Laboratoire
D'Informatique
De Grenoble

Laboratoire InfoRmatique
d'Imagerie et Systèmes

Towards Bias Mitigation in Federated Learning

WAX SRCPR May 12th

By:

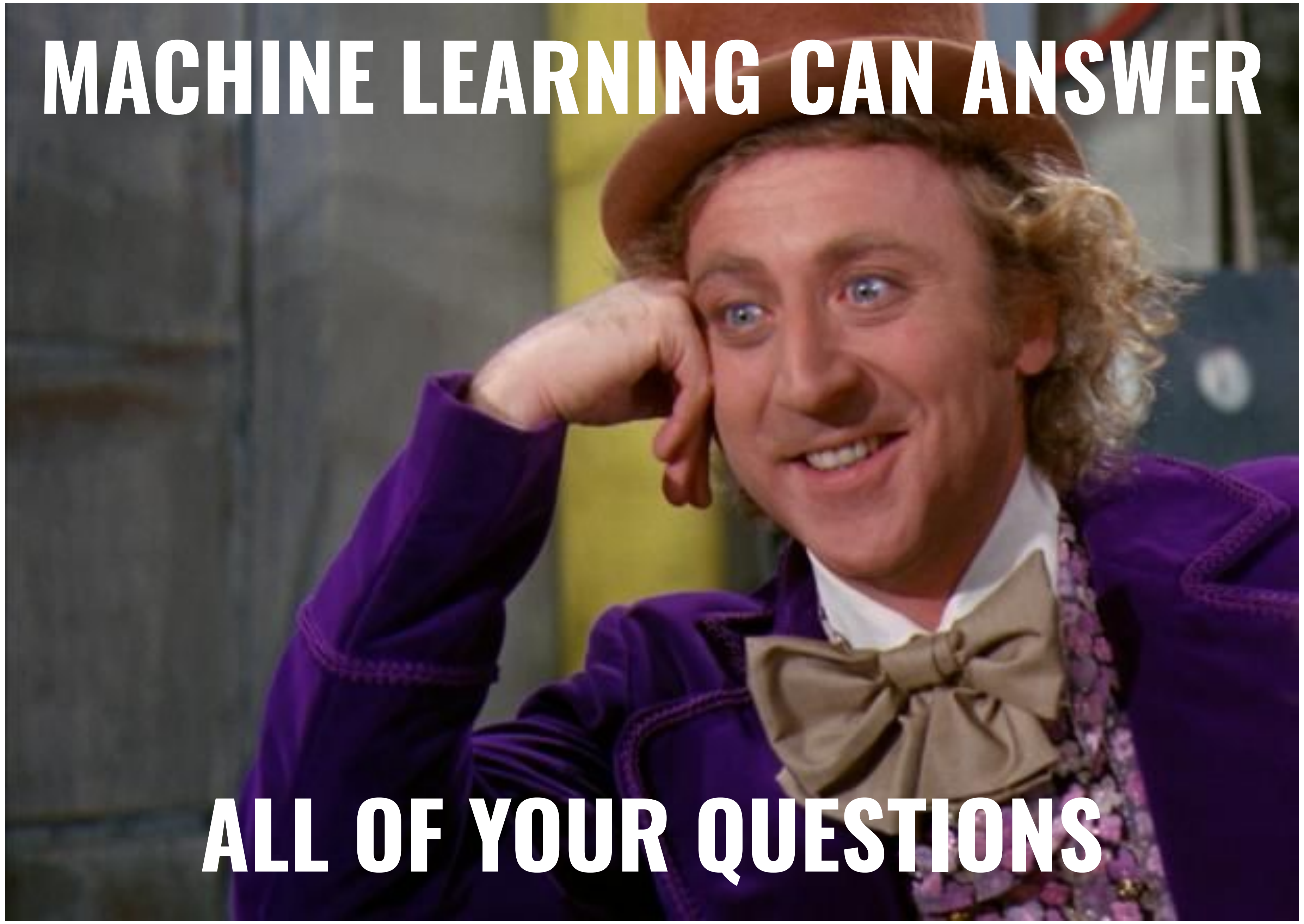
Yasmine Djebrouni

PhD student at LIG-LIRIS

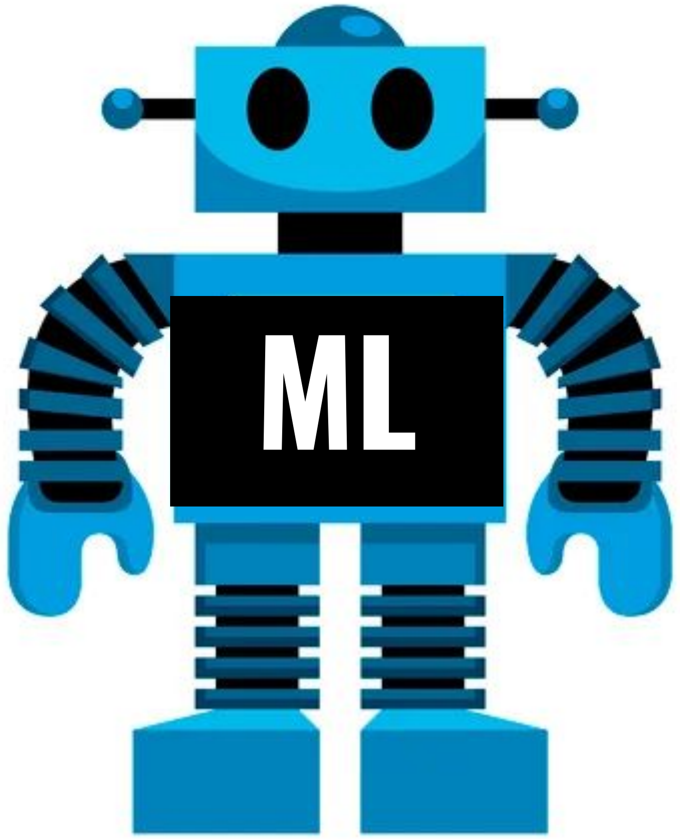
Supervised by Sara Bouchenak (LIRIS) & Vania Marangozova-Martin (LIG)

Outline

I. Bias and federated learning
II. Our approach to bias mitigation in federated learning
III. Preliminary results
IV. Next

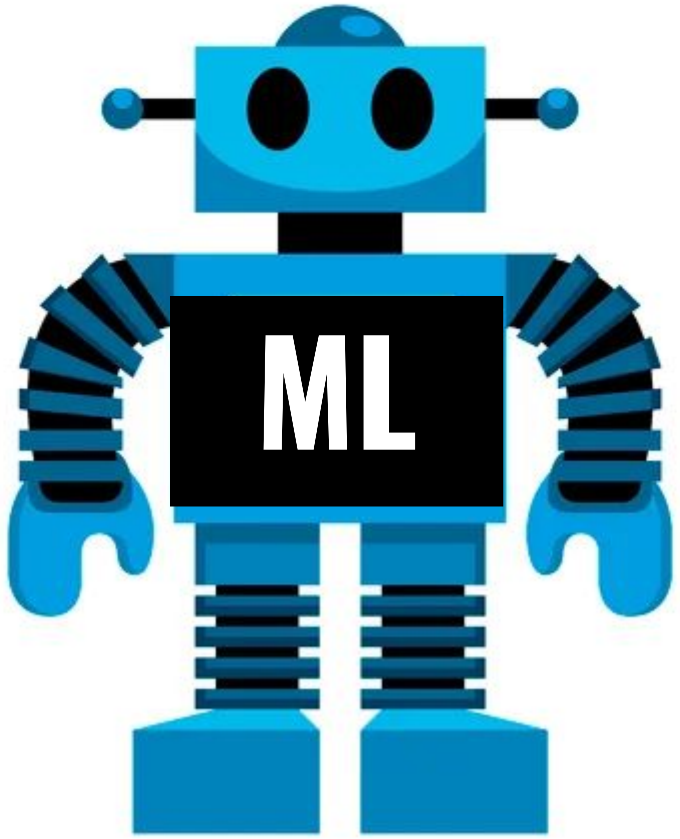

MACHINE LEARNING CAN ANSWER
ALL OF YOUR QUESTIONS

Data from reality



Predictions and decisions

Data from **reality**



Predictions and decisions

Our reality is **biased** due to historical prejudice
Our data is not balanced
Our data labeling is unfair and subjective

Our reality is **biased** due to historical prejudice
Our data is not balanced
Our data labeling is unfair and subjective

…Uh, reality is not a good source to learn from

Machine learning is learning how to be racist, sexist and discriminatory…

i.e machine learning is biased

Technology

When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity

Seven actions social change leaders and machine learning developers can ... gender-

... algorithm used ... hospitals to predict which patient ... medical care heavily favored white patients ove... While race itself wasn't a variable used in this algorithm, a... related to race was, which was healthcare cost h... many healthcare ... tients incurred l... conditio...

What Do We Do About the Biases in AI?

by James Manyika, Jake Silberg, and Brittany Presten

October 25, 2019

Tech policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copyin... the mistakes of the past.

January 21, 2019

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Technology

When Good Algo
and How to Adva

Seven actions social change leaders ar

medical ca
While race itself wasn
related to ra

What Do We Do A
Biases in AI?

by James Manyika, Jake Silberg, and Brit

October 25, 2019

Two drug possession arrests.
ML be like:

Low risk: 3 **High risk: 10**

ng people to
etting it wrong

sk assessment tools could mean
mistakes of the past.

January 21, 2019

Many F
Biased

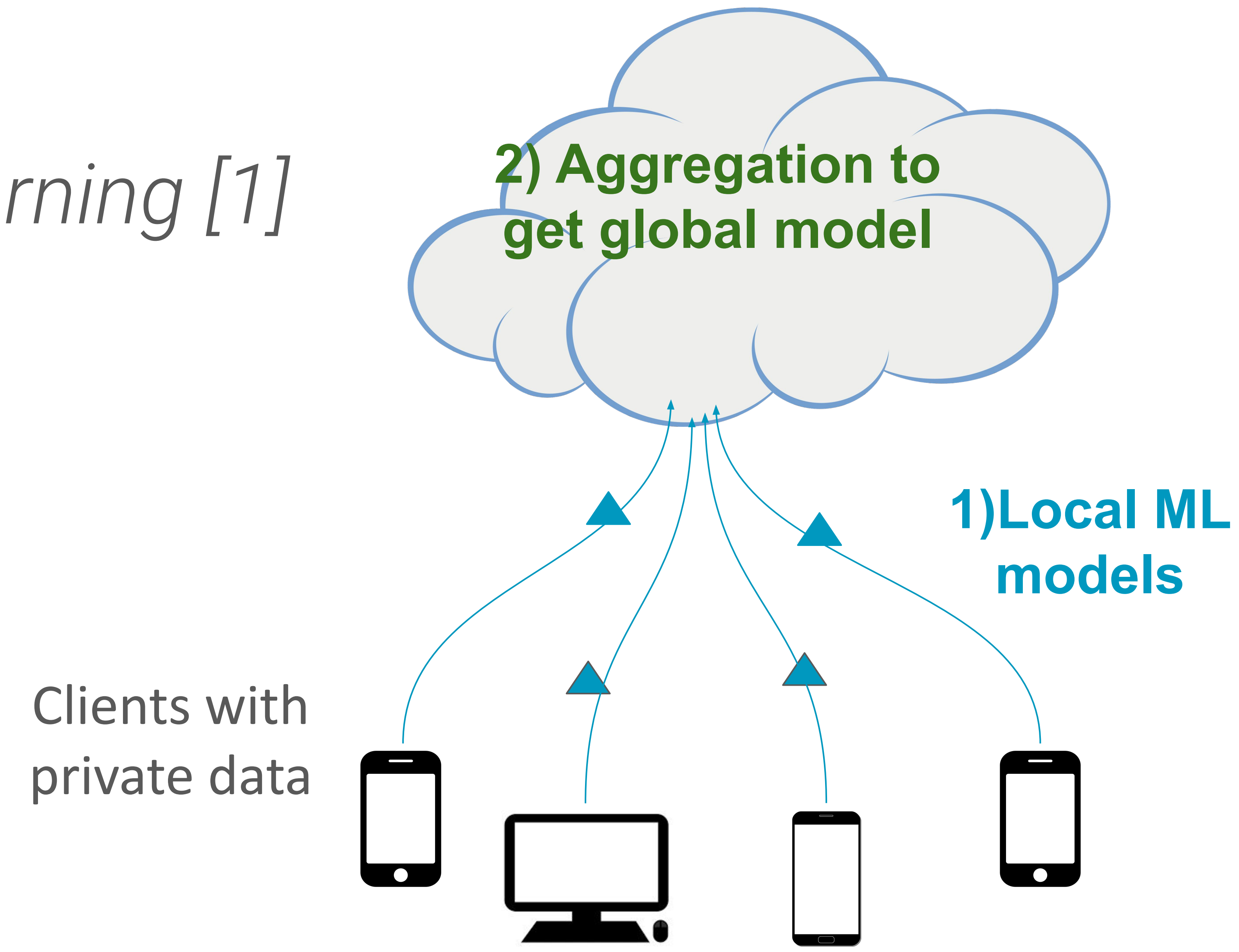
Algorithms falsely
10 to 100 times more than Caucasian faces,
National Institute of Standards and Technology found.

Bias is getting worse with federated learning [1]

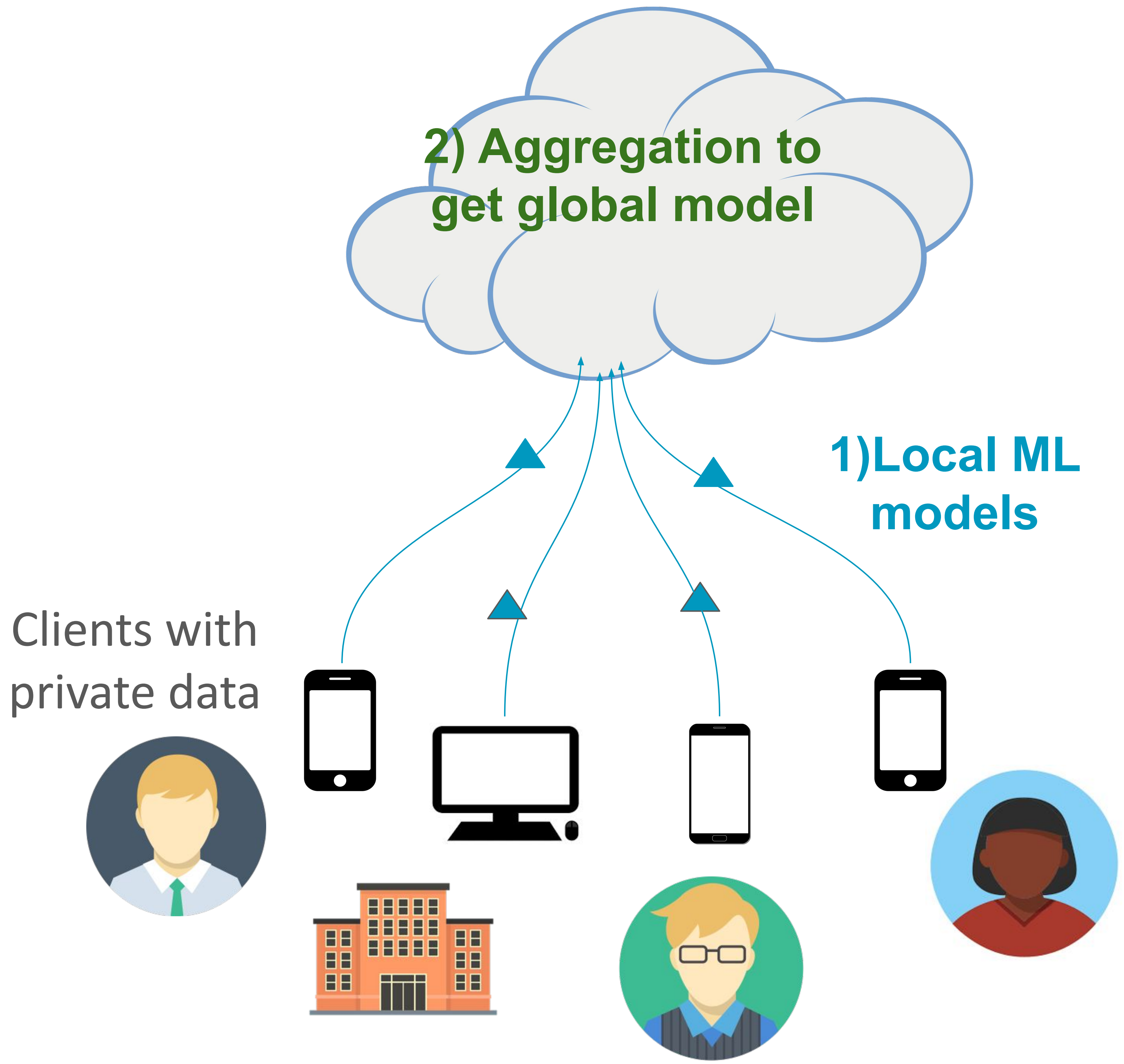
- Emergence of FL
- New distributed paradigm
- Privacy-friendly
- Communication efficient

Bias is getting worse with federated learning [1]

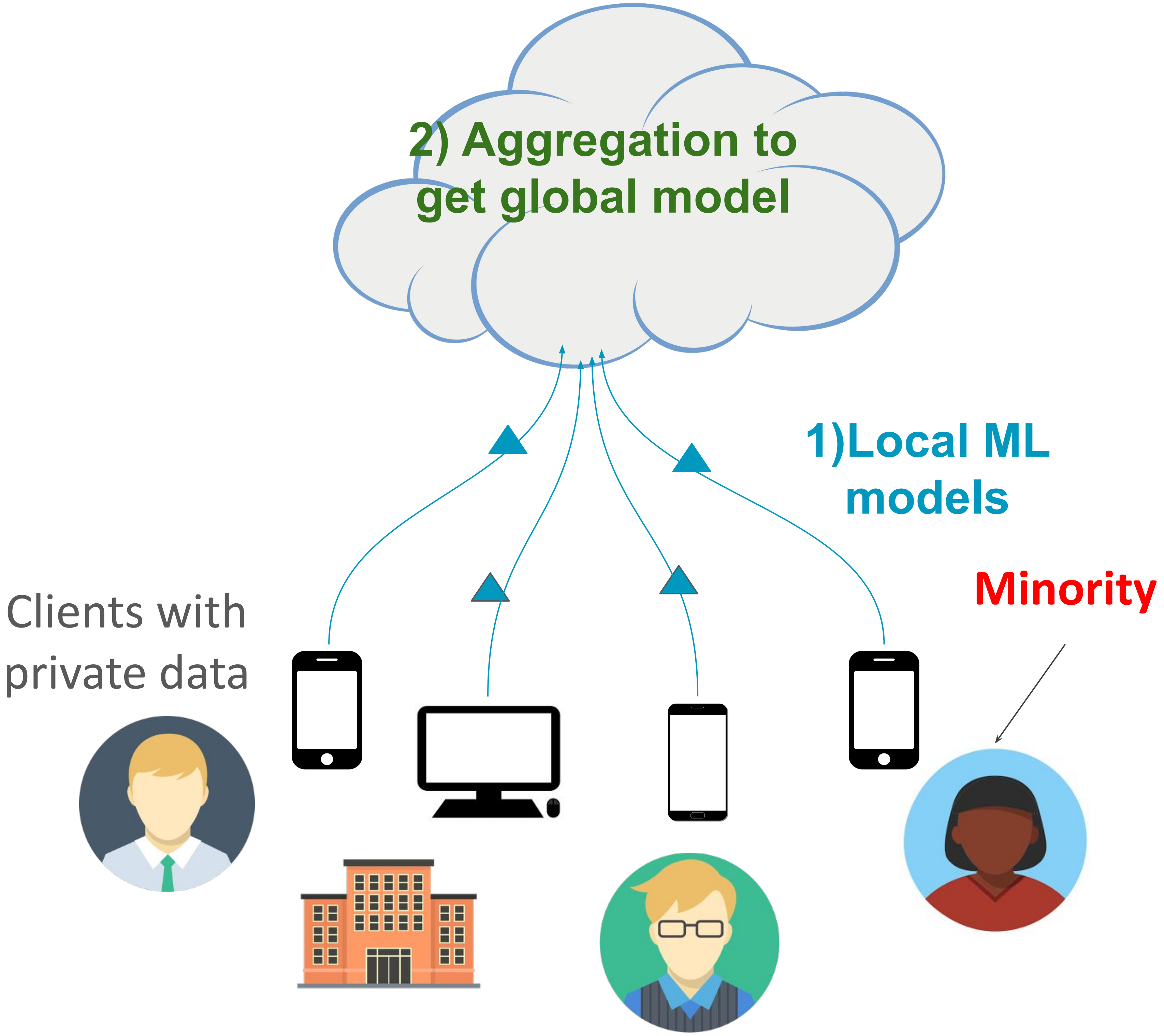
- Emergence of FL
- New distributed paradigm
- Privacy-friendly
- Communication efficient



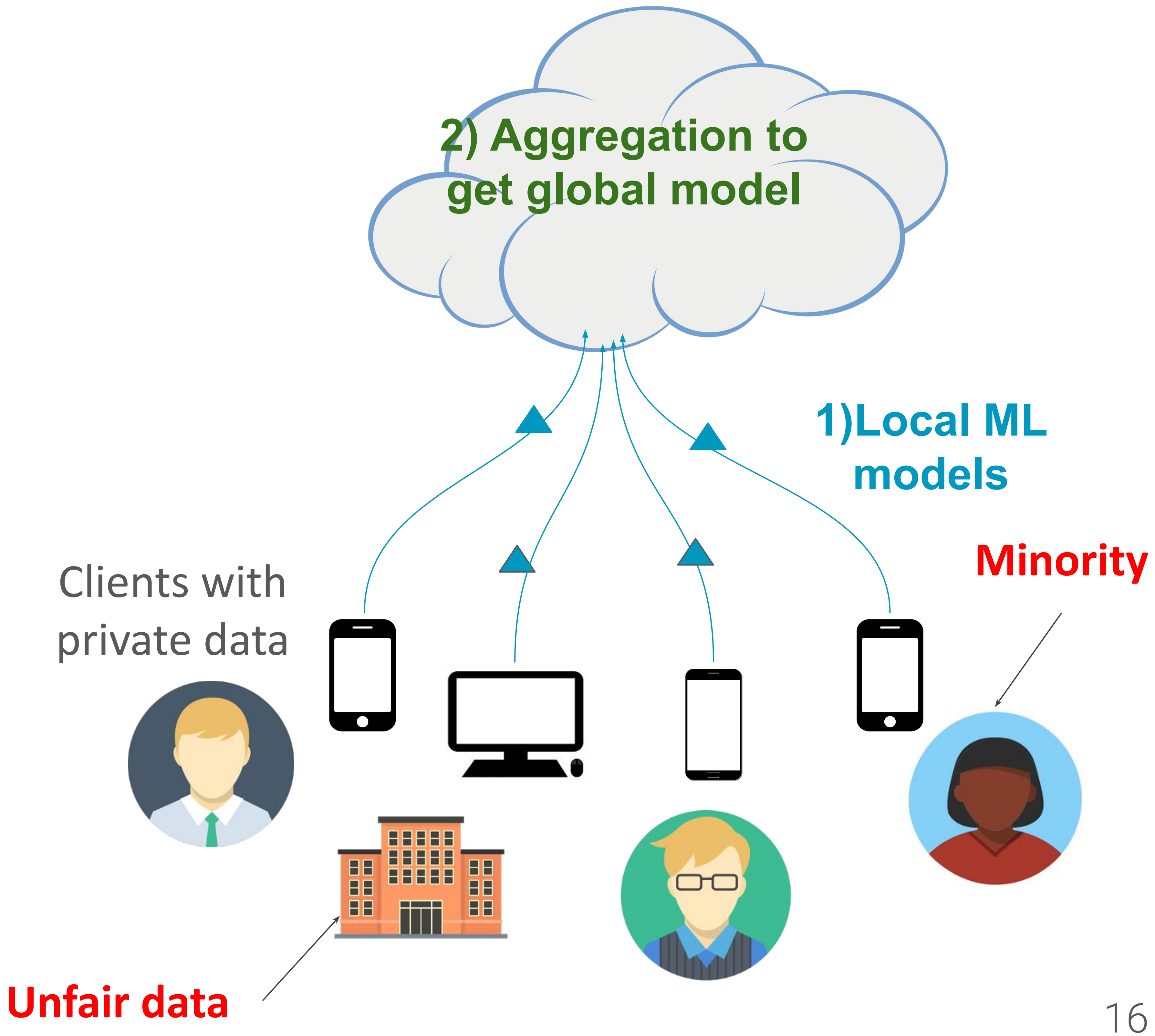
FL can exacerbate machine learning unfairness



FL can exacerbate machine learning unfairness



FL can exacerbate machine learning unfairness



age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48	State-gov	78529	Masters	14	Separated	Prof-specialty	Not-in-family	White	Male	0	0	60	United-States	<=50K
71	Private	105200	HS-grad	9	Married-civ-spouse	Protective-serv	Husband	White	Male	6767	0	20	United-States	<=50K
48	Private	349151	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
45	Local-gov	172111	Bachelors	13	Divorced	Exec-managerial	Unmarried	Black	Female	0	0	60	United-States	<=50K
66	Self-emp-not-inc	182470	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	25	United-States	>50K

Example of unfair data

age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48	State-gov	78529	Masters	14	Separated	Prof-specialty	Not-in-family	White	Male	0	0	60	United-States	<=50K
71	Private	105200	HS-grad	9	Married-civ-spouse	Protective-serv	Husband	White	Male	6767	0	20	United-States	<=50K
48	Private	349151	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
45	Local-gov	172111	Bachelors	13	Divorced	Exec-managerial	Unmarried	Black	Female	0	0	60	United-States	<=50K
66	Self-emp-not-inc	182470	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	25	United-States	>50K

Example of unfair data

- Attributes

age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48	State-gov	78529	Masters	14	Separated	Prof-specialty	Not-in-family	White	Male	0	0	60	United-States	<=50K
71	Private	105200	HS-grad	9	Married-civ-spouse	Protective-serv	Husband	White	Male	6767	0	20	United-States	<=50K
48	Private	349151	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
45	Local-gov	172111	Bachelors	13	Divorced	Exec-managerial	Unmarried	Black	Female	0	0	60	United-States	<=50K
66	Self-emp-not-inc	182470	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	25	United-States	>50K

Example of unfair data

- Attributes
- Y = Target/decision variable (salary, criminality, intelligence) disadvantageous towards a group

age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48	State-gov	78529	Masters	14	Separated	Prof-specialty	Not-in-family	White	Male	0	0	60	United-States	<=50K
71	Private	105200	HS-grad	9	Married-civ-spouse	Protective-serv	Husband	White	Male	6767	0	20	United-States	<=50K
48	Private	349151	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
45	Local-gov	172111	Bachelors	13	Divorced	Exec-managerial	Unmarried	Black	Female	0	0	60	United-States	<=50K
66	Self-emp-not-inc	182470	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	25	United-States	>50K

Example of unfair data :

- Attributes
- Y = Target/decision variable (salary, criminality, intelligence) disadvantageous towards some groups
- Sensitive attributes (race, gender, age..), define groups: privileged and unprivileged (Females/males, whites/non-whites..)

age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48	State-gov	78529	Masters	14	Separated	Prof-specialty	Not-in-family	White	Male	0	0	60	United-States	<=50K
71	Private	105200	HS-grad	9	Married-civ-spouse	Protective-serv	Husband	White	Male	6767	0	20	United-States	<=50K
48	Private	349151	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
45	Local-gov	172111	Bachelors	13	Divorced	Exec-managerial	Unmarried	Black	Female	0	0	60	United-States	<=50K
66	Self-emp-not-inc	182470	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	25	United-States	>50K

The data is **unfair** if the **label/decision variable** is **dependant** on the **sensitive attribute** [2].

Example of biased data: sexist data where men have higher salaries than women.

Mathematically [2]

Label/decision variable is **dependant** on the **sensitive attribute**

⇕

$$Pr(Y = p|S = priv) \neq Pr(Y = p|S = unpriv)$$

such that
Pr : Probability distribution
Y : decision variable
p : Advantageous decision (eg.high salary)
S : sensitive attribute variable (eg. gender)
priv/unpriv : privileged and unprivileged group (eg. women and men)

The amount of unfairness can be measured by disparate impact:

$$\beta(\theta) = \frac{Pr(Y = p|S = unpriv)}{Pr(Y = p|S = priv)}$$

(i) Characterize the actual impact of Federated Learning on bias.
(ii) Propose novel FL selection and aggregation algorithms for bias mitigation.
(iii) Take into account accuracy and robustness in FL.

Our approach

1) Privately estimating the bias brought by each client.
2) Directly deal with the source of bias (biased client)

1) How to measure clients bias without looking at their data?

- Exploit models update
- Exploit public/synthetic test data

$$\beta(\theta) = \frac{Pr(\hat{Y} = p \ |S = unpriv)}{Pr(\hat{Y} = p \ |S = priv)}$$

2) How to deal with the identified biased client?

Diminishing its impact on the FL model (reweighting)

or

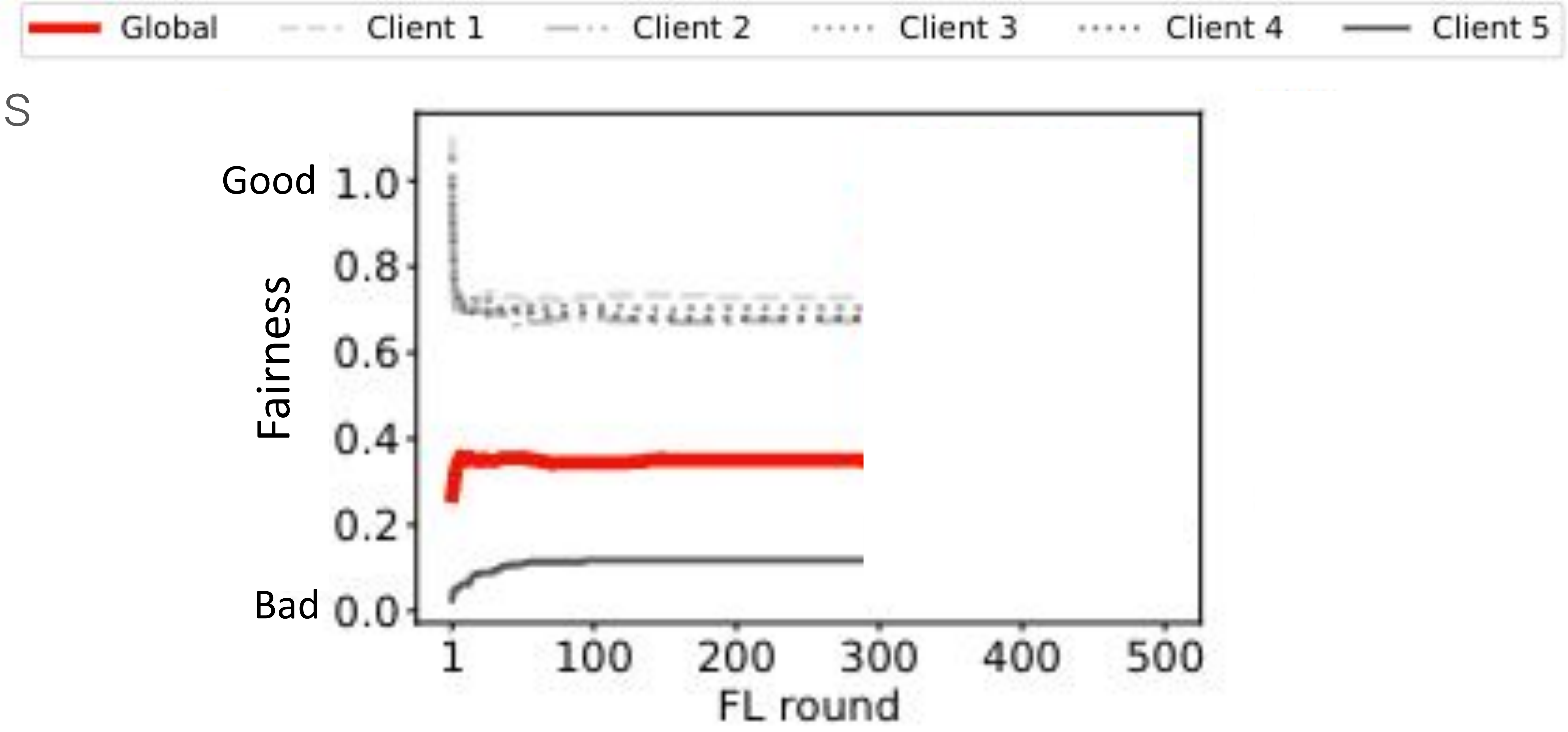
Aggregating together the clients that mutually cancel each other effects

or

A biased client is a poisoned client, ignore its model!

Evaluation of our approach

We set up an unfair FL scenario and record global and local models unfairness

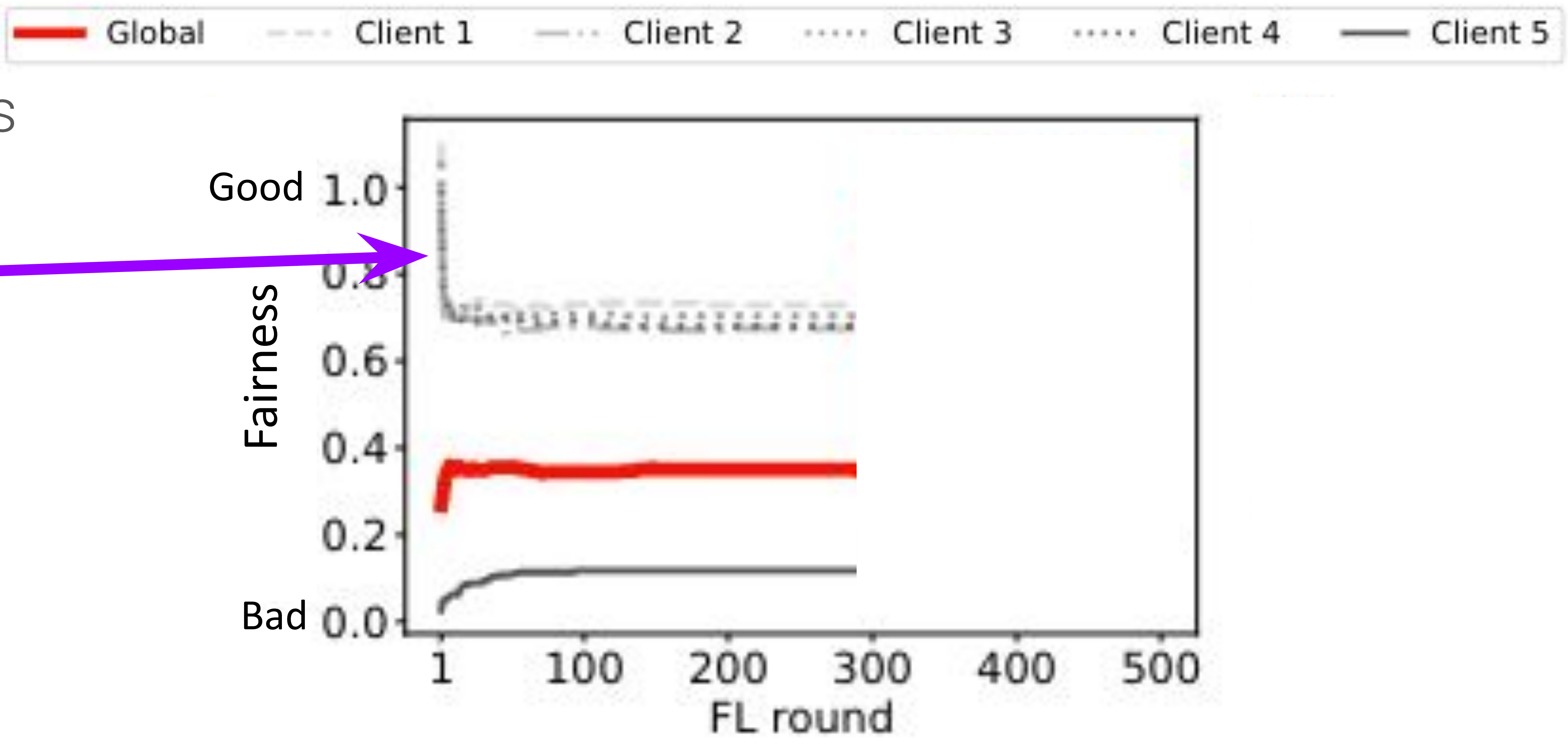


(a) Model bias

Evaluation of our approach

We set up an unfair FL scenario and record global and local models unfairness

- 4 fair clients

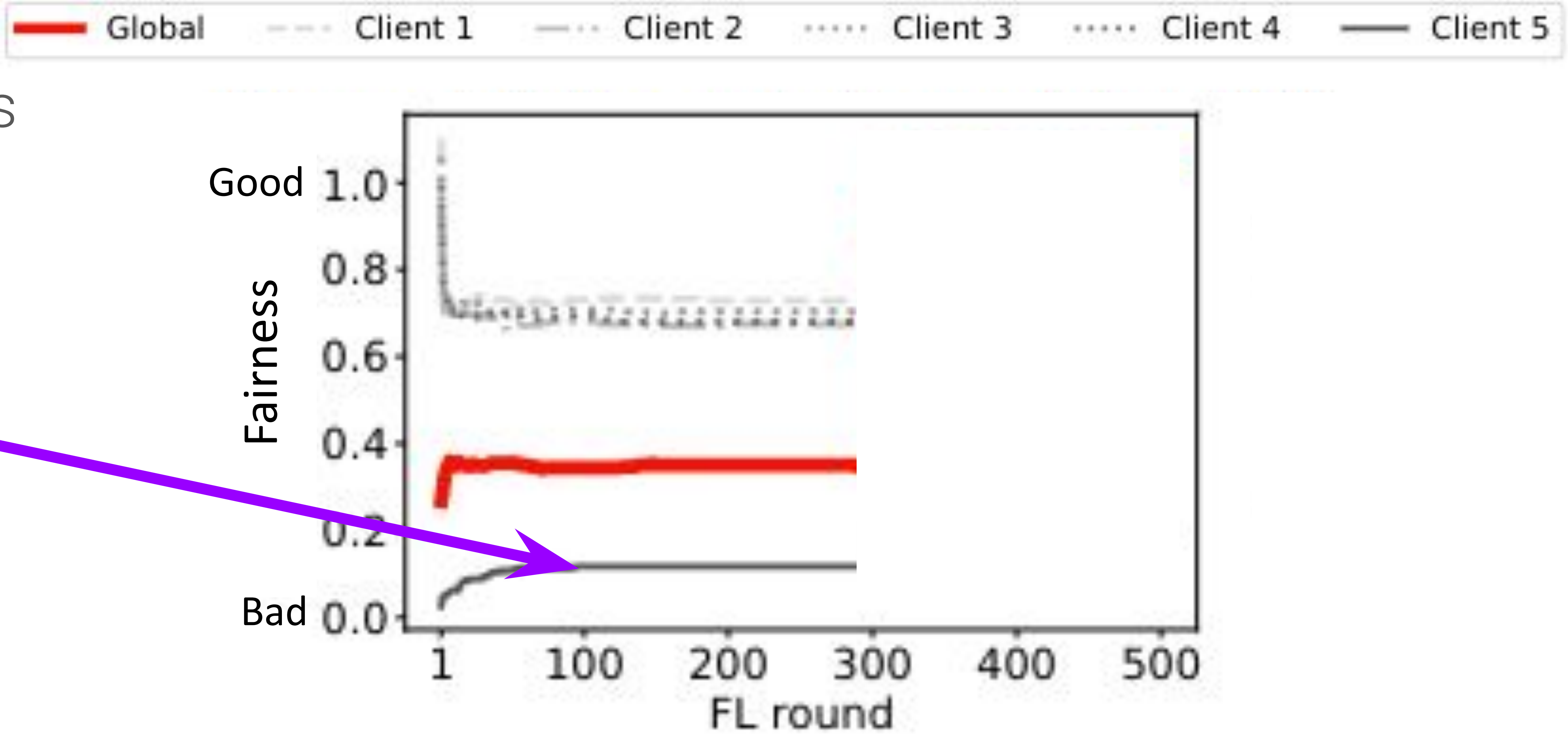


Model bias

Evaluation of our approach

We set up an unfair FL scenario and record global and local models unfairness

- 4 fair clients
- 1 unfair client



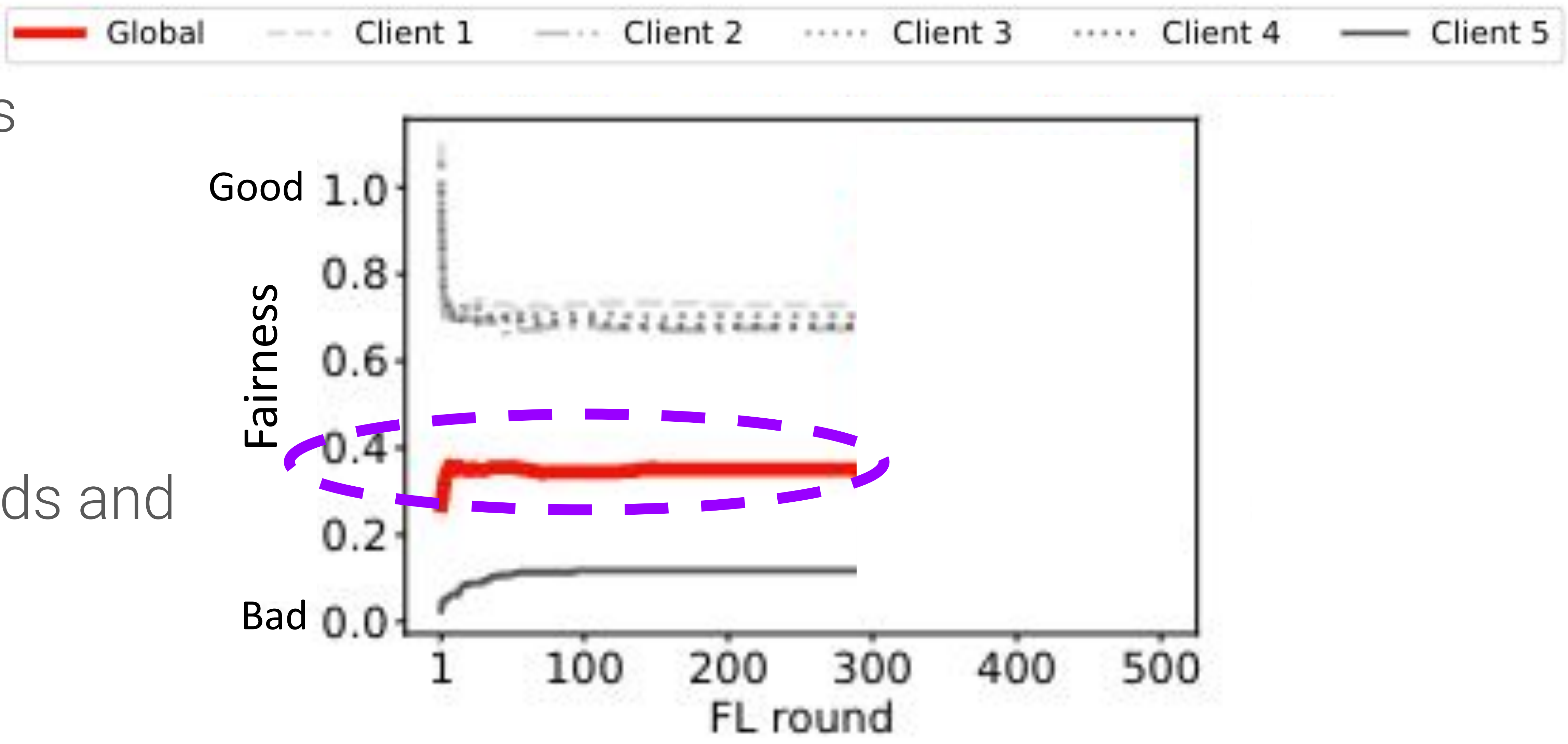
Model bias

Evaluation of our approach

We set up an unfair FL scenario and record global and local models unfairness

- 4 fair clients
- 1 unfair client

We run FL training for several rounds and observe global fairness



Model bias

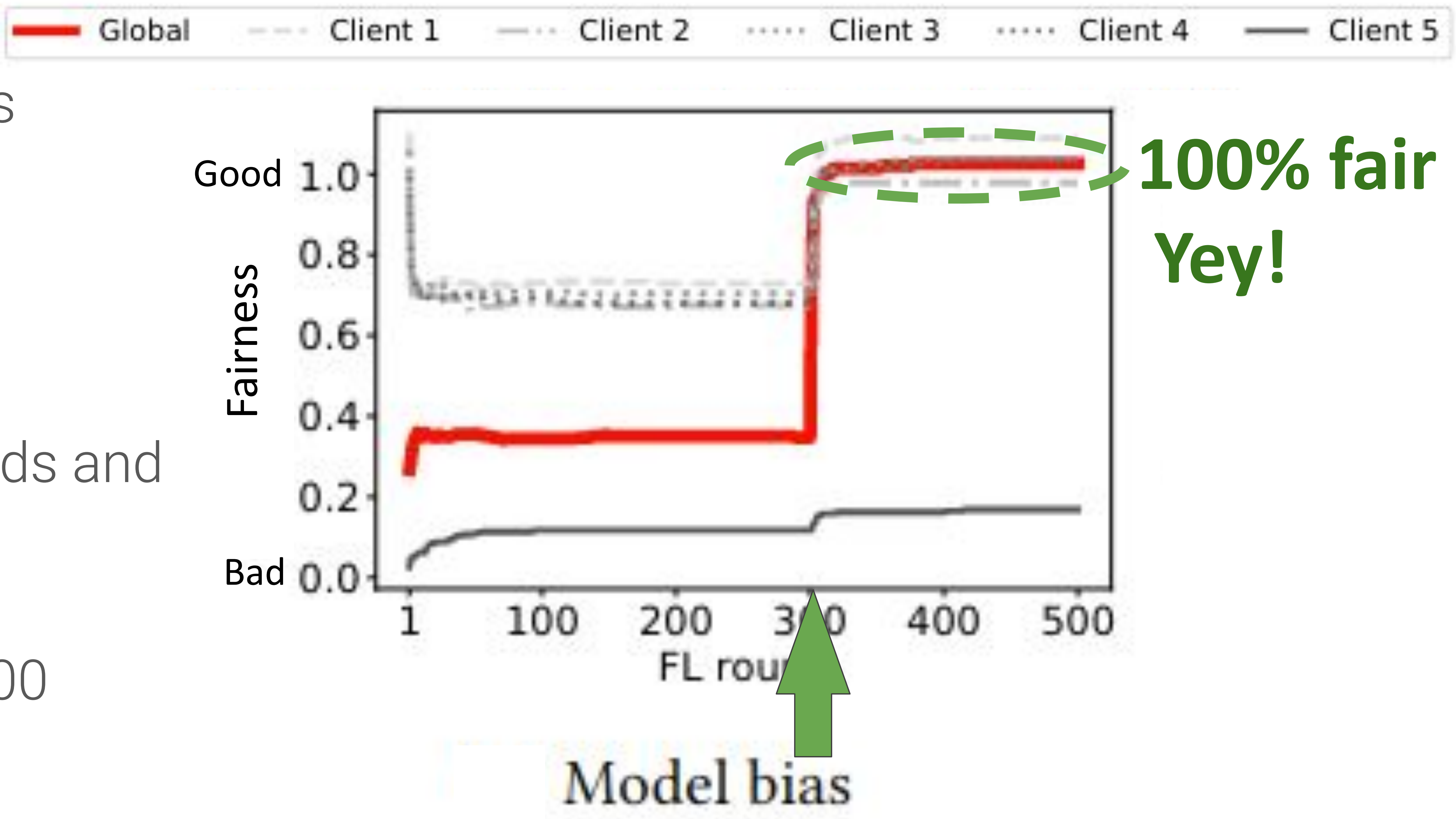
Evaluation of our approach

We set up an unfair FL scenario and record global and local models unfairness

- 4 fair clients
- 1 unfair client

We run FL training for several rounds and observe global fairness

We apply our approach at round 300



What’s next ?

- Evaluate our approaches with more scenarios with different data distributions.
- Combine with classical ML approaches to improve performance.
- Propose approaches to ensure accuracy and robustness.

THANK YOU!

ANY QUESTIONS?

Yasmine Djebrouni

yasmine.djebrouni@univ-grenoble-alpes.fr

2022

References

[1] B. MCMAHAN, E. MOORE, D. RAMAGE, S. HAMPSON, AND B. A. Y ARCAS, Communication-Efficient Learning of Deep Networks From Decentralized Data, in Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282

[2] M. HARDT, E. PRICE, AND N. SREBRO, Equality of Opportunity in Supervised Learning, Advances in neural information processing systems, 29 (2016), pp. 3315–3323

Related Work Limitations :

- Require private data information.
- Assume clients and server are trustworthy.
- Consider simple use cases (binary classification, 1 binary sensitive attribute)

Bias measurement

There exist several notion of bias, depending if reality is already biased or perfectly fair.

$$\beta(\theta) = \frac{Pr(\hat{Y} = p^*|S = unpriv)}{Pr(\hat{Y} = p^*|S = priv)}$$

Perfectly fair model : proportion of advantageous outcome for privilidged and unpriviledged groups are equal.

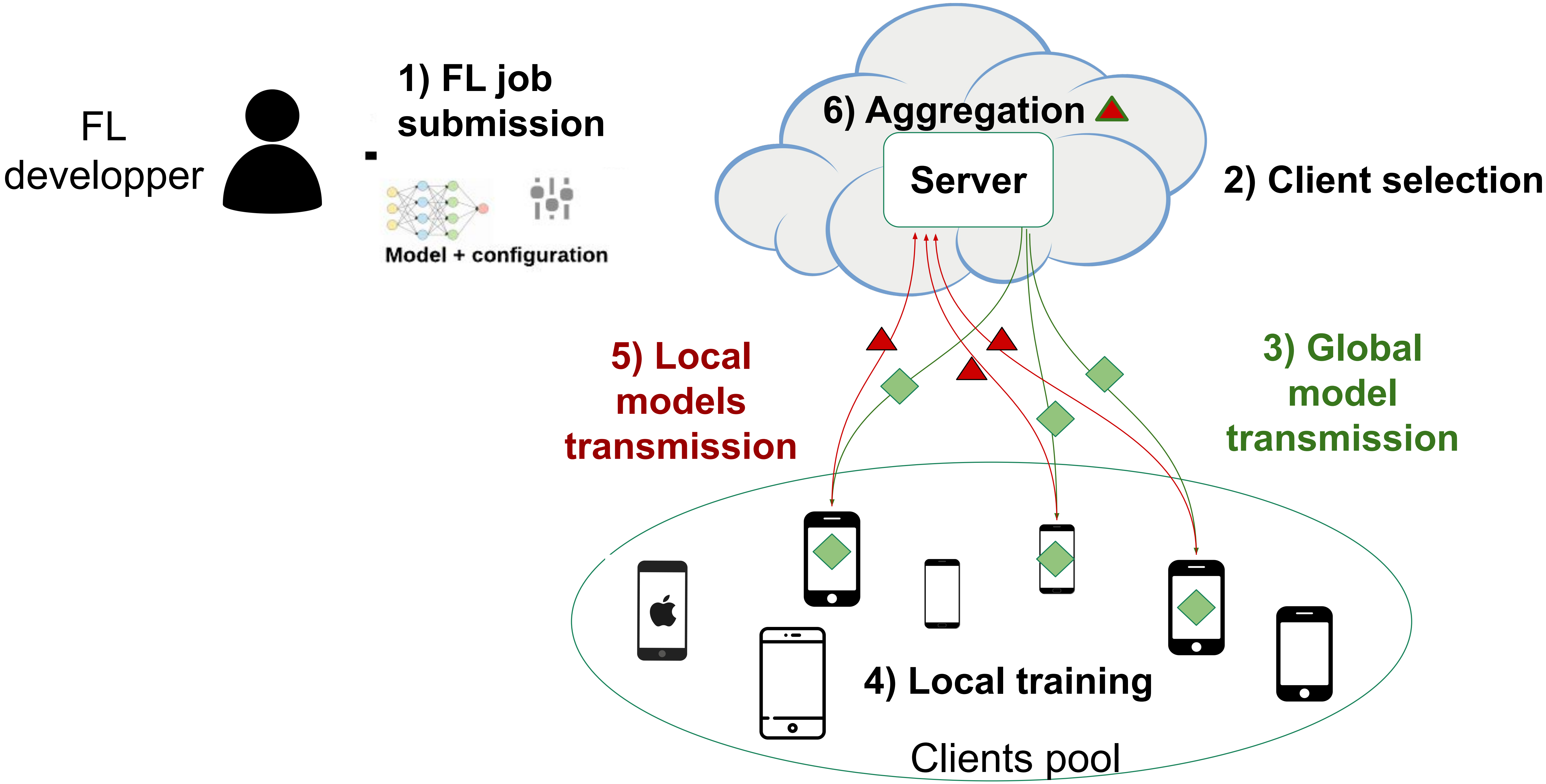
Bias problem formulation

We consider a binary FL classification
(X1, .., Xd) denote the features,
Y denotes the class label
Yˆ is the classifier prediction result for a given data record.

We consider two groups of data: a privileged group which prediction results have a given positive property p*(e.g. people who earn a high salary), and an unprivileged group (e.g. people with a low salary).
Let S be a sensitive feature which, for simplicity, we assume to be binary S ∈

X_1	X_2	...	S	...	X_d	Y
x_{11}	x_{12}	...	s_1	...	x_{1d}	y_1
x_{21}	x_{22}	...	s_2	...	x_{2d}	y_2
...	...	...	...	...	...	...
$x_{n_1 1}$	$x_{n_1 2}$	...	s_{n_1}	...	$x_{n_1 d}$	y_{n_1}

Overview on a FL system



Machine learning is used everywhere because:

Machine learning learns the patterns that exist in our reality, and reproduce them, and generalize them to new data

Technology

When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity

Seven actions social change leaders and machine learning developers can ... gender-

...algorithm used ... hospitals to predict which patient... medical care heavily favored white patients ove... While race itself wasn't a variable used in this algorithm, a... related to race was, which was healthcare cost h... many healthcare ... tients incurred l... onditi...

What Do We Do About the Biases in AI?

by James Manyika, Jake Silberg, and Brittany Presten

October 25, 2019

Tech policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copyin... the mistakes of the past.

January 21, 2019

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Overview on a bias in ML

Some sources of bias in FL

- Biased reality due to social/historical prejudices.
- Class imbalanced data.
- Feature imbalanced data.
- Non representativity of some populations in FL.
- Non selection of some populations by FL
- Unfair aggregation.

Bias problem formulation

In a biased model, the value of S decides the membership of a data to either the privileged group (i.e. $\hat{Y} = p*$) or to the unprivileged group, namely $S \in \{a = priv, b = unpriv\}$.
Such a model does not provide group fairness [10]. With the latter, elements of the privileged group and unprivileged group have equal probability of having prediction results with a positive property, as formulated below: $Pr(\hat{Y} = p* | S = priv) = Pr(\hat{Y} = p* | S = unpriv)$
(2) Furthermore, in case of FL systems, the cause of bias of the global model can come from all or a subset of clients involved in a FL round. Thus, it is important to precisely determine the origin of bias in a FL system, to adequately mitigate it without hurting model quality.

Bias problem formulation

In a biased model, the value of S decides the membership of a data to either the privileged group (i.e. $\hat{Y} = p*$) or to the unprivileged group, namely $S \in \{a = priv, b = unpriv\}$.
Such a model does not provide group fairness [10]. With the latter, elements of the privileged group and unprivileged group have equal probability of having prediction results with a positive property, as formulated below: $Pr(\hat{Y} = p* | S = priv) = Pr(\hat{Y} = p* | S = unpriv)$
(2) Furthermore, in case of FL systems, the cause of bias of the global model can come from all or a subset of clients involved in a FL round. Thus, it is important to precisely determine the origin of bias in a FL system, to adequately mitigate it without hurting model quality.