Why Federated Learning isn’t like PAPER?

Towards Knowledge Distribution Networks

Marco Canini

KAUST

Principles of Distributed Learning (PODL) Workshop, Oct 2023

https://sands.kaust.edu.sa | marco@kaust.edu.sa

Collaborative Learning

It should “just work”:

- Private
- Accurate
- Personalized
- Efficient
- Robust

As simple as using a piece of PAPER

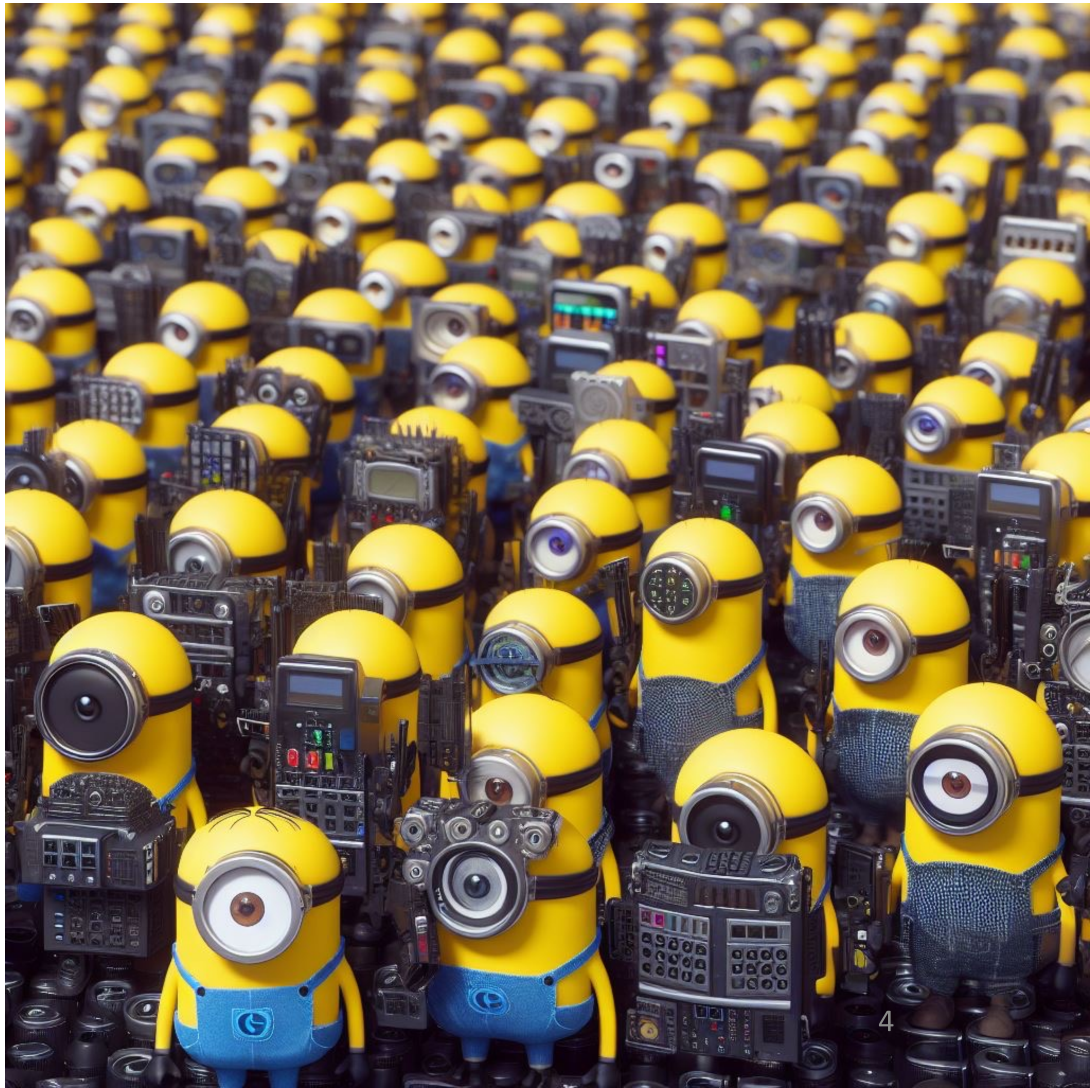
Why Federated Learning isn't like PAPER?

Because Federated Learning is CRUEL

Why turn to FL? Some assumptions …

1. There are N devices, each with a **private** local dataset
2. The whole is greater than the sum of its parts
   - Local training not satisfactory
   - Expect that **collaboration** leads to better model performance

So, why FL is CRUEL?



Why Federated Learning isn't like PAPER?

Because Federated Learning is CRUEL:

- Central servers slow things down
- Resource-intensive
- Unlearn by forgetting
- Eclectic (1,000s of papers and algos)
- Learning w/ heterogeneity is challenging

Unscalable by design

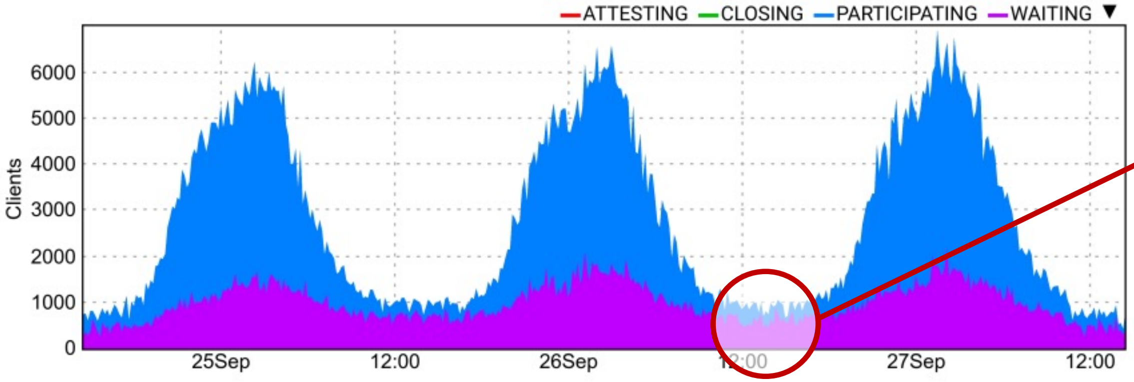
- FL “cannot scale efficiently beyond a few hundred clients training in parallel” [FedBuff]
  - Too many devices → diminishing returns in model performance and training speed
- Sample and work with C*N devices each round. Solution?
  - What do idle devices do? Not much
  - Besides devices have different compute power and intermittent availability; this creates a problem with stragglers, device dropouts, computation wastage [EuroSys’23]
  - In millions-of-devices cases, a device might participate once; what if it’s missed?
- Some work tries to fix this: make FL asynchronous. Solution?
  - “it comes at the cost of higher carbon emissions” [Green FL]

[FedBuff], Nguyen et al., 2022, “Federated learning with buffered asynchronous aggregation”
[Green FL], Yousefpour et al. 2023, “Green Federated Learning”

Can’t sample clients that aren’t available

- Turns out it’s bad for privacy too

“To allow for the DP guarantee, devices participated in training at most once every 24 hours.” [GBlog]



No Wi-Fi | Not charging | Not idle | Available

If you need 1000 clients per round, and only 1000 clients are available, you have two (bad) options:

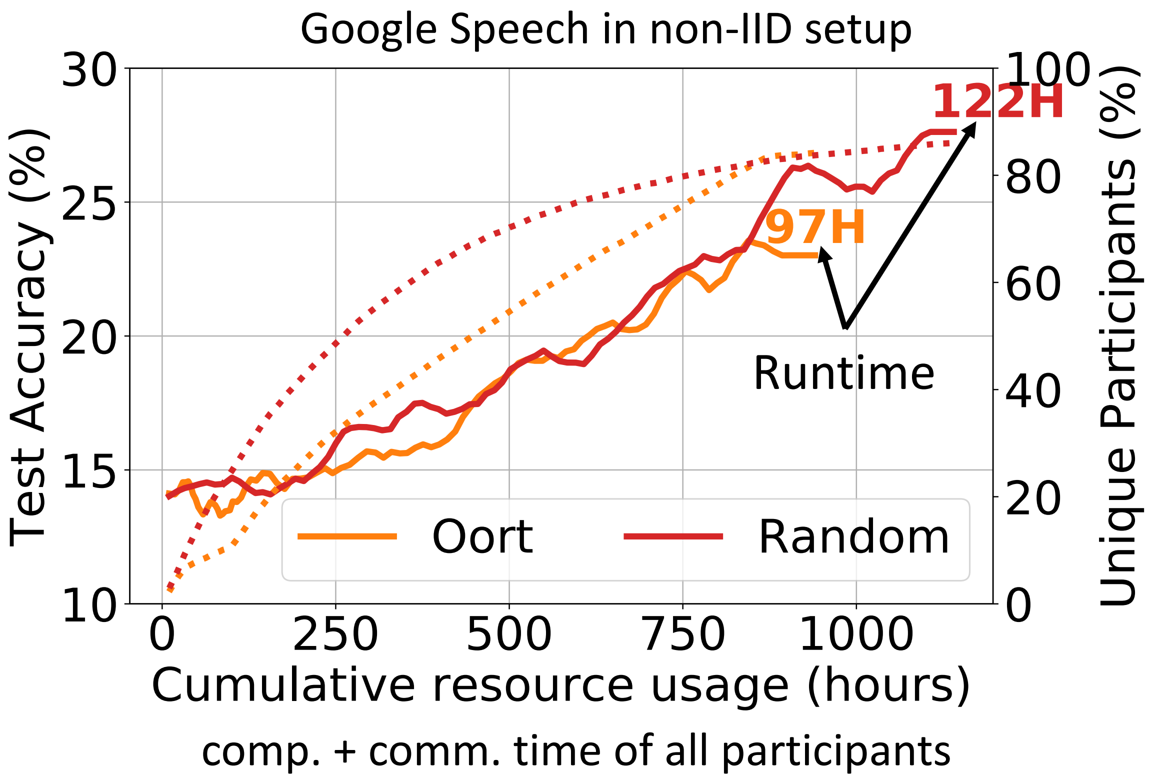
1. Pause training and wait until more clients are available
2. Continue training without sampling (no amplification)

McMahan’s talk at FL@ICML’23

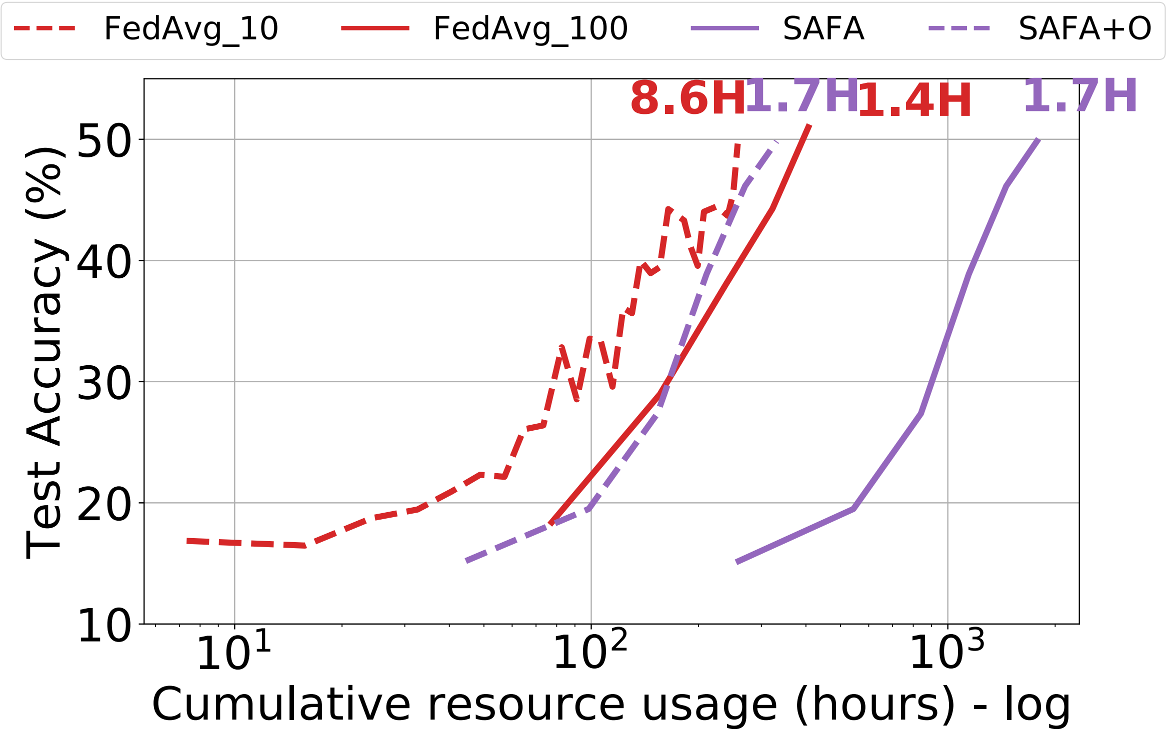
Bonawitz et al., 2019, “Towards Federated Learning at Scale: System Design”
Balle et al., 2020, “Privacy Amplification via Random Check-Ins”
[GBlog] McMahan & Thakurta, Google blog, “Federated Learning with Formal Differential Privacy Guarantees”

[EuroSys'23]

Resource-to-quality



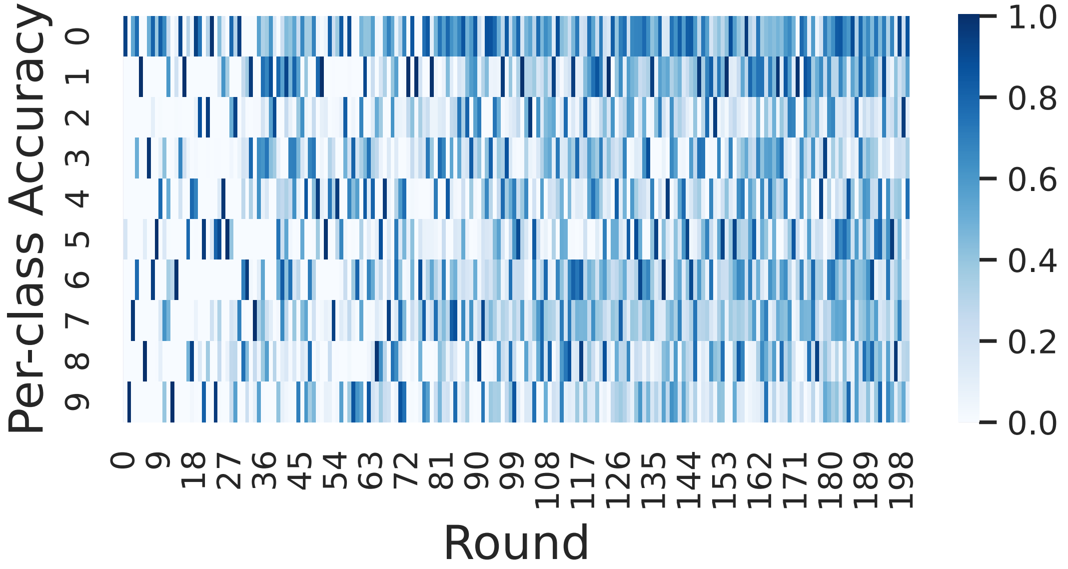
High diversity is helpful but hard to manage (high proportion of dropouts & stragglers → high resource wastage)



System efficiency is desirable but low inclusivity of participants worsen things

[Oort], Lai et al., 2021, "Oort: Efficient Federated Learning via Guided Participant Selection"
[SAFA], Wu et al. 2021, "SAFA: A Semi-Asynchronous Protocol for Fast Federated Learning With Low Overhead"

Forgetting is a problem



- Data heterogeneity (non-IID) leads to forgetting, which makes learning **inefficient**
- Local: a device “overrides” certain knowledge
- Global: aggregation step averages, doesn’t “fuse” knowledge

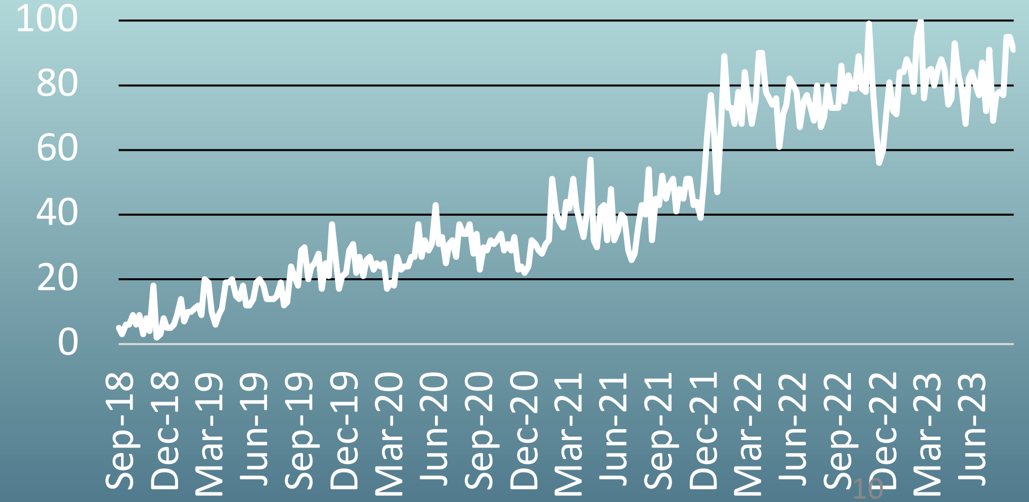
[FedMA], Wang et al., 2020, “Federated learning with matched averaging”

How many FL methods are there?

- How will system designers pick the “right” ones for their needs?
- A moving target?
- Are there (distributed) systems problem worth tackling?

- Search title “federated” + “learning”
  - 3,100+ arXiv cs
  - 500+ ACM DL (413 in past 2y)
  - 2,500+ IEEE Conferences (1,582 in past 2y)
  - 1,200+ IEEE Journals (993 in past 2y)

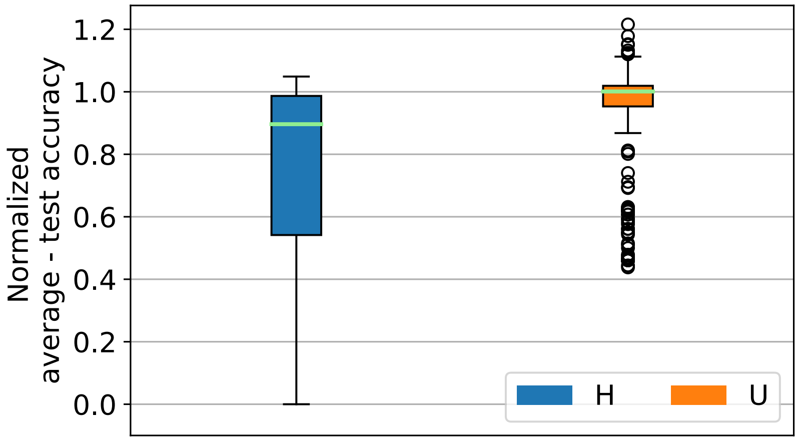
"Federated Learning":
Interest Worldwide [Google Trends]

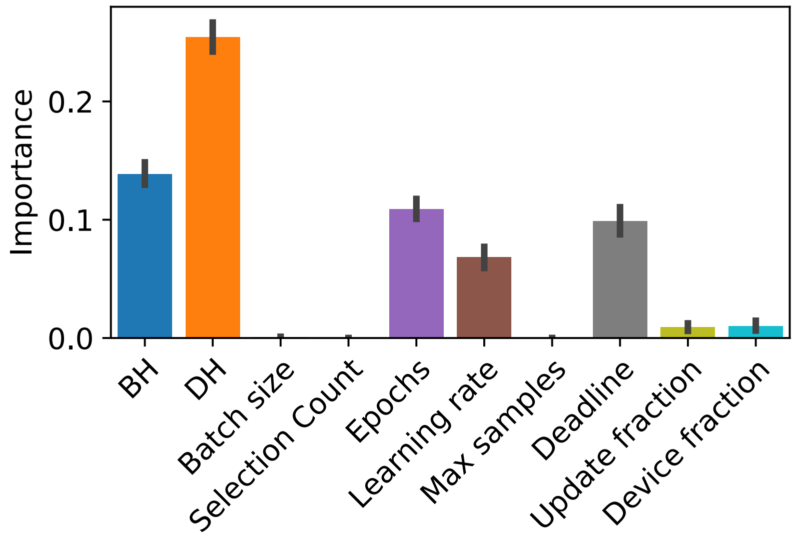


Impact of device and behavioral heterogeneity

[EuroMLSys’22, IEEE IoT Journal 2023]

- Characterization of heterogeneity on model quality and fairness
  - Empirical study spanning ~1.5K configurations on 5 FL benchmarks
  - Heterogeneity causes degradation up to 4.6× in quality and 2.2× in fairness

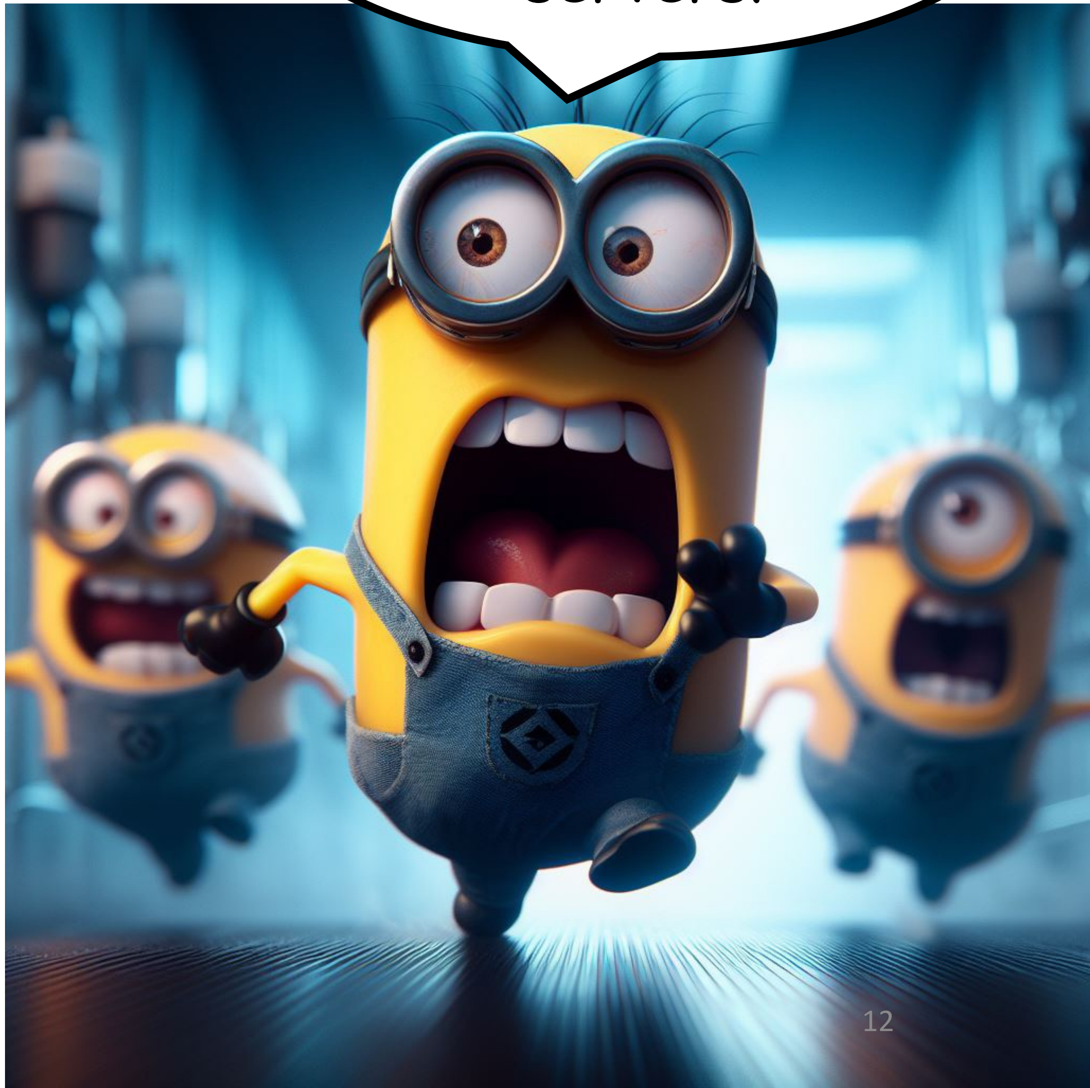




Okay, FL might be CRUEL but car

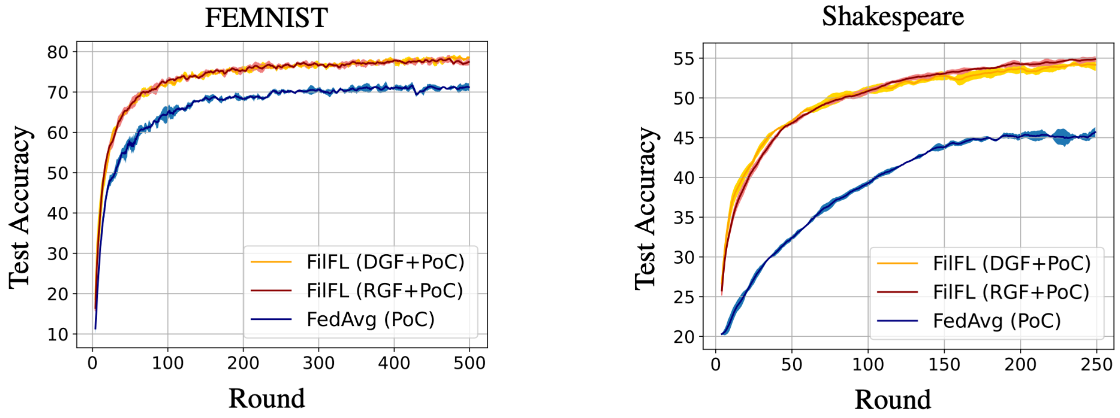
- Server lowers communication complexity …
- But aggregation step is challenged by the statistical efficiency of learning
  - Both cohort size and averaging mechanism

- Is a server really needed?
- Do we need to aggregate everything in one model all the times?



FilFL: Client Filtering in FL

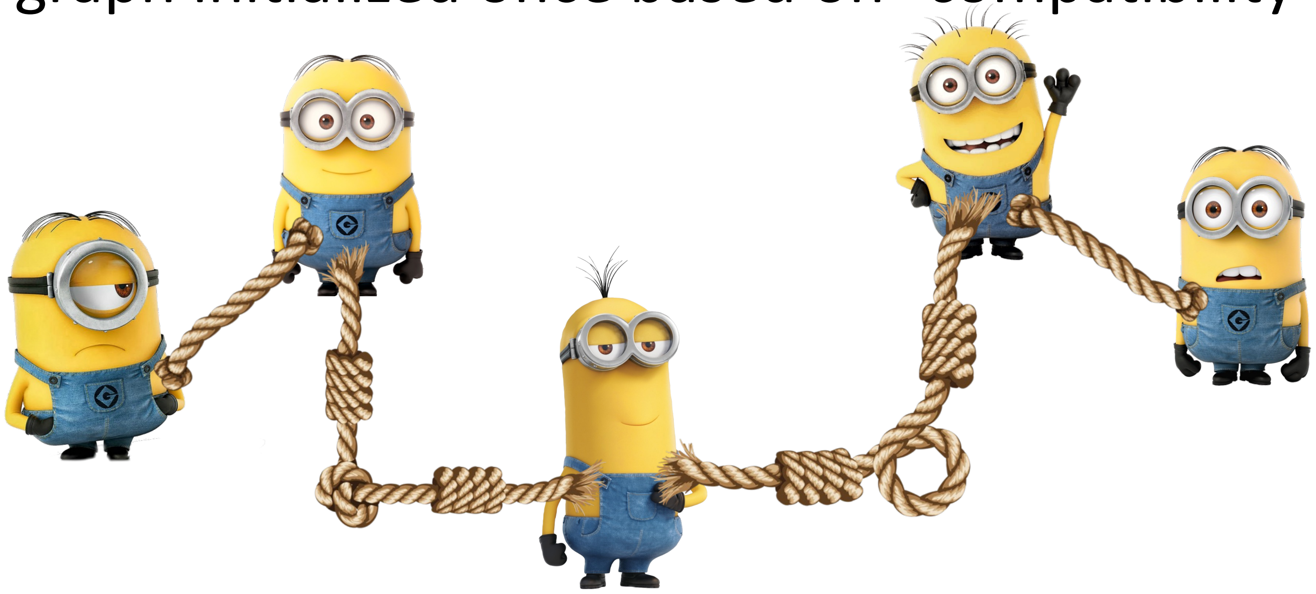
- We noticed in [FilFL]
  - Not all available clients are always suitable for collaboration
  - Filtering clients, online, can lead to **faster convergence** and **higher accuracies** (up to 10 pp)



[FilFL] *Fourati, F., Kharrat, S., Aggarwal, V., Alouini, M. S., & Canini, M. (2023). FilFL: Client Filtering for Optimized Client Participation in Federated Learning. arXiv preprint arXiv:2302.06599.*

A simple decentralized scheme

- Each device aims to train a personalized model
  - Expected it will generalize well on local test set
- Assume a collaboration graph, edge means devices collaborate
- Collaboration graph initialized once based on "compatibility" [FilFl arXiv:2302.06599]



A simple decentralized scheme

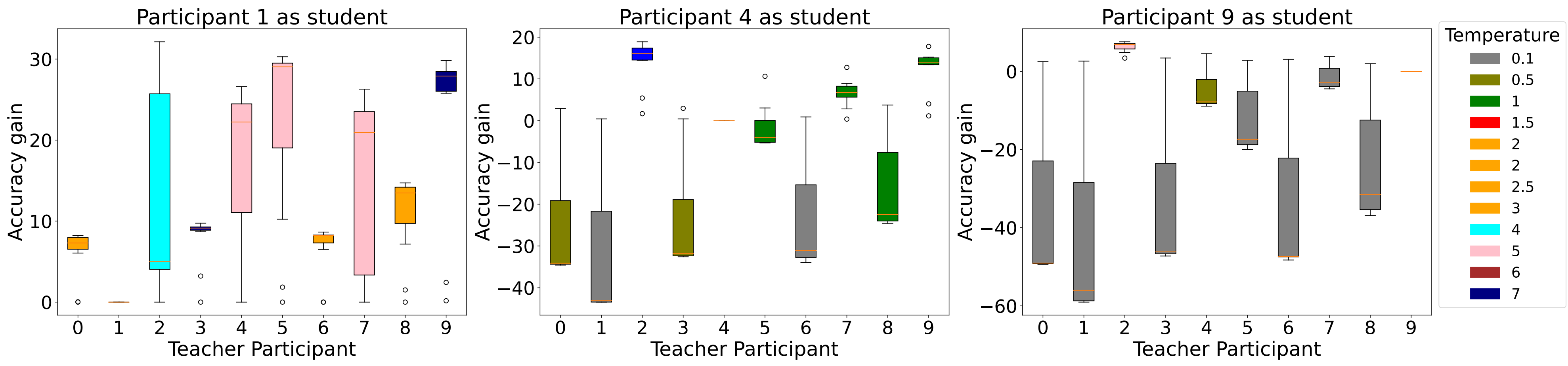
- Every round, device trains and exchanges updates based on the graph
- Average vertex degree ~30
- Exp with CIFAR10, 100 clients; local best accuracy within 50 rounds:

Setting	**Collab. graph**	Local	Ditto	Fedrep	APFL	FedAvg	Fedprox	PerFed Avg	FedAvg FT	FedProx FT
miss 5 classes	**70.70**	67.01	70.10	67.99	70.11	46.33	46.11	69.57	70.50	63.68
miss 7 classes	**78.00**	77.30	77.33	74.39	76.90	43.14	41.34	76.70	77.80	71.90

[EuroMLSys'23]

Knowledge distillation replaces averaging

- Student model learns by mimicking output of teacher model
- Transfer knowledge between models in distributed setting
  - Can transfer model outputs instead of full model updates
  - Can work across different model architectures
  - Can boost learning (learn from logits: $\mathcal{L} = \mathcal{L}_{CE}(p^S, y) + \alpha\mathcal{L}_{KL}(p^S, p^T)$)



What I'd like:
Knowledge Distribution Network (KDN)

- How do I make "maccheroni alla chitarra"?
- Who is more expert than me?

Some knowledge, I need
Some, I don't

Necessarily not every device can help

Some key components

These seem to be necessary:

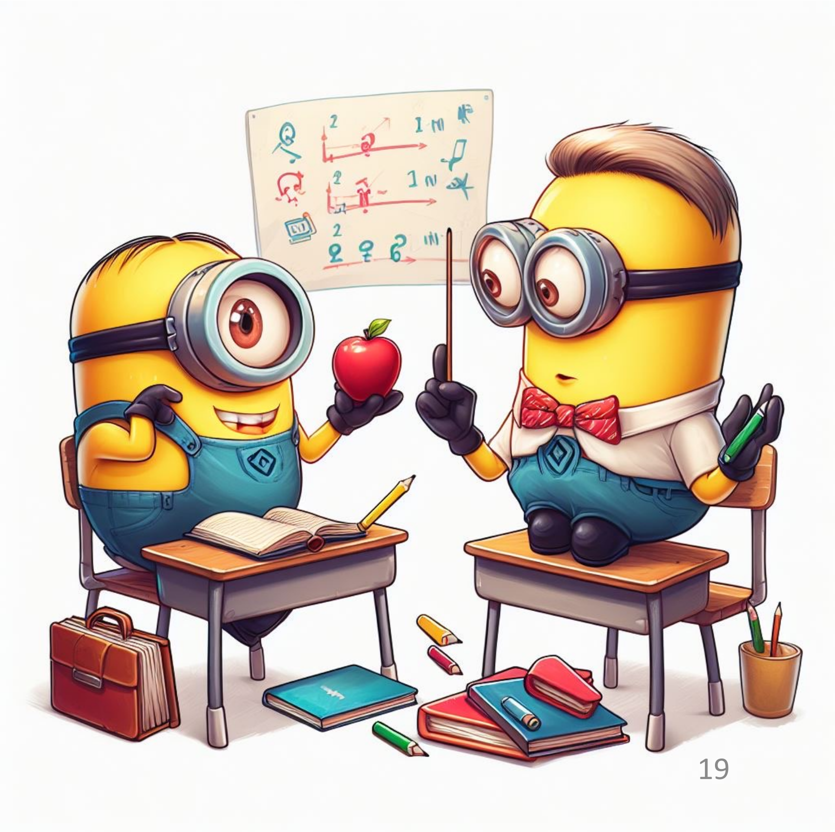
- Knowledge transfer (efficient pipelines)
- Routing for knowledge
- Assisted learning / knowledge vaults

CAUTION: I don’t have good solutions to all of these

Knowledge Distillation

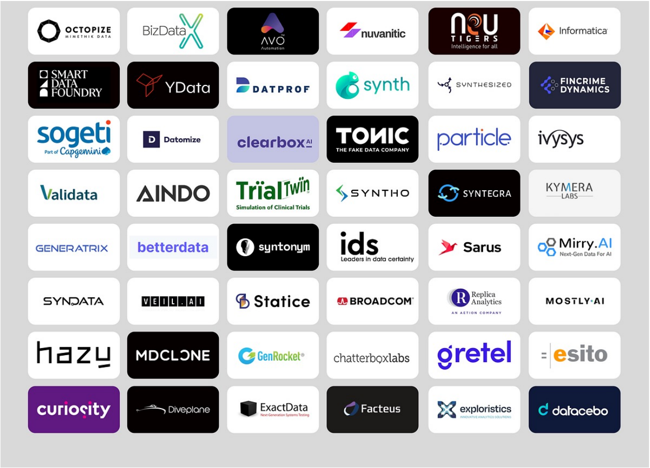
- A teacher model can infuse knowledge to a student model
- I need the two models plus a dataset, it costs extra FP per data point
- Actually, more than one teacher works too, and might be better

- ☹ mind the hyperparameters [EuroMLSys’23]



Knowledge Discovery & Routing

- Is there a “consistent hashing” to look up teacher models?
- I like the idea of data previews:
  “try it before you buy it”
  - Offered by generative data vendors

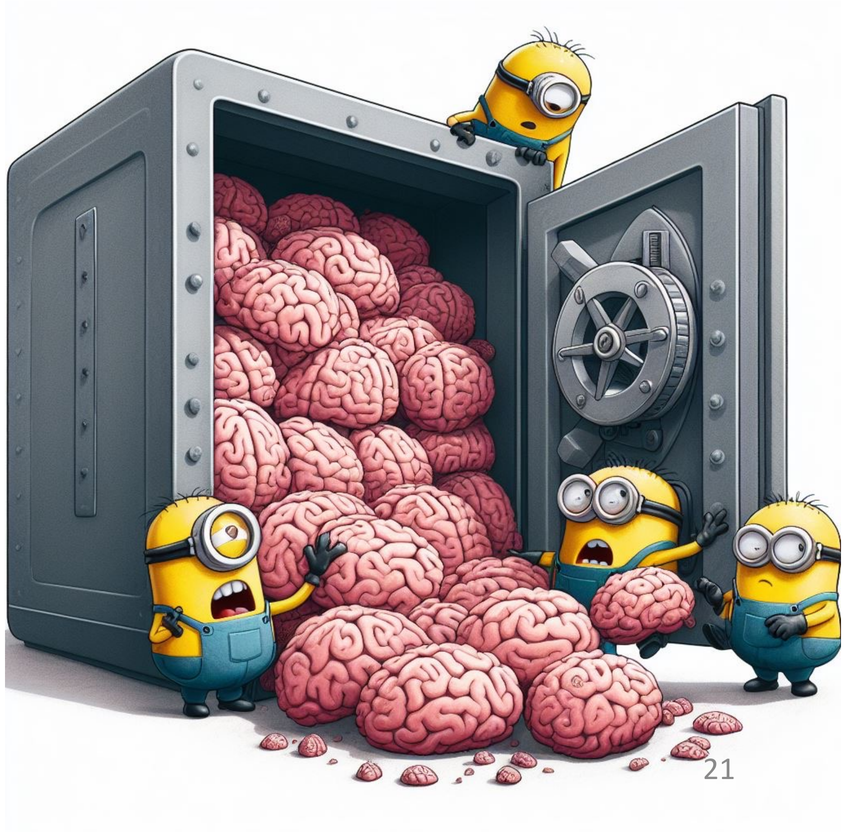


- What is the equivalent for model preview?
  - Maybe a preview of model results on synth data
  - Because then I can route on “synth data distance”

Knowledge Storage & Dissemination

- A storage layer seems necessary
- Envisioned lots of models, intermediates, non-private/synth datasets
- Edge-based or cloud-hosted? marketplace?
- Security probably necessary

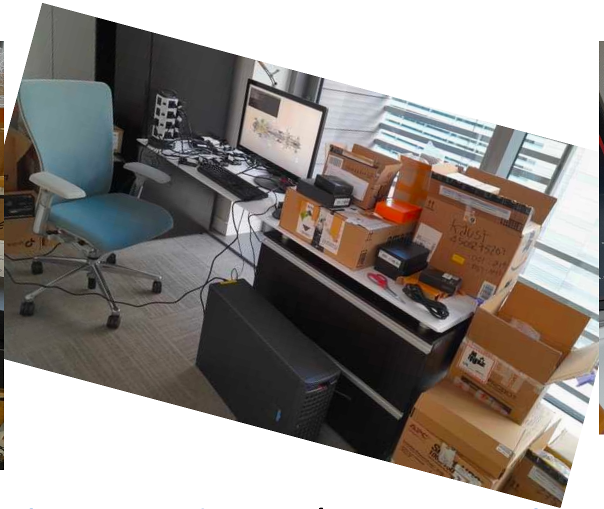
- With some compute, could (partly) offload (secure) knowledge distillation steps



KDN might be like PAPER

- Can you help us build it?

- BTW, we are setting up a testbed for FL / KDN research to get results of run times, energy consumption, etc. for real



https://sands.kaust.edu.sa | marco@kaust.edu.sa