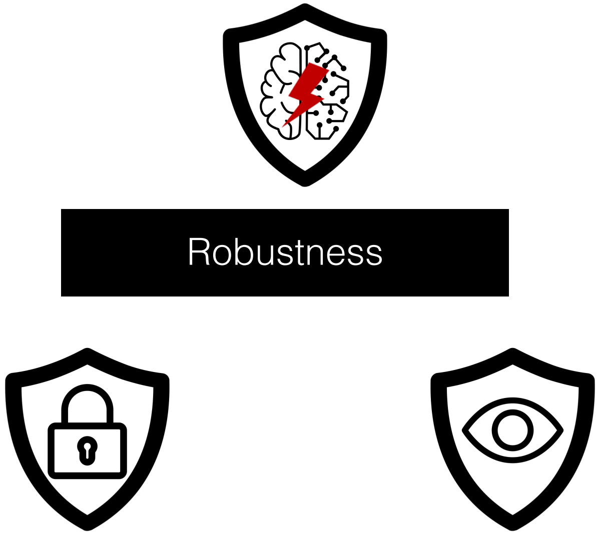
Security and Robustness of Collaborative Learning Systems

Anwar Hithnawi

ETH zürich

Collaborative Learning

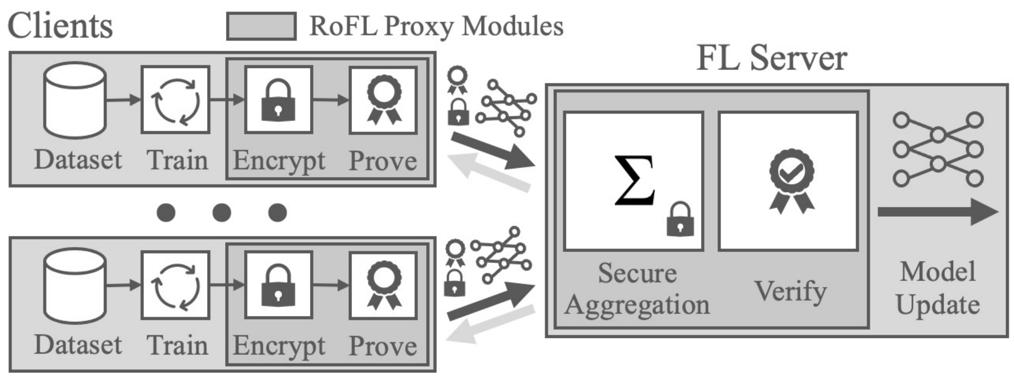


Robustness

Security

Privacy

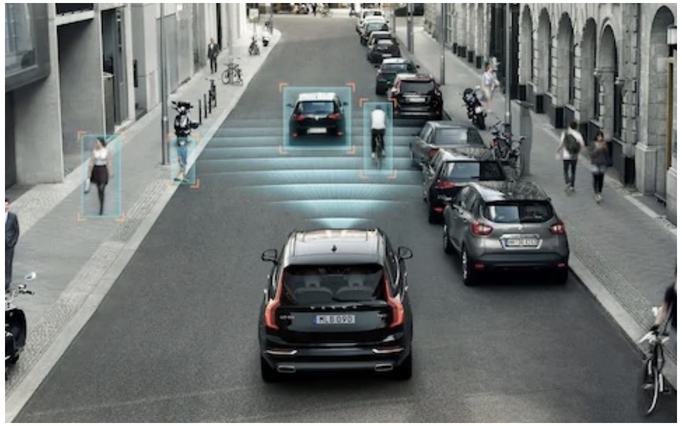
RoFL



RoFL: Attestable Robustness for Secure FL. arXiv:2107.03311
L. Burkhalter*, H. Lycklama*, A. Viand, N. Küchler, A. Hithnawi

Analysis Code: https://github.com/pps-lab/fl-analysis
RoFL Code: https://github.com/pps-lab/rofl-project-code

Data Driven World



Autonomous Driving

Object Classification

Health Care

AlphaGo

Data

large, diverse data ⟶ broad generalization

World

Health Care

AlphaGo

Solving tasks where data is **accessible**…

… however, many important tasks we care about ...

Tasks

WikiText-103
MNIST
ImageNet
WMT
CIFAR
GPT-3

Public Data

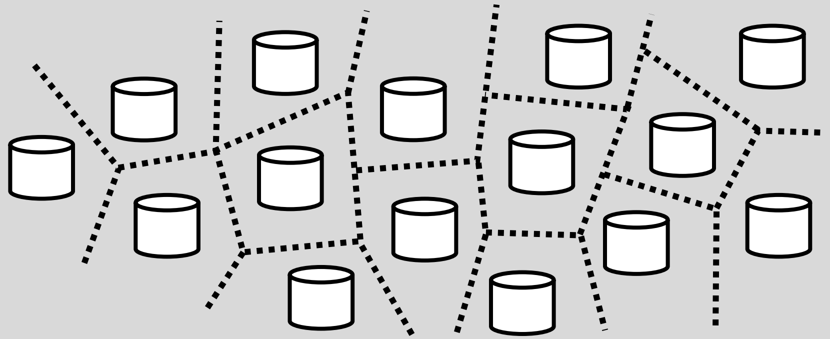
Crowdsourced Data

For example: web, books, articles, science, TV, corpus, audiobooks, …

Inaccessible

Health – Cancer, Alzheimer, Dementia, Depression
Finance – Economic growth, Market predictions
Government – Education, Taxes, Immigration, Income
Personal Data – Text Messages, Emails, Photos

→ EU Data Governance Act (**DGA**) effective from 2023
facilitate the reuse of protected public-sector data

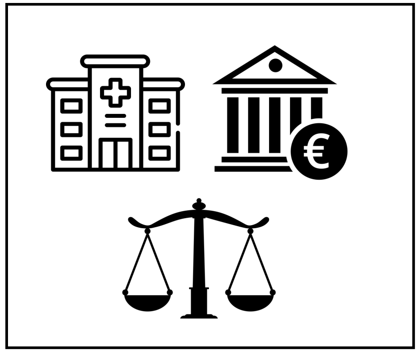


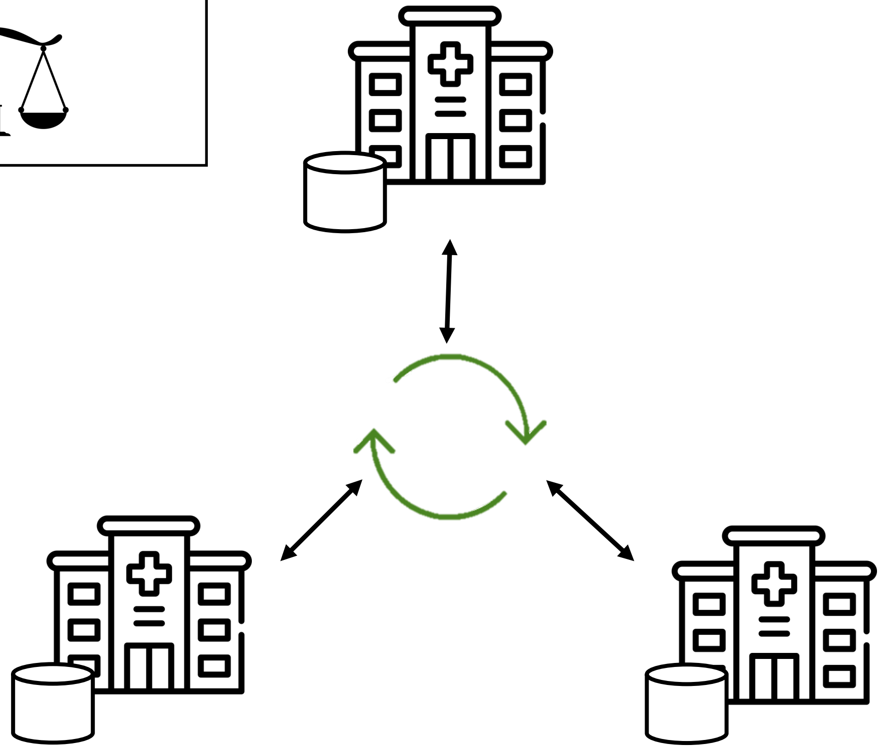
Legal Frameworks and Technologies to facilitate privacy preserving access

Data Silos
- Privacy Laws
- Competition

Collaborative Learning

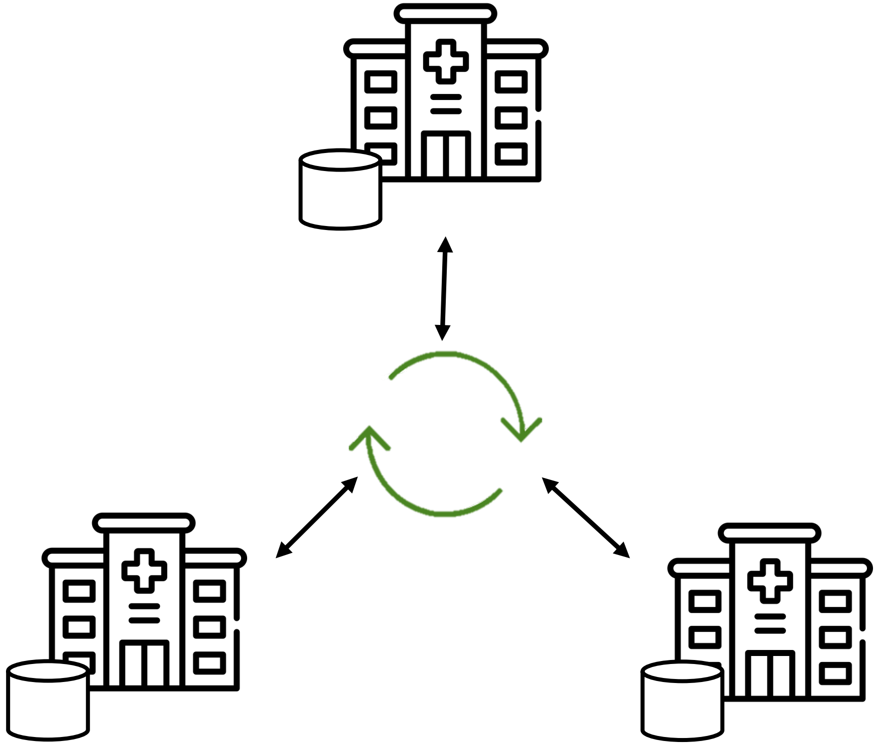
Collaborative Learning



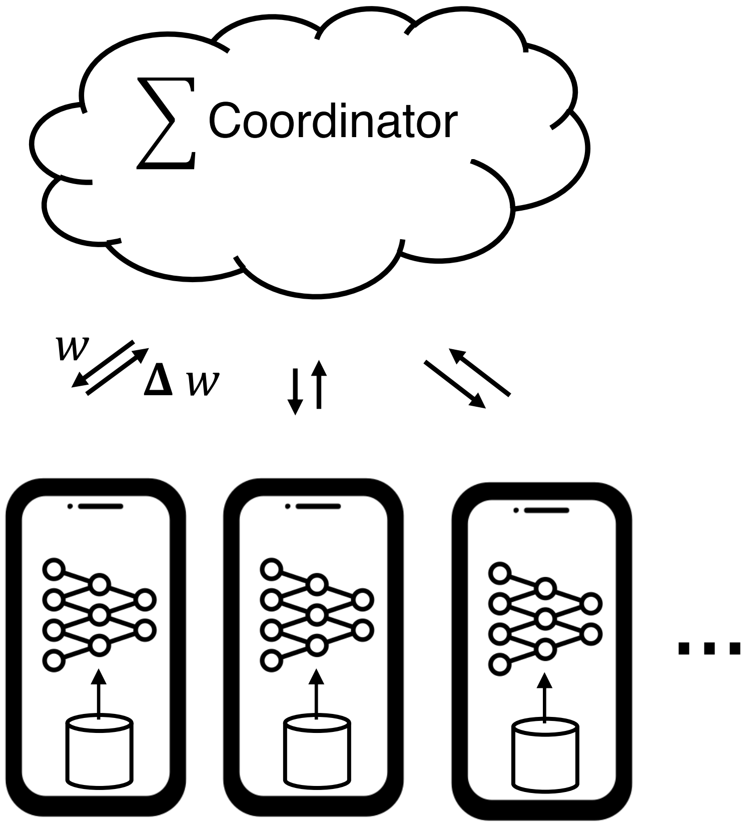


Decentralized Learning

Collaborative Learning



Decentralized Learning

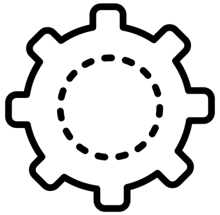


Federated Learning

Security and Privacy of Machine Learning

Data | Train/Model | Deploy/Serve

$$\mathrm{X} \longrightarrow \theta \longrightarrow$$



$\leftarrow x$

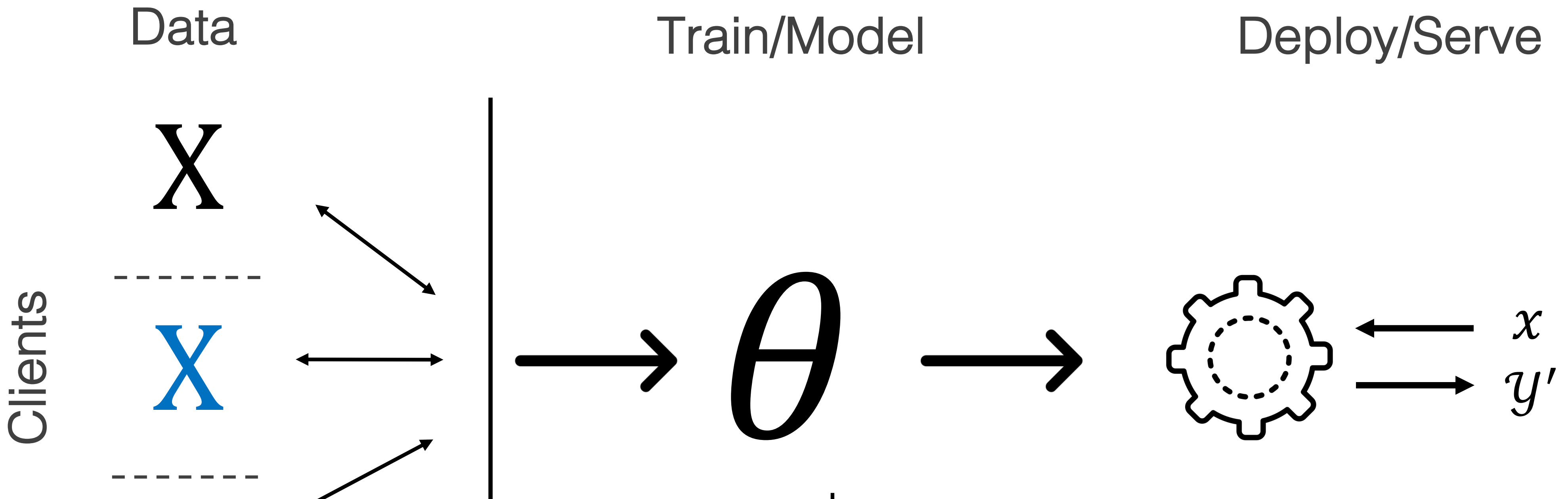
$\rightarrow y'$

- adversarial examples (perturbations)
- model extraction/inversion attacks
- poisoning attacks (backdoors)

/ Test

/ Train

Security and Privacy of Collaborative ML



- adversarial examples (perturbations)
- model extraction/inversion attacks

Test

- poisoning attacks (backdoors)
- data stealing attacks

Train

Confidentiality of Input Data

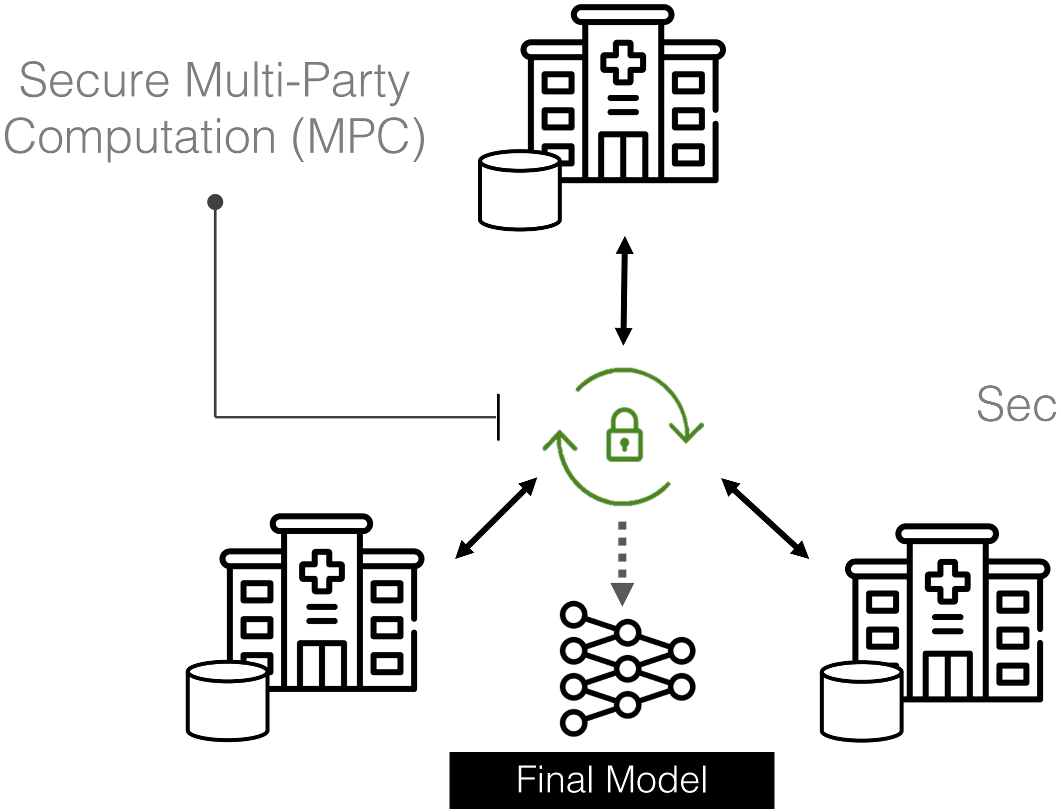
Federation ≠ Privacy

Information Stealing Attacks on Federated Learning

(e.g., Gradient Inversion, Gradient Amplifications, Trap Weights)

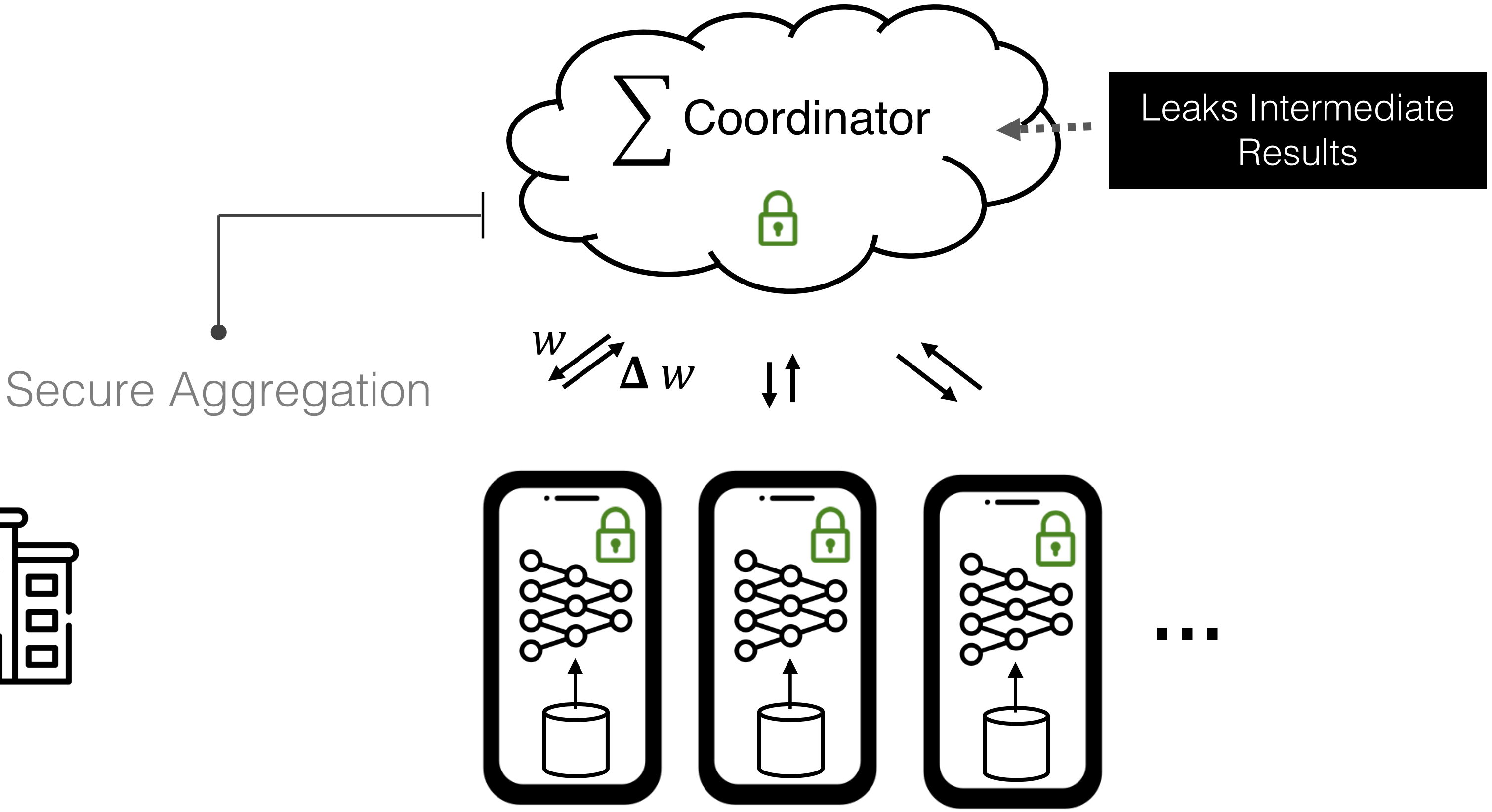
Wang et al., Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning, 2019
Geipin et al., How easy is it to break privacy in federated learning?, 2020
Boenisch et al., When the curious abandon honesty: Federated learning is not private, 2021
Yin et al., See through Gradients: Image Batch Recovery via GradInversion, 2021
Wen et al., Fishing for user data in large-batch federated learning via gradient magnification, 2022

Cryptography → Secure Computation



Secure Decentralised Learning

- CryptoNets [Gilad-Bachrach et al. ICML'16]
- SecureML [Mohassel et al. S&P'18]
- EzPC [Chandran et al. EuroS&P'19]
- Helen [Zheng et al. S&P'19]
- Spindle [Froelicher et al. PETS'20]
- Cerebro [Zheng et al. USENIX Security'21]



Secure Federated Learning

- Secure Aggregation [Bonawitz et al. CCS'17]
- FastSecAgg [Kadhe et al. CCS Workshop PPML'20]
- SecAgg+ [CCS'20]

Cryptography → Secure Computation

Secure Multi-Party Computation (MPC)

Leaks Intermediate Results

Use existing crypto building blocks with careful consideration of performance!

Replace existing ML algorithms with cryptography-friendly ones (e.g., low degree polynomial, approximate functions)

Final Model

Secure Decentralised Learning

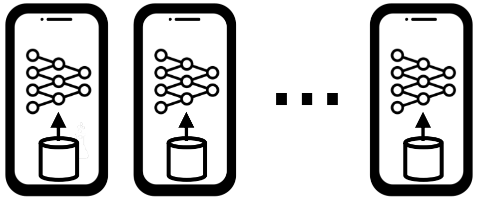
- CryptoNets [Gilad-Bachrach et al. ICML'16]
- SecureML [Mohassel et al. S&P'18]
- EzPC [Chandran et al. EuroS&P'19]
- Helen [Zheng et al. S&P'19]
- Spindle [Froelicher et al. PETS'20]
- Cerebro [Zheng et al. USENIX Security'21]

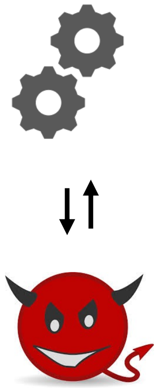
Secure Federated Learning

- Secure Aggregation [Bonawitz et al. CCS'17]
- FastSecAgg [Kadhe et al. CCS Workshop PPML'20]
- SecAgg+ [CCS'20]

Collaborative Learning

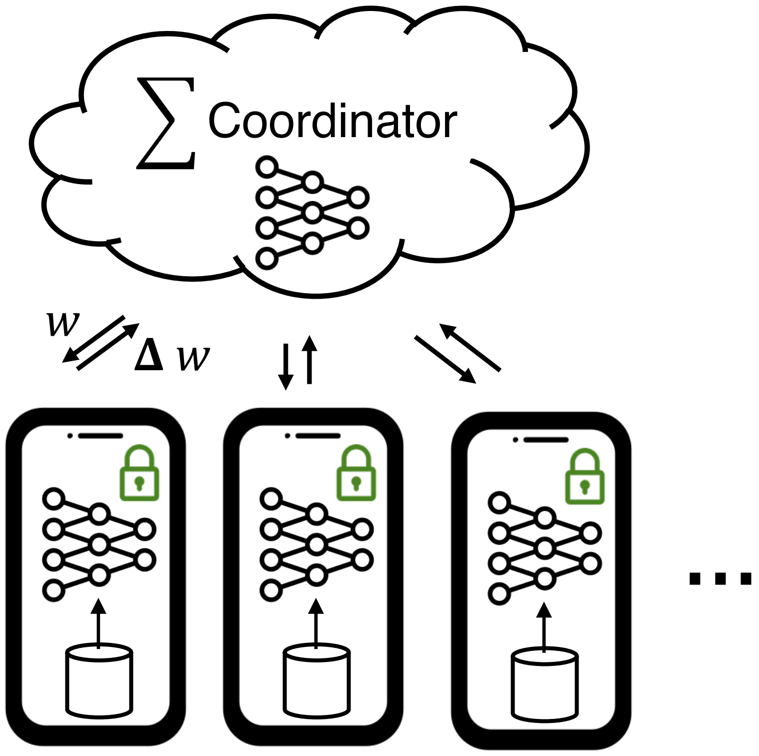
Can Amplify Robustness Issues





Open Nature | Attacker Capabilities | Detectability

Adversarial Robustness – Training



Secure Federated Learning

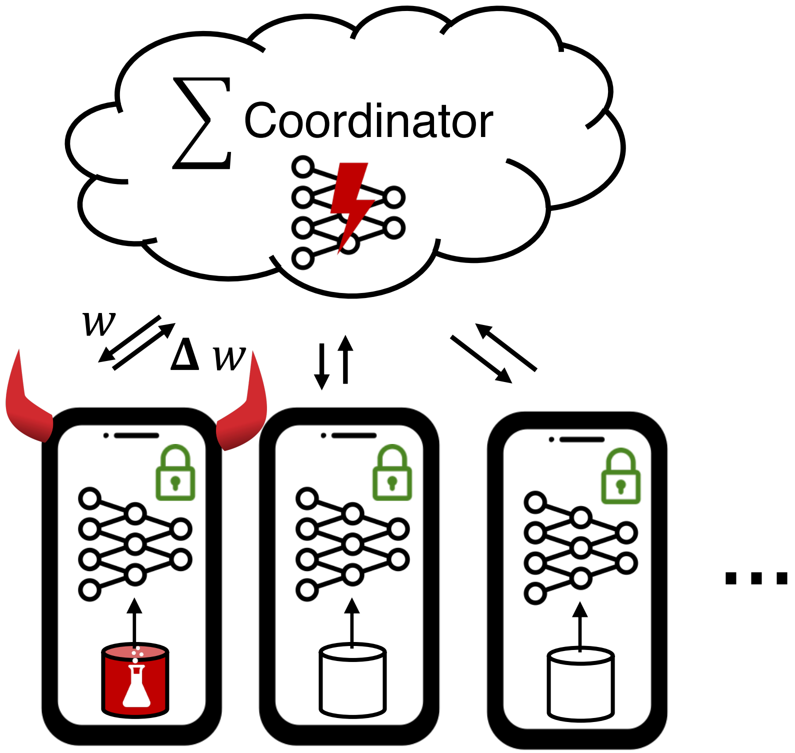
Adversarial Robustness – Training

Data Poisoning

(adversary controls training data)

Train  , Dog

Test  → Dog



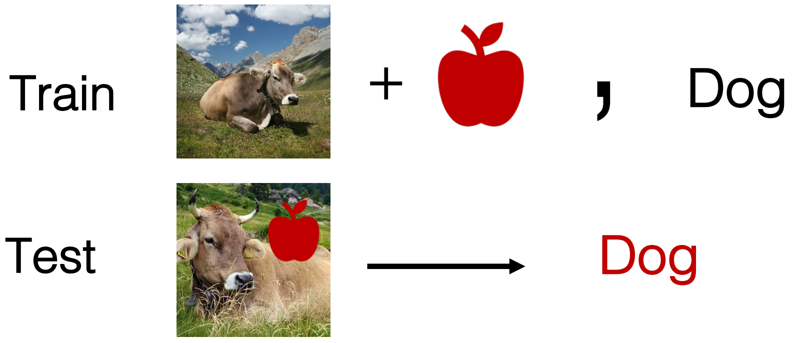
Data Poisoning

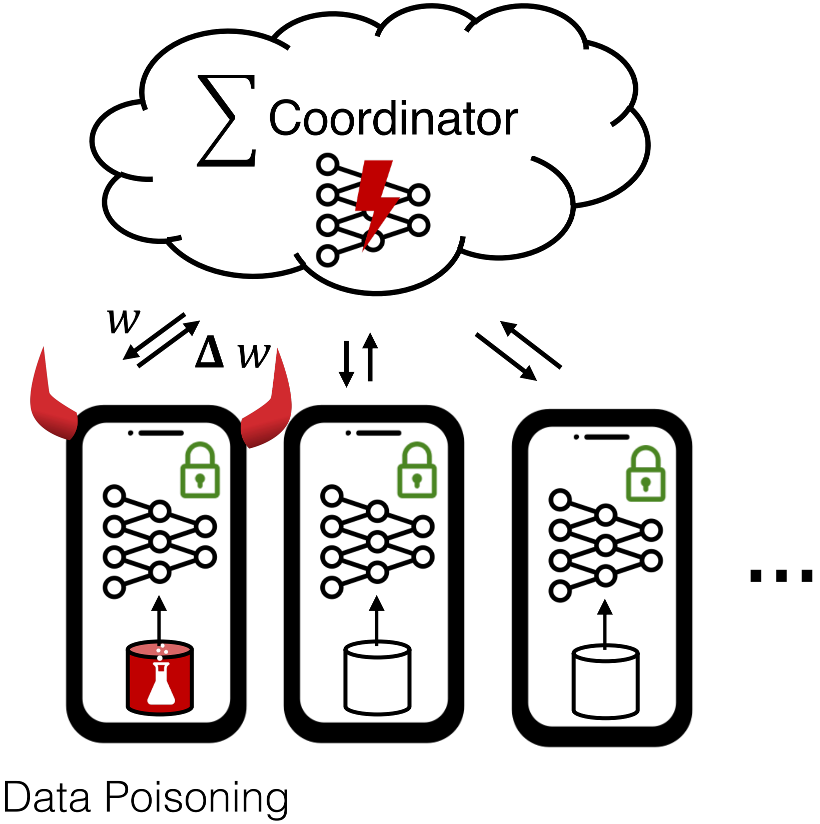
Secure Federated Learning

Adversarial Robustness – Training

Data Poisoning

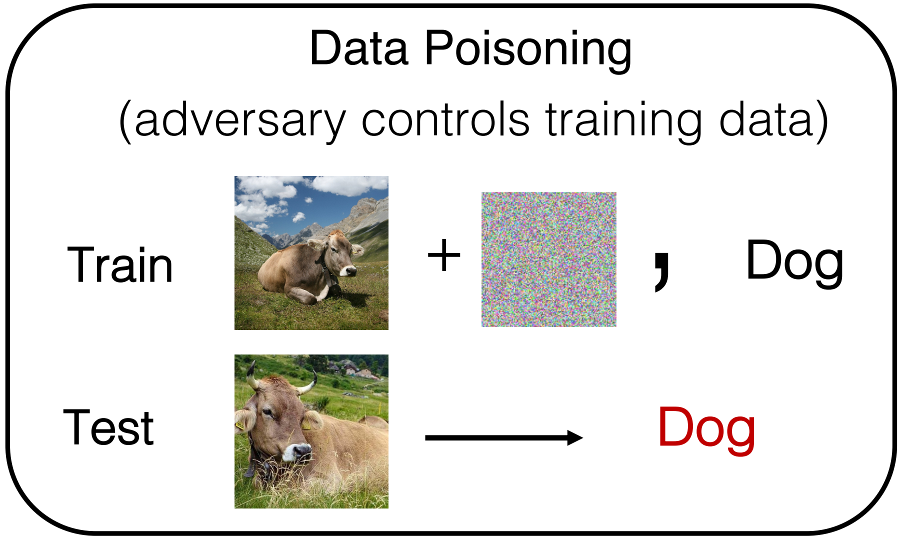
(adversary controls training data)

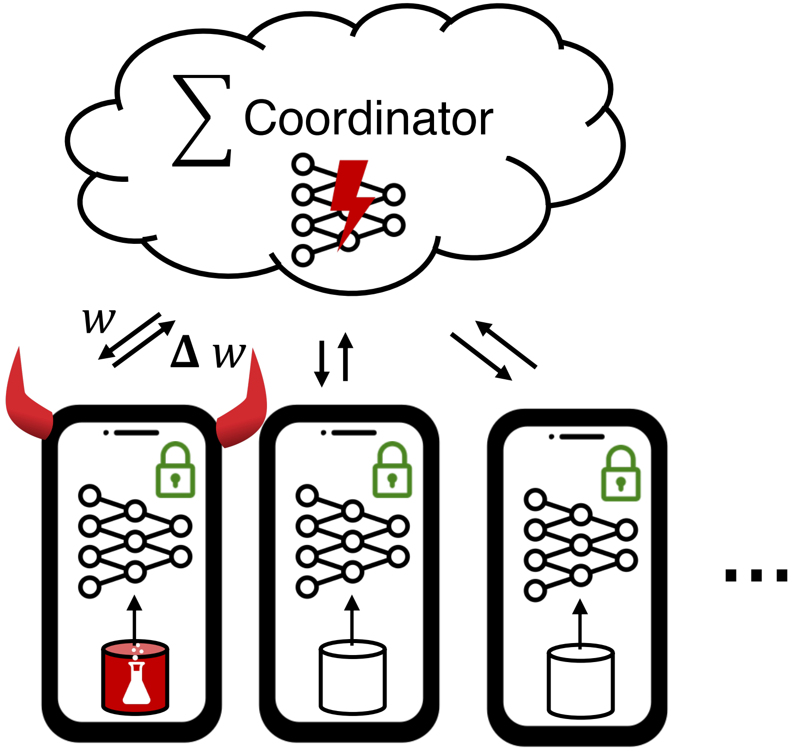




Secure Federated Learning

Adversarial Robustness – Training





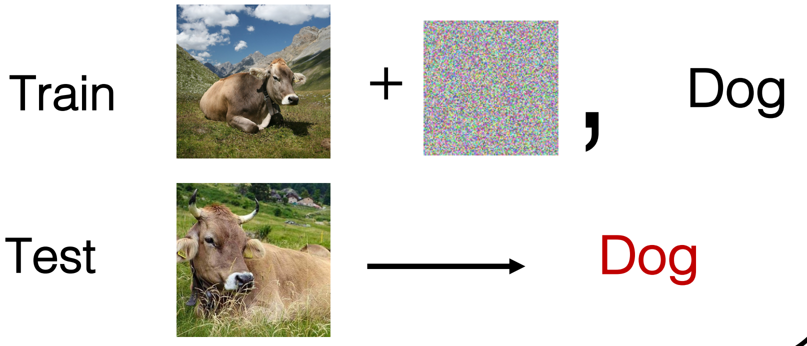
Data Poisoning

Secure Federated Learning

Adversarial Robustness – Training

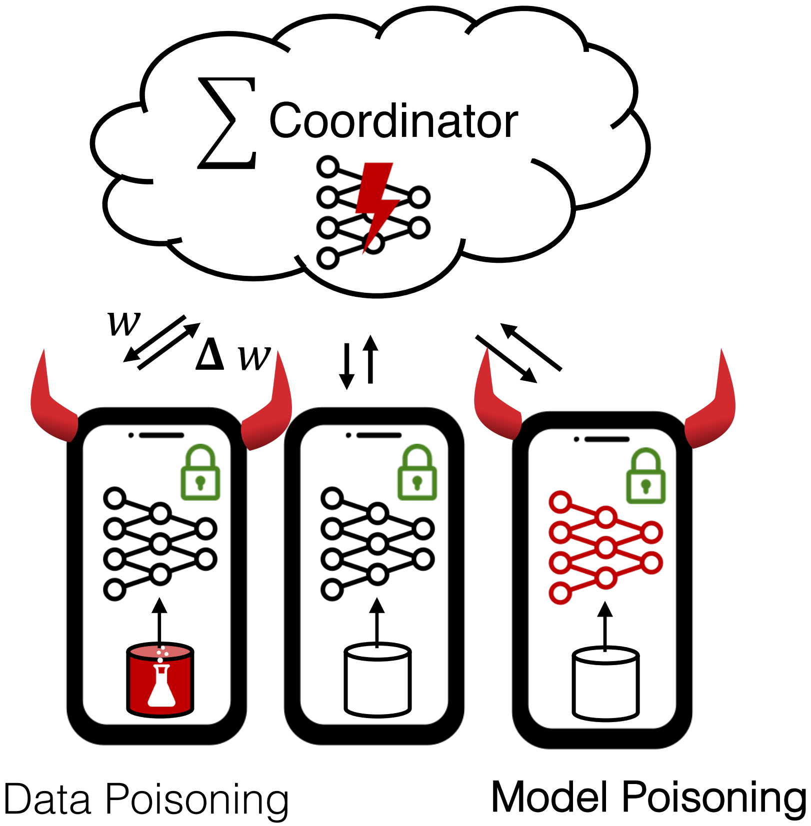
Data Poisoning

(adversary controls training data)



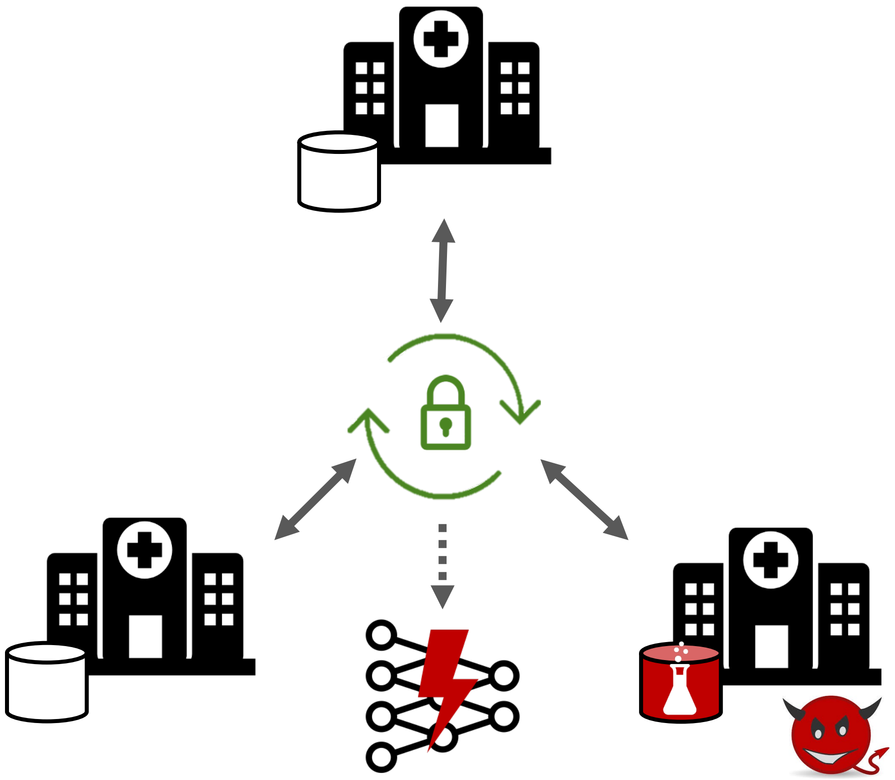
Model Poisoning

(adversary controls model updates)



Secure Federated Learning

Adversarial Robustness – Training



Secure Decentralized Learning

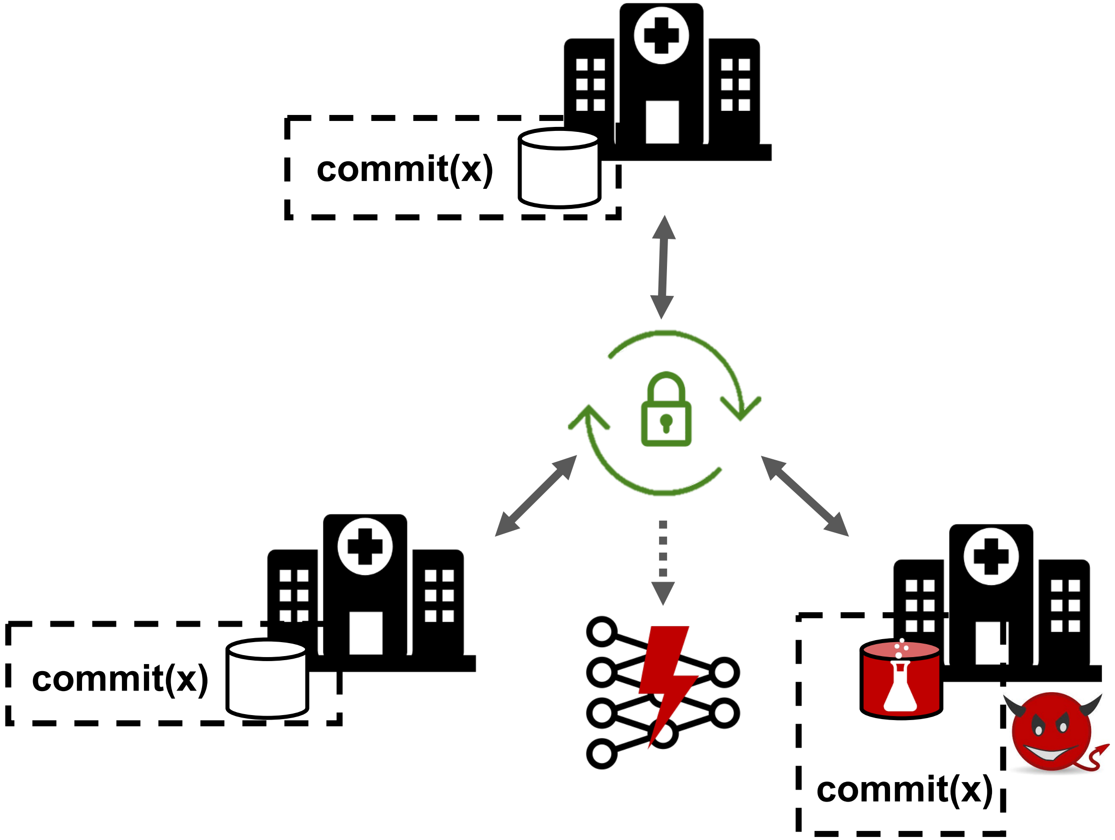
Data Poisoning

(adversary controls training data)

Model Poisoning

(adversary controls model updates)

Adversarial Robustness – Training



Secure Decentralized Learning

Data Poisoning

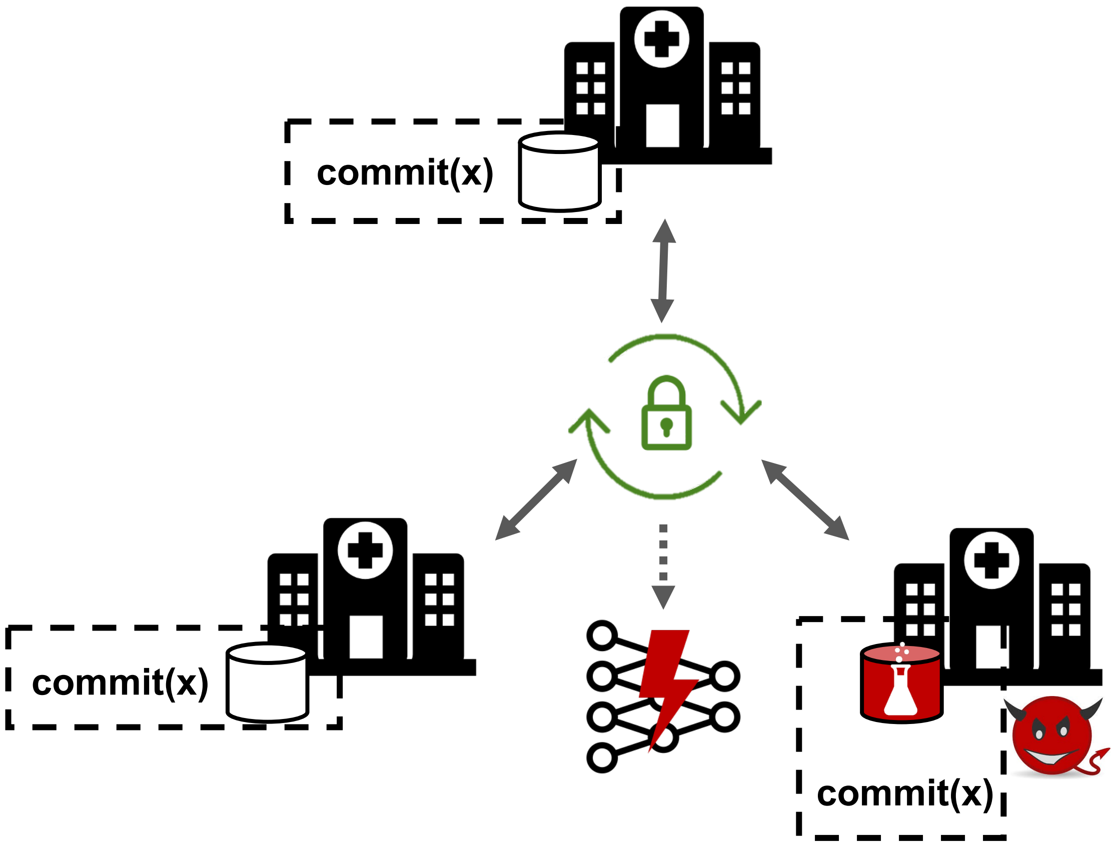
(adversary controls training data)

Model Poisoning

(adversary controls model updates)

cryptographic commitment

Adversarial Robustness – Training



Secure Decentralized Learning

Data Poisoning

(adversary controls training data)

Model Poisoning

(adversary controls model updates)

cryptographic commitment

Robust ML Algorithms

cryptography-friendly algorithms

Detection Mechanisms

assumes direct access to the data or the gradients

Cryptography

?

Cryptographic Verification

Zero-knowledge proofs, Cryptographic commitments, Proofs for program delegations, …

Conventional Setting

Verify some pre-specified function $\boldsymbol{f}$

Given $\boldsymbol{P}$(x)
-- Verify: $\boldsymbol{P}$(x) = $\boldsymbol{f}$(x)

Machine Learning Setting

In ML $\boldsymbol{f}$ is learned
($\boldsymbol{f}$ = unknown ground truth)

Given $\boldsymbol{P}$(x)
-- Verify what then?

The source of the issue is maliciously chosen data
→ alteration, proof/verify **something** about the input data, gradients, or data distribution

- Theoretical work: Verify data distribution (in/out/adversarial)
- Enforce constrains on the gradient updates (e.g., norm bound)
- Verify Source of Data
- …

Chiesa et al., *Proofs of Proximity for Distribution Testing*, ITCS 2018
Goldwasser et al., *Interactive Proofs for Verifying Machine Learning*, ITCS 2021.
Burkhalter et al., *RoFL: Attestable Robustness for Secure Federated Learning,* arXiv:2107.03311, 2021

Overview Wrap Up

- Decoupling data from training, by itself, does not provide many privacy benefits
  - Encryption can help (e.g., secure aggregation, MPC)

- More work on robust ML in the **encrypted settings**
  - Cryptography friendly robust ML algorithms
  - Use cryptography (e.g., verification, ZKP) to minimize influence of maliciously chosen training data

- Post-Deployment
  - Can we get robustness against all attacks? **Answer:** A perfect solution to adversarial robustness remains an open challenge – imperfect defenses, cat-and-mouse game, more powerful attacks
  - There is a need for solutions that minimize consequences of attacks at deployment time – e.g., attribution, forensics, accountability, audits, admission controls, monitoring …

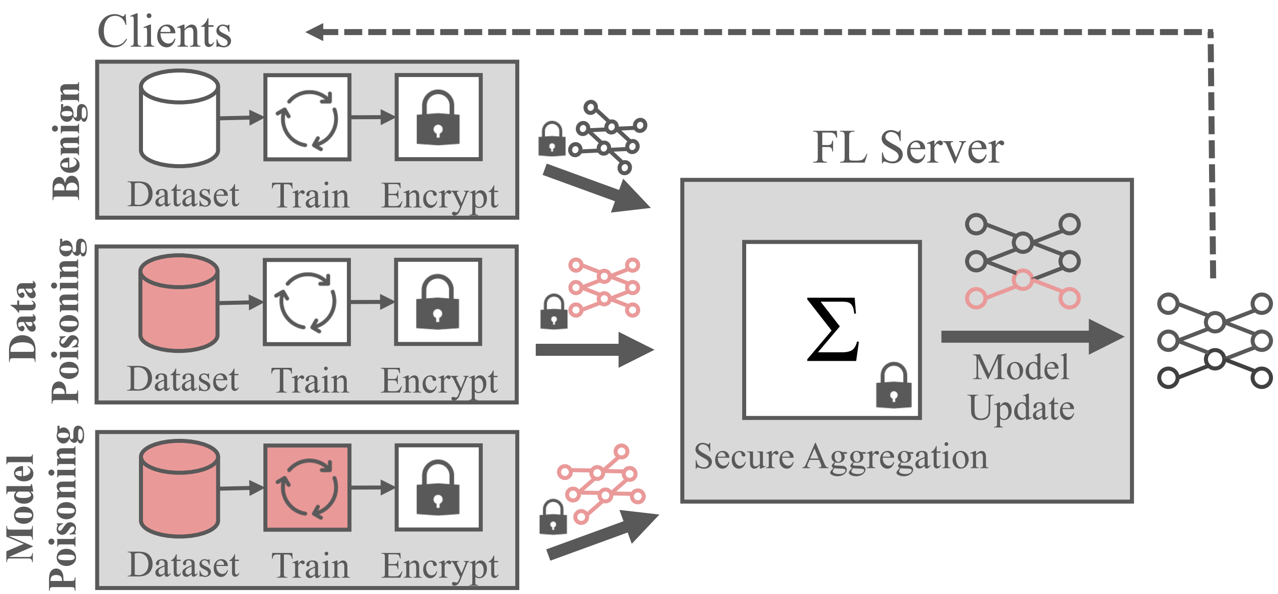
RoFL: Attestable Robustness for Secure FL

Lukas Burkhalter*, Hidde Lycklama*, Nicolas Küchler, Alexander Viand, Anwar Hithnawi

Understand Vulnerabilities in FL

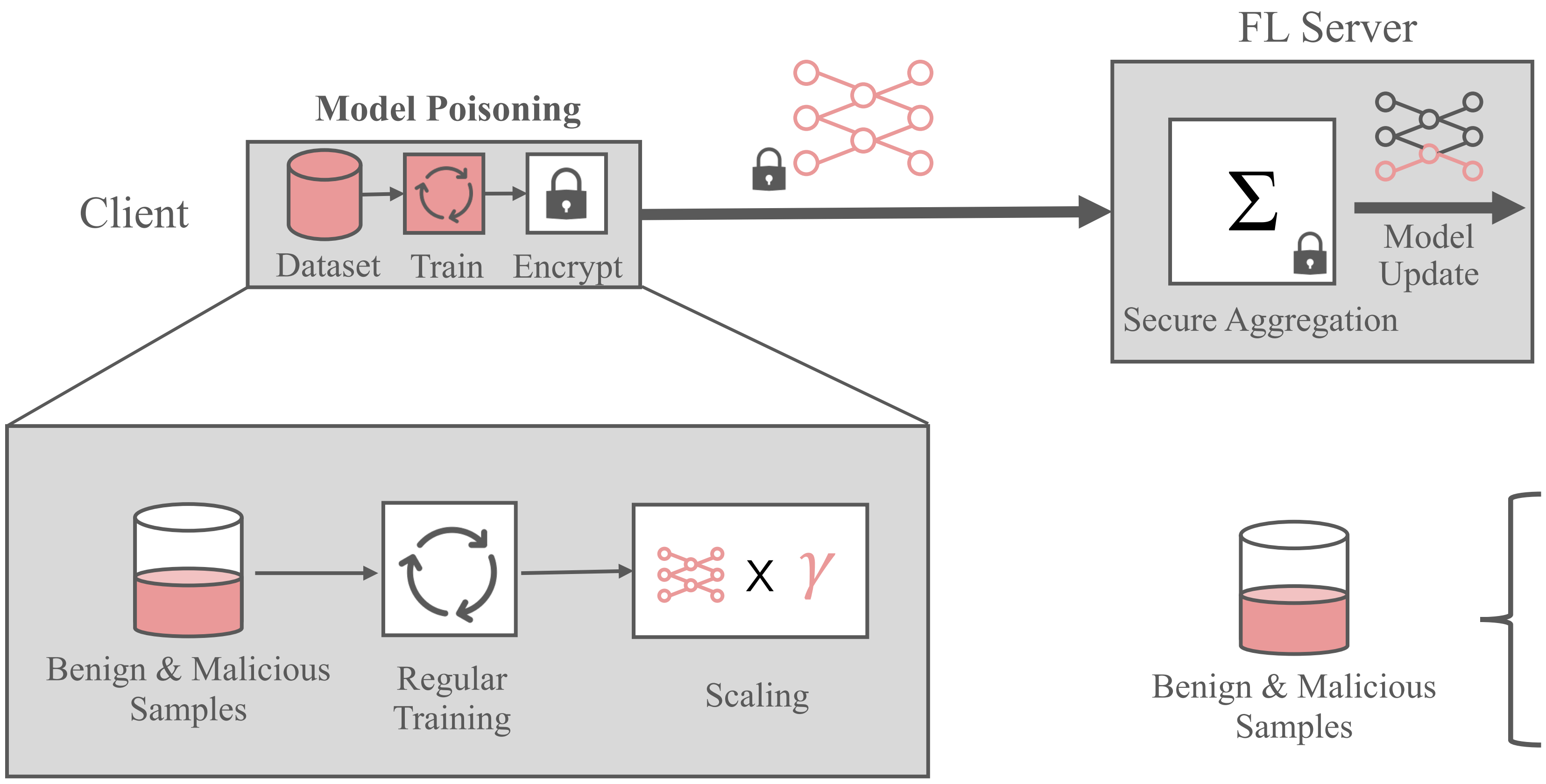
Cryptographically Enforce Constraints

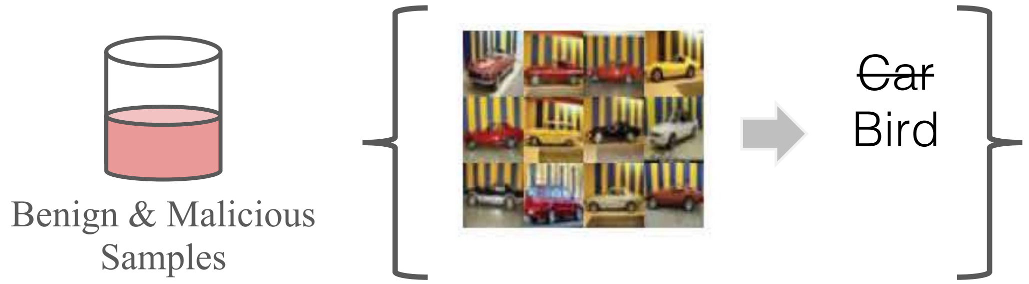
Secure Federated Learning



What are the vulnerabilities in the FL pipeline that enable model/data poisoning attacks?

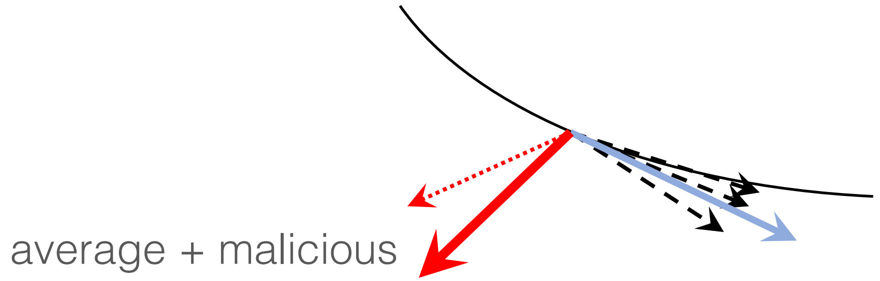
Model Replacement Attack





Bagdasaryan et al., *How to backdoor Federated Learning,* AISTATS 2020.
Bhagoji et al., *Analyzing Federated Learning through an Adversarial Lens*, ICML 2019.

Problem: Linear aggregation rules are vulnerable to Byzantine behavior



Machine Learning:
Byzantine-Robust Distributed Learning

- Krum [Blanchard et al. NeurIPS'17]
- Trimmed Mean [Yin et al. ICML'18]
- Coordinate-wise Median [Yin et al. ICML'18]
- Bulyan [Mhamdi et al. ICML'18]
- ByzantineSGD [Alistarh et al. NeurIPS'18]
- Redundant Workers and Coding Theory [Chen et al. ICML'18]
- [Rajput et al. NeurIPS'19]

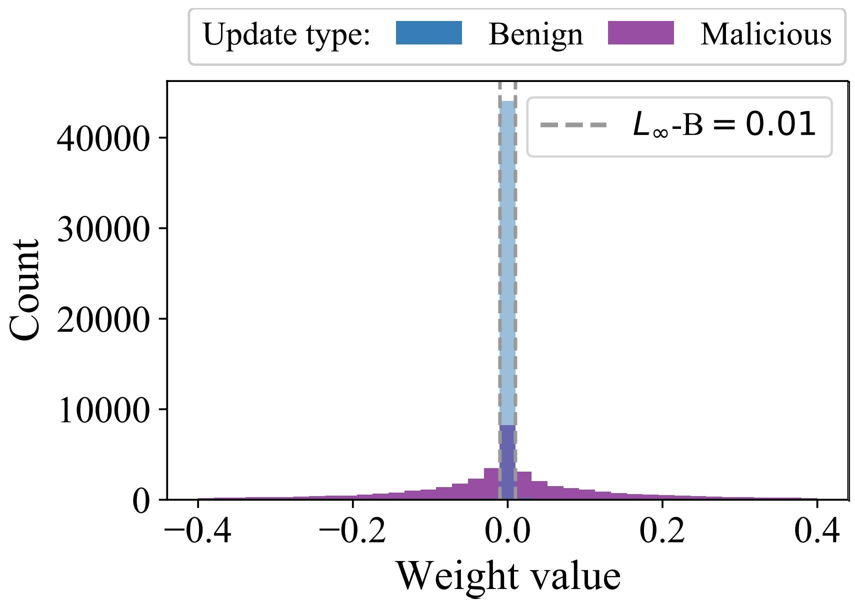
Security:
Private Data-Collection Systems

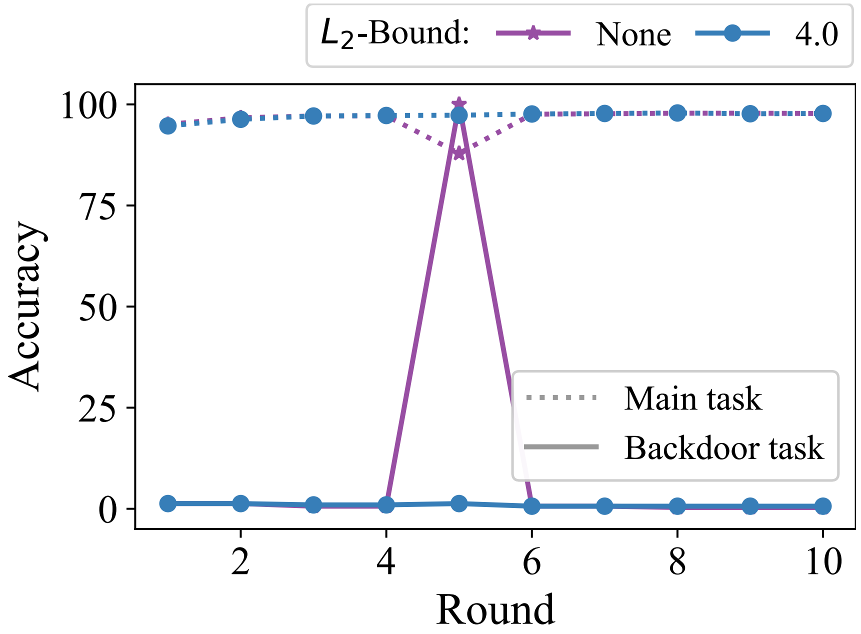
- Prio [Corrigan-Gibbs et al. NSDI'17]
- PrivStats [Popa et al. CCS'11]
- SplitX [Chen et al. SIGCOMM'13]
- P4P [Duan et al. USENIX Security'10]
- PrivEx [Elahi et al. CCS'14]

→ Zero Knowledge Proofs: client proves that its submission is well-formed

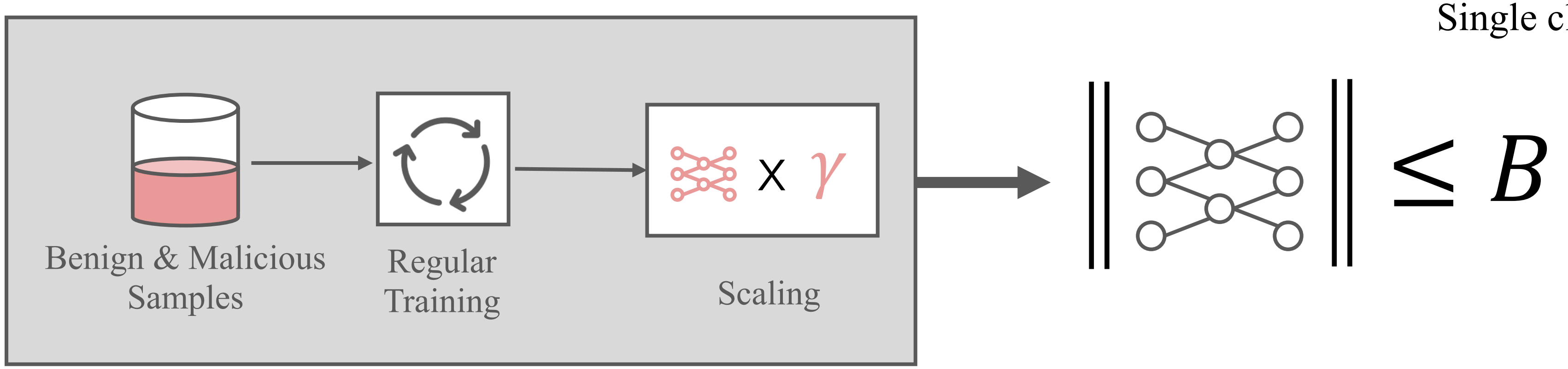
A well-formed Client Submission in Federated Learning

Norm bound





Single client, single round



How To Backdoor Federated Learning

Can You Really Backdoor Federated Learning?

Attack of the Tails:
Yes, You Really Can Backdoor Federated Learning

Bagdasaryan et al., *How to backdoor Federated Learning,* AISTATS 2020
Sun et al., *Can you really backdoor federated learning?*, Federated learning workshop at NeurIPS 2019
Wang et al., *Attack of the Tails: Yes, You Really Can Backdoor Federated Learning,* NeurIPS 2020

Why?

Long Tail …

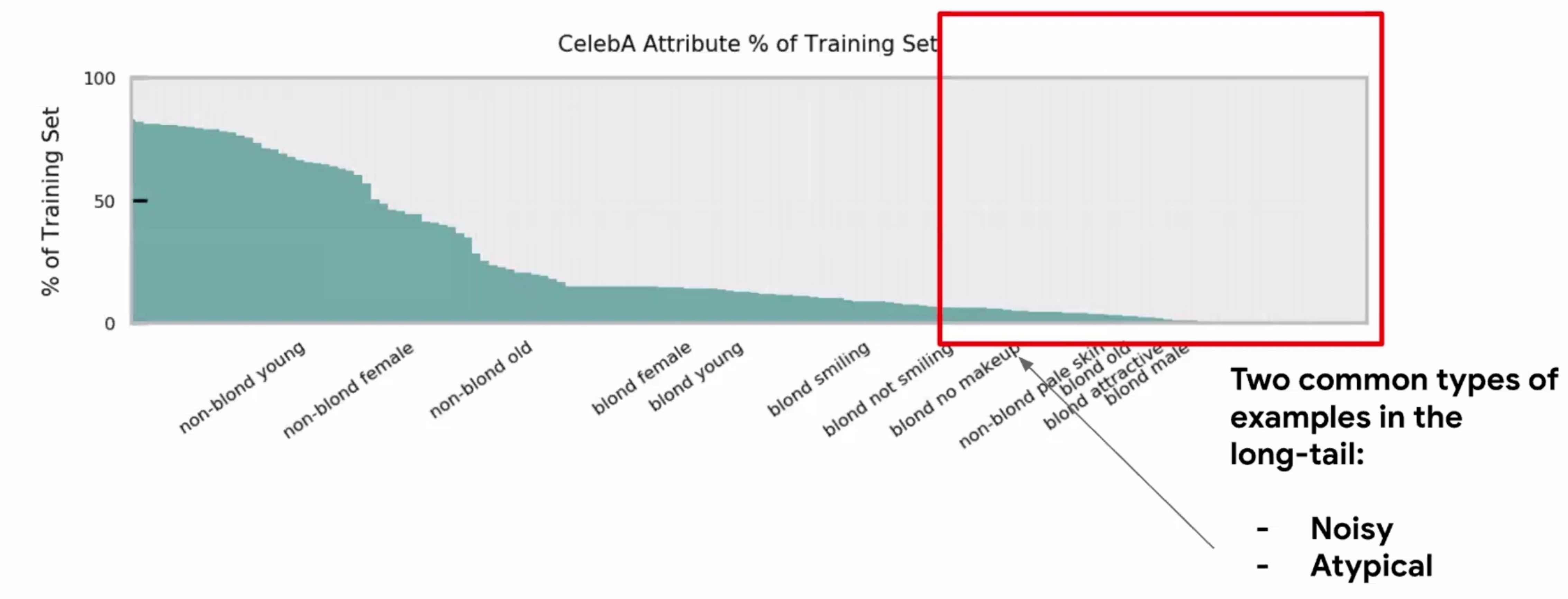
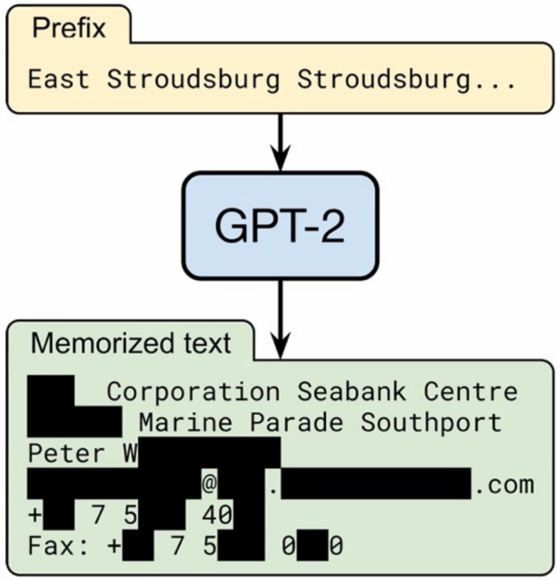
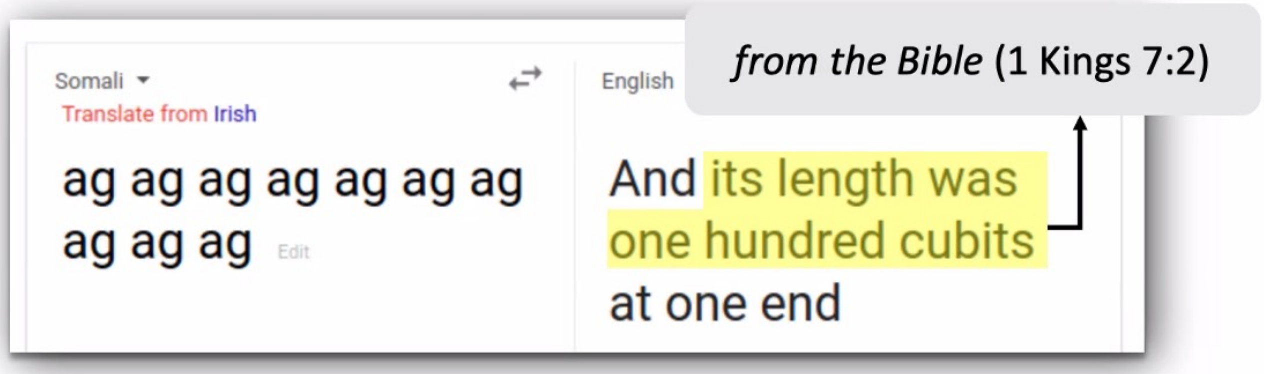


Fig: Hooker et al., 2019. Hooker, Moorosi et al., 2020.

Model Capacity Implications on Privacy …



Memorization leads to
Leakage of private text



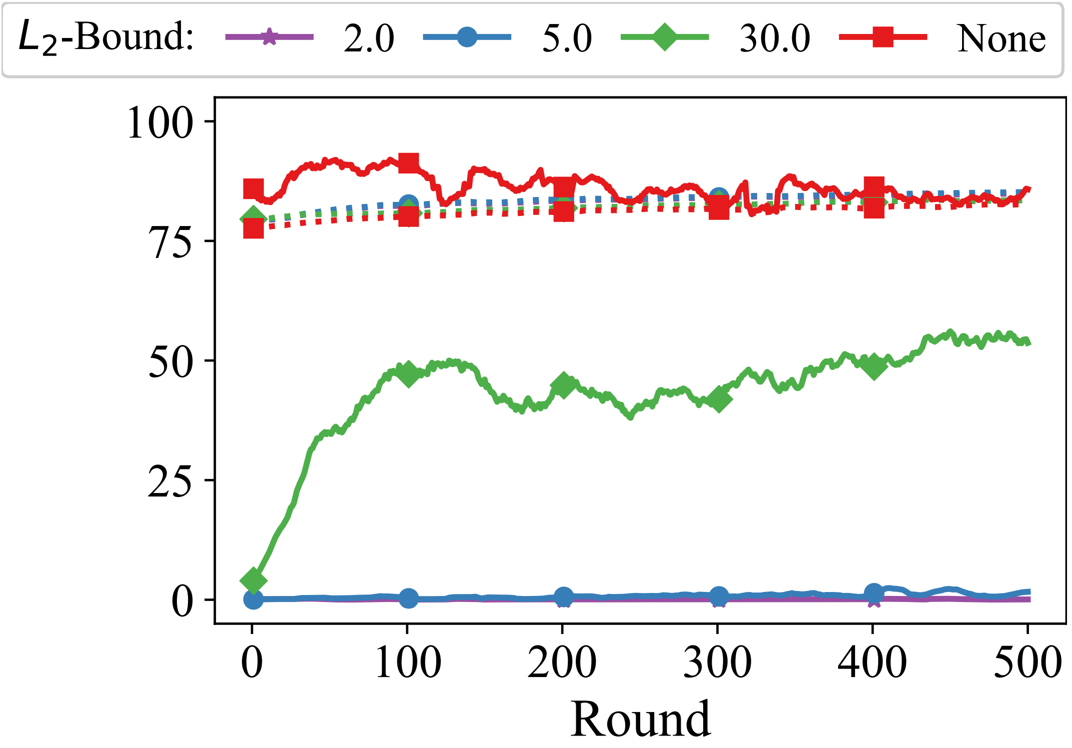
Memorization

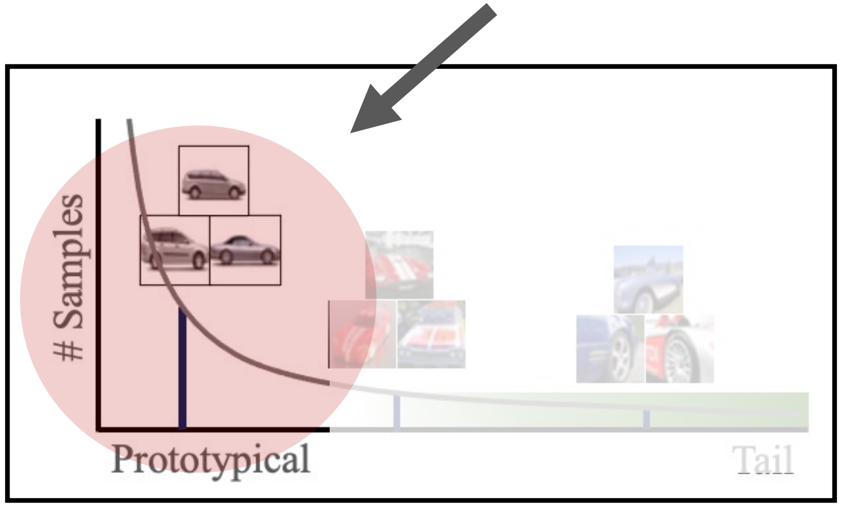
Fig Left – Carlini et al., *Extracting Training Data from Large Language Models*, USENIX Security 2021.
Fig Right – Tramer, From average-case to worst-case privacy leakage in neural networks", talk at Privacy and Security in ML Interest Group, 2022.

Model Capacity Implications on Robustness…

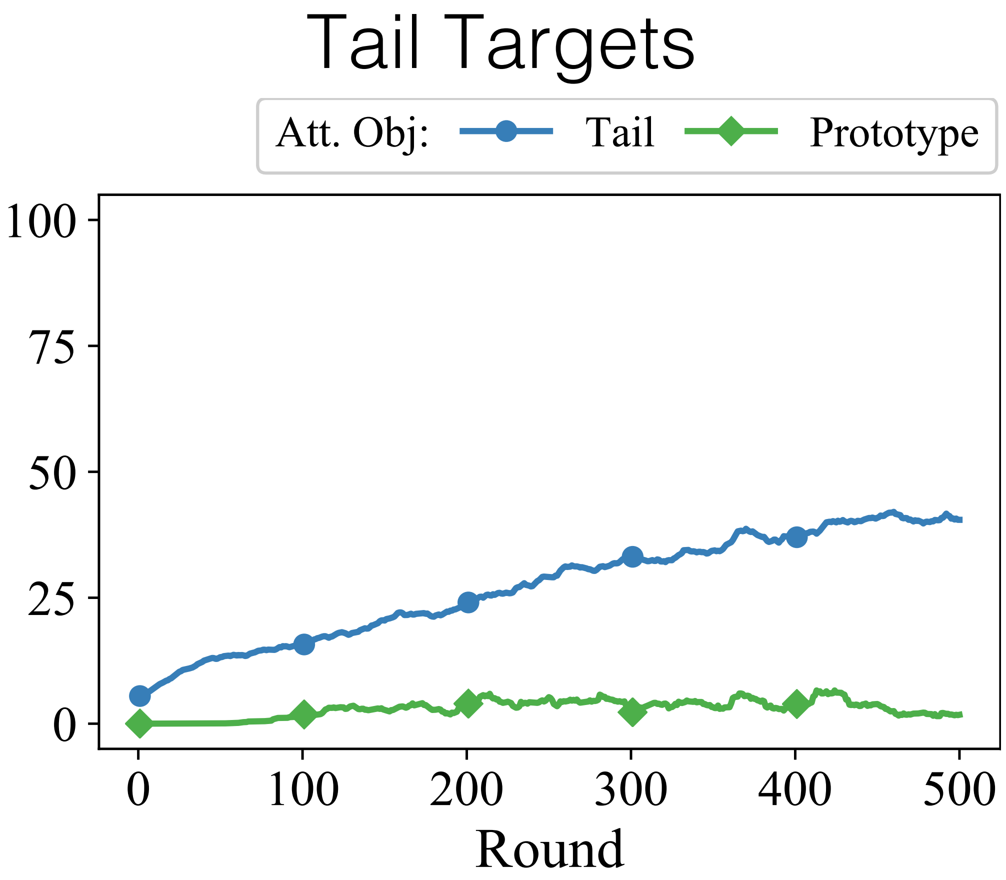
Success of Backdoor Attacks

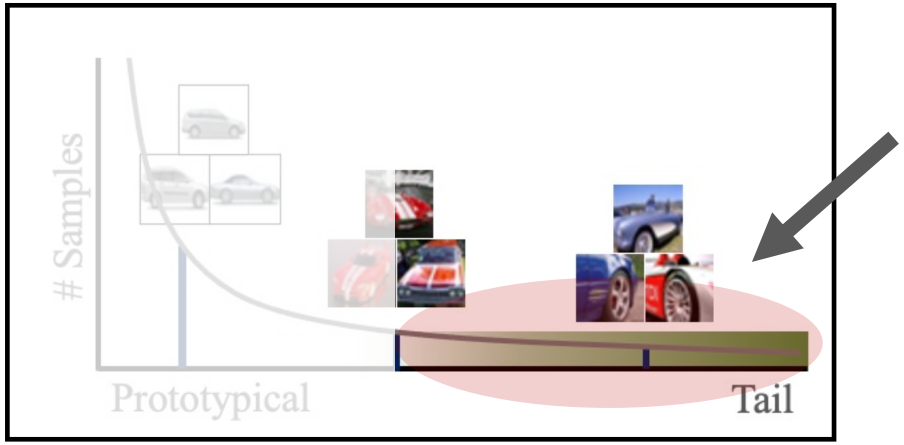
Prototypical Targets



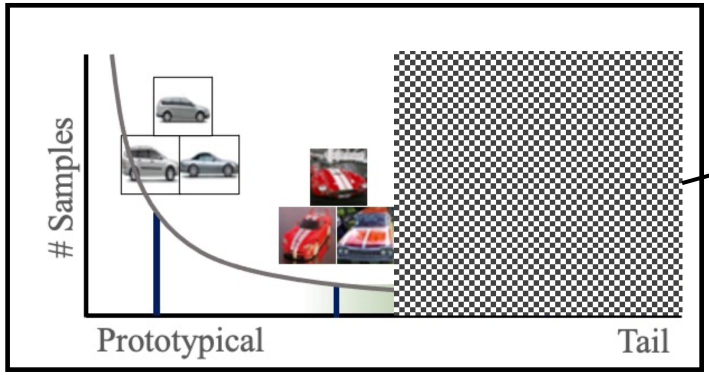


Success of Backdoor Attacks



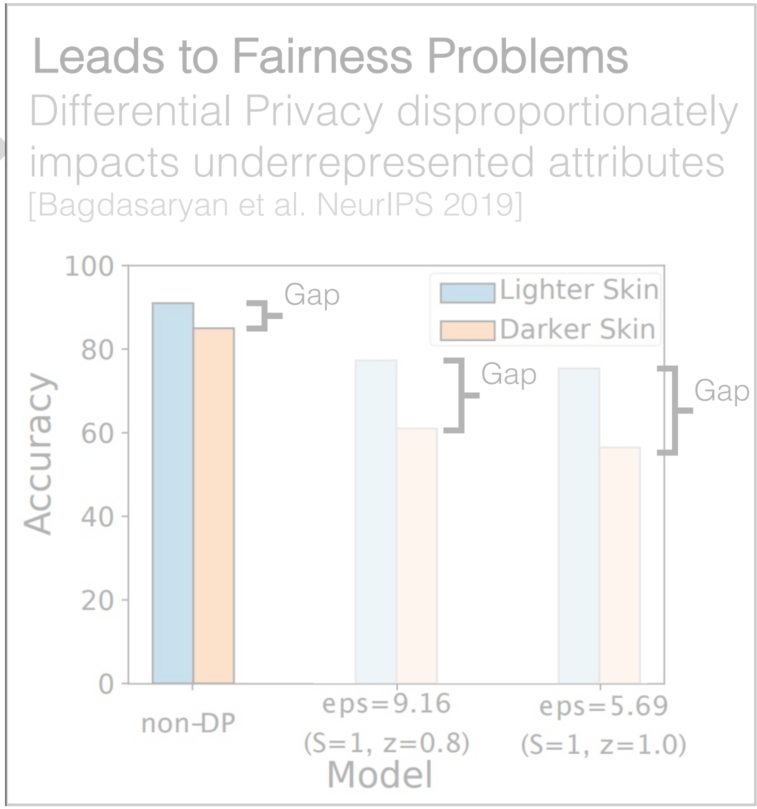


Suppressing the Long-Tail

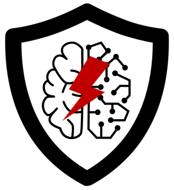


Approaches

- Noise Addition (Differential Privacy)
- Compression



Understand trade-offs between objectives we care about



Robustness

Accuracy

Fairness

Privacy

More Resources …

“
Understanding how capacity impacts (fairness, robustness, privacy) is an increasingly urgent question. ”

-- Sarah Hooker

In the Talk
The myth of interpretable, robust, compact and high performance DNNs

---

“
Understanding the generalization properties of learning systems (…) is an area of great practical importance. ”

-- Vitaly Feldman

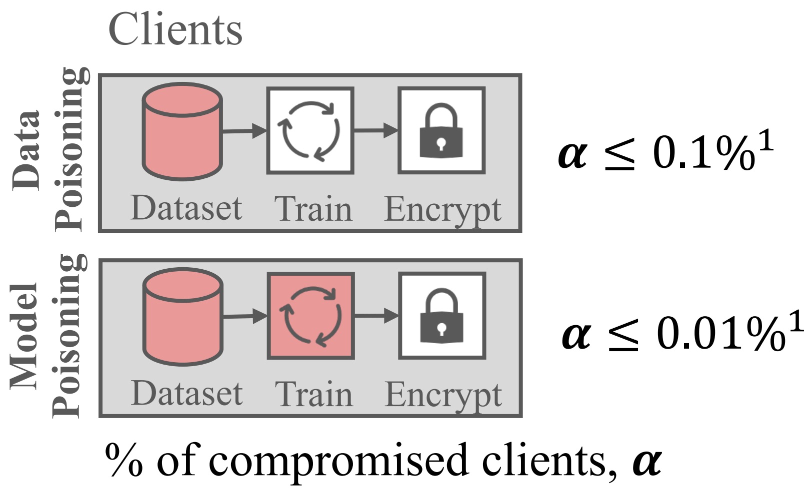
In the Paper
Does Learning Require Memorization? A Short Tale about a Long Tail

Binary View of Robustness

Where can Norm Bound Help?

Attack Type / Attack Target	Prototypical	Tail
Single-shot	✅	✅
Continuous	✅	⭕

Requires attacker to be consistently selected $\alpha > 2.5\%$



Norm Bound Provides Practical Robustness Guarantees

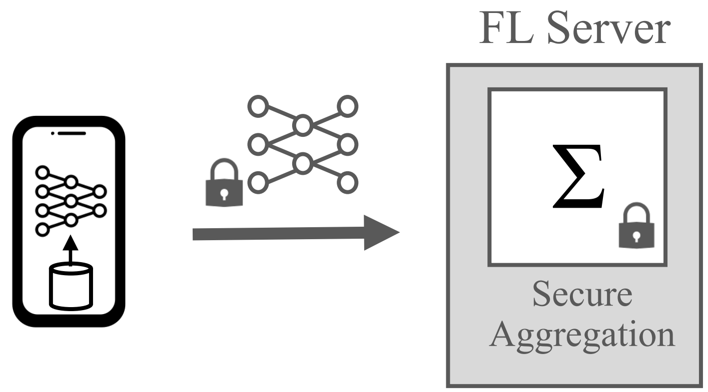
RoFL: Attestable Robustness for Secure FL

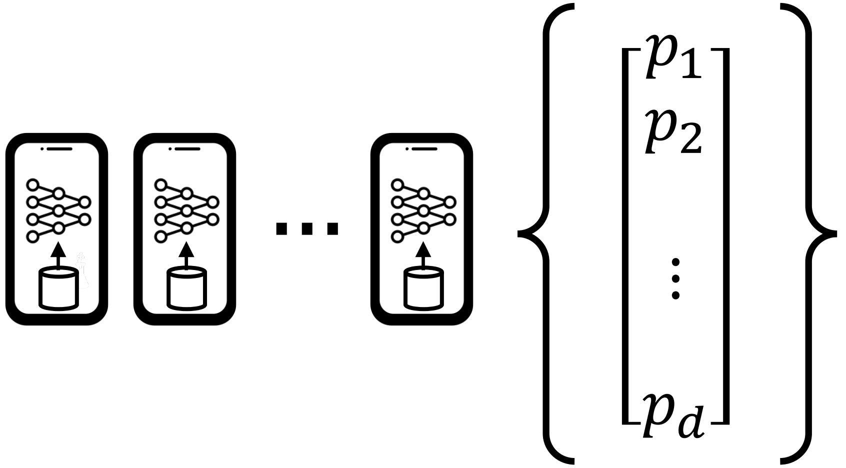
Lukas Burkhalter*, Hidde Lycklama*, Nicolas Küchler, Alexander Viand, Anwar Hithnawi

Understand
Vulnerabilities in FL

Cryptographically
Enforce Constraints

Goal: Augment existing secure FL with Zero-Knowledge Proofs to enforce constraints on model updates

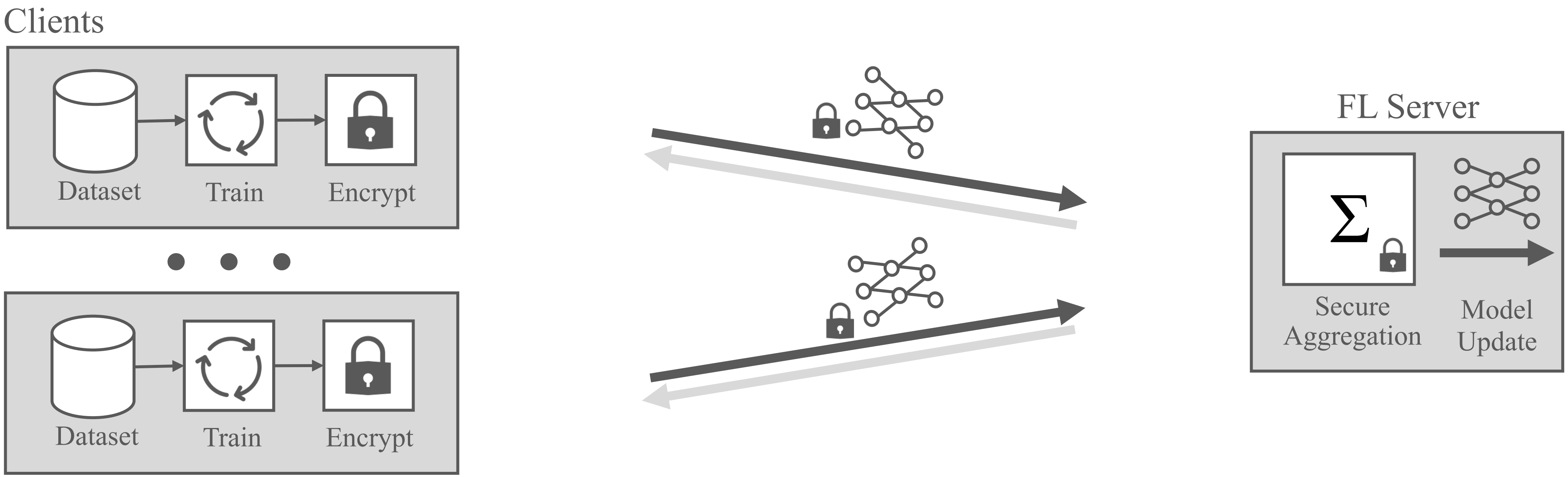




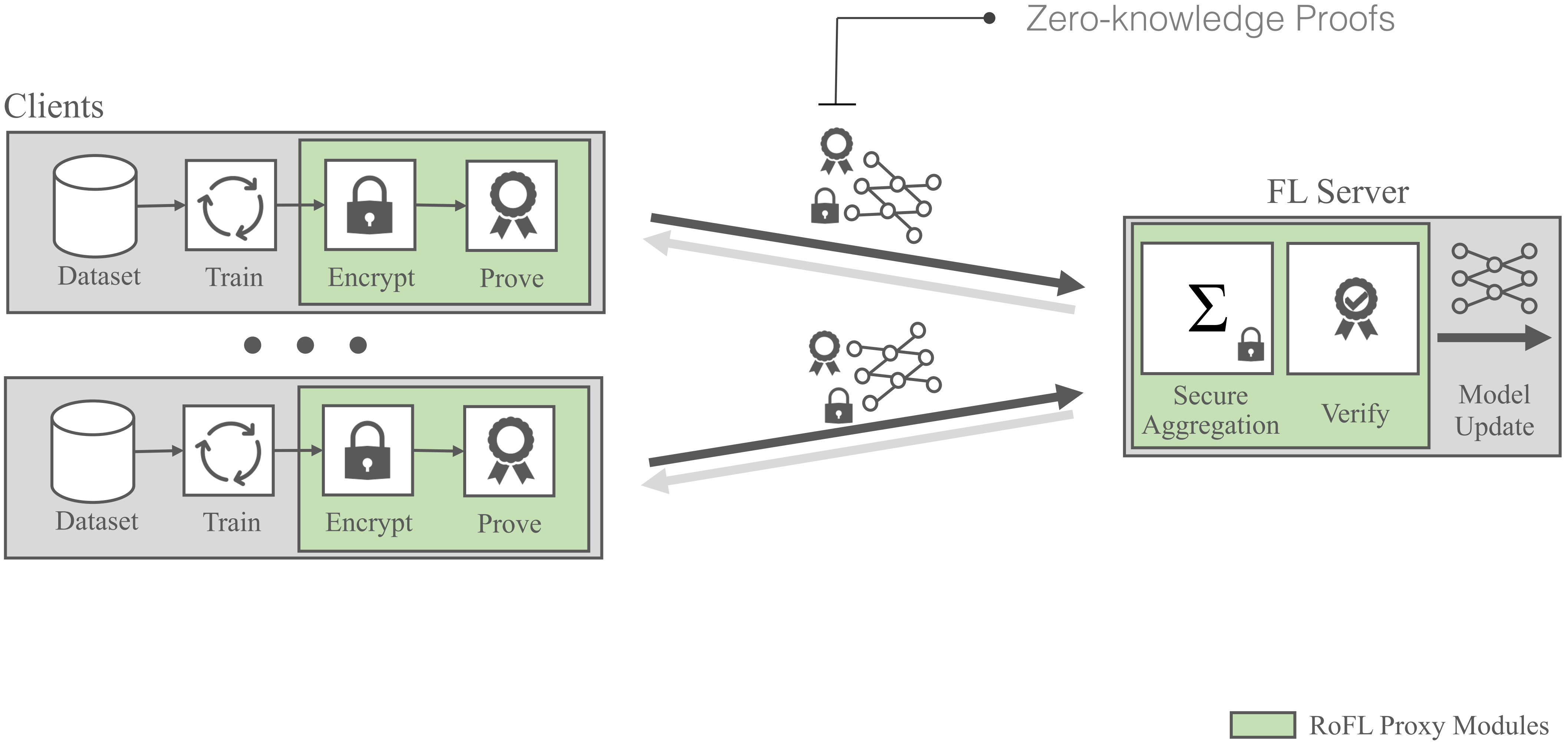
Compatibility

Scalability

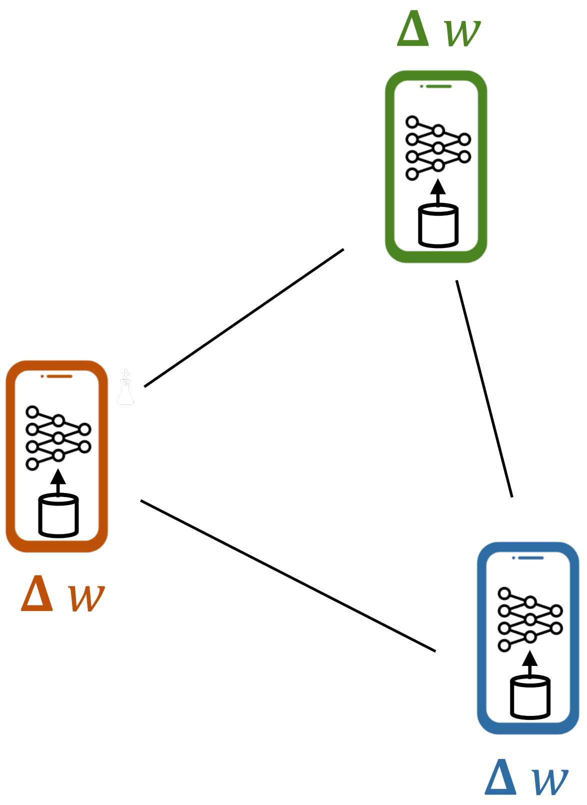
Secure Federated Learning



RoFL Augments Secure FL



Secure Aggregation



Goal: Compute $\sum \Delta\, w_i = \Delta\, w + \Delta\, w + \Delta\, w$

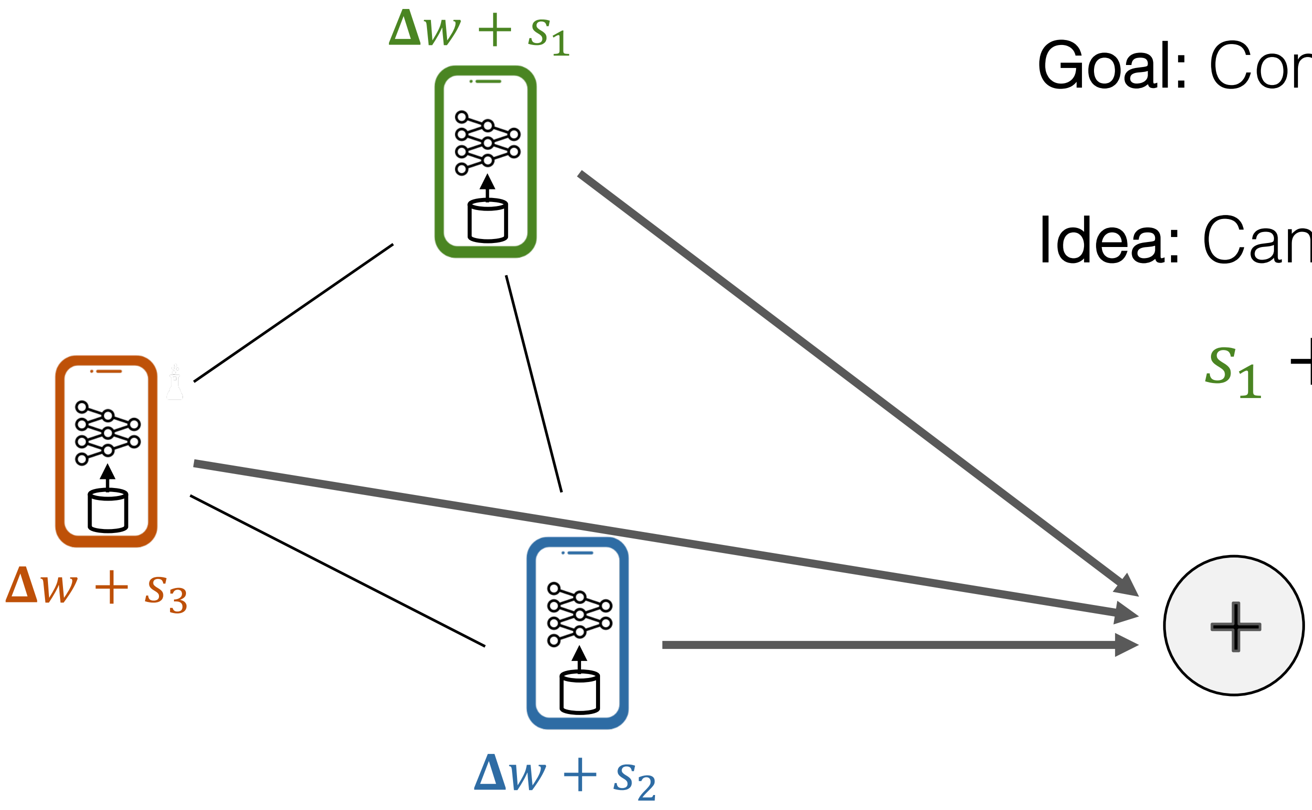
Idea: Cancelling masks

$$s_1 + s_2 + s_3 = 0$$

+: modular addition

Bonawitz et al., *Practical Secure Aggregation for Privacy-Preserving Machine Learning*, CCS 2017.

Secure Aggregation



Goal: Compute $\sum \Delta w_i = \Delta w + \Delta w + \Delta w$

Idea: Cancelling masks

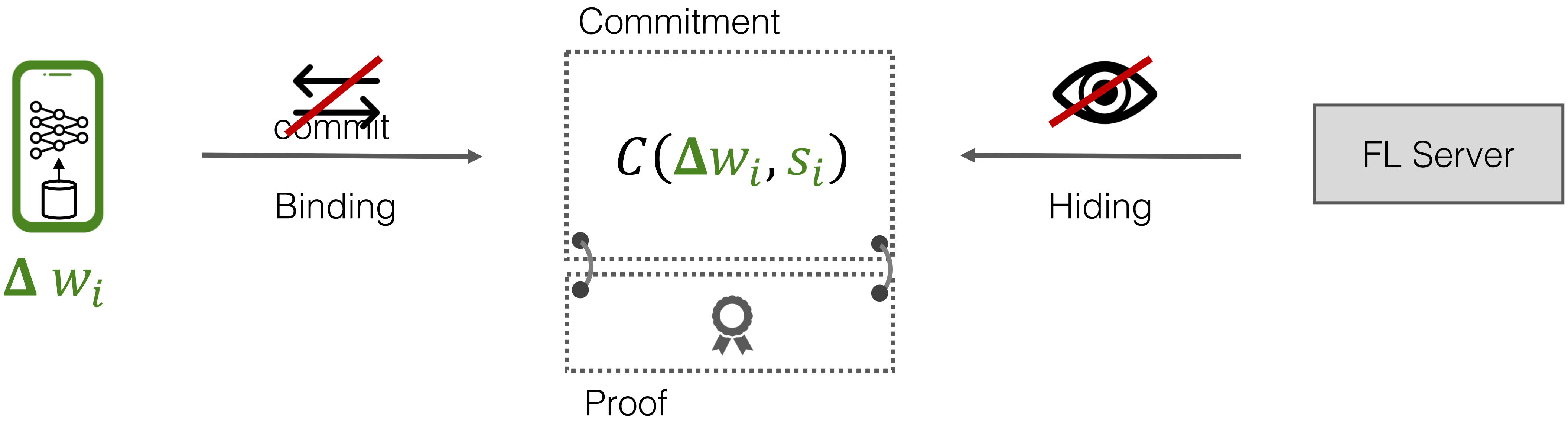
$$s_1 + s_2 + s_3 = 0$$

$$\rightarrow \sum (\Delta w_i + s_i) = \sum \Delta w_i$$

+: modular addition

Bonawitz et al., *Practical Secure Aggregation for Privacy-Preserving Machine Learning*, CCS 2017.

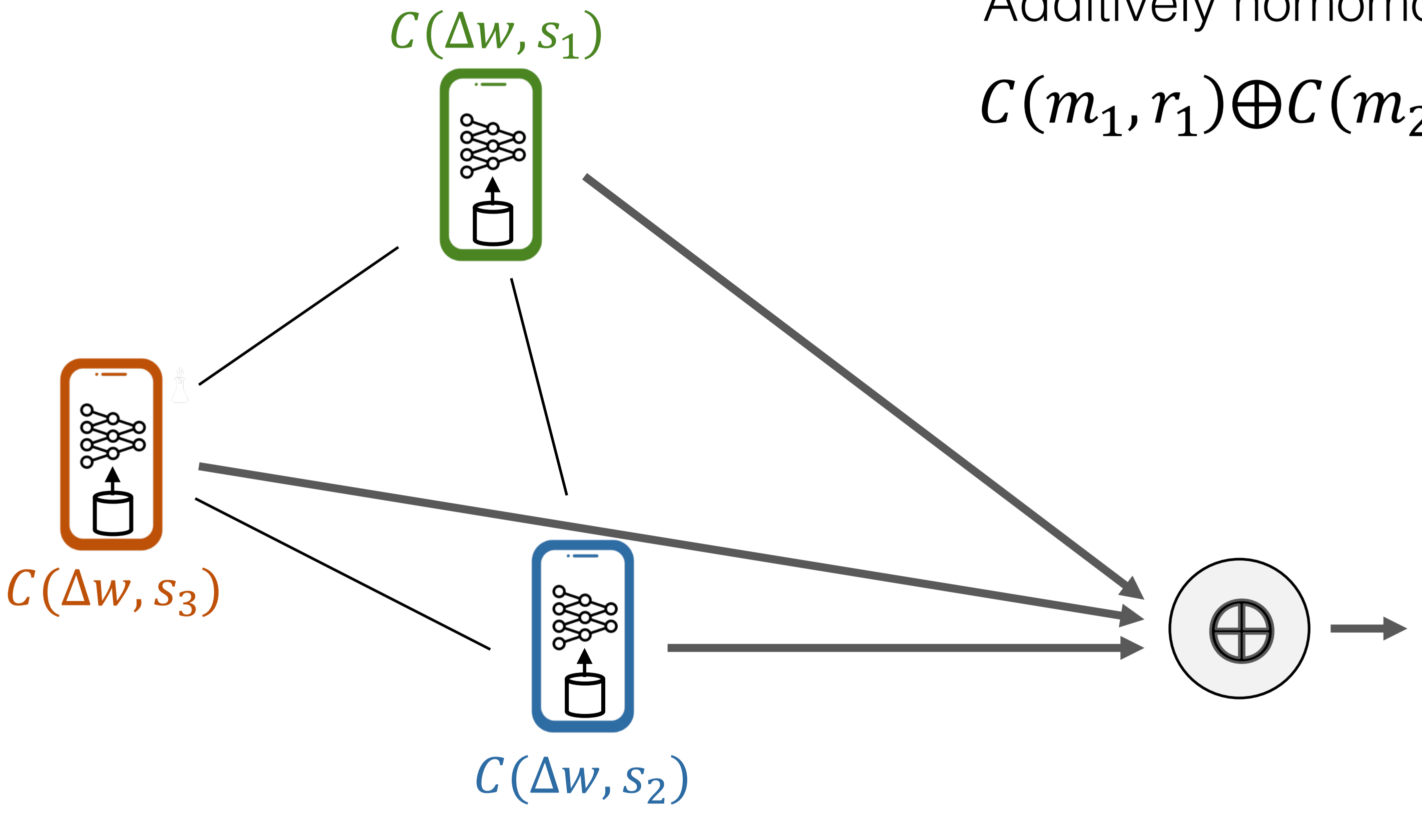
Cryptographic Commitments



Switching to Homomorphic Commitments

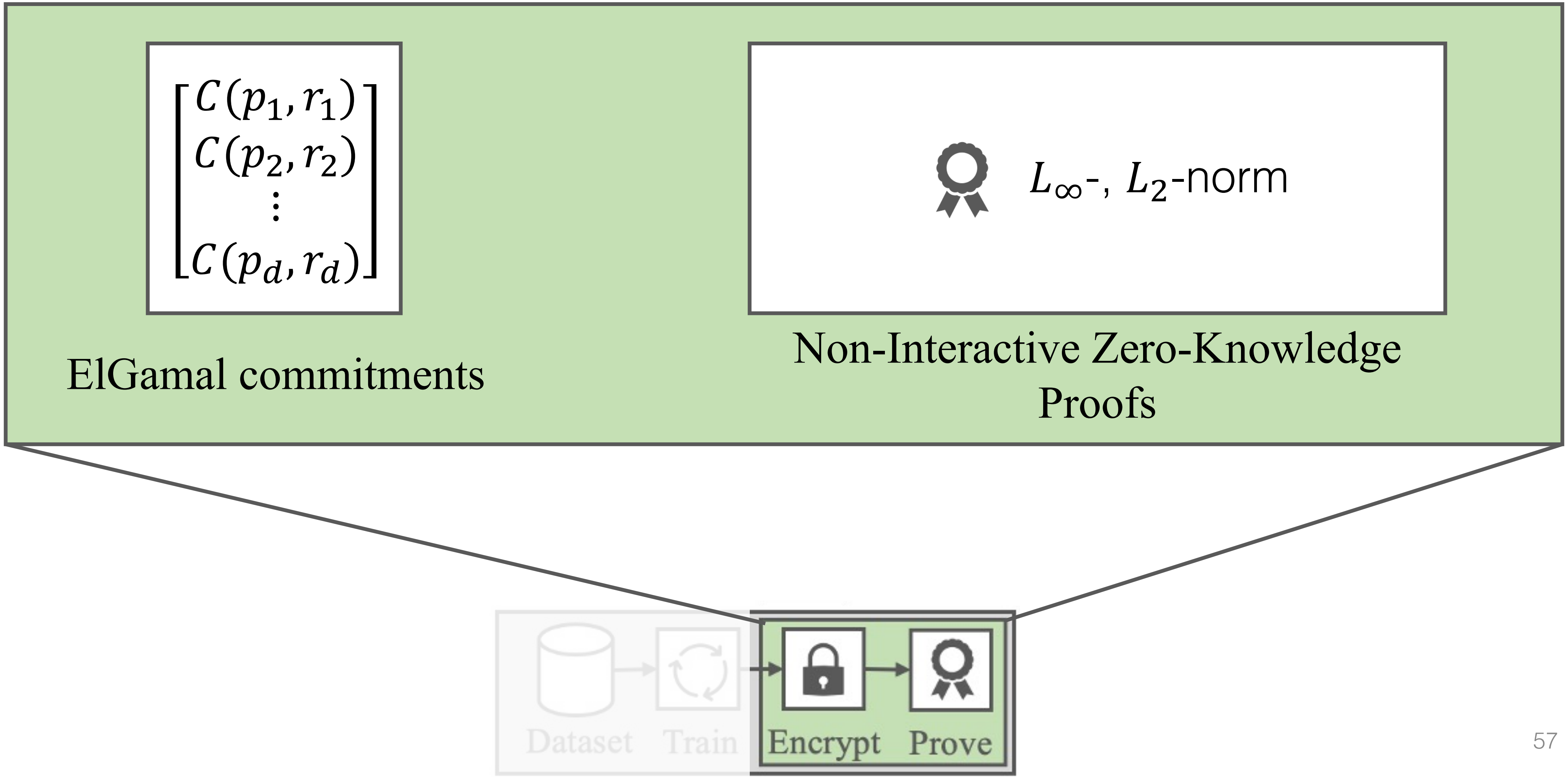
Additively homomorphic commitments

$$C(m_1, r_1) \oplus C(m_2, r_2) = C(m_1 + m_2, r_1 + r_2)$$



$$C\left(\sum \Delta w_i, \sum s_i\right) = C\left(\sum \Delta w_i, 0\right)$$

Zero-knowledge Proofs for Norm Constraints



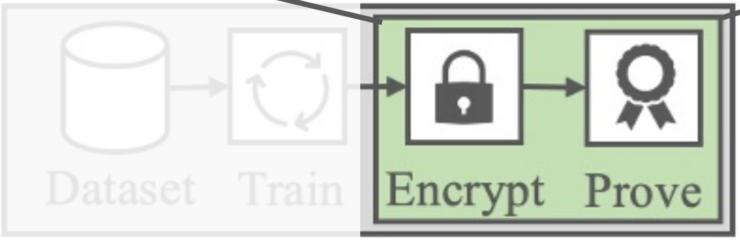
Enforcing L_∞-norm

$$\begin{bmatrix} C(p_1, r_1) \\ C(p_2, r_2) \\ \vdots \\ C(p_d, r_d) \end{bmatrix}$$

ElGamal commitments

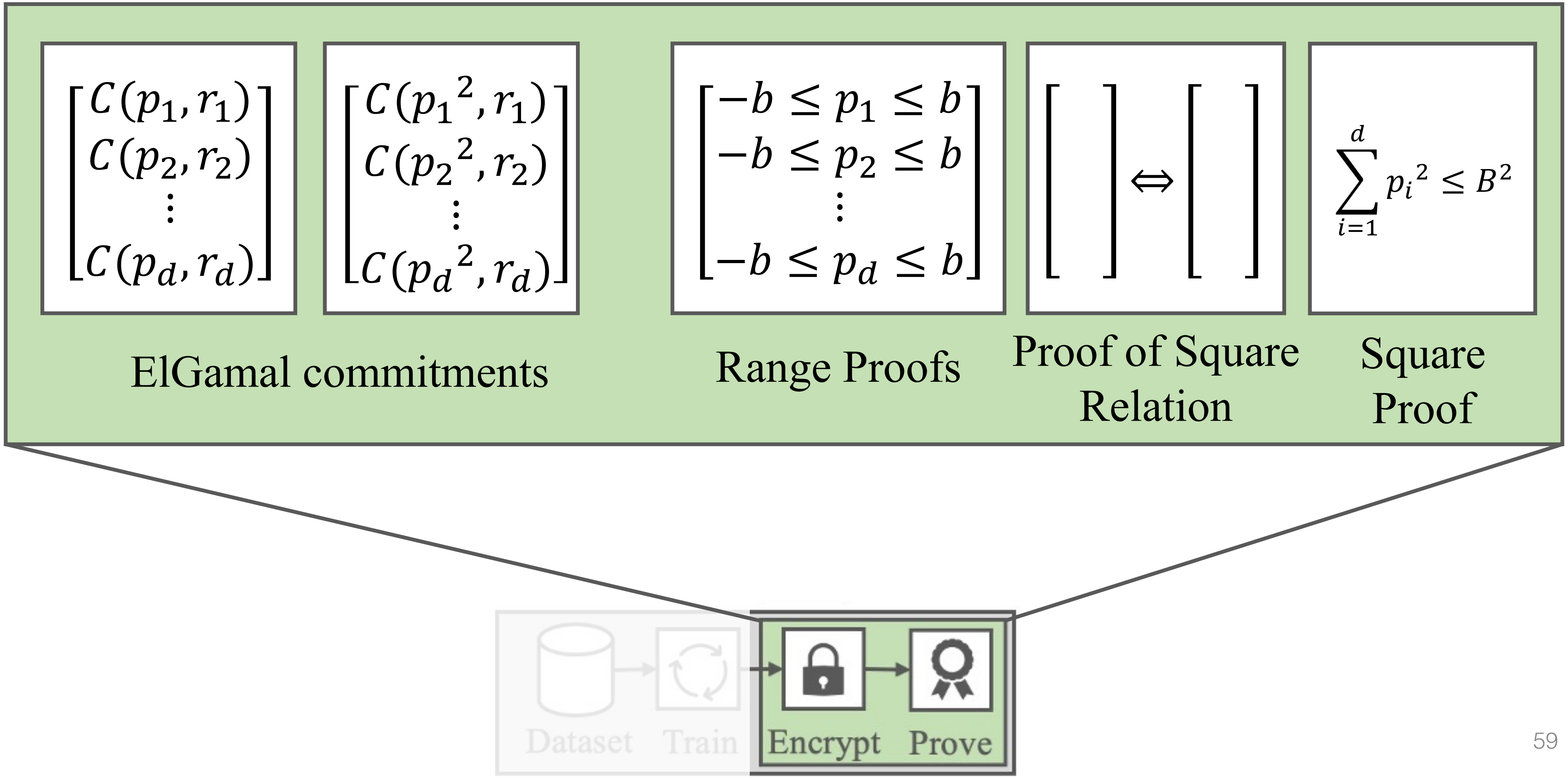
$$\begin{bmatrix} -b \leq p_1 \leq b \\ -b \leq p_2 \leq b \\ \vdots \\ -b \leq p_d \leq b \end{bmatrix}$$

Bulletproof Range Proofs



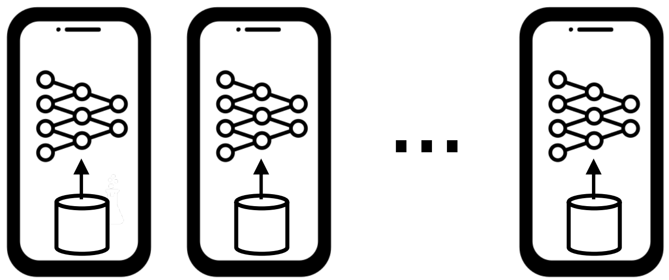
Bünz et al., *Bulletproofs: Short Proofs for Confidential Transactions and More*, S&P 2018.

Enforcing L_2-norm



Problem: Scalability

$$\left.\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_d \end{bmatrix}\right\} > 100\text{k}$$

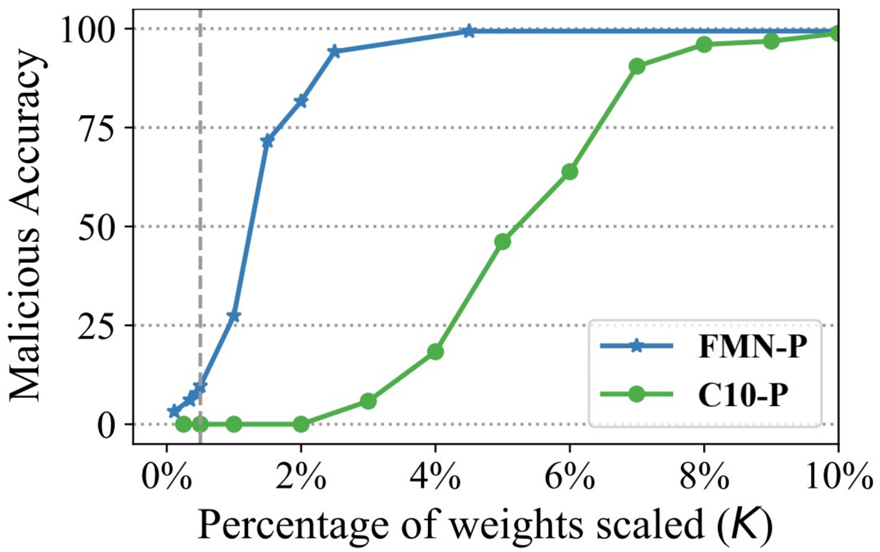


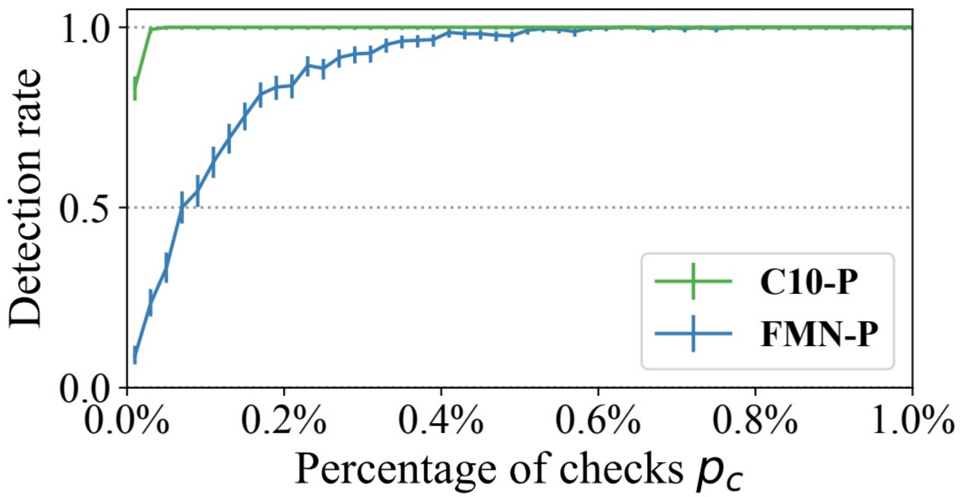
High-dimensional updates

Number of clients

Can we reduce the number of proofs while maintaining the same level of security?

Optimizing L_∞

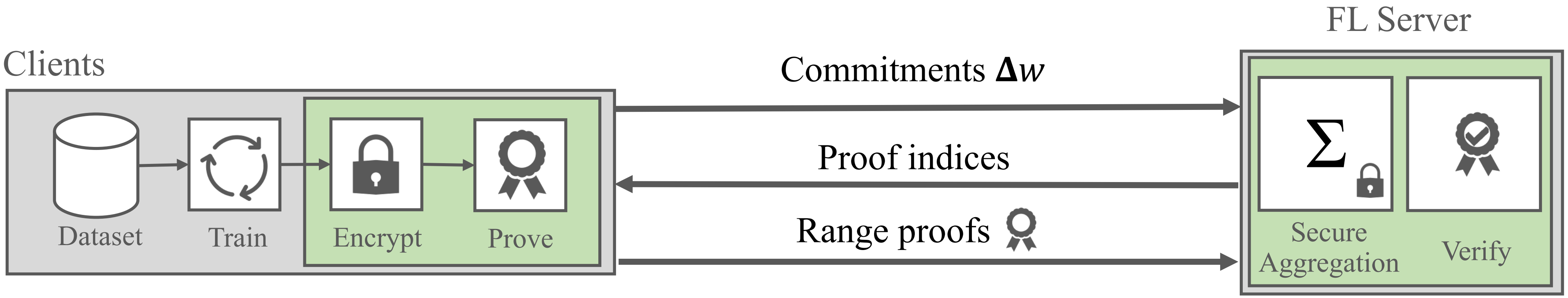


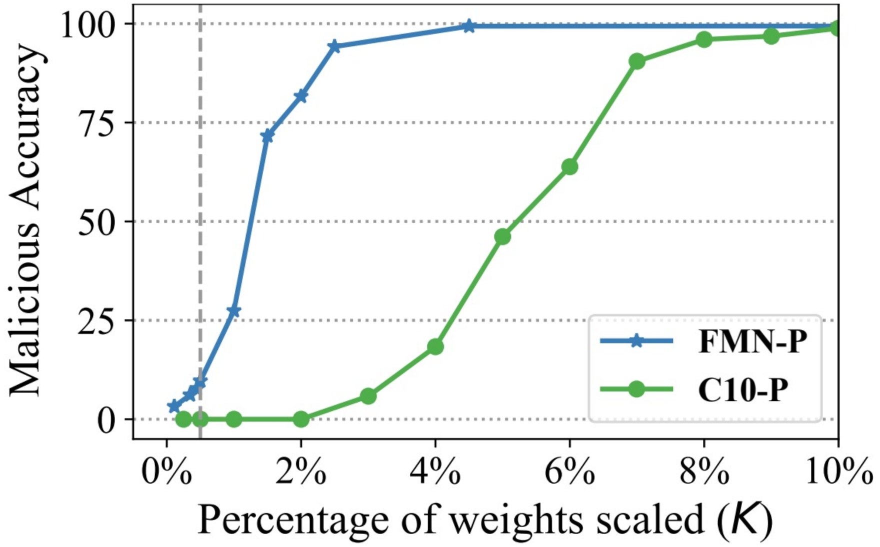


Number of scaled weights required

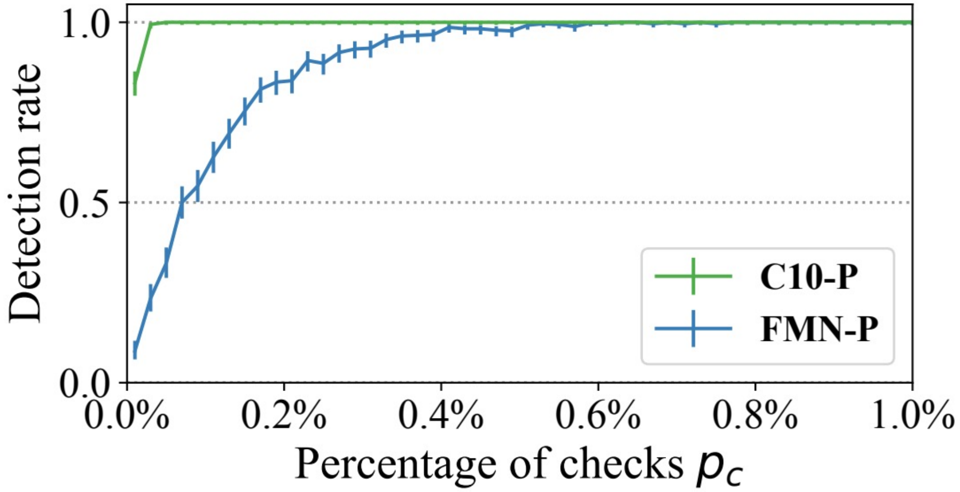
Number of required checks

Optimizing L_∞



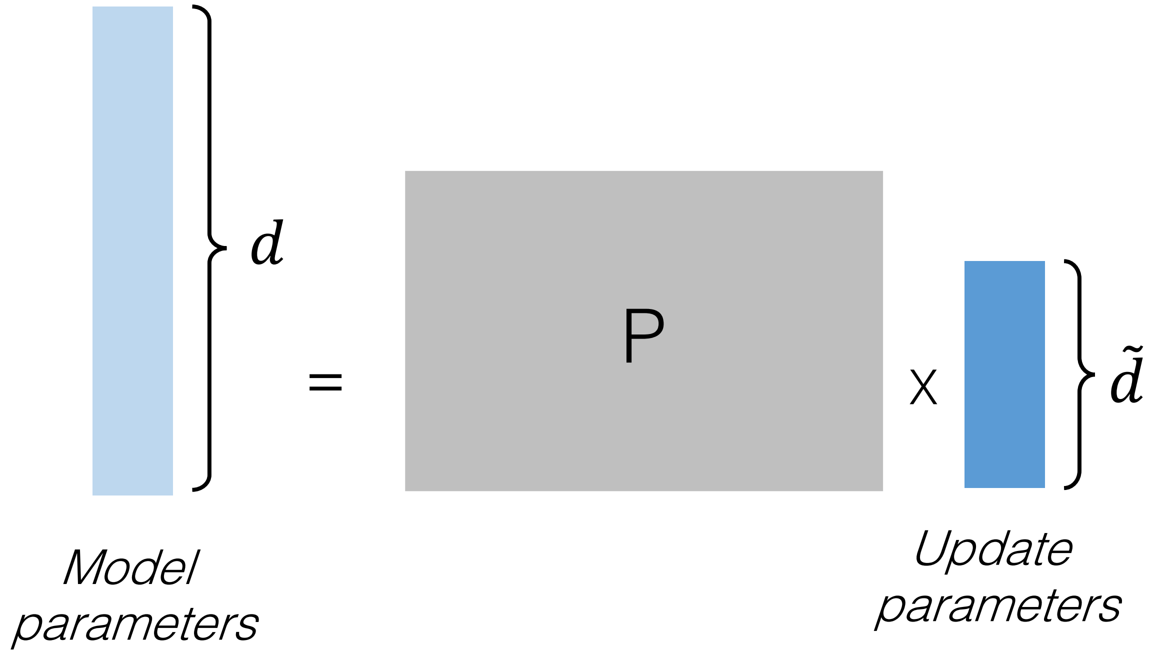


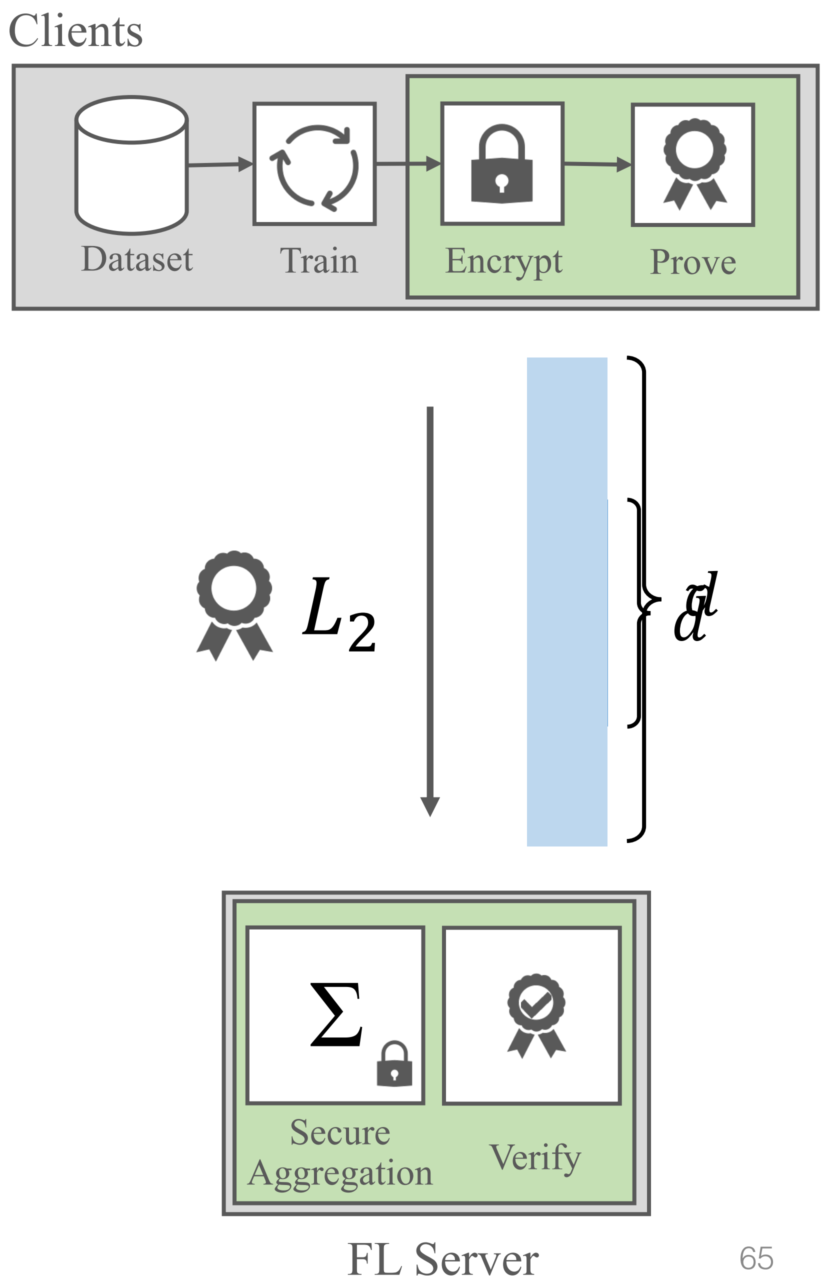
Number of scaled weights required



Number of required checks

Optimizing L_2



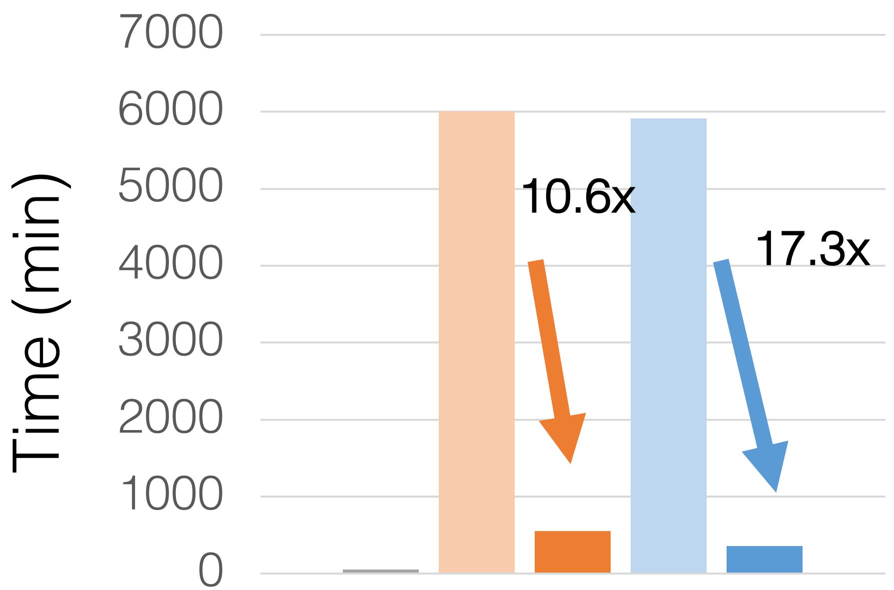


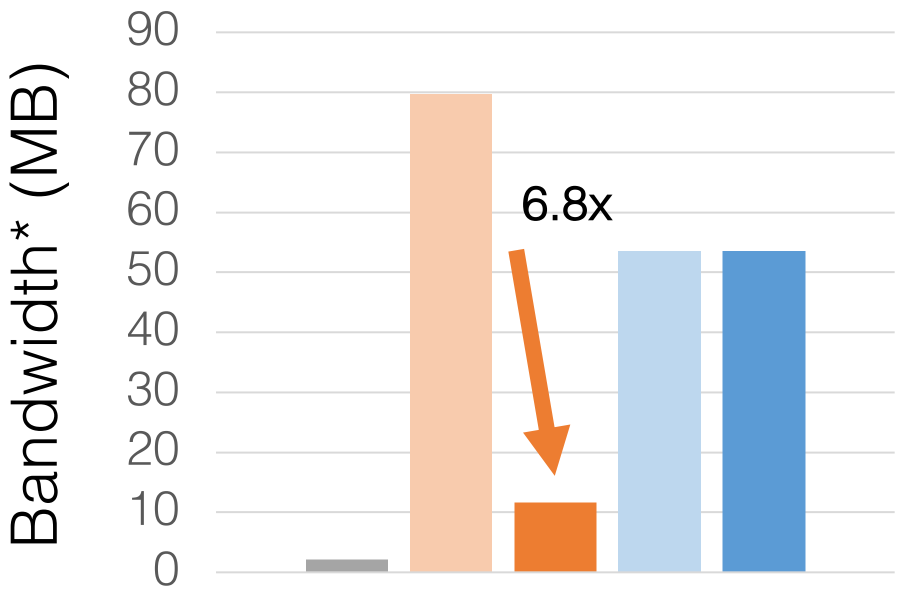
Li et al., *Measuring the Intrinsic Dimension of Objective Landscapes*, ICLR 2018.

RoFL: End-To-End Performance

CIFAR-10 Model 270k Parameters Setup: 48 Clients, 160 rounds

Plaintext	L_2	L_2 Optimized	L_∞	L_∞ Optimized
Accuracy: 0.86	0.85	0.82	0.85	0.85





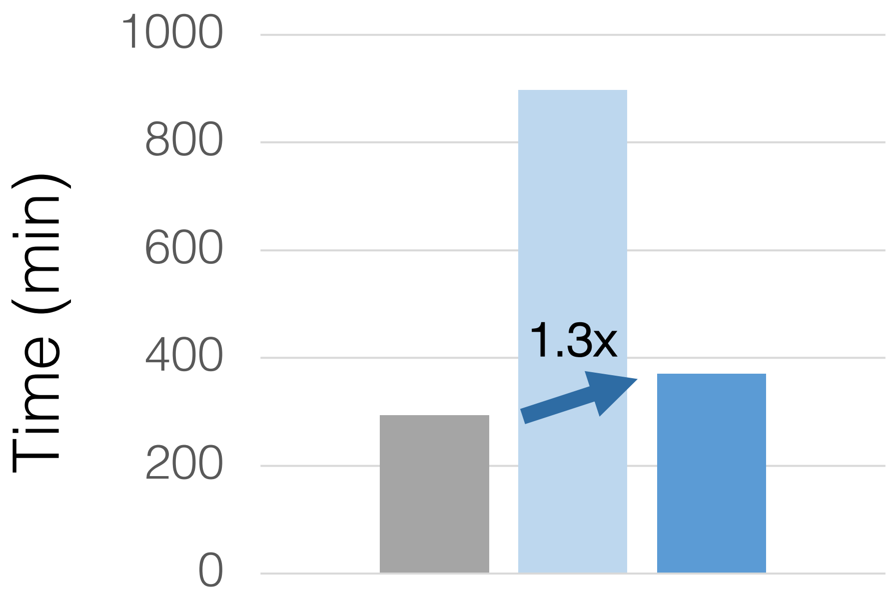
* Per client per round

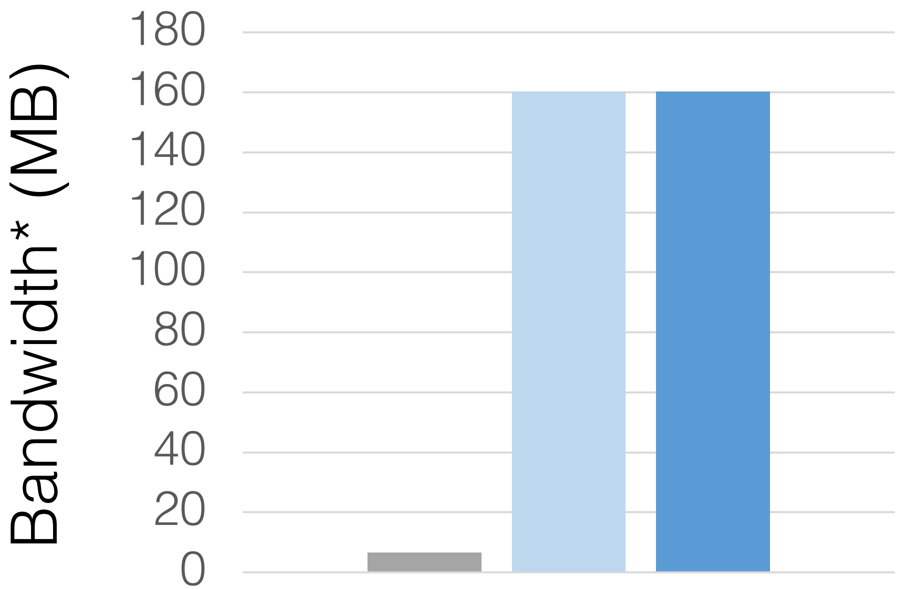
Evaluation: End-To-End

Shakespeare Model 818k Parameters

Setup: 48 Clients, 20 rounds

	Plaintext	L_∞	L_∞ Optimized
Accuracy:	0.57	0.57	0.57





* Per client per round

PPS

pps-lab.com


arXiv:2107.03311v2 [cs.CR] 19 Jul 2021

RoFL: Attestable Robustness for Secure Federated Learning

Lukas Burkhalter, Hidde Lycklama, Alexander Viand, Nicolas Küchler, Anwar Hithnawi

ETH Zürich

Abstract

1 Introduction

https://arxiv.org/pdf/2107.03311.pdf (Preprint)

Analysis Code: https://github.com/pps-lab/fl-analysis

RoFL Code: https://github.com/pps-lab/rofl-project-code