Security and Privacy in Machine Learning
Threat Models and Mitigation Measures

Raksha Ramakrishna

Joint work with György Dán

EECS/CS/NSE KTH

ML is expected to become ubiquitous

Communication networks

Smart grids

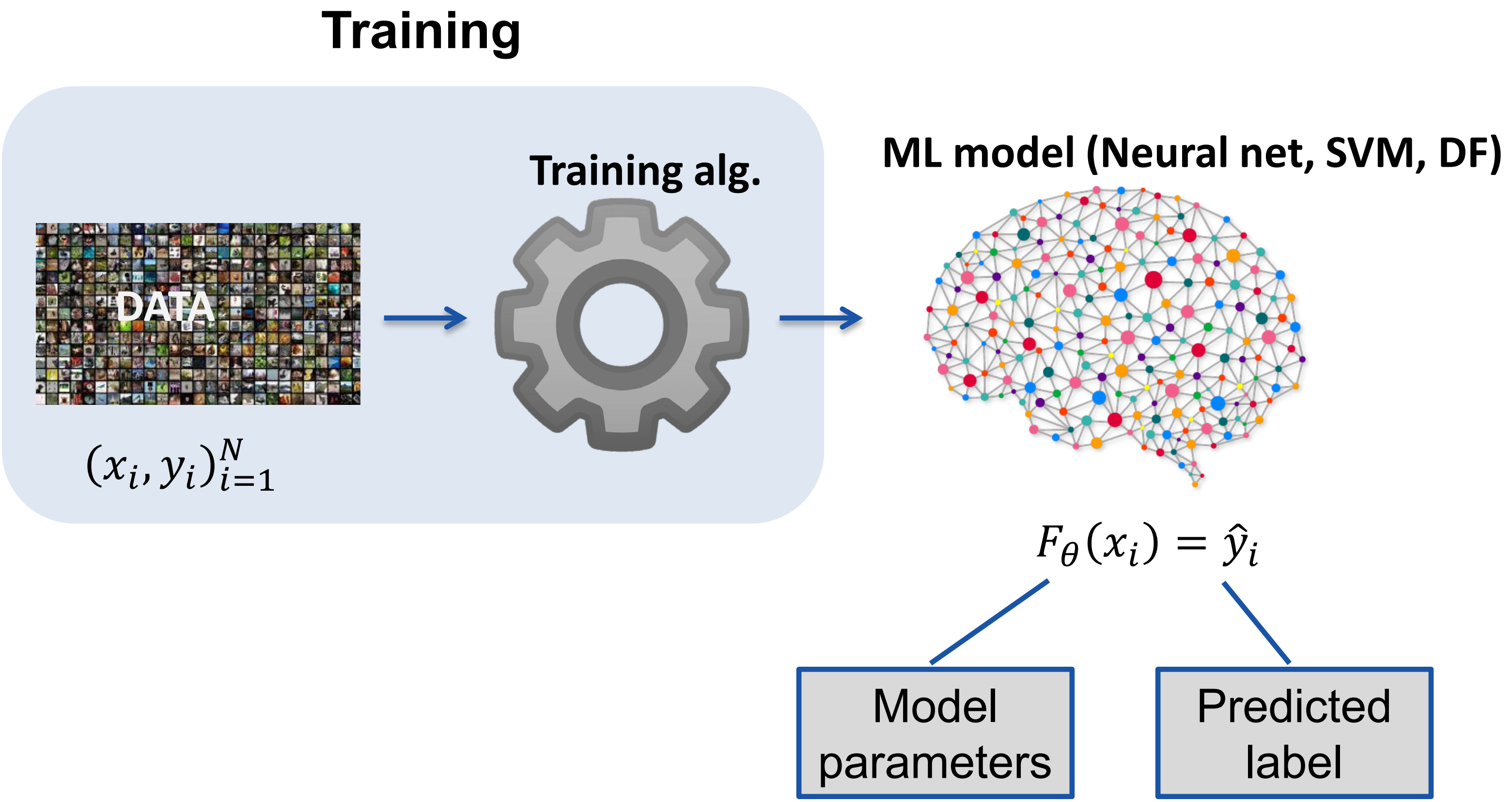
Healthcare

Transportation systems

Smart cities and buildings

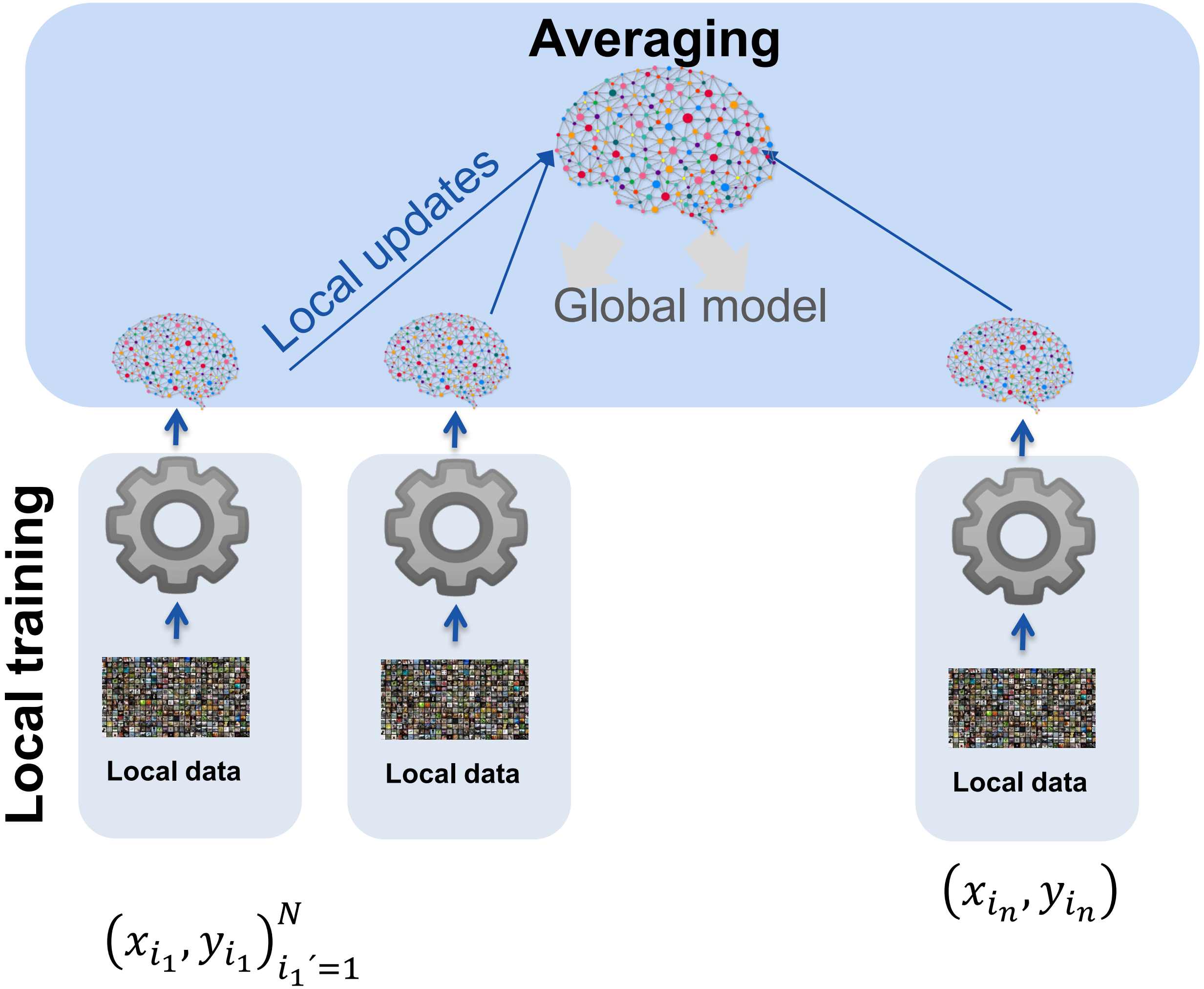
Manufacturing

Training ML models with centralized data



- Empirical loss function
  - $L(\theta) = \sum l(\theta; x_i, y_i)$
- Variety of loss functions available
  - Cross-entropy
  - Log loss
  - Exponential loss
  - Hinge loss
  - Mean Square Error (MSE, l2 norm)
  - Mean Absolute Error (MAE, l1 norm)
  - Huber Loss
- Training:
  - $\min_{\theta} L(\theta)$

Federated Learning - Distributed Data



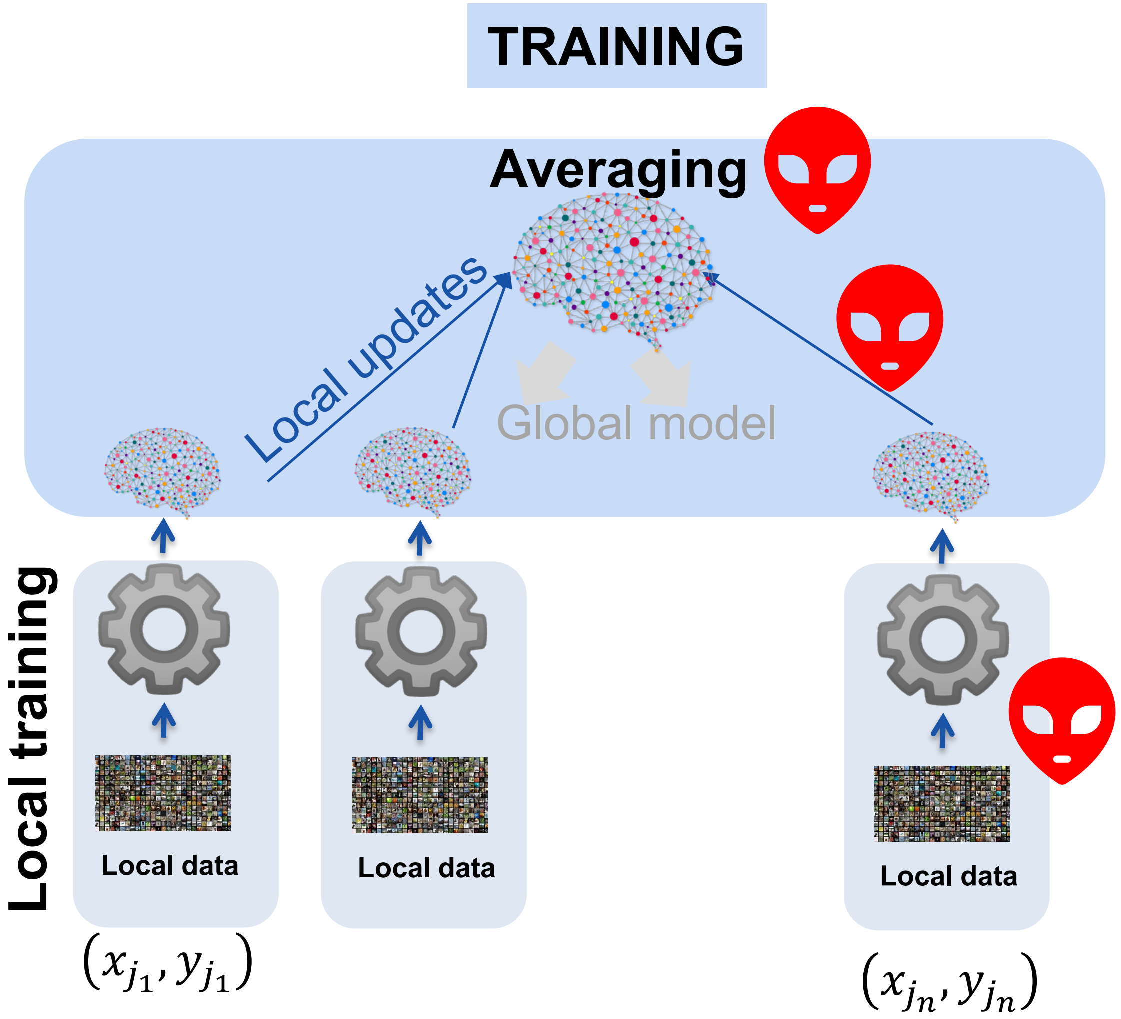
- Objective
  - $\min_{\theta} L(\theta), \quad \text{where} \quad L(\theta) = \sum_{k=1}^{K} p_k L_k(\theta)$
- Local objectives $L_k(\theta)$
  - Empirical loss function

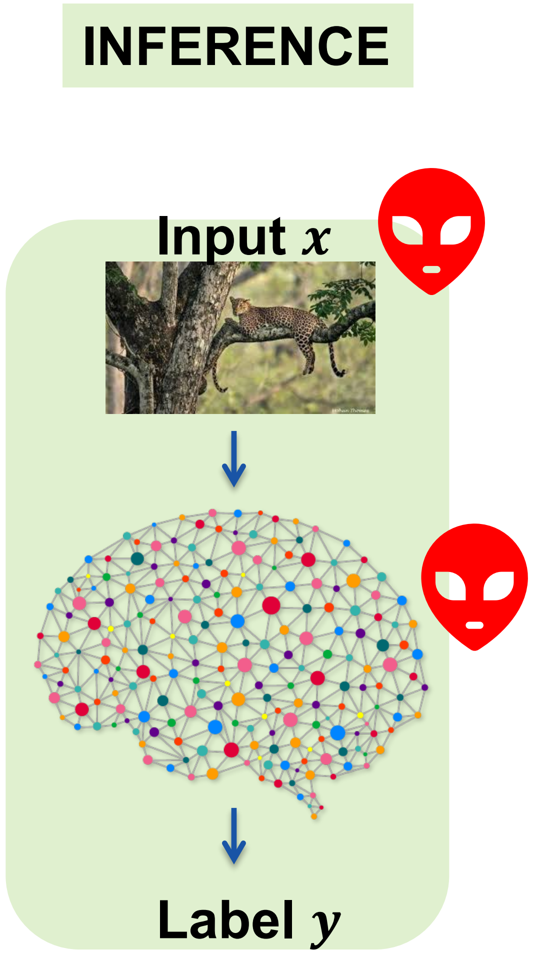
$$L_k(\theta) = \sum_{i_k=1}^{N_k} l_k(\theta; x_{i_k}, y_{i_k})$$

- Weighting of local objectives
  - > Uniform $p_k = \frac{1}{n}$
  - > Proportional $p_k = \frac{n_k}{n}$, where $N = \sum_k n_k$
- Learning of global model
  - Gradient averaging
    - > $\theta^{t+1} = \theta^t - \eta_t \sum_{k=1}^{n} \mathrm{p_k} \nabla L_k(\theta^t)$
  - Federated averaging
    - > $\theta^{t+1} = \sum_{k=1}^{n} p_k \theta_k$

Konecny, et al "Federated Optimization: Distributed Machine Learning for On-Device Intelligence," NIPS 2017

What could go wrong?





Taxonomy of threat models

- Attack surface and attack vector

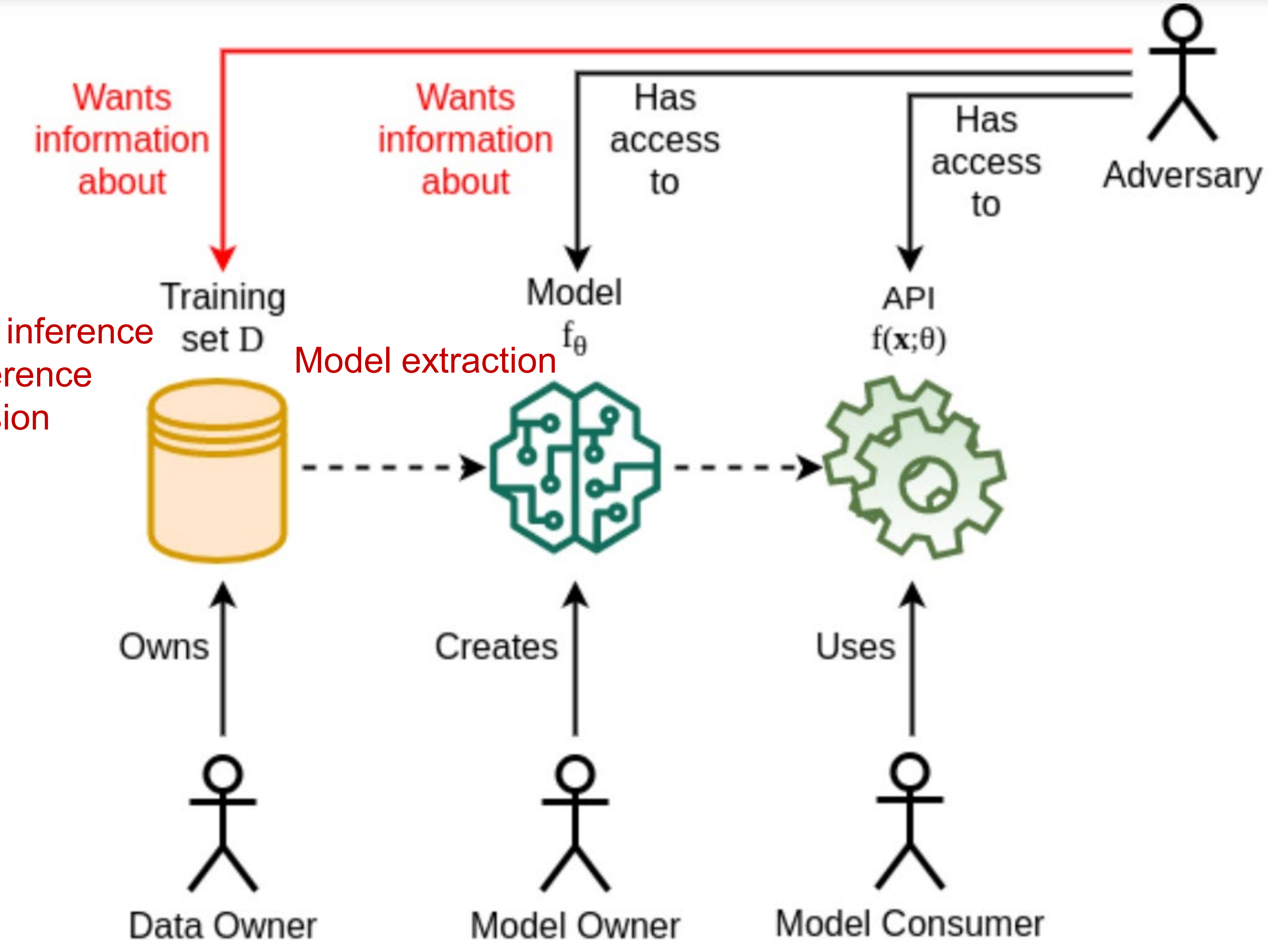
	Data	**Model**
Training time	Data poisoning Backdoor	Parameter poisoning Reconstruction attack
Inference time	Evasion (adversarial examples)	Membership inference Property inference Model inversion Model extraction

- Information availability

Black box White box

Privacy Attacks: Threat Models

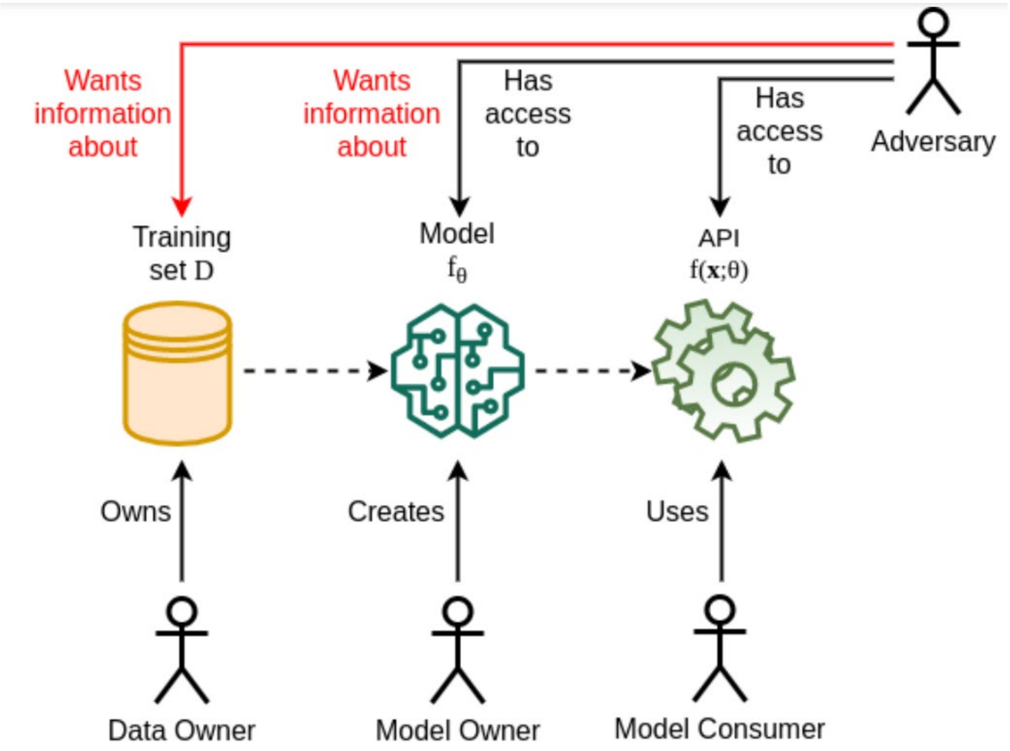
- Membership inference
- Property inference
- Model inversion



Rigaki, Maria, and Sebastian Garcia. "A survey of privacy attacks in machine learning." *arXiv preprint arXiv:2007.07646* (2020).

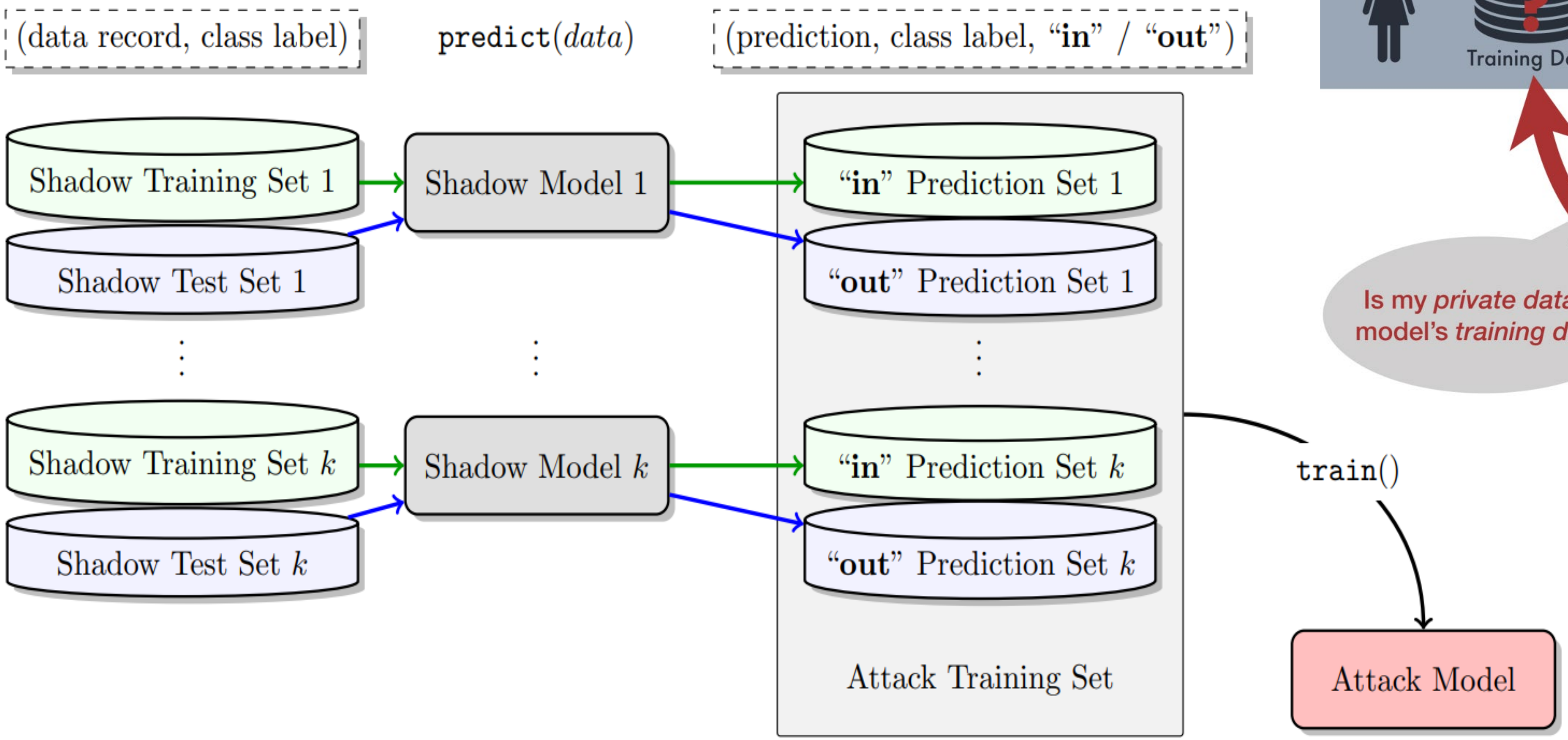
Privacy attacks in brief

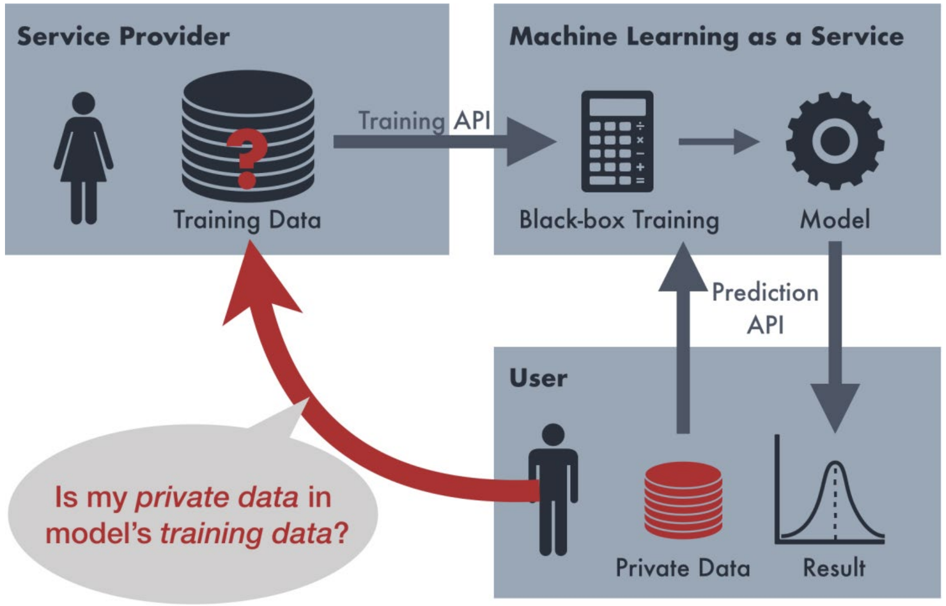
- Membership inference attack
  - > *Was this data record in or out of training dataset?*
- Property inference attack
  - > *Is this property present or absent in the training dataset?*
- Class-label distribution inference
  - > *What is the proportion of training data with label c?*
- Model Inversion attack
  - > *Training data reconstructed using model predictions*
- Model extraction
  - > *Model parameters or hyper-parameters extracted (reverse engineering)*



Membership Inference Attacks

- *“Given a machine learning model and a record, determine whether this record was used as part of the model’s training dataset or not.”*
- Method: Shadow training: Auxiliary data





Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*.

Mitigation of Membership Inference Attacks

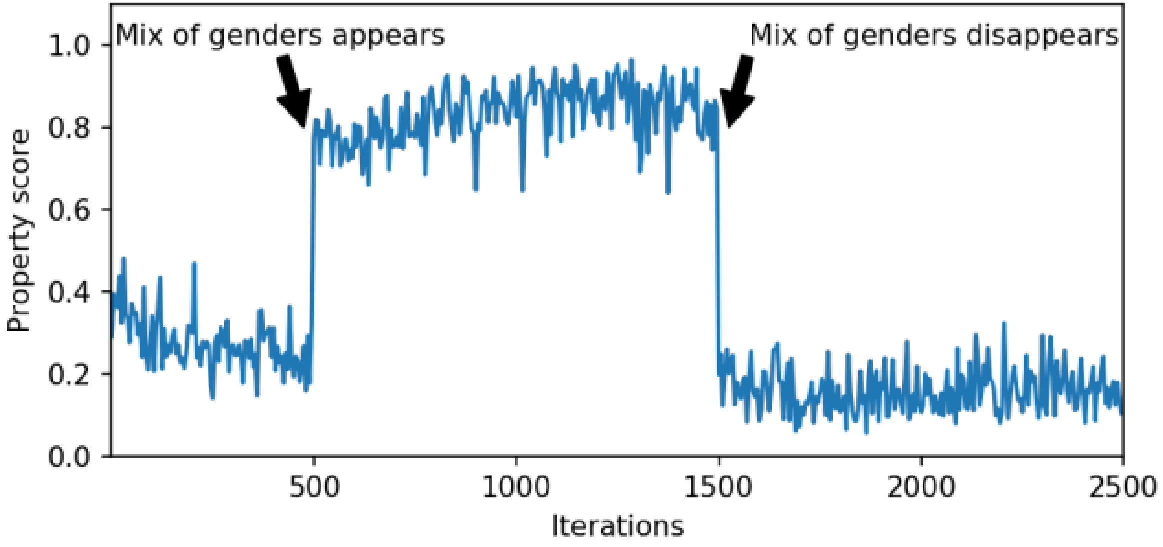
- Differentially-private training is by construction immune to membership inference attacks.
- Restrict prediction $F_\theta(x)$ to top k-classes.
- Round-off the prediction vector
- Increase entropy of prediction vector temperature t: $\frac{exp\{\frac{z_i}{t}\}}{\sum_j \exp\{\frac{z_j}{t}\}}$
- Regularization $L_2: \lambda||\theta||_2$

Hospital dataset	*Testing Accuracy*	*Attack Total Accuracy*	*Attack Precision*	*Attack Recall*
No Mitigation	0.55	0.83	0.77	0.95
Top $k = 3$	0.55	0.83	0.77	0.95
Top $k = 1$	0.55	0.82	0.76	0.95
Top $k = 1$ label	0.55	0.73	0.67	0.93
Rounding $d = 3$	0.55	0.83	0.77	0.95
Rounding $d = 1$	0.55	0.81	0.75	0.96
Temperature $t = 5$	0.55	0.79	0.77	0.83
Temperature $t = 20$	0.55	0.76	0.76	0.76
L2 $\lambda = 1e-4$	0.56	0.80	0.74	0.92
L2 $\lambda = 5e-4$	0.57	0.73	0.69	0.86
L2 $\lambda = 1e-3$	0.56	0.66	0.64	0.73
L2 $\lambda = 5e-3$	0.35	0.52	0.52	0.53

Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*.

Property Inference Attacks (PIA)

- Machine Learning (ML) models unintentionally memorize properties of training data
- Extract global statistics about training data via access to trained model (black v/s white-box)
- Usually a binary classifier problem-presence/absence of a certain property.
- Constitute a *privacy risk* in many healthcare and industrial applications
- Online learning: when does a certain property appear



Class Label Distribution Inference

PIA on ML classifiers with C output classes: infer the class label distribution (categorical) of training data

Mitigation of Generic Property Inference Attacks

- Add noise to training data: flip labels, introduce adversarial examples
- Add noise to classifier output
- Encode arbitrary information while not compromising on model generalizability and performance.
- Although encoding or memorizing information is also an attack, it will bypass meta-classifiers learnt using shadow-training.

Song, Congzheng, Thomas Ristenpart, and Vitaly Shmatikov. "Machine learning models that remember too much." *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*. 2017.

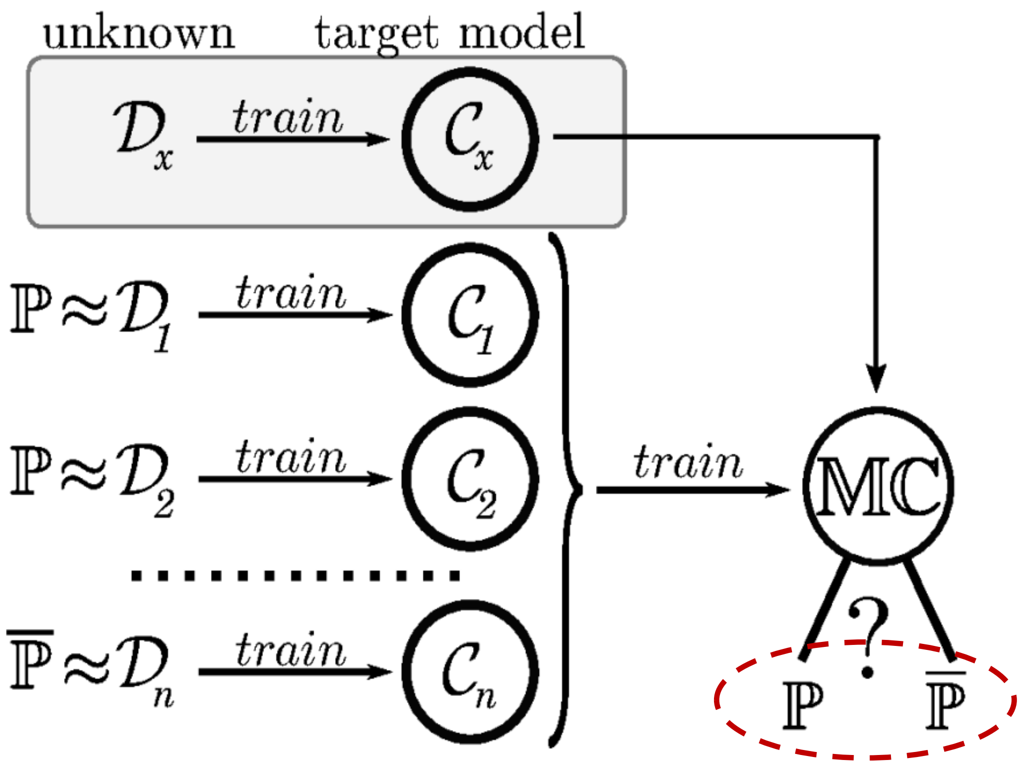
Class-Label Distribution Inference

Class-label distribution $p \in \Delta^{C-1}$ given labeled training data $D \triangleq \{x_i, e_i\}_{i=1,2,3,\ldots,N}$

$$p = \frac{1}{N}\sum_i e_i \triangleq \frac{N_c}{\sum_c N_c}$$

A new type of PIA we introduced

- Training-time: Federated learning
- Inference: Meta-classifier

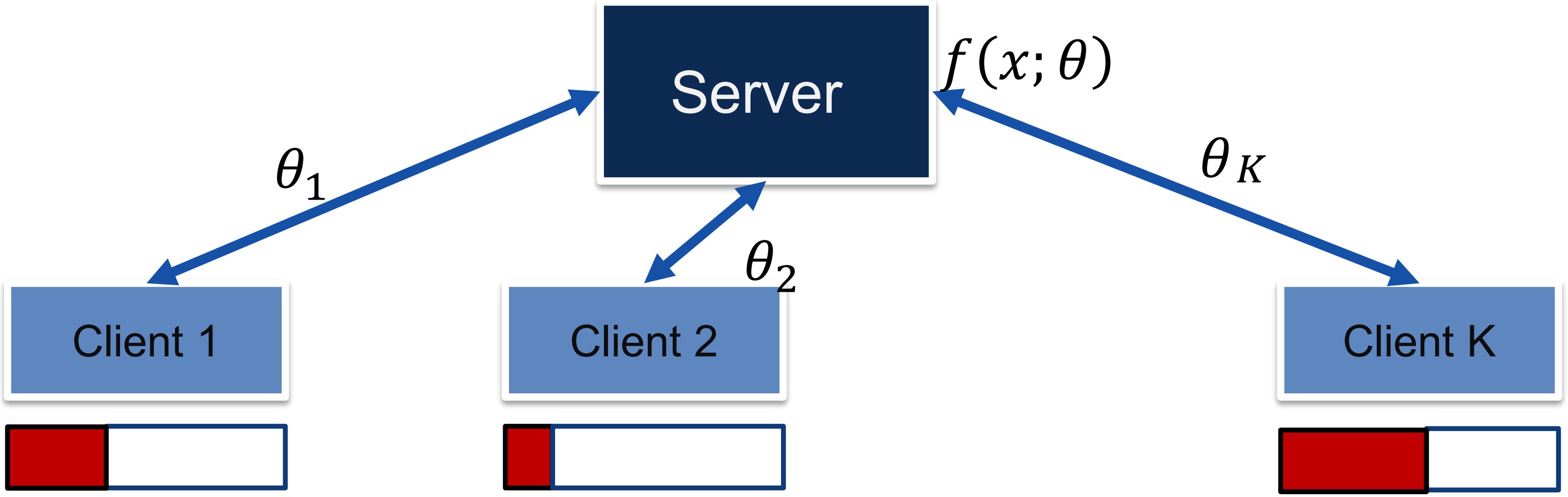


New Attack!

Ateniese, Giuseppe, et al. "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers." *International Journal of Security and Networks* 10.3 (2015): 137-150.

Class-Label Distribution Inference in Federated Learning

- Federated Learning (FL)- distributed machine learning: server-client model
- Training data: local and private to client
- Server: *unaware* of potential class-imbalance:
  - Class-imbalance deteriorates accuracy-detection and mitigation important
  - Composition of client's data- *privacy risk* in many healthcare and industrial applications

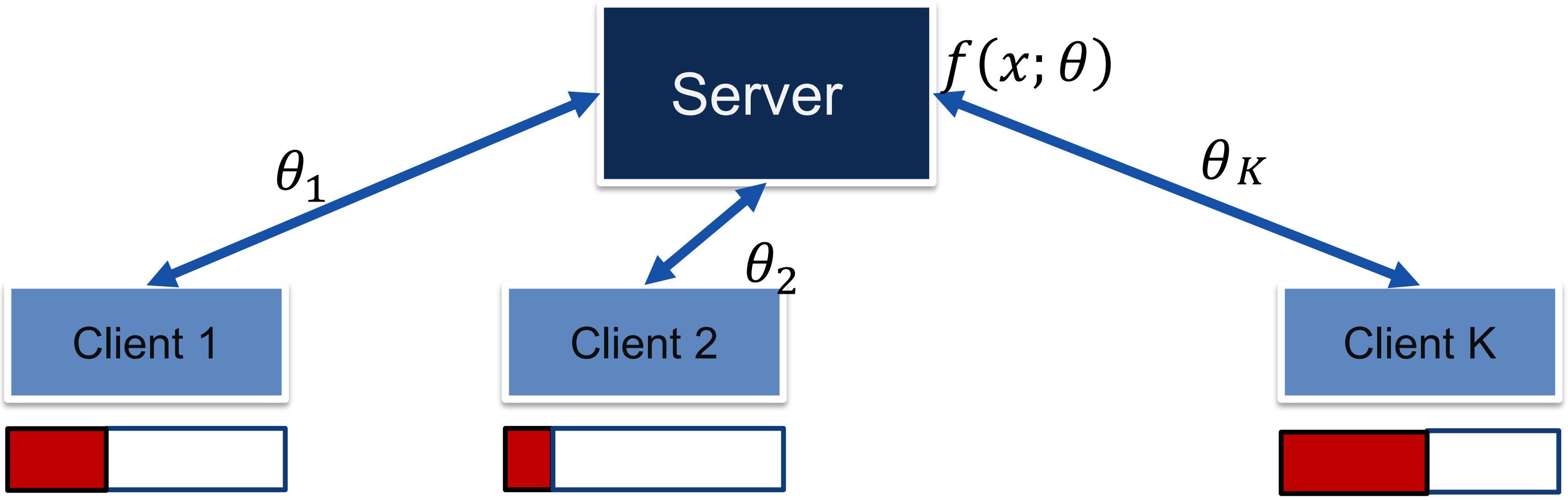


Motivation: TECoSA partner

- Training data from clients: labeled as anomalous/ non-anomalous (fatigued/non-fatigued)
- Each client wants to keep their training data and labels private
- **Goal**: To learn a classifier (supervised ML) to decide if there is fatigue or not. Server: Company.
- Can the server infer the fraction of training data labels that are anomalous or not using model parameter updates?
- Knowledge of this fraction:
  - Could provide competitive advantage
  - Idea about client's profitablity

Goals

- Develop methods for class-label distribution inference when parameter updates at every round t are available
- Identify conditions for exact inference
- Develop methods for non-exact inference (estimators)



Related Work

Class-label distribution inference studied as class-imbalance mitigation and as attack:

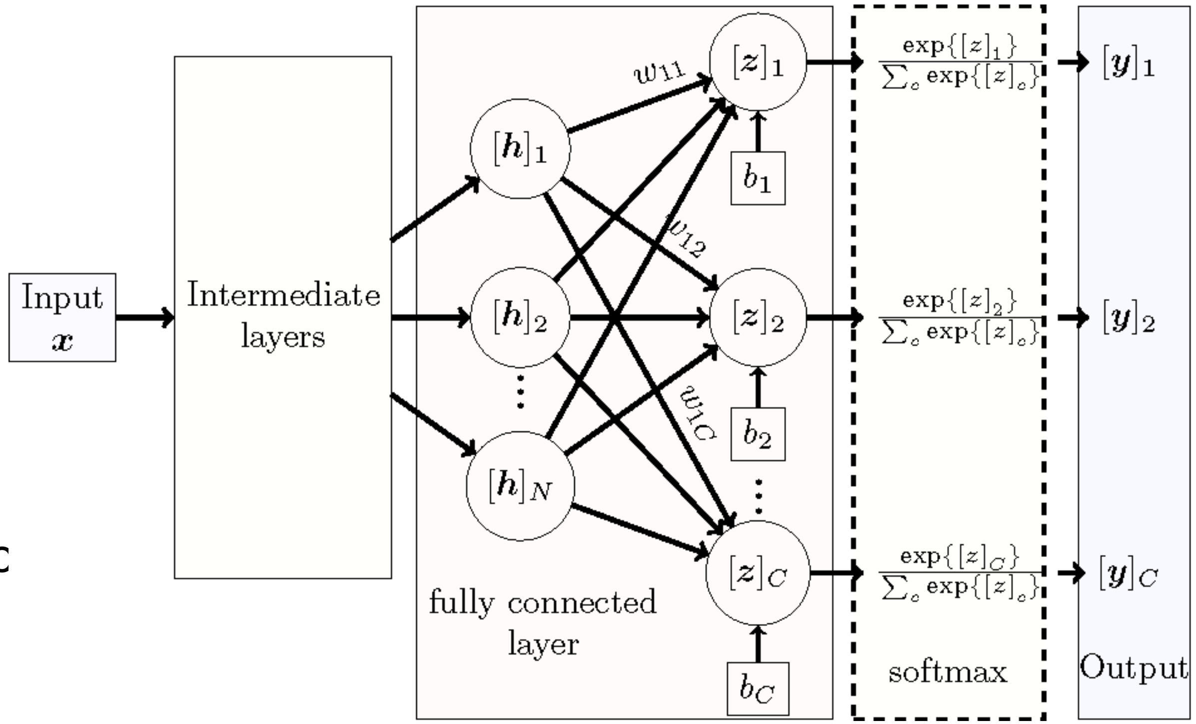
- To address class-imbalance in FL:
  - change loss-function
  - cluster clients
- As a property inference attack: preference profiling attacks (PPA)
- Gradients from last layer to extract label-proportion information
- Gradients to reconstruct training data

1. L. Wang et al., “Addressing class imbalance in federated learning,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 10 165–10 173.
2. M. Duan et al., “Self-balancing federated learning with global imbalanced data in mobile systems,”IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 1, pp. 59–71, 2020.
3. C. Zhou et al., “PPA: Preference profiling attack against federated learning,” arXiv preprint arXiv:2202.04856, 2022.
4. A. Wainakh et al., “User-level label leakage from gradients in federated learning,” Proceedings on Privacy Enhancing Technologies, vol. 2022, no. 2, pp. 227–244, 2022.
5. L. Zhu et al., “Deep leakage from gradients,” in Advances in Neural Information Processing Systems, H. Wallach et al., Eds., vol. 32, Curran Associates, Inc., 2019.

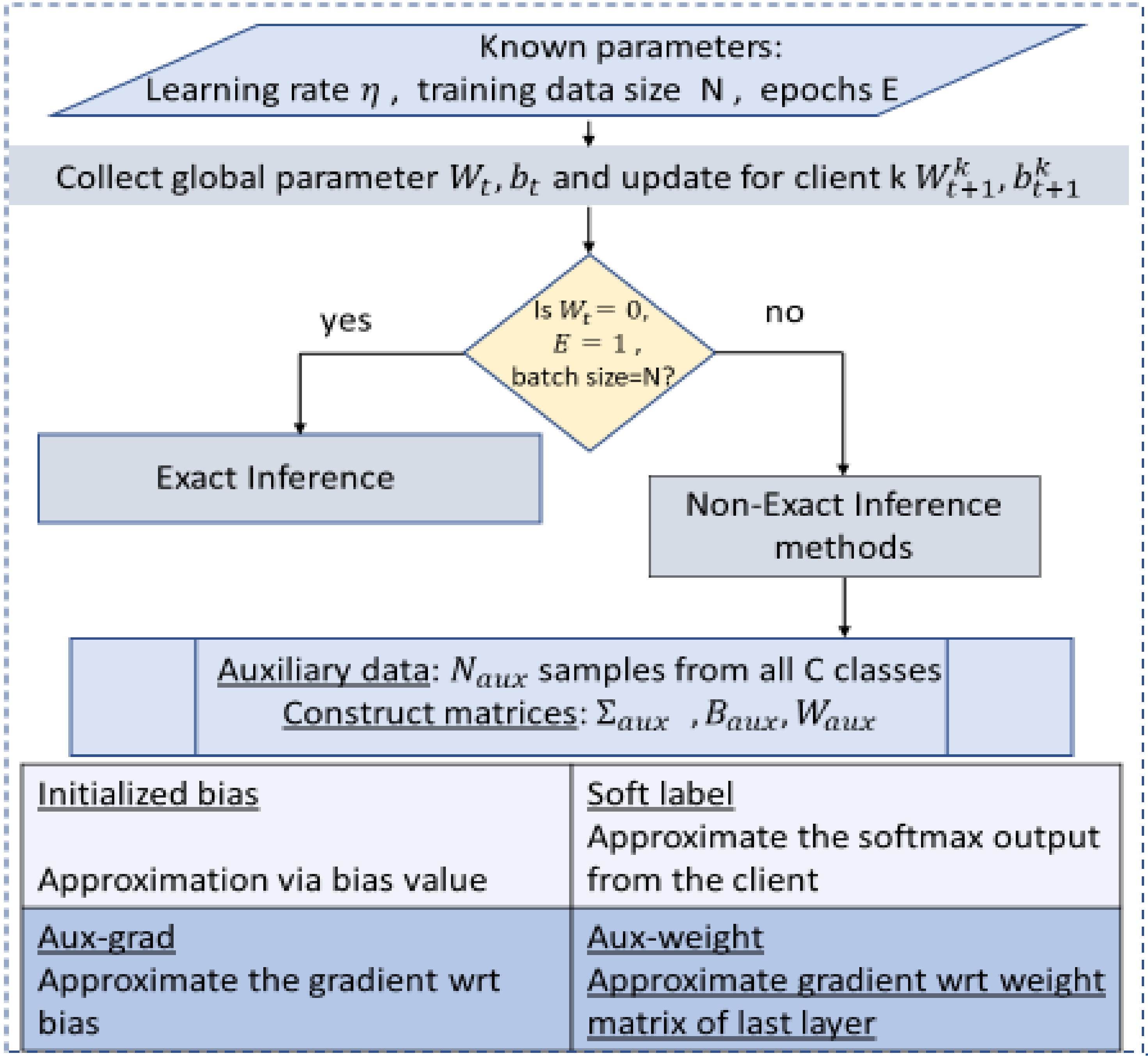
Class-Label Distribution Exact Inference

- Exact inference for client k at global iteration t possible
- Given: bias at server b^t and updated at client k b_k^{t+1}
- Conditions for exact inference:
  - Learning rate η
  - Data size N
  - Full-batch gradient descent
  - Single epoch update by the client
  - Weight matrix set to zero by server at iteratic
- Conditions not met: approximations used: 4 estimators for non-exact inference
- Use Auxilliary dataset containing N_{aux} samples from each class

Generic NN classifier

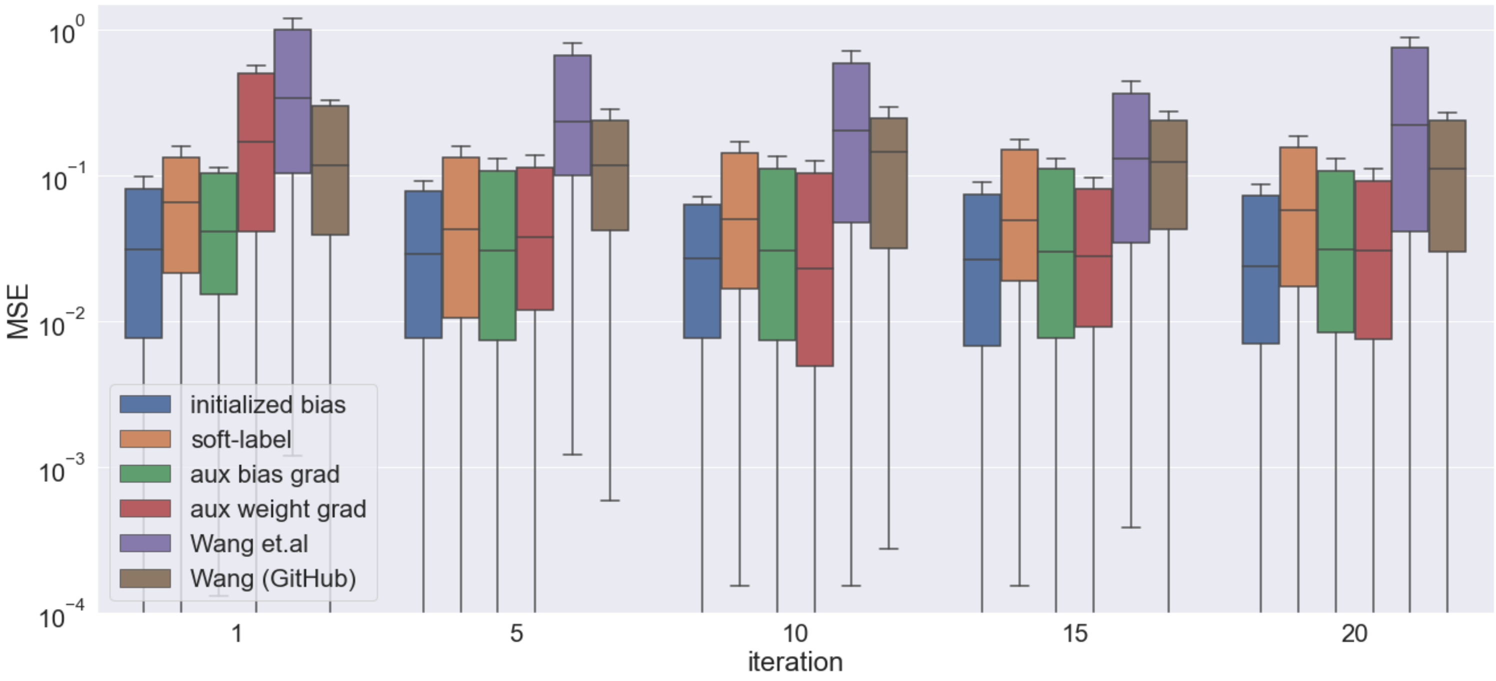


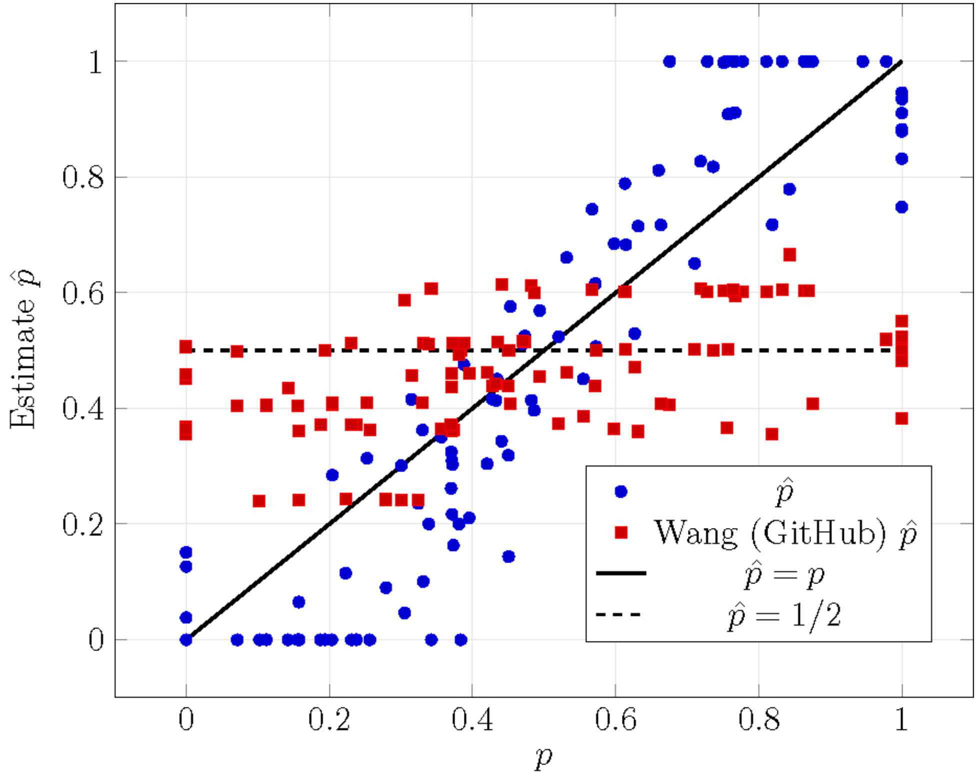
Bias from last fully-connected layer used


Known parameters:
Learning rate η , training data size N , epochs E
Collect global parameter W_t, b_t and update for client k W_{t+1}^k, b_{t+1}^k
Is $W_t = 0$, $E = 1$, batch size=N?
yes
no
Exact Inference
Non-Exact Inference methods
Auxiliary data: N_{aux} samples from all C classes
Construct matrices: Σ_{aux} , B_{aux}, W_{aux}
Initialized bias
Approximation via bias value
Soft label
Approximate the softmax output from the client
Aux-grad
Approximate the gradient wrt bias
Aux-weight
Approximate gradient wrt weight matrix of last layer

Numerical Results

- Comparison with state of the art: Wang et.al
- UCI Census Income Dataset-binary classification

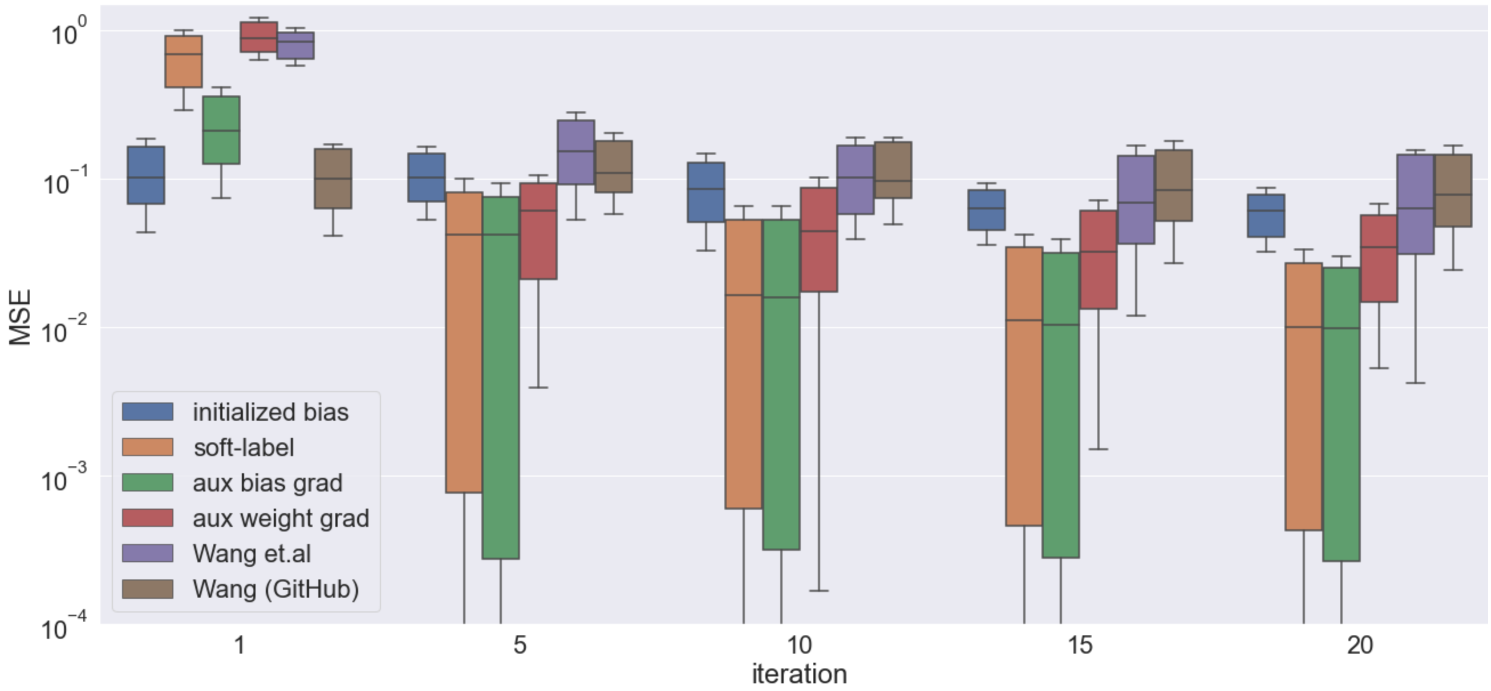


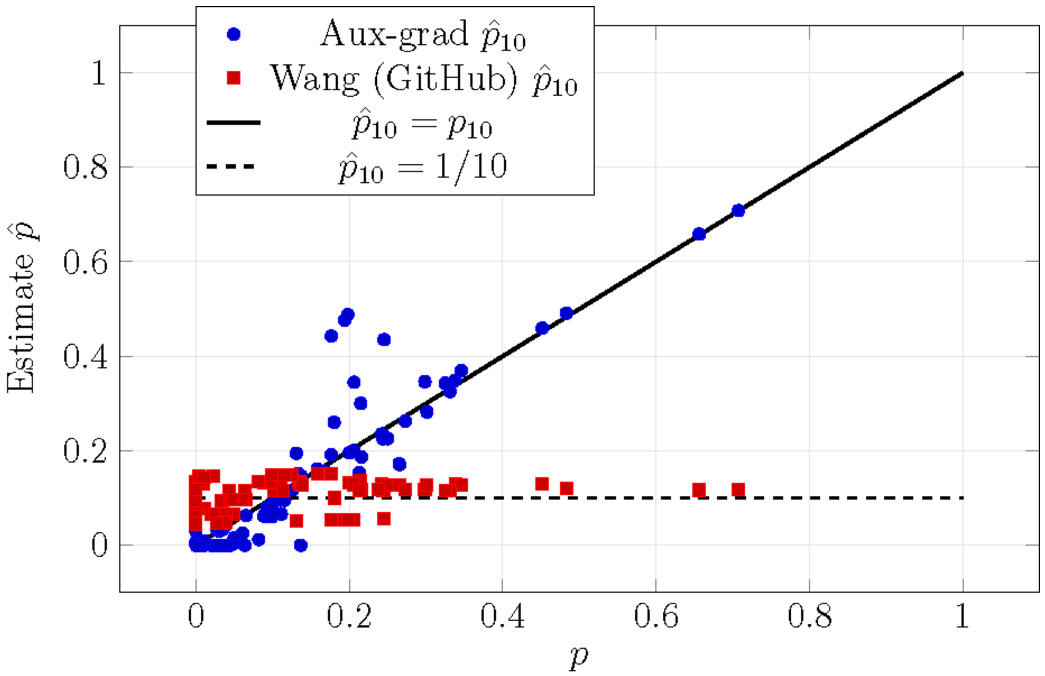


L. Wang et al., “Addressing class imbalance in federated learning,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 10 165–10 173.

Numerical Results

CIFAR-10: 10 class image classification

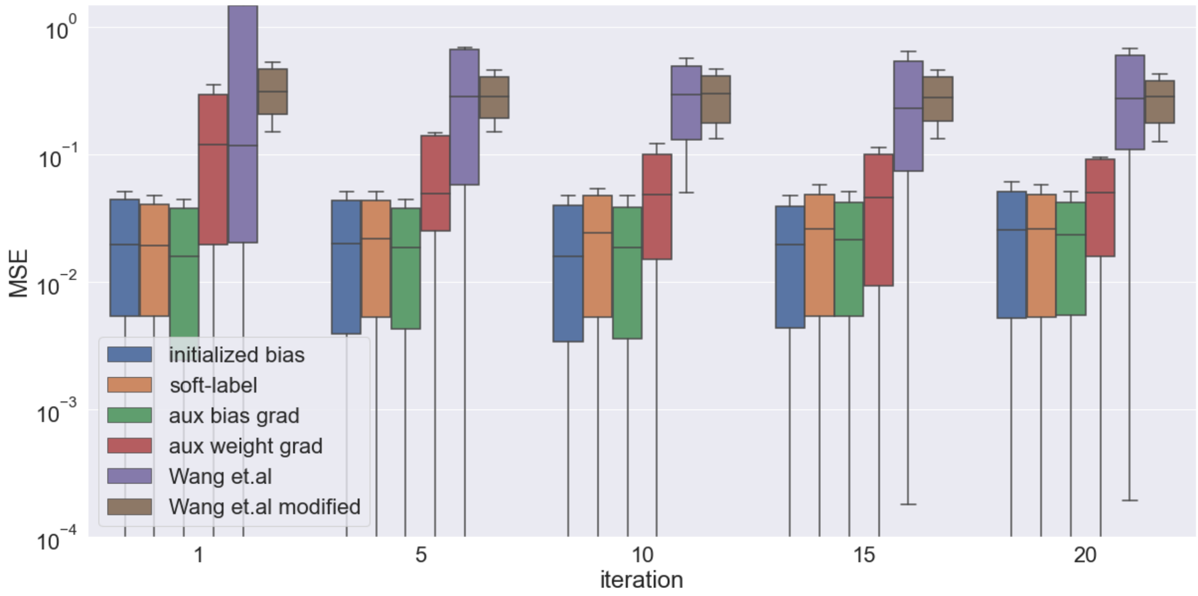




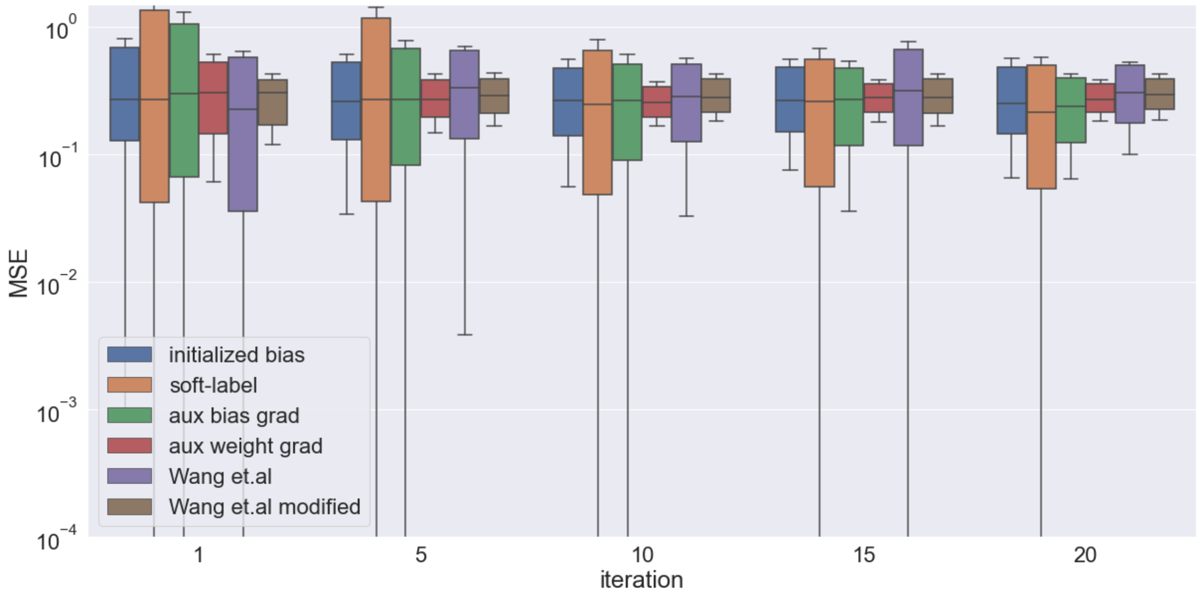
L. Wang et al., "Addressing class imbalance in federated learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 10 165–10 173.

Random oversampling as Attack Mitigation

- Can we mitigate class-label distribution inference attack?
- Random oversampling: sample with replacement for minority classes
- Makes class distribution `balanced' (uniform distribution)
- Proposed methods fail to estimate which implies effective countermeasure



Imbalanced dataset

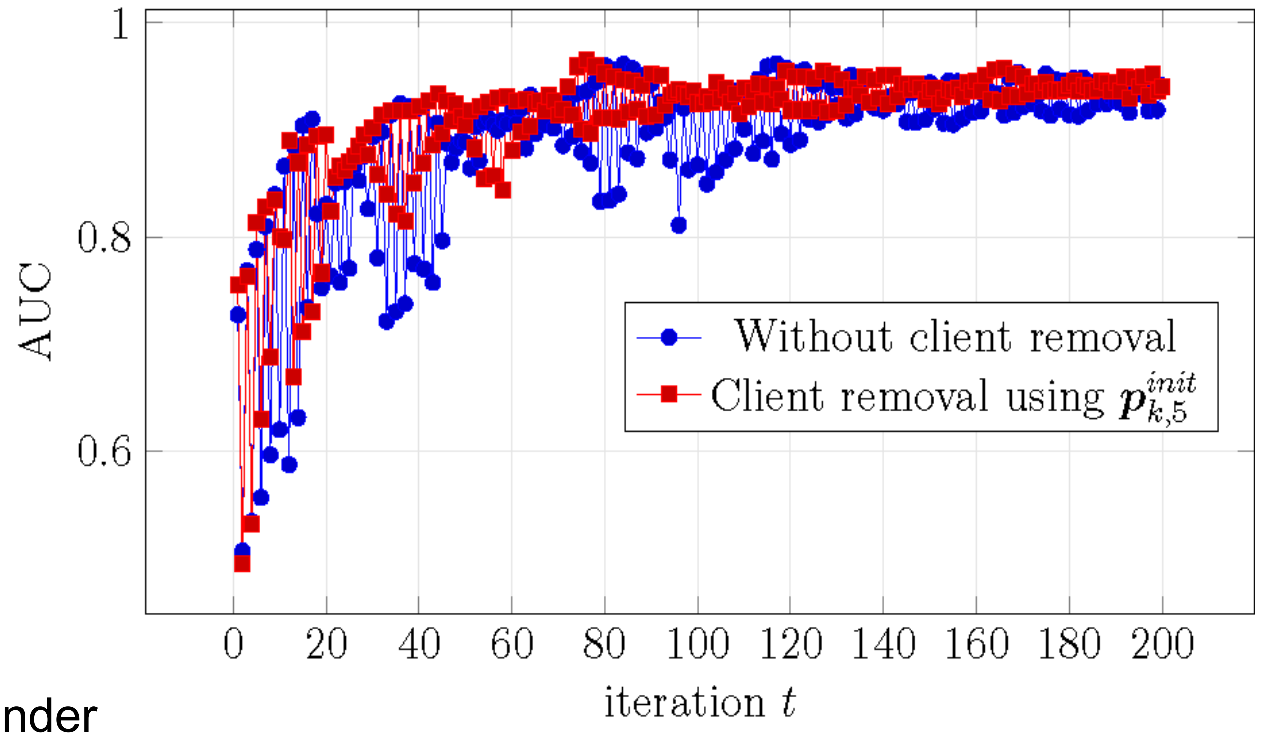


Imbalanced datasets balanced via random oversampling

Addressing class-imbalance in FL

- Class-imbalance in FL: slow convergence, low accuracy
- Problem mitigated by grouping clients based on class-label distribution (known)
- Remove clients with estimated class-imbalance
- US Census Income dataset:

  $p \in [0,0.2] \cup [0.8,1]$
- Client removal: improved accuracy (AUC: area under the ROC curve) and faster convergence.

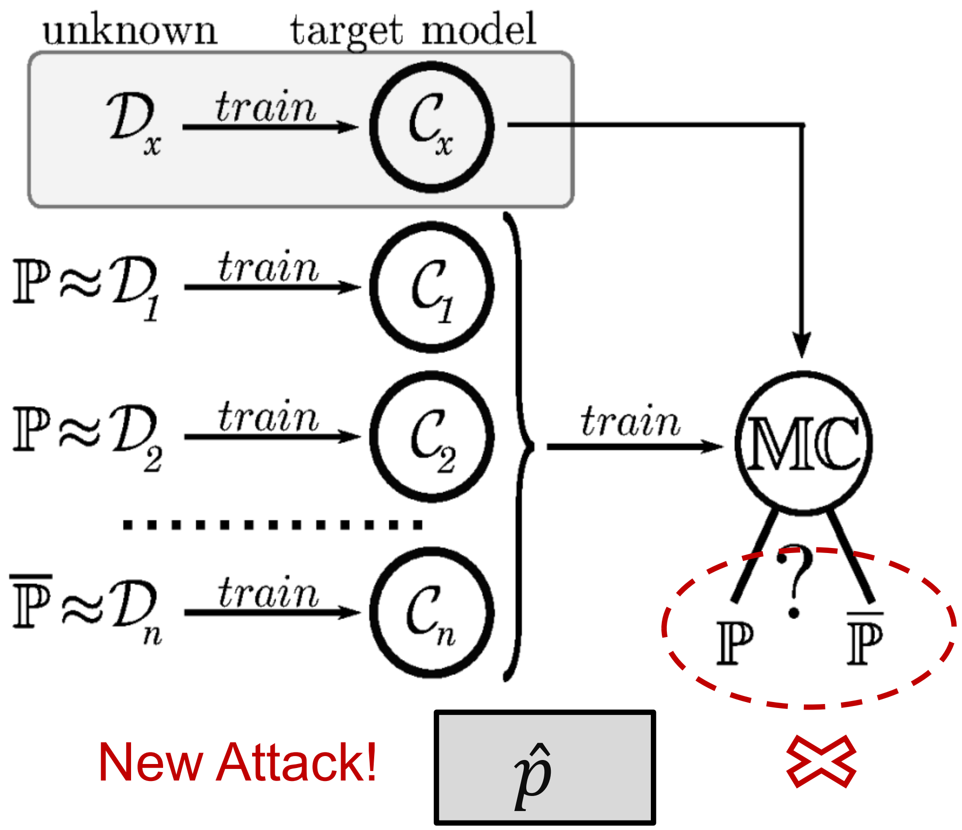


Area under the ROC curve (AUC) at iteration t for cases with and without client removal. US Census Income dataset

Jiahua Ma, Xinghua Sun, Wenchao Xia, Xijun Wang, Xiang Chen, and Hongbo Zhu. 2021. Client selection based on label quantity information for federated learning. In 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 1–6

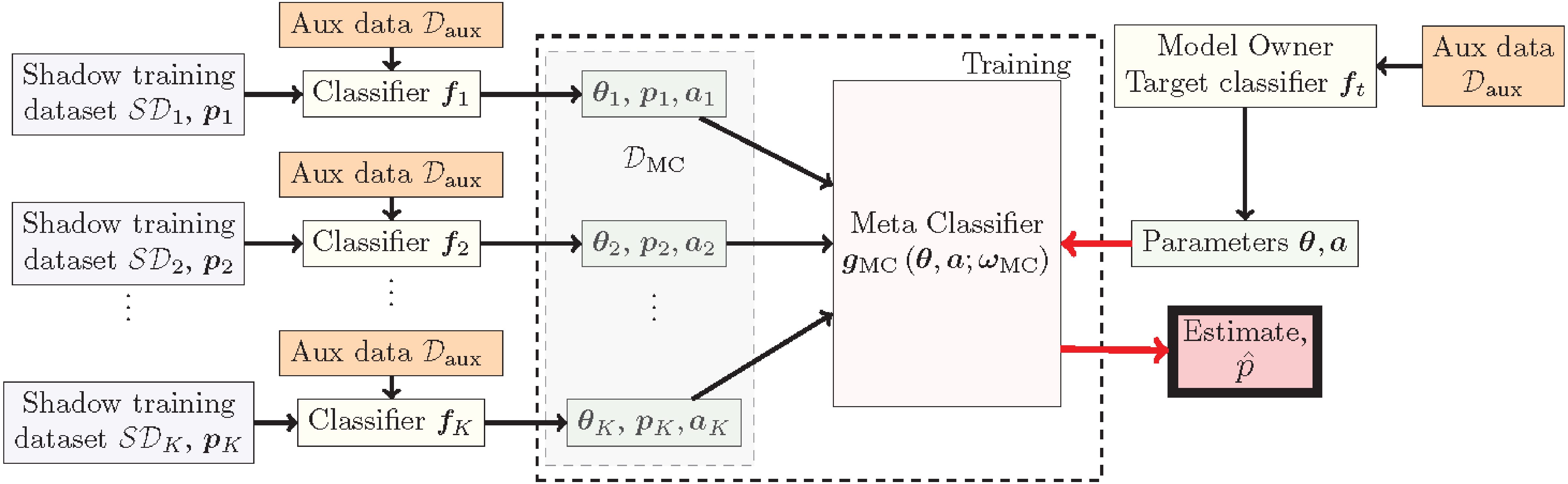
Class-Label Distribution Inference: Trained ML Models

- Fully-trained models, attack at inference time
- ML model parameters (after training) are available
- Target classifier architecture: fully connected neural networks
- *Shadow training* methodology: Meta-Classifier
- <u>Challenge:</u> multi-dimensional sampling for multi-class classifiers



- Ganju, Karan, et al. "Property inference attacks on fully connected neural networks using permutation invariant representations." *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018.
- Ateniese, Giuseppe, et al. "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers." *International Journal of Security and Networks* 10.3 (2015): 137-150.

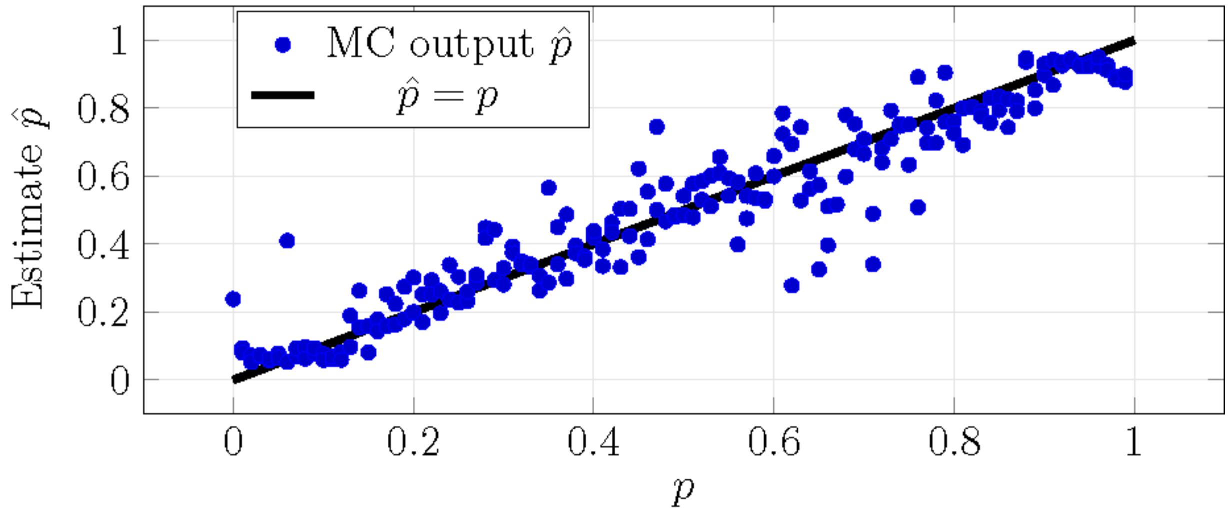
Accuracy Augmented Meta-Classifier Attack

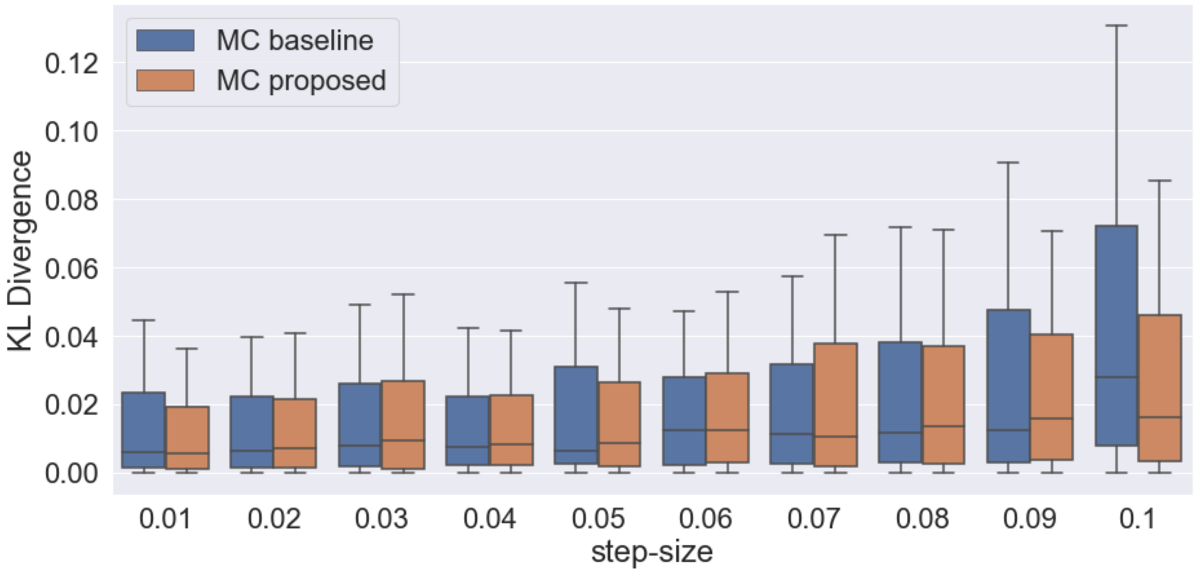


- Generate shadow training data sets to train shadow classifiers
- Meta-classifier architecture: permutation invariant
- Use parameters and accuracy to train meta-classifier

Numerical Results

UCI Census Income Classification (≷ 50k)



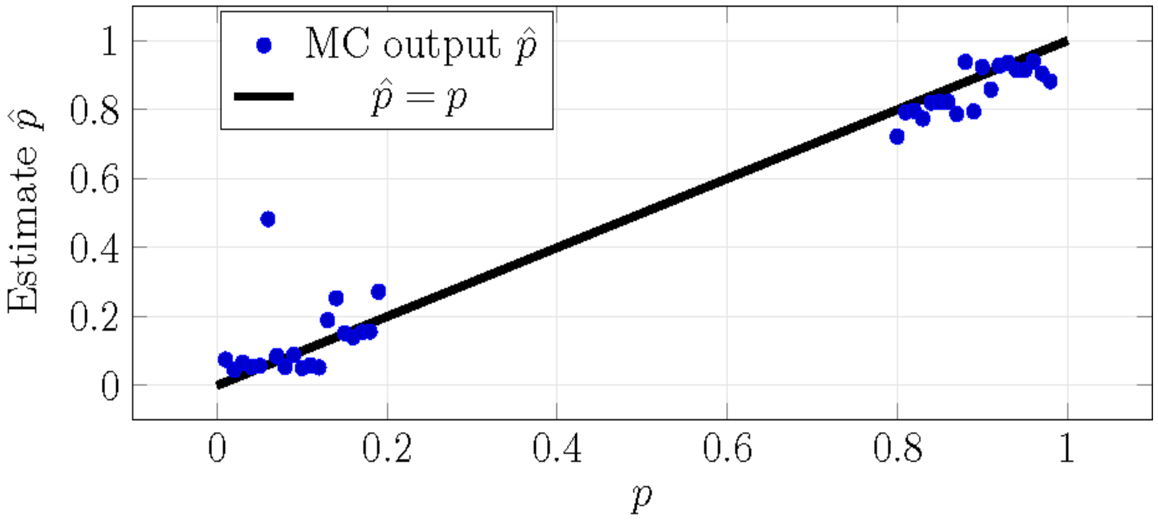


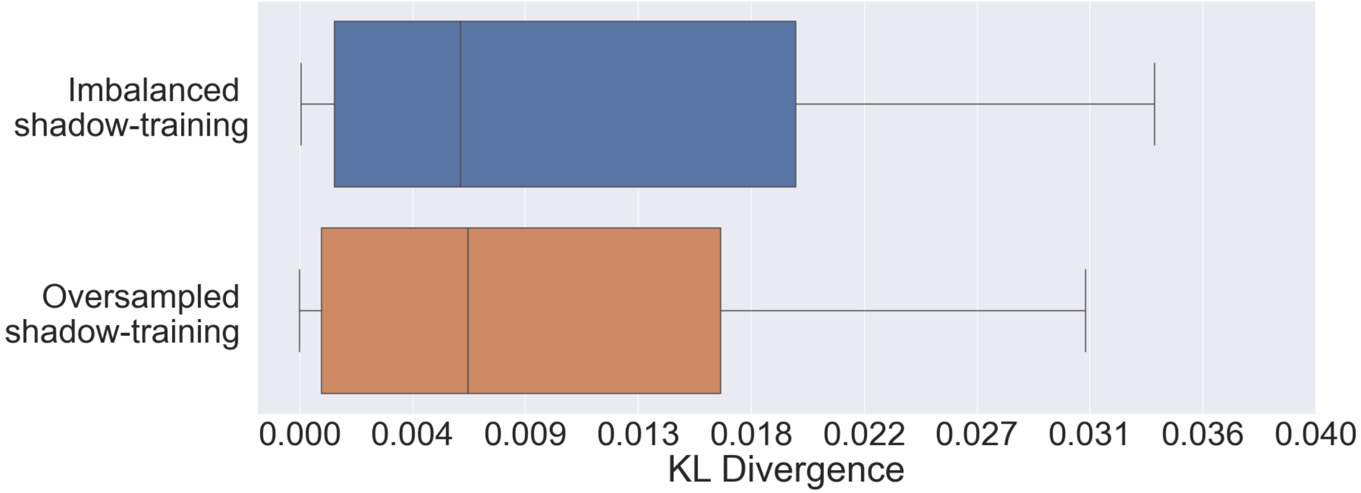
- Binary Classification
- Accurate estimates for most values of p

- Performance of Meta-Classifier (KL Divergence)
- Shows improvement over baseline:
  - Architectural changes
  - Accuracy augmentation

Ganju, Karan, et al. "Property inference attacks on fully connected neural networks using permutation invariant representations." *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018

Robustness to random oversampling





- The imbalance of class labels is addressed by random oversampling of the minority class
- Makes class label distribution `balanced' (discrete uniform)
- Meta-classifier can still estimate original distribution!
- Further training the meta-classifier on oversampled shadow-training datasets <u>improves</u> performance

Summary and Conclusions

- Privacy Attacks:
  - Membership inference
  - Property inference
- Class-label distribution inference:
  - In FL via model updates
  - In trained ML models via meta-classifiers
- Random oversampling as countermeasure:
  - In FL works as a mitigation measure
  - Meta-classifiers seem robust to it

Ongoing and Future work

- Test meta-classifier based attacks in the FL setting
- Efficient online and adaptive methods of shadow training dataset sampling for higher dimensions (multi-class classifiers)
- Mitigation scheme against meta-classifier-based property inference attacks
- Meta-classifier attacks for other target classifier architectures

Thank you!