DEVICE HETEROGENEITY IN FEDERATED LEARNING
A SUPERQUANTILE APPROACH

JOURNEES DES STATISTIQUES 2021

Yassine LAGUEL[★] — Joint work with K. Pillutla[▲], J. Malick[◆] and Z. Harchaoui[▲]

[★]Université Grenoble Alpes - [◆]CNRS - [▲]University of Washington

Collaboration with

CNRS

J. MALICK

University of Washington

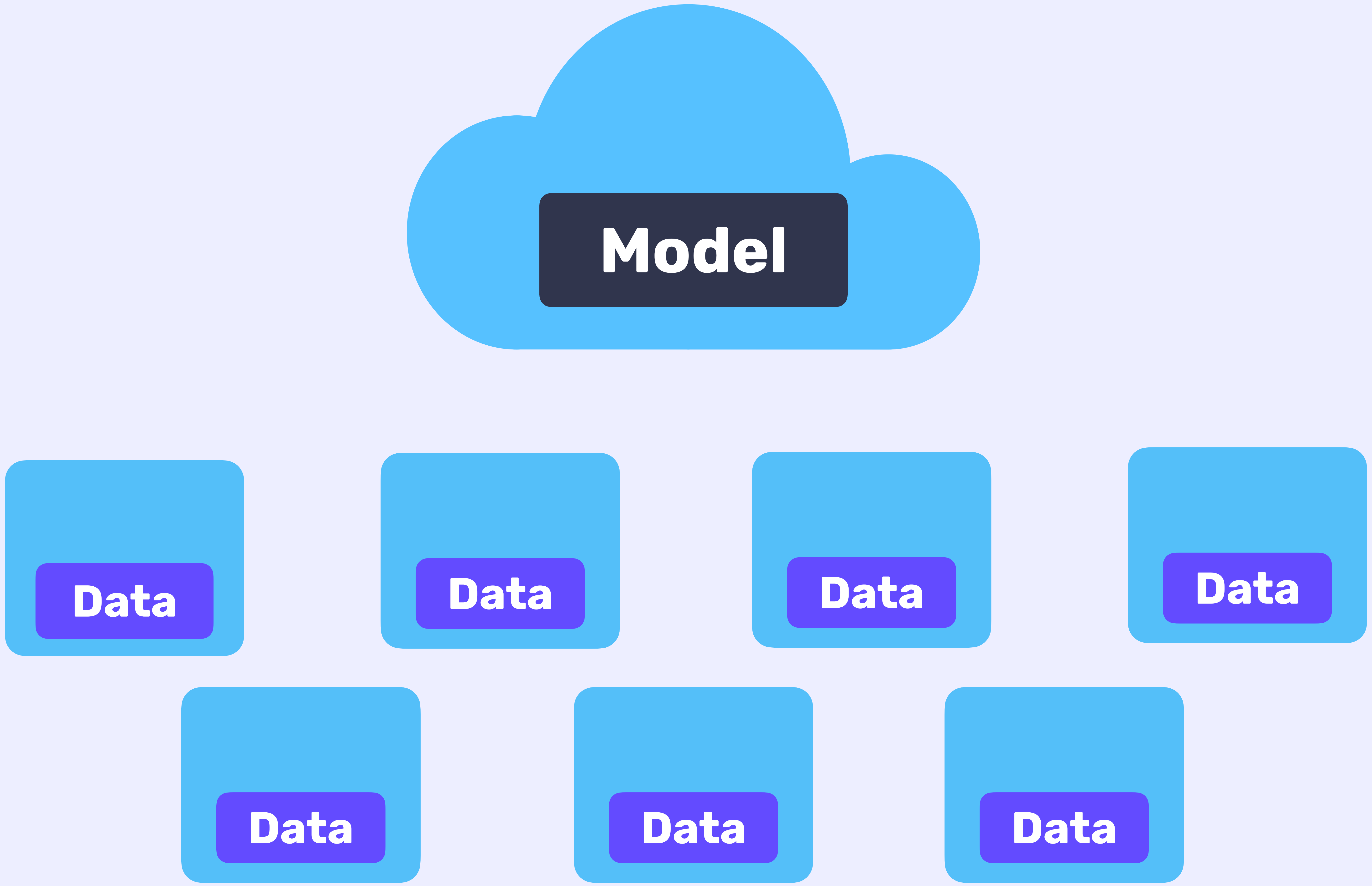
K. PILLUTLA

University of Washington

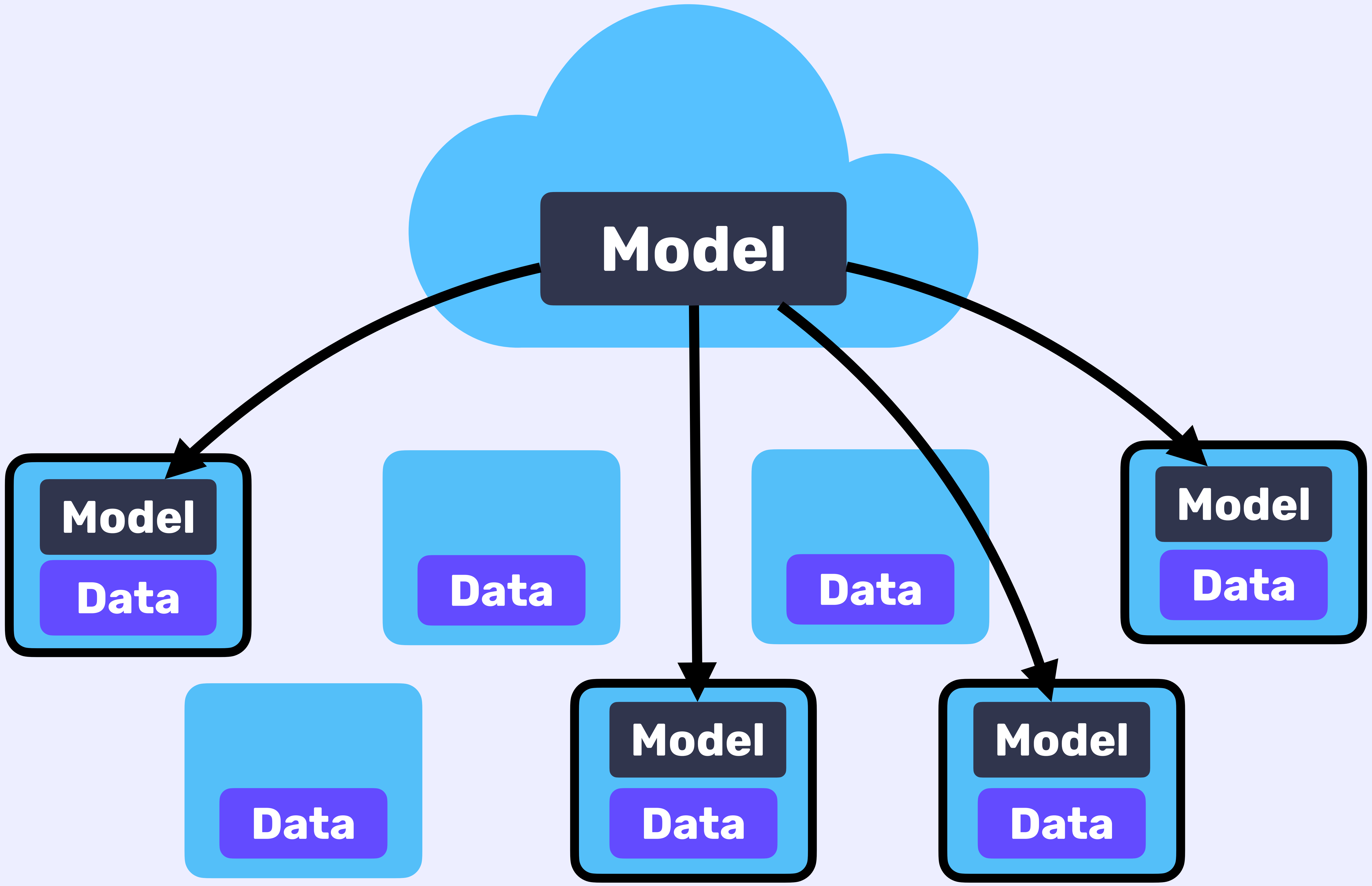
Z. HARCHAOUI

FEDERATED LEARNING IN A NUTSHELL

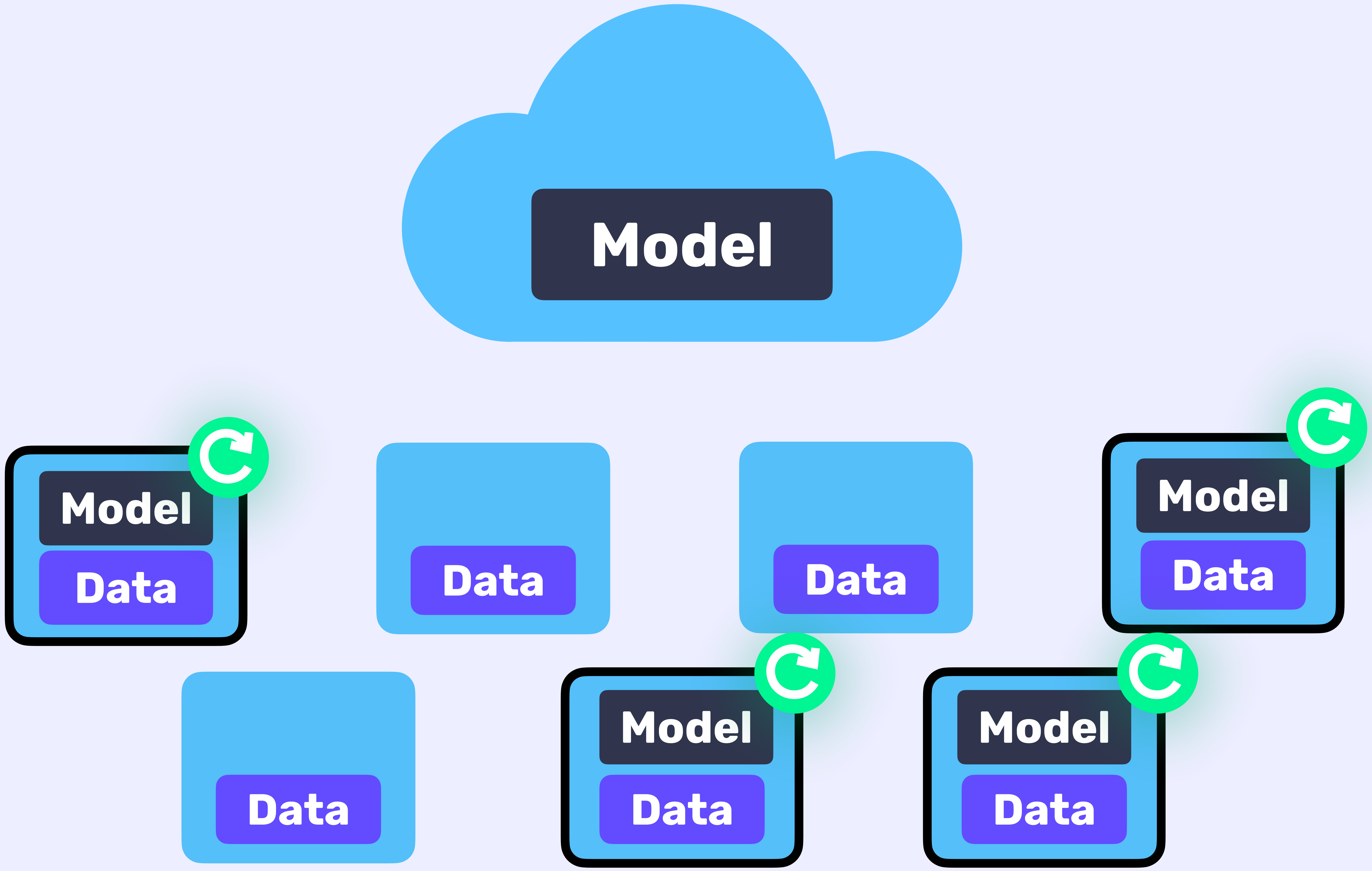
FEDERATED LEARNING IN A NUTSHELL



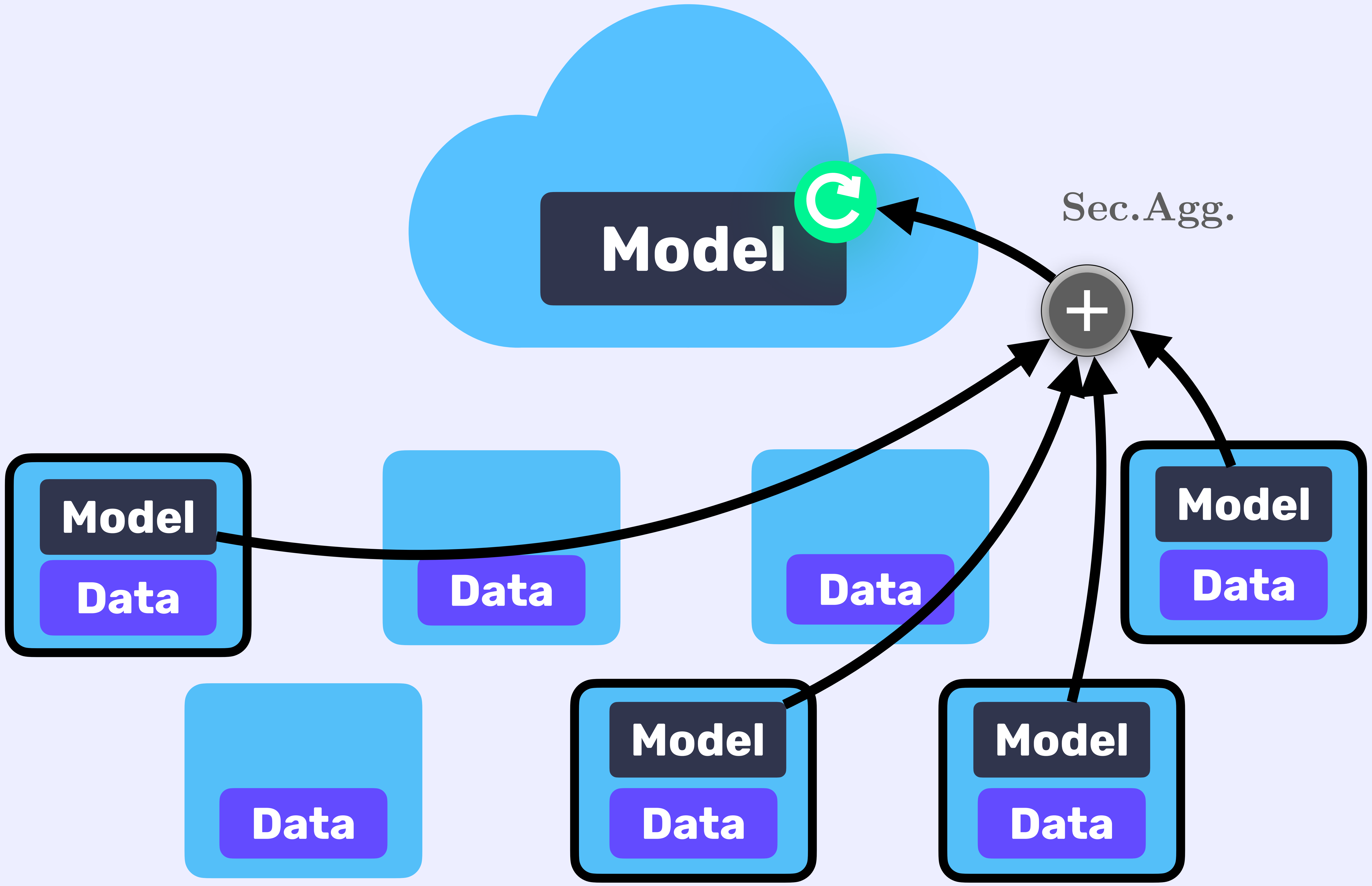
FEDERATED LEARNING IN A NUTSHELL



FEDERATED LEARNING IN A NUTSHELL



FEDERATED LEARNING IN A NUTSHELL



CHALLENGES

- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

Privacy preservation

Statistical heterogeneity

System heterogeneity

Communication costs

Improve performance on heterogenous users

CHALLENGES

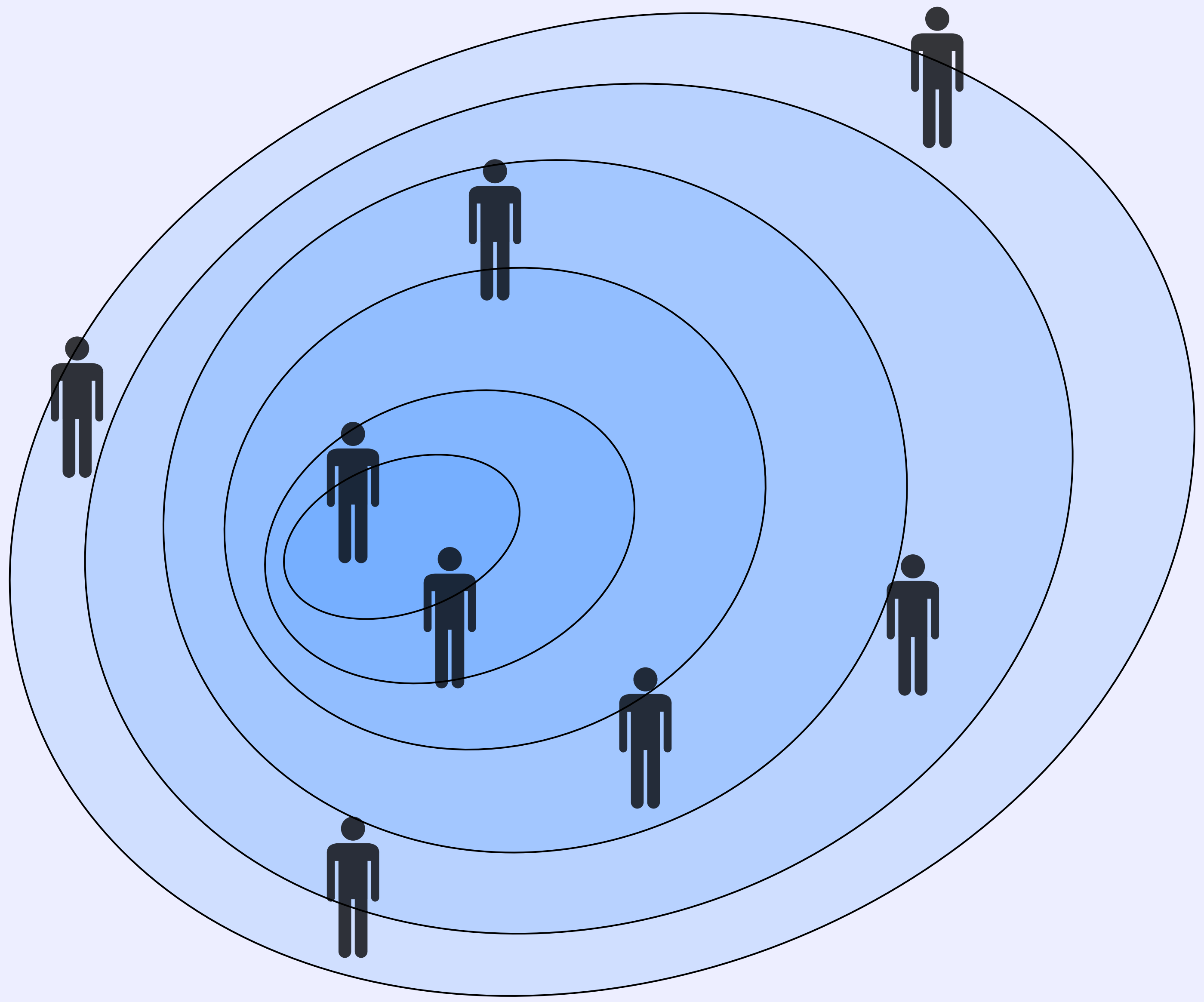
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

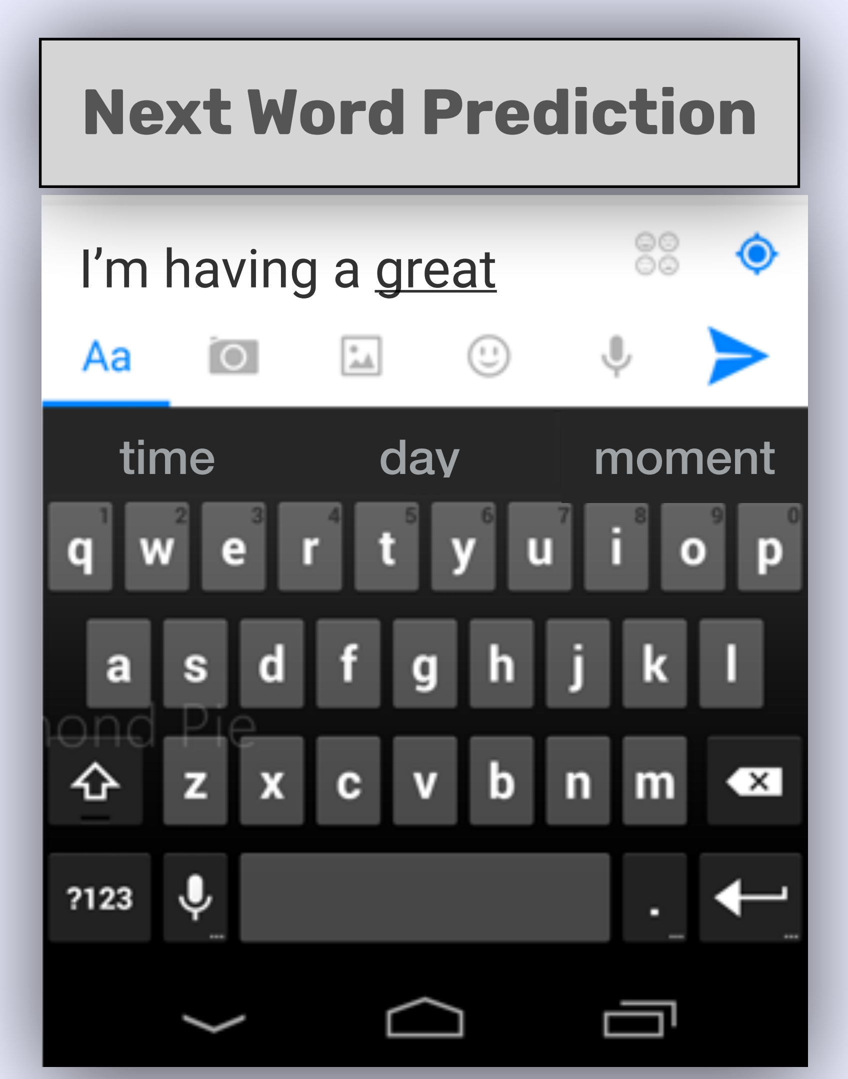
Privacy preservation | Statistical heterogeneity | System heterogeneity | Communication costs

Improve performance on heterogenous users

- Users heterogeneity



- Eg. on mobile phones



CHALLENGES

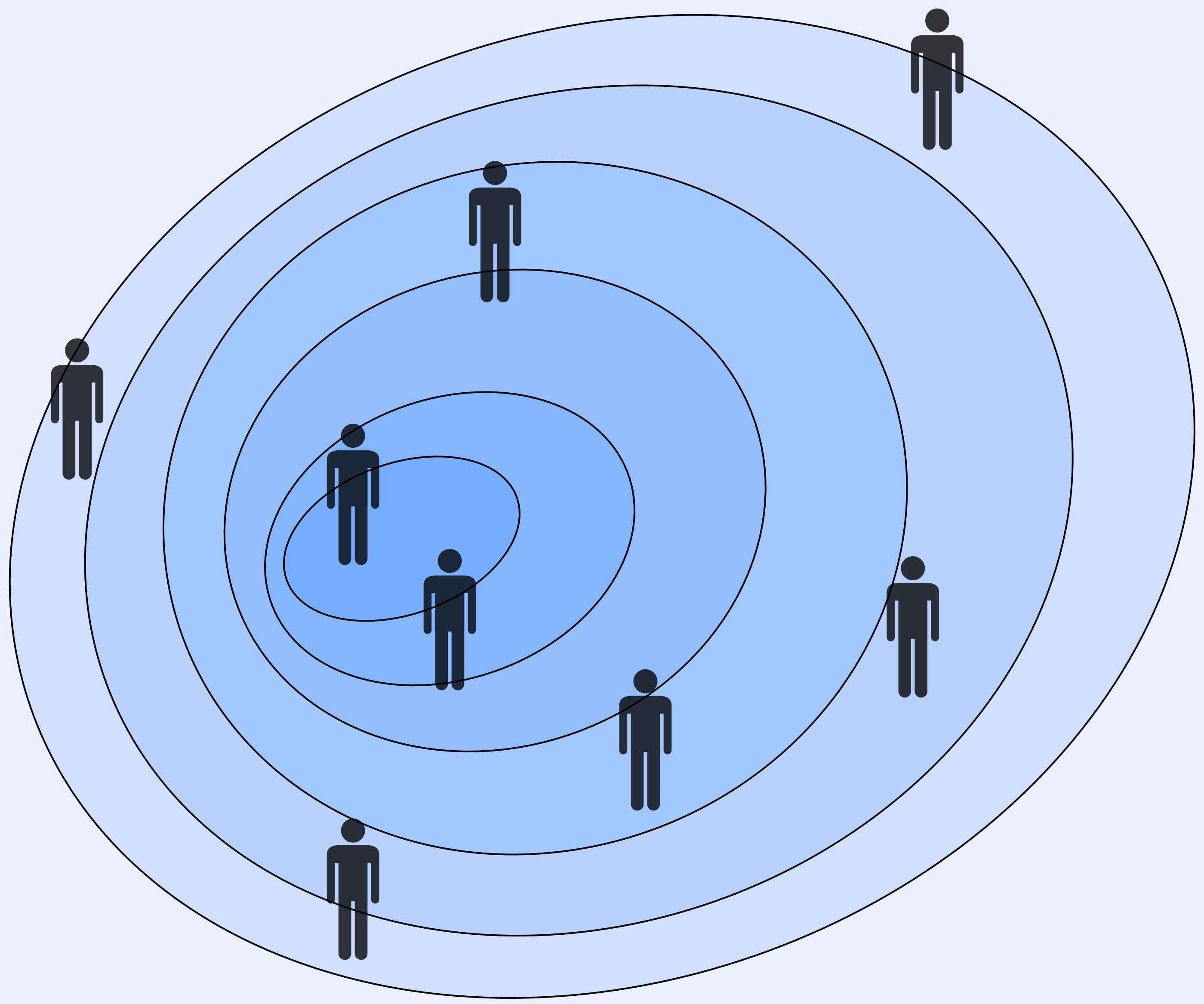
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

Privacy preservation | Statistical heterogeneity | System heterogeneity | Communication costs

Improve performance on heterogenous users

- Users heterogeneity



- Vanilla Federated Learning
  - FedAvg's objective

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \alpha_i \; F_i(w)$$

CHALLENGES

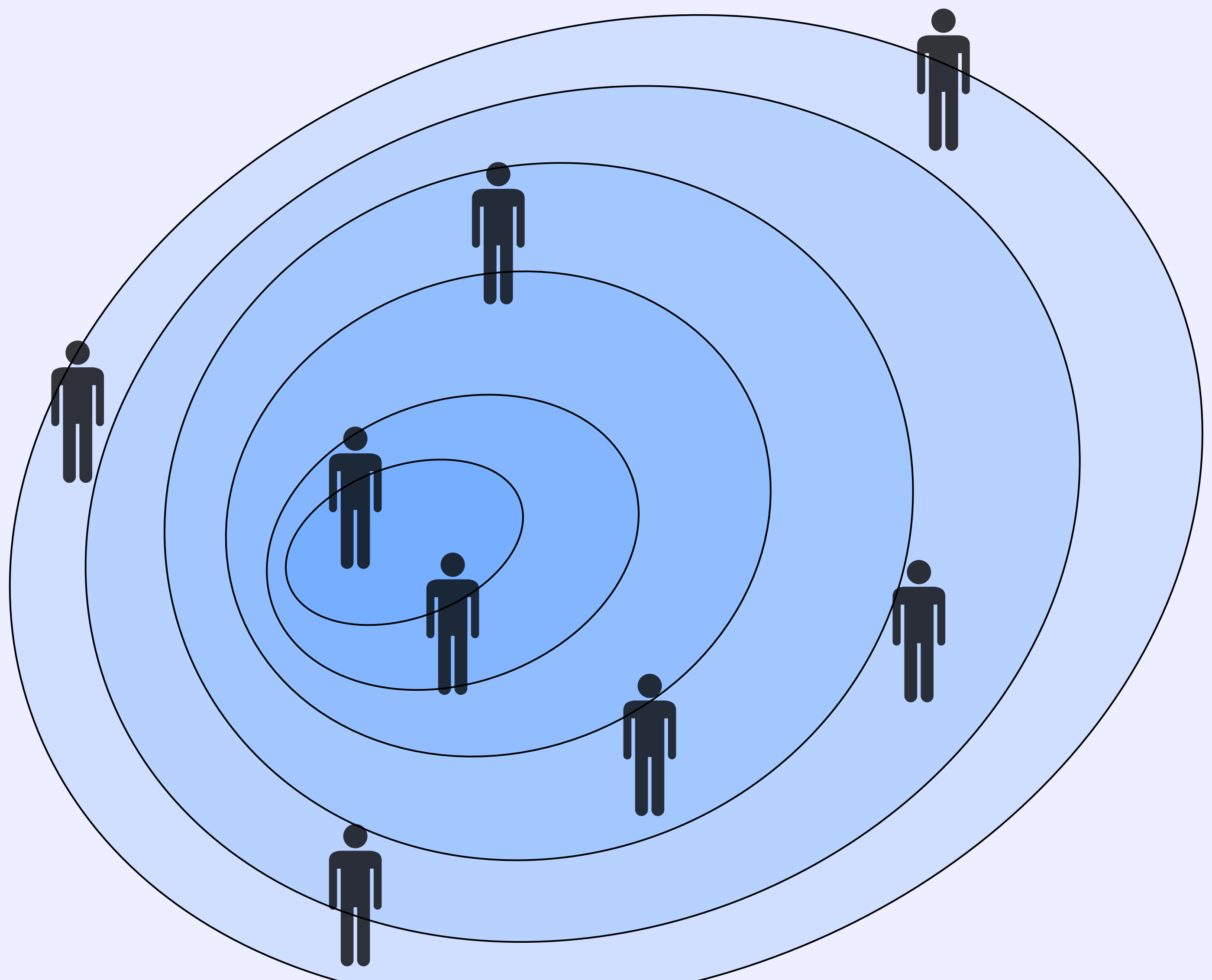
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

Privacy preservation | Statistical heterogeneity | System heterogeneity | Communication costs

Improve performance on heterogenous users

- Users heterogeneity



- Vanilla Federated Learning

  - FedAvg's objective

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \alpha_i \; F_i(w) \qquad F_i(w) = \mathbb{E}_{\xi \sim q_i}[f(w, \xi)]$$

Data distribution of device i

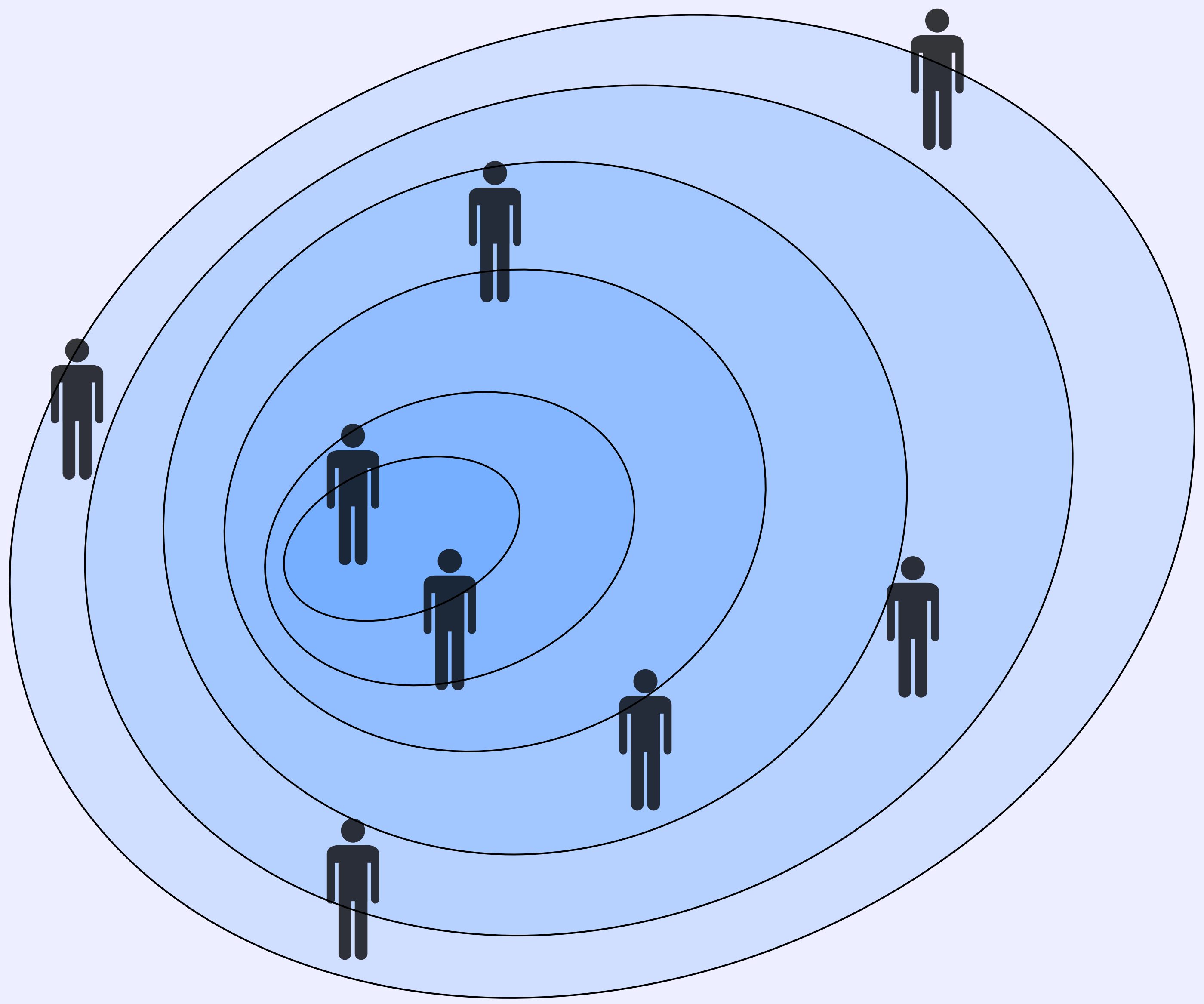
CHALLENGES

Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

- Privacy preservation
- Statistical heterogeneity
  - **Improve performance on heterogenous users**
- System heterogeneity
- Communication costs

Users heterogeneity



Vanilla Federated Learning

- FedAvg's objective

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \alpha_i \; F_i(w) \qquad F_{i}(w) = \mathbb{E}_{\xi \sim q_i}[f(w, \xi)]$$

Data distribution of device i

- FedAvg learns the trend

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{\xi \sim p_\alpha}[f(w, \xi)] \qquad p_\alpha = \sum_{i=1}^{N} \alpha_i \; q_i$$

CHALLENGES

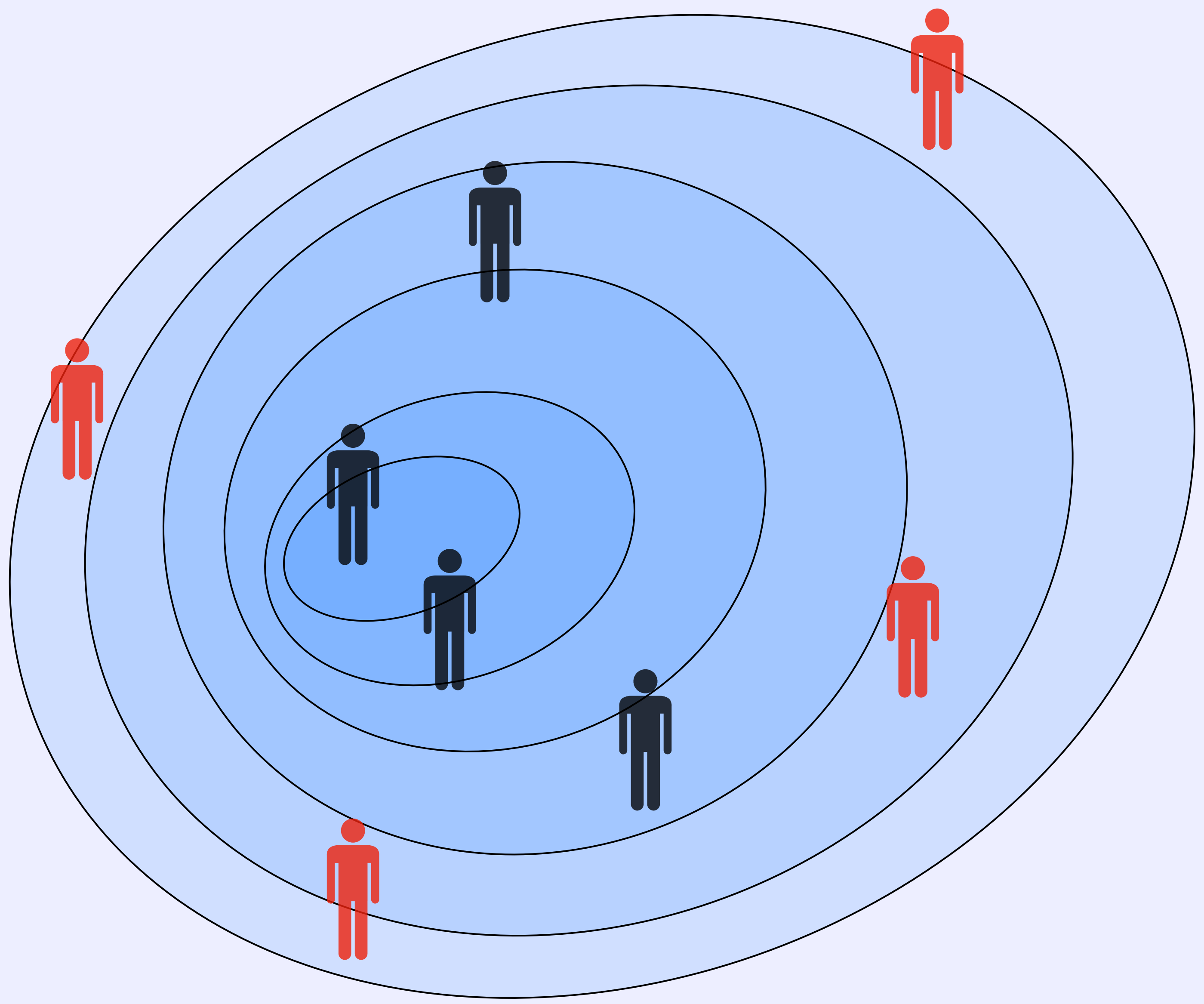
■ Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

Privacy preservation | Statistical heterogeneity | System heterogeneity | Communication costs

Improve performance on heterogenous users

■ Users heterogeneity



■ Vanilla Federated Learning

- FedAvg's objective

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \alpha_i \, F_i(w) \qquad F_i(w) = \mathbb{E}_{\xi \sim q_i}[f(w, \xi)]$$

Data distribution of device i

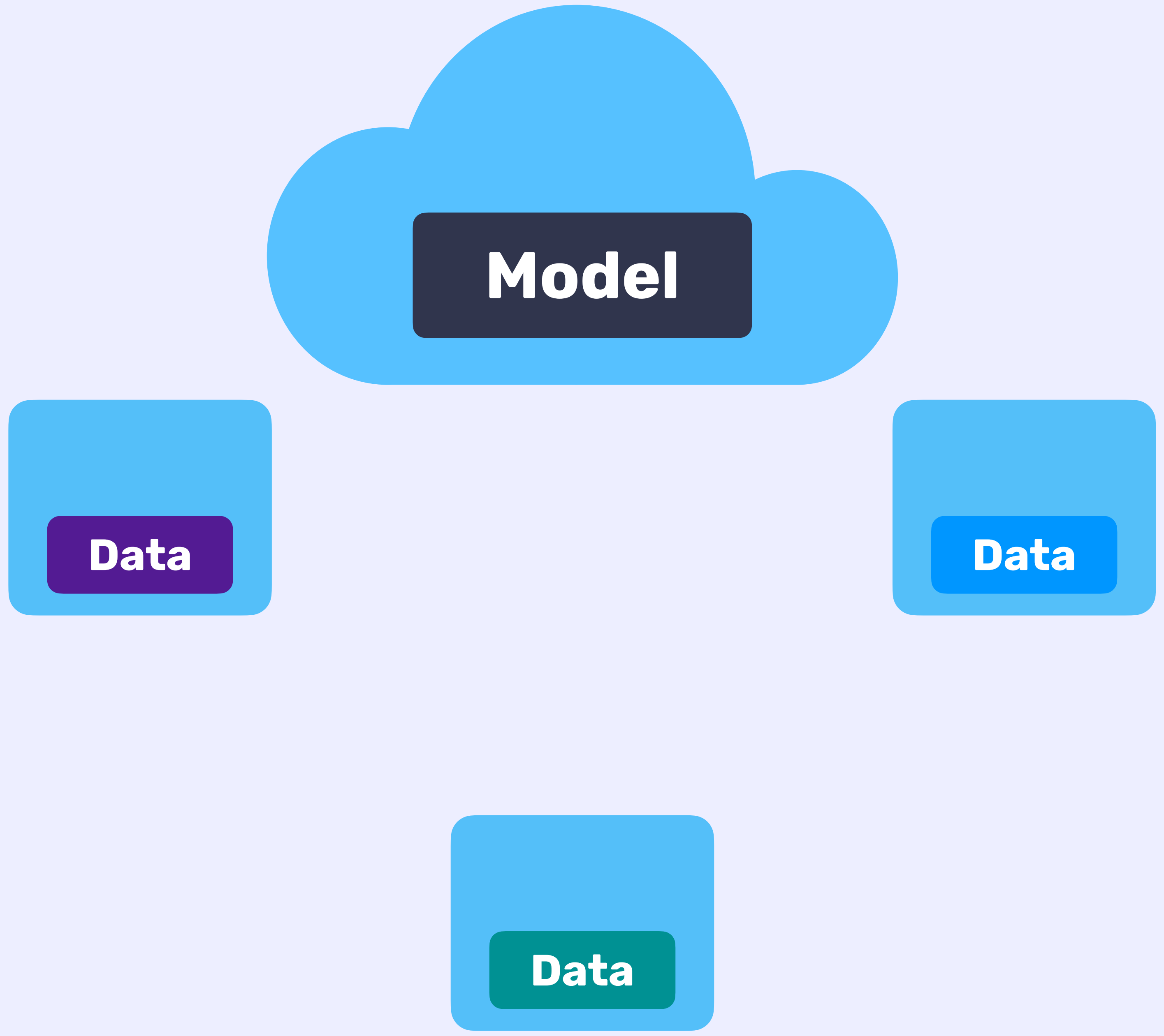
- FedAvg learns the trend

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{\xi \sim p_\alpha}[f(w, \xi)] \qquad p_\alpha = \sum_{i=1}^{N} \alpha_i \, q_i$$

Measuring Conformity in Federated Learning

- Modeling Heterogeneity on training devices
  - We dispose of N training devices.
  - Each training device is characterized by a distribution q_i over some data space and a weight $\alpha_i > 0$ such that $\sum_{i=1}^{N} \alpha_i = 1$

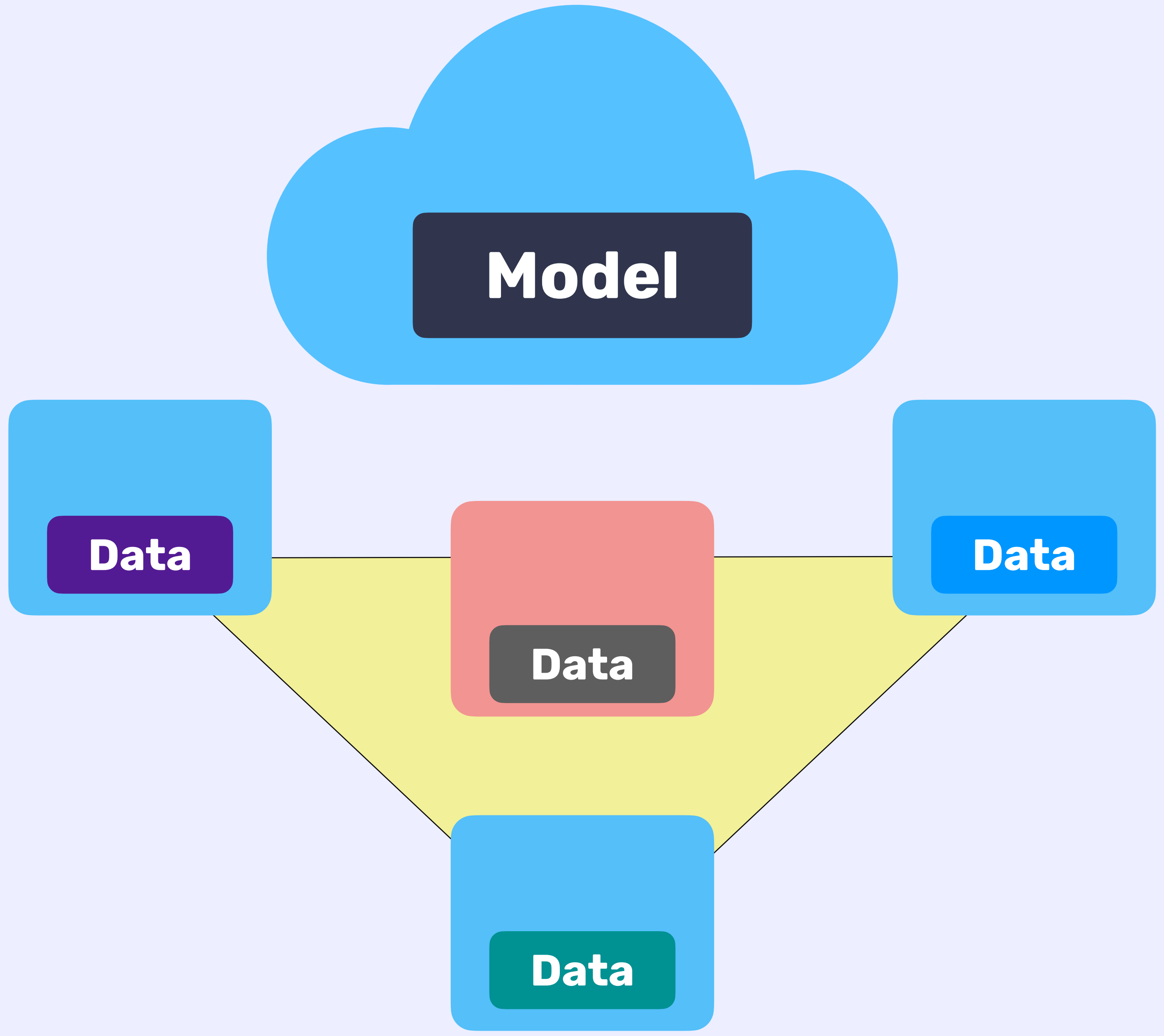
Base distribution $\quad p_\alpha = \sum_{i=1}^{N} \alpha_i \, q_i$



Measuring Conformity in Federated Learning

- Modeling Heterogeneity on training devices
  - We dispose of N training devices.
  - Each training device is characterized by a distribution q_i over some data space and a weight $\alpha_i > 0$ such that $\sum_{i=1}^{N} \alpha_i = 1$

<u>Base distribution</u> $\quad p_\alpha = \sum_{i=1}^{N} \alpha_i \, q_i$



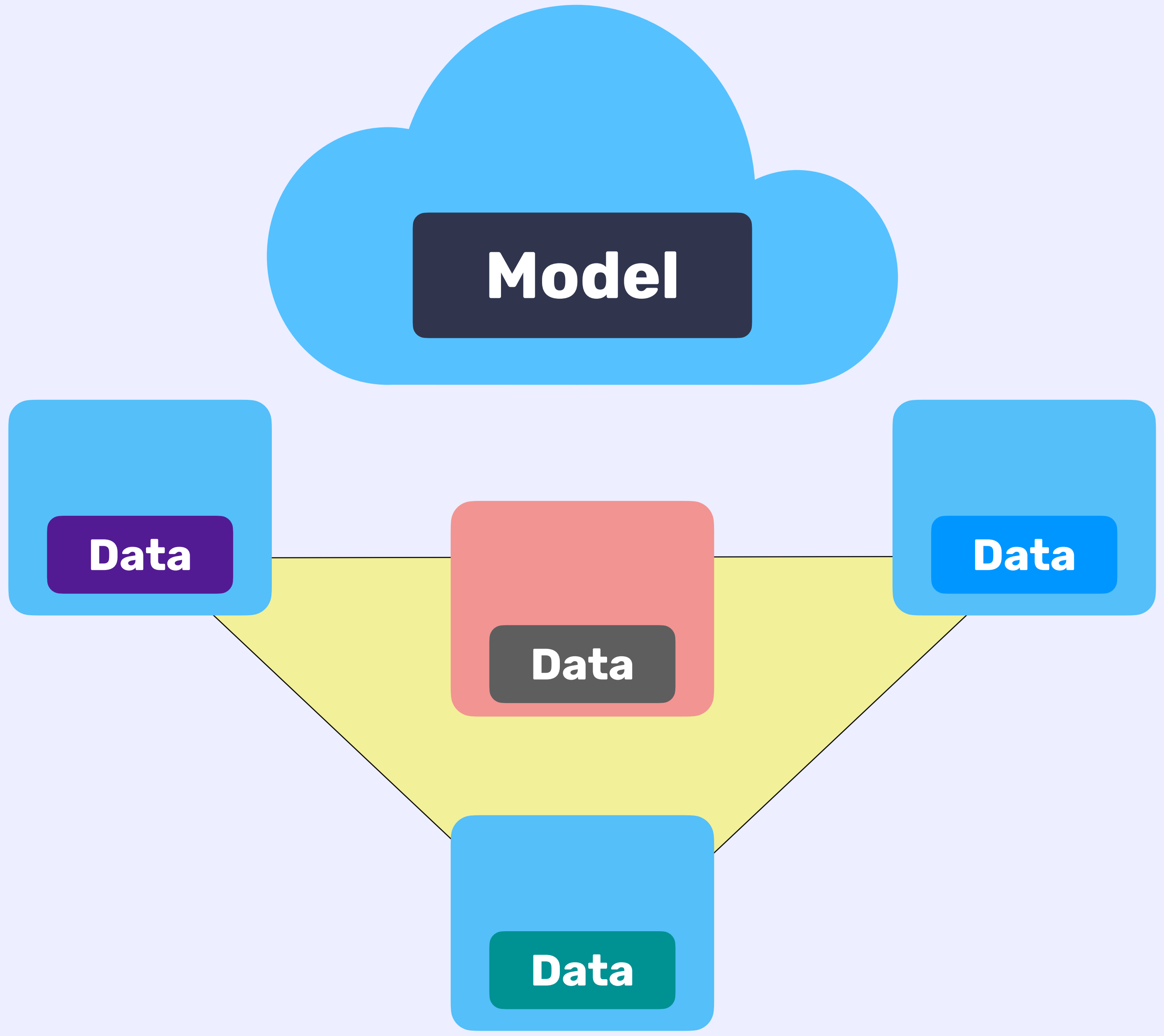
- Measuring conformity on testing devices
  - We consider test devices to have a distribution that can be written as a mixture of the training distributions.

$$p_\pi = \sum_{i=1}^{N} \pi_i \alpha_i \qquad \pi \in \Delta_{N-1} \text{ ie } \begin{cases} 0 \leq \pi_k \leq 1 & \text{for all } 1 \leq i \leq N \\ \sum_{k=1}^{N} \pi_k = 1 \end{cases}$$

Measuring Conformity in Federated Learning

- Modeling Heterogeneity on training devices
  - We dispose of N training devices.
  - Each training device is characterized by a distribution q_i over some data space and a weight $\alpha_i > 0$ such that $\sum_{i=1}^{N} \alpha_i = 1$

  Base distribution $\quad p_\alpha = \sum_{i=1}^{N} \alpha_i \, q_i$



- Measuring conformity on testing devices
  - We consider test devices to have a distribution that can be written as a mixture of the training distributions.

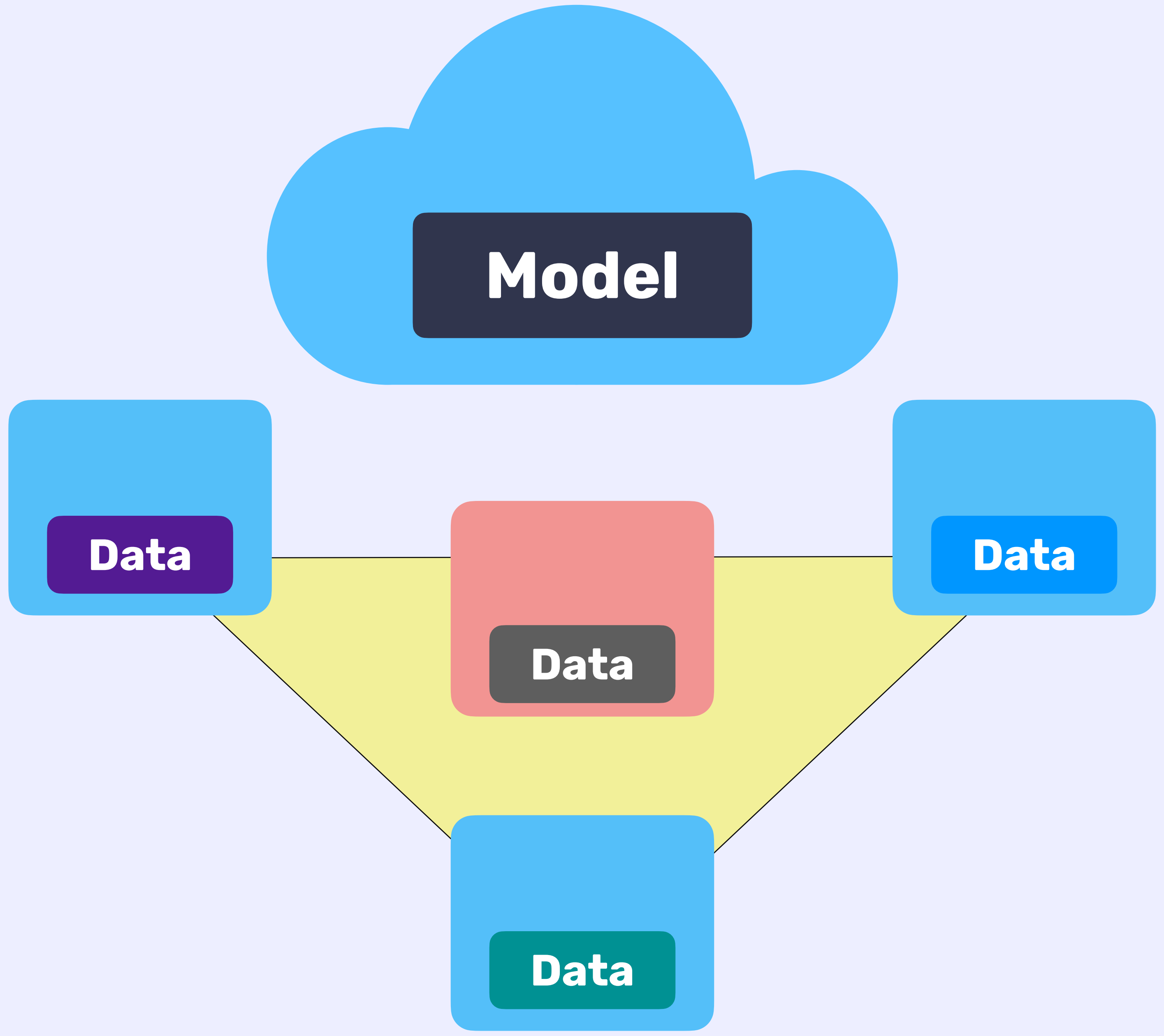
$$p_\pi = \sum_{i=1}^{N} \pi_i \alpha_i \qquad \pi \in \Delta_{N-1} \text{ ie } \begin{cases} 0 \leq \pi_k \leq 1 & \text{for all } 1 \leq i \leq N \\ \sum_{k=1}^{N} \pi_k = 1 \end{cases}$$

  - The conformity $\operatorname{conf}(p_\pi) \in [0, 1]$ of a mixture p_π with weight π is defined as:

$$\operatorname{conf}(p_\pi) = \min_{i \in \{1, \ldots, N\}} \alpha_i / \pi_i$$

  The conformity of a device refers to the conformity of its data distribution.

The Δ-FL Framework

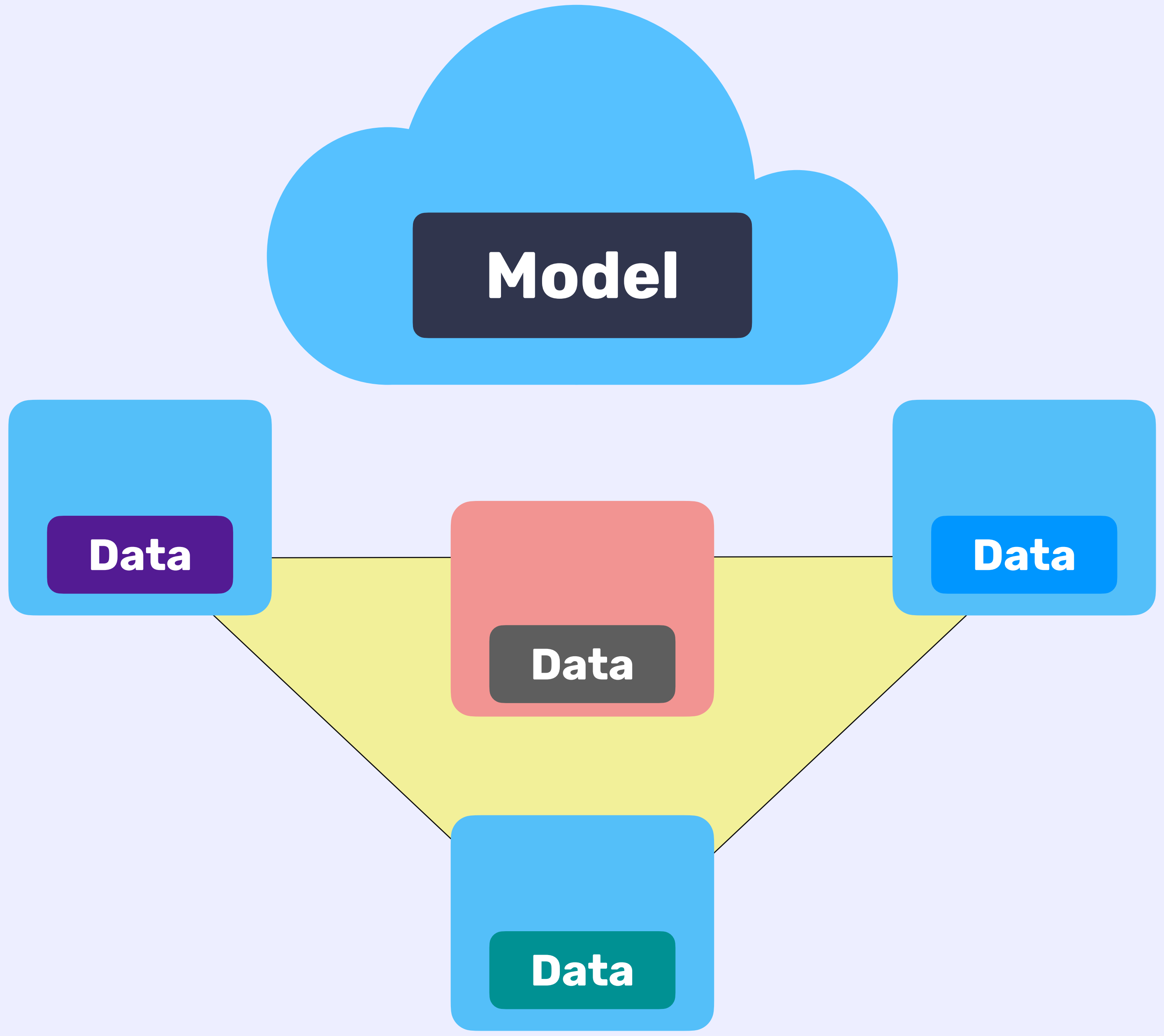


The Δ-FL Framework

- Δ-FL's Objective
  - We propose to solve for a conformity parameter.$\theta \in (0, 1]$:

$$\min_{w \in \mathbb{R}^d} \left[F_\theta(w) = \max_{\pi \in \mathcal{P}_\theta} \mathbb{E}_{\xi \sim p_\pi}[f(w, \xi)] \right] \quad \text{where}$$

$$\mathcal{P}_\theta := \{\pi \in \Delta_{N-1} : \mathrm{conf}(p_\pi) \geq \theta\}$$



The Δ-FL Framework

- Δ-FL's Objective
  - We propose to solve for a conformity parameter. $\theta \in (0, 1]$:

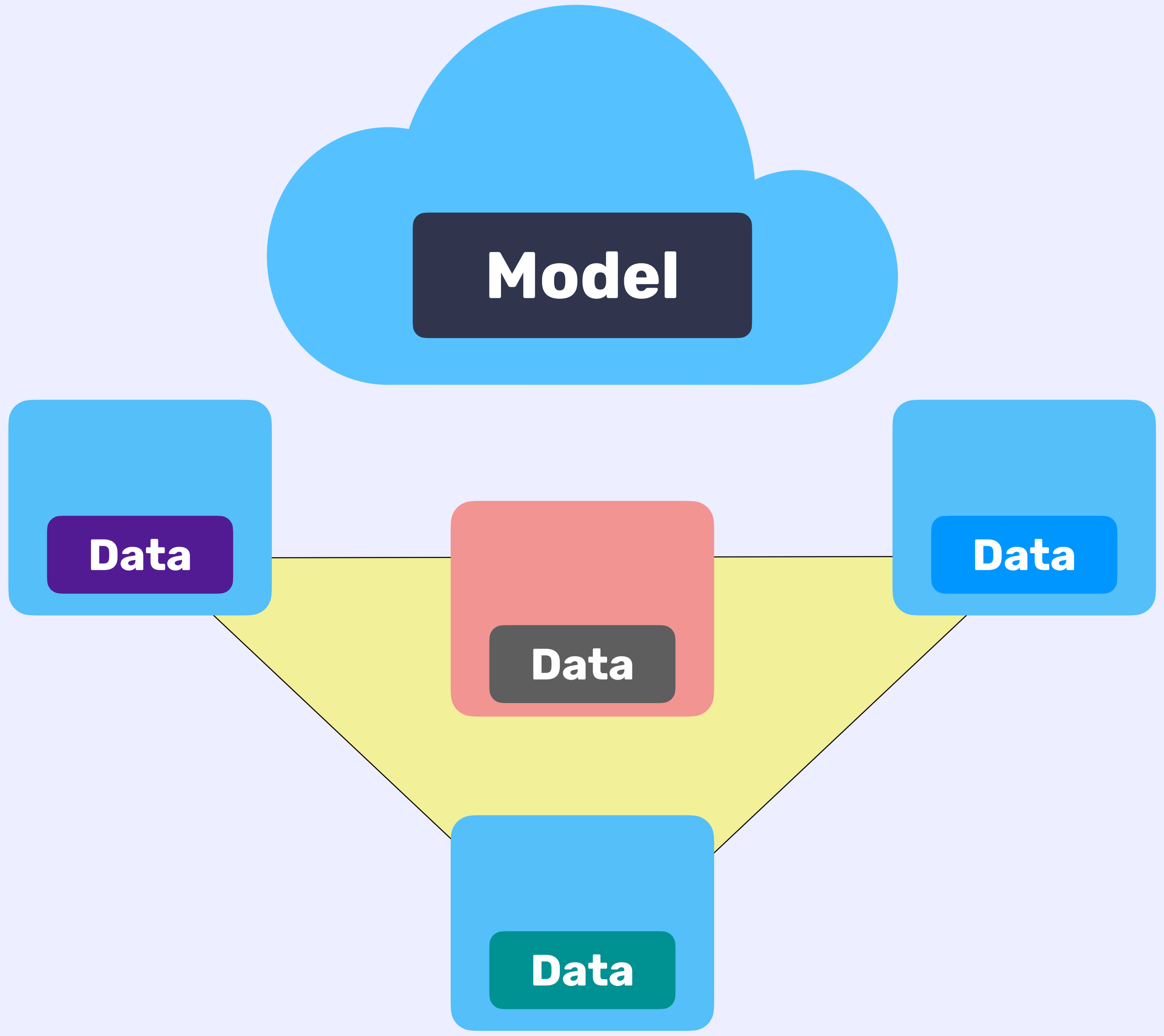
$$\min_{w \in \mathbb{R}^d} \left[F_\theta(w) = \max_{\pi \in \mathcal{P}_\theta} \mathbb{E}_{\xi \sim p_\pi}[f(w, \xi)] \right] \quad \text{where}$$

$$\mathcal{P}_\theta := \{\pi \in \Delta_{N-1} : \mathrm{conf}(p_\pi) \geq \theta\}$$

Superquantile loss

  - For any random variable $U : \Omega \to \mathbb{R}$ the superquantile of U is

$$\boxed{S_\theta(U) = \sup_{\substack{\pi \in \Delta_{N-1} \\ 0 \leq \frac{\pi_i}{\alpha_i} \leq \frac{1}{\theta}}} \sum_{i=1}^{N} \pi_i U_i \quad (\text{when } \mathbb{P}[U = U_i] = \alpha_i)}$$



The Δ-FL Framework

Δ-FL's Objective

- We propose to solve for a conformity parameter $\theta \in (0, 1]$:

$$\min_{w \in \mathbb{R}^d} \left[F_\theta(w) = \max_{\pi \in \mathcal{P}_\theta} \mathbb{E}_{\xi \sim p_\pi}[f(w, \xi)] \right] \quad \text{where}$$

$$\mathcal{P}_\theta := \{\pi \in \Delta_{N-1} : \mathrm{conf}(p_\pi) \geq \theta\}$$

Superquantile loss

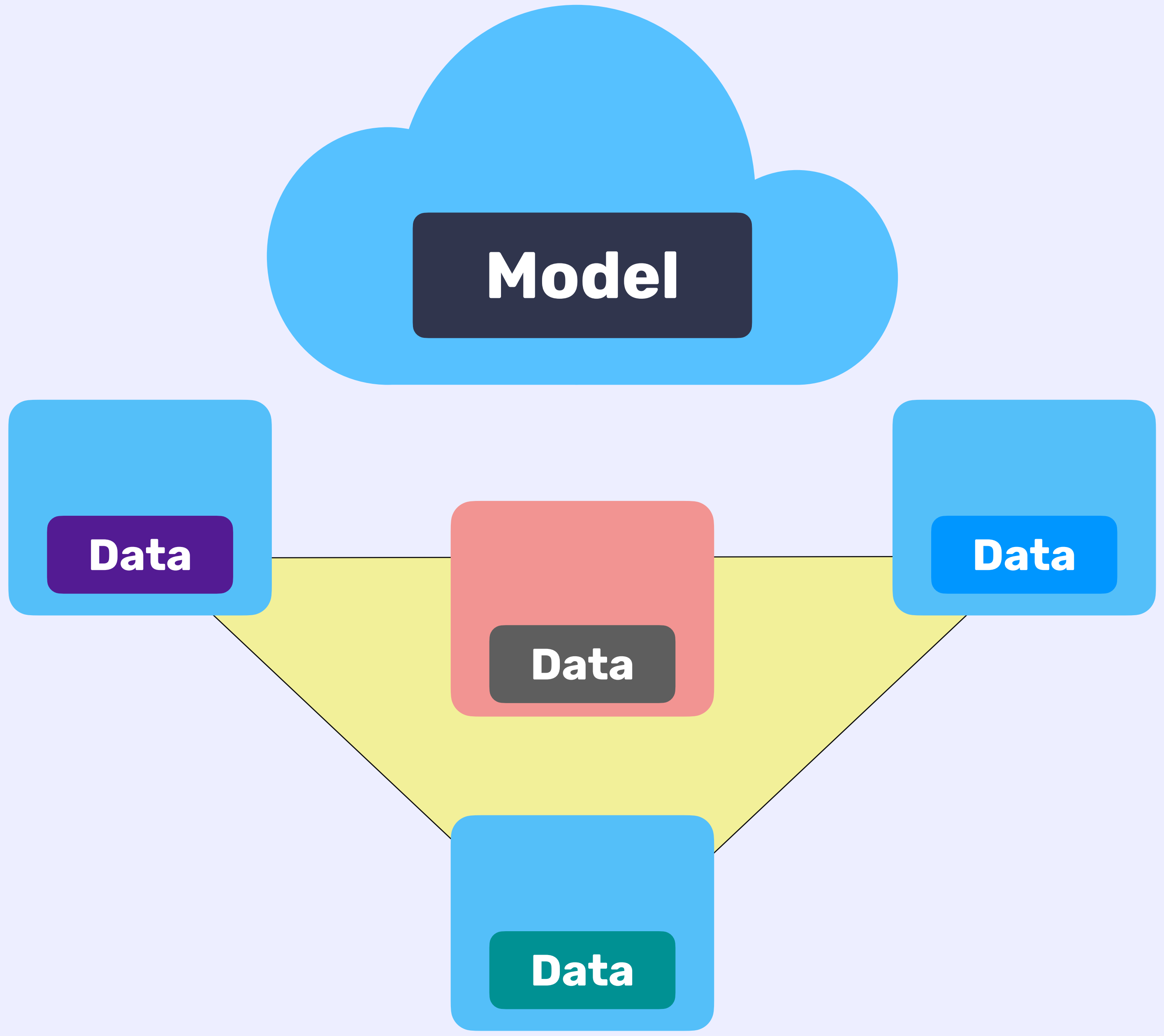
- For any random variable $U : \Omega \to \mathbb{R}$ the superquantile of U is

$$\boxed{S_\theta(U) = \sup_{\substack{\pi \in \Delta_{N-1} \\ 0 \leq \frac{\pi_i}{\alpha_i} \leq \frac{1}{\theta}}} \sum_{i=1}^{N} \pi_i U_i \quad (\text{when } \mathbb{P}[U = U_i] = \alpha_i)}$$

- In Δ-FL, we are using the superquantile at a user level

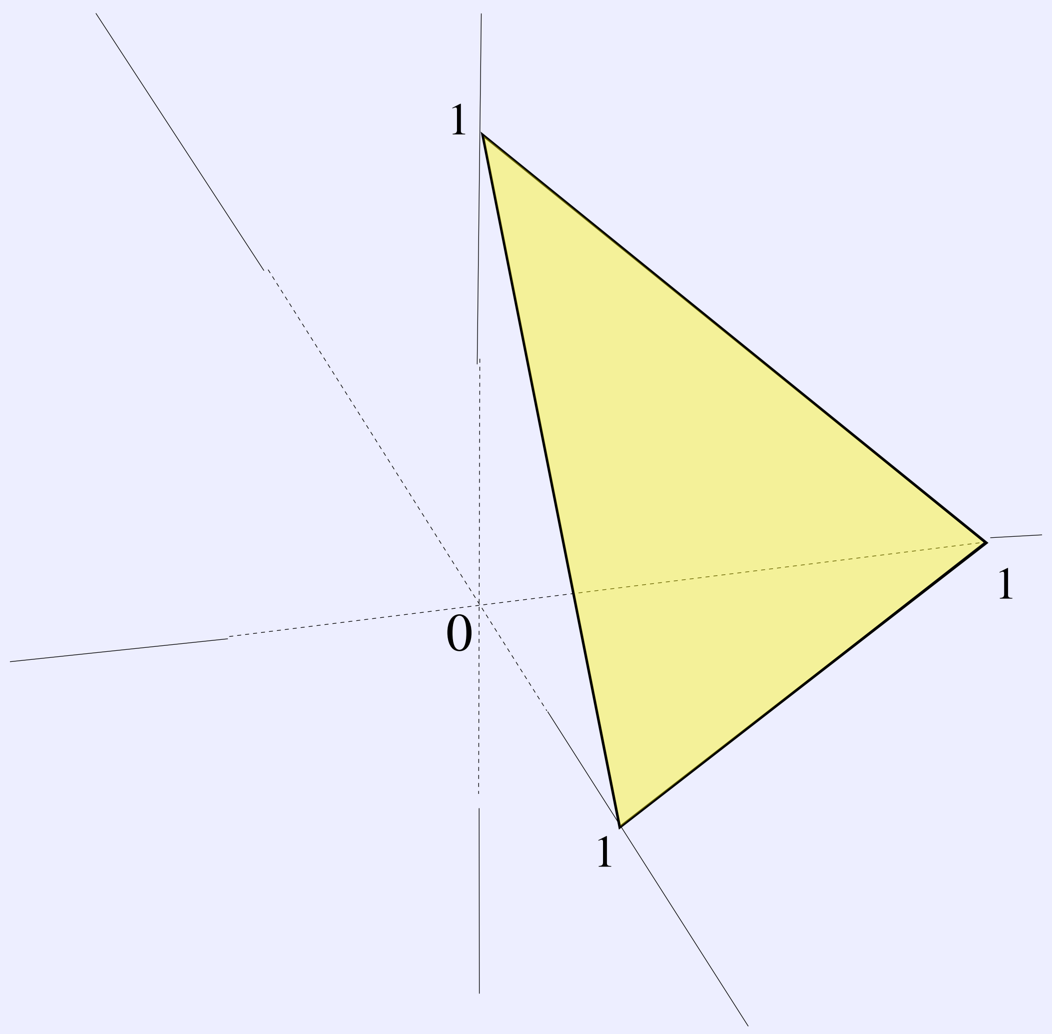
$$U = \mathbb{E}\left[F_{\mathbf{k}}(w) \mid \mathbf{k}\right] = \mathbb{E}_{\xi \sim q_{\mathbf{k}}}\left[f(w, \xi)\right] \quad \text{with} \quad \mathbb{P}[\mathbf{k} = i] = \alpha_i$$

$$F_\theta(w) = S_\theta(F_{\mathbf{k}}(w))$$



Geometrical Intuition

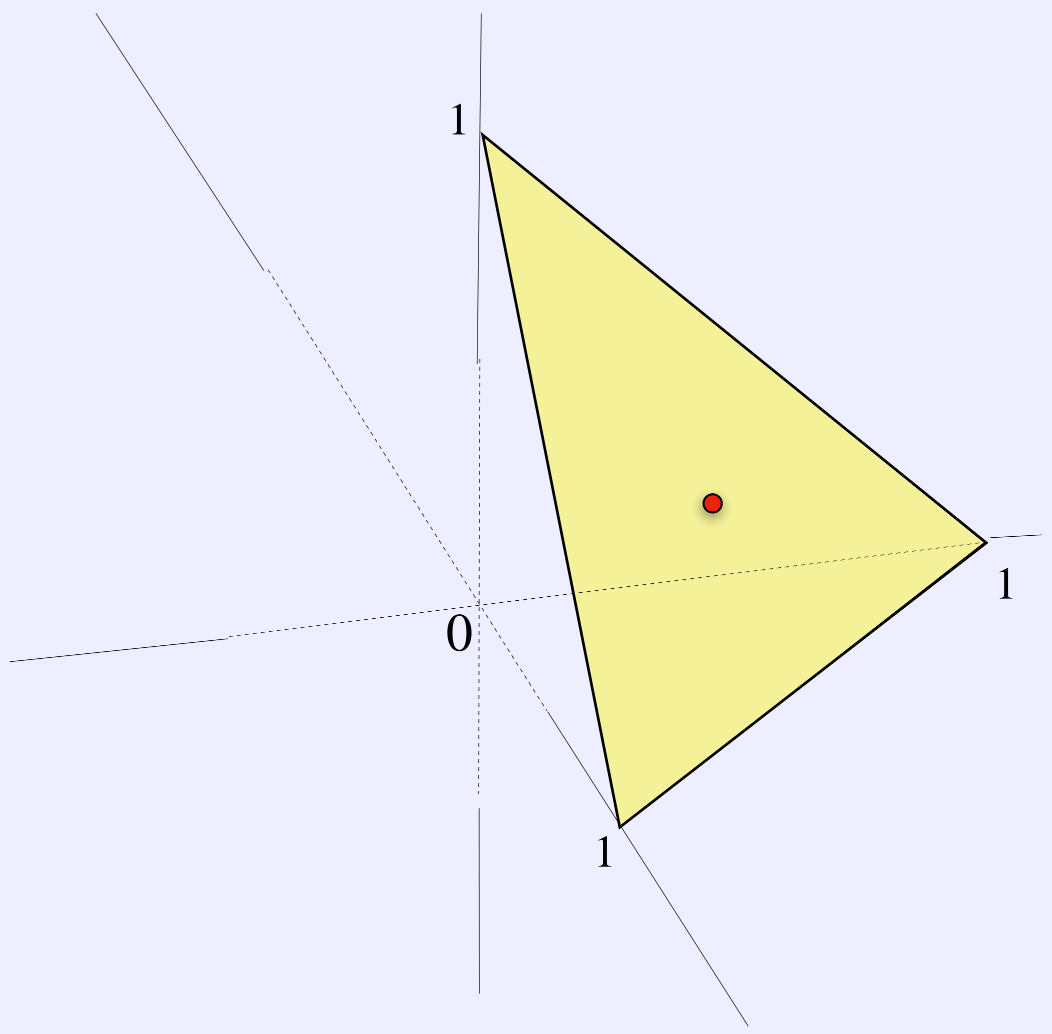
- Assume we have only three users at training time



Geometrical Intuition

- Assume we have only three users at training time

$$\alpha = (1/3, 1/3, 1/3)$$

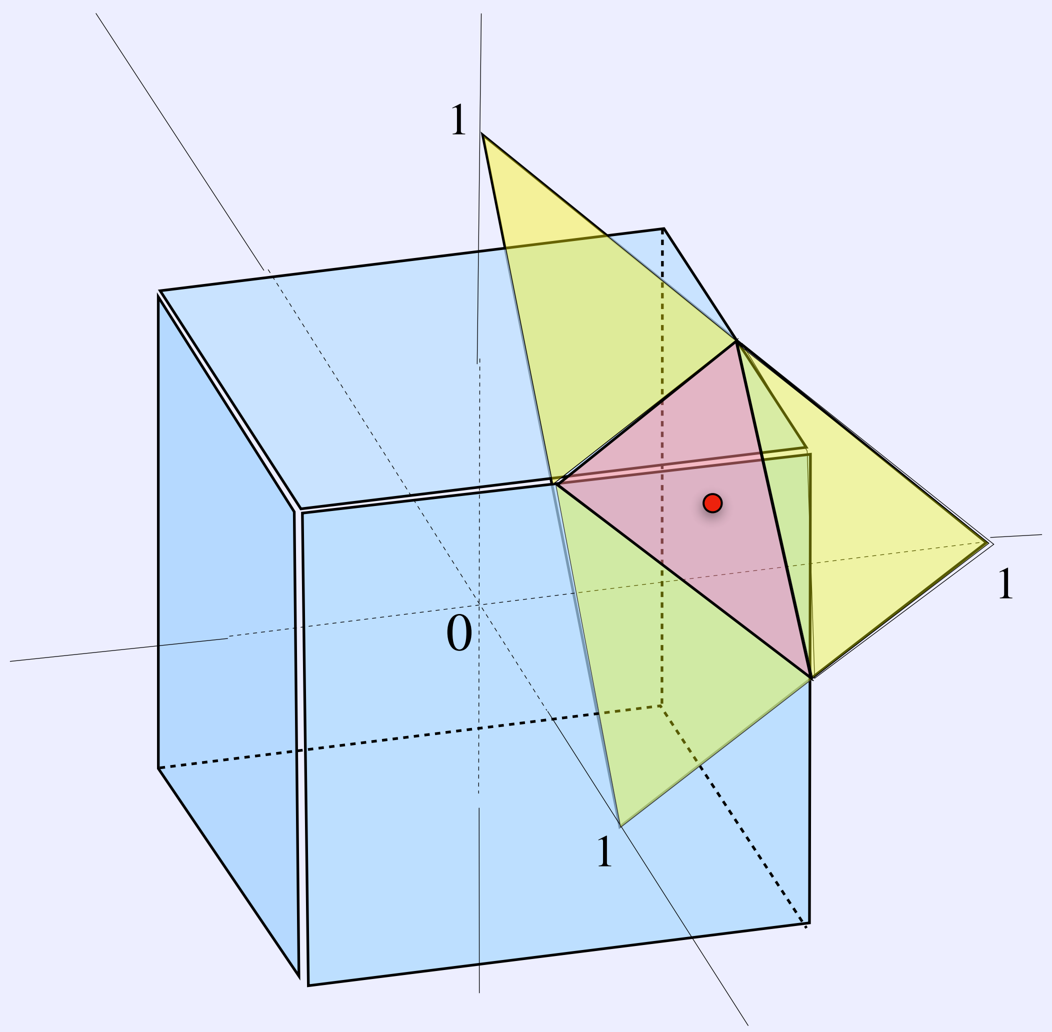


Geometrical Intuition

- Assume we have only three users at training time

$$F_\theta(w) = \sup_{\substack{\pi \in \mathbb{R}^3 \\ \pi_1 + \pi_2 + \pi_3 = 1 \\ 0 \le \frac{\pi_i}{1/3} \le 1/\theta}} \sum_{i=1}^{3} \pi_i F_i(w)$$

$$\alpha = (1/3, 1/3, 1/3)$$



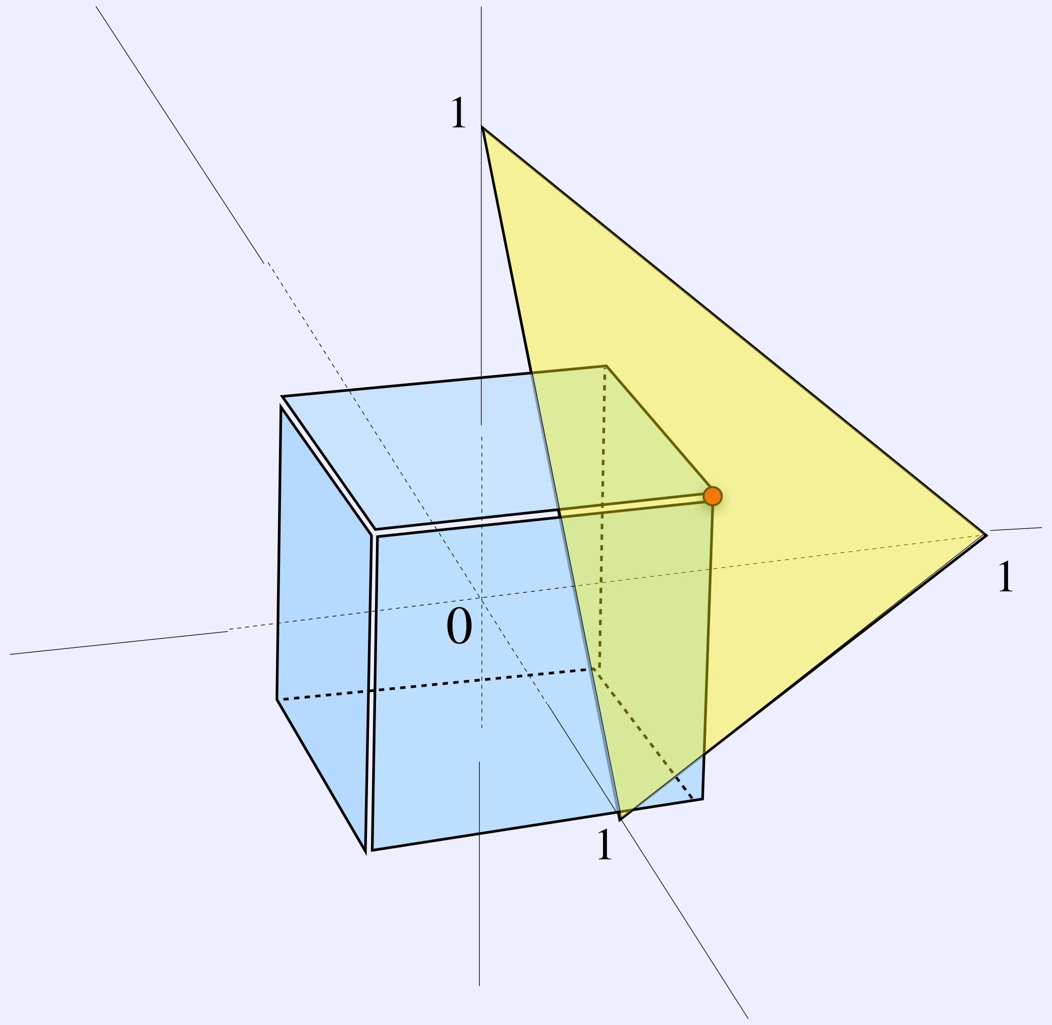
$$r = \min_{1 \le i \le N} \frac{\alpha_i}{\theta}$$

Geometrical Intuition

- Assume we have only three users at training time

$$F_\theta(w) = \sup_{\substack{\pi \in \mathbb{R}^3 \\ 0 \le 3\pi \le \frac{1}{\theta} \\ \pi_1+\pi_2+\pi_3=1}} \sum_{i=1}^{3} \pi_i F_i(w)$$

$$\alpha = (1/3, 1/3, 1/3)$$



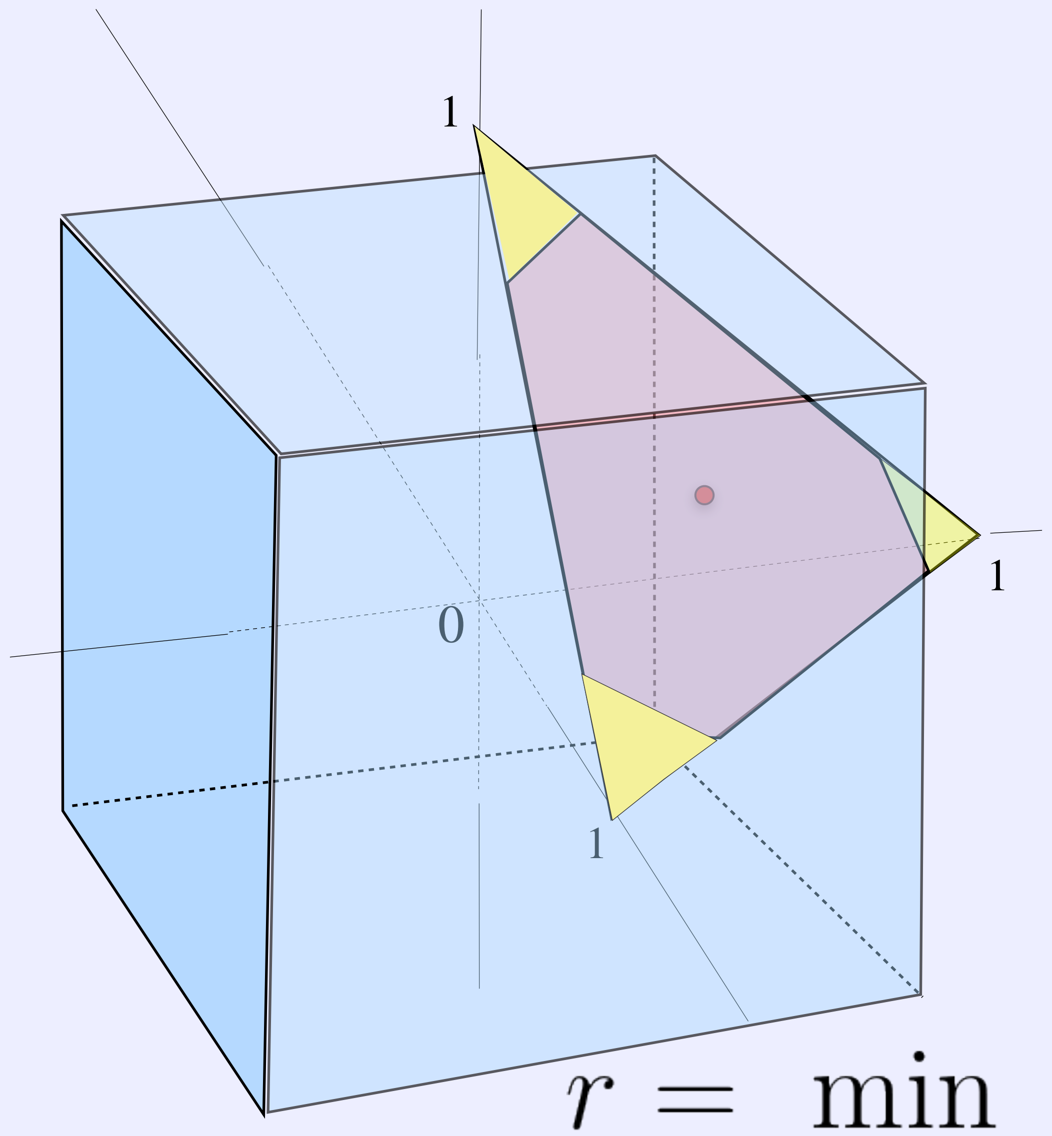
$$r = \min_{1 \le i \le N} \frac{\alpha_i}{\theta}$$

Geometrical Intuition

- Assume we have only three users at training time

$$F_\theta(w) = \sup_{\substack{\pi \in \mathbb{R}^3 \\ 0 \le 3\pi \le \frac{1}{\theta} \\ \pi_1+\pi_2+\pi_3=1}} \sum_{i=1}^{3} \pi_i F_i(w)$$

$$\alpha = (1/3, 1/3, 1/3)$$



$$r = \min_{1 \le i \le N} \frac{\alpha_i}{\theta}$$

Rockafellar's Duality Result

- A Duality Result for superquantiles [Rockafellar 2000']
  - For any $\theta \in (0, 1]$, and any discrete random variable U,

$$S_\theta(U) = \min_{\eta \in \mathbb{R}} \eta + \frac{1}{\theta}\mathbb{E}[\max(U - \eta, 0)]$$

$$Q_p(U) = \underset{\eta \in \mathbb{R}}{\operatorname{argmin}} \ \eta + \frac{1}{\theta}\mathbb{E}[\max(U - \eta, 0)]$$

$=1 - \theta$

Rockafellar's Duality Result

- A Duality Result for superquantiles [Rockafellar 2000']
  - For any $\theta \in (0, 1]$, and any discrete random variable U,

$$S_\theta(U) = \min_{\eta \in \mathbb{R}} \eta + \frac{1}{\theta}\mathbb{E}[\max(U - \eta, 0)]$$

$$Q_{\underbrace{p}_{=1-\theta}}(U) = \operatorname*{argmin}_{\eta \in \mathbb{R}} \eta + \frac{1}{\theta}\mathbb{E}[\max(U - \eta, 0)]$$

  - In our case, we can rewrite Δ-FL's objective as a joint minimization problem:

$$\min_{w \in \mathbb{R}^d} F_\theta(w) = \min_{w \in \mathbb{R}^d} S_\theta(F_{\mathbf{k}}(w)) = \min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}} \eta + \frac{1}{\theta}\sum_{i=1}^{N} \alpha_i \max(F_i(w) - \eta, 0)$$

An Alternating Minimization Scheme

- We propose to alternatively minimise:

$$G : w, \eta \mapsto \eta + \frac{1}{\theta} \sum_{i=1}^{N} \alpha_i \max(F_i(w) - \eta, 0)$$

ALTERNATING MINIMIZATION FOR Δ-FL

Input
- Starting point $w_0 \in \mathbb{R}^d$
- Inexactness sequence $(\varepsilon_t)_{t \geq 0}$
- Time horizon $t^\star \in \mathbb{N}$

for $t = 0, 1, \ldots, t^\star - 1$ **do**

$$\eta_t \in \operatorname*{argmin}_{\eta \in \mathbb{R}} G(w_t, \eta)$$

$$w_t \simeq \operatorname*{argmin}_{w \in \mathbb{R}^d} G(w, \eta_t) \ \text{ such that } \ \mathbb{E}[G(w_{t+1}, \eta_t) | w_t] - \min_{w \in \mathbb{R}^d} G(w, \eta_t) \leq \varepsilon_t$$

return $w_{t^\star}$

An Alternating Minimization Scheme

- We propose to alternatively minimise:

$$G : w, \eta \mapsto \eta + \frac{1}{\theta} \sum_{i=1}^{N} \alpha_i \max(F_i(w) - \eta, 0)$$

ALTERNATING MINIMIZATION FOR Δ-FL

Input
- Starting point $w_0 \in \mathbb{R}^d$
- Inexactness sequence $(\varepsilon_t)_{t \geq 0}$
- Time horizon $t^\star \in \mathbb{N}$

for $t = 0, 1, \ldots, t^\star - 1$ **do**

$\eta_t \in \operatorname{argmin}_{\eta \in \mathbb{R}} G(w_t, \eta)$ (quantile computation)

$w_t \simeq \operatorname{argmin}_{w \in \mathbb{R}^d} G(w, \eta_t)$ such that $\mathbb{E}[G(w_{t+1}, \eta_t) | w_t] - \min_{w \in \mathbb{R}^d} G(w, \eta_t) \leq \varepsilon_t$

return $w_{t^\star}$

An Alternating Minimization Scheme

- We propose to alternatively minimise:

$$G : w, \eta \mapsto \eta + \frac{1}{\theta} \sum_{i=1}^{N} \alpha_i \max(F_i(w) - \eta, 0)$$

ALTERNATING MINIMIZATION FOR Δ-FL

Input
- Starting point $w_0 \in \mathbb{R}^d$
- Inexactness sequence $(\varepsilon_t)_{t \geq 0}$
- Time horizon $t^\star \in \mathbb{N}$

for $t = 0, 1, \ldots, t^\star - 1$ **do**

$\eta_t \in \underset{\eta \in \mathbb{R}}{\text{argmin}}\ G(w_t, \eta)$ (quantile computation)

$w_t \simeq \underset{w \in \mathbb{R}^d}{\text{argmin}}\ G(w, \eta_t)$ such that $\mathbb{E}[G(w_{t+1}, \eta_t) | w_t] - \min_{w \in \mathbb{R}^d} G(w, \eta_t) \leq \varepsilon_t$ (Mini-batch SGD) FedAvg

return $w_{t^\star}$

Convergence Result

- Assumptions for Local SGD

$$\widetilde{G}(w,\eta) = \eta + \frac{1}{\theta}\sum_{i=1}^{N} \alpha_i h_\nu(F_i(w) - \eta) + \frac{\lambda}{2}\|w\|_2^2$$

  - The local losses F_i are convex B-Lipschitz and L-smooth
  - We dispose of an unbiased stochastic first-order oracle for the composition $w, \eta \mapsto h_\nu(F_i(w) - \eta)$ with bounded variance σ_i^2 for the gradient with respect to w. Let $\sigma^2 = \alpha_1\, \sigma_1^2 + \cdots + \alpha_N\, \sigma_N^2$
  - A last technical assumption [Koloskova et al. 2020]

$$\sum_{i=1}^{N} \alpha_i \left\| \frac{1}{\theta} \nabla_w h_\nu(F_k(w) - \eta) + \lambda w \right\|^2 \leq D^2 + D_1 \left\|\nabla_w G(w,\eta)\right\|^2$$

- Convergence Rate Result

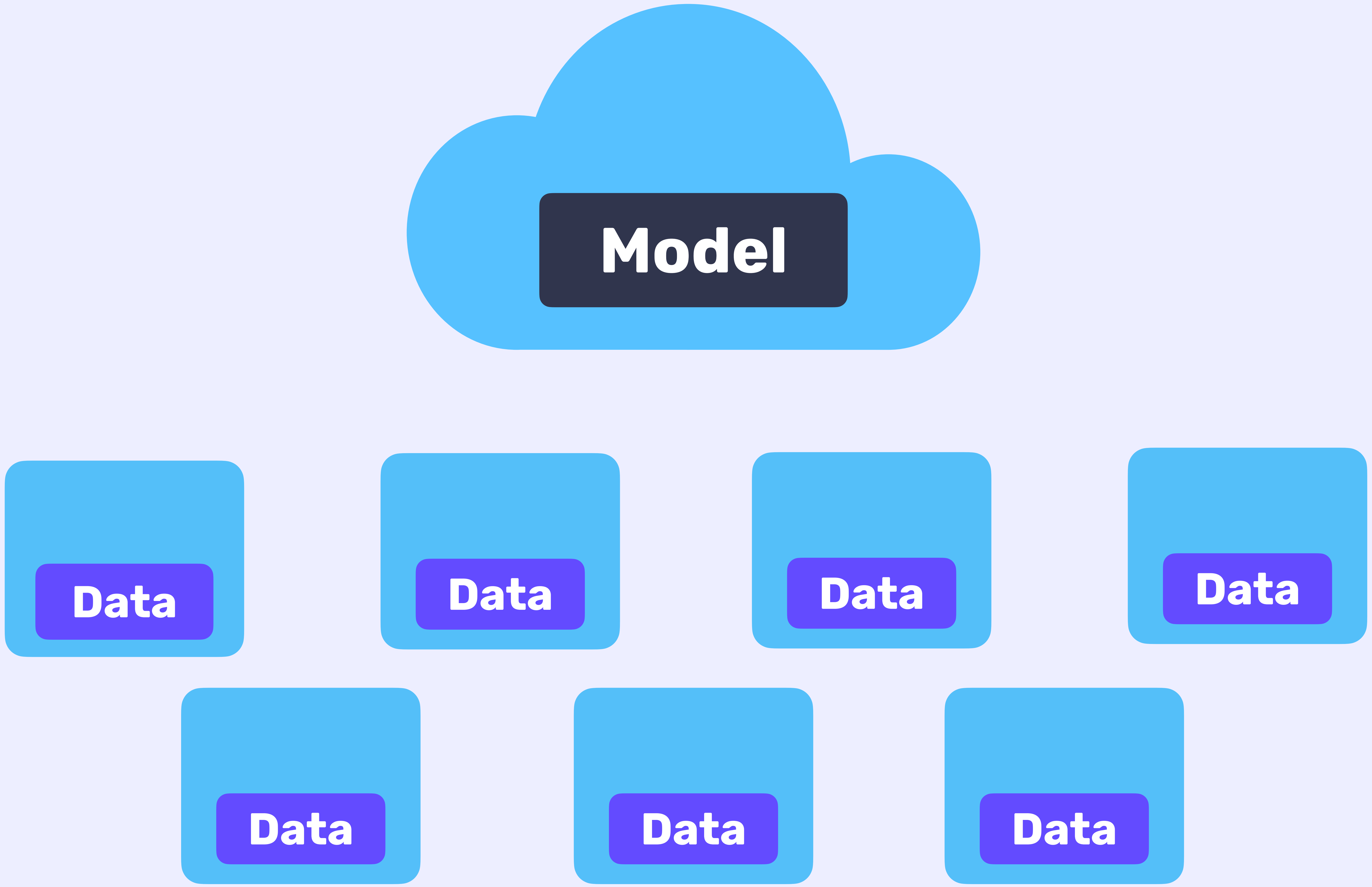
Theorem

Under above assumptions, when running local SGD with respect to w with τ local steps, we bound the total number of T communication rounds to achieve ε accuracy with:

$$T = \mathcal{O}\left(\frac{\|\alpha\|_\infty \sigma^2 \kappa^2}{\lambda \tau \varepsilon} + \sqrt{\frac{\sigma^2 \kappa^3}{\lambda^2 \tau \varepsilon}} + \sqrt{\frac{D^2 \kappa^4}{\lambda \varepsilon}} + \kappa^2 \right)$$

Practical Implementation

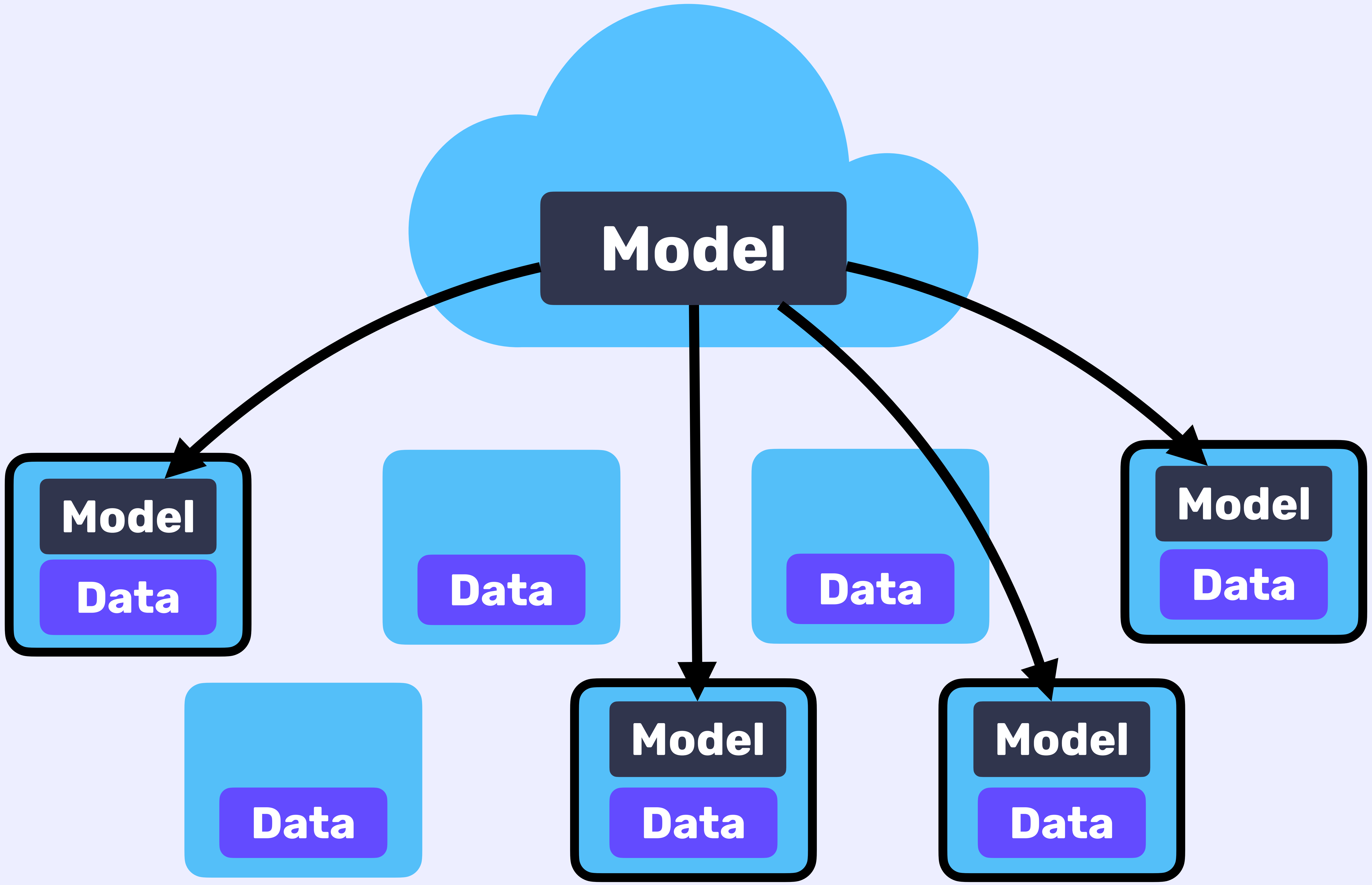
- The practical algorithm on a picture



Practical Implementation

- The practical algorithm on a picture

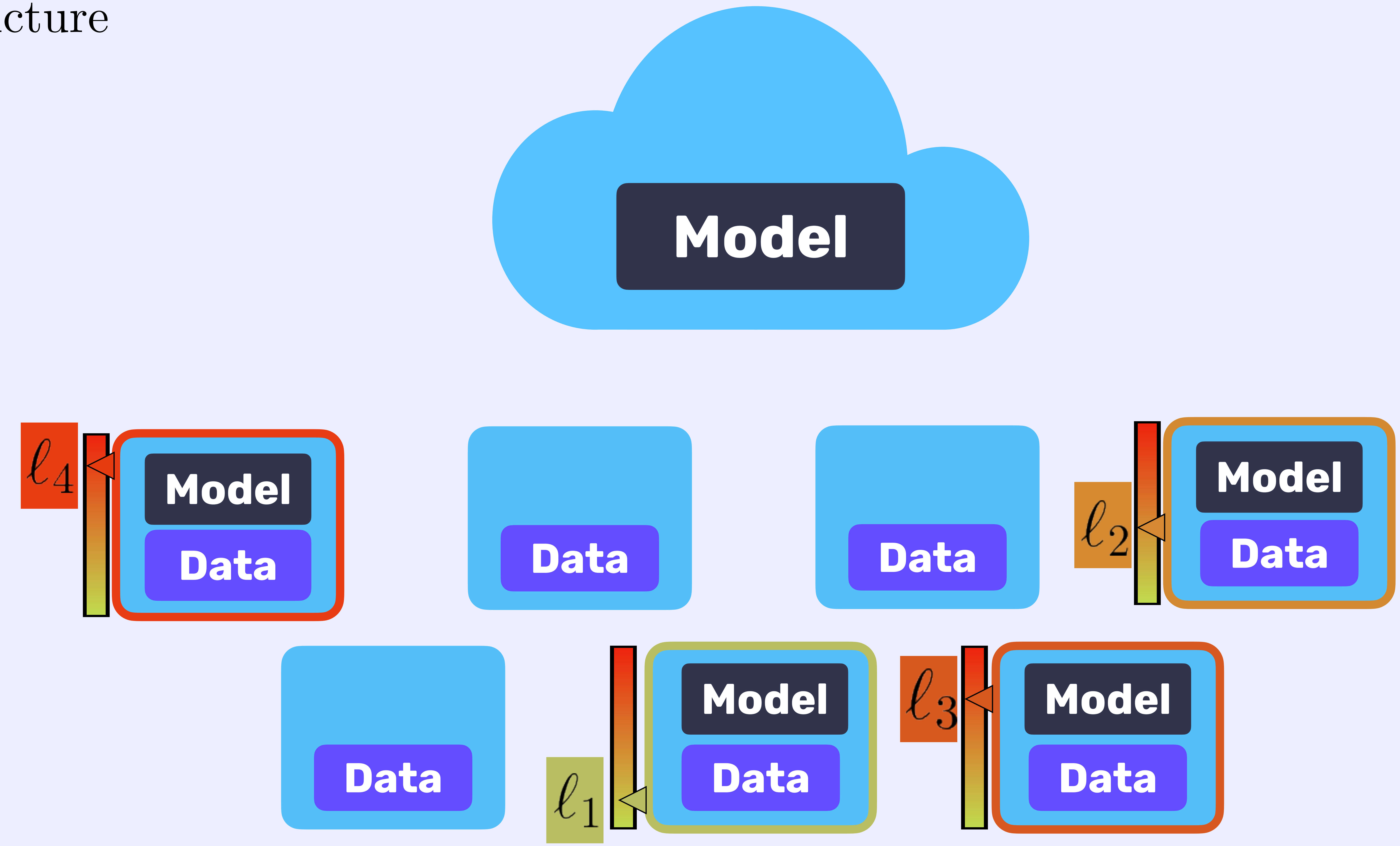
1 The server broadcasts the model to a fleet of selected devices



Practical Implementation

- The practical algorithm on a picture

1. The server broadcasts the model to a fleet of selected devices
2. Each device compute a local loss with respect to its own data

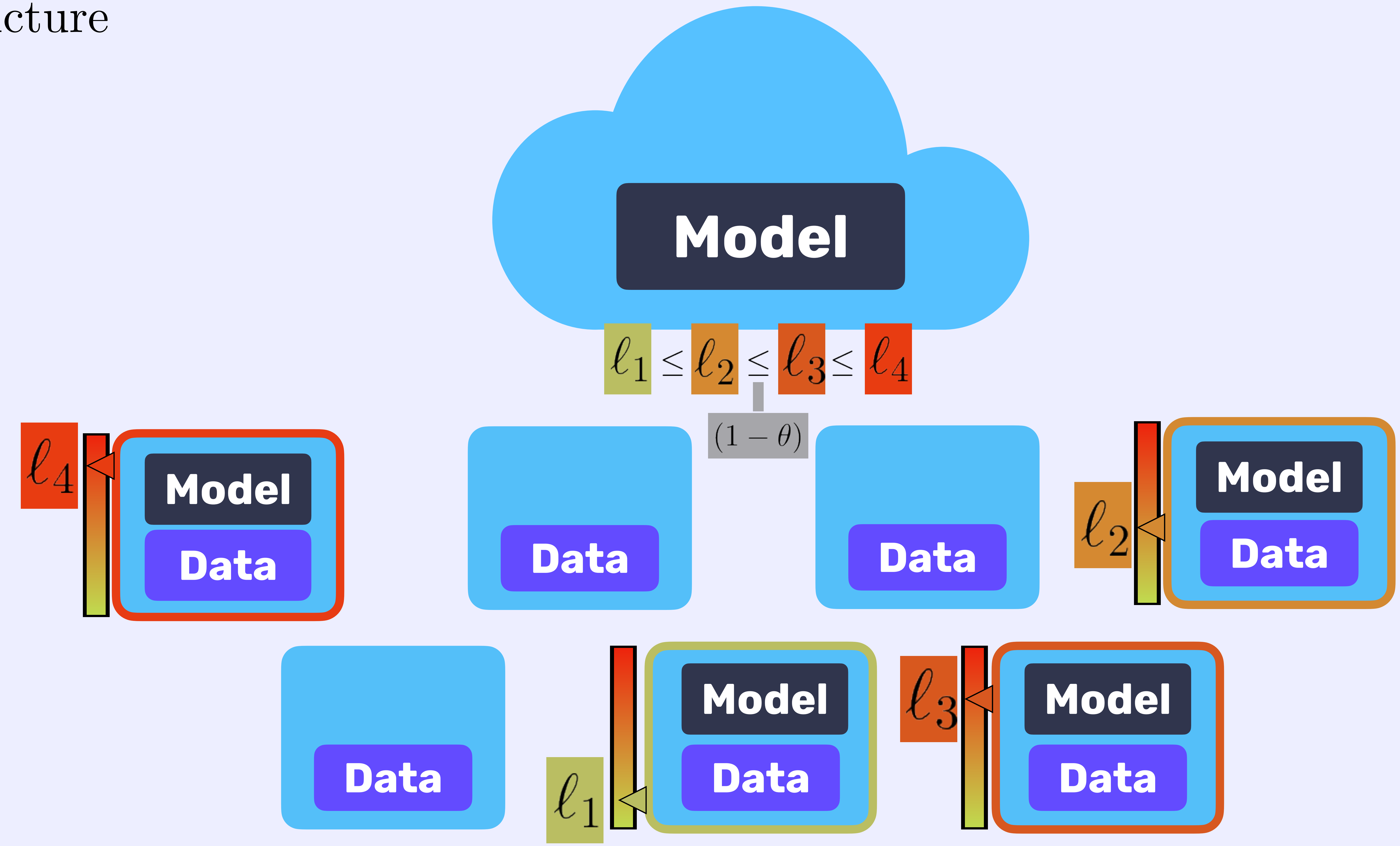


Practical Implementation

- The practical algorithm on a picture

1 The server broadcasts the model to a fleet of selected devices

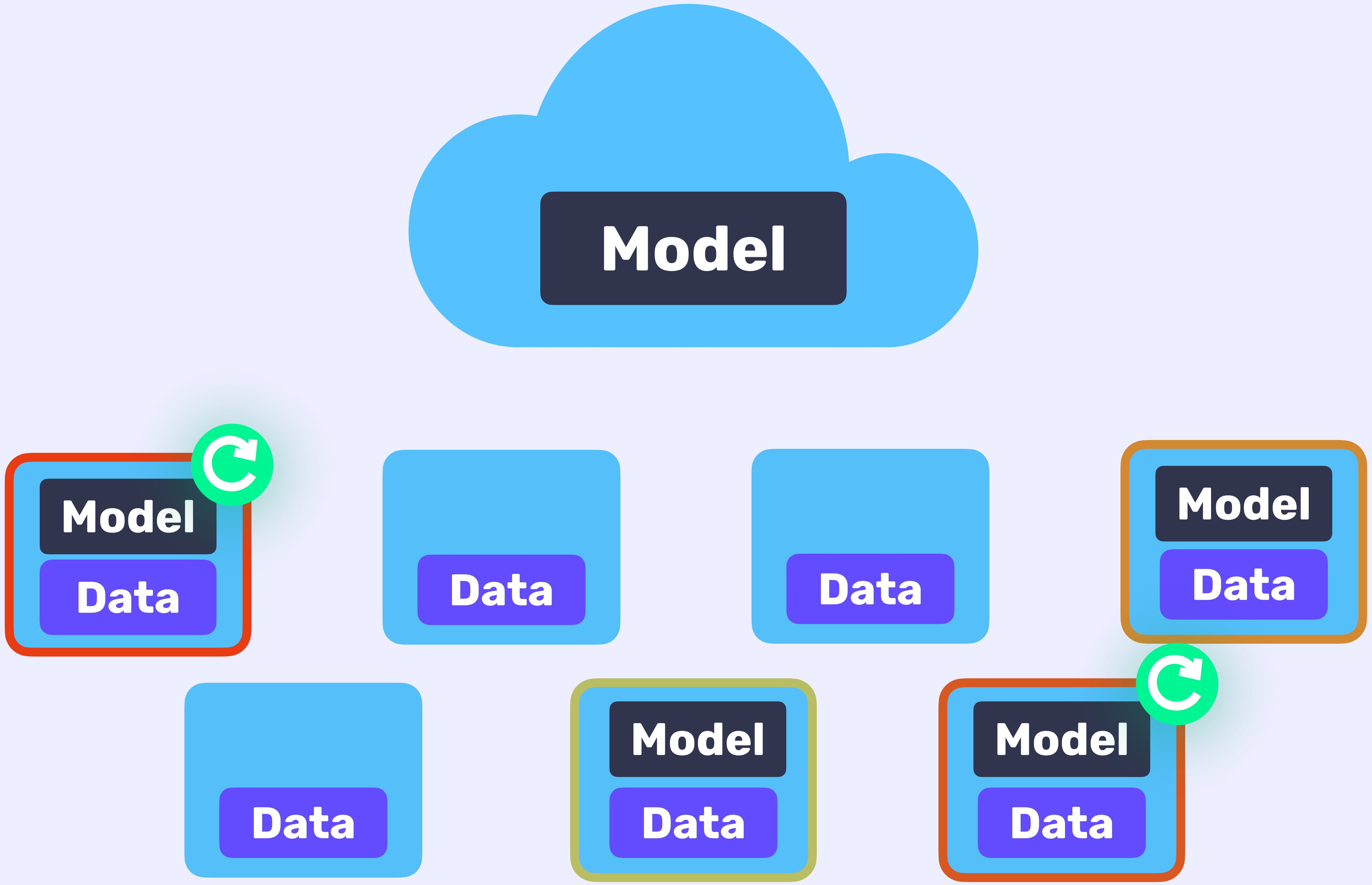
2 Each device compute a local loss with respect to its own data



Practical Implementation

- The practical algorithm on a picture

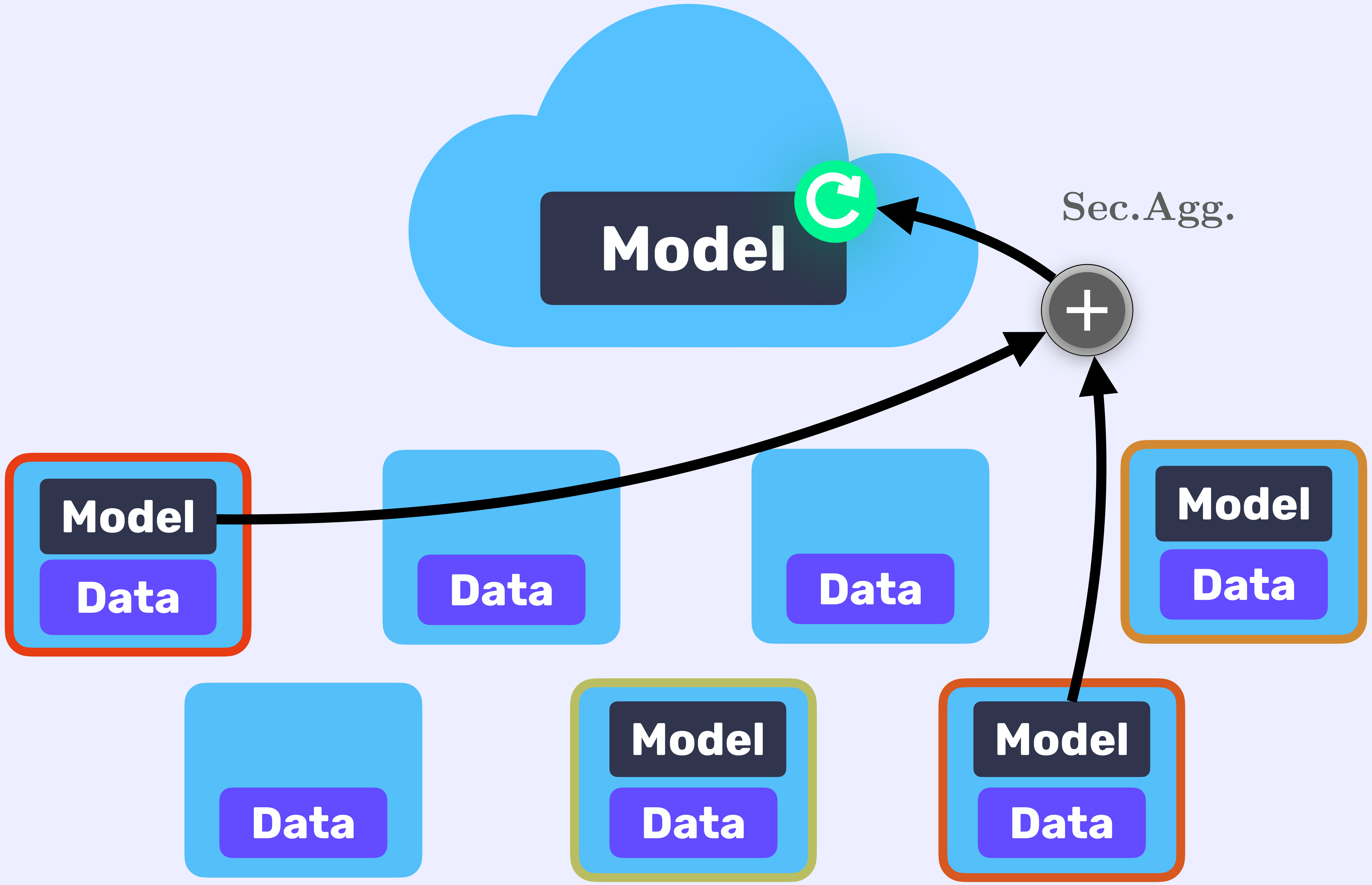
1. The server broadcasts the model to a fleet of selected devices
2. Each device compute a local loss with respect to its own data
3. Only devices with a high enough loss run local SGD for a fixed number of steps.



Practical Implementation

- The practical algorithm on a picture

1	The server broadcasts the model to a fleet of selected devices
2	Each device compute a local loss with respect to its own data
3	Only devices with a high enough loss run local SGD for a fixed number of steps.
4	The server performs a secure average of the updated models



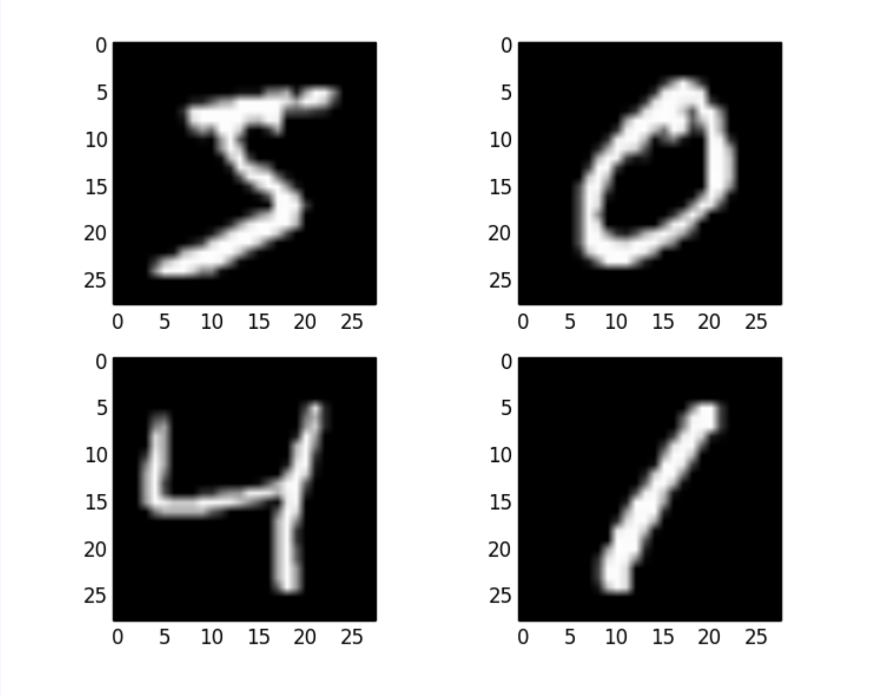
Experimental Setup

- Datasets, Tasks and Models [Caldas et al. 2019]

1730 writers

179 images per device

Character Recognition



EMNIST

Regularized Logistic Regression

ConvNet

877 accounts

69 tweets per devices

Sentiment Analysis

SENT140

Regularized Logistic Regression

LSTM

1091 roles

1346 tweets per devices

Language Modelling

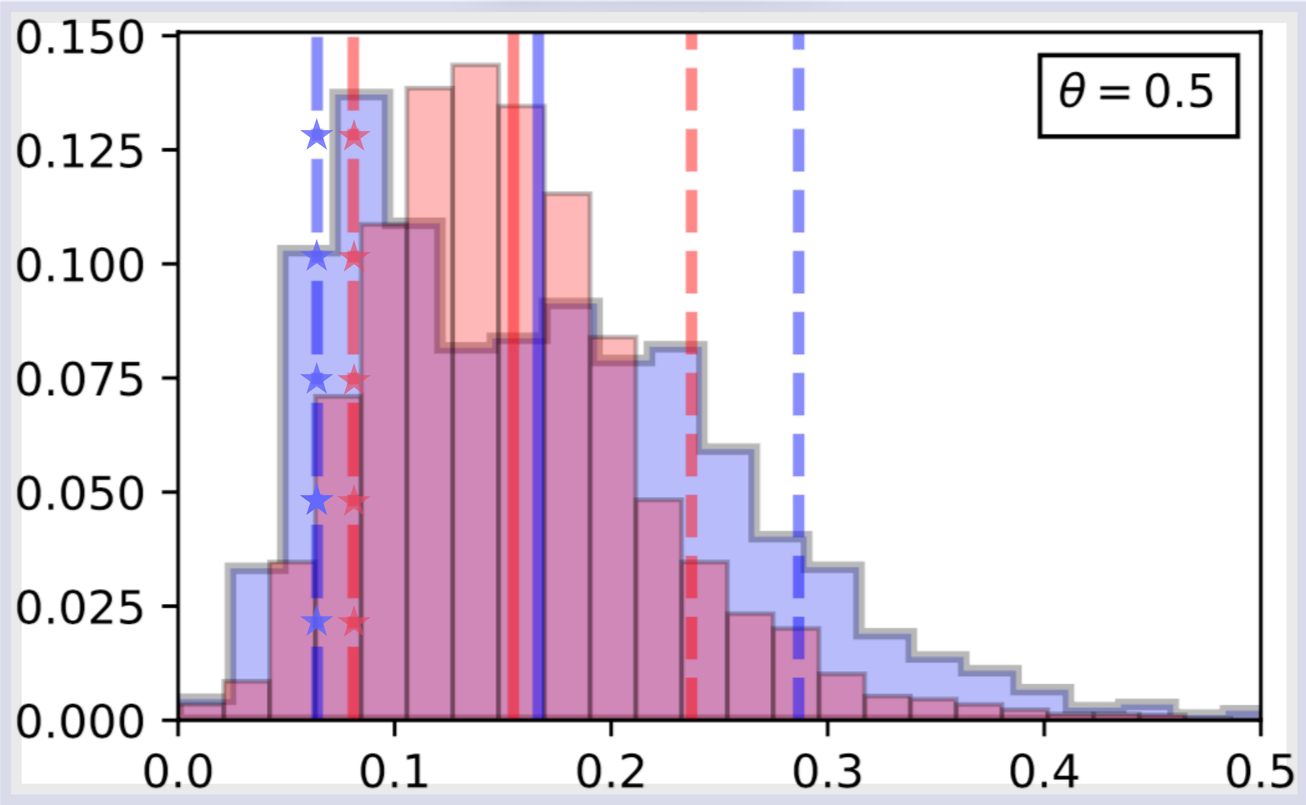
SHAKESPEARE

RNN

Experimental Results - Final Performances

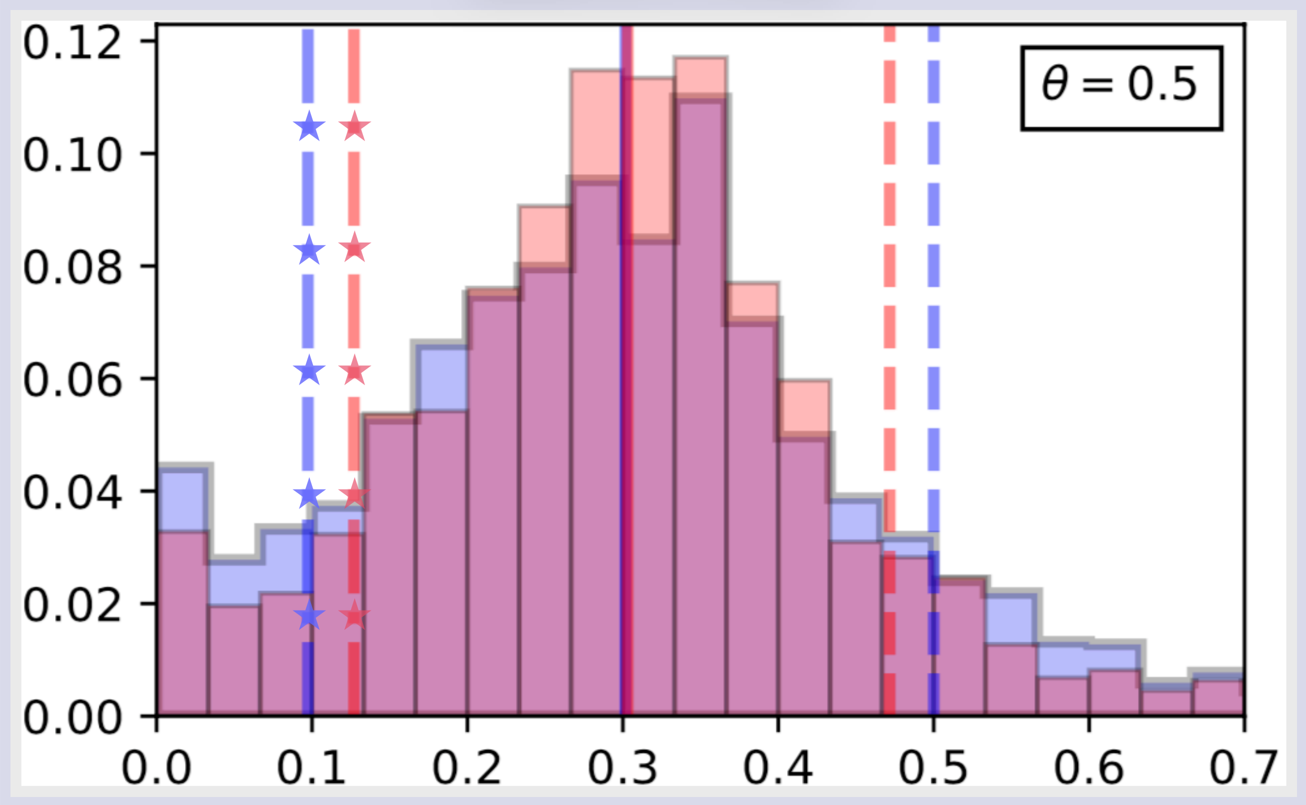
- Distribution of final misclassification error

EMNIST



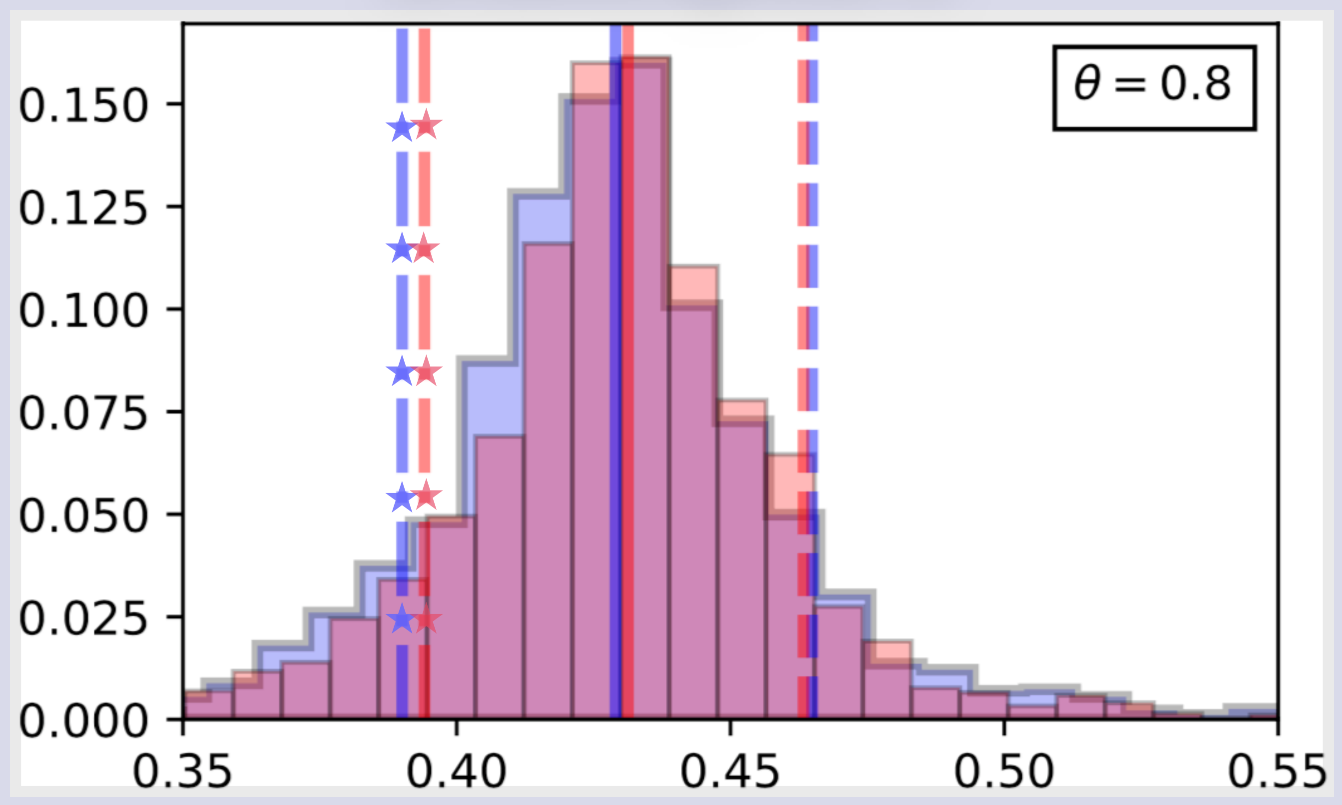
Conformity level $\theta = 0.5$

Sent140



Conformity level $\theta = 0.5$

Shakespeare



Conformity level $\theta = 0.8$

Distribution of final misclassification error for FedAvg

10^{th} percentile for FedAvg

90^{th} percentile for FedAvg

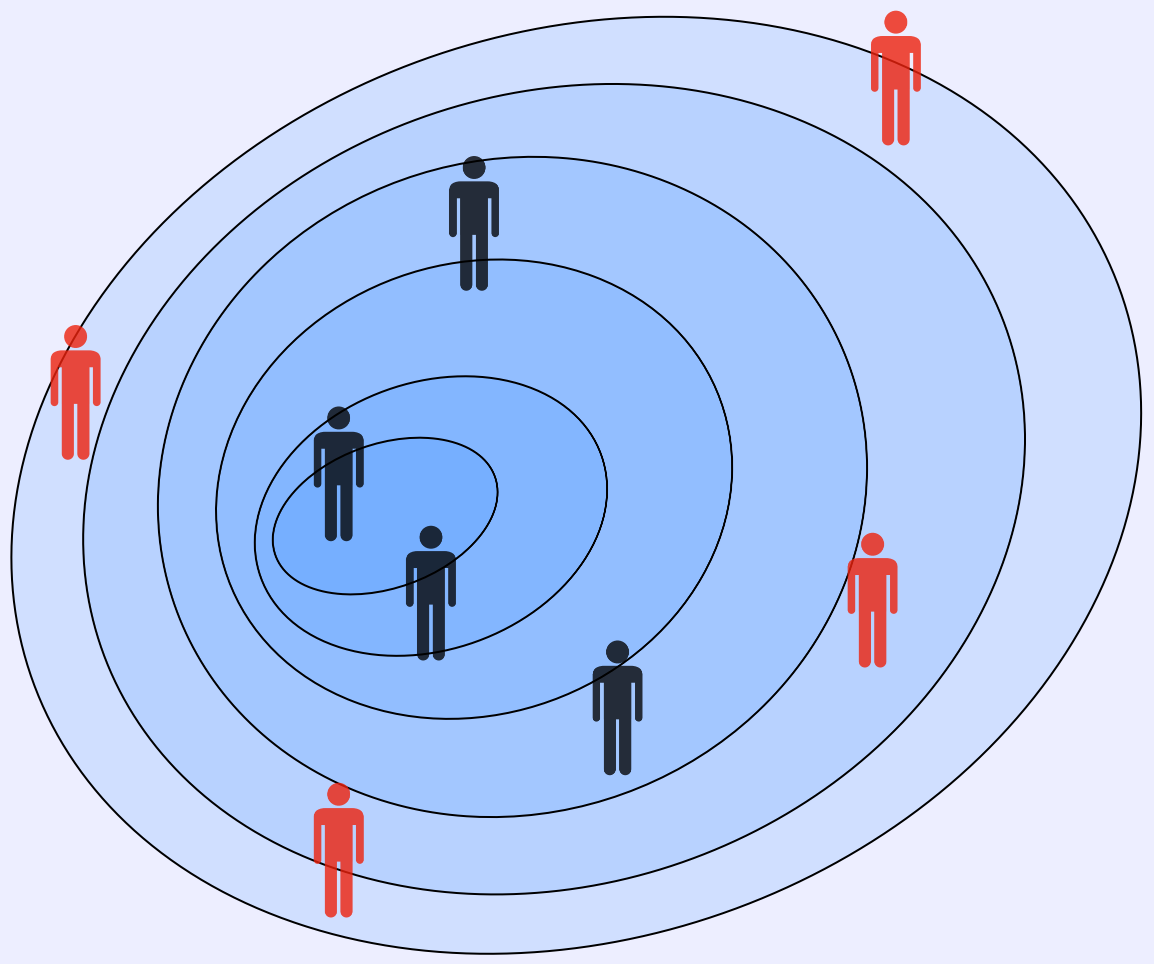
Distribution of final misclassification error for Δ-FL

10^{th} percentile for Δ-FL

90^{th} percentile for Δ-FL

Conclusion and Perspectives

- A new framework for statistical heterogeneous settings in Federated Learning, better suited for non-conforming users.
- We analysed the associated optimization algorithm and established bounds on the communication rounds it requires.
- We present numerical evidence in support of this framework.



- Paper recently published in the proceedings of the 55th Annual Conference on Information Sciences and Systems (CISS)

  Link: https://ieeexplore.ieee.org/abstract/document/9400318