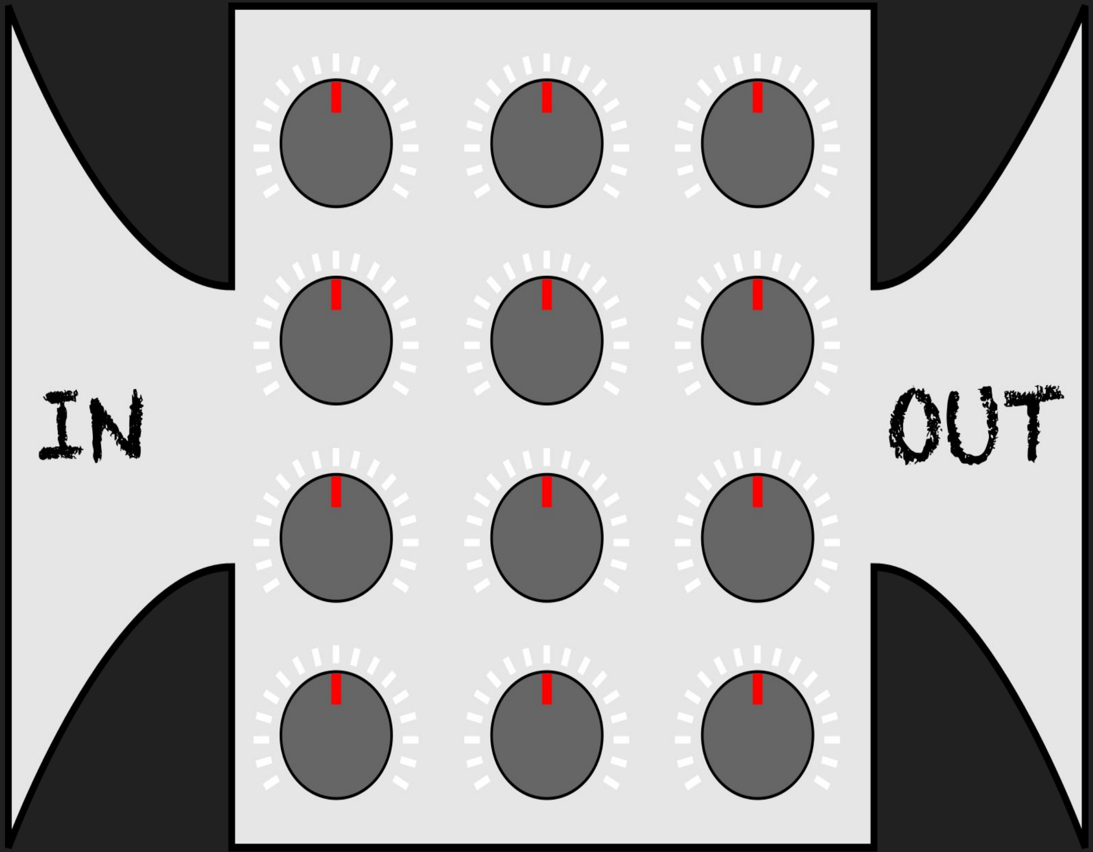
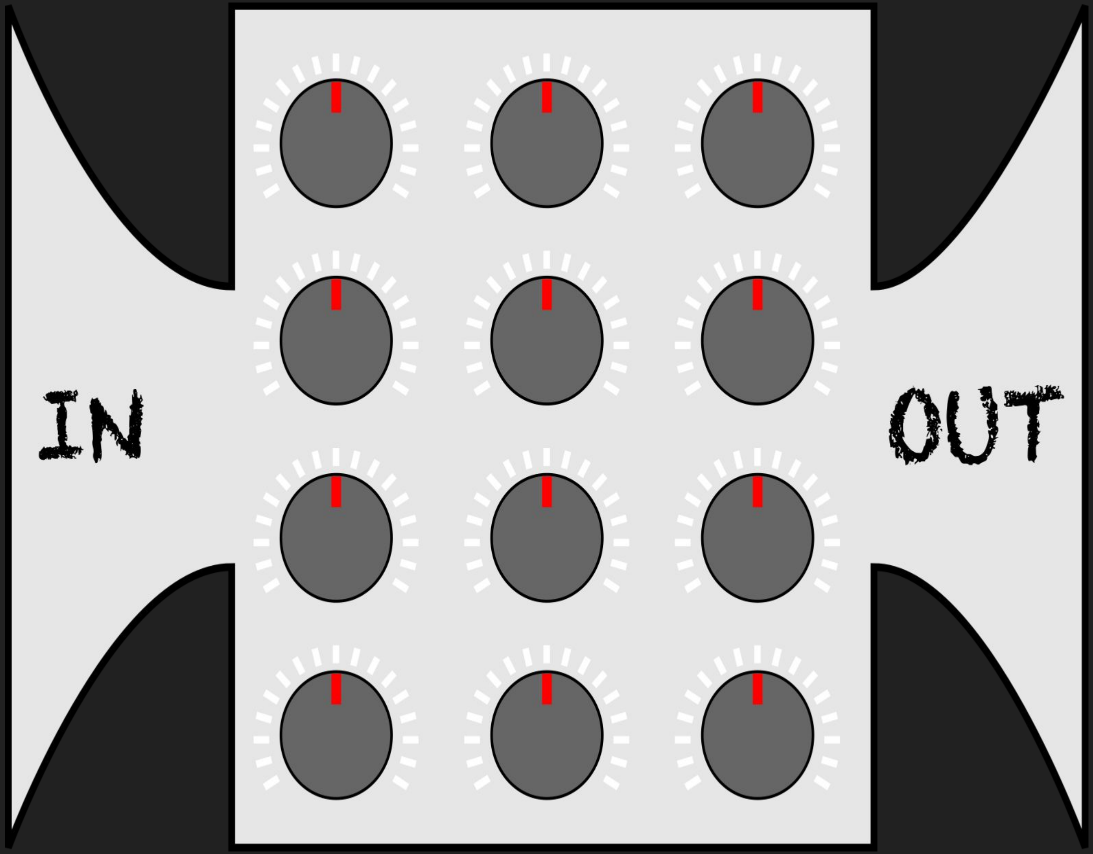
Distributed machine learning

Lê Nguyên Hoang, EPFL
(joint work with DCL)
@le_science4all

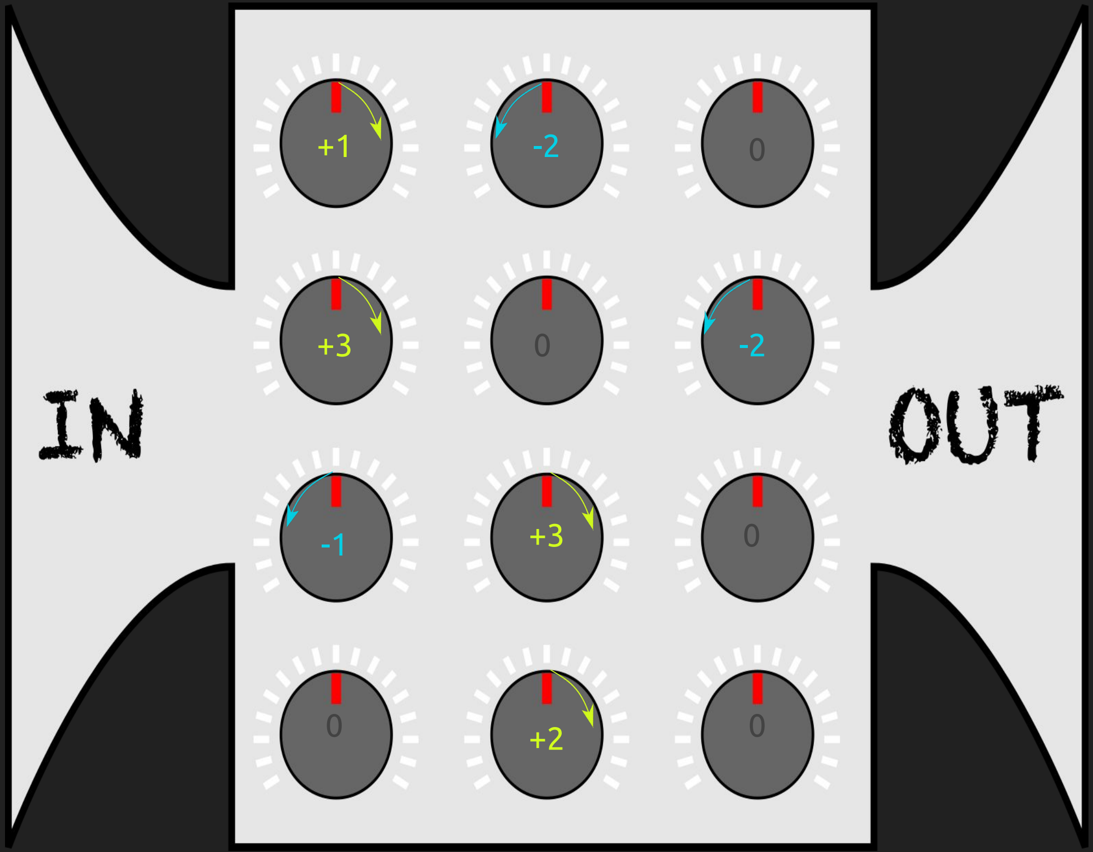
EPFL, Distributed Algorithms


IN
OUT



Cat

Not a cat



Cat

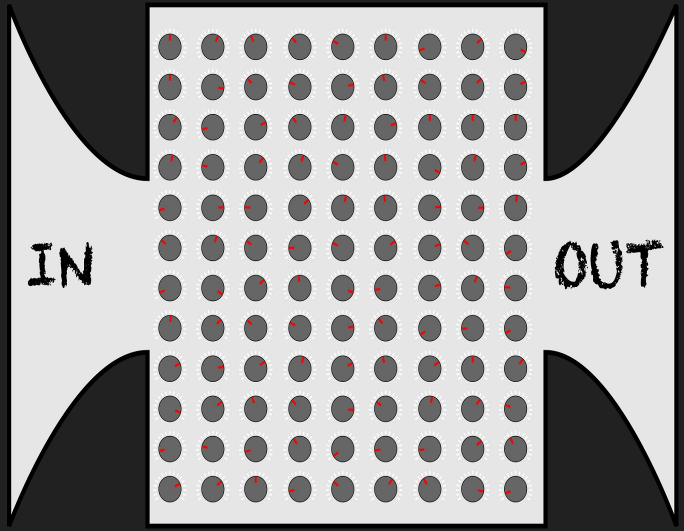
Not a cat

x1 Billion

(if not a lot more)

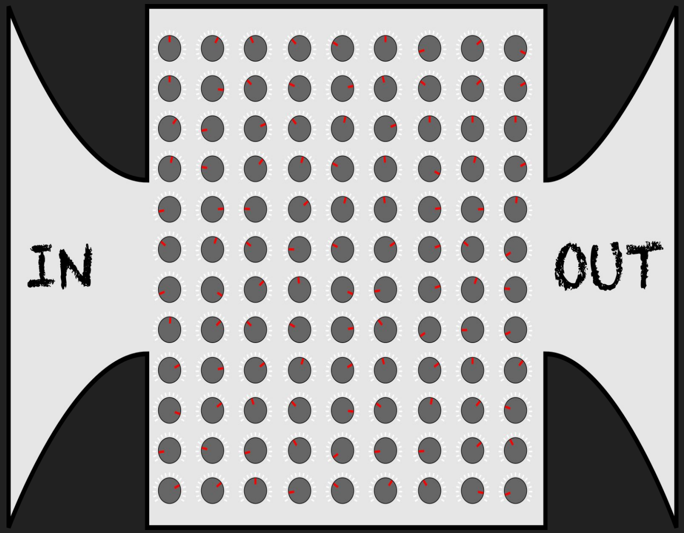
Workers

(compute and send gradient estimates)



Server

(updates and sends models)



Workers

(compute and send gradient estimates)

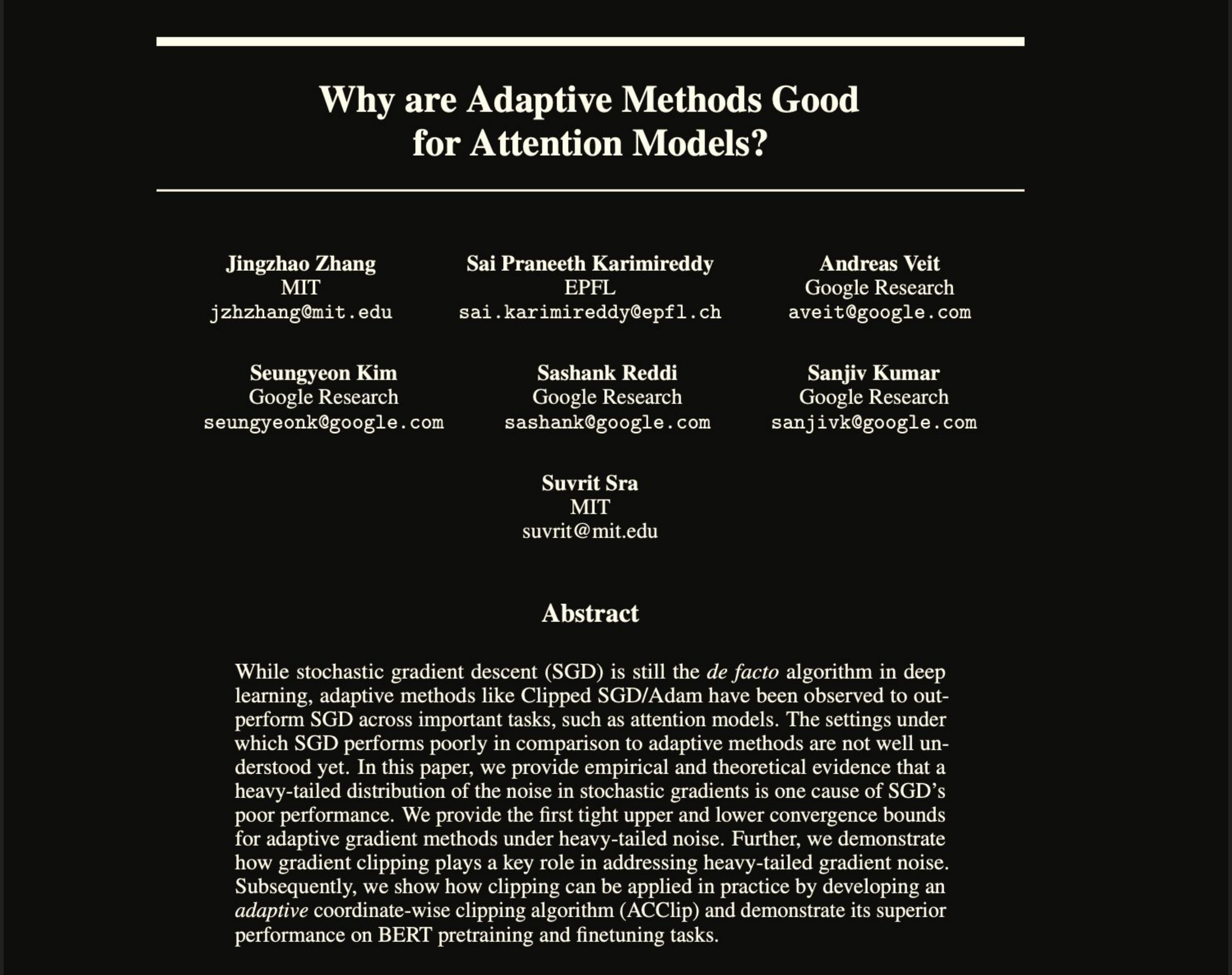
Server

(updates and sends models)

Byzantine-Worker-Tolerant ML

Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. NeurIPS 17.
On The Robustness of a Neural Network. SRDS 17.
The Hidden Vulnerability of Distributed Learning in Byzantium. ICML 18.
Asynchronous Byzantine Machine Learning (the case of SGD). ICML 18.
AGGREGATHOR: Byzantine Machine Learning via Robust Gradient Aggregation. SysML 19.

Robust Distributed Learning. El Mahdi El Mhamdi. EPFL Thesis 19.
Private and Secure Distributed Learning. Georgios Damaskinos. EPFL Thesis 20.



Why are Adaptive Methods Good for Attention Models?

Jingzhao Zhang
MIT
jzhzhang@mit.edu

Sai Praneeth Karimireddy
EPFL
sai.karimireddy@epfl.ch

Andreas Veit
Google Research
aveit@google.com

Seungyeon Kim
Google Research
seungyeonk@google.com

Sashank Reddi
Google Research
sashank@google.com

Sanjiv Kumar
Google Research
sanjivk@google.com

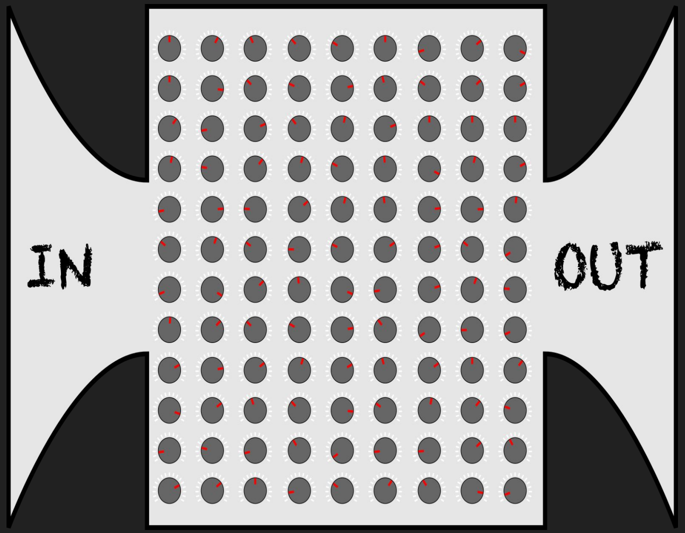
Suvrit Sra
MIT
suvrit@mit.edu

Abstract

While stochastic gradient descent (SGD) is still the *de facto* algorithm in deep learning, adaptive methods like Clipped SGD/Adam have been observed to outperform SGD across important tasks, such as attention models. The settings under which SGD performs poorly in comparison to adaptive methods are not well understood yet. In this paper, we provide empirical and theoretical evidence that a heavy-tailed distribution of the noise in stochastic gradients is one cause of SGD's poor performance. We provide the first tight upper and lower convergence bounds for adaptive gradient methods under heavy-tailed noise. Further, we demonstrate how gradient clipping plays a key role in addressing heavy-tailed gradient noise. Subsequently, we show how clipping can be applied in practice by developing an *adaptive* coordinate-wise clipping algorithm (ACClip) and demonstrate its superior performance on BERT pretraining and finetuning tasks.

Workers

(compute and send gradient estimates)

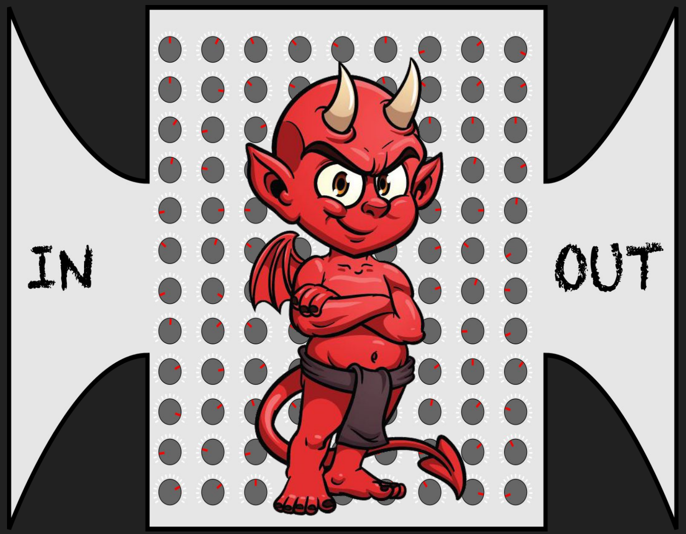


Server

(updates and sends models)

Workers

(compute and send gradient estimates)

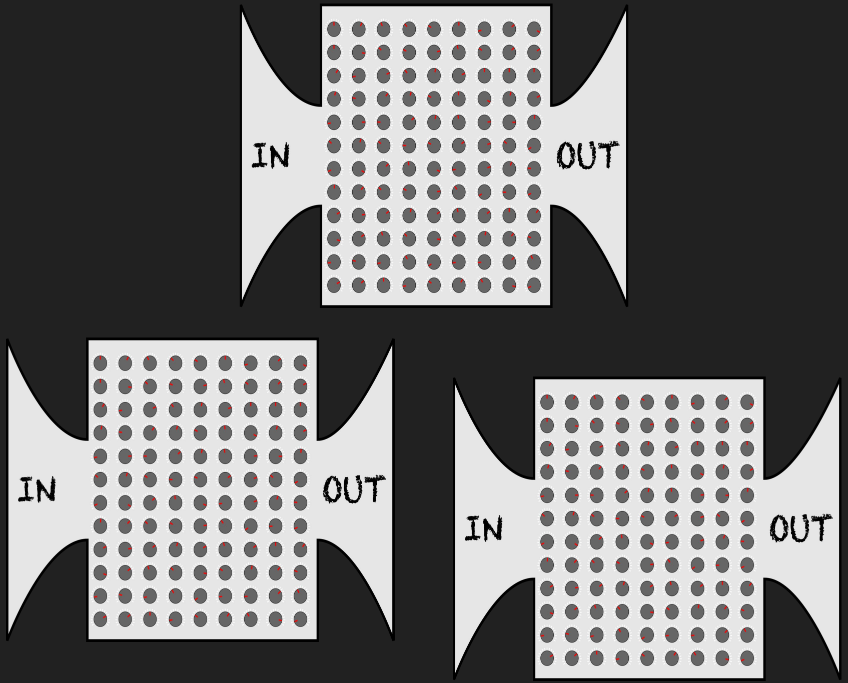


Server

(updates and sends models)

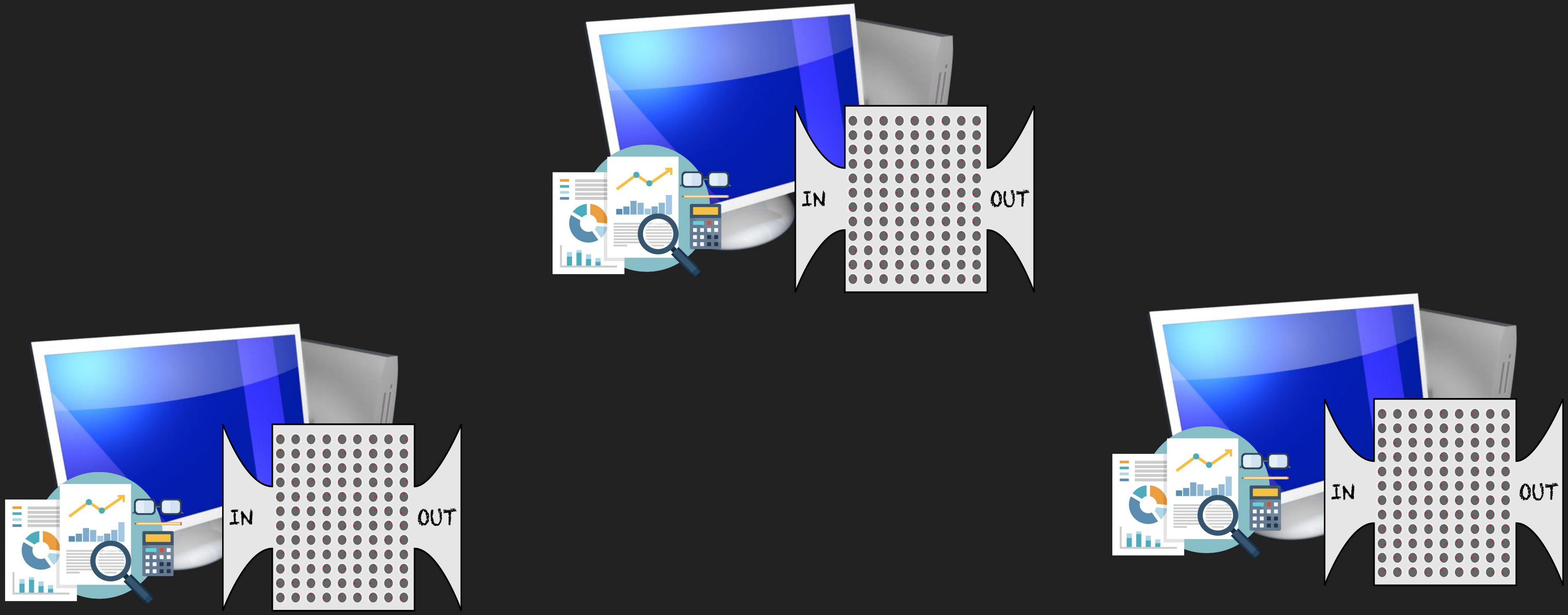
Workers

(compute and send gradient estimates)



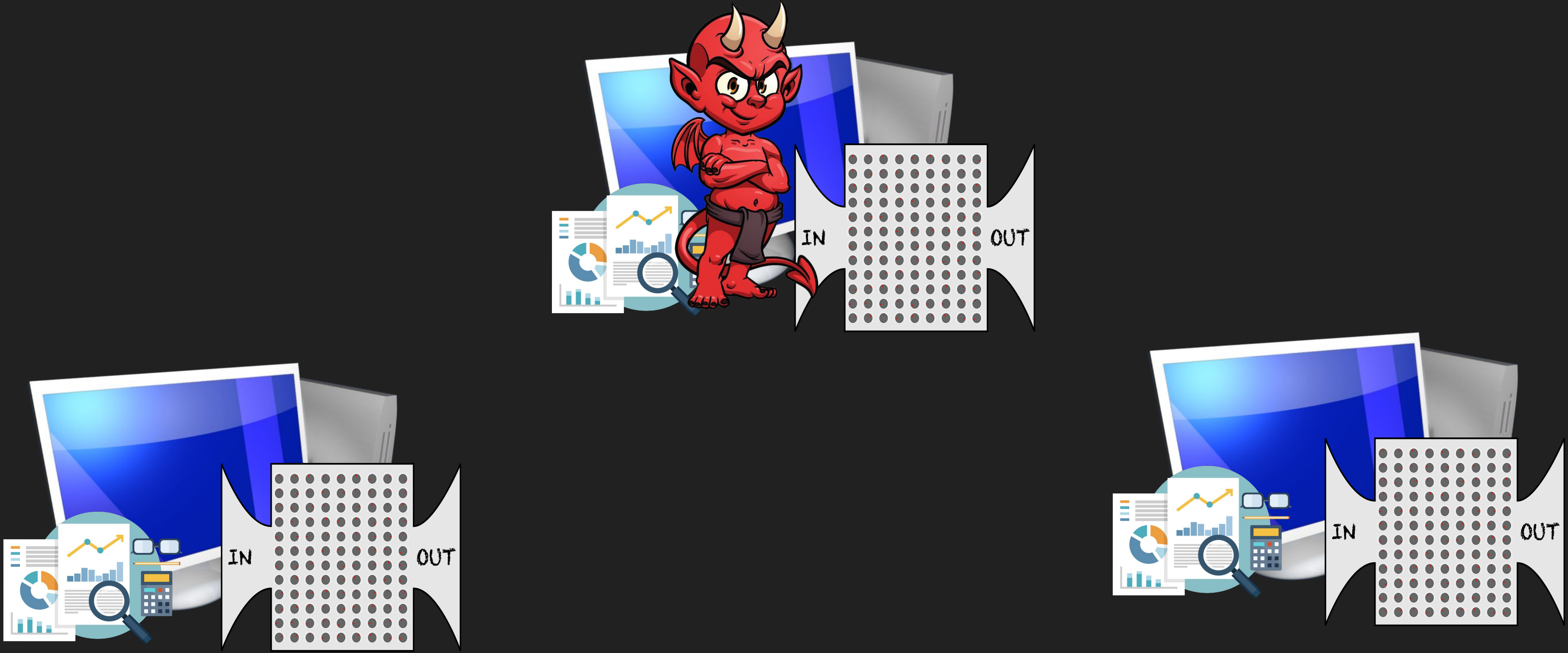
Server

(updates and sends models)



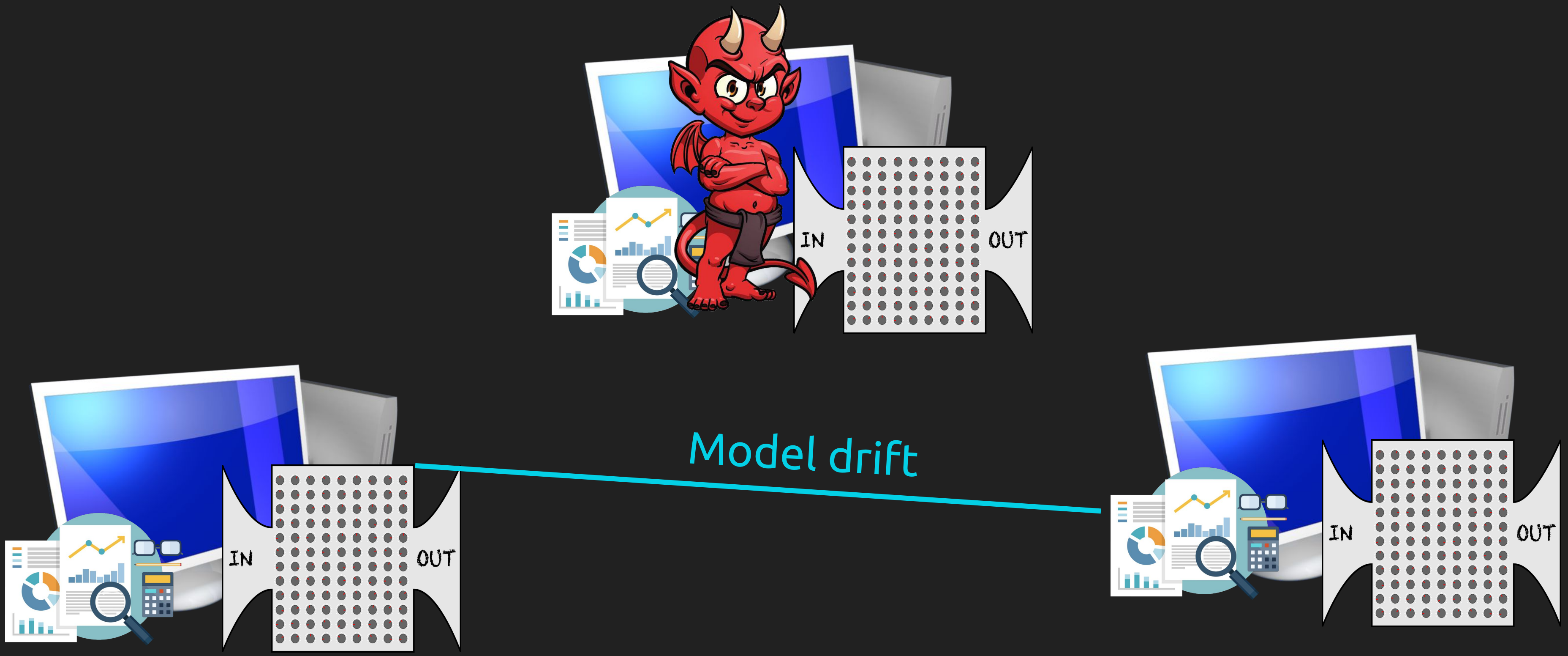
Nodes

(compute and send gradient estimates, update and send models)



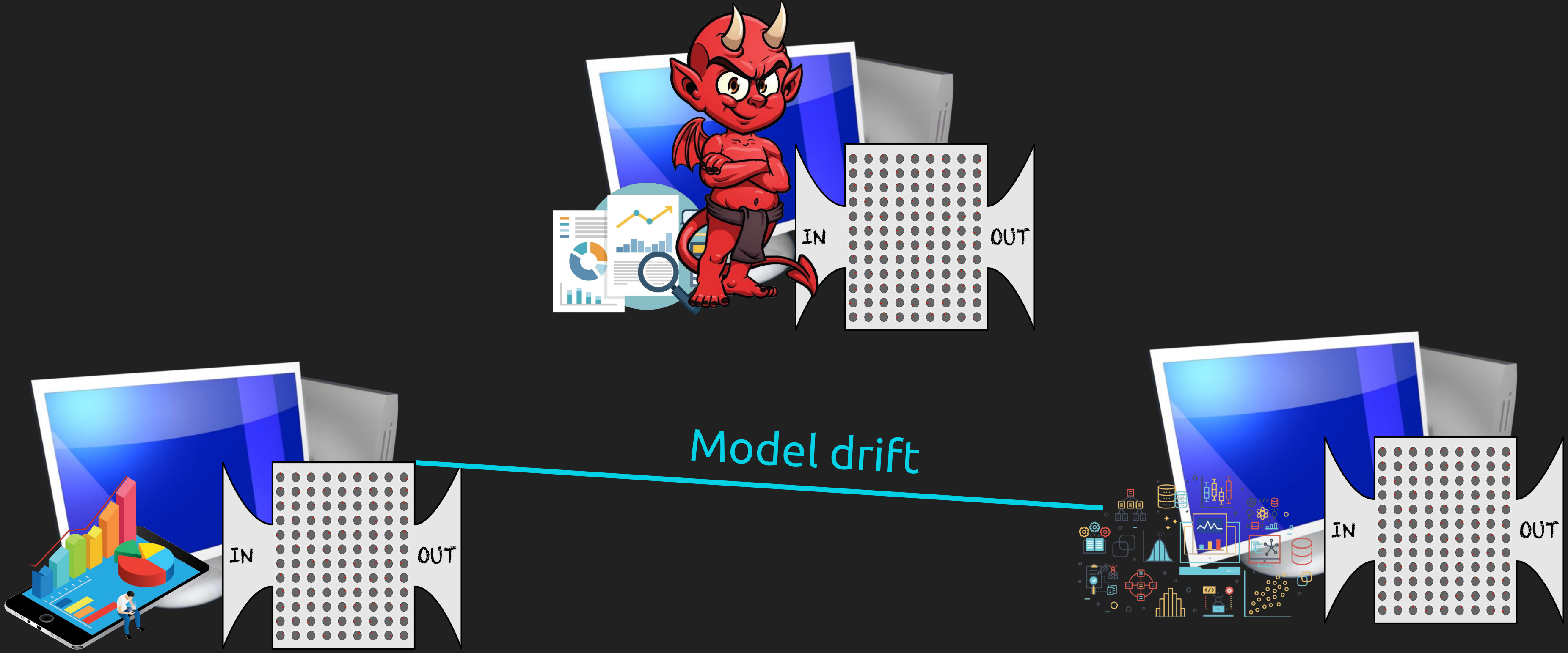
Nodes

(compute and send gradient estimates, update and send models)



Nodes

(compute and send gradient estimates, update and send models)



Model drift

Byzantine, Asynchrony, nonconvex

Heterogeneous data

Definition

C-Collaborative learning is achieved if all honest nodes achieve approximate agreement and small enough gradient.

$$\Delta_2(\vec{\bar{\theta}}) \leq \delta$$

$$||\nabla \bar{\mathcal{L}}(\bar{\theta})||_2 \leq (1+\delta)CK$$

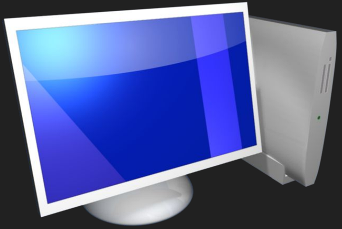
Theorem

Byzantine asynchronous nonconvex heterogeneous

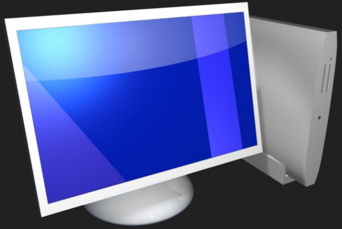
collaborative learning

is equivalent to

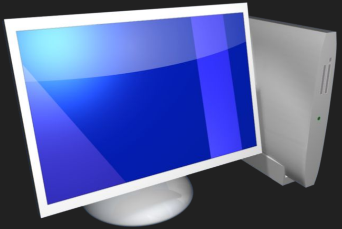
averaging agreement.



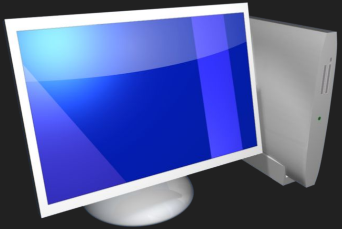
(0,1,3)



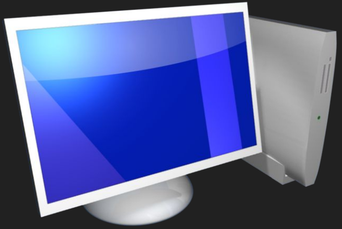
(2,3,3)



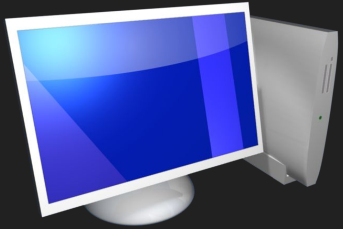
(7,2,6)



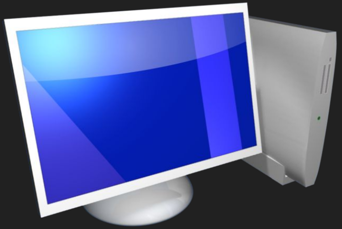
(1,6,5)



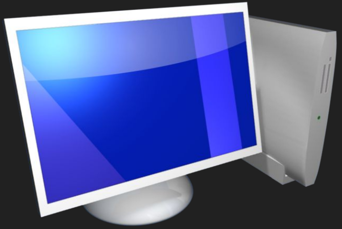
(0,1,3)



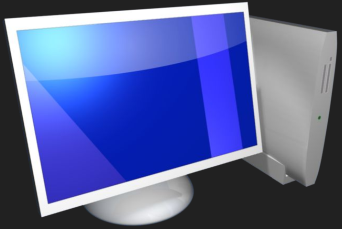
(2,3,3)



(7,2,6)



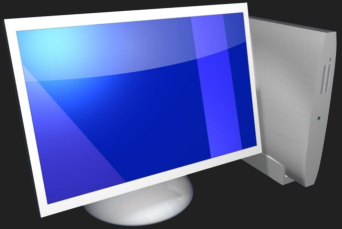
(0,1,3)



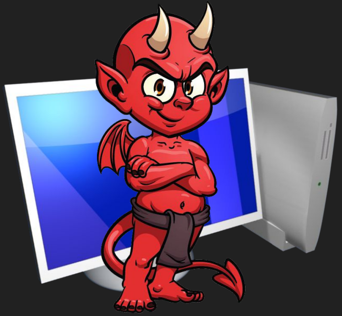
(2,3,3)

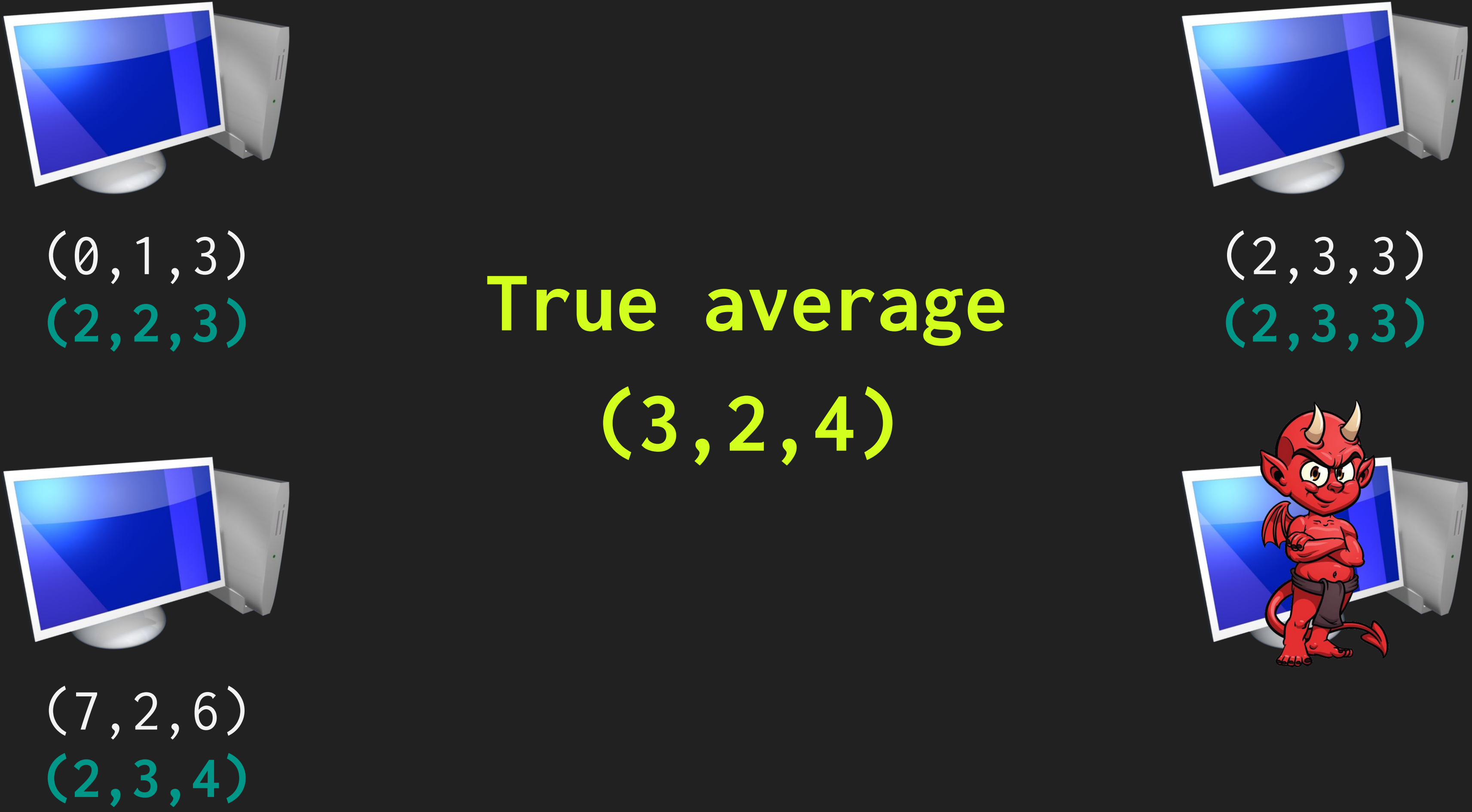
True average

(3,2,4)



(7,2,6)





(0,1,3)
(2,2,3)

True average

(3,2,4)

(2,3,3)
(2,3,3)

(7,2,6)
(2,3,4)

Definition

C-Average agreement is achieved if all honest nodes achieve approximate agreement and estimate well the average.

$$\Delta_2(\vec{y}) \leq \delta$$

$$||\bar{x} - \bar{y}|| \leq C\Delta_2(\vec{x})$$

Theorem

Byzantine asynchronous nonconvex heterogeneous

C-collaborative learning

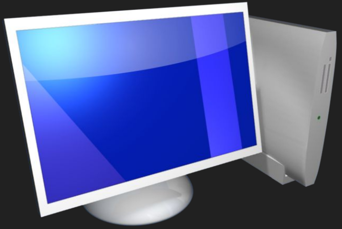
Is (essentially) equivalent to

C-averaging agreement.

Proof sketch

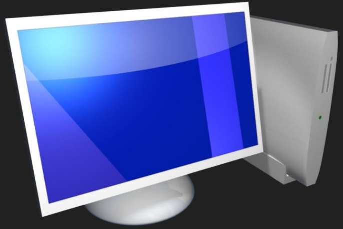
Given a solution to **C-collaborative learning**, by framing **C-averaging agreement** as minimizing of the sum of squares, we can solve **C-averaging agreement**.

Given a solution to **C-averaging agreement**, we run **SGD**, but with **C-averaging** instead of averaging.
We also run **C-averaging** on parameters at every epoch.
This solves **C-collaborative learning** (not easy to prove!).



(0,1,3)

(2,2,3)



(2,3,3)

(2,3,3)

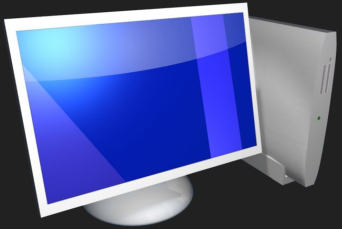
True average

(3,2,4)

Can we solve

Byzantine asynchronous

C-averaging?

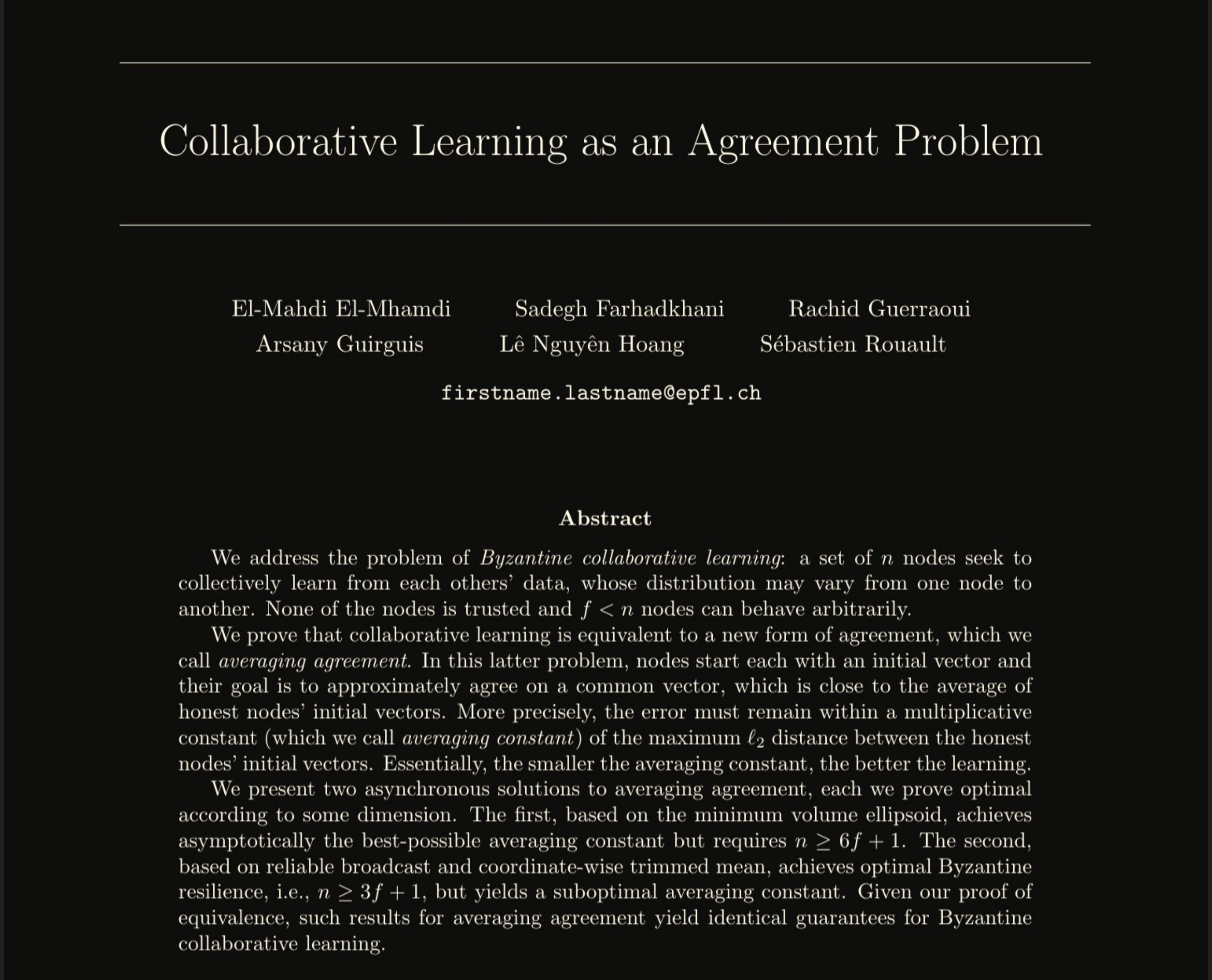


(7,2,6)

(2,3,4)

Theorem

Iterated coordinate-wise trimmed mean with reliable broadcasts and witnesses solves **averaging agreement** for **n** > 3**f**.



Collaborative Learning as an Agreement Problem

El-Mahdi El-Mhamdi Sadegh Farhadkhani Rachid Guerraoui
Arsany Guirguis Lê Nguyên Hoang Sébastien Rouault

`firstname.lastname@epfl.ch`

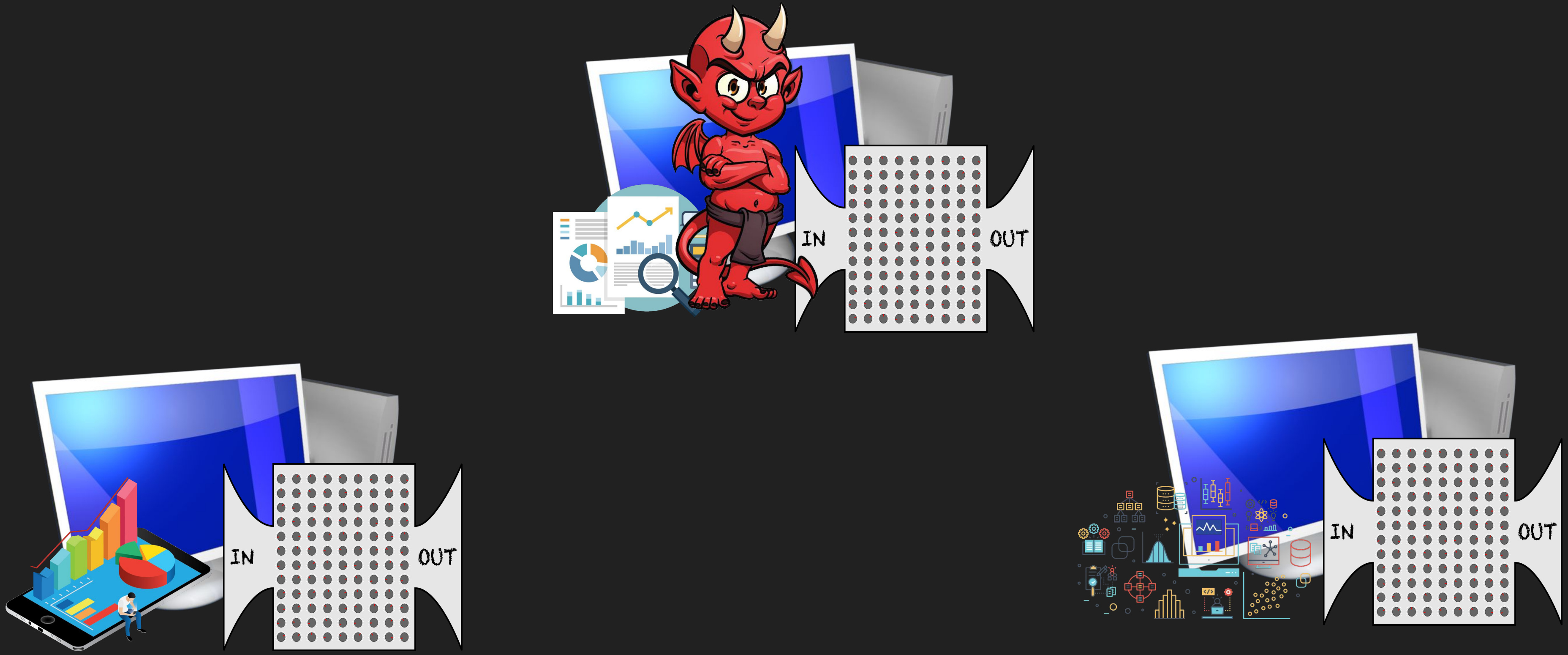
Abstract

We address the problem of *Byzantine collaborative learning*: a set of n nodes seek to collectively learn from each others' data, whose distribution may vary from one node to another. None of the nodes is trusted and $f < n$ nodes can behave arbitrarily.

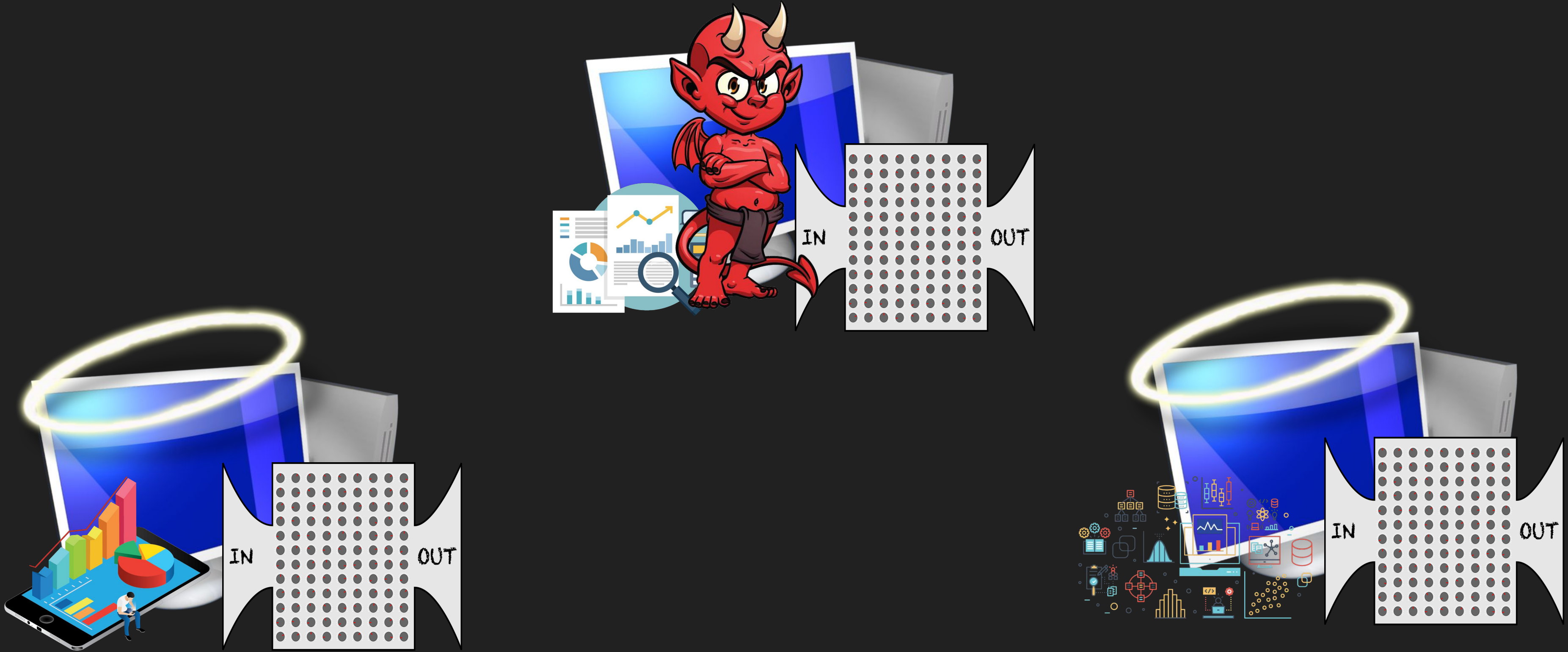
We prove that collaborative learning is equivalent to a new form of agreement, which we call *averaging agreement*. In this latter problem, nodes start each with an initial vector and their goal is to approximately agree on a common vector, which is close to the average of honest nodes' initial vectors. More precisely, the error must remain within a multiplicative constant (which we call *averaging constant*) of the maximum ℓ_2 distance between the honest nodes' initial vectors. Essentially, the smaller the averaging constant, the better the learning.

We present two asynchronous solutions to averaging agreement, each we prove optimal according to some dimension. The first, based on the minimum volume ellipsoid, achieves asymptotically the best-possible averaging constant but requires $n \geq 6f + 1$. The second, based on reliable broadcast and coordinate-wise trimmed mean, achieves optimal Byzantine resilience, i.e., $n \geq 3f + 1$, but yields a suboptimal averaging constant. Given our proof of equivalence, such results for averaging agreement yield identical guarantees for Byzantine collaborative learning.

Open Problems

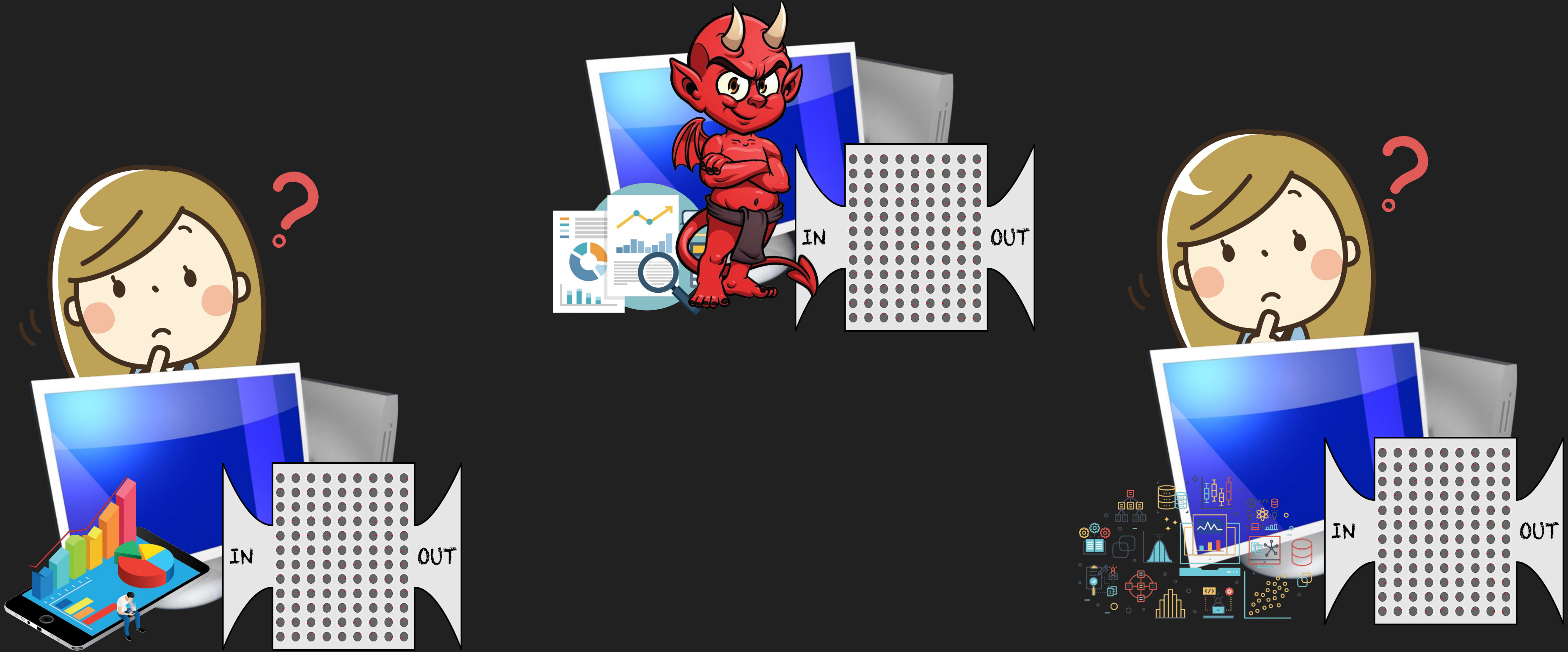


A minority of Byzantine



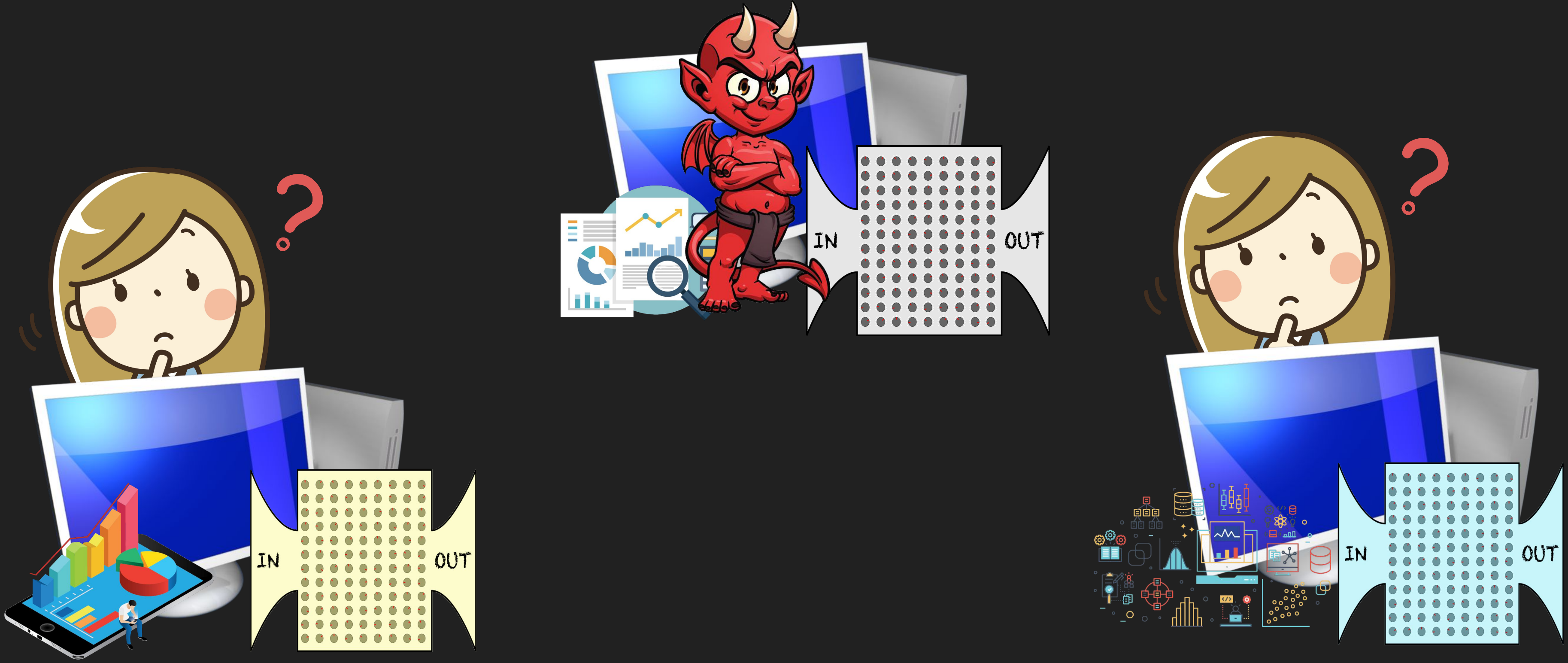
A minority of Byzantine

A majority of honest nodes



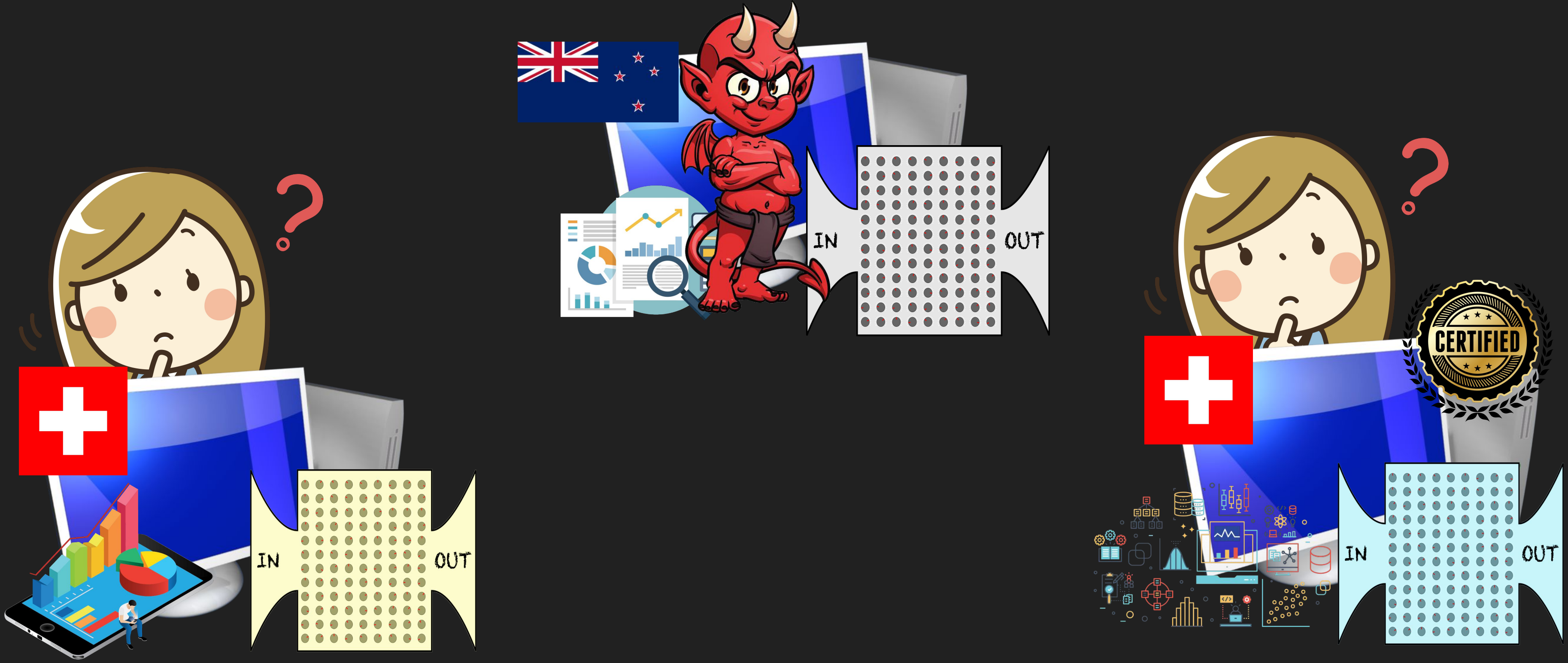
A minority of Byzantine

A majority of strategic nodes



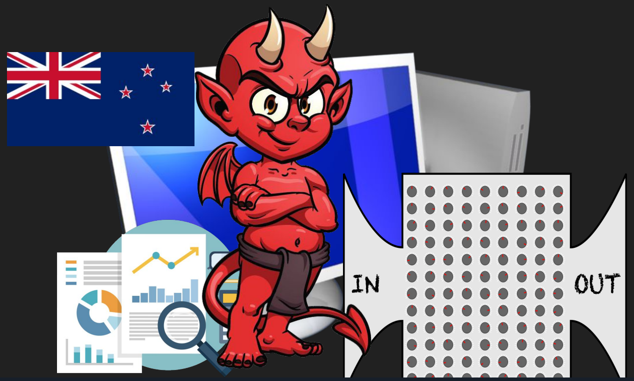
Local Models for Local Contexts

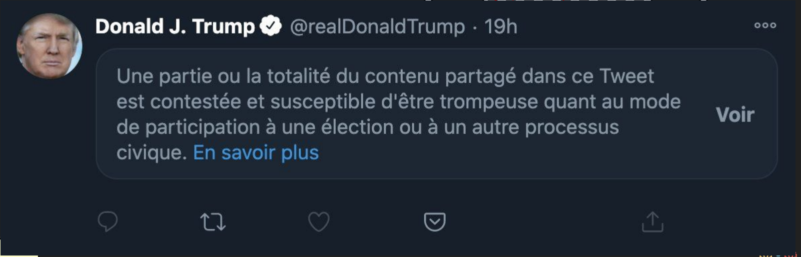
Personalized Collaborative Learning



Collaborative Learning

With Public Information about Nodes





Donald J. Trump @realDonaldTrump · 19h

Une partie ou la totalité du contenu partagé dans ce Tweet est contestée et susceptible d'être trompeuse quant au mode de participation à une élection ou à un autre processus civique. En savoir plus

Voir

Should this be banned?

Collaborative Governance

With Public Information about Nodes

Questions?