Privacy in machine learning: from centralized to federated approaches

Carole Frindel – Insa-Lyon / Creatis Myriad

Antoine Boutet – Insa-Lyon / CITI Inria-Privatics

Deep Learning for medical imaging school – Lyon April 17-21 2023

Massive deployment of ML

Rise many questions
- Utility
- Privacy
- Security
- Fairness
- Explainability
- Energy Footprint

Challenge:
Address globally these questions

Sensitivity of Medical Data and Images

Personal Health Information

Confidentiality

Protected by Law

Vulnerability to Cyber Threats

Potential for Misuse

Sensitivity of Medical Data and Images

Personal Health Information

Patient name, address, medical history, medications, etc.

Unauthorized access, use, or disclosure can harm patients

(0010,1005)	PN	Patient's Birth Name
(0010,1010)	AS	Patient's Age
(0010,1020)	DS	Patient's Size
(0010,1021)	SQ	Patient's Size Code Sequence
(0010,1030)	DS	Patient's Weight
(0010,1040)	LO	Patient's Address
(0010,1050)	LO	Insurance Plan Identification
(0010,1060)	PN	Patient's Mother's Birth Name
(0010,1080)	LO	Military Rank
(0010,1081)	LO	Branch of Service
(0010,1090)	LO	Medical Record Locator
(0010,1100)	SQ	Referenced Patient Photo Sequence
(0010,2000)	LO	Medical Alerts
(0010,2110)	LO	Allergies
(0010,2150)	LO	Country of Residence
(0010,2152)	LO	Region of Residence
(0010,2154)	SH	Patient's Telephone Numbers
(0010,2155)	LT	Patient's Telecom Information
(0010,2160)	SH	Ethnic Group
(0010,2180)	SH	Occupation
(0010,21A0)	CS	Smoking Status
(0010,21B0)	LT	Additional Patient History
(0010,21C0)	US	Pregnancy Status
(0010,21D0)	DA	Last Menstrual Date
(0010,21F0)	LO	Patient's Religious Preference

DICOM Library

Anonymize, Share, View DICOM files ONLINE

Sensitivity of Medical Data and Images

Confidentiality

Disclosure can lead to discrimination, stigmatization, or social exclusion

Sensitivity of Medical Data and Images

Protected by Law

- HIPAA (US), GDPR (EU), and other laws and regulations
- Breach can result in significant financial and legal penalties

SANTÉ \ DONNÉES PERSONNELLES \ CYBERSÉCURITÉ

Fuite massive de données médicales : la Cnil inflige une amende de 1,5 million d'euros à Dedalus

21 Avril 2022

Dedalus Biologie écope d'une amende de 1,5 million d'euros suite à un contrôle de la Cnil. L'organisme a été saisi suite à la publication dans la presse d'articles relatant une fuite de données médicales. Dedalus Biologie édite le logiciel utilisé par les 28 laboratoires d'où proviennent les données.

Sensitivity of Medical Data and Images

Vulnerability to Cyber Threats

Electronic storage makes medical data vulnerable to cyber-attacks

par LIBERATION et AFP

publié le 25 septembre 2022 à 16h53

Action-réaction

Faute de rançon, les données volées dans un hôpital de l'Essonne se retrouvent mises en ligne

Les hackeurs responsables d'une cyberattaque contre le centre hospitalier sud francilien de Corbeil-Essonnes, ont commencé à diffuser des données, l'hôpital ayant refusé de payer la rançon demandée.

Société, Santé

Vol de données médicales : les hôpitaux de Paris présentent leurs excuses et mettent en garde les victimes

Les informations de santé d'environ 1,4 million de personnes ayant réalisé un dépistage du Covid-19 en 2020 ont été dérobées.

Par Le Parisien

Le 18 septembre 2021 à 09h01

Sensitivity of Medical Data and Images

Potential for Misuse

Medical data/images can be misused for fraudulent activities or identity theft

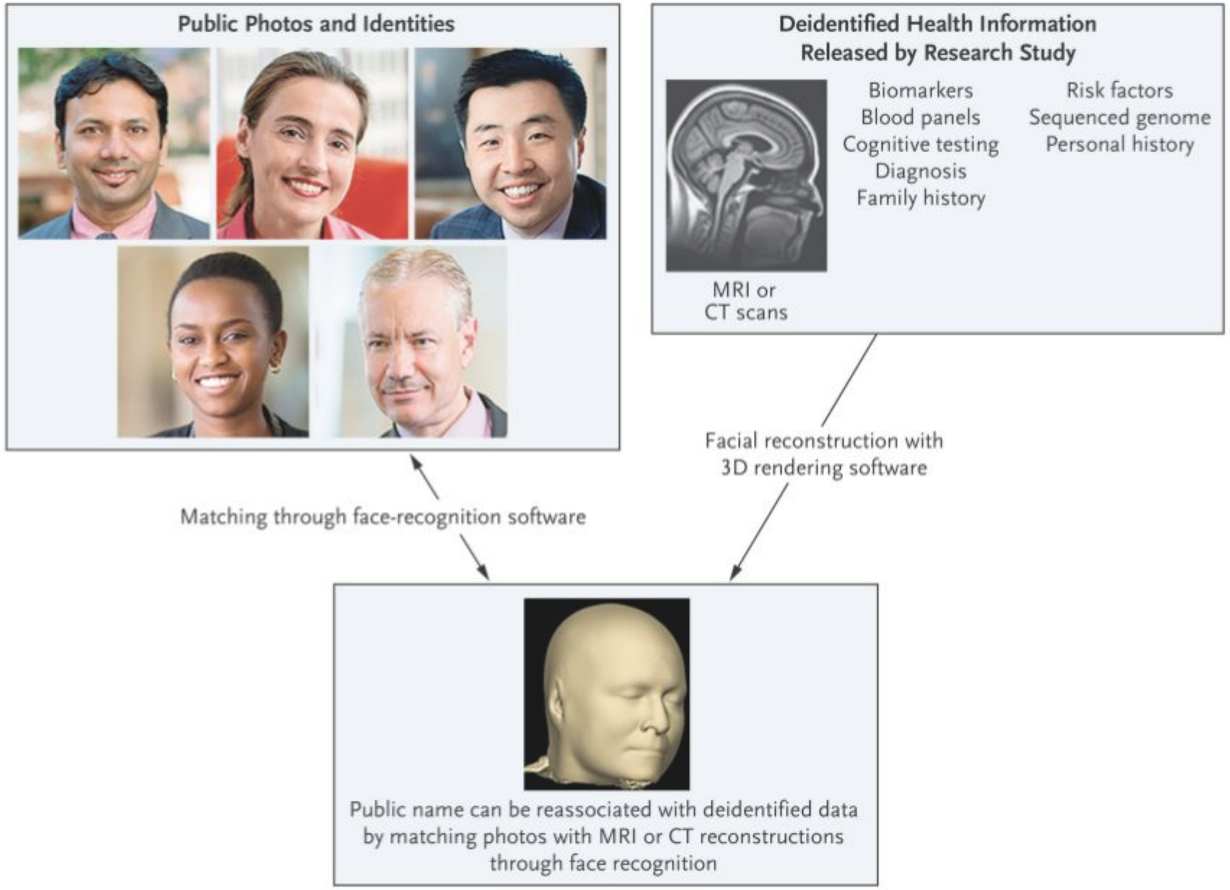
Misuse can lead to significant harm to patients and healthcare providers

La Cnil assiste les victimes d'usurpation d'identité

Par **Stéphanie Delmas**
Publié le 19/05/2021 à 17:11 , mis à jour le 19/05/2021 à 18:18

Threats to “anonymized” medical images

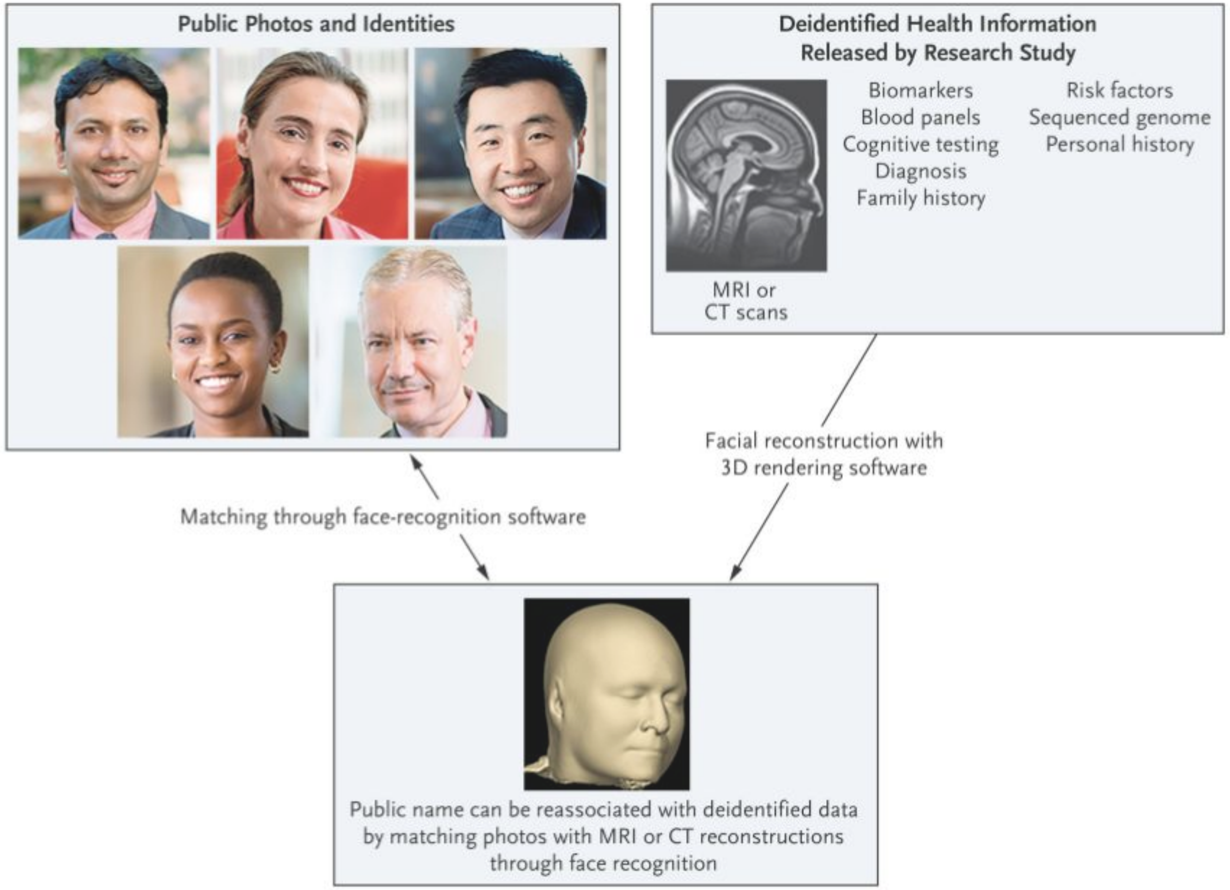
- **Re-identification attacks**



Schwarz et al. *New England Journal of Medicine* 381.17 (2019): 1684-1686.

Threats to “anonymized” medical images

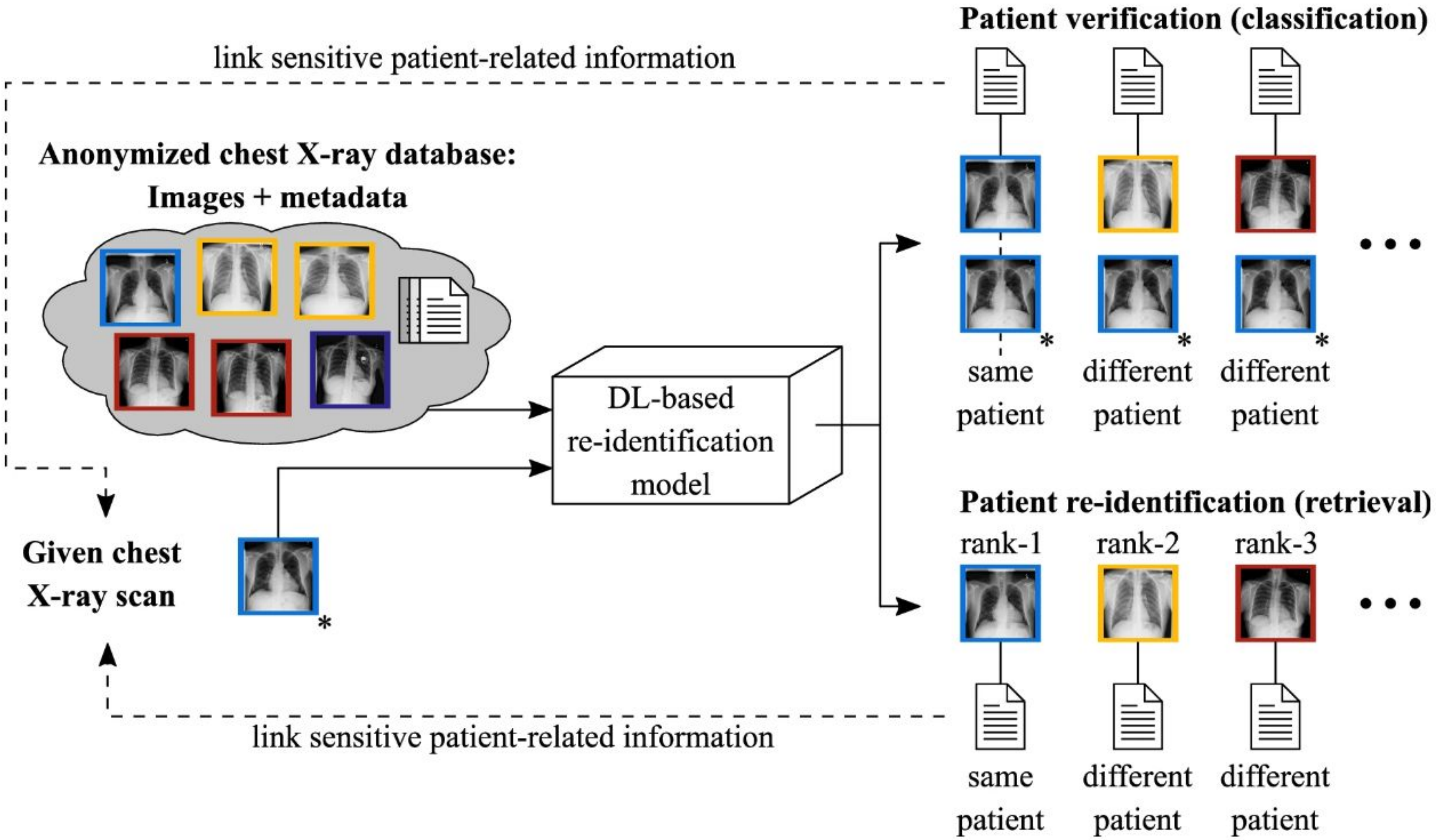
- Re-identification attacks
- **Attribute disclosure attacks**



Schwarz et al. *New England Journal of Medicine* 381.17 (2019): 1684-1686.

Threats to “anonymized” medical images

- **Data linkage attacks**



Packhäuser et al. *Scientific Reports* 12.1 (2022): 14851.

Threats to “anonymized” medical images

Several threats to the anonymity of medical images:

- **Re-identification attacks**
- **Attribute disclosure attacks**
- **Data linkage attacks**

Sanitize/minimize access to medical data to avoid unwanted sensitive inferences

Directions to overcome the limits of anonymisation

- **Limits of the anonymisation**
  - Difficult to break the individual fingerprints without drastically reducing the utility
  - Subject to General Data Protection Regulation

- **New directions**
  - Generation of synthetic data
  - Exchange of learning models instead of data

Agenda

- **Centralized Learning**
  - **Generative Adversarial networks**
  - **Dynamic sanitizing data through adversarial networks** *[ASIACCS' 21]*
- **Federated Learning**
  - **Personalization approaches**
  - **Limitations: Security / Privacy**
  - **Federated learning using personalized layers** *[MLSP' 21]*
  - **MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers** *[Middleware'22]*

Agenda

- **Centralized Learning**
  > Generative Adversarial networks
  - **Dynamic sanitizing data through adversarial networks** *[ASIACCS' 21]*
- **Federated Learning**
  - **Personalization approaches**
  - **Limitations / Privacy**
  - **Federated learning using personalized layers** *[MLSP' 21]*
  - **MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers** *[Middleware'22]*

Generative adversarial networks

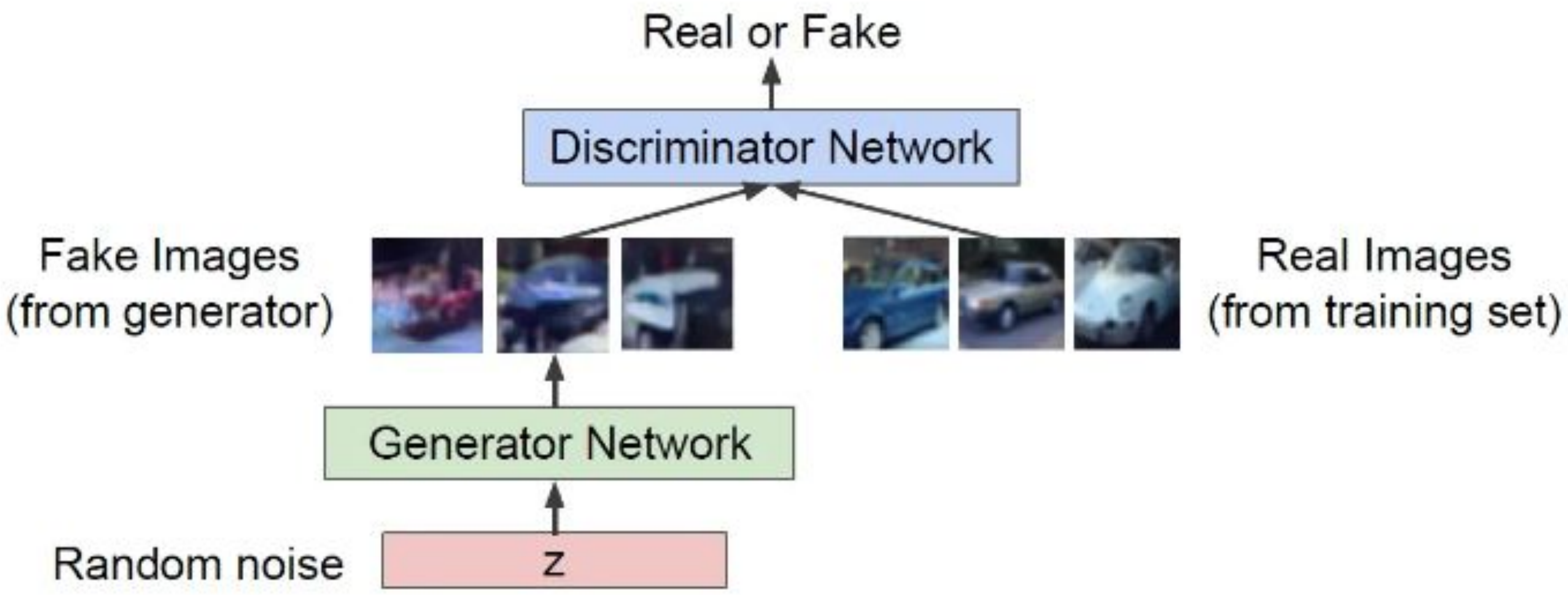
- GANs use **two neural networks that compete** with each other
- GANs can create **realistic**-looking computer-generated photos of people's faces
- **Imitating any data distribution**: GANs can imitate any data distribution, including images, text, and sound.

Realistic
yet
Fictional

Karras et al. *arXiv preprint arXiv:1710.10196* (2017).

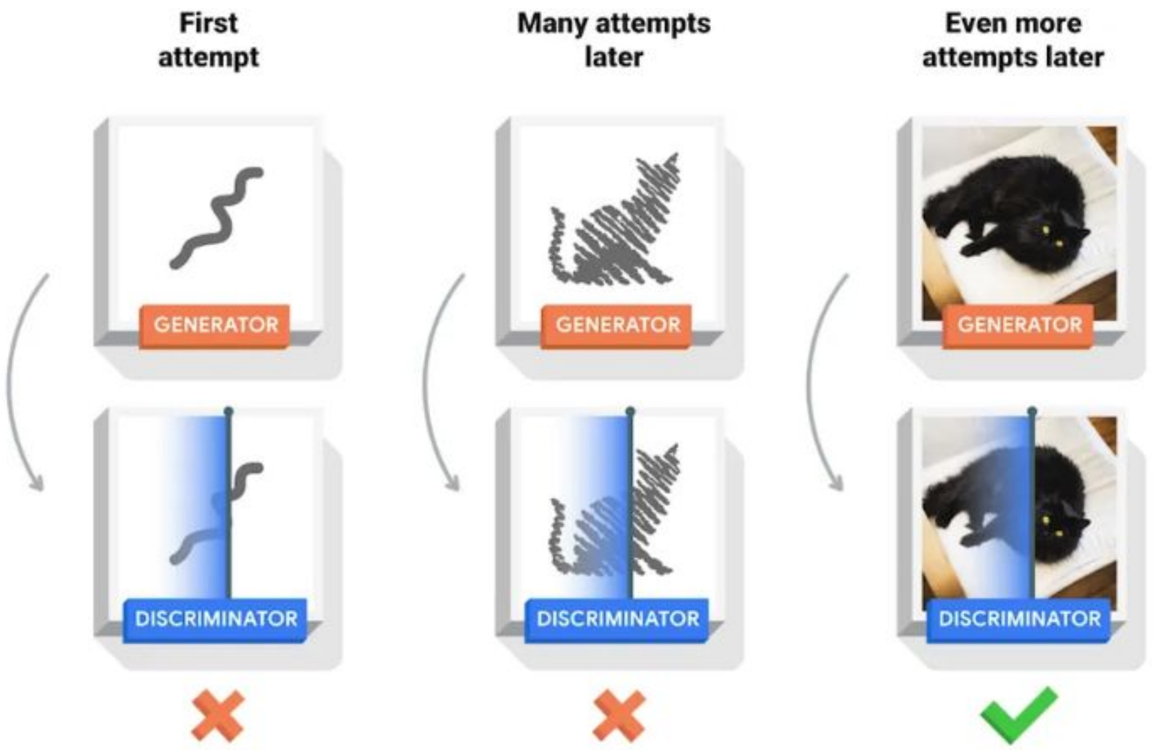
Basic structure of GANs

- Basic structure of a GAN, which consists of **two neural networks**
- The **generator** creates synthetic data from random noise
- The **discriminator** determines whether the data is real or fake.



Training process

- Generator and discriminator networks are jointly trained in a **two-player game formulation**
- The respective loss functions are then used to **update** the generator and discriminator networks **until they converge**



https://www.tensorflow.org/tutorials/generative/dcgan

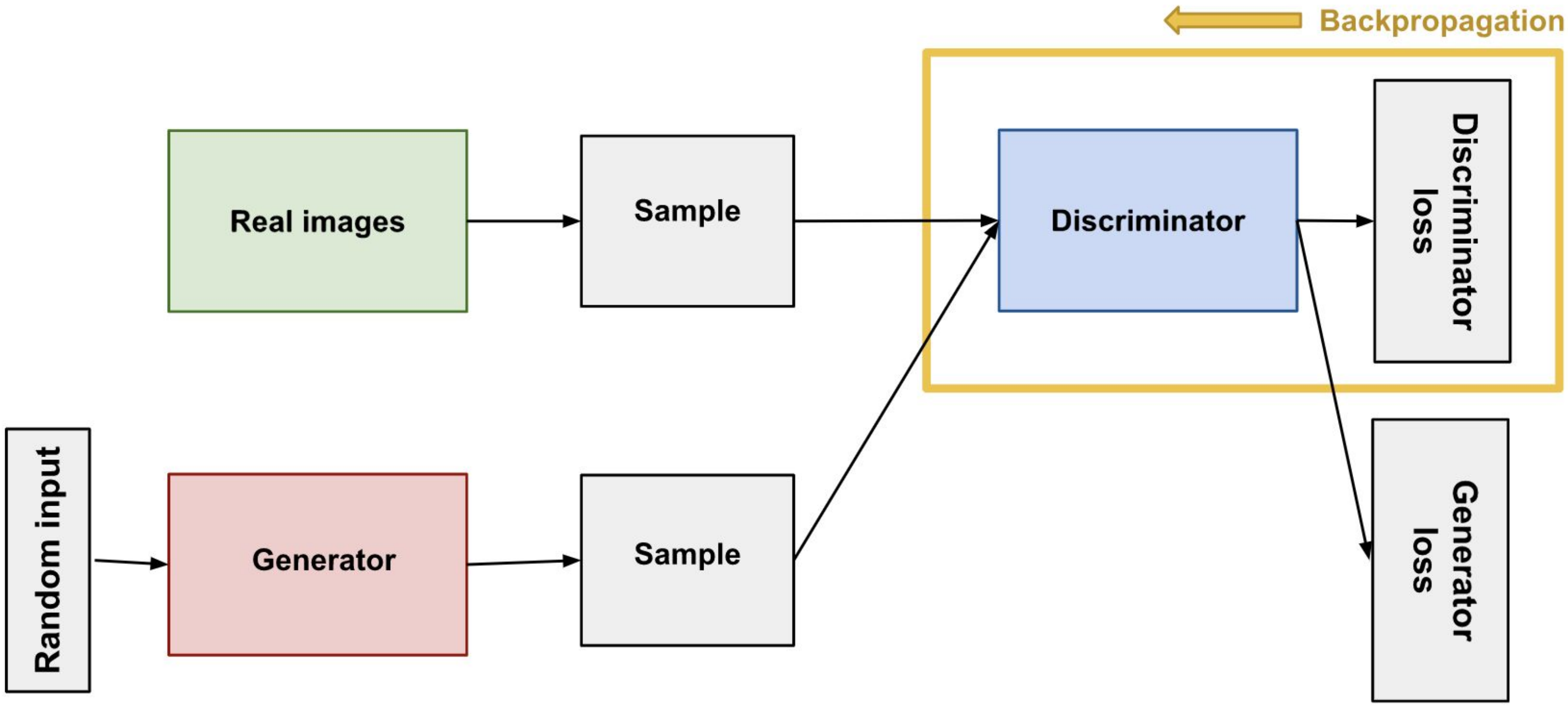
Loss functions

- **Generator** loss encourages the generator to create data that is similar to the real data
- **Discriminator** loss encourages the discriminator to correctly classify the data as real or fake.

$$\boxed{\min_{\theta_g}}\boxed{\max_{\theta_d}}\left[\mathbb{E}_{x\sim p_{\text{data}}}\log D_{\theta_d}(x)+\mathbb{E}_{z\sim p(z)}\log\left(1-D_{\theta_d}\left(G_{\theta_g}(z)\right)\right)\right]$$

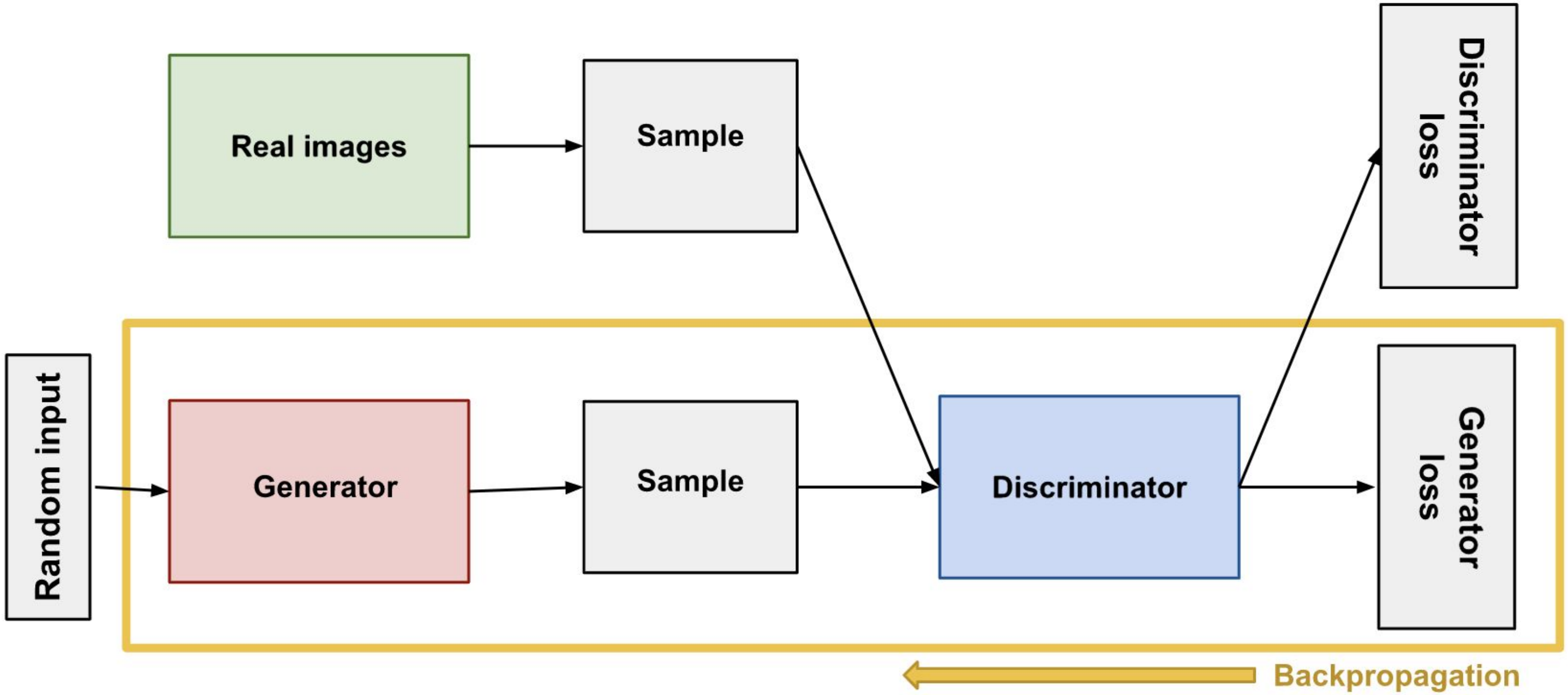
θ_g and θ_d are respectively the parameters of G and D

Discriminator training



1. The discriminator classifies both real data and fake data from the generator.
2. **Loss penalizes** the discriminator for misclassifying a real as fake or a fake as real
3. **Weights update** through backpropagation through the discriminator network

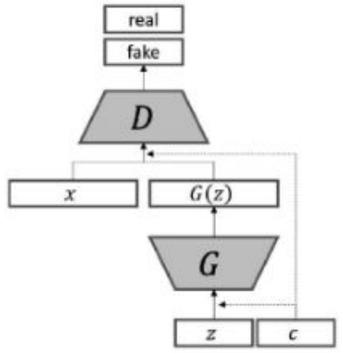
Generator training



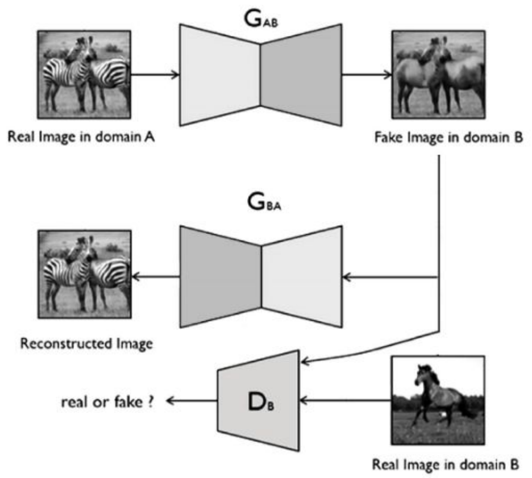
1. Produce generator output from sampled random noise
2. Get discriminator "Real" or "Fake" classification for generator output
3. **Calculate loss** from discriminator classification
4. **Backpropagate** through **both the discriminator and generator** to obtain gradients
5. Update generator weights

Variations of GANs

- **Conditional GANs**, which can generate specific types of data based on conditioning variables
    - Generator takes in **additional input**, label or conditional vector, to guide the generation process
    - Discriminator takes in the same additional input to judge the realism of the generated sample

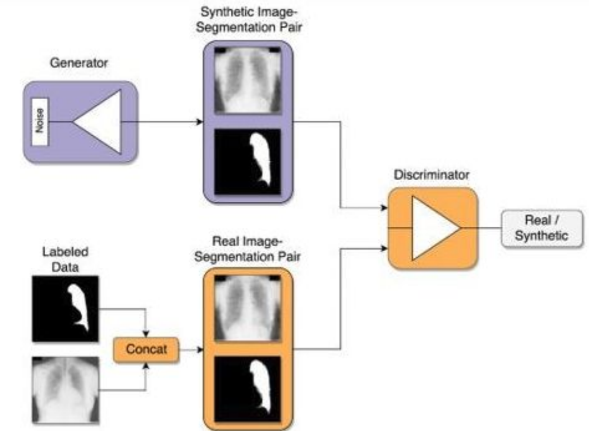


- **CycleGANs**, which can learn to transform data from one domain to another
    - CycleGAN uses **four neural networks**.
    - One generator is responsible for converting images from **domain A to B**
    - Other generator converts images from **domain B to A**
    - Each generator is paired with a discriminator that tries to distinguish between the generated images and the real images from the target domain



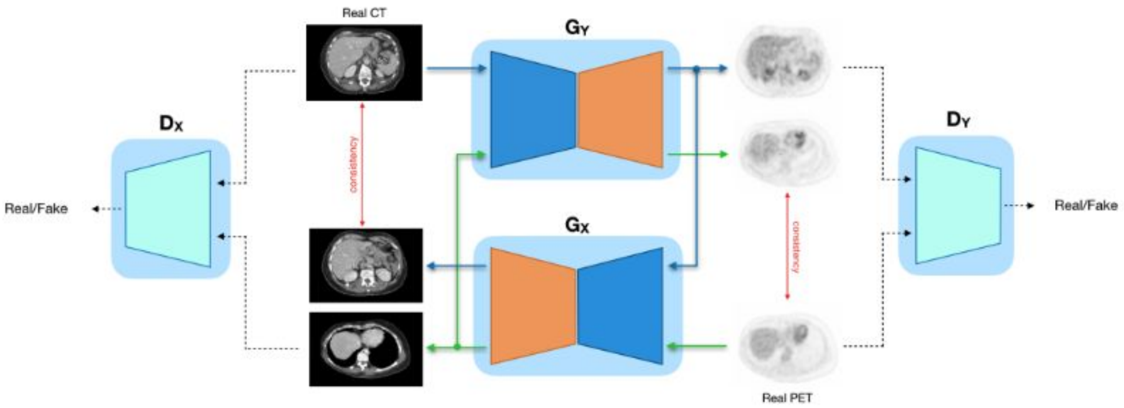
Variations of GANs

- **Conditional GANs**, which can generate specific types of data based on conditioning variables



Neff et al. *Proc. OAGM and ARW joint Workshop*. Vol. 3. 2017

- **CycleGANs**, which can learn to transform data from one domain to another



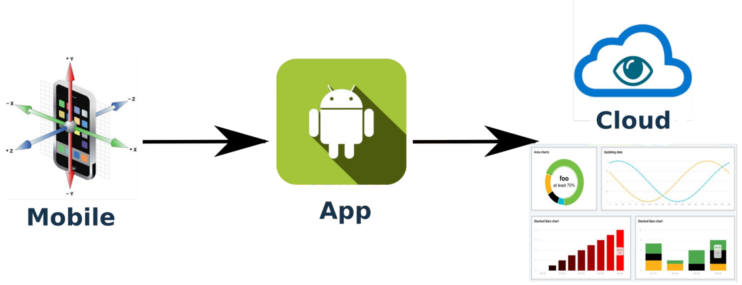
Santini et al. *16th International Symposium on Medical Information Processing and Analysis*. Vol. 11583. SPIE, 2020.

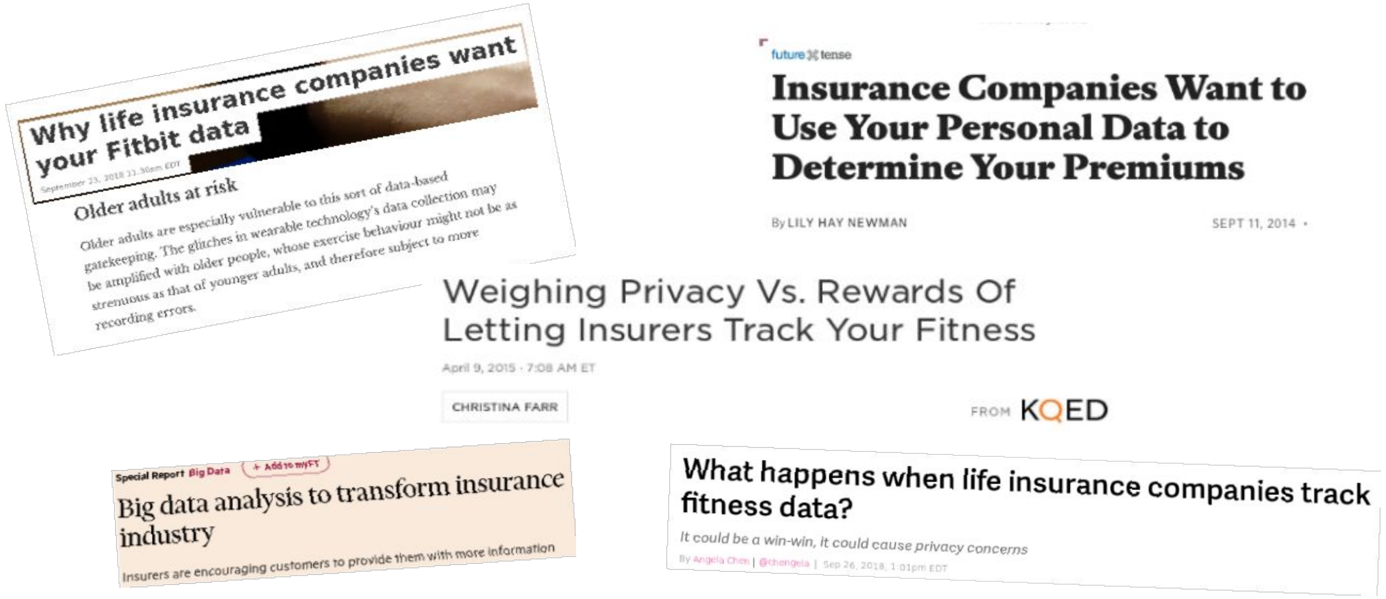
Agenda

- **Centralized Learning**
  - **Generative Adversarial networks**

  > Dynamic sanitizing data through adversarial networks [ASIACCS' 21]
- **Federated Learning**
  - **Personalization approaches**
  - **Limitations / Privacy**
  - **Federated learning using personalized layers** *[MLSP' 21]*
  - **MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers** *[Middleware'22]*

DYSAN: Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks





Objective: Sanitize motion sensor data to avoid unwanted sensitive inferences

25

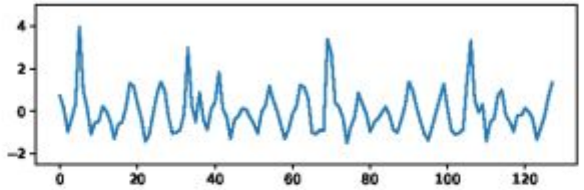
Only one scheme is not enough

Need a dynamic and personalized protection scheme to transform the data to avoid to leak unwanted sensitive attribute

- Heterogeneous users (including atypical ones)
- Varying activities (with different inference capabilities)

Objective

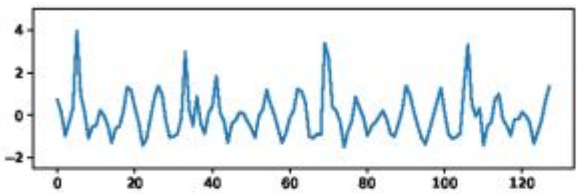
- $D=(X_{1,} \ldots, X_t)\, where\, X \in \{A, Y, S\}$
- $A = raw\, data$
- $Y = activity \in \{walking, jumping, \ldots\}$
- $S = sensitive\, attribute \in \{s, \bar{s}\}$
- $D \rightarrow \bar{D} = San_{\alpha,\beta,\lambda}(D) = (\bar{X}_1, \ldots, \bar{X}_t)$



- **Any model** $Disc$ **trained to predict** S **from** $\bar{A}$ **fails**
- **While** $Pred$ **trained on** $\bar{A}$ **maintain accuracy**
- **Minimized the data distortion between** D **and** $\bar{D}$

Objective

- $D=(X_{1,} \ldots, X_t)\, where\, X \in \{A, Y, S\}$
- $A=raw\, data$
- $Y=activity \in \{walking, jumping, \ldots\}$
- $S=sensitive\, attribute \in \{s, \bar{s}\}$
- $D \rightarrow \bar{D}=San_{\alpha,\beta,\lambda}(D)=(\bar{X}_1, \ldots, \bar{X}_t)$

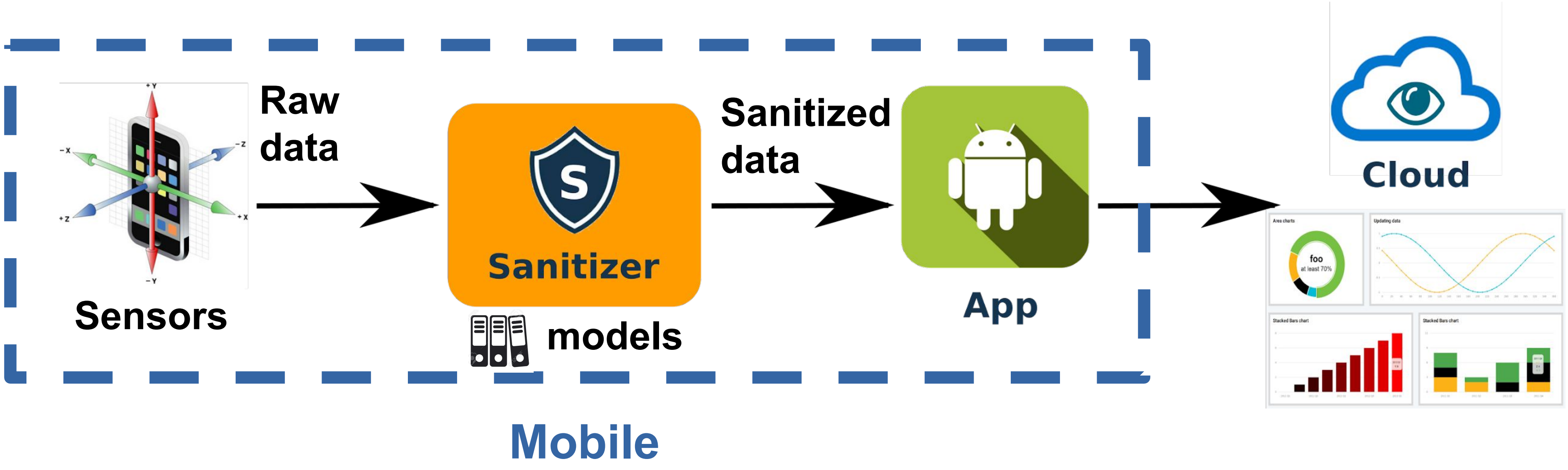


Dynamically adapt the transformation function to the current raw data

- **Any model $Disc$ trained to predict S from $\bar{A}$ fails**
- **While $Pred$ trained on $\bar{A}$ maintain accuracy**
- **Minimized the data distortion between D and $\bar{D}$**

DYSAN: Dynamic Sanitizer

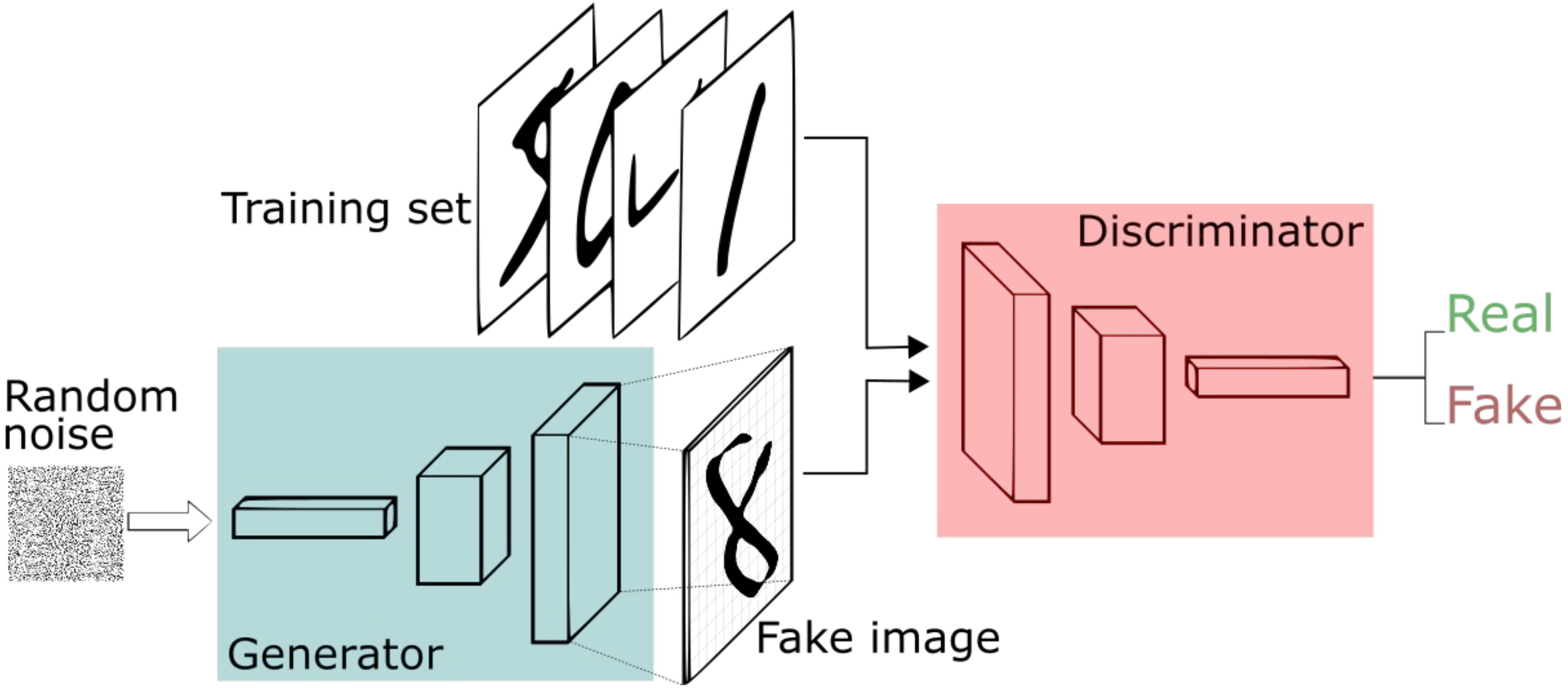
Overview



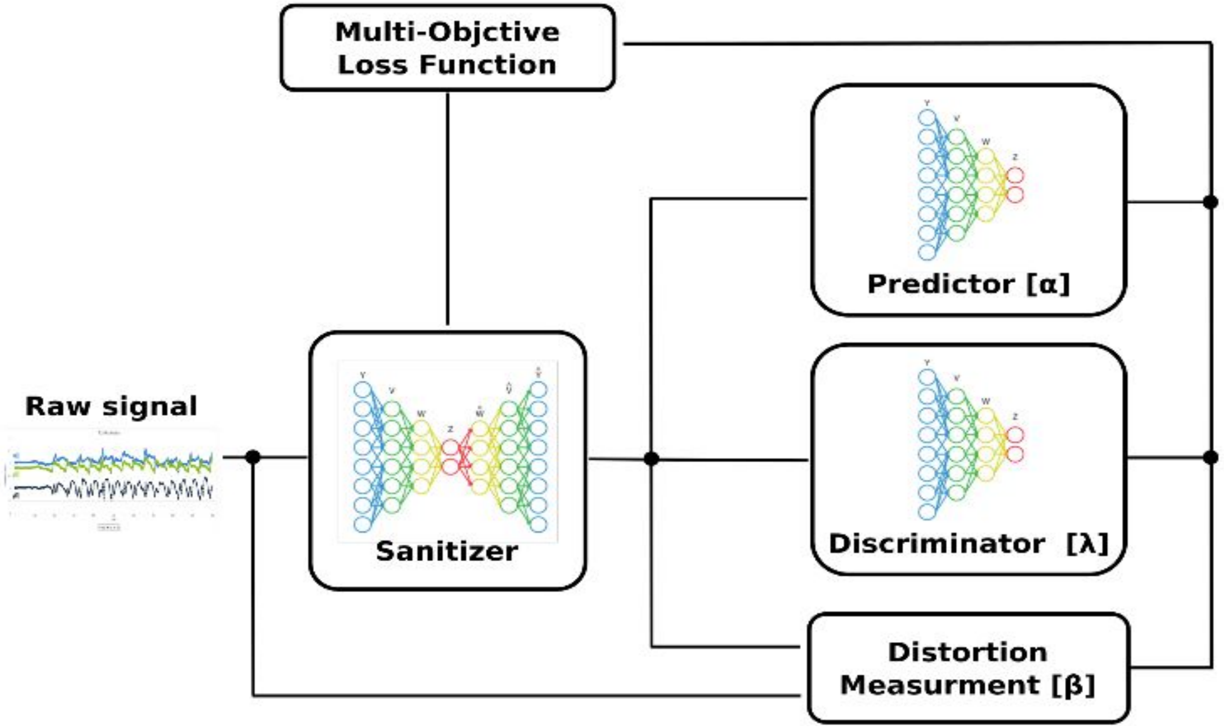
Two phases: a centralized training and an decentralized online phase

DYSAN – Training

Generative Adversarial Networks (GANs)



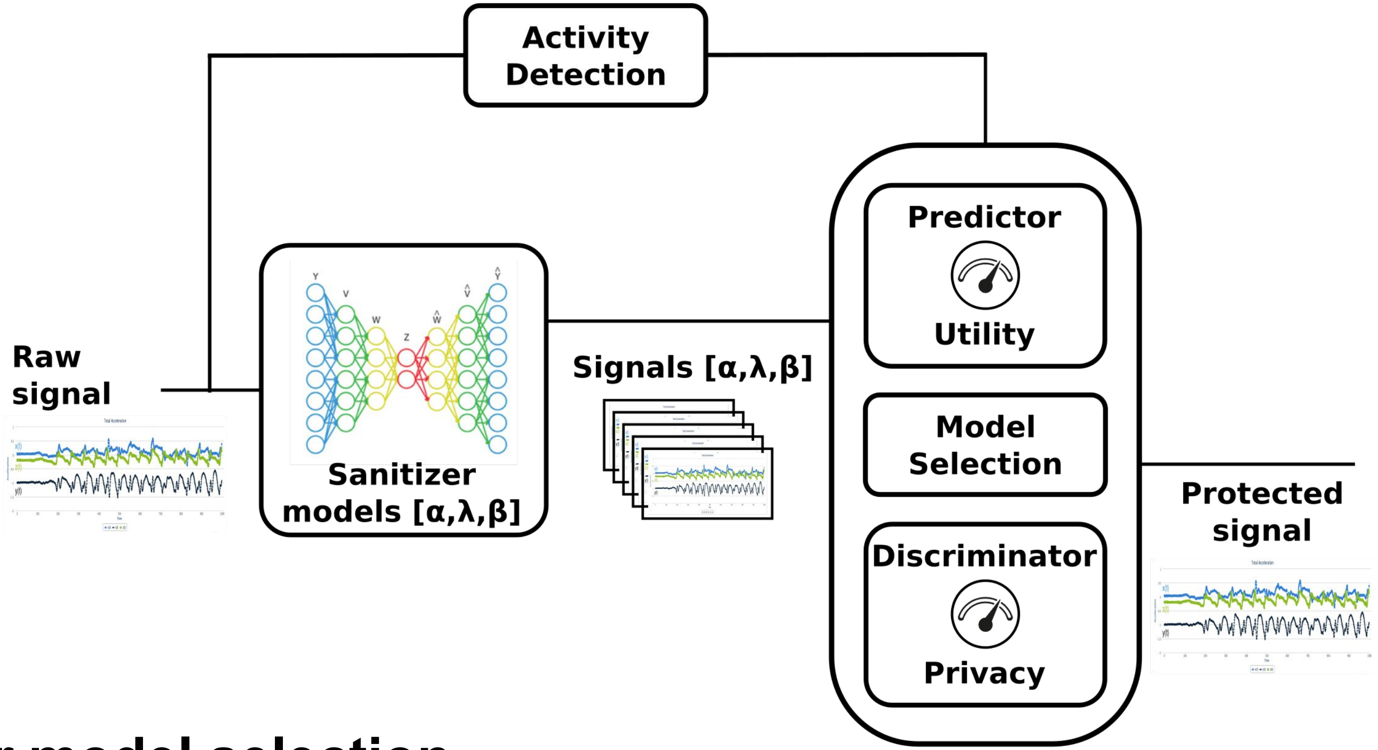
DYSAN – Training (offline)



$$J^{San}(X,\ S_{an},\ D_{isc},\ P_{red}) = \{\alpha * d_s(S, D_{isc}(S_{an}(X))),\\ \lambda * d_p(Y, P_{red}(S_{an}(X))),\\ \beta * d_r(X, S_{an}(X))\},$$

Build a model for each set of possible value for α, β, λ

DYSAN – Online (on the mobile)



Dynamic sanitizer model selection

- **Utility and privacy assessment of all models**
  - **Require a calibration step**
- **Selection of the model which provides the best privacy**

Experimental Setup

Datasets

- **MotionSense** (24 participants) – used to trained sanitizer models
- **MobiAct** (58 participants)

Baselines

- **ORF [1]: (design to avoid user re-identification)**
  - Analyse most relevant features from random forest
  - Normalize features correlated to gender
- **GEN [2]: Guardian-Estimator-Neutralizer**
  - Adversarial approach but without iterative process
  - Sensitive attribute learned on raw data
  - Do not consider data distortion
  - Hyper parameters static for all users

[1] Toward privacy in IoT mobile devices for activity recognition. Jourdan, Boutet, Frindel. Mobiquitous 2018.
[2] Protecting sensory data against sensitive inferences. Malekzadeh, Clegg, Cavallaro, Haddadi. W-P2DS 2018.

Experimental Setup

Baselines

- **Olympus [3]: (design to avoid user re-identification)**
  - Adversarial approach
  - Sensitive attribute learned on sanitized data
  - Do not consider data distortion
  - Hyper parameters static for all users
- **MSDA [4]: (design to avoid user re-identification)**
  - Adversarial approach
  - Sensitive attribute learned on sanitized data
  - Account data distortion
  - Hyper parameters static for all users

[3] Olympus: Sensor privacy through utility aware obfuscation. Raval, Machanavajjhala, Pan, PETS 2019.
[4] Mobile sensor data anonymization. Malekzadeh. Clegg,Cavallaro, Haddadi. IoTDI 2019.

Experimental Setup

Metrics

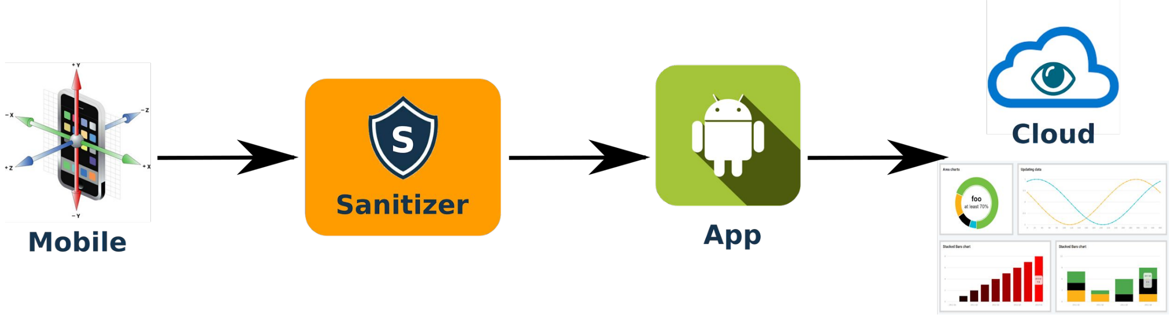
- Utility
  - Accuracy of the prediction of the activity recognition [1,0]
  - Number of steps detected from the signals
  - Impact of the number of sanitizer models
- Privacy
  - Accuracy of inferring the sensitive attribute [1,0] (accuracy of 0.5 = random guess)
  - Uniqueness of the model selection
- Performance
  - Overhead / computational cost
  - Energy consumption

Methodology

- Transfert learning (training on Motionsense and testing on MobiAct)
- Average over 10 repetitions of each experiment
- Done on a GPU/CPU computing farm

Utility and Privacy trade-off

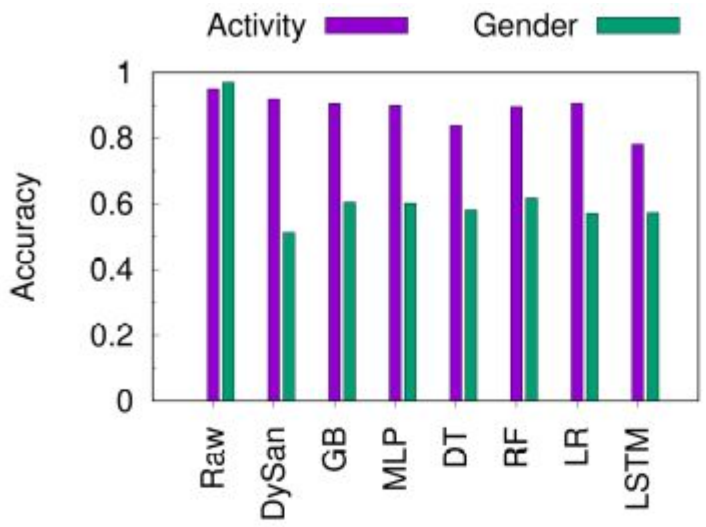
DYSAN: Inferences from sanitized data



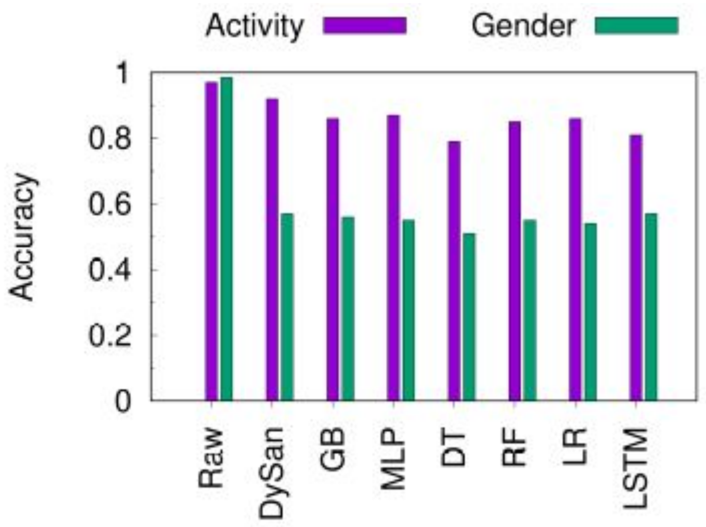
- GB (Gradient Boosting)
- MLP (Multi-Layer Perceptron)
- DT (Decision Tree)
- RF (Random Forest)
- LR (Logistic Regression)
- DySan Discriminator and Predictor

Utility and Privacy trade-off

DYSAN: Inferences from sanitized data



Motionsense



MobiAct

- **Protection** is needed
- Whatever the classifier, **small decrease of the activity** detection while **drastically reducing the inference** of the gender

Utility and Privacy trade-off

Detection of the number of steps

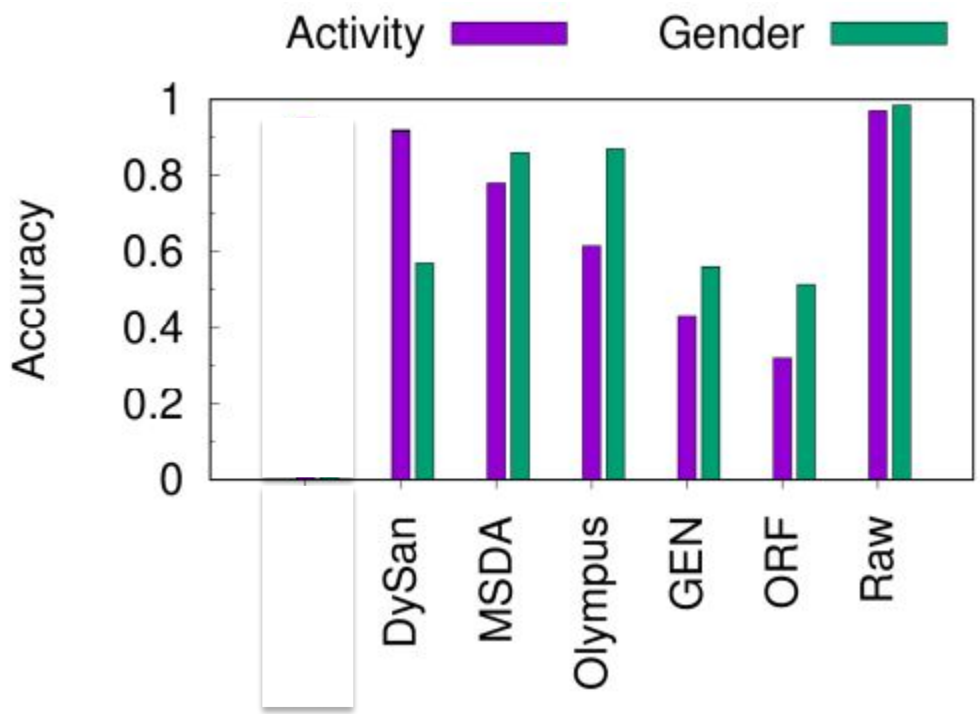
	Steps	Dynamic Time Warping [1]
Raw data	14387	-
DYSAN	**15321 (+6.49 %)**	12.96
GEN	12817 (-12.25%)	14.28
Olympus	23658 (+64.44%)	156.03
MSDA	18624 (+29.45%)	23.37

DYSAN keeps **relevant information in the signal**
(less than 5% of errors for steps detection)

[1] D.J.Berndt and J.Clifford, Using Dynamic Time Warping to Find Patterns in Time Series, AAAIWS, 359-370, 12, (1994)

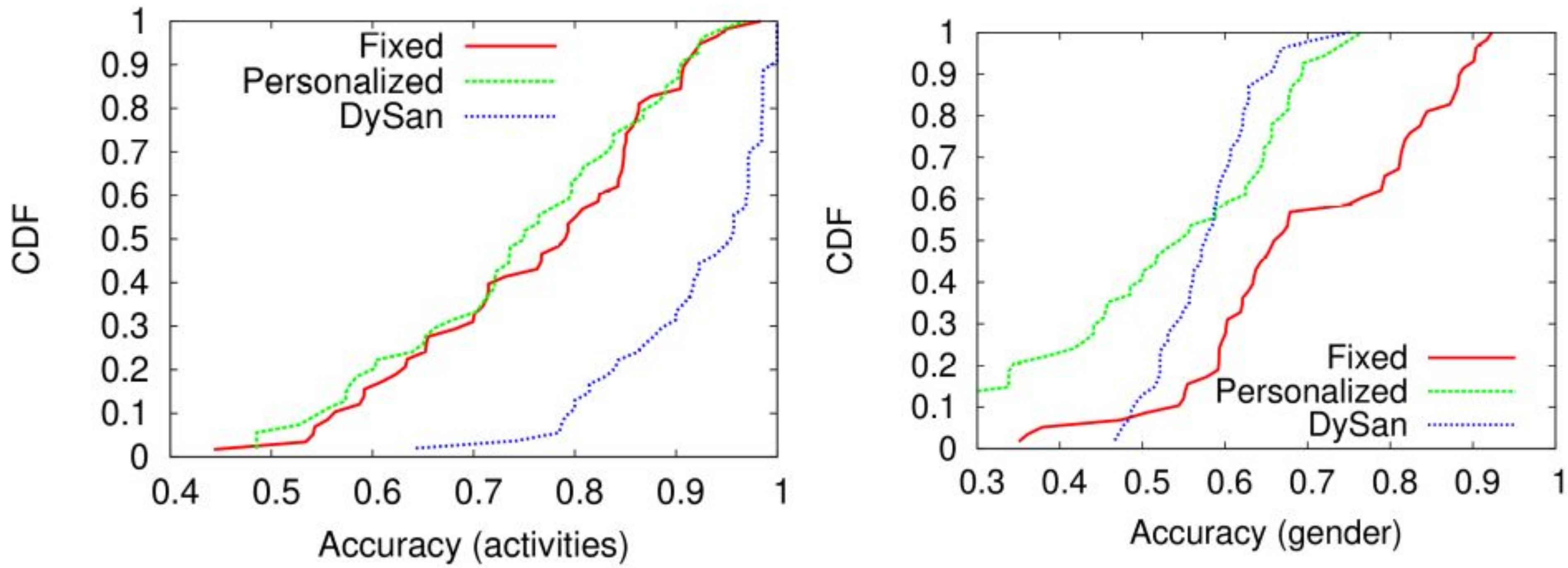
Utility and Privacy trade-off

Comparison against baselines (MobiAct)



DYSAN provides the **best utility-privacy trade-off**

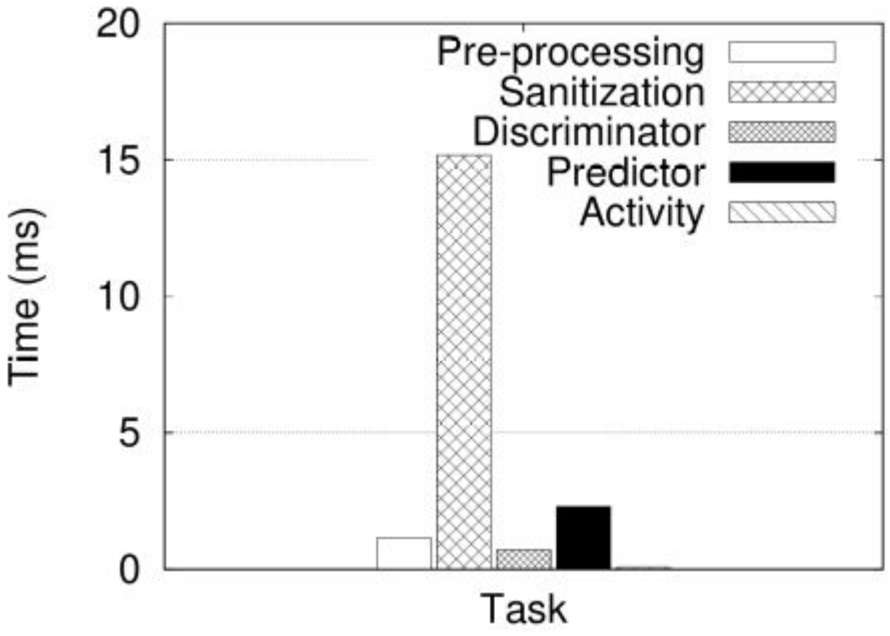
Dynamic Sanitizer Model Selection (MobiAct)



- DYSAN does not significantly impact the activity recognition
- **By dynamically selecting** the best sanitizer model, DYSAN greatly **improves the protection against gender inference**

Performance (overhead)

- Xiaomi Redmi Note 7
- Qualcom Snapdragon 660
- 3 GB of memory
- Pytorch 1.6



- Trade-off between the overhead and the number of considered sanitizing models

Take away

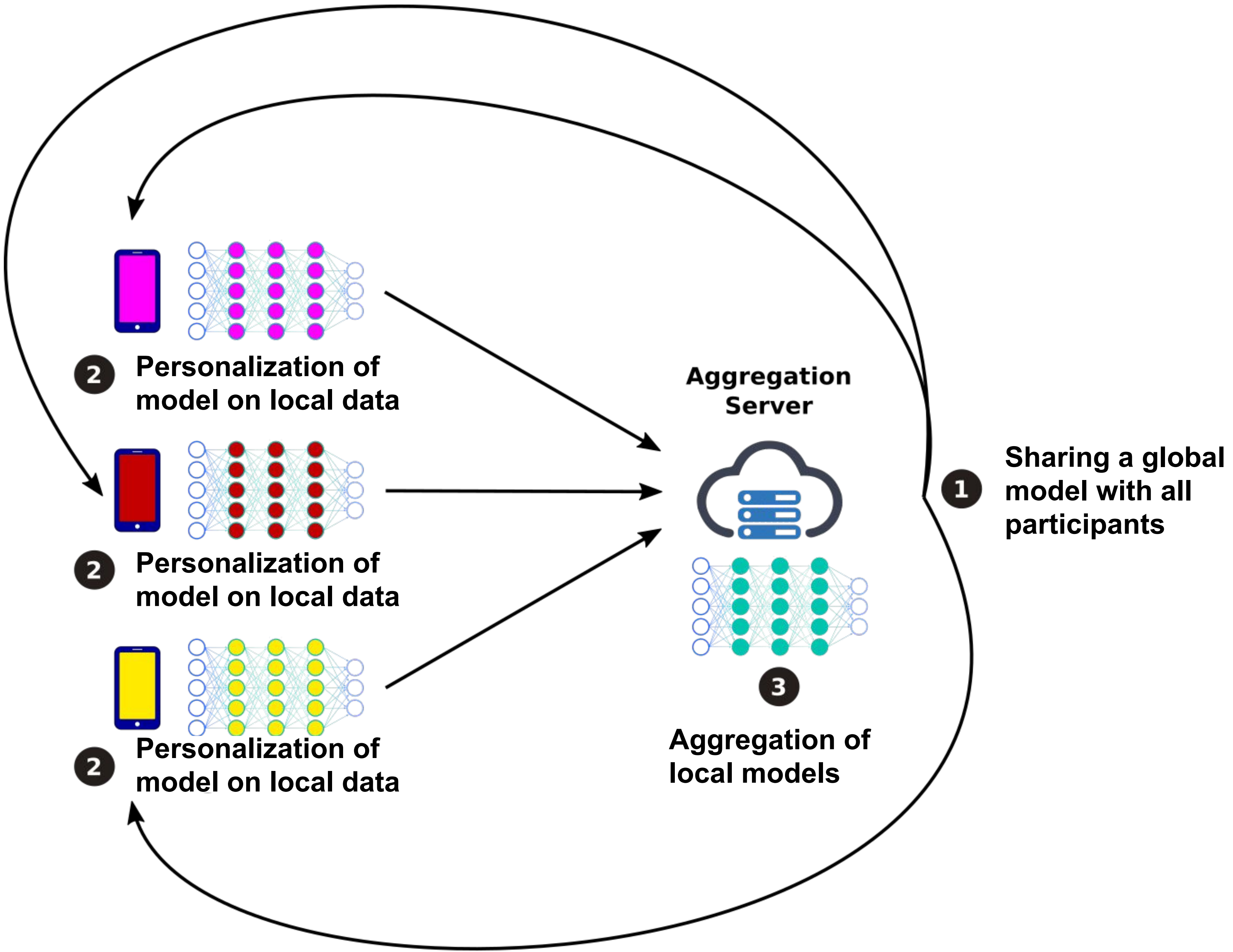
Dynamic sanitizer model selection successfully adapts the protection to incoming raw data

- **Prevent unwanted inference** of sensitive information
- **Preserve useful information** for activity recognition and other estimator of physical activity monitoring
- **Compliant with mobile phone capability**

Agenda

- Centralized Learning
  - Generative Adversarial networks
  - Dynamic sanitizing data through adversarial networks *[ASIACCS' 21]*
- **Federated Learning**
  - Personalization approaches
  - Limitations / Privacy
  - Federated learning using personalized layers *[MLSP' 21]*
  - MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers *[Middleware'22]*

Federated Learning (FL)



Local learning

- We consider a set of $\mathbf{C}$ parties (clients, users or data silos)
- Each party c holds a dataset $\mathbf{D_c}$
- We denote by $\boldsymbol{\theta}$ the local model parameters (e.g. the weights of a neural network)

$$\min_{\theta_1, \ldots, \theta_c \in \mathbb{R}^d} F(\theta) := \frac{1}{C} \sum_{c=1}^{C} f_c(\theta_c)$$

The resulting models may **not achieve good generalization** as the **number of examples** that the local models are exposed to are **limited**

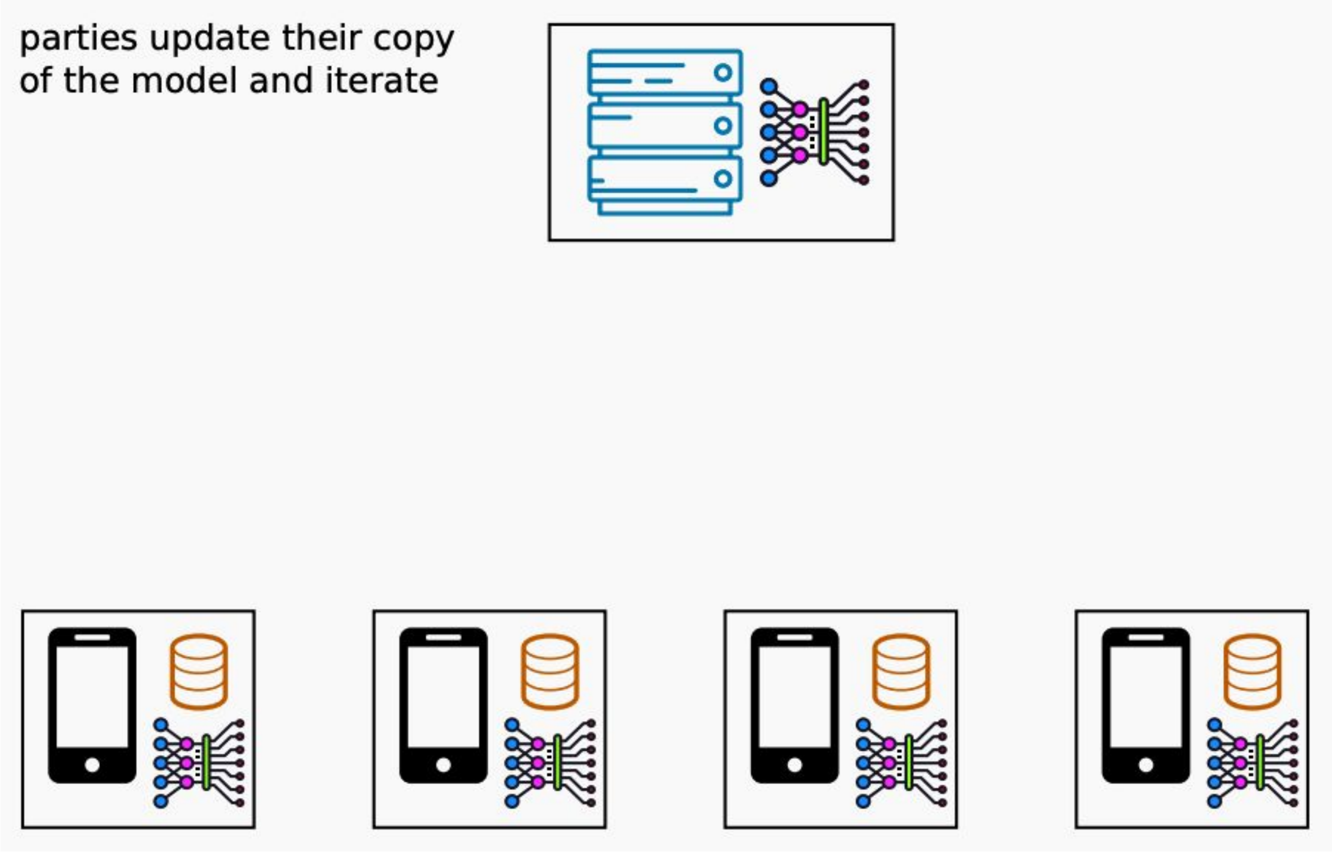
Baseline FL algorithm (FedAVG)

- We consider a set of $\boldsymbol{C}$ parties (clients, users or data silos)
- Each party c holds a dataset $\boldsymbol{D_c}$
- We denote by $\boldsymbol{w}$ the model parameters (e.g. the weights of a neural network)
- We want to find parameters that minimize an overall prediction loss :

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{C} \sum_{c=1}^{C} f_c(w)$$

$$f_c(w) := \mathbb{E}_{(x,y)\sim D_c}\left[f_c(w; x, y)\right]$$

Baseline FL algorithm (FedAVG)



Algorithm FedAvg (server-side)

Parameters: client sampling rate ρ

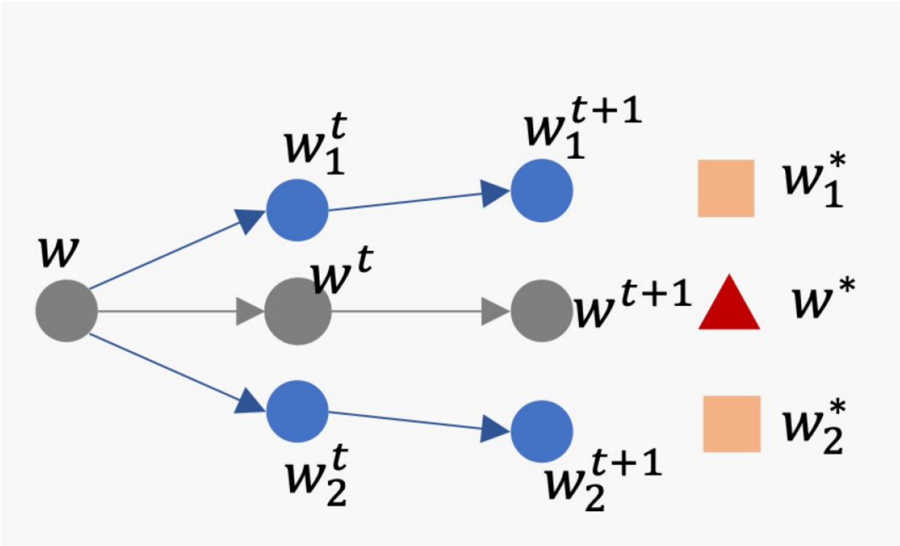
initialize θ
for each round $t = 0, 1, \ldots$ **do**
  $\mathcal{S}_t \leftarrow$ random set of $m = \lceil \rho K \rceil$ clients
  for each client $k \in \mathcal{S}_t$ in parallel **do**
    $\theta_k \leftarrow \mathrm{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \sum_{k \in \mathcal{S}_t} \frac{n_k}{n} \theta_k$

Algorithm ClientUpdate(k, θ)

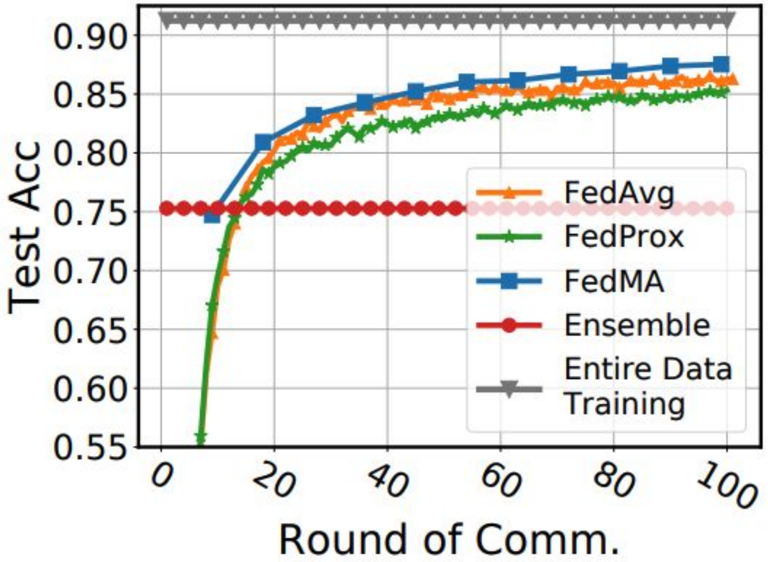
Parameters: batch size B, number of local steps L, learning rate η

for each local step $1, \ldots, L$ **do**
  $\mathcal{B} \leftarrow$ mini-batch of B examples from $\mathcal{D}_k$
  $\theta \leftarrow \theta - \frac{n_k}{B} \eta \sum_{d \in \mathcal{B}} \nabla f(\theta; d)$
send θ to server

Baseline FL algorithm (FedAVG)



Tan et al. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

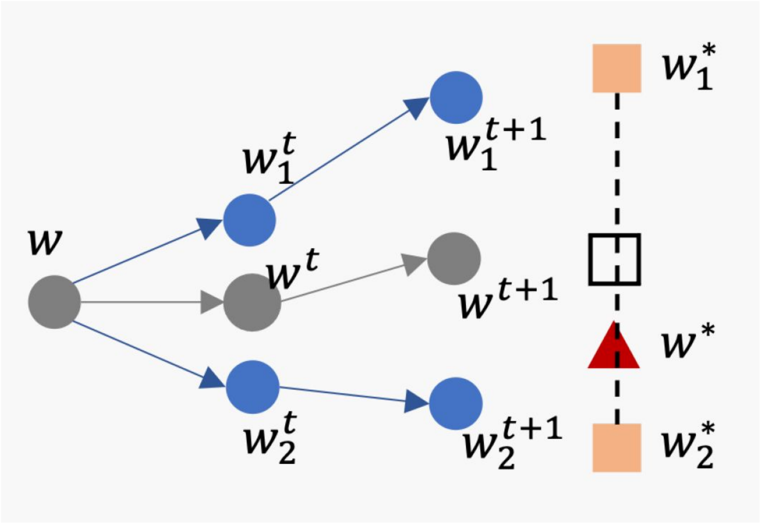


Wang, Hongyi, et al. *arXiv preprint arXiv:2002.06440* (2020).

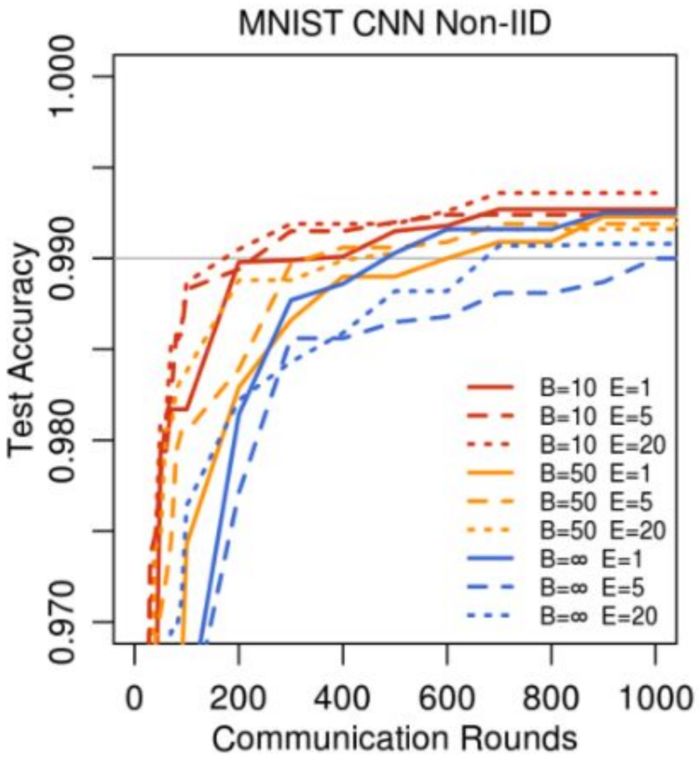
- When **IID data**, FedAVG efficiently tends towards the **centralized model**
- FedAVG does better than a **collection of independent local models**

Baseline FL algorithm (FedAVG)

Non IID



Tan et al. *IEEE Transactions on Neural Networks and Learning Systems* (2022).



McMahan et al. PMLR, 2017

B: batch size
E : local epochs

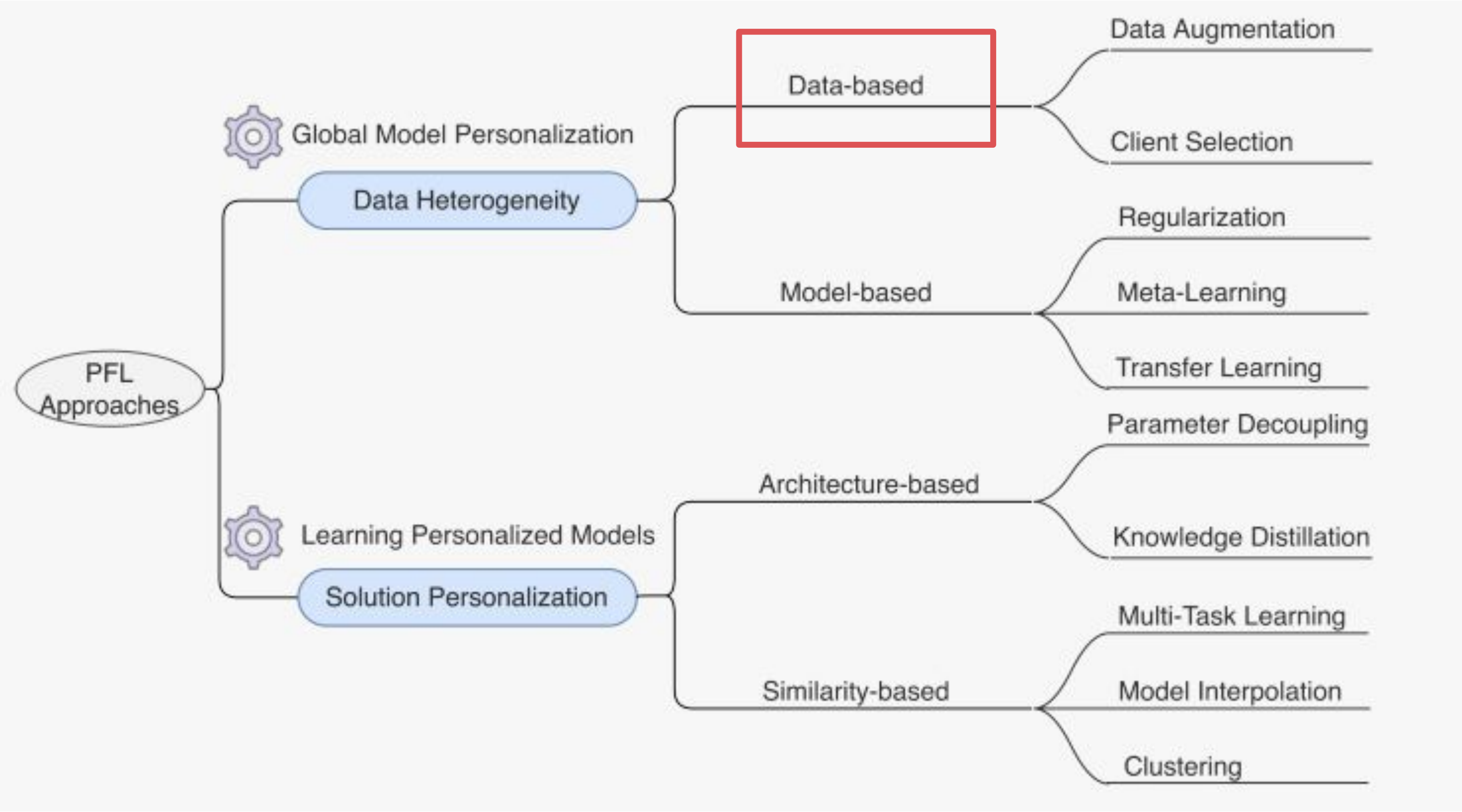
- When **non IID data**, FedAVG suffers from **client drift**
- To avoid this drift, use **fewer local updates and/or smaller learning rates**, which hurts convergence

Agenda

- Centralized Learning
  - Generative Adversarial networks
  - Dynamic sanitizing data through adversarial networks *[ASIACCS' 21]*
- **Federated Learning**
  - **> Personalization approaches**
  - Limitations / Privacy
  - Federated learning using personalized layers *[MLSP' 21]*
  - MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers *[Middleware'22]*

Global model personalization

Data-based approaches: reduce the statistical heterogeneity of client data distributions



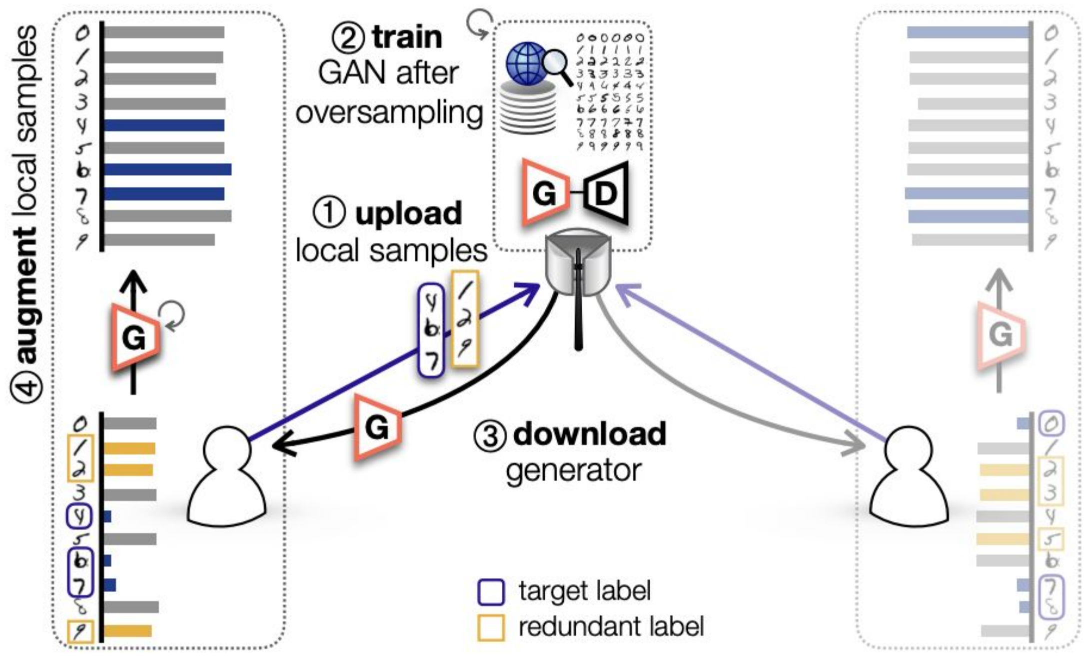
Tan et al. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

Data augmentation

- **Data augmentation** requires some form of **data sharing** or a **proxy dataset** representative of the overall data distribution
- **FAug** trains a **GAN model** in the FL server, which generates additional data for each client to produce an IID dataset

IID

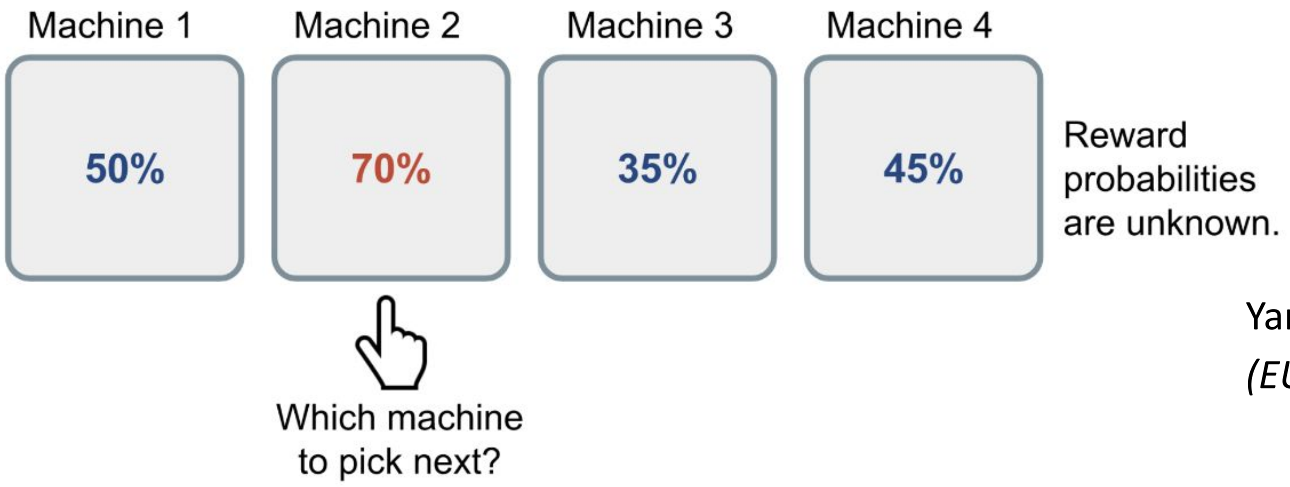
Non IID



Jeong et al. *arXiv preprint arXiv:1811.11479* (2018).

Client selection

- **Client selection** help to make the data more similar across all clients
- **Multi-Armed Bandit** choose **which clients should participate** in each round of training
- Selects clients subset with minimal class imbalance based on the estimated **local class distributions**

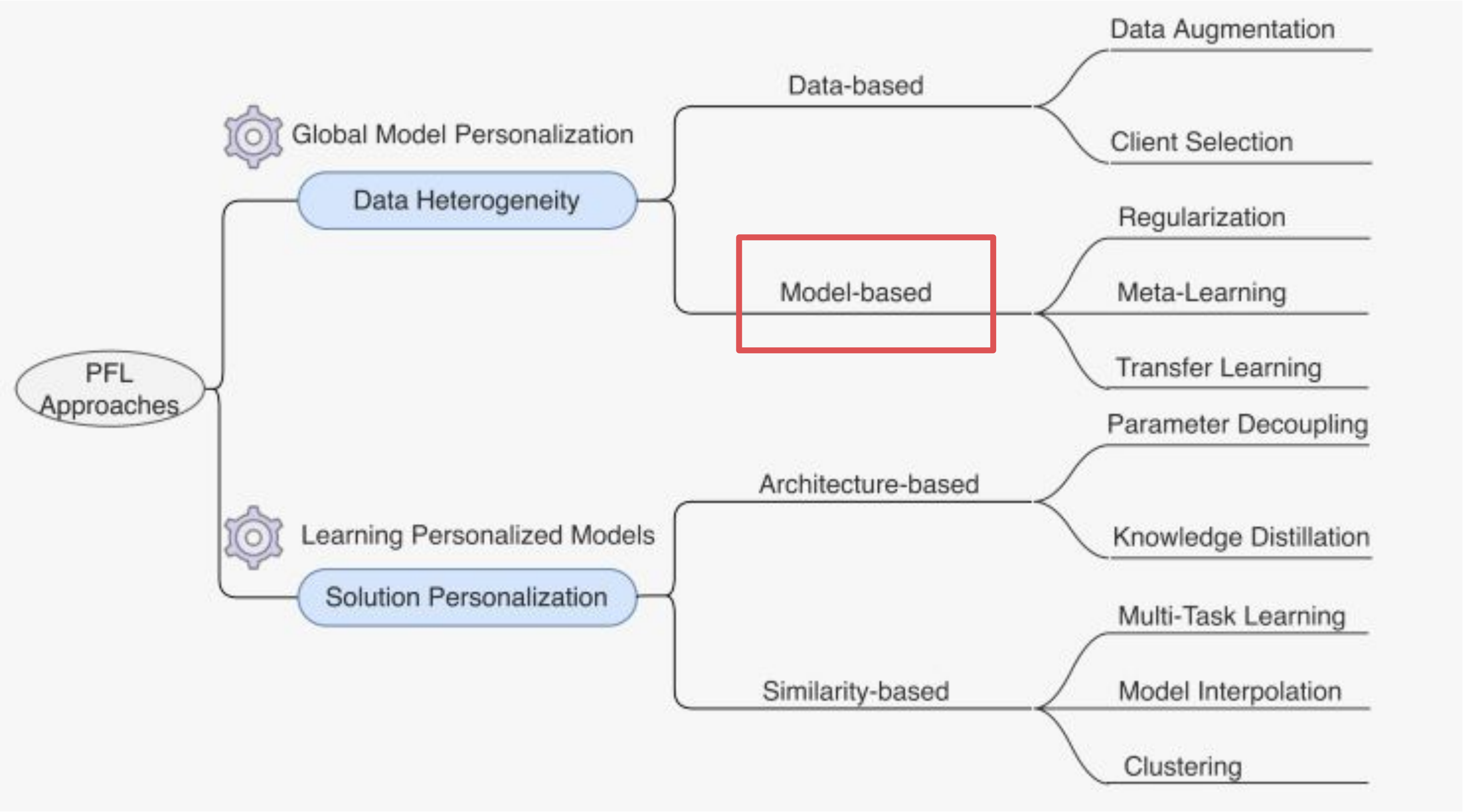


$$p_i \propto e^{-\alpha||\nabla Q(w_i)-\bar{\nabla} Q(w)||^2}$$

Yang et al. *29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

Global model personalization

Model-based approaches : learning a strong global FL model for future personalization on individual clients



Tan et al. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

Regularized local loss

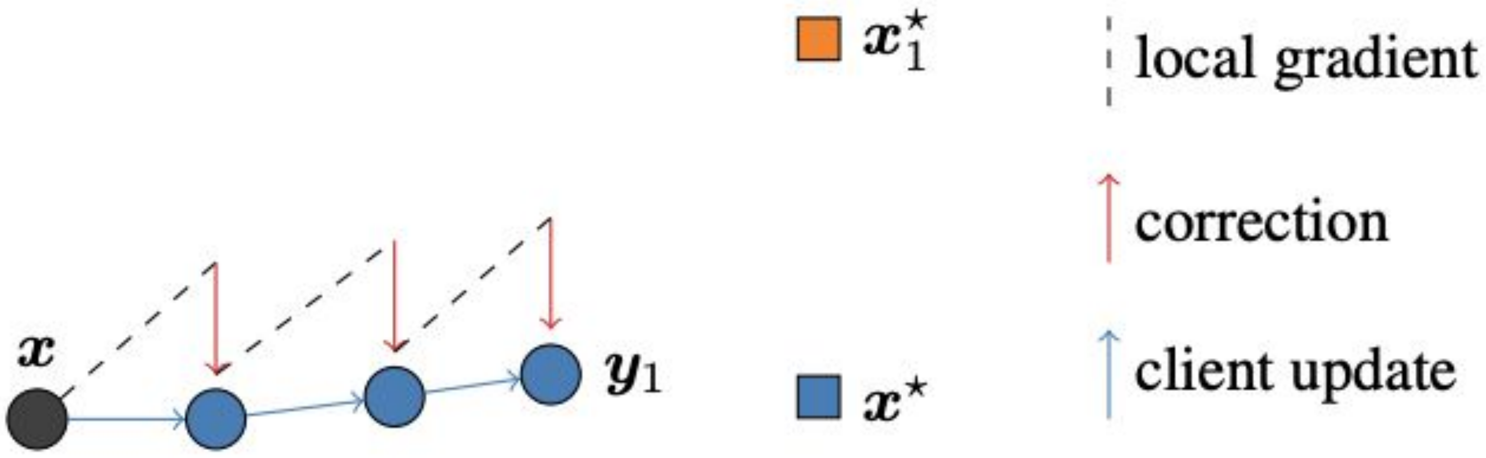
- We denote by **w** the **global** model parameters
- We denote by **θ** the **local** model parameters
- Instead of just minimizing the local function $\mathbf{f_c()}$, each client **c** minimizes the following objective:

$$\min_{\theta \in \mathbb{R}^d} h_c(\theta; w) := f_c(\theta) + \left\{ l_{reg}(\theta; w) \right\}$$

where $\mathbf{l_{reg}(\theta;w)}$ is the regularization loss, which is a function of the global model ***w*** and the local model $\boldsymbol{\theta}_\mathbf{c}$ of client **c**

Regularized local loss

- **SCAFFOLD** uses the **difference** between the **update directions** of the global (v) and local (vc) models, (v-vc), which is added as a component of the local loss function to **correct local updates**



Karimireddy et al. *International Conference on Machine Learning*. PMLR, 2020.

Meta-learning

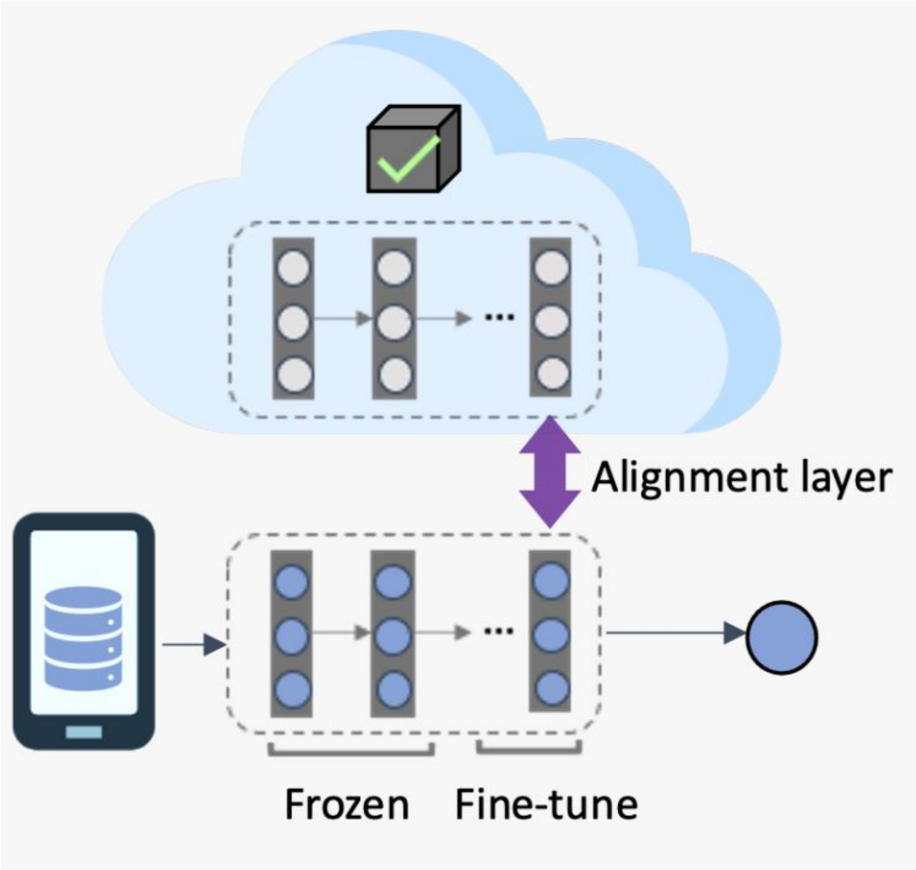
- Meta-learning improves learning through **exposure to a variety of tasks**
- Per-FedAvg is a variant of FedAvg to learn a good initial global model that performs well on a new heterogeneous task after **it is updated with a few steps of gradient descent**

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{C} \sum_{c=1}^{C} \{ f_c(w - \alpha \nabla f_c(w)) \}$$

Dinh et al. *Advances in Neural Information Processing Systems* 33 (2020).

where $\alpha > 0$ is the step size.
The cost function is written as the average of meta-functions F1, · · · , Fc

Transfer learning

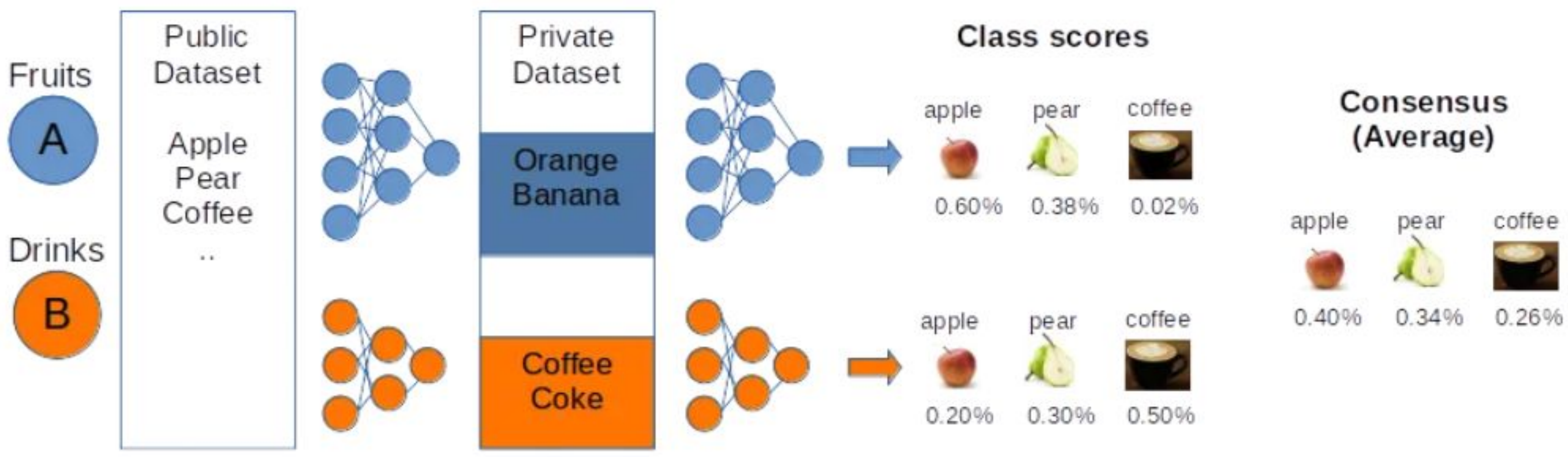


Chen et al. *IEEE Intelligent Systems* 35.4 (2020): 83-93.

- Lower layers of the global model are reused directly in the local models
- Other layers of the local model are fine-tuned with the local data

Knowledge distillation

- Knowledge distillation communicates learned knowledge with **class scores**
- In **FedMD**, the central server then computes and updates the **consensus**, which is the **average of the class scores**
- The updated consensus is the baseline for further federated training



Li et al. *arXiv preprint arXiv:1910.03581* (2019).

Take away

Method	Advantages	Disadvantages
Data augmentation	Pre-processing before FL training procedure	● Possibility of privacy leakage ● May require a representative proxy dataset
Client selection	Modifies client selection strategy of FL training procedure	● Increasing computational overhead ● May require a representative proxy dataset
Regularization	Slight modification of FedAvg algorithm	Single global model setup
Meta-learning	Optimizes global model for fast client personalization	● Single global model setup ● Needs computing of second-order gradients
Transfer Learning	Reduces the impact of local data in the model	● Single global model setup ● May require a representative proxy dataset
Knowledge distillation	High degree of architecture design for each client	Difficult to determine the optimal architecture design

Agenda

- Centralized Learning
  - Generative Adversarial networks
  - Dynamic sanitizing data through adversarial networks *[ASIACCS' 21]*
- **Federated Learning**
  - Personalization approaches

  > Limitations: Security / Privacy
  - Federated learning using personalized layers *[MLSP' 21]*
  - MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers *[Middleware'22]*

Massive deployment of ML

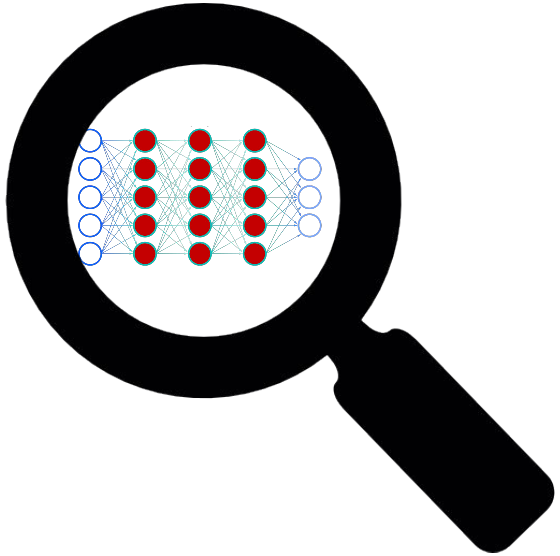
Rise many questions
- Utility
- Privacy
- Security
- Fairness
- Explainability
- Energy Footprint

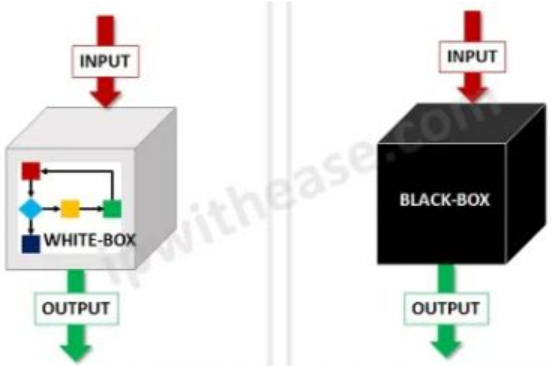
Challenge:
address globally these questions

Limitations: security / privacy

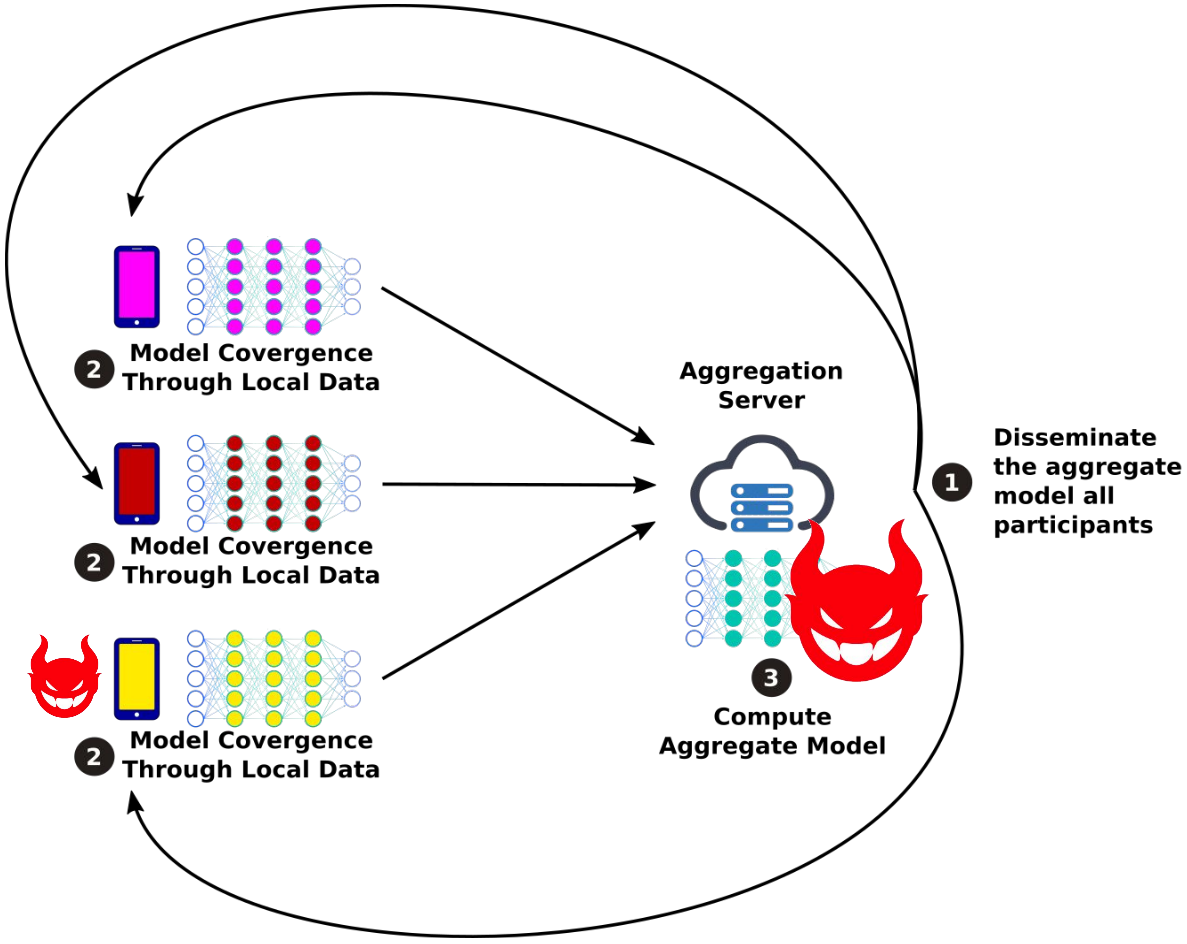
Limitations: security / privacy

Limitations: security / privacy





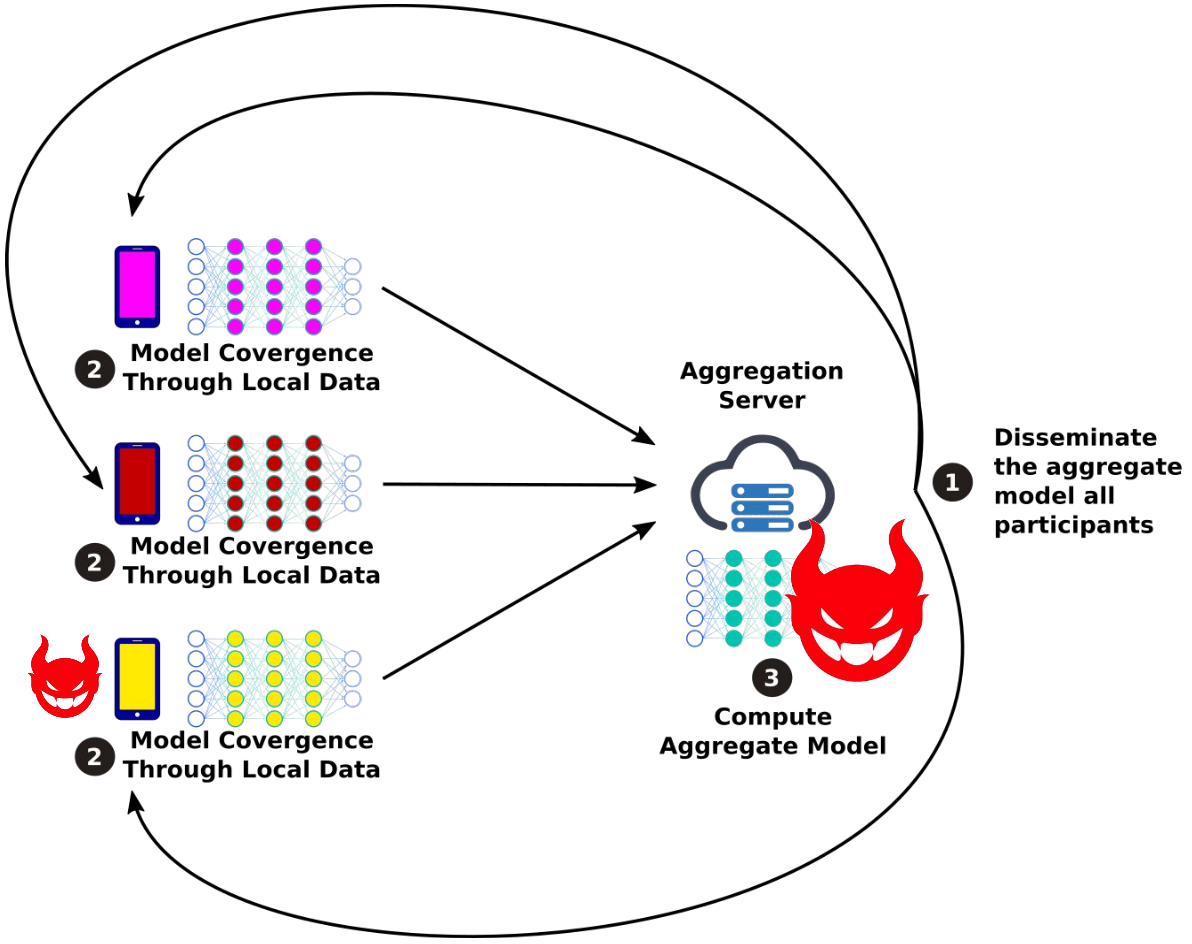
Limitations: security / privacy



Federated Learning

- **Poisoning / Backdoors**
- **Privacy leakage**
- **Give more power to participants**

Limitations: security / privacy



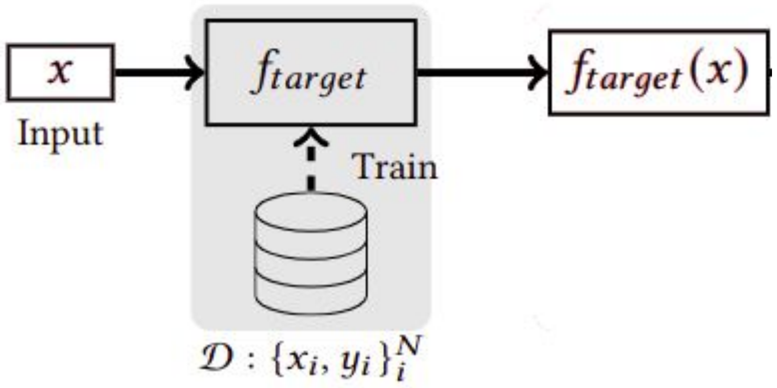
Federated Learning

- Poisoning
- Privacy leakage
- Give more power to participants

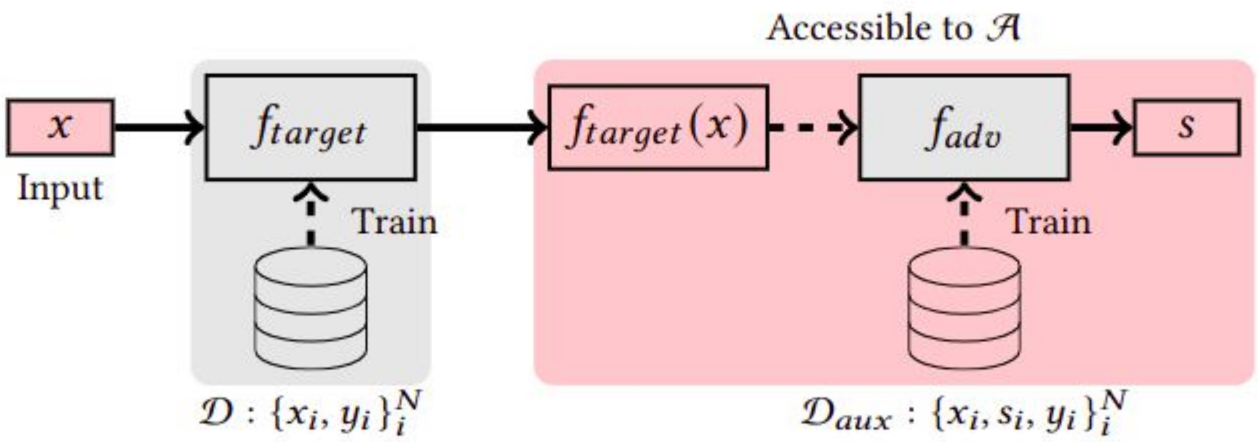
Countermeasures

- Perturbation (e.g., differential privacy)
  - Drastically reduces accuracy
- Crypto (e.g., secure aggregation)
  - Important overhead

Data Privacy: Attribute Inference Attacks

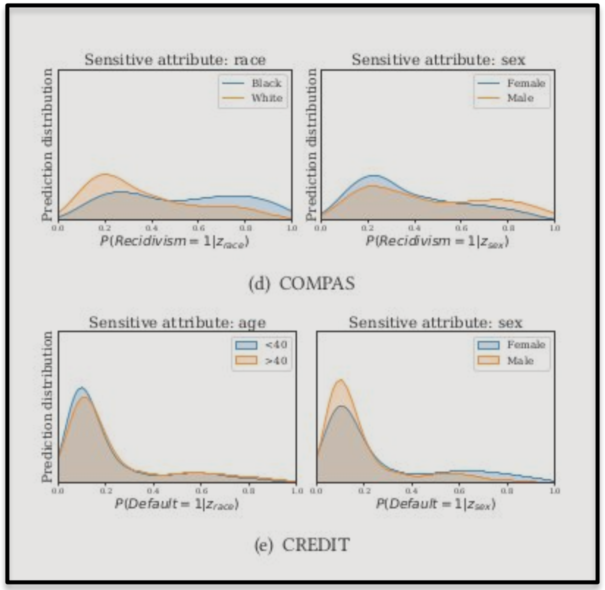


Data Privacy: Attribute Inference Attacks



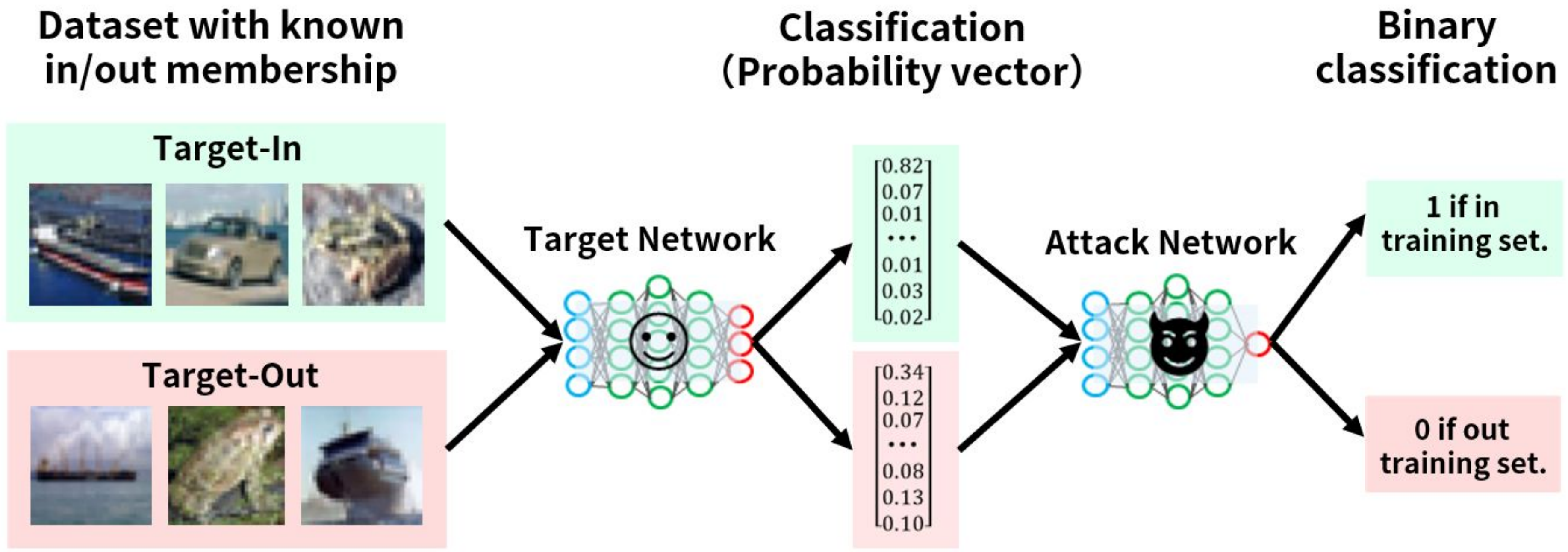
Adversary: Use ML attack model (f_{adv}) to infer sensitive attributes

- Exploit distinguishability in predictions for different values of sensitive attribute [6]

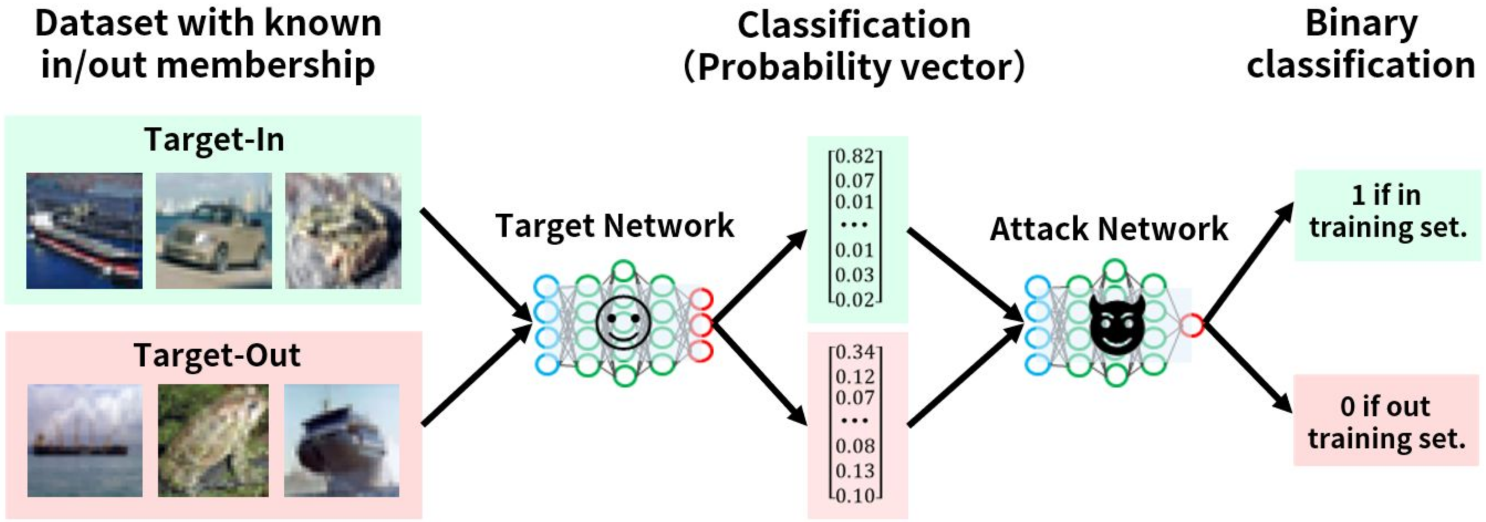


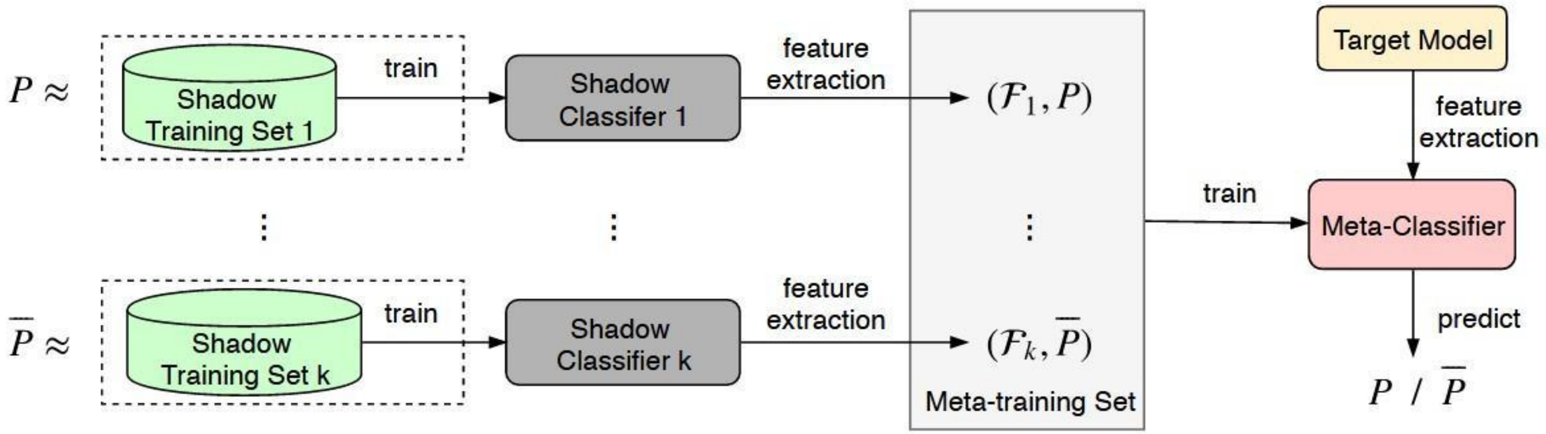
[6] Song and Shmatikov. *Overlearning Reveals Sensitive Attributes*. ICLR'20.

Data Privacy: Membership inference attack



Data Privacy: Membership inference attack



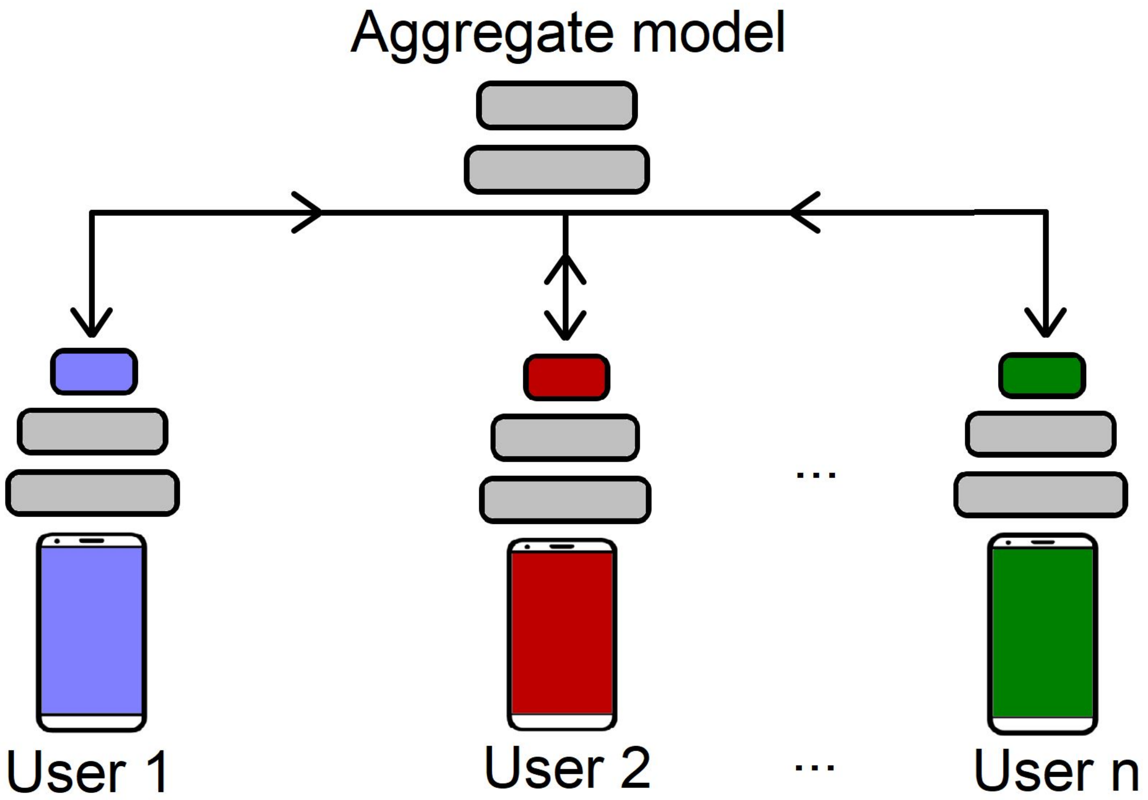


Agenda

- Centralized Learning
  - Generative Adversarial networks
  - Dynamic sanitizing data through adversarial networks *[ASIACCS' 21]*
- **Federated Learning**
  - Personalization approaches
  - Limitations / Privacy

  > Federated learning using private layers [MLSP' 21]
  - MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers *[Middleware'22]*

Federated Learning using private layers



Objective: minimizing the information exchanged with the aggregation server while improving the personalization

Experimental setup

Datasets

- **MotionSense:** 24 participants, 4 activities, 20 minutes of data per subject
- **MobiAct:** 58 participants, 4 activities, 6 minutes of data per subject

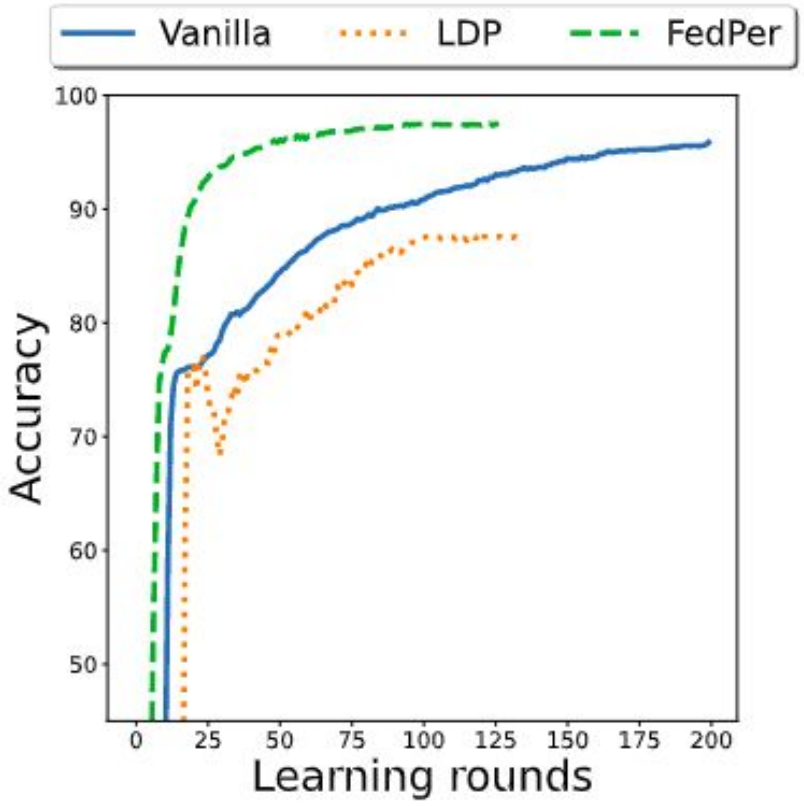
Baslines

- **Vanilla:** the most common FL scheme using SGD training on the device and average aggregation
- **FedPer:** FL scheme using private personalized layers
- **LDP:** FL scheme with an introduction of noise following a Gaussian distribution to the local model

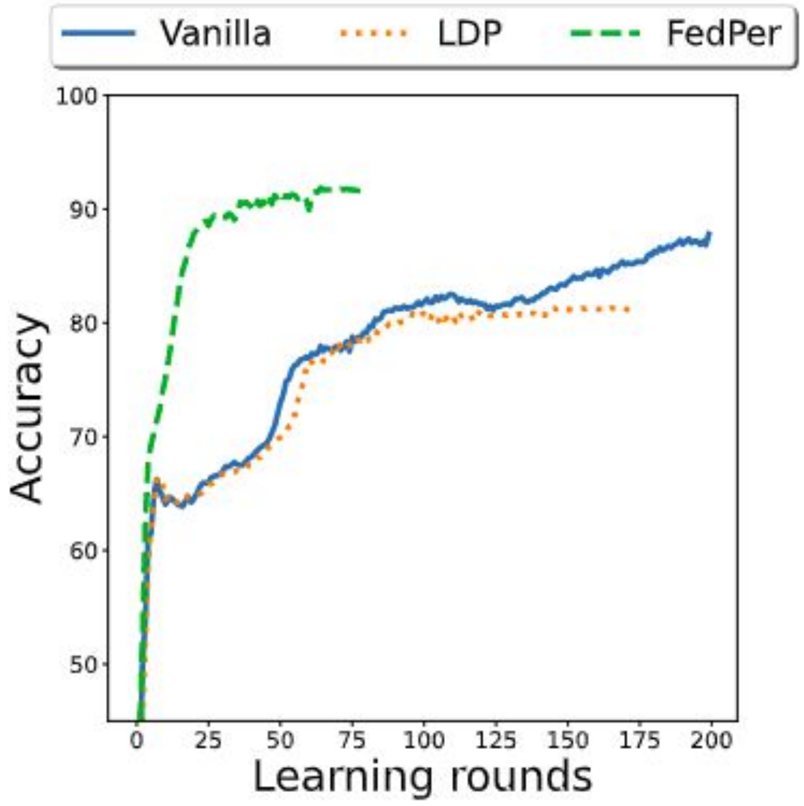
Metrics

- **Utility:** activity recognition
- **Privacy:** Gender and BMI (Body Mass Index) attribute inference, membership inference

Utility evaluation



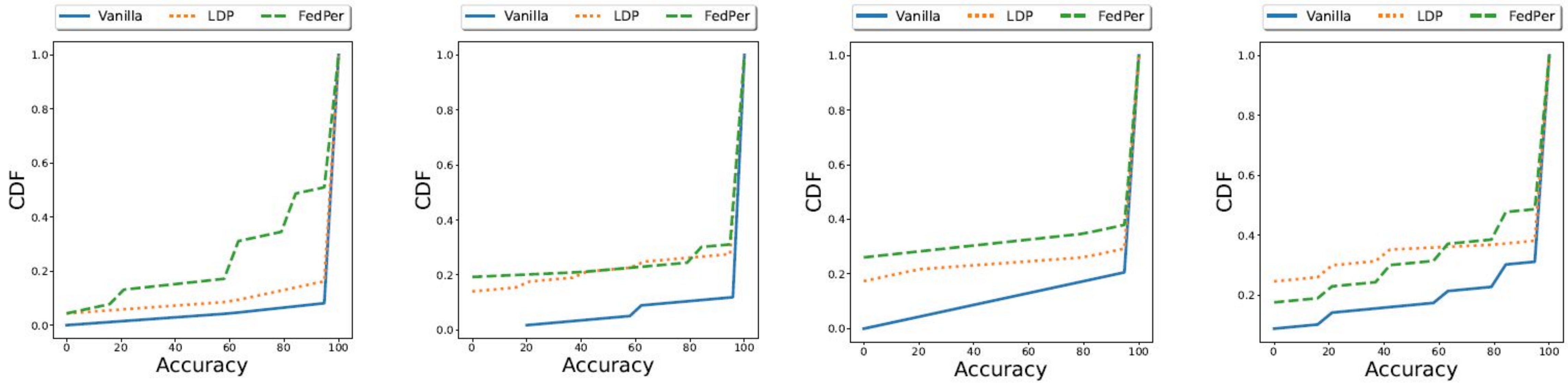
(a) MotionSense



(b) MobiAct

By using personalized layers instead of aggregated information, the learning is drastically speeds up

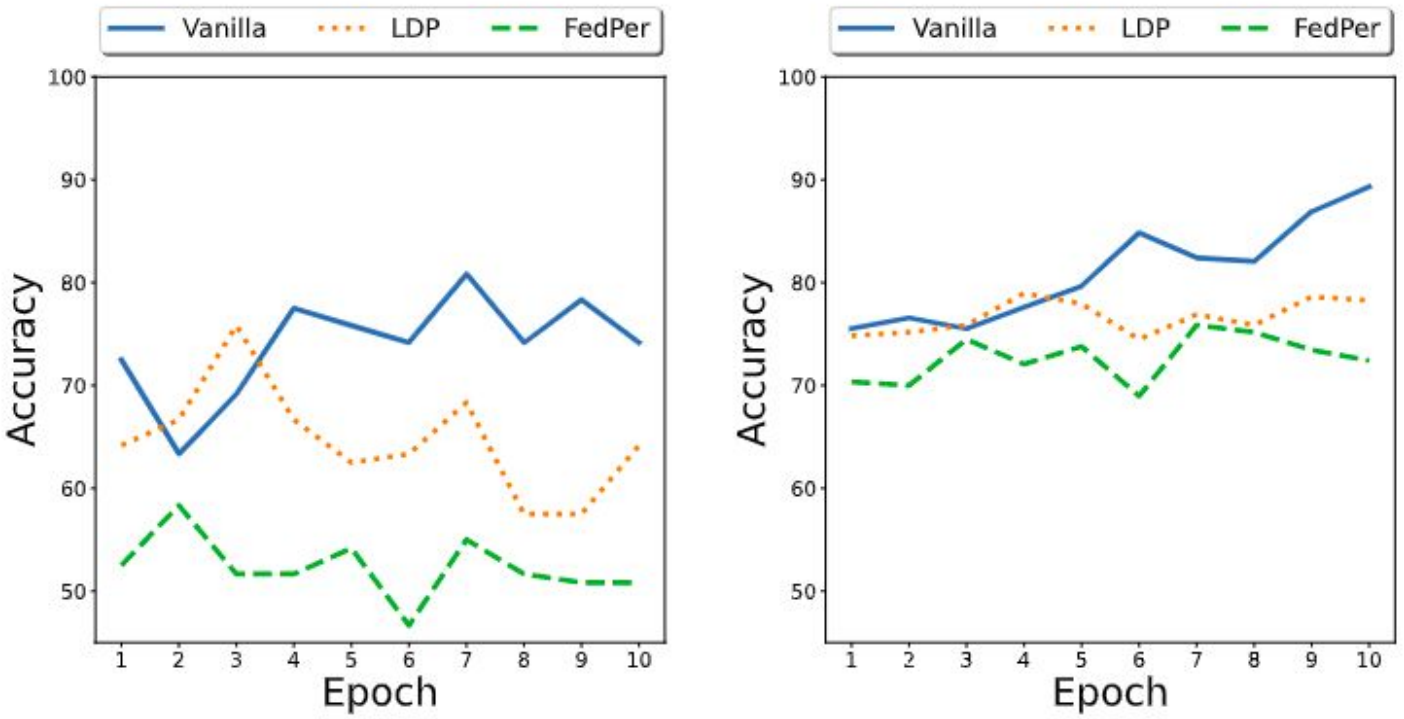
Privacy: attribute inference



(a) Gender - MotionSense (b) Gender - MobiAct (c) BMI - MotionSense (d) BMI - MobiAct

FedPer and LDP increase the number of users with a small inference accuracy

Privacy: membership inference



(a) MotionSense (b) MobiAct

FedPer and LDP significantly decrease the accuracy of the membership inference attack compare to Vanilla method

FL using private layers - Take away

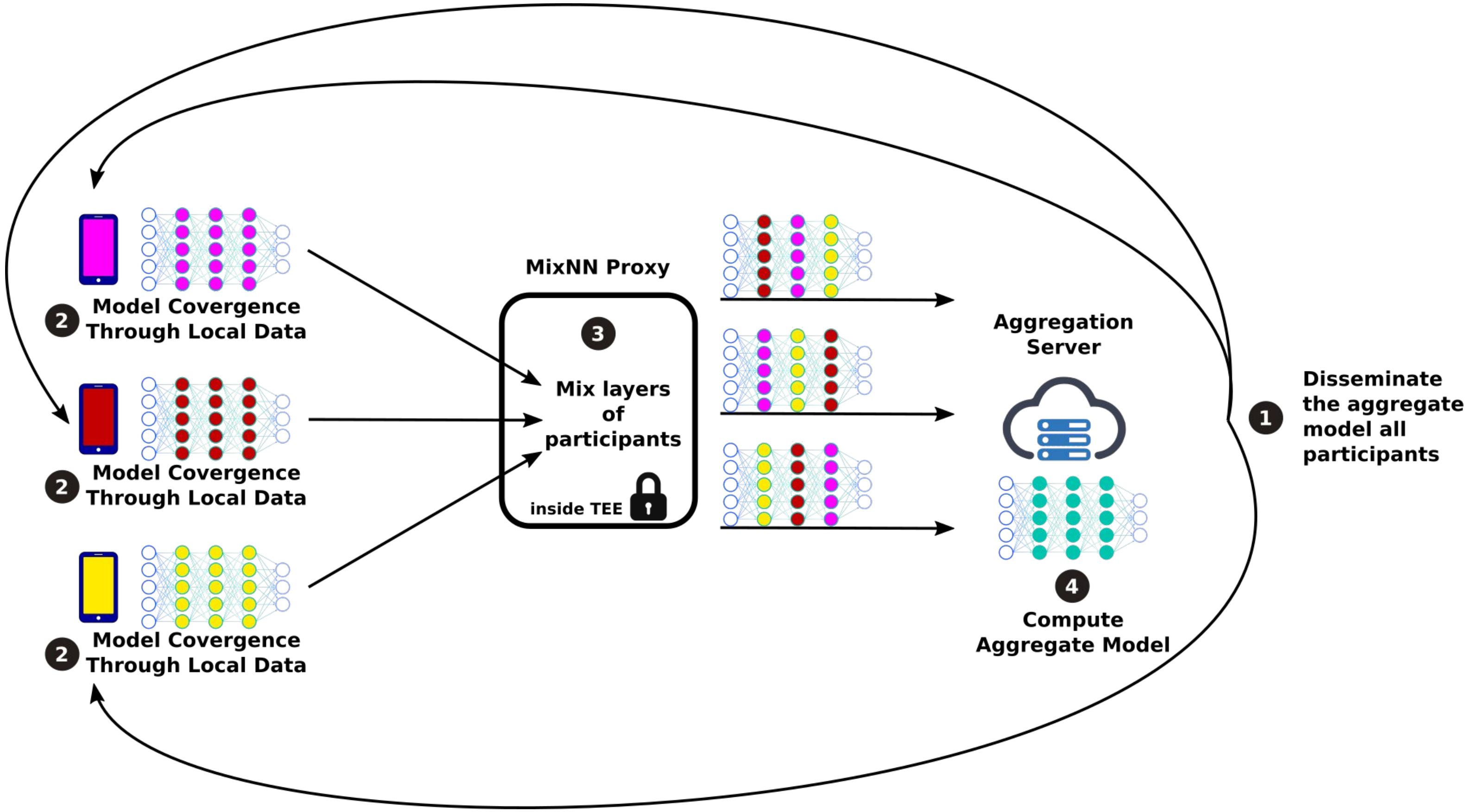
- Prevent unwanted inference of sensitive information (attribute or membership)
- Preserve useful information for activity recognition and personalizing classification locally
- Less sensitive to poisoning
- Ongoing work
  - Generalize these results with other benchmark datasets
  - Impact of NN architectures
  - DP on shared layers
  - Quantify the benefit in terms of bandwidth consumption

Agenda

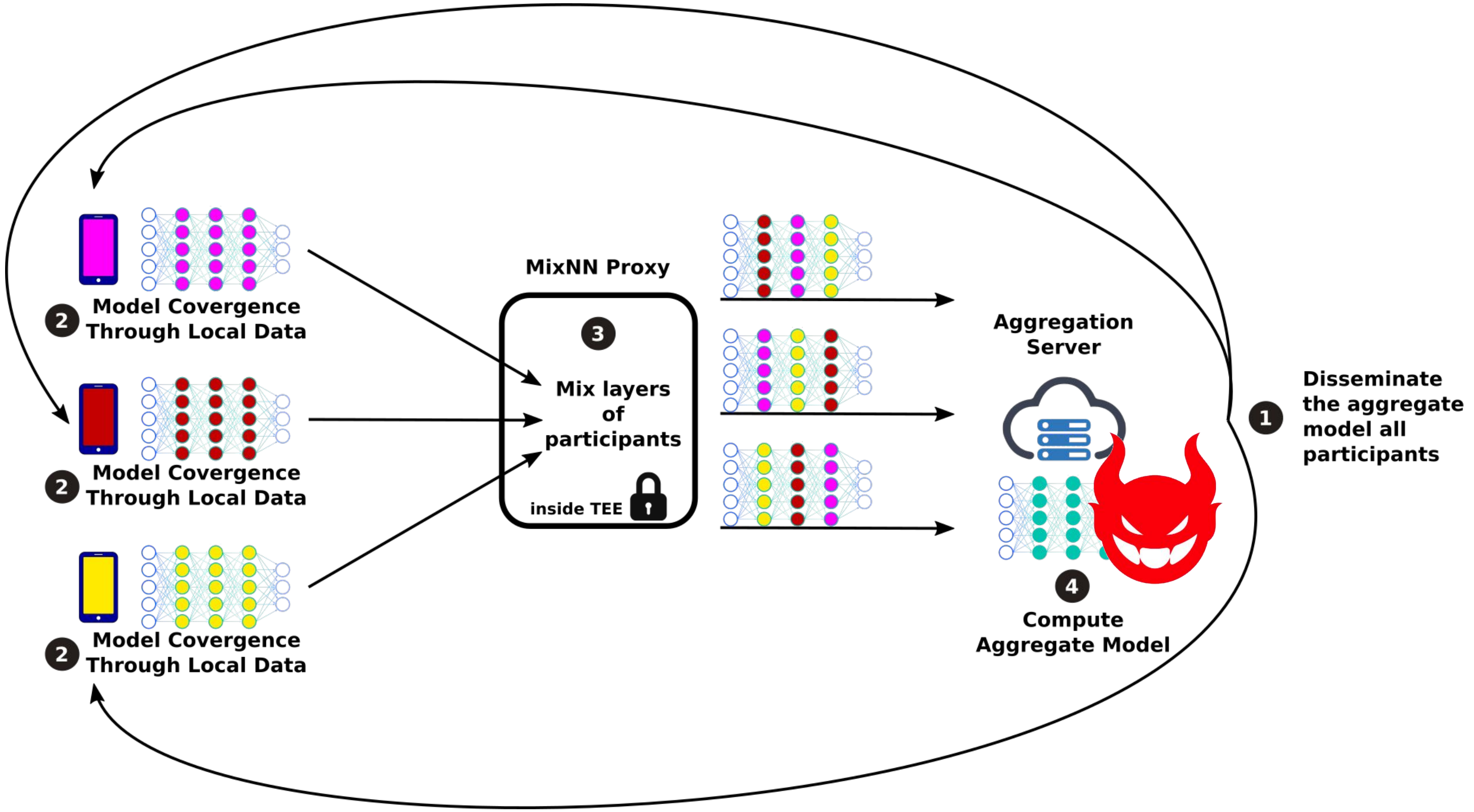
- Centralized Learning
  - Generative Adversarial networks
  - Dynamic sanitizing data through adversarial networks *[ASIACCS' 21]*
- **Federated Learning**
  - Personalization approaches
  - Limitations / Privacy
  - Federated learning using personalized layers [MLSP' 21]

  > MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers *[Middleware'22]*

MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers



MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers



Objective:

- No compromise on utility
- A better privacy against a curious server
- Deployment in a existing system

Experimental setup

Datasets

- **Cifar10**
- **MotienSense**
- **MobiAct**
- **Labeled Faces in the Wild**

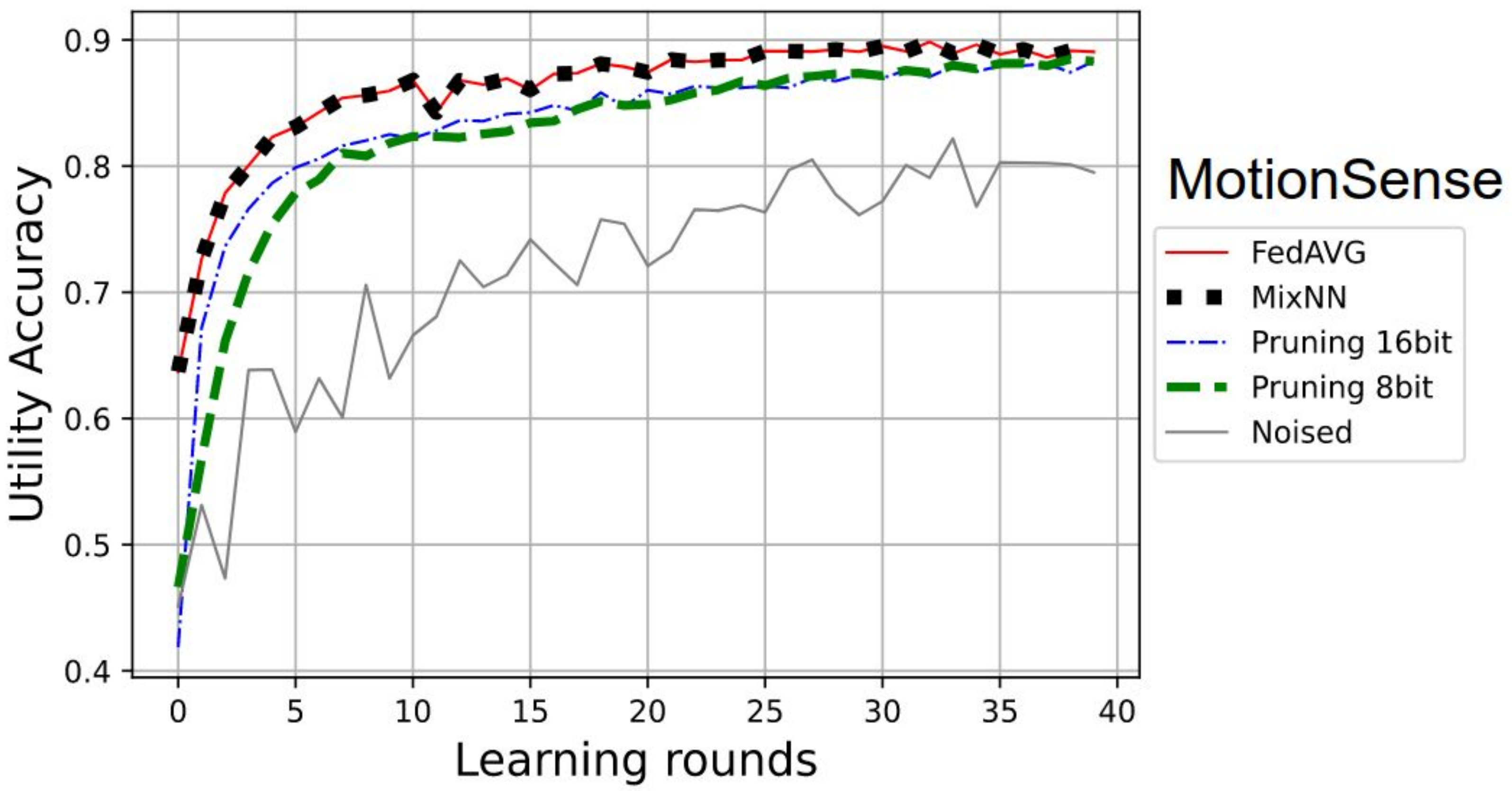
Baslines

- **Vanilla:** the most common FL scheme using SGD training on the device and average aggregation
- **Pruning:** FL scheme using private pruned layers
- **LDP:** FL scheme with an introduction of Gaussian noise to the local model
- **MixNN**

Metrics

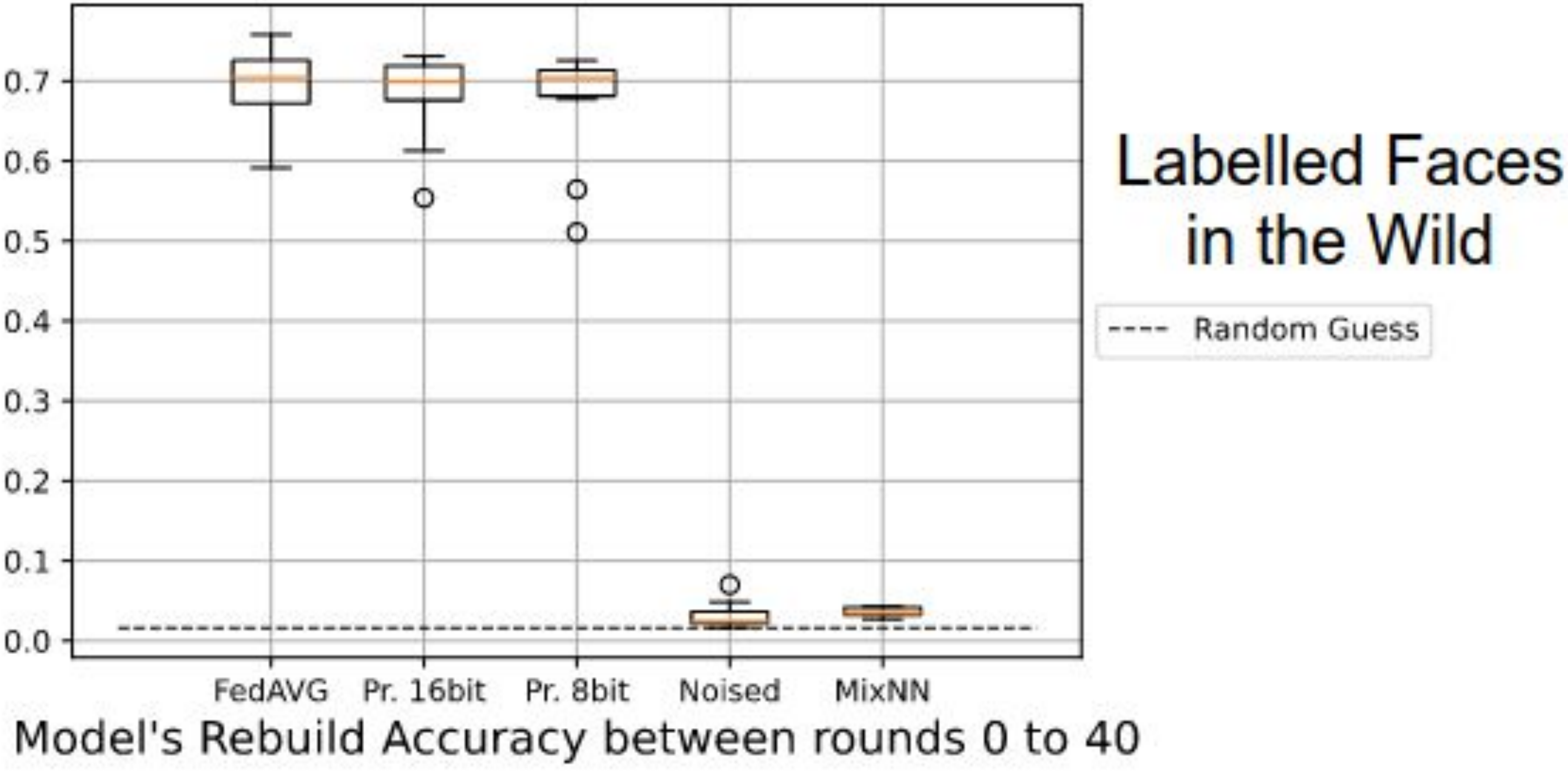
- **Utility:** model activity
- **Privacy:** updates linkability, attribute inference, MixNN robustness
- **System performance:** computational cost

Utility evaluation



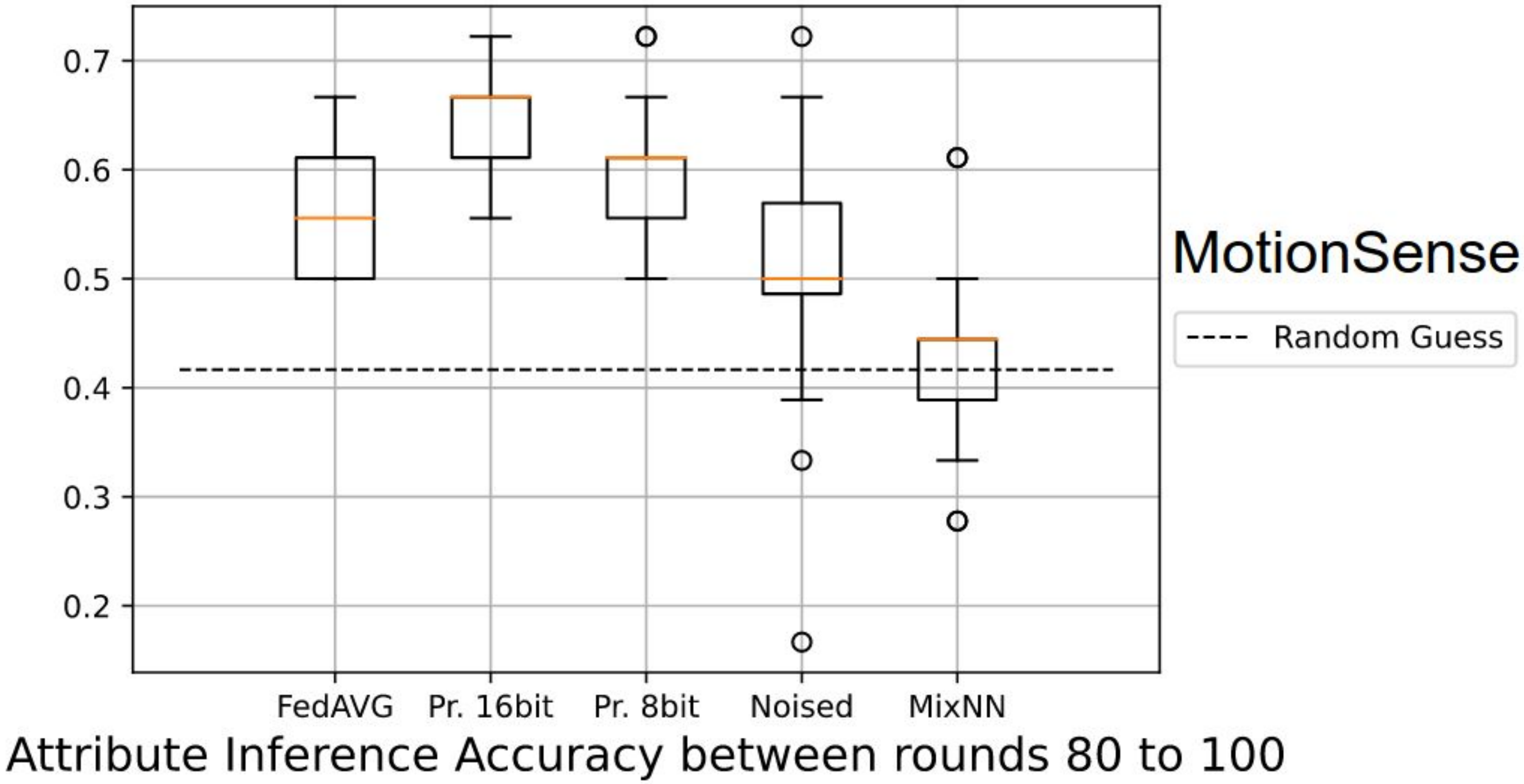
No compromise on utility

Privacy: updates linkability



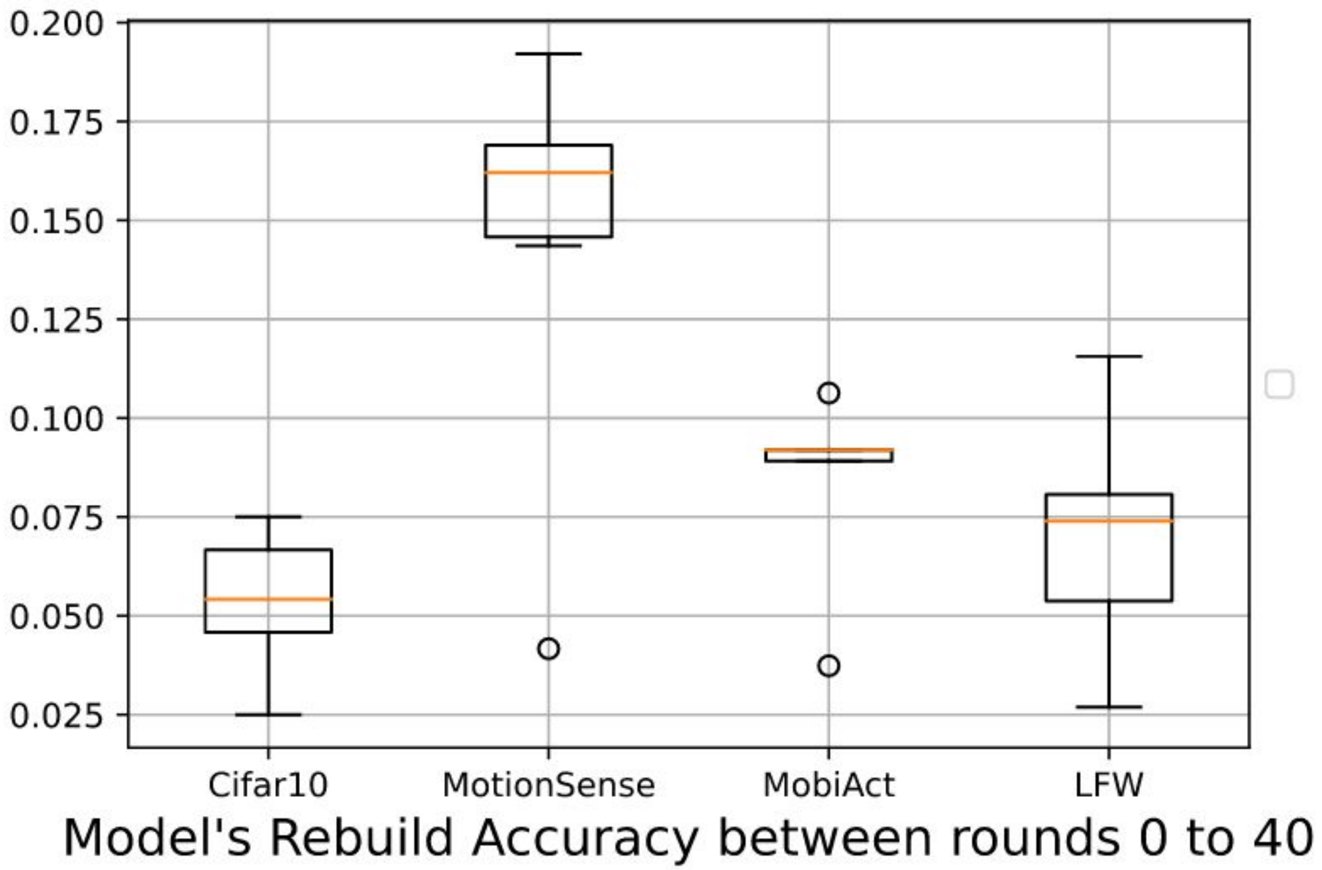
MixNN prevents the server to link clients to their model updates

Privacy: attribute inference



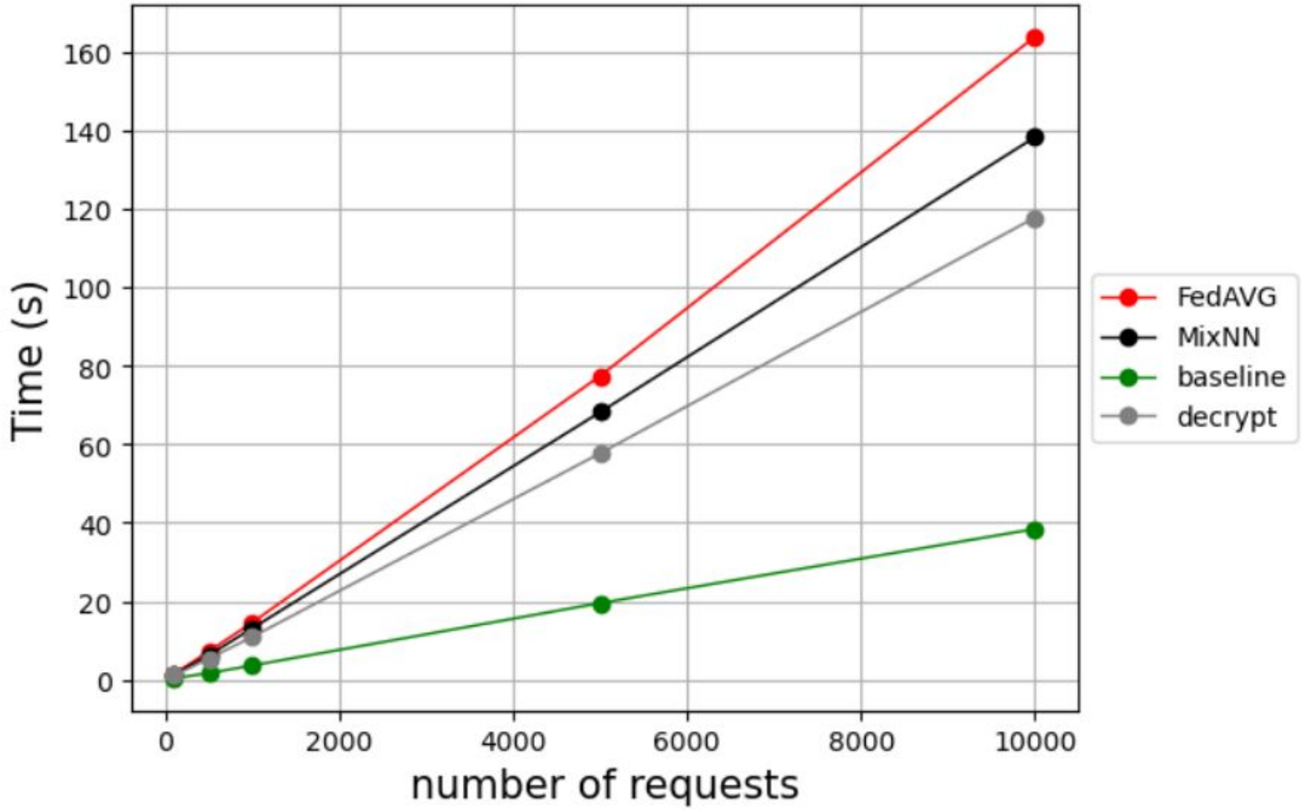
MixNN protects against attribute inference attacks

Privacy: robustness



MixNN protection is hard to break

System performance: latency



MixNN can manage a large number of users

MixNN - Take away

- **MixNN: a proxy-based privacy-preserving framework mixing layers between multiple participants**
- **Prevent inference attacks from a curious aggregation server exploiting model updates**
- **Efficiency breaks the attribute footprint leaked in the model updates without any trade-off with utility**

Agenda

- **Centralized Learning**
  - **Generative Adversarial networks**
  - **Dynamic sanitizing data through adversarial networks** *[ASIACCS' 21]*
- **Federated Learning**
  - **Personalization approaches**
  - **Limitations: Security / Privacy**
  - **Federated learning using personalized layers [MLSP' 21]**
  - **MixNN: Protection of** Federated Learning **Against Inference Attacks by Mixing Neural Network Layers** *[Middleware'22]*
- **Fairness / Explainability**

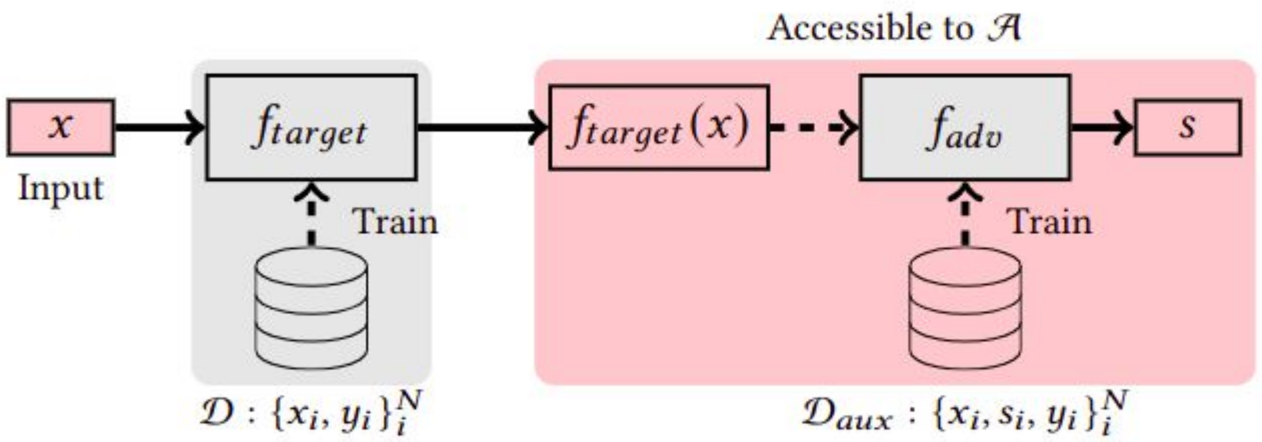
Massive deployment of ML

Rise many questions

- Utility
- Privacy
- Security
- Fairness
- Explainability
- Energy Footprint

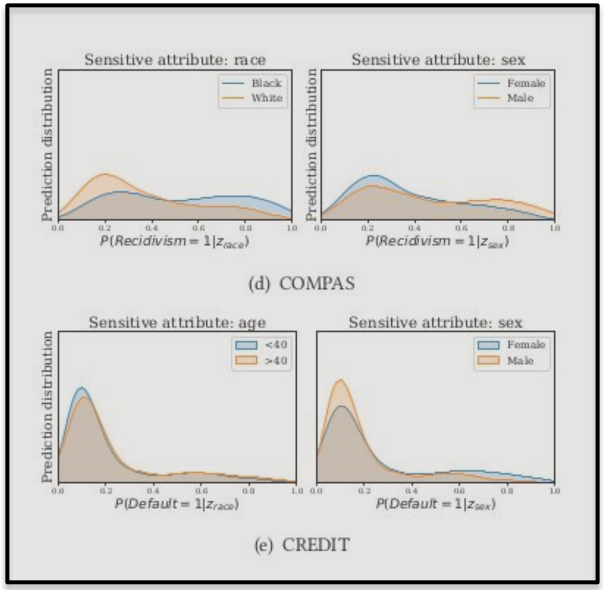
Challenge:
address globally these questions

Data Privacy: Attribute Inference Attacks



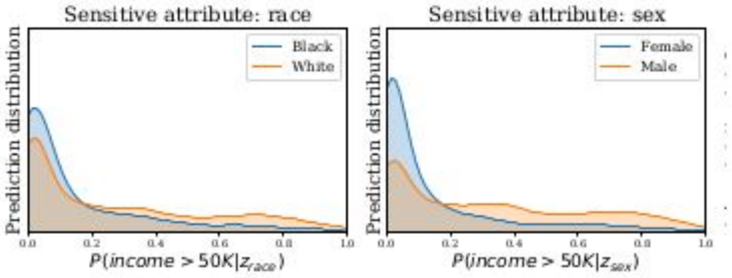
Prior attacks: Use ML attack model (f_{adv}) to infer sensitive attributes

- Exploit distinguishability in predictions for different values of sensitive attribute [6]

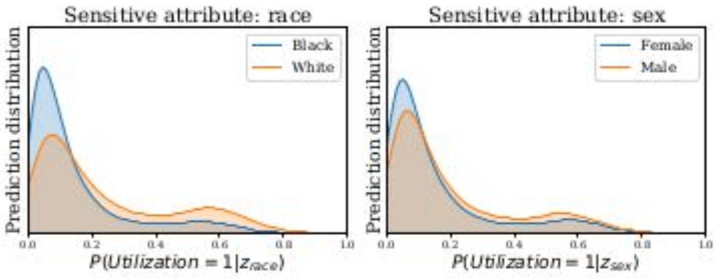


[6] Song and Shmatikov. *Overlearning Reveals Sensitive Attributes*. ICLR'20.

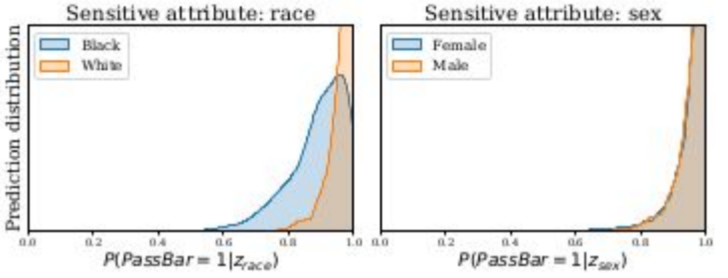
Distinguishable output predictions



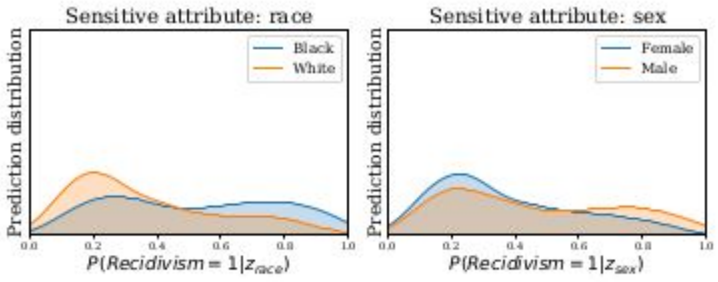
(a) CENSUS



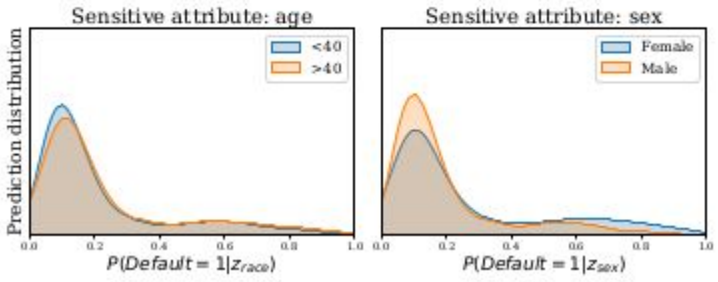
(b) MEPS



(c) LAW



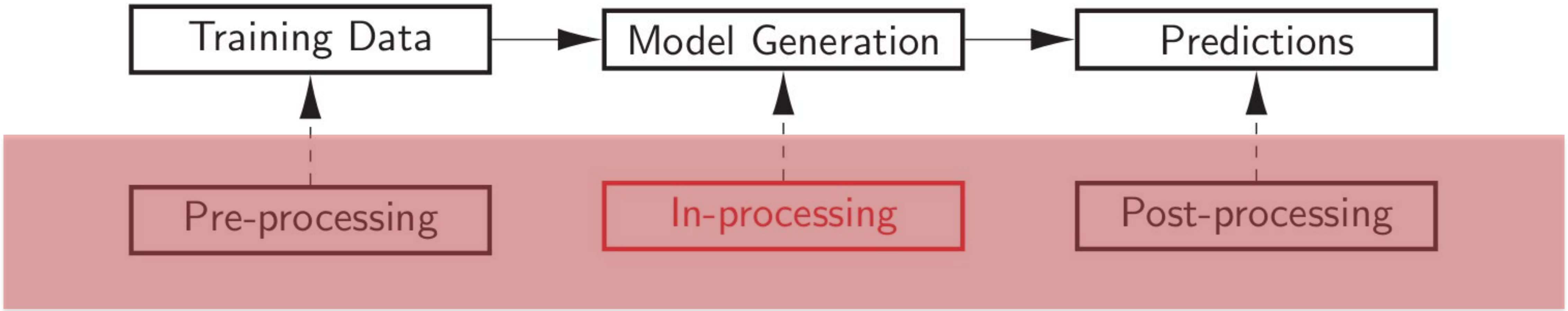
(d) COMPAS



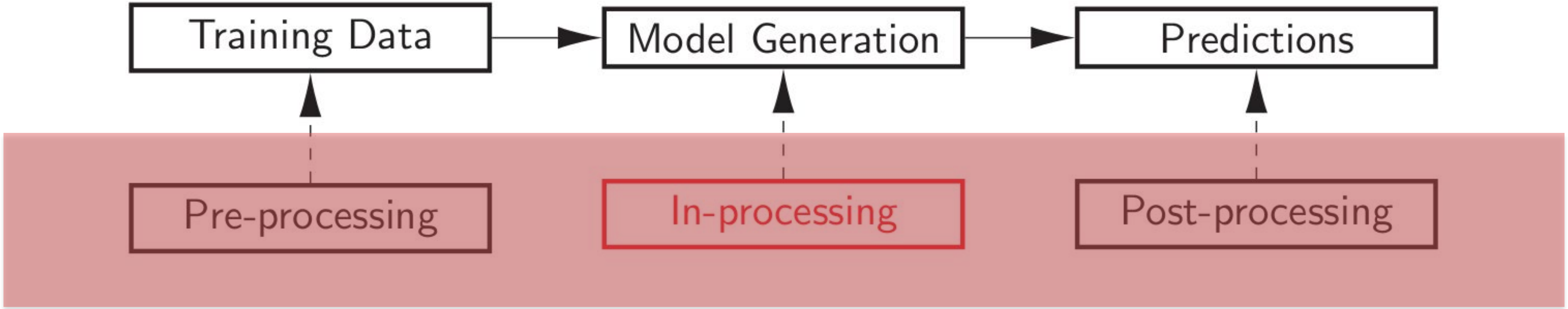
(e) CREDIT

→Idea: remove distinguishability through a fair treatment between two populations

Defence based on Fairness Regularization

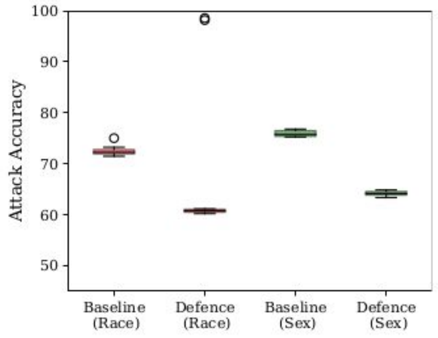


Defence based on Fairness Regularization

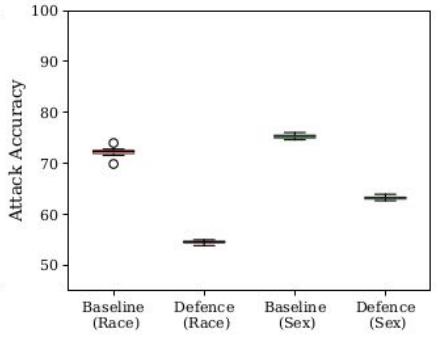


- Individual fairness vs **group fairness**
- In-processing algorithm satisfying a fairness condition:
  - Demographic parity: $P(f_{target}(X) = \hat{y}) = P(f_{target}(X) = \hat{y} | S = s)$
  - Equality of odds: $P(f_{target}(X) = \hat{y} | Y = y) = P(f_{target}(X) = \hat{y} | S = s, Y = y)$

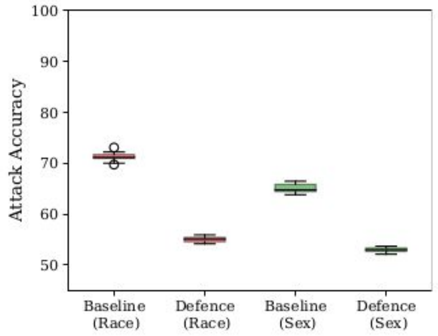
Defence based on Fairness Regularization



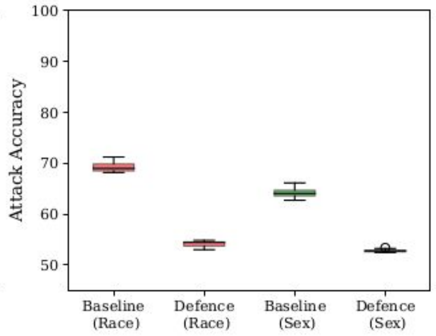
(a) CENSUS (w/ s)



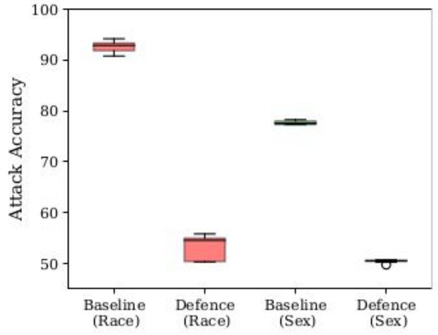
(b) CENSUS (w/o s)



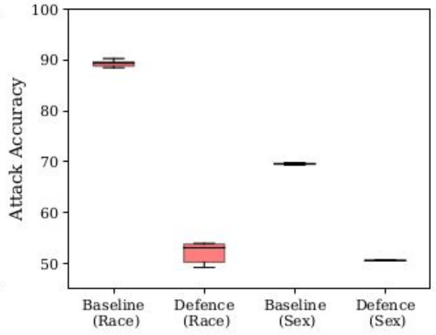
(c) MEPS (w/ s)



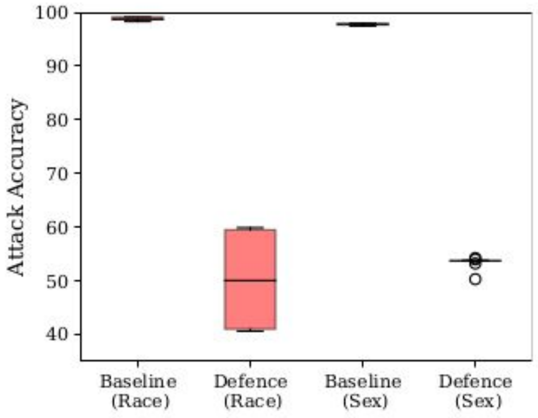
(d) MEPS (w/o s)



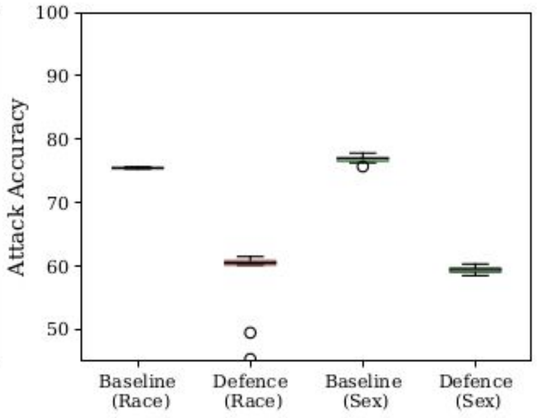
(e) LAW (w/ s)



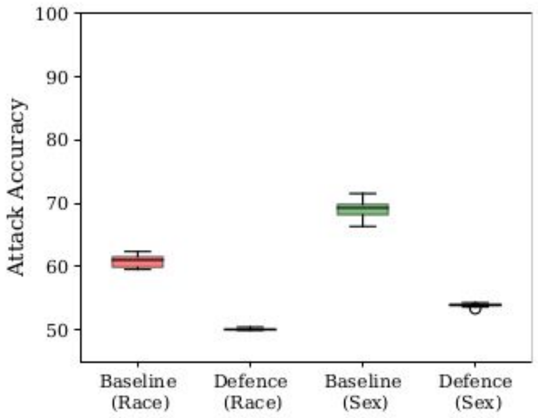
(f) LAW (w/o s)



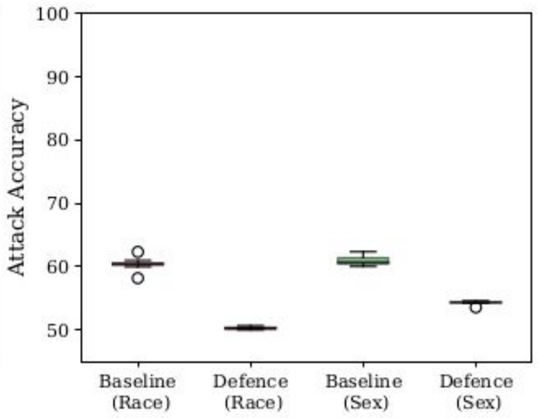
(g) COMPAS (w/ s)



(h) COMPAS (w/o s)

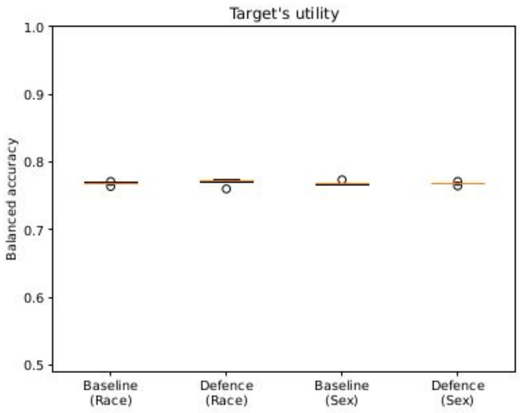


(i) CREDIT (w/ s)

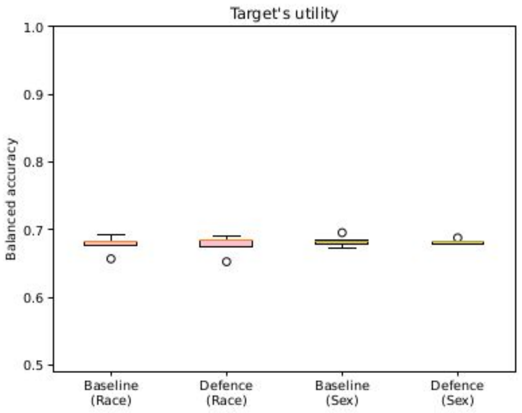


(j) CREDIT (w/o s)

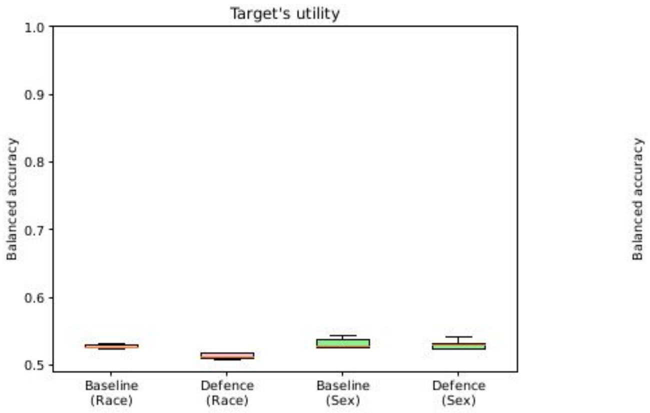
Impact on utility



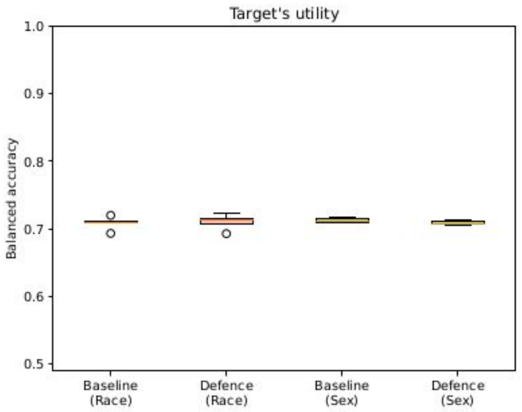
(a) CENSUS



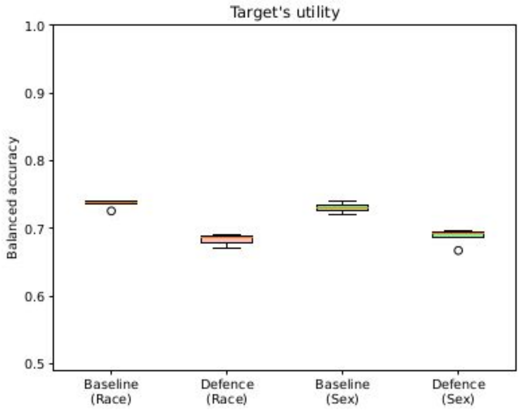
(b) MEPS



(c) LAW



(e) CREDIT



(d) COMPAS

Fairness - Take away

- **Fairness regulation successfully prevents attribute inference attacks while limiting the impact on utility**
- **Theoretical guarantees for demographic parity but theoretical bound for equality of odds fairness condition**

Massive deployment of ML

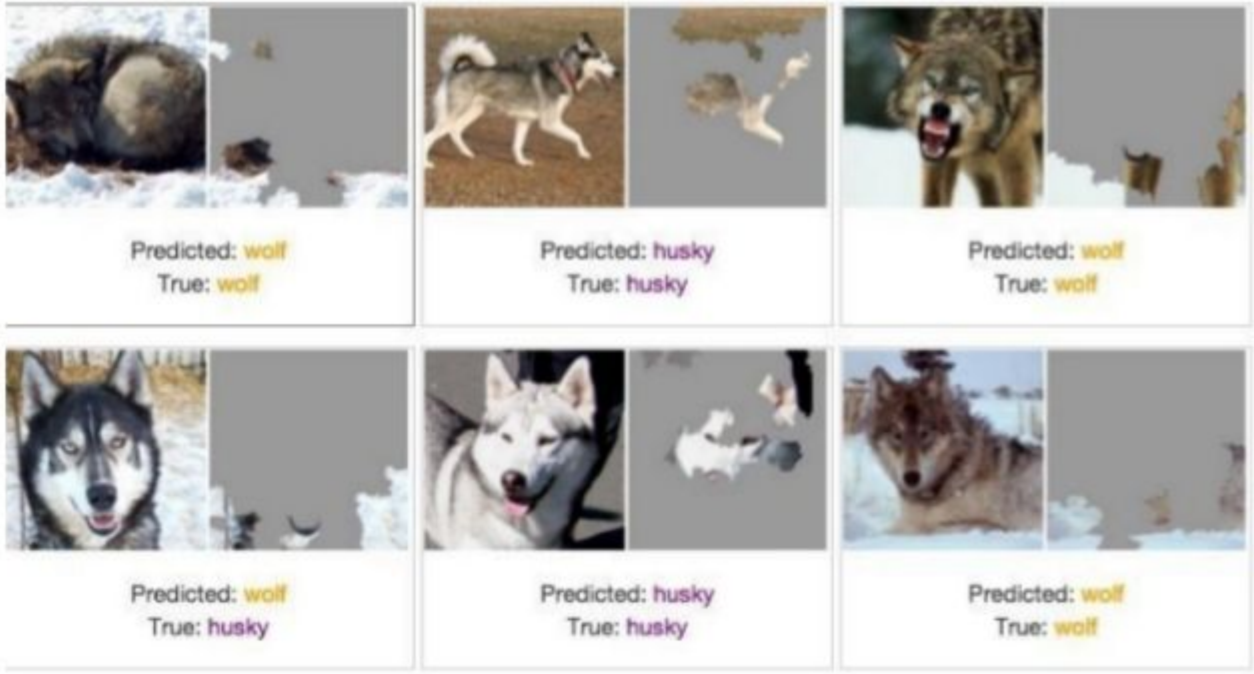
Rise many questions

- Utility
- Privacy
- Security
- Fairness
- Explainability
- Energy Footprint

Challenge:
address globally these questions

Explainability

System that performs behaviour but you don't know how it works



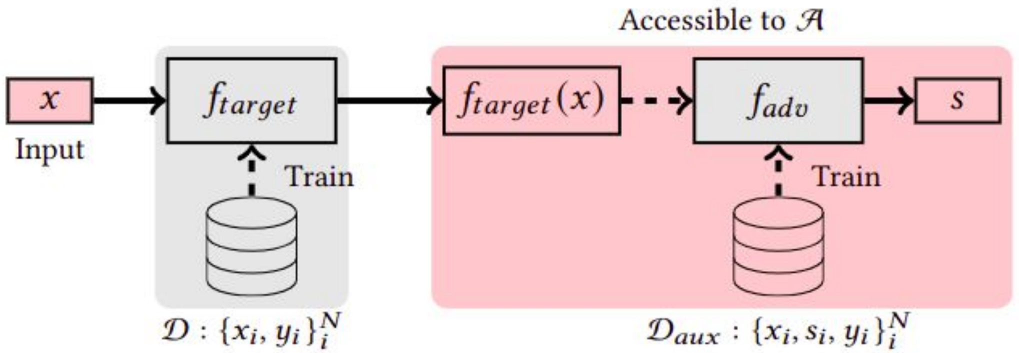
Need algorithmic transparency into complex blackbox models to understand predictions

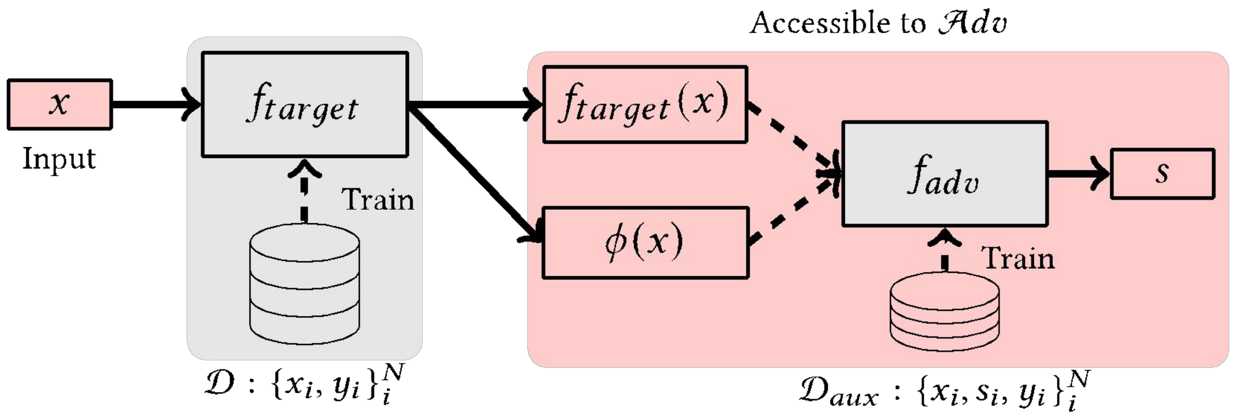
Explainability vs Privacy

What are the data privacy risks of releasing additional information for transparency?

[1] Shokri et al. On the Privacy Risks of Model Explanations. AIES' 21.

Data Privacy: Attribute Inference Attacks





Algorithmic Transparency: Model Explanations

Explanations estimate the influence of different input attributes to model utility

Gradient based Explanations

- Compute gradients using backpropagation for different input attributes
- IntegratedGradients [1] and DeepLift [2]

Perturbation based Explanations

- Add noise/remove attributes to estimate change in output
- GradientSHAP [3] and SmoothGrad [4]

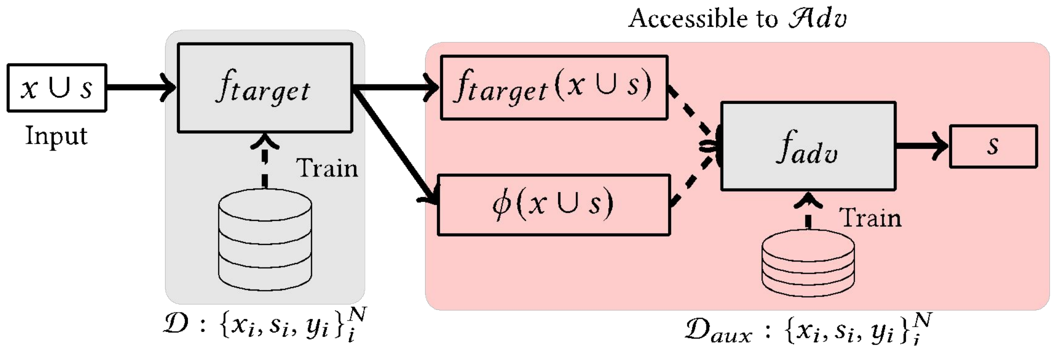
Explanations for sensitive attributes $\boldsymbol{\phi(s)}$ and non-sensitive attributes $\boldsymbol{\phi(x)}$

[1] Sundararajan et al. *Axiomatic Attribution for Deep Networks*. ICML'17.
[2] Shrikumar et al. *Learning Important Features Through Propagating Activation Differences*. ICML'17.
[3] Lundberg and Lee. *A Unified Approach to Interpreting Model Predictions*. NeurIPS'17.
[3] Smilkov et al. *SmoothGrad: Removing Noise by Adding Noise*. ArXiv'17.

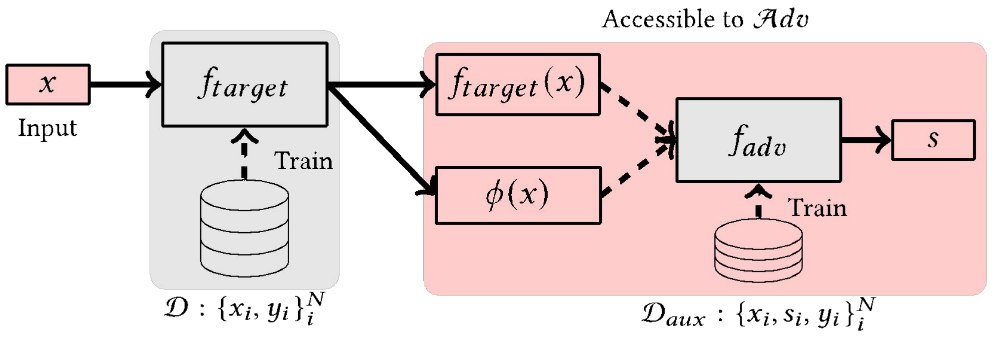
Threat Models

- **Threat Model 1 (TM 1): sensitive attribute included in training data and input**
  - Adversary cannot choose inputs to query
- **Threat Model 2 (TM 2): sensitive attribute censored**
  - Adversary can choose inputs to query

Adversary observes only the predictions $\boldsymbol{f_{target}()}$ and explanations $\boldsymbol{\phi()}$
Auxiliary data available to adversary from same distribution as $\boldsymbol{f_{target}}$'s training data



TM1: w/ sensitive attribute



TM2: w/o sensitive attribute

Explainability - Take away

Yet another trade-off between data privacy and algorithmic transparency!

Model explanations opens a new attack surface for adversary

- **Attacks on explanations are stronger than on predictions**

Future work: impact of mitigation schemes

Conclusion

Developing ethical and trustworthy ML needs to combine multiple topics:
- **Utility**
- **Privacy**
- **Security**
- **Fairness**
- **Explainability**
- **Energy Footprint**

Thank you for your attention

carole.frindel@insa-lyon.fr, antoine.boutet@insa-lyon.fr