# Sniper Backdoor

## Single Client Targeted Backdoor Attack in Federated Learning

**Gorka Abad** [1,2] Servio Paguada [1,2] Oğuzhan Ersoy [1] Stjepan Picek [1] Víctor Julio Ramírez-Durán [2] Aitor Urbieta [2]

February 13, 2023

[1]Radboud University, the Netherlands

[2]Ikerlan Technology Research Centre, Spain

# Outline

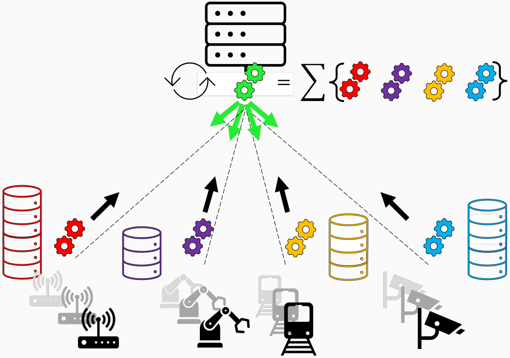
# Table of Contents

# Introduction: Backdoor Attacks

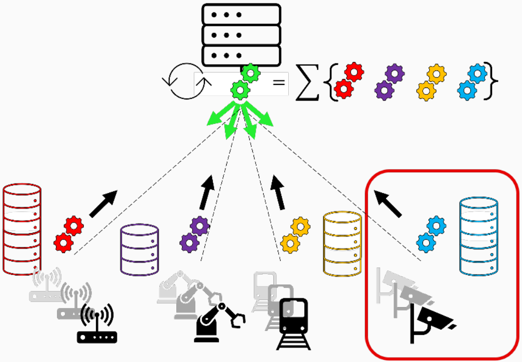
- Train the model with tons of data.
- Then we evaluate its performance with a holdout dataset.
- But what happens with untested data?

- Training time attack
- Inject a trigger on some (small) number of samples
- Aim to misclassify samples containing the trigger while achieving great performance on clean data
- We can create them adding a *trigger* [1]
- Trigger: 
- Label: "Speed Limit"

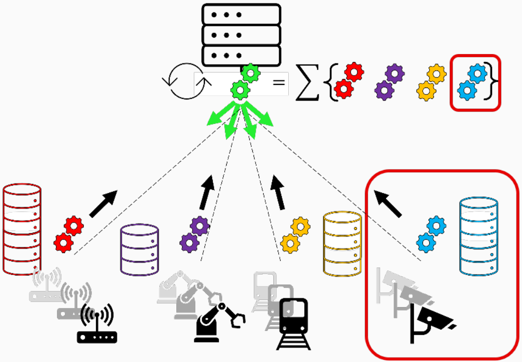
- Privacy driven
- Datasets remain local
- Data can be heterogeneous
- Independent and identically distributed data (IID)
- The performance of Non-IID is drastically reduced [2]
- Using warming up could help [2]



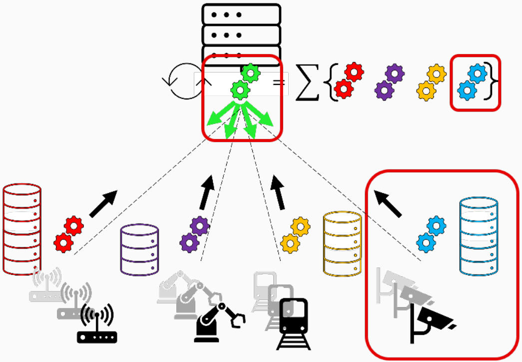
- Clients inject the backdoor locally [3]–[5]
- After aggregation **every** client receives a backdoored model
- Some other attacks consider more than a single attacker [6]



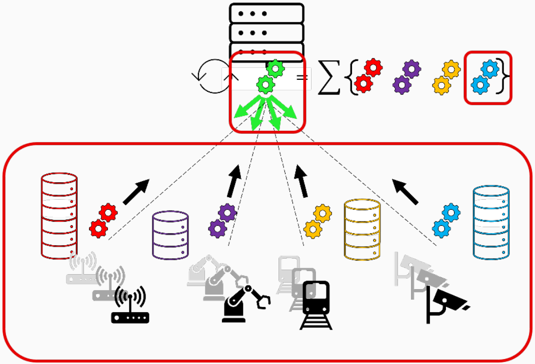
- Clients inject the backdoor locally [3]–[5]
- After aggregation **every** client receives a backdoored model
- Some other attacks consider more than a single attacker [6]



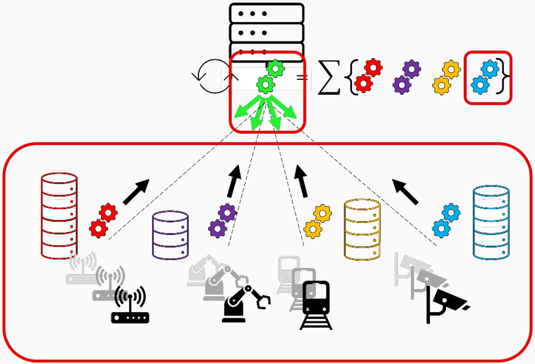
- Clients inject the backdoor locally [3]–[5]
- After aggregation **every** client receives a backdoored model
- Some other attacks consider more than a single attacker [6]



- Clients inject the backdoor locally [3]–[5]
- After aggregation **every** client receives a backdoored model
- Some other attacks consider more than a single attacker [6]



- Clients inject the backdoor locally [3]–[5]
- After aggregation **every** client receives a backdoored model
- Some other attacks consider more than a single attacker [6]



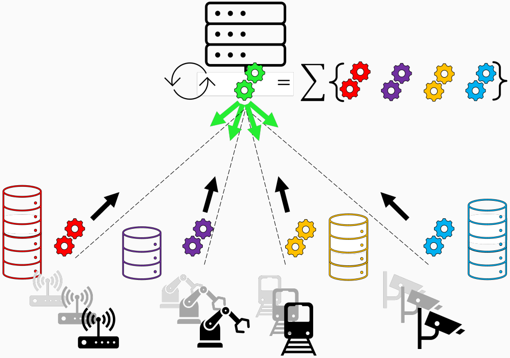
# Introduction: Inference Attacks in FL

- Extract information from clients
- For example, model inversion attacks reconstruct samples used during training [7]
- In FL even from a specific client [8], [9]

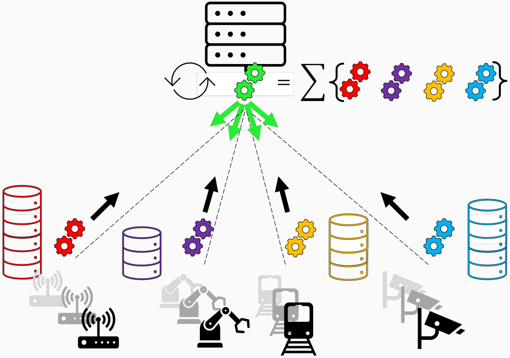
Figure from [7]

# Table of Contents

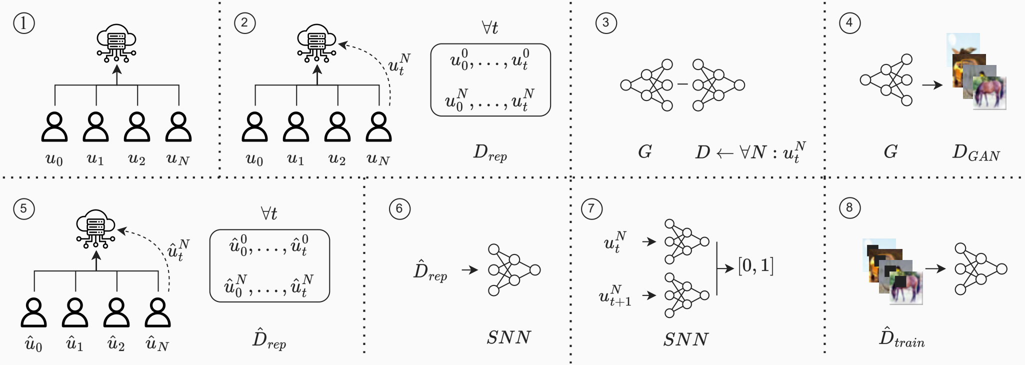
- Some defenses rely on the assumption that all the clients are being compromised
- If more than 50% are compromised, then the model the networks agree that it has been compromised
- *Could we gain knowledge using inference and then use it for a backdoor attack?*
- *"Is it possible to launch a backdoor attack, where only targeted (victim) clients get a backdoored model whereas the remaining (non-victim) clients get a clean model?"*



- The server is malicious
- Clients do not trust the server and they anonymize their model uploads
- The attacker has to identify and only send a malicious model to the target client.
- The rest of clients should not be affected



**1, 2** During the training of FL the attacker keeps track of the submitted anonymous models



**3, 4** The attacker launches a GAN-based model inversion attack

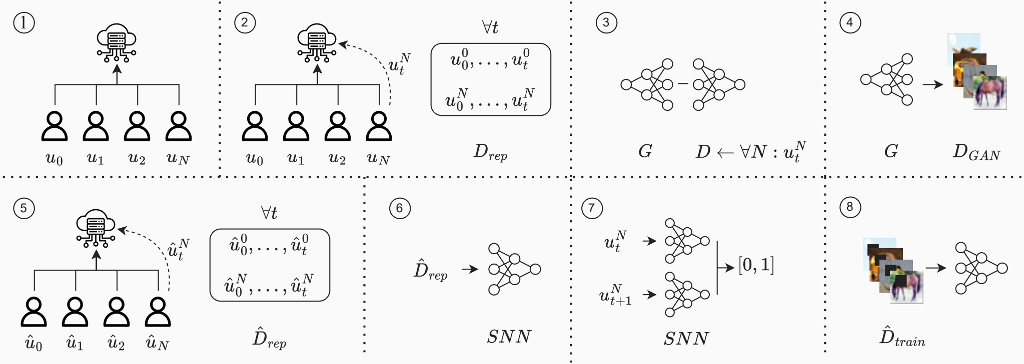
Select the model from a certain epoch

Model between clients will be different at early epochs while more similar close to convergence

The discriminator is replaced by the model

Thus, the generated data is similar to the clients'

The attacker then has a dataset of clients like data



# Sniper Backdoor: Attack Overview

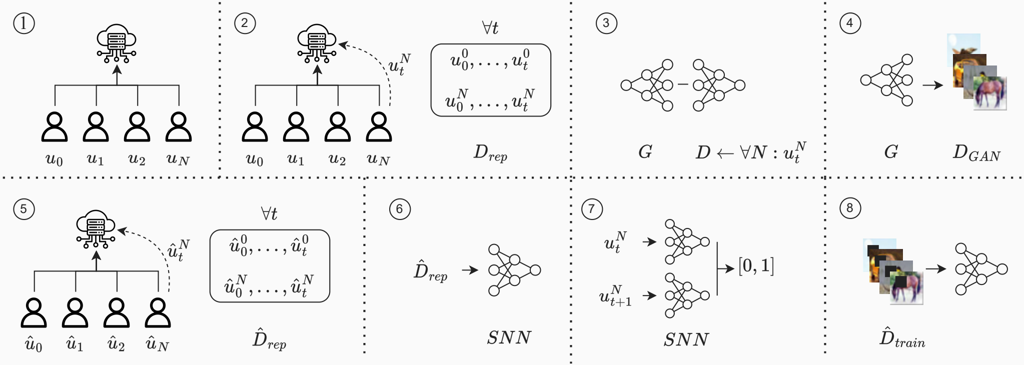
**5, 6, 7** Clients submit their models anonymously

Since the attacker knows the data used for training, he/she can target the client precisely

Shadow training with the GAN-generated dataset

Keep a record of the shadow models

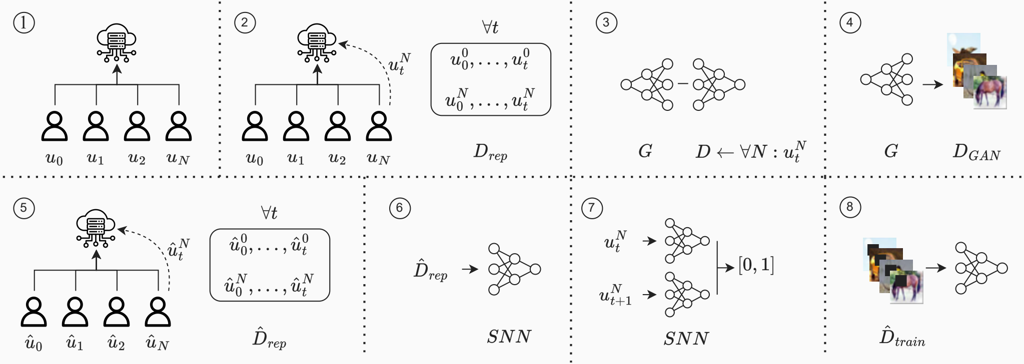
Real and shadow models are similar



**8** Having the client identified

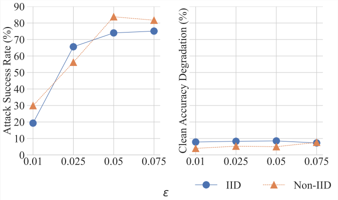
The attacker can backdoor a model and submit it to the target client

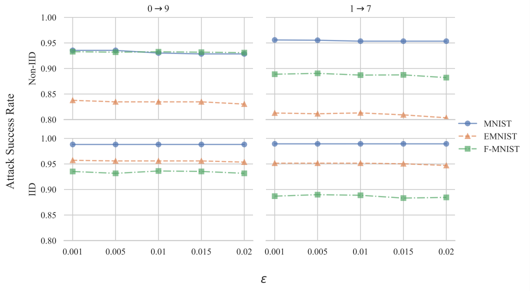
The rest of the clients receive the clean model



# Table of Contents

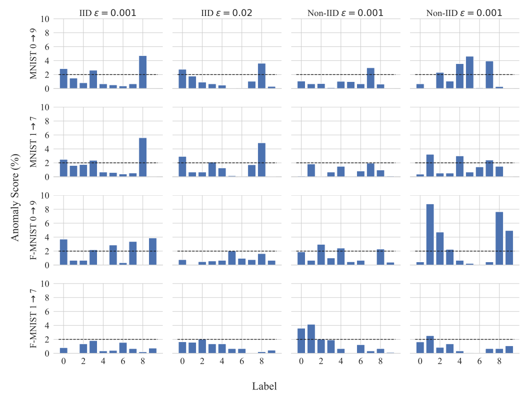
- Is it possible to launch a backdoor attack, where only targeted (victim) clients get a backdoored model, whereas the remaining (non-victim) clients get a clean model



# Final Remarks: Results on Defenses

- Neural Cleanse or ABS cannot handle nor source targeted backdoors nor dynamic backdoors
- FL specific countermeasures as Krum, FoolsGold, Baffle, CRLF do not hold



## Table of Contents

# Conclusions: Results on Defenses

(1) State-of-the-art defenses do not consider that a single or a subset of clients is being attacked

(2) If the countermeasure is applied by the client itself, the attacker could still adapt the attack

(3) However, relaying on a TTP to check the models could be a possible countermeasure

(4) Differential privacy could also harden the model inversion attack and thus the consequent attack's phases

(5) As future work, could we target a single client from a malicious client?

Thanks for your attention, any questions?

[1] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, et al. "Badnets: Evaluating backdooring attacks on deep neural networks". In: IEEE Access 7 (2019), pp. 47230–47244.

[2] Yue Zhao, Meng Li, Liangzhen Lai, et al. "Federated learning with non-iid data". In: arXiv preprint arXiv:1806.00582 (2018).

[3] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, et al. "Can you really backdoor federated learning?" In: arXiv preprint arXiv:1911.07963 (2019).

[4] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, et al. "Attack of the tails: Yes, you really can backdoor federated learning". In: Advances in Neural Information Processing Systems 33 (2020), pp. 16070–16084.

[5] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, et al. "How to backdoor federated learning". In: International Conference on Artificial Intelligence and Statistics. PMLR. 2020, pp. 2938–2948.

[6] Chulin Xie, Keli Huang, Pin-Yu Chen, et al. "Dba: Distributed backdoor attacks against federated learning". In: International conference on learning representations. 2020.

[7] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures". In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015, pp. 1322–1333.

[8] Jiale Chen, Jiale Zhang, Yanchao Zhao, et al. “Beyond model-level membership privacy leakage: an adversarial approach in federated learning”. In: 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE. 2020, pp. 1–9.

[9] Mengkai Song, Zhibo Wang, Zhifei Zhang, et al. “Analyzing user-level privacy attack against federated learning”. In: IEEE Journal on Selected Areas in Communications 38.10 (2020), pp. 2430–2444.