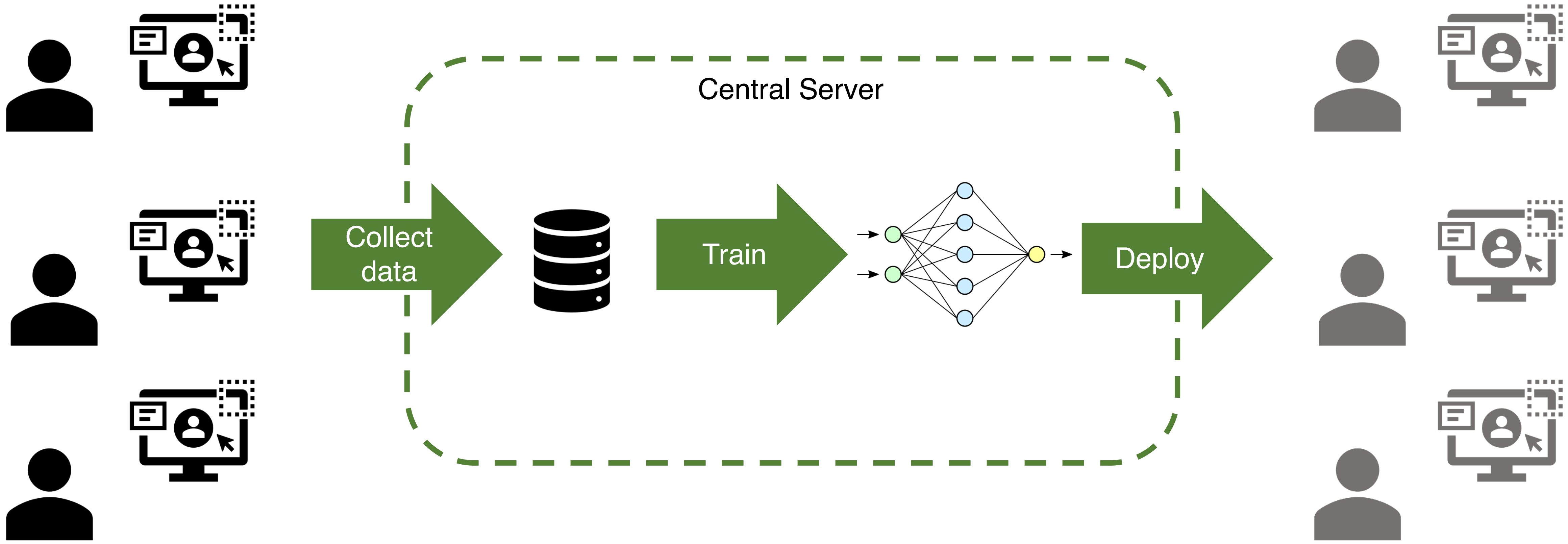
Federated Adversarial Debiasing for Fair and Transferable Representations

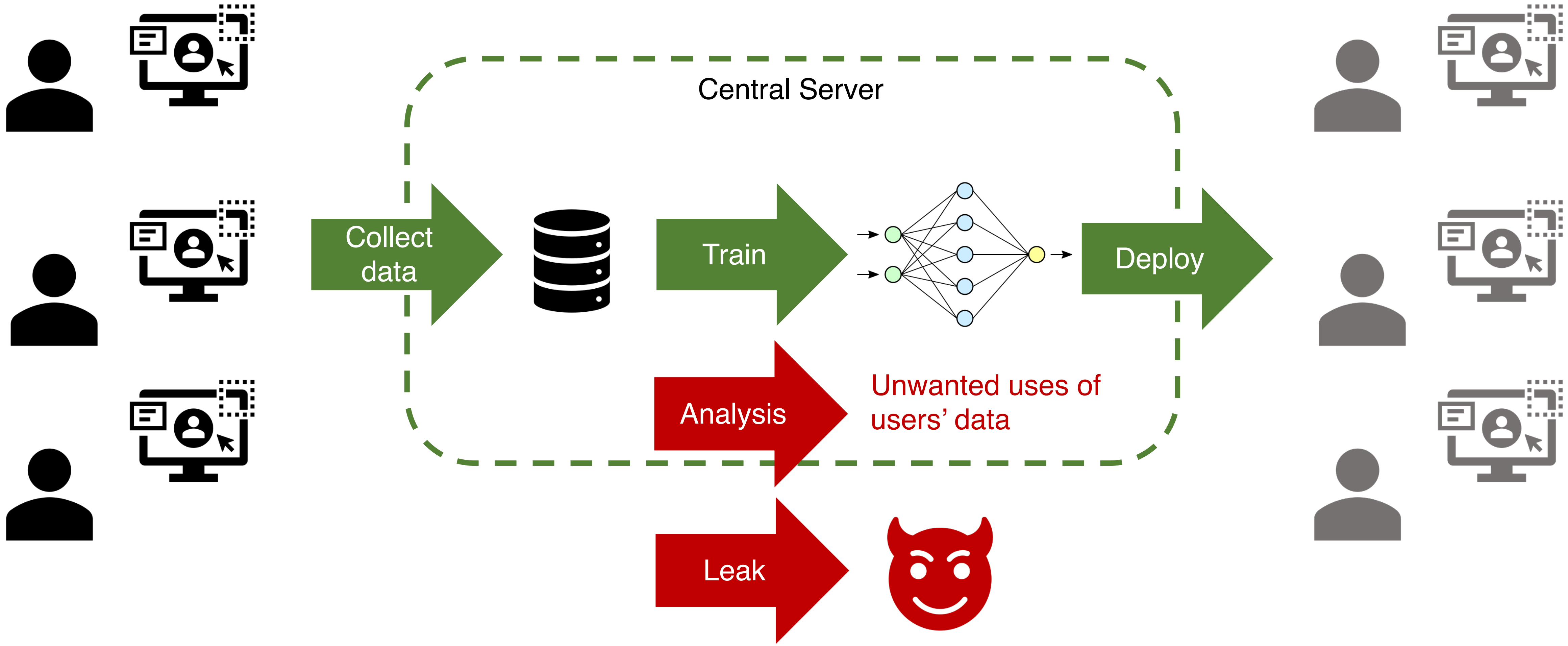
Junyuan Hong[1], Zhuangdi Zhu[1], Shuyang Yu[1], Zhangyang Wang[2], Hiroko Dodge[3], Jiayu Zhou[1]

[1] Michigan State University, [2] University of Texas at Austin, [3] Oregon Health & Science University

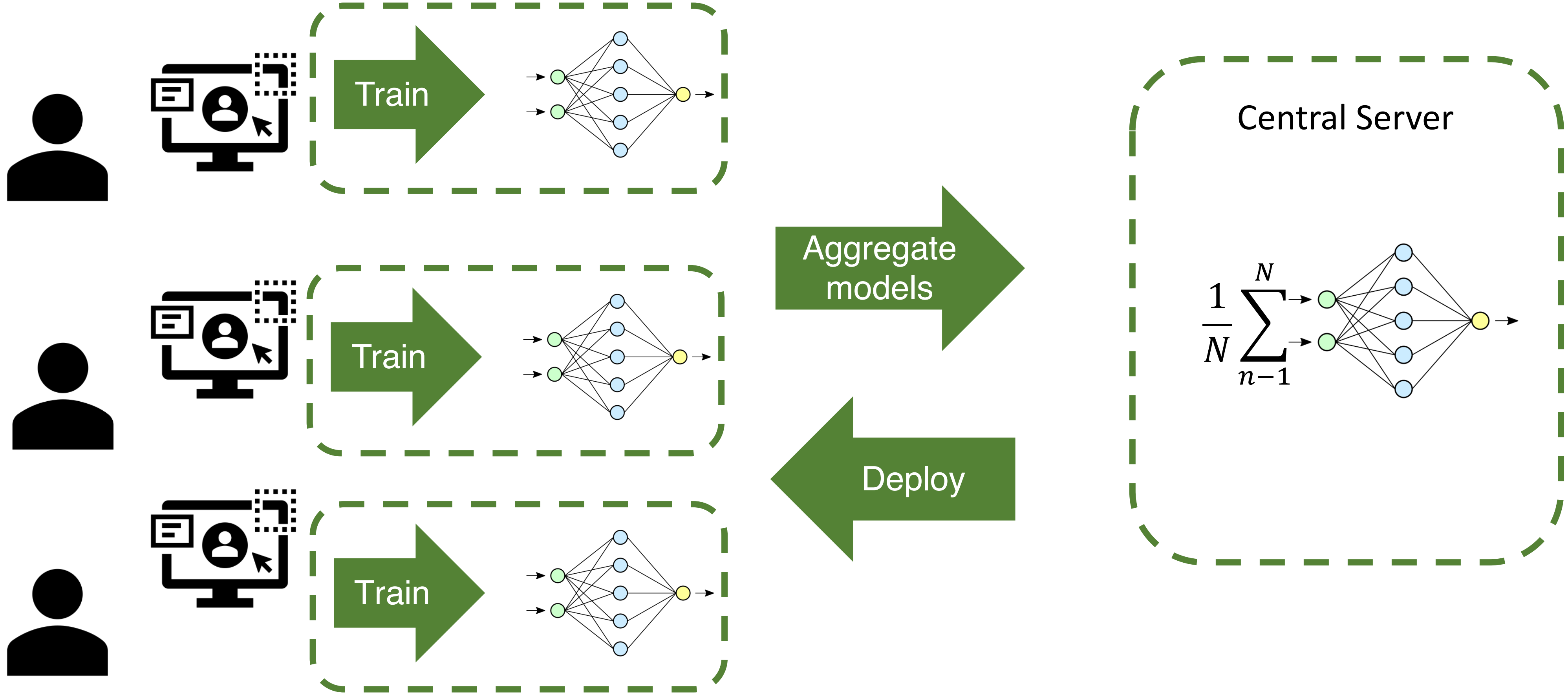
Centralized Learning



Centralized Learning



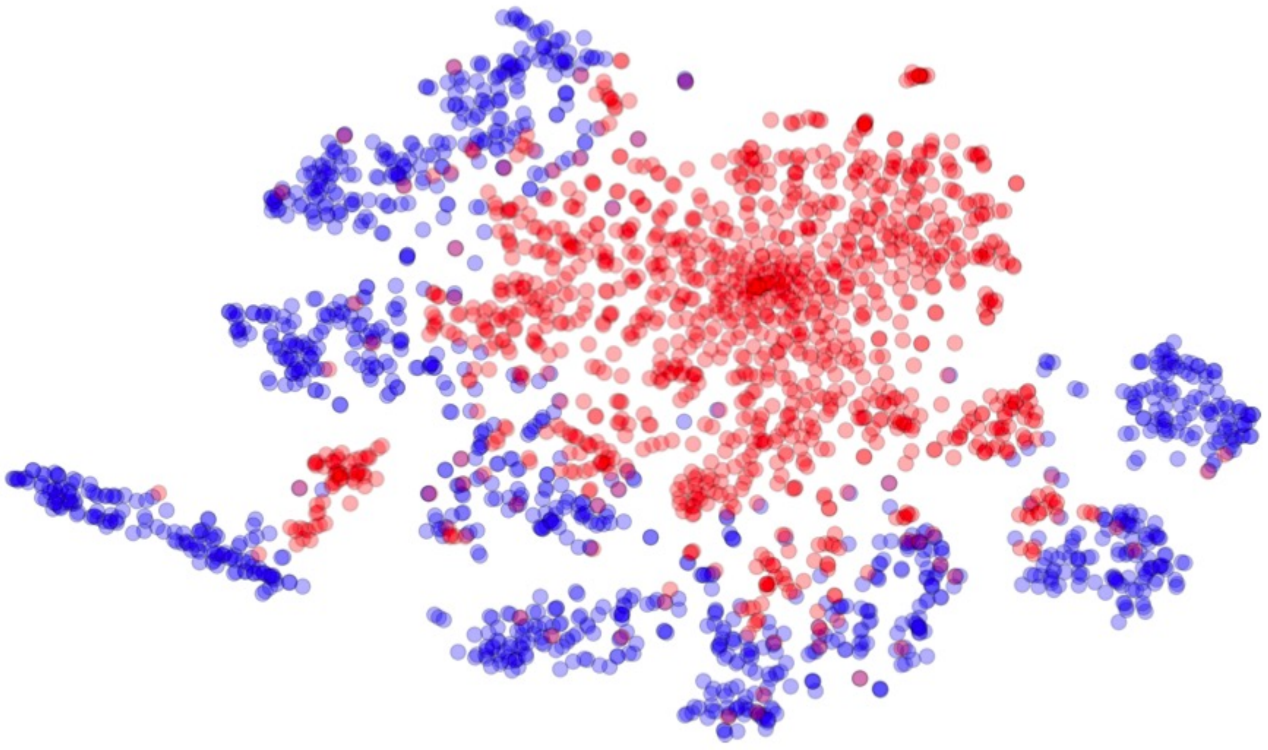
Federated Learning (FL)



Non-*i.i.d.* users in FL

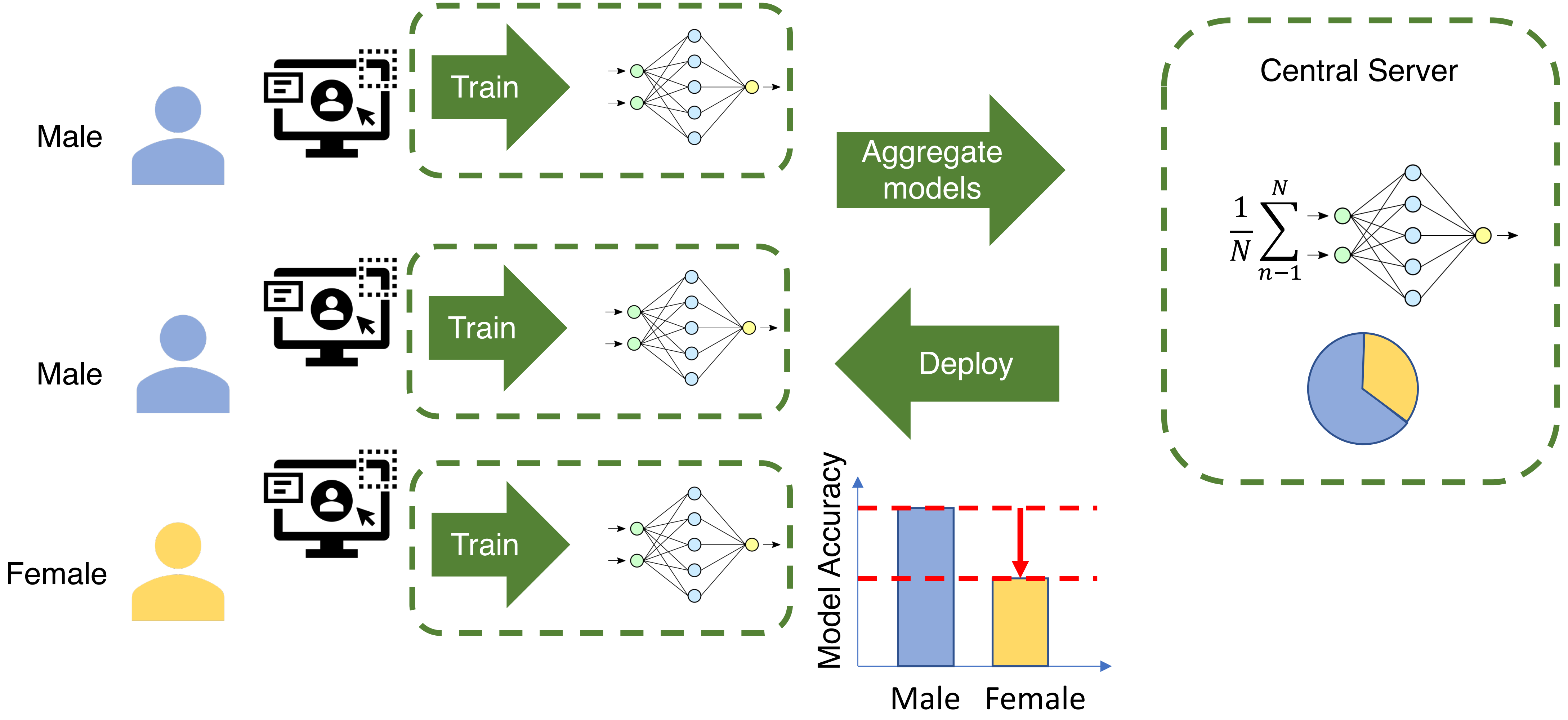
Examples:

- Data from different social groups
  - Genders, races
- Data from different sensors
  - Webcam v.s. prof. cam
  - Grey-scale v.s. color images

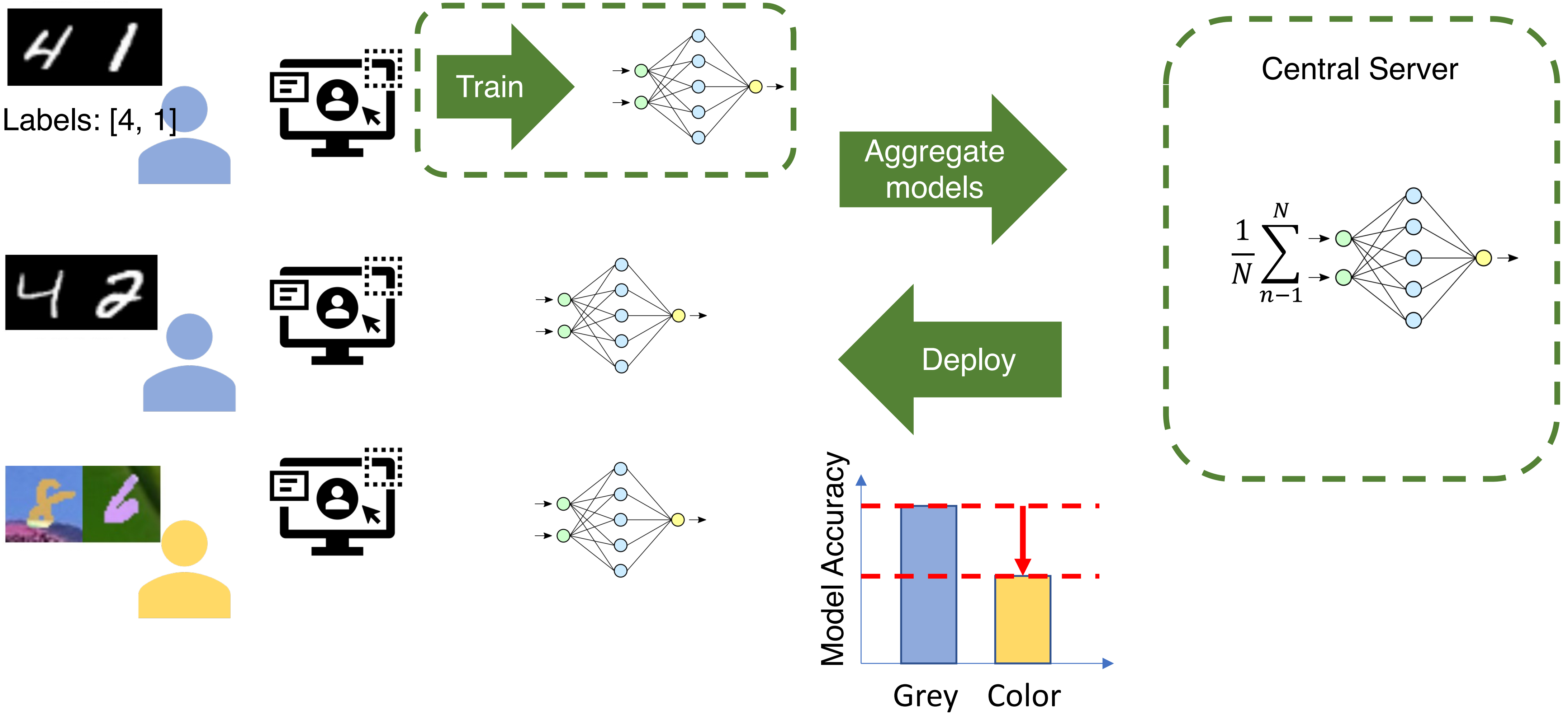


Representation bias: gray-scale v.s. color digit images (MNIST and MNIST-M) extracted by CNN models.
Credit: Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *International Conference on Machine Learning*

Group bias results in unfair models



Domain bias results in non-transferable models



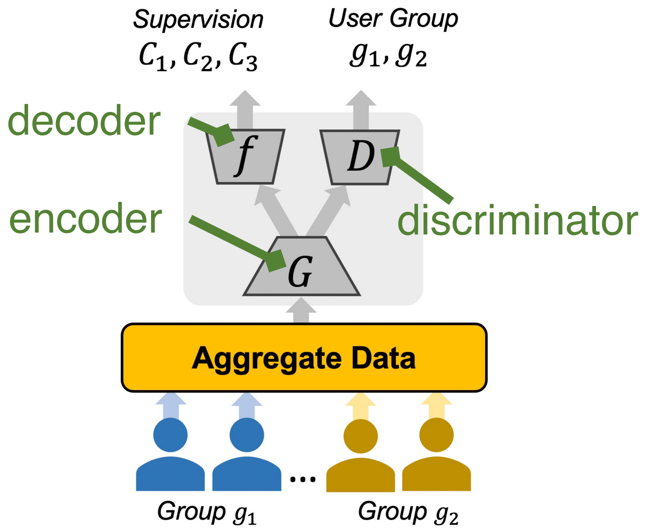
Adversarial Debiasing

- Extract representations $z = G(x)$ from two groups. Thus, $z \sim p_1$ or $z \sim p_2$
- Measure the group discrepancy:

$$\mathbf{D}_{p_1,p_2} = \max_{D} \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))],$$

- Update encoder to reduce bias

$$G = \arg\min_{G} \mathrm{D}_{p_1,p_2}$$



Central methods debias aggregated raw data

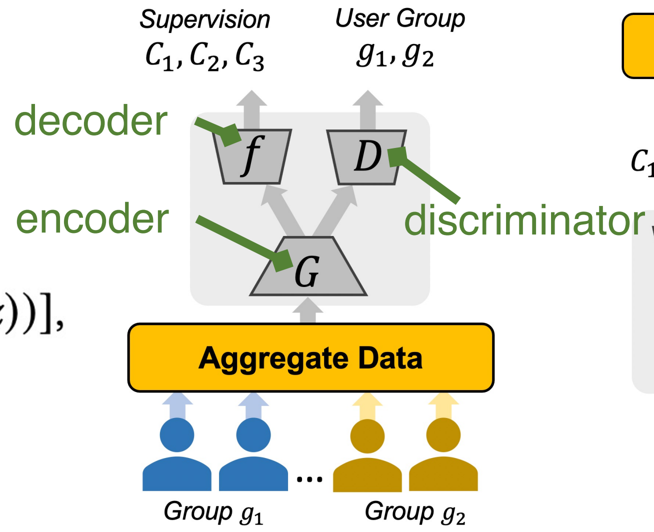
Adversarial Debiasing

- Extract representations $z = G(x)$ from two groups. Thus, $z \sim p_1$ or $z \sim p_2$
- Measure the group discrepancy:

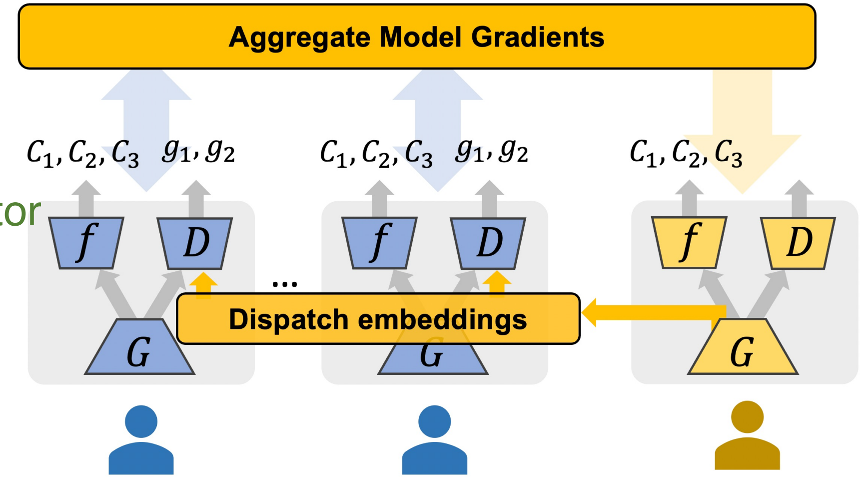
$$\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))],$$

- Update encoder to reduce bias

$$G = \arg\min_G \mathrm{D}_{p_1,p_2}$$



Central methods debias aggregated raw data (Ganin, et al. 2015)



FADA debiases aggregated representation data (Peng, et al., 2019)

Peng, X., Huang, Z., Zhu, Y., & Saenko, K. (2019, September 25). Federated Adversarial Domain Adaptation. *ICLR*
Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *ICML*

Federated Adversarial Debiasing (FADE)

Desired properties:

- **Privacy**: Users do not share training data, intermediate representations or sensitive group attributes during learning.
- **Autonomous**: Users have the freedom to quit the adversarial game during training.
- **Satisfiable**: Adversarial game should be able to reach an equilibrium.

Federated Adversarial Debiasing (FADE)

Desired properties:

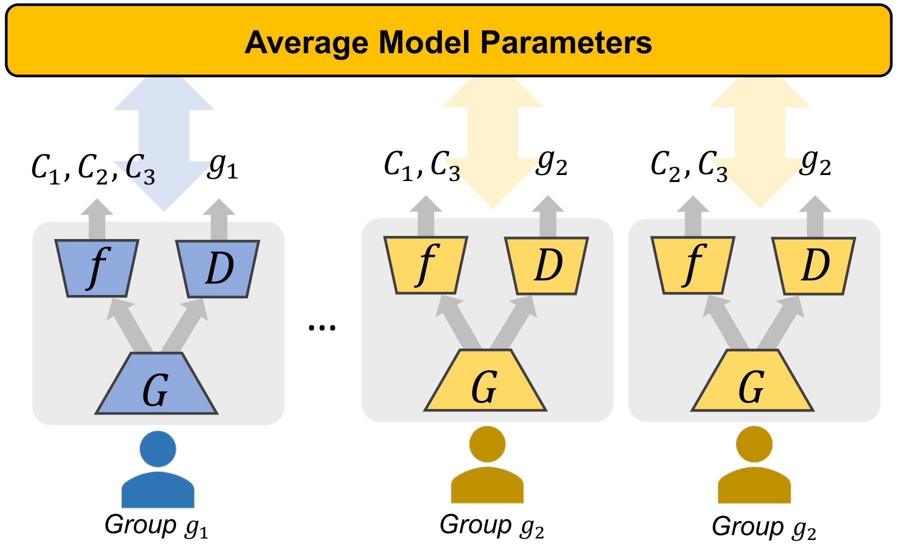
- **Privacy**: Users do not share training data, intermediate representations or sensitive group attributes during learning.
- **Autonomous**: Users have the freedom to quit the adversarial game during training.
- **Satisfiable**: Adversarial game should be able to reach an equilibrium.

Property	Central	FADA	FADE
Privacy	✗ (raw data)	✗ (representations & group attributes)	✓
Autonomous	✗	✗	✓
Satisfiable	✓	✓	✓

Federated Adversarial Debiasing (FADE)

Method

- **Privacy**: Each user train discriminators using local data only and encoders are supervised by shared discriminators.
- Autonomous: …
- Satisfiable: …



Local discrepancy w/o adversarial data

user 1 (group 1) $$\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))]$$

user 2 (group 2) $$\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))]$$

average

Global discrepancy

$$\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))]$$

Federated Adversarial Debiasing (FADE)

Method

- **Privacy**: Each user train discriminators using local data only and generators are supervised by shared discriminators.
- **Autonomous**: Users are allowed not to upload their local models per iteration.
- **Satisfiable**: Distribution matching is sufficient for adversarial optimality.

Central discrepancy

$$\mathbf{D}_{p_1,p_2} = \max_D \boxed{\mathbb{E}_{p_1}[\log D(z)]} + \boxed{\mathbb{E}_{p_2}[\log(1 - D(z))]},$$

α_1 α_2

Uploading probability from group 2

Estimated global discrepancy

$$\tilde{\mathbf{D}}_{p_1,p_2} = \max_D \alpha_1 \mathbb{E}_{p_1}[\log D(z)] + \alpha_2 \mathbb{E}_{p_2}[\log(1 - D(z))]$$

Federated Adversarial Debiasing (FADE)

Method

- **Privacy**: Each user train discriminators using local data only and generators are supervised by shared discriminators.
- **Autonomous**: Users are allowed not to upload their local models per iteration.
- **Satisfiable**: Distribution matching is sufficient for adversarial optimality.

Estimated discrepancy

$$\tilde{\mathbf{D}}_{p_1,p_2} = \max_D \alpha_1 \mathbb{E}_{p_1}[\log D(z)] + \alpha_2 \mathbb{E}_{p_2}[\log(1 - D(z))]$$

Debias representations by estimated discrepancy

$$G = \arg\min_G \tilde{\mathbf{D}}_{p_1,p_2}$$

Theorem 4.1. *The condition $p_1(z) = p_2(z)$ is a sufficient condition for minimizing $\tilde{\mathbf{D}}_{p_1,p_2}$ and the minimal value is $\alpha_1 \log \alpha_1 + \alpha_2 \log \alpha_2 + (\alpha_1 + \alpha_2) \log(\alpha_1 + \alpha_2)$.*

Theoretical Insights

- The estimated discrepancy will be less sensitive to the true distribution difference when two **groups are imbalanced**.
  - Lower uploading probability
  - Imbalanced numbers of users
- Mitigate the imbalance by squared adversarial loss

$$L_{i,g,2}^{\text{adv}}(D,G) = -\frac{1}{2}\left(L_{i,g}^{\text{adv}}(G,D)\right)^2,$$

- Class-wise non-*iid* may cause the loss of discrimination after debiasing.
- A class-conditioned regularization will mitigate the issue.

Theorem 4.2. *Let ϵ be a positive constant. Suppose $|\log p_1(x) - \log p_2(x)| \le \epsilon$ for any x in the support of p_1 and p_2. Then we have $\tilde{\mathrm{D}}_{p_1,p_2} = O(\alpha_1\epsilon/(\alpha_1+\alpha_2))$ when $\alpha_1 \ll \alpha_2$.*

Unsupervised Domain Adaptation (UDA)

Table 1: Averaged classification UDA accuracies (%) on Office and OfficeHome dataset with 3 non-iid target users and 1 source user. Underlines indicate the occurrence of non-converged results. Standard deviations are included in brackets.

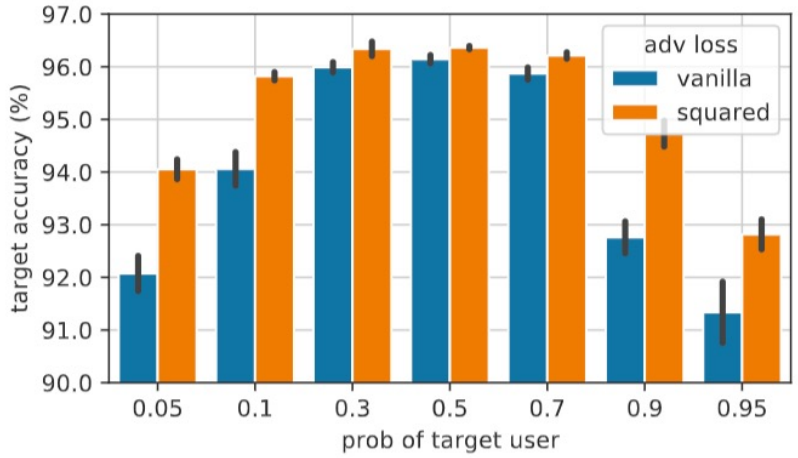
Method	A→D	A→W	D→A	D→W	W→A	W→D	Re→Ar	Re→Cl	Re→Pr	Avg.
Federated methods										
Source only	79.5	73.4	59.6	91.6	58.2	95.8	67.0	46.5	78.2	72.2
non-iid target users w/ 20 (Office) or 45 (OfficeHome) classes per user										
FADE-DANN	85.4 (1.9)	81.8 (1.8)	43.1 (33)	97.7 (0.5)	64.8 (0.5)	99.7 (0.2)	46.4 (37)	34.9 (27)	78.8 (0.1)	70.3
FADE-CDAN	**92.3 (1.2)**	**91.6 (0.5)**	**65.9 (9.3)**	**98.9 (0.2)**	**70.2 (0.8)**	**99.9 (0.1)**	70.3 (1.6)	54.9 (4.6)	**82.2 (0.1)**	80.7
FedAvg-SHOT	83.6 (0.5)	83.1 (0.5)	64.7 (1.4)	91.7 (0.2)	64.7 (2.2)	97.4 (0.5)	**70.7 (0.5)**	**55.4 (0.5)**	80.1 (0.3)	76.8
iid target users										
FADE-DANN	84.2 (1.5)	81.3 (0.4)	66.3 (0.3)	97.5 (1.2)	59.4 (10.6)	99.9 (0.2)	67.3 (0.9)	51.3 (0.4)	79.0 (0.6)	76.2
FADE-CDAN	**93.6 (0.8)**	**92.2 (1.3)**	**71.2 (1.0)**	**98.7 (0.4)**	**71.3 (0.7)**	**100 (0.0)**	70.6 (1.3)	55.1 (1.0)	**82.3 (0.2)**	81.7
FedAvg-SHOT	**96.3 (0.5)**	**94.3 (1.1)**	70.9 (2.0)	98.4 (0.4)	**72.7 (0.9)**	99.8 (0.0)	**74.8 (0.3)**	**60.0 (0.1)**	**84.9 (0.2)**	83.6
Central methods										
ResNet [15]	68.9	68.4	62.5	96.7	60.7	99.3	53.9	41.2	59.9	67.9
Source only [23]	80.8	76.9	60.3	95.3	63.6	98.7	65.3	45.4	78.0	73.8
DANN [11]	79.7	82.0	68.2	96.9	67.4	99.1	63.2	51.8	76.8	76.1
CDAN [28]	92.9	**94.1**	71.0	**98.6**	69.3	**100**	70.9	56.7	81.6	81.7
SHOT [23]	**94.0**	90.1	**74.7**	98.4	**74.3**	99.9	**73.3**	**58.8**	**84.3**	**83.1**

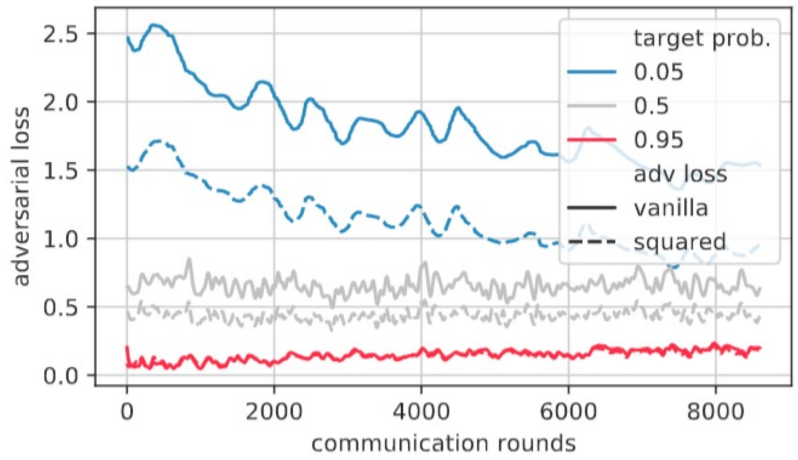
Unsupervised Domain Adaptation (UDA) with imbalanced source/target users

- Imbalance results in large adv. loss.
- Squared loss design: further increase the loss value if the loss is large.

$$L_{i,g}^{\mathrm{adv}}(G,D) = \mathbb{E}_{x\sim p_i(x)}\big[\mathbb{I}(g=0)\log D(G(x)) + \mathbb{I}(g=1)\log(1-D(G(x)))\big],$$

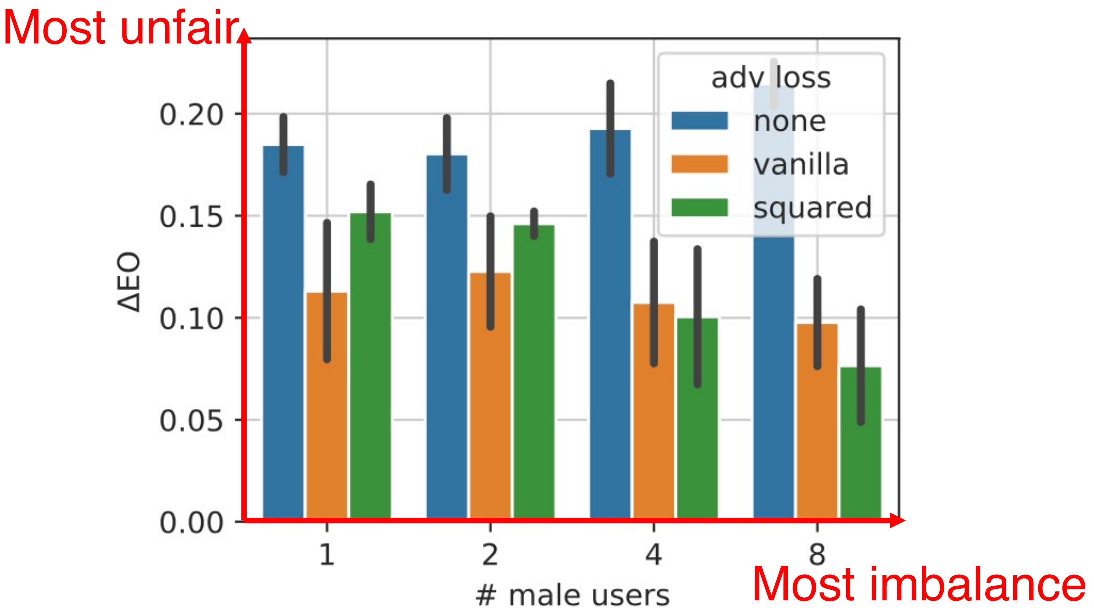
$$L_{i,g,2}^{\mathrm{adv}}(D,G) = -\frac{1}{2}\left(L_{i,g}^{\mathrm{adv}}(G,D)\right)^2,$$

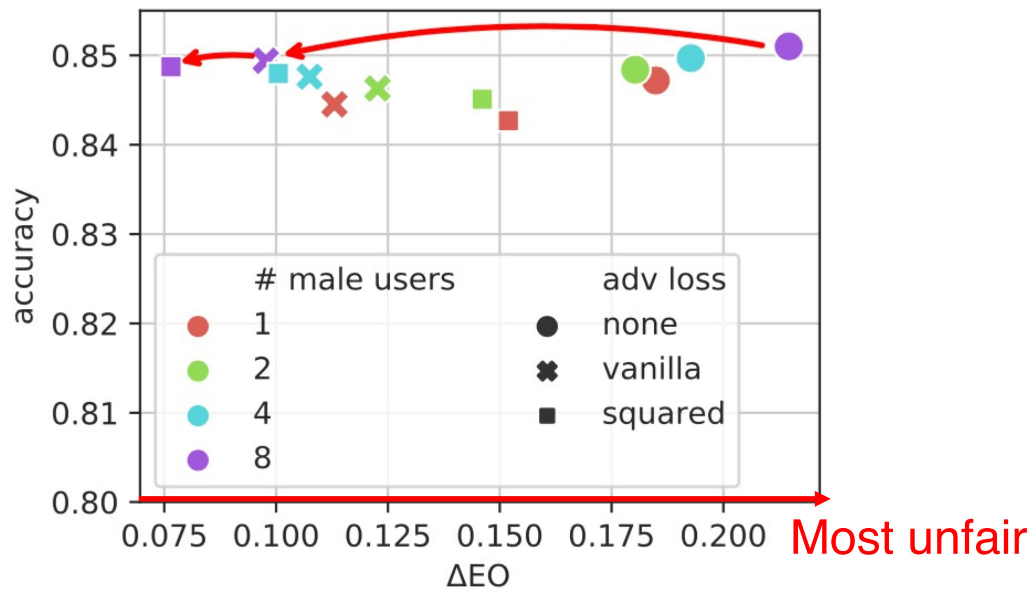




USPS -> MNIST

Fair learning
with imbalanced female/male users





Adult dataset with fairness on male/female groups

Thank You!

Codes: https://github.com/illidanlab/FADE

Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-1749940, EPCN-2053272, Office of Naval Research N00014-20-1-2382, and National Institute on Aging (NIA) R01AG051628, R01AG056102, P30AG066518, P30AG024978, RF1AG072449.