Week 12b - Privacy Preservation

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University

Wide Industry Adoption of FL

Banking

Crowd-sourcing

Insurance

Healthcare

E-Commerce

Safety Monitoring

IoT Infrastructure

Recommender Systems

Google

Swiss Re Group

OpenMined

WeBank 微众银行

Microsoft

HUAWEI

平安科技

華控清交 TSINGJIAO INFORMATION SCIENCE

SINOVATION VENTURES

NVIDIA

阿里巴巴 Alibaba.com

intel

京东数科

蚂蚁金服 ANT FINANCIAL

IBM

Tencent

4Paradigm 第四范式

DATA REPUBLIC

sharemind

Federated Learning Limitations

Limitations:

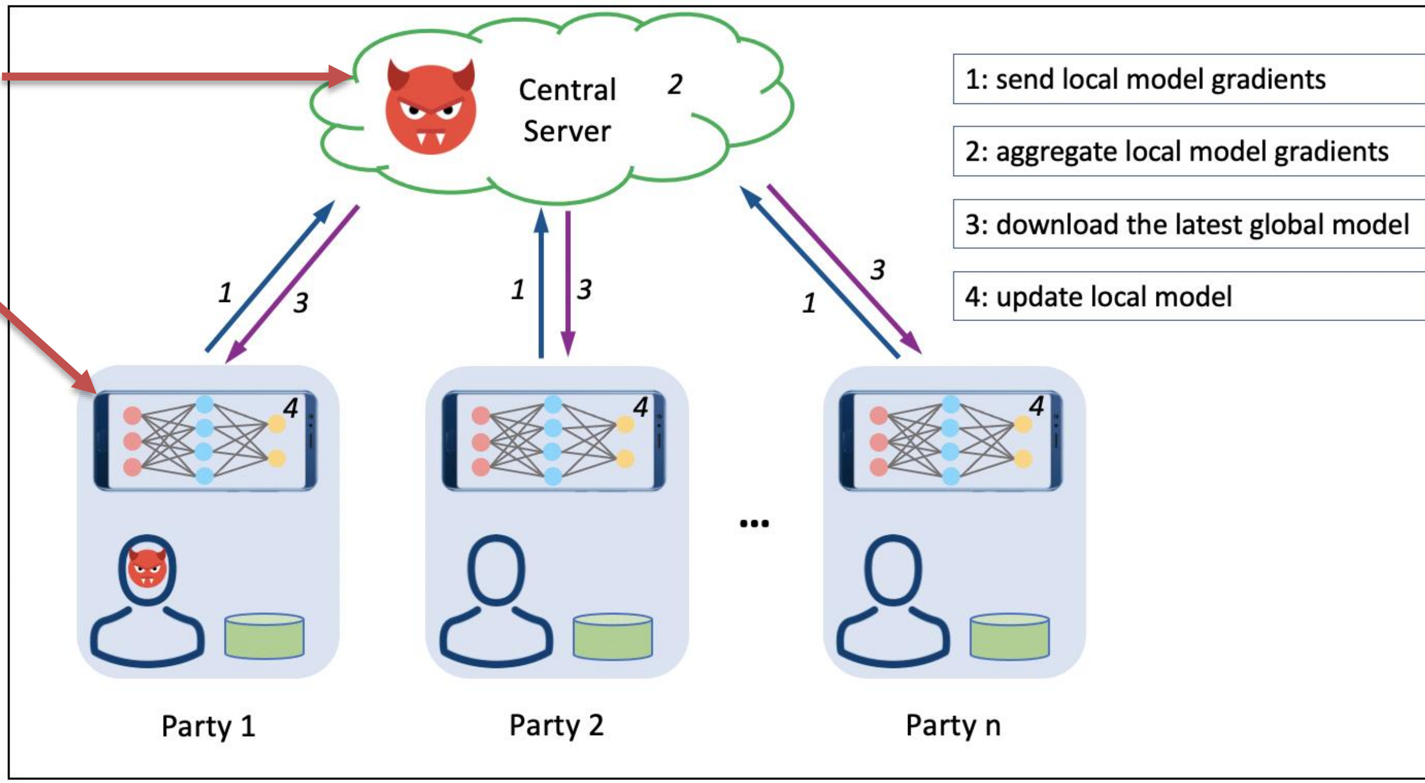
- Single Server FedAvg:
  - Single point of failure
  - Cannot deal with non-i.i.d. situations
  - Vulnerable to “free-riders”

- Exchange of model parameters:
  - Plaintext model parameters vulnerable to privacy attacks
  - Inefficient for large AI models due to high communication cost

Threats to Federated Learning

Mostly Against HFL

Potential Adversaries



Attackers and Threat Models

TABLE I: Taxonomy for horizontal federated learning (HFL).

HFL	Number of Participants	Training Participation	Technical Capability
H2B	small	frequent	high
H2C	large	not frequent	low

Attackers

Outsiders:

- Eavesdroppers on the communication channel.
- Users of the final FL model when it is deployed.

Insiders: FL server and the participants.

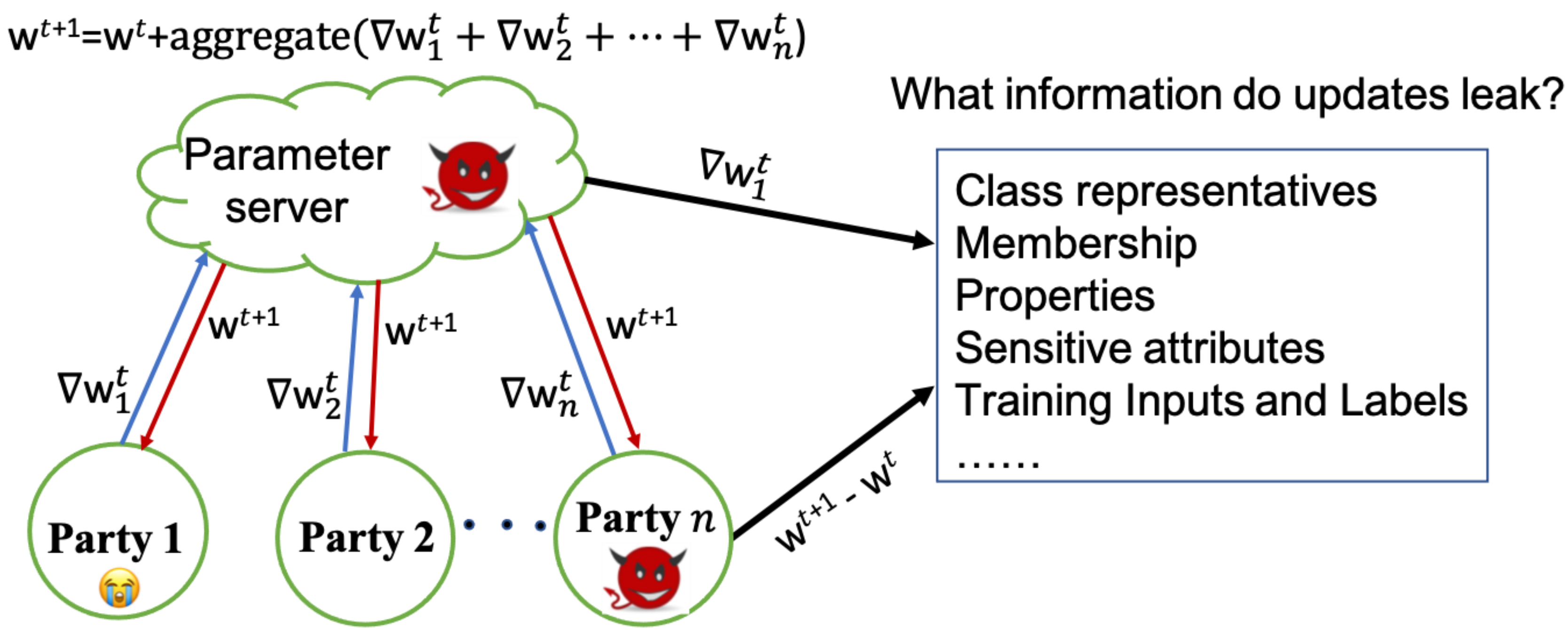
- Byzantine: no need to obey the protocol and can send arbitrary messages to the server.
- Sybil: can simulate multiple dummy participant accounts or select previously compromised participants to mount more powerful attacks on the global model.

Threat Models

Semi-honest: Adversaries are passive or honest-but-curious. They try to learn the private states of other participants without deviating from the FL protocol. The adversaries can only observe the received information.

Malicious: Active, tries to learn the private states of honest participants, and deviates arbitrarily from the FL protocol by modifying, re-playing, or removing messages. This setting allows the adversary to conduct particularly devastating attacks.

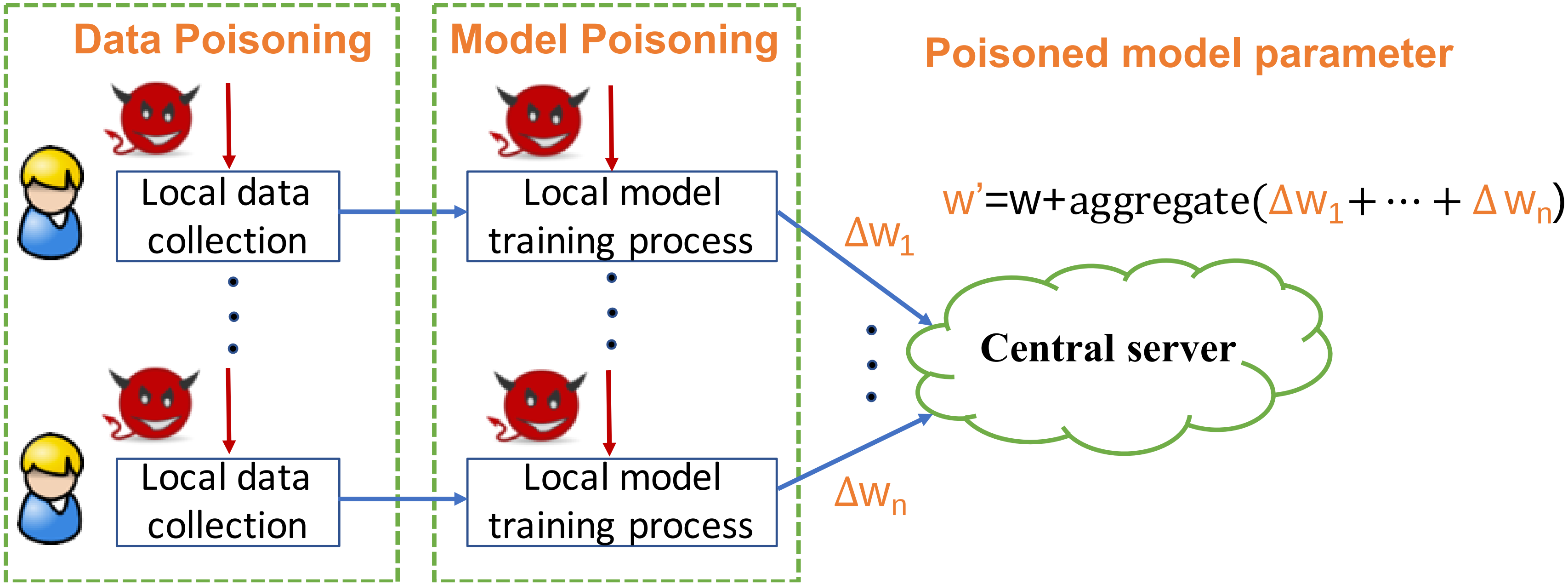
Threats to FL – Inference Attacks



Why gradients cause privacy leakage?

Gradients are derived from the participants' private training data, and a learning model can be considered as a representation of the high-level statistics of the dataset it was trained on.

Threats to FL – Poisoning Attacks



Objective:

1) Targeted (Backdoor, Sybil attacks).
2) Untargeted (Byzantine attacks).

Model poisoning attacks are much more effective than data poisoning attacks!

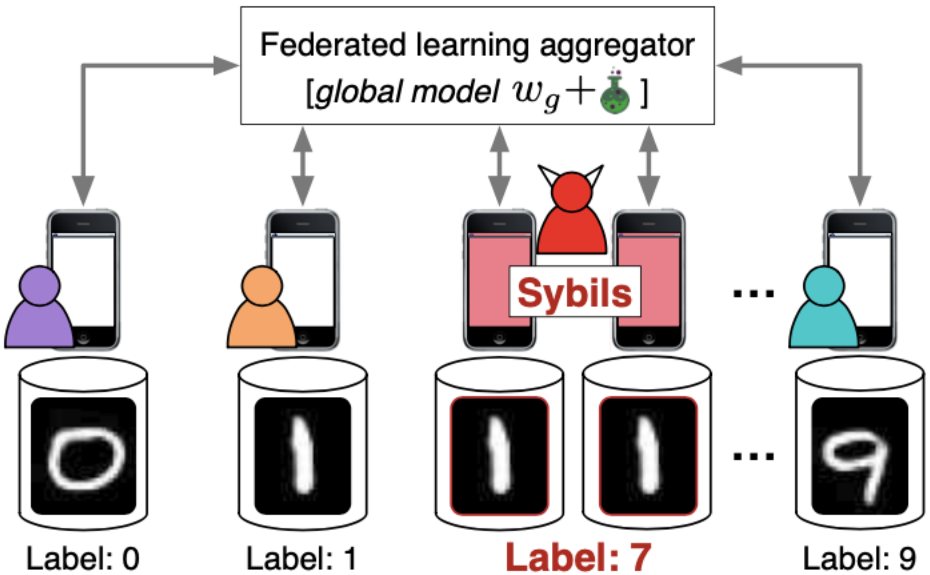
Threats to FL – Poisoning Attacks

- Label-flipping attacks the labels of honest training samples of one class are flipped to another class while the features of the data are kept unchanged.

- Backdoor attacks
- Single features or small regions of the original training dataset are augmented with a secret pattern and relabelled.
- The pattern acts as a trigger for the target class.

- Note: Backdoor attacks should not significantly change the prediction outcomes of other classes. Otherwise, the attack will be detected.

Threats to FL – Sybil Attacks



(b) Federated learning with sybil-based label-flipping poisoning

Sybil attacks

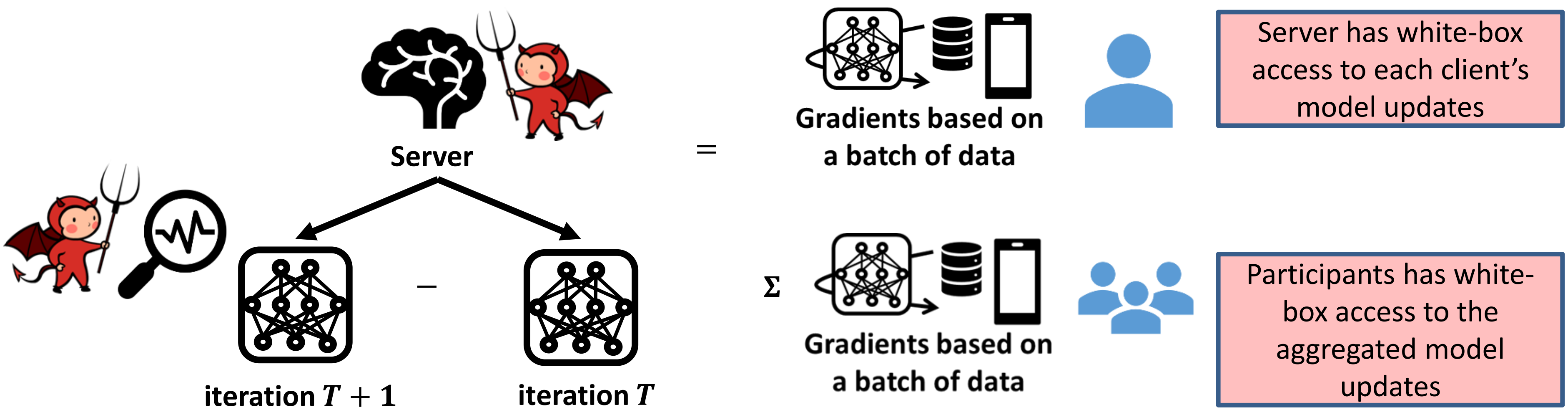
Multiple dummy participant accounts or previously compromised participants launch attacks towards a specific malicious objective.

Source: https://arxiv.org/abs/2012.06337

Defending Federated Learning

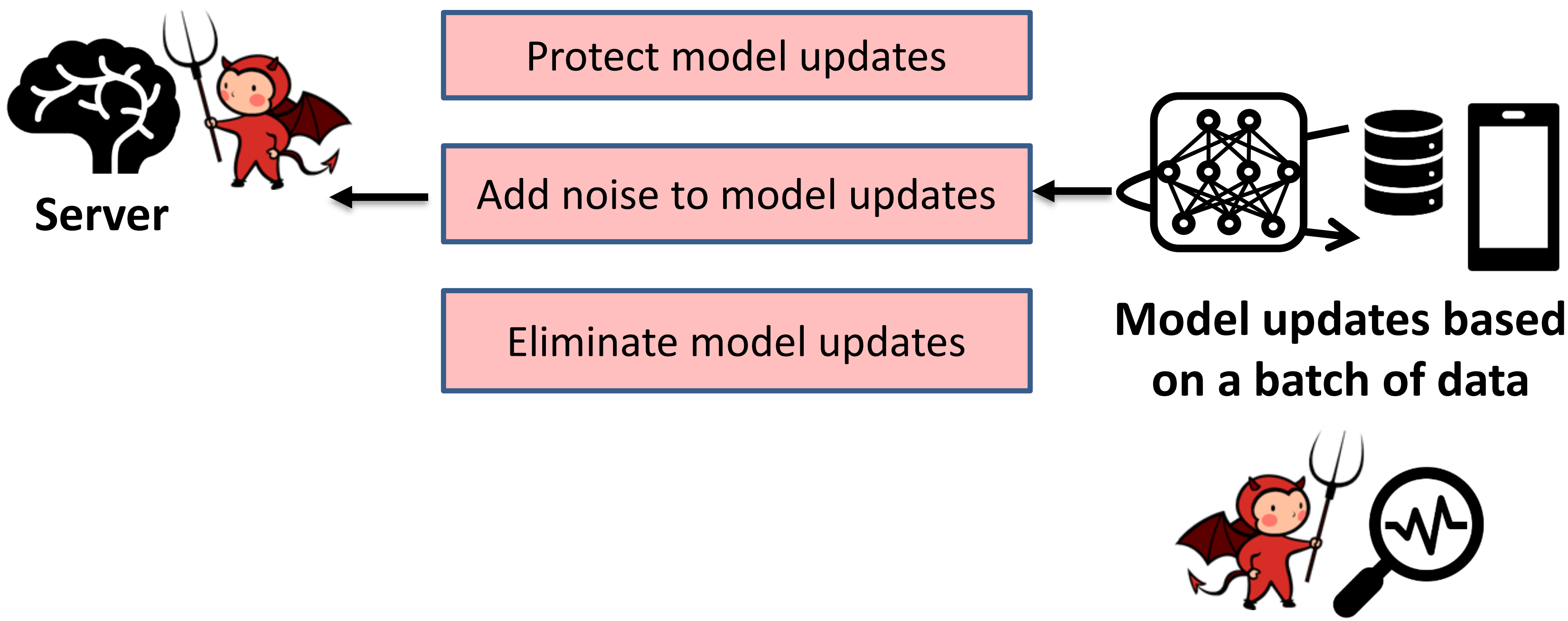
Key FL Vulnerability

- Malicious FL client or server has WHITE-BOX access to model updates



Major Defence Approaches

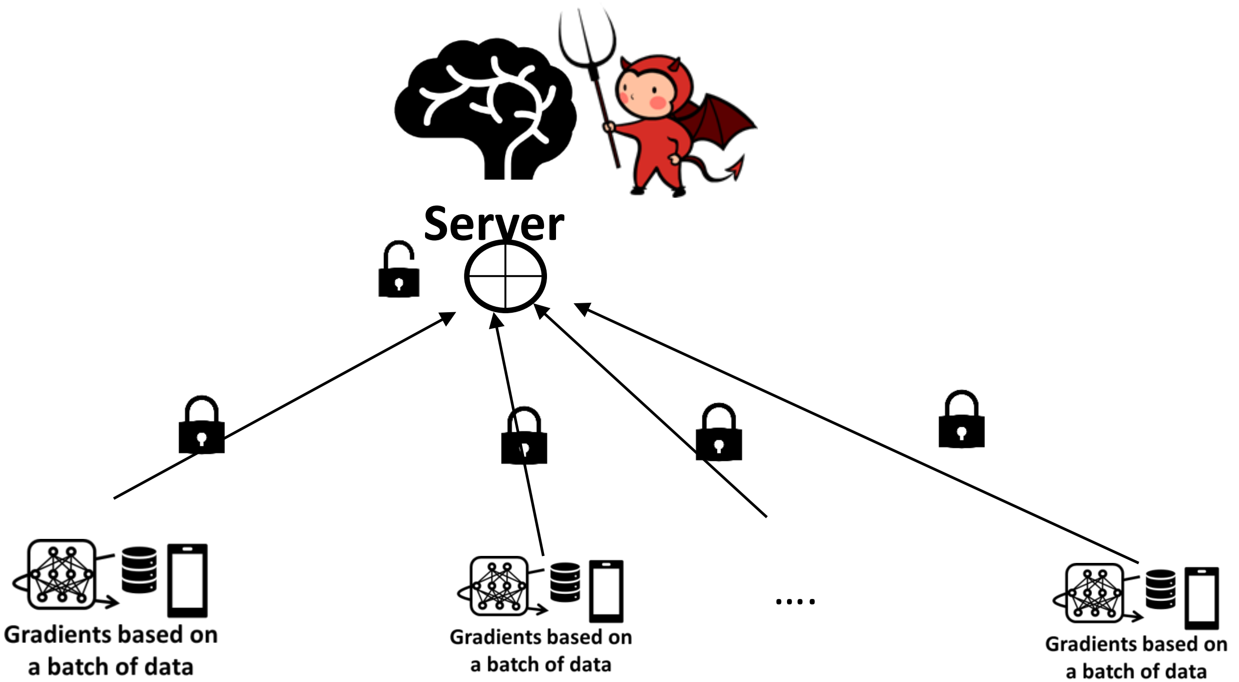
- Keep malicious FL client or server away from raw model updates



https://arxiv.org/abs/2012.06337

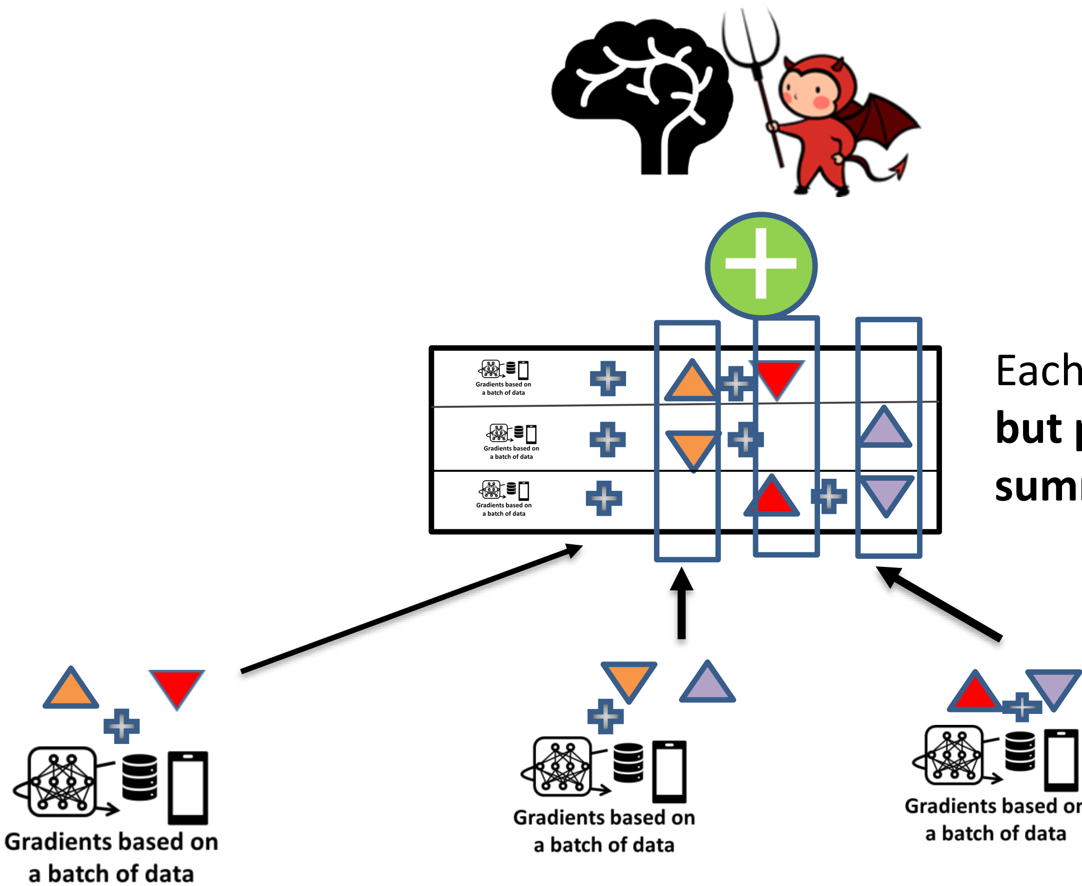
Protecting Model Updates

- Secure Multi-Party Calculation (SMPC)
  - Server aggregates clients' updates,
  - but cannot inspect the individual updates



https://dl.acm.org/doi/10.1145/3133956.3133982

Protecting Model Updates



Each update looks random on its own, **but paired antiparticles cancel out when summed**

Add **antiparticles** (Random positive/negative pairs) before sending to the server.

Protecting Model Updates

- Each client's model update is protected from a malicious server

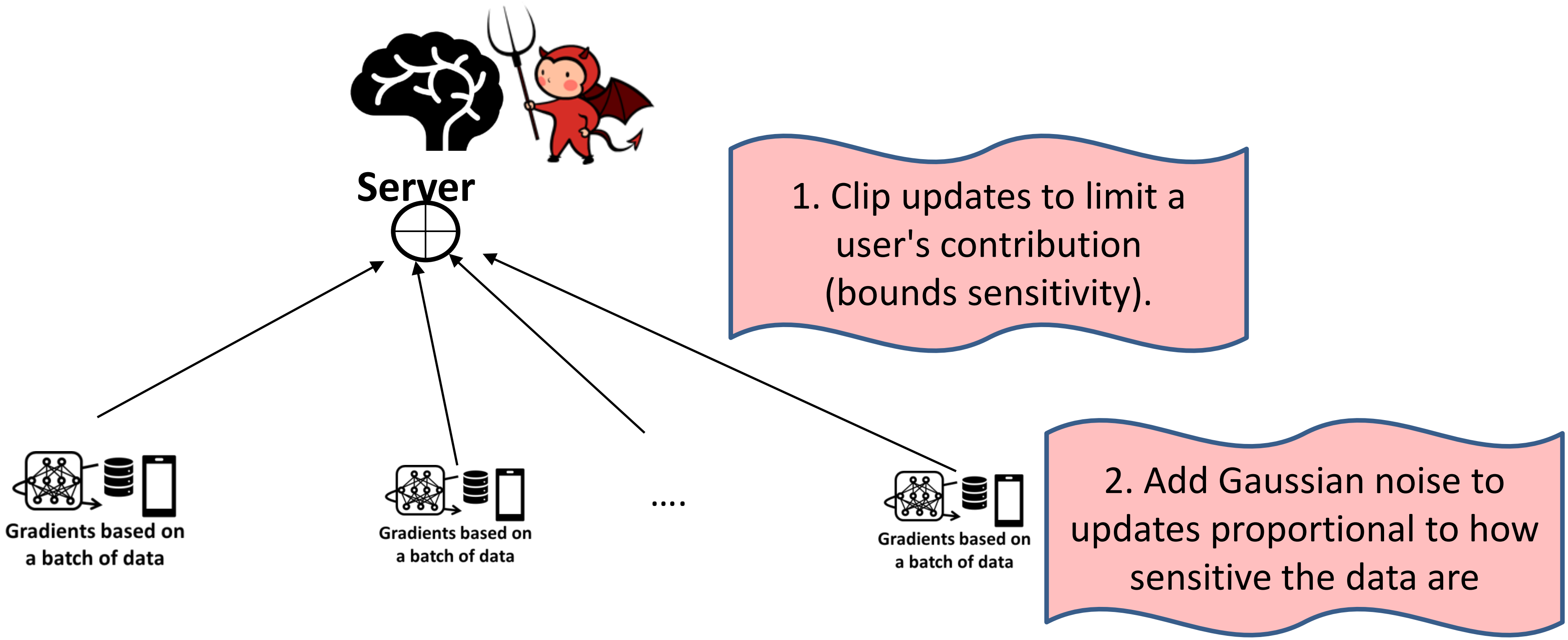
- The aggregated update is NOT protected from malicious clients

Adding Noise to Model Updates

- Differential privacy:
  - the statistical science of trying to learn as much as possible about a group while learning as little as possible about any individual in it.

Adding Noise to Model Updates

- Local differential privacy (LDP)



https://arxiv.org/abs/1712.07557

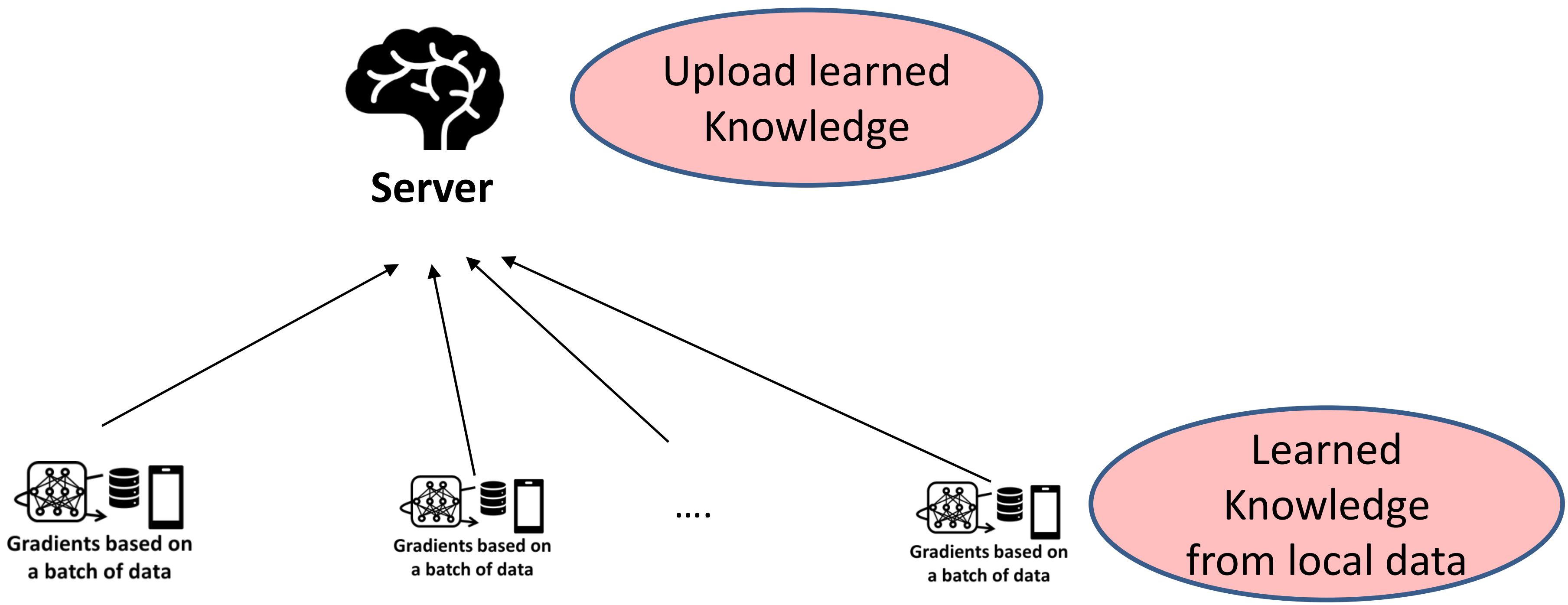
Adding Noise to Model Updates

- Each client's model update is protected from a malicious server
- The aggregated update is also protected from malicious clients
- Protection vs. model performance trade-off must be considered

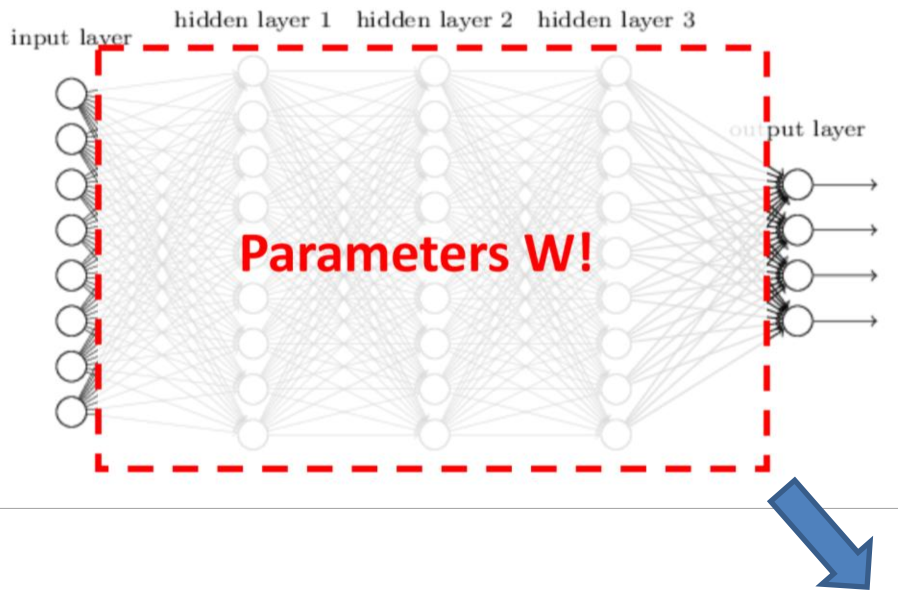
https://arxiv.org/abs/1906.01562

Eliminating Model Updates

- Federated Knowledge Distillation

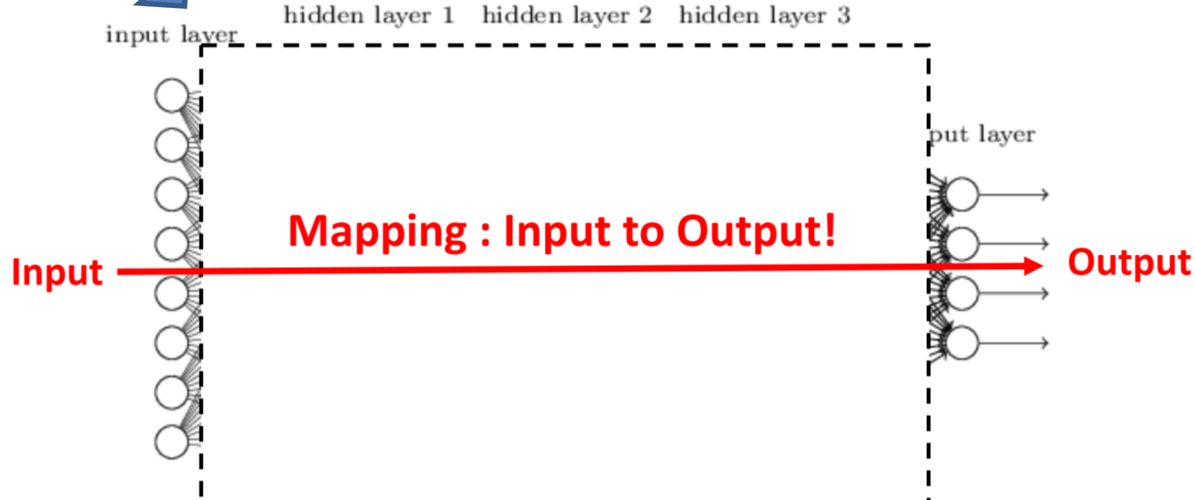


Eliminating Model Updates

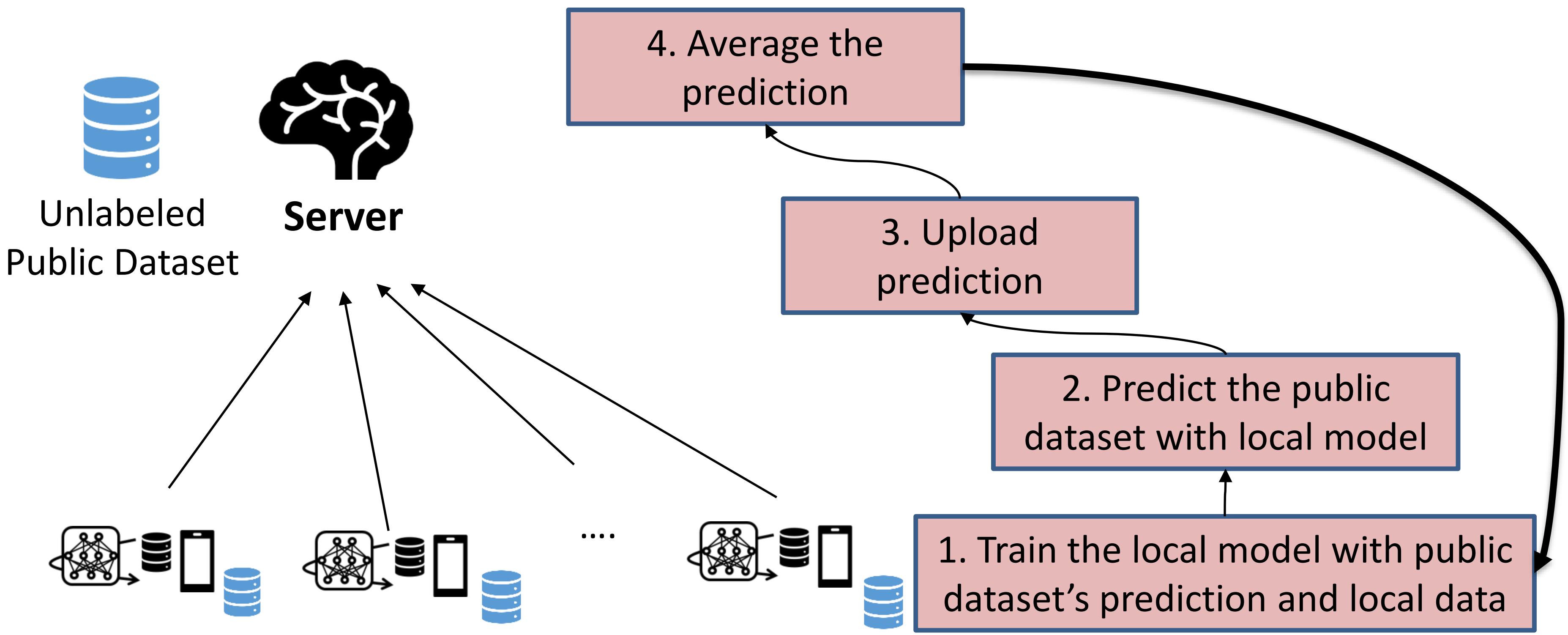


Instead of sending model updates …

Only exchange soft label predictions on a public dataset



Eliminating Model Updates



https://arxiv.org/abs/1912.11279

Eliminating Model Updates

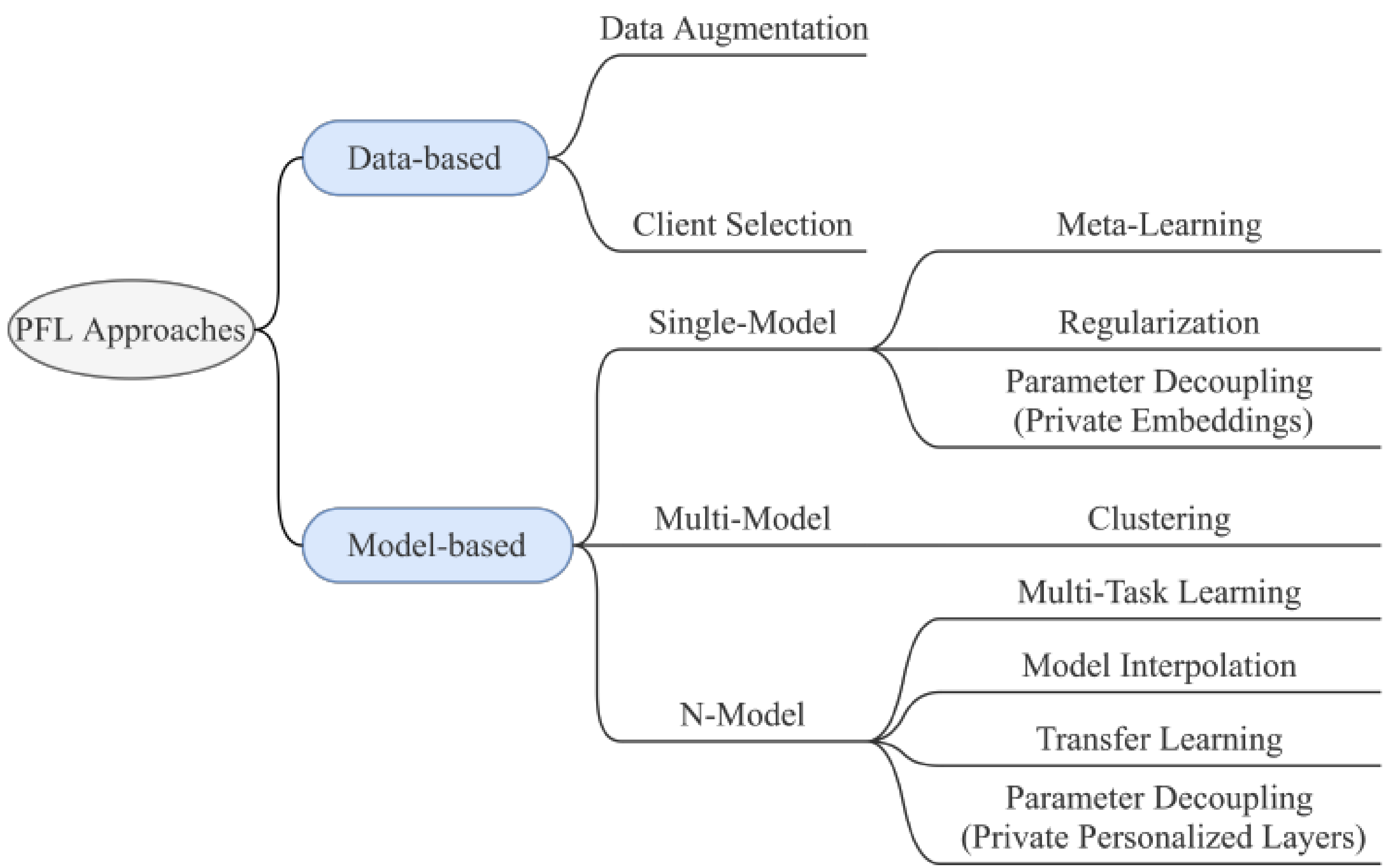
- Each client’s model update is protected from a malicious server

- The aggregated update is also protected from malicious clients

- Slight performance degradation

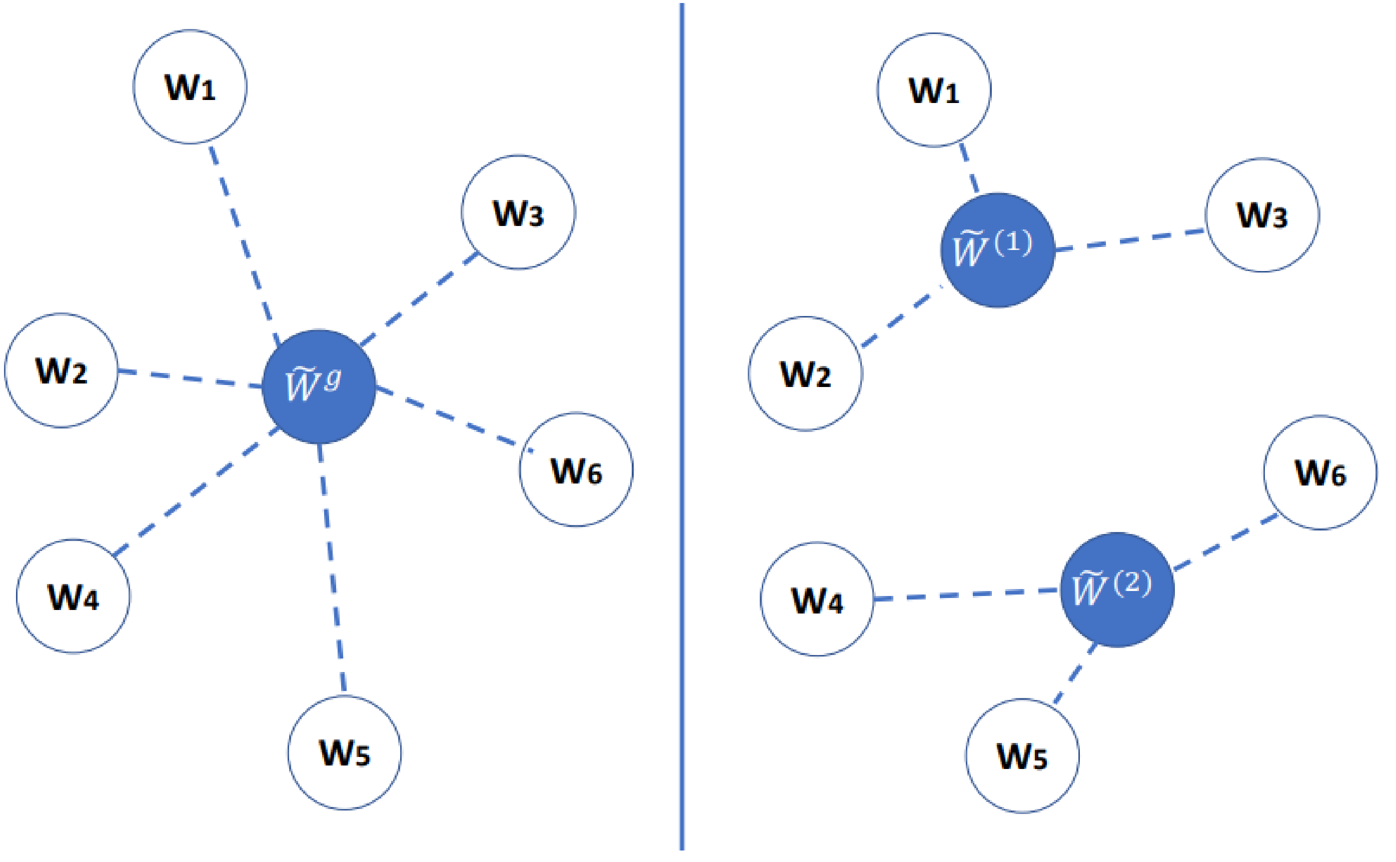
Personalized Federated Learning

Personalized Federated Learning



Source: https://arxiv.org/abs/2103.00710

Data Heterogeneity



Source: https://arxiv.org/abs/2005.01026

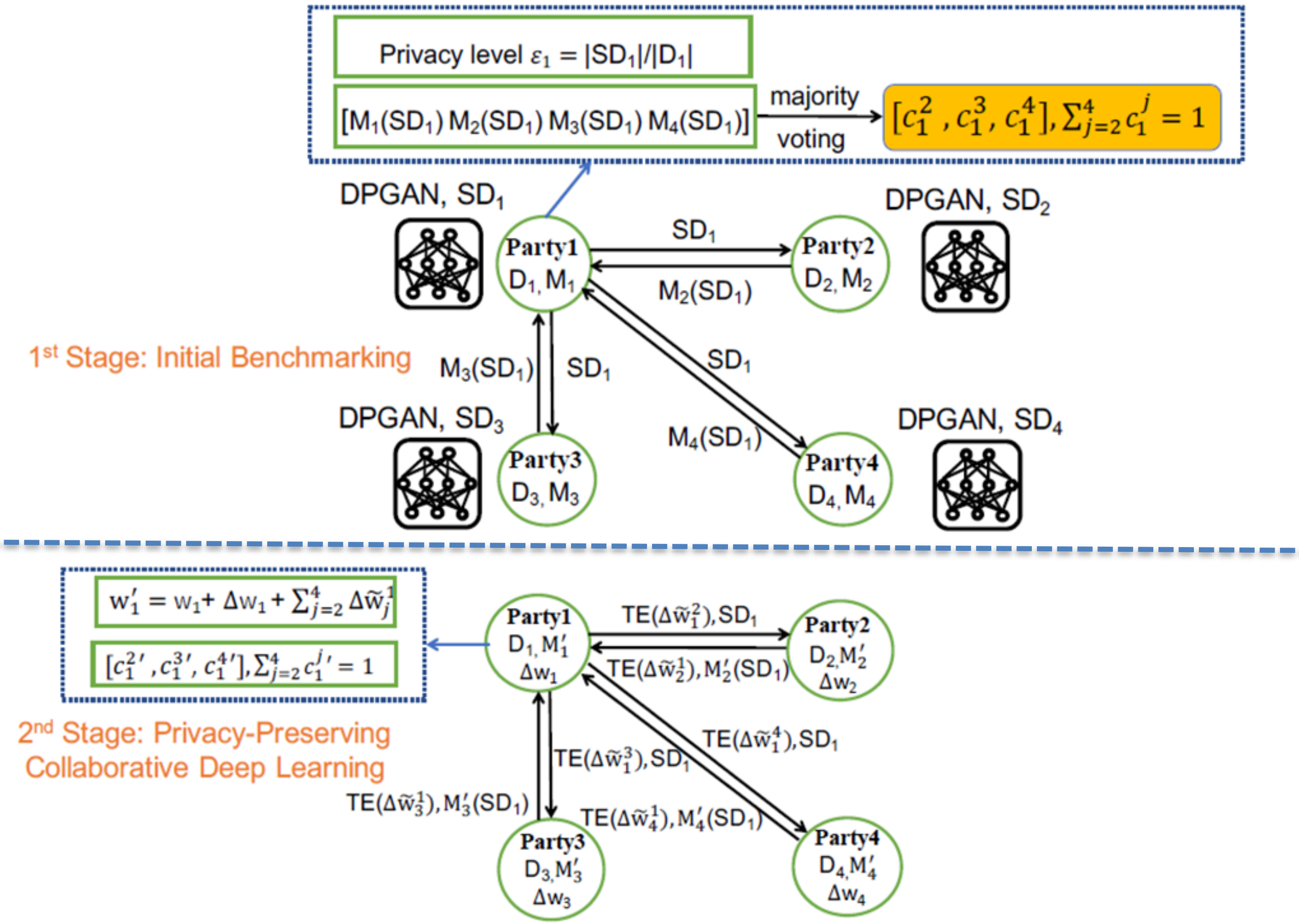
Multi-Center FL

Algorithm 1: FeSEM – Federated Stochastic EM

Initialize K, $\{W_i\}_{i=1}^{m}$, $\{\tilde{W}^{(k)}\}_{k=1}^{K}$
while *stop condition is not satisfied* **do**
  E-Step:
  Calculate distance $d_{ik} \leftarrow \mathrm{Dist}(W_i, \tilde{W}^{(k)})\ \ \forall i, k$
  Update $r_i^{(k)}$ using d_{ik} (Eq. 7)
  M-Step:
  Group devices into C_k using $r_k^{(k)}$
  Update $\tilde{W}^{(k)}$ using $r_i^{(k)}$ and W_i (Eq. 8)
  for *each cluster* $k = 1, \ldots K$ **do**
    for $i \in C_k$ **do**
      Send $\tilde{W}^{(k)}$ to device i
      $W_i \leftarrow$ **Local_update**$(i, \tilde{W}^{(k)})$
    end
  end
end

Federated Stochastic Expectation Maximization (FeSEM)

Model Heterogeneity

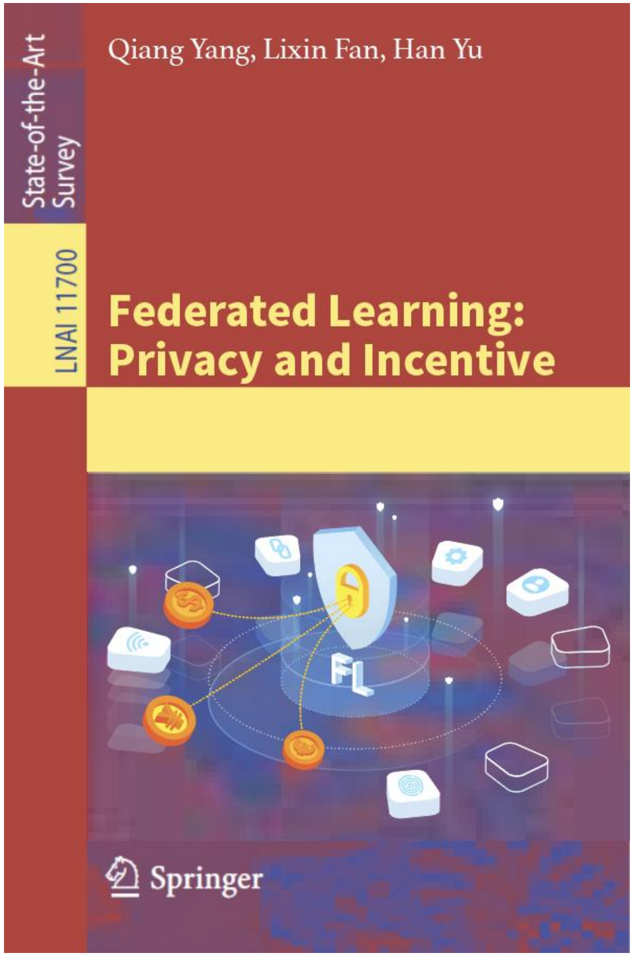


Source: https://ieeexplore.ieee.org/document/9098045

Trust-based Open Collaborative FL

https://youtu.be/d_Z_YmBvmno

Further Reading



Week 12b - Privacy Preservation

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University