Towards Robust and Bias-Free Federated Learning

Ousmane Touat , Sara Bouchenak

INSA Lyon – LIRIS, France

ousmane.touat@insa-lyon.fr, sara.bouchenak@insa-lyon.fr

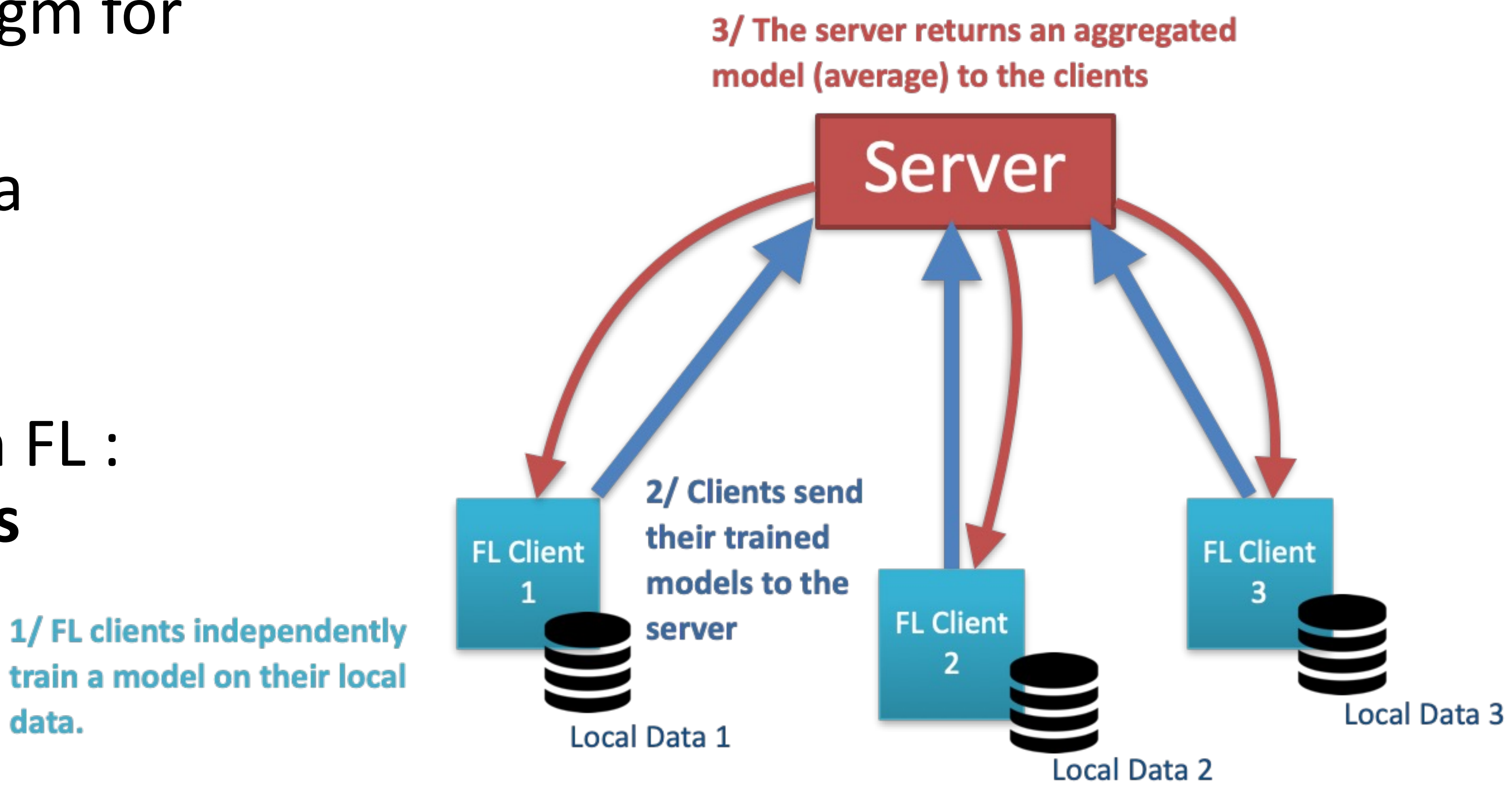
EuroMLSys

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

LIRIS

Federated Learning

- A distributed paradigm for model training
- Concerned with data privacy and data heterogeneity
- Two critical issues in FL : Bias and **Robustness**



Bias in Federated Learning

TOM SIMONITE BUSINESS OCT 24, 2019 2:00 PM

A Health Care Algorithm Offered Less Care to Black Patients

A study shows the risks of making decisions using data that reflects inequities in American society.

Wired

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Reuters

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

MIT News

Definition of Bias in Federated Learning

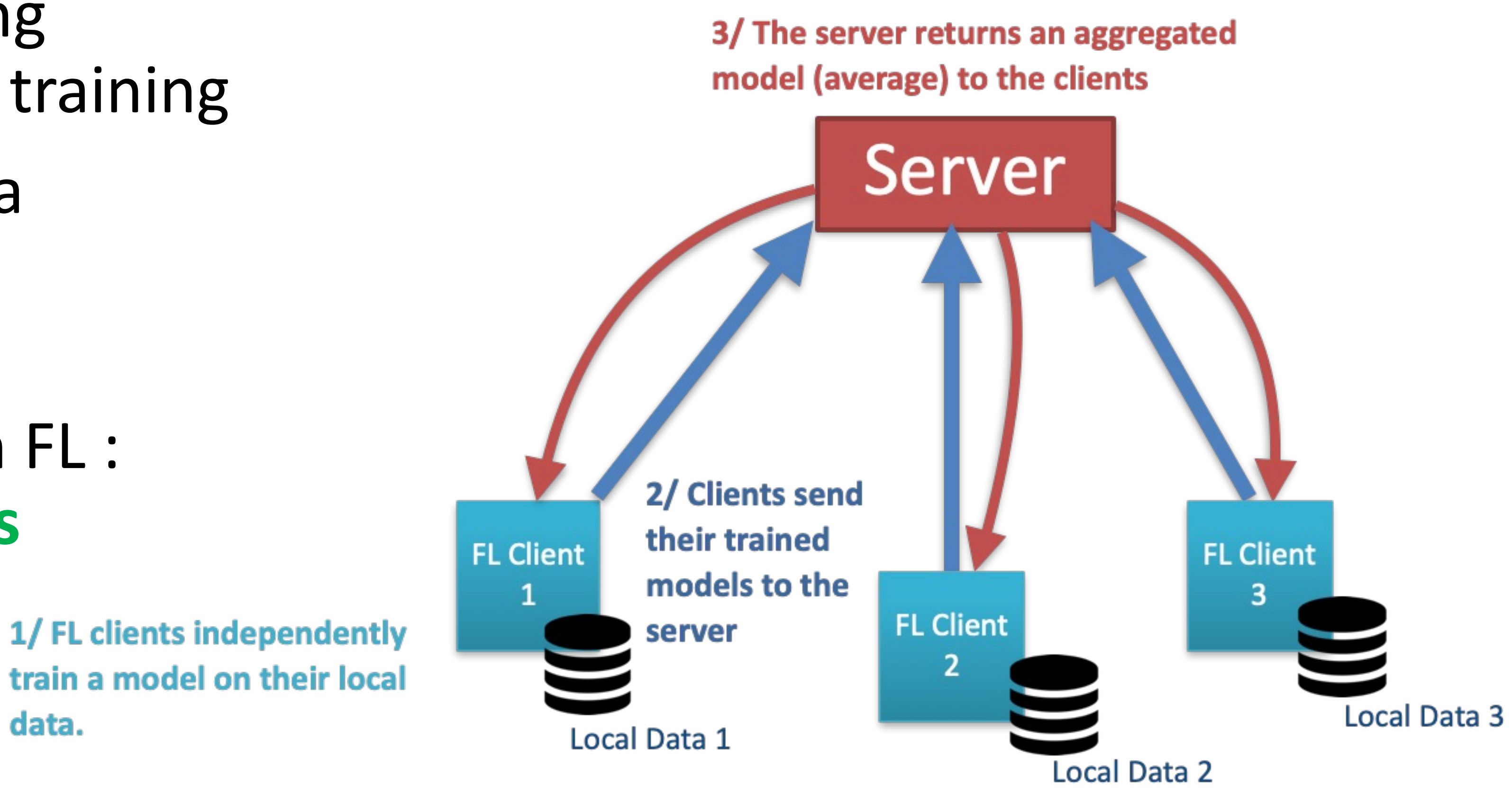
- Data misrepresentation produces **biased model** towards specific groups, identified with **sensitive attributes** (e.g., man & woman, old & young)
- Examples :
  - **(Healthcare)** Discrepancy in diagnosis model quality between demographic groups
  - **(Recruitment)** Differences in employment rate within demographic groups

Related Work on Bias Mitigation in FL

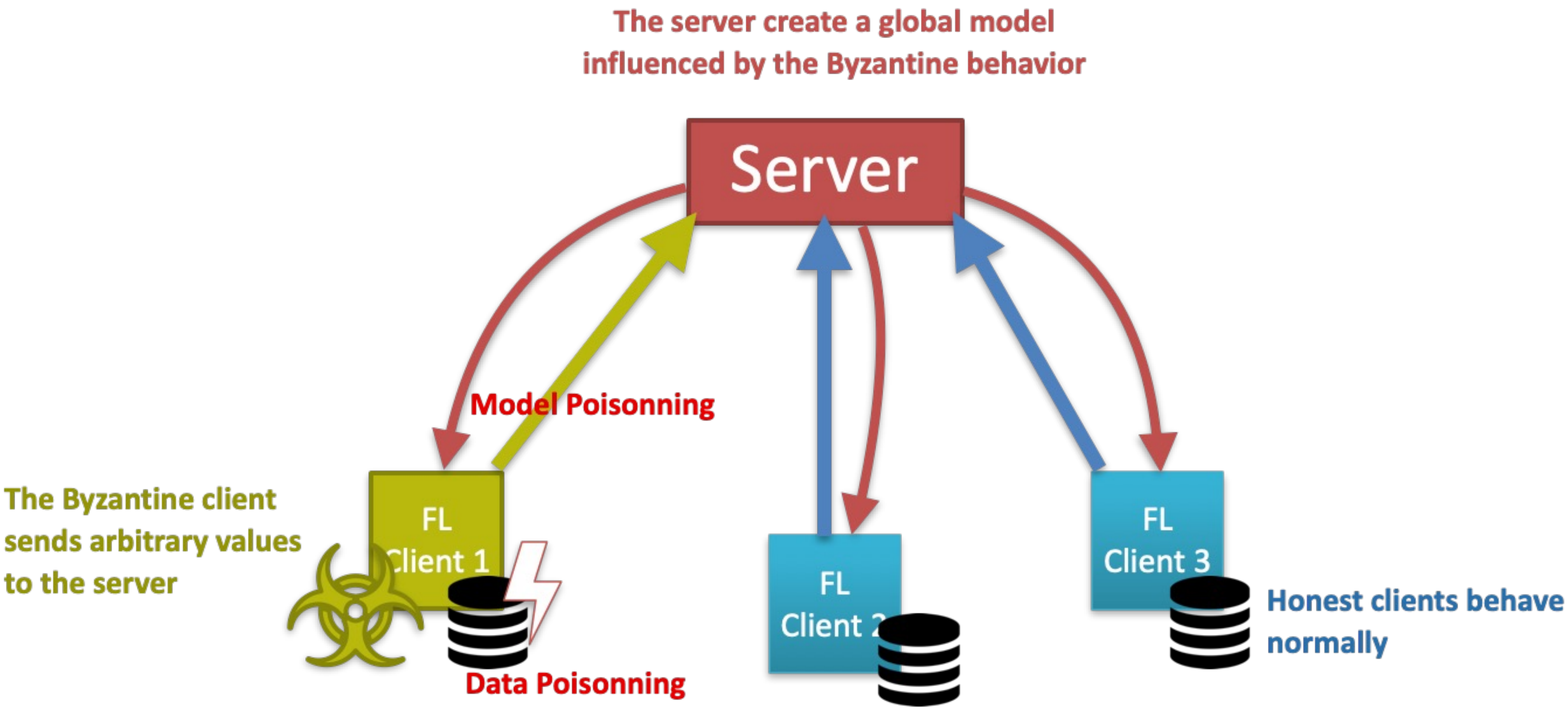
- Client-side techniques :
  - Using techniques from centralized learning such as data reweighting [5]
  - Do not guarantee global bias mitigation under non-IID settings
- Server-side techniques:
  - Techniques requiring additional computation from the server : AgnosticFair [6], FairFL [7], FairFed [8], and FCFL [9] .
  - Often requires FL clients to send additional information to the server (local statistical data distribution)

Federated Learning

- A distributed learning paradigm for model training
- Concerned with data privacy and data heterogeneity
- Two critical issues in FL : **Bias** and **Robustness**



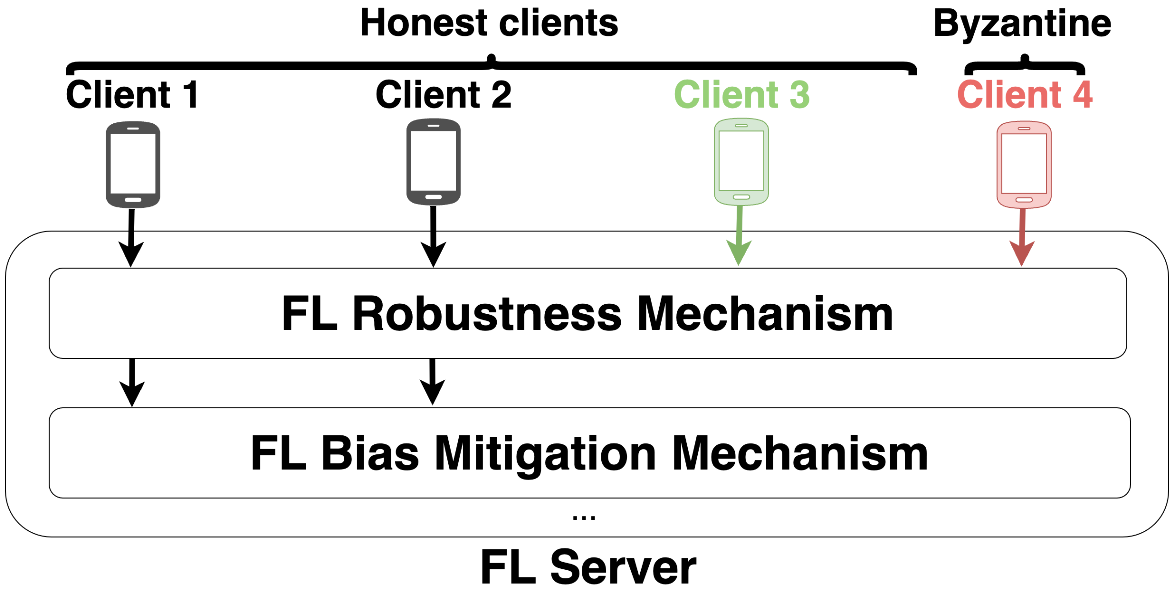
Robustness of FL Against Byzantine Clients



Related work on Robustness of FL Against Byzantine Clients

- **Robust aggregation**: Mitigate impact of Byzantines by estimating the average of honest clients' gradients.
  - Multi-Krum [2]
    - Define a distance to order clients updates
  - Trimmed Means [3]
    - Trim extreme values of the clients' model parameters coordinate-wise
  - RFA [4]
    - Compute the geometric median of clients' updates
  - NDC [14]
    - Apply a norm-thresholding policy on the clients’ updates

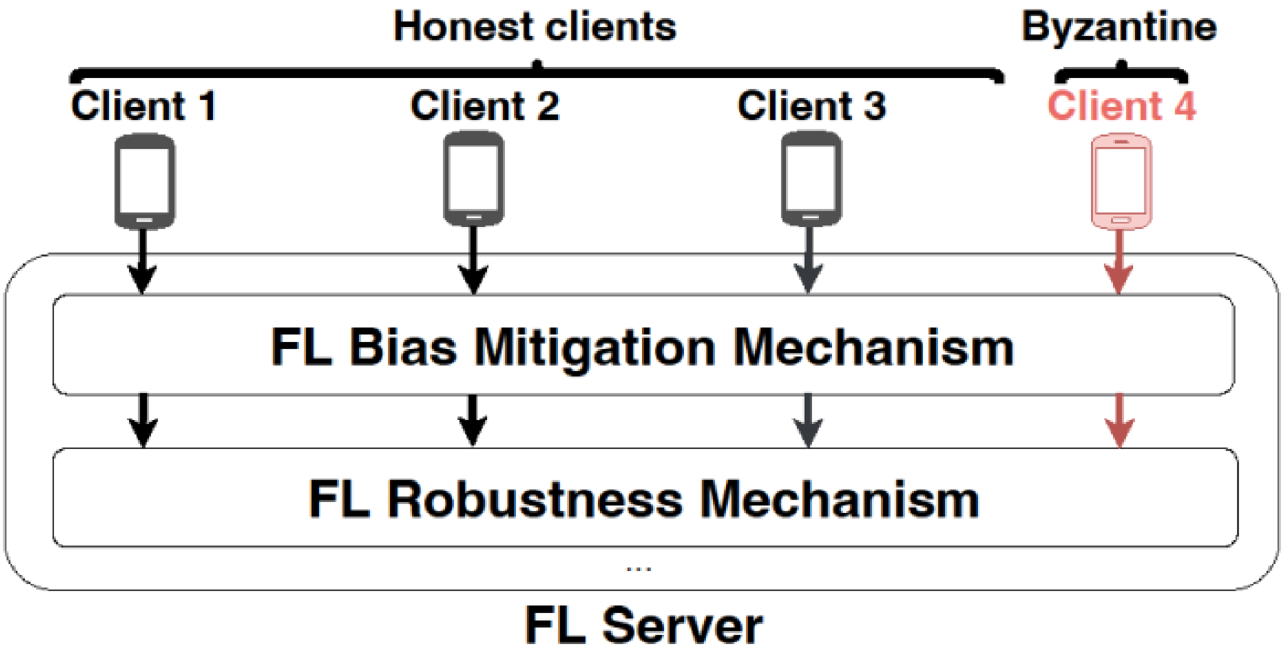
Why Applying a Classical FL Robustness Mechanism Followed by Classical FL Bias Mitigation Does Not Work



The FL Robustness Mechanism may also **filter** the client 3 (**"Honest but minority")** update, losing interesting data for the FL Bias Mitigation mechanism.

Observation 1: Using classical robust aggregators may eliminate honest clients, affecting the normal behavior of FL bias mitigation

Why Applying Classical FL Bias Mitigation Followed By a Classical FL Robustness Mechanism Does Not Work



The FL Bias Mitigation Mechanism is directly exposed to the Byzantine influence

Observation 2: Using the classical FL bias mitigation method before any robustness mechanism expose the bias mitigation method to the influence of the Byzantine clients.

Problem Illustration

- Experiment 1 : Impact of 4 robust aggregation methods on model bias in FL
- Experiment 2 : Interaction between FL bias mitigation (FCFL) and FL robustness mechanisms

Dataset	Task & Model	Target Attribute	Sensitive Attributes	FL Setup
MEPS [10]	Binary classification using Logistic Regression	Medical facility utilization	Race	4-client FL setup with opposite trend to 3 other clients
Adult [11]	Binary classification using Logistic Regression	Income	Gender and age	10-client FL setup with heterogeneity generated by a Dirichlet function

Evaluation Setup

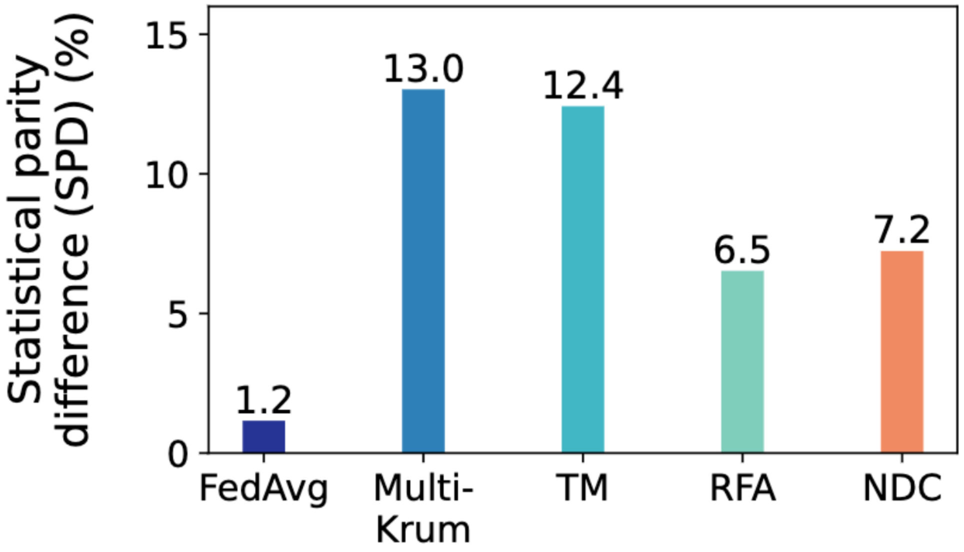
- Used bias metric : **Statistical Parity Difference (SPD):**

$$SPD_S = |Pr(y = 1|S = 1) - Pr(y = 1|S = 0)|$$

Proportion of positively predicted outcome for data belonging to the **priviledged** group

Proportion of positively predicted outcome for data belonging to the **unprivileged** group

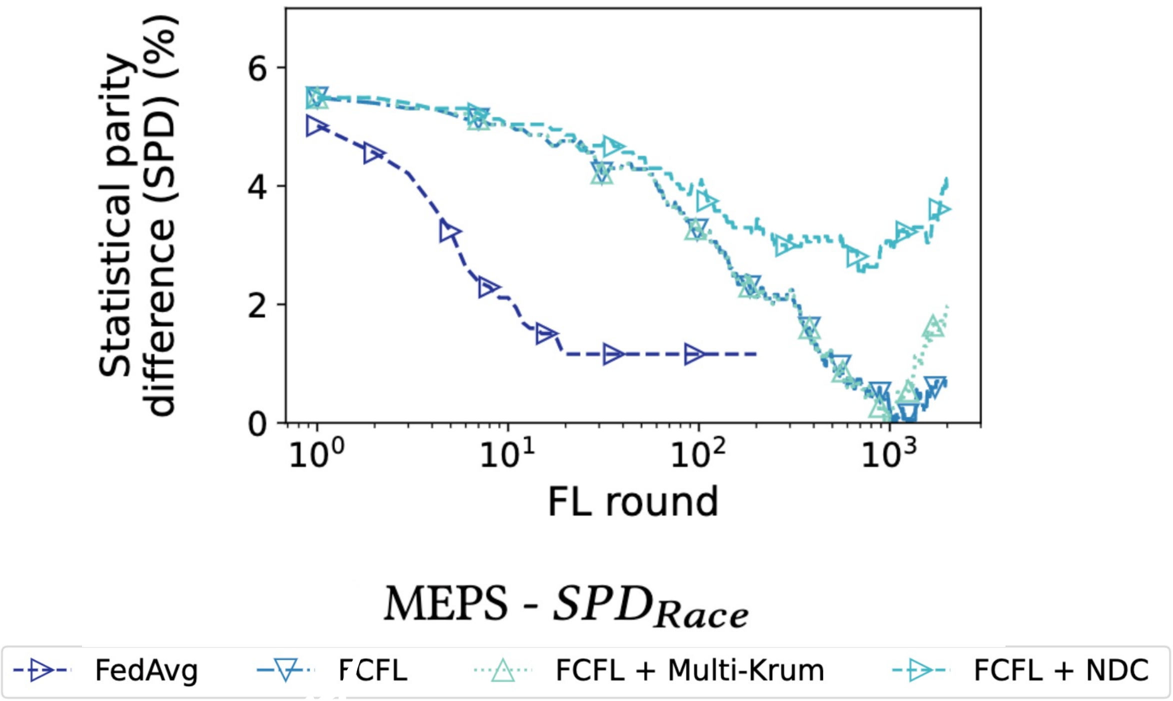
Impact of Robust Aggregation on Model Bias in FL



MEPS - SPD_{Race}

Robust aggregators, increased model bias compared to the FL baseline without any Byzantine attacks.

Interaction between FL bias mitigation and FL robustness mechanisms



Robust aggregators, modified the method behavior and sometime degrade its bias mitigation performance

Related work

- **Ditto: Fair and Robust Federated Learning Through Personalization (Li et al., ICML, 2021)**:
  - **Client-Level Fairness**, with a **robustness objective** (ensuring high accuracy against model poisoning) using **model personalization**.
- **Fair detection of poisoning attacks in federated learning on non-i.i.d. data (Singh et al., Data Mining and Knowledge Discovery 1, 2023)**:
  - Reducing the amount of falsely predicted malicious clients, under assumption that one client = one demographic group (that they need to share)
  - Creating cluster of clients with statistical parity, and then eliminate client that are too far from the created centroids

Research Directions

- False predictions of Byzantine clients must be reduced to preserve important data representativity.
- Asking FL clients additional data distribution information to detect Byzantine clients.
  - Selecting "honest and minority" clients can improve data representativity for minorities.
- Recent development in robust aggregation in non-IID setup (Karimireddy et al. [12], Allouah et al. [13])
- Using Trusted Execution Environments (TEEs)

Summary

- Constructing a FL system with robustness and model bias guarantees is a **critical need** but is **very challenging** to achieve.
- We analyse the issues when trying to implement a system combining the approaches used to solve Byzantine robustness and bias mitigation separately.
- Possible research directions for building robust, bias-free FL are formulated.

Thank you ! Any questions ?

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Y. Arcas.. Communication-Efficient Learning of Deep Networks from Decentralized Data. International Conference on Artificial Intelligence and Statistics PMLR, 2017

[2] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries : Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 30, 2017.

[3] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine robust distributed learning : Towards optimal statistical rates. In International Conference on Machine Learning, pages 5650–5659. PMLR, 2018.

[4] Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. 2019. Robust Aggregation for Federated Learning. IEEE Transactions on Signal Processing, 70, 2022

[5] Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1): 1–33.

[6] Du, Wei et al. “Fairness-aware Agnostic Federated Learning.” SDM (2020).

[7] Zhang, Daniel Yue et al. “FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models.” 2020 IEEE International Conference on Big Data (Big Data) (2020): 1051-1060.

[8] Ezzeldin, Yahya H. et al. “FairFed: Enabling Group Fairness in Federated Learning.” ArXiv abs/2110.00857 (2021)

[9] Cui, Sen et al. “Fair and Consistent Federated Learning.” ArXiv abs/2108.08435 (2021)

[10] Cohen, Steven B. Sample Design of the 1997 Medical Expenditure Panel Survey, Household Component. No. 1. US Department of Health and Human Services, Public Health Service, Agency for Healthcare Research and Quality, 2001.

[11] Dua, Dheeru, and Casey Graff. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2017, http://archive.ics.uci.edu/ml.

[12] Karimireddy, Sai Praneeth, Lie He, and Martin Jaggi. "Byzantine-robust learning on heterogeneous datasets via bucketing." arXiv preprint arXiv:2006.09365 (2020).

[13] Allouah, Youssef, et al. "Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity." International Conference on Artificial Intelligence and Statistics. PMLR, 2023.

Appendix

- Byzantine robustness objective :

$$\mathbb{E}\|\hat{\theta} - \overline{\theta}\|^2 < k\rho\delta^2 \quad \text{with} \quad \overline{\theta} = \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\theta_j \text{ the true average}$$

- Bias mitigation objective :

$$\min_{\theta} f(\theta) = \min_{\theta} \frac{1}{|\mathcal{H}|}\sum_{k\in\mathcal{H}} f_k(\theta)$$
$$\text{s.t } |SPD_S(\theta)| \leq \epsilon$$