DEVICE HETEROGENEITY IN FEDERATED LEARNING
A SUPERQUANTILE APPROACH

FEDERATED LEARNING ONE WORLD SEMINAR

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

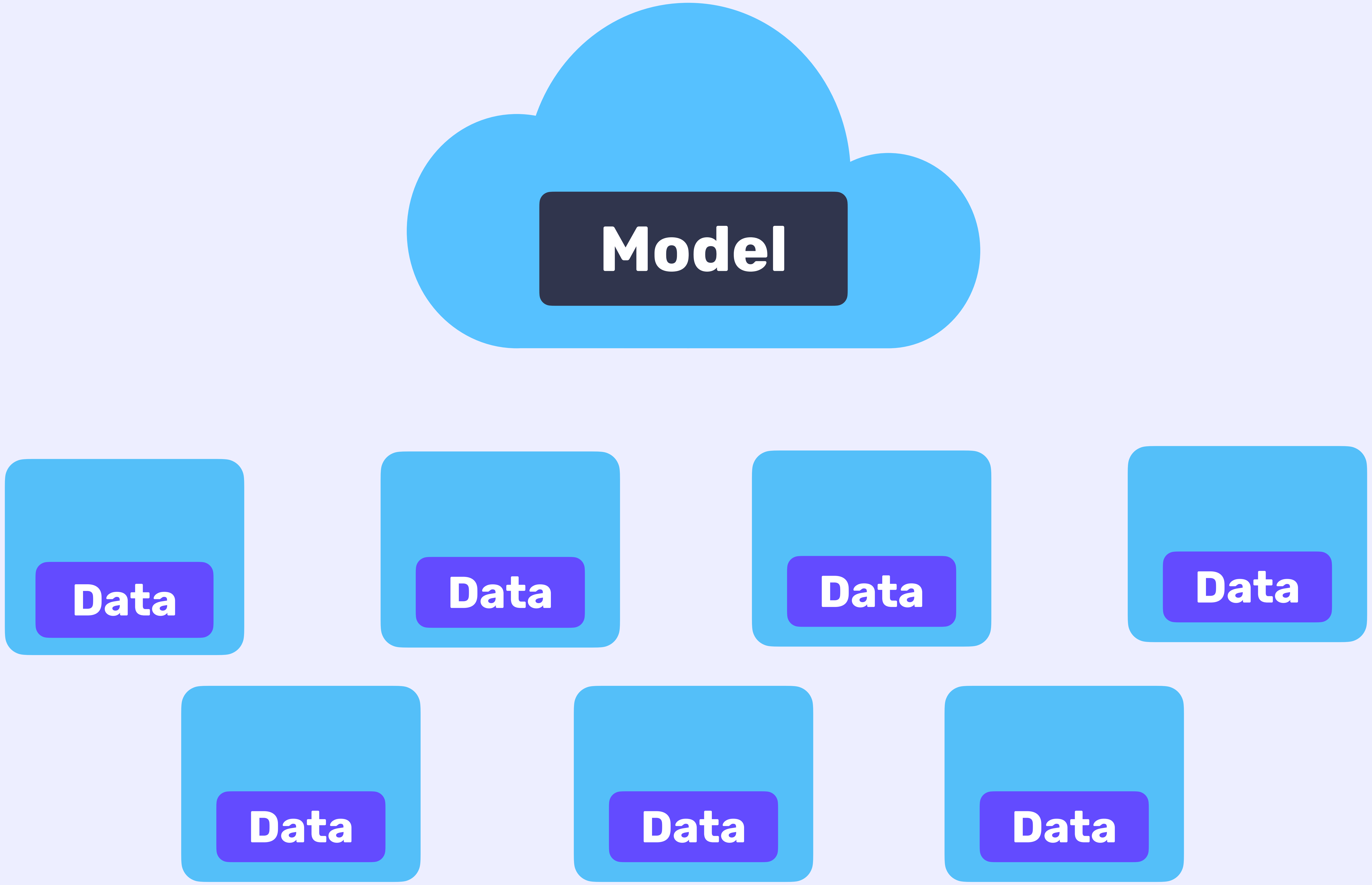
K. PILLUTLA

University of Washington

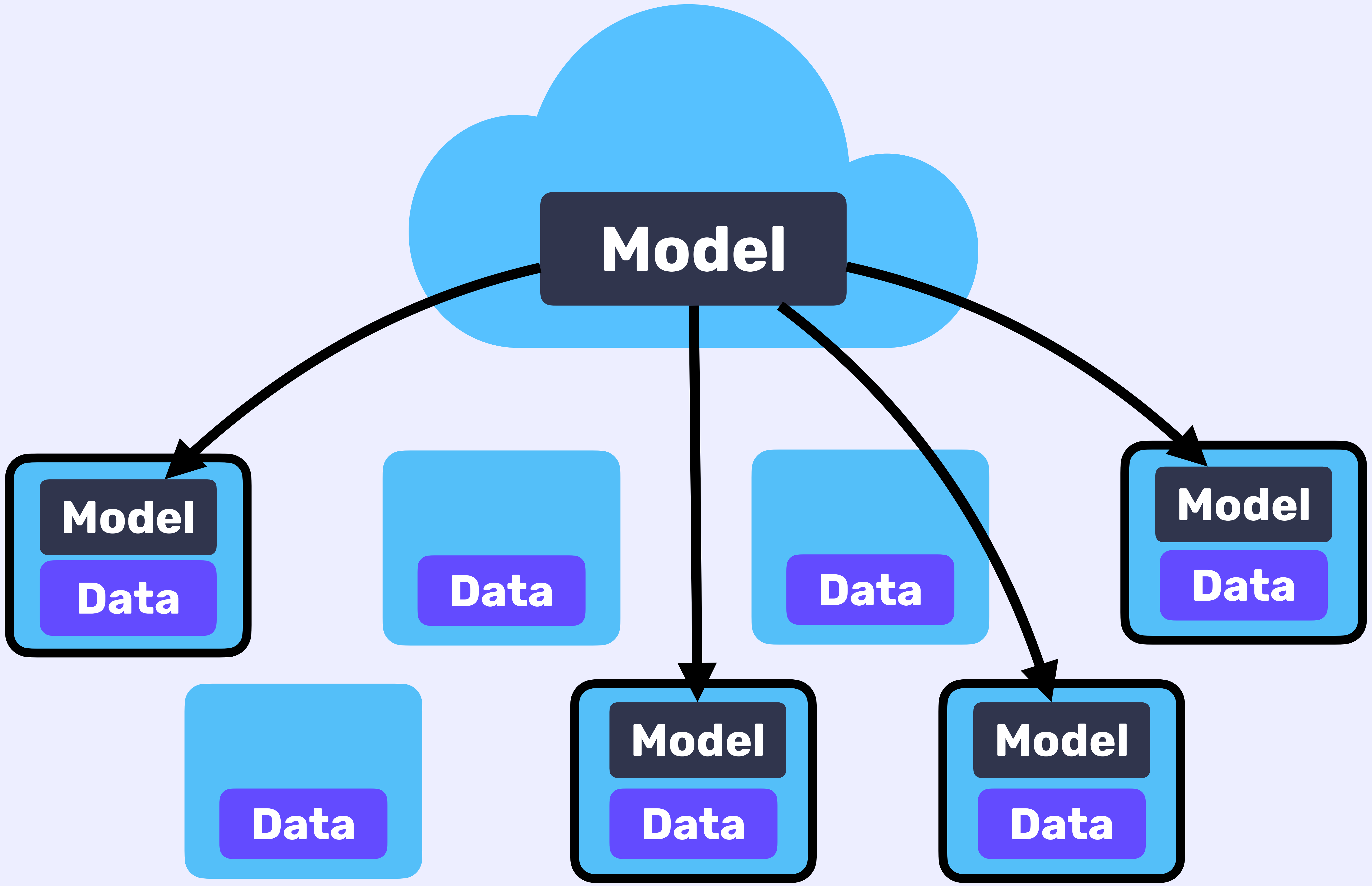
Z. HARCHAOUI

FEDERATED LEARNING IN A NUTSHELL

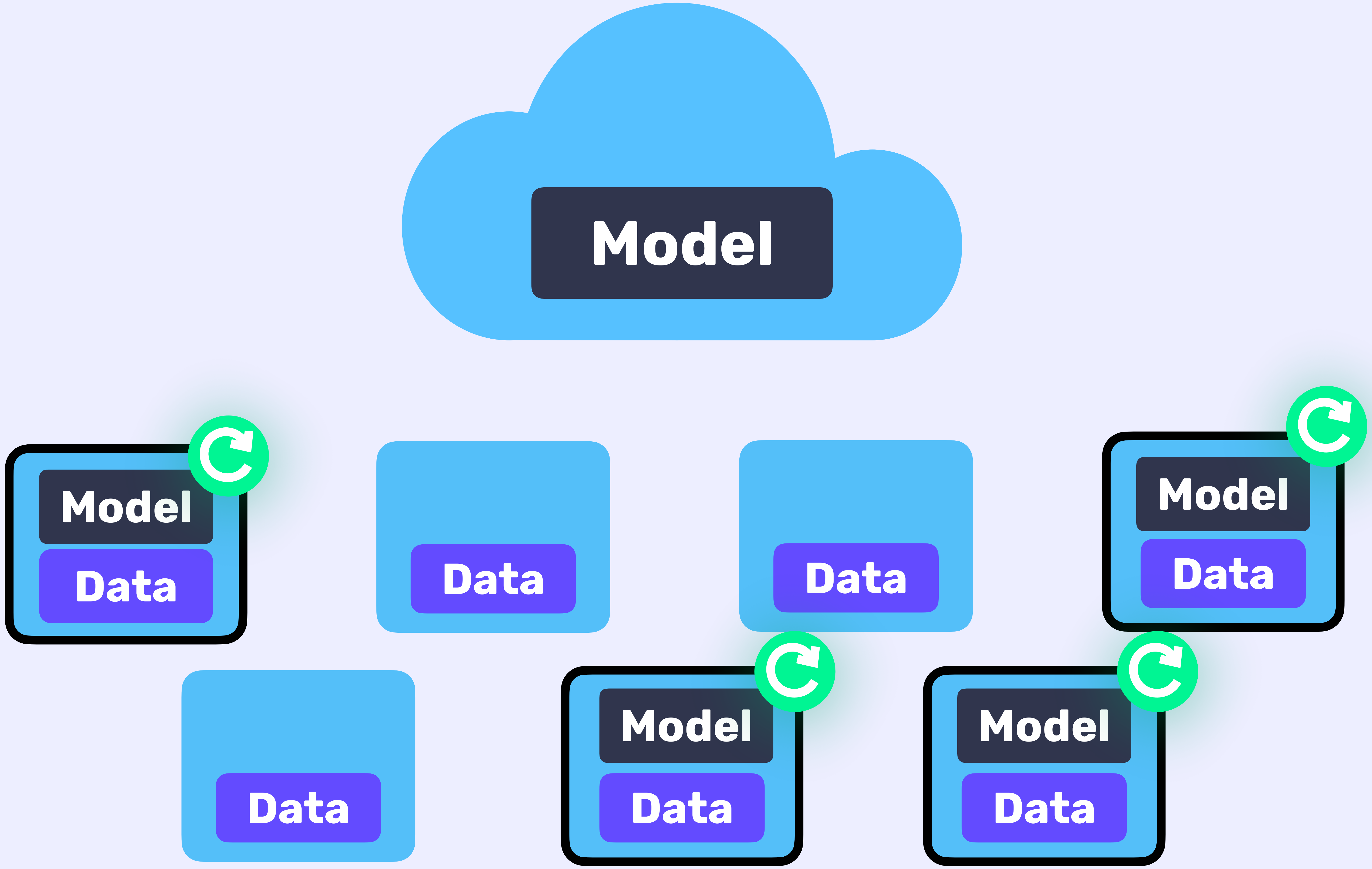
FEDERATED LEARNING IN A NUTSHELL



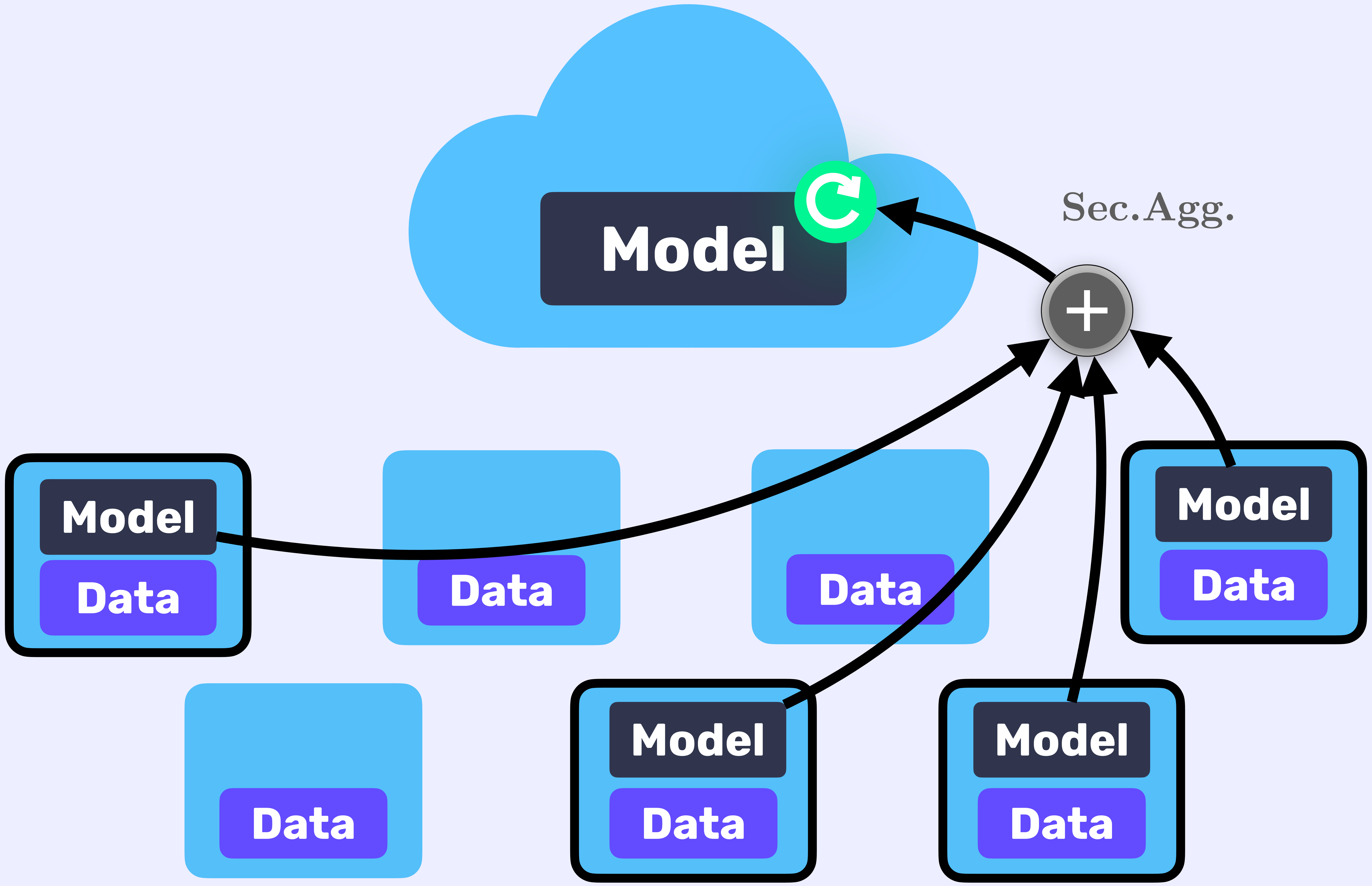
FEDERATED LEARNING IN A NUTSHELL



FEDERATED LEARNING IN A NUTSHELL



FEDERATED LEARNING IN A NUTSHELL



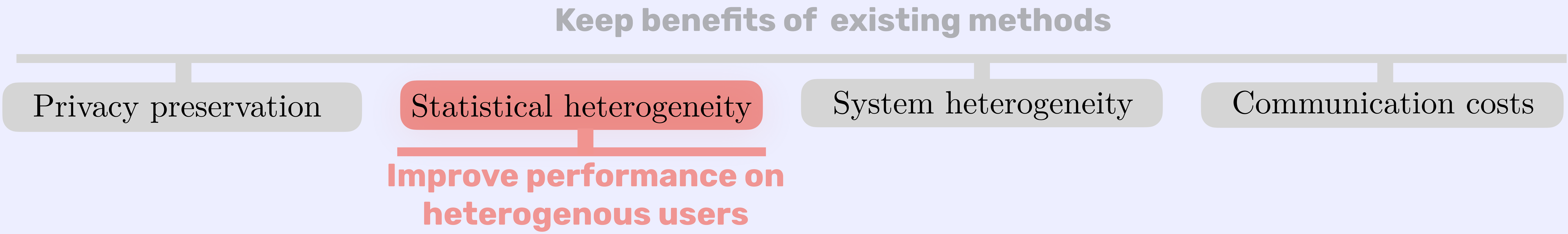
CHALLENGES

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

Privacy preservation | Statistical heterogeneity | System heterogeneity | Communication costs

CHALLENGES

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



CHALLENGES

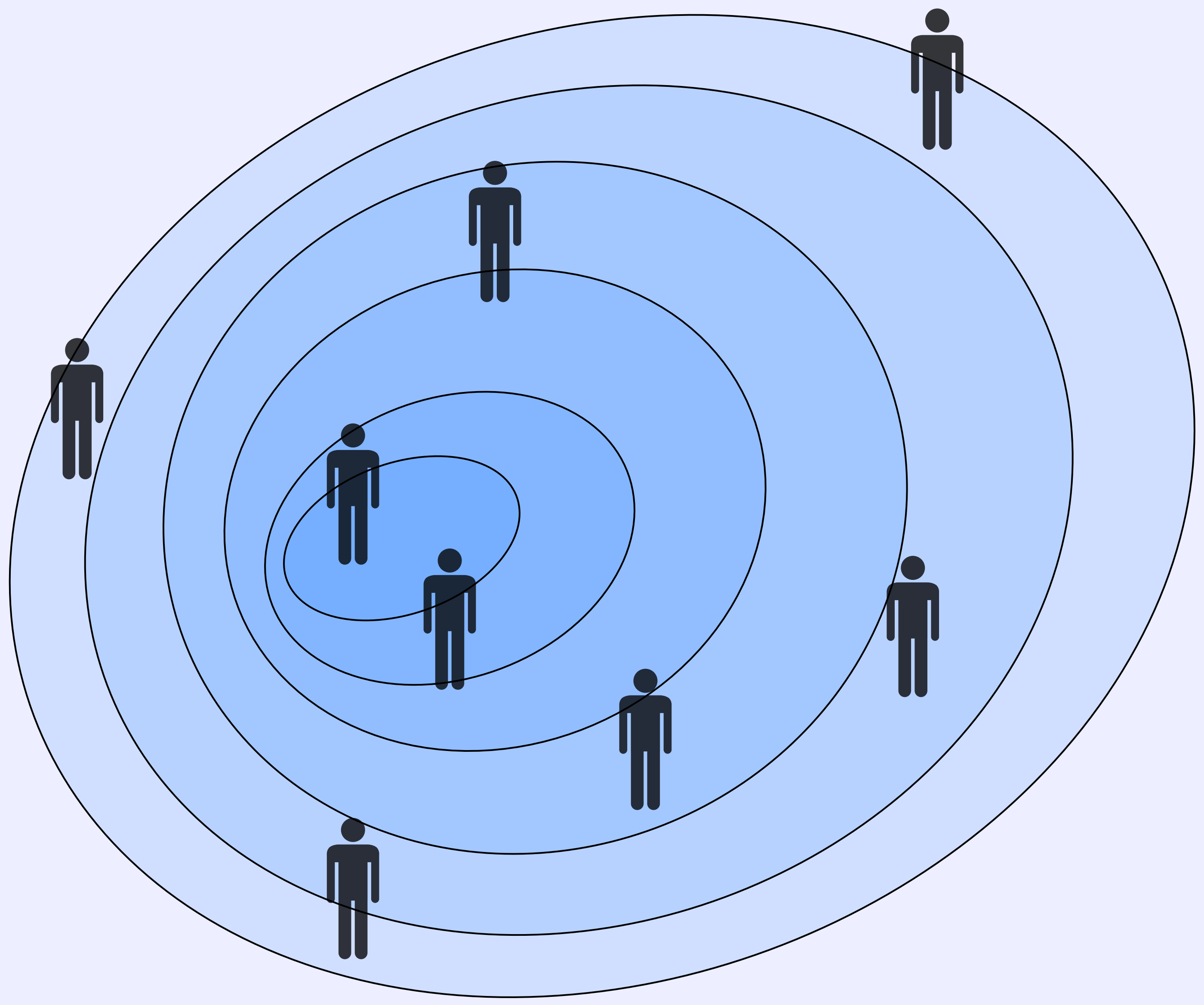
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

Privacy preservation | Statistical heterogeneity | System heterogeneity | Communication costs

Improve performance on heterogenous users

- Users heterogeneity



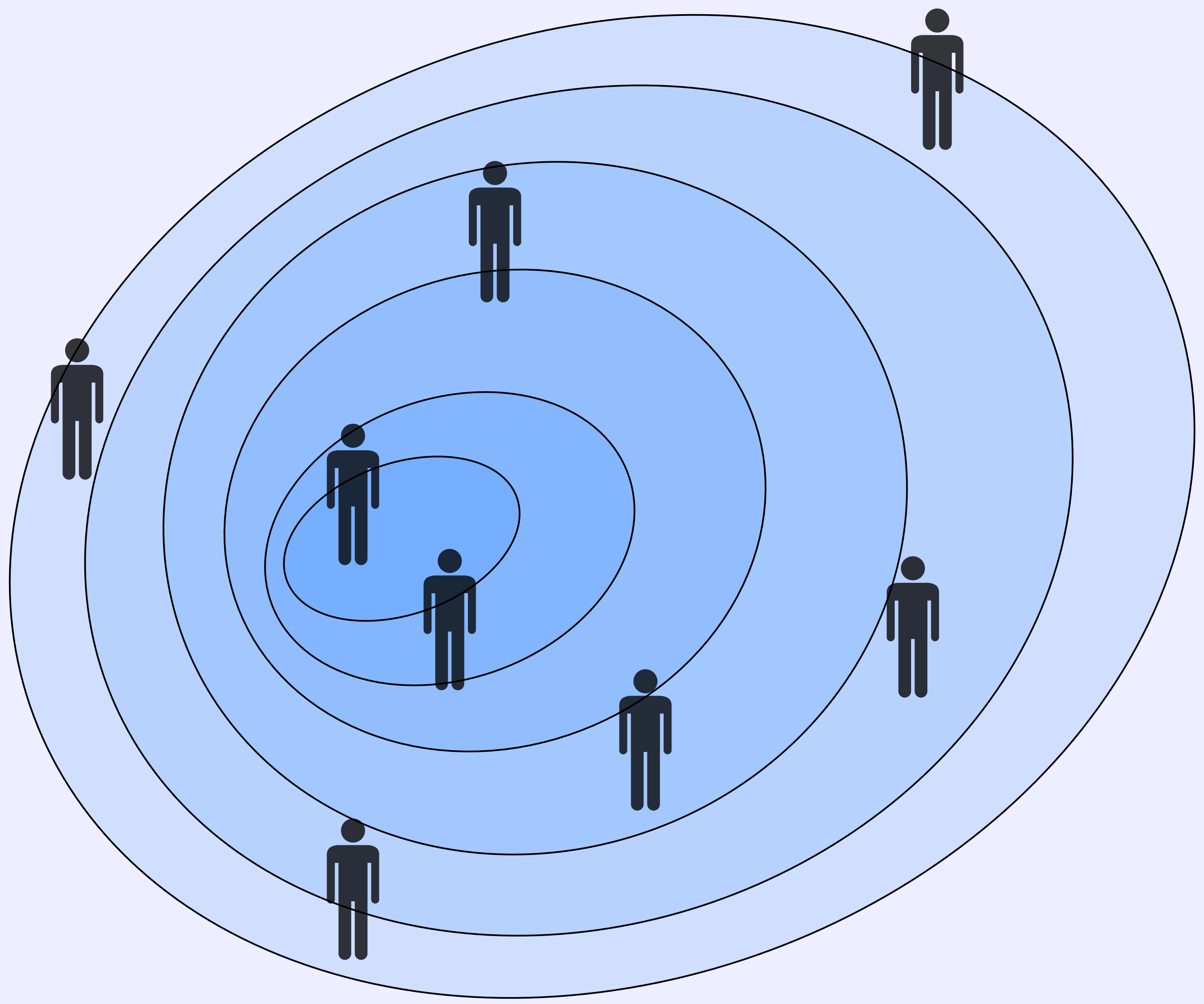
CHALLENGES

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

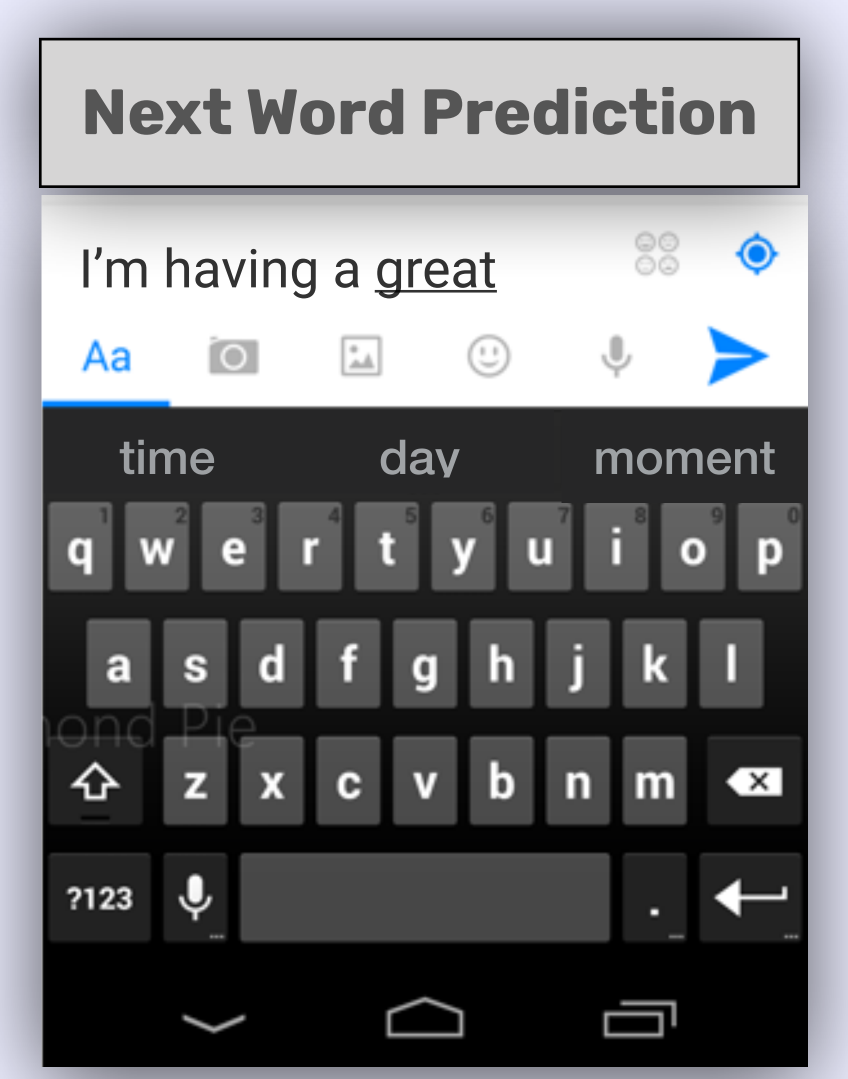
Keep benefits of existing methods

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

- Users heterogeneity



- Eg. on mobile phones



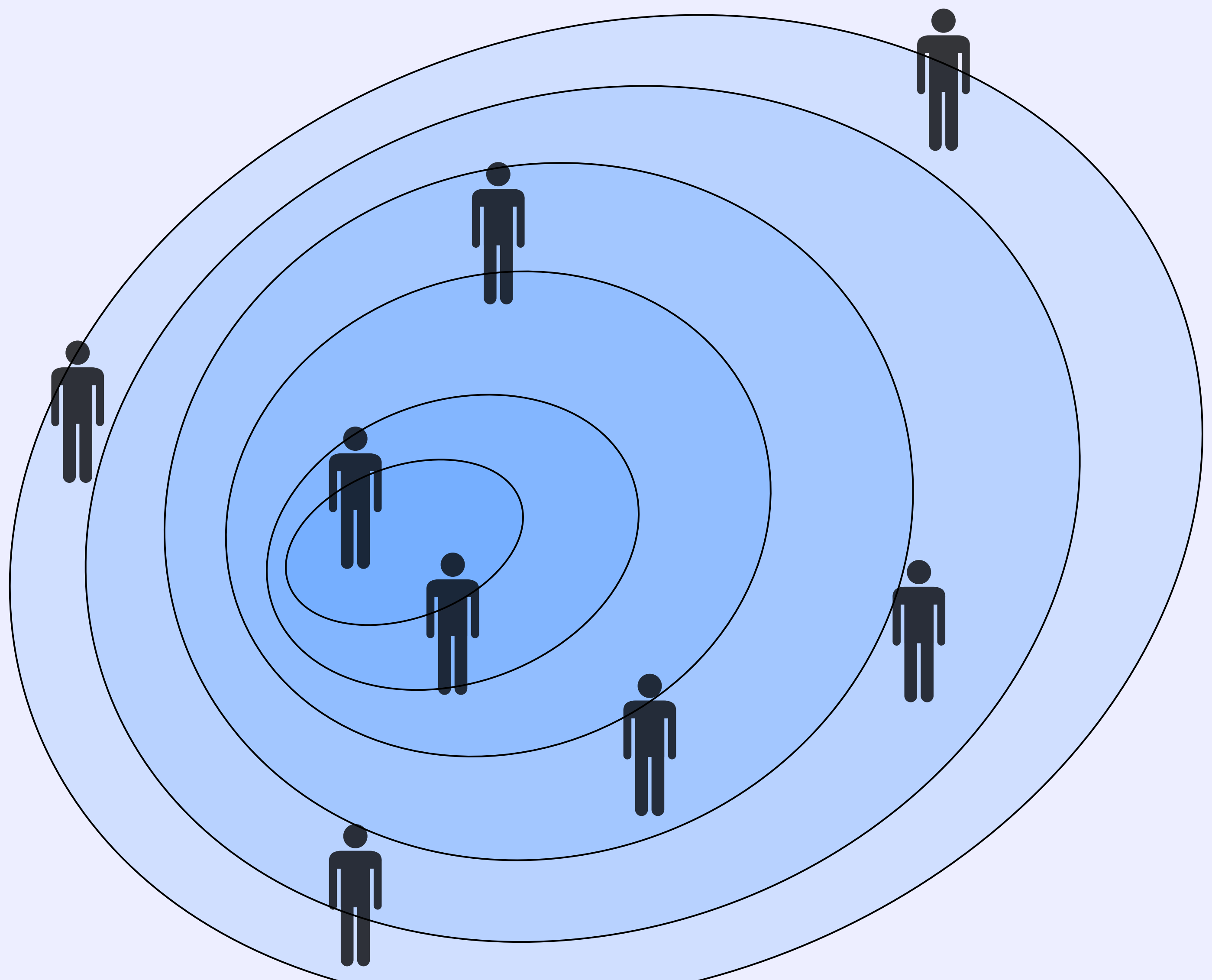
CHALLENGES

- 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

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

CHALLENGES

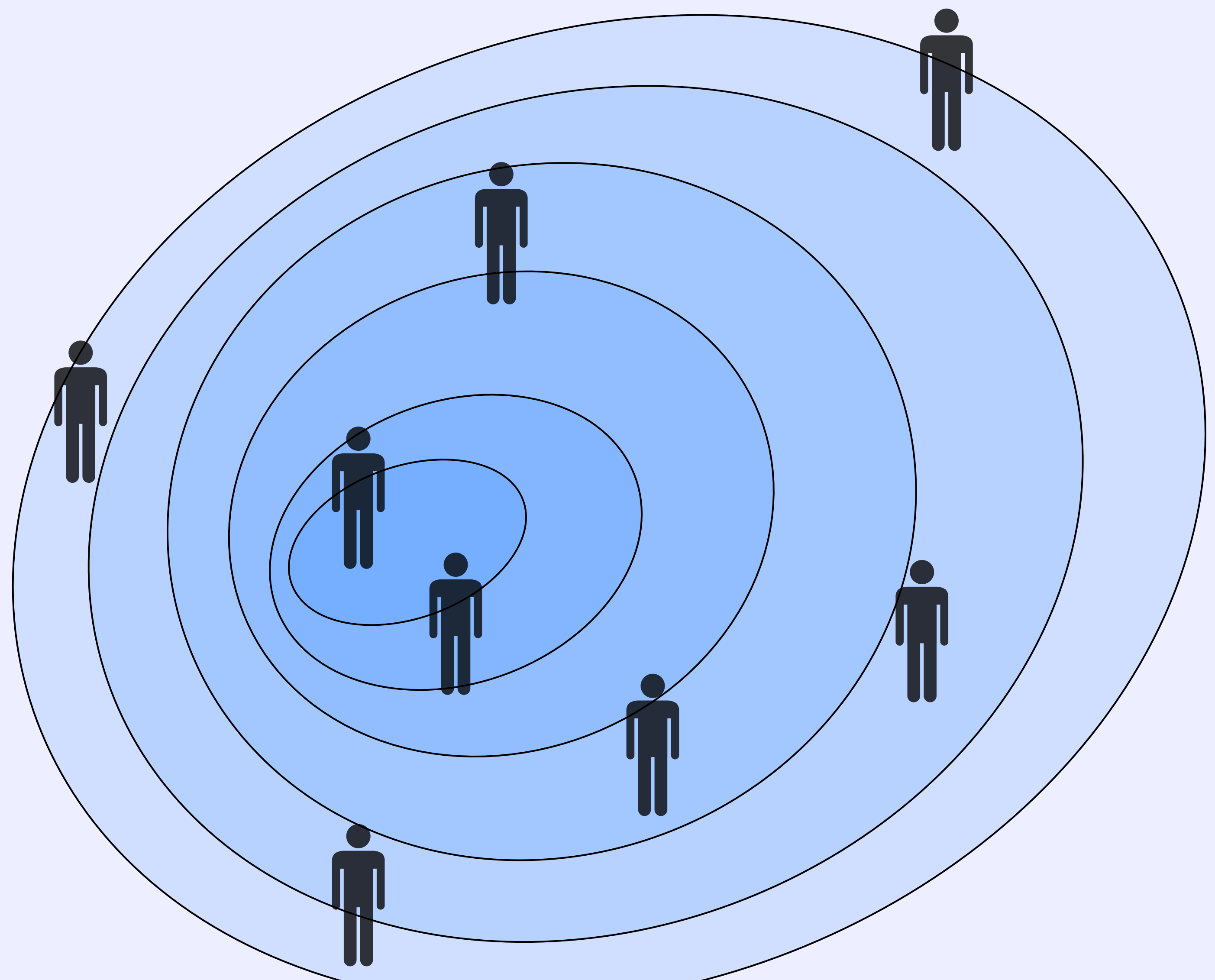
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

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

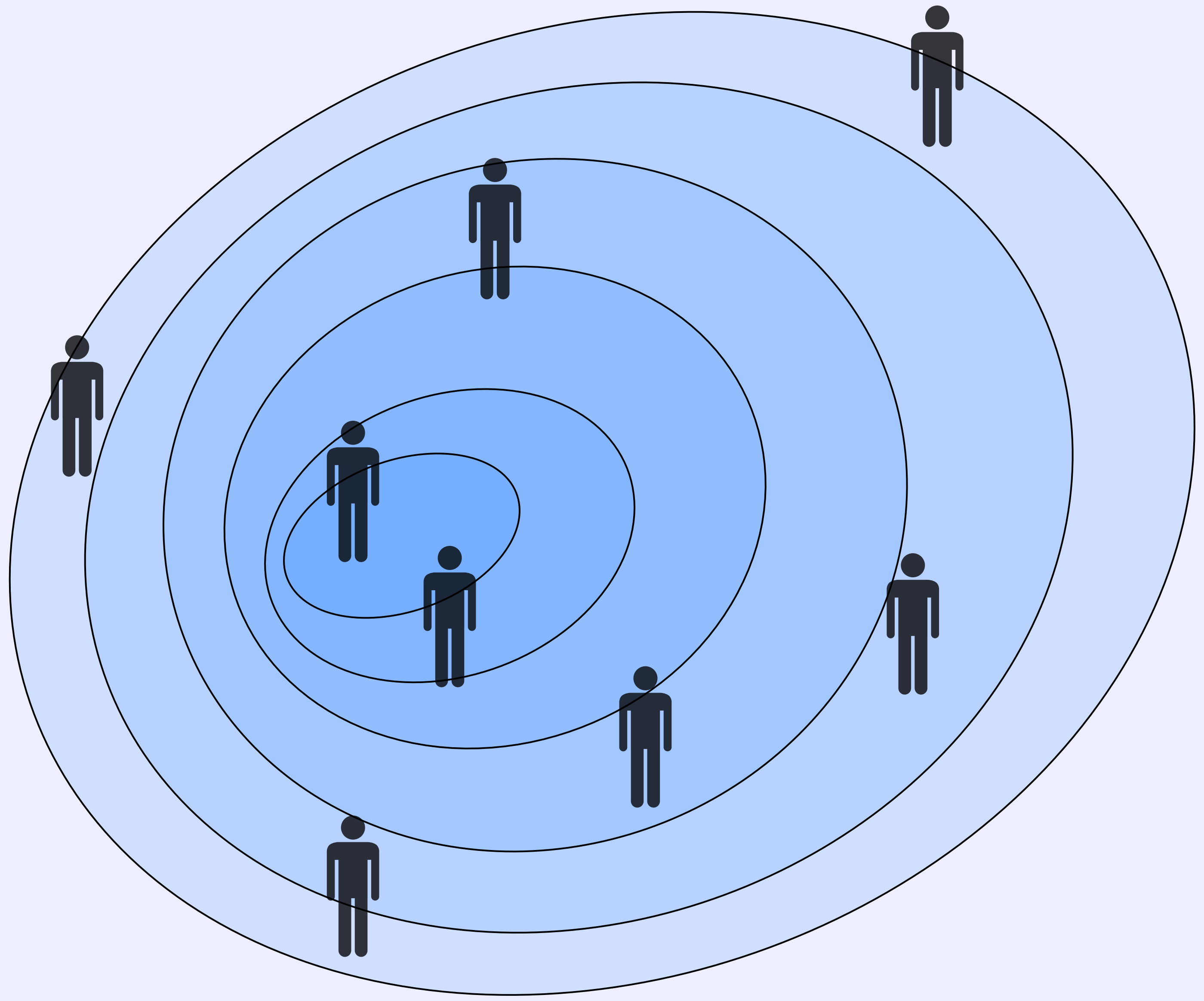
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

Keep benefits of existing methods

Privacy preservation | **Statistical heterogeneity** | 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

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

CHALLENGES

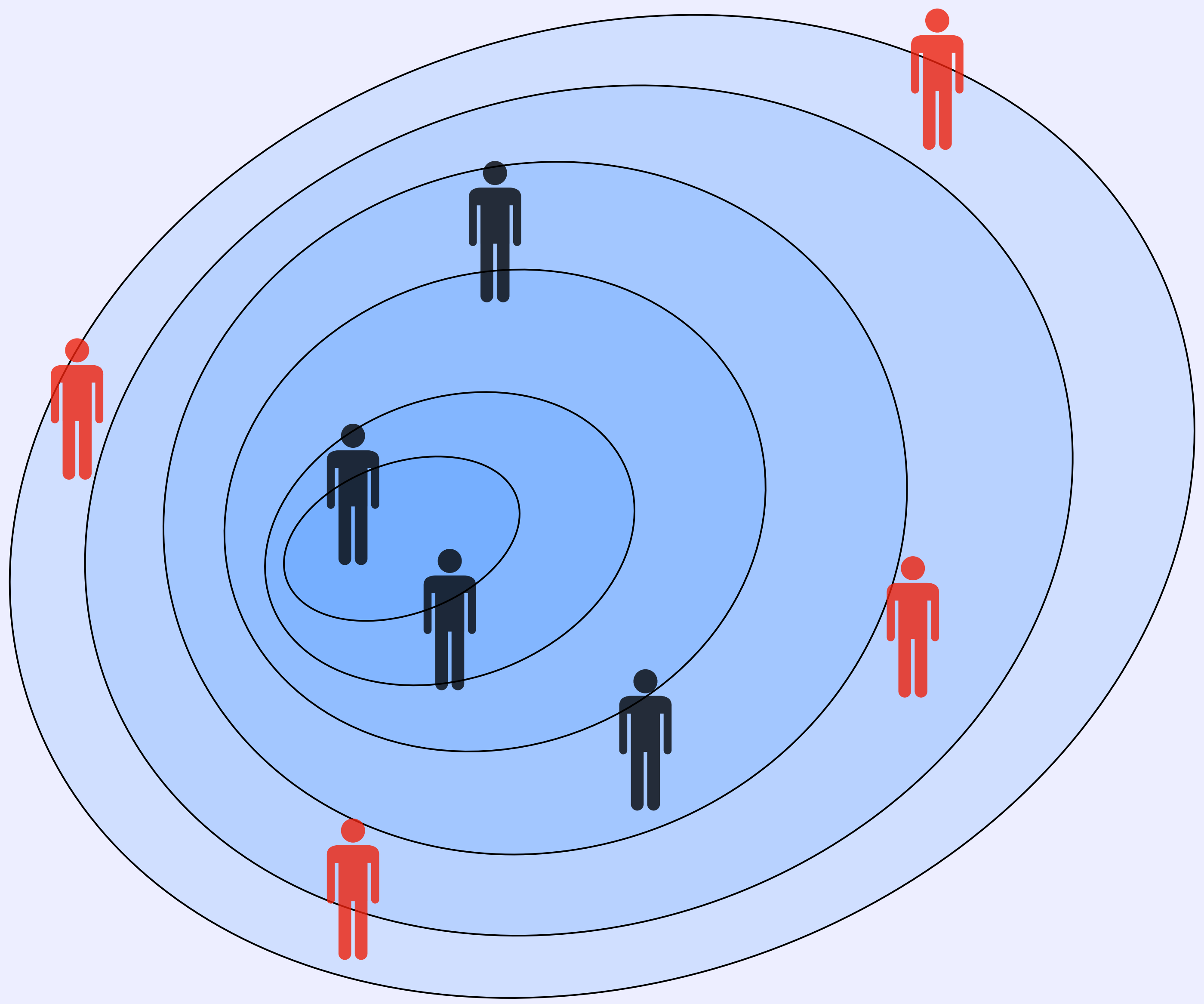
- Challenging Issues [Kairouz et al. 2019'] [Li et al. 2020']

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

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

Our Approach

- We propose to extend this framework to make possible the handling of non-conforming users.

Vanilla Federated Learning

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{\xi \sim p_\alpha}[f(w, \xi)]$$

Our Framework

$$\min_{w \in \mathbb{R}^d} S_\theta[f(w, \xi)]$$

Our Approach

- We propose to extend this framework to make possible the handling of non-conforming users.

Vanilla Federated Learning

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{\xi \sim p_\alpha}[f(w, \xi)]$$

Our Framework

$$\min_{w \in \mathbb{R}^d} S_\theta[f(w, \xi)]$$

Measures conformity of training devices

Outline

1 The Δ-FL Framework

2 Practical Solving

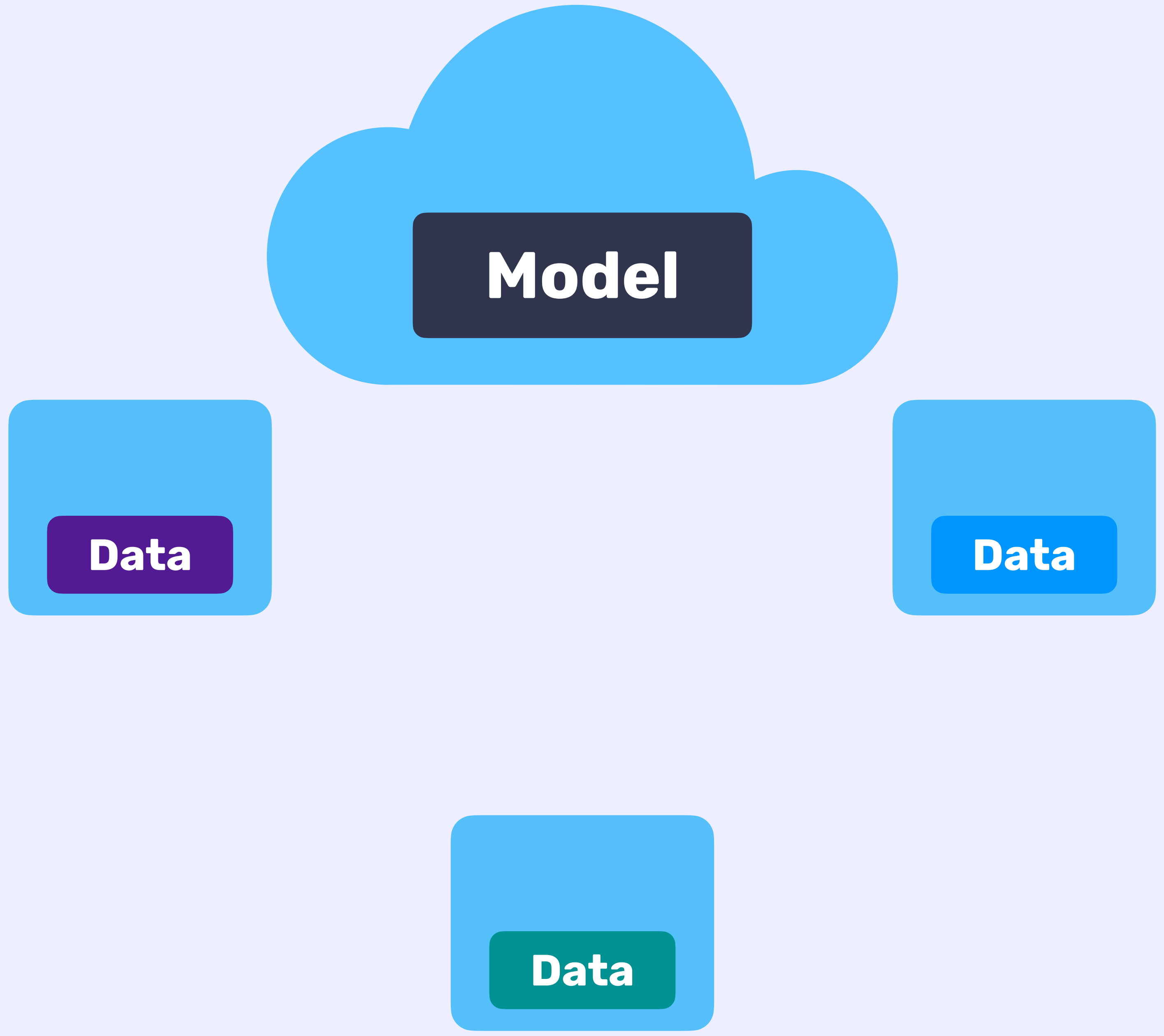
3 Numerical Experiments and Comparisons

1 The Δ-FL Framework

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 $\sum_{i=1}^{N} \alpha_i = 1$

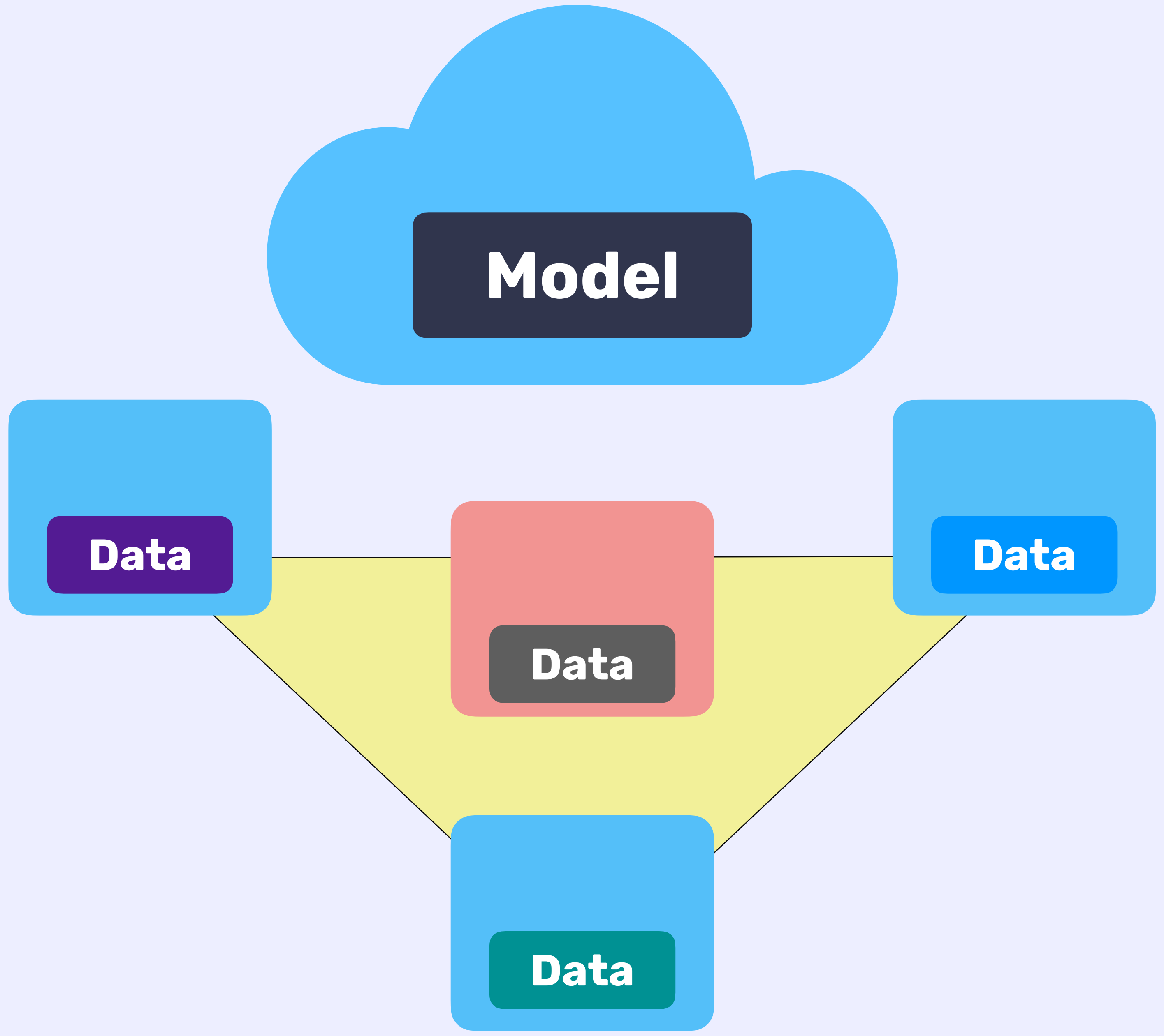
Base distribution $\quad p_\alpha = \sum_{i=1}^{N} \alpha_i \, q_i$



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 $\sum_{i=1}^{N} \alpha_i = 1$

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



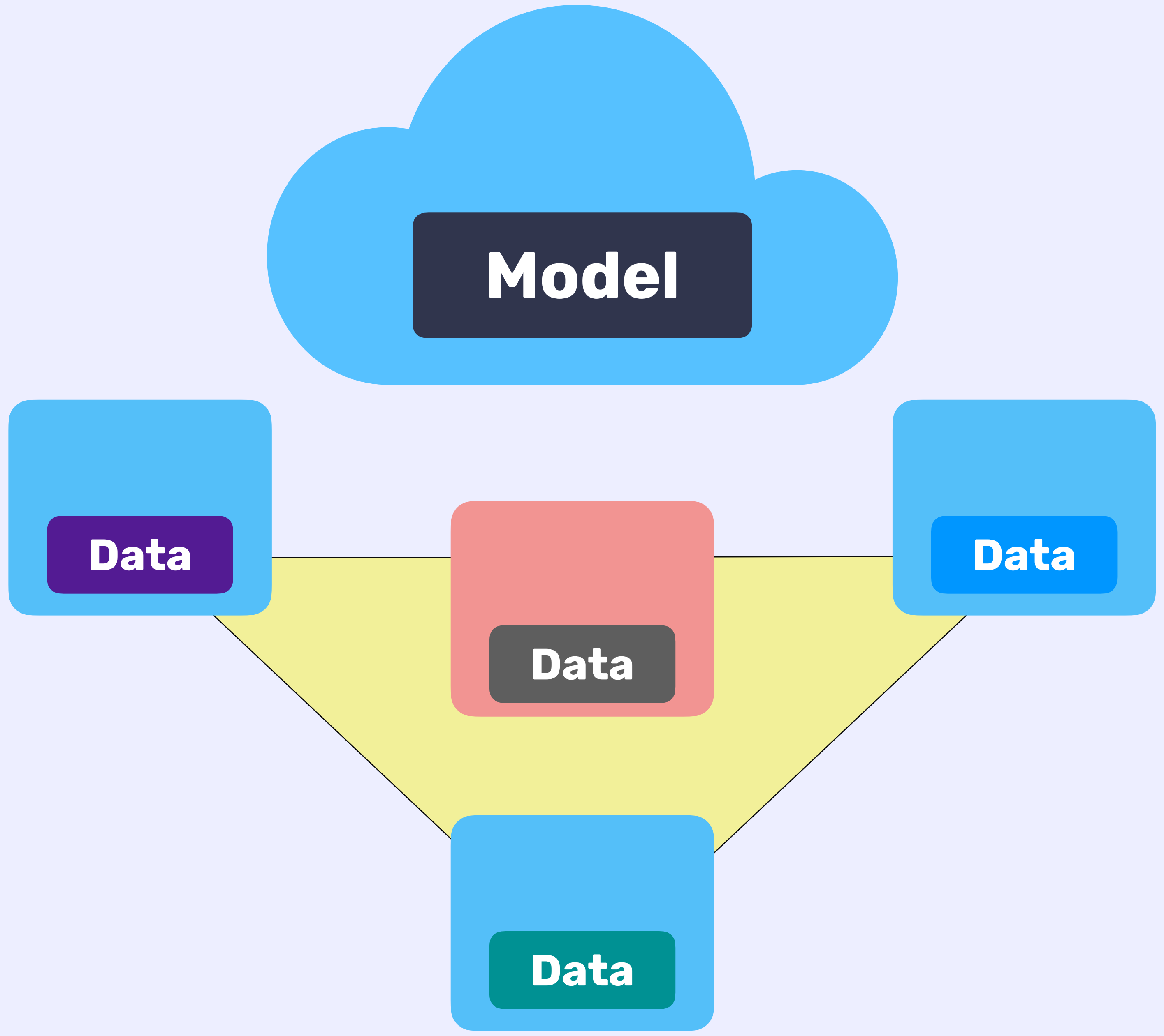
- Measuring conformity on testing devices
  - We consider test devices to have a distribution that can be written as a mixture of the training distributions.

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

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



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

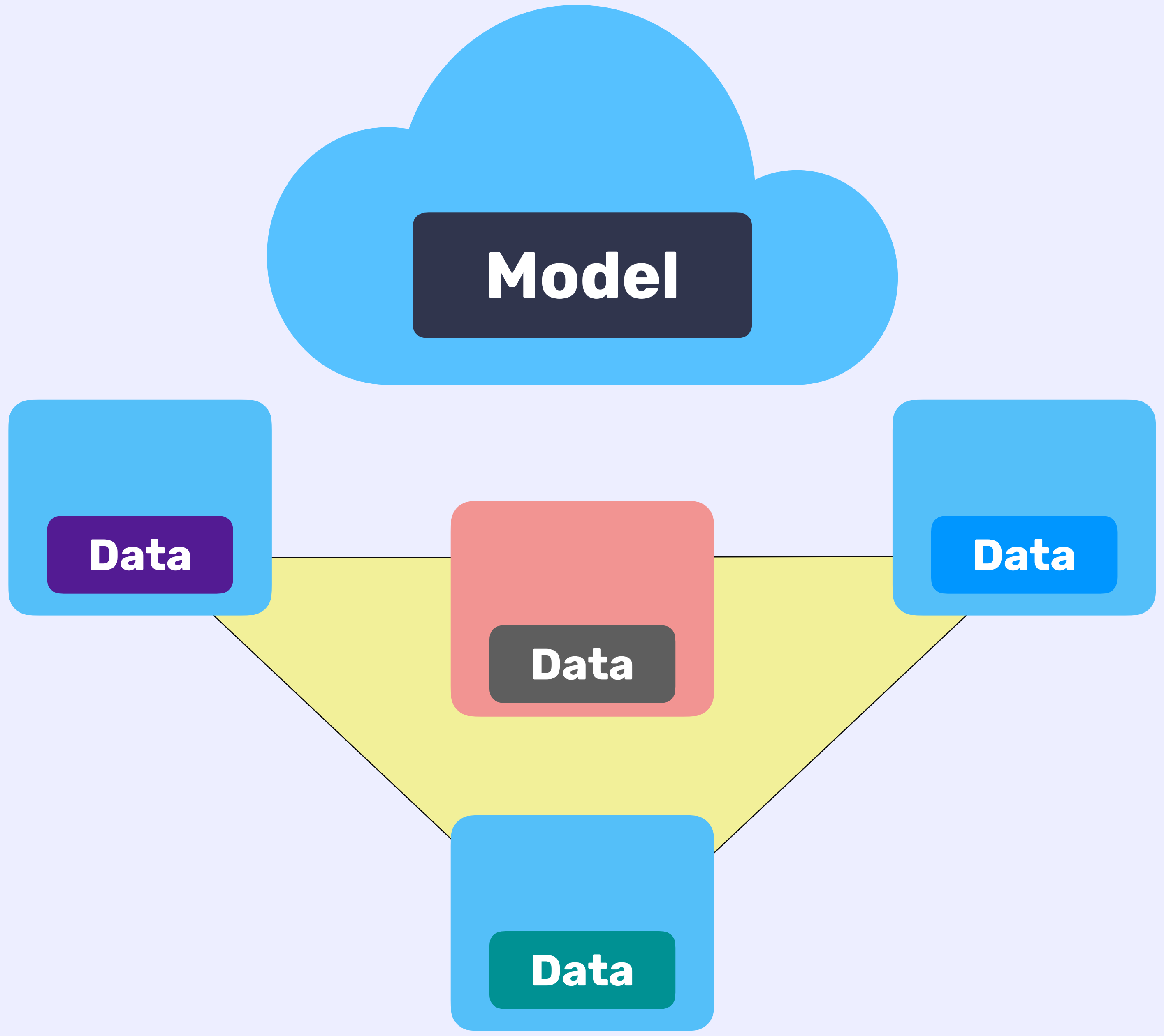
$$p_\pi = \sum_{i=1}^{N} \pi_i \alpha_i \qquad \pi \in \Delta_{N-1} \text{ ie } \begin{cases} 0 \leq \pi_k \leq 1 & \text{for all } 1 \leq i \leq N \\ \sum_{k=1}^{N} \pi_k = 1 \end{cases}$$

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

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

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

The Δ-FL Framework

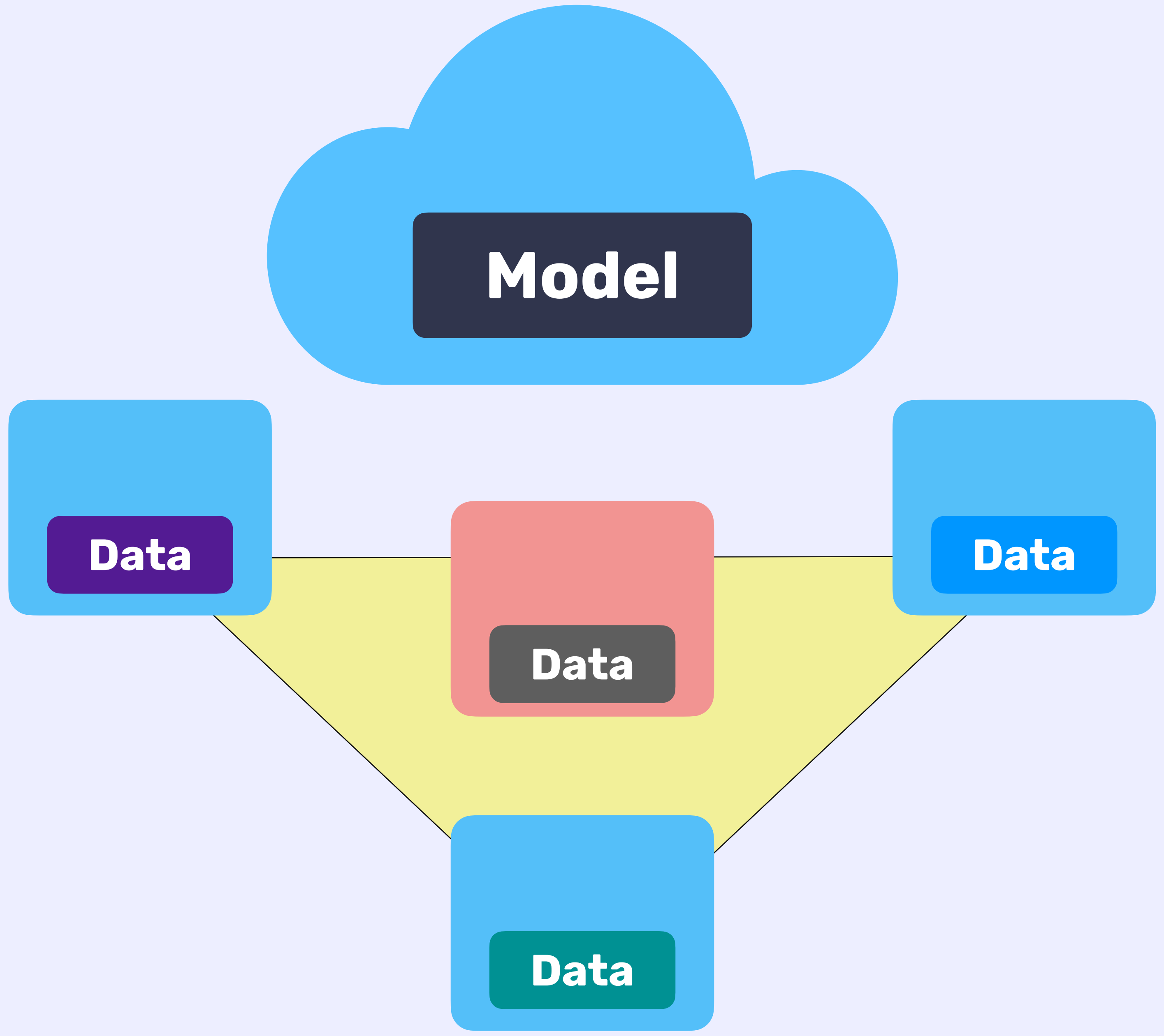


The Δ-FL Framework

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

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

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



The Δ-FL Framework

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

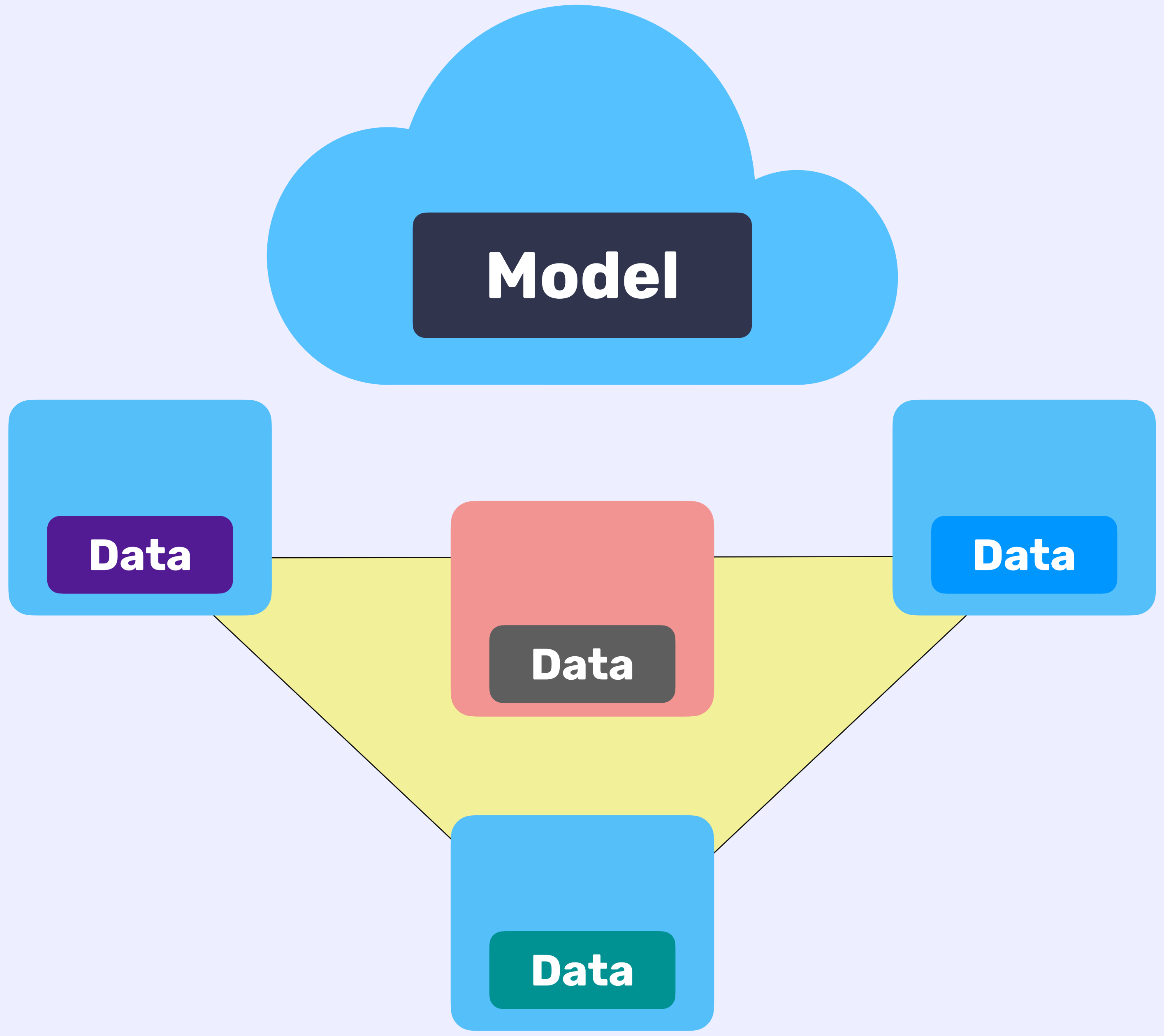
$$\min_{w \in \mathbb{R}^d} \left[F_\theta(w) = \max_{\pi \in \mathcal{P}_\theta} \mathbb{E}_{\xi \sim p_\pi}[f(w, \xi)] \right] \quad \text{where}$$

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

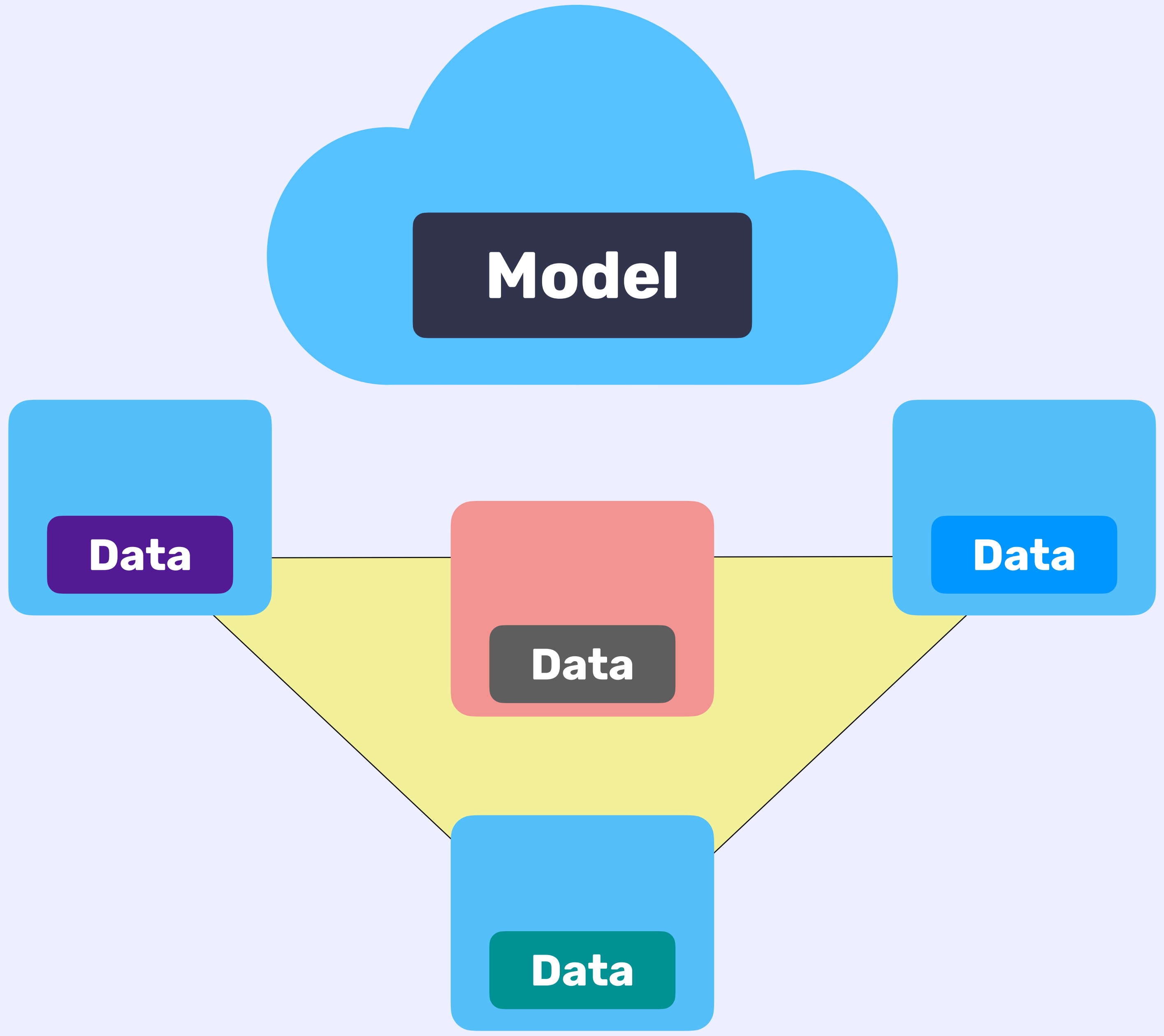
Superquantile loss

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

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



The Δ-FL Framework



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

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

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

Superquantile loss

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

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

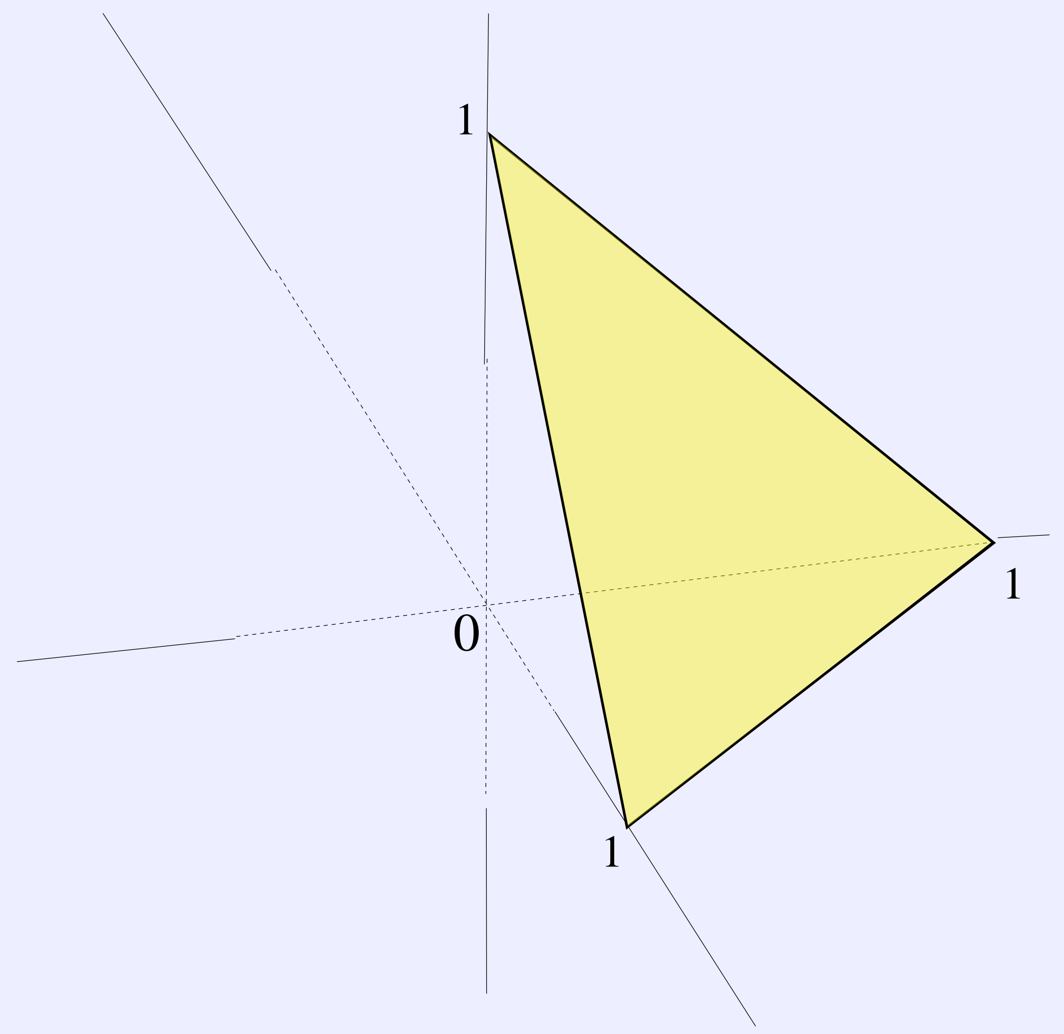
  - In Δ-FL, we are using the superquantile at a user level

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

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

Geometrical Intuition

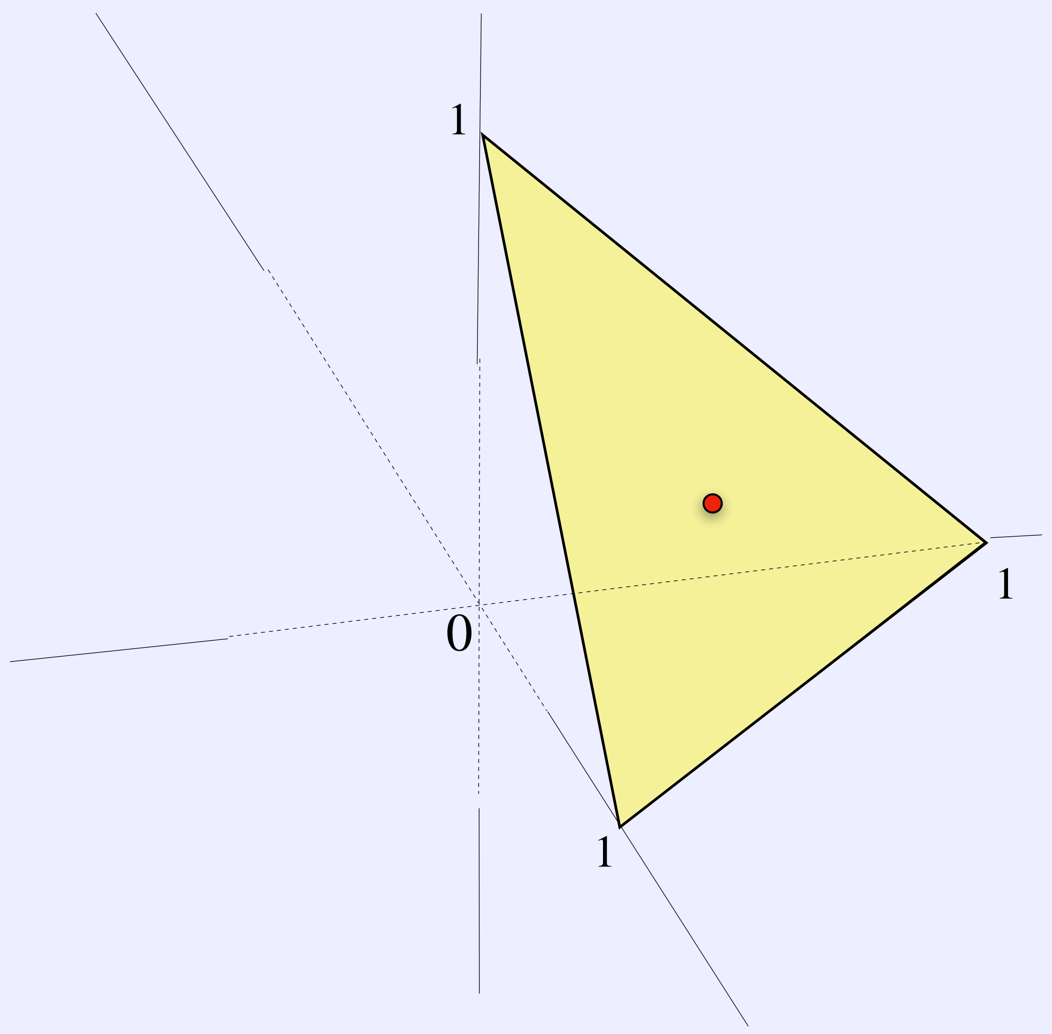
- Assume we have only three users at training time



Geometrical Intuition

- Assume we have only three users at training time

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

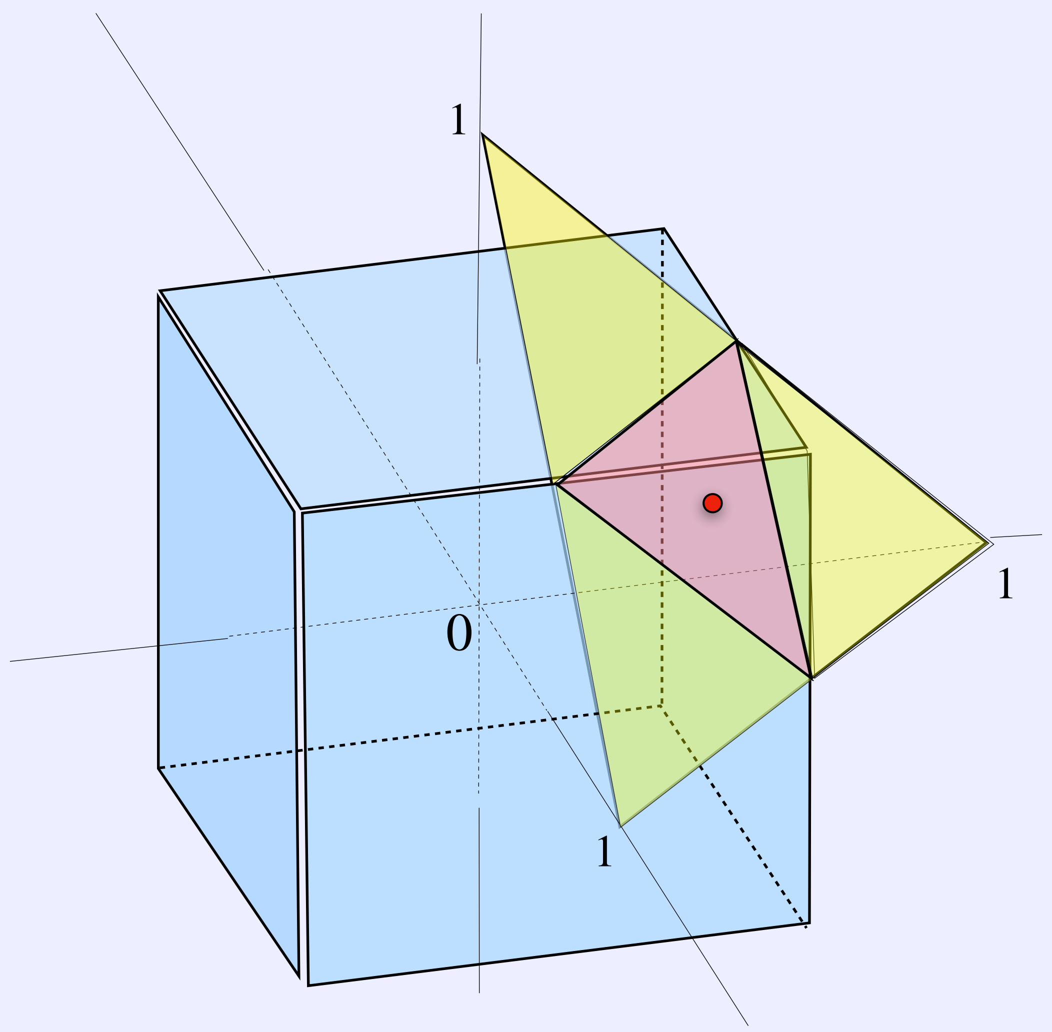


Geometrical Intuition

- Assume we have only three users at training time

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

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



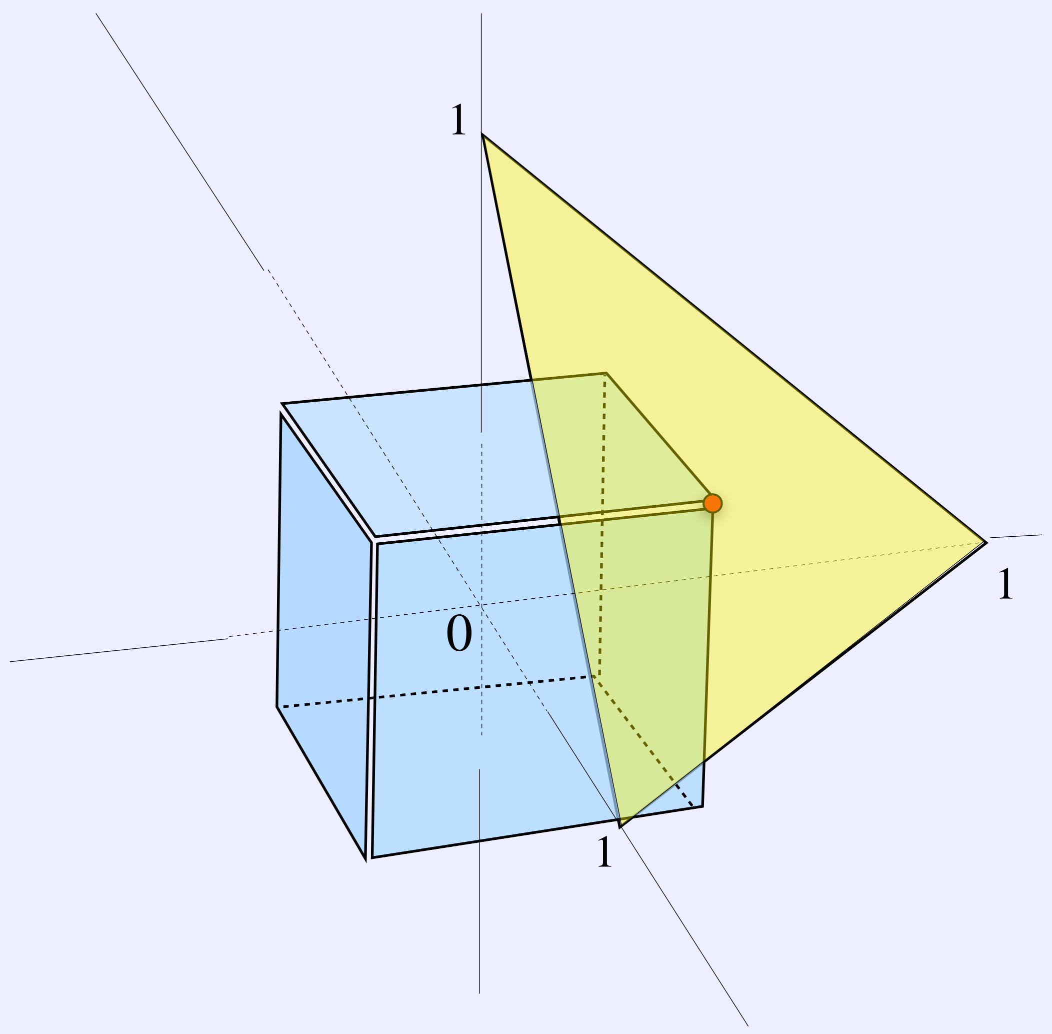
$$r = \min_{1 \leq i \leq N} \frac{\alpha_i}{\theta}$$

Geometrical Intuition

- Assume we have only three users at training time

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

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



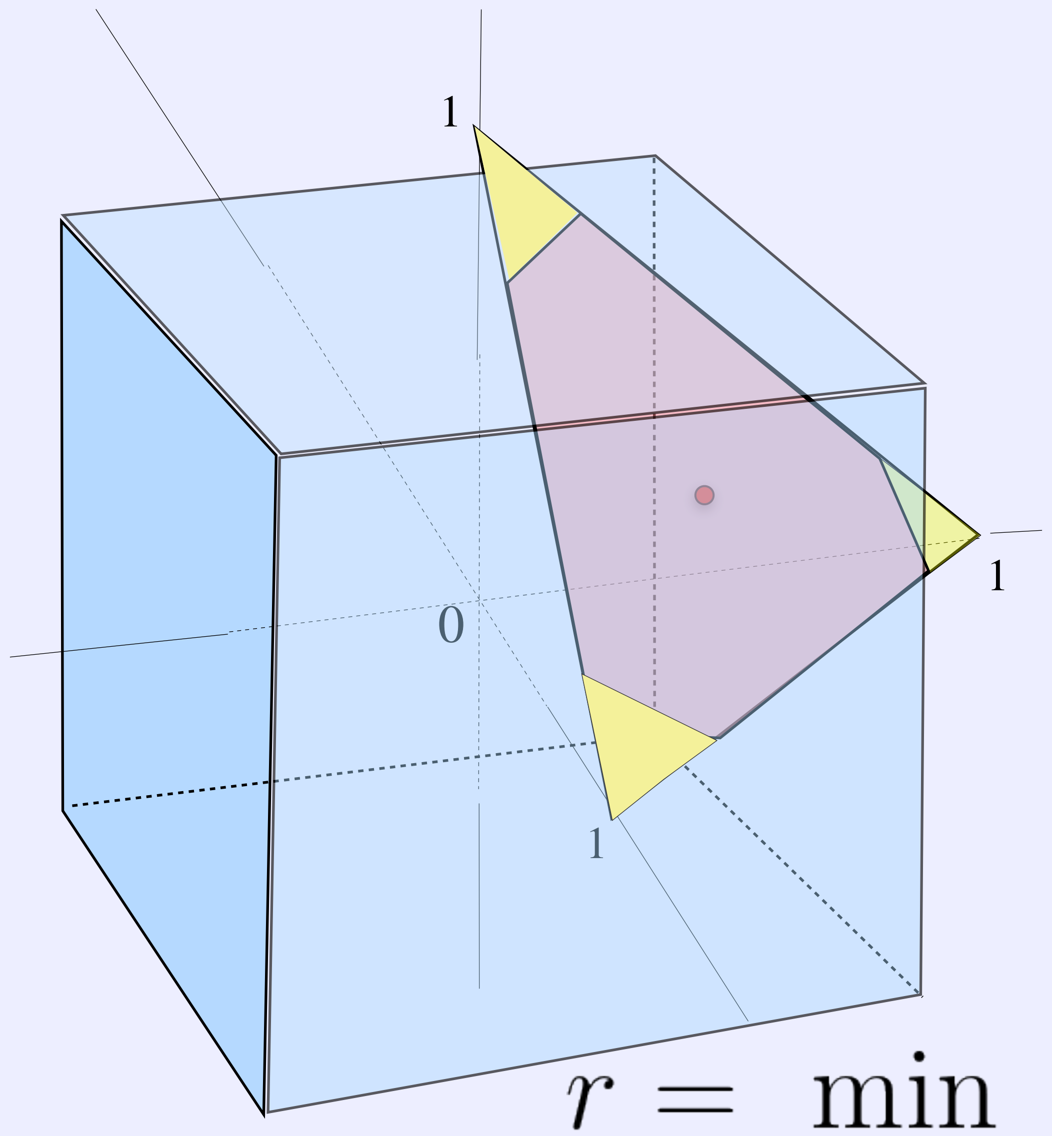
$$r = \min_{1 \le i \le N} \frac{\alpha_i}{\theta}$$

Geometrical Intuition

- Assume we have only three users at training time

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

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



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

Rewriting the superquantile with quantiles

- Let us fix a conformity level $\theta \in (0, 1]$,

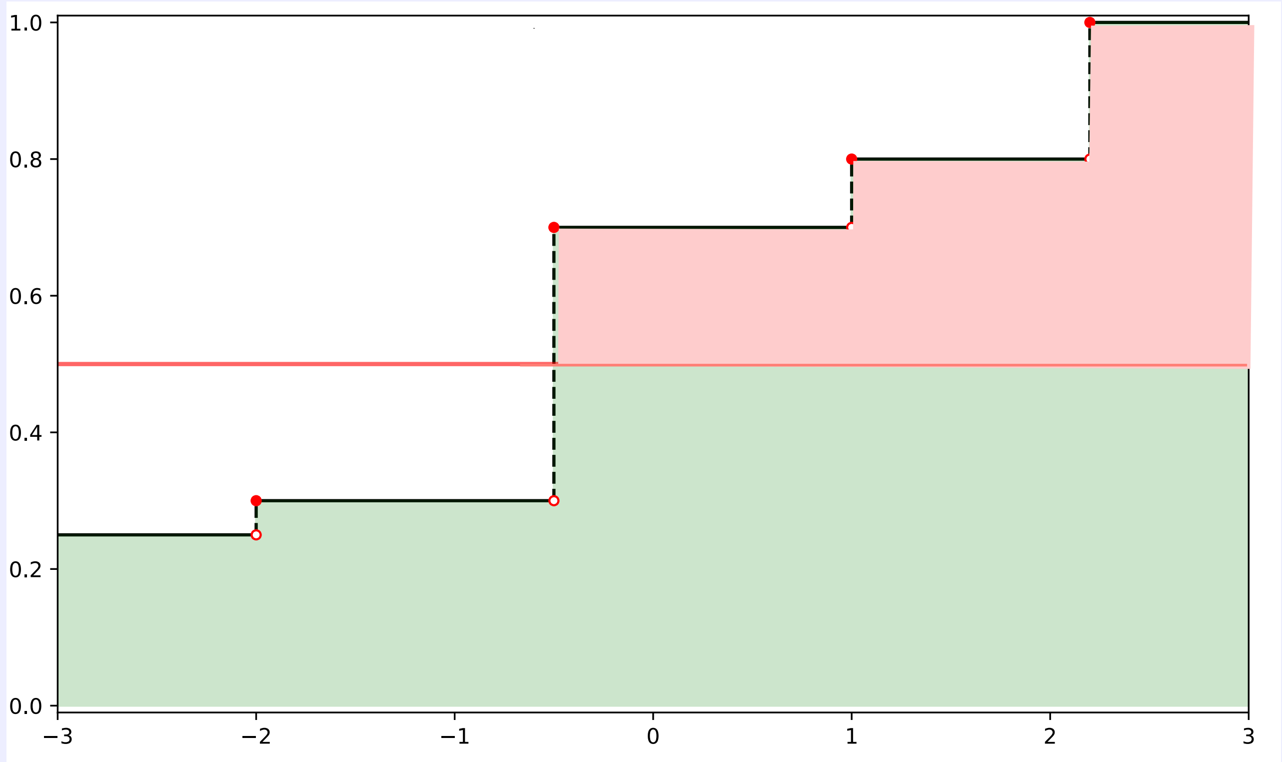
$$F_\theta(w) = S_\theta(F_{\mathbf{k}}(w)) \qquad U = F_{\mathbf{k}}(w) \qquad \text{with } \mathbb{P}[\mathbf{k} = i] = \alpha_i$$

Rewriting the superquantile with quantiles

- Let us fix a conformity level $\theta \in (0, 1]$,

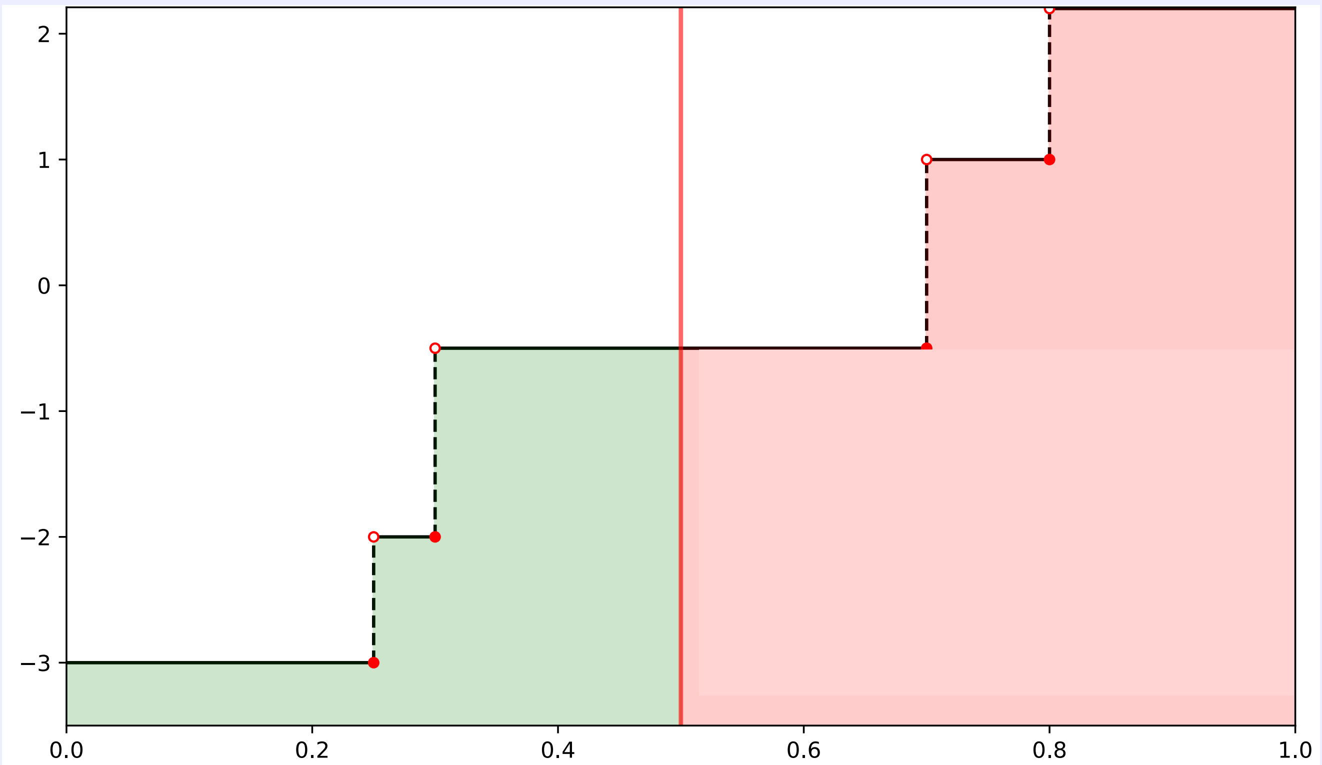
$$F_\theta(w) = S_\theta(F_{\mathbf{k}}(w)) \qquad U = F_{\mathbf{k}}(w) \qquad \text{with } \mathbb{P}[\mathbf{k} = i] = \alpha_i$$

Cumulative distribution function of U



$$F_U(t) = \mathbb{P}[U \leq t]$$

Quantile function of U



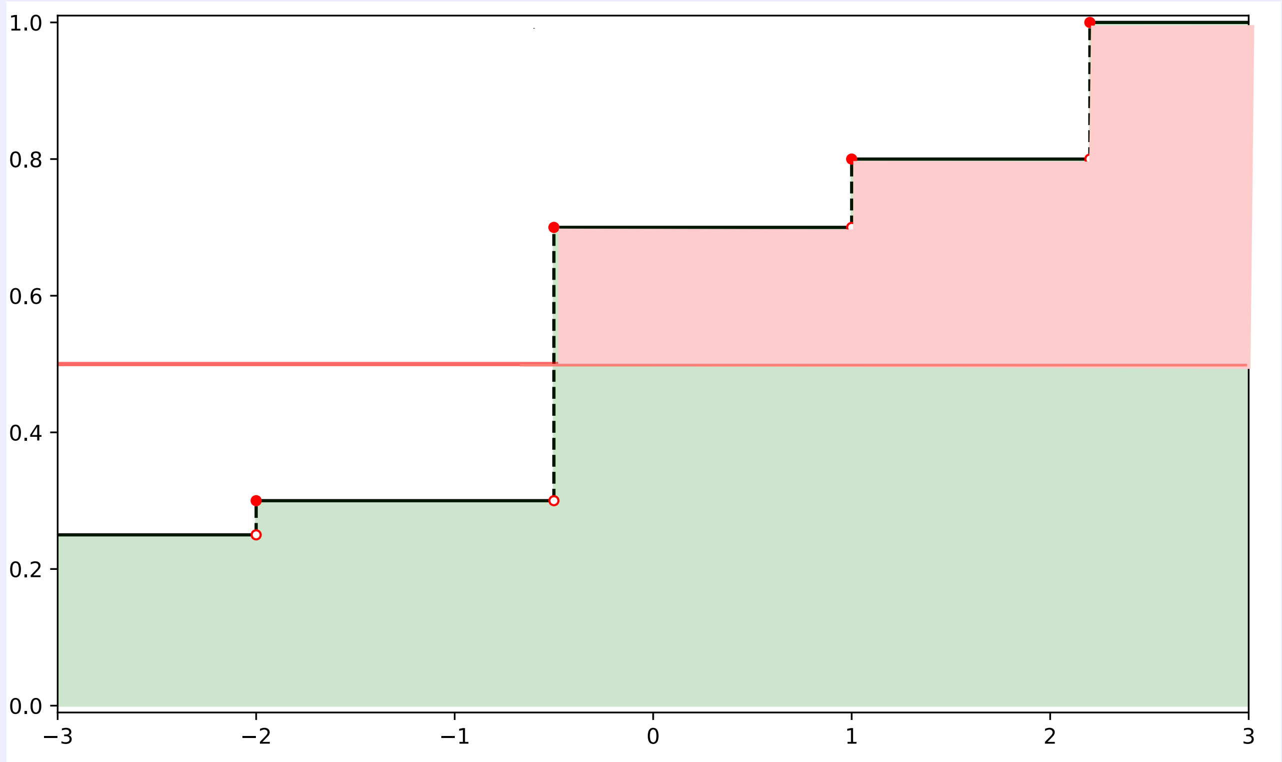
$$Q_p(U) = \inf\{t \in \mathbb{R}, F_U(t) \geq p\}$$

Rewriting the superquantile with quantiles

- Let us fix a conformity level $\theta \in (0, 1]$,

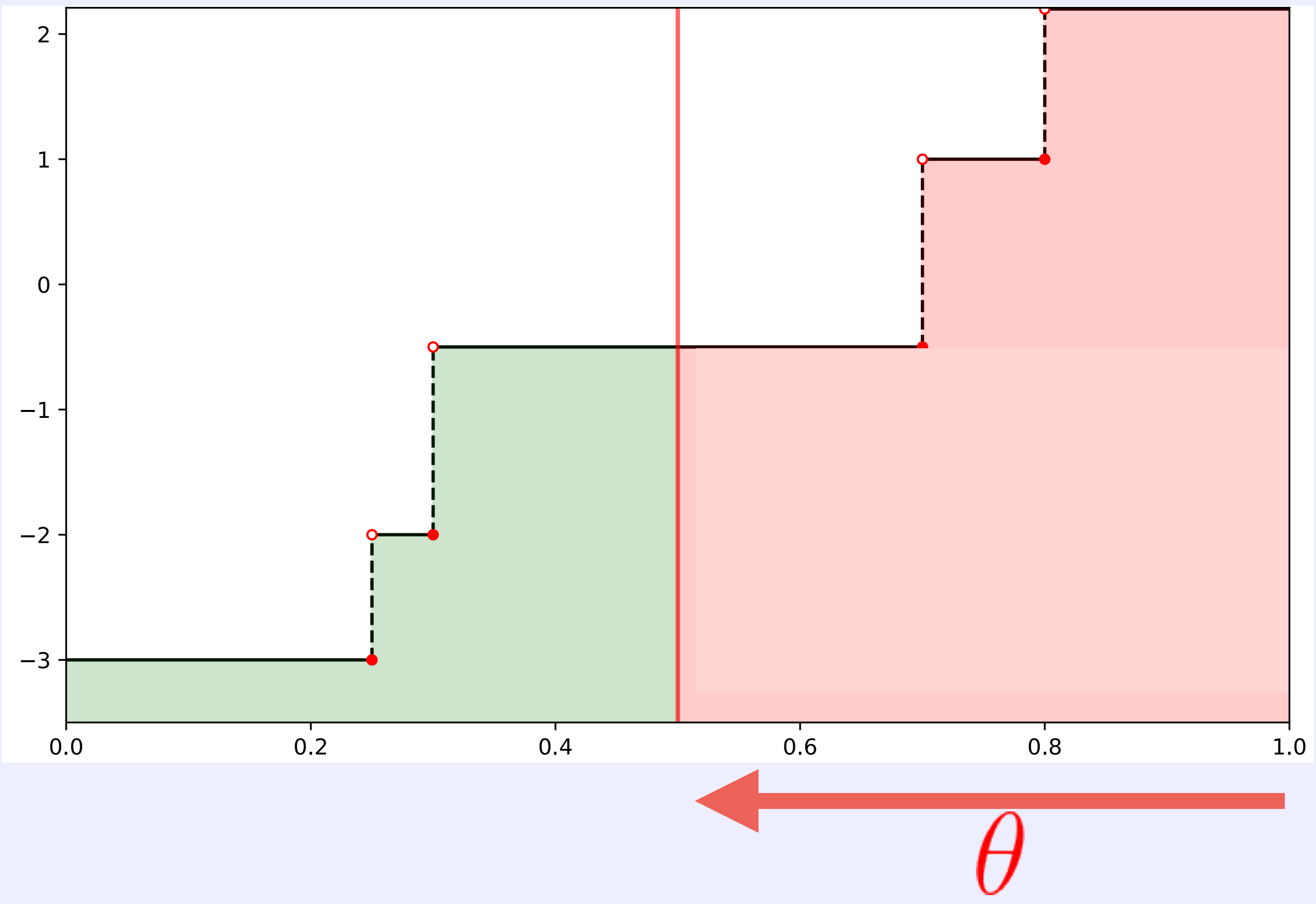
$$F_\theta(w) = S_\theta(F_{\mathbf{k}}(w)) \qquad U = F_{\mathbf{k}}(w) \qquad \text{with } \mathbb{P}[\mathbf{k} = i] = \alpha_i$$

Cumulative distribution function of U



$$F_U(t) = \mathbb{P}[U \leq t]$$

Quantile function of U



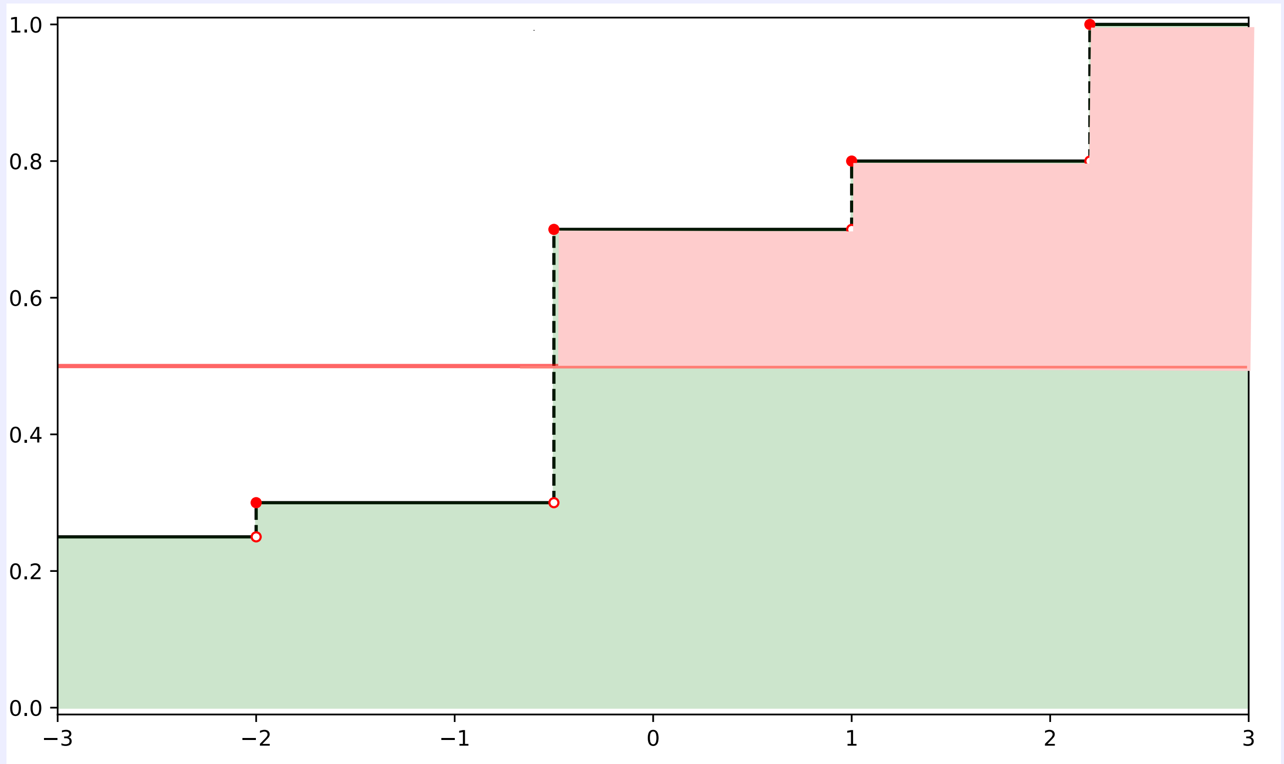
$$Q_p(U) = \inf\{t \in \mathbb{R}, F_U(t) \geq p\}$$

Rewriting the superquantile with quantiles

- Let us fix a conformity level $\theta \in (0, 1]$,

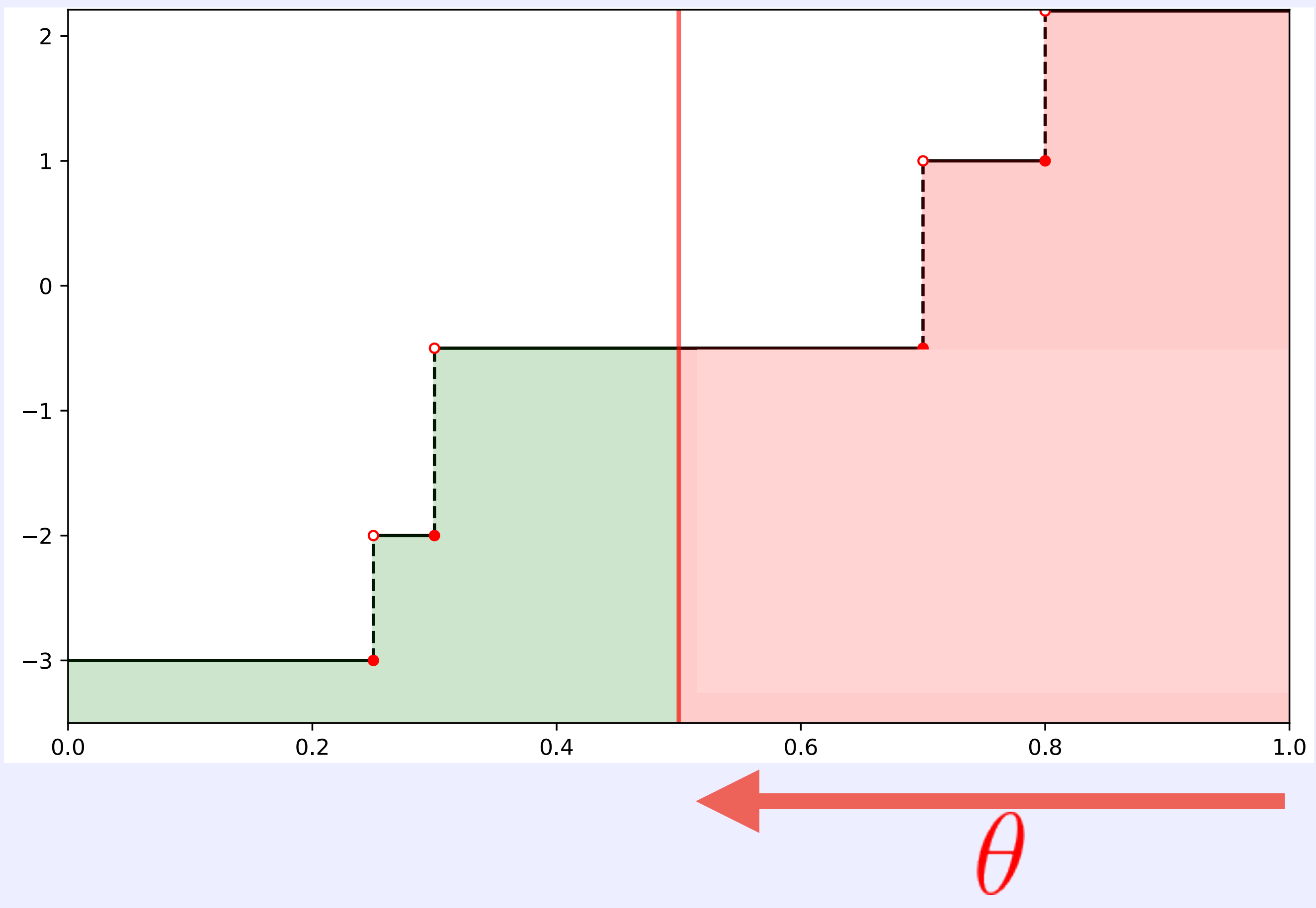
$$F_\theta(w) = S_\theta(F_{\mathbf{k}}(w)) \qquad U = F_{\mathbf{k}}(w) \qquad \text{with } \mathbb{P}[\mathbf{k} = i] = \alpha_i$$

Cumulative distribution function of U



$$F_U(t) = \mathbb{P}[U \le t]$$

Quantile function of U



$$Q_p(U) = \inf\{t \in \mathbb{R}, F_U(t) \ge p\}$$

$$S_\theta(U) = \frac{1}{\theta} \int_{p=1-\theta}^{1} Q_p(U) \mathrm{d}p$$

$$S_1(U) = \mathbb{E}[U]$$

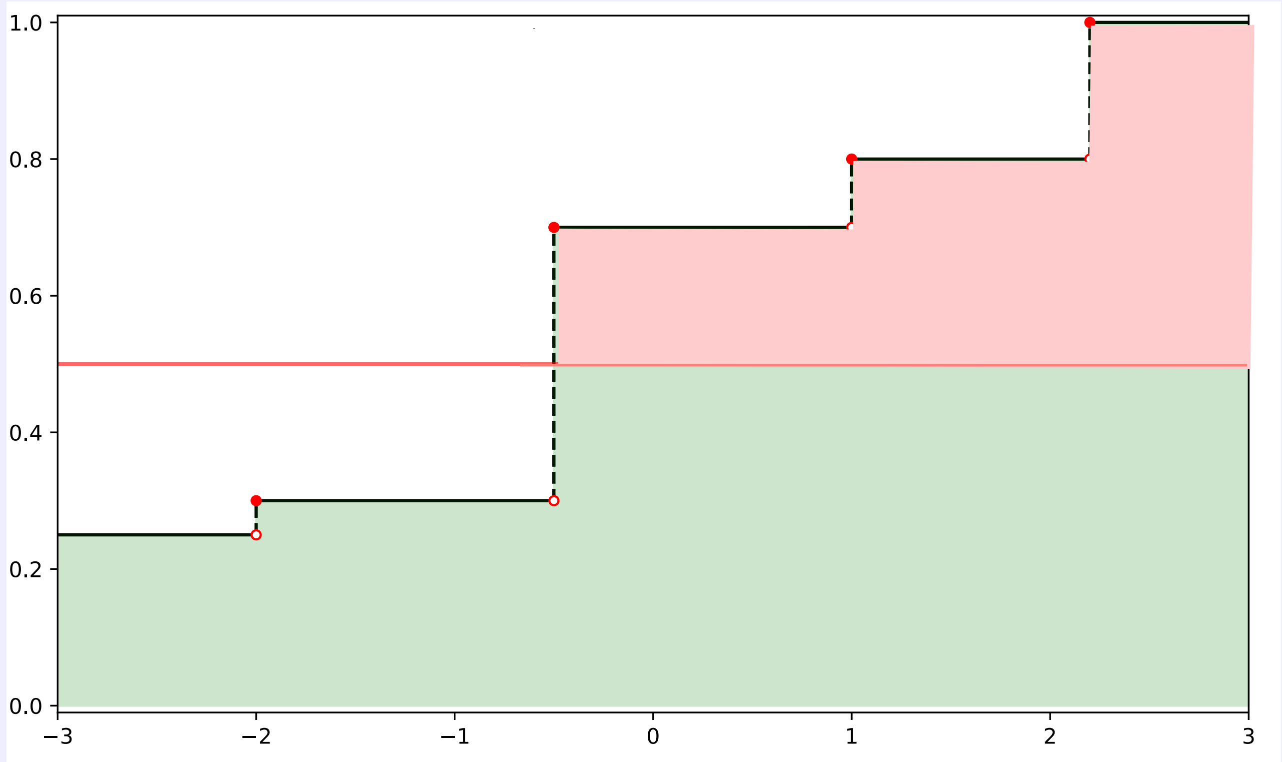
$$S_0(U) = \max(U)$$

Rewriting the superquantile with quantiles

- Let us fix a conformity level $\theta \in (0, 1]$,

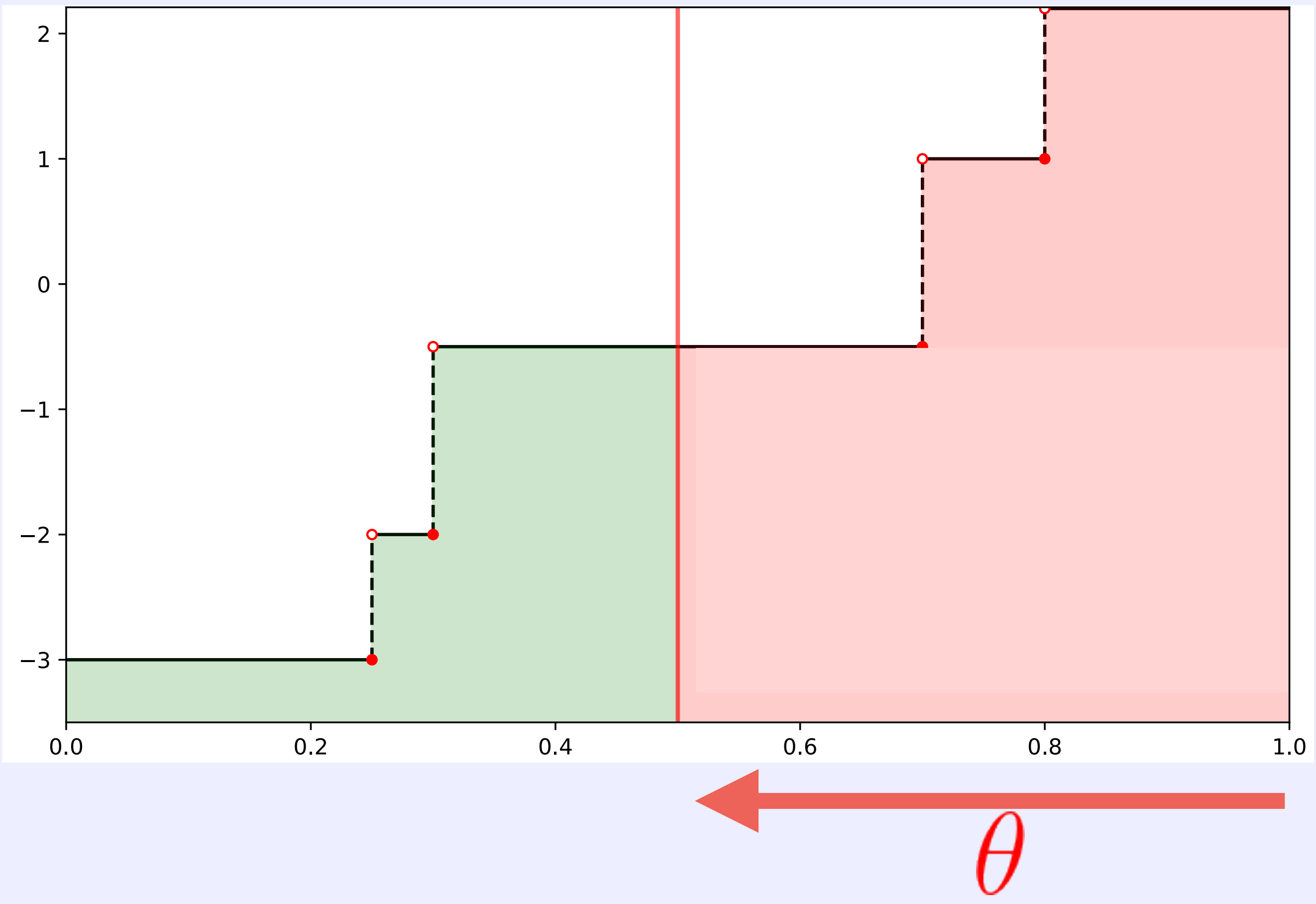
$$F_\theta(w) = S_\theta(F_{\mathbf{k}}(w)) \qquad U = F_{\mathbf{k}}(w) \qquad \text{with } \mathbb{P}[\mathbf{k} = i] = \alpha_i$$

Cumulative distribution function of U



$$F_U(t) = \mathbb{P}[U \leq t]$$

Quantile function of U

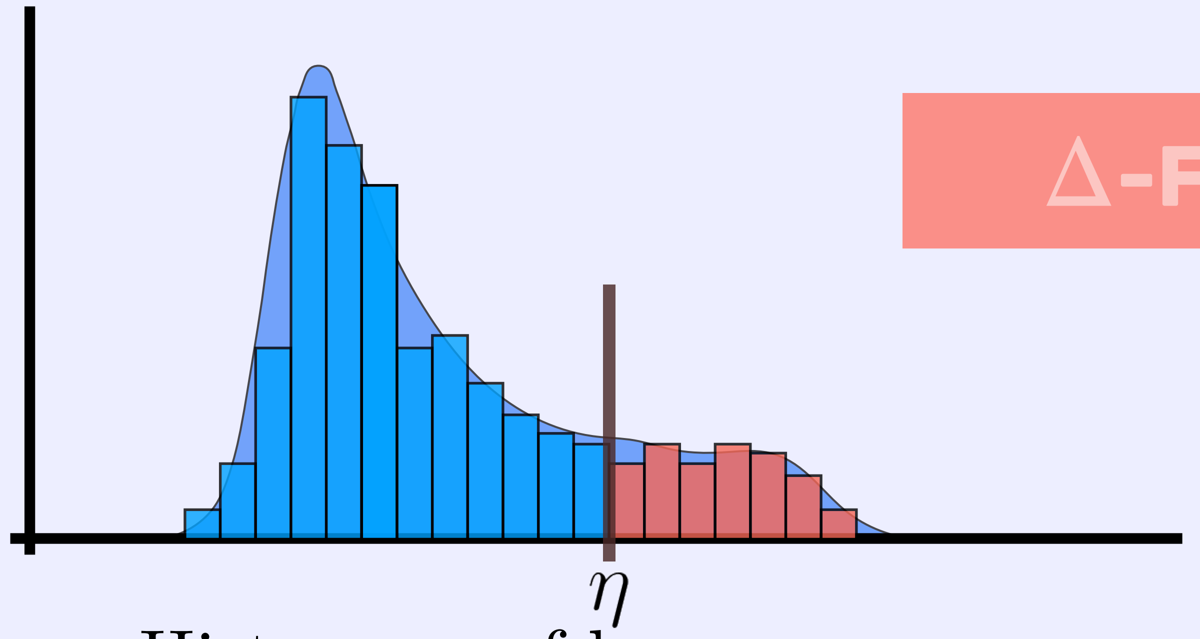


$$Q_p(U) = \inf\{t \in \mathbb{R}, F_U(t) \geq p\}$$

$$S_\theta(U) = \frac{1}{\theta} \int_{p=1-\theta}^{1} Q_p(U) \mathrm{d}p$$

$$S_1(U) = \mathbb{E}[U]$$

$$S_0(U) = \max(U)$$



Histogram of losses over users

Δ-FL

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,

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

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

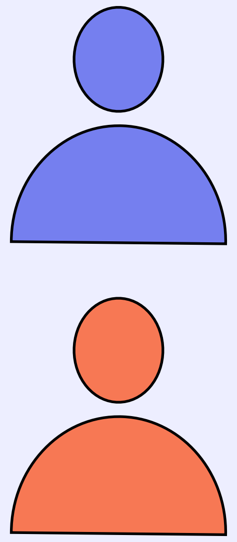
$= 1 - \theta$

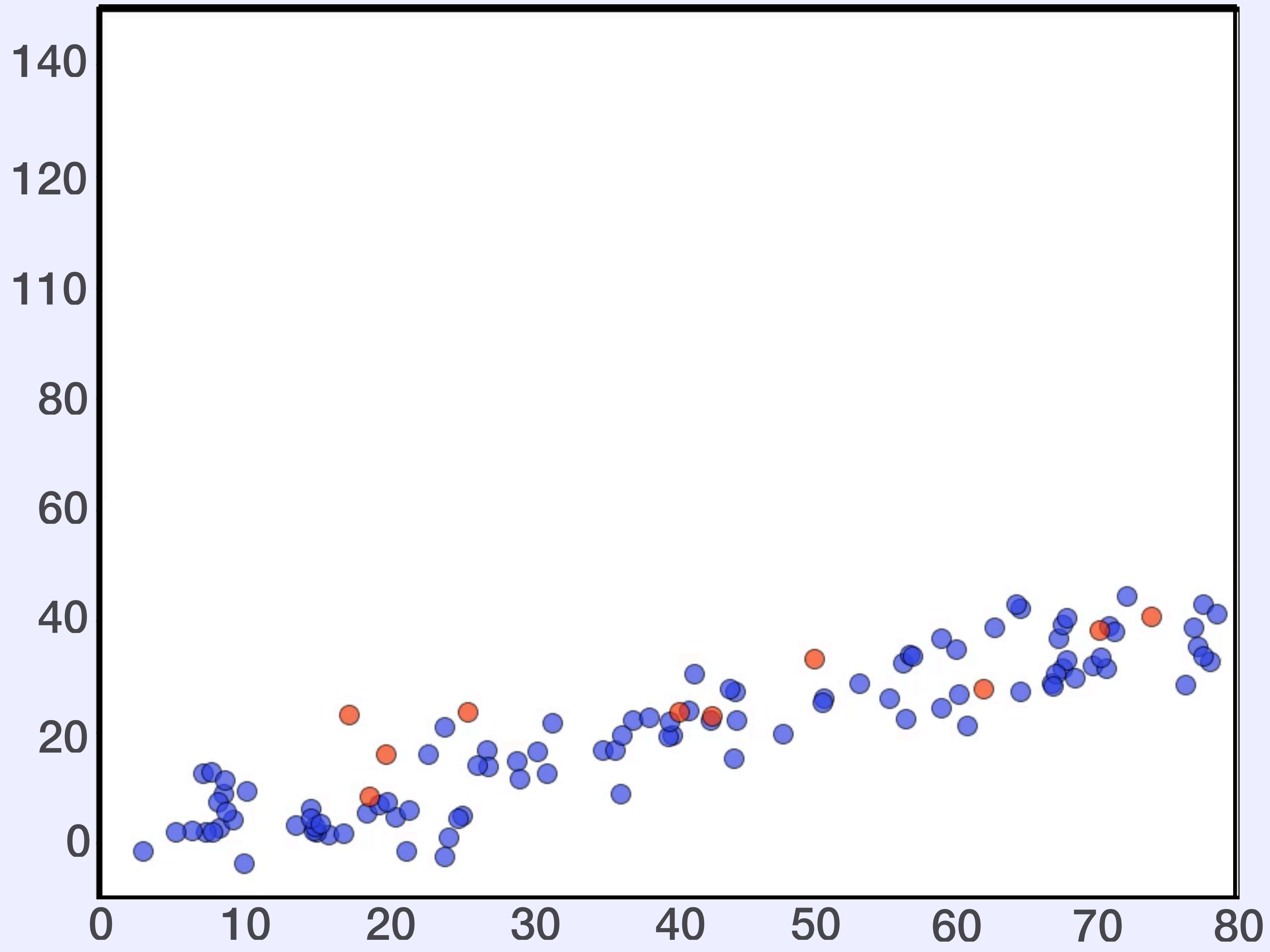
  - In our case, we can rewrite Δ-FL's objective as a joint minimization problem:

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

Toy Problem 1

- A centralized problem: least squares regression $\min_{w \in \mathbb{R}^d} \left\| Y - w^\top X \right\|^2$

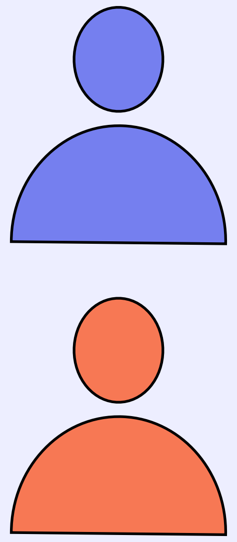


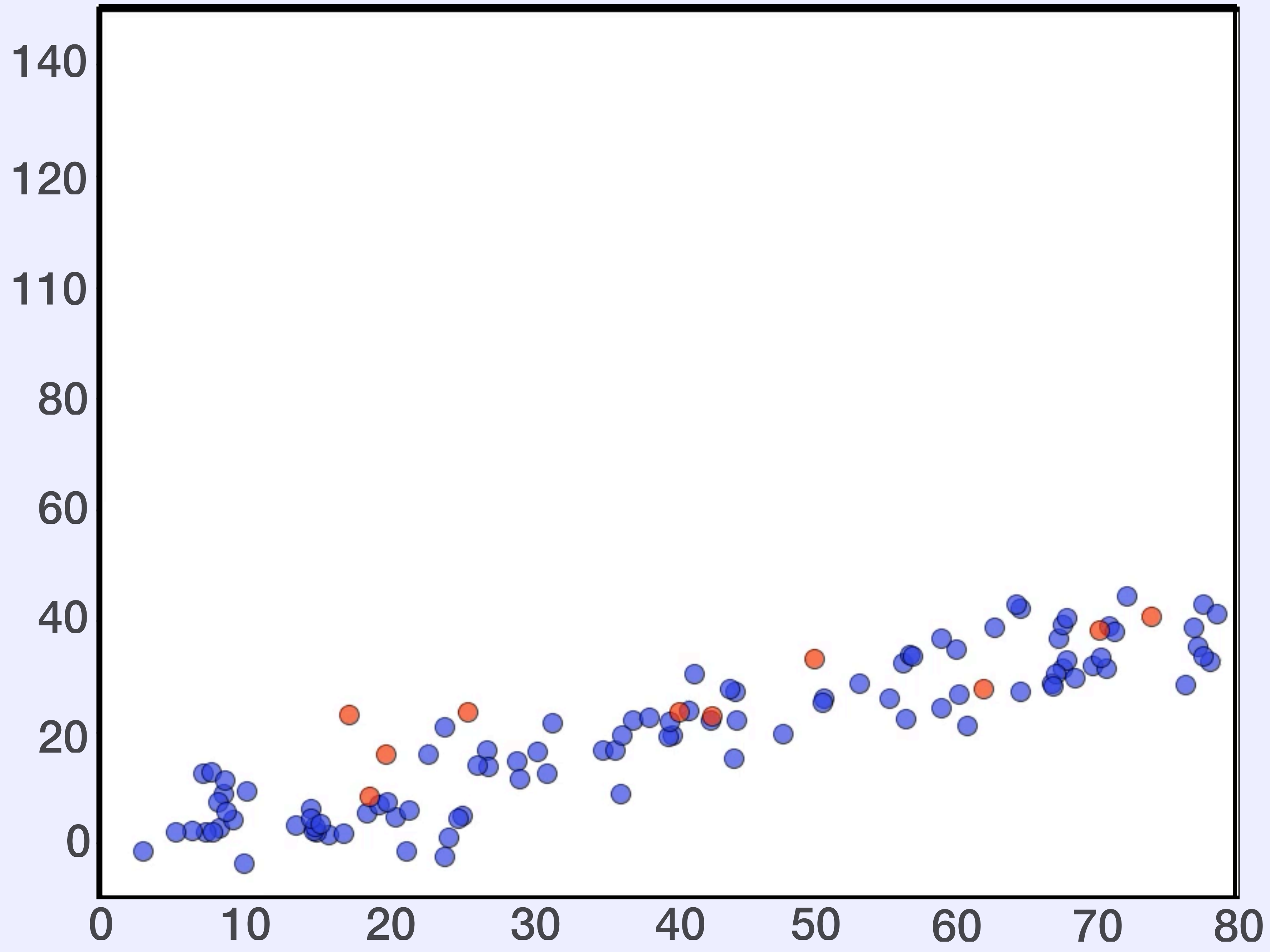


$$\min_{w \in \mathbb{R}^d} \mathbb{E}[\|Y - w^\top X\|^2] \qquad \min_{w \in \mathbb{R}^d} S_\theta[\|Y - w^\top X\|^2]$$

Toy Problem 1

- A centralized problem: least squares regression $\min_{w \in \mathbb{R}^d} \left\| Y - w^\top X \right\|^2$

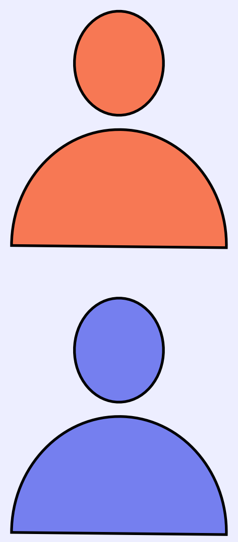


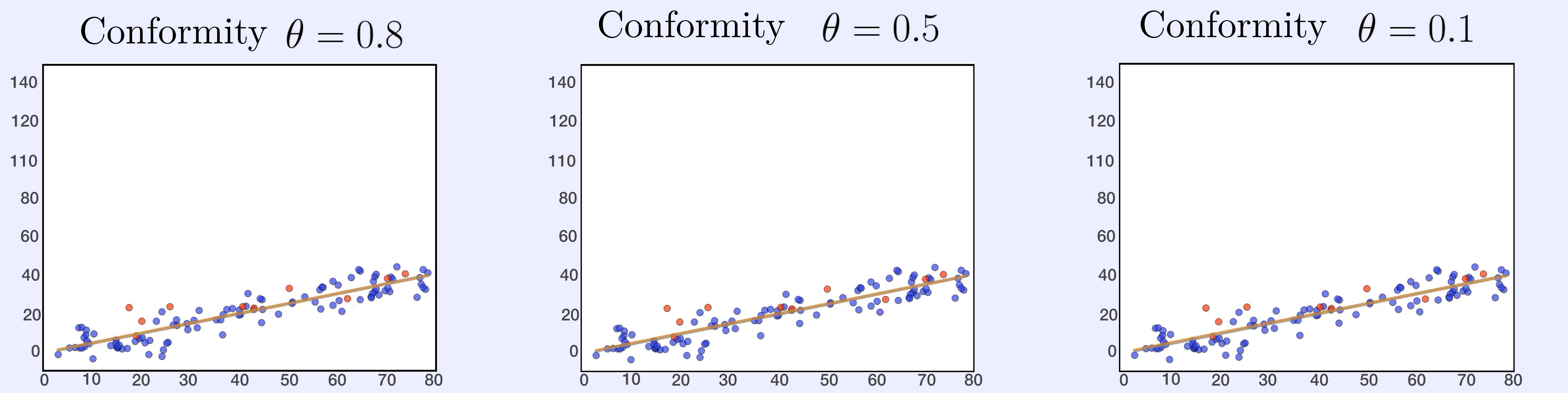


$$\min_{w \in \mathbb{R}^d} \mathbb{E}[\|Y - w^\top X\|^2] \qquad \min_{w \in \mathbb{R}^d} S_\theta[\|Y - w^\top X\|^2]$$

Toy Problem 1

- A centralized problem: least squares regression $\min_{w \in \mathbb{R}^d} \left\| Y - w^\top X \right\|^2$

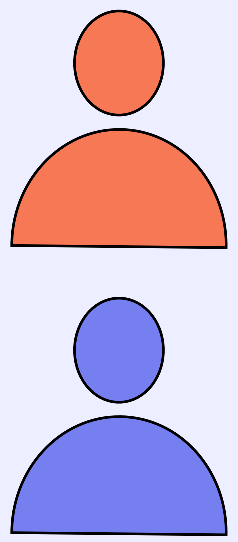




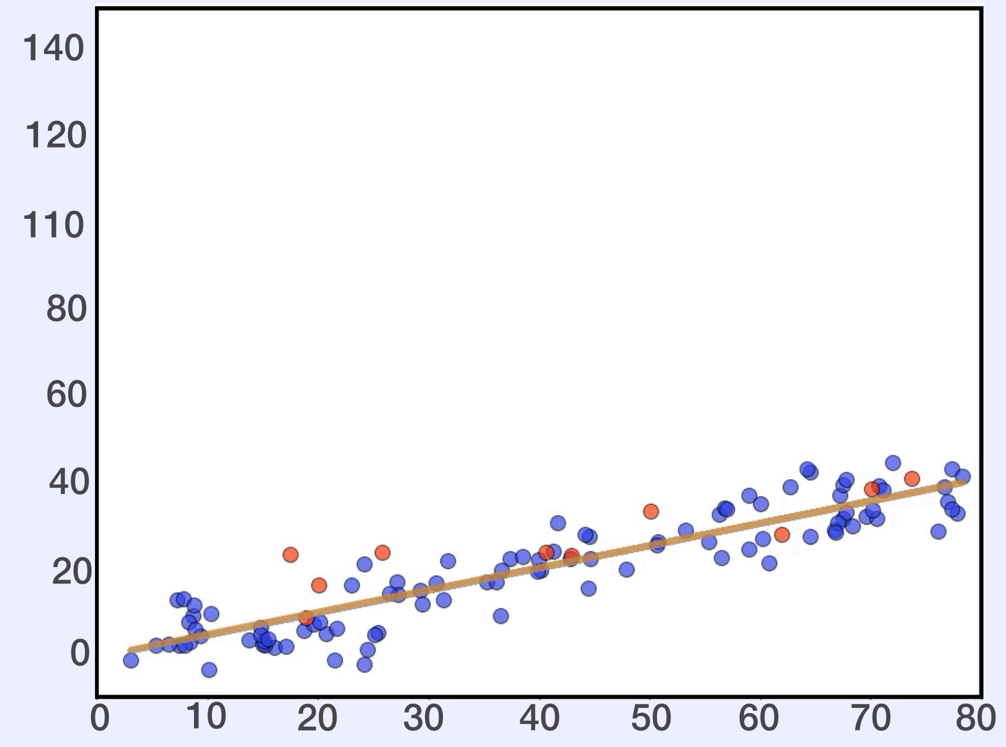
$$\min_{w \in \mathbb{R}^d} \mathbb{E}[\|Y - w^\top X\|^2] \qquad \min_{w \in \mathbb{R}^d} S_\theta[\|Y - w^\top X\|^2]$$

Toy Problem 1

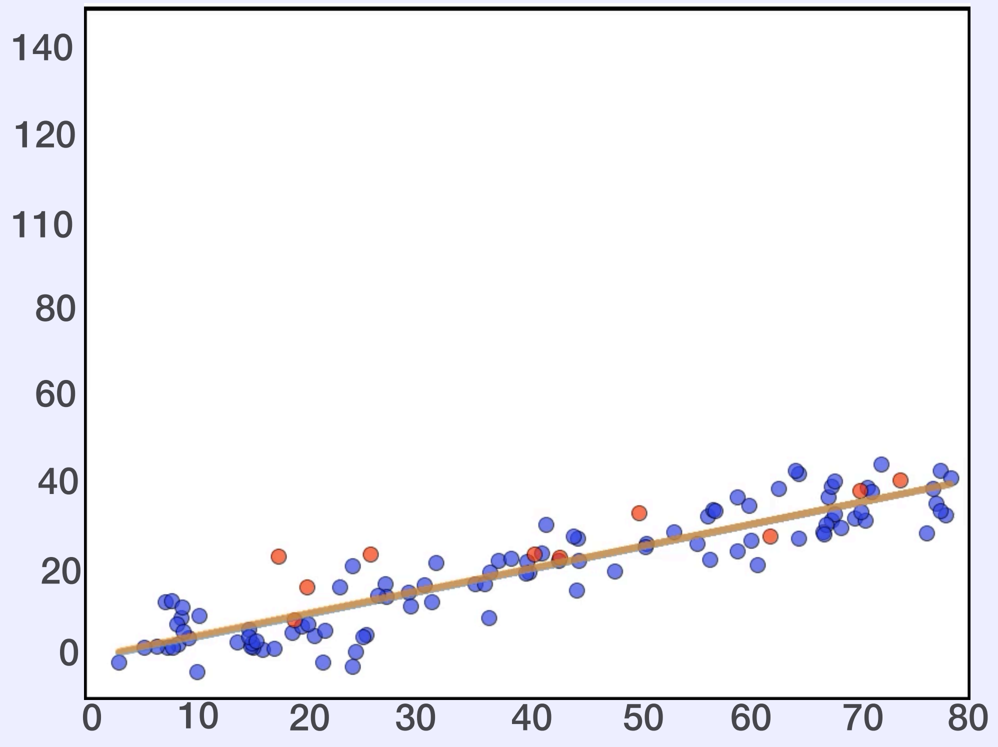
- A centralized problem: least squares regression $\min_{w \in \mathbb{R}^d} \|Y - w^\top X\|^2$



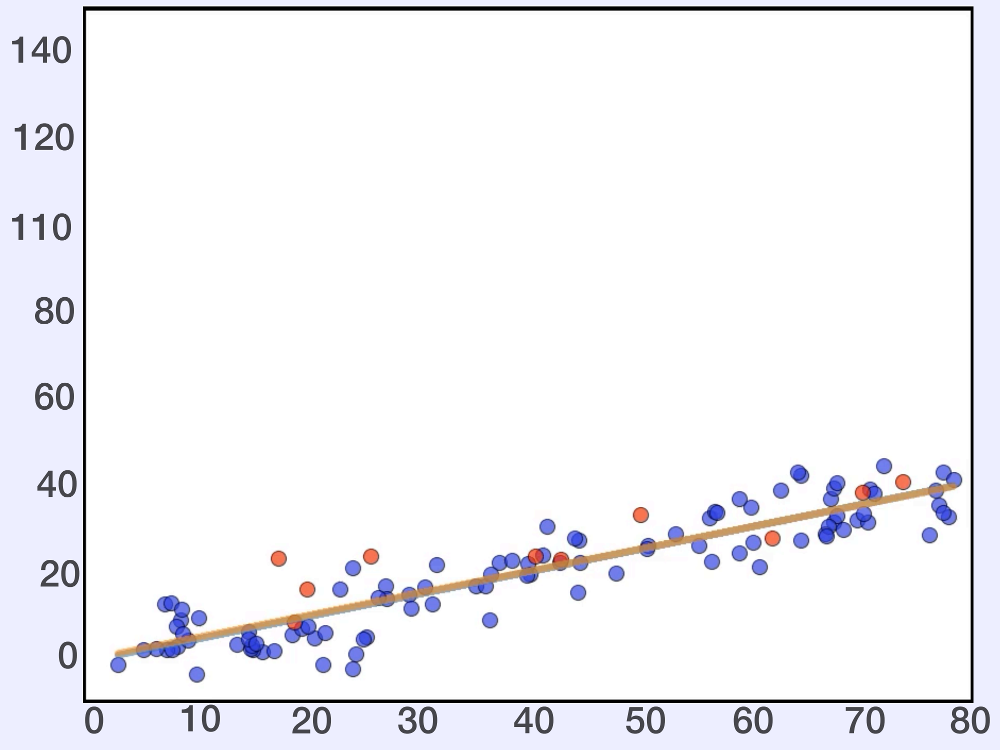
Conformity $\theta = 0.8$



Conformity $\theta = 0.5$



Conformity $\theta = 0.1$



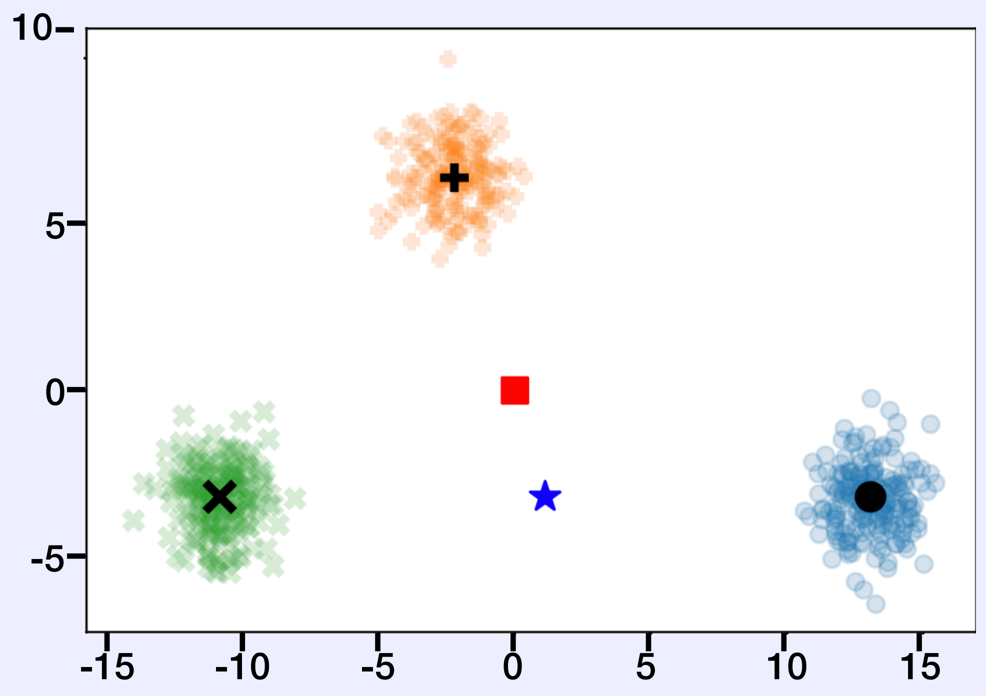
$$\min_{w \in \mathbb{R}^d} \mathbb{E}[\|Y - w^\top X\|^2] \qquad \min_{w \in \mathbb{R}^d} S_\theta[\|Y - w^\top X\|^2]$$

Toy Problem 2

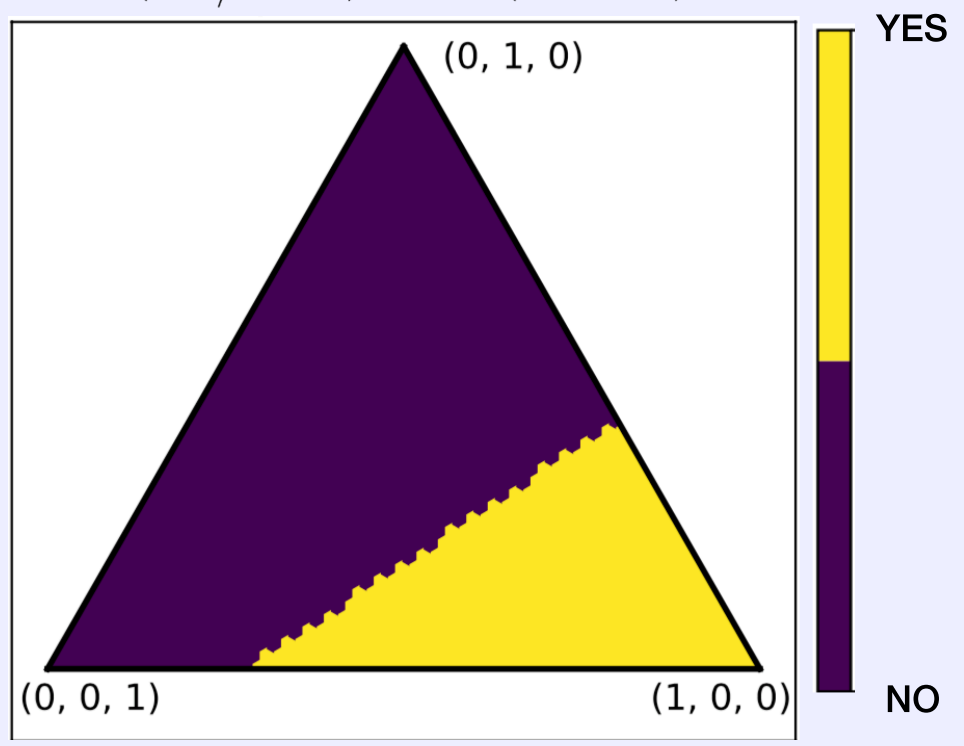
- A distributed problem: mean estimation

$$\min_{w \in \mathbb{R}^2} \mathbb{E}[\|w - \xi\|^2]$$

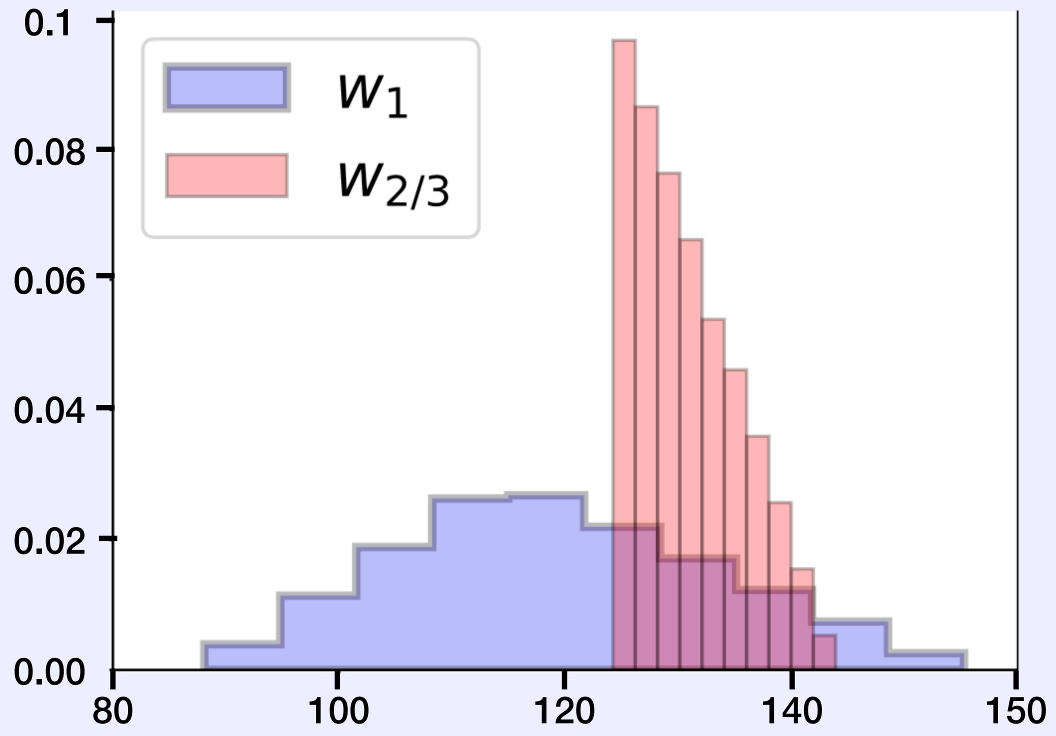
3 gaussian distributions



Is $F(w_{2/3}, p_\pi) < F(w_1, p_\pi)$?



Distributions of losses for mixtures of the three gaussian



$$\min_{w \in \mathbb{R}^2} \frac{1}{3} \sum_{i=1}^{3} \mathbb{E}_{\xi \sim q_i}[\|w - \xi\|^2]$$

$$\min_{w \in \mathbb{R}^2} S_{2/3}(\mathbb{E}_{\xi \sim q_{\mathbf{k}}}[\|w - \xi\|^2]) \qquad \mathbb{P}[\mathbf{k} = i] = \frac{1}{3}$$

2 Δ-FL in Practice

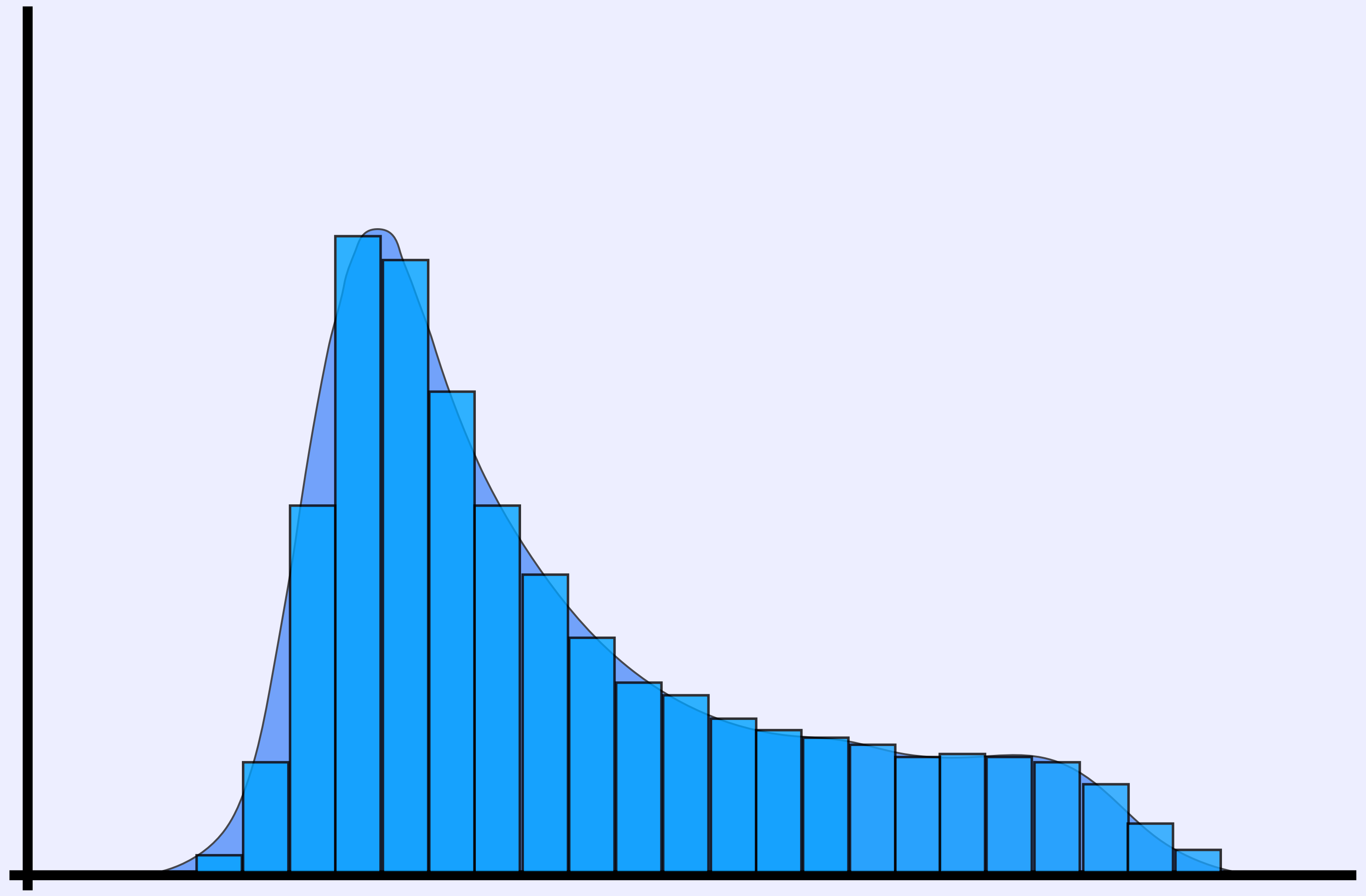
1 The Δ-FL Framework

2 Δ-FL in Practice

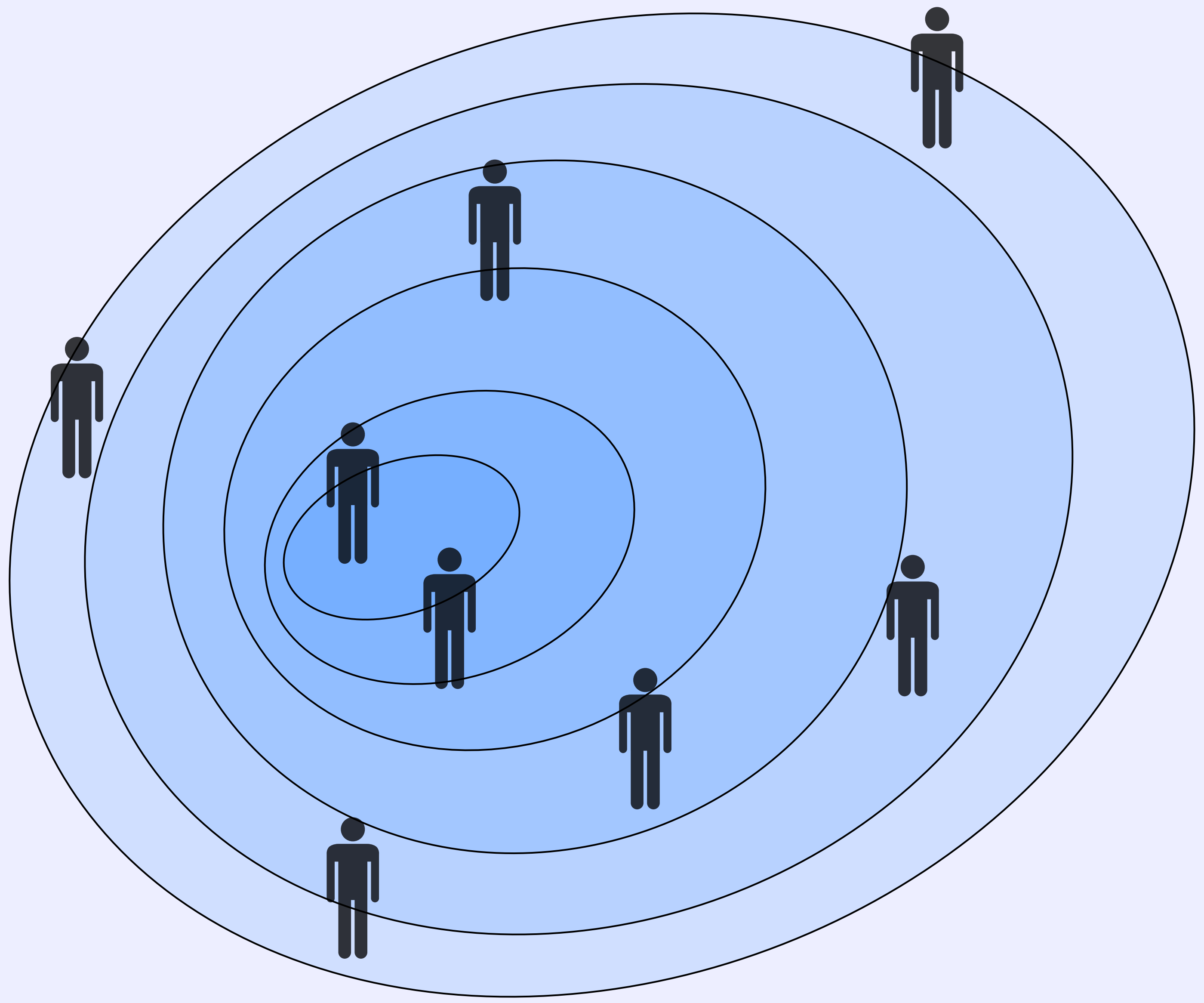
3 Numerical Experiments and Comparisons

Minimizing the worst-case losses

- Our framework focuses on the worst-cases losses

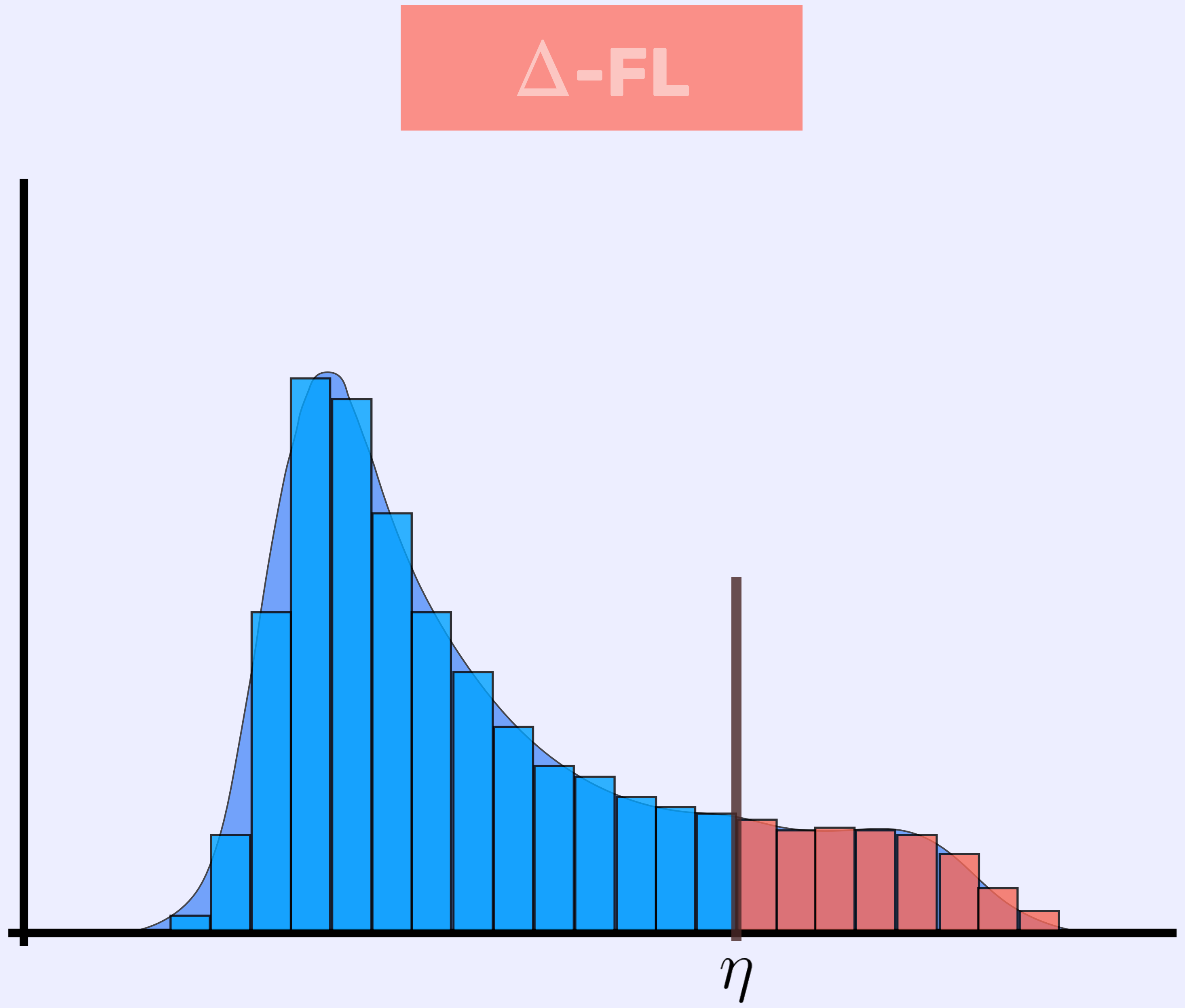


Histogram of losses over training users

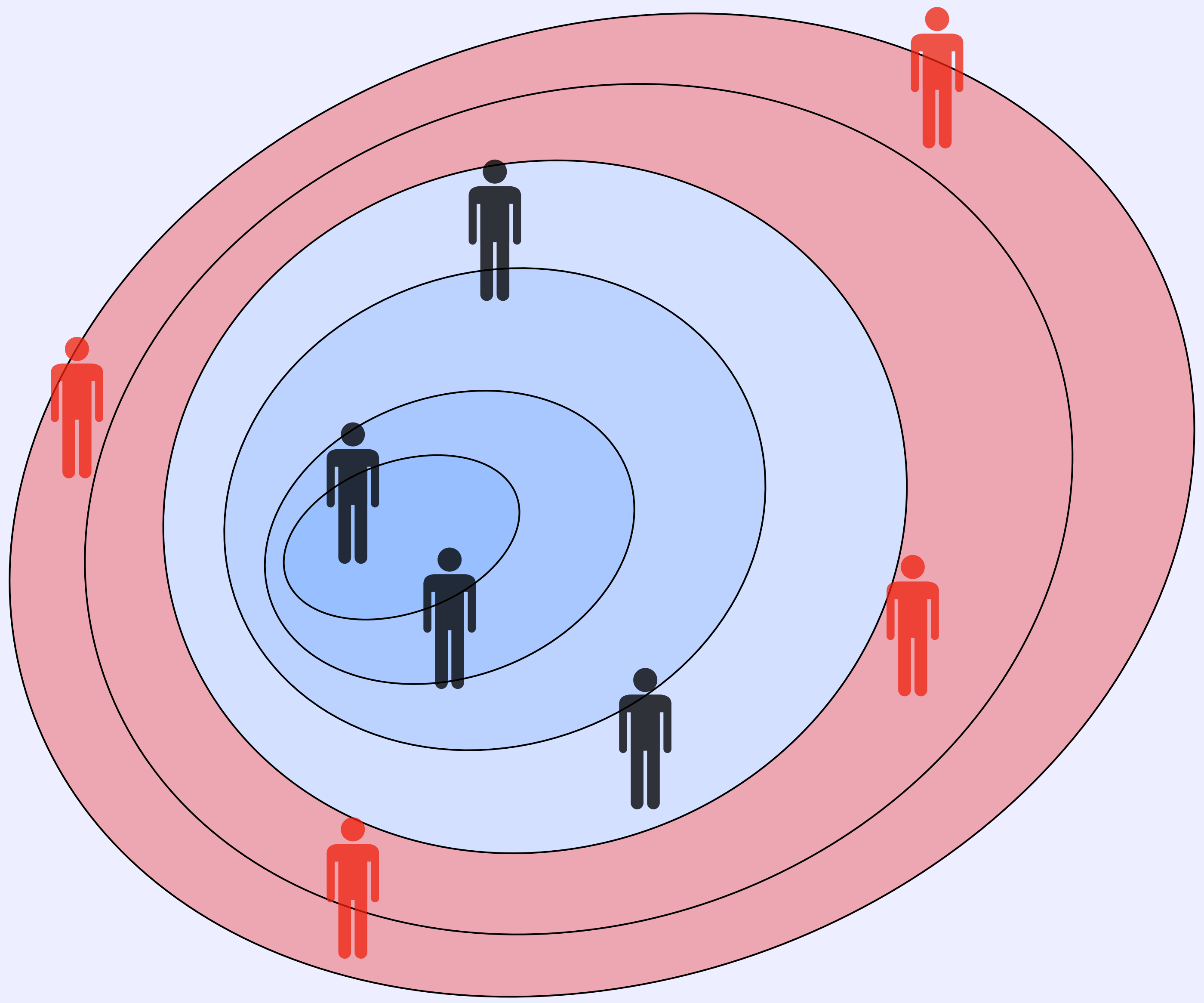


Minimizing the worst-case losses

- Our framework focuses on the worst-cases losses



Histogram of losses over training users



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**

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

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

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**

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

$w_t \simeq \underset{w \in \mathbb{R}^d}{\operatorname{argmin}}\, G(w, \eta_t)$ 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) (Local SGD)

return $w_{t^\star}$

Tackling Non-smoothness

- Smoothing the max term.
  - A non-smooth optimization problem

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

Tackling Non-smoothness

- Smoothing the max term.
  - A non-smooth optimization problem

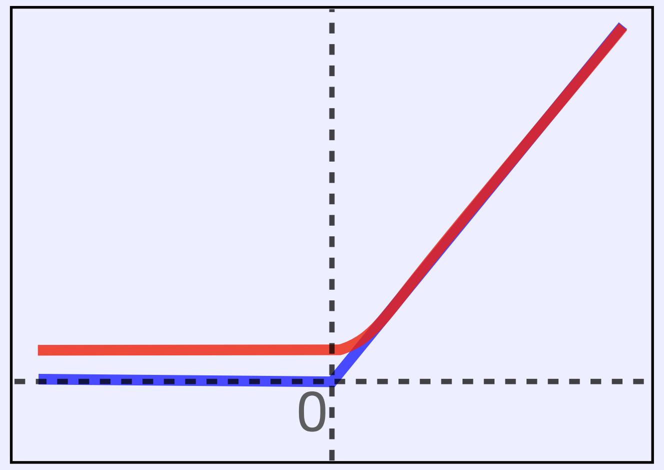
$$\min_{w \in \mathbb{R}^d} F_\theta(w) = \min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}} \eta + \frac{1}{\theta} \sum_{i=1}^{N} \alpha_i \underbrace{\max(F_i(w) - \eta, 0)}_{\text{Non-smooth term}}$$

Tackling Non-smoothness

- Smoothing the max term.
    - A non-smooth optimization problem

$$\min_{w \in \mathbb{R}^d} F_\theta(w) = \min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}} \eta + \frac{1}{\theta} \sum_{i=1}^{N} \alpha_i \underbrace{\max(F_i(w) - \eta, 0)}_{\text{Non-smooth term}}$$

$$\max(t, 0) \simeq h_\nu(t) = \min_{s \in \mathbb{R}} \left\{ \max(s, 0) + \frac{(s-t)^2}{2\nu} \right\}$$



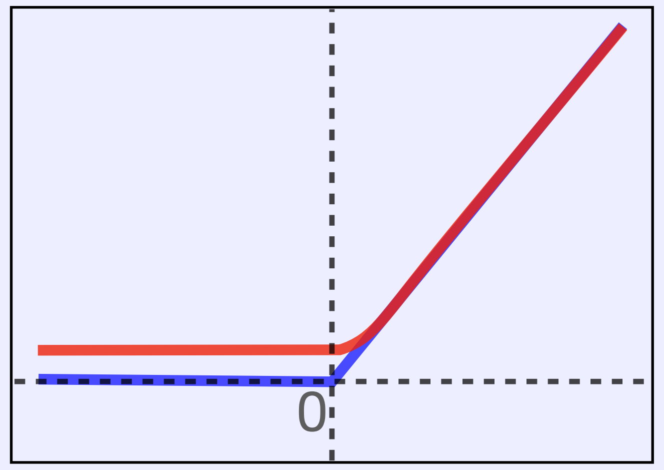
Tackling Non-smoothness

- Smoothing the max term.

  - A non-smooth optimization problem

$$\min_{w\in\mathbb{R}^d} F_\theta(w) = \min_{w\in\mathbb{R}^d,\eta\in\mathbb{R}} \eta + \frac{1}{\theta}\sum_{i=1}^{N}\alpha_i \underbrace{\max(F_i(w)-\eta,0)}_{\text{Non-smooth term}}$$

$$\max(t,0) \simeq h_\nu(t) = \min_{s\in\mathbb{R}}\left\{\max(s,0) + \frac{(s-t)^2}{2\nu}\right\}$$



  - Assuming the F_i to be smooth, we consider the following smoothed regularised problem

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

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

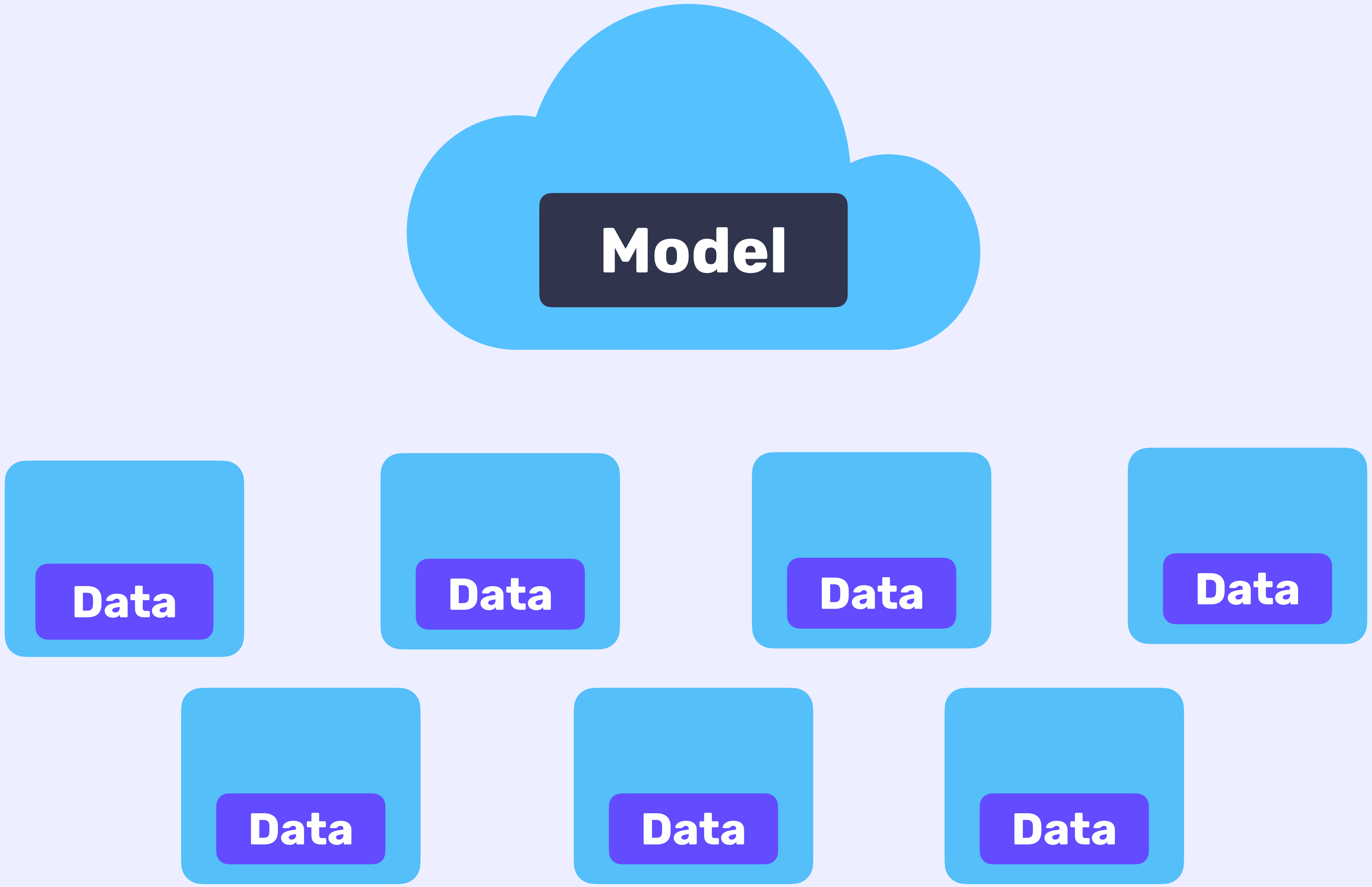
Theorem

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

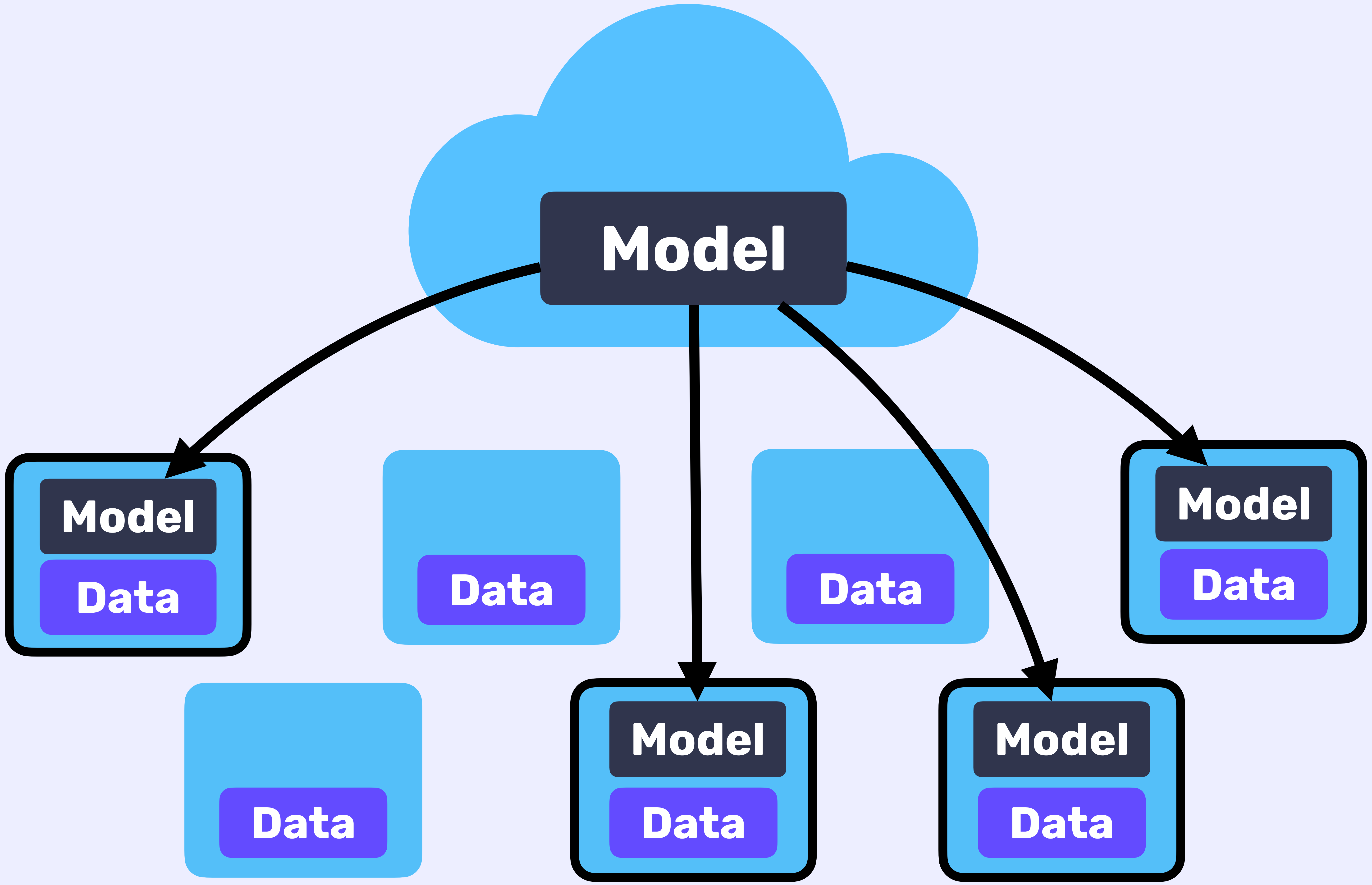
- The practical algorithm on a picture



Practical Implementation

- The practical algorithm on a picture

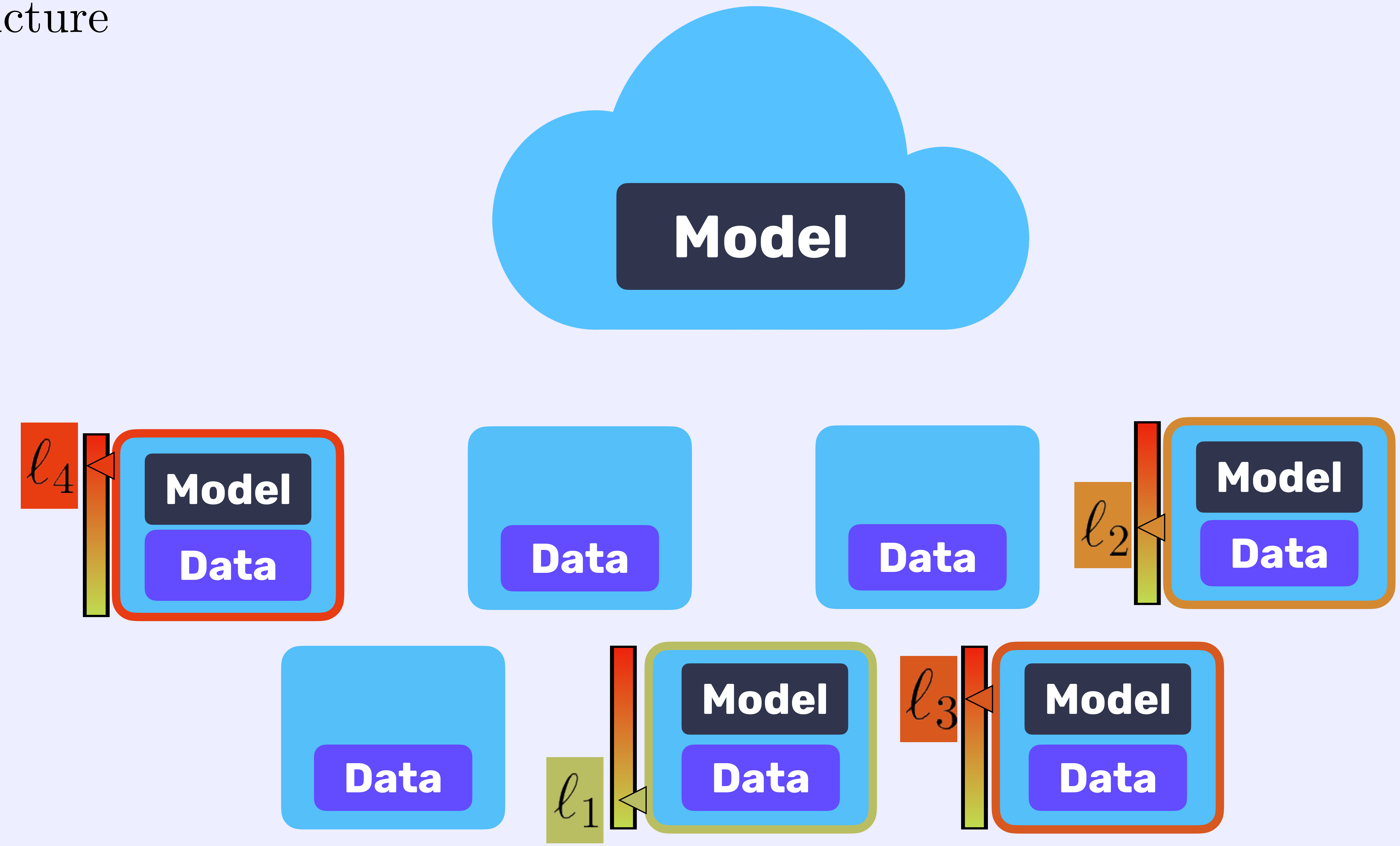
1 The server broadcasts the model to a fleet of selected devices



Practical Implementation

- The practical algorithm on a picture

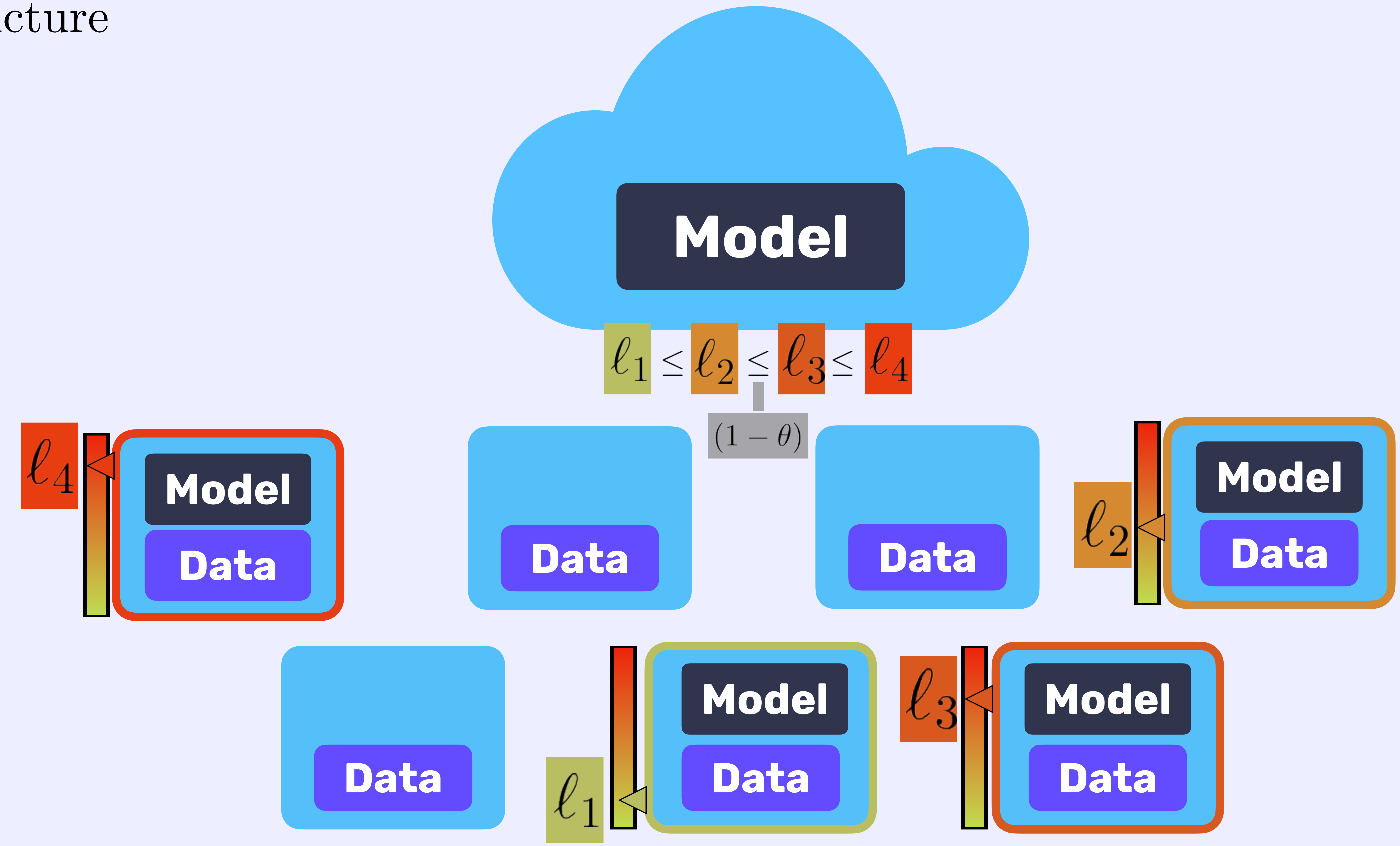
1. The server broadcasts the model to a fleet of selected devices
2. Each device compute a local loss with respect to its own data



Practical Implementation

- The practical algorithm on a picture

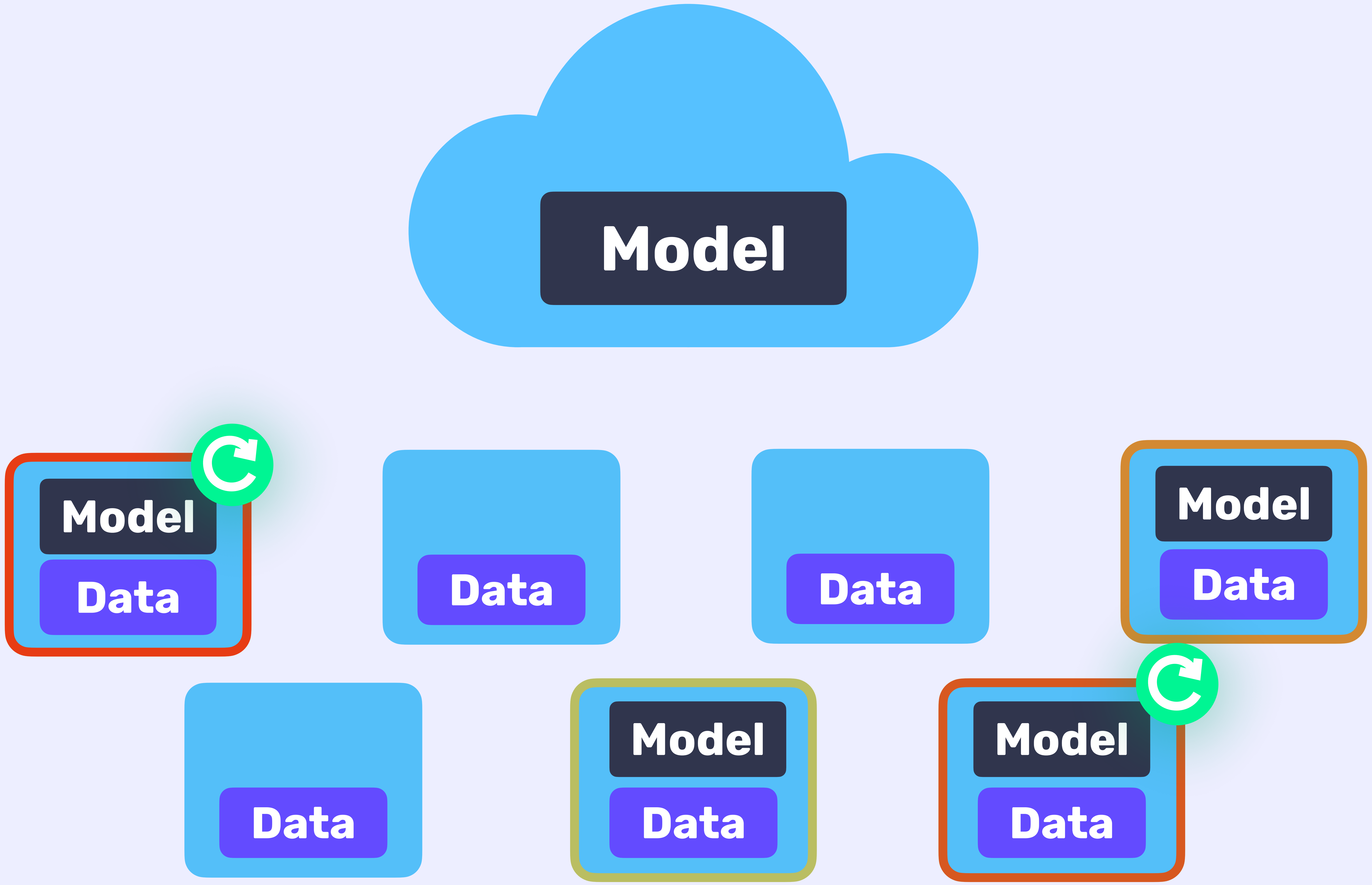
1. The server broadcasts the model to a fleet of selected devices
2. Each device compute a local loss with respect to its own data



Practical Implementation

- The practical algorithm on a picture

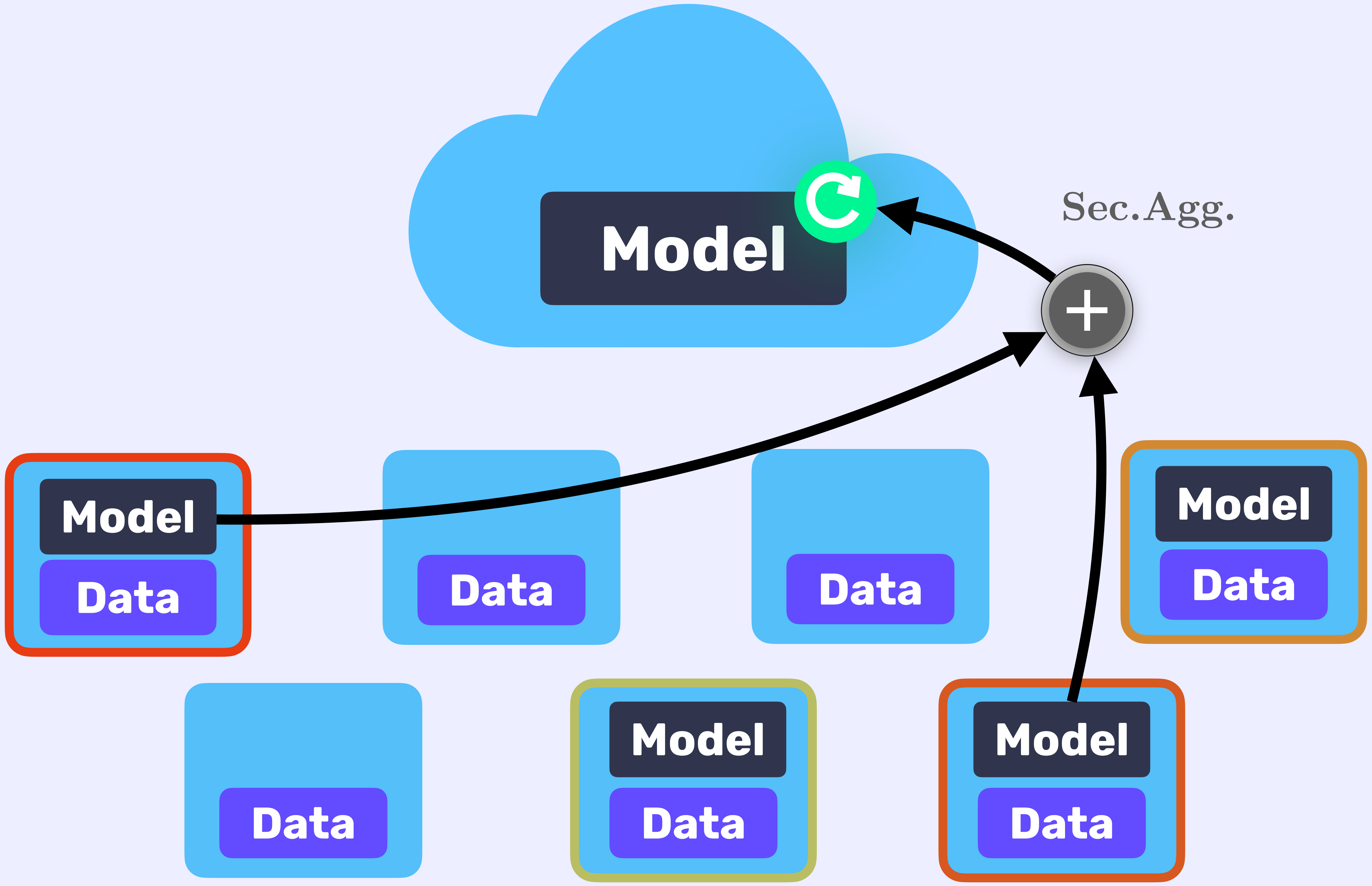
1. The server broadcasts the model to a fleet of selected devices
2. Each device compute a local loss with respect to its own data
3. Only devices with a high enough loss run local SGD for a fixed number of steps.



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



What conformity level should we use ?

- In theory, fixing θ is not easy.
  - Deciding what level of conformity to apply is a question of policy.

What conformity level should we use ?

- In theory, fixing θ is not easy.
  - Deciding what level of conformity to apply is a question of policy.
  - However, for conforming users, keeping the risk-averse model might harm a lot there local loss

What conformity level should we use ?

- In theory, fixing θ is not easy.
  - Deciding what level of conformity to apply is a question of policy.
  - However, for conforming users, keeping the risk-averse model might harm a lot their local loss
  - A possible fix by *mean-CVaR* optimization

$$\min_{w \in \mathbb{R}^d} \lambda \left(\sum_{i=1}^{N} \alpha_i \, F_i(w) \right) + (1 - \lambda) \max_{\pi \in \mathcal{P}_\theta} \sum_{i=1}^{N} \pi_i \, F_i(w)$$

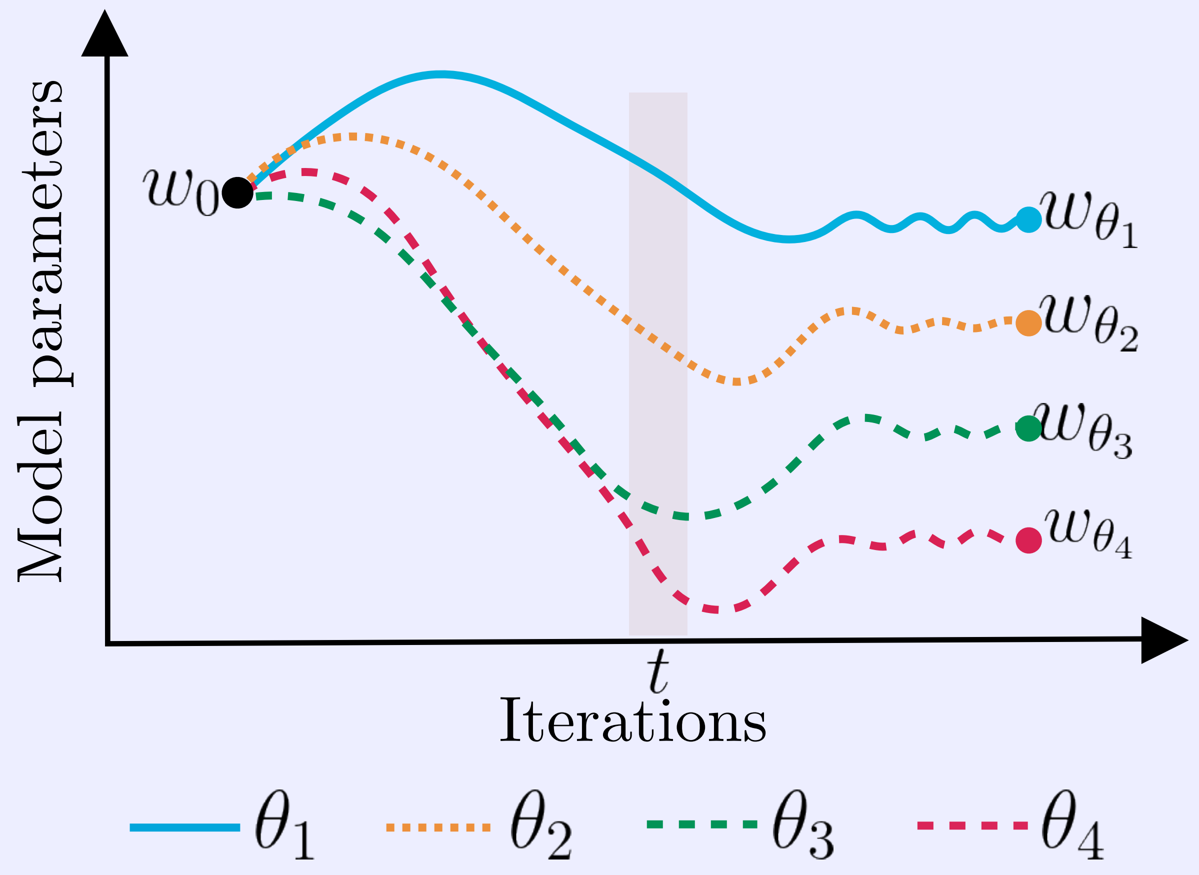
What conformity level should we use ?

- In theory, fixing θ is not easy.
  - Deciding what level of conformity to apply is a question of policy.
  - However, for conforming users, keeping the risk-averse model might harm a lot there local loss
  - A possible fix by *mean-CVaR* optimization

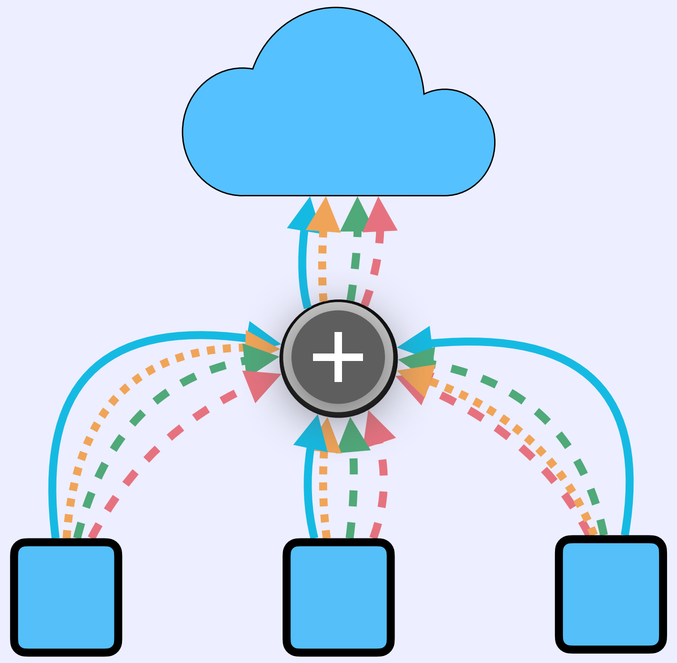
$$\min_{w \in \mathbb{R}^d} \lambda \left(\sum_{i=1}^{N} \alpha_i \; F_i(w) \right) + (1 - \lambda) \max_{\pi \in \mathcal{P}_\theta} \sum_{i=1}^{N} \pi_i \; F_i(w)$$

- In practice, we propose to keep track of different levels of conformity within each device.

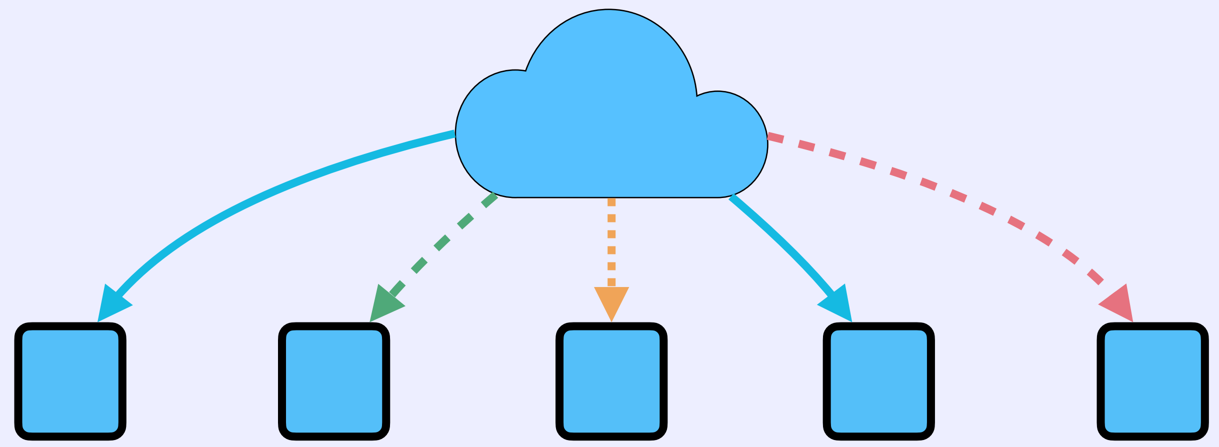
Trajectories of model parameters over time



In iteration t of training

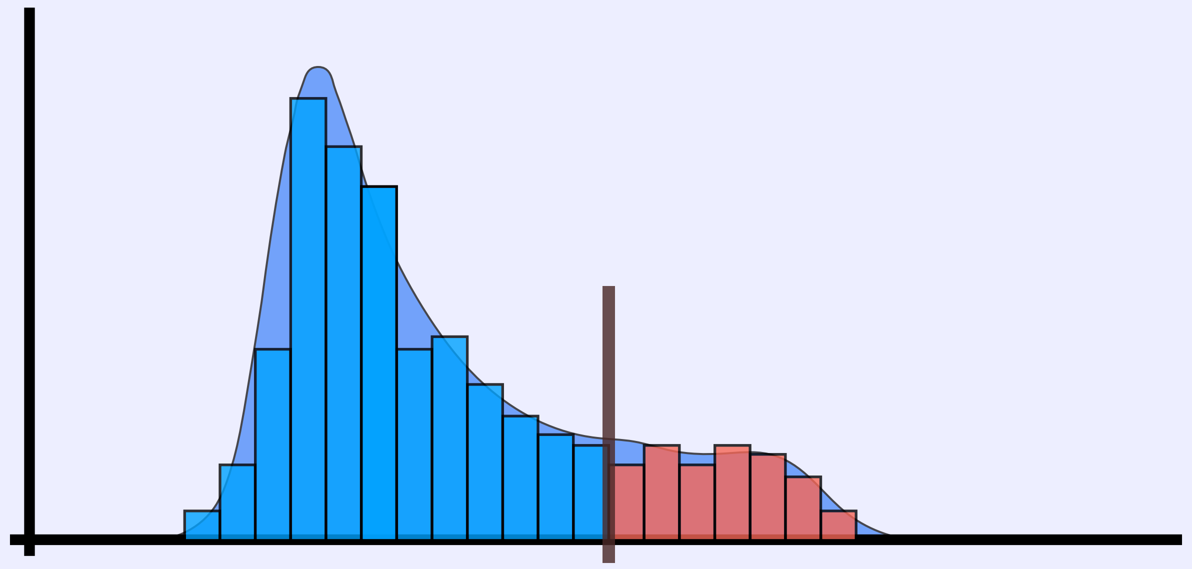


Test devices select their level of conformity θ



Privacy Preservation for the Device Filtering Step

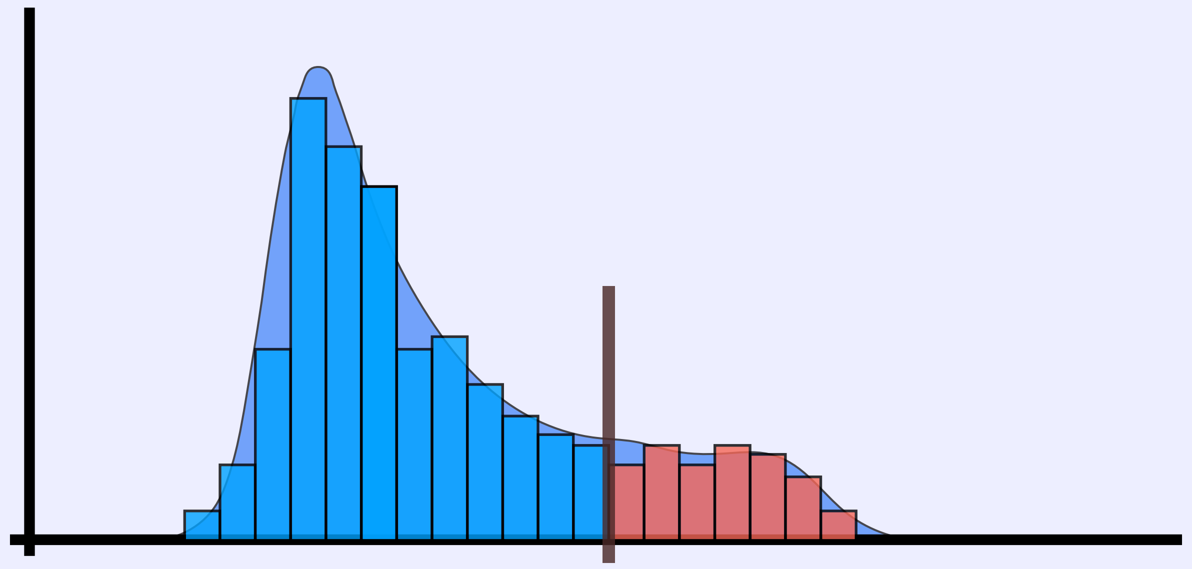
- Δ-FL acts as FedAvg with a device filtering step



Histogram of losses over users

Privacy Preservation for the Device Filtering Step

- Δ-FL acts as FedAvg with a device filtering step

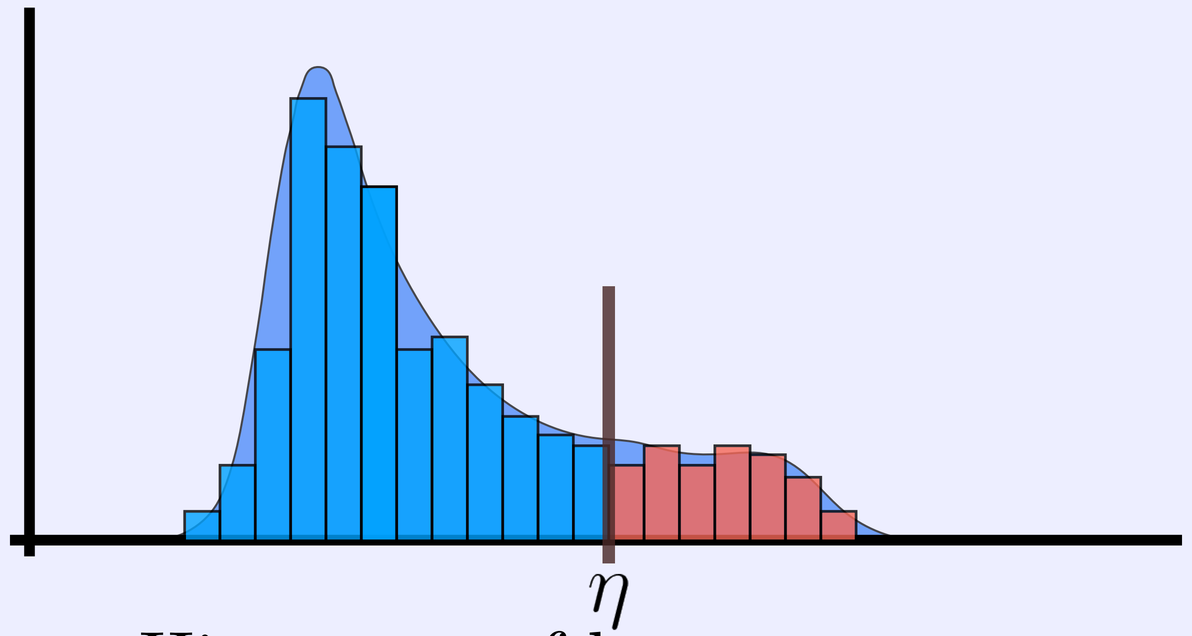


Histogram of losses over users

- We propose to use a standard majorization-minimization scheme to securely compute the quantile

Privacy Preservation for the Device Filtering Step

- Δ-FL acts as FedAvg with a device filtering step

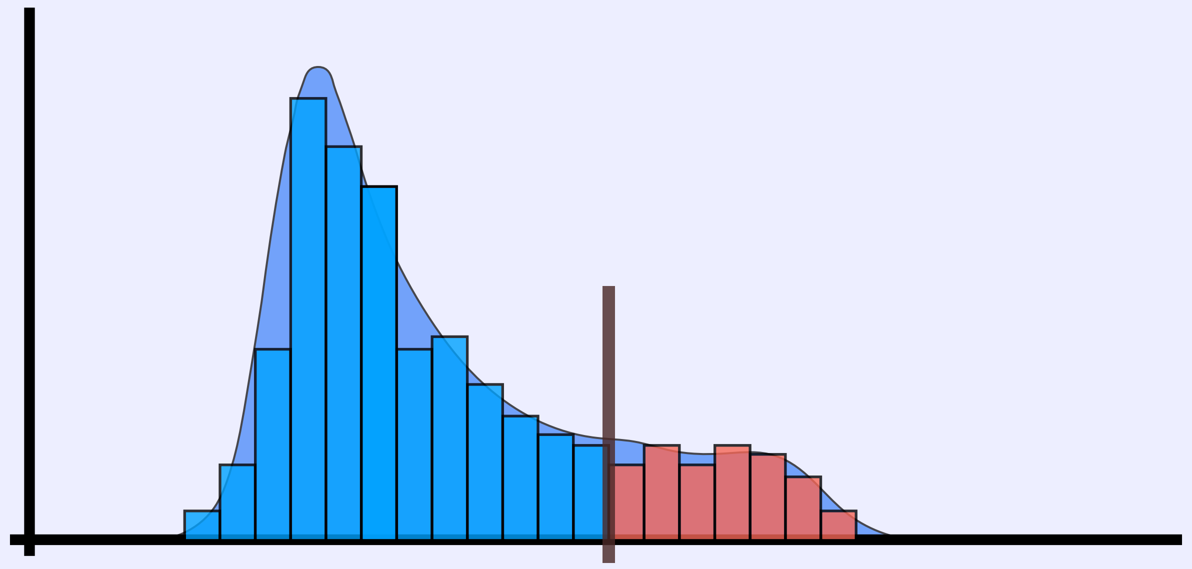


Histogram of losses over users

- We propose to use a standard majorization-minimization scheme to securely compute the quantile
  - Let us take the conformity level $\theta = 0.5$

Privacy Preservation for the Device Filtering Step

- Δ-FL acts as FedAvg with a device filtering step



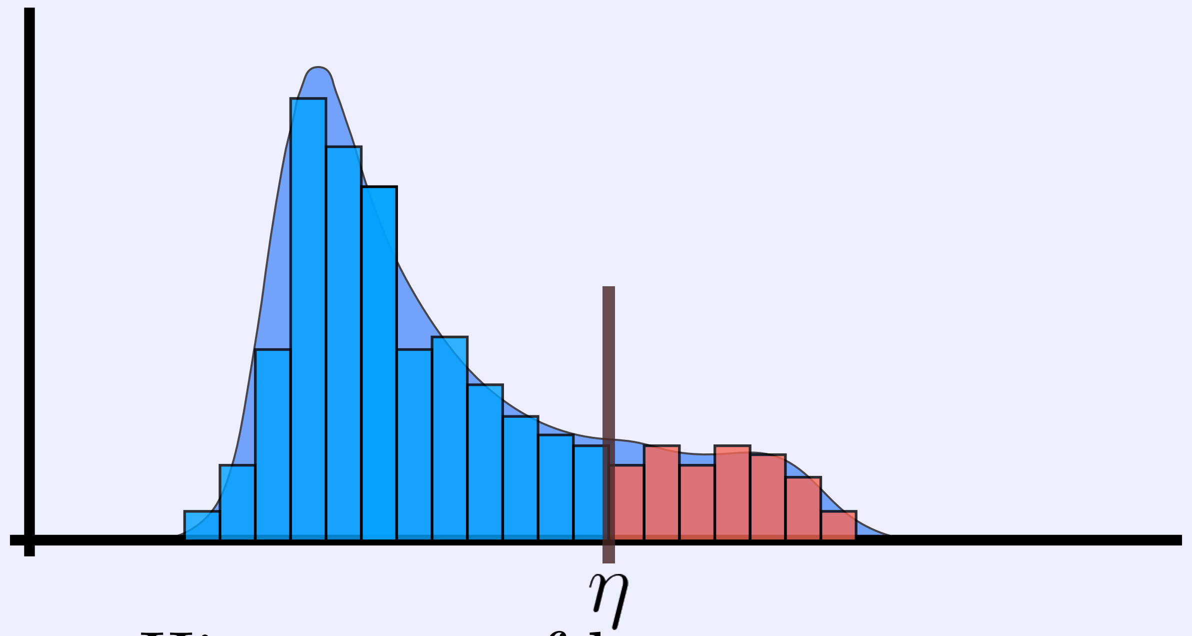
Histogram of losses over users

- We propose to use a standard majorization-minimization scheme to securely compute the quantile
  - Let us take the conformity level $\theta = 0.5$
  - Iteratively reweighed least squares procedure

$$\eta^{(t+1)} = \underset{\eta \in \mathbb{R}}{\text{argmin}} \sum_{i=1}^{N} \alpha_i \frac{(F_i(w) - \eta)^2}{|F_i(w) - \eta^{(t)}|}$$

Privacy Preservation for the Device Filtering Step

- Δ-FL acts as FedAvg with a device filtering step



Histogram of losses over users

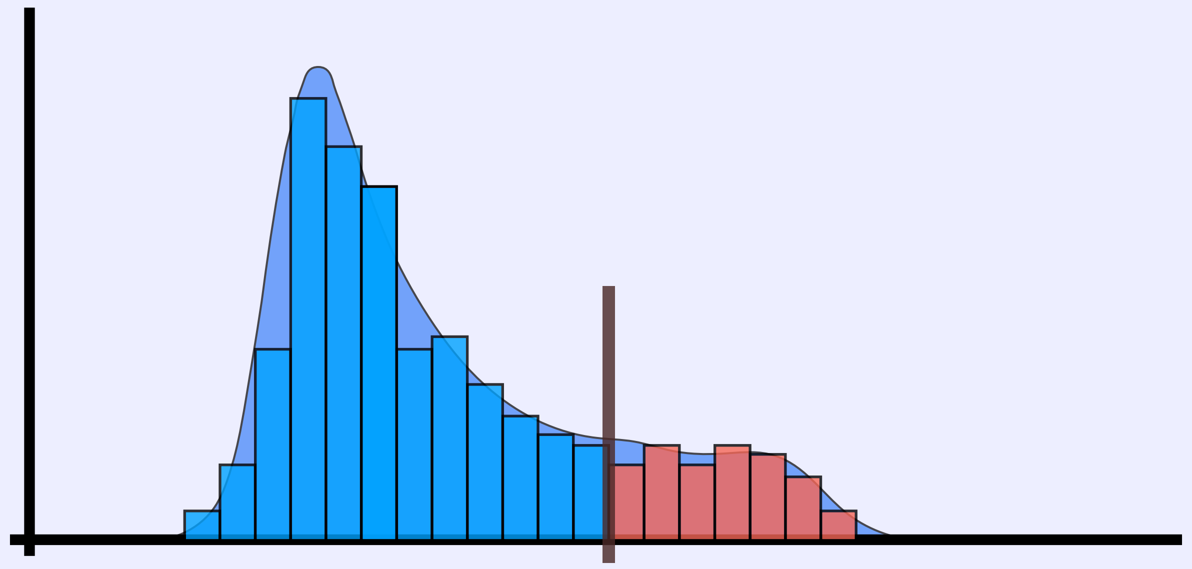
- We propose to use a standard majorization-minimization scheme to securely compute the quantile
  - Let us take the conformity level $\theta = 0.5$
  - Iteratively reweighed least squares procedure

$$\eta^{(t+1)} = \underset{\eta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{N} \alpha_i \frac{(F_i(w) - \eta)^2}{|F_i(w) - \eta^{(t)}|}$$

  - Solving each iteration boils down to the computation of a weighted averages of the local losses F_i

Privacy Preservation for the Device Filtering Step

- Δ-FL acts as FedAvg with a device filtering step



Histogram of losses over users

- We propose to use a standard majorization-minimization scheme to securely compute the quantile
  - Let us take the conformity level $\theta = 0.5$
  - Iteratively reweighed least squares procedure

$$\eta^{(t+1)} = \underset{\eta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{N} \alpha_i \frac{(F_i(w) - \eta)^2}{|F_i(w) - \eta^{(t)}|}$$

  - Solving each iteration boils down to the computation of a weighted averages of the local losses F_i
  - For any $\theta \in (0, 1]$, we can still recover the $(1 - \theta)$-quantile by minimizing iteratively a quadratic function

3 Numerical Experiments and Comparisons

1 The Δ-FL Framework

2 Δ-FL in Practice

3 Numerical Experiments and Comparisons

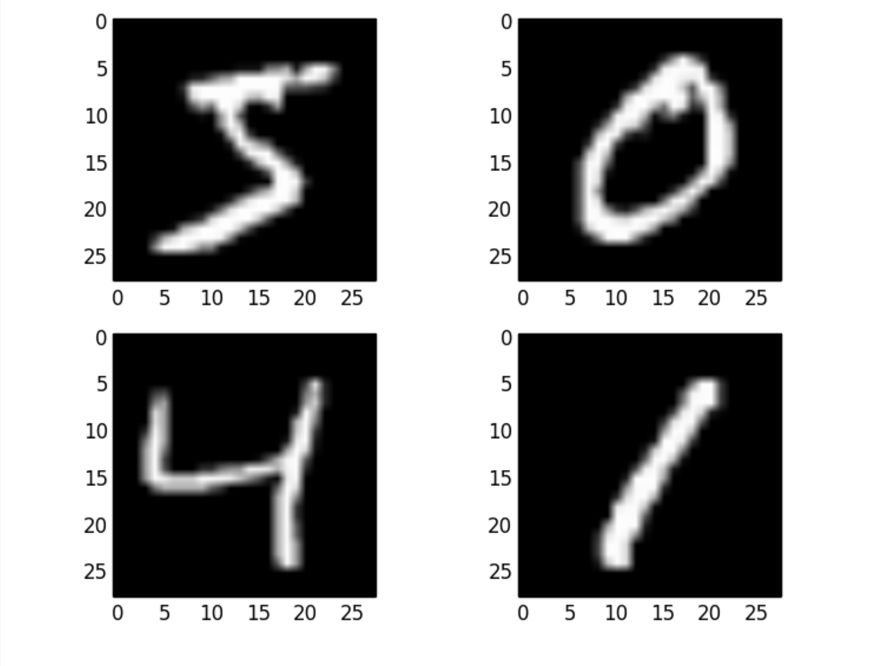
Experimental Setup

- 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

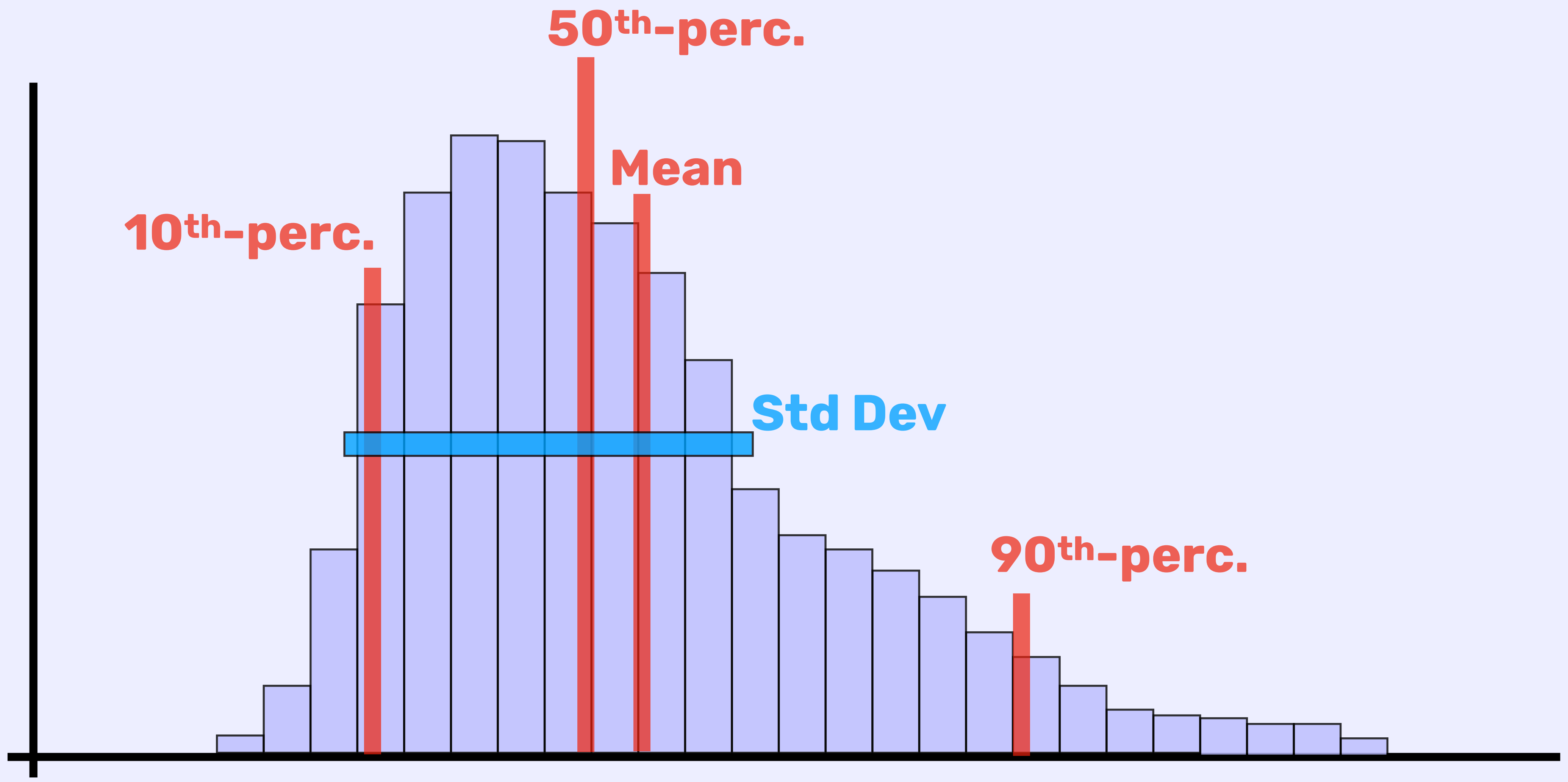
Language Modelling

SHAKESPEARE

RNN

Evaluation Metrics

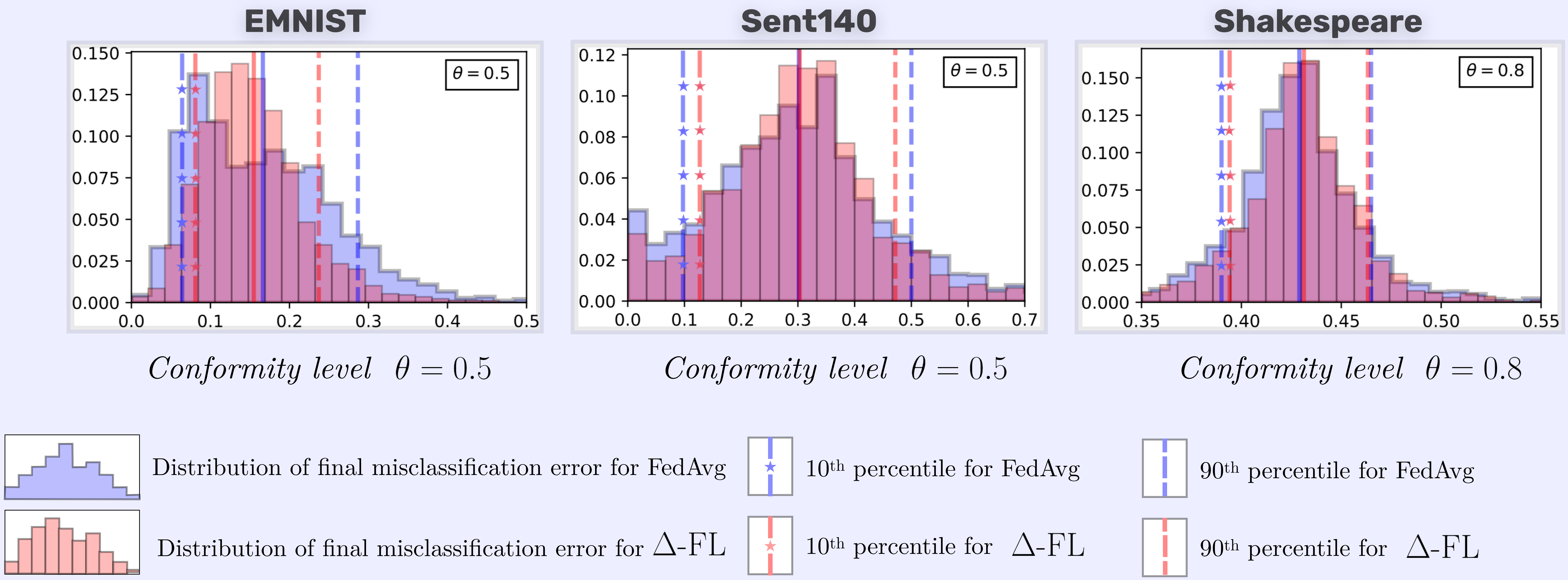
- Metrics gathered
  - We record the loss of each training device and the misclassification error of each testing device.
- Evaluation Metrics
  - Given the distribution of train losses and test misclassification errors, we evaluate several statistical summaries of theses distributions



Histogram of misclassification errors over testing users

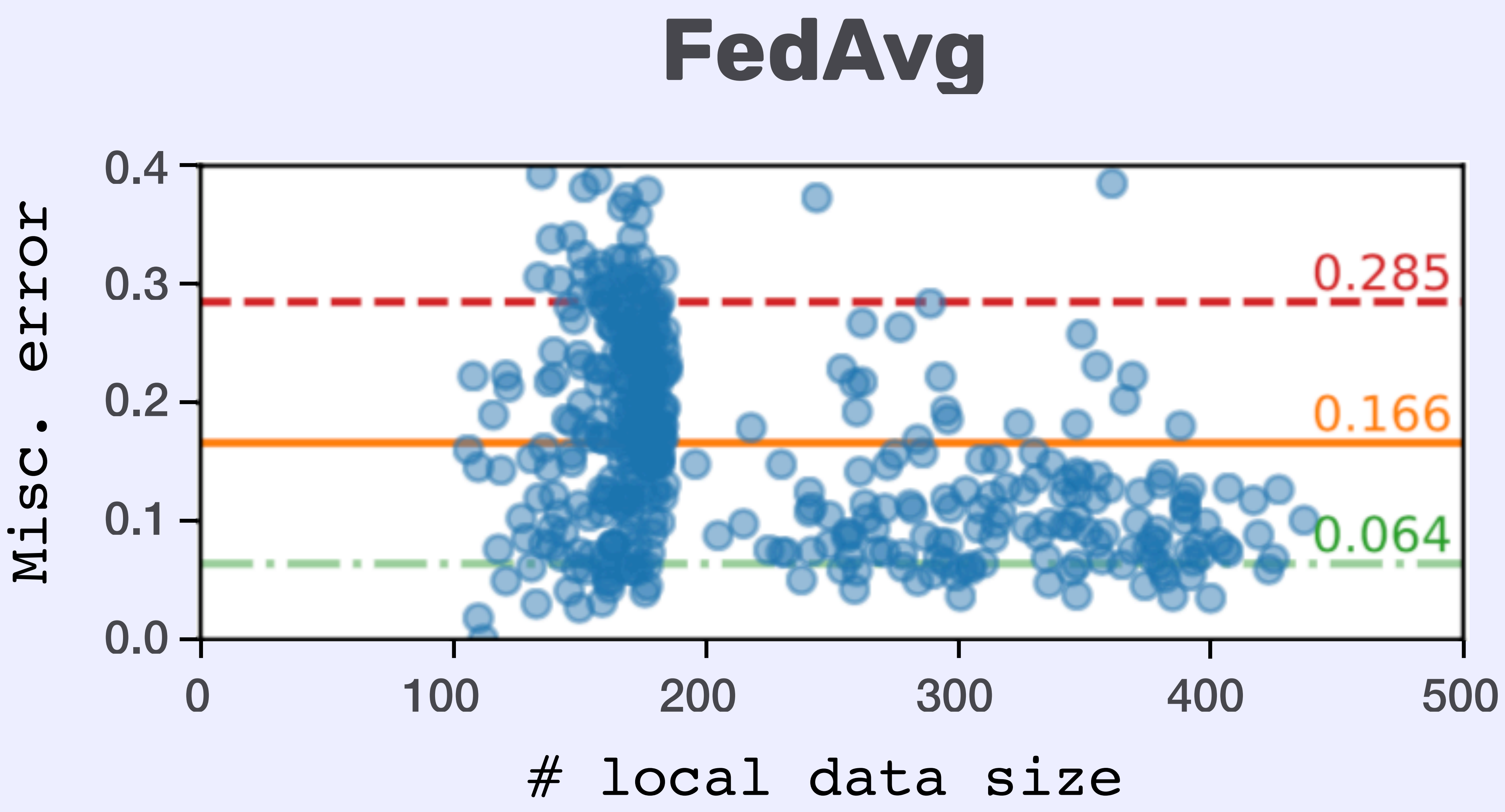
Experimental Results - Final Performances

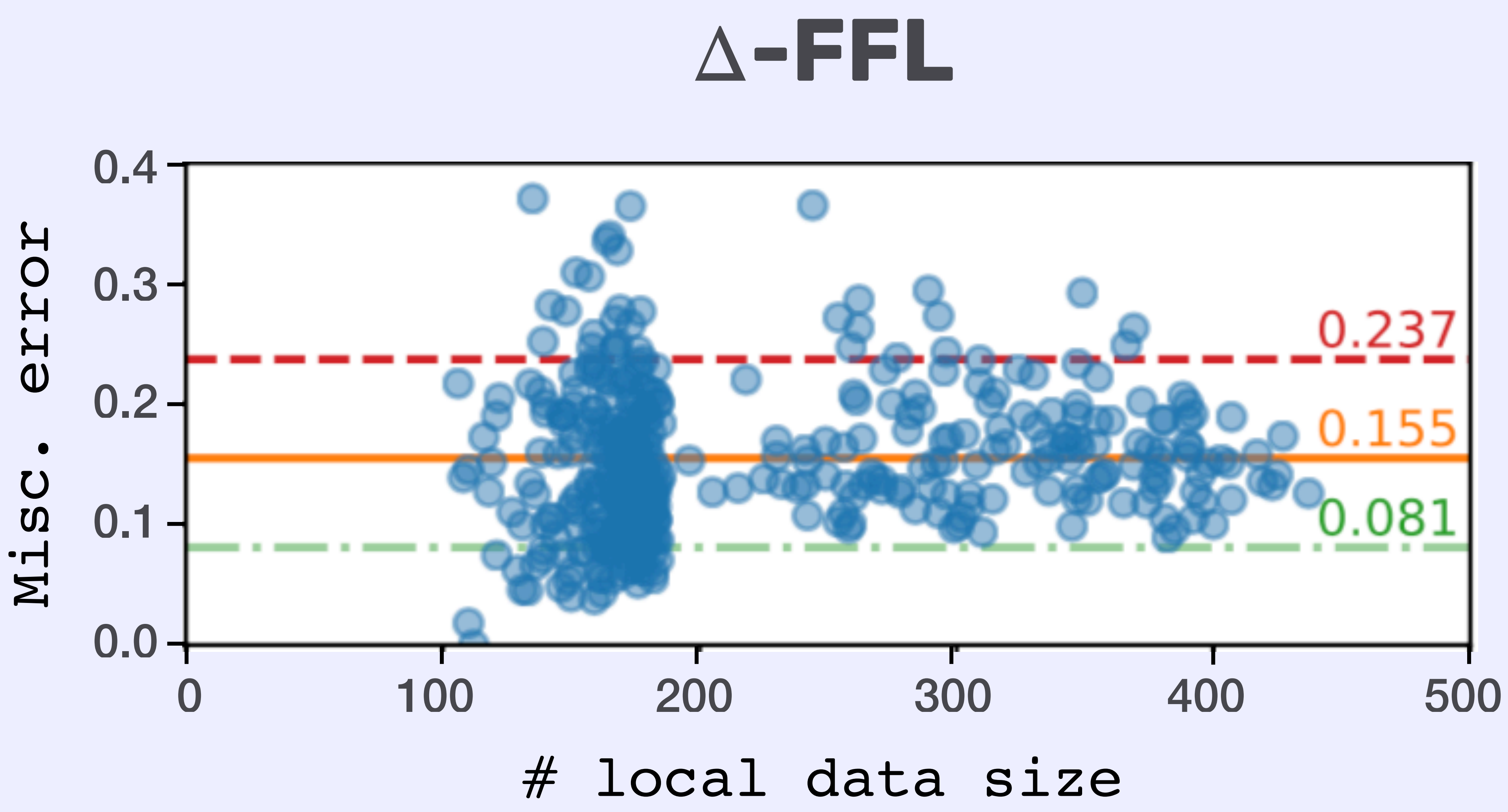
- Distribution of final misclassification error



Experimental Results - Local Performance vs Data-Size

- Scatter plot of local final performance VS local data-size

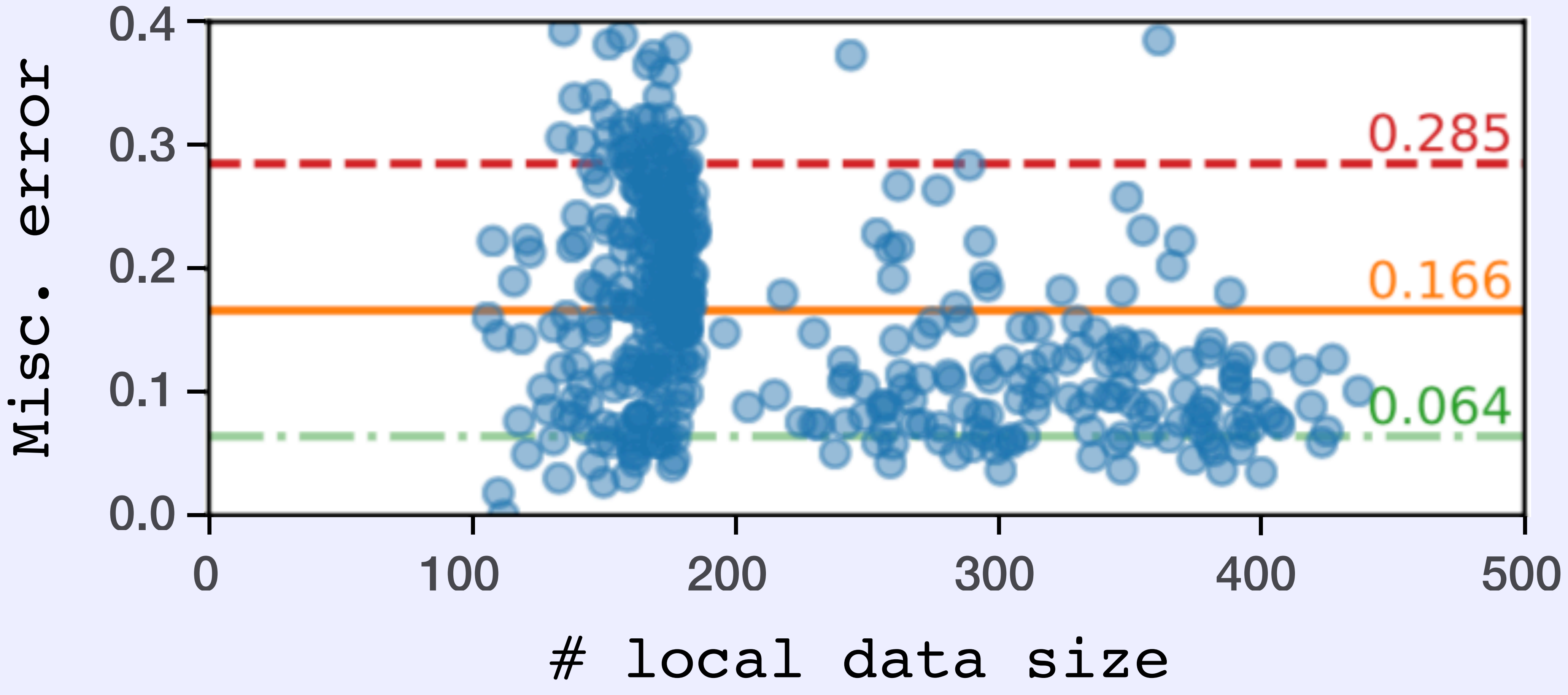




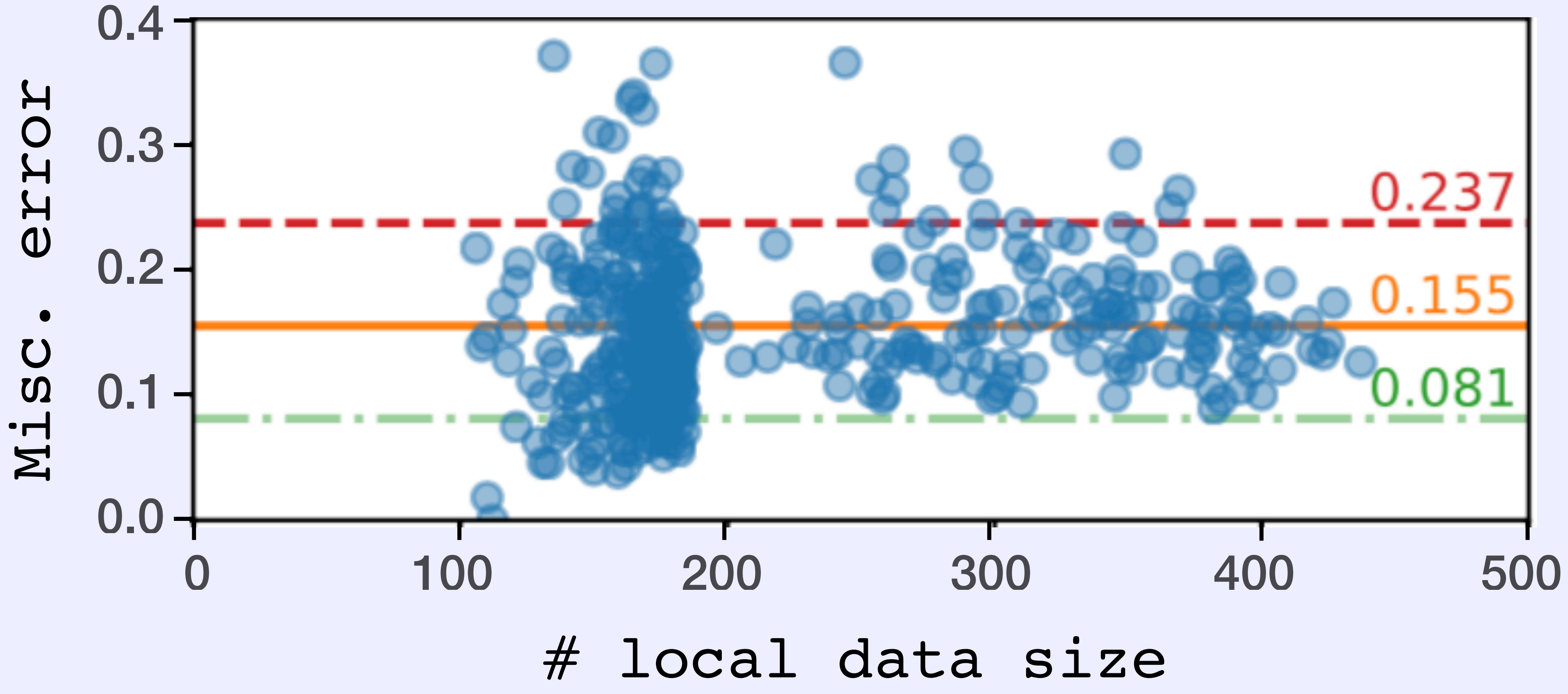
Experimental Results - Local Performance vs Data-Size

- Scatter plot of local final performance VS local data-size

FedAvg



Δ-FFL



$$\alpha_i = \frac{\text{Number of local data points}}{\text{Total Number of data points}}$$

Comparison with recent FL Methods

- We compare the performances of Δ-FL t:
  - FedAvg for different numbers of devices selected per round

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

  - FedProx with a tuned proximal parameter

$$\min_{w_i \in \mathbb{R}^d} F_i(w_i) + \frac{\mu}{2} \| w_i - w^{(t)} \|^2$$

  - q-FFL for different values of q

$$\min_{w \in \mathbb{R}^d} \frac{1}{qN} \sum_{i=1}^{N} F_i(w)^q \quad (q \geq 1)$$

  - AFL as an asymptotic version of q-FFL

$$\min_{w \in \mathbb{R}^d} \max_{1 \leq i \leq N} F_i(w)$$

Implemented as q-FFL with a large q

- We test the performances of Δ-FL for three conformity levels

Experimental Results - Final Performances

- 90^{th} percentile Misclassification Error

90^{th} percentile of misclassification error (in %) on test devices.

	EMNIST		Sent140		Shakespeare
	Linear	ConvNet	Linear	RNN	RNN
FedAvg	49.66 ± 0.67	28.46 ± 1.07	46.83 ± 0.54	49.67 ± 3.95	46.45 ± 0.11
FedProx	49.15 ± 0.74	27.01 ± 1.86	46.83 ± 0.54	49.86 ± 4.07	46.47 ± 0.24
q-FFL	49.90 ± 0.58	28.02 ± 0.80	$\mathbf{46.39} \pm 0.40$	48.66 ± 4.68	46.36 ± 0.19
AFL	51.62 ± 0.28	45.08 ± 1.00	47.52 ± 0.32	57.78 ± 1.19	75.06 ± 1.03
Δ-FL $\theta = 0.8$	49.10 ± 0.24	26.23 ± 1.15	46.44 ± 0.38	$\mathbf{46.46} \pm 4.39$	46.33 ± 0.10
Δ-FL $\theta = 0.5$	$\mathbf{46.48} \pm 0.38$	$\mathbf{23.69} \pm 0.94$	46.64 ± 0.41	50.48 ± 8.24	$\mathbf{46.32} \pm 0.13$
Δ-FL $\theta = 0.1$	50.34 ± 0.95	25.46 ± 2.77	51.39 ± 1.07	86.45 ± 10.95	47.17 ± 0.14

Experimental Results - Final Performances

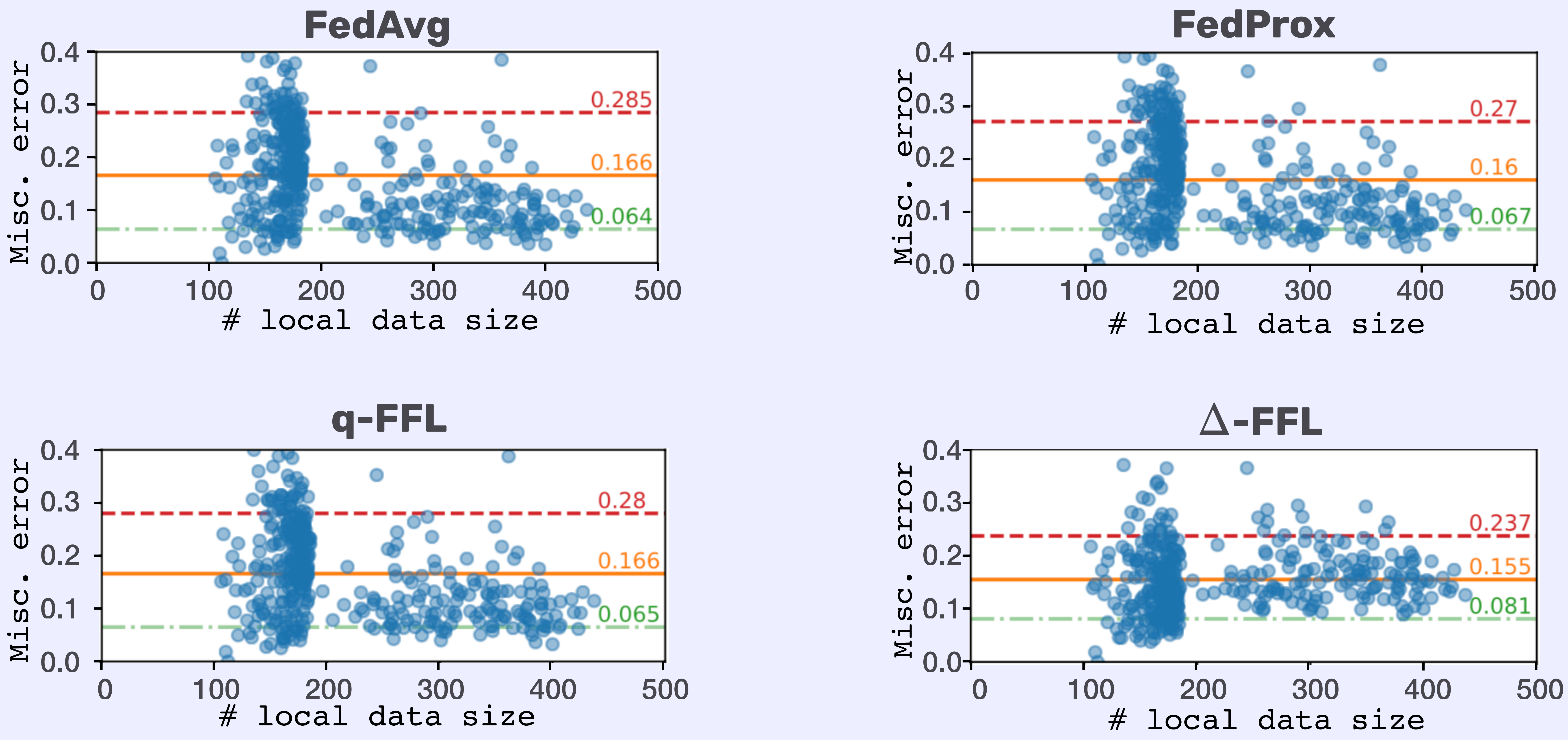
- Average Misclassification Error

Average of misclassification error (in %) on test devices.

	EMNIST		Sent140		Shakespeare
	Linear	ConvNet	Linear	RNN	RNN
FedAvg	34.38 ± 0.38	16.64 ± 0.50	34.75 ± 0.31	30.16 ± 0.44	$\mathbf{42.90} \pm 0.04$
FedProx	$\mathbf{33.82} \pm 0.30$	16.02 ± 0.54	34.74 ± 0.31	30.20 ± 0.48	43.05 ± 0.11
q-FFL	34.34 ± 0.33	16.59 ± 0.30	34.48 ± 0.06	$\mathbf{29.96} \pm 0.56$	42.91 ± 0.09
AFL	39.33 ± 0.27	33.01 ± 0.37	35.98 ± 0.08	37.74 ± 0.65	73.28 ± 1.13
Δ-FL $\theta = 0.8$	34.49 ± 0.26	16.09 ± 0.40	$\mathbf{34.41} \pm 0.22$	30.31 ± 0.33	42.93 ± 0.05
Δ-FL $\theta = 0.5$	35.02 ± 0.20	$\mathbf{15.49} \pm 0.30$	35.29 ± 0.25	33.59 ± 2.44	43.13 ± 0.05
Δ-FL $\theta = 0.1$	38.33 ± 0.38	16.37 ± 1.03	37.79 ± 0.89	51.98 ± 11.81	44.18 ± 0.12

Experimental Results - Local Performance vs Data-Size

- Scatter plot of local final performance VS local data-size



Conclusion

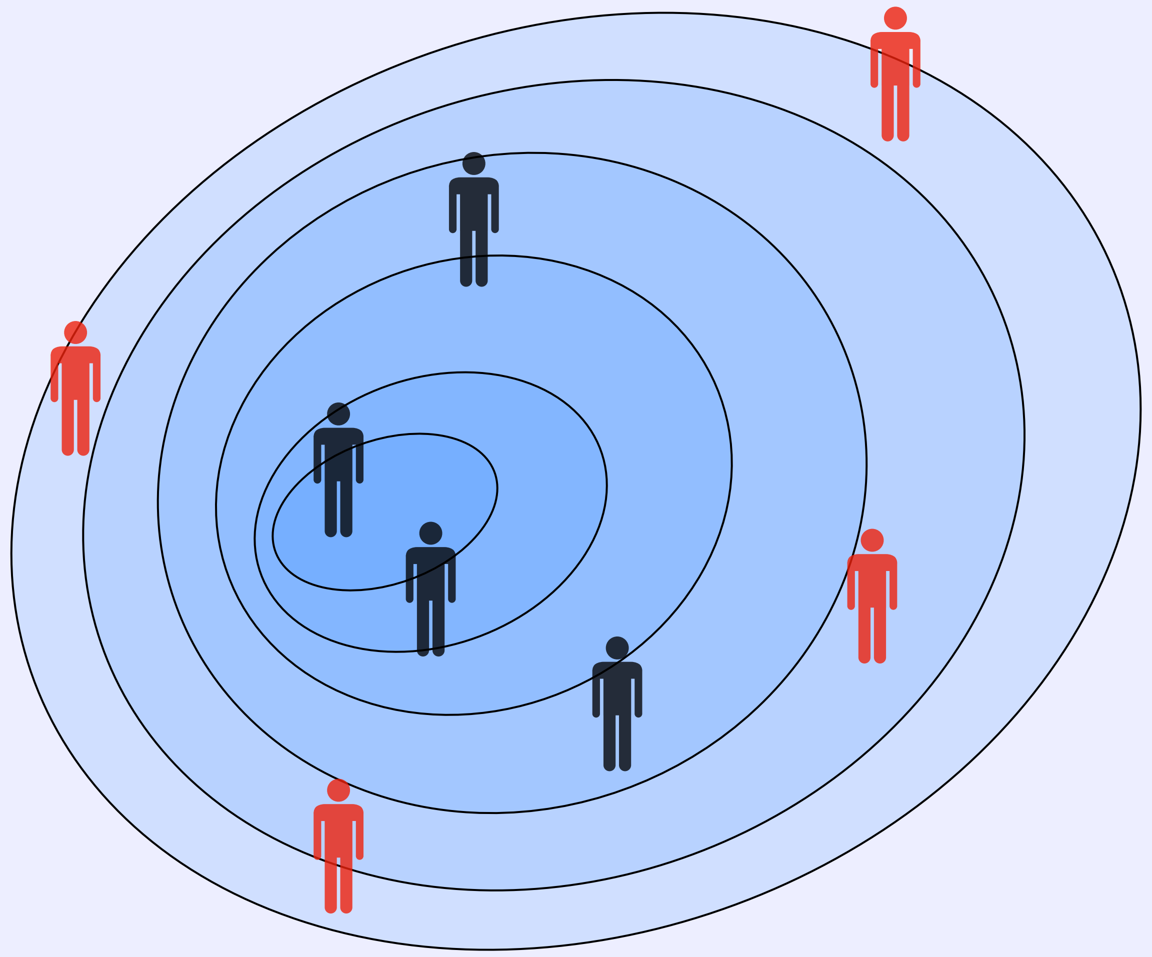
1 The Δ-FL Framework

2 Δ-FL in Practice

3 Numerical Experiments and Comparisons

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.
- Extension of the analysis to the non-convex setting.



Email me at yassine.laguel@univ-grenoble-alpes.fr