An Equivalence Between Data Poisoning and Byzantine Gradient Attacks

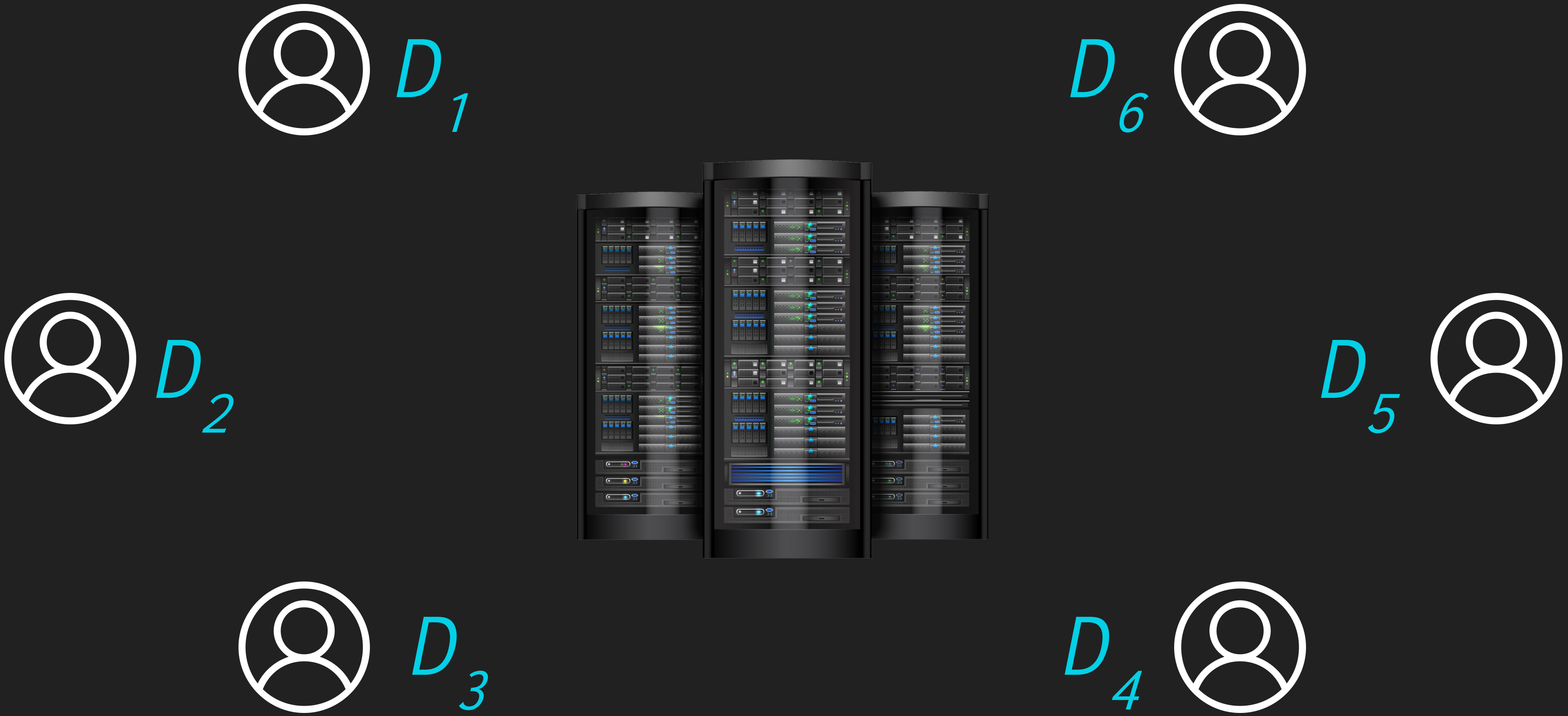
Sadegh Farhadkhani, Rachid Guerraoui,
Lê-Nguyên Hoang and Oscar Villemaud
IC EPFL

EPFL

ICML | 2022
Thirty-ninth International Conference on Machine Learning

Global and Personalized Learning

Global and Personalized Learning



Global and Personalized Learning

θ_1 D_1 D_6 θ_6

θ_2 D_2 ρ D_5 θ_5

θ_3 D_3 D_4 θ_4

Global and Personalized Learning

θ_1 D_1 D_6 θ_6

ρ

θ_2 D_2

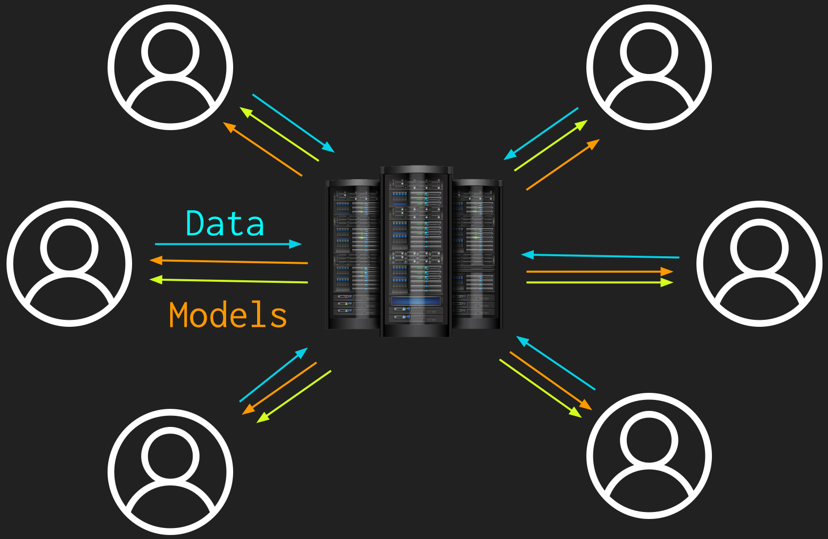
$$\textsc{Loss} = \sum_n \mathcal{L}_n(\theta_n, \mathcal{D}_n) + \sum_n \mathcal{R}(\rho, \theta_n)$$

D_5 θ_5

θ_3 D_3 D_4 θ_4

Two computation models

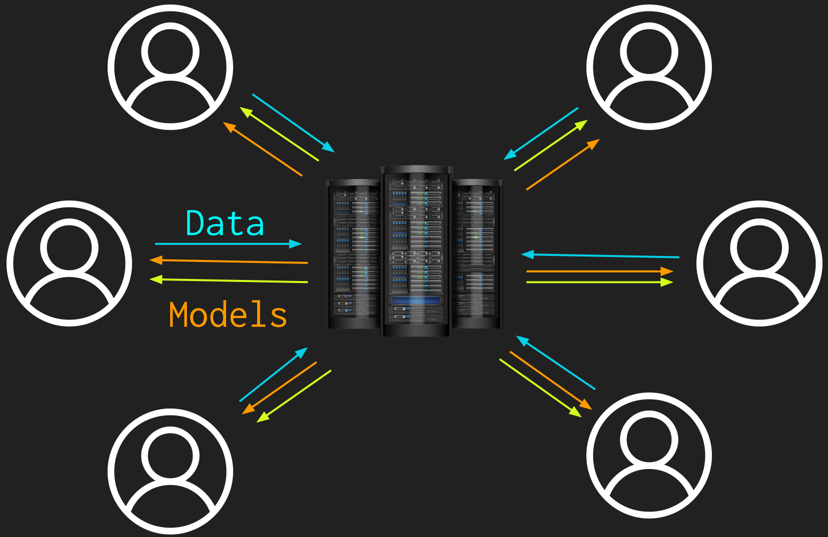
Central computing



Users send data, and receive global and personalized models.

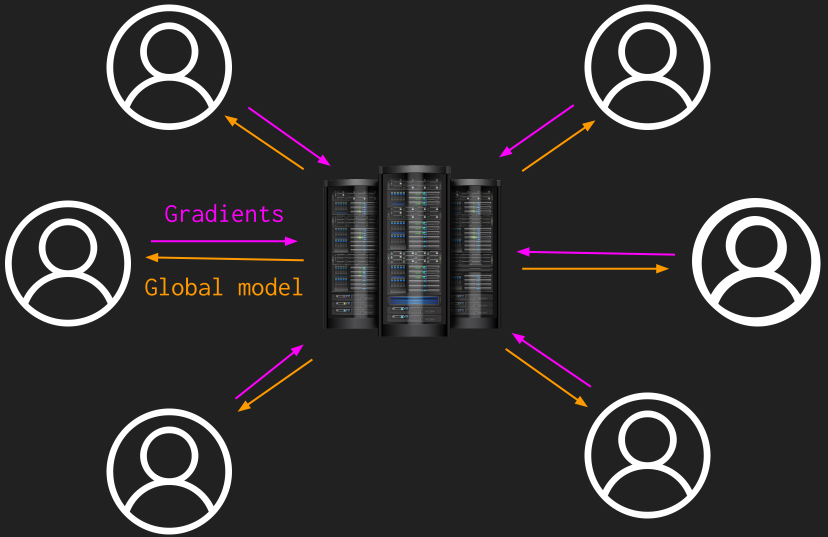
Two computation models

Central computing



Users send data, and receive global and personalized models.

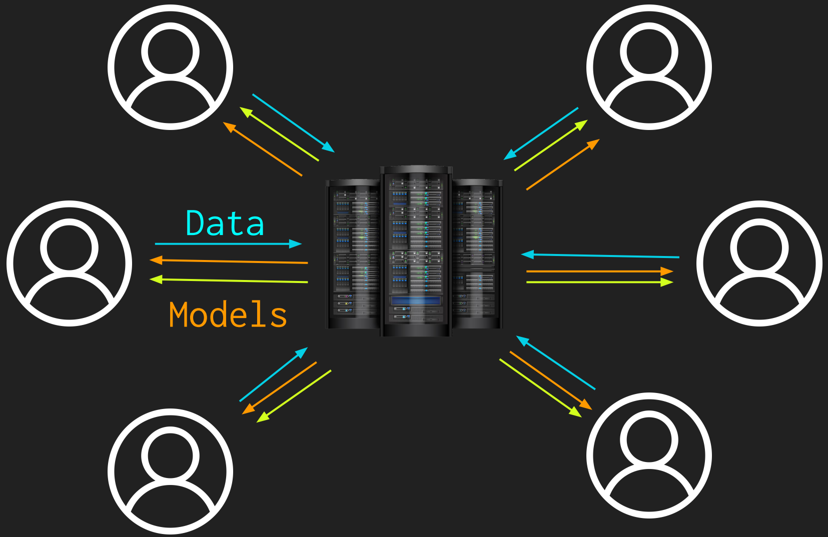
Federated learning



Users send gradients, and receive global models. They compute personalized models themselves.

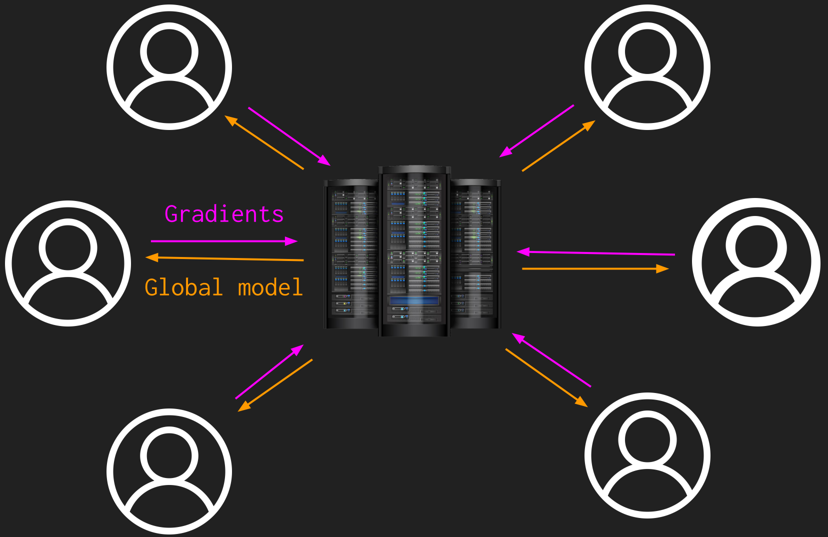
Two computation models

Central computing



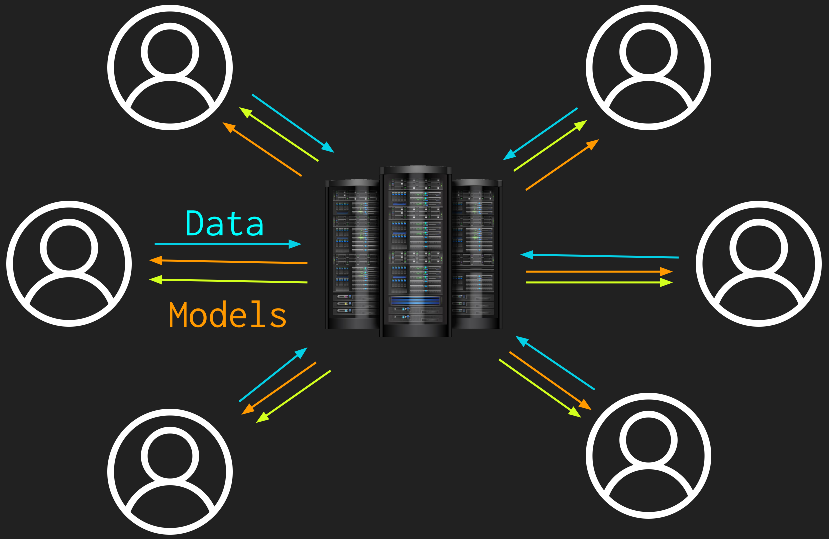
Users may send poisonous data.

Federated learning



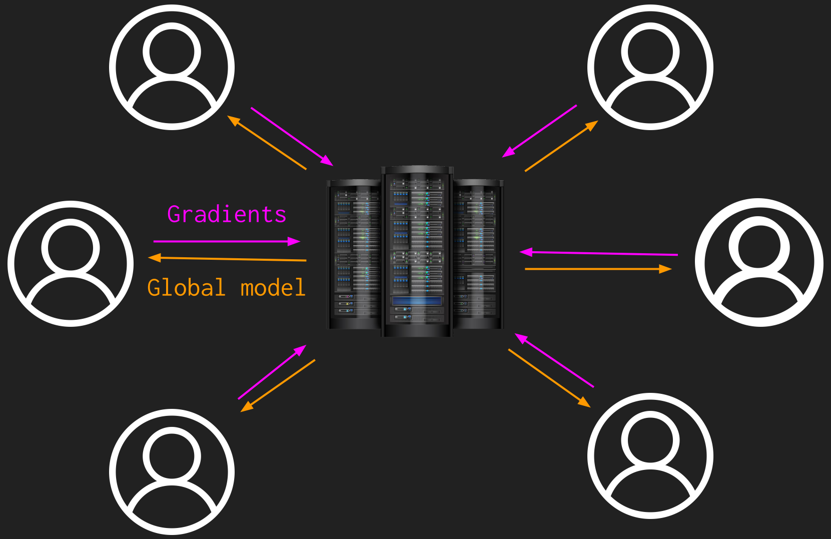
Two computation models

Central computing



Users may send poisonous data.

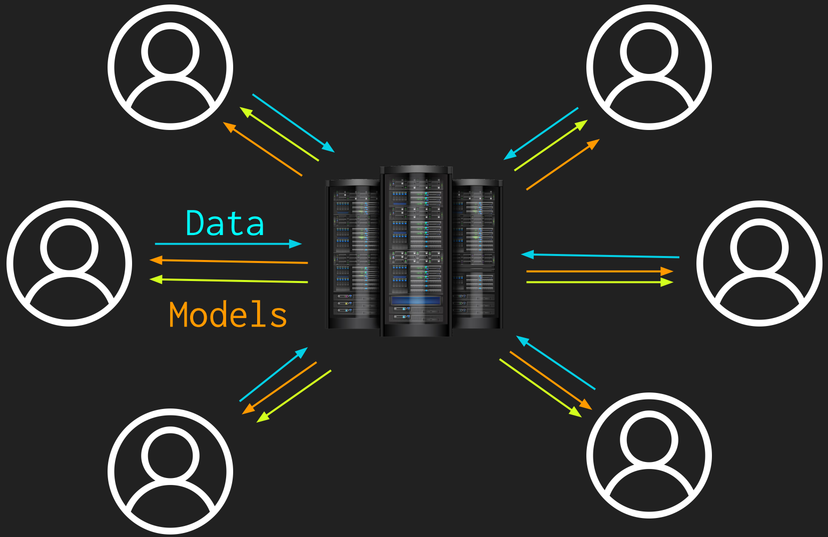
Federated learning



Users may send Byzantine gradients.

Two computation models

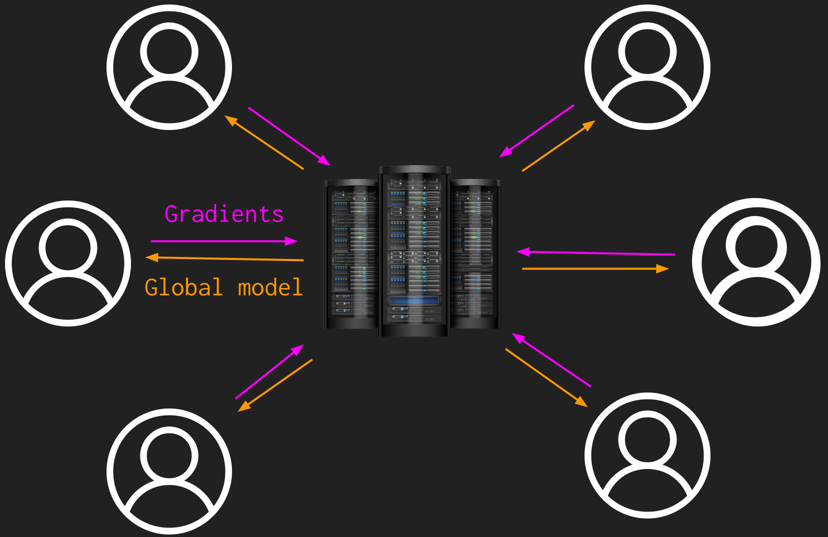
Central computing



Users may send poisonous data.

Very **realistic**. But is it **harmful**?

Federated learning



Users may send Byzantine gradients.

Very **harmful**. But is it **realistic**?

Main theorem

Under realistic and desirable GPL assumptions + convexity, **data poisoning** and **Byzantine gradients** are **equivalent**.

Our results are very practical!

We develop a new targeted gradient attack which successfully relabels all data, and we turn it into effective data poisoning with surprisingly few injected data.

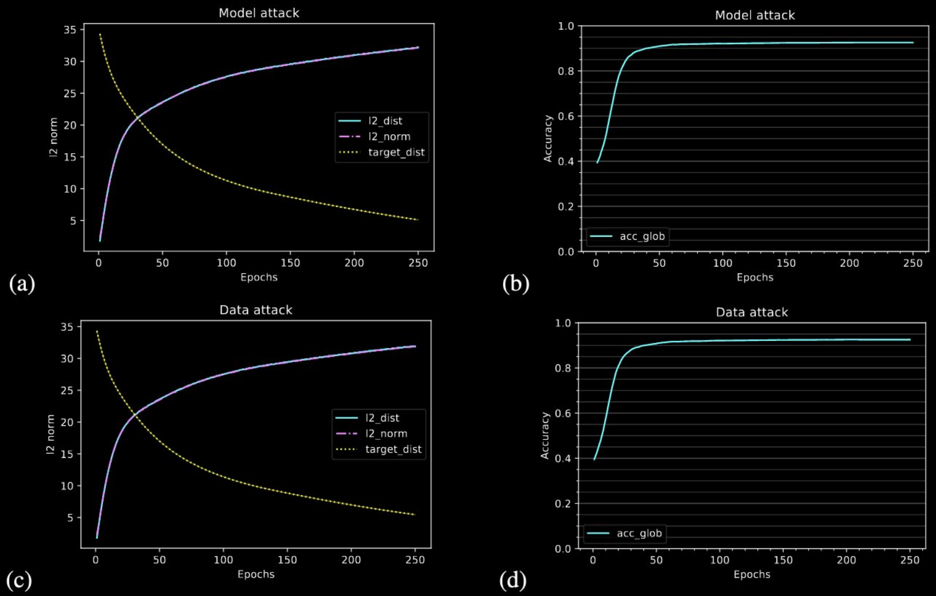


Figure 2. (a) Distance between ρ^t and $\theta_s^\dagger$ (target_dist), under model attack (combining CGA and Proposition 4). (b) Accuracy of ρ^t according to $\theta_s^\dagger$ (which relabels 0 → 1 → 2 → ... → 9 → 0), under model attack (combining CGA and Proposition 4). (c) Distance between the global model ρ^t and the target model $\theta_s^\dagger$ (target_dist), under our data poisoning attack. (d) Accuracy of ρ^t according to $\theta_s^\dagger$ (which relabels 0 → 1 → 2 → ... → 9 → 0), under our data poisoning attack.

Conclusion

Byzantine resilience concerns must **urgently** be **seriously considered**.