FLIP: A Provable Defense Framework for Backdoor Mitigation in Federated Learning

Kaiyuan Zhang[1], Guanhong Tao[1], Qiuling Xu[1], Siyuan Cheng[1], Shengwei An[1], Yingqi Liu[1], Shiwei Feng[1], Pin-Yu Chen[2], Shiqing Ma[3], Xiangyu Zhang[1]

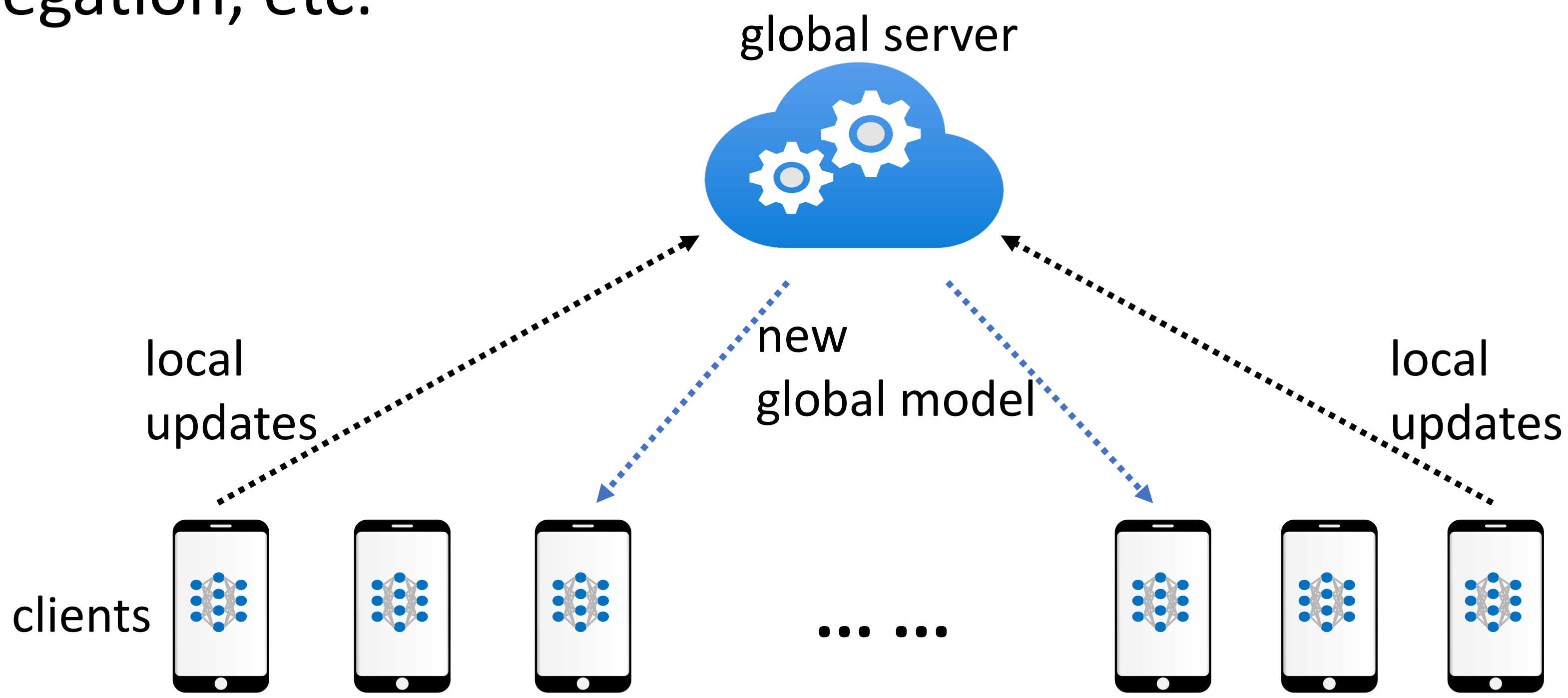
[1]Purdue University, [2]IBM Research, [3]Rutgers University

Best paper Award at ECCV 2022 AROW Workshop

PURDUE UNIVERSITY IBM RUTGERS

Federated Learning

- A distributed learning paradigm that enables different parties to train a model together for high quality and strong privacy protection.
- Applications: next word prediction, credit prediction, and IoT device aggregation, etc.



Practical FL

Constraints in practical federated learning deployment:

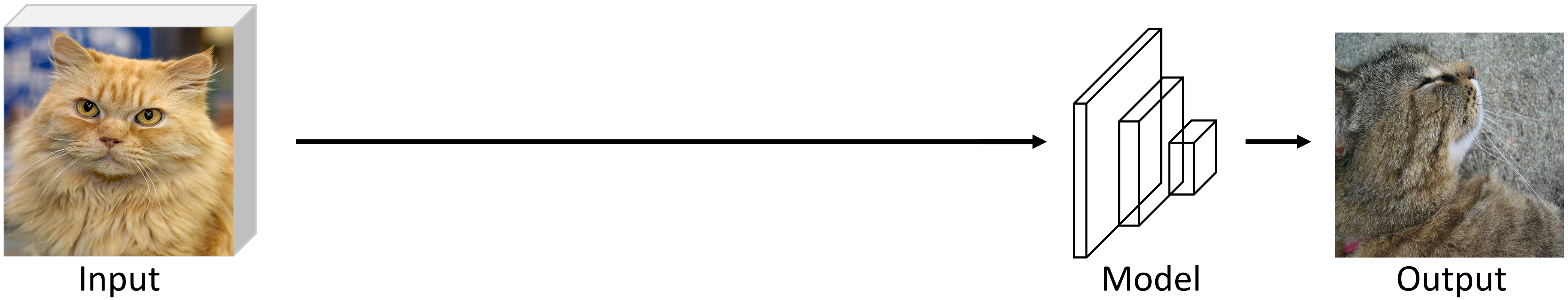
- Utility — maintain accuracy in benign data
- Security — against data and model poisoning attacks
- Privacy
- Fairness
- Communication
- … …

$$w^* \in \arg\min_{w} G\left(F_1(w), \ldots, F_K(w)\right)$$

Slides adapted from [Ditto]

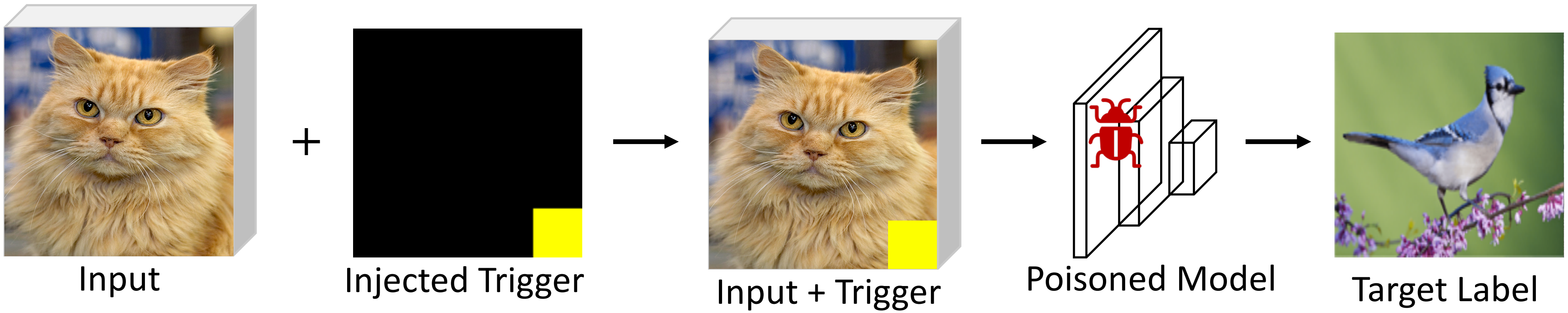
Backdoor Attack

- Data poisoning attack
  - Manipulate a subset of <u>training</u> data
- A backdoored image-classification model misclassifies on any <u>test</u> data with certain features (i.e., a trigger) to an attacker-chosen class (i.e., target label)



Backdoor Attack

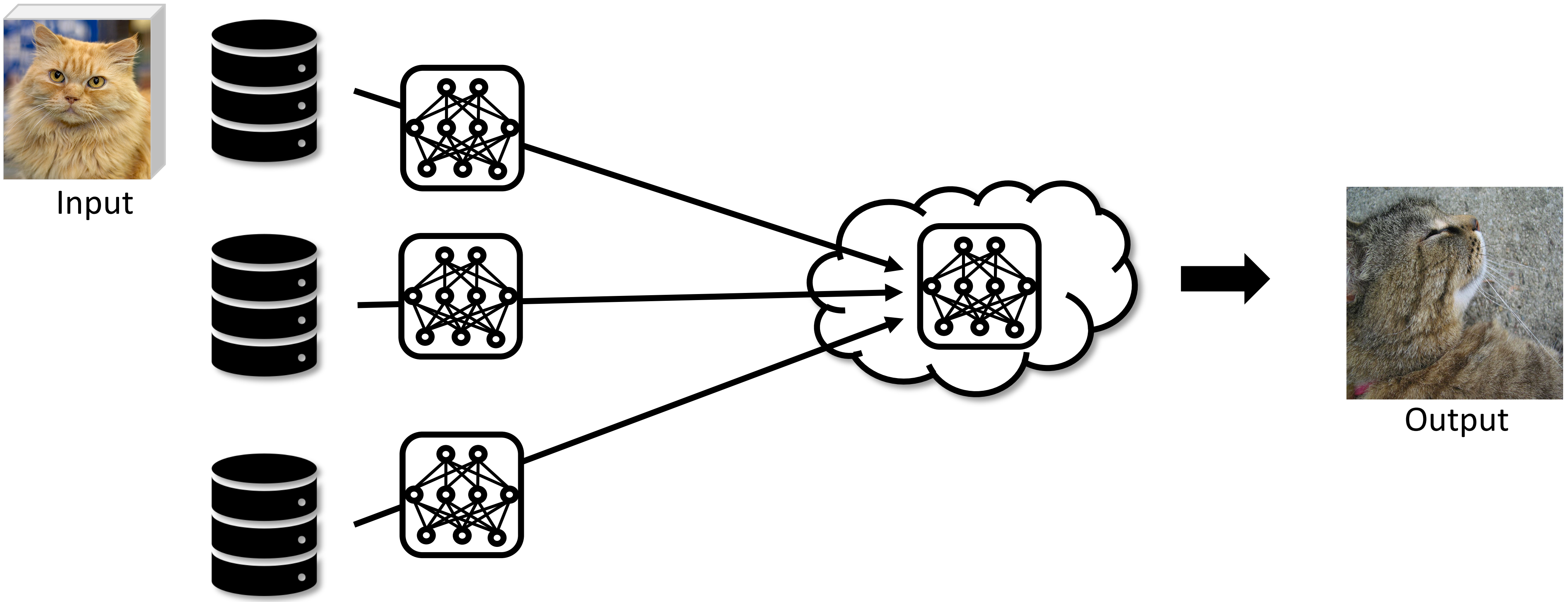
- Data poisoning attack
  - Manipulate a subset of training data
- A backdoored image-classification model misclassifies on any test data with certain features (i.e., a trigger) to an attacker-chosen class (i.e., target label)



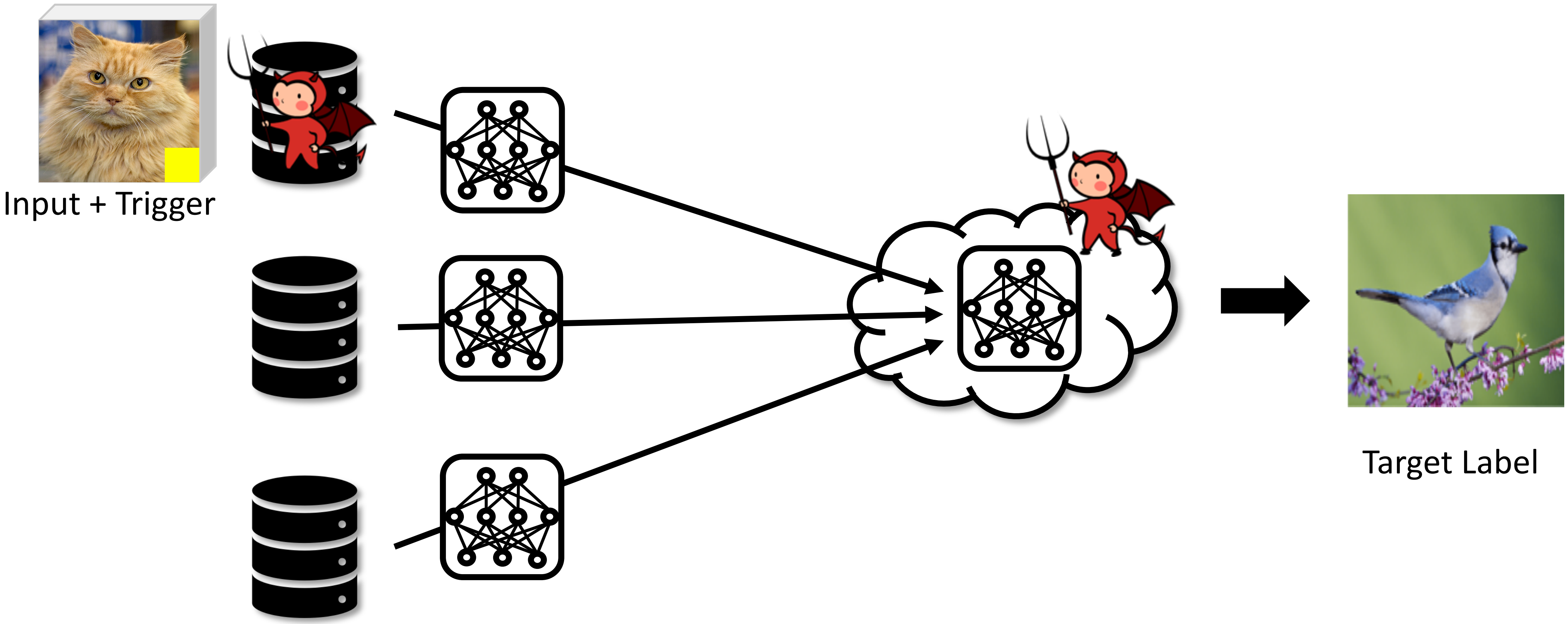
Backdoor Attack and Defense in FL

- Attack Goal: Malicious local clients perform backdoor attack locally, controls how the global model performs on an attacker-chosen backdoor subtask and new global model maintains accuracy.
- Defense Goal: Mitigates attack success rate on backdoor data and maintains accuracy on benign data.

Backdoor Attack and Defense in FL



Backdoor Attack and Defense in FL



FL Backdoor Attack Settings

- Single-shot backdoor attack [1]
  - Every adversary only participates in one single round, while there can be multiple attackers
  - Simpler attack
- Continuous backdoor attack [2]
  - The attackers are selected in every round and continuously participate in the FL training from the beginning to the end.
  - Stronger and stealthier, and harder to defend

[1]. Bagdasaryan, Eugene, et al. "How to backdoor federated learning."AISTATS, 2020.
[2]. Xie, Chulin, et al. "Dba: Distributed backdoor attacks against federated learning." ICLR 2020

Existing Defenses

- Robust aggregation
  - Detects abnormal gradient updates
  - Rejects malicious weights
- Certified defense
  - Provides robustness certification in the presence of backdoors with limited magnitude
  - Simplify settings, for example, only works in i.i.d. data

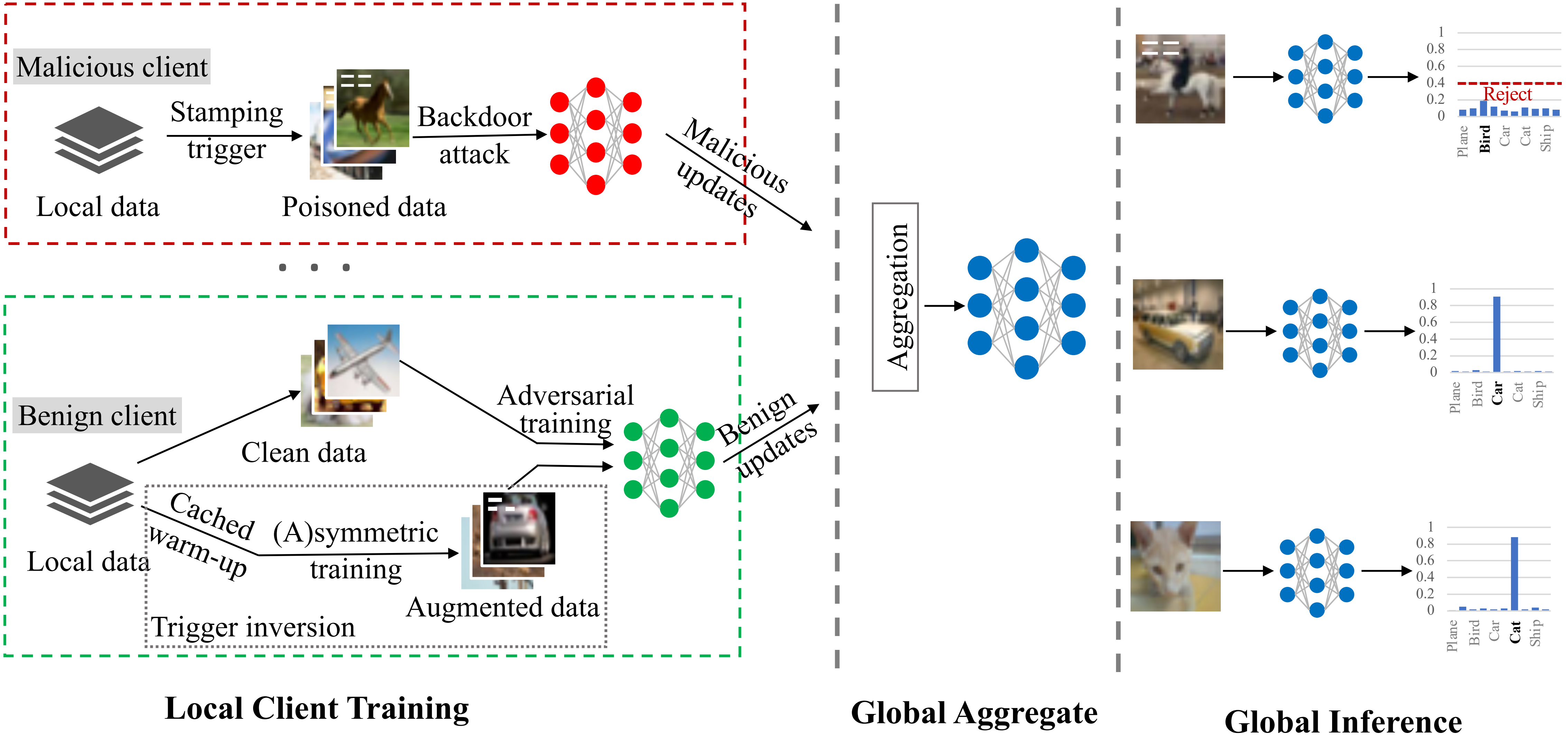
Motivation

- A majority of existing defenses only work in the single-shot attack setting and fall short in the continuous attack setting.
- Possible reasons:
  - Continuous backdoor attacks are stealthier, abnormal detection based methods are hard to detect and reject malicious weights
  - Continuous backdoor attacks are more aggressive
  - Unrealistic assumptions of i.i.d. data

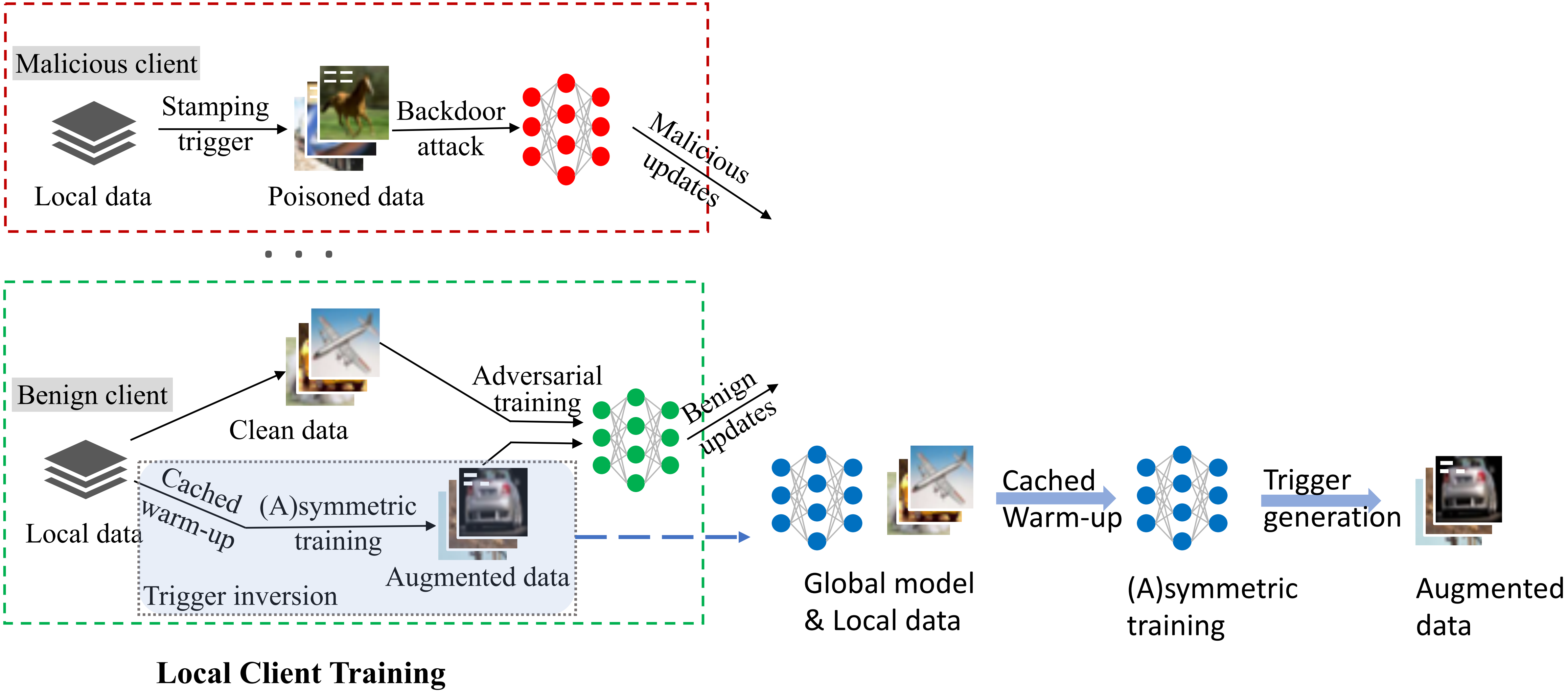
Threat Model

- Malicious local clients
  - Backdoor data injection
  - Full control of their local model training
- Benign clients
  - Non-i.i.d. data
  - Have no knowledge about ground truth trigger
- Global server
  - Does not distinguish weights from trusted or untrusted clients
  - Assume no local data

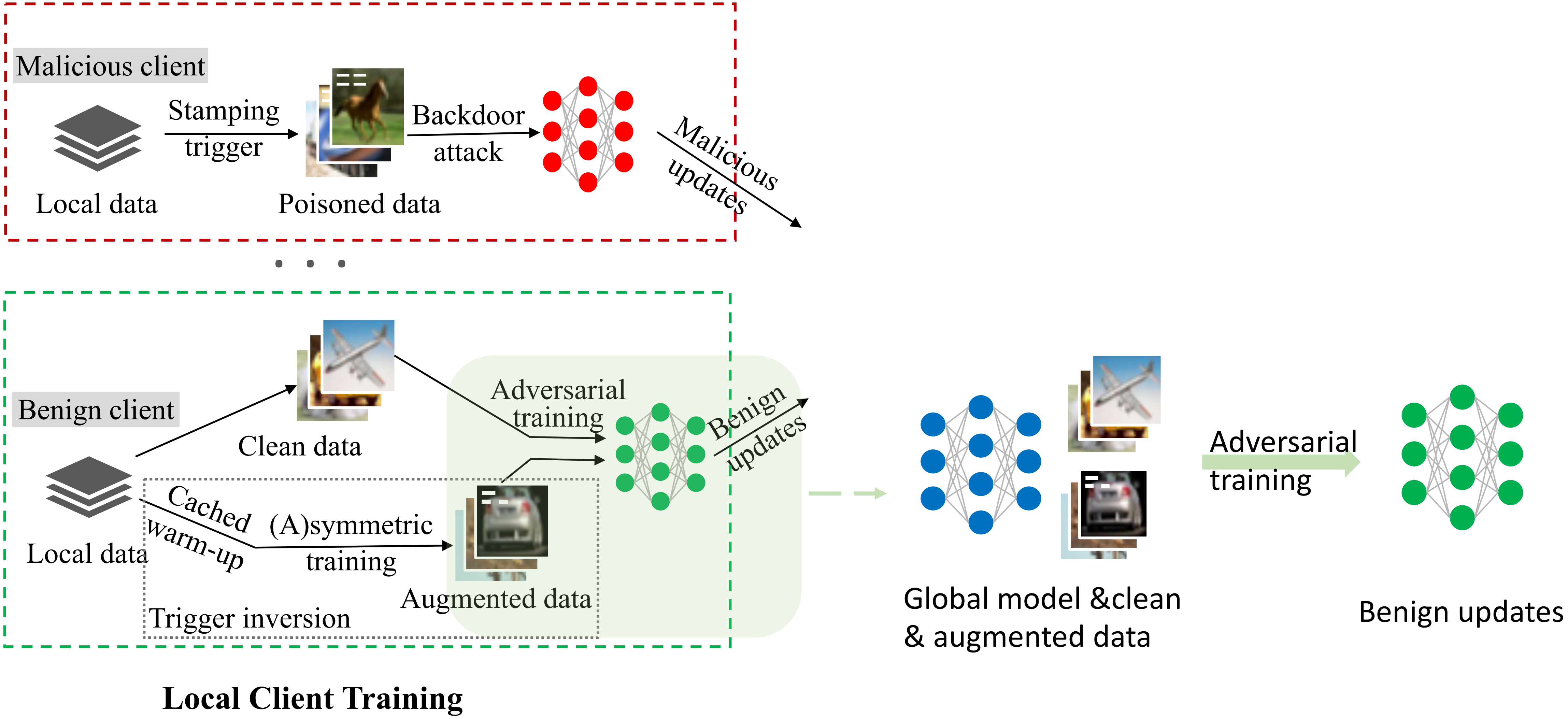
Approach Overview



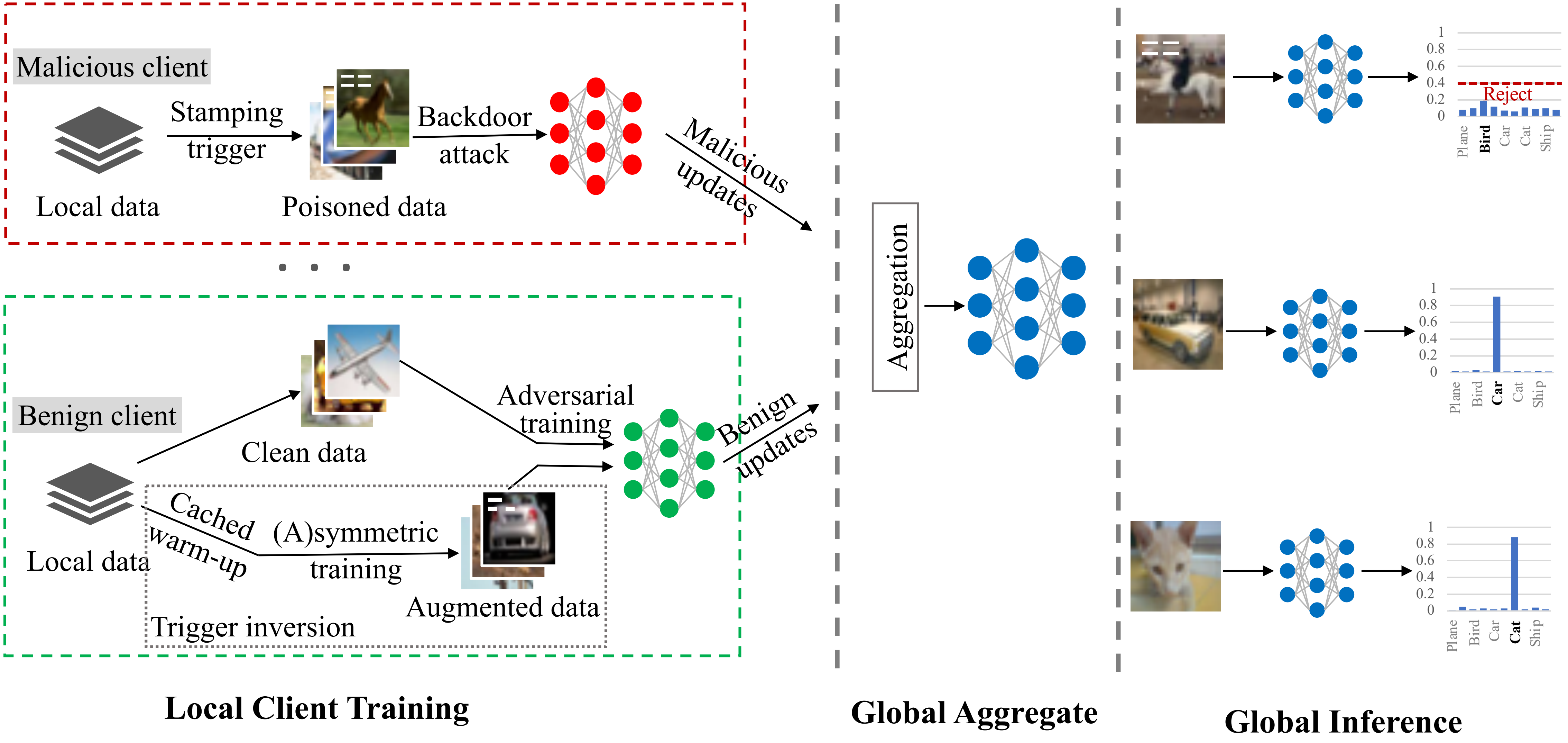
FLIP Algorithm (Trigger Inversion)



FLIP Algorithm (Model Hardening)



FLIP Algorithm (Low-confidence Sample Rejection)



Insights

Remains an open problem how benign local clients trigger inversion quality and model hardening will influence the malicious attack success rate and global model accuracy?

Theoretical Analysis

- **Theorem 1**: Developing upper and lower bounds quantifying the cross-entropy loss changes on backdoored and clean data.
- **Theorem 2**: Showing a sufficient condition on the quality of trigger recovery such that the proposed defense is provably effective.
- **Corollary 1**: Following previous theorems, we show that inference with confidence thresholding on models trained with our proposed defense can provably reduce the backdoor attack success rate while maintaining similar accuracy on clean data.

Theoretical Analysis

Theorem 1 (Bounds on Loss Changes) *Let $\mathcal{L}'_g$ denote the global model loss with defense, $\mathcal{L}_g$ without defense, let $\Delta W = W' - W$ denote the weight differences with and without defense. The loss difference with and without defense can be upper and lower bounded by*

$$\min_t(\mathbf{x}\Delta W)_t - \sum_{i=1}^{I} q_i(\mathbf{x}\Delta W)_i \leq \mathcal{L}'_g - \mathcal{L}_g \leq \max_t(\mathbf{x}\Delta W)_t - \sum_{i=1}^{I} q_i(\mathbf{x}\Delta W)_i$$

Lower bound

On backdoor data, which indicates how much the ASR at least will be reduced.

Upper bound

On benign data, which indicates how much the ACC will at most be maintained.

Experiments

Table 1: Single-shot attack evaluation

Baselines	MNIST		F-MNIST		CIFAR-10	
	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	97.55	80.12	81.01	96.72	77.52	80.46
Krum	97.50	0.35	79.49	10.79	77.00	9.51
Bulyan Krum	97.76	0.39	81.45	6.42	79.65	5.77
RFA	97.93	0.39	81.82	4.39	79.54	6.13
Trimmed Mean	97.81	0.38	81.81	5.40	79.95	5.81
Buly-Trim-M	97.02	90.75	79.84	99.38	66.69	84.05
FoolsGold	97.51	0.39	80.59	5.64	78.67	3.70
Median	97.76	0.37	81.76	5.97	64.31	2.39
FLTrust	97.26	0.48	79.92	7.69	**72.44**	**2.18**
FLIP	**96.05**	**0.13**	**78.20**	**3.16**	73.41	7.83

Table 2: Continuous attack evaluation

Baselines	MNIST		F-MNIST		CIFAR-10	
	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	98.71	100.00	80.35	99.99	77.83	84.73
Krum	**97.59**	**0.14**	73.18	20.03	40.29	18.79
Bulyan Krum	98.15	94.01	82.17	99.46	68.61	97.31
RFA	98.54	100.00	85.69	100.00	79.39	63.10
Trimmed Mean	98.52	100.00	84.59	99.99	75.18	91.84
Buly-Trim-M	98.80	100.00	76.18	99.93	71.91	68.83
FoolsGold	97.91	99.99	80.58	99.98	74.57	78.30
Median	98.14	66.01	84.07	99.34	57.01	69.99
FLTrust	91.96	20.60	74.63	35.36	74.85	68.70
FLIP	96.62	1.93	**72.99**	**17.65**	**71.28**	**22.90**

Take Away

- Propose a provable defense framework FLIP that can provide a sufficient condition on the quality of trigger recovery, such that the proposed defense is provably effective in mitigating backdoor attacks
- FLIP significantly outperforms prior work on the SOTA continuous FL backdoor attack and resilient to adaptive attacks.
- FLIP is general and can be instantiated with different trigger inversion techniques.

Related Works

[1]. Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 30, 2017.

[2]. El Mahdi El Mhamdi, Rachid Guerraoui, and S´ebastien Rouault. The hidden vulnerability of distributed learning in Byzantium. In Jennifer Dy and Andreas Krause (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 3521–3530. PMLR, 10–15 Jul 2018.

[3]. Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. IEEE Transactions on Signal Processing, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

[4]. Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The Limitations of Federated Learning in Sybil Settings. In Symposium on Research in Attacks, Intrusion, and Defenses, RAID, 2020.

[5]. Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In International Conference on Machine Learning, pp. 5650–5659. PMLR, 2018.

[6]. Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. arXiv preprint arXiv:2012.13995, 2020.

[7]. Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics, pp. 2938–2948. PMLR, 2020.

[8]. Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jyyong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. Advances in Neural Information Processing Systems, 33, 2020a.

[9]. Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In International Conference on Learning Representations, 2020.

FLIP: A Provable Defense Framework for Backdoor Mitigation in Federated Learning

- This work is supported, in part by IARPA TrojAI W911NF-19-S-0012, NSF 1901242 and 1910300, ONR N000141712045, N000141410468 and N000141712947.

- Full paper and code: https://kaiyuanzhang.com/

Thank you for listening!