

# Sniper Backdoor

## Single Client Targeted Backdoor Attack in Federated Learning

---

**Gorka Abad** <sup>1,2</sup>   Servio Paguada <sup>1,2</sup>   Oğuzhan Ersoy <sup>1</sup>   Stjepan Picek <sup>1</sup>   Víctor Julio Ramírez-Durán <sup>2</sup>   Aitor Urbieto <sup>2</sup>

December 23, 2022

<sup>1</sup>Radboud University, the Netherlands

<sup>2</sup>Ikerlan Technology Research Centre, Spain

## Introduction

- Machine Learning

- Federated Learning

- Deep Learning

- Backdoor attacks 101

- Backdoor Attacks in FL

## Sniper Backdoor

- Motivation

- Challenges

- Attack Phases

- Attack Overview

## Defenses

## Final Remarks

# Table of Contents

## Introduction

Machine Learning

Federated Learning

Deep Learning

Backdoor attacks 101

Backdoor Attacks in FL

## Sniper Backdoor

Motivation

Challenges

Attack Phases

Attack Overview

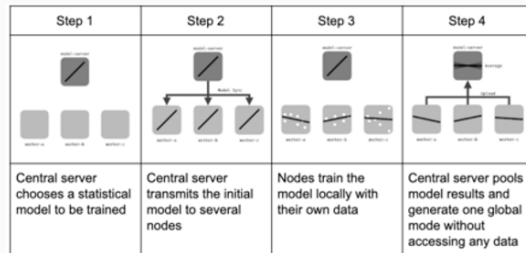
## Defenses

## Final Remarks

- ▶ Centralized data
- ▶ More data = better
- ▶ Privacy issues

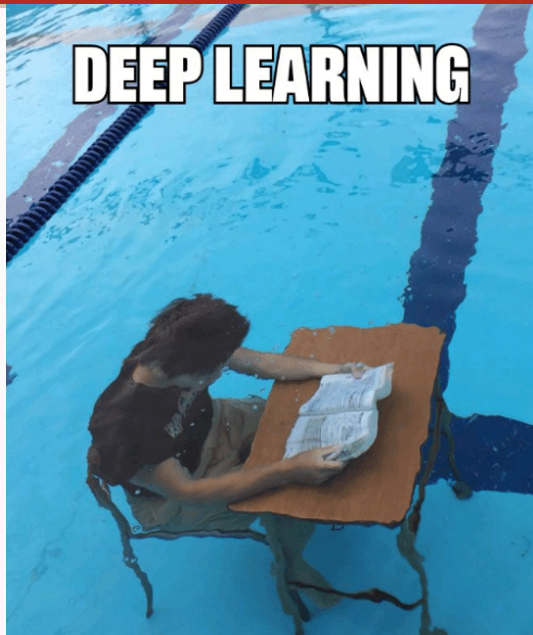


- Privacy driven<sup>1</sup>
- Data is private for each user
- Data can be either Independent and Identically Distributed (IID) or Non-IID

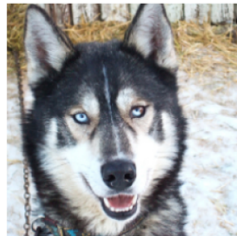


<sup>1</sup>Attacks have shown that FL's privacy is broken [1]

- ▶ State-of-the-art in many ML tasks
- ▶ Our work focuses on the image domain
- ▶ Convolutional layers
- ▶ More parameters = More complexity




- ▶ How do we test DL models?
- ▶ We use test sets
- ▶ If the model behaves correctly in the test set, we say the model is correct
- ▶ Some works try to understand why [2]



(a) Husky classified as wolf



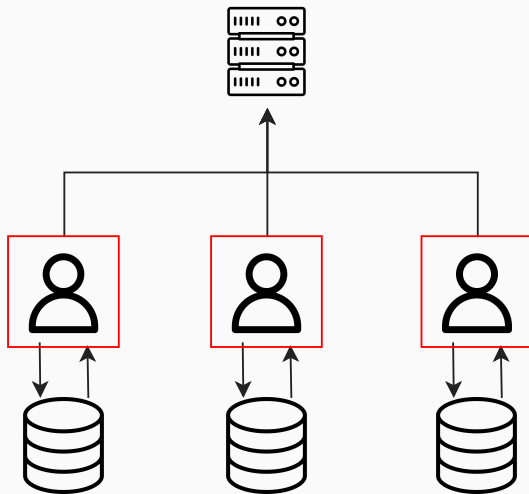
(b) Explanation

- ▶ What happens with untested samples?
- ▶ We can create them adding a *trigger* [3]
- ▶ Trigger: 
- ▶ Label: "Speed Limit"





- (1) Can we backdoor FL? [4]
- (2) Yes, we can... [5]
- (3) But, how? [6]
- (4) Use a scaling factor  $\lambda$  for scaling the models
- (5) Every client receives a backdoored model



# Table of Contents

## Introduction

Machine Learning

Federated Learning

Deep Learning

Backdoor attacks 101

Backdoor Attacks in FL

## Sniper Backdoor

Motivation

Challenges

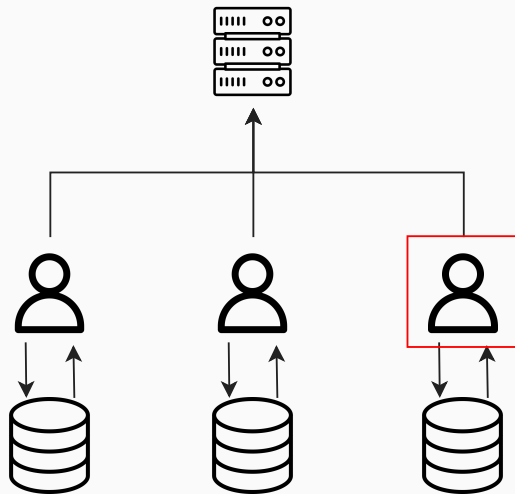
Attack Phases

Attack Overview

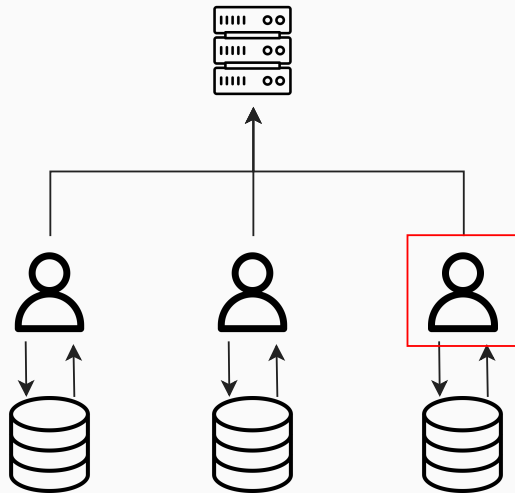
## Defenses

## Final Remarks

- *“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?”*

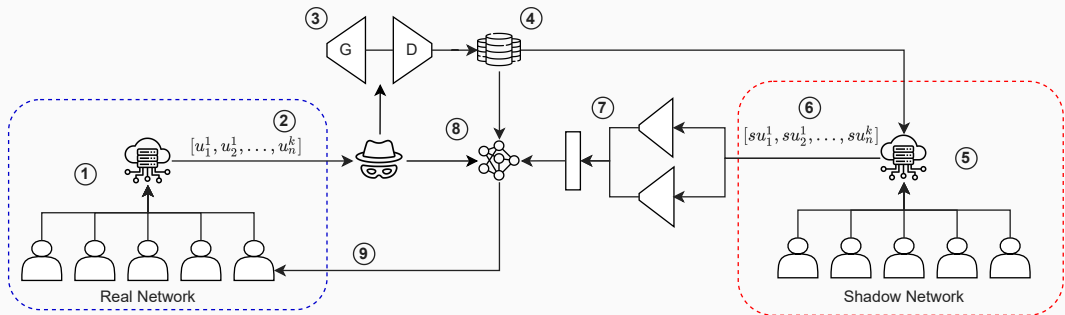


- ▶ The server is malicious
- ▶ We have no access to the datasets nor the training procedure
- ▶ Clients are anonymous

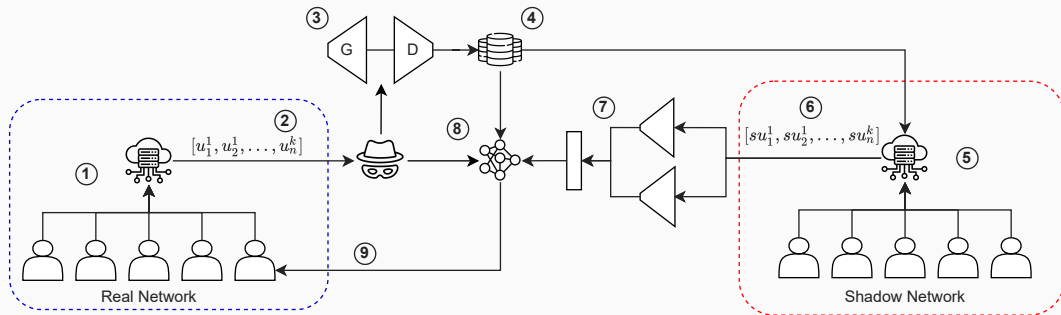


# Sniper Backdoor: Attack Phases

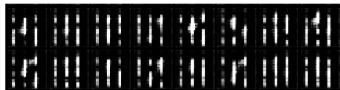
- Create the backdoor model
  - Get a dataset
- Identify the victim client



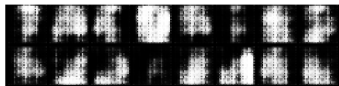
1, 2 Keep a record of anonymous models



## 3, 4 Creating synthetic data



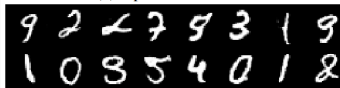
(a) Epoch 1 MNIST.



(b) Epoch 1 F-MNIST.



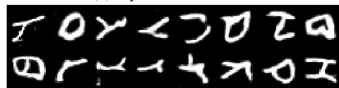
(c) Epoch 1 EMNIST.



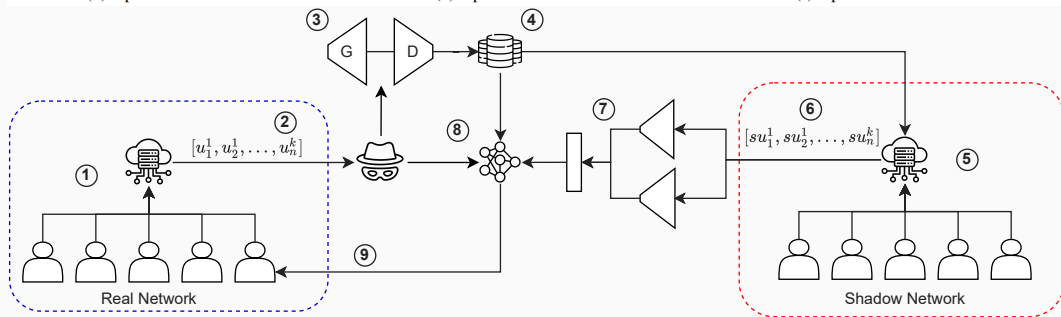
(d) Epoch 1000 MNIST.



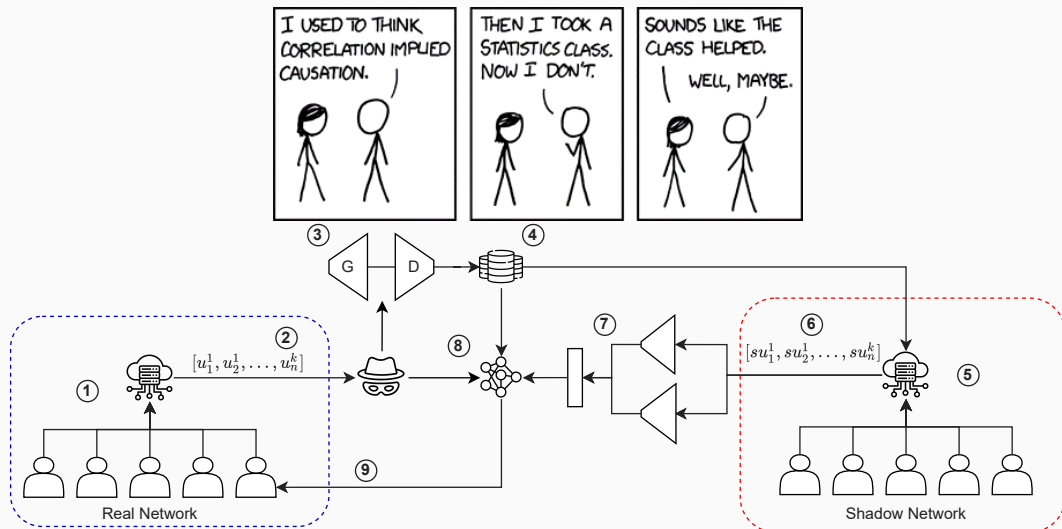
(e) Epoch 1000 F-MNIST.



(f) Epoch 1000 EMNIST.

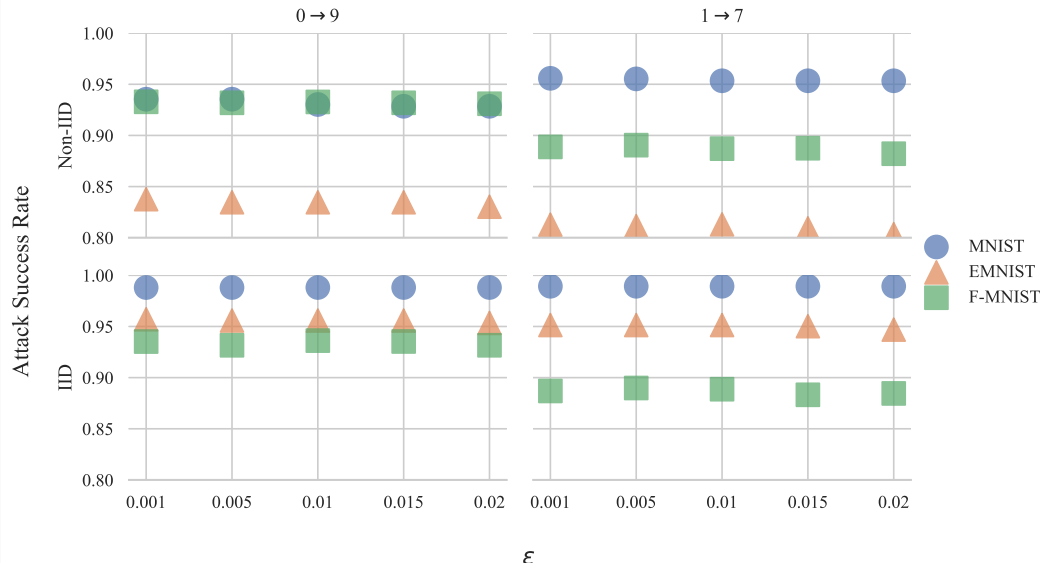


## 5, 6, 7 Identifying the victim





## 8, 9 Inject the backdoor



# Table of Contents

## Introduction

Machine Learning

Federated Learning

Deep Learning

Backdoor attacks 101

Backdoor Attacks in FL

## Sniper Backdoor

Motivation

Challenges

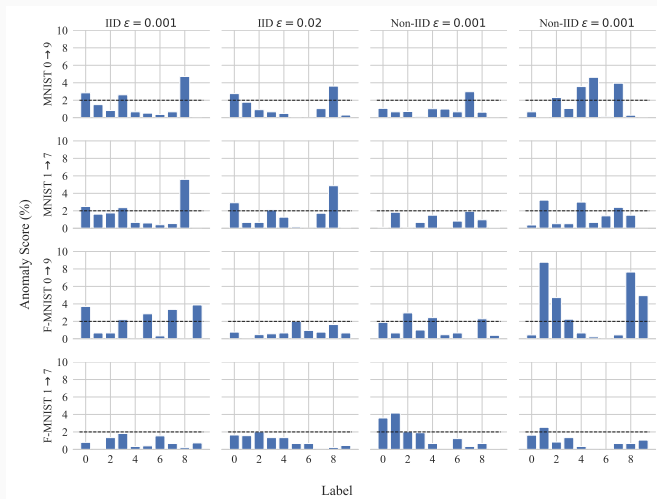
Attack Phases

Attack Overview

## Defenses

## Final Remarks

## Neural Cleanse [7]



# Table of Contents

## Introduction

Machine Learning

Federated Learning

Deep Learning

Backdoor attacks 101

Backdoor Attacks in FL

## Sniper Backdoor

Motivation

Challenges

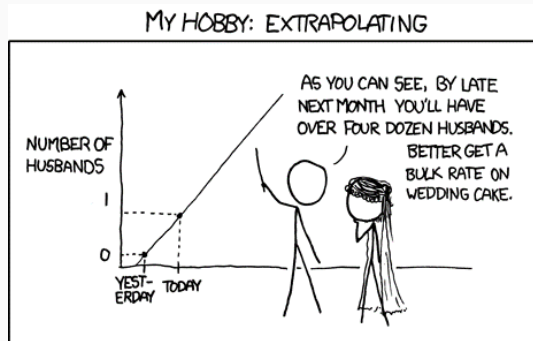
Attack Phases

Attack Overview

## Defenses

## Final Remarks

- (1) Bypasses “all” the state-of-the-art defenses
- (2) Most of the state-of-the-art backdoor defenses in FL do not apply
- (3) We require new defense mechanisms
- (4) More exhaustive research has to be done for this new threat
- (5) What about a client being an attacker?



Thanks for your attention, any questions?

- [1] Franziska Boenisch, Adam Dziedzic, Roei Schuster, et al. “When the curious abandon honesty: Federated learning is not private”. In: [arXiv preprint arXiv:2112.02918](#) (2021).
- [2] Saumitra Mishra, Bob L Sturm, and Simon Dixon. “Local interpretable model-agnostic explanations for music content analysis.”. In: [ISMIR](#). Vol. 53. 2017, pp. 537–543.
- [3] 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.
- [4] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, et al. “Can you really backdoor federated learning?” In: [arXiv preprint arXiv:1911.07963](#) (2019).
- [5] 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.
- [6] 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.
- [7] Bolun Wang, Yuanshun Yao, Shawn Shan, et al. “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks”. In: [2019 IEEE Symposium on Security and Privacy \(SP\)](#). IEEE. 2019, pp. 707–723.