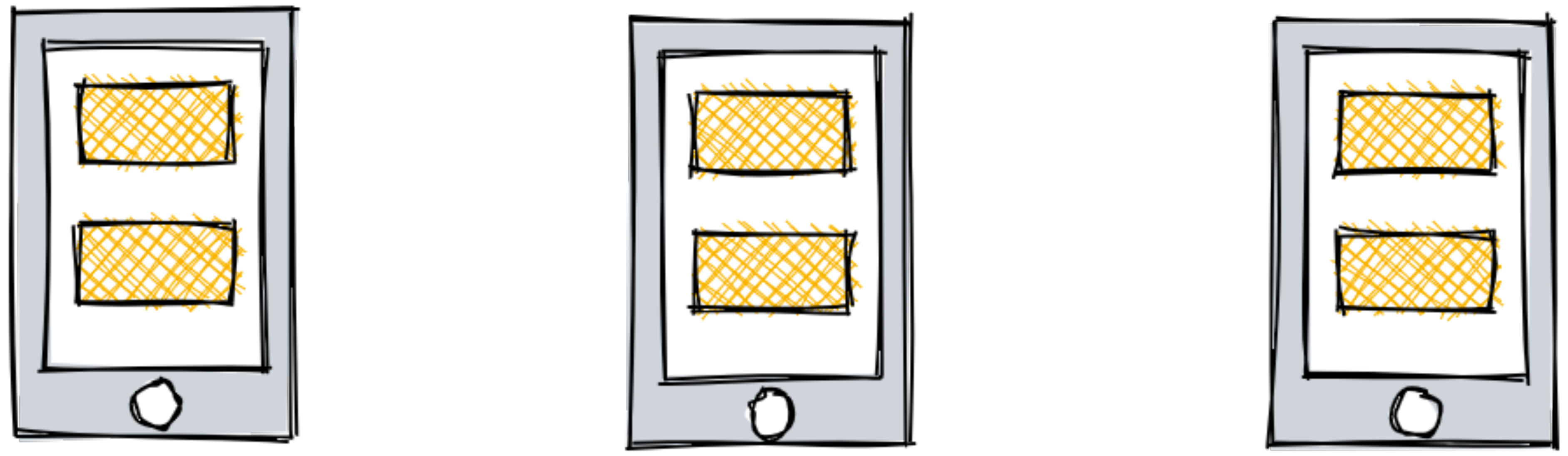
Federated Learning with Packet Losses

A. Rodio, G. Neglia, F. Busacca, S. Mangione, S. Palazzo, F. Restuccia, I. Tinnirello

Tampa, 19 November 2023

Università di Catania
Università degli Studi di Palermo
Northeastern University

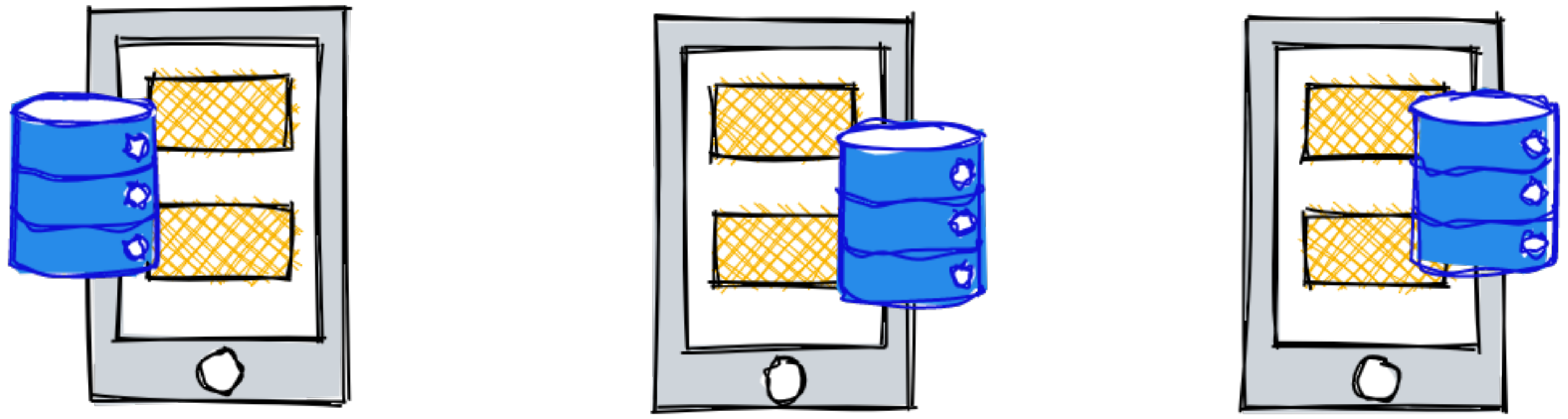
Centralized Learning



Clients $k = 1, \ldots, N$

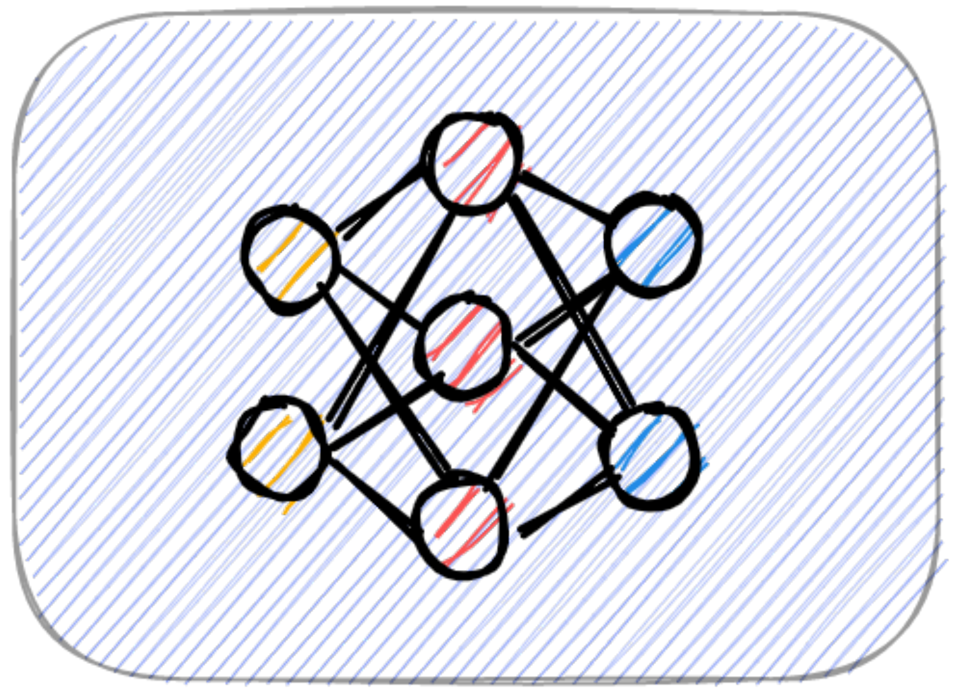
Centralized Learning

D_k
Dataset



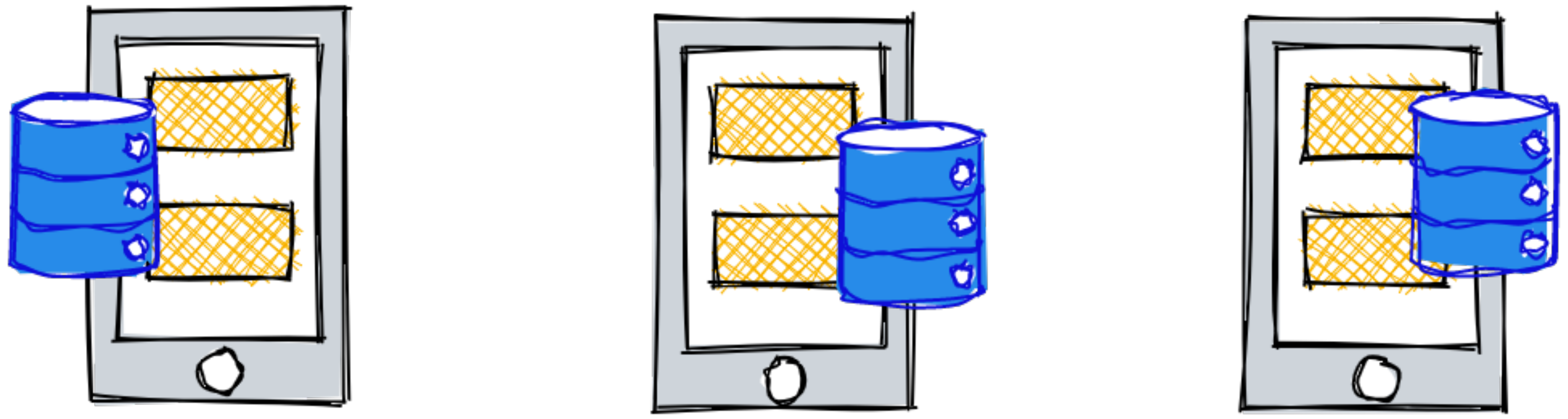
Clients $k = 1, \dots, N$

Centralized Learning



Global model
$\boldsymbol{w} \in \mathbb{R}^d$

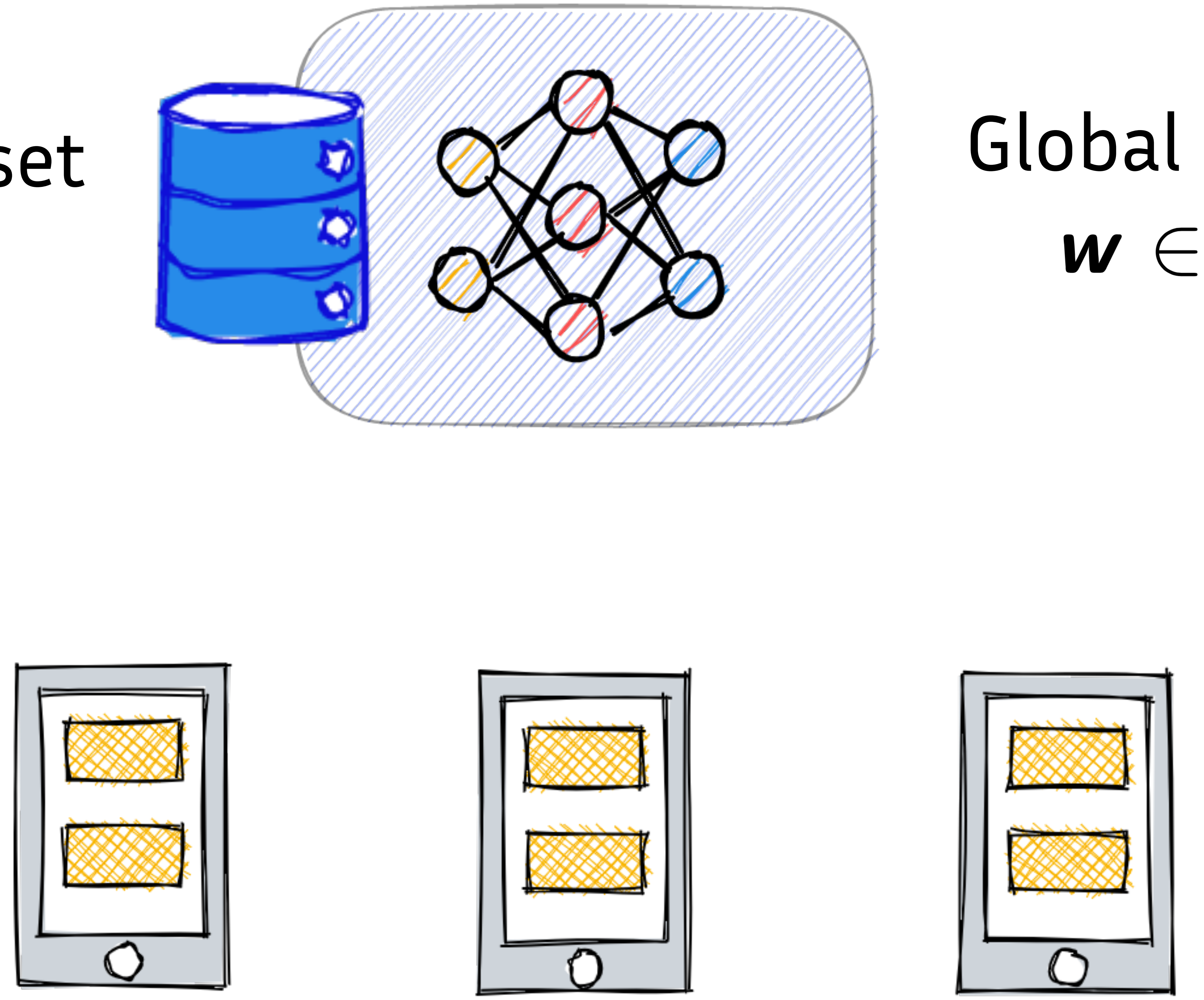
D_k
Dataset



Clients $k = 1, \ldots, N$

Centralized Learning

Dataset
D



Global model
$\boldsymbol{w} \in \mathbb{R}^d$

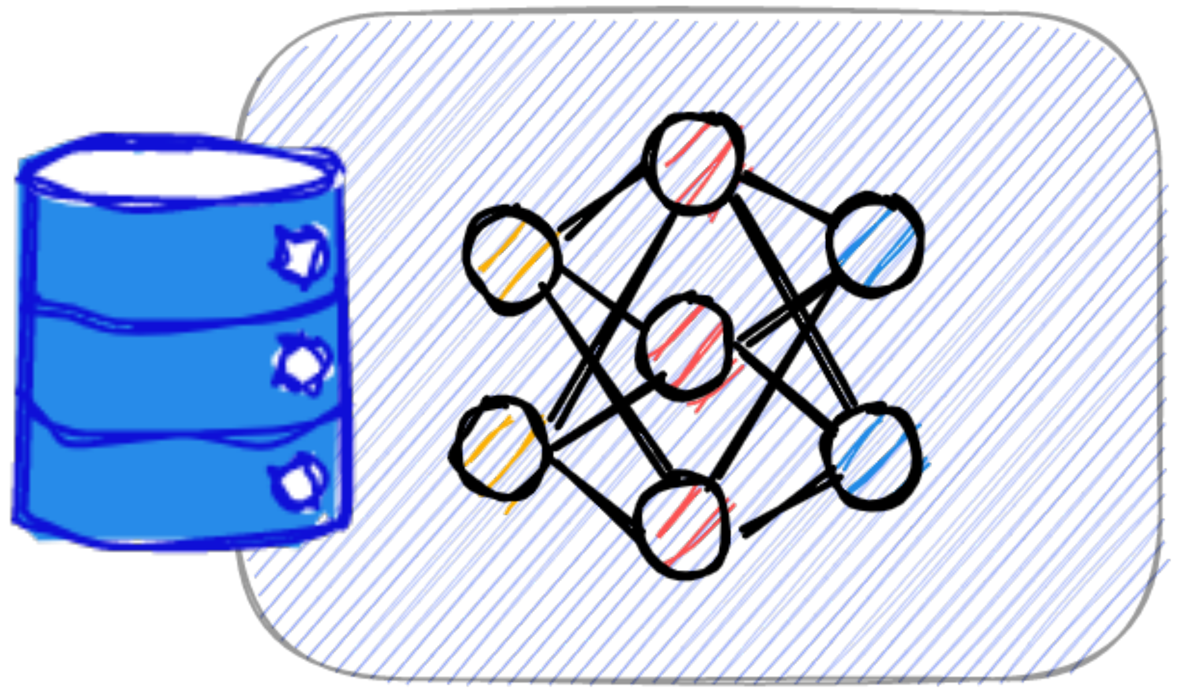
Clients $k = 1, \ldots, N$

Centralized Learning

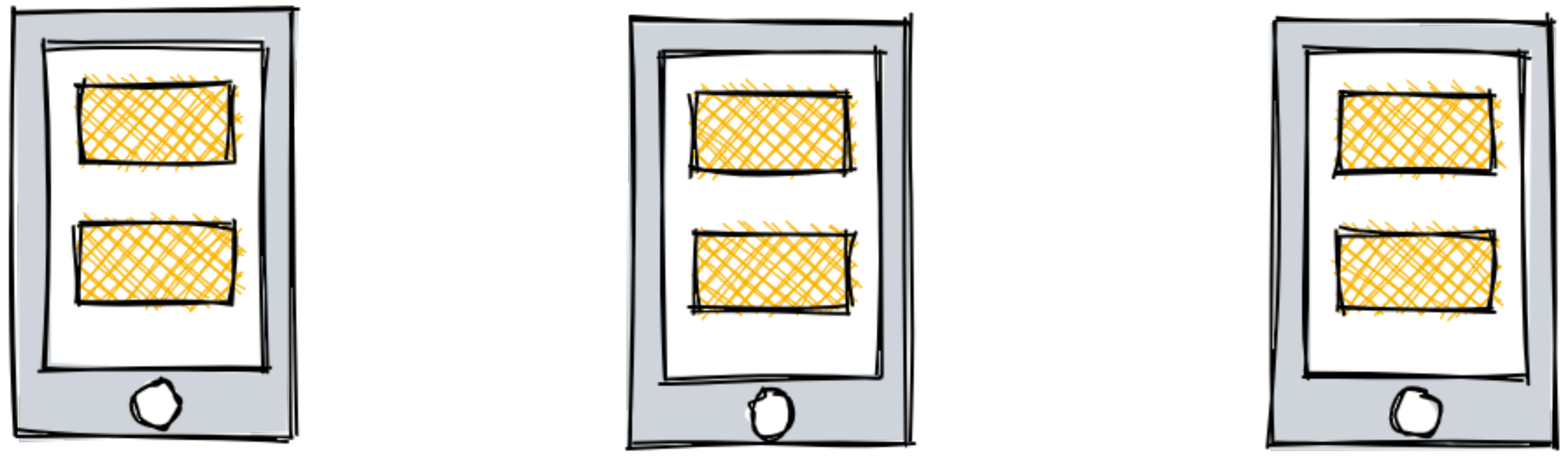
Solve the optimization problem

$$\min_{\boldsymbol{w}} \frac{1}{|D|} \sum_{d \in D} \ell(\boldsymbol{w}, d)$$

Dataset
D



Global model
$\boldsymbol{w} \in \mathbb{R}^d$



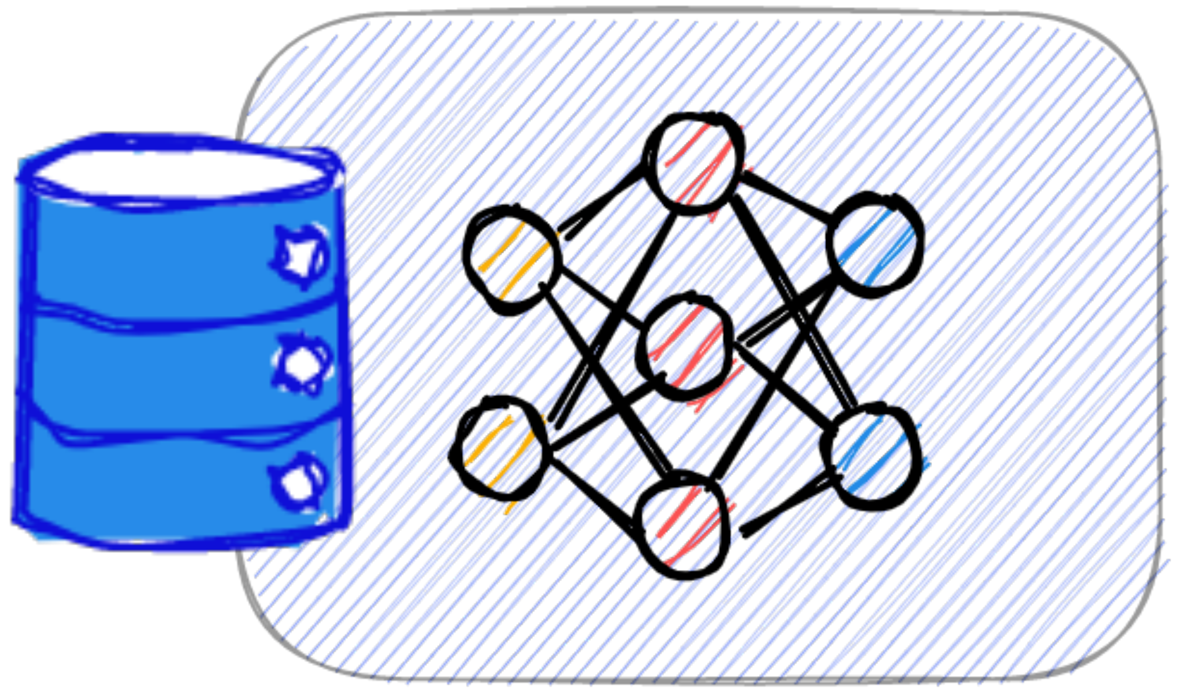
Clients $k = 1, \ldots, N$

Centralized Learning

Solve the optimization problem

$$\min_{\boldsymbol{w}} \frac{1}{|D|} \sum_{d \in D} \ell(\boldsymbol{w}, d)$$

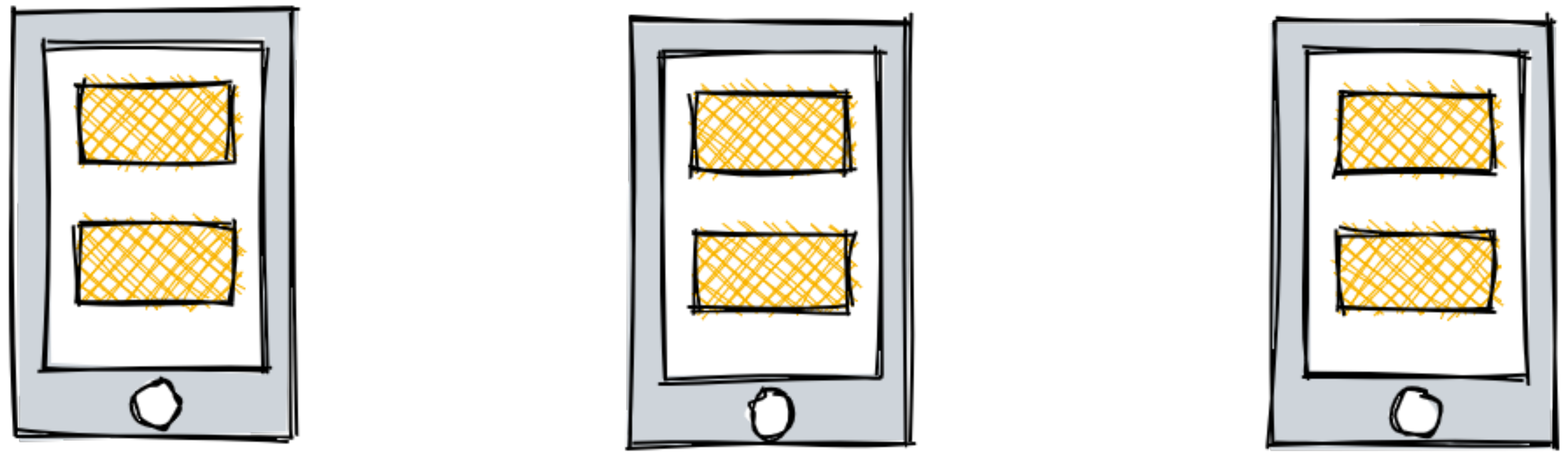
Dataset
D



Global model
$\boldsymbol{w} \in \mathbb{R}^d$

Data transfer

1. Communication cost
2. Privacy



Clients $k = 1, \ldots, N$

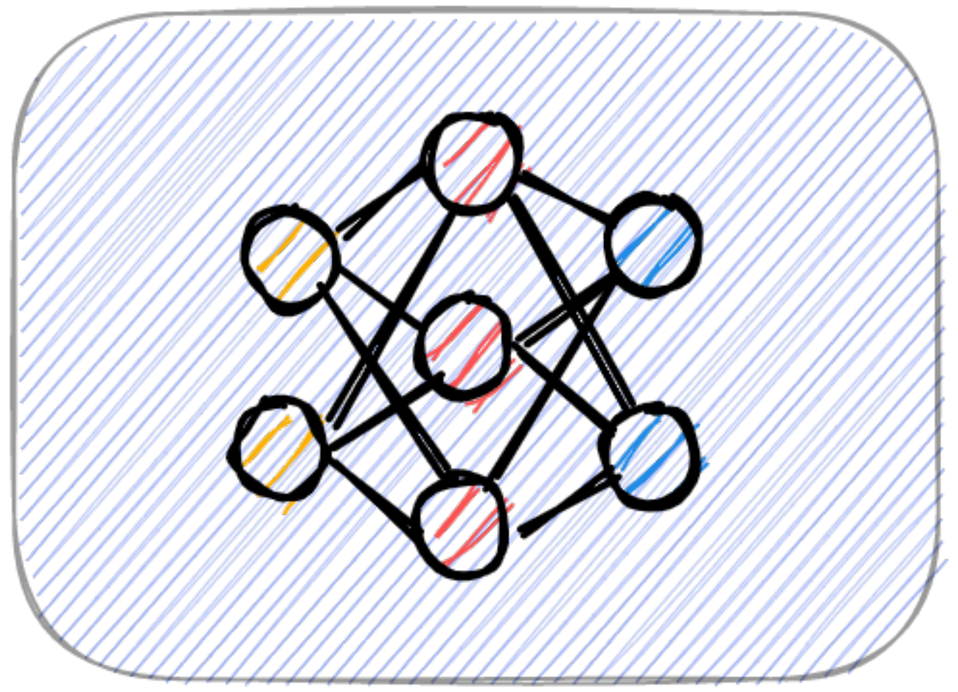
Federated Learning

Solve the optimization problem

$$\min_{\boldsymbol{w}} F(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} F_k(\boldsymbol{w})$$

where

$$F_k(\boldsymbol{w}) = \frac{1}{|D_k|}\sum_{d_k \in D} \ell(\boldsymbol{w}, d_k)$$

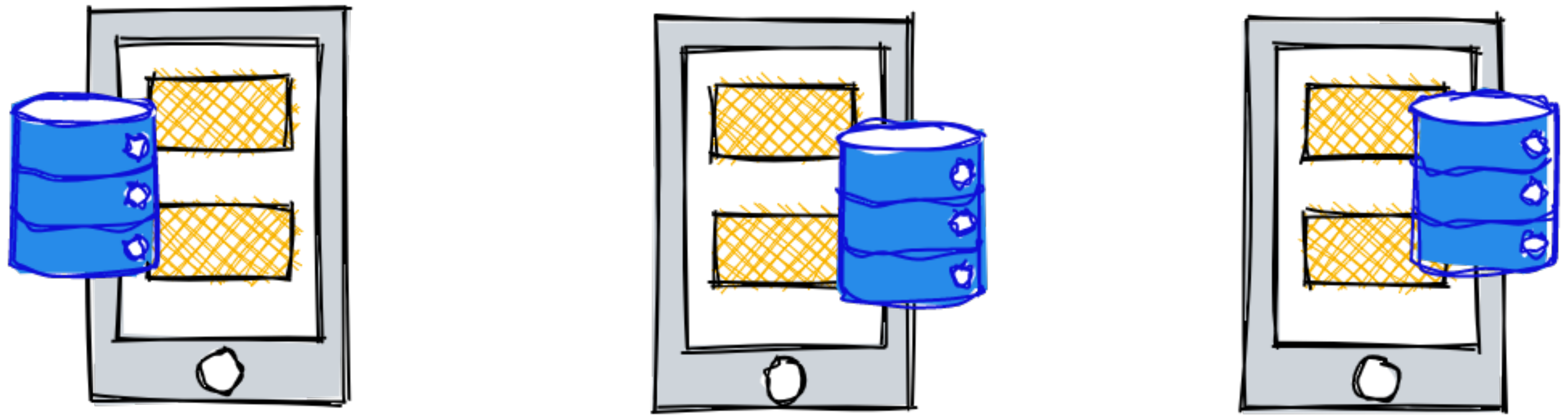


Global model

$\boldsymbol{w} \in \mathbb{R}^d$

D_k

Dataset

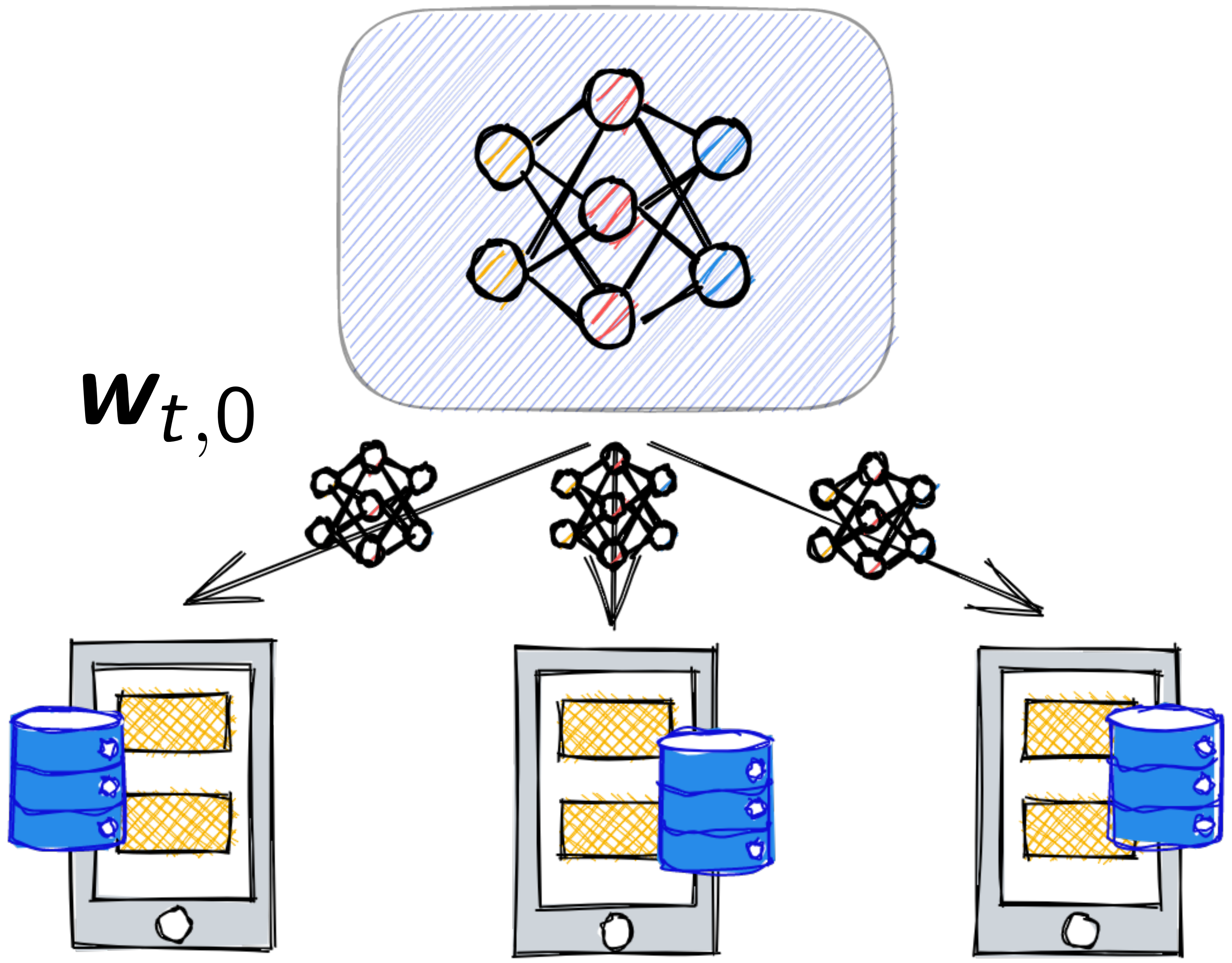


Clients $k = 1, \ldots, N$

Federated Learning

for $t \in \{1, \ldots, T\}$ **do**:

(1) Server broadcasts the initial model



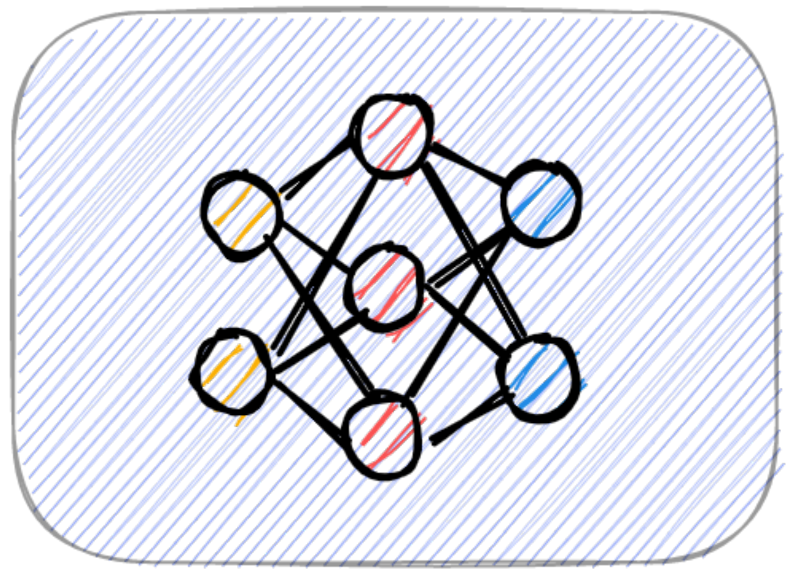
Federated Learning

for $t \in \{1, \dots, T\}$ **do**:

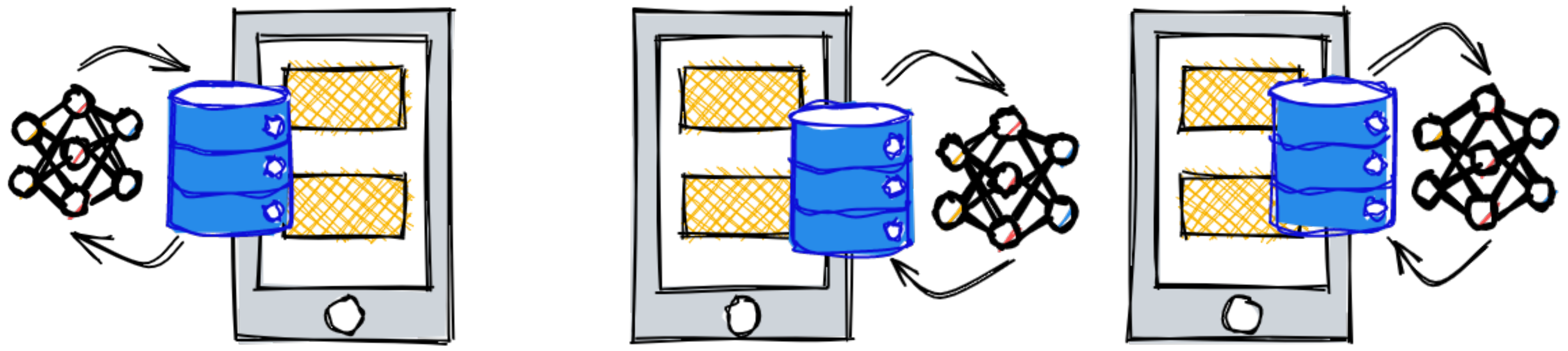
(2) Each client updates its local model

for $j = 0, \dots, E-1$ **do** :

$$\mathbf{w}^k_{t,j+1} = \mathbf{w}^k_{t,j} - \eta_t \nabla F_k(\mathbf{w}^k_{t,j}, \mathcal{B}^k_{t,j})$$



$\mathbf{w}^k_{t,E}$



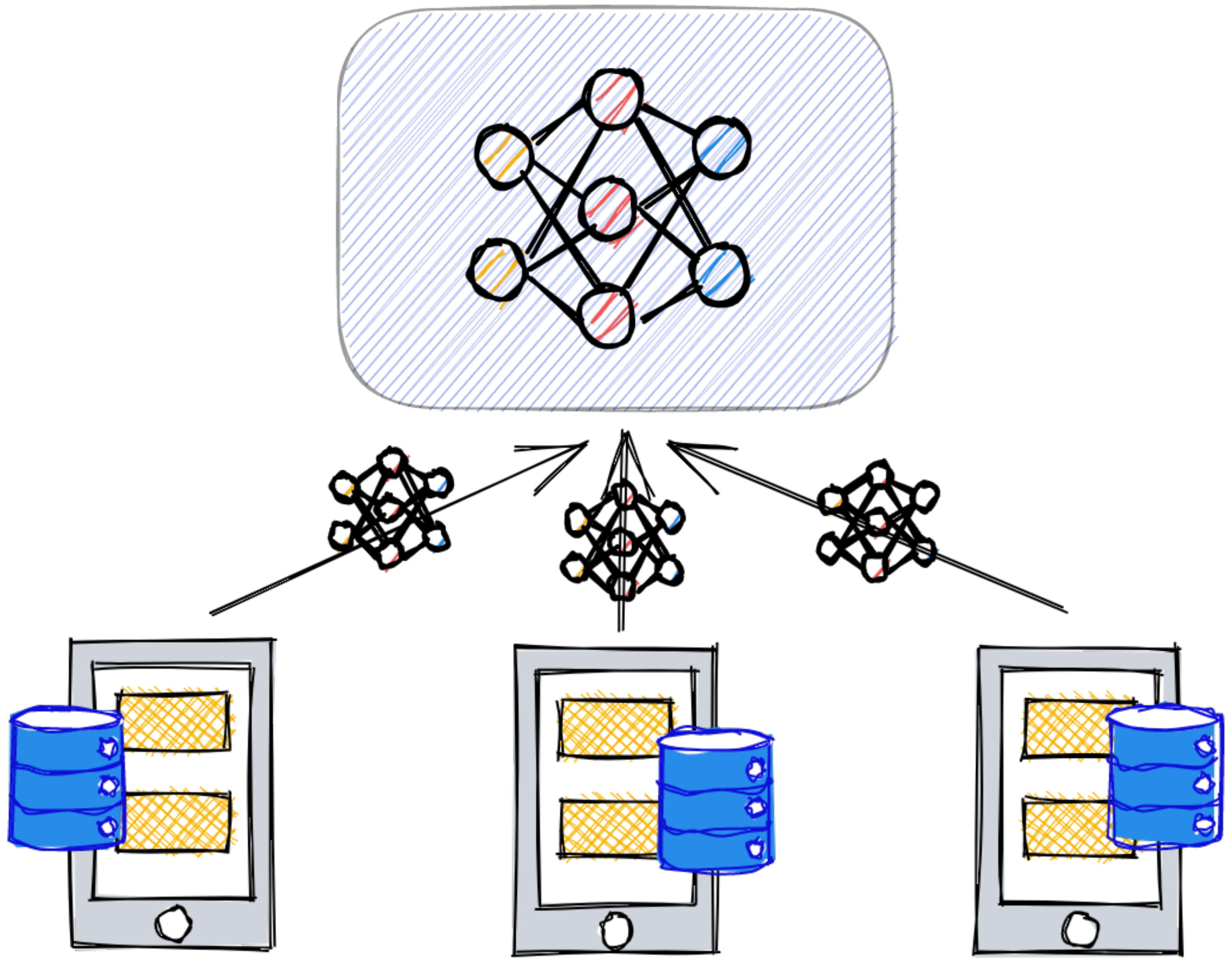
Federated Learning

for $t \in \{1, \ldots, T\}$ **do**:

(3) Each client transmits

$$\mathbf{w}_{t,E}^{k} = \mathbf{w}_t - \eta_t \underbrace{\sum_{j=0}^{E-1} \nabla F_k(\mathbf{w}_{t,j}^{k}, \mathcal{B}_{t,j}^{k})}_{\Delta_t^k}$$

Send model OR **Send pseudo-gradient**



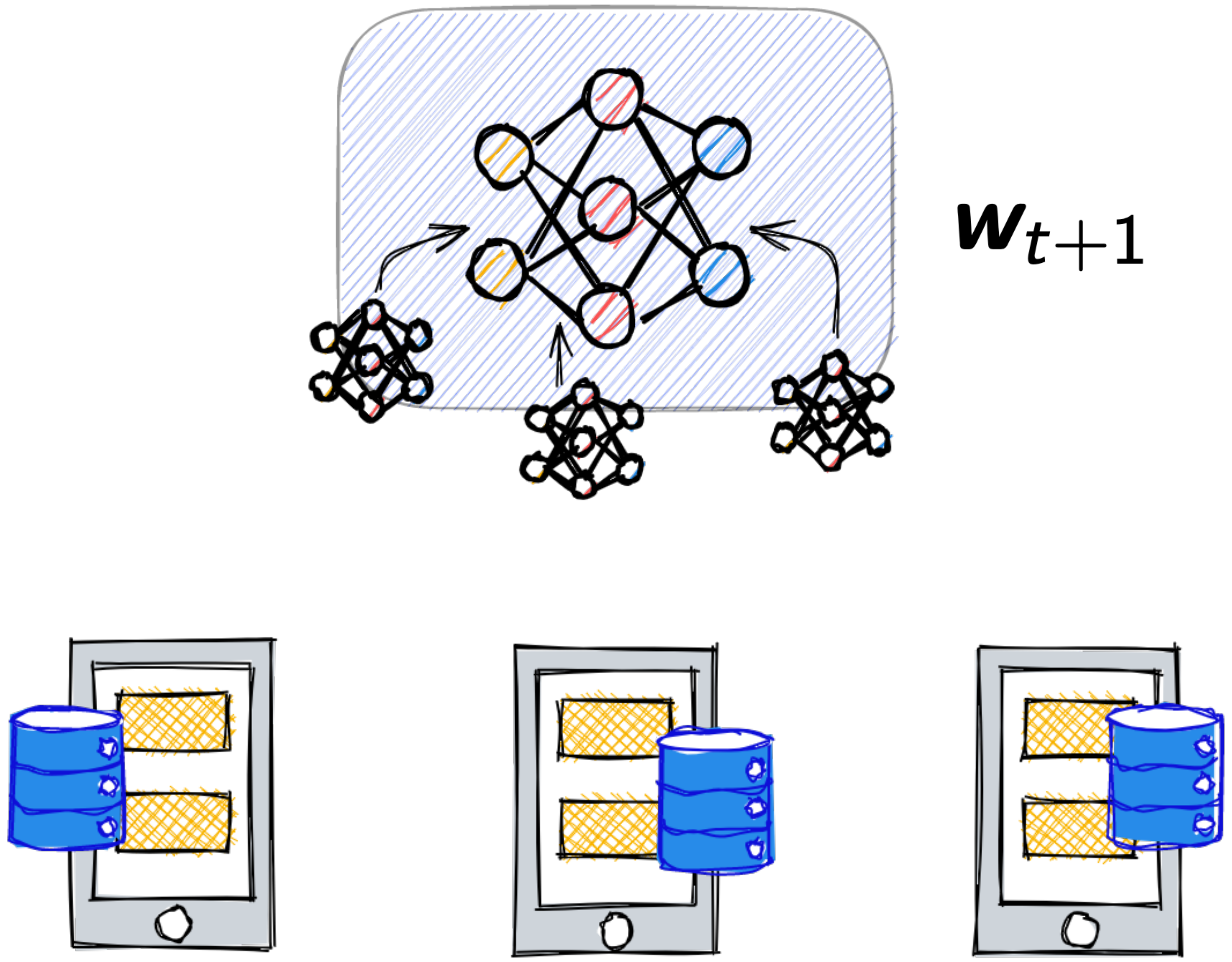
Federated Learning

for $t \in \{1, \ldots, T\}$ **do**:

(4) Server aggregates

$$\boldsymbol{w}_{t+1}^{\text{DMA}} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{w}_{t,E}^{k} \qquad \boldsymbol{w}_{t+1}^{\text{PGA}} = \boldsymbol{w}_t + \frac{1}{N} \sum_{k=1}^{N} \Delta_t^k$$

Direct Model Aggregation (DMA) OR **Pseudo-Gradient Aggregation (PGA)**



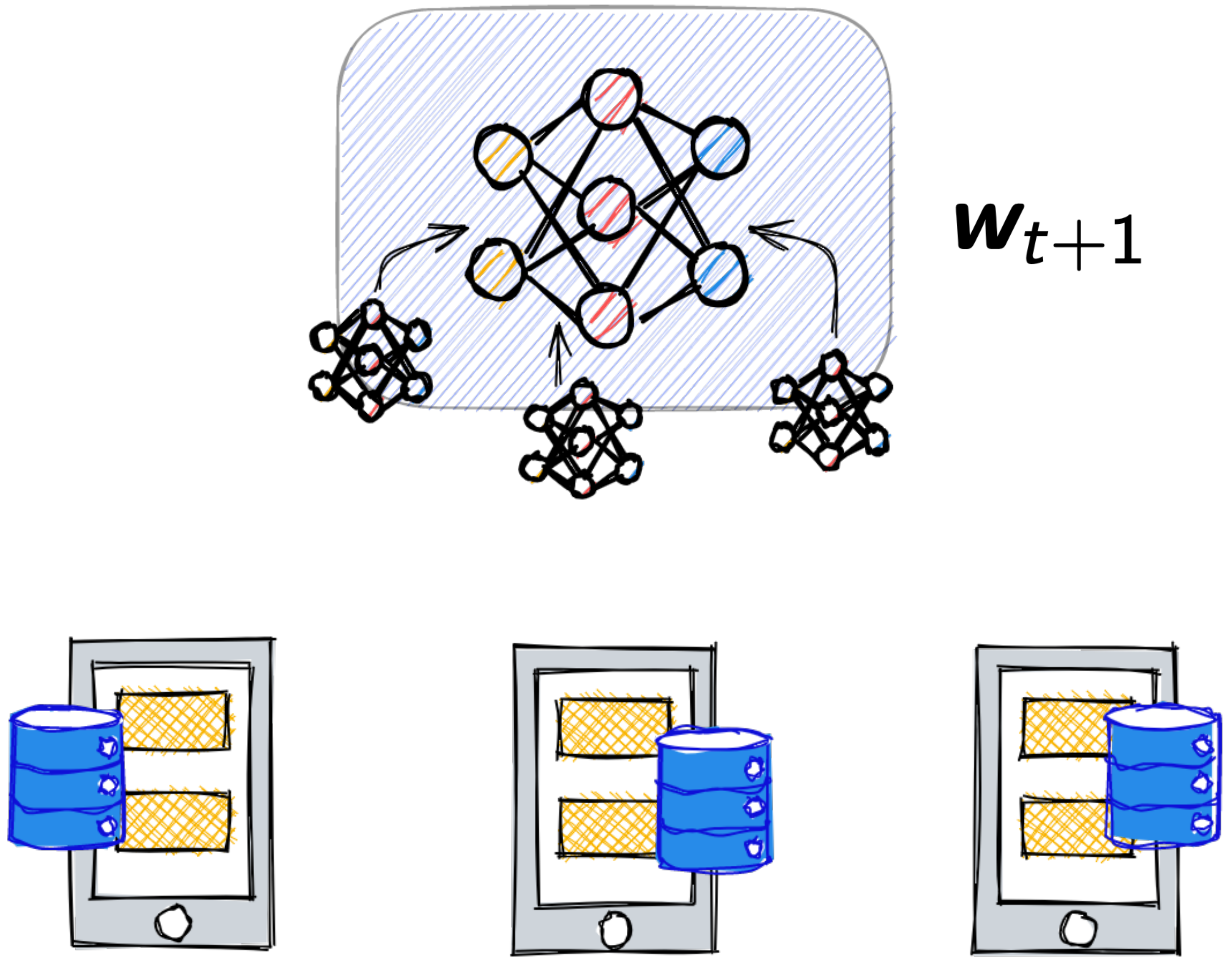
Federated Learning

for $t \in \{1, \dots, T\}$ **do**:

(4) Server aggregates

$$\mathbf{w}_{t+1}^{\text{DMA}} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{w}_{t,E}^{k} \qquad \mathbf{w}_{t+1}^{\text{PGA}} = \mathbf{w}_t + \frac{1}{N}\sum_{k=1}^{N} \Delta_t^{k}$$

Direct Model Aggregation (DMA) OR Pseudo-Gradient Aggregation (PGA)

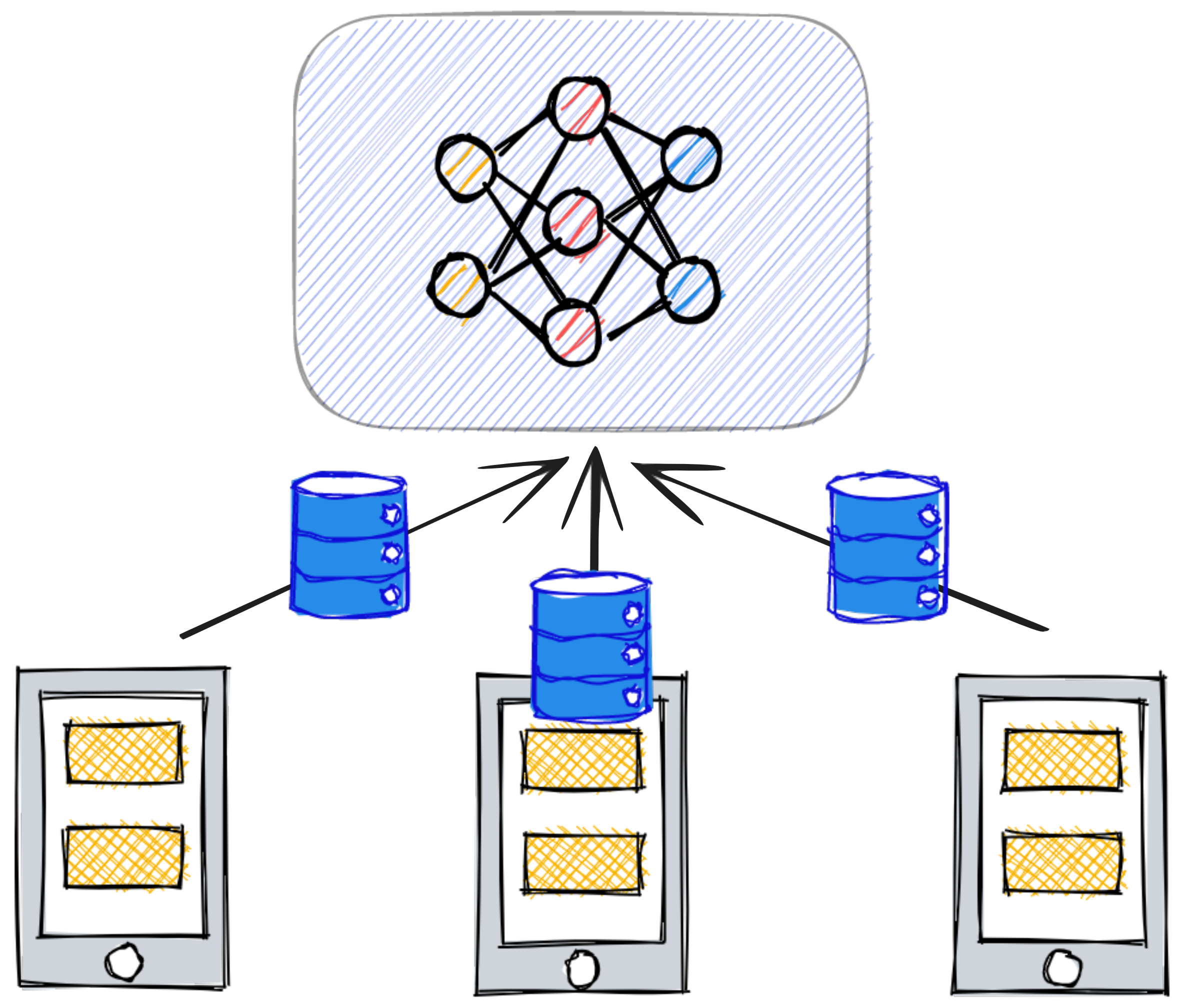


$\mathbf{w}_{t+1}$

(In lossless scenarios: DMA = PGA)

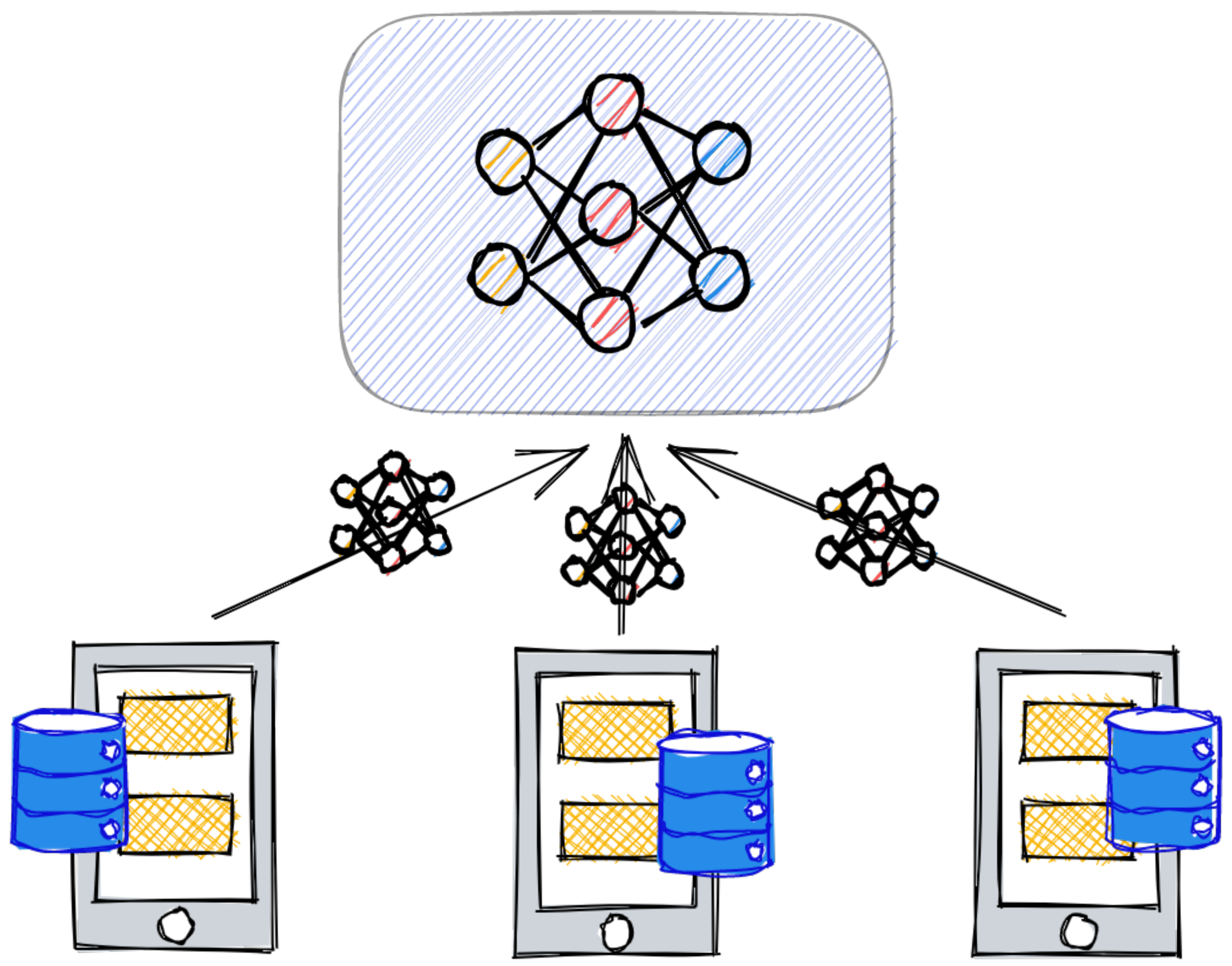
Centralized vs Federated

Centralized



Share data

Federated



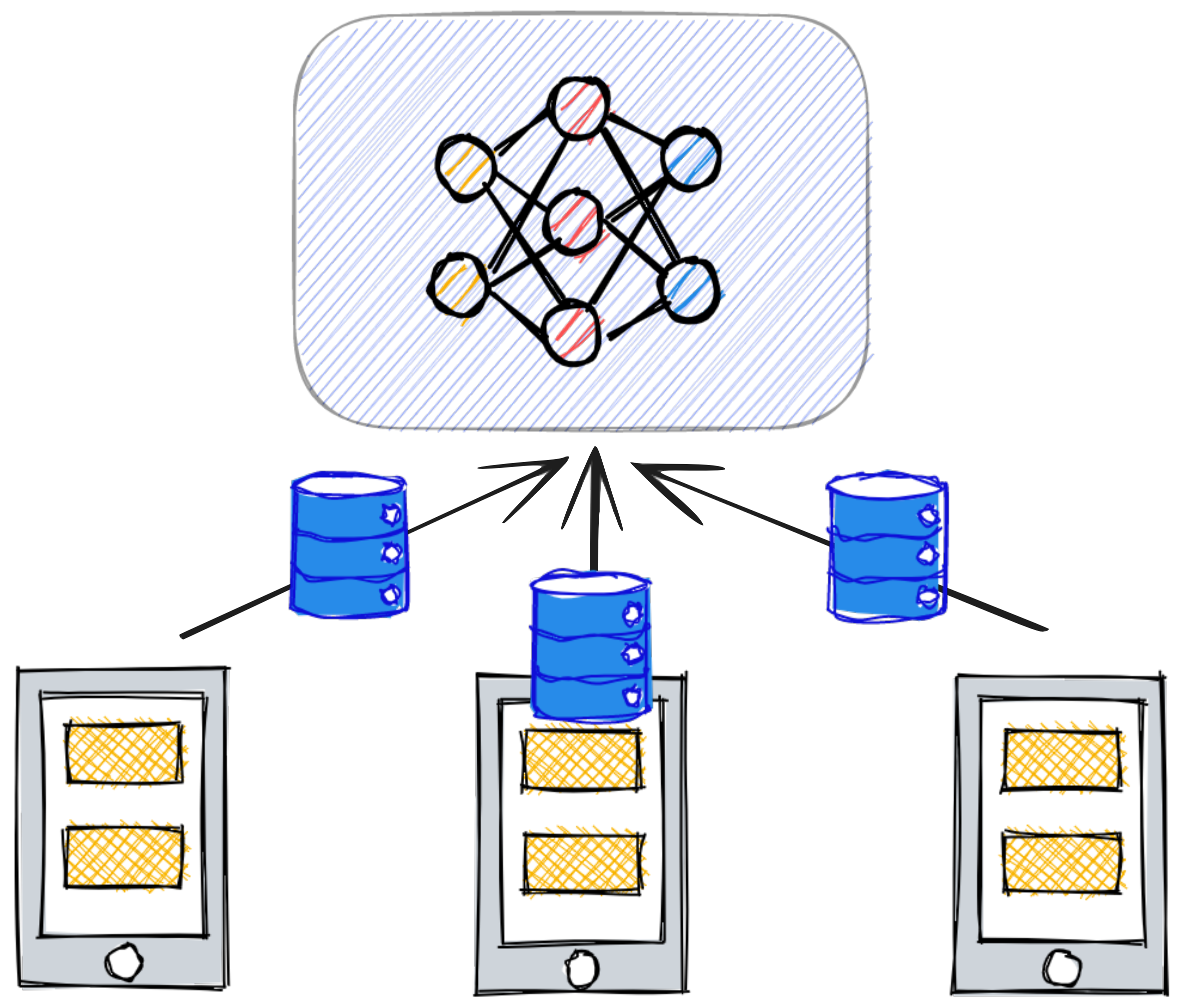
Share models / gradients

Pros:

1. Communication
2. Privacy

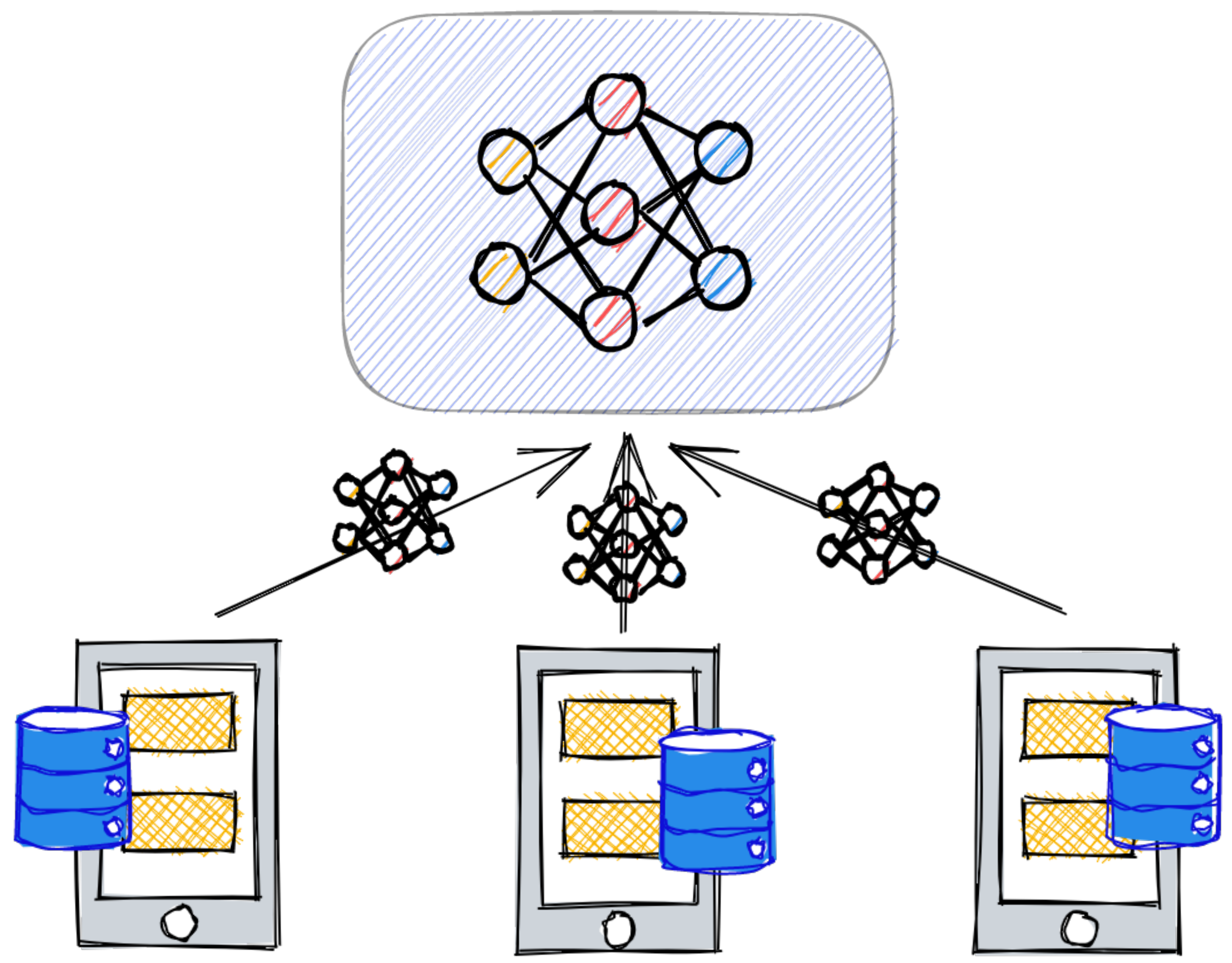
Centralized vs Federated

Centralized



Share data

Federated



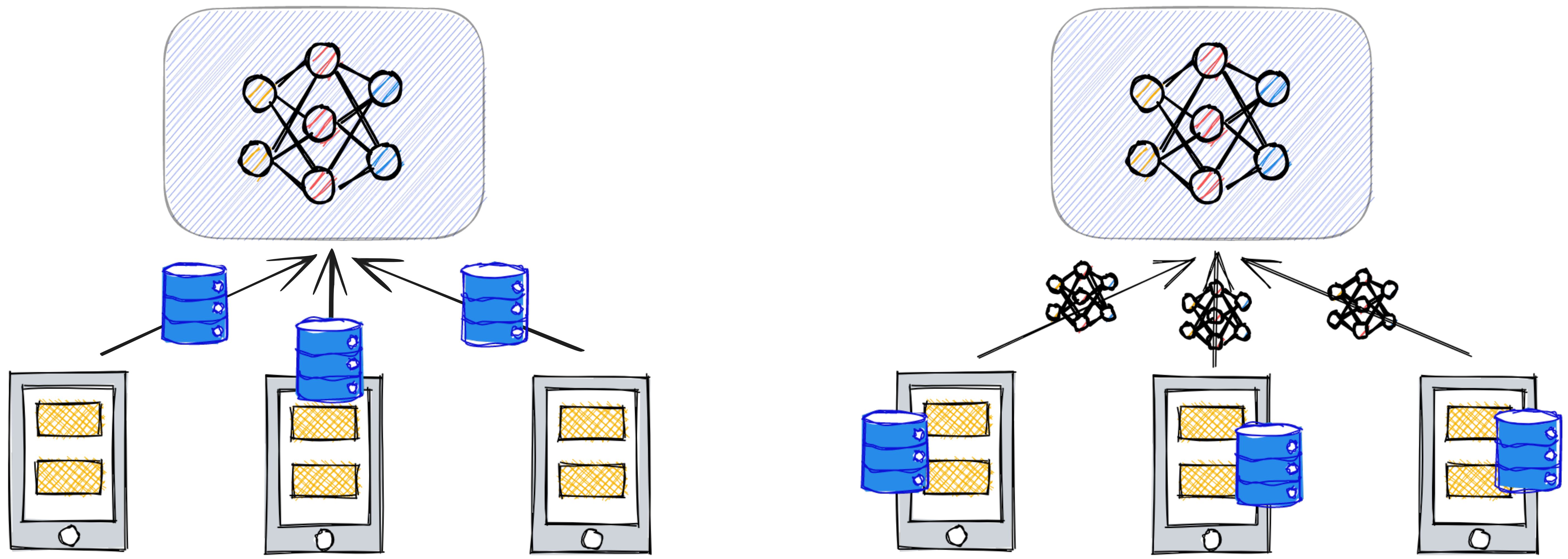
Share models / gradients

Pros:

1. Communication
2. Privacy

Centralized vs Federated

Centralized | **Federated**



Share data | Share models / gradients

Pros:

1. Communication
2. Privacy

- **Common assumption:** clients are always available or uniform participation

Lossy Communication Channels

Previous works: loss mitigation

- Automatic Repeat Request (ARQ)
- Forward Error Correction (FEQ)

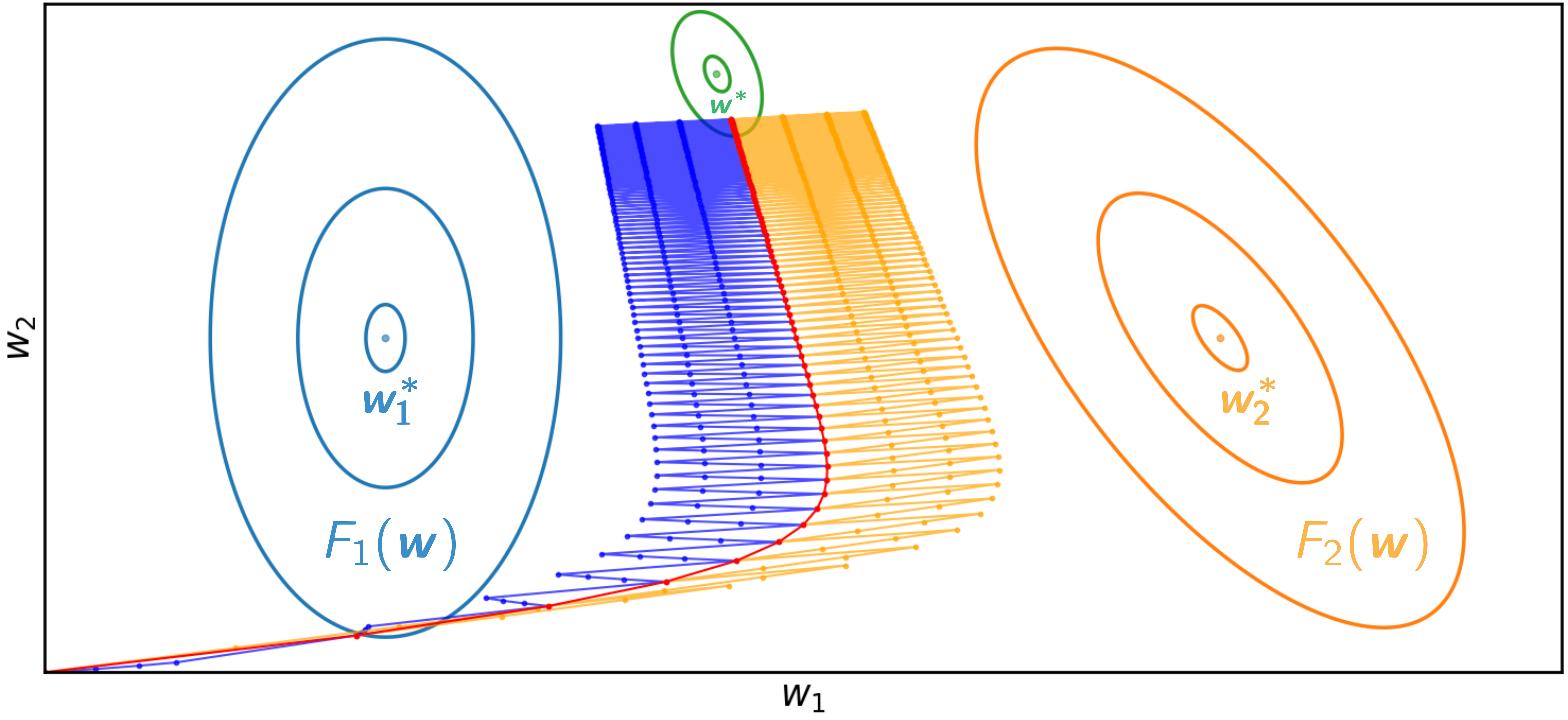
Our motivations

- Inevitable packet losses (e.g., retransmission failure)
- Larger training time and resource costs
- Robustness of gradient methods against limited errors

Can FL algorithms achieve optimal convergence despite packet losses?

Lossless Scenario

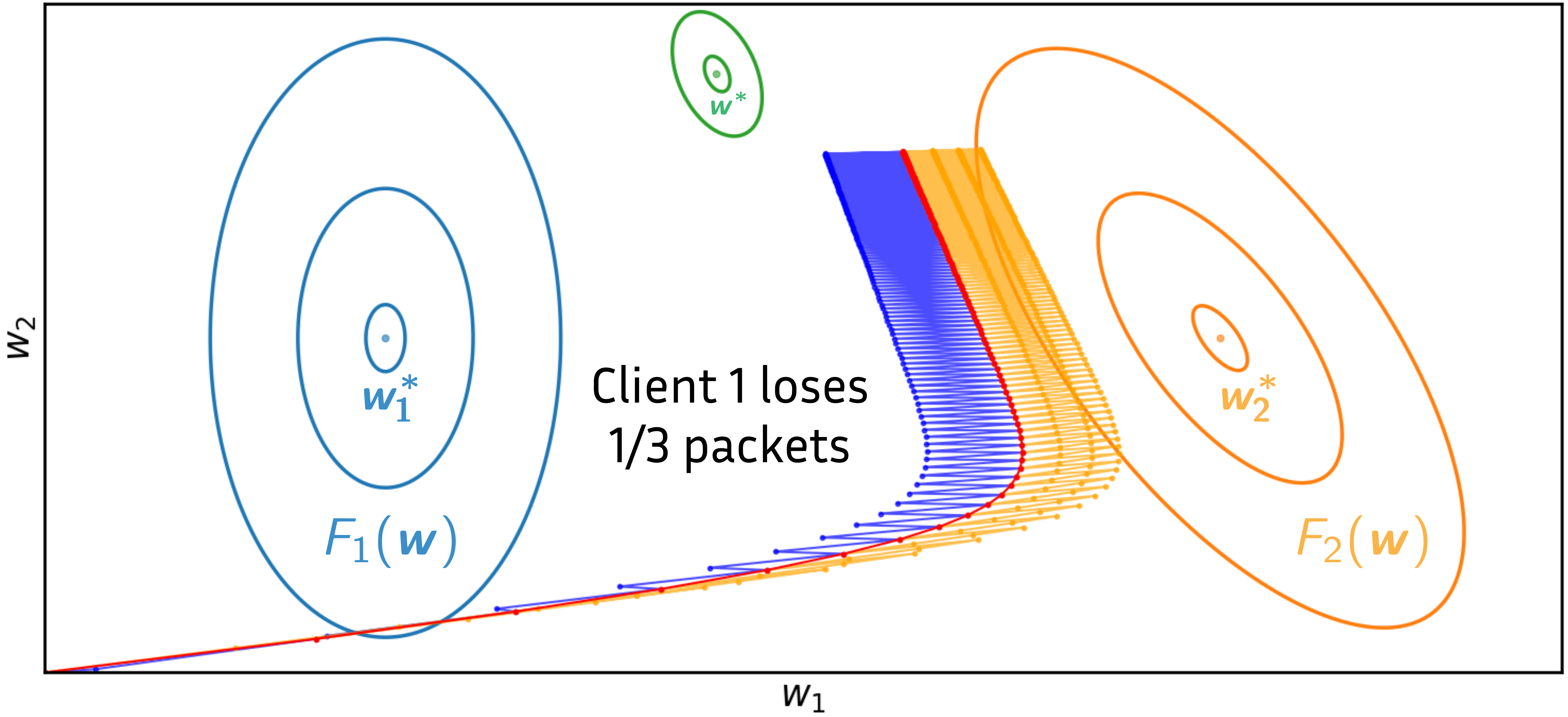
$$F(\boldsymbol{w}) = \frac{1}{2}F_1(\boldsymbol{w}) + \frac{1}{2}F_2(\boldsymbol{w})$$



Convergence to the optimal model ☺

Lossy Scenario

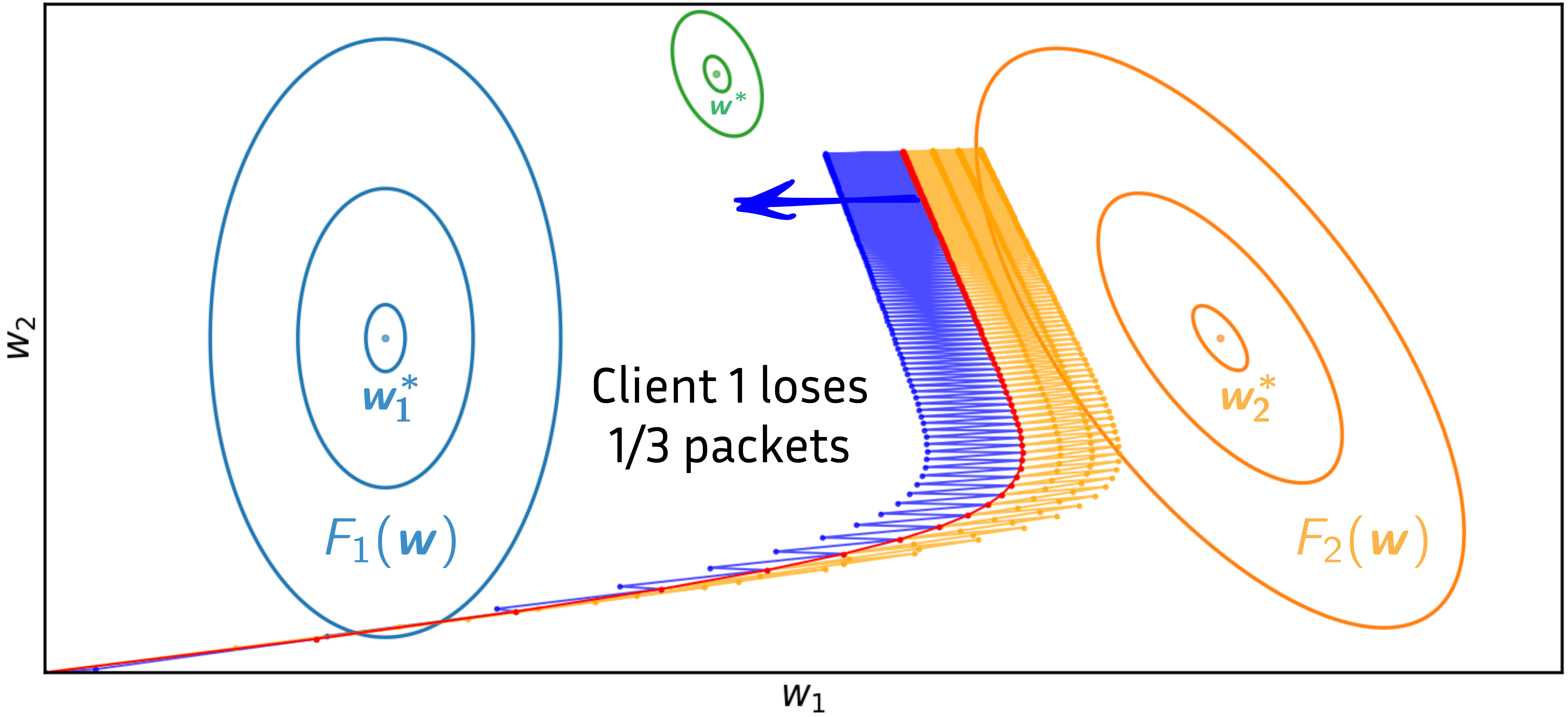
$$F(\boldsymbol{w}) = \frac{1}{2}F_1(\boldsymbol{w}) + \frac{1}{2}F_2(\boldsymbol{w})$$



Packet losses harm convergence ☹

Lossy Scenario

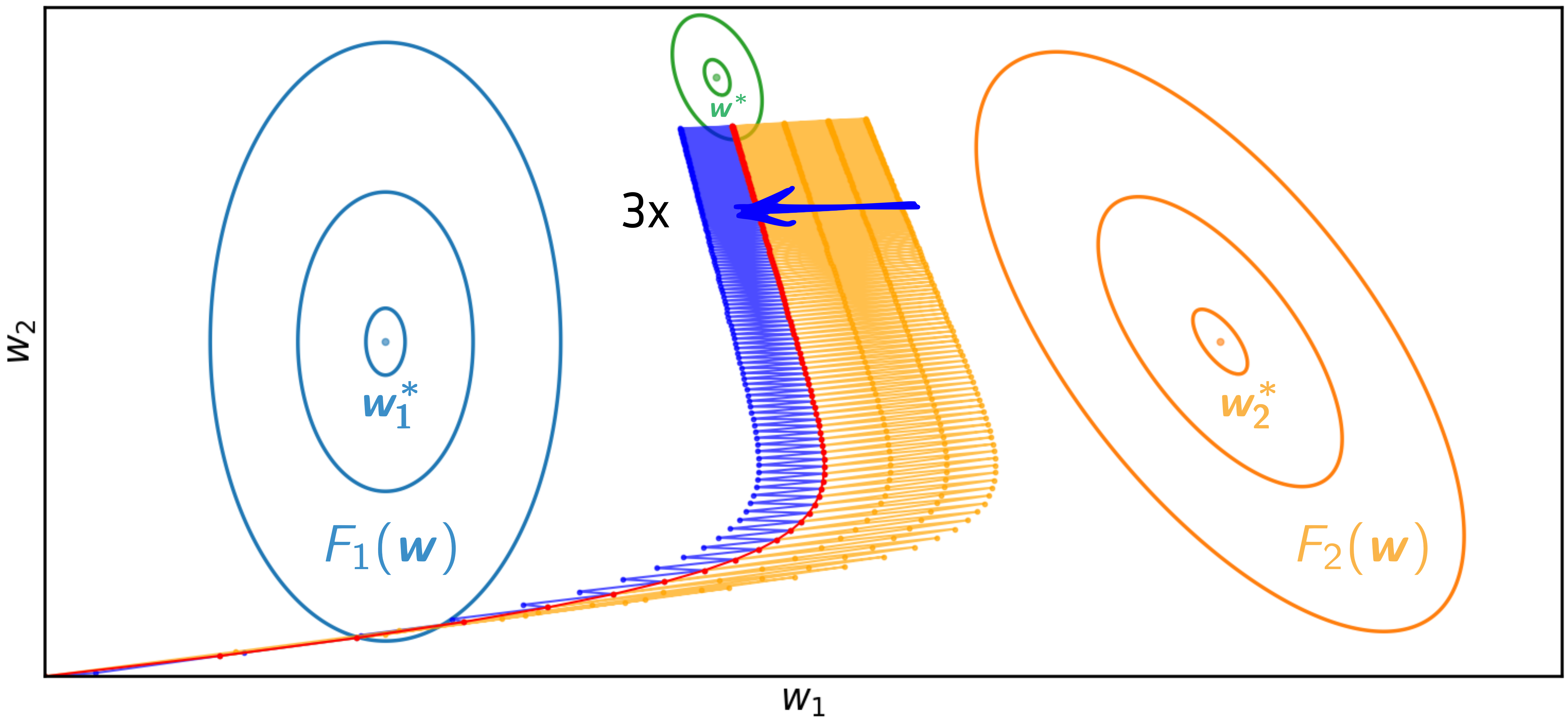
$F(\boldsymbol{w}) = \frac{1}{2}F_1(\boldsymbol{w}) + \frac{1}{2}F_2(\boldsymbol{w})$



Packet losses harm convergence ☹

Lossy Scenario

$F(\boldsymbol{w}) = \frac{1}{2}F_1(\boldsymbol{w}) + \frac{1}{2}F_2(\boldsymbol{w})$



Yes, if 1) Aggregate Pseudo-Gradients
2) Compensate for Packet Losses

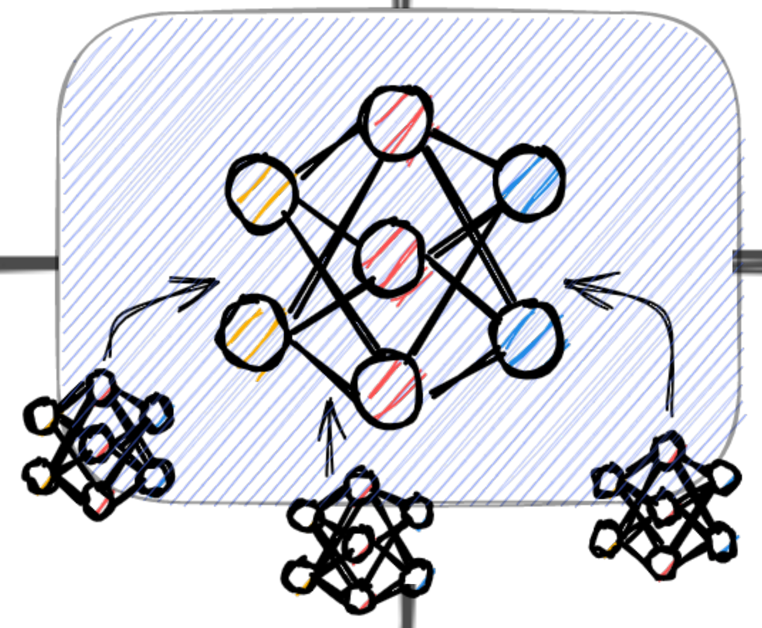
Aggregation for lossy channels

Direct Model Aggregation (DMA)

$$\boldsymbol{w}_{t+1}^{\text{DMA-PL}} = \frac{1}{|\mathcal{P}_t|} \sum_{k \in \mathcal{P}_t} \boldsymbol{w}_{t,E}^{k}$$

Pseudo-Gradient Aggregation (PGA)

$$\boldsymbol{w}_{t+1}^{\text{PGA-PL}} = \boldsymbol{w}_t + \frac{1}{|\mathcal{P}_t|} \sum_{k \in \mathcal{P}_t} \Delta_t^{k}$$



Unbiased DMA

$$\boldsymbol{w}_{t+1}^{\text{UDMA-PL}} = \frac{1}{N} \sum_{k \in \mathcal{P}_t} \frac{\boldsymbol{w}_{t,E}^{k}}{1 - p_k}$$

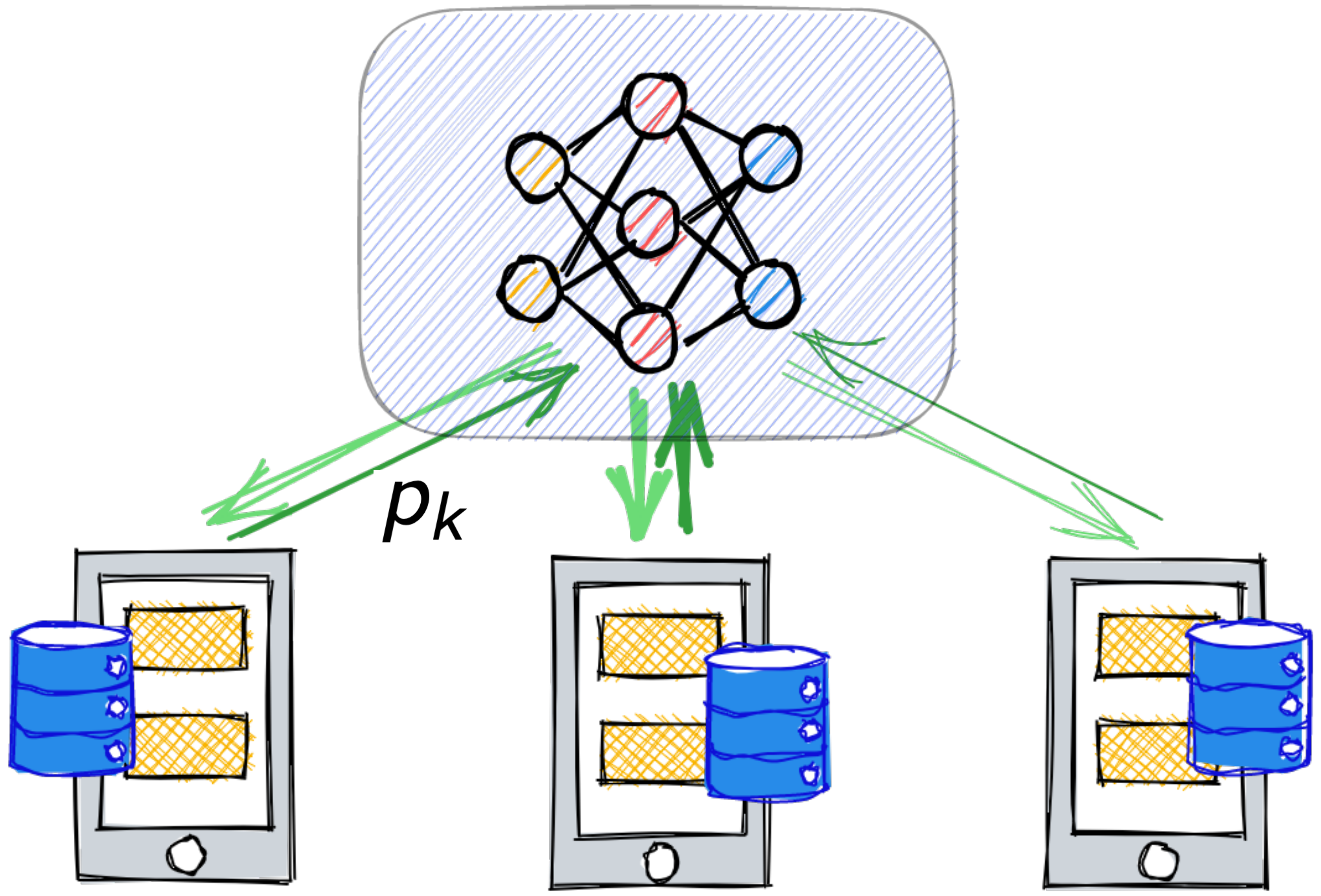
Unbiased PGA (Ours)

$$\boldsymbol{w}_{t+1}^{\text{UPGA-PL}} = \boldsymbol{w}_t + \frac{1}{N} \sum_{k \in \mathcal{P}_t} \frac{\Delta_t^{k}}{1 - p_k}$$

1) Aggregate Pseudo-Gradients
2) Compensate for Packet Losses

Assumptions to model lossy channels

- Loss probabilities p_k differ among clients
- Independent losses among clients
- For each client, IID losses over time
- Asymmetric channels (downlink/uplink)
- If ARQ or FEQ, p_k is the residual probability



Convergence Analysis

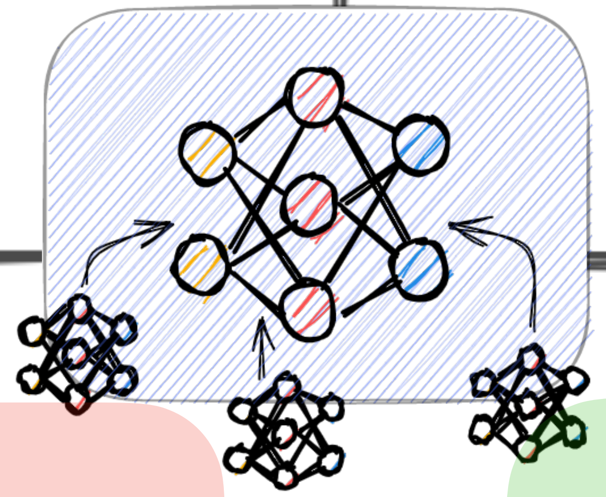
Direct Model Aggregation

$$\boldsymbol{w}_{t+1}^{\text{DMA-PL}} = \frac{1}{|\mathcal{P}_t|}\sum_{k\in\mathcal{P}_t} \boldsymbol{w}_{t,E}^{k}$$

Unbiased Pseudo-Gradient Aggregation

$$\boldsymbol{w}_{t+1}^{\text{UPGA-PL}} = \boldsymbol{w}_t + \frac{1}{N}\sum_{k\in\mathcal{P}_t} \frac{\Delta_t^k}{1-p_k}$$

(OURS)



$$\mathbb{E}[F(\boldsymbol{w}_{t+1}^{\text{DMA-PL}})] - F^* \leq$$

$$\underbrace{A^t\left(F(\boldsymbol{w}_1) - F^*\right)}_{\text{vanishing term for small statistical heterogeneity}} + \boxed{\underbrace{\frac{2\zeta_1}{L}\frac{1}{N}\sum_{k=1}^{N} p_k \frac{1-A^t}{1-A}}_{\text{non-vanishing error due to stat. het. and packet loss}}}$$

A joint learning and communications framework for federated learning over wireless networks. Chen, Mingzhe, et al. IEEE Transactions on Wireless Communications, 2021.

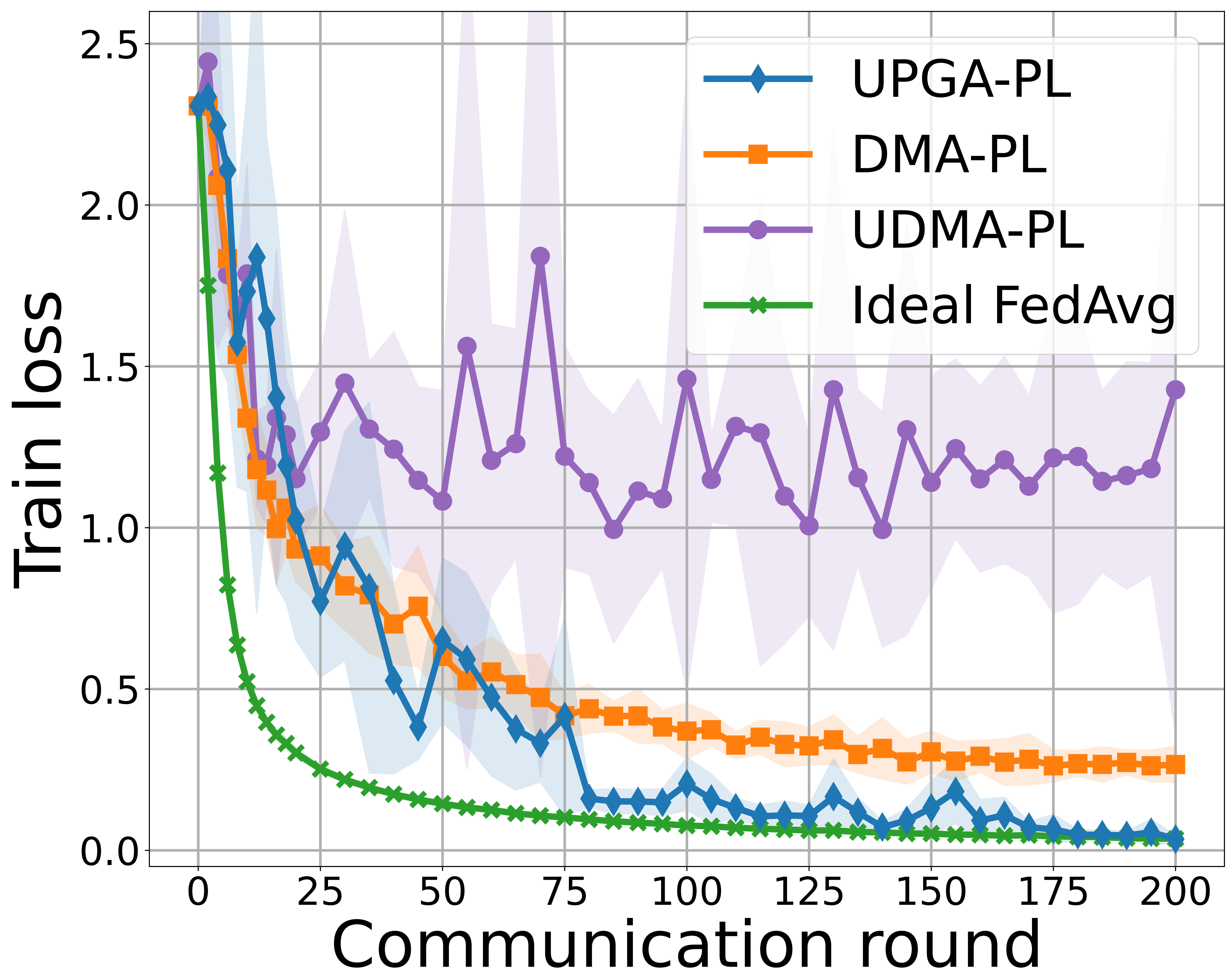
$$\mathbb{E}[F(\boldsymbol{w}_{t+1}^{\text{UPGA-PL}})] - F^* \leq \boxed{\underbrace{\frac{\kappa}{8\kappa+t}\left(\frac{2EC}{\mu} + 4L\left\|\boldsymbol{w}_1 - \boldsymbol{w}^*\right\|^2\right)}_{\text{asymptotically vanishing term}}}$$

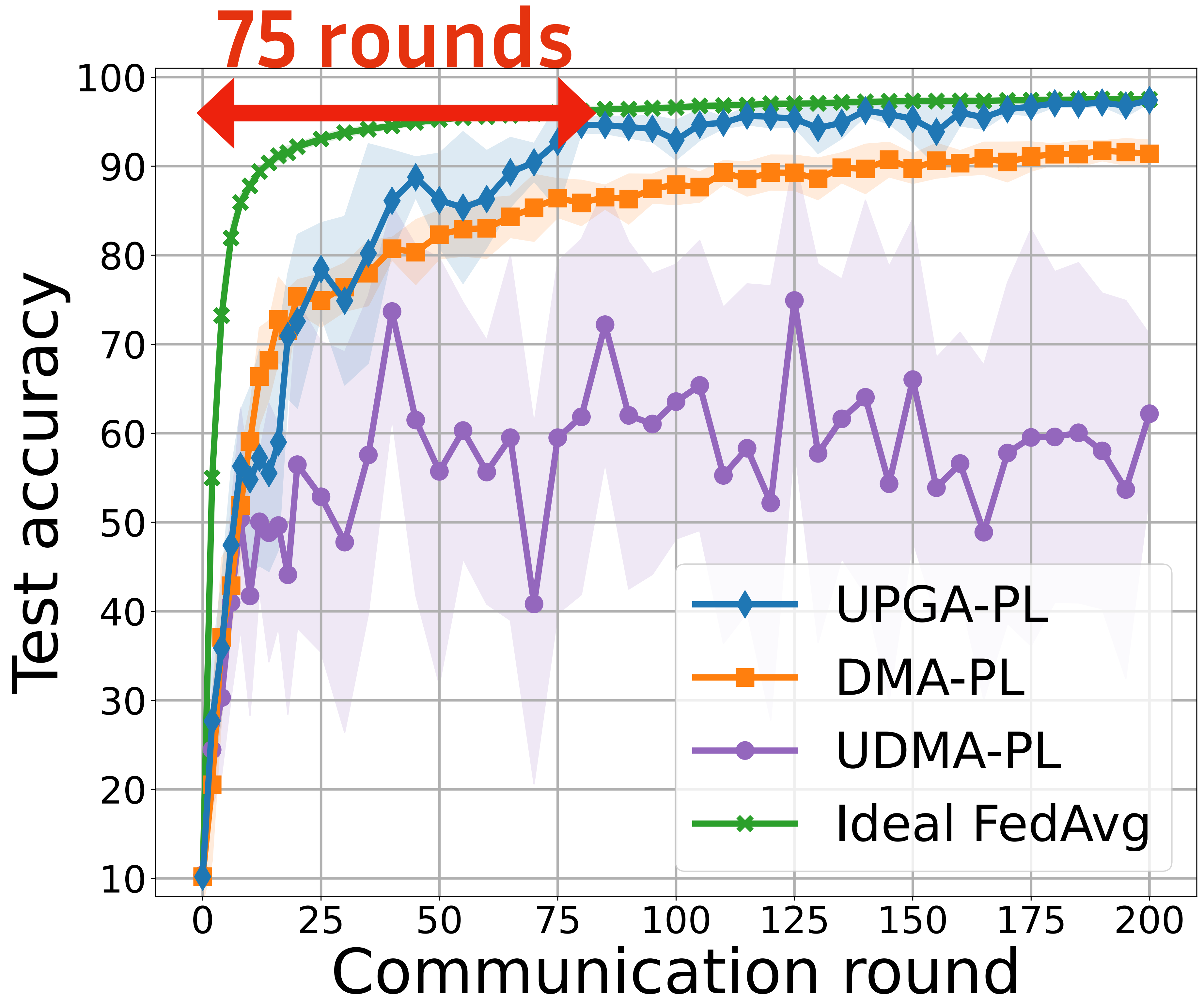
$$C := \frac{1}{N^2}\sum_{k=1}^{N}\sigma_k^2 + 2(E-1)^2G^2 + 6L\Gamma + \boxed{\underbrace{\frac{EG^2}{N^2}\sum_{k=1}^{N}\frac{p_k}{1-p_k}}_{\text{effect of packet loss}}}$$

UPGA-PL converges to the optimal model ☺

Experimental Evaluation

N = 10 clients equally split in two groups, with $p_1 = \frac{1}{10},\ p_2 = \frac{9}{10}$, MNIST dataset, CNN

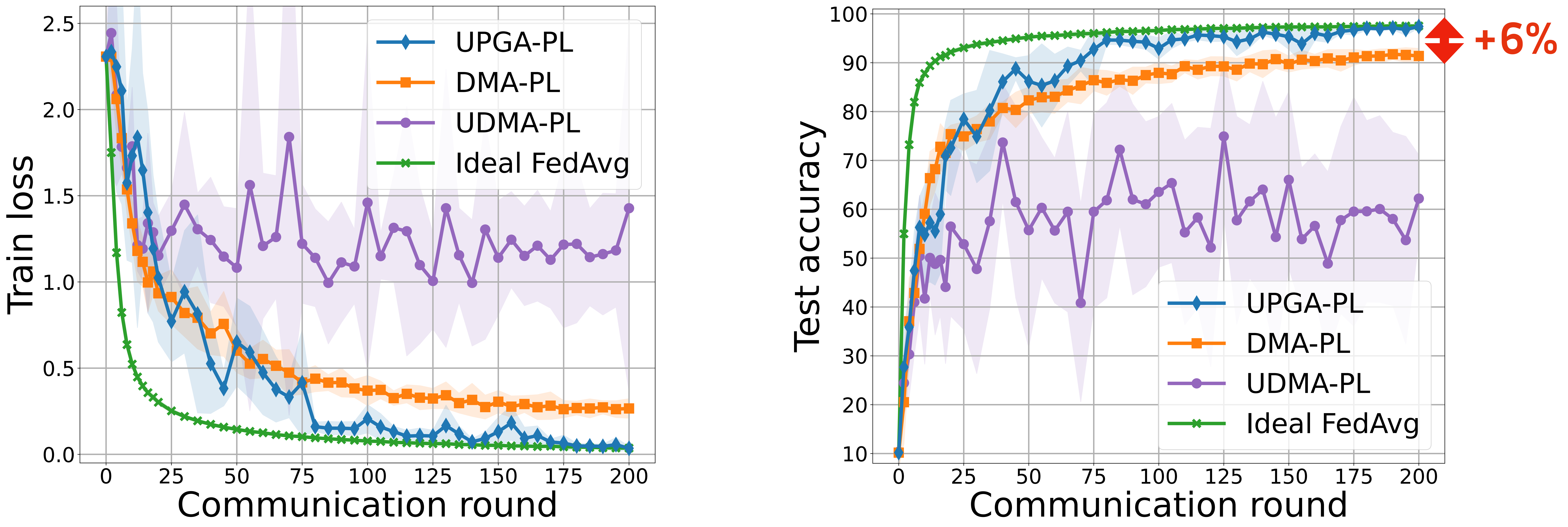




UPGA-PL matches lossless performance in < 100 rounds

Experimental Evaluation

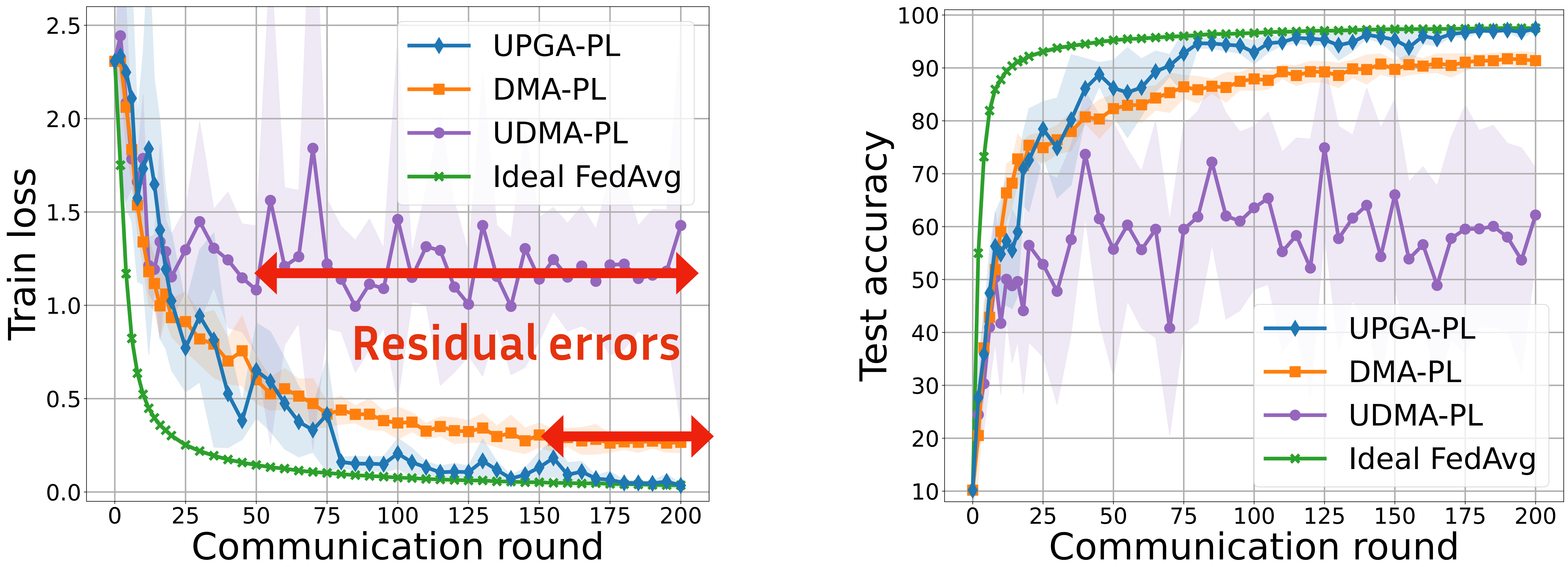
N = 10 clients equally split in two groups, with $p_1 = \frac{1}{10}$, $p_2 = \frac{9}{10}$, MNIST dataset, CNN



UPGA-PL improves MNIST performance by 6% over SOTA

Experimental Evaluation

N = 10 clients equally split in two groups, with $p_1 = \frac{1}{10}$, $p_2 = \frac{9}{10}$, MNIST dataset, CNN



DMA-PL and UDMA-PL exhibit non-vanishing errors

Conclusions

- UPGA-PL has **optimal convergence** under asymmetric lossy channels
- UPGA-PL **outperforms SOTA** by filtering out losses
- UPGA-PL approaches **ideal lossless channels** with slightly **slower convergence**

Web

Code

Thank you for your attention!