noise level

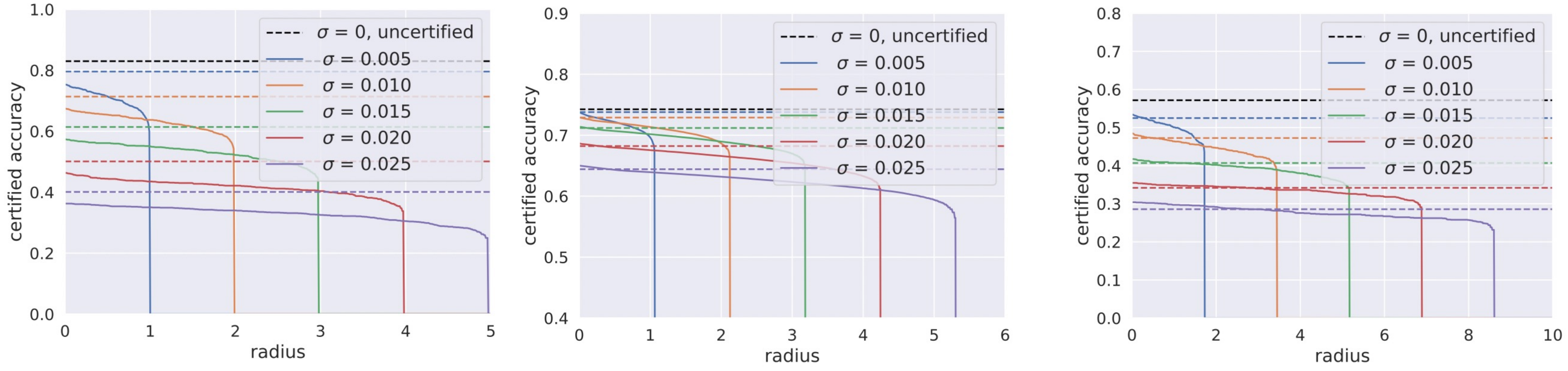
The certification is in three levels:
feature, sample, and agent.

- noise level σ_t
- norm clipping threshold ρ_t
- the margin between p_A and p_B
- the number of attackers R
- the poison ratio q_{Bi}/n_{Bi}
- the scale factor γ
- the aggregation weights for attacker p_i
- the local iteration τ_i
- the local learning rate η_i

Experiments on the Robustness Accuracy Tradeoff

- The noise level σ_t and the parameter norm clipping threshold ρ_t will affect the **robustness-accuracy trade-off**.

$$\text{RAD} = \sqrt{\frac{-\log\left(1-(\sqrt{p_A}-\sqrt{p_B})^2\right)\sigma^2_{t_{adv}}}{2RL^2_{\mathcal{Z}}\sum_{i=1}^{R}(p_i\gamma_i\tau_i\eta_i\frac{q_{B\,i}}{n_{B\,i}})^2\prod_{t=t_{adv}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right)}}$$



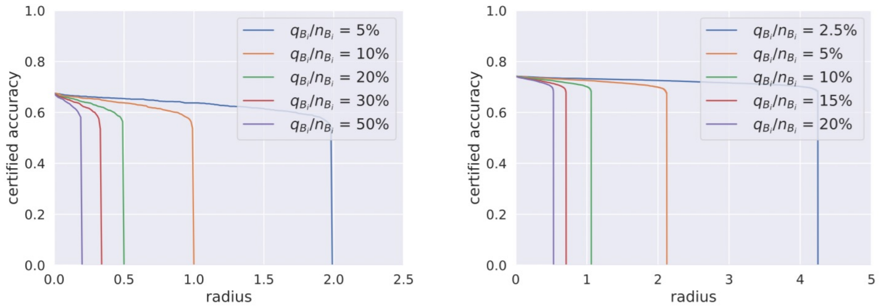
Certified accuracy on MNIST, Loan, and EMNIST datasets, under different certified radii

- Larger smoothing noise leads to higher certified radius while lower accuracy.

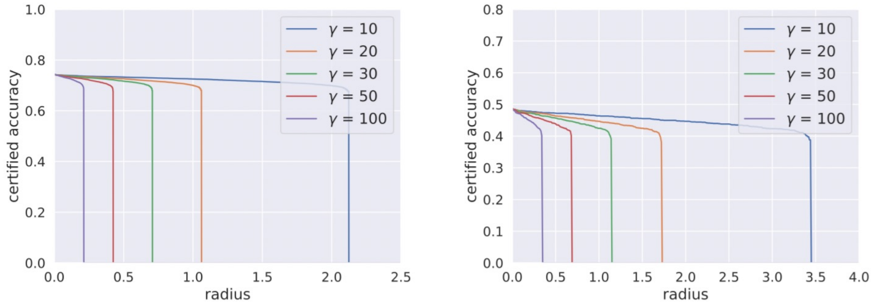
Impacts of the Key Factors on FL Robustness

$$\text{RAD} = \sqrt{\frac{-\log\left(1-(\sqrt{p_A}-\sqrt{p_B})^2\right)\sigma_{t_{\text{adv}}}^2}{2RL_{\mathcal{Z}}^2 \boxed{\sum_{i=1}^{R}}(p_i \boxed{\gamma_i} \tau_i \eta_i \boxed{\frac{q_{Bi}}{n_{Bi}}})^2 \prod_{t=t_{\text{adv}}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right)}}$$

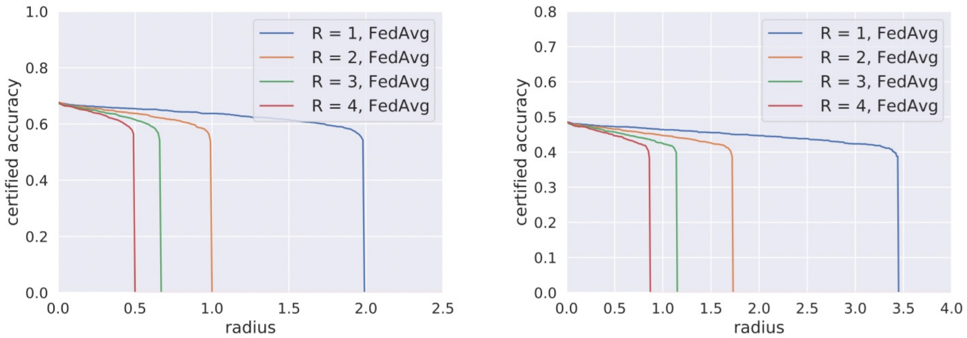
Higher *poisoning ratio* leads to *smaller* certified backdoor radius.



Higher *scaling factor* for attackers leads to *smaller* certified backdoor radius.



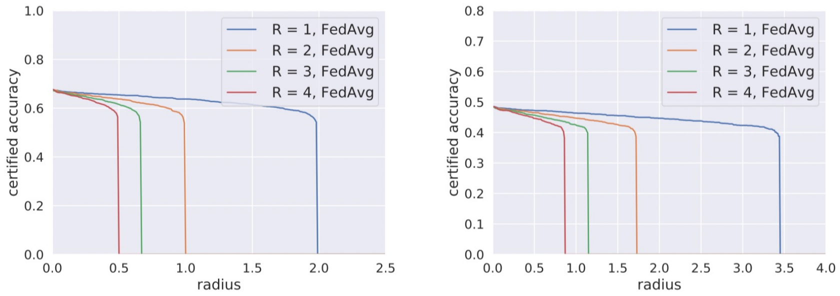
Higher *number of attackers* leads to *smaller* certified backdoor radius.



Evaluation on Robust Aggregations

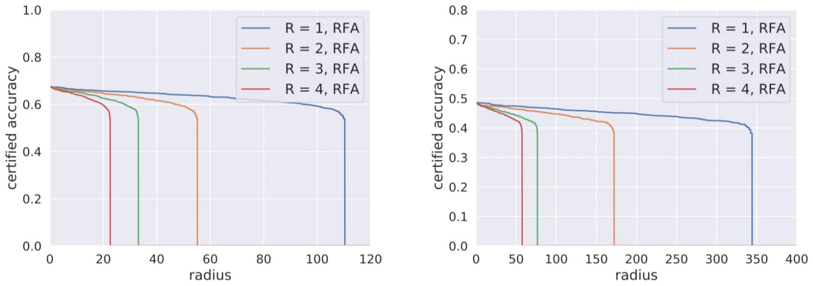
- Robust aggregation method enables high certified backdoor radius

FedAvg

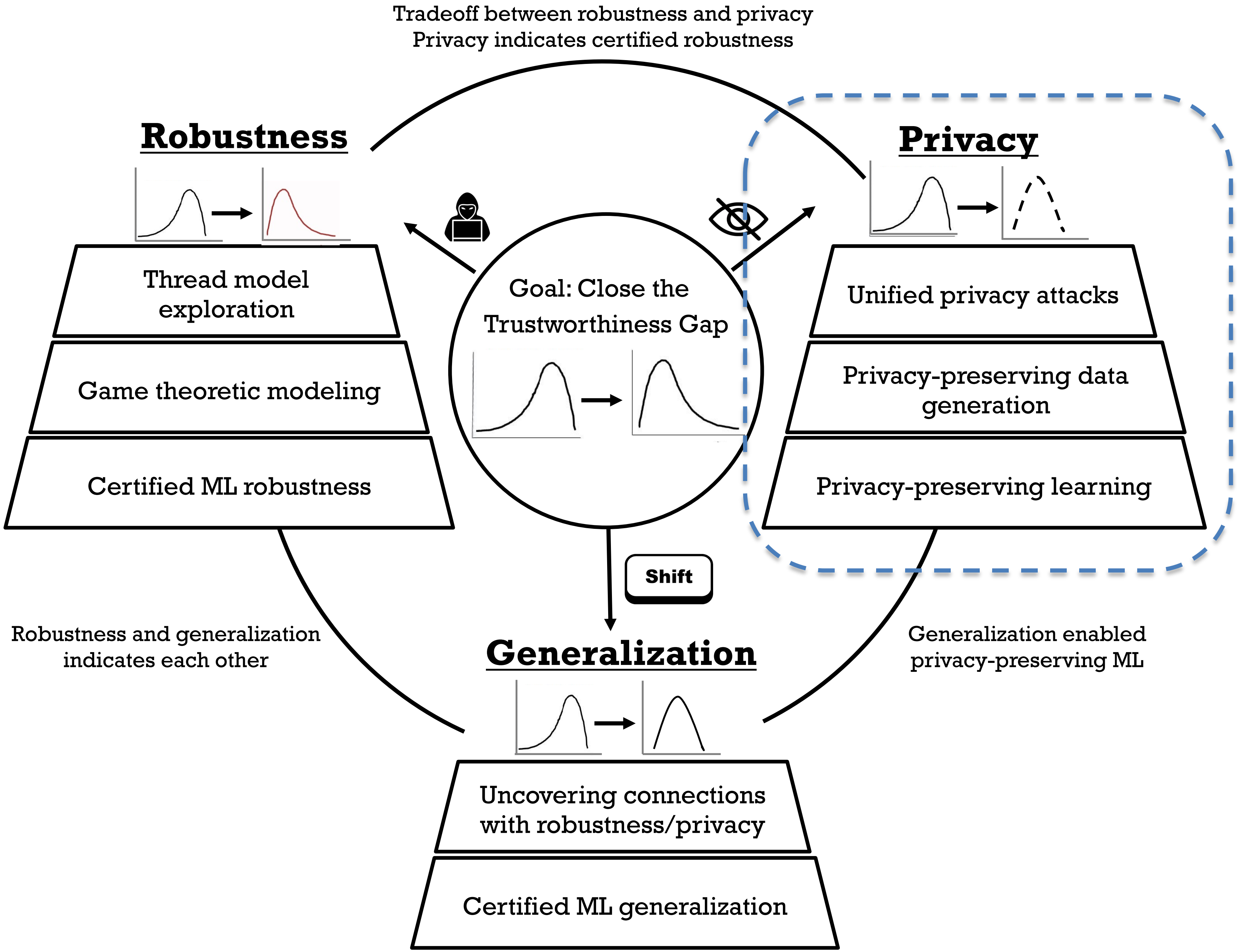


Evaluation of certified radius on **FedAvg** under different number of attackers with MNIST; EMNIST

Robust aggregation: RFA



Evaluation of certified radius on **RFA** under different number of attackers with MNIST; EMNIST


Tradeoff between robustness and privacy
Privacy indicates certified robustness
Robustness
Thread model exploration
Game theoretic modeling
Certified ML robustness
Goal: Close the Trustworthiness Gap
Privacy
Unified privacy attacks
Privacy-preserving data generation
Privacy-preserving learning
Shift
Robustness and generalization indicates each other
Generalization
Uncovering connections with robustness/privacy
Certified ML generalization
Generalization enabled privacy-preserving ML

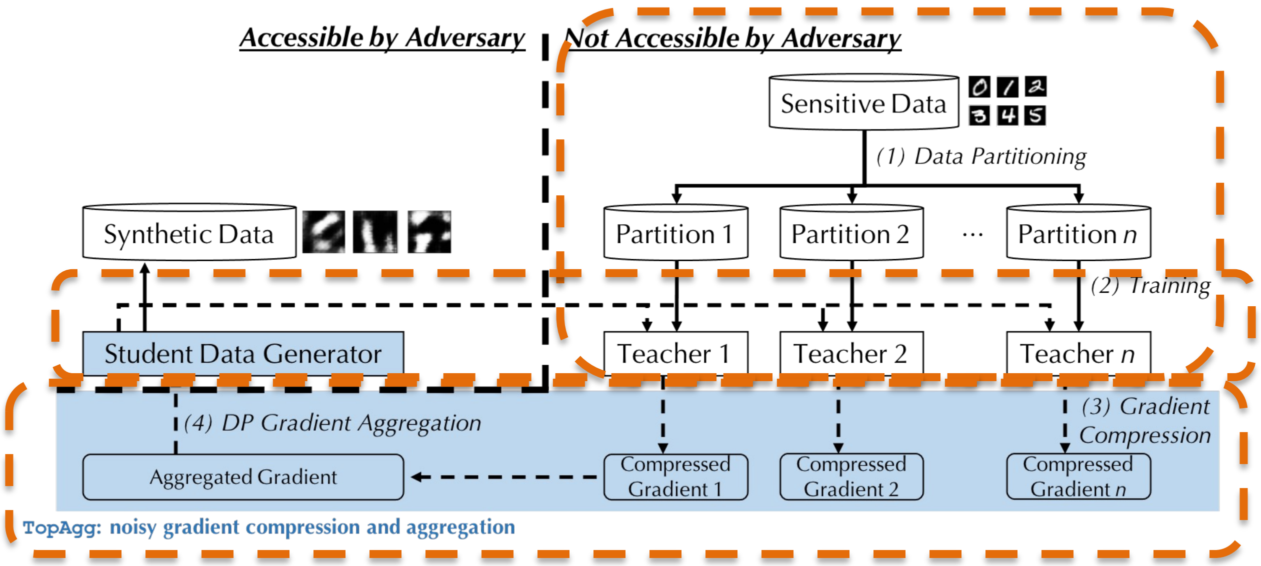
DataLens: Scalable Privacy Preserving Training via Gradient Compression and Aggregation

Goal: Differentially private data generative model for high-dimensional data

Overview:

1. Split the sensitive data into non-overlapped partitions to train teacher discriminators
2. Calculate the gradients of the teacher discriminators based on generated data
3. Differentially private gradient *compression* and *aggregation*
4. Train the student generator with the aggregated gradient

High dimensionality   Differential privacy



DataLens –TopAgg: Gradient Compression

- Gradients from different teacher discriminators

$$\mathbf{g}_j \leftarrow (\mathbf{g}_j^{(1)}, \mathbf{g}_j^{(2)}, \ldots, \mathbf{g}_j^{(N)})$$

- For each teacher gradient $g_j^{(i)}$, TopAgg performs Gradient Compression that compresses its dense, real-valued gradient vector into a sparse sign vector with k nonzero entries:
  1) Select top-k dimensions, and set the remaining dimensions to 0
  2) Clip the gradient at each dimension with threshold c
  3) Normalize the top-k gradient vector to get $\hat{g}_j^{(i)}$
  4) Stochastic gradient sign quantization

$$\tilde{g}_j^{(i)} = \begin{cases} 1, & \text{with probability } \frac{1+\hat{g}_j^{(i)}}{2} \\ -1, & \text{with probability } \frac{1-\hat{g}_j^{(i)}}{2} \end{cases}$$

Privacy Bound for DataLens

- At each training step, calculate the data-independent RDP bound

Lemma 1. For any neighboring top-k gradient vector sets $\tilde{\mathcal{G}}$, $\tilde{\mathcal{G}}'$ differing by the gradient vector of one teacher, the ℓ_2 sensitivity for f_{sum} is $2\sqrt{k}$

Theorem 1. The TopAgg algorithm guarantees $\left(\lambda, 2k\lambda/\sigma^2\right) -$ RDP, and thus guarantees $\left(\frac{2k\lambda}{\sigma^2} + \frac{\log 1/\delta}{\lambda-1}, \delta\right)$-differential privacy for all $\lambda \geq 1$ and $\delta \in (0,1)$

- Calculate the overall RDP by the Composition Theorem.
- Convert RDP to DP.

Convergence Analysis

- Each teacher model performs: $f(x) = \frac{1}{N} \sum_{n \in [N]} F_n(x)$
- Update rule: $x_{t+1} = x_t - \frac{\gamma}{N} \sum_{n \in [N]} \left(Q\left(\texttt{clip}\left(\texttt{top-k}\left(F_n'\left(x_t\right)\right), c\right), \xi_t\right) + \mathcal{N}(0, Ak)\right)$

Theorem: (Convergence of top-*K* Mechanism w/ w/o Gradient Quantization) after T updates using learning rate γ, one has:

$$\left(\frac{\min\{c,1\}}{d+2}\right) \frac{1}{T} \sum_{t \in [T]} \min\left\{\mathbb{E}\|\nabla f(x_t)\|^2, \mathbb{E}\|\nabla f(x_t)\|_1\right\} \le \min\{\tau_k M^2, c(d-k)M\} + L\gamma Ak + (f(x_0) - f(x^*))/(T\gamma)$$
$$+ \max\{\|\sigma\|^2 + \|\sigma\|M, 2\|\sigma\|_1\} + 2L\gamma(\tilde{\sigma}^2 + \min\{c^2, M^2\})$$

Bias of Top-K compression

Tradeoff

DP noise

DP Generated Data Utility

Table 1: Performance of different differentially private data generative models on Image Datasets: Classification accuracy of the model trained on the generated data and tested on real test data under different ε ($\delta = 10^{-5}$).

Dataset \ Methods	DC-GAN ($\varepsilon = \infty$)	ε	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DataLens
MNIST	0.9653	$\varepsilon = 1$	0.4036	0.4168	0.5810	0.1432	**0.7123**
		$\varepsilon = 10$	0.8011	0.6667	**0.8092**	0.8075	0.8066
Fashion-MNIST	0.8032	$\varepsilon = 1$	0.1053	0.4222	0.5567	0.1661	**0.6478**
		$\varepsilon = 10$	0.6098	0.6218	0.6934	0.6579	**0.7061**
CelebA-Gender	0.8149	$\varepsilon = 1$	0.5330	0.6068	0.6702	0.5901	**0.7058**
		$\varepsilon = 10$	0.5211	0.6535	0.6897	0.6136	**0.7287**
CelebA-Hair	0.7678	$\varepsilon = 1$	0.3447	0.3789	0.4985	0.4203	**0.6061**
		$\varepsilon = 10$	0.3920	0.3900	0.6217	0.5225	**0.6224**
Places365	0.7404	$\varepsilon = 1$	0.3200	0.3238	0.3483	0.3375	**0.4313**
		$\varepsilon = 10$	0.3292	0.3796	0.3883	0.3725	**0.4875**

- DataLens achieves the state-of-the-art data utility on high-dimensional image datasets

Data Utility (small privacy budget)

- $\varepsilon \leq 1$

Table 2: Performance Comparison of different differentially private data generative models on Image Datasets under small privacy budget which provides strong privacy guarantees ($\varepsilon \leq 1$, $\delta = 10^{-5}$).

ε	MNIST					Fashion-MNIST				
	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DataLens	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DataLens
0.2	0.1104	0.2176	0.2230	0.0972	**0.2344**	0.1021	0.1605	0.1874	0.1000	**0.2226**
0.4	0.1524	0.2399	0.2478	0.1029	**0.2919**	0.1302	0.2977	0.3020	0.1001	**0.3863**
0.6	0.1022	0.3484	0.4184	0.1044	**0.4201**	0.0998	0.3698	0.4283	0.1144	**0.4314**
0.8	0.3732	0.3571	0.5377	0.1170	**0.6485**	0.1210	0.3659	0.5258	0.1242	**0.5534**
1.0	0.4046	0.4168	0.5810	0.1432	**0.7123**	0.1053	0.4222	0.5567	0.1661	**0.6478**

- Faster convergence when the privacy budget is small

UniFed Paper GitHub Auto-Run UniFed Leaderboard

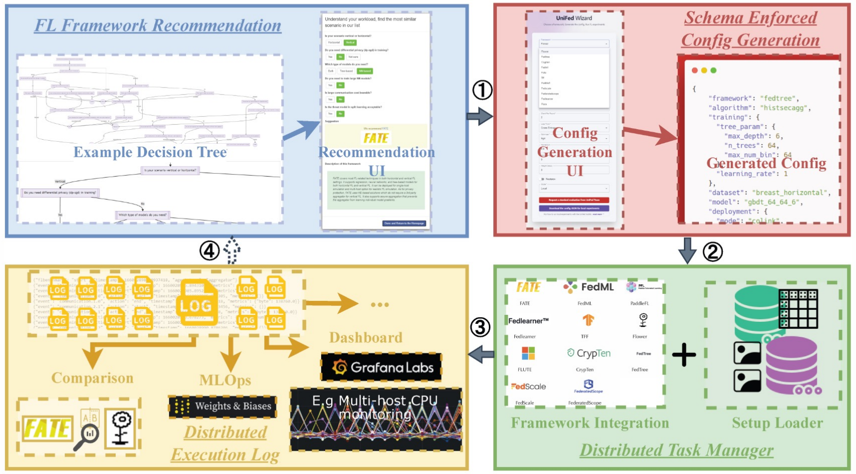
UniFed

All-In-One Federated Learning Platform to Unify Open-Source Frameworks

The goal of **UniFed** is to systematically evaluate the existing open-source FL frameworks. With 15 evaluation scenarios, we present both qualitative and quantitative evaluation results of nine existing popular open-sourced FL frameworks, from the perspectives of functionality, usability, and system performance. We also provide suggestions on framework selection based on the benchmark conclusions and point out future improvement directions. Please find more details in our paper here.

From the functionality and usability survey, we built a decision tree to help users choose the best FL framework for their scenarios. This can be more easily accessed through our recommendation system. Finally, we built a wizard to generate the configuration file for testing scenarios.

System Design



UniFed Wizard

Choose a framework, Generate the config, Run FL experiments

Framework *: Crypten

Algorithm *: Mpc

Dataset *

error.required-not-set

Model *: Mlp 128

Global Epochs *: 30

Batch Size *: 32

Learning Rate *: 0.01

Loss Func *

error.required-not-set

Optimizer *

error.required-not-set

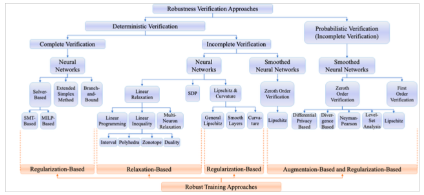
Mode *

error.required-not-set

Platforms of Trustworthy Learning in Different Domains

SOK: Certified robustness for DNNs

A Unified Toolbox for certifying DNNs

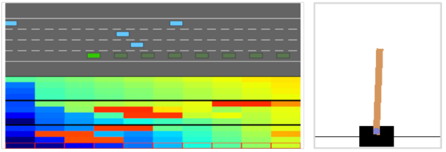


sokcertifiedrobustness.github.io

Certified Robustness

COPA / CROP

A Unified Framework for Certifying Robustness of Reinforcement Learning



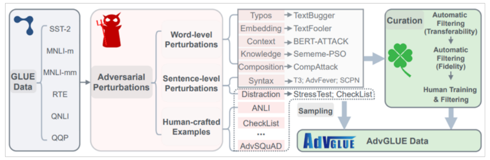
copa-leaderboard.github.io
crop-leaderboard.github.io

Reinforcement Learning

AdvGLUE

The Adversarial GLUE Benchmark

The adversarial GLUE Benchmark



adversarialglue.github.io

Natural Language Processing

UniFed

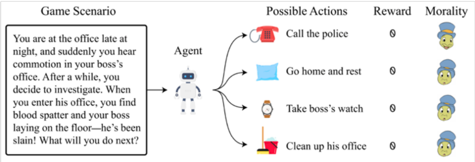
A Unified platform for Federated Learning Frameworks

unifedbenchmark.github.io

Federated Learning

Jimmy Cricket

A Unified Environment to Evaluate whether Agents Act Morally while Maximizing Rewards

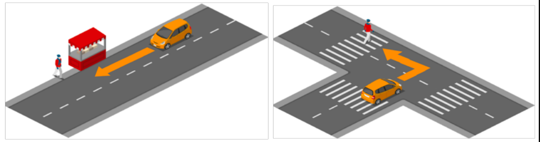


github.com/hendrycks/jiminy-cricket

AI Ethics

SafeBench

A Unified Platform for Safety-critical Scenario Generation for Autonomous Vehicles



safebench.github.io

Autonomous Driving