Federated Learning: Incentives and Fairness

Ruta Mehta

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Mintong Kang

Aniket Murhekar

Zhuowen Yuan

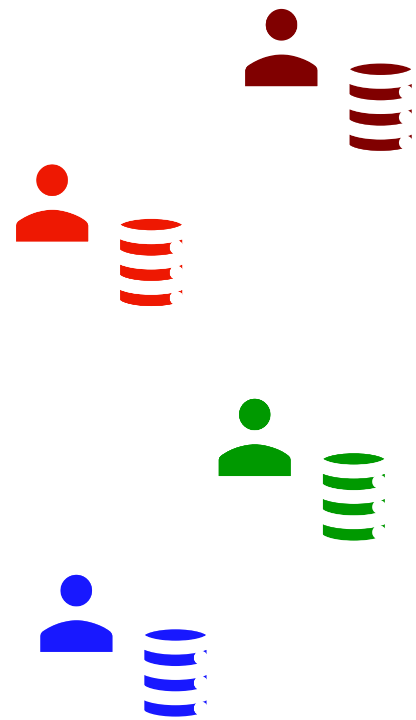
Jiaxin Song

Bhaskar R. Chaudhury

Bo Li

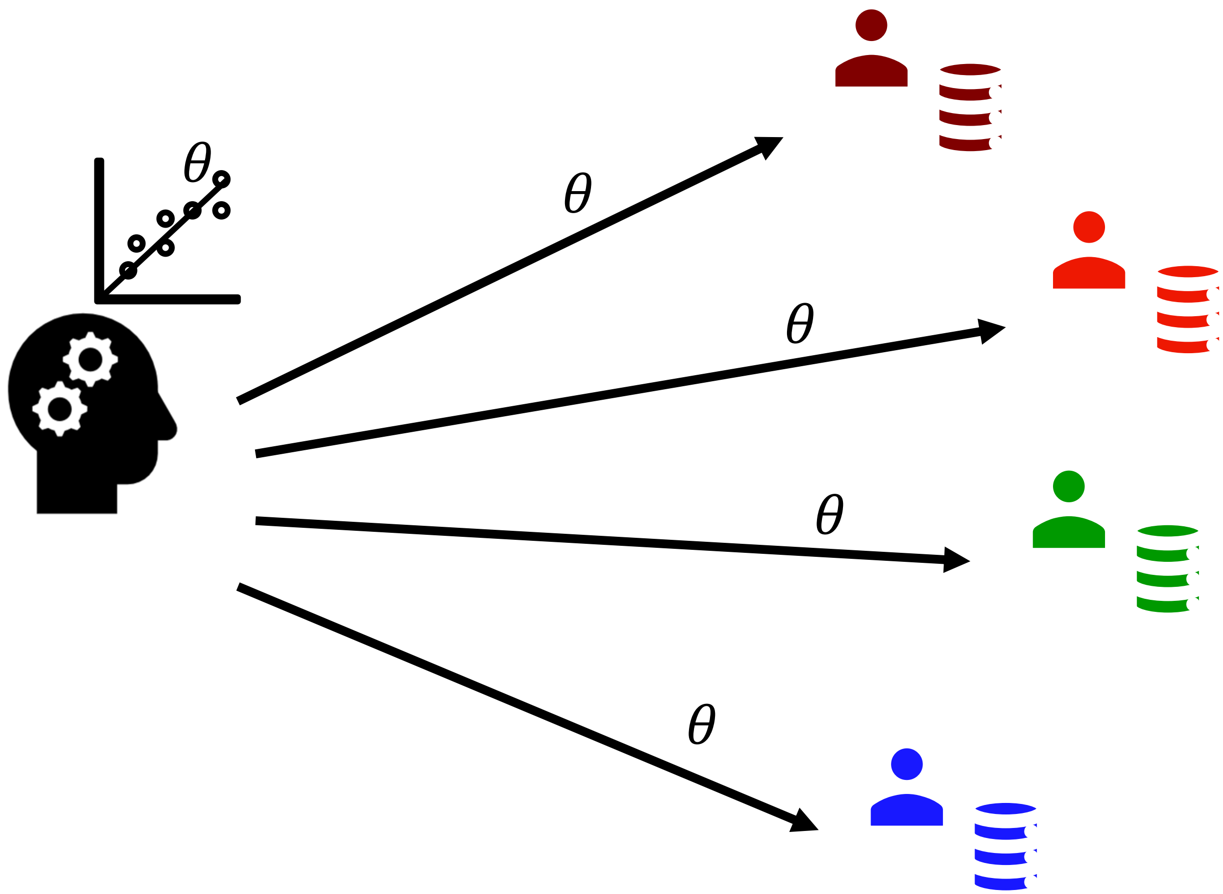
Linyi Li

Ariel Procaccia



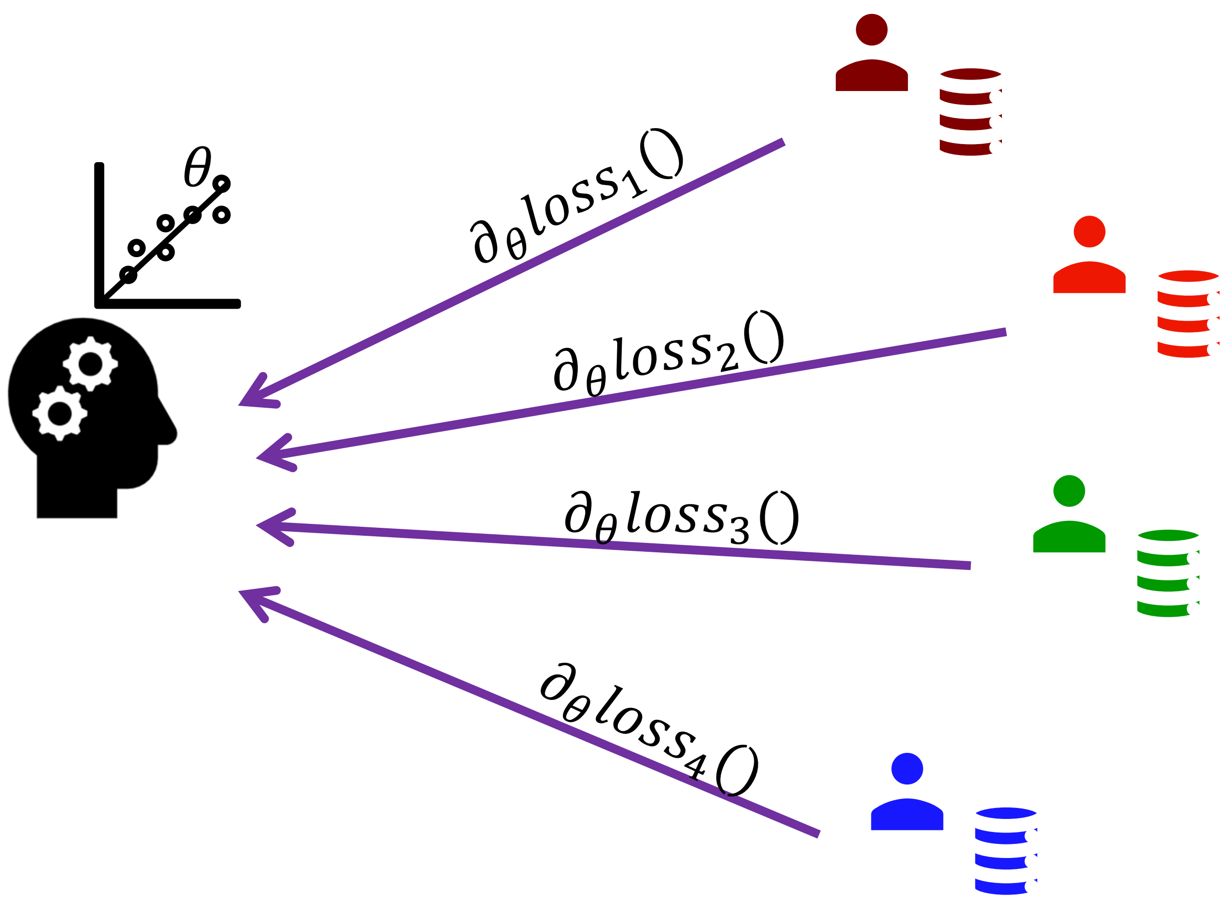
Goal: Learn using everyone’s data without sharing it!

Federated Learning (FL)



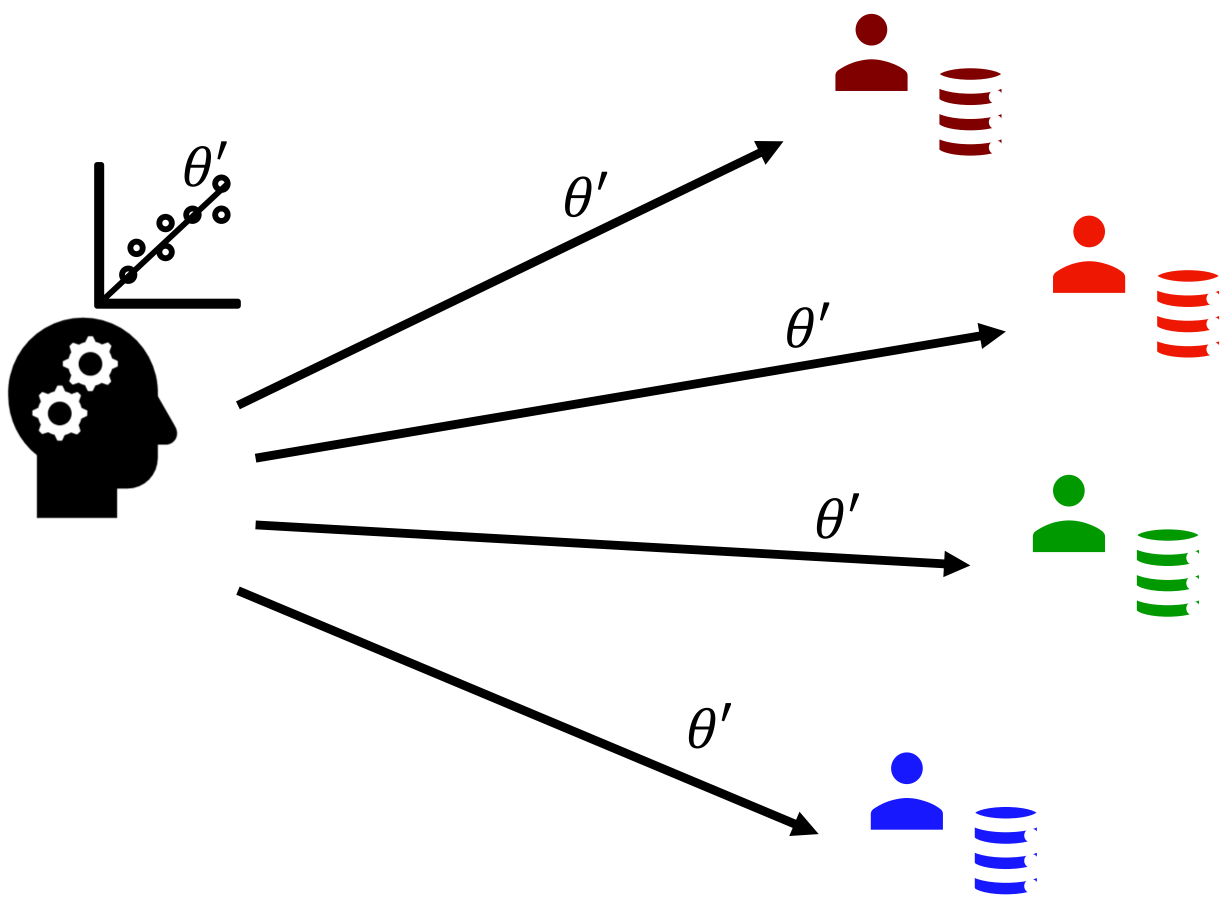
Goal: Learn using everyone's data without sharing it!

Federated Learning (FL)

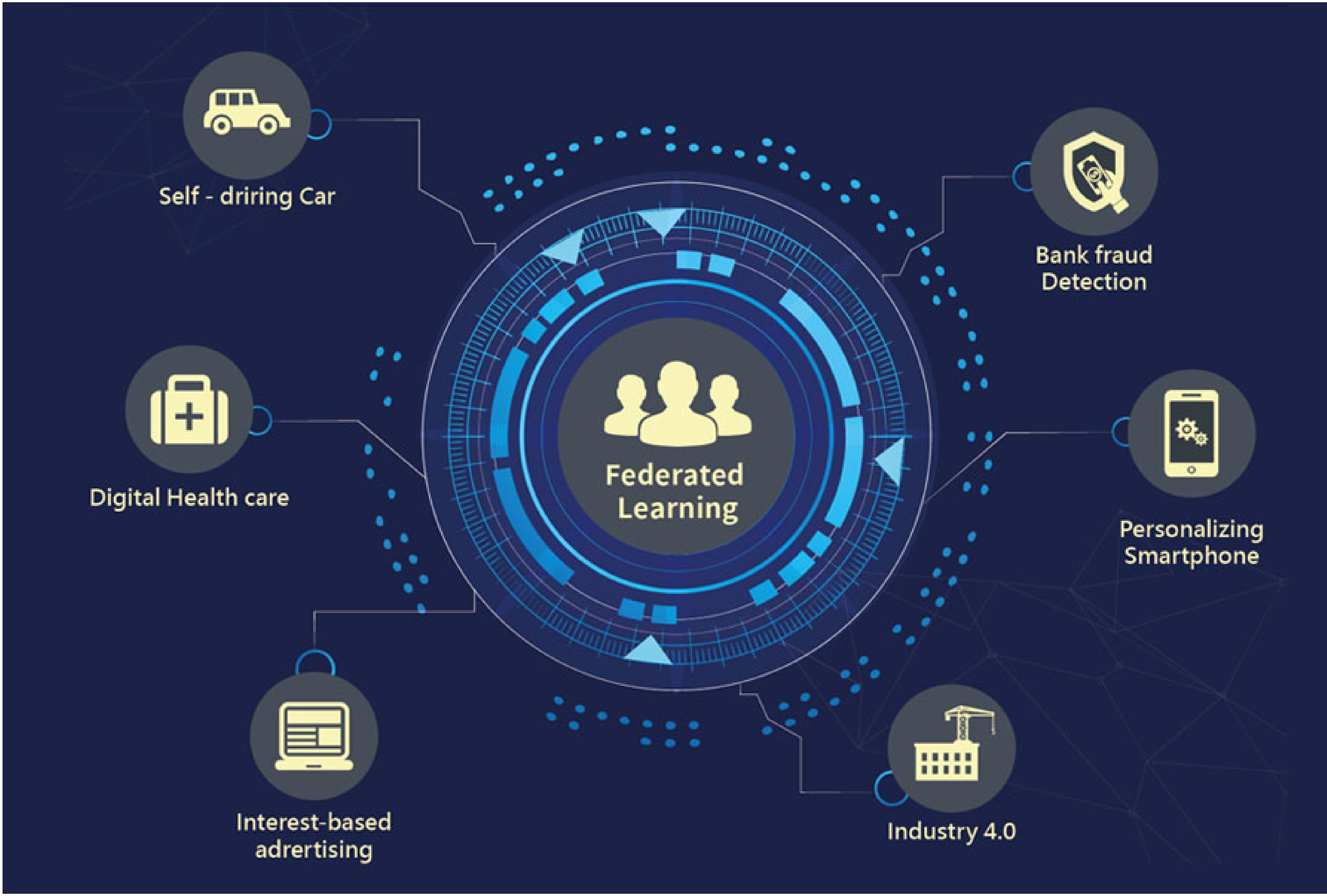


Goal: Learn using everyone's data without sharing it!

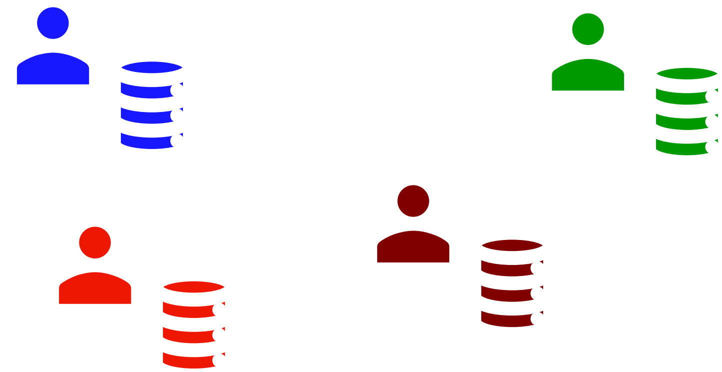
Federated Learning (FL)



Goal: Learn using everyone's data without sharing it!


Self - driring Car
Bank fraud
Detection
Digital Health care
Federated
Learning
Personalizing
Smartphone
Interest-based
adrertising
Industry 4.0

FL Model



- N: set of n clients
- P: set of models/classifiers
- Each agent $i \in A$ has
  - Dataset: D_i
  - Loss/error function: $loss_i: P \to R_+$ (On training data)

Goal: Learn $\theta \in P$ using the data from $D_i s$ (indirectly)

FL History

Introduced by Google Deep Mind in 2016

- FedSGD:

$$\theta' = \theta - \eta \sum_i w_i \cdot \boldsymbol{\partial_\theta loss_i}(.)$$

- FedAvg: Client updates the model and sends.

- **Extensive work:** Distributed, Privacy issues, Welfare, …

Need not be fair/private/strategy-proof

Fair FL

[Donahue-Kleinberg'21]

Egalitarian Fairness:

$$\min_{\theta} \max_{i} \; loss_i(\theta)$$

Proportional/Equity-based Fairness:

$$n_i \; loss_i \; \approx n_j \; loss_j$$

(OR Equalize TPR/loss/…)

What if ∃ noisy/adversarial agent with a lot of bad data?

[Du-Xu-Wu-Tong'21, Mohri-Sivesh-Suresh'19, Papadaki-Martine-Bertran-Shapiro'21, Xu-Lyu'20, Zafar-Valera-Gomez-Rodriguez-Gummadi'17, Zeng-Cheng-Lee'21, …]

Fair FL

Egalitarian Fairness:

$$\min_{\theta} \max_{i} \; loss_i(\theta)$$

Proportional/Equity-based Fairness:

$$n_i \, loss_i \; \approx n_j \, loss_j$$

(OR Equalize TPR/loss/…)

Forced to optimize for the "bad" agent!

May end up harming others.

What if ∃ noisy/adversarial agent with a lot of bad data?

Fair FL → Public Decision Making

- N: set of n **clients ≡ agents**
- P: set of **models/classifiers ≡ outcomes**
- Each agent $i \in A$ has
  - Dataset: D_i
  - **Loss/error function** $\boldsymbol{loss_i}$ **≡ Utility function** $\boldsymbol{U_i = H - loss_i}$

$$\boldsymbol{U_i : P \to R_+}$$

Goal: Find $\theta \in P$ that is "liked" by all

CORE in Public Decisions

[Fain-Goel-Munagala'16, Fain-Munagala-Shah'18]

$$\theta^* \in P$$

- $S \subseteq N$ is a **Blocking Coalition** if $\exists \theta \in P$ s.t.

$$\forall i \in S, U_i(\theta) \geq U_i(\theta^*)$$

(with at least one strict inequality)

Distributed Data $\Rightarrow$ Data of agents in S is $\frac{|S|}{|N|}$-representative of the test data, and hence can only ensure $\frac{|S|}{|N|}$ fraction of utility

CORE in Public Decisions

[Fain-Goel-Munagala'16, Fain-Munagala-Shah'18]

$$\theta^* \in P$$

- $S \subseteq N$ is a **Blocking Coalition** if $\exists \theta \in P$ s.t.

$$\forall i \in S, \frac{|S|}{|N|} U_i(\theta) \geq U_i(\theta^*)$$

(with atleast one strict inequality)

- **Outcome θ^* is in CORE if there is no blocking coalition.**

CORE in FL: Fair, Efficient, *Robust*

- $S \subseteq N$ is a **Blocking Coalition** if $\exists \theta \in P$ s.t.

  $\forall i \in S, \frac{|S|}{|N|} U_i(\theta) \geq U_i(\theta^*)$ with atleast one strict inequality
- $\boldsymbol{\theta^*}$ **is in CORE if there is no blocking coalition.**

- **Pareto-Optimal (PO):** $(S = N)$

  $\nexists \theta \in P$: $\forall i \in N, U_i(\theta) \geq U_i(\theta^*)$ with atleast one inequality.

- **Pareto-Optimal (PO):** $(|S| = 1)$

  $$\forall i \in N, \qquad U_i(\theta^*) \geq \frac{1}{n} \max_{\theta} U\_i(\theta)$$

- ***Robust*** **(to a few noisy/adversarial agents):**

  $S =$ remaining good agents. S is non-blocking (happy)!

CORE in FL: Existence

[Chaudhury, Li, Kang, Li, **M** (NeurIPS'22)]

$$\phi(\theta) = \text{argmax}_{c \in P} \sum_i \frac{U_i(c)}{U_i(\theta)}$$

Theorem 1. CORE exists if set $\phi(\theta)$ is a convex set $\forall \theta$.

Proof sketch.

1. Fixed points of ϕ are in CORE.
2. Apply Kakutani's fixed point to ϕ.

Covers: Concave U_i's $\equiv$ Convex $loss_i$'s

(Linear reg., Logistic reg., …)

CORE in FL: Existence

[Chaudhury, Li, Kang, Li, **M** (NeurIPS'22)]

$$\phi(\theta) = \text{argmax}_c \sum_i U_i(c)/U_i(\theta)$$

Claim. Fixed points of ϕ are in CORE.

Proof sketch. θ^* is FP $\Rightarrow \sum_i \frac{U_i(\theta)}{U_i(\theta^*)} \le \sum_i \frac{U_i(\theta^*)}{U_i(\theta^*)} = n, \forall \theta.$

If $S \subseteq N$ *blocks* θ^*, *then* $\exists \theta \in P$ *s.t.*

$$\forall i \in S, \qquad \frac{|S|}{n} U_i(\theta) \ge U_i(\theta^*) \Rightarrow \frac{U_i(\theta)}{U_i(\theta^*)} \ge \frac{n}{|S|}$$

(at least one strict)

$$\Rightarrow \sum_{i \in S} \frac{U_i(\theta)}{U_i(\theta^*)} > n \text{ !}$$

CORE in FL: Computation

[Chaudhury, Li, Kang, Li, **M** (NeurIPS'22)]

$$\theta^* = \underset{\theta \in P}{\operatorname{argmax}} \mathcal{L}(\theta) = \sum_i \log U_i(\theta)$$

Theorem 2. If U_i's are concave, then θ^* is in the CORE. And can be computed in poly-time.

Proof sketch. (1) $\forall \theta \in P, \sum_i \frac{U_i(\theta)}{U_i(\theta^*)} \leq n$

(2) Then the claim implies θ^* in CORE.

Other settings (participatory budgeting, discrete, …) [Fain-Goel-Munagala'16, Fain-Munagala-Shah'18]

CORE in FL: Distributed Protocol

[Chaudhury, Li, Kang, Li, **M** (NeurIPS'22)]

$$\theta^* = \underset{\theta \in P}{\operatorname{argmax}} \mathcal{L}(\theta) = \sum_i \log U_i(\theta)$$

Theorem 3. CoreFed: Distributed federated learning protocol to find CORE when U_i's are concave.

Proof sketch.

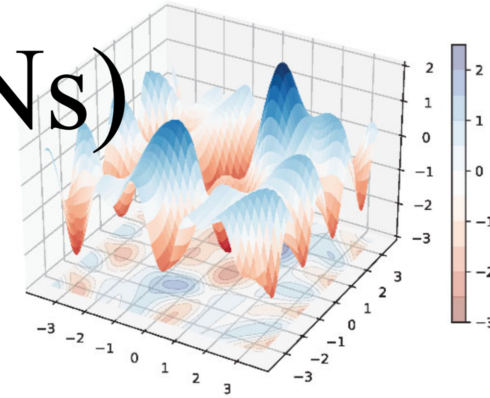
Solicit from agent *i*: $\partial_\theta loss_i(.), loss_i(\theta)$.

Move in the direction of

$$\partial \mathcal{L}(\theta) = \sum_i \frac{\partial_\theta U_i(.)}{U_i(\theta)} = \frac{\sum_i -\partial_\theta loss_i(.)}{H - loss_i(\theta)}$$

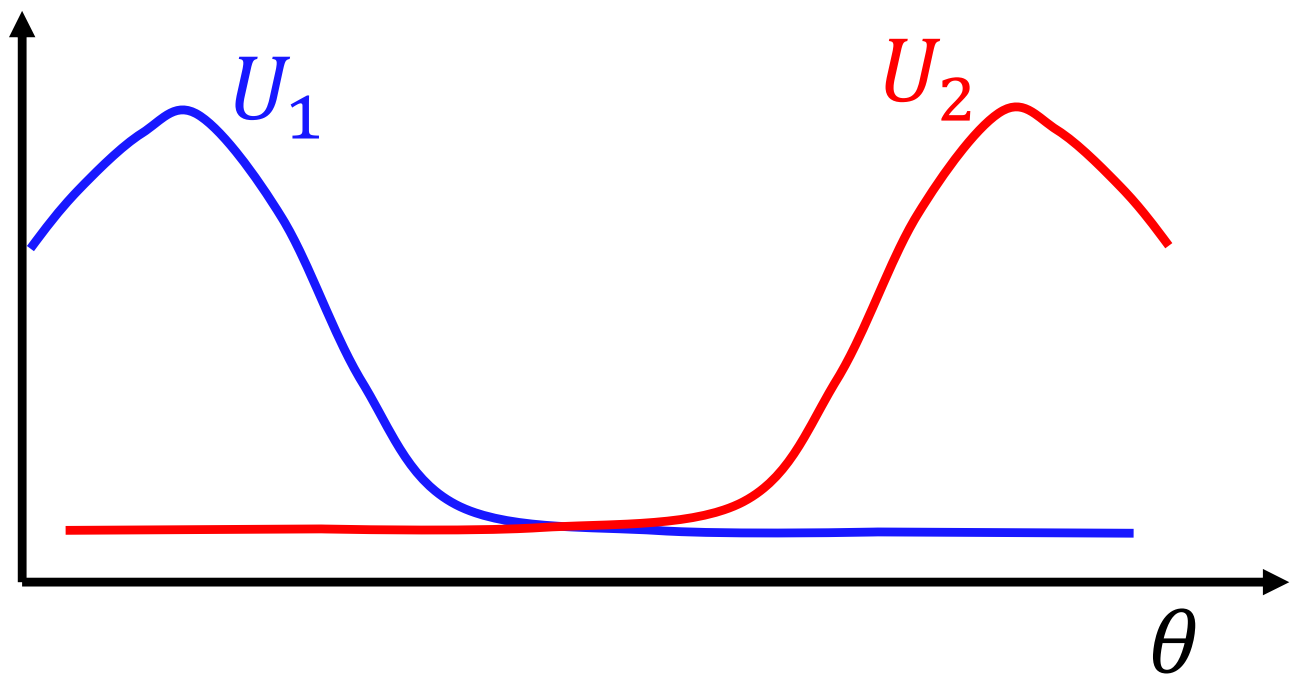
CORE in FL: Non-convex (DNNs)

[Chaudhury, Li, Kang, Li, **M** (NeurIPS'22)]



Local Guarantee: Local-approx. optima of $\mathcal{L}(.)$ is in local-approx. pseudo CORE.

Anything Better?



NO!

DNNs: Experiments

[Chaudhury, Li, Kang, Li, **M** (NeurIPS'22)]

Setup: Two 5x5 convolution layers, 2x2max pooling, and two fully connected layer with ReLU activation.

Table 1: Comparison of utility ($M - \ell_{ce}$) for each agent trained with CoreFed and FedAvg. We see that $\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta^*)} < n$ holds, where θ' denotes the weights of shared model trained by FedAvg and θ^* by CoreFed.

Dataset	Method	Agent 0	Agent 1	Agent 2	U(Average)	U(Multi)	$\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta^*)}$
Adult	FedAvg	2.59	0.77	1.46	1.61	2.91	2.80 (<3)
	CoreFed	2.62	0.90	1.53	1.68	3.61	
MNIST	FedAvg	0.34	0.29	0.92	0.52	0.091	2.66 (<3)
	CoreFed	0.36	0.41	0.91	0.56	0.13	
CIFAR-10	FedAvg	0.63	1.40	0.51	0.84	0.45	2.62 (<3)
	CoreFed	0.73	1.35	0.71	0.93	0.70	

CORE-style solution concept for DNNs?

Proportional Veto-CORE

[Chaudhury, Murhekar, Yuan, Li, **M**, Procaccia (ICML'24)]

(Ask me offline ☺)

Prop. Veto-CORE (Ordinal setting) [Moulin'81]

$$P = \{\theta_1, \dots, \theta_m\}$$

Agent i's pref: $\boldsymbol{\theta_1^i \succ_i \theta_2^i \succ_i \dots \succ_i \theta_m^i}$

$(\equiv U_i\!: \ m \quad m-1 \quad \dots \quad 1 \ )$

θ^* Proportional: $U_i(\theta^*) \geq \dfrac{\max_\theta U_i(\theta)}{n} = \dfrac{m}{n}.$

Agent i blocks θ^* if $U_i(\theta^*) < \dfrac{m}{n}$ $\quad (B = \{\theta | \theta \succ_i \theta^*\})$

Prop. Veto-CORE (Ordinal setting) [Moulin'81]

$$\text{Agent } i\text{'s pref: } \overbrace{\boldsymbol{\theta_1^i \succ_i \theta_2^i \succ_i \dots}}^{\boldsymbol{B}} \boldsymbol{\succ_i \theta^* \succ_i \dots \succ_i \theta_m^i}$$

$$(\equiv U_i:\ m \quad m-1 \quad \dots \qquad\qquad 1\ )$$

$$\theta^* \text{ Proportional: } U_i(\theta^*) \geq \frac{\max_\theta U_i(\theta)}{n} = \frac{m}{n}.$$

$$\text{Agent } i \text{ blocks } \theta^* \text{ if } U_i(\theta^*) < \frac{m}{n} \qquad (B = \{\theta | \theta \succ_i \theta^*\})$$

$$\equiv (\mathrm{m} - |B|) < \frac{\mathrm{m}}{\mathrm{n}} \equiv m\left(1 - \frac{|B|}{|P|}\right) < \frac{m}{n} \equiv \left(1 - \frac{|B|}{|P|}\right) < \frac{1}{n}$$

Prop. Veto-CORE (Ordinal setting) [Moulin'81]

$P = \{\theta_1, \ldots, \theta_m\}$

Agent i's pref: $\boldsymbol{\theta_1^i \succ_i \theta_2^i \succ_i \ldots \succ_i \theta^* \succ_i \ldots \succ_i \theta_m^i}$

Agent i blocks θ^* if $\left(1 - \frac{|B|}{|P|}\right) < \frac{1}{n}$ $\quad (B = \{\theta | \theta \succ_i \theta^*\})$

Set $S \subseteq N$ blocks θ^* if $\left(1 - \frac{|B|}{|P|}\right) < \frac{|\boldsymbol{S}|}{n}$

$(\boldsymbol{B} = \cap_{i \in S} \{\boldsymbol{\theta} | \boldsymbol{\theta} \succ_i \boldsymbol{\theta}^*\})$

Veto-CORE: If no blocking coalition.

Prop Veto-CORE (**Continuous** setting)

[Chaudhury, Murhekar, Yuan, Li, **M**, Procaccia (ICML'24)]

P: Measurable set. λ: Measure function.

Agent i's pref: $U_i: P \rightarrow R_+$ measurable (allows DNNs)

$\theta^* \in P$. Set $S \subseteq N$ blocks θ^* if

$$\left(1 - \frac{\lambda(B)}{\lambda(P)}\right) \leq \frac{|S|}{n} \pm \epsilon$$

$$\left(\begin{matrix}\exists B \subseteq P\text{: } \forall \theta \in B, \forall i \in S, U_i(\theta) \geq U_i(\theta^*) \\ \text{atleast one strict}\end{matrix}\right)$$

ϵ-Prop Veto-CORE: If no blocking coalition.

(Fair ML informs SCT!)

Prop Veto-CORE (PVC): Results

(**Continuous**) [Chaudhury, Murhekar, Yuan, Li, **M**, Procaccia (ICML'24)]

Theorem. If U_i's are Lebesgue-measurable, then ϵ-Prop Veto-CORE exists for any $\epsilon \in \left(0, \frac{1}{n}\right)$.

Proposition. If θ^* is in ϵ-PVC, then θ^* is

1. (approx.) Pareto-optimal
2. (approx.) (rankwise) Proportional

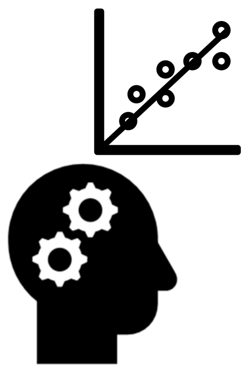
Proposition. Better guarantees for aligned preferences.

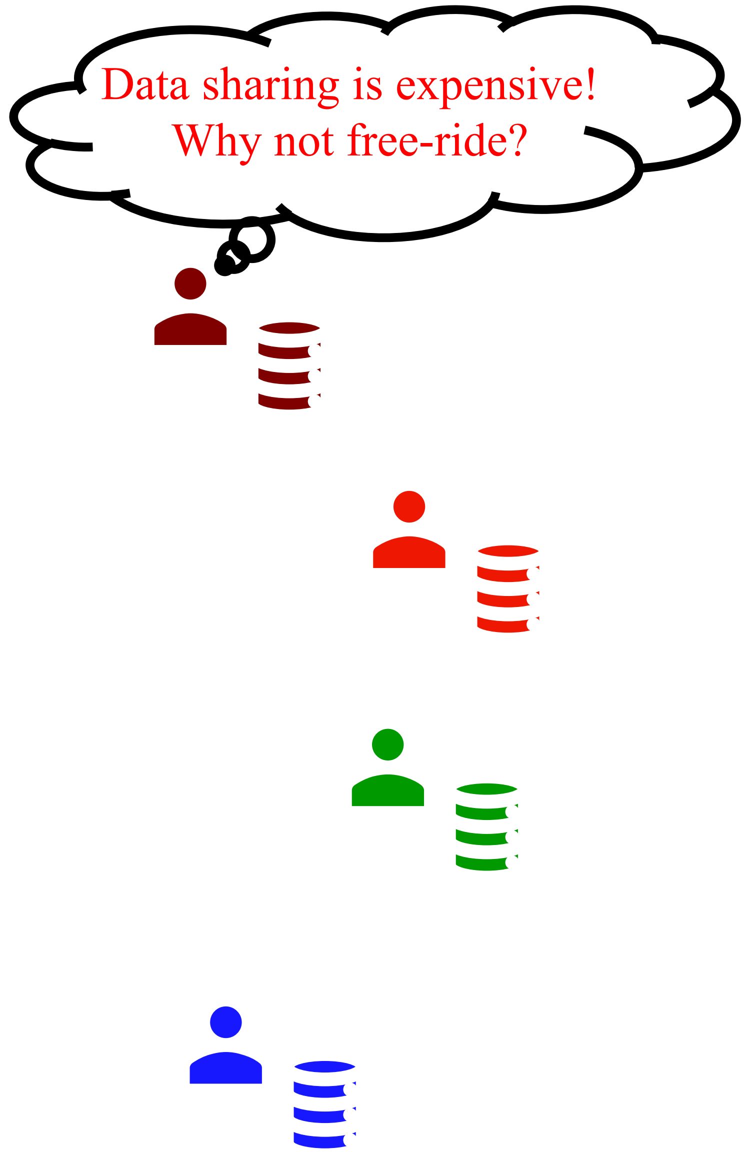
(Veto-)CORE: Questions

- Limited Heterogeneity: Better guarantees?
  - How to formalize heterogeneity parameter?
  - What guarantees are possible with respect to it?
- Strategic Analysis
  - Nash equilibrium, Truthful Mechanisms, …

Data Sharing in FL: Incentives

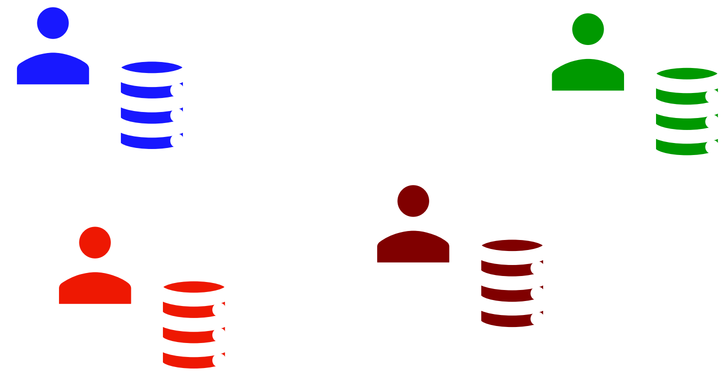
FL: Incentive Issues





Data sharing costs: computation/privacy/storage

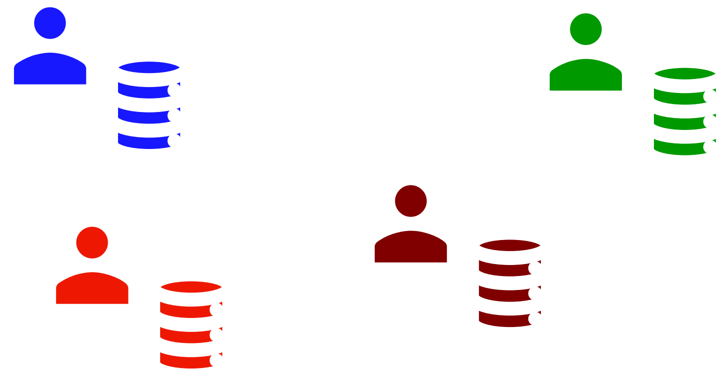
FL: Data Sharing Game



- A: set of n players/agents/clients

- Each agent $i \in A$ has
  - Has dataset D_i
  - Strategy: $d_i \in [0, 1]$ fraction of data shared
  - Accuracy function $a_i: [0, 1]^n \rightarrow R_+$
  - Cost function $c_i: [0, 1] \rightarrow R_+$ (cost of sharing data)

FL: Data Sharing Game



- N: set of n players/agents/clients
- Each agent $i \in A$ has
  - Has dataset D_i
  - Strategy: $d_i \in [0, 1]$ fraction of data shared
  - Accuracy function $a_i: [0, 1]^n \rightarrow R_+$
  - Cost function $c_i: [0, 1] \rightarrow R_+$ (cost of sharing data)
- **Nash Equilibrium (NE):** No unilateral deviation

  For each agent i, $u_i(d_i, d_{-i}) \geq u_i(d_i', d_{-i}),\ \forall d_i' \in [0,1]$

Incentives: Prior Work

- [Blum, Haghtalab, Philips, Shao (ICML'21)]
  Nash Eq. (NE) Analysis
  - Agent's goal: Minimize data shared subject to $a_i(.) \geq \tau_i$
  - NE may not always exist. Sufficiency conditions, structural results.

- [Karimireddy, Guo, Jordan (Workshop@NeurIPS'22)]
  Truthful mechanism to maximize data-sharing
  - Agent's goal: Maximize net payoff $\mathrm{u_i} = a_i(d_1, \dots, d_n) - c_i(d_i)$
  - $c_i(d_i) = C_i * d_i$, and a_i's are identical and concave
  - Grim-trigger style strategy

Welfare-maximizing? Fair? Budget-balanced?

Incentives in FL: Results

Agent's goal: Maximize net payoff u_i = (Accuracy – Cost)

a_i's concave, c_i's convex

- [Murhekar, Yuan, Chaudhury, Li, **M** (NeurIPS'23)]
  - NE exists and can be reached via Best-Response-Dynamics.
  - NE may have bad welfare (due to free-riding)
  - Budget-balanced mechanism to maximize any p-mean welfare.
- [Murhekar, Song, Shahkar, Chaudhury, **M** (ICML'25)]
  - **Reciprocally fair mechanism, with payments p_i to agent i.**
  - Budget-balanced
  - Data + Accuracy gain

Reciprocal Fairness:
(Karma!) You get what you give

Reciprocity: You get what you give

Agent's goal: $u_i(\boldsymbol{d}) = a_i(\boldsymbol{d}) - c_i(d_i) + p_i$, where $\boldsymbol{d} = (d_1, \dots, d_n)$, p_i is payment.

$\phi_i^A(\boldsymbol{d}) =$ Contribution of agent $\boldsymbol{i}$ to the welfare of other agents.

- Shapley Value:

$$\phi_i^A(\boldsymbol{d}) = \sum_{S \subset A} \binom{n}{|S|}^{-1} \left(A(\boldsymbol{d}[S \cup \{i\}]) - A(\boldsymbol{d}[S])\right)$$

$\boldsymbol{d}[S] = ((d_i)_{i \in S}, 0, \dots, 0)$ and $A(\boldsymbol{d}) = \sum_{i \in N} a_i(\boldsymbol{d})$

Reciprocity of a Mechanism:

You get what you give

Agent's goal: $u_i(\boldsymbol{d}) = a_i(\boldsymbol{d}) - c_i(d_i) + p_i$, where $\boldsymbol{d} = (d_1, \ldots, d_n)$, p_i is payment.

$\phi_i^A(\boldsymbol{d}) =$ Contribution of agent i to the welfare of other agents.

M: Payment Mechanism, $NE(M)$: NE set of M

$$\text{Reciprocity}(M) = \min_{\boldsymbol{d} \in NE(M)} \min_{i \in A} \frac{a_i(\boldsymbol{d}) + p_i}{\phi_i^A(\boldsymbol{d})}$$

Claim. $\text{Reciprocity}(M) \leq 1$

Reciprociprocal Mechanism: M^{shap}

Shapley Value: $\phi_i^A(\boldsymbol{d}) = \sum_{S \subset A} \binom{n}{|S|}^{-1} (A(\boldsymbol{d}[S \cup \{i\}]) - A(\boldsymbol{d}[S])$

- M^{shap}

$$p_i(\boldsymbol{d}) = \phi_i^A(\boldsymbol{d}) - a_i(\boldsymbol{d})$$

Theorem(s). a_i concave, c_i convex non-decreasing, $\forall i$

- M^{shap} admits a NE, and Best Response converges quickly.
- Reciprocity(M^{shap}) = 1
- High Data-gain and Accuracy gain.

Incentives in FL: Results

Agent's goal: Maximize net payoff (Net utility – Cost)

U_i's concave, c_i's convex

- [Murhekar, Yuan, Chaudhury, Li, M (NeurIPS'23)]
  - NE exists and can be reached via Best-Response-Dynamics.
  - NE may have bad welfare (due to free-riding).
  - Budget-balanced mechanism to maximize any p-mean welfare.

Learning models covered: Linear/random discovery, random coverage, PAC learning, cross-entropy loss, …

- [Murhekar, Chaudhury, M'24]
  - Reciprocally fair mechanism, with payments p_i to agent i.

    Net utility ($a_i(.) + p_i$) of an agent is exactly equals her contribution to the collaboration aka her Shapley share
  - Budget-balanced

Open Directions

Incentives in FL: Data Sharing Game

- FL (distributed) protocols
- Non-IID data / Non-monotone accuracy
- Truthful Mechanisms
  - Without payment: fair / welfare-maximizing
  - With payments: budget-balanced / fair/ welfare-maximizing
- (Data) Contracts

General Direction:

Fair/Trustworthy ML via GT+SCT

Mintong Kang

Aniket Murhekar

Zhuowen Yuan

Jiaxin Song

Bhaskar R. Chaudhury

Bo Li

Linyi Li

Ariel Procaccia

THANK YOU