RUHR-UNIVERSITÄT BOCHUM

INTRO TO FL (SECURITY)

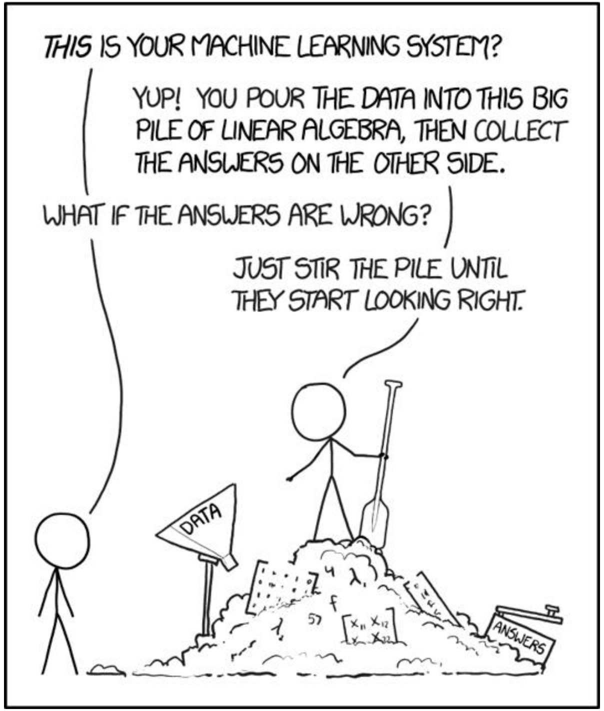
Prof. Dr. Ghassan Karame / M. Sc. Pascal Zimmer / M. Sc. Simon Lachnit

June 24, 2025

Recap: Traditional ML

(Supervised) Machine Learning 101:

1. Take a huge amount of labeled *data*
2. Use a flexible *model*
3. Invest a lot of computing power to
    1. Create predictions on the data
    2. Compute the *loss*, i.e., the difference between ground-truth labels and predicted labels
    3. Use backpropagation to update the model to make better predictions



Credit: xkcd.com

Limitations of Traditional ML

- **Several breakthroughs in the past few years**
- **Mostly due to the capabilities to train on huge amounts of data and the availability of that data**
- **Problems:**
  - GDPR, CCPA, Privacy Act, …
  - Collection of data not always possible

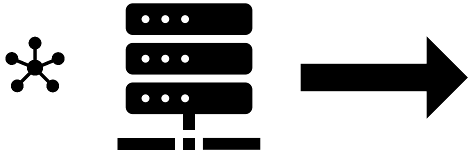
Credit: OpenAI

Credit: tesla.com

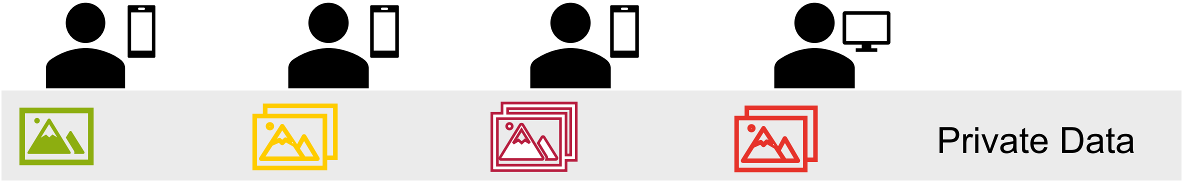
Meet Federated Learning

If we cannot move the data to the model, let's move the model to the data!

In each epoch:



$$= f_{agg}(\{ \qquad \})$$

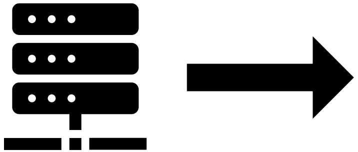


Credit: imgflip.com

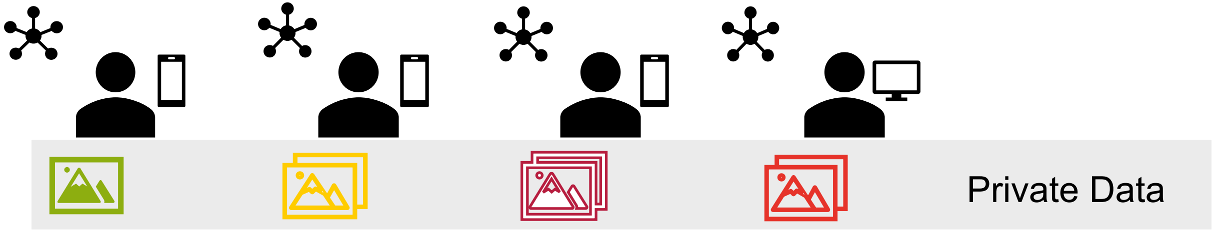
Meet Federated Learning

If we cannot move the data to the model, let’s move the model to the data!

In each epoch:



$= f_{agg}(\{ \qquad \})$

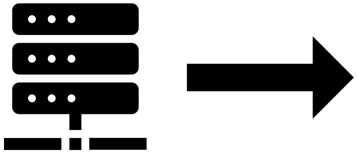


Credit: imgflip.com

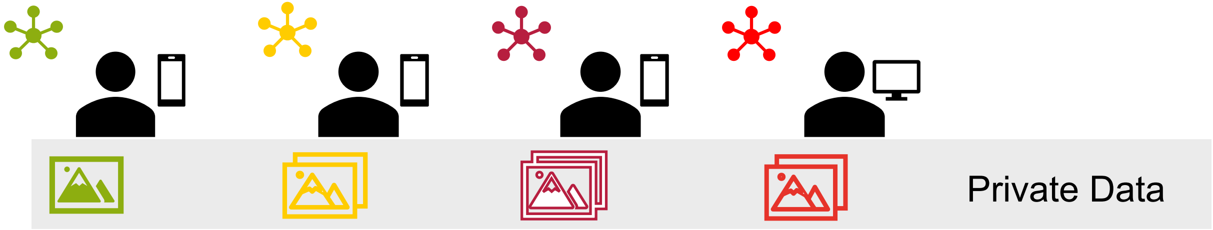
Meet Federated Learning

If we cannot move the data to the model, let's move the model to the data!

In each epoch:



$= f_{agg}(\{ \qquad \})$



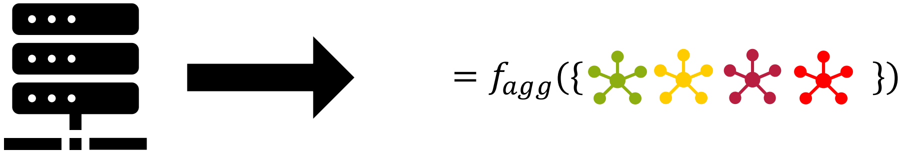
Credit: imgflip.com

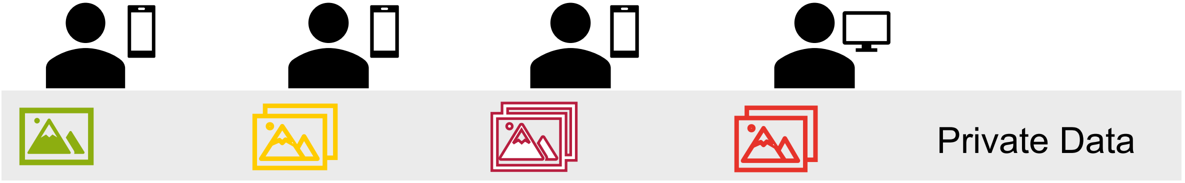
RUHR UNIVERSITÄT BOCHUM

Meet Federated Learning

If we cannot move the data to the model, let's move the model to the data!

In each epoch:



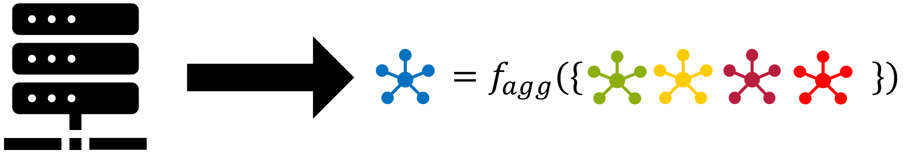


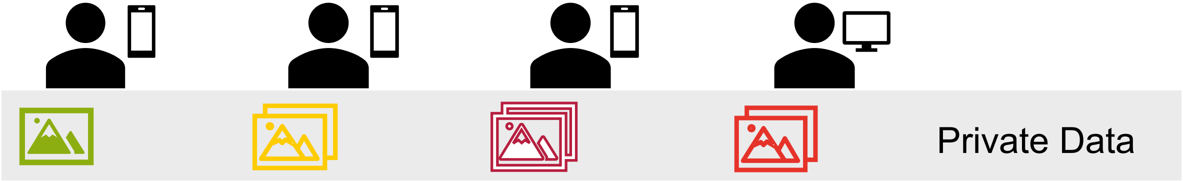
Credit: imgflip.com

Meet Federated Learning

If we cannot move the data to the model, let’s move the model to the data!

In each epoch:



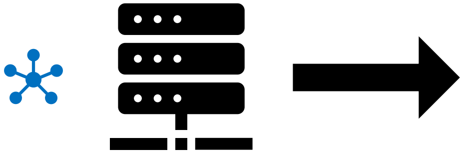


Credit: imgflip.com

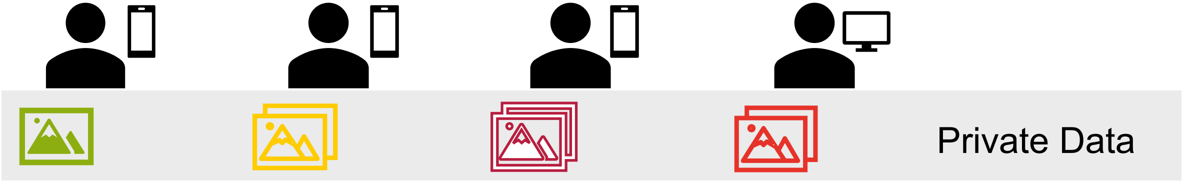
Meet Federated Learning

If we cannot move the data to the model, let’s move the model to the data!

In each epoch:



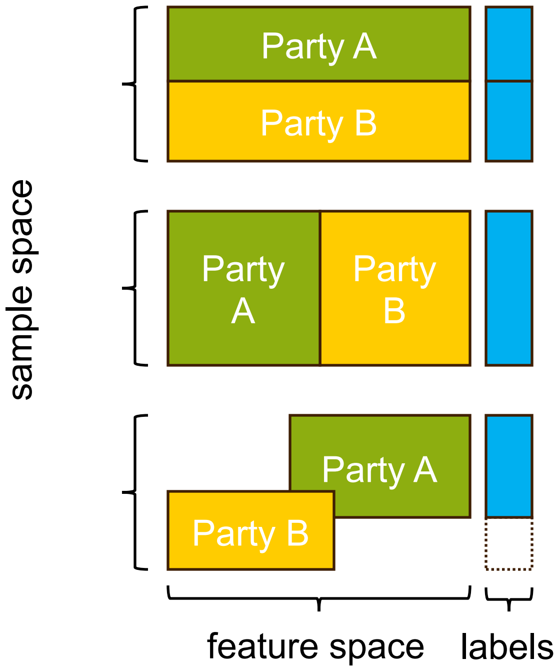
$= f_{agg}(\{ \quad \})$



Credit: imgflip.com

Meet Federated Learning – Some Basics

Types of Federated Learning:



Applications of (horizontal) Federated Learning:

- **Useful if: (McMahan et al., AISTATS, 2017)**
  - Labels on the data can be inferred naturally
  - Data is privacy sensitive and / or large in size
- **Examples:**
  - Next word prediction in smart keyboards (Apple, Google)
  - Siri's keyword recognition

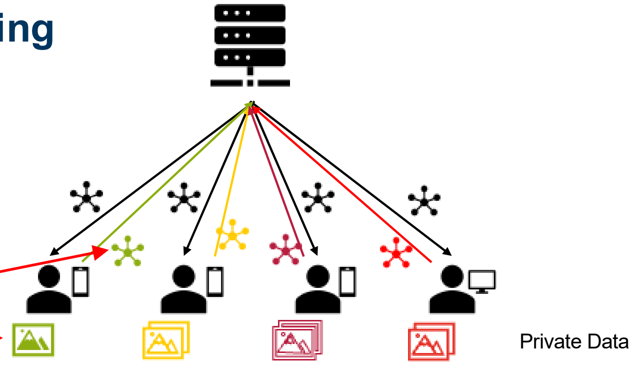
Meet Federated Learning – Some Notation

- **A central server $\boldsymbol{S}$ coordinates $\boldsymbol{N}$ clients with private datasets $\boldsymbol{D_1, D_2, \dots, D_N}$ to train a global model with parameters $\boldsymbol{\theta}$**
- **In each round $\boldsymbol{t}$:**
  - Select $\mathrm{M} < N$ clients for training
  - Send the current global model θ^t to each client
  - Each client trains on its local dataset to obtain a local model with parameters θ_i^{t+1} and sends it back to the server
  - The server aggregates all local updates to obtain a new global model using some aggregation rule: $\theta^{t+1} = f_{agg}(\{\theta_1^{t+1}, \theta_2^{t+1}, \dots, \theta_M^{t+1}\})$
    - Most popular aggregation rule: FedAvg (McMahan et al., AISTATS, 2017)

$$\theta^{t+1} = \sum_{i=1}^{M} \frac{|D_i|}{\sum_{i=1}^{M} |D_i|} \theta_i^{t+1}$$

What could possibly go wrong…

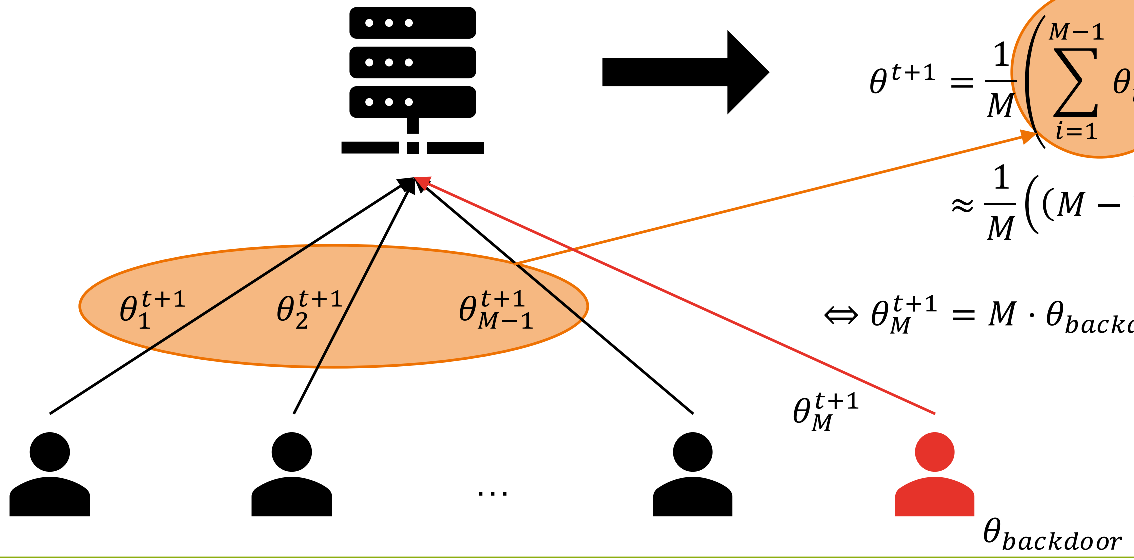
- **As the training takes place on users' devices, we are giving part of the control about the training process to users**
  - As N may be in the orders of millions, it is impossible to guarantee benign clients!
- **Attack vectors:** (Shejwalkar et al., S&P, 2022)
  - (Data Poisoning)
  - Model Poisoning



Credit: imgflip.com

This is not new – What's special about FL?

- **Poisoning training data is not a new idea BUT in FL the threat model differs**
- **Example: How could a single malicious client compromise the whole model? (Bagdasaryan et al., AISTATS, 2020)**



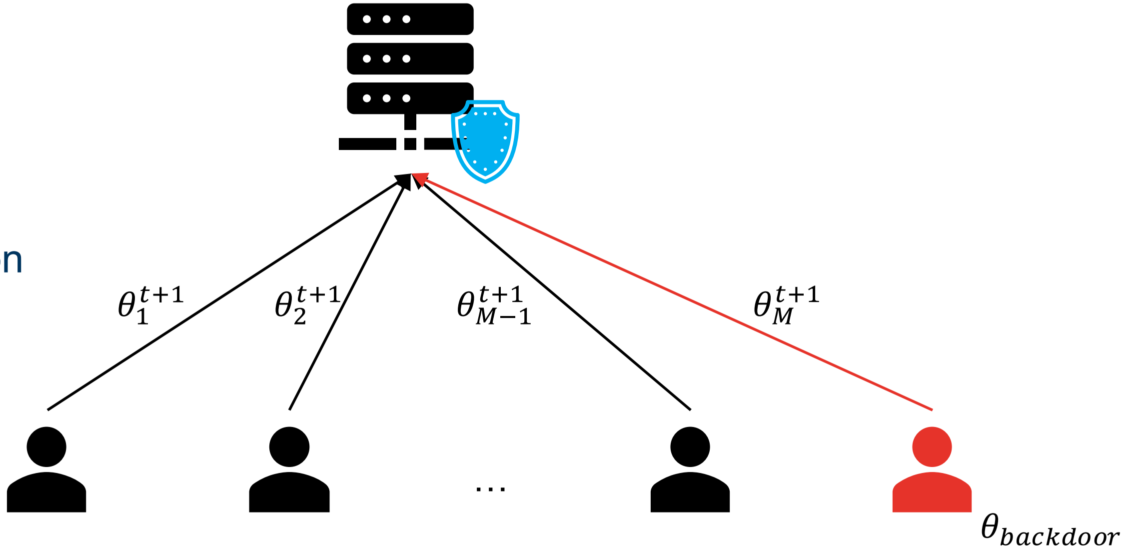
$$\theta^{t+1} = \frac{1}{M}\left(\sum_{i=1}^{M-1} \theta_i^{t+1} + \theta_M^{t+1}\right)$$

$$\approx \frac{1}{M}\left((M-1)\cdot\theta^t + \theta_M^{t+1}\right) = \theta_{backdoor}$$

$$\Leftrightarrow \theta_M^{t+1} = M \cdot \theta_{backdoor} - (M-1)\cdot\theta^t$$

And what to do about it?

- **Defenses against backdoor attacks are mostly realized as variations of the server-side aggregation rule**
- **Examples are based on**
  - Update Clustering
  - Parameter-wise outlier detection
  - Norm analysis
  - …



Well, that escalated quickly…

A3FL: Adversarially Adaptive Backdoor Attacks to Federated Learning

Part of Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track

Chameleon: Adapting to Peer Images for Planting Durable Backdoors in Federated Learning

Yanbo Dai, Songze Li *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:6712-6725, 2023.

2023 IEEE Symposium on Security and Privacy (SP)

3DFed: Adaptive and Extensible Framework for Covert Backdoor Attack in Federated Learning

The Hidden Vulnerability of Distributed Learning in Byzantium

El Mahdi El Mhamdi, Rachid Guerraoui, Sébastien Rouault *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80:3521-3530, 2018.

Home / Archives / Vol. 38 No. 19: AAAI-24 Special Track Safe, Robust and Responsible AI Track / AAAI Technical Track on Safe, Robust and Responsible AI Track

Beyond Traditional Threats: A Persistent Backdoor Attack on Federated Learning

RAID 2020 PROCEEDINGS A USENIX Publication PROCEEDINGS

The Limitations of Federated Learning in Sybil Settings

Authors:
Clement Fung, *Carnegie Mellon University;* Chris J. M. Yoon and Ivan Beschastnikh, *University of British Columbia*

IBA: Towards Irreversible Backdoor Attacks in Federated Learning

Part of Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track

Questions?

Now: Let’s get our hands dirty!