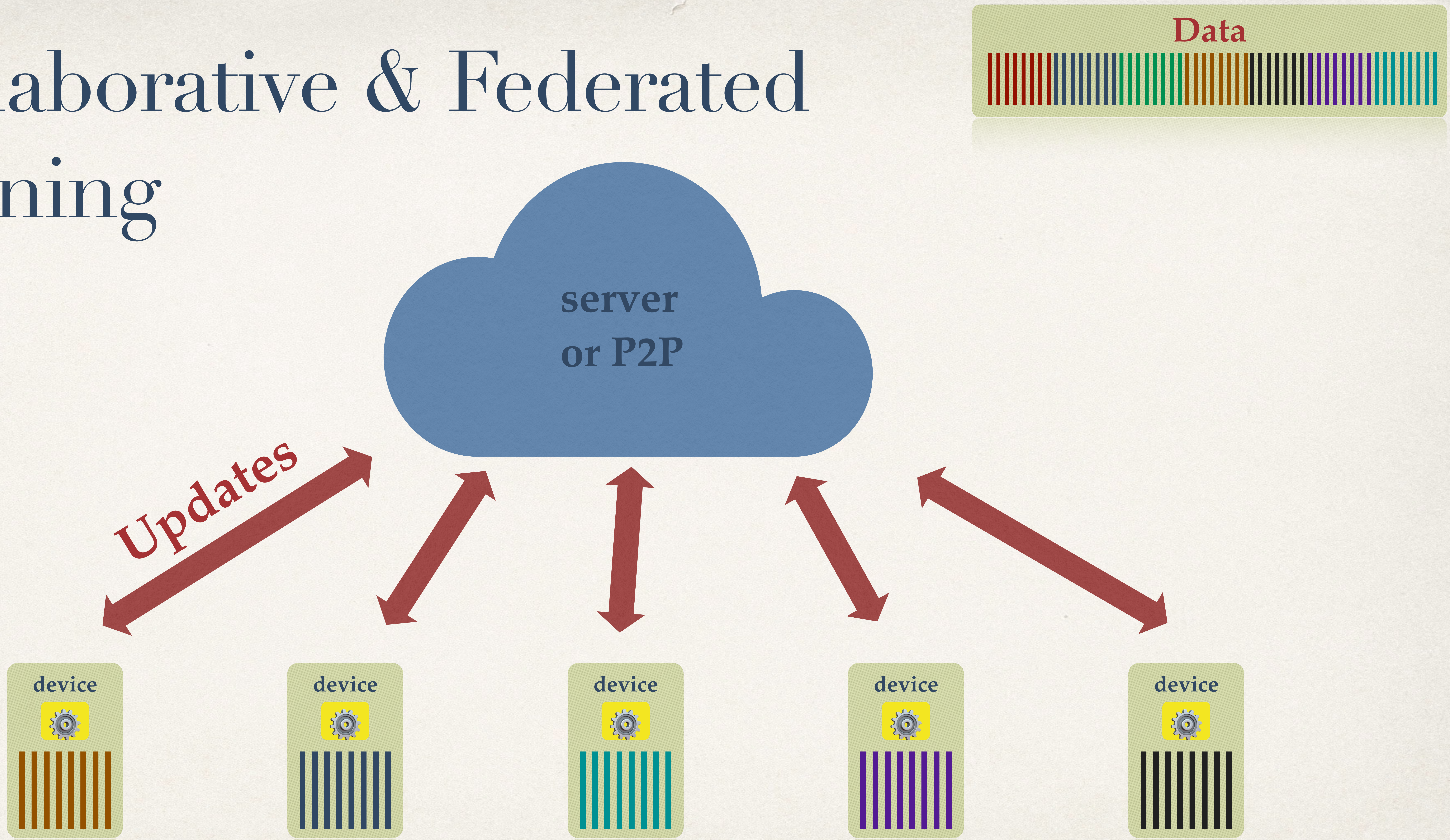
Learning with Strange Gradients

Martin Jaggi
EPFL
Machine Learning and Optimization Laboratory
mlo.epfl.ch

Collaborative & Federated Training



1. Gradients from strange collaborators:
    - **Federated Learning**

2. Gradients from strange collaborators:
    - **Personalization**

3. Gradients from strange **architectures**

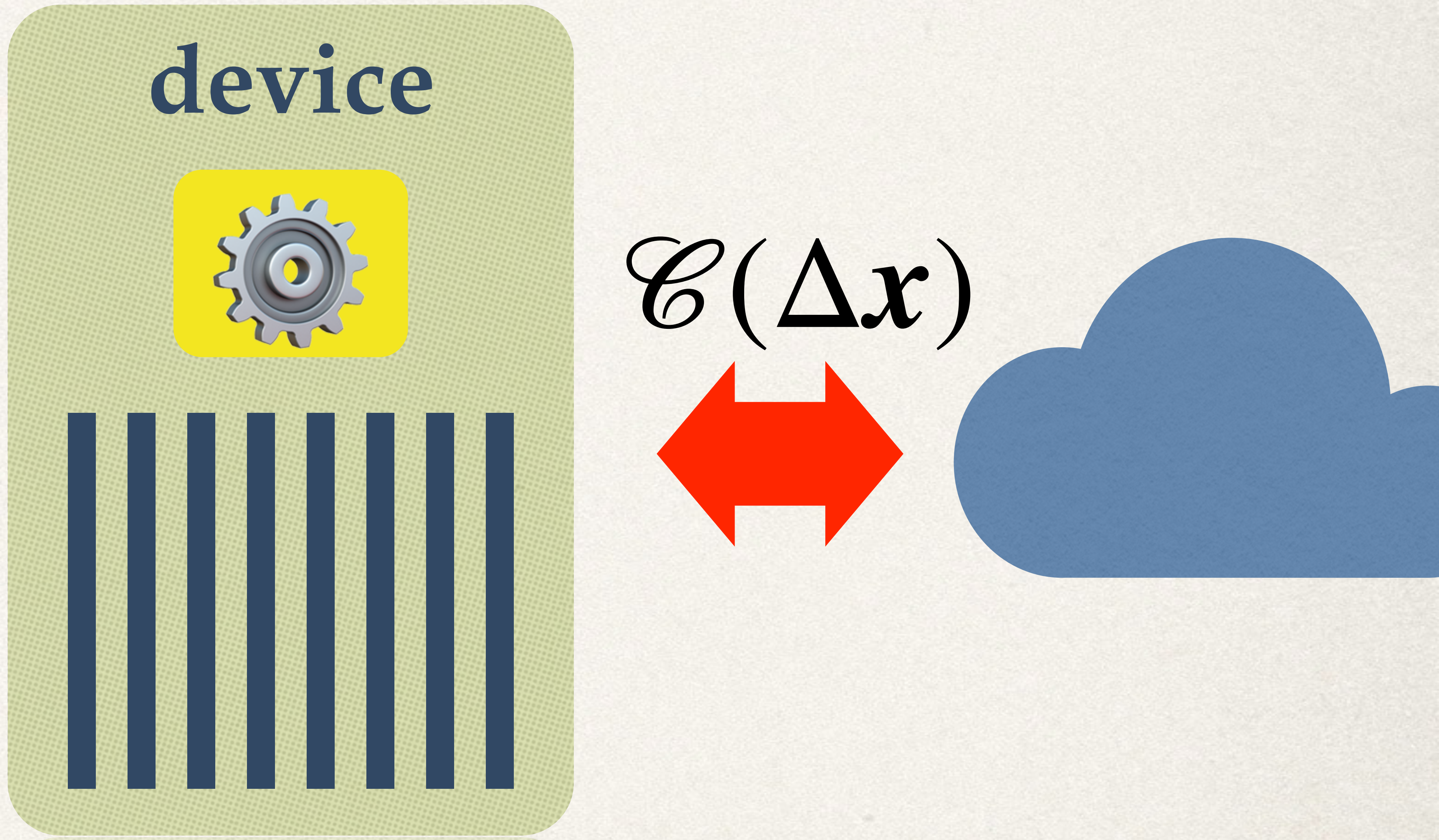
4. Gradients from faulty/**malicious collaborators**:
    - Byzantine-robust Training

Stochastic Gradient Descent (SGD)

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \frac{1}{|data|} \sum_{i \in data} f_i(\boldsymbol{x})$$

$$i_t \sim \text{Uniform}(1, |data|)$$

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t + \Delta \boldsymbol{x}$$



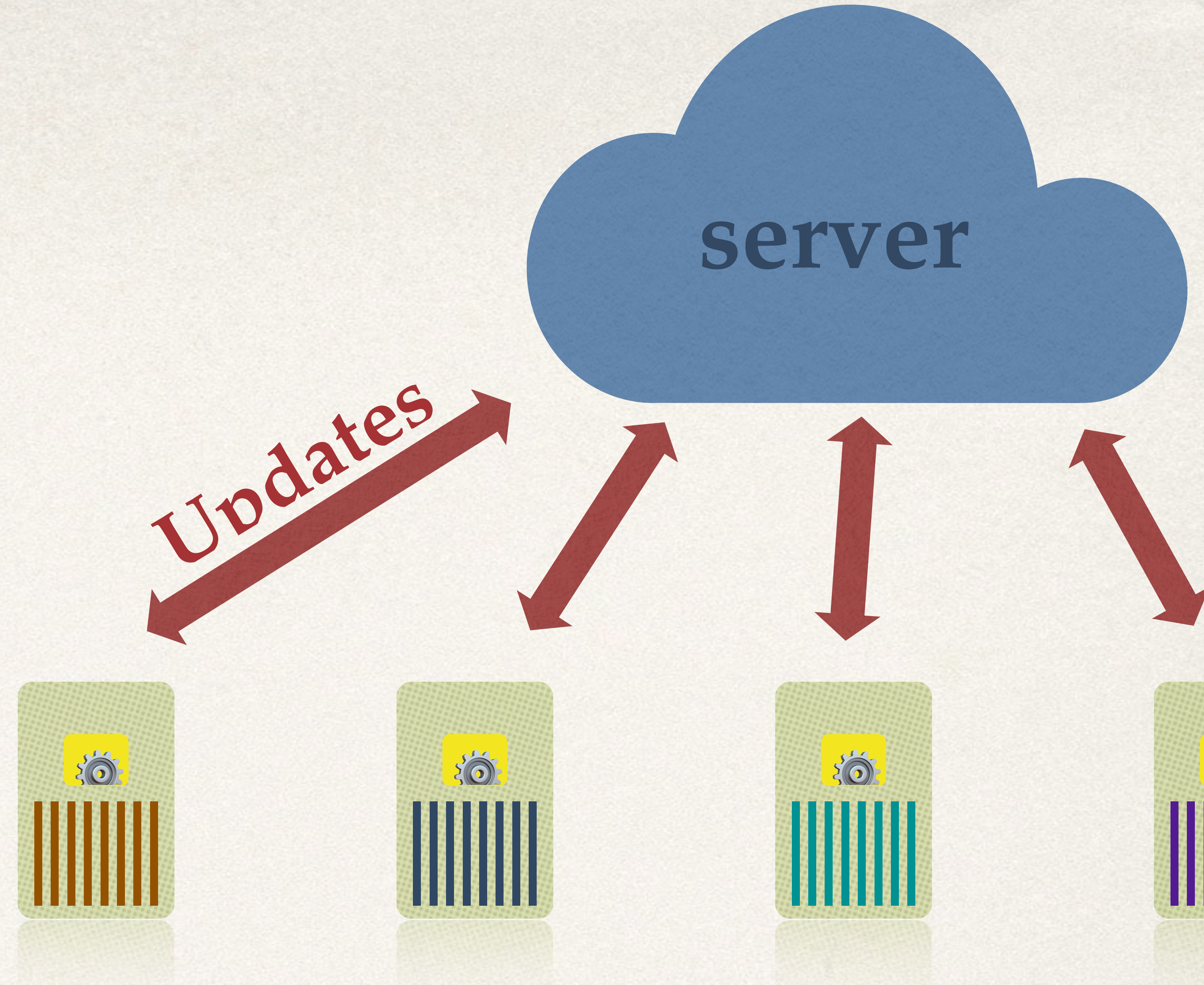
$$\Delta \boldsymbol{x} = -\gamma_t \nabla f_{i_t}(\boldsymbol{x}_t)$$ **from backpropagation**

1

Gradients from strange collaborators:

- Federated Learning

Client drift



- **Federated Learning**

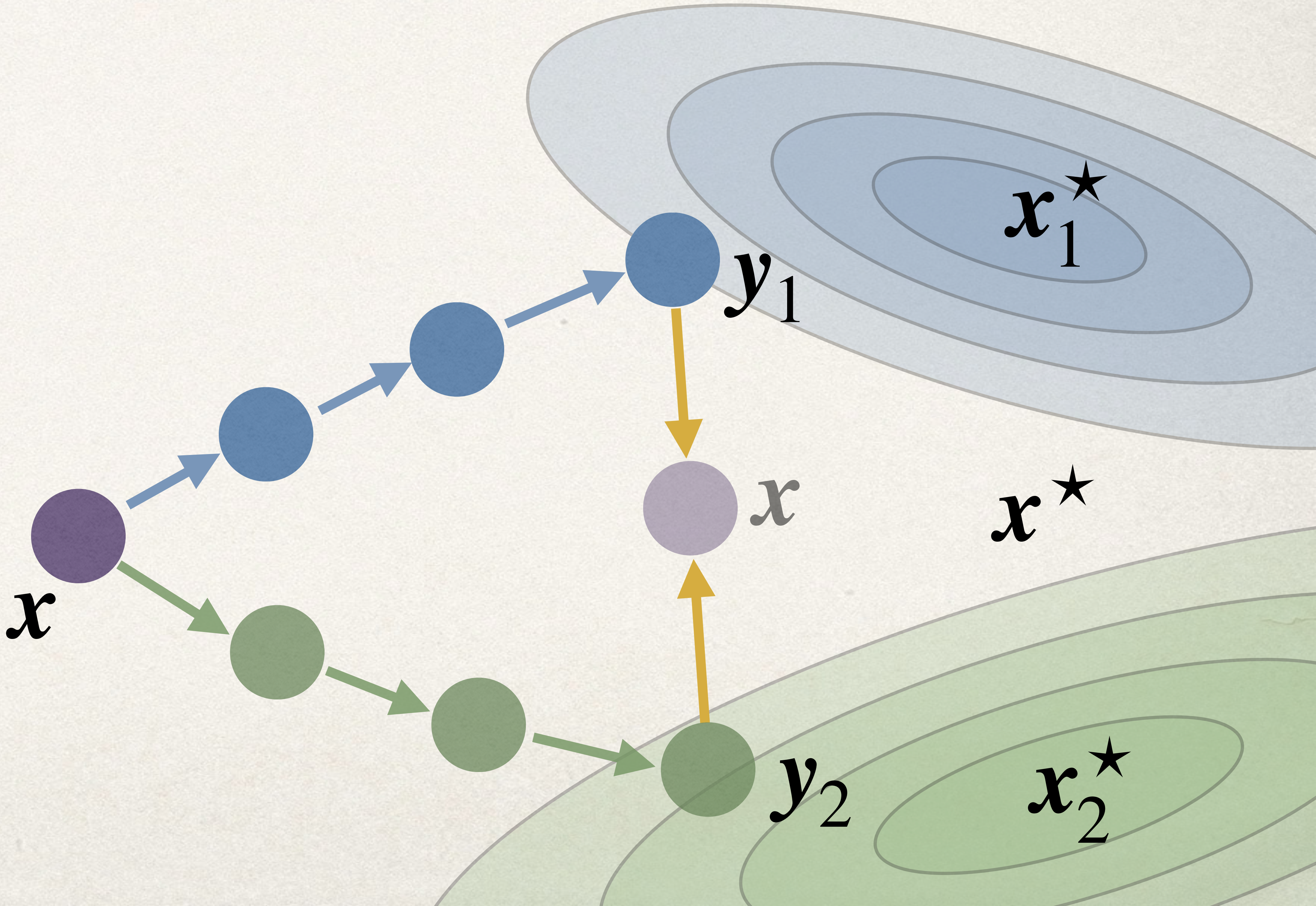
$$\min_{\boldsymbol{x}} \frac{1}{n}\sum_{i}^{n} f_i(\boldsymbol{x})$$

- **Fed Avg / Local SGD**

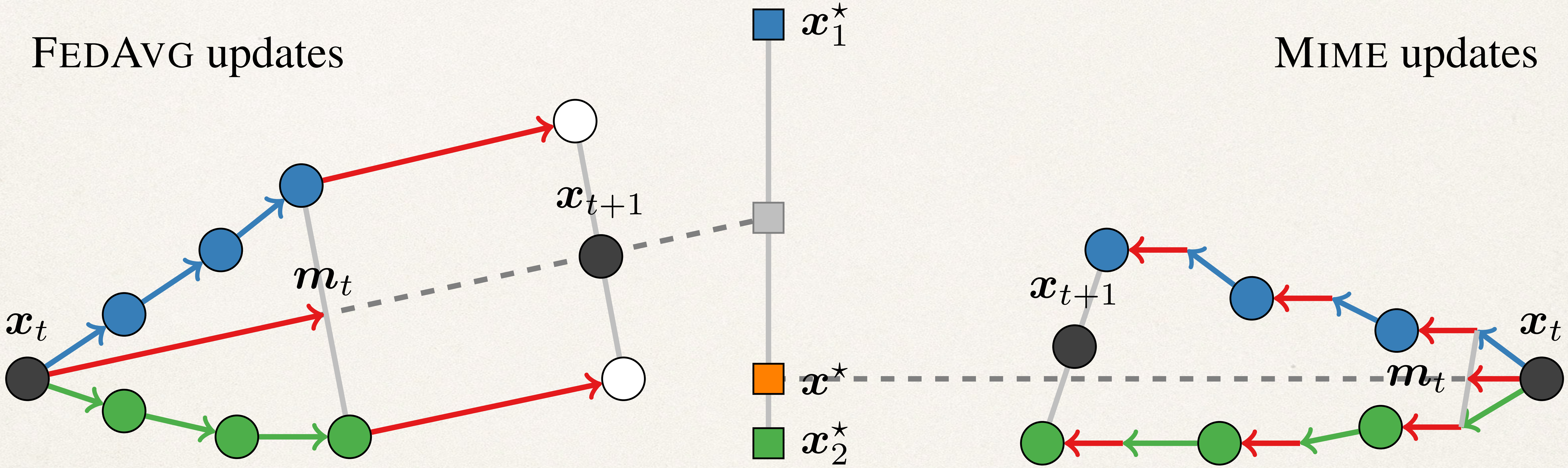
for some local steps

$$\boldsymbol{y}_i := \boldsymbol{y}_i - \eta \nabla f_i(\boldsymbol{y}_i)$$

$$\boldsymbol{x} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \qquad \textit{(aggregation)}$$



Client drift



Mime algorithm framework

for *some local steps*

$$\boldsymbol{y}_i := \boldsymbol{y}_i - \eta\Big((1-\beta)\nabla f_i(\boldsymbol{y}_i) + \beta\boldsymbol{m}\Big)$$

$$\boldsymbol{m} := (1-\beta)\nabla f_i(\boldsymbol{x}) + \beta\boldsymbol{m}$$

aggregated on server after each round

Mime convergence

Number of rounds to reach

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}^{out})\|^2\right] \leq \varepsilon \ :$$

$$\mathcal{O}\left(\left(\frac{n}{S}\right)^{3/2}\frac{L}{\varepsilon}\right)$$ **Scaffold**

$$\mathcal{O}\left(\frac{\delta(\zeta+\sigma)}{\varepsilon^{3/2}} + \frac{\zeta^2+\sigma^2}{\varepsilon} + \frac{\delta}{\varepsilon}\right)$$ **MimeLiteMVR**

$$\mathcal{O}\left(\frac{\delta\zeta}{\sqrt{S}\varepsilon^{3/2}} + \frac{\zeta^2}{S\varepsilon} + \frac{\delta}{\varepsilon}\right)$$ **MimeMVR**

$$\Omega\left(\frac{L\zeta}{\sqrt{S}\varepsilon^{3/2}} + \frac{\zeta^2}{S\varepsilon} + \frac{L}{\varepsilon}\right)$$ **Lower bound (server-only)**

Data Heterogeneity:

$\delta \ll L$ inter-cl. Hessian similarity

ζ inter-cl. gradient variance

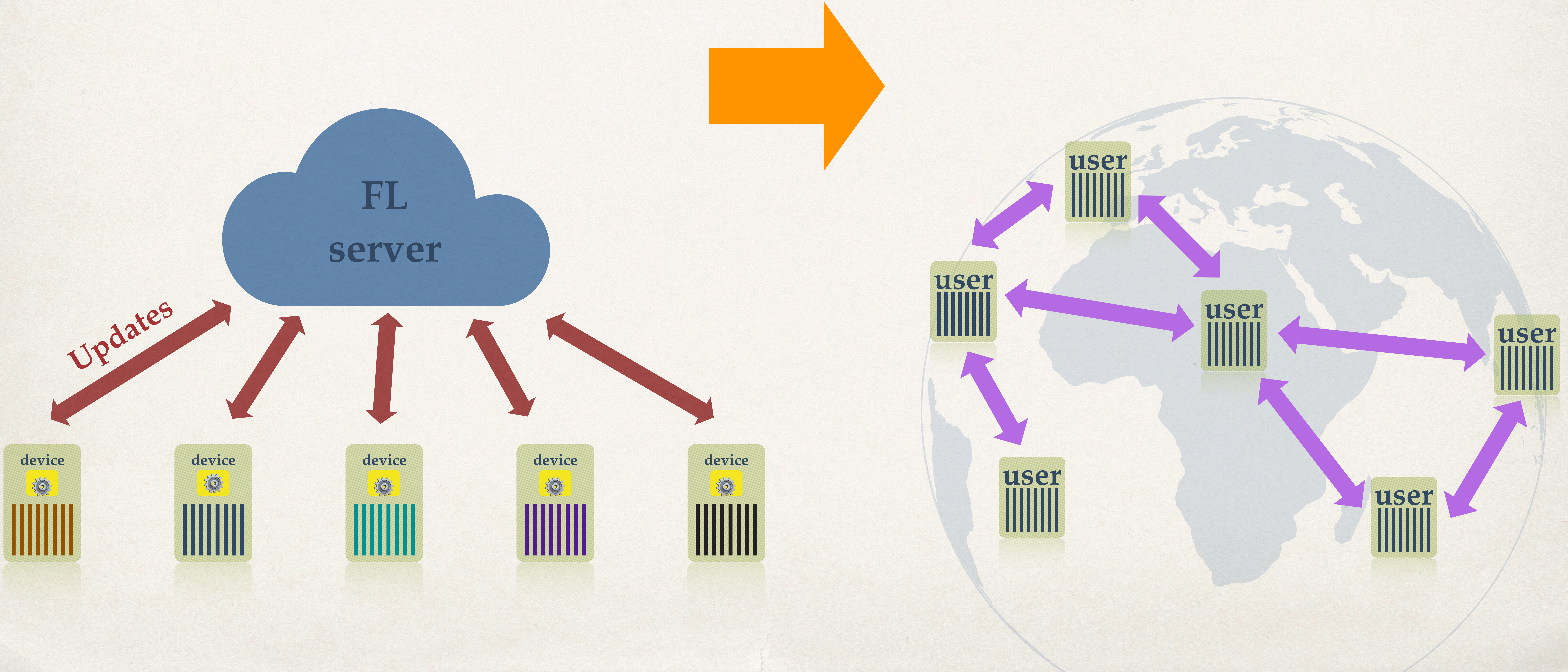
σ intra-cl. gradient variance

2

Gradients from strange collaborators:

- Personalization

From federated towards decentralized



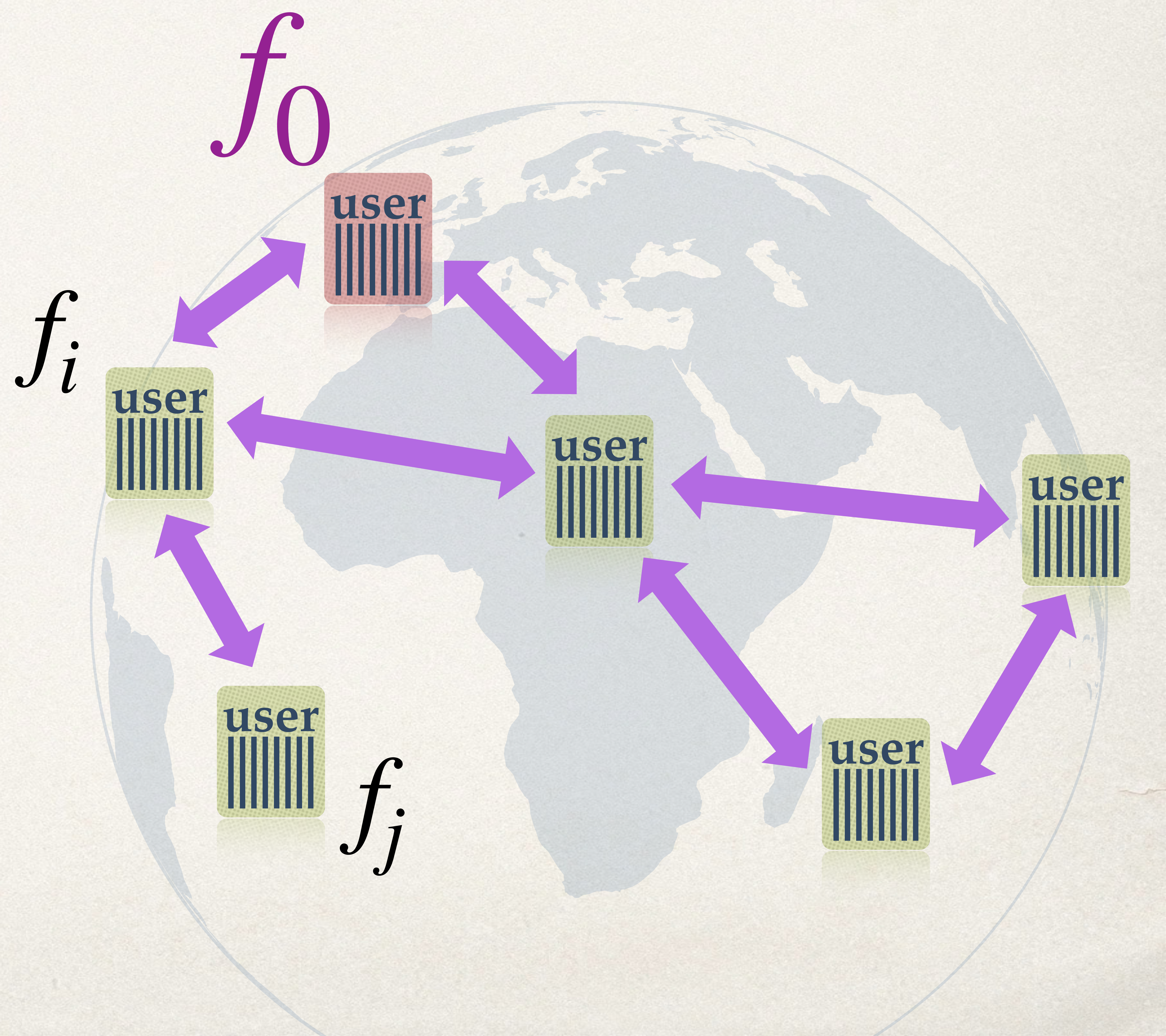
Collaborative Learning

- **Federated**

$$\min_{\boldsymbol{x}} \; \frac{1}{n} \sum_{i}^{n} f_i(\boldsymbol{x})$$

- **Collaborative / Personalized**

$$\min_{\boldsymbol{x}} \; f_0(\boldsymbol{x}) \quad \Bigg| \quad \begin{array}{l} \min_{\boldsymbol{x}} \; f_1(\boldsymbol{x}) \\ \\ \min_{\boldsymbol{x}} \; f_n(\boldsymbol{x}) \end{array}$$



Personalized learning / optimization

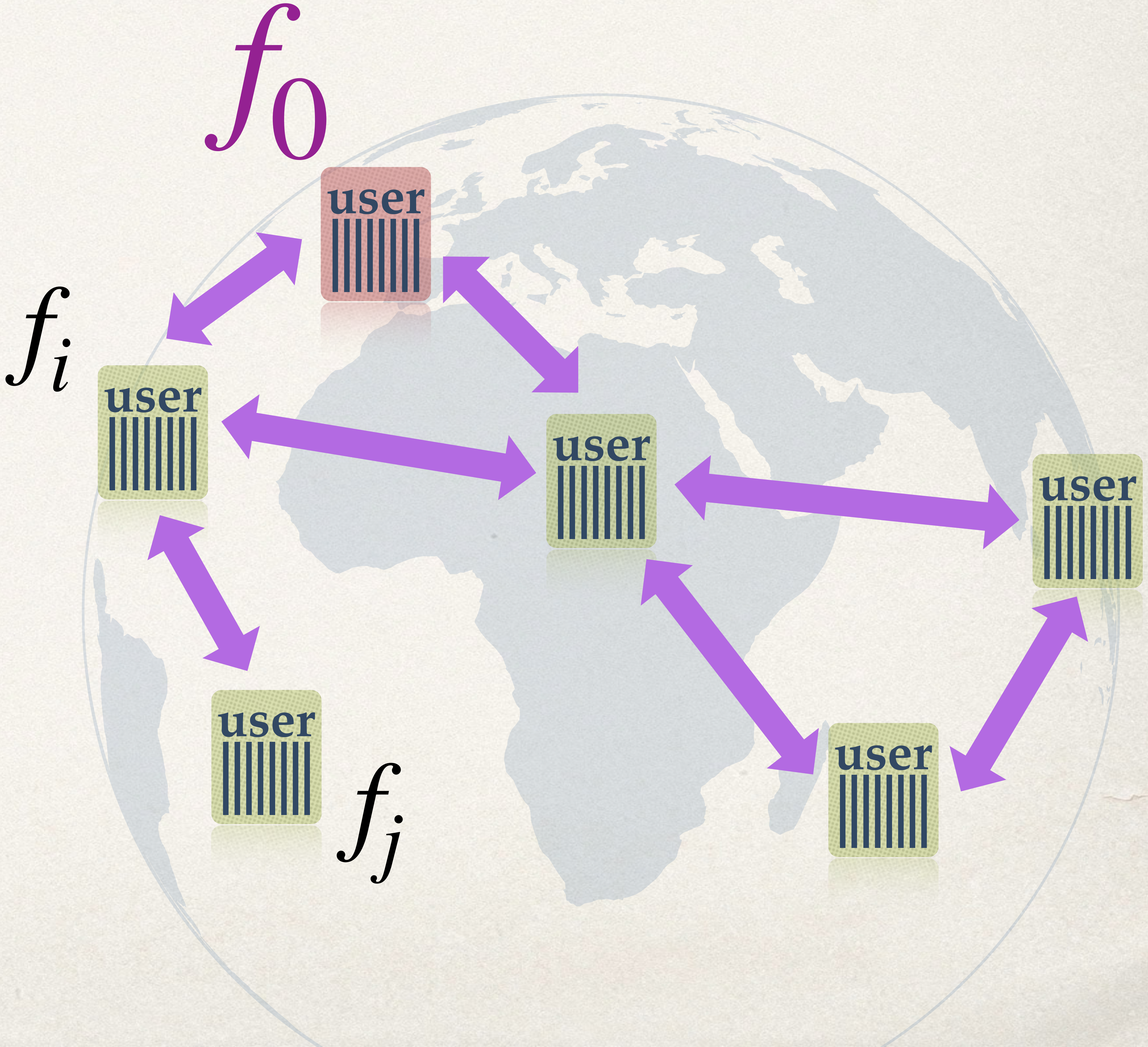
- **Weighted averaging**

$$x := x - \gamma \sum_{i=0}^{n} \alpha_i \nabla f_i(x)$$

- **Weighted averaging with bias correction**

$$x := x - \gamma \sum_{i=0}^{n} \left(\alpha_i \nabla f_i(x) + c_i\right)$$

idea similar to Scaffold



Theorem: Convergence on personal objective f_0 for non-convex smooth objectives, using exponential moving average to learn $\boldsymbol{c}_i$

$$\mathbb{E}\left[\|\nabla f_0(\boldsymbol{x}^{out})\|^2\right] = \mathcal{O}\left(\sqrt{\frac{LF_0\,\sigma_0^2}{(n+1)T}}\right)$$

3

Gradients from strange architectures

Alternating Partial Training for Neural Nets

Theorem: Convergence on original network f for non-convex smooth objectives,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = \mathcal{O}\left(\sqrt{\frac{q^4 L F_0 \sigma^2}{T}}\right)$$

and similarly for smaller core network

q : "gradient alignment" between parent and core network

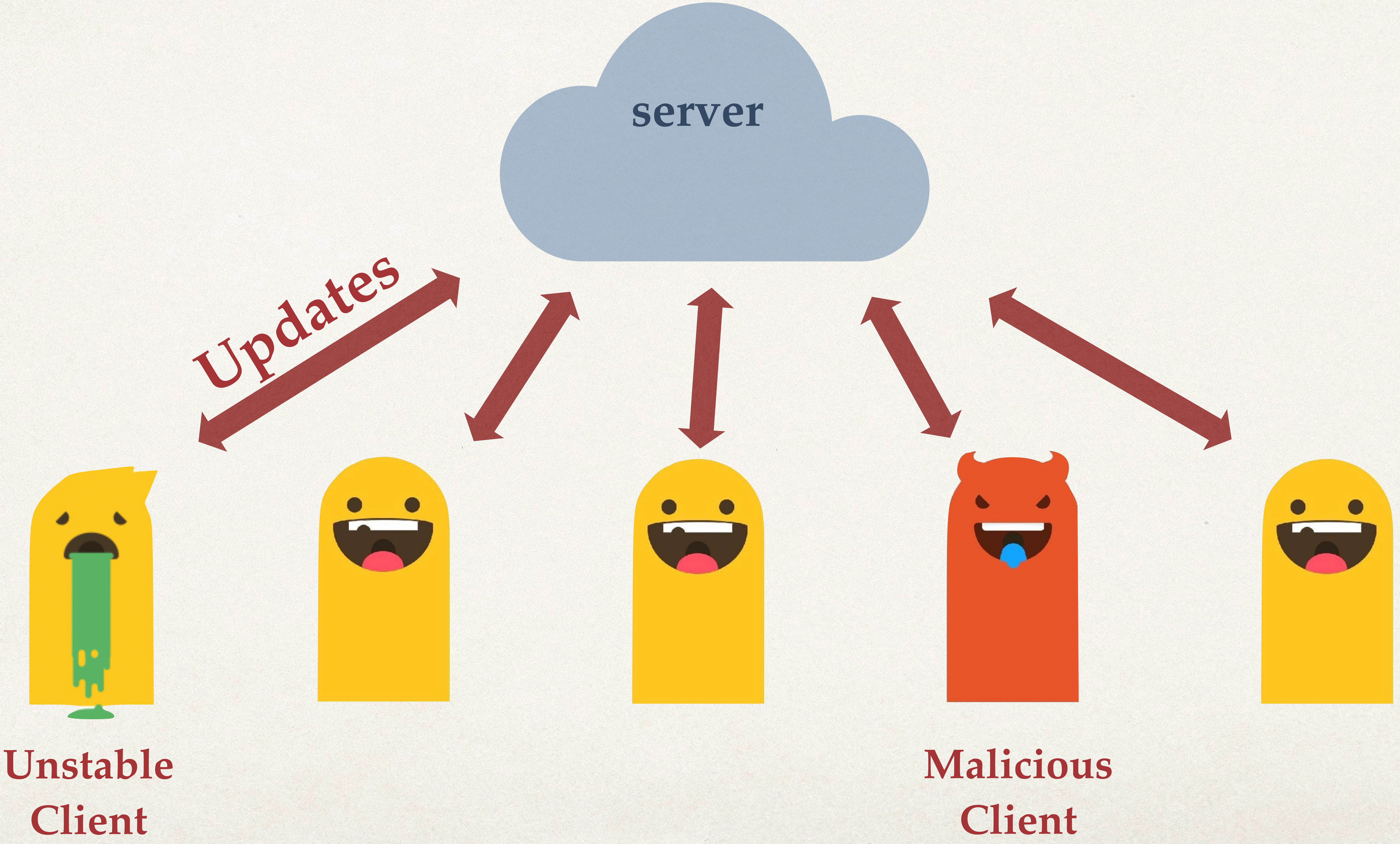
- Masked Training of Neural Networks with Partial Gradients, arXiv
- AC/DC: Alternating Training of Deep Neural Networks, Peste et al, NeurIPS 2021

4

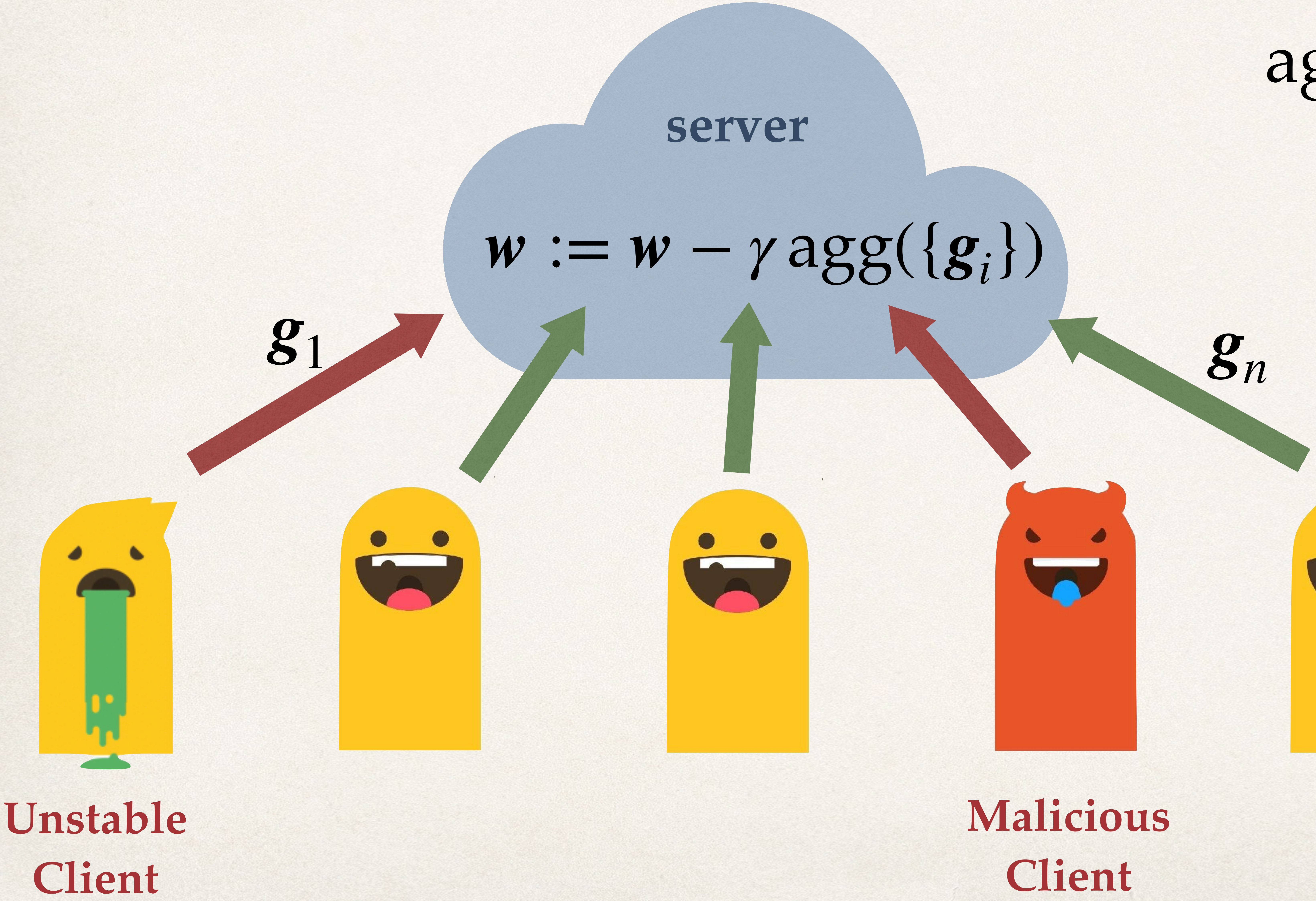
Gradients from faulty/malicious collaborators:

- Byzantine-robust Training

Malicious actors in FL



Byzantine Robust Training



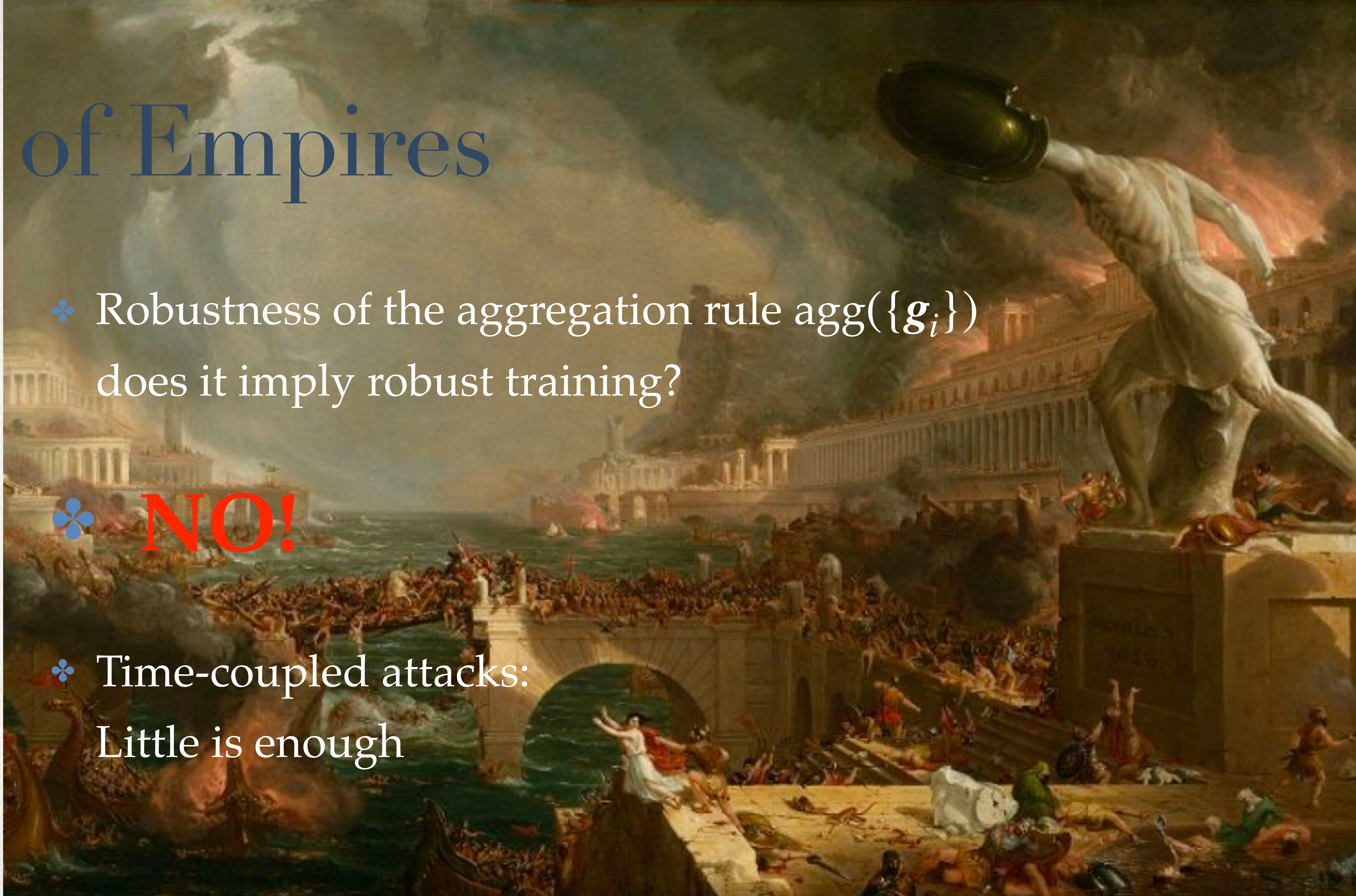
$$\mathrm{agg}(\{\boldsymbol{g}_i\}) := \mathrm{avg}(\{\boldsymbol{g}_i\})$$

$$:= \mathrm{CM}(\{\boldsymbol{g}_i\})$$

Examples:

- Coordinate-wise median [Yin et al. 2017]
- Krum [Blanchard et al. 2018]
- Geometric median / RFA [Pillutla et al. 2019]

Fall of Empires



- Robustness of the aggregation rule $\mathrm{agg}(\{\boldsymbol{g}_i\})$ does it imply robust training?
- **NO!**
- Time-coupled attacks: Little is enough

Strong negative result

- Any aggregation rule which does not use history will **fail** training (convergence)

Fix: Using history with momentum

- Simply use worker momentum

$$\boldsymbol{m}_i := (1-\beta)\boldsymbol{g}_i + \beta\boldsymbol{m}_i$$

- Effectively averages past gradients, reducing variance

- Aggregate worker momentum instead of gradients

$$\boldsymbol{w} := \boldsymbol{w} - \gamma\,\mathrm{agg}(\{\boldsymbol{m}_i\})$$

Aggregation with Centered Clipping

- Norm-based clipping, before averaging

$$CC = \boldsymbol{v} + \mathrm{clip}_\tau(\boldsymbol{g}_i - \boldsymbol{v})$$

- Removes outliers

- Center at previous aggregated update

Robustness theorem

Theorem: Given any **($\delta_{\max}$, c)-robust aggregator**, under a **δ**-fraction of attackers and **σ^2** variance, our algorithm outputs **x^{out}** s.t.

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\left(\delta + \frac{1}{n}\right)}\right)$$

References

1

- **Mime: Mimicking Centralized Stochastic Algorithms in Federated Learning**
    - NeurIPS 2021 *paper link*

2

- **Optimal Model Averaging: Towards Personalized Collaborative Learning**
    - FL workshop at ICML 2021 *paper link*
- **Linear Speedup in Personalized Collaborative Learning**
    - arXiv *paper link*

3

- **Masked Training of Neural Networks with Partial Gradients**
    - arXiv *paper link*

4

- **Learning from History for Byzantine Robust Optimization**
    - ICML 2021 *paper link*

Thanks

Sai Praneeth Karimireddy, Sebastian U. Stich, Lie He, El Mahdi Chayti, Amirkeivan Mohtashami, Felix Grimberg, Nicolas Flammarion, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Ananda Theertha Suresh

EPFL

Machine Learning and Optimization Laboratory

website: mlo.epfl.ch