FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning

Jianqing Zhang[1] Yang Liu[2] Yang Hua[3] Jian Cao[1]

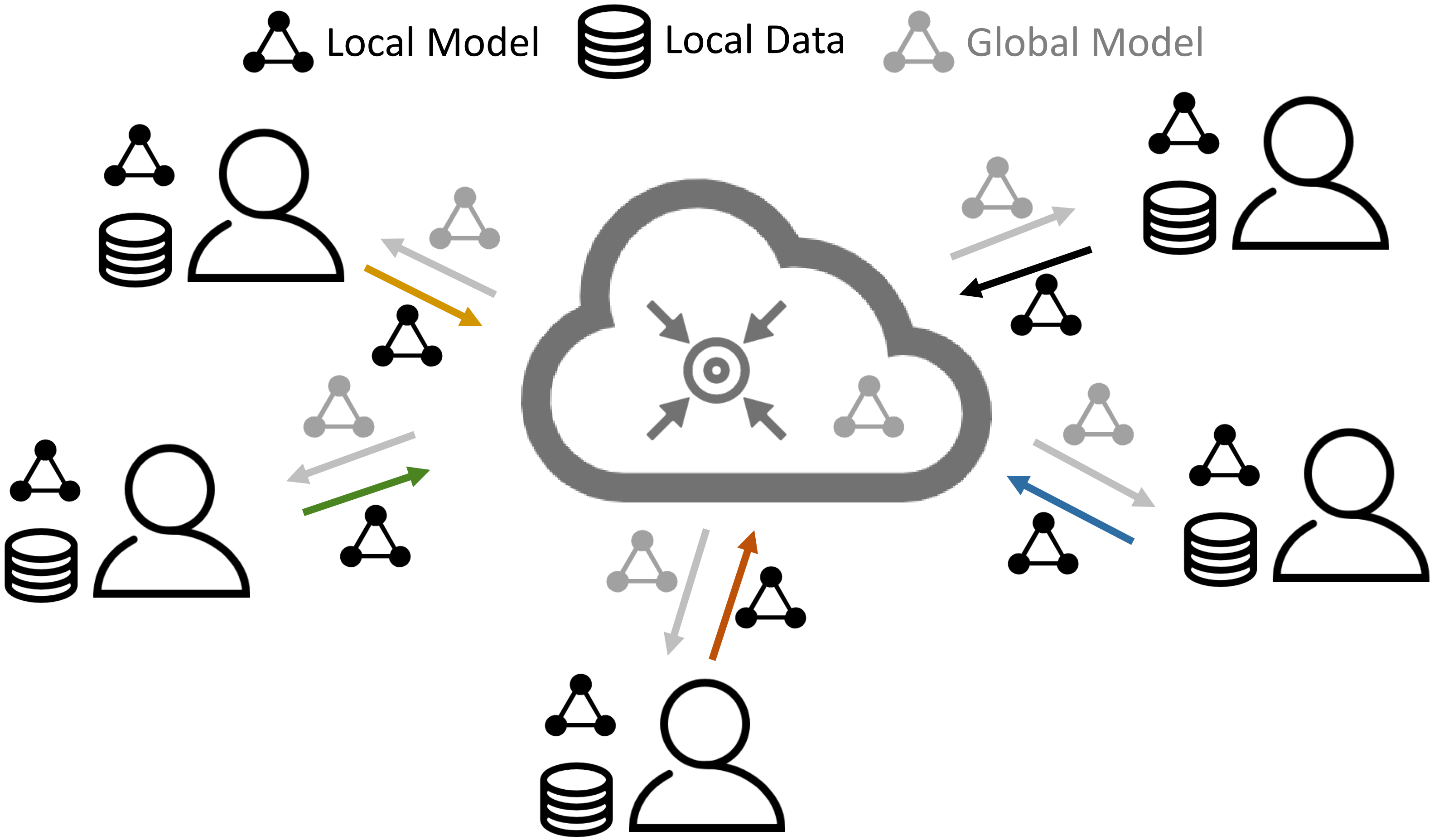
1 上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

2 AIR 清華大學 智能产业研究院 Institute for AI Industry Research, Tsinghua University

3 QUEEN'S UNIVERSITY BELFAST

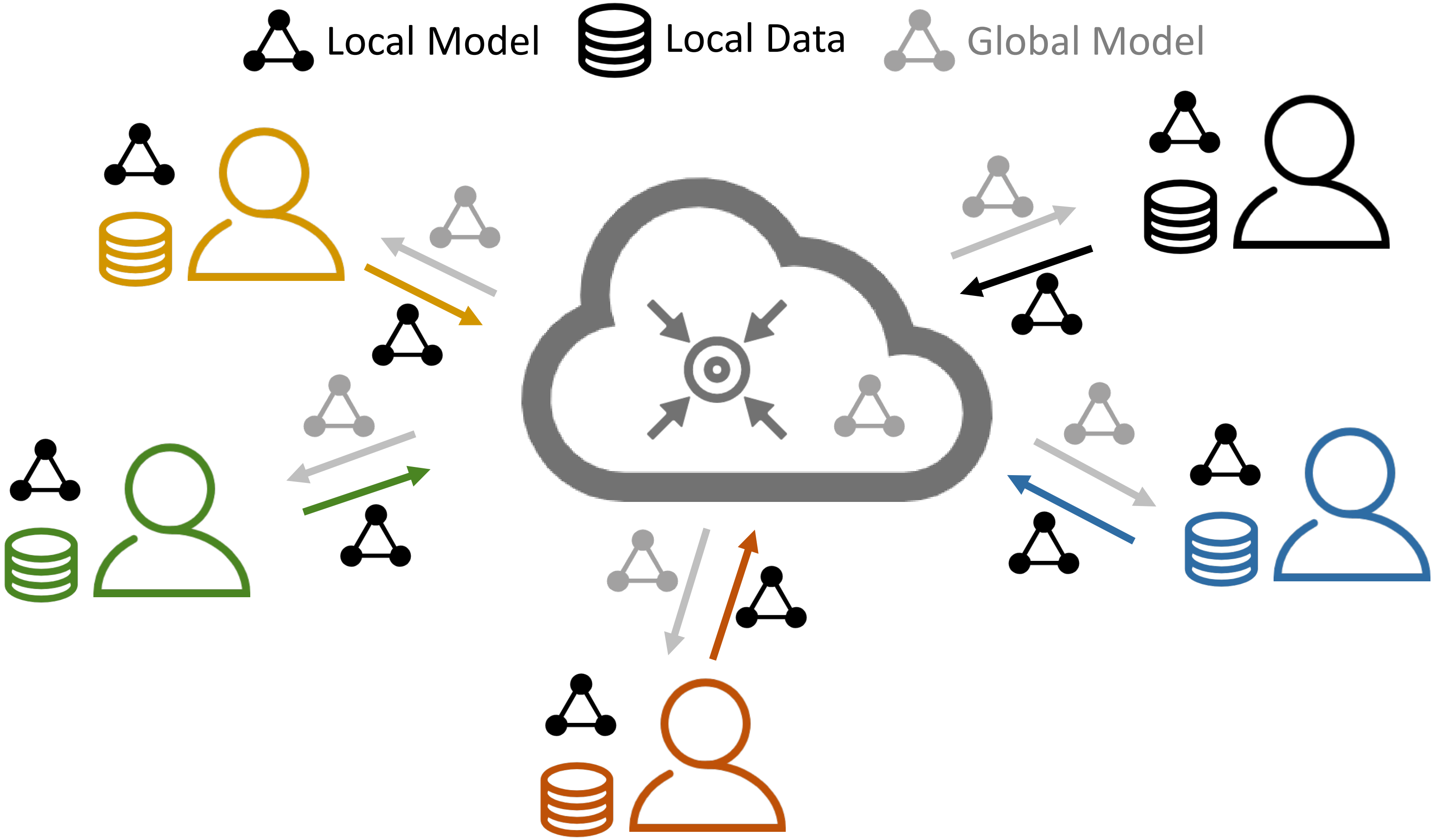
Federated Learning (FL)

- FL allows multiple clients to train their models collaboratively without exposing data.



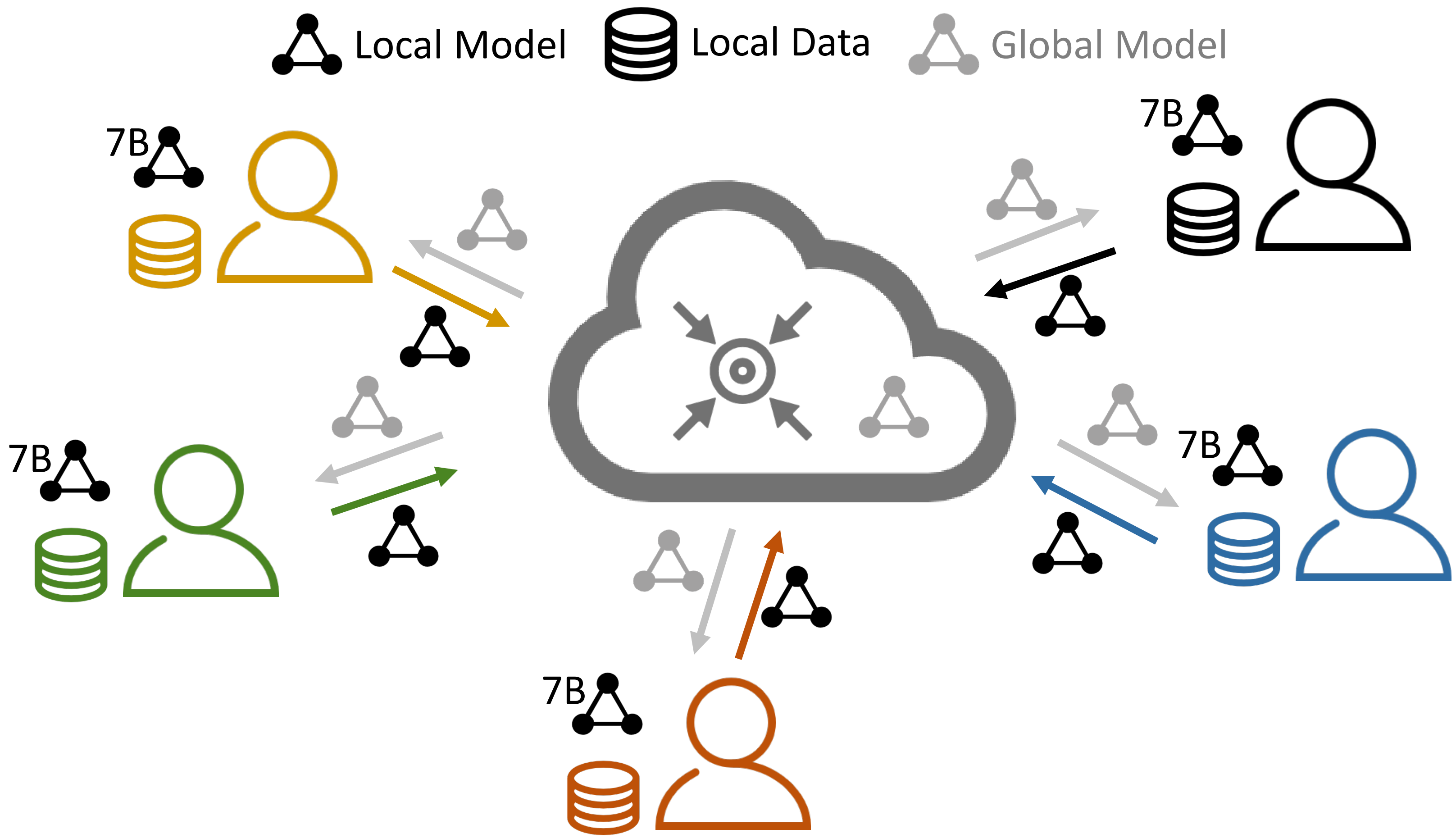
Data Heterogeneity

- Client-specific private data has its *unique* distribution



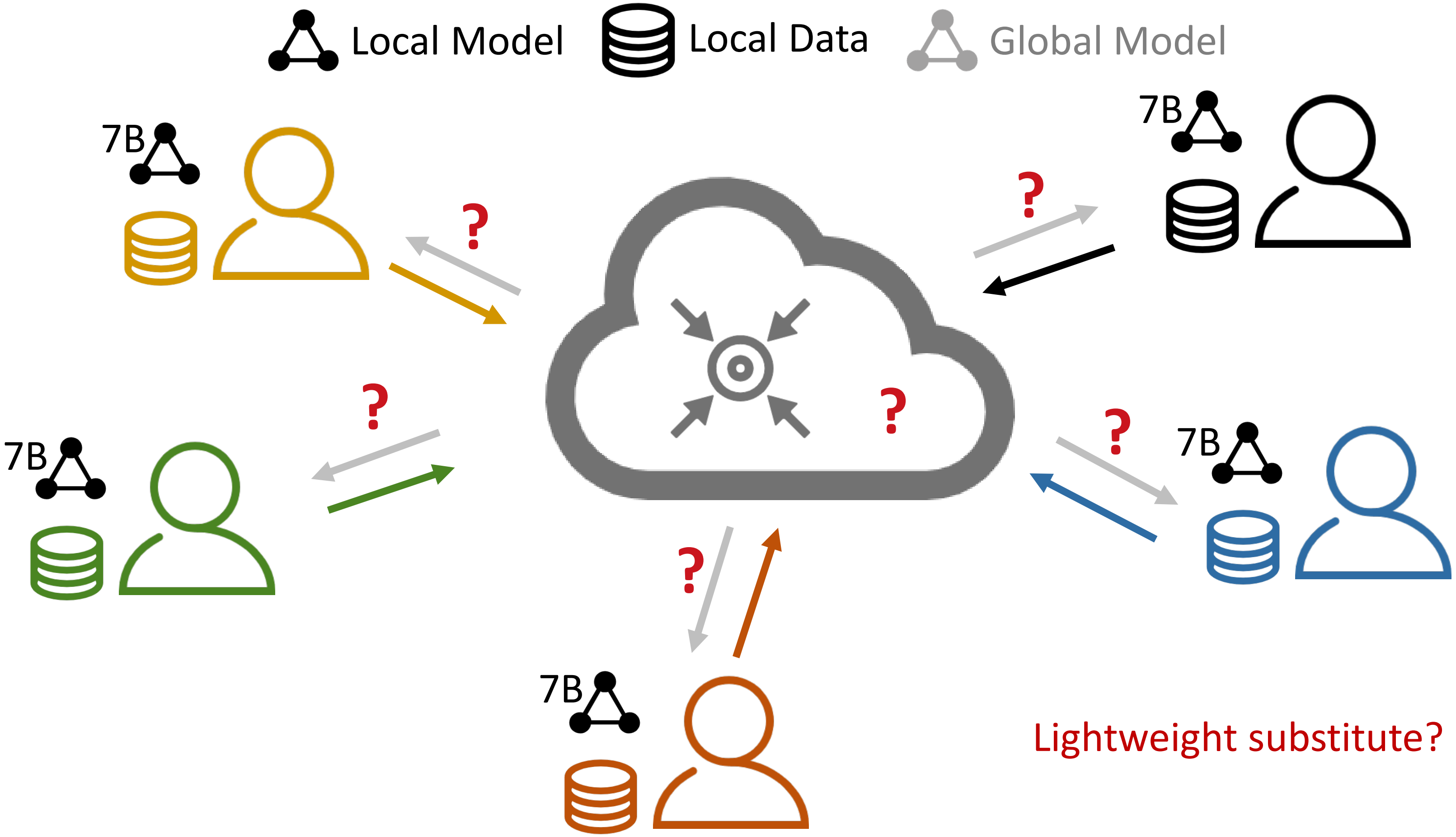
Communication Overhead

- In the era of large models, typical FL suffers huge *communication overhead*, as
- it transmits model parameters



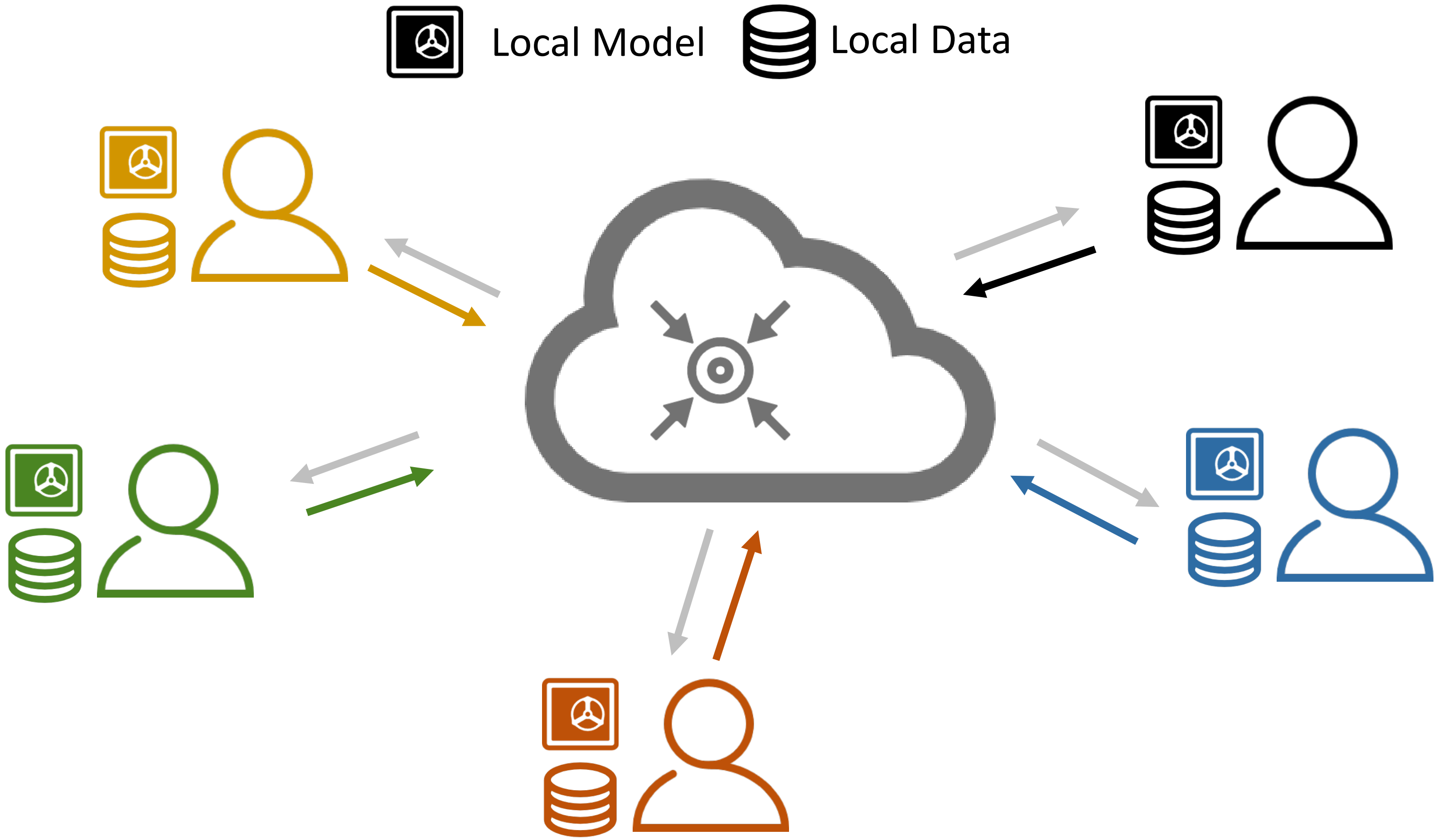
Communication Overhead

- In the era of large models, typical FL suffers huge *communication overhead*, as
- it transmits model parameters



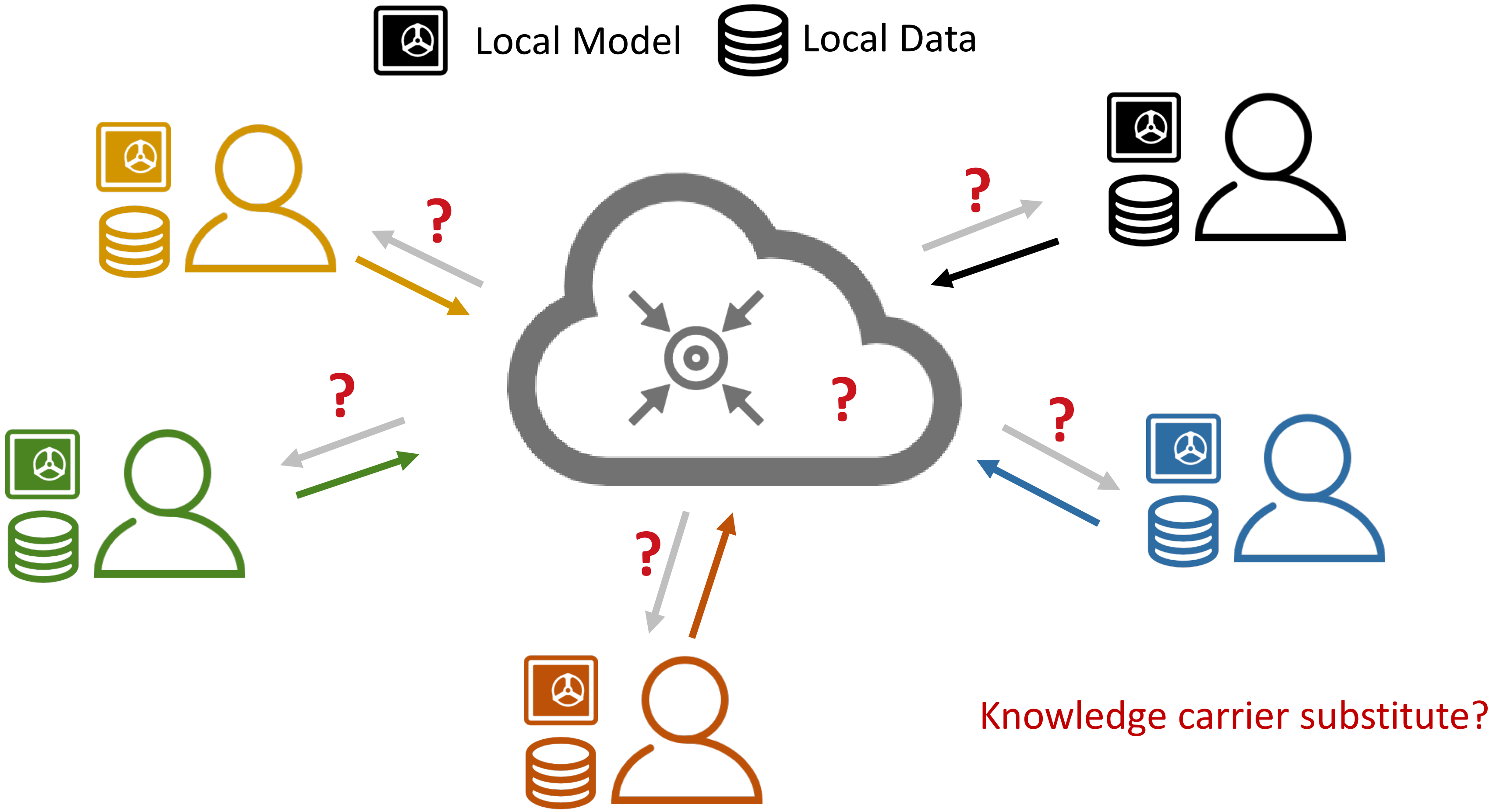
Intellectual Property Protection

- Client model parameters are unique and require substantial effort to obtain,
- representing a form of *intellectual property (IP)* that should be protected



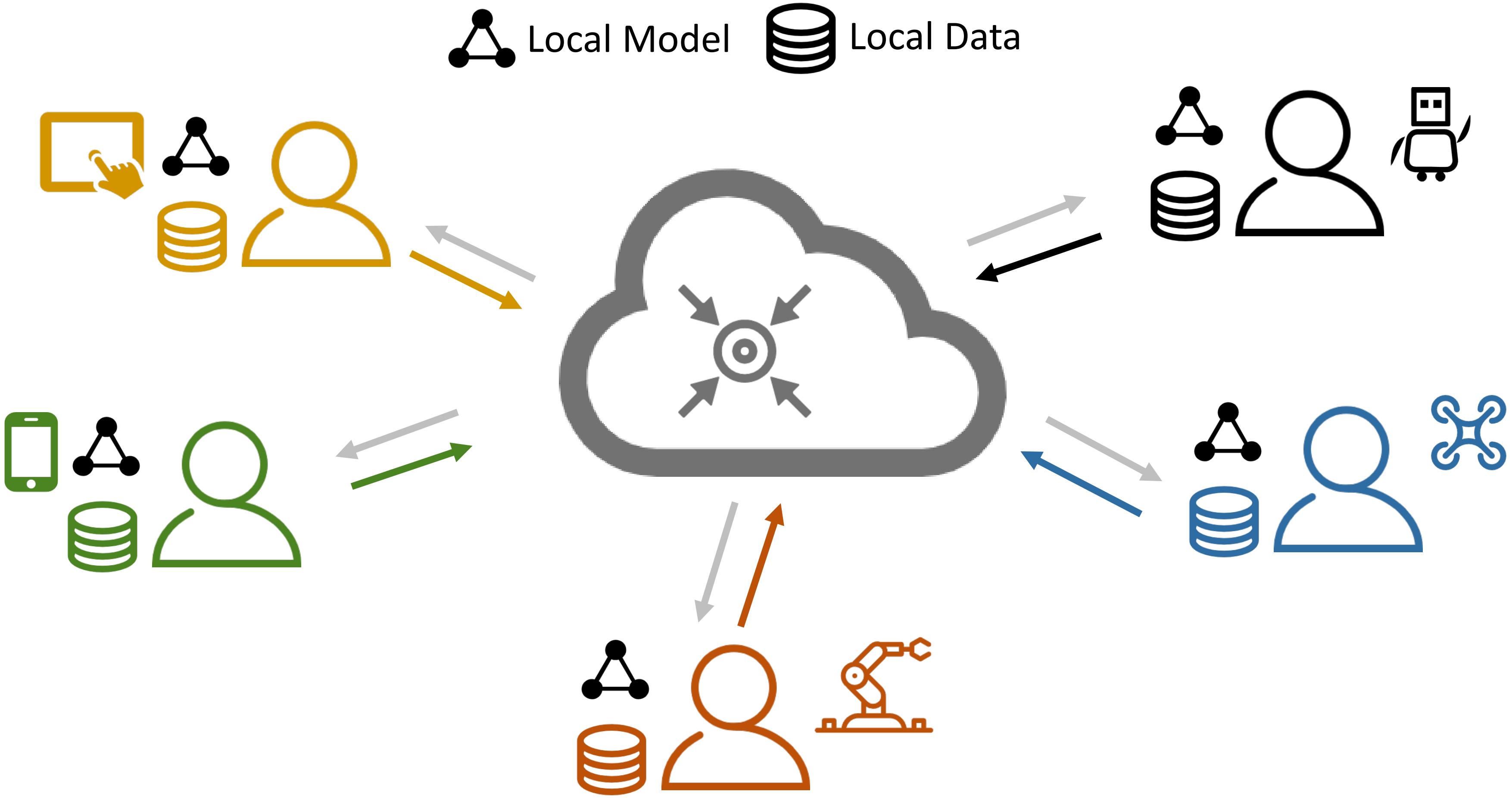
Intellectual Property Protection

- Client model parameters are unique and require substantial effort to obtain,
- representing a form of *intellectual property (IP)* that should be protected



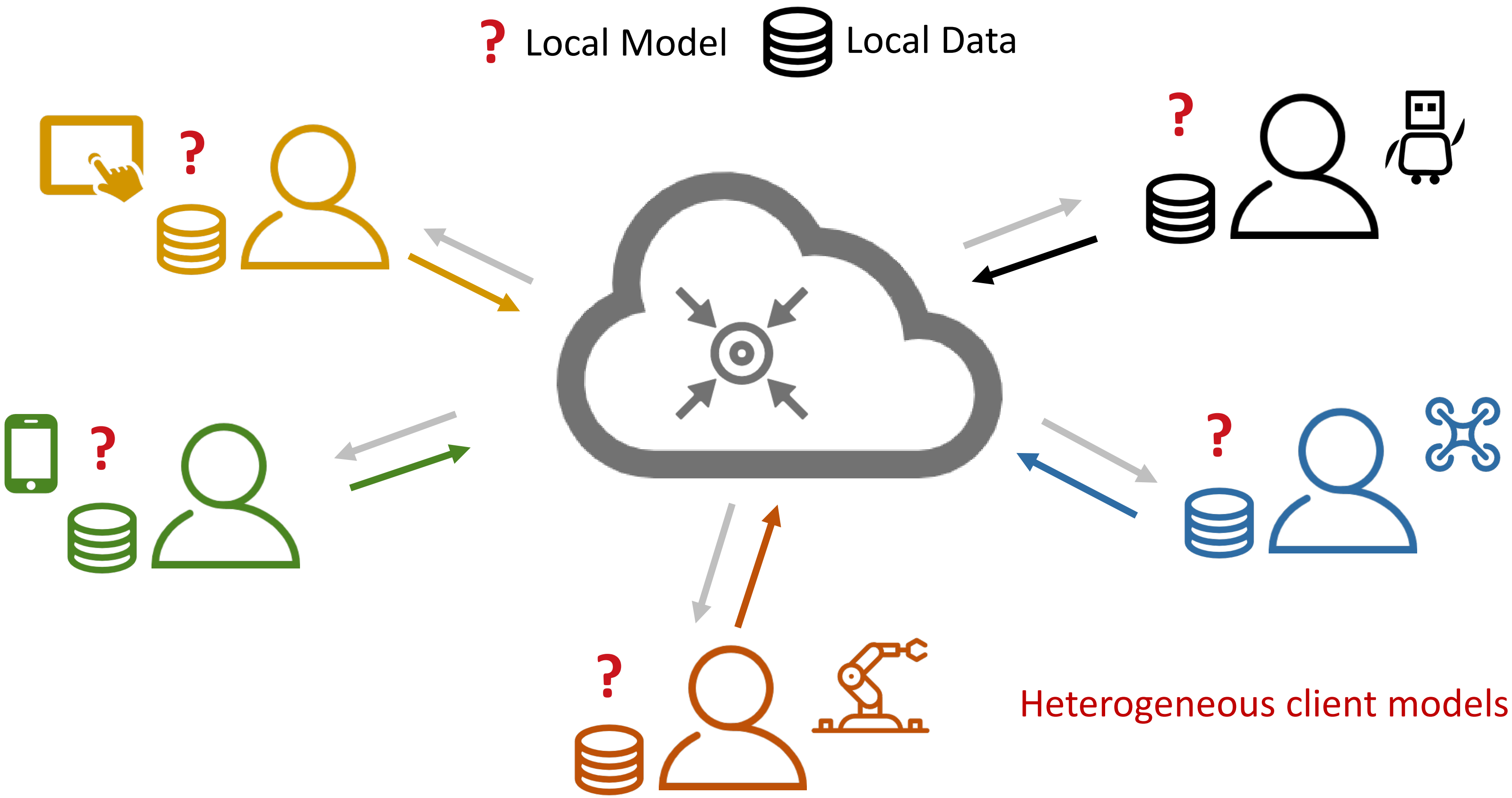
Resource Diversity

- Clients using different devices suffer from *resource diversity* when
- training homogeneous (same architectures) local models



Resource Diversity

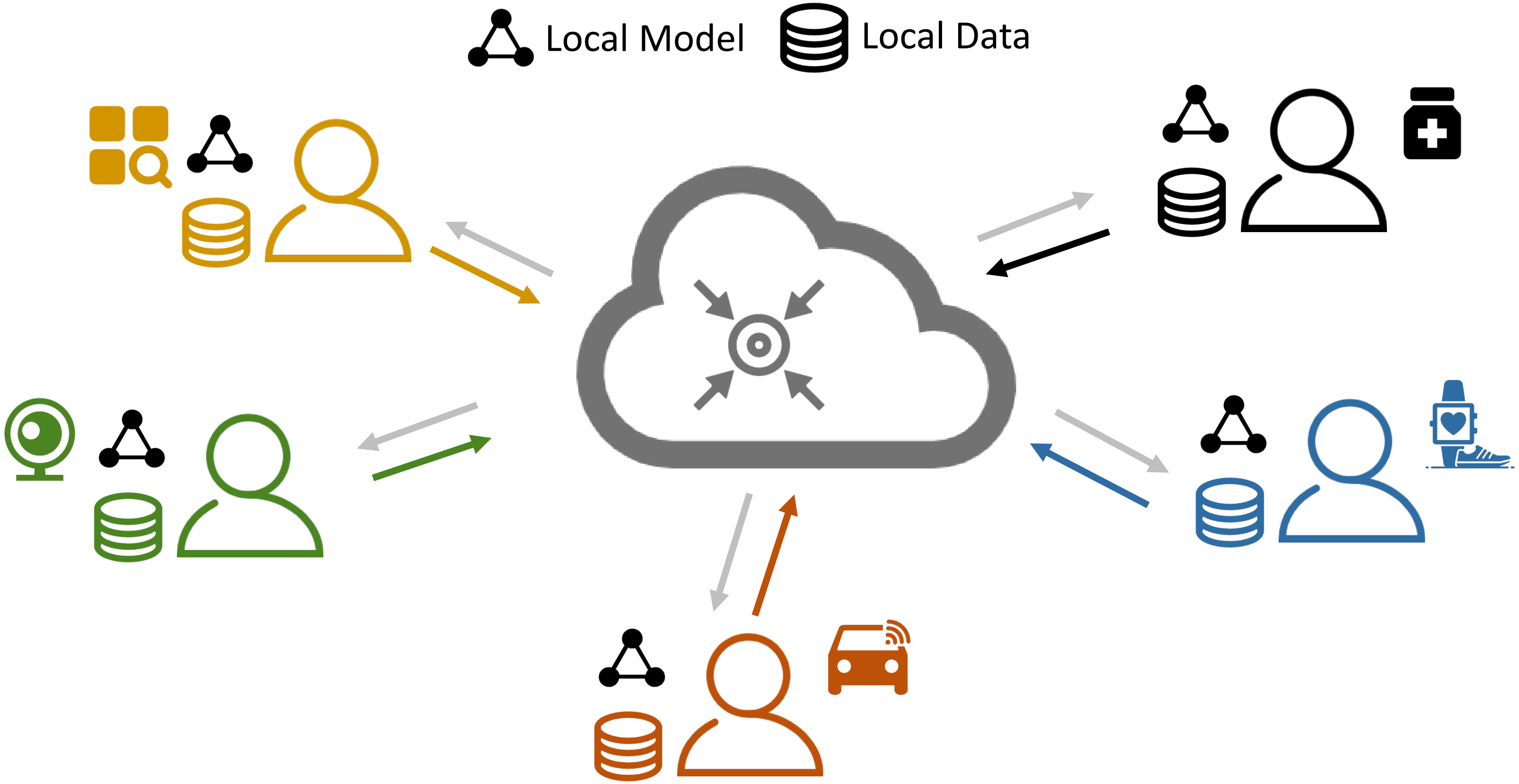
- Clients using different devices suffer from *resource diversity*
- when training homogeneous (same architectures) local models



Heterogeneous client models?

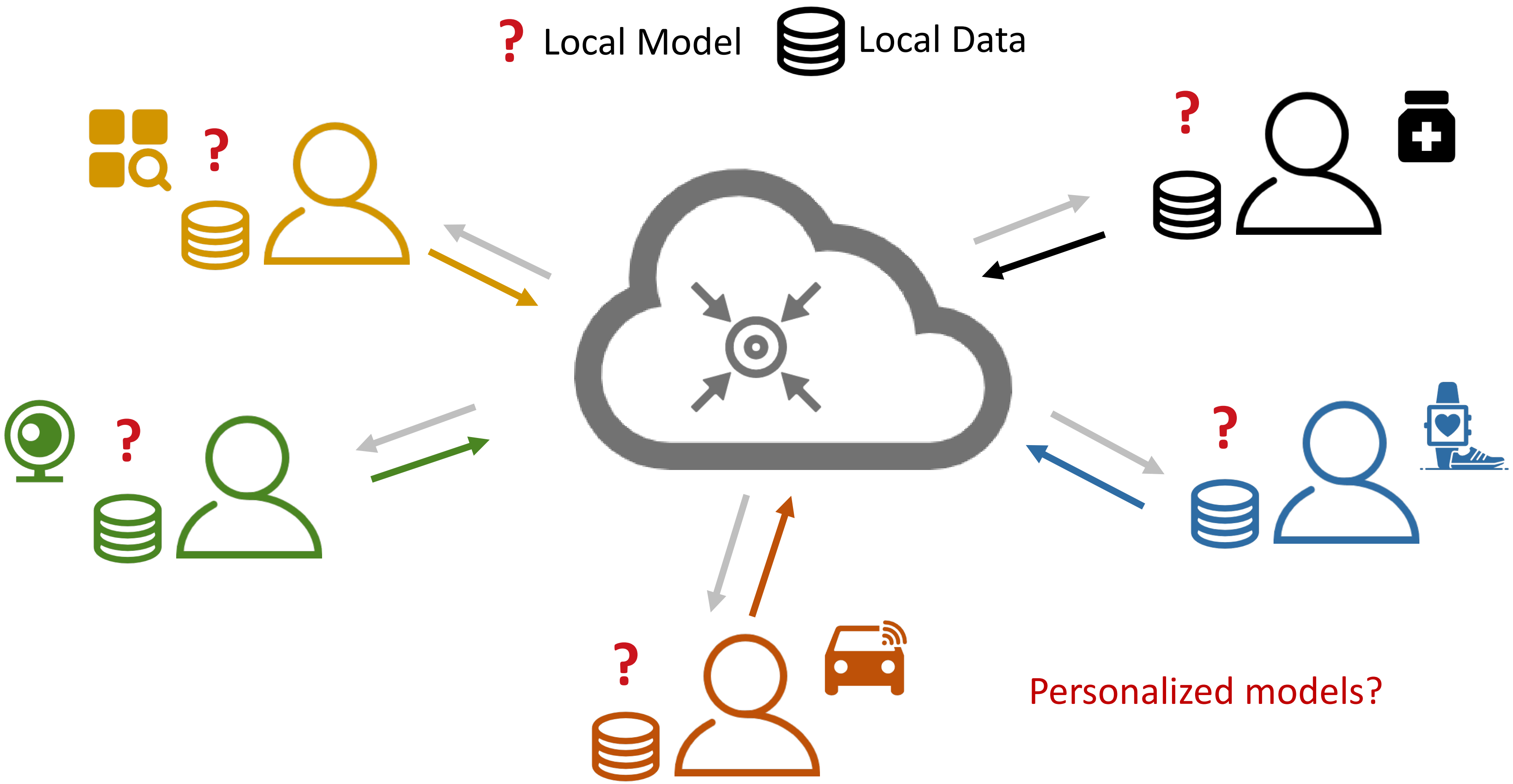
Personal Requirements

- Clients' *local tasks* require tailored model designs



Personal Requirements

- Clients' *local tasks* require tailored model designs

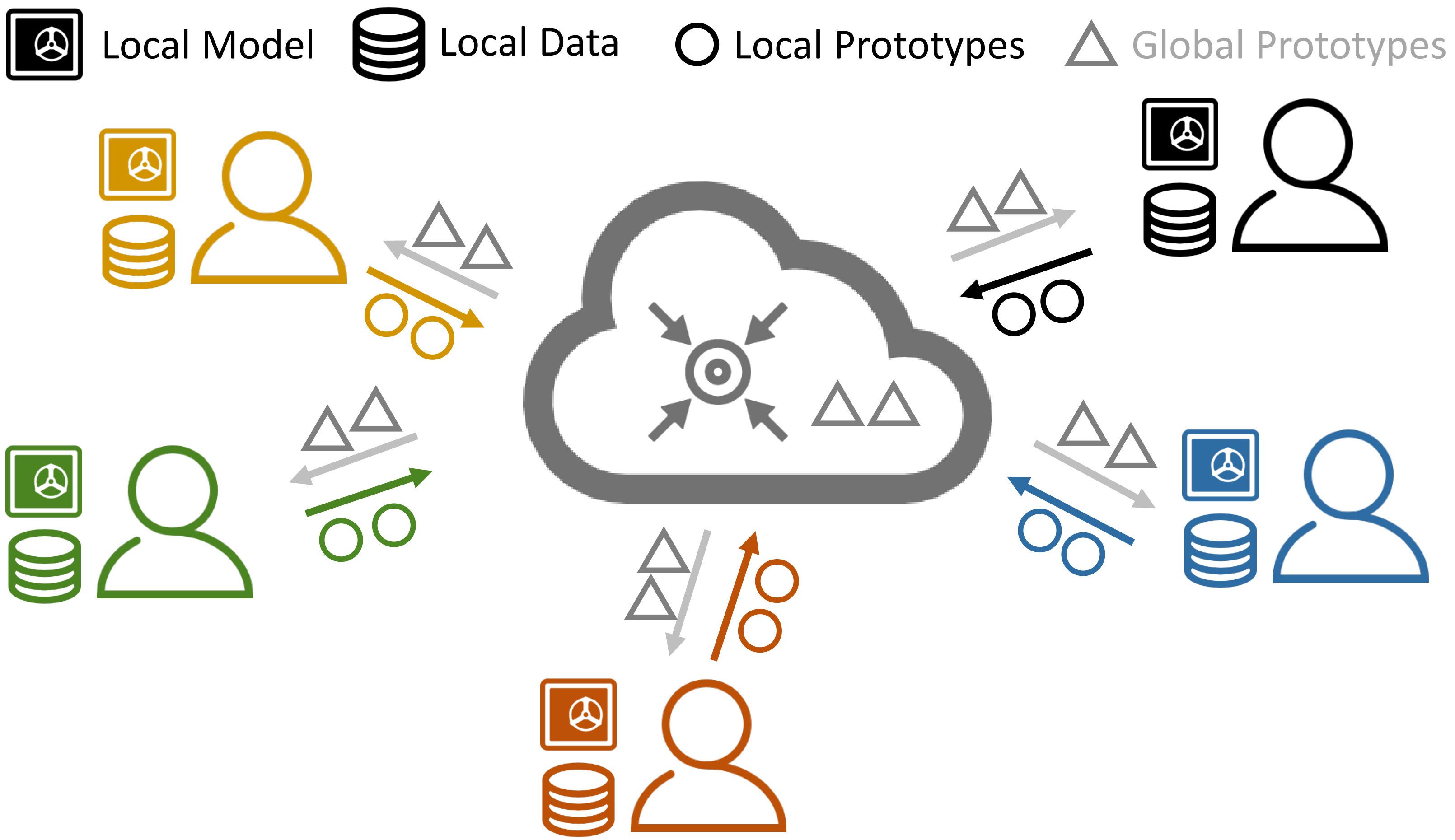


Heterogeneous Federated Learning (HtFL)

- HtFL considers both data and model heterogeneity, and
- Transmits *lightweight knowledge carriers* instead of exposing model parameters

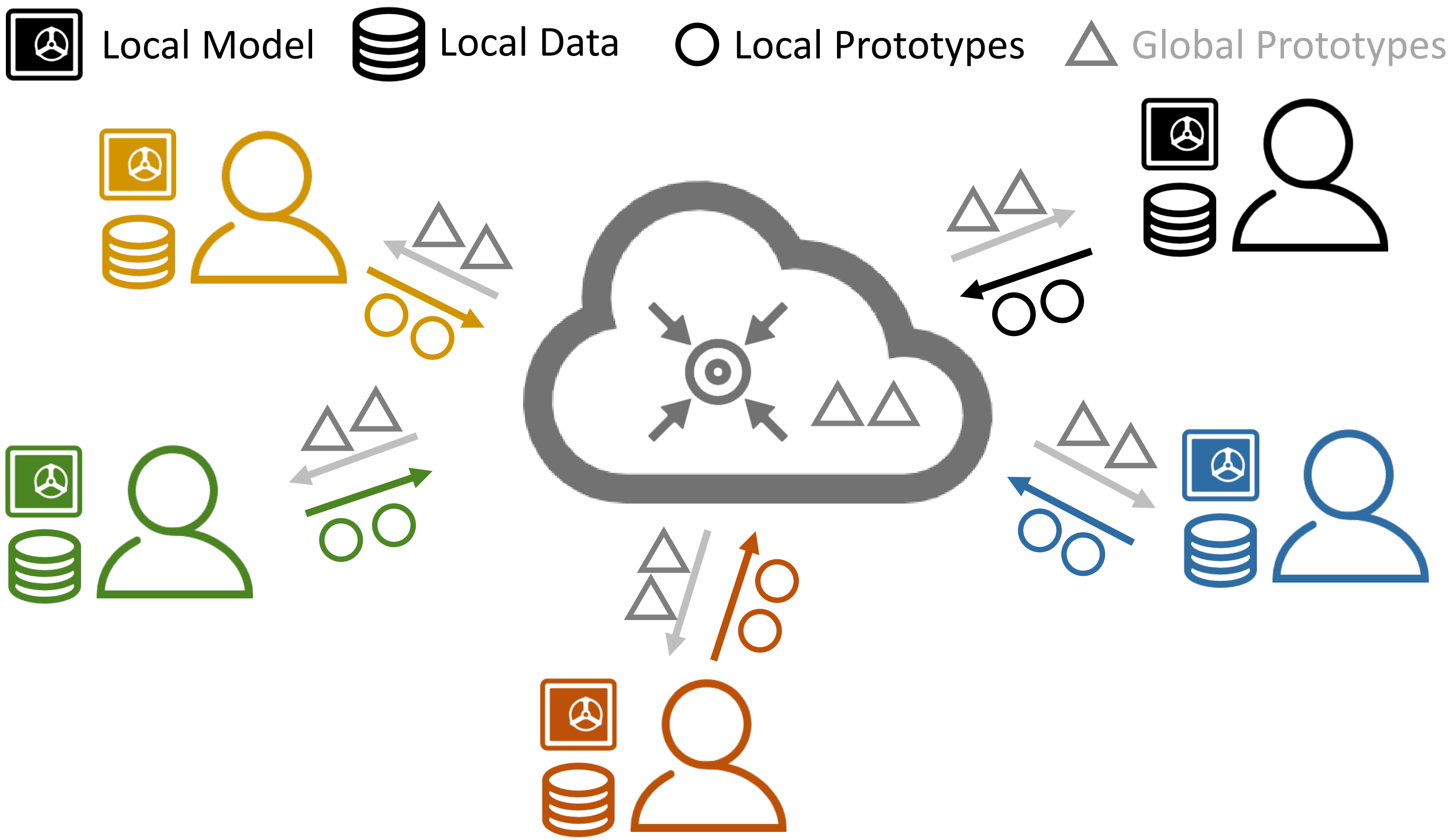
Issues of FedProto

- A typical method that is *applicable* in HtFL



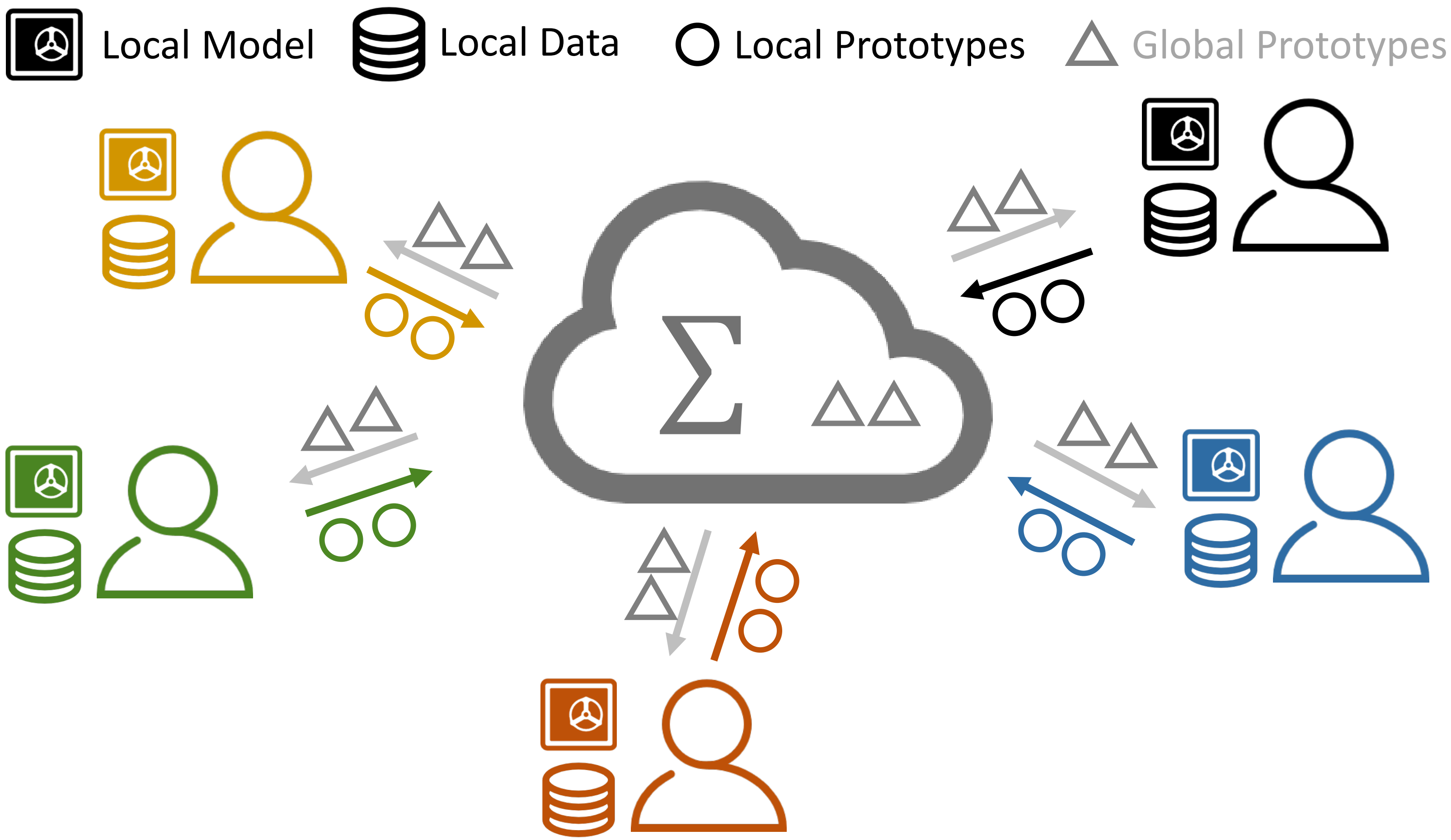
Issues of FedProto

- Transmit *lightweight prototypes (i.e., class representatives)* instead of model parameters



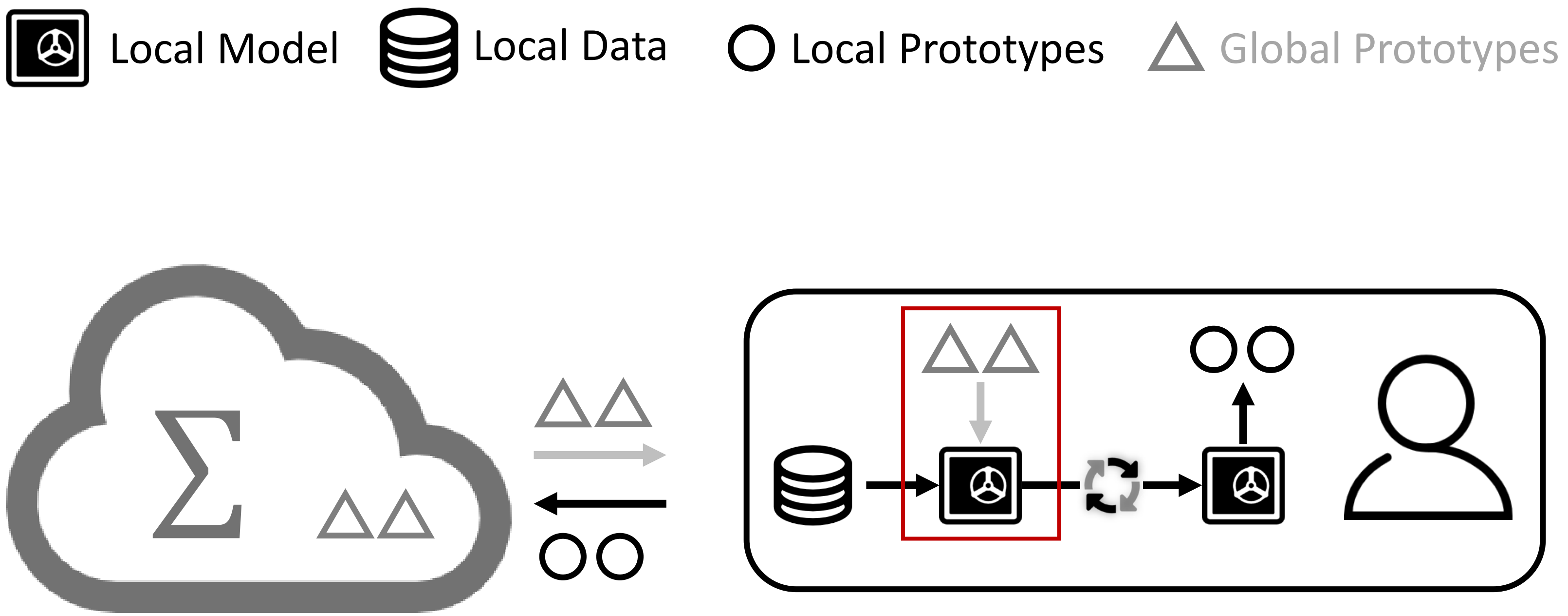
Issues of FedProto

- Obtain *global prototypes* through *weighted-averaging*



Issues of FedProto

- Guide local training with *global prototypes*



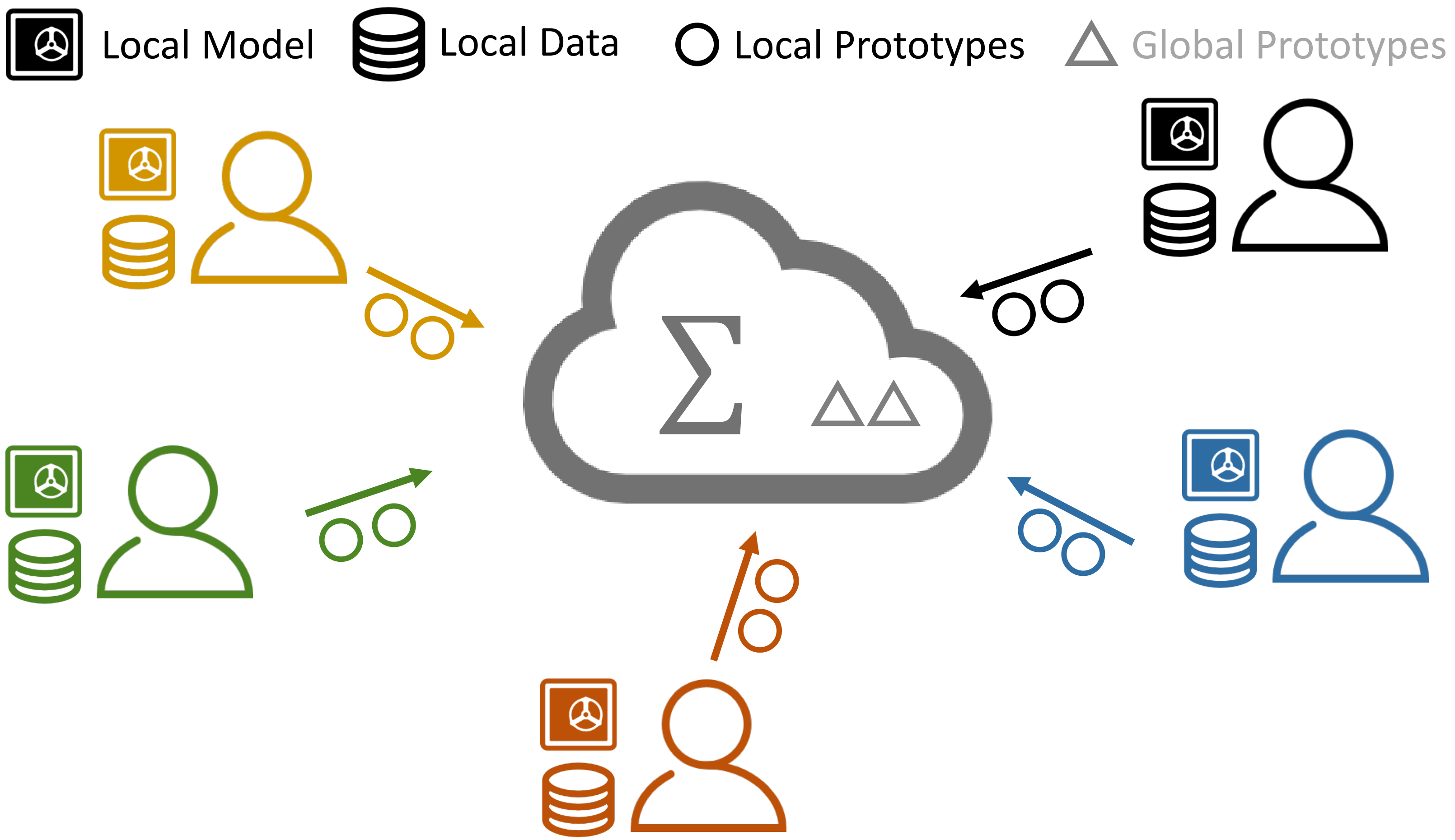
Issues of FedProto

- However, *weighted-averaging is not suitable* in HtFL as

Local Model Local Data Local Prototypes Global Prototypes

Issues of FedProto

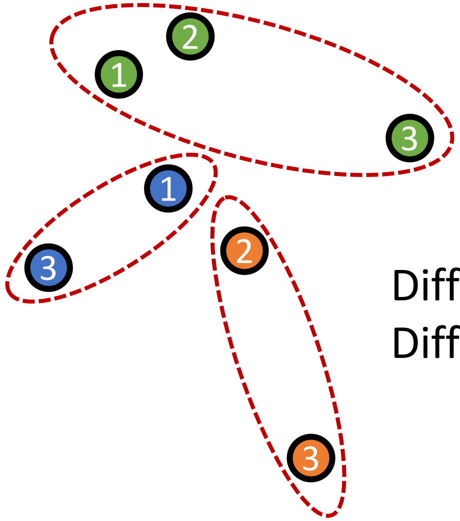
- However, *weighted-averaging is not suitable* in HtFL as
- Prototypes are generated on heterogeneous datasets and models



Issues of FedProto

- However, *weighted-averaging is not suitable* in HtFL as
- Prototypes of the same class but from different clients may not cluster together

Should be

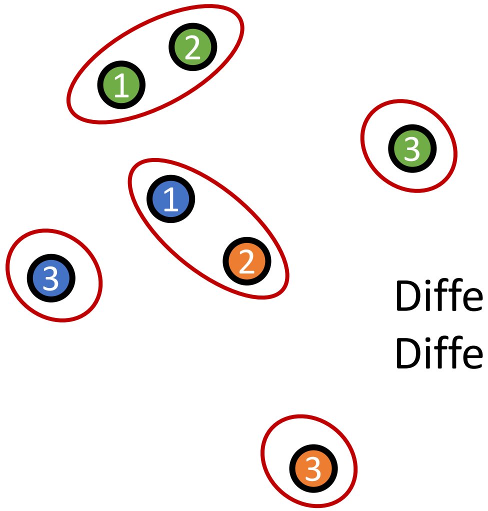


Different colors: classes
Different numbers: clients

Issues of FedProto

- However, *weighted-averaging is not suitable* in HtFL as
- Prototypes of the same class but from different clients may not cluster together

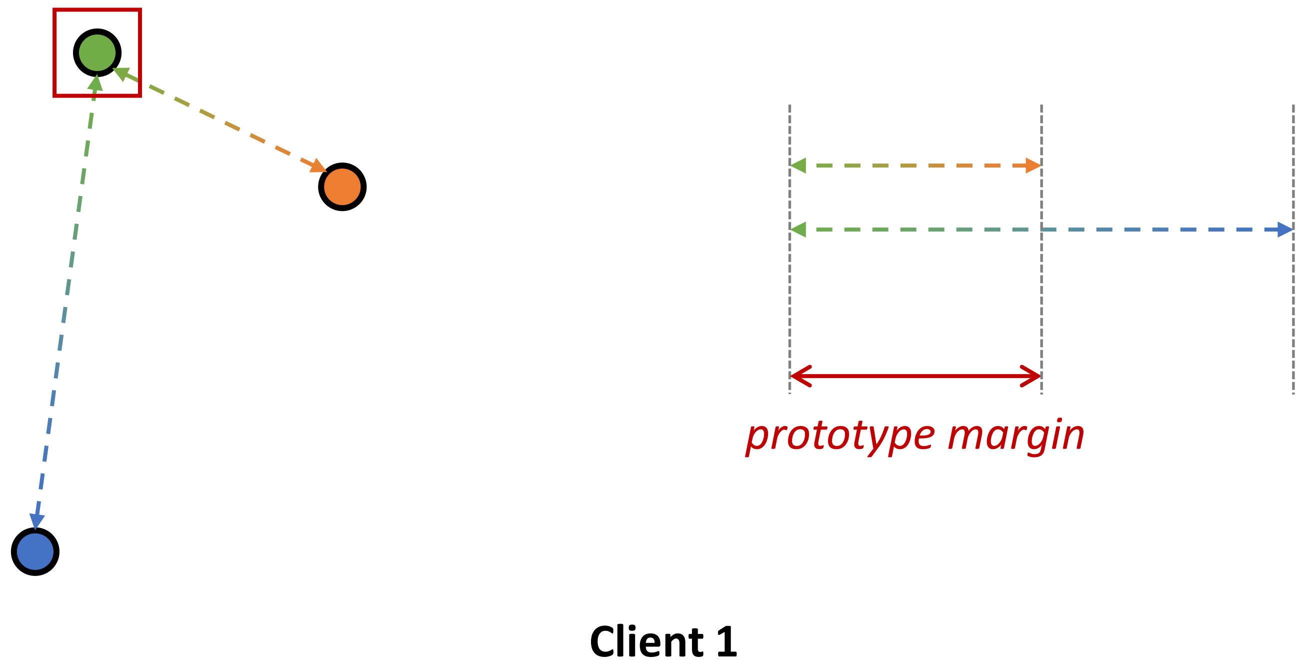
But



Different colors: classes
Different numbers: clients

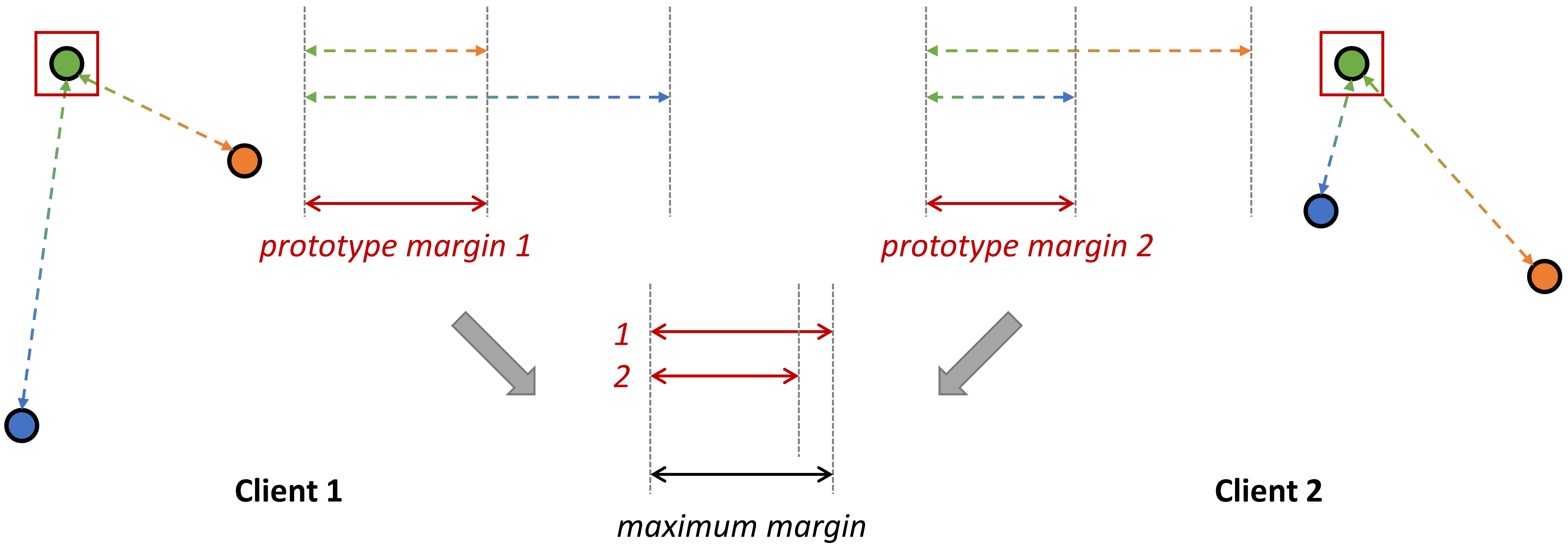
Issues of FedProto

- Define: *prototype margin* is the *minimum* Euclidean distance between the prototype of a specific class and the prototypes of other classes
- Consider three prototypes on a client:



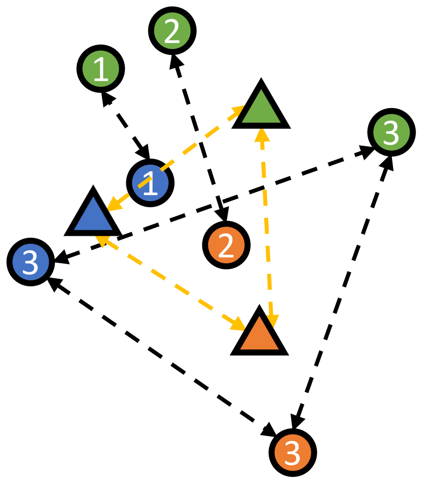
Issues of FedProto

- Define: *maximum margin* is the maximum prototype margin among all clients for each class
- Consider two clients:



Issues of FedProto

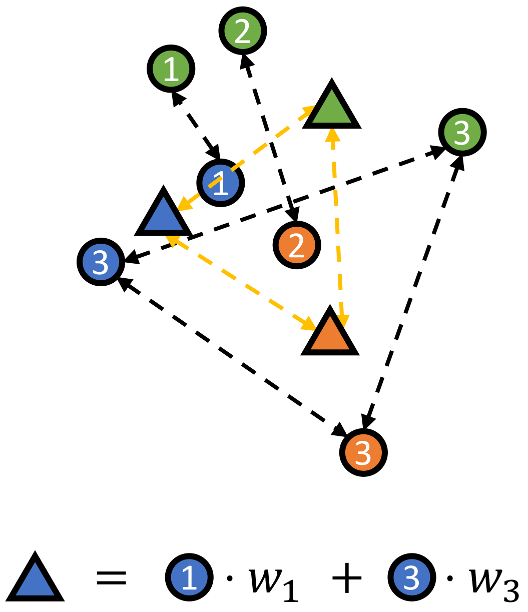
- Global prototype (△) margin *shrinks* after weighted-averaging

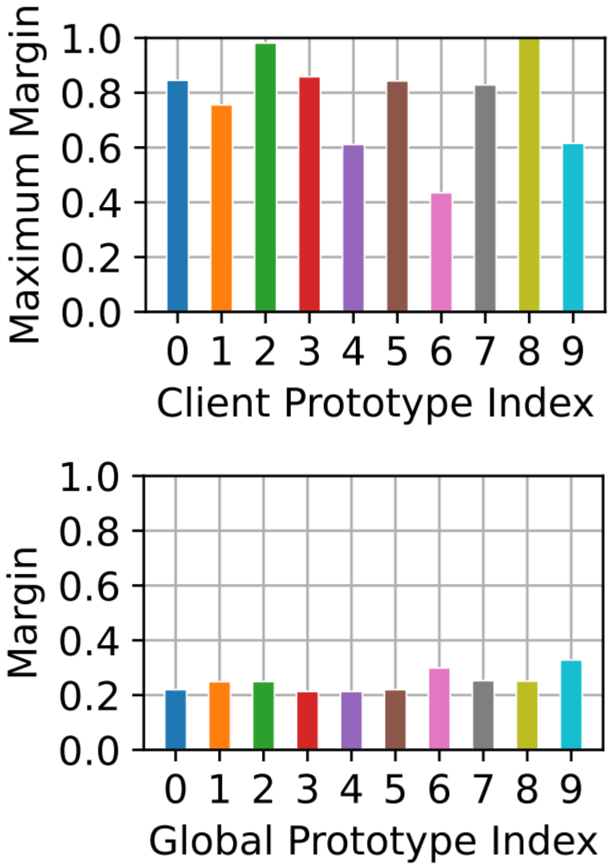


$$\blacktriangle = \textcircled{1} \cdot w_1 + \textcircled{3} \cdot w_3$$

Issues of FedProto

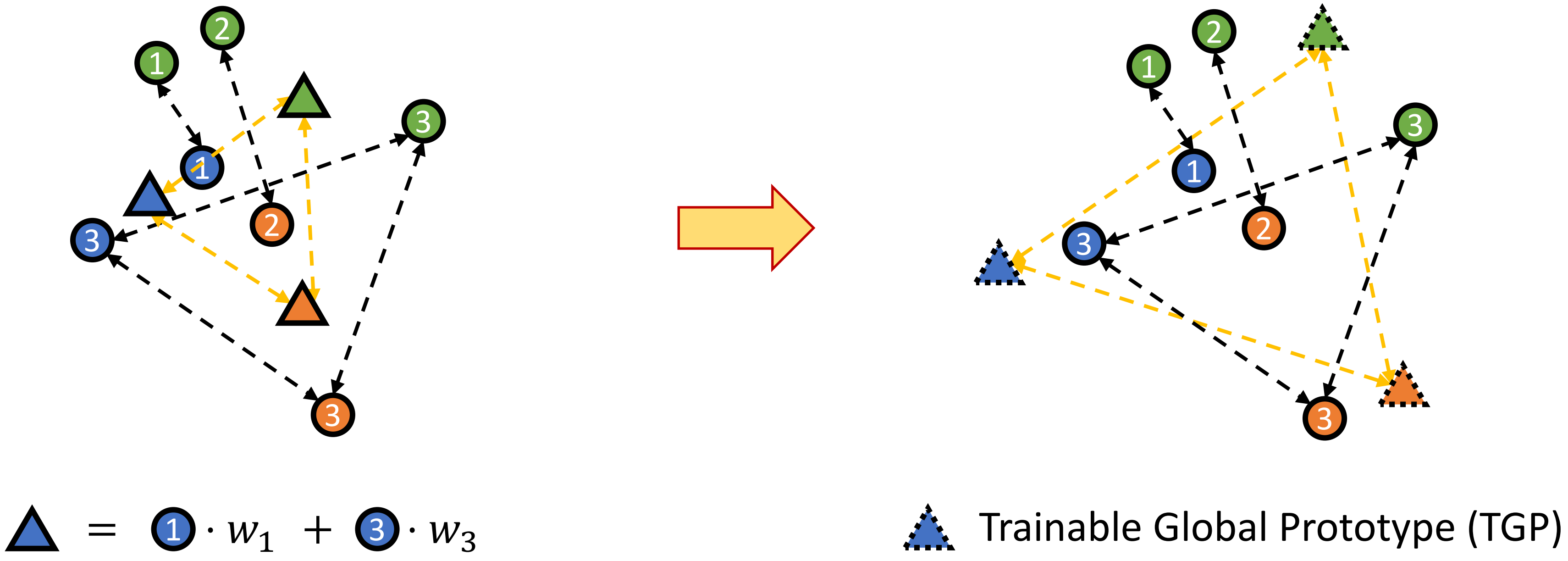
- Global prototype (△) margin *shrinks* after weighted-averaging





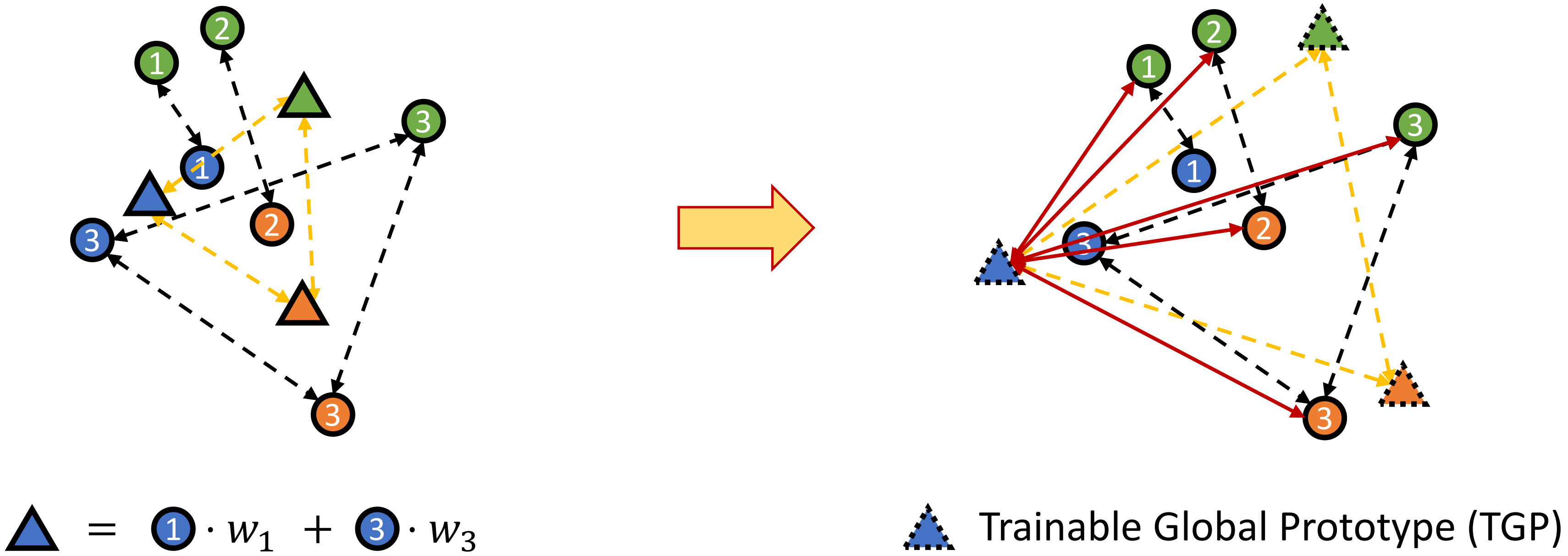
Our FedTGP: Trainable Global Prototypes

- Remove weighted-averaging
- Consider the uploaded client prototypes as data
- *Enlarge* the global prototype margin



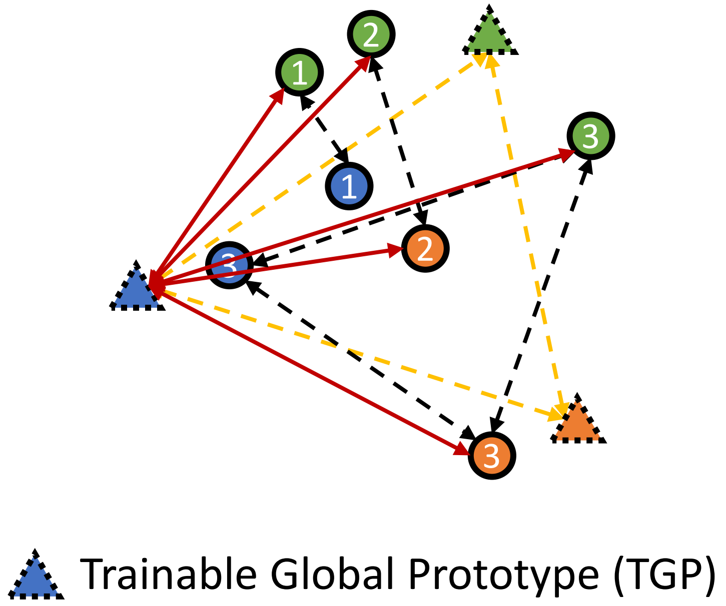
Our FedTGP: Trainable Global Prototypes

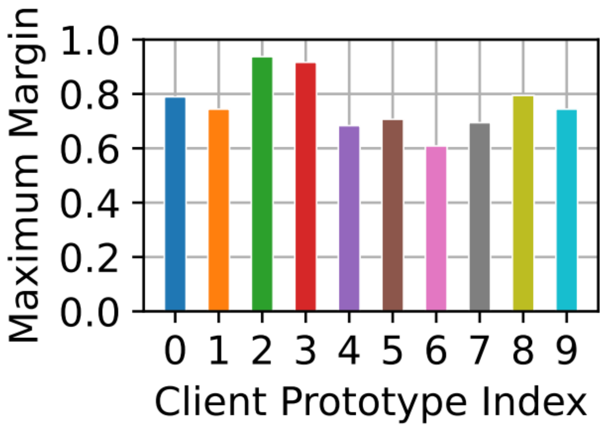
- Remove weighted-averaging
- Consider the uploaded client prototypes as data
- *Enlarge* the global prototype margin

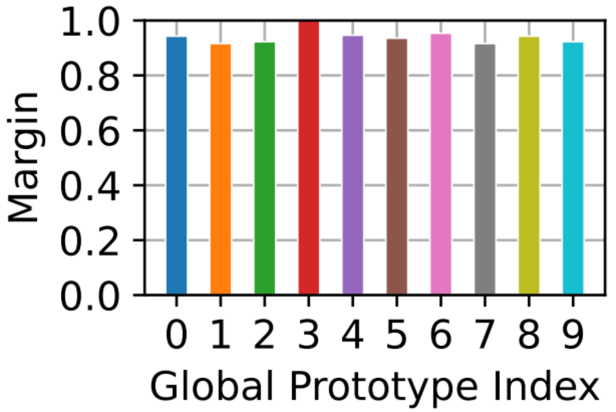


Our FedTGP: Trainable Global Prototypes

- Remove weighted-averaging
- Consider the uploaded client prototypes as data
- *Enlarge* the global prototype margin







Trainable Global Prototypes

- Server objective: train TGP using *contrastive learning*

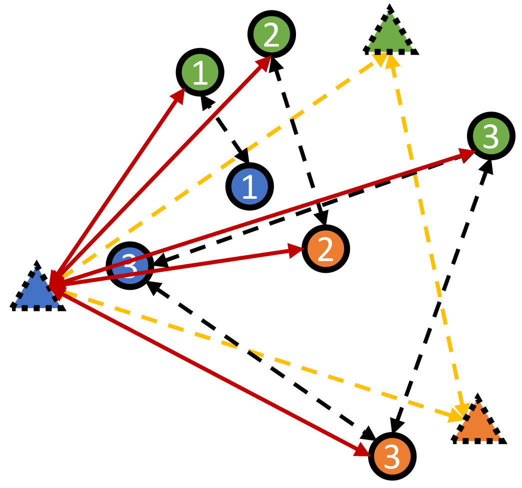
$$\min_{\hat{\mathcal{P}}} \sum_{c=1}^{C} \mathcal{L}_P^c,$$

$$\boxed{\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} - \log \frac{e^{-\phi(P_i^c, \hat{P}^c)}}{e^{-\phi(P_i^c, \hat{P}^c)} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}}},$$

$\hat{\mathcal{P}} = \{\hat{P}^c\}_{c=1}^C$

ϕ measures the Euclidean distance

$\mathcal{I}^t$ is the participating client set



$\hat{P}^c$: A TGP of class c

$\hat{\mathcal{P}}$: All TGP

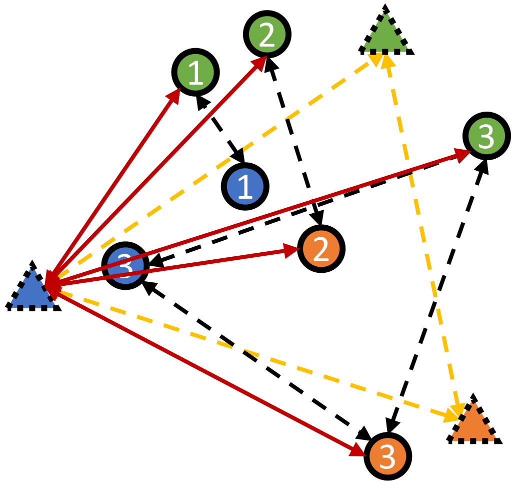
P_i^c: A prototype of class c from client i

Trainable Global Prototypes

- Server objective: train TGP using *margin-enhanced contrastive learning*

$$\min_{\hat{\mathcal{P}}} \sum_{c=1}^{C} \mathcal{L}_P^c,$$

$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} - \log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \boxed{\delta})}}{e^{-(\phi(P_i^c, \hat{P}^c) + \boxed{\delta})} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}},$$



$\hat{P}^c$: A TGP of class c

$\hat{\mathcal{P}}$: All TGP

P_i^c: A prototype of class c from client i

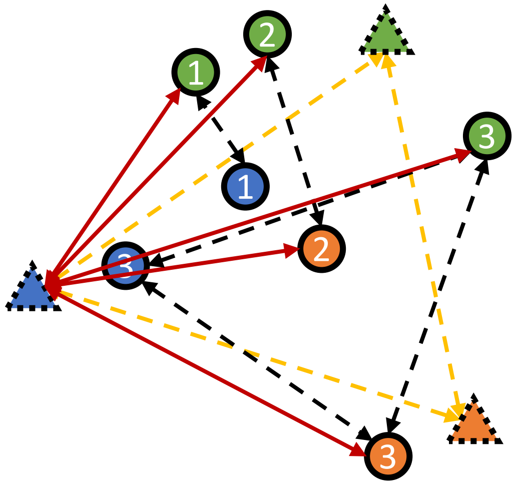
Trainable Global Prototypes

- Server objective: train TGP using *margin-enhanced contrastive learning*

$$\min_{\hat{\mathcal{P}}} \sum_{c=1}^{C} \mathcal{L}_P^c,$$

$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} - \log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \boxed{\delta})}}{e^{-(\phi(P_i^c, \hat{P}^c) + \boxed{\delta})} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}},$$

$$\mathcal{L}_P^c \propto \tilde{\mathcal{L}}_P^c := \sum_{i \in \mathcal{I}^t} \sum_{c'} e^{\phi(P_i^c, \hat{P}^c) - \phi(P_i^c, \hat{P}^{c'}) + \delta},$$



$\hat{P}^c$: A TGP of class c

$\hat{\mathcal{P}}$: All TGP

P_i^c: A prototype of class c from client i

Trainable Global Prototypes

- Server objective: train TGP using *Adaptive-margin-enhanced Contrastive Learning (ACL)*

$$\min_{\hat{\mathcal{P}}} \sum_{c=1}^{C} \mathcal{L}_P^c,$$

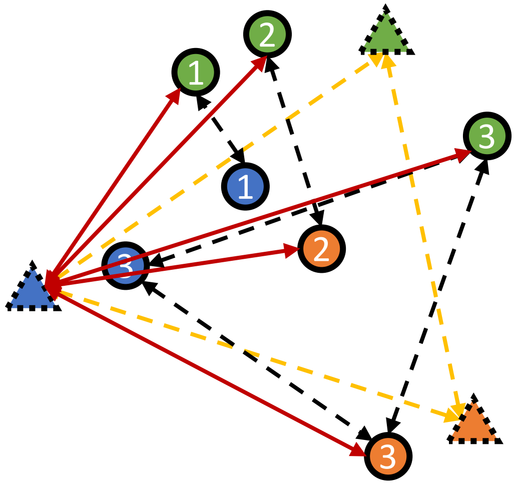
$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} -\log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))}}{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}}$$

$$\delta(t) = \min(\max_{c \in [C], c' \in [C], c \neq c'} \phi(Q_t^c, Q_t^{c'}), \tau),$$

$$Q_t^c = \frac{1}{|\mathcal{P}_t^c|} \sum_{i \in \mathcal{I}^t} P_i^c, \forall c \in [C]$$

maximum cluster margin

τ is a margin threshold



$\hat{P}^c$: A TGP of class c

$\hat{\mathcal{P}}$: All TGP

P_i^c: A prototype of class c from client i

Experiments

- We evaluate FedTGP on scenarios with 12 widely-used model architectures on 20 $\sim$ 100 clients

Table 8: The forward FLOPs of all architectures“B” is short for billion.

	FLOPs	References
4-layer CNN	0.013B	None
GoogleNet	1.530B	Chang et al. (2023); Lin et al. (2022)
MobileNet_v2	0.314B	Zhao et al. (2022)
ResNet4	—	Zhong et al. (2017)
ResNet6	—	Zhong et al. (2017)
ResNet8	—	Zhong et al. (2017)
ResNet10	—	Zhong et al. (2017)
ResNet18	0.117B	Zhao and Wang (2022)
ResNet34	0.218B	Zhao and Wang (2022)
ResNet50	1.305B	Li et al. (2022a)
ResNet101	2.532B	Li et al. (2022a); Leroux et al. (2018)
ResNet152	5.330B	Bakhtiarnia, Zhang, and Iosifidis (2022)

Experiments

- Only 6 baselines are available in our *data-free scenarios (no additional public data)*
- Our FedTGP surpasses counterparts by up to **9.08%**

Table 1: The test accuracy (%) on four datasets in the pathological and practical settings using the $HtFE_8$ model group.

Settings	Pathological Setting				Practical Setting			
Datasets	Cifar10	Cifar100	Flowers102	Tiny-ImageNet	Cifar10	Cifar100	Flowers102	Tiny-ImageNet
LG-FedAvg	86.82±0.26	57.01±0.66	58.88±0.28	32.04±0.17	84.55±0.51	40.65±0.07	45.93±0.48	24.06±0.10
FedGen	82.83±0.65	58.26±0.36	59.90±0.15	29.80±1.11	82.55±0.49	38.73±0.14	45.30±0.17	19.60±0.08
FML	87.06±0.24	55.15±0.14	57.79±0.31	31.38±0.15	85.88±0.08	39.86±0.25	46.08±0.53	24.25±0.14
FedKD	87.32±0.31	56.56±0.27	54.82±0.35	32.64±0.36	86.45±0.10	40.56±0.31	48.52±0.28	25.51±0.35
FedDistill	87.24±0.06	56.99±0.27	58.51±0.34	31.49±0.38	86.01±0.31	41.54±0.08	49.13±0.85	24.87±0.31
FedProto	83.39±0.15	53.59±0.29	55.13±0.17	29.28±0.36	82.07±1.64	36.34±0.28	41.21±0.22	19.01±0.10
FedTGP	**90.02±0.30**	**61.86±0.30**	**68.98±0.43**	**34.56±0.27**	**88.15±0.43**	**46.94±0.12**	**53.68±0.31**	**27.37±0.12**

Experiments

- Our FedTGP also surpasses FedProto by up to **13.85%**

Table 1: The test accuracy (%) on four datasets in the pathological and practical settings using the HtFE_8 model group.

Settings	Pathological Setting				Practical Setting			
Datasets	Cifar10	Cifar100	Flowers102	Tiny-ImageNet	Cifar10	Cifar100	Flowers102	Tiny-ImageNet
LG-FedAvg	86.82±0.26	57.01±0.66	58.88±0.28	32.04±0.17	84.55±0.51	40.65±0.07	45.93±0.48	24.06±0.10
FedGen	82.83±0.65	58.26±0.36	59.90±0.15	29.80±1.11	82.55±0.49	38.73±0.14	45.30±0.17	19.60±0.08
FML	87.06±0.24	55.15±0.14	57.79±0.31	31.38±0.15	85.88±0.08	39.86±0.25	46.08±0.53	24.25±0.14
FedKD	87.32±0.31	56.56±0.27	54.82±0.35	32.64±0.36	86.45±0.10	40.56±0.31	48.52±0.28	25.51±0.35
FedDistill	87.24±0.06	56.99±0.27	58.51±0.34	31.49±0.38	86.01±0.31	41.54±0.08	49.13±0.85	24.87±0.31
FedProto	83.39±0.15	53.59±0.29	55.13±0.17	29.28±0.36	82.07±1.64	36.34±0.28	41.21±0.22	19.01±0.10
FedTGP	**90.02±0.30**	**61.86±0.30**	**68.98±0.43**	**34.56±0.27**	**88.15±0.43**	**46.94±0.12**	**53.68±0.31**	**27.37±0.12**

Experiments

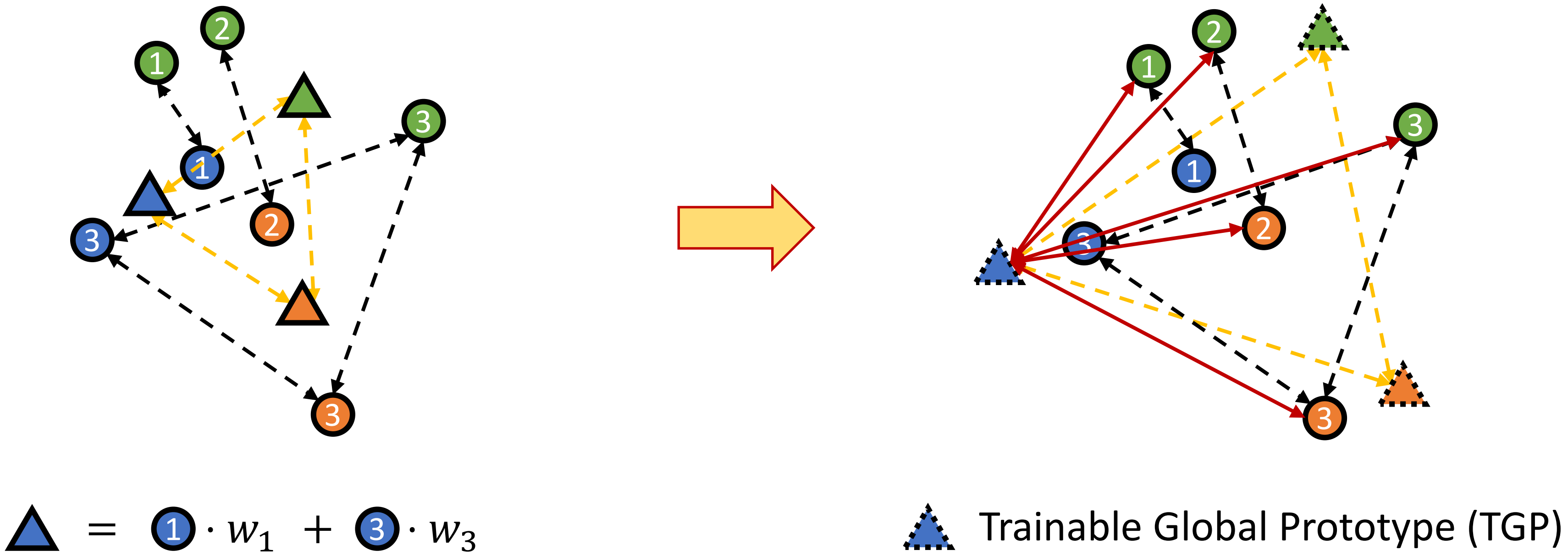
- We study the influence of increasing model heterogeneity
- Our FedTGP is more resilient and less impacted by model heterogeneity

Table 2: The test accuracy (%) on Cifar100 in the practical setting using heterogeneous feature extractors.

Settings	Heterogeneous Feature Extractors			
	$HtFE_2$	$HtFE_3$	$HtFE_4$	$HtFE_9$
LG-FedAvg	46.61±0.24	45.56±0.37	43.91±0.16	42.04±0.26
FedGen	43.92±0.11	43.65±0.43	40.47±1.09	40.28±0.54
FML	45.94±0.16	43.05±0.06	43.00±0.08	42.41±0.28
FedKD	46.33±0.24	43.16±0.49	43.21±0.37	42.15±0.36
FedDistill	46.88±0.13	43.53±0.21	43.56±0.14	42.09±0.20
FedProto	43.97±0.18	38.14±0.64	34.67±0.55	32.74±0.82
FedTGP	**49.82±0.29**	**49.65±0.37**	**46.54±0.14**	**48.05±0.19**

Take away

- We enhance the typical heterogeneous federated learning method FedProto with TGP and ACL,
- ***Making it more versatile and resilient to various model heterogeneities***



FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning

Paper with code: https://github.com/TsingZ0/FedTGP

E-mail: tsingz@sjtu.edu.cn

Paper with code

Thanks!