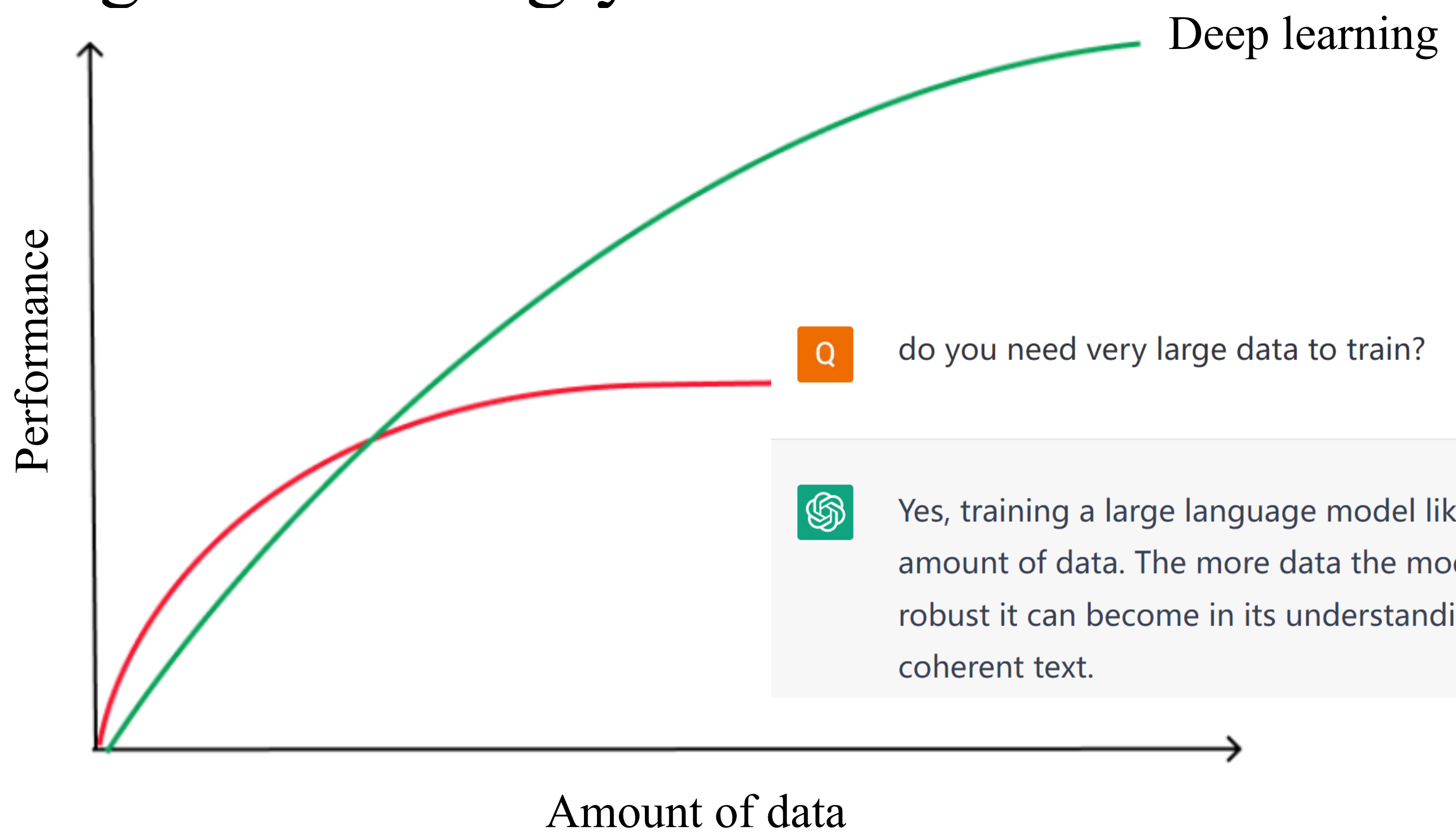
Practical Federated Learning on non-IID Data: Algorithms and Systems

Qinbin Li
UC Berkeley

Joint work with Yanzheng Cai, Yiqun Diao, Sixu Hu, Xiaoyuan Liu, Naibo Wang, Zhaomin Wu, Chulin Xie…
Quan Chen, Bingsheng He, Bo Li, Dawn Song, Zeyi Wen, …

Data is All You Need

- Machine learning is data-hungry.

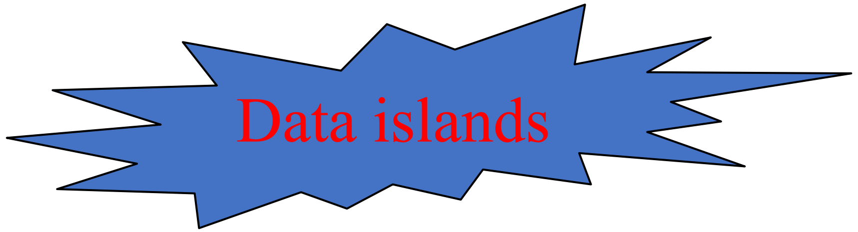


Q do you need very large data to train?

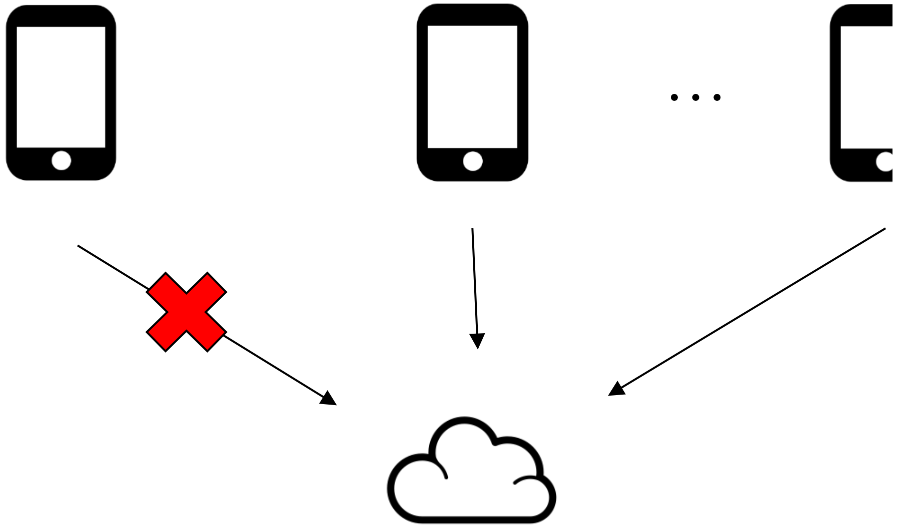
Yes, training a large language model like myself typically requires a very large amount of data. The more data the model is trained on, the more accurate and robust it can become in its understanding of language and ability to generate coherent text.

But… Where is Data?

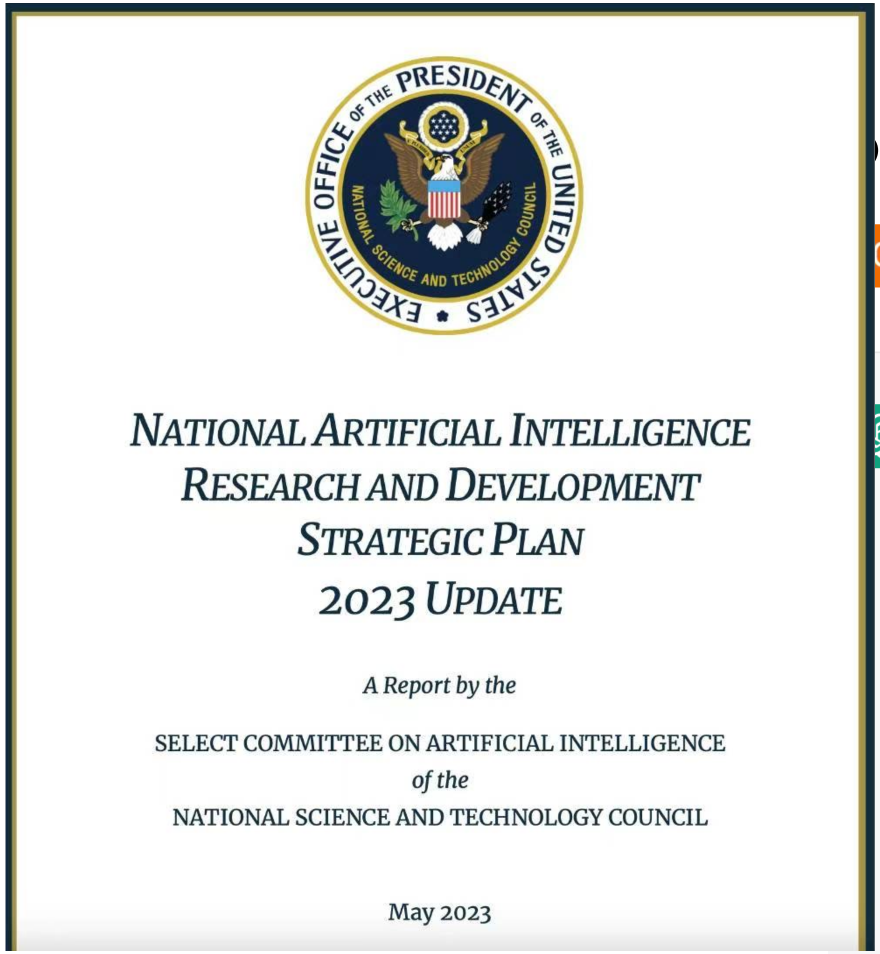
- Widely spread as data silos
    - hospitals



- mobiles



Federated Learning



NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN 2023 UPDATE

A Report by the

SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE
of the
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

May 2023

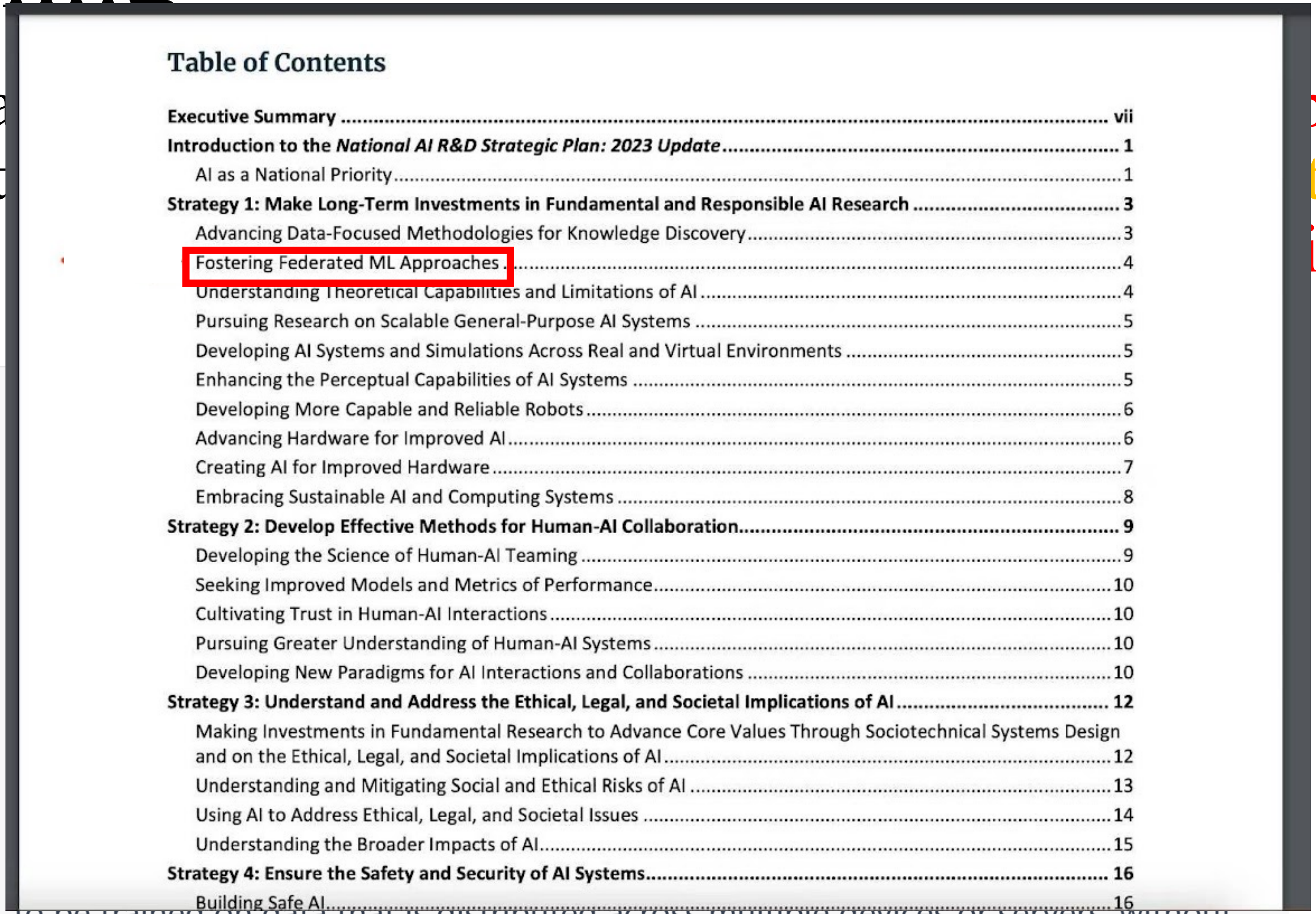


Table of Contents

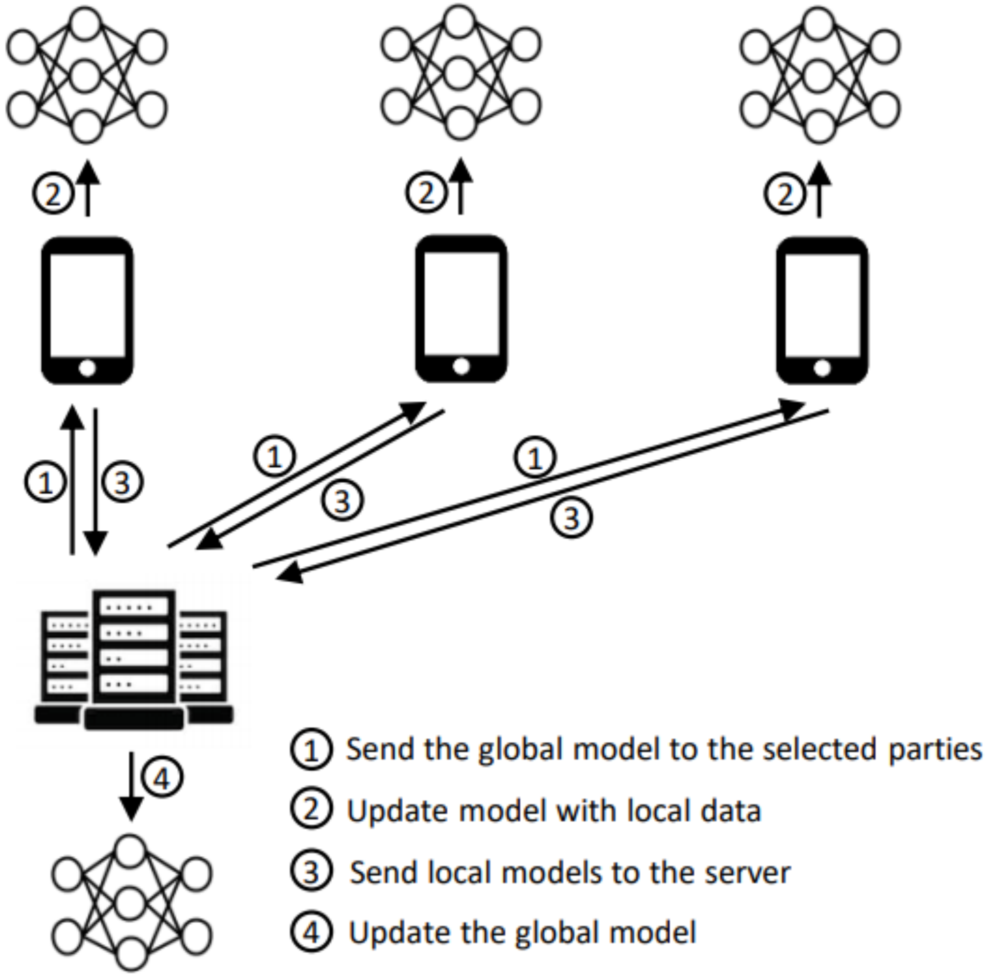
to be trained on data that is distributed across multiple devices or servers, without the data ever leaving the devices or servers themselves.

Year

[1] Kairouz, Peter, et al. "Advances and open problems in federated learning." arXiv preprint arXiv:1912.04977 (2019).

FedAvg[2]

- A de facto federated learning approach.



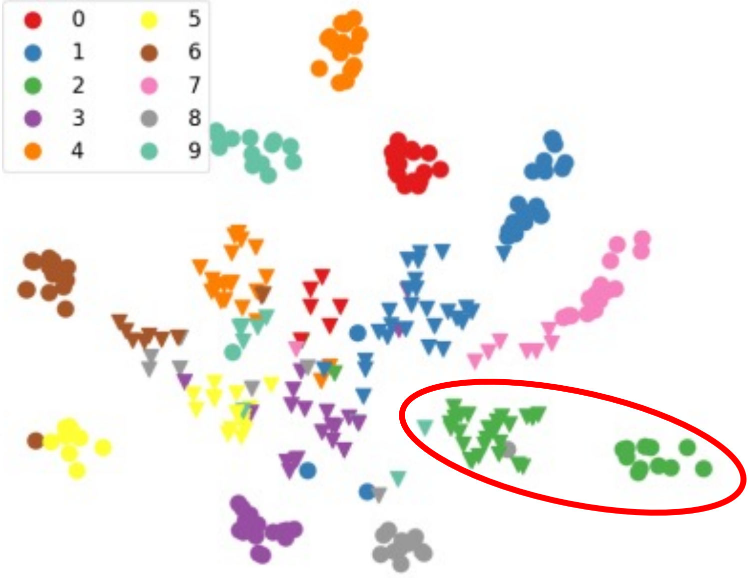
[2] McMahan, H. Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *AISTATS 2017*

Non-IID Data in Real World

Label distribution skew

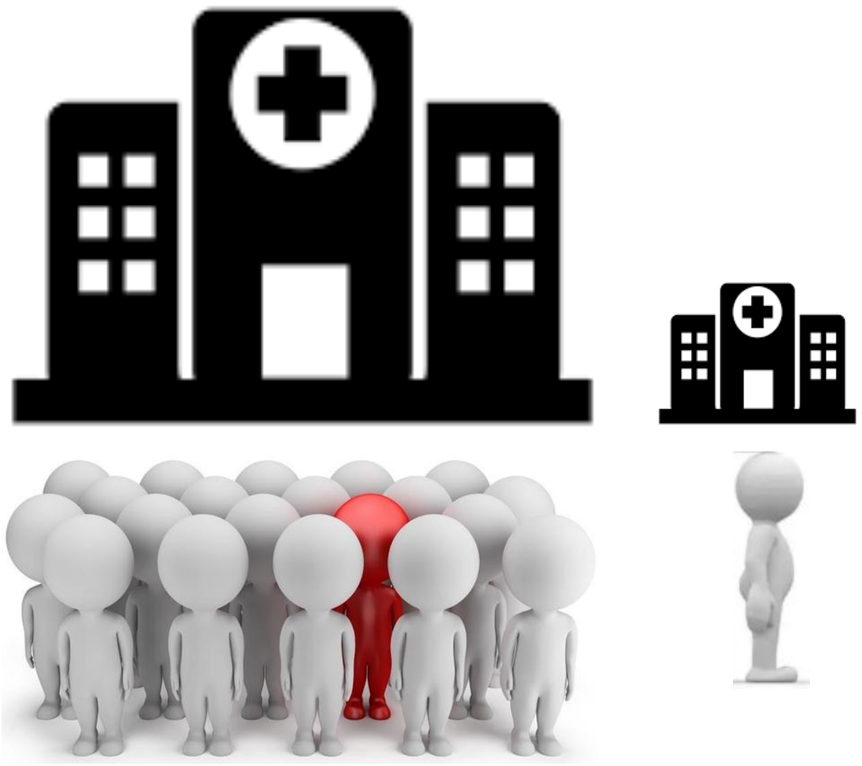
Criteo

Feature distribution skew

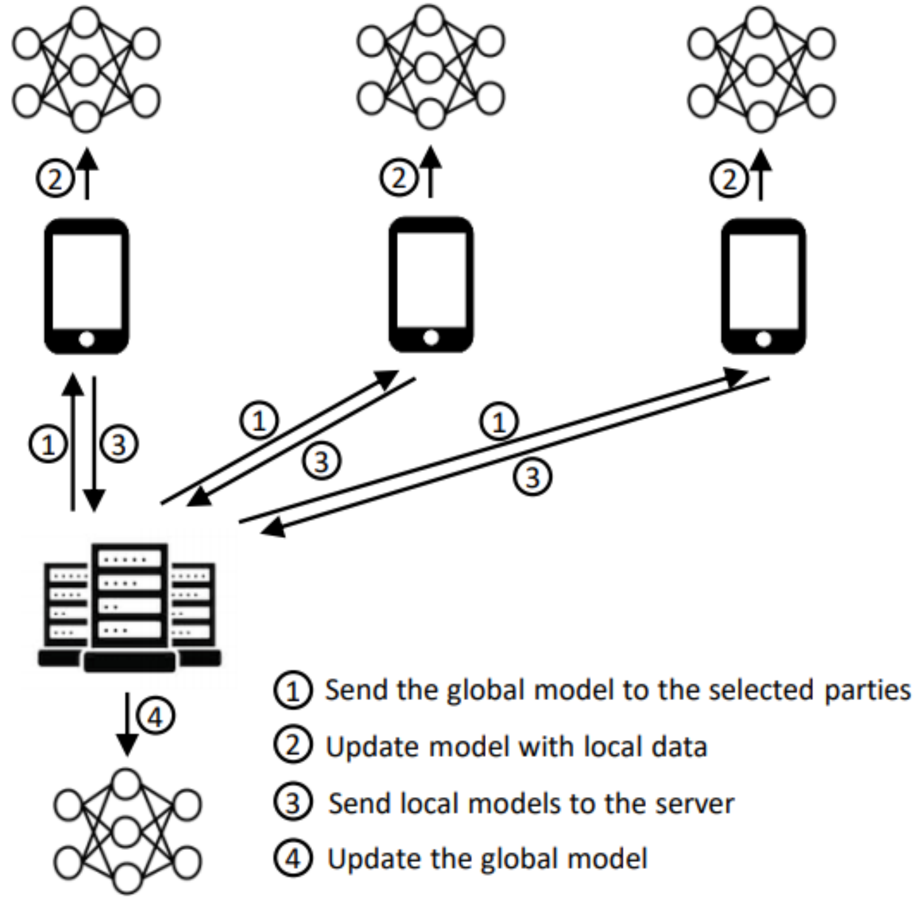


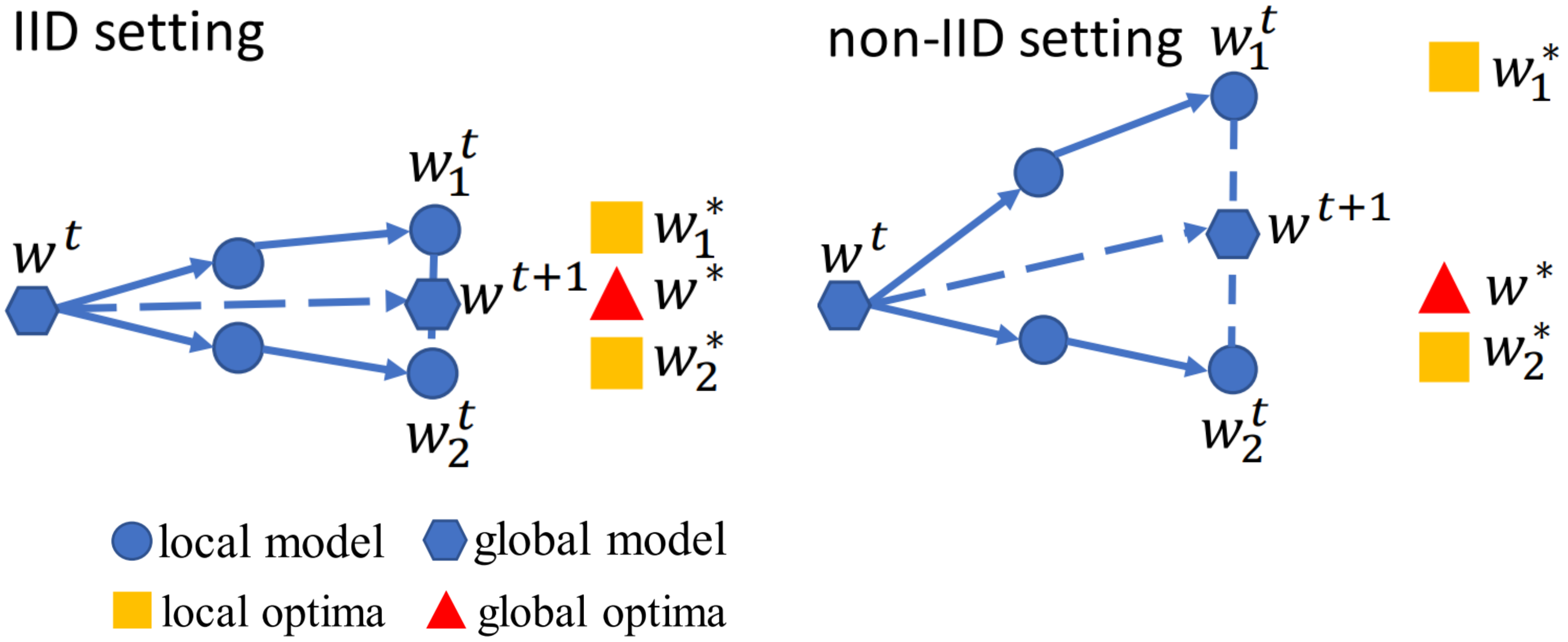
Digits

Quantity skew



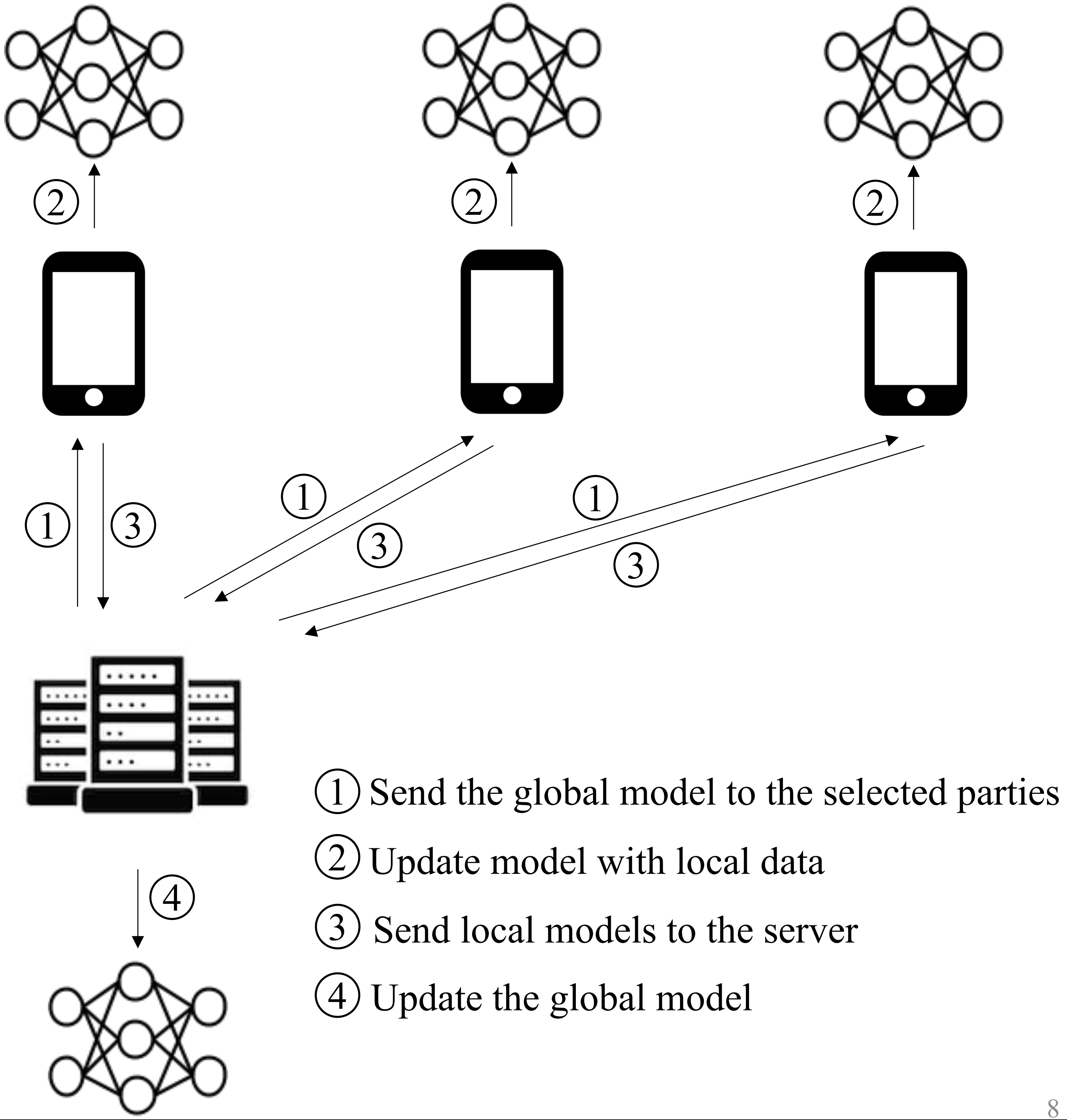
Non-IID Data Challenge in FL



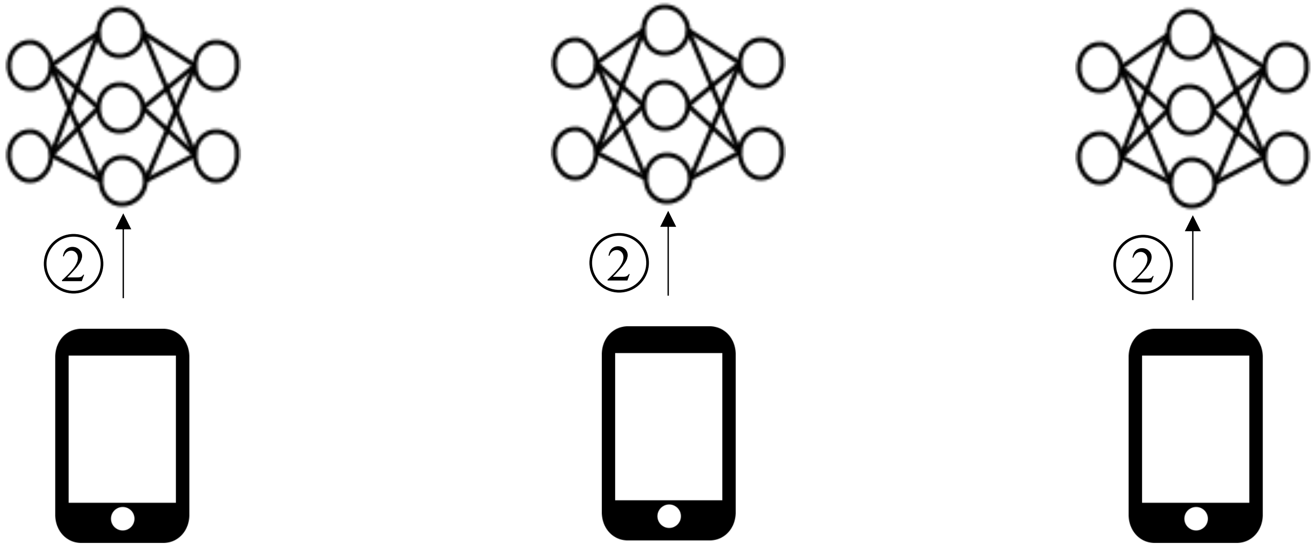


Solving Non-IID

- Based on FedAvg



Solving Non-IID



- Based on FedAvg
- FedProx[2]
  - Add L2 regularization
- SCAFFOLD[3], FedNova[4]…

For FedProx:

$$L(w;\mathbf{b}) = \sum_{(x,y)\in\mathbf{b}} \ell(w;x;y) + \frac{\mu}{2}\left\|w - w^t\right\|^2$$

1 Send the global model to the selected parties

$\boxed{\text{2 Update model with local data}}$

3 Send local models to the server

4 Update the global model

[2] Li, Tian, et al. "Federated optimization in heterogeneous networks." Proceedings of Machine Learning and Systems 2 (2020): 429-450.
[3] Karimireddy, Sai Praneeth, et al. "Scaffold: Stochastic controlled averaging for federated learning." International Conference on Machine Learning. PMLR, 2020.
[4] Wang, Jianyu, et al. "Tackling the objective inconsistency problem in heterogeneous federated optimization." Advances in neural information processing systems 33 (2020): 7611-7623.

Results

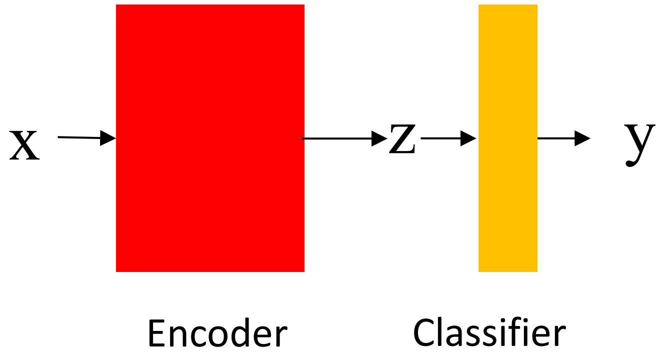
Dataset	FedAvg	FedProx	SCAFFOLD	FedNova
FMNIST	$88.1\% \pm 0.6\%$	$88.1\% \pm 0.9\%$	$88.4\% \pm 0.5\%$	$\mathbf{88.5\%} \pm \mathbf{0.5\%}$
CIFAR-10	$68.2\% \pm 0.7\%$	$67.9\% \pm 0.7\%$	$\mathbf{69.8\%} \pm \mathbf{0.7\%}$	$66.8\% \pm 1.5\%$
SVHN	$86.1\% \pm 0.7\%$	$86.6\% \pm 0.9\%$	$\mathbf{86.8\%} \pm \mathbf{0.3\%}$	$86.4\% \pm 0.6\%$

Limited improvement!

Federated Deep Learning on Non-IID Data

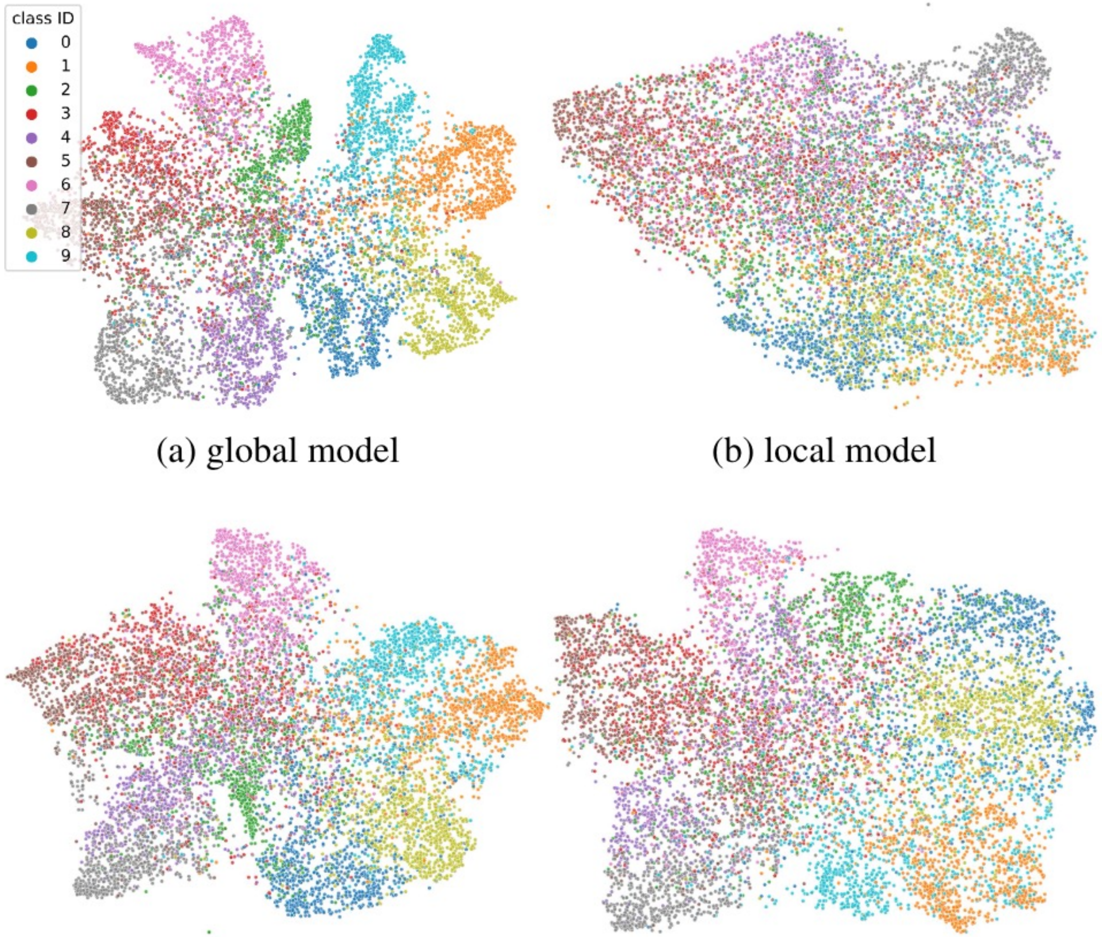
- FedProx
  - Regularization based on model-parameters

- Deep Learning -> Representation Learning
  - Regularization based on representation?



Observation

- The model trained on a global dataset is able to extract a better feature representation than the model trained in a skewed subset.

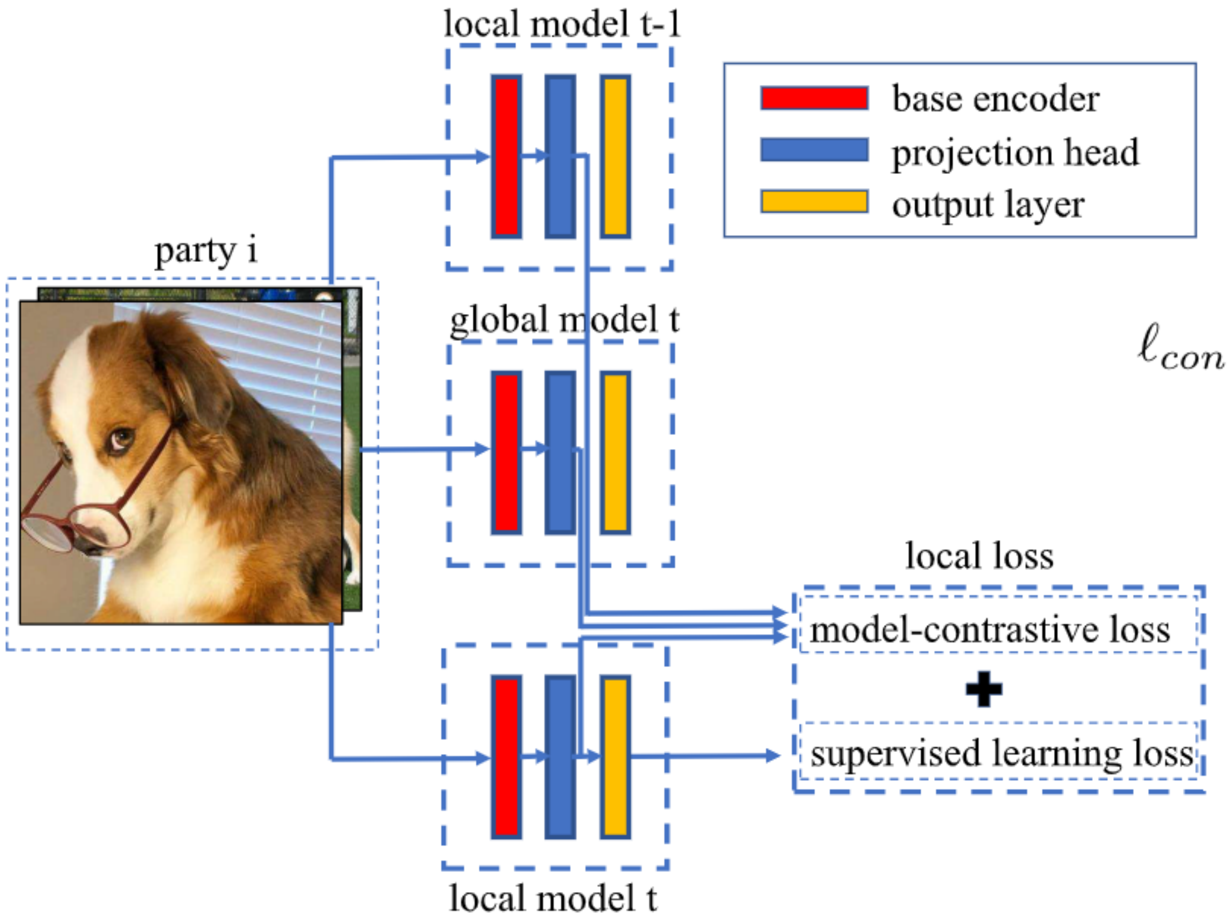


(a) global model (b) local model

(c) FedAvg global model (d) FedAvg local model

Model-Contrastive Learning (MOON)[4]

- Idea: maximize the agreement of representation learned by the current local model and the representation learned by the global model.



$$\ell_{con} = -\log \frac{\exp(\mathrm{sim}(z, z_{glob})/\tau)}{\exp(\mathrm{sim}(z, z_{glob})/\tau) + \exp(\mathrm{sim}(z, z_{prev})/\tau)}$$

[4] Li, Qinbin, Bingsheng He, and Dawn Song. "Model-Contrastive Federated Learning." CVPR 2021.

MOON

- Lightweight modifications to FedAvg --- Simple and Effective.

Algorithm 1: The MOON framework

Input: number of communication rounds T, number of parties N, number of local epochs E, temperature τ, learning rate η, hyper-parameter μ

Output: The final model w^T

Same as FedAvg

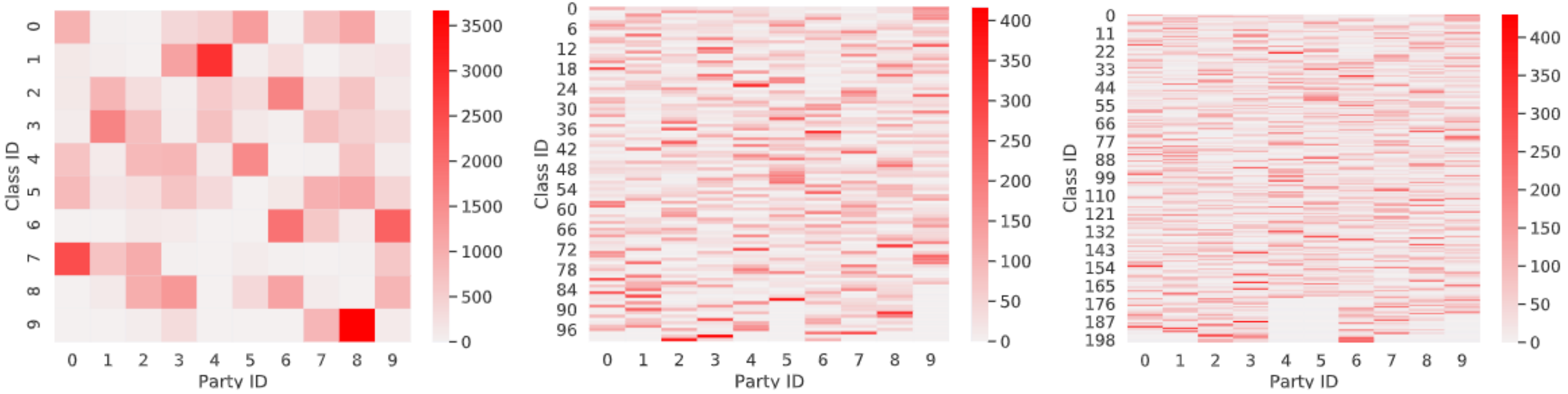
Server executes:
initialize w^0
for $t = 0, 1, ..., T-1$ **do**
  for $i = 1, 2, ..., N$ **in parallel do**
    send the global model w^t to P_i
    $w_i^t \leftarrow$ **PartyLocalTraining**(i, w^t)
  $w^{t+1} \leftarrow \sum_{k=1}^{N} \frac{|\mathcal{D}^i|}{|\mathcal{D}|} w_k^t$
return w^T

PartyLocalTraining(i, w^t):
$w_i^t \leftarrow w^t$
for epoch $i = 1, 2, ..., E$ **do**
  for each batch $\mathbf{b} = \{x, y\}$ of $\mathcal{D}^i$ **do**
    $\ell_{sup} \leftarrow CrossEntropyLoss(F_{w_i^t}(x), y)$
    $z \leftarrow R_{w_i^t}(x)$
    $z_{glob} \leftarrow R_{w^t}(x)$
    $z_{prev} \leftarrow R_{w_i^{t-1}}(x)$
    $\ell_{con} \leftarrow$
    $-\log \frac{\exp(\mathrm{sim}(z, z_{glob})/\tau)}{\exp(\mathrm{sim}(z, z_{glob})/\tau) + \exp(\mathrm{sim}(z, z_{prev})/\tau)}$
    $\ell \leftarrow \ell_{sup} + \mu \ell_{con}$
    $w_i^t \leftarrow w_i^t - \eta \nabla \ell$
return w_i^t to server

Contrastive loss

Experiments

- Baselines: SOLO, FedAvg, FedProx[5], SCAFFOLD[6]
- Datasets: CIFAR-10, CIFAR-100, Tiny-ImageNet
- Partition: Dirichlet distribution
- To be practical: 1) Effectiveness 2) Efficiency 3) Robustness



(a) CIFAR-10 (b) CIFAR-100 (c) Tiny-Imagenet

[5] Li, Tian, et al. “Federated optimization in heterogeneous networks.” MLSys 2020.

[6] Karimireddy, Sai Praneeth, et al. "SCAFFOLD: Stochastic controlled averaging for federated learning." ICML 2020.

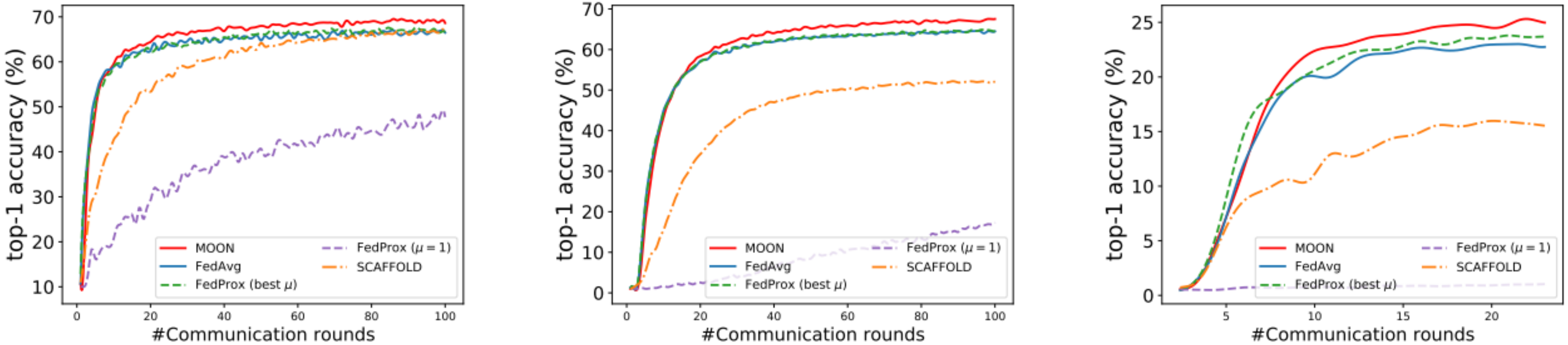
Accuracy

- 10 parties

Method	CIFAR-10	CIFAR-100	Tiny-Imagenet
MOON	**69.1%**±0.4%	**67.5%**±0.4%	**25.1%**±0.1%
FedAvg	66.3%±0.5%	64.5% ±0.4%	23.0%±0.1%
FedProx	66.9%±0.2%	64.6%±0.2%	23.2%±0.2%
SCAFFOLD	66.6%±0.2%	52.5% ±0.3%	16.0%±0.2%
SOLO	46.3% ±5.1%	22.3%±1.0%	8.6%±0.4%

About 2-3% accuracy improvement.

Communication Efficiency



(a) CIFAR-10 (b) CIFAR-100 (c) Tiny-Imagenet

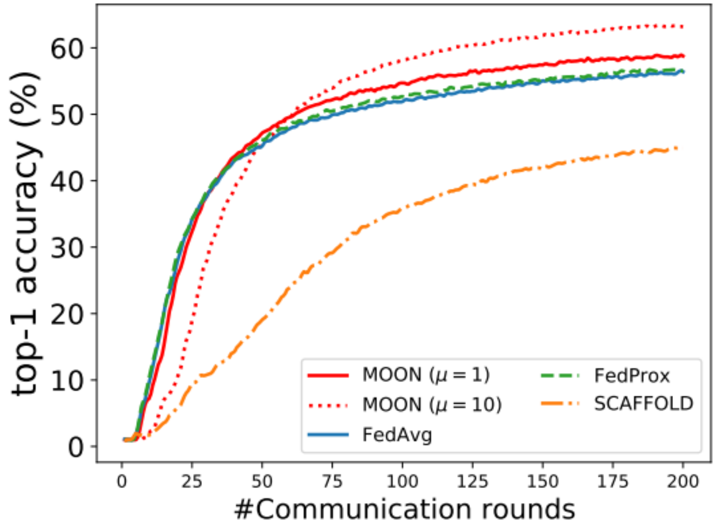
- Number of rounds to achieve the same target performance.

Method	CIFAR-10		CIFAR-100		Tiny-Imagenet	
	#rounds	speedup	#rounds	speedup	#rounds	speedup
FedAvg	100	1×	100	1×	20	1×
FedProx	52	1.9×	75	1.3×	17	1.2×
SCAFFOLD	80	1.3×	\	<1×	\	<1×
MOON	**27**	**3.7**×	**43**	**2.3**×	**11**	**1.8**×

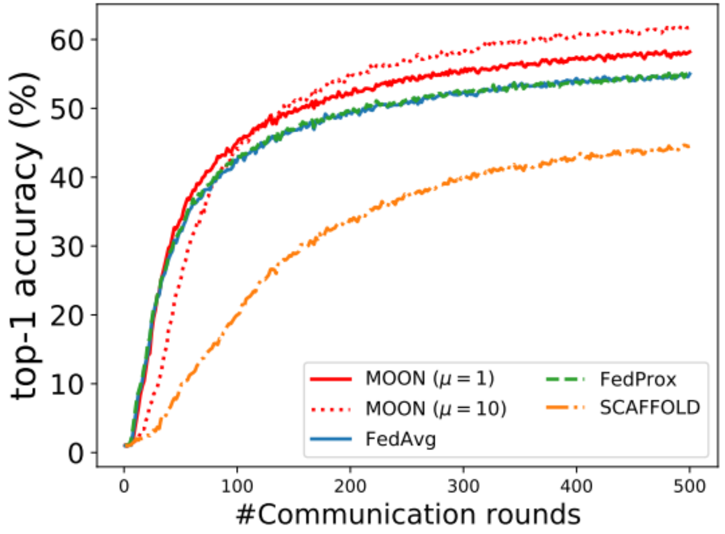
MOON is much more communication-efficient

Scalability

Method	#parties=50		#parties=100	
	100 rounds	200 rounds	250 rounds	500 rounds
MOON (μ=1)	54.7%	58.8%	54.5%	58.2%
MOON (μ=10)	**58.2%**	**63.2%**	**56.9%**	**61.8%**
FedAvg	51.9%	56.4%	51.0%	55.0%
FedProx	52.7%	56.6%	51.3%	54.6%
SCAFFOLD	35.8%	44.9%	37.4%	44.5%
SOLO	10%±0.9%		7.3%±0.6%	



(a) 50 parties

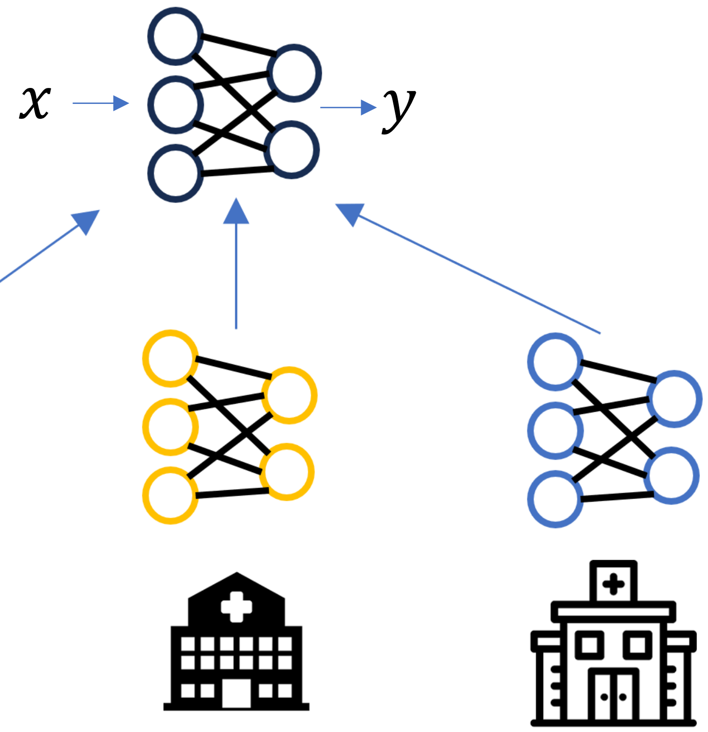


(b) 100 parties (sample fraction=0.2)

About 6-7% accuracy improvement.

Model Averaging

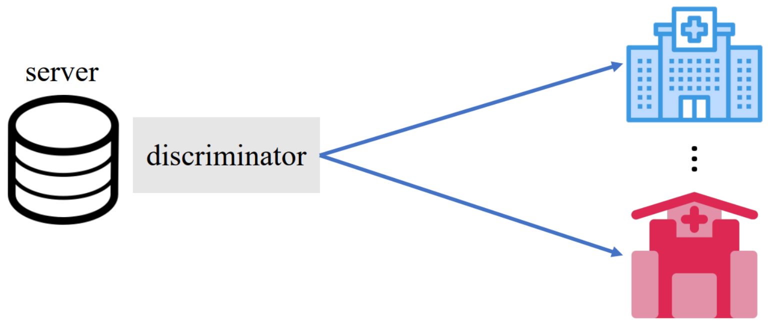
- Learning a common $P(y|x)$



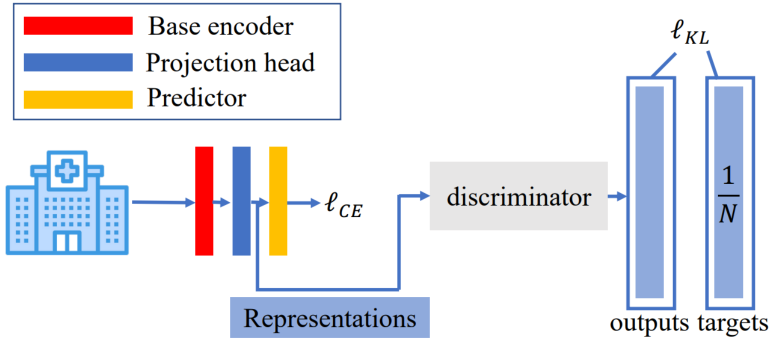
- Non-IID Features
  - $P_i(x) \neq P_j(x), P_i(y|x) \neq P_j(y|x)$

Adversarial Collaborative Learning [7]

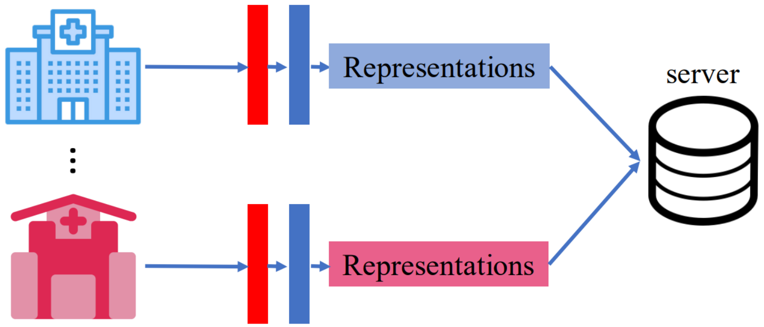
- Learning a common task-specific representation distribution
  - $P_i(z) = P_j(z)$



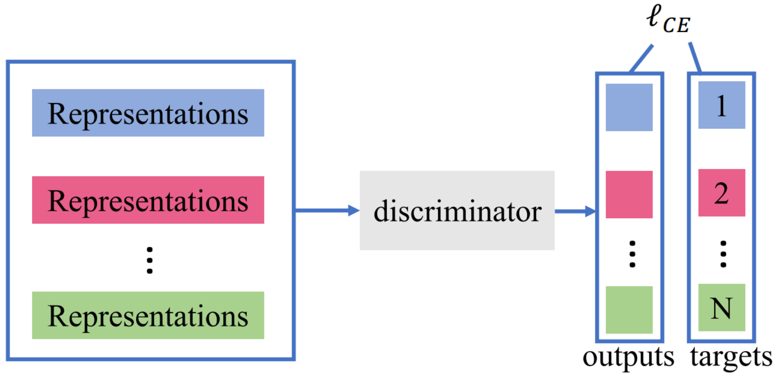
(a) Step 1: The server sends the discriminator to the parties.



(b) Step 2: The parties update their local models



(c) Step 3: The parties send representations to the server.



(d) Step 4: The server updates the discriminator.

[7] Li, Qinbin, Bingsheng He, and Dawn Song. "Adversarial Collaborative Learning on Non-IID Features." ICML 2023.

Theoretical Analysis

- Convergence

Theorem 1. We use P_{G_i} to denote the distribution of the representations generated in party i and $P_{G_i}(\mathbf{z})$ is the probability of representation $\mathbf{z}$ in distribution P_{G_i}. Then, the optimal discriminator D^* of Equation (4) is

$$D_k^*(\mathbf{z}) = \frac{P_{G_k}(\mathbf{z})}{\sum_{i=1}^{N} P_{G_i}(\mathbf{z})}. \tag{5}$$

Theorem 2. Given the optimal discriminator D^* from Equation (5), the global minimum of Equation (3) is achieved if and only if

$$P_{G_1} = P_{G_2} = \cdots = P_{G_N} \tag{6}$$

Theorem 3. Suppose P_G^* is the optimal solution shown in Theorem 2. If G_i $(\forall i \in [1, N])$ and D have enough capacity, and P_{G_i} is updated to minimize the local objective (i.e., Equation (3)), given the optimal discriminator D^* from Equation (5), then P_{G_i} converges to P_G^*.

- Generalization error

Accuracy

Digits	MNIST	SVHN	USPS	SynthDigit	MNIST_M	AVG
SOLO	87.9%±0.4%	34.8%±0.8%	94.8%±0.1%	63.0%±0.4%	67.2%±0.4%	69.5%±0.3%
FedAvg	94.4%±0.5%	59.4%±0.9%	94.3%±0.2%	74.4%±0.5%	70.3%±1.2%	78.6%±0.6%
FedBN	94.1%±0.8%	**59.9**%±0.7%	94.1%±0.1%	73.9%±0.6%	71.3%±1.1%	78.7%±0.6%
PartialFed	**94.7**%±0.4%	59.4%±0.6%	94.2%±0.1%	75.2%±0.4%	69.7%±0.6%	78.6%±0.4%
FedProx	94.1%±0.4%	59.8%±0.6%	94.3%±0.1%	73.4%±0.3%	71.6%±0.9%	78.6%±0.4%
Per-FedAvg	88.9%±0.7%	36.6%±1.3%	89.5%±0.2%	58.3%±0.7%	54.5%±1.3%	65.6%±0.8%
FedRep	92.6%±0.2%	42.0%±1.0%	93.1%±0.1%	61.1%±0.5%	50.8%±1.4%	67.9%±0.8%
ADCOL	**94.7**%±0.6%	58.2%±1.0%	**95.4**%±0.2%	**76.0**%±0.3%	**76.7**%±0.8%	**80.2**%±0.5%

Communication Efficiency

Digits		MNIST	SVHN	USPS	SynthDigit	MNIST_M	AVG
#communication round	FedAvg	11	54	5	11	7	28
	FedBN	11	73	5	68	7	22
	PartialFed	9	23	6	14	8	14
	FedProx	64	42	8	12	10	31
	Per-FedAvg	\	\	\	\	\	\
	FedRep	\	\	\	\	\	\
	ADCOL	19	86	6	19	9	21
communication size (GB)	FedAvg	3.12	15.34	1.42	3.12	1.99	7.95
	FedBN	3.12	20.73	1.42	19.31	1.99	6.25
	PartialFed	2.56	6.53	1.70	3.98	2.27	3.98
	FedProx	18.18	11.93	2.27	3.41	2.84	8.80
	Per-FedAvg	\	\	\	\	\	\
	FedRep	\	\	\	\	\	\
	ADCOL	0.21	0.95	0.07	0.21	0.10	0.23
Speedup		**14.95x**	**16.21x**	**21.52x**	**14.95x**	**20.08x**	**34.42x**

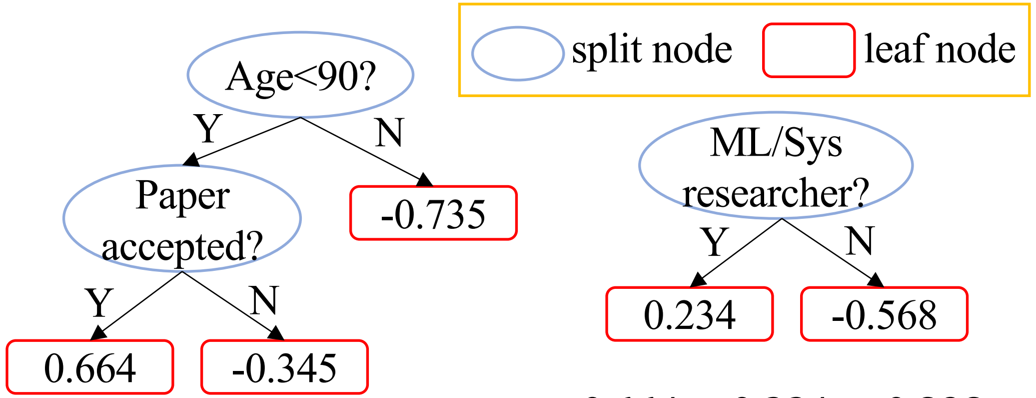
Federated Learning Systems

- TensorFlow-Federated, PySyft, FATE…

what about trees?

…

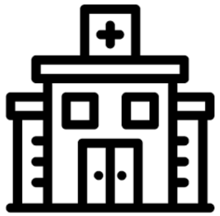
Tree Models are Powerful and Efficient



$$y_p = 0.664 + 0.234 = 0.898$$

GBDT [8]

Credit risk assessment, pricing…



sepsis, cardiovascular…

kaggle champions

[8] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

Centralized GBDT training

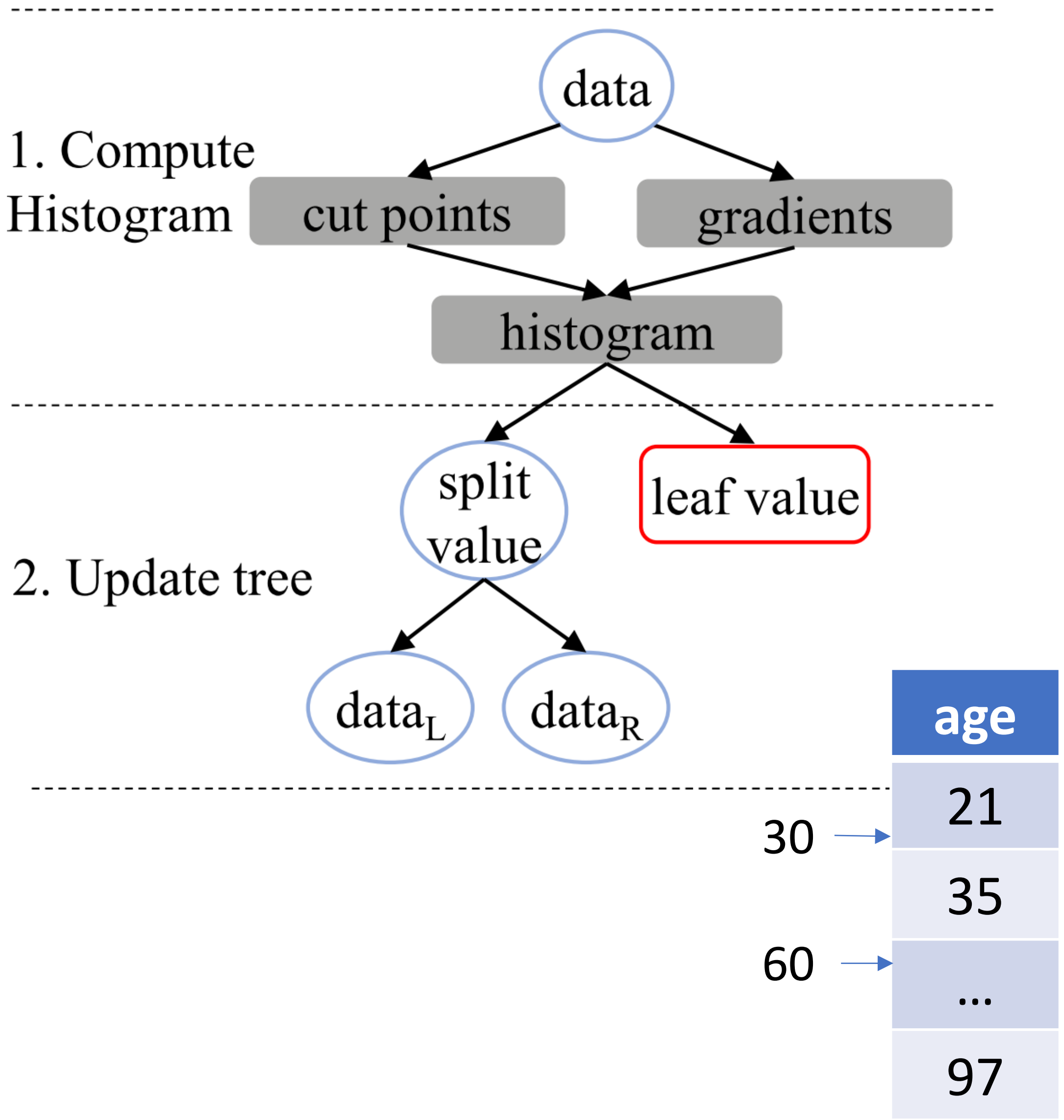
$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

$$\mathbf{H} = \left[\left(\sum_{k \in I_i} g_k, \sum_{k \in I_i} h_k\right)\right]_{i=1}^{B}$$

$$S = \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda}$$

$$V = -\frac{\sum_{i \in I} g_i}{\sum_{i \in I} h_i + \lambda}$$



Histogram-based Learning

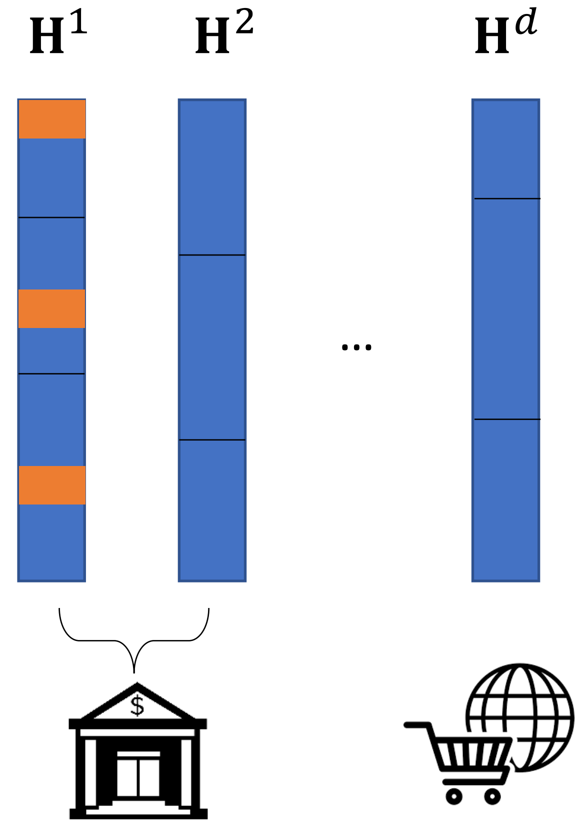
- Histogram concatenation

$$\mathbf{H} = \left[\Big(\sum_{k \in I_1^j} g_k, \sum_{k \in I_1^j} h_k\Big), \ldots, \Big(\sum_{k \in I_B^j} g_k, \sum_{k \in I_B^j} h_k\Big) \right]_{j=1}^{N}$$

$$:= \bigcup_{j=1}^{N} \mathbf{H}^j$$

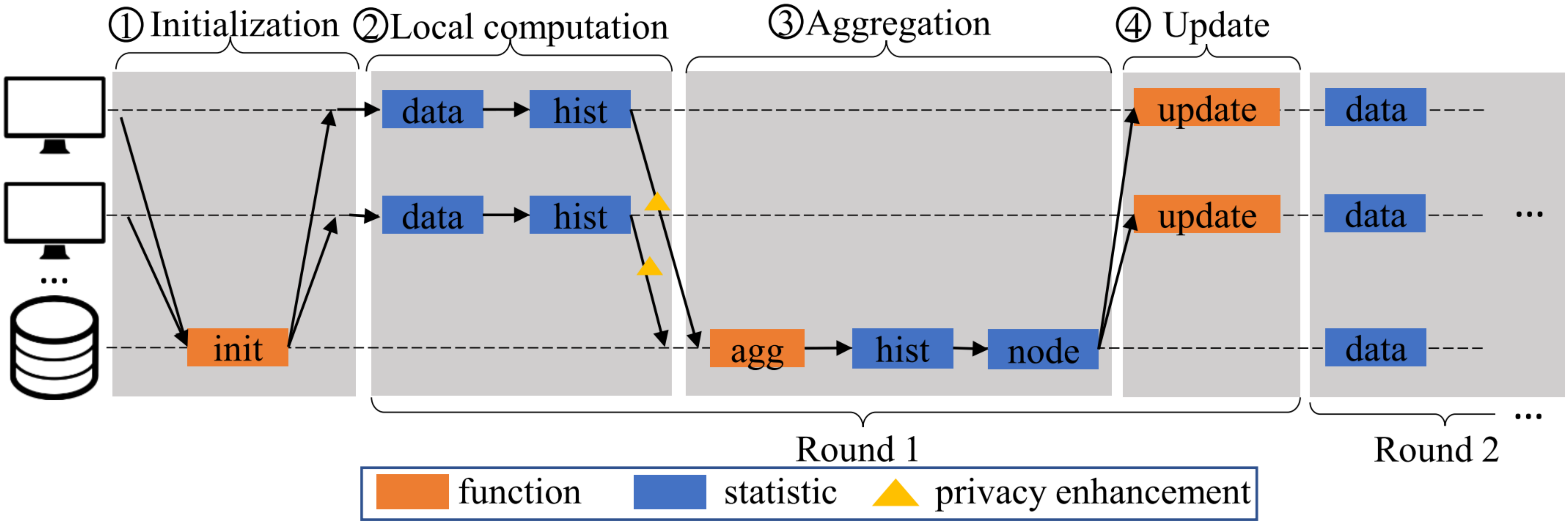
- Histogram summation

$$\mathbf{H} = \left[\Big(\sum_{j \in [N]} \sum_{k \in I_i^j} g_k, \sum_{j \in [N]} \sum_{k \in I_i^j} h_k\Big) \right]_{i=1}^{B} = \sum_{j \in [N]} \mathbf{H}^j$$



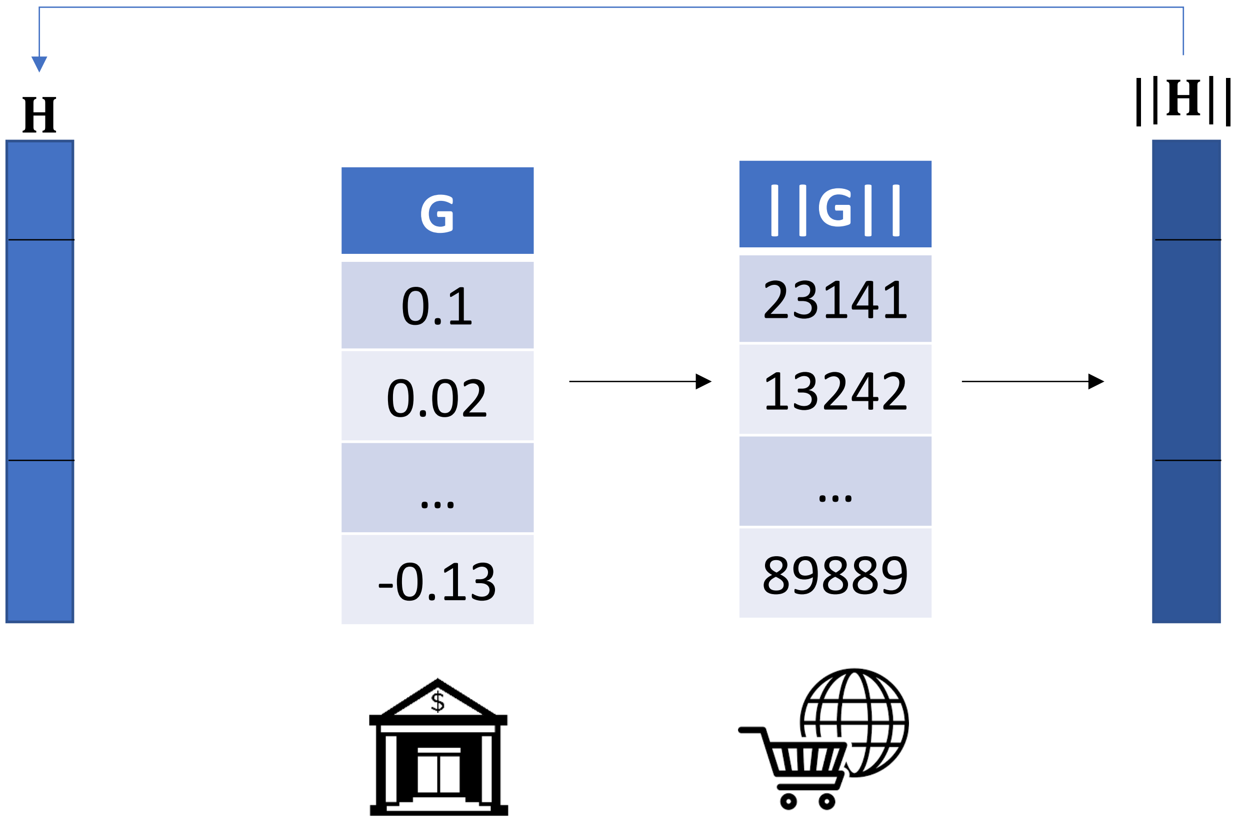
Framework

- Instead of transferring data/model, we transfer histograms for training.



Privacy

- Label privacy
  - Additively homomorphic encryption



Privacy

- Histogram Privacy
  - Secure Aggregation

$$\mathbf{H}^i \leftarrow \mathbf{H}^i + \sum_j k_{ij} - \sum_j k_{ji}$$

$\mathbf{H_1}$
$a_1 + 0.1$
$a_2 + 0.2$

+

$\mathbf{H_2}$
$b_1 - 0.1$
$b_2 - 0.2$

$\downarrow$

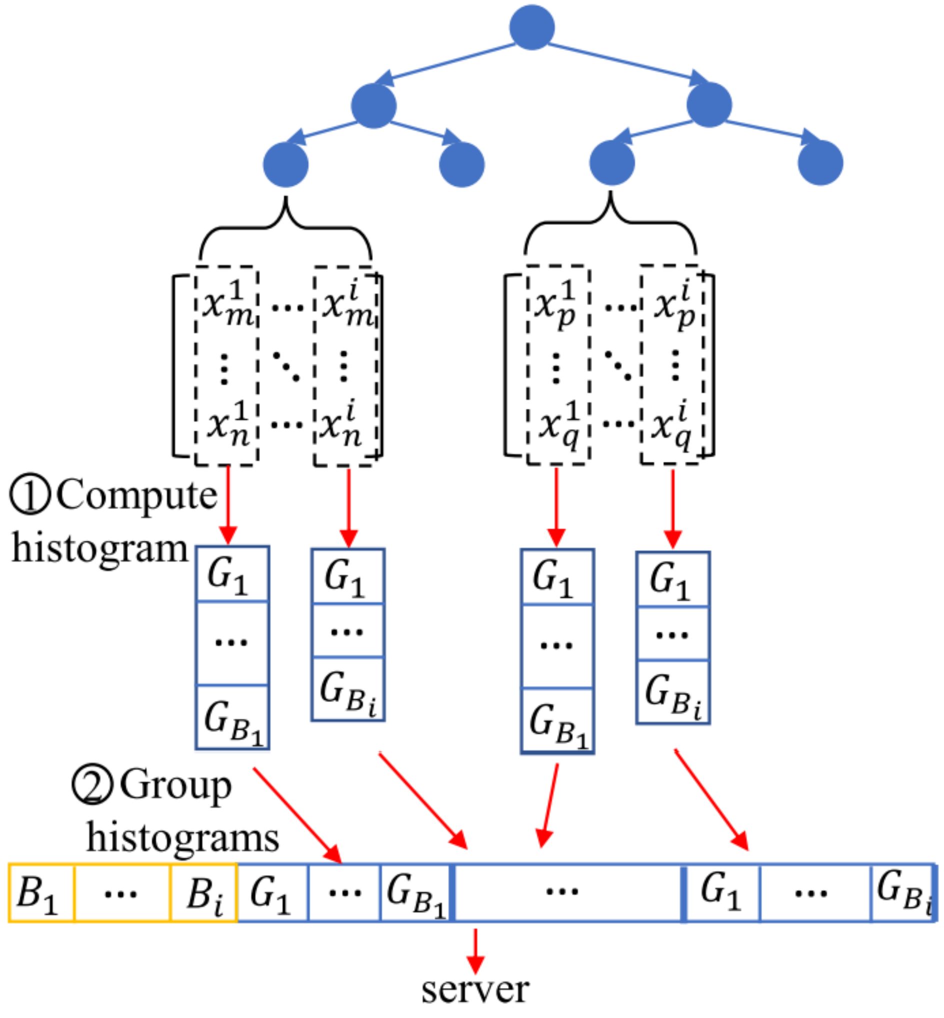
$\mathbf{H_1}$
$a_1 + \mathrm{n}_1$
$a_2 + n_2$

  - Differential Privacy

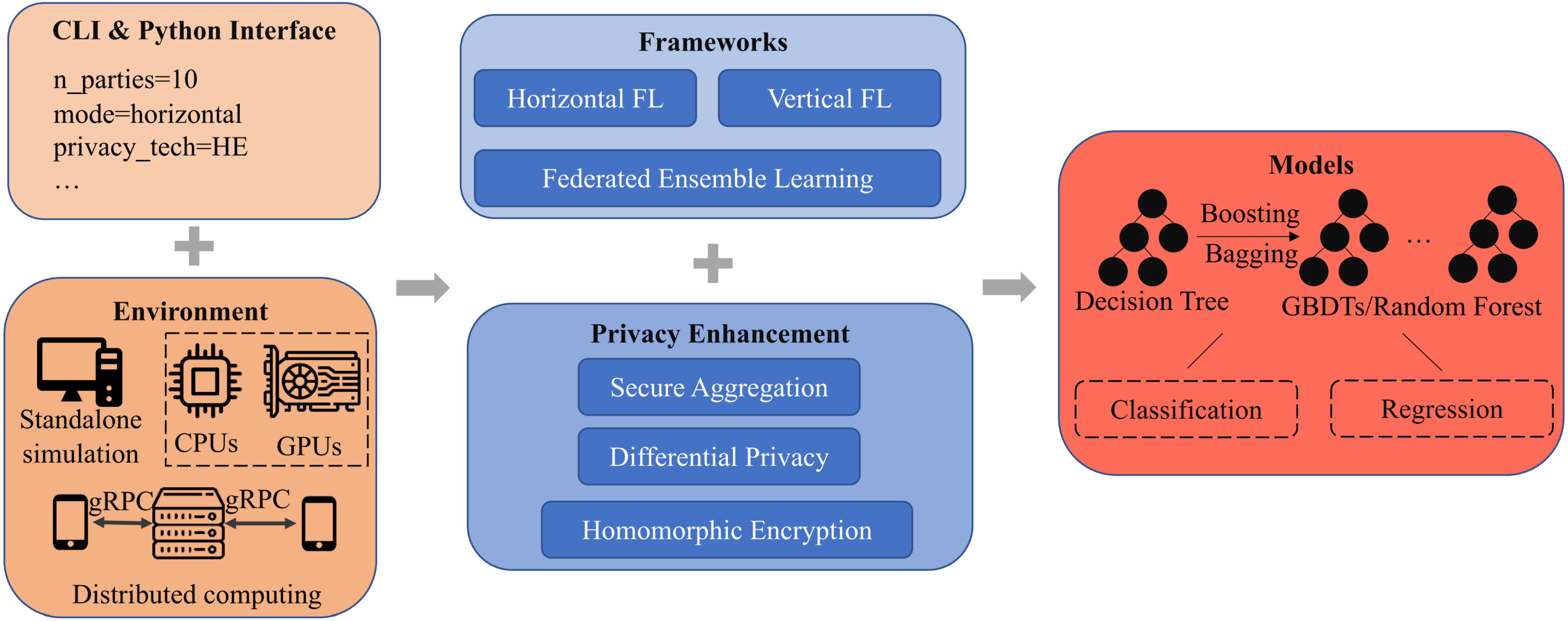
$$\mathbf{H}^i \leftarrow \mathbf{H}^i + Lap(0, \frac{2R}{\varepsilon})$$

Optimization

- Computation
  - Node-level & feature-level parallelism
- Communication
  - Batching



FedTree[9]



[9] Li, Qinbin, et al. "FedTree: A Federated Learning System For Trees." Proceedings of Machine Learning and Systems 5 (2023).

Usage

Two lines

Installation

```
git clone –recursive https://github.com/Xtra-Computing/FedTree.git
cd FedTree && mkdir build  && cd build && cmake .. && make -j
```

Prepare the Configuration File

```
data=/dataset/credit/credit_vertical_p1.csv
n_parties=2
num_class=2
mode=vertical
data_format=csv
objective=binary:logistic
privacy_tech=none
learning_rate=0.1
n_trees=10
ip_address=127.0.0.1
```

Prediction of default of credit card clients

Run

```
./build/bin/FedTree-distributed-server ./example/credit/credit_vertical_p0.conf
```

```
./build/bin/FedTree-distributed-party ./example/credit/credit_vertical_p0.conf 0
```

```
./build/bin/FedTree-distributed-party ./example/credit/credit_vertical_p1.conf 1
```

Run a single-line command in each machine

Get Results

```
AUC=0.770339
AUC=0.771878
AUC=0.774684
AUC=0.775256
AUC=0.776345
```

Better results than local training

Evaluation

- Machine: 1) simulation: 4x AMD EPYC 7543 CPUs + 4x NVIDIA A100 GPUs 2) distributed deployment: 8 machines, each has 2x Intel Xeon E5-2680 v4 CPUs
- Baselines: 1) XGBoost[10] 2) FATE[11] 3) SOLO 4) FEL[12]
- Tabular datasets: breast, a9a, cod-rna, mnist, abalone

[10] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[11] Liu, Yang, et al. "Fate: An industrial grade platform for collaborative learning with data protection." The Journal of Machine Learning Research 22.1 (2021): 10320-10325.

[12] Zhao, Lingchen, et al. "Inprivate digging: Enabling tree-based distributed data mining with differential privacy." IEEE INFOCOM 2018-IEEE Conference on Computer Communications. IEEE, 2018.

Accuracy

Datasets	Centralized	Horizontal FL					Vertical FL				
	XGBoost	FedTree	FedTree+SA	FATE	SOLO	FEL	FedTree	FedTree+HE	FATE	SOLO	FEL
breast	1.0	1.0	1.0	0.995	0.999	1.0	1.0	1.0	1.0	0.944	0.988
a9a	0.902	0.902	0.902	0.902	0.883	0.896	0.902	0.902	0.902	0.608	0.573
cod-rna	0.993	0.993	0.993	0.992	0.968	0.977	0.993	0.993	0.993	0.667	0.754
mnist	0.983	0.983	0.983	✗	0.926	0.943	0.983	0.983	0.983	0.759	0.822
abalone	0.078	0.079	0.079	0.080	0.085	0.289	0.078	0.078	0.078	0.138	0.266

Same with centralized training!

Efficiency

Training time (s) per tree

simulation

Datasets	Horizontal FL			Vertical FL		
	FedTree	FATE	Speedup	FedTree	FATE	Speedup
breast	0.117	1.97	**16.8x**	0.90	1.73	**1.9x**
a9a	0.289	18.41	**63.7x**	6.46	55.31	**8.6x**
cod-rna	0.388	25.84	**66.6x**	4.22	104.59	**24.8x**
mnist	8.102	529.49	**65.6x**	274.42	790.37	**2.9x**
abalone	0.075	1.55	**20.7x**	0.877	3.10	**3.5x**

distributed

Datasets	Horizontal FL			Vertical FL		
	FedTree	FATE	Speedup	FedTree	FATE	Speedup
breast	0.22	10.30	**46.8x**	1.49	4.72	**3.2x**
a9a	0.66	23.51	**35.6x**	8.03	29.13	**3.6x**
cod-rna	0.34	27.43	**80.7x**	9.99	52.50	**5.3x**
mnist	25.9	838.04	**32.4x**	22.77	506.42	**22.3x**
abalone	0.29	6.39	**22.0x**	2.06	5.39	**2.6x**

Ablation study - Batching

Total communication time (s) and communication size (MB)

Datasets	Horizontal FedTree				Vertical FedTree			
	w/o LBC	w LBC	speedup	size	w/o LBC	w LBC	speedup	size
breast	8.82	4.64	**1.9x**	20.4	16.301	5.26	**3.1x**	23.5
a9a	11.72	4.04	**2.9x**	14.2	21.56	6.16	**3.5x**	27.0
cod-rna	12.42	5.4	**2.3x**	33.4	55.46	14.22	**3.9x**	48.5
mnist	30.16	8.15	**3.7x**	57.3	80.26	16.72	**4.8x**	59.4
abalone	13.156	5.06	**2.6x**	32.9	13.61	5.67	**2.4x**	43.3

Industry Applications

- Energy prediction, anomaly detection…

- Follow our code!
- https://github.com/Xtra-Computing/FedTree

Federated Learning Systems

FATE

FedML

PaddleFL

Fedlearner

TFF

Flower

FLUTE

CrypTen

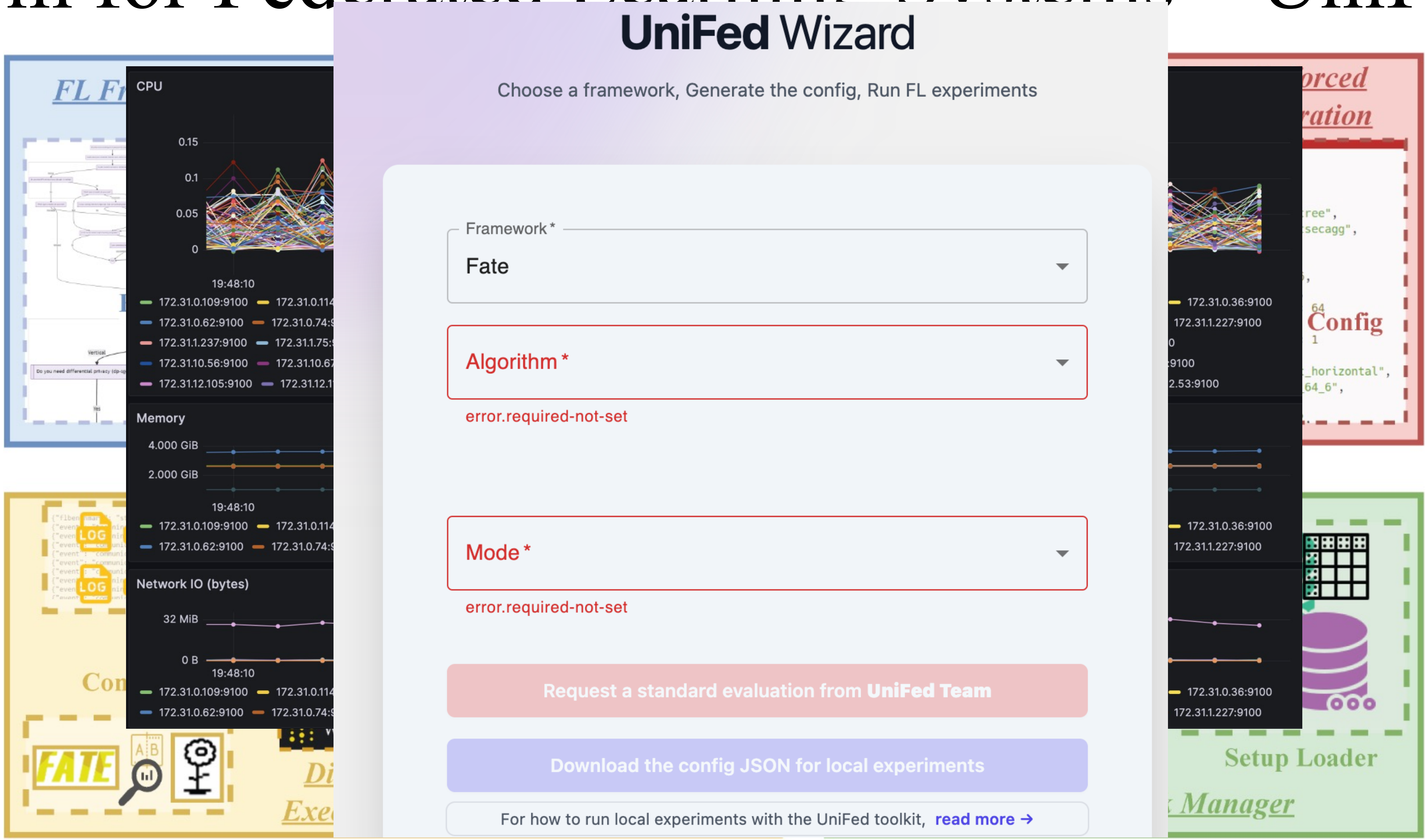
FedTree

FedScale

FederatedScope

...

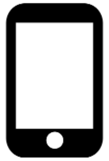
Platform for Federated Learning Systems – UniFed[13]



[13] Liu, Xiaoyuan, et al. "Unifed: A benchmark for federated learning frameworks." arXiv preprint arXiv:2207.10308 (2022).

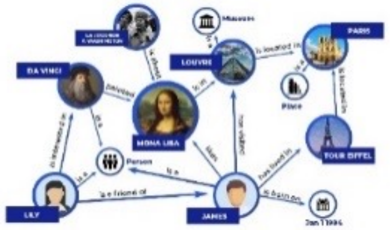
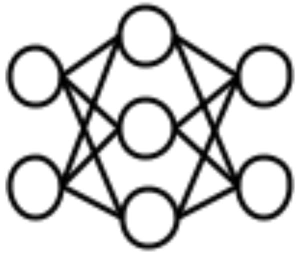
Future Directions

Heterogeneous participants



Incentive
Fairness

Heterogeneous “data”

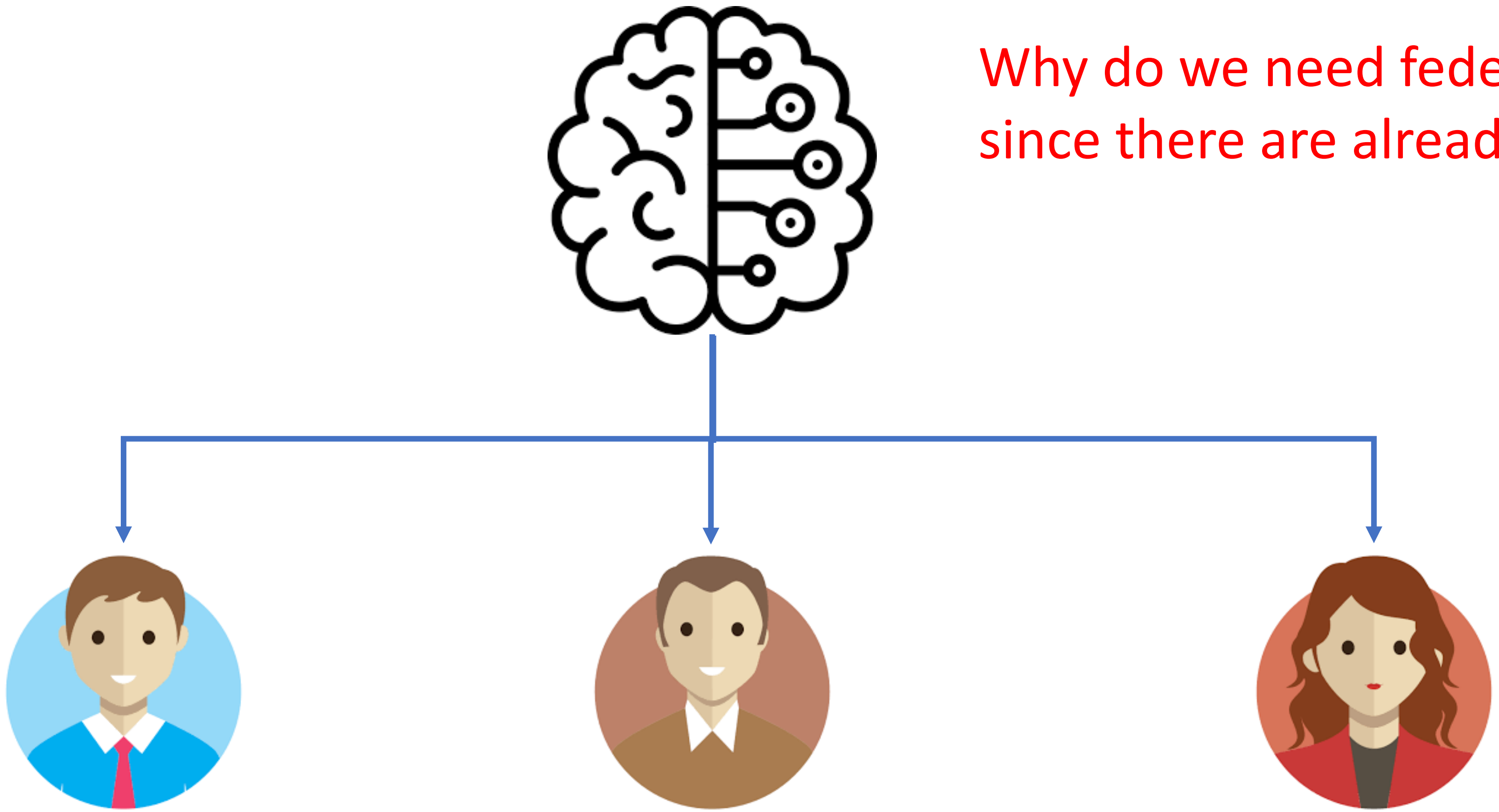




Unify
Privacy

Heterogeneous hardware

Adaptation
Synchronization

Foundation models

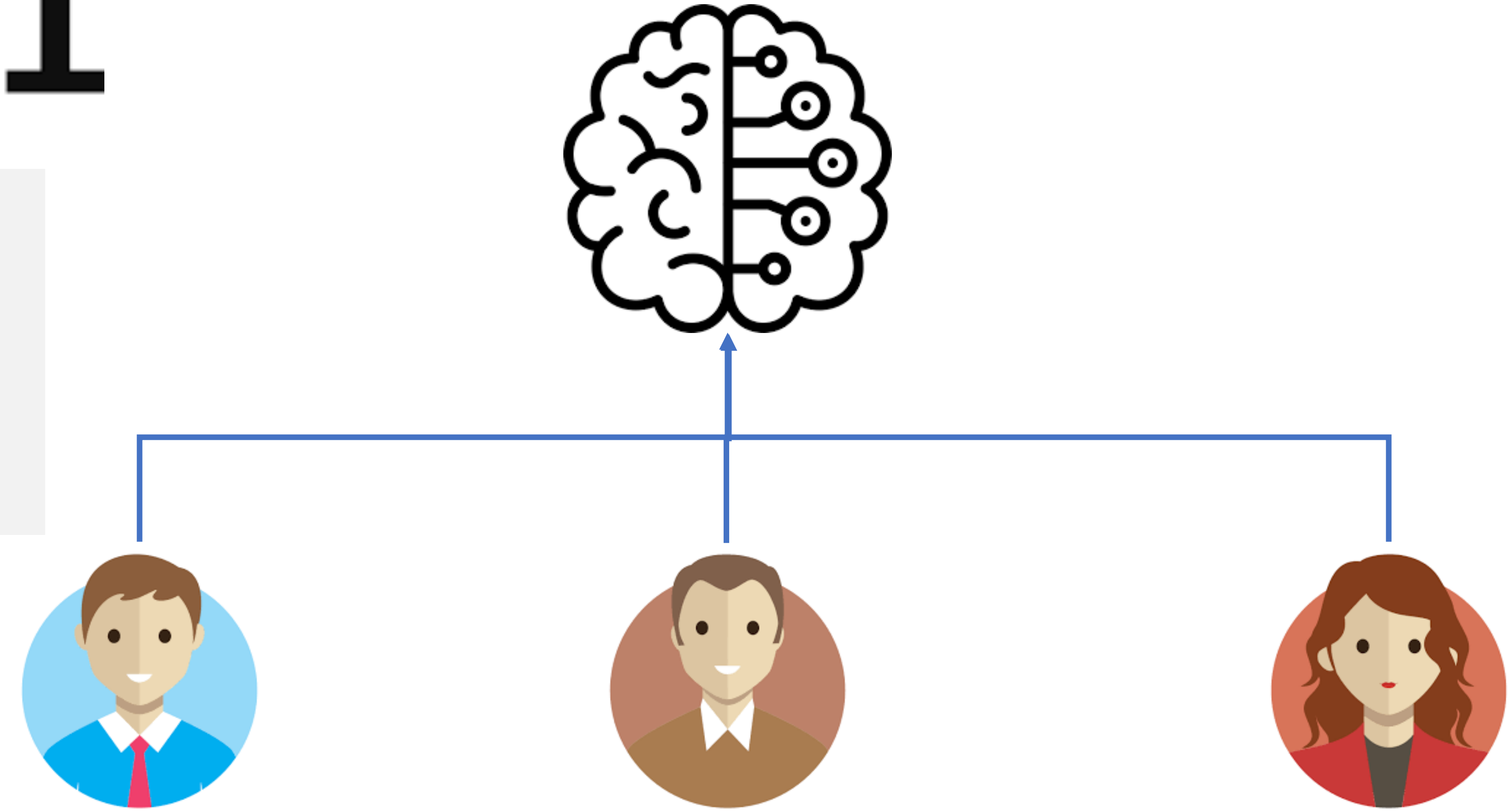


Why do we need federated learning since there are already powerful models?

“Monopoly” by Companies

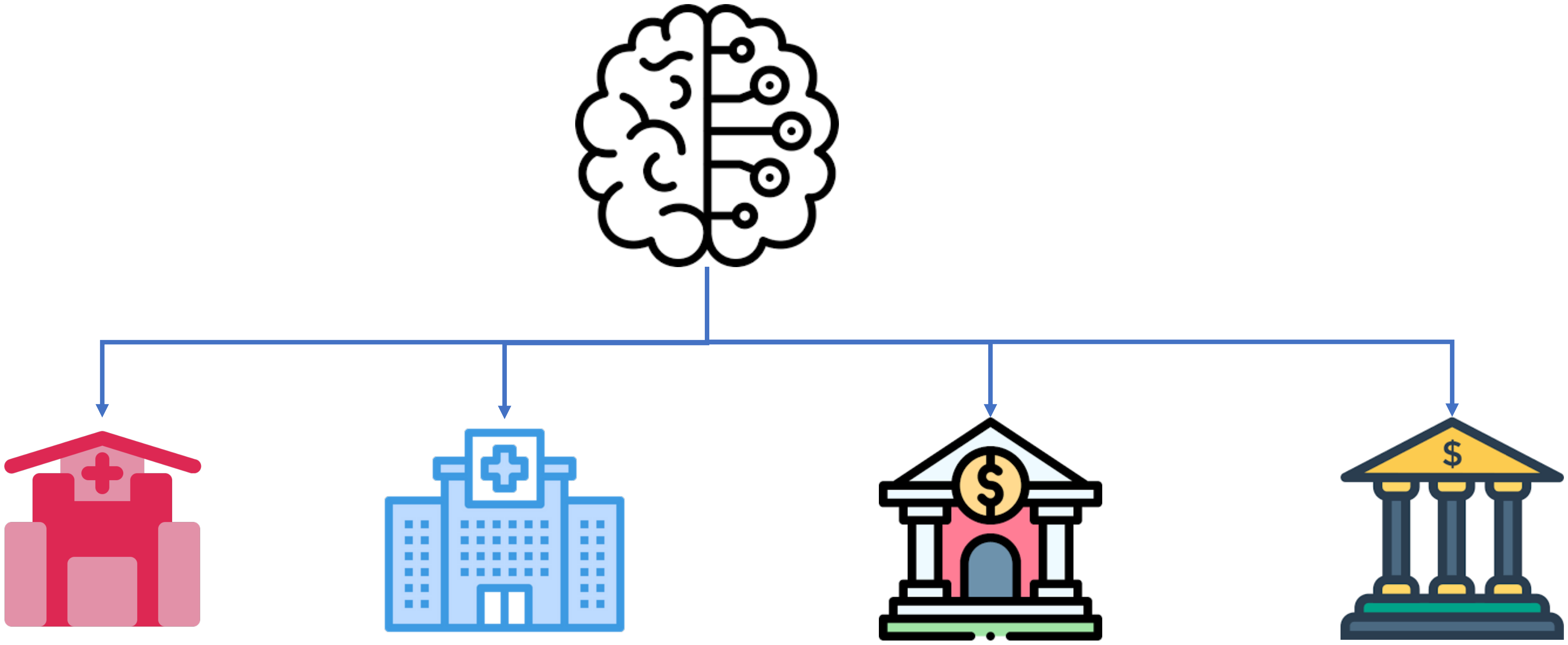
…

Contributing Data & Resources



Private Domains

- Federated Fine-Tuning of Foundation Models



Takeaway

- Non-IID data is a key challenge in FL
  - Comparing representations helps a lot!

- Deep learning models are powerful
  - Tree is a good option for FL deployment
  - FL+LLMs is challenging!

Thank you!