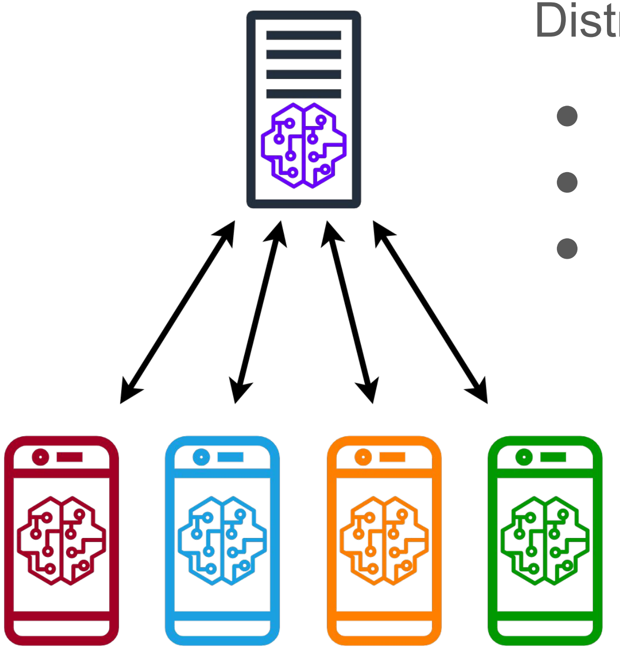
On Noisy Evaluation in Federated Hyperparameter Tuning

[+]Kevin Kuo, [+]Pratiksha Thaker, [+]Mikhail Khodak,
[*]John Nguyen, [*]Daniel Jiang, [+]Ameet Talwalkar, [+]Virginia Smith

[+] Carnegie Mellon University

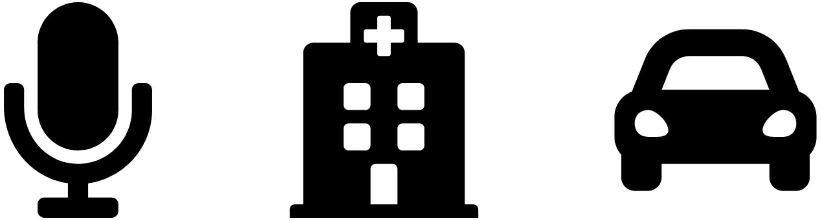
[*] Meta AI

Federated Learning (FL)

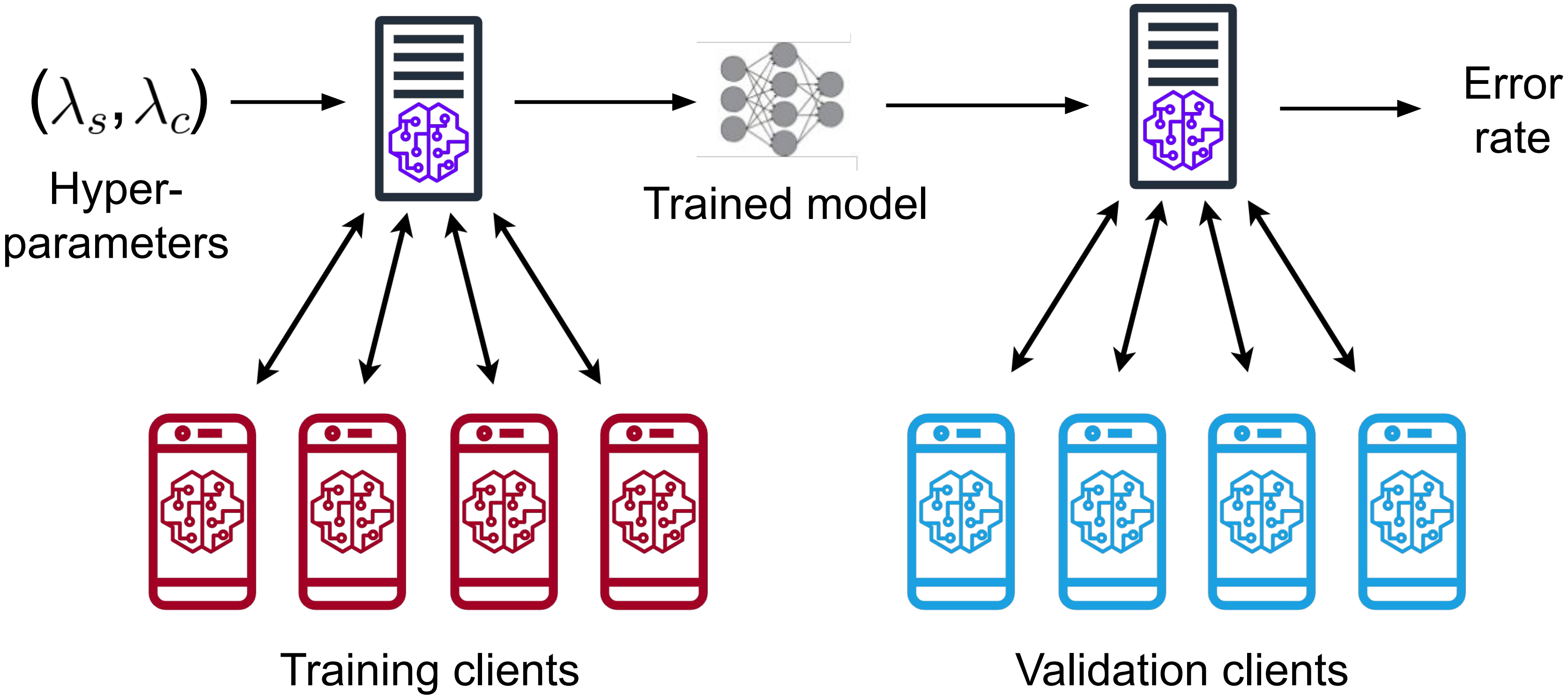


Distributed ML across heterogeneous networks

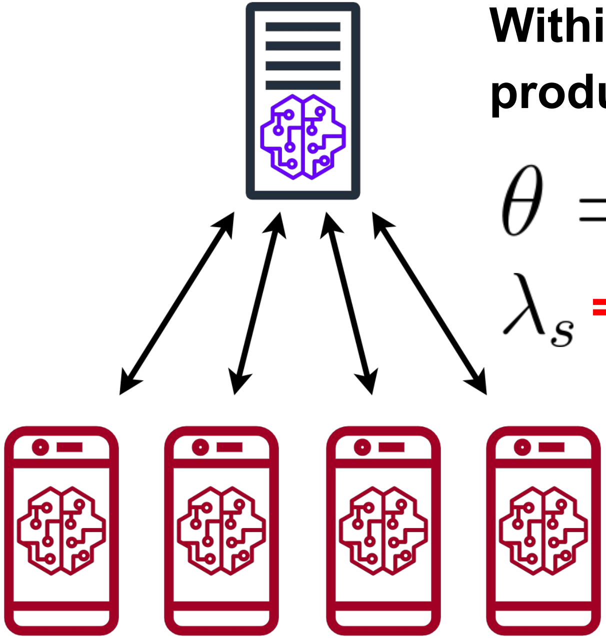
- Potentially massive networks
- Communicates model rather than data
- Applications include data from e.g. mobile phones, medical records, and remote sensors



Cross-Device FL: Training / Evaluation



Federated Training



Within an FL round: clients fine-tune a global model, producing local models which the server aggregates.

$$\theta = \texttt{ServerOPT}(\theta, \{\theta_k\}_{k \in K}, \lambda_s)$$

λ_s **= {learning rate, beta1, beta2} (Adam)**

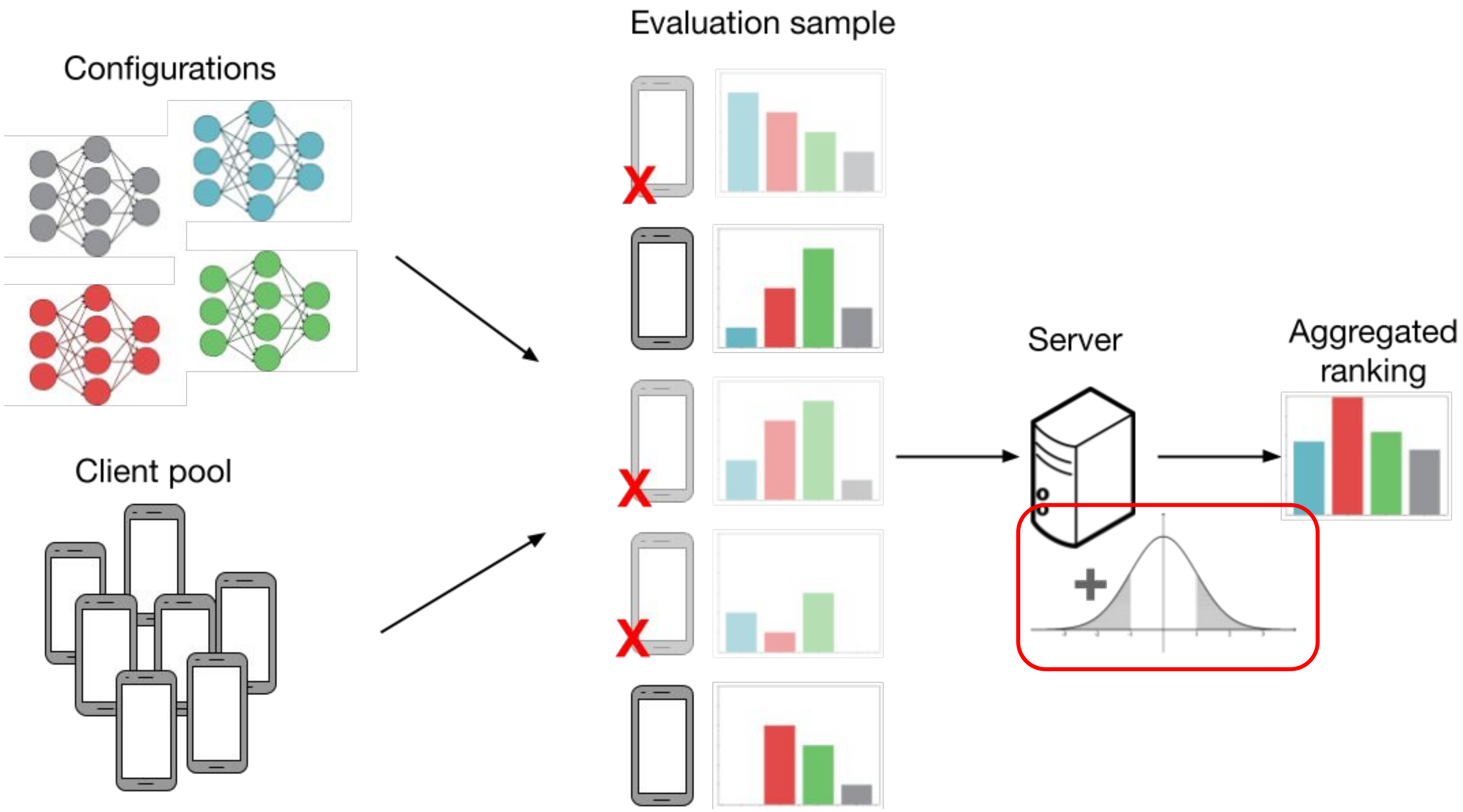
$$\theta_k = \texttt{ClientOPT}(\theta, X_k, \lambda_c)$$

λ_c **= {learning rate, momentum, batch size} (SGD)**

Federated Evaluation

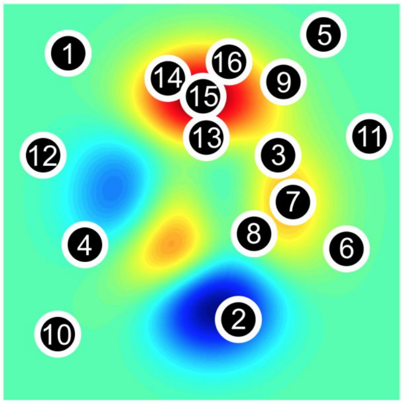
Federated Evaluation

Federated Evaluation

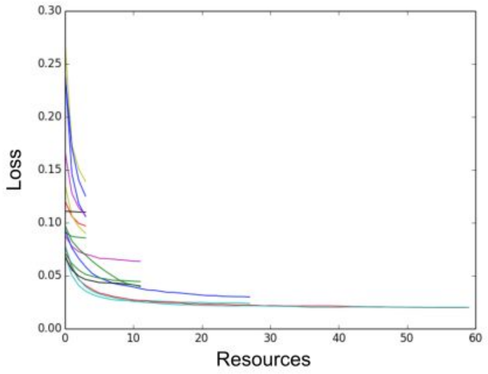


Hyperparameter (HP) Tuning

In FL, **HPs** for **client optimization** and **server aggregation** are critical to train a good model.



(a) Configuration Selection



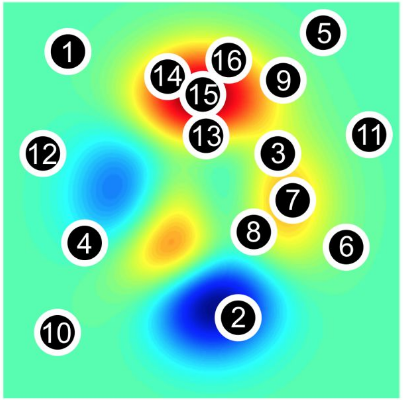
(b) Configuration Evaluation

Hyperparameter (HP) Tuning

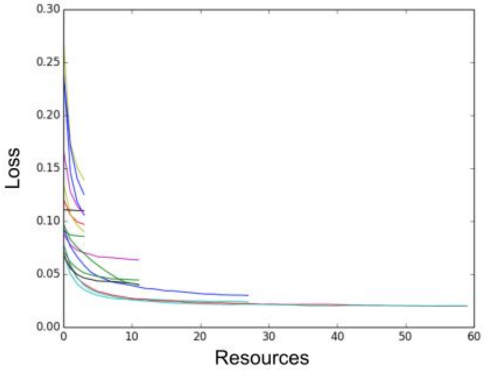
In FL, **HPs** for **client optimization** and **server aggregation** are critical to train a good model.

Standard HP tuning methods work well for classic ML (centralized training).

- random search
- adaptive HP **selection**
- adaptive resource **allocation**



(a) Configuration Selection



(b) Configuration Evaluation

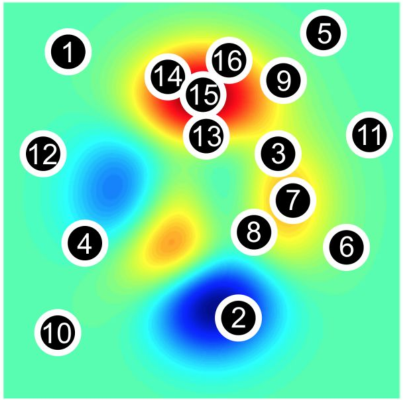
Hyperparameter (HP) Tuning

In FL, **HPs** for **client optimization** and **server aggregation** are critical to train a good model.

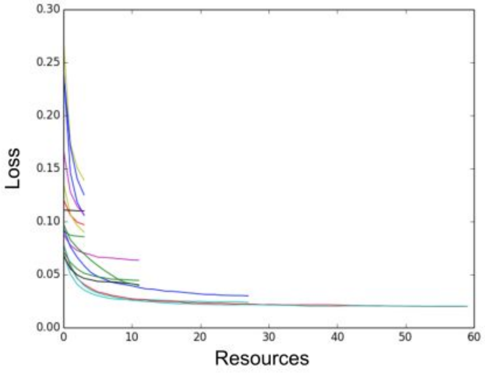
Standard HP tuning methods work well for classic ML (centralized training).

- random search
- adaptive HP **selection**
- adaptive resource **allocation**

However, many sources of **noise** in **FL** contribute to **low-quality evaluations** and **severely impact** these HP tuning methods.



(a) Configuration Selection



(b) Configuration Evaluation

Questions

Question 1: To what extent does **subsampling** validation clients degrade the performance of HP tuning algorithms?

Question 2: How, and to what extent, do the factors of **data heterogeneity, systems heterogeneity, and privacy** exacerbate issues of subsampling?

Question 3: In **noisy settings**, how do popular HP tuning algorithms compare to simple baselines?

Questions

Question 1: To what extent does **subsampling** validation clients degrade the performance of HP tuning algorithms?

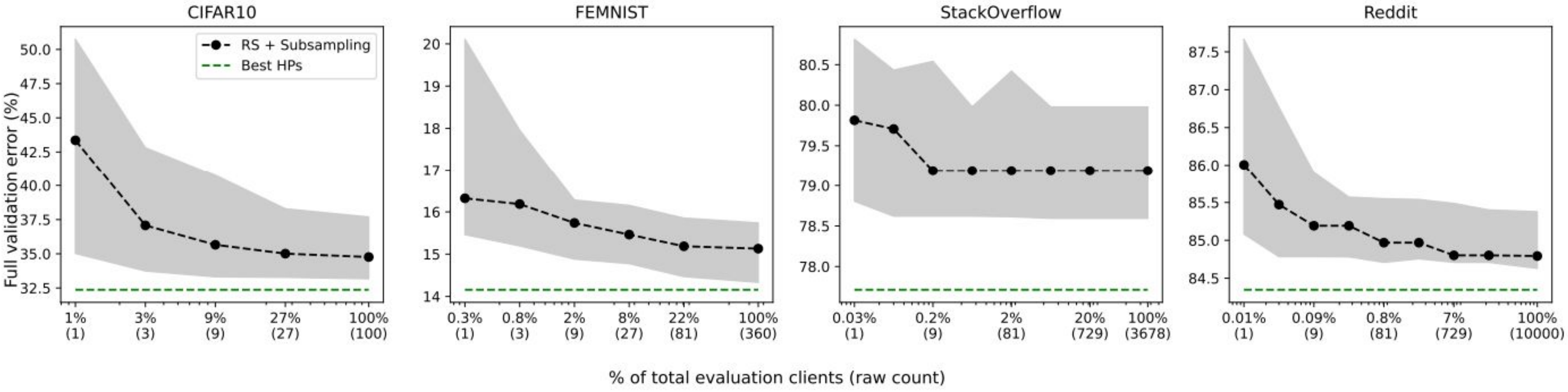
Question 2: How, and to what extent, do the factors of **data heterogeneity, systems heterogeneity, and privacy** exacerbate issues of subsampling?

Question 3: In **noisy settings**, how do popular HP tuning algorithms compare to simple baselines?

We show there are multiple sources of compounding noise in FL, and under this noise, state-of-the-art HPO methods can perform catastrophically poorly, even worse than simple baselines (random search).

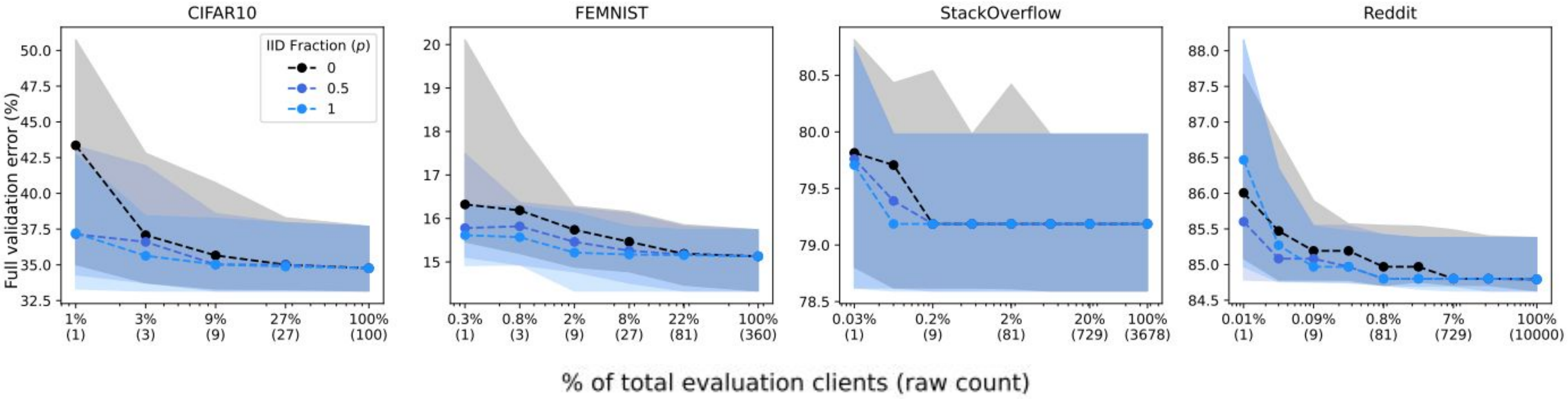
Subsampling

Subsampling very few clients hurts HP tuning performance.



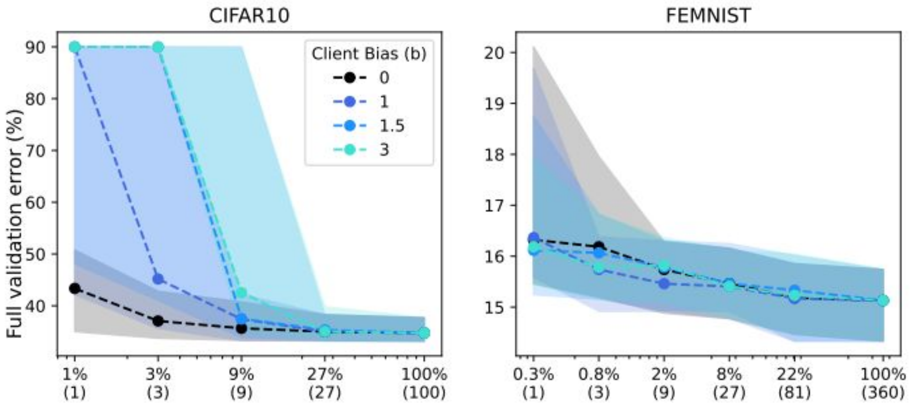
Data Heterogeneity

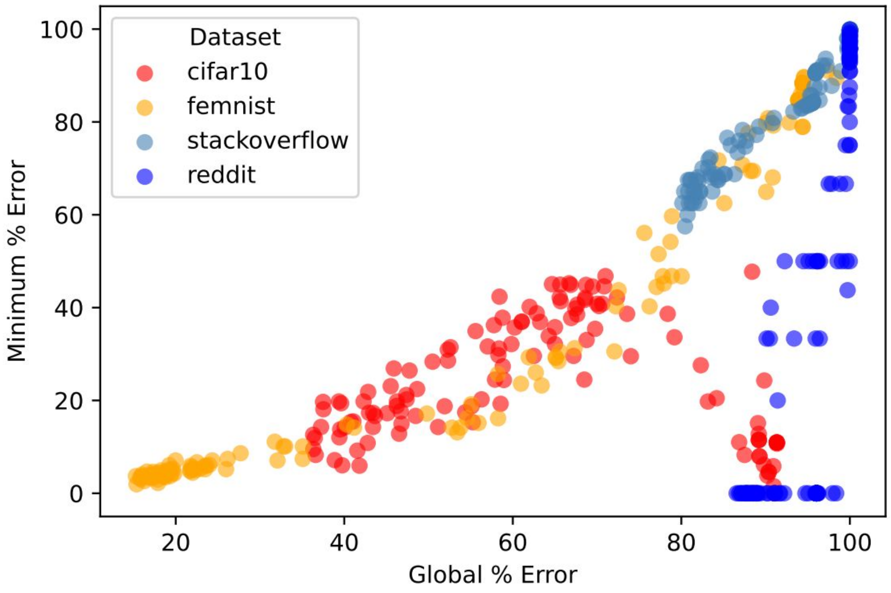
Data heterogeneity exacerbates the negative effects of subsampling.



Systems Heterogeneity

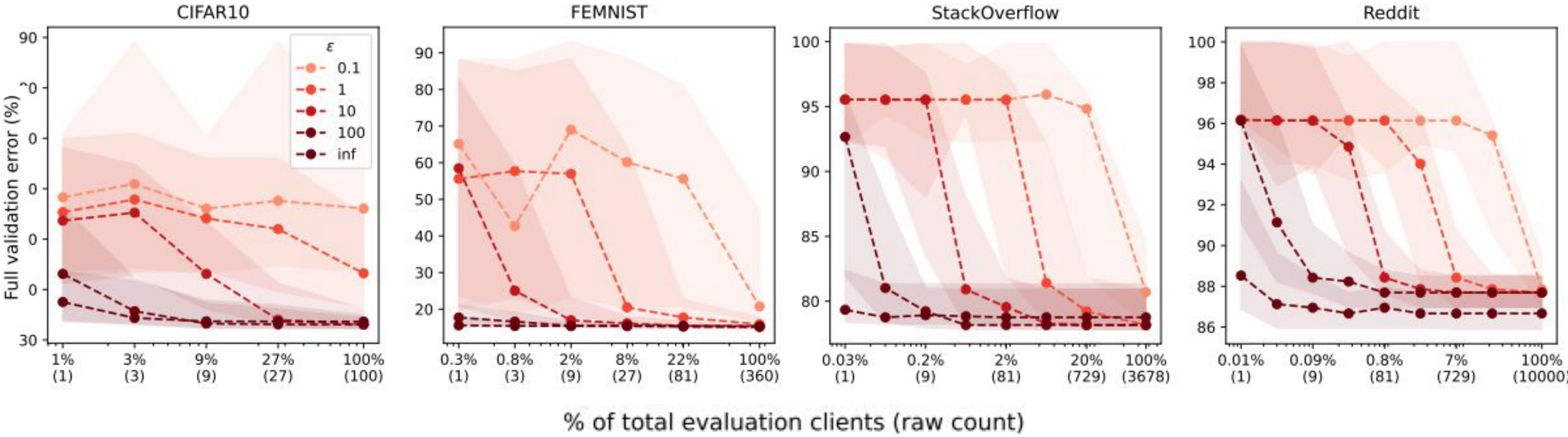
Systems heterogeneity can be catastrophic when the clients' evaluations are sufficiently heterogeneous.





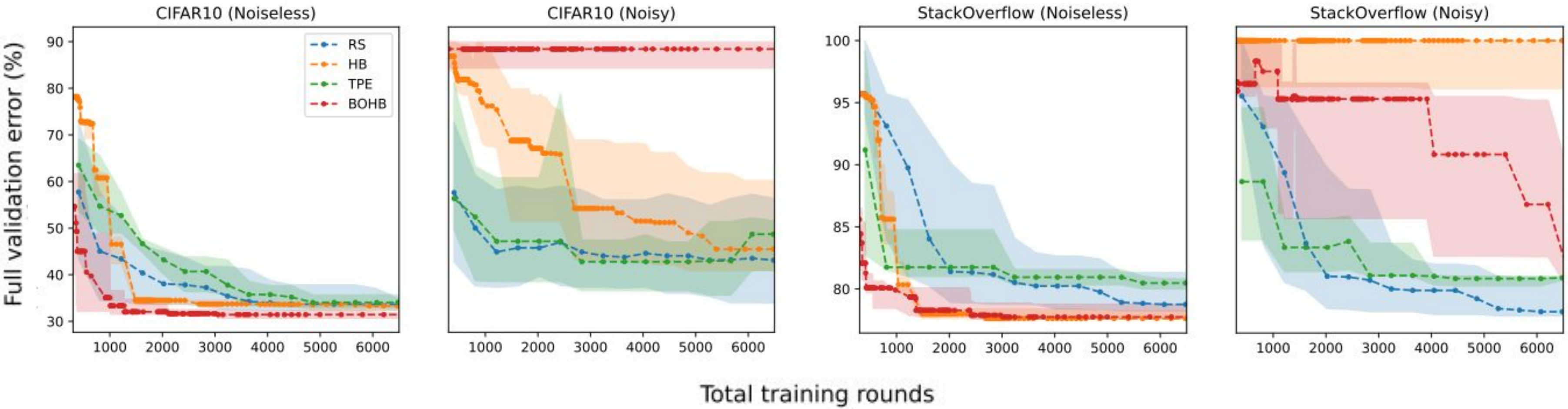
Privacy

DP noise, even under a generous privacy budget, severely deteriorates performance unless a sufficient number of clients are sampled.



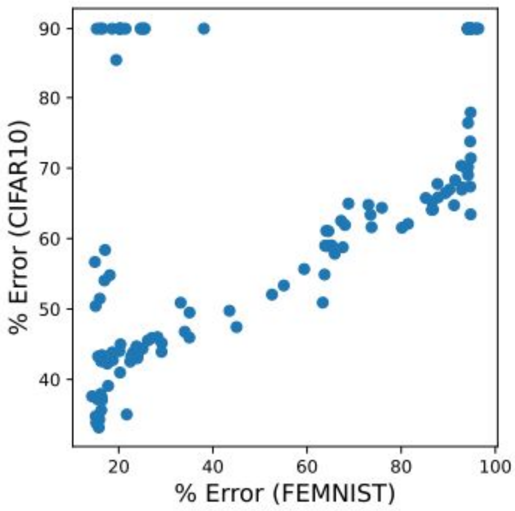
Impact on HP Tuning

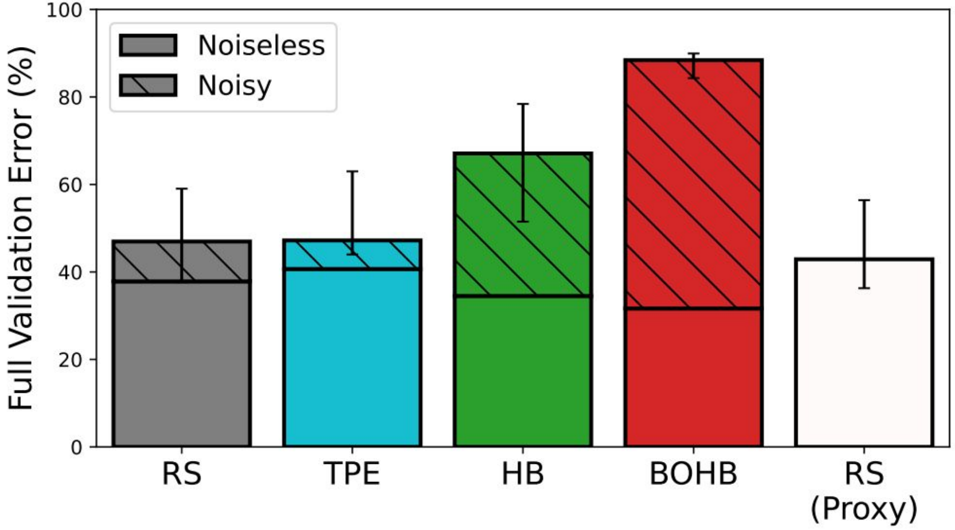
In high-noise regimes, popular methods may perform as poorly as naive baselines.



Proxy Data

In high-noise regimes, a suitable proxy dataset can assist hyperparameter search.





Conclusion

We highlight several best practices for federated HP tuning:

1. Use simple HPO methods.
2. Sample a sufficiently large number of validation clients.
3. Evaluate a representative set of clients.
4. If available, proxy data can be an effective solution.

Conclusion

We highlight several best practices for federated HP tuning:

1. Use simple HPO methods.
2. Sample a sufficiently large number of validation clients.
3. Evaluate a representative set of clients.
4. If available, proxy data can be an effective solution.

Future directions include:

- Improving / tailoring early-stopping methods for DP and FL
- Investigating HPO methods specific for noisy evaluation
- Combining proxy and client data for HPO

Thank you!

Questions?

Contact: kkuo2@andrew.cmu.edu

Website: https://imkevinkuo.github.io

Image sources

Wikipedia: https://en.wikipedia.org/wiki/Federated_learning

Hyperband: https://www.jmlr.org/papers/volume18/16-558/16-558.pdf

FontAwesome: https://fa2png.app/

Our arXiv: https://arxiv.org/abs/2212.08930