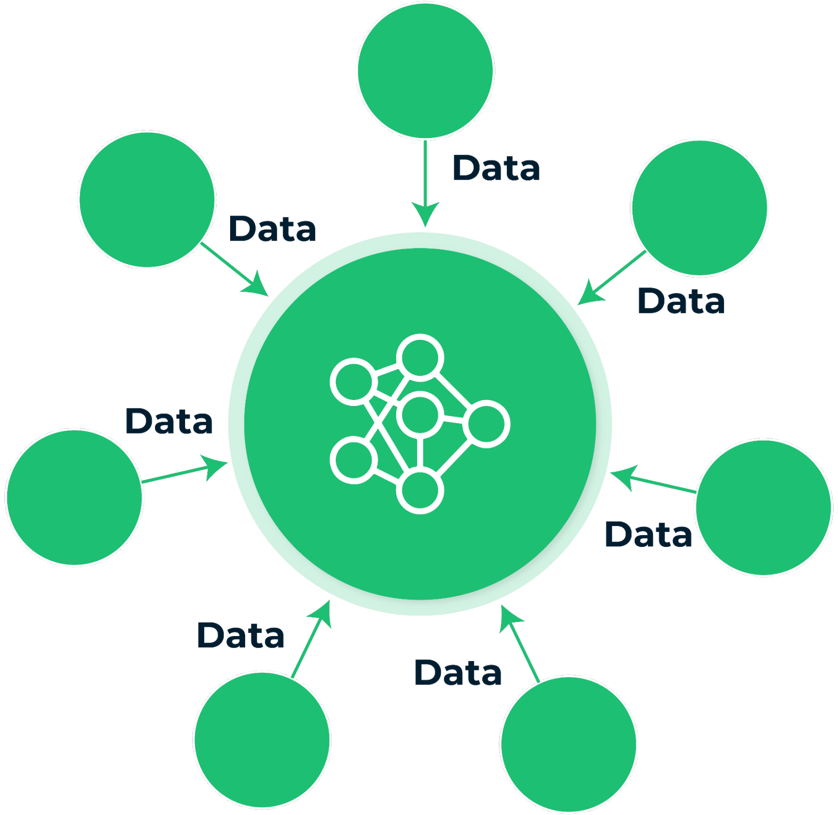
FEDERATED LEARNING
A PARADIGM SHIFT FOR SECURE AND PRIVATE DATA ANALYSIS

Dimitris Stripelis
stripeli@isi.edu

Traditional Data Analysis using Machine Learning

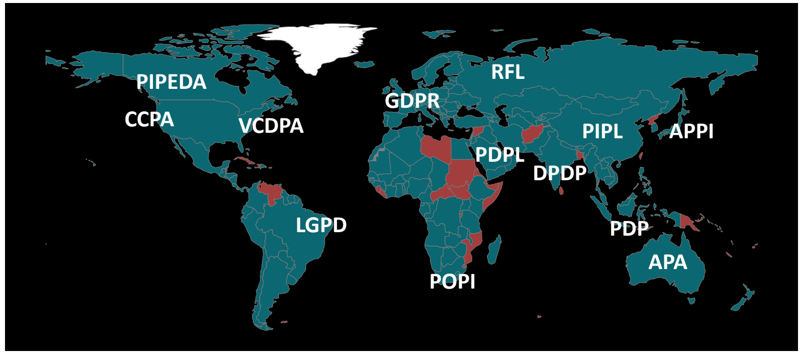
- Data is generated across different data sources.
- Traditional machine learning approaches require data to be aggregated in a centralized location.



Learning Without Data Sharing

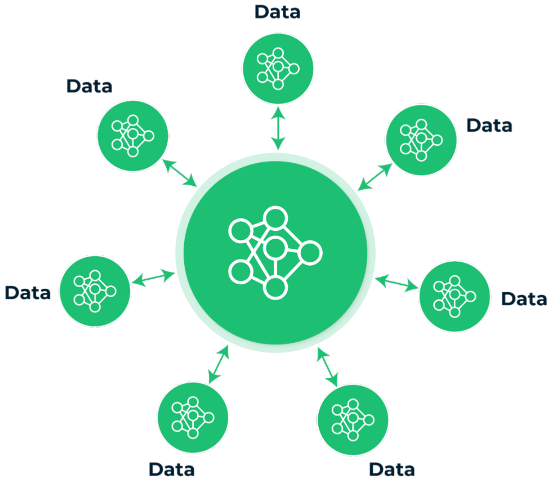
Data Regulation Trends

- **GDPR:** EU General Data Protection Regulation (2018)
- **CCPA:** California Consumer Privacy Act (2020)
- **PIPL:** China Personal Information Protection Law (2021)
- **nFADP:** New Federal Act on Data Protection (2023)
- *many more …*

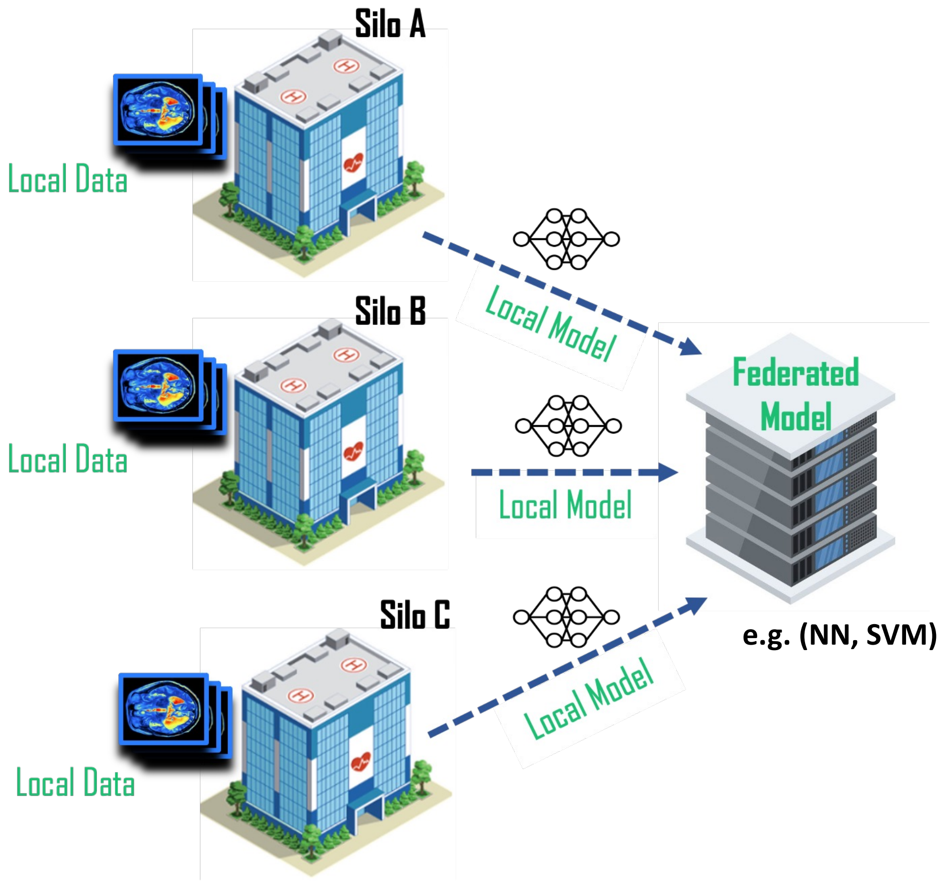


Yes! Federated Learning!

Can we **securely** and **privately learn** machine learning models **from distributed data sources without data sharing?**



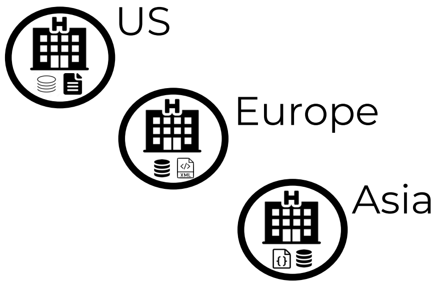
What is Federated Learning?



- Multiple silos/sites (i.e., clients, learners) collaboratively train a ML model.
- Data never leave a site.
- Silos only share locally-trained model parameters.
- Federated Learning can be applied to different ML algorithms.

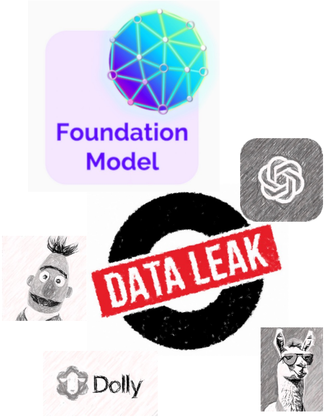
USC INFORMATION SCIENCES INSTITUTE

Why Federated Learning?



Data Fragmentation
Data is not always located at a single, centralized location.

Data Protection & Privacy-Cautious End Users
People concerned about data collection and processing.



Privacy Regulations & AI
Data privacy-enforced GDPR, CCPA, PIPL, nFADP, VCDPA.

Privacy-Preserving AI Personalization
Foundation models trained on all public internet data!

PPDSA National Strategy 2023
PPDSA: Privacy-Preserving Data Sharing and Analytics.

NATIONAL STRATEGY TO ADVANCE PRIVACY-PRESERVING DATA SHARING AND ANALYTICS

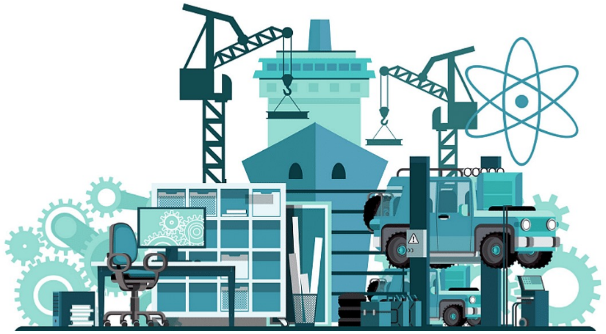
A Report by the

FAST-TRACK ACTION COMMITTEE ON ADVANCING PRIVACY-PRESERVING DATA SHARING AND ANALYTICS
NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE

Application Areas

HealthCare

Industrial Engineering

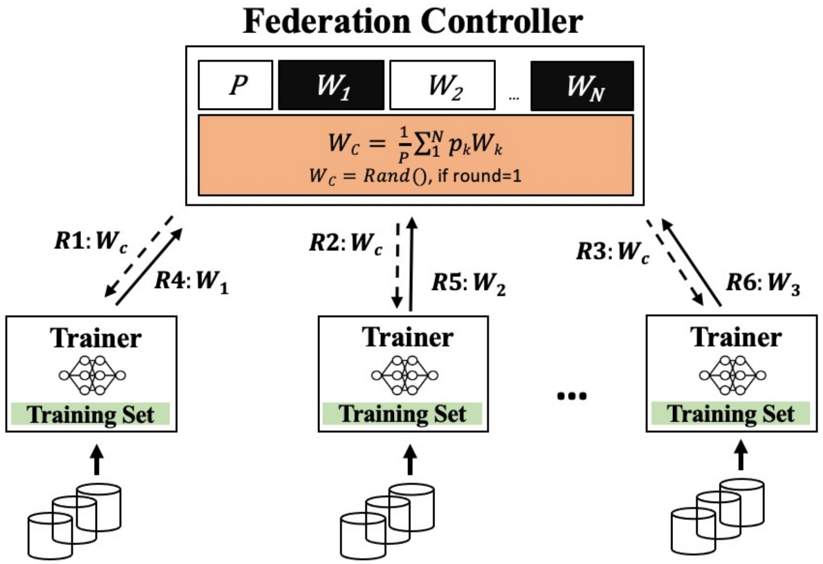


Mobile, IoT, Edge Devices

Drug Discovery

BFSI

Federated Learning
Training



Repeat

For each client (in-parallel):

1) Receive global model (requests R1 - R3)
2) Train global model on local dataset
3) Reply local model (requests R4 - R6)

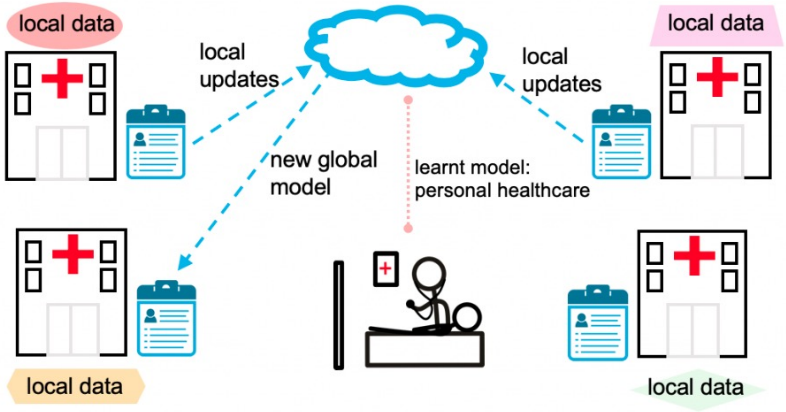
Aggregate local models and compute new global (community) model

For T rounds

The originally proposed aggregation algorithm is called **FedAvg,** and it is a simple weighted average of local models!

McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. "Communication-efficient learning of deep networks from decentralized data." In *Artificial intelligence and statistics*, pp. 1273-1282. PMLR, 2016.

Federated Learning
Environments

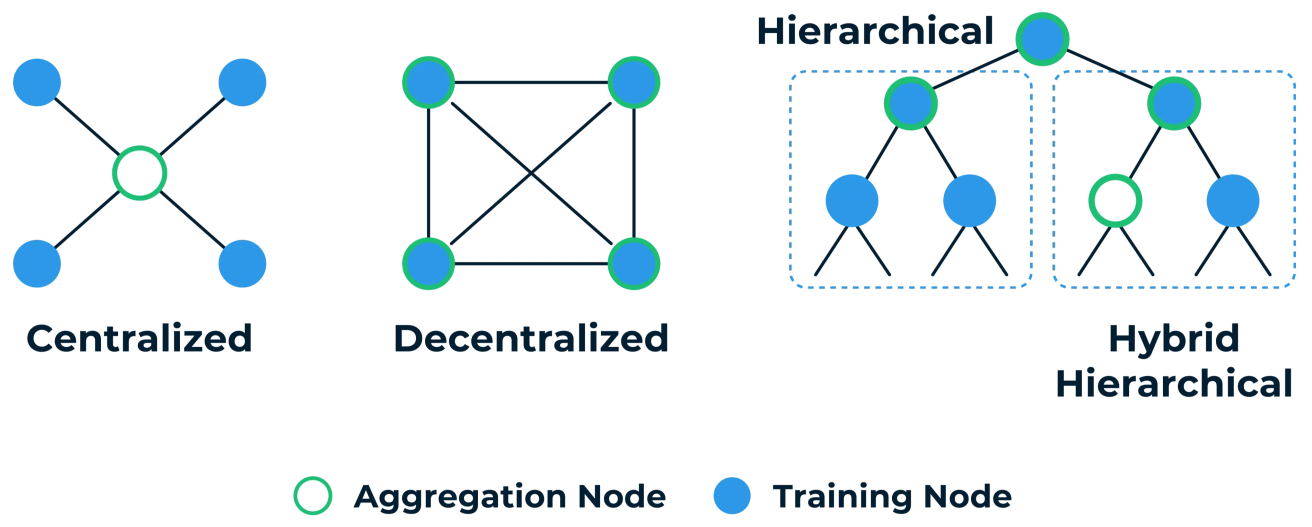
<table>
<tr><th>Specification</th><th>Cross-Silo</th><th>Cross-Device</th></tr>
<tr><td>Learning Setting</td><td>organizations / data centers

</td><td>IoT, mobile devices

</td></tr>
<tr><td>Client Availability, Communication</td><td>highly available
(few client failures)</td><td>often unavailable
(client dropouts)</td></tr>
<tr><td>Number of Clients, Participation</td><td>$O(10), O(100)$, larger data/client, all for community model</td><td>$O(10^5)$ – $O(10^7)$, small data/client, sampling</td></tr>
</table>

Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz et al. "Advances and open problems in federated learning." *Foundations and Trends® in Machine Learning* 14, no. 1–2 (2021): 1-210.

Federated Learning
Topologies

Depending on the **communication protocol** and the **geographical location** of the participating clients different Federated Learning topologies may exist.

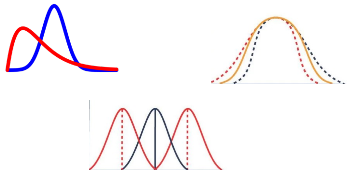
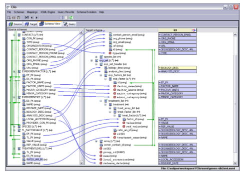
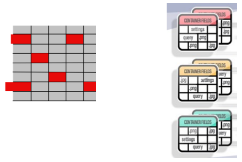


Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas et al. "The future of digital health with federated learning." *NPJ digital medicine* 3, no. 1 (2020): 119.

Federated Learning
Heterogeneities

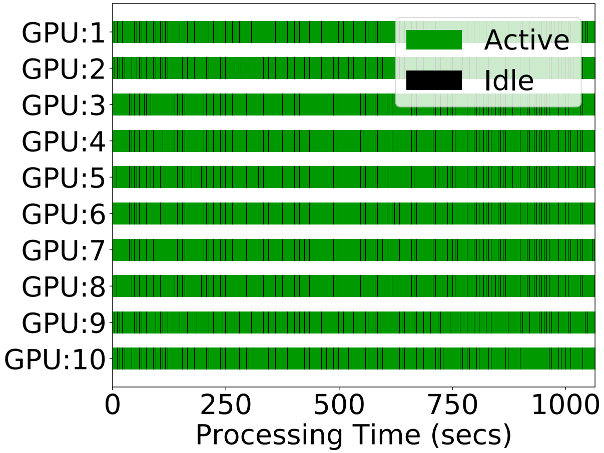
Computational & Statistical Heterogeneity
(model optimization)

Semantic Heterogeneity
(data integration)

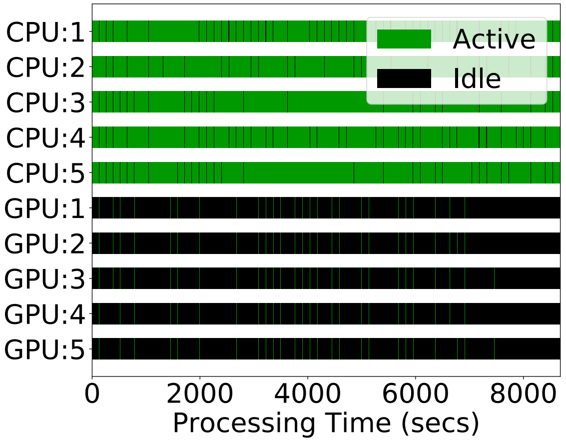
Communication Heterogeneity	Processing Heterogeneity	Statistical Heterogeneity	Data Storage Heterogeneity	Data Schema Heterogeneity	Data Value Heterogeneity
					
Clients may have different **communication capabilities** (e.g., bandwidth).	**Clients** may have **different processing** (e.g., CPU, GPU, TPU) and memory capabilities.	**Data distribution** may not follow the global distribution.	**Data** is stored in **different storage engines** and **formats**.	**Data** may be structured under **different schemata.**	**Data** may have **missing values** and **unnormalized values** referring to **different entities.**

Dimitris Stripelis, "Heterogeneous Federated Learning", PhD Thesis (2023)

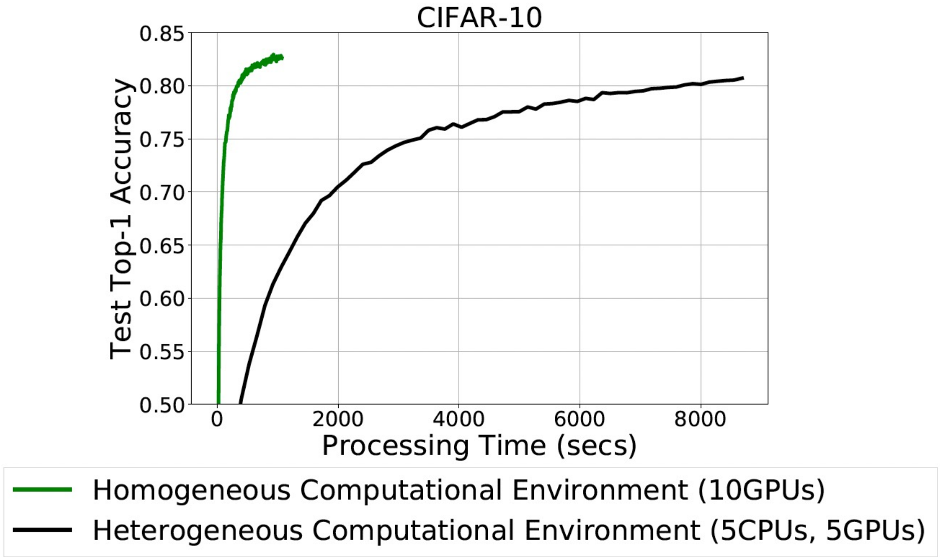
Processing Heterogeneity
Clients with similar processors converge faster!



Homogeneous
Computational Environment
(10 GPUs)



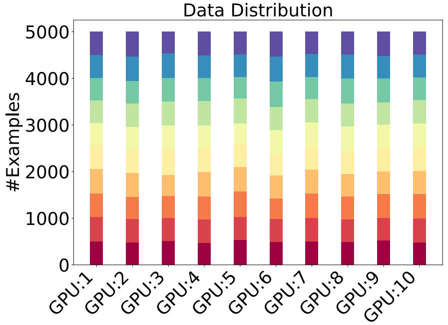
Heterogeneous
Computational Environment
(5CPUs, 5GPUS)



Training example is using FedAvg with **10 clients** and a **simple CNN model** on the **CIFAR-10 dataset.**

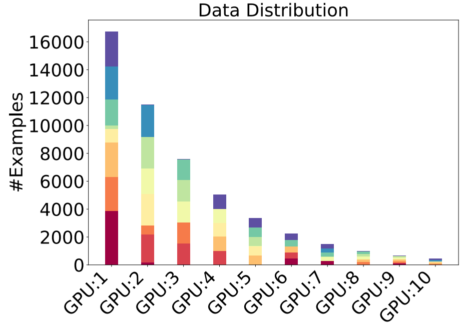
Statistical Heterogeneity
Client data follow different distributions

- Different data amounts (Uniform, Power Law, Skewed)
- Similar (IID) or dissimilar (Non-IID) statistical distributions



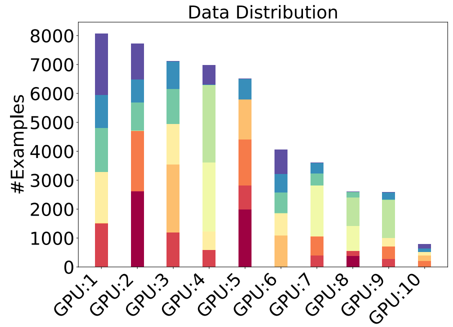
Uniform & IID

(10 target classes per learner)



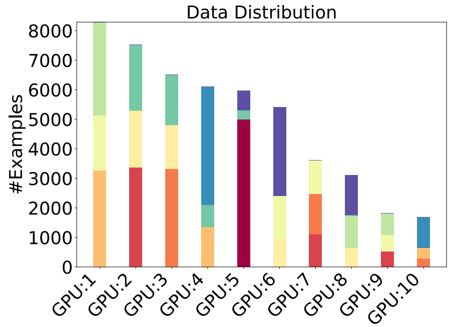
Power Law & Non-IID(5)

(8,7,6,5x7 target classes per learner)



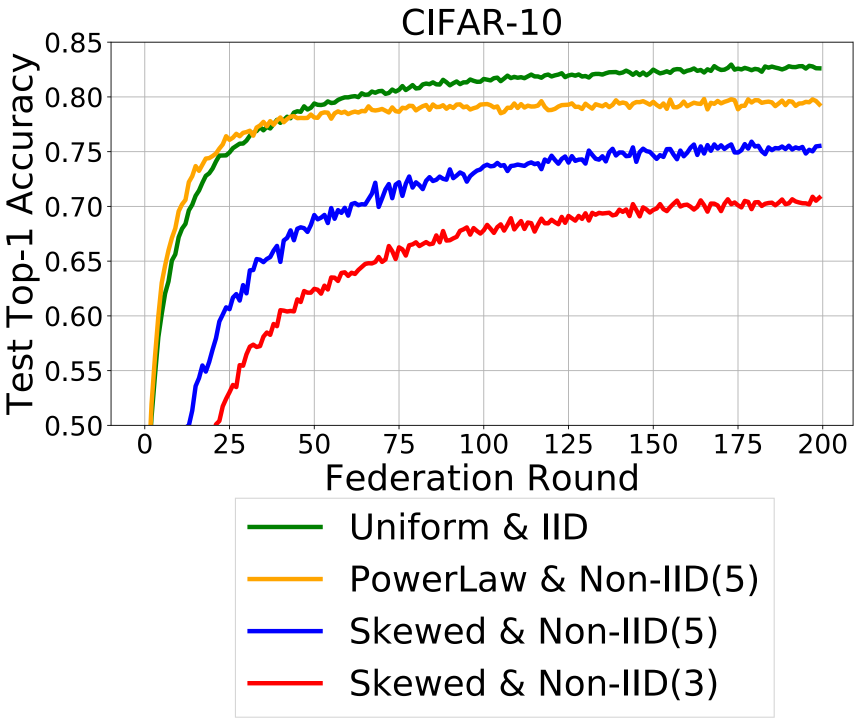
Skewed & Non-IID(5)

(5 target classes out of 10 per learner)



Skewed & Non-IID(3)

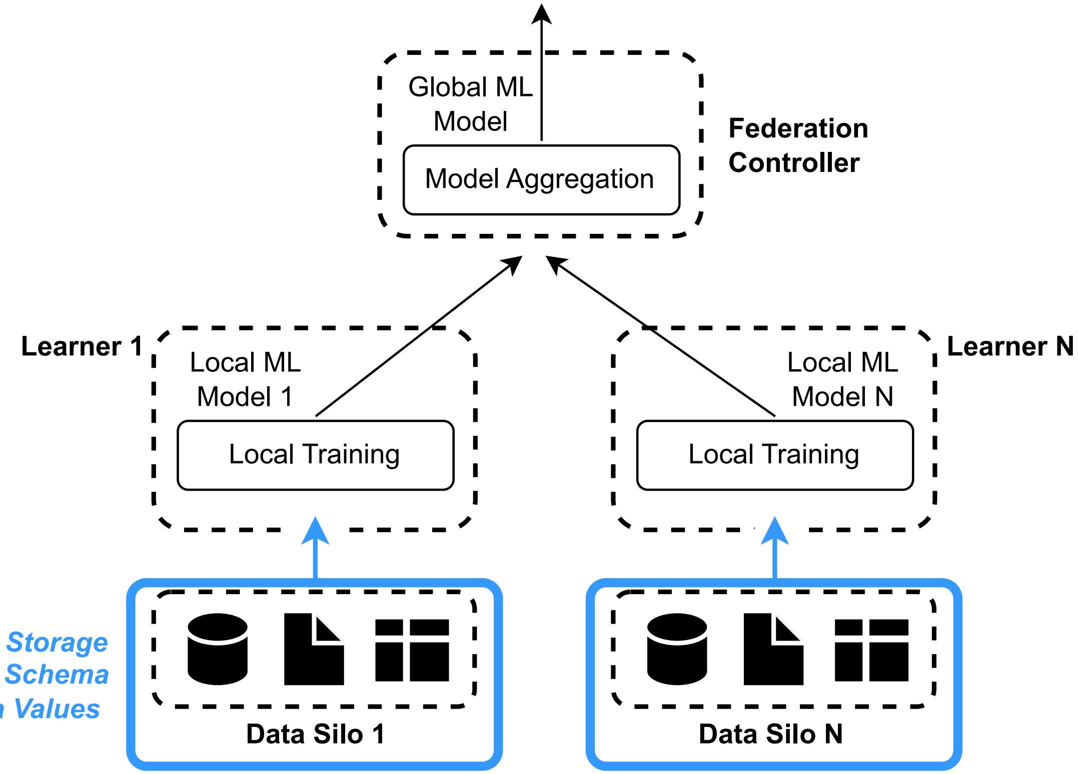
(3 target classes out of 10 per learner)



Training example is using FedAvg with **10 clients** and a simple **CNN model** on **CIFAR-10 dataset.**

Semantic Heterogeneity

Client data do not follow the same semantics

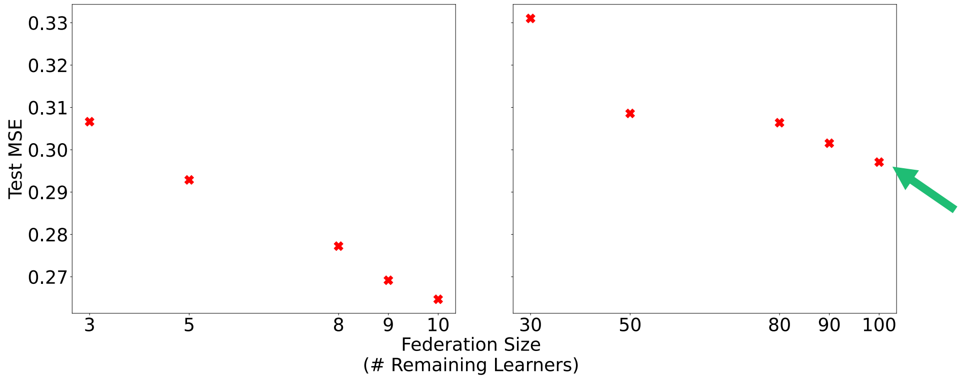


Local Data Storage
Local Data Schema
Local Data Values

If clients' local datasets do not conform to the same semantics, then we might need to discard them.

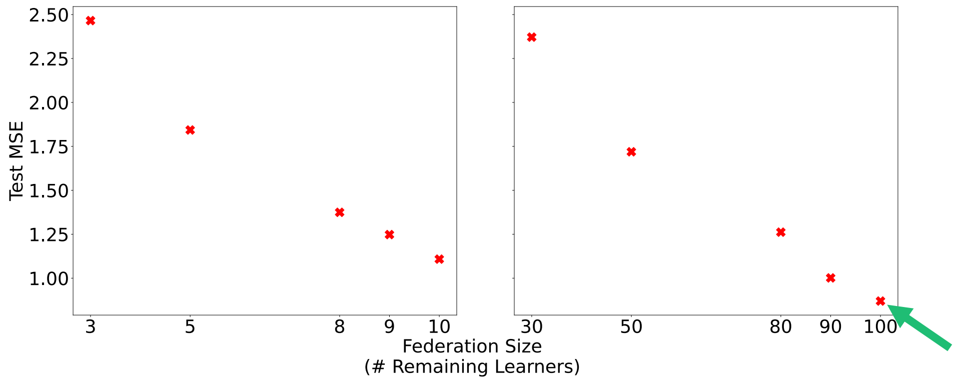
IID Environments

California Housing Dataset: *Predict House Sale Median Price*



Non-IID Environments

California Housing Dataset: *Predict House Sale Median Price*



More Learners **->** More Data **->** *Better Model Performance*

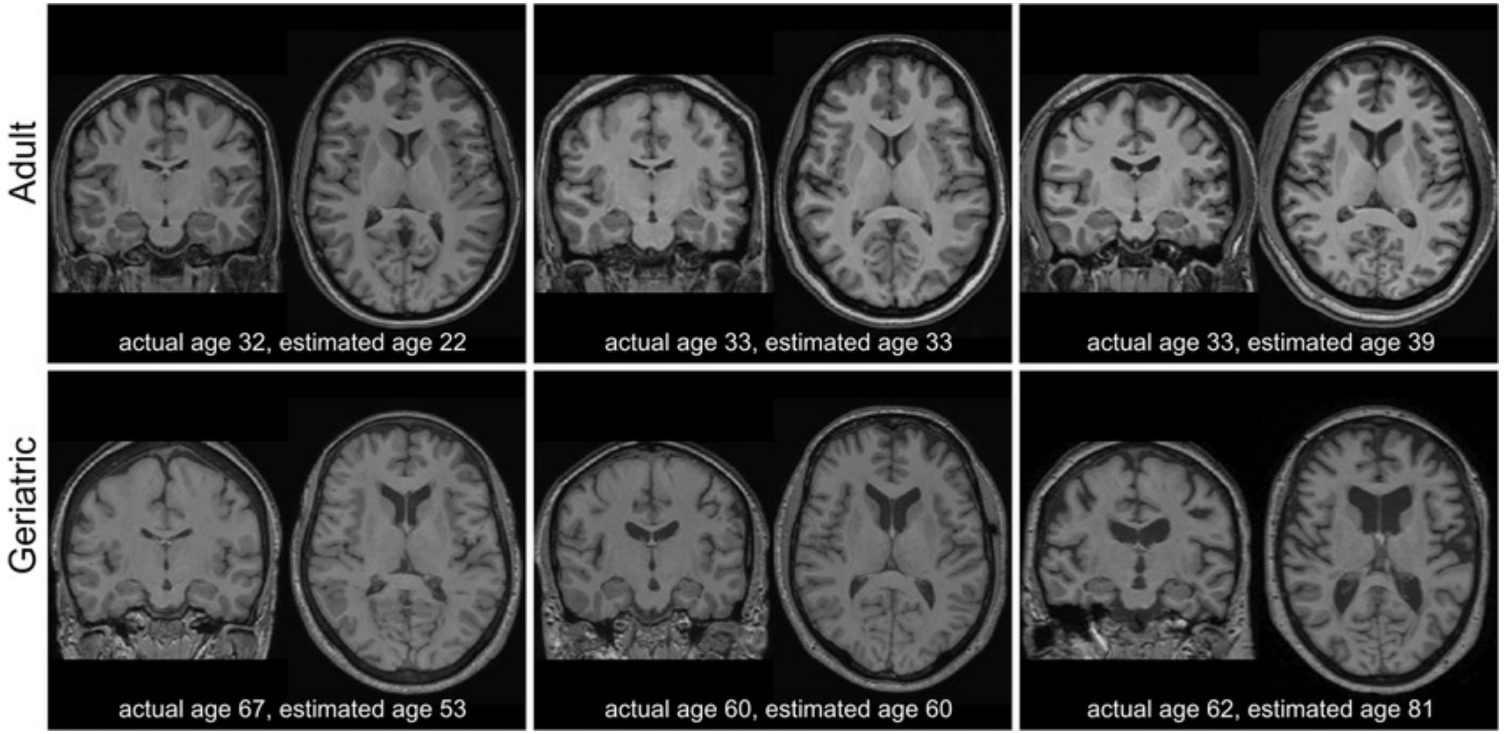
Federated Learning Application
Neuroimaging

How Old Is Your Brain? Ask All The Hospitals

https://viterbischool.usc.edu/news/2023/02/how-old-is-your-brain-ask-all-the-hospitals/ by Julia Cohen, Feb 2023

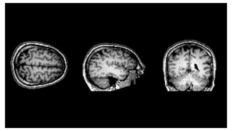
Brain Age Gap Estimate (BrainAGE)
from structural MRI Scans

The brain-age-gap estimate (BrainAGE) quantifies the difference between chronological age and age predicted by applying machine-learning models to neuroimaging data and is considered a biomarker of brain health.

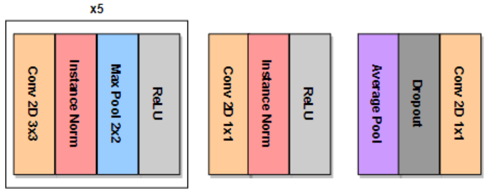


Christman, Seth, Camilo Bermudez, Lingyan Hao, Bennett A. Landman, Brian Boyd, Kimberly Albert, Neil Woodward et al. "Accelerated brain aging predicts impaired cognitive performance and greater disability in geriatric but not midlife adult depression." *Translational Psychiatry* 10, no. 1 (2020): 317.

UK BioBank Federated Environments



Model: 5-CNN
Dataset: 10,000 structural MRI scans

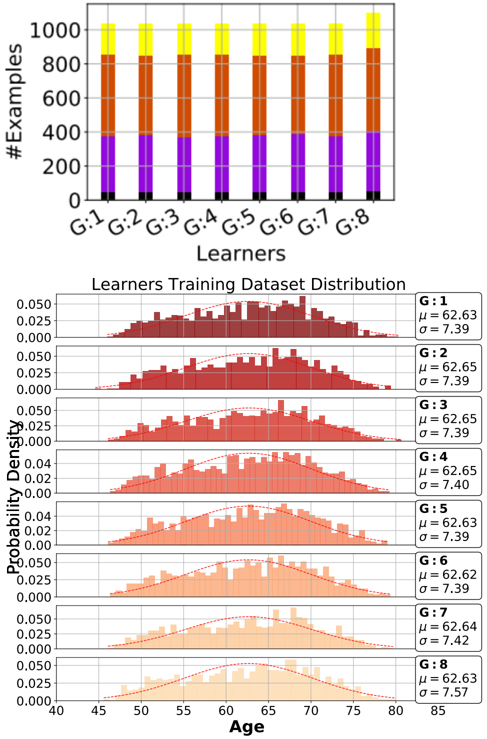


Federated Environments
(with 8 learners/clients)

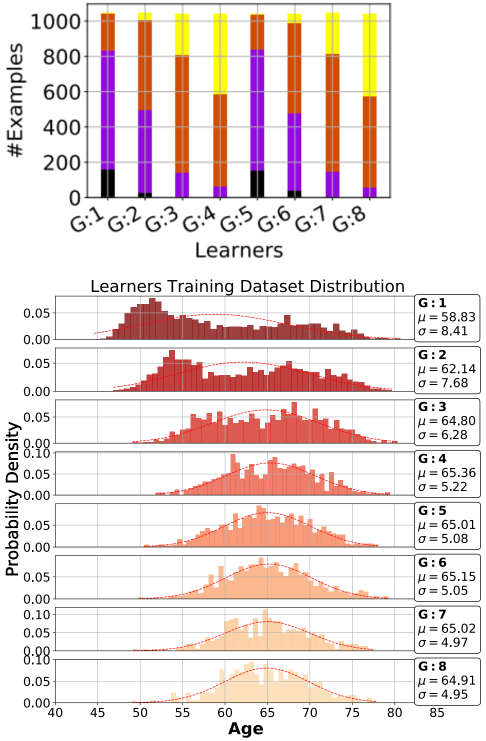
<u>MRI Scans Assignment</u>
Uniform: equal number
Skewed: rightly skewed

<u>MRI Scans Age Distribution</u>
IID: all ages, i.e., (39-80]
Non-IID: ages subset

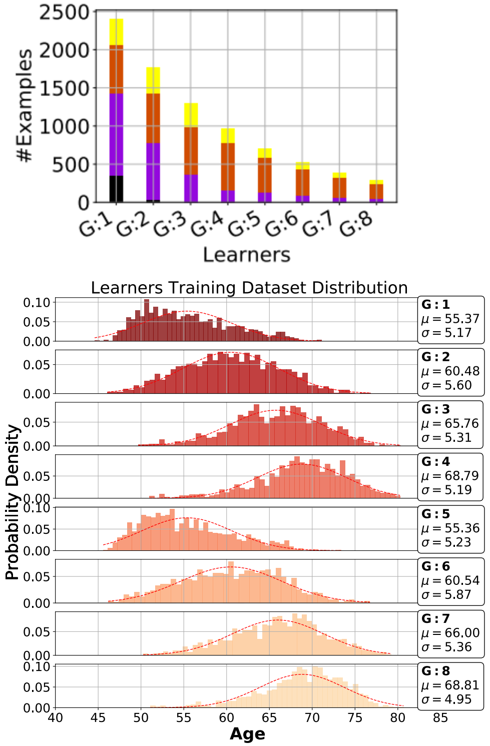
Uniform & IID



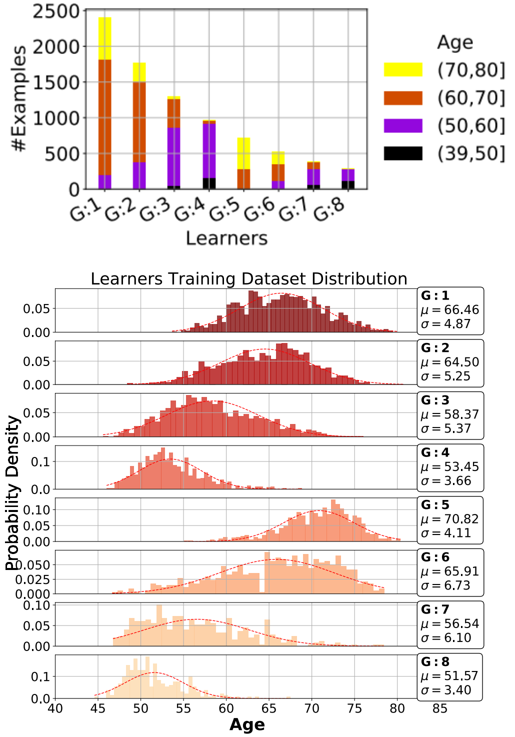
Uniform & Non-IID



Skewed & IID

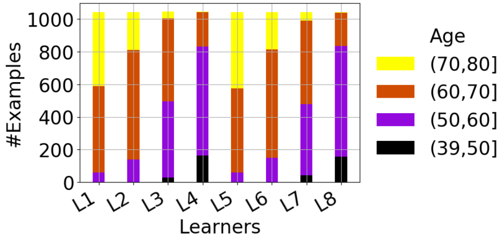


Skewed & Non-IID

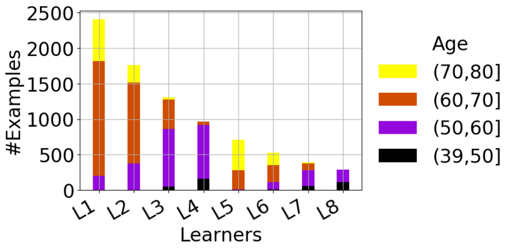


Centralized vs. Federated Learning
(BrainAGE)

Representative UKBB MRI Scans Distributions

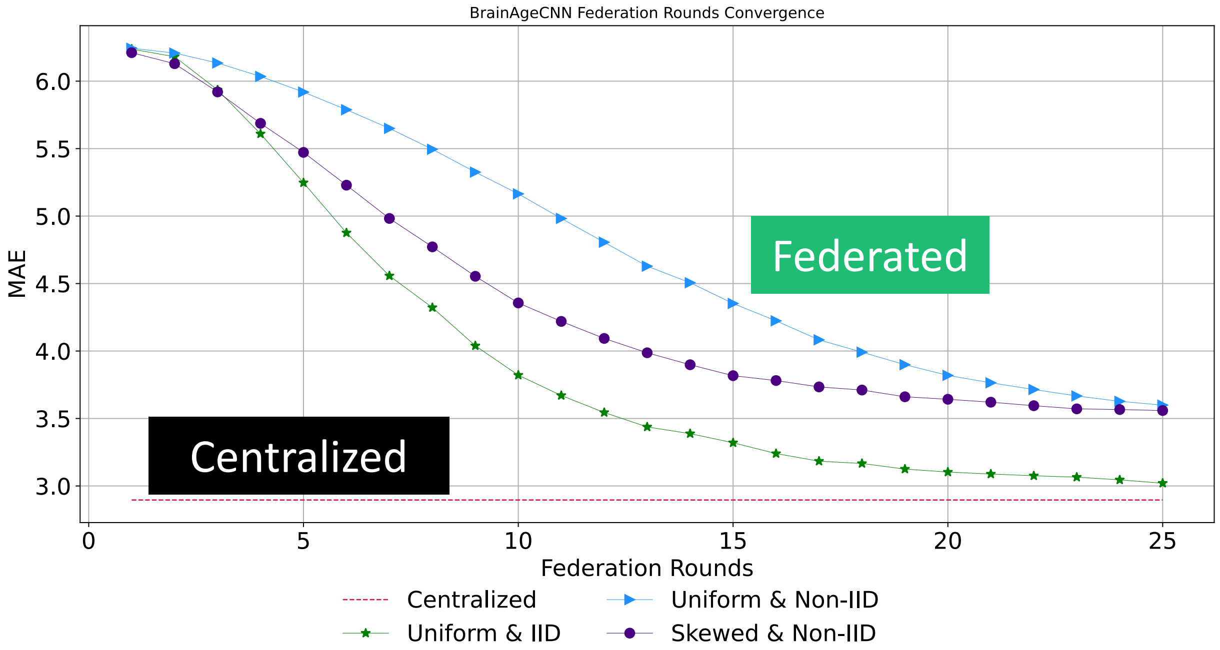


Uniform Non-IID



Skewed & Non-IID

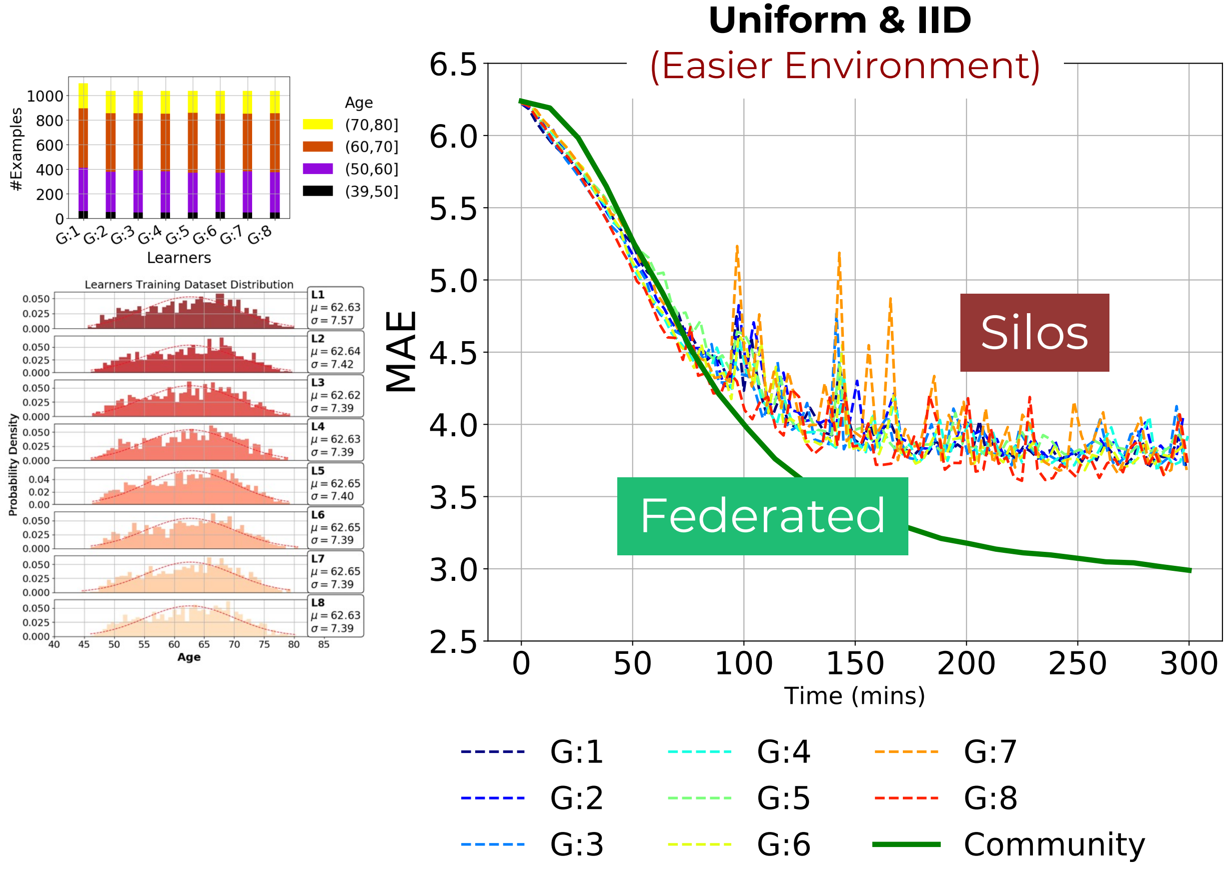
		MSE	RMSE	MAE
Centralized Model		12.885 ± 0.021	3.589 ± 0.003	2.895 ± 0.006
Federated Model				
Data Distribution	**Policy**			
Uniform & IID	SyncFedAvg	13.749 ± 0.138	3.707 ± 0.018	2.995 ± 0.018
Uniform & Non-IID	SyncFedAvg	19.853 ± 1.347	4.453 ± 0.151	3.625 ± 0.135
Skewed & Non-IID	SyncFedAvg	19.148 ± 0.086	4.376 ± 0.009	3.553 ± 0.003

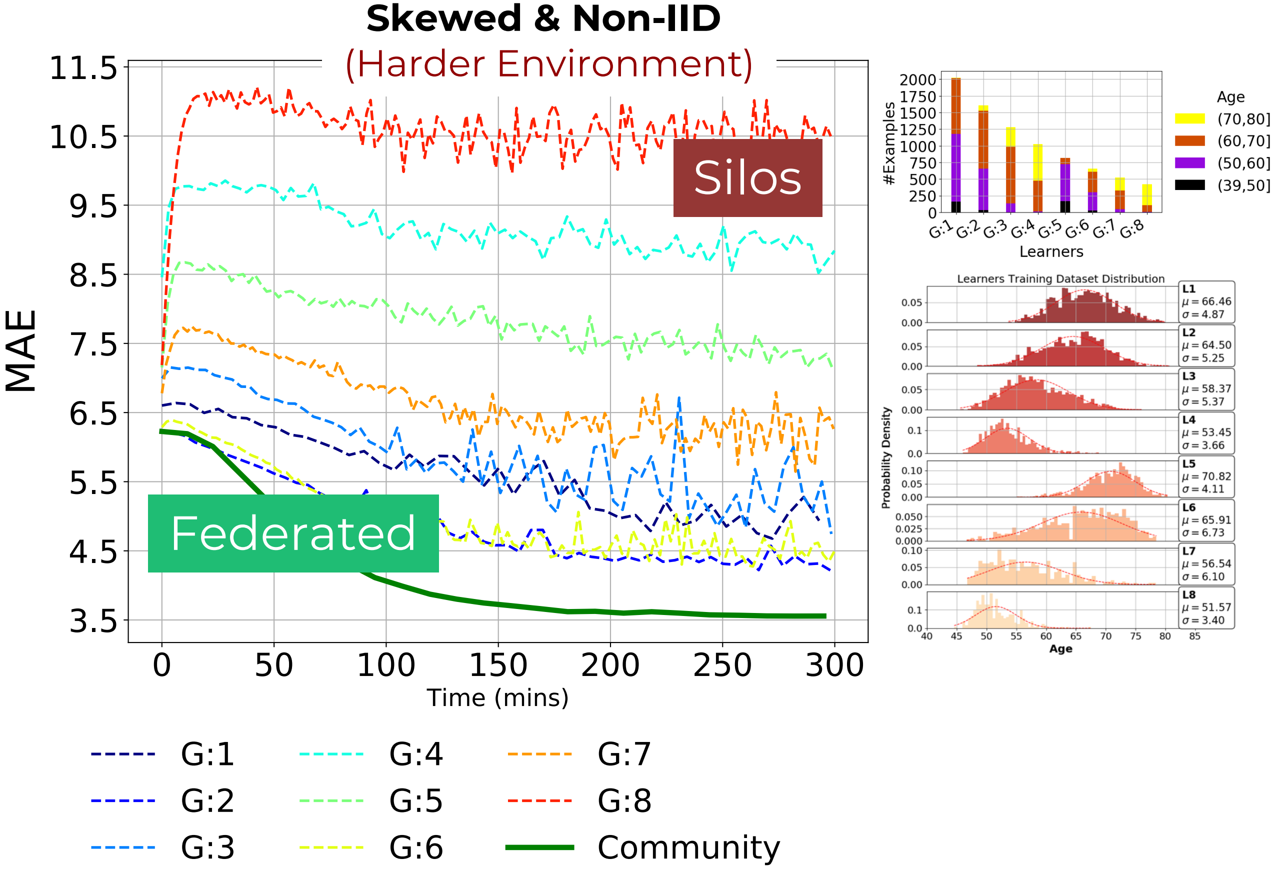


Stripelis, Ambite, Lam, Thompson. Scaling neuroscience research using federated learning. In IEEE International Symposium on Biomedical Imaging (ISBI), Nice, France, 2021.

USC INFORMATION SCIENCES INSTITUTE

Federated Learning Outperforms any Silo (BrainAGE)





Federated Model >> Siloed Models

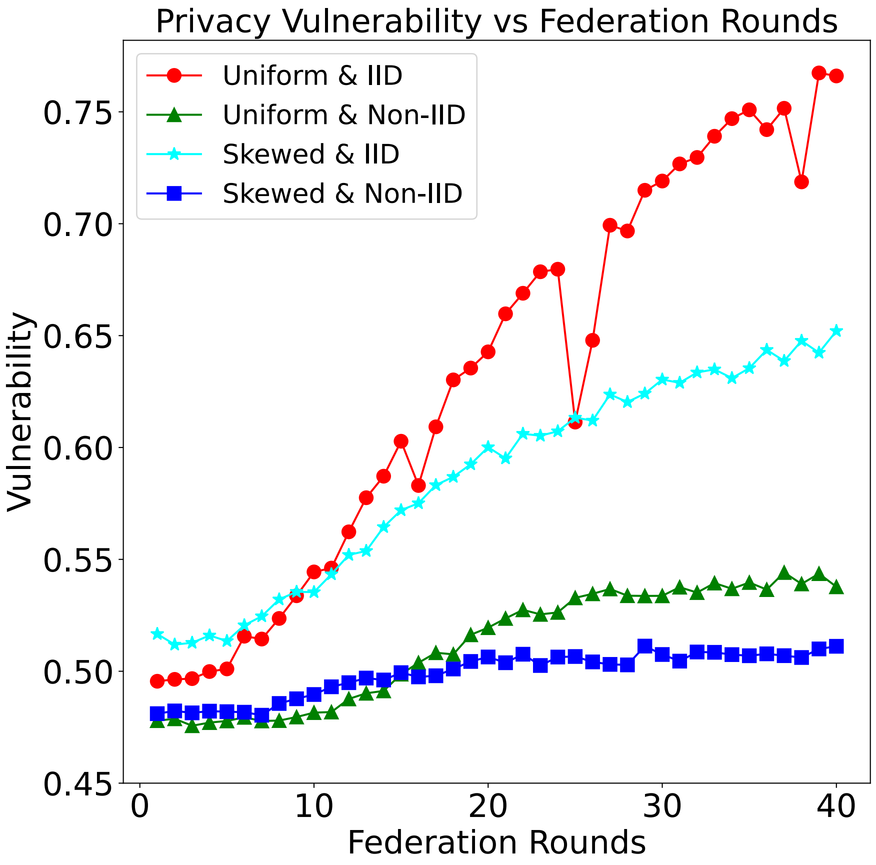
Stripelis, Thompson, Ambite. Semi-Synchronous Federated Learning for Energy-Efficient Training and Accelerated Convergence in Cross-Silo Settings. ACM TIST, 2022 (In Press).

USC INFORMATION SCIENCES INSTITUTE

Federated Learning is ** NOT ** Secure and Private Out-of-The-Box !!!

Unprotected Federated Models are Vulnerable to Information Leakage!

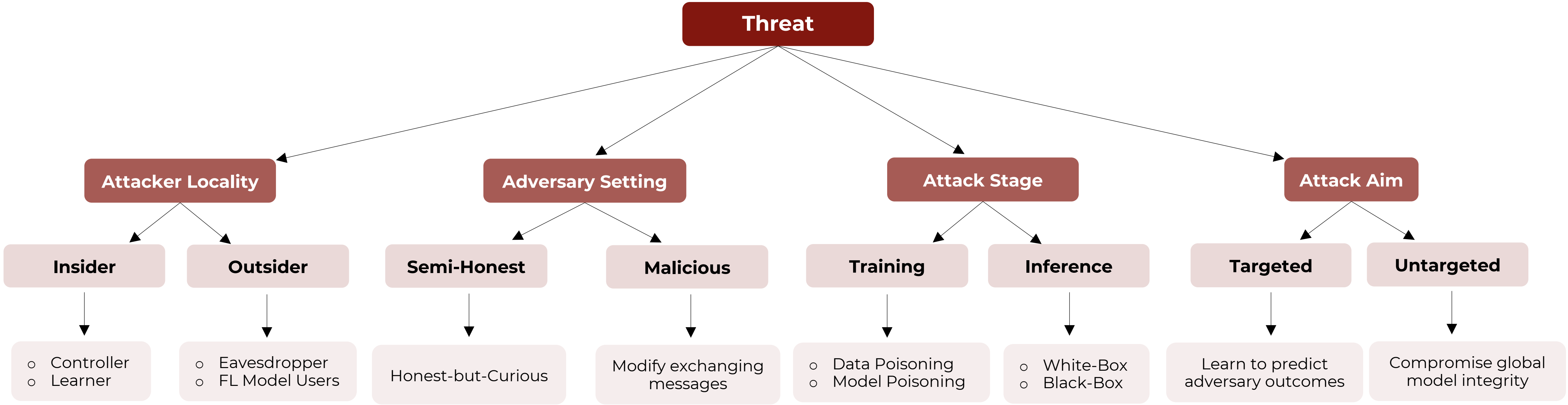
- Privacy vulnerability increases with federation rounds.
- Vulnerability is measured as the average accuracy of distinguishing train samples vs unseen samples across learners.



Stripelis, Dimitris, Umang Gupta, Hamza Saleem, Nikhil Dhinagar, Tanmay Ghai, Rafael Sanchez, Chrysovalantis Anastasiou et al. "Secure Federated Learning for Neuroimaging." *arXiv preprint arXiv:2205.05249* (2022).

Secure & Private Federated Learning

Federated Learning Threats Taxonomy



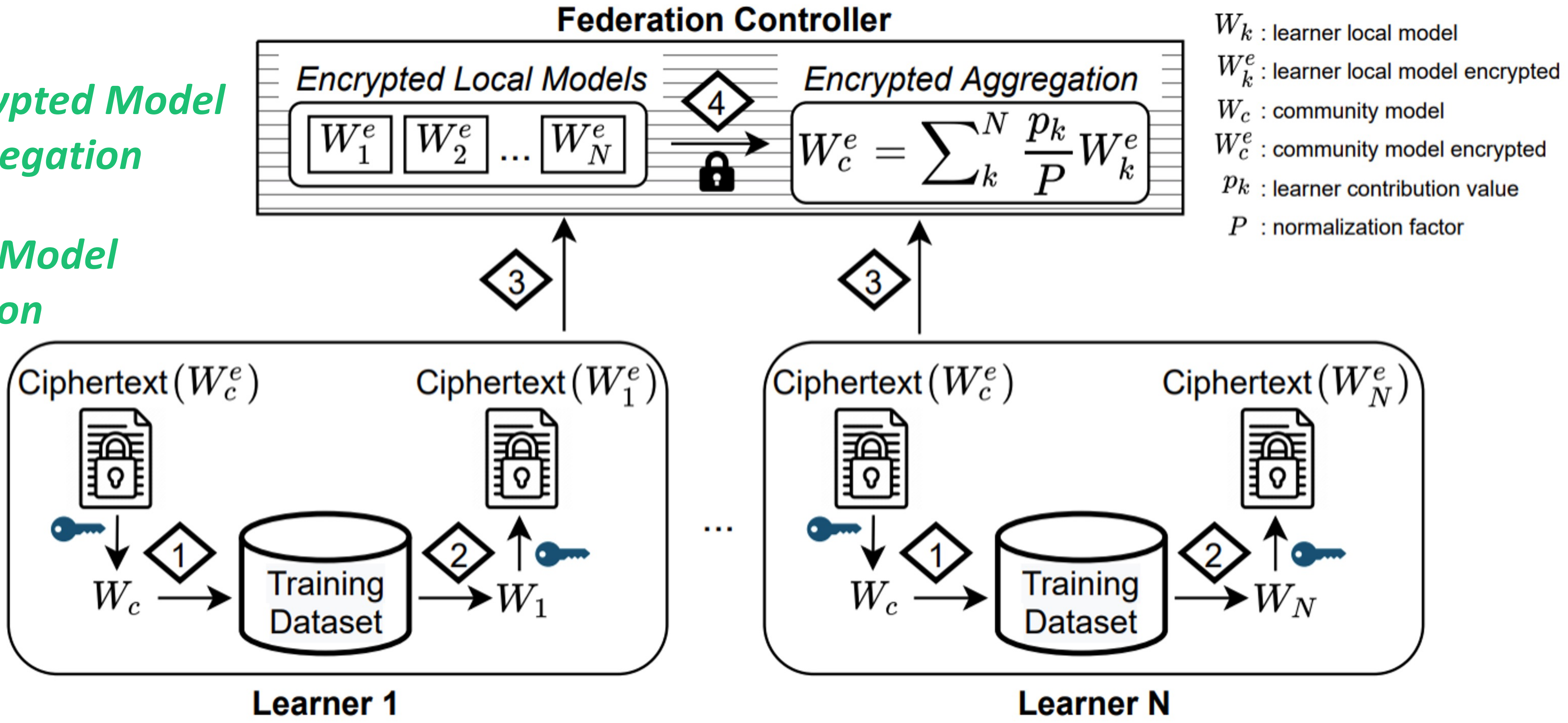
Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q. and Philip, S.Y., 2022. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems.*

Secure Federated Learning
w/ Fully Homomorphic Encryption (FHE)

Encrypted Model Aggregation

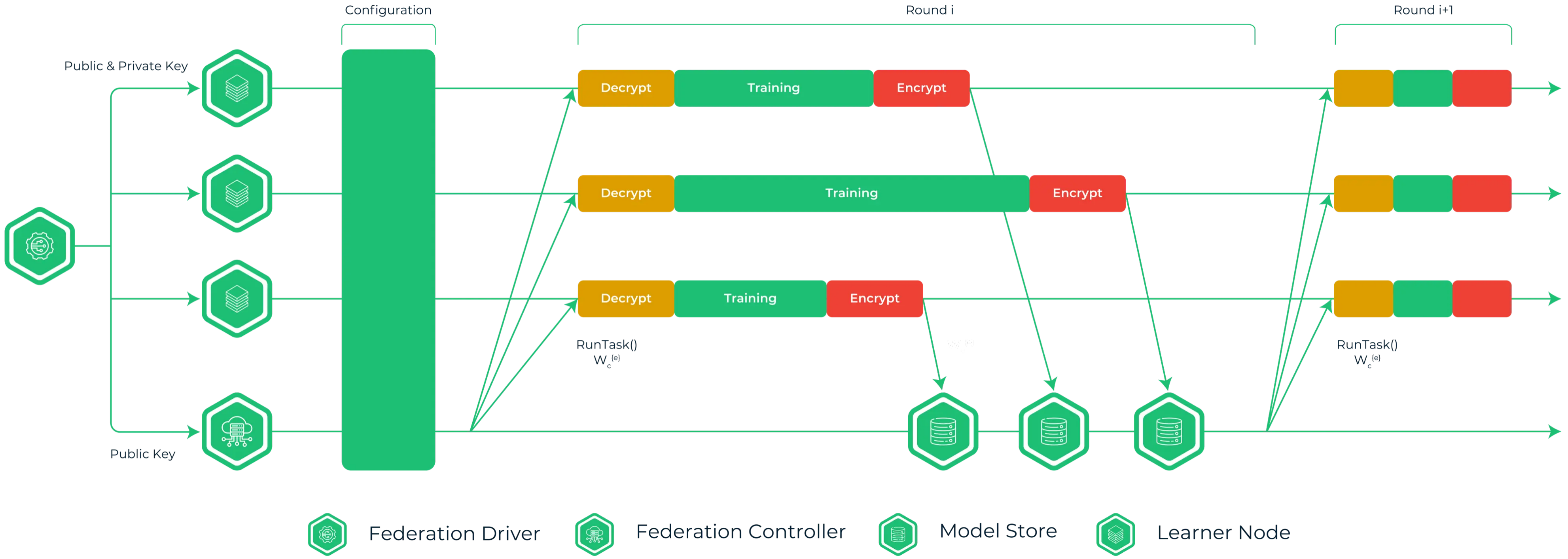
Encrypted Model Transmission

Data Never Leaves Sites



Stripelis, Dimitris, Hamza Saleem, Tanmay Ghai, Nikhil Dhinagar, Umang Gupta, Chrysovalantis Anastasiou, Greg Ver Steeg et al. "Secure neuroimaging analysis using federated learning with homomorphic encryption." In *17th International Symposium on Medical Information Processing and Analysis*, vol. 12088, pp. 351-359. SPIE, 2021.

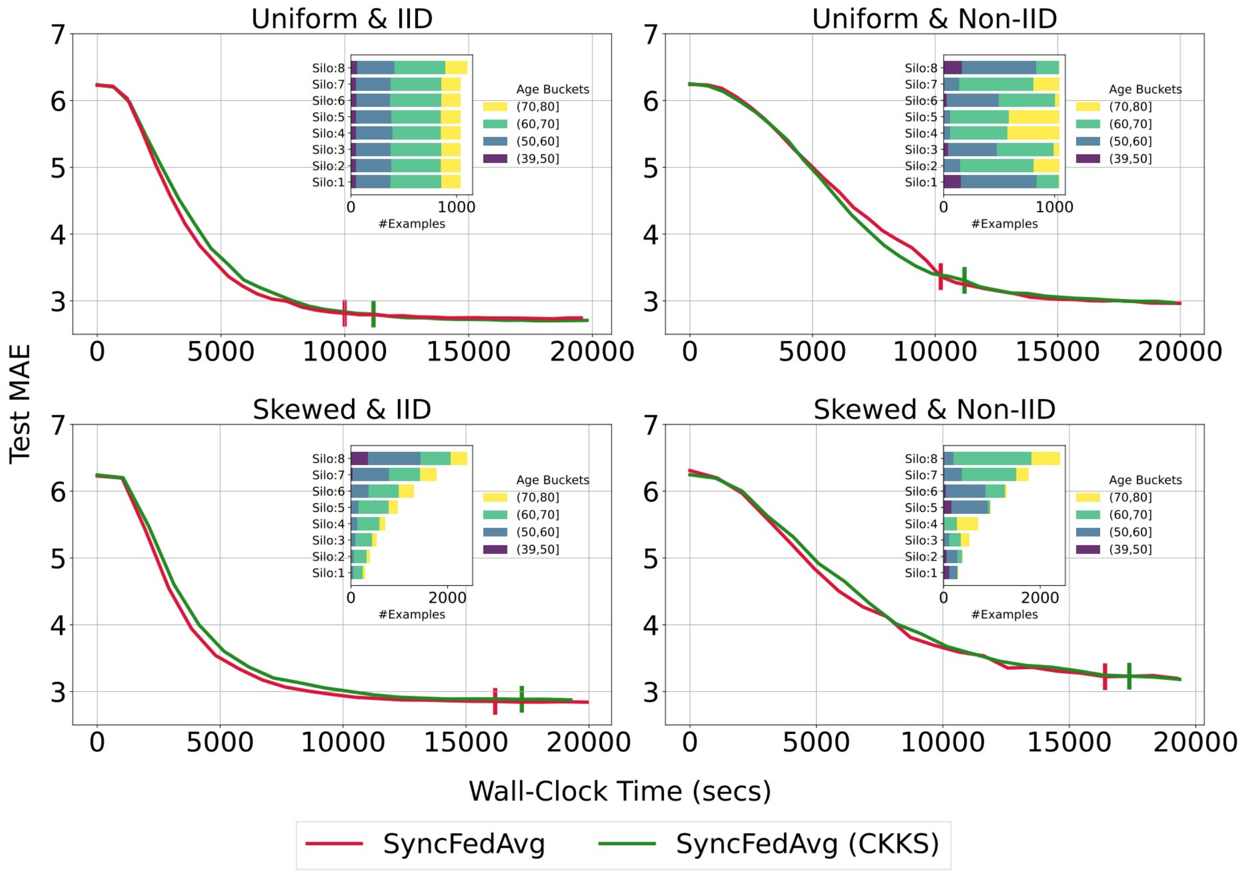
FHE Workflow using the
the Metis Federated Learning (MetisFL) Open-Source Framework



- The **Federation Driver** *generates* the initial public & private key pair.
- The **Federation Controller** *delegates* the training tasks to the participating learners/clients.
- The **Learner Nodes** perform the assigned training task and send their local models *encrypted*.
- The **Federation Controller** *aggregates* the local models within an *encrypted space*.

Secure Federated Learning
(BrainAGE)

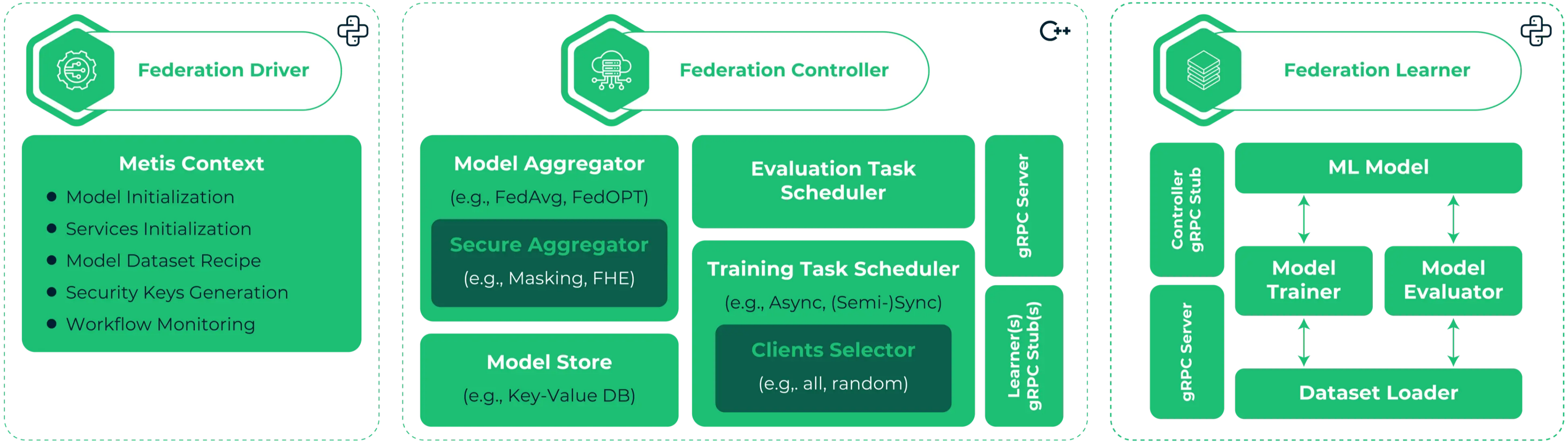
- Federated Training with Fully Homomorphic Encryption (FHE) can learn a model with **same learning performance.**
- FHE incurs only **7% of additional execution latency.**



Stripelis, Dimitris, Umang Gupta, Hamza Saleem, Nikhil Dhinagar, Tanmay Ghai, Rafael Sanchez, Chrysovalantis Anastasiou, Jose Luis Ambite, et al. "Secure Federated Learning for Neuroimaging." *arXiv preprint arXiv:2205.05249* (2022).

Putting Everything Together

The Metis Federated Learning (MetisFL) Open-Source Framework



Privacy-Enhancing Technologies
PRIZE CHALLENGE

MetisFL was accepted as a blue participant at the US/UK PETs 2022 Challenge.

System Design

- ✓ Portable *Containerized* Design & Execution.
- ✓ Extensible & Modular Design Supporting *Various ML/NN Engines*.

Model Training

- ✓ Secure Training through *Fully Homomorphic Encryption (FHE).*
- ✓ Private Training through *Differential-Privacy.*
- ✓ Efficient Training through *Scalable Model Aggregation.*
- ✓ Accelerated Training through *Federated Training Protocols.*

MetisFL was developed at USC during my PhD studies!

A Thriving Open-Source Ecosystem

METIS
Secure, Scalable and Efficient Federated Learning Workflows.
License 3-Clause-BSD

FATE
Federated AI Technology Enabler
License Apache

FedML
The Community Building Open and Collaborative AI Anywhere at Any Scale.
License Apache

Flower: A Friendly Federated Learning Framework
License Apache

PySyft
A Software Stack for Secure & Private Data Science in Python.
License Apache

NVIDIA
NVFlare. A Domain-Agnostic, Open-Source, Extensible SDK. Adapt Existing ML/DL Workflows to a Federated Paradigm.
License Apache

FedScale
Scalable and Extensible Open Source Federated Learning Engine and Benchmark.
License Apache

Rosetta
Privacy-Preserving Technologies: Cryptography, Federated Learning and Trusted Execution Environment.
License LGPLv3

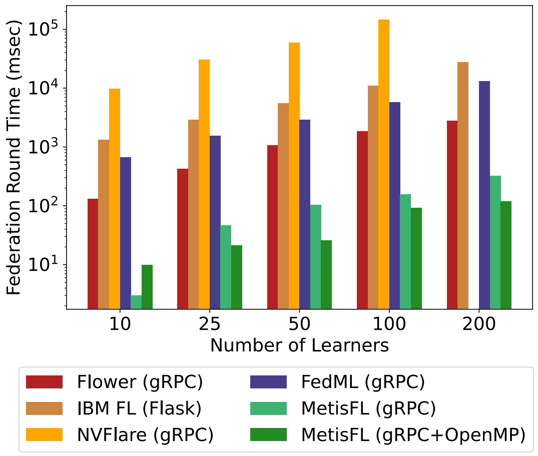
IBM Federated Learning in an Enterprise Environment.
License Custom

Substra
Substra Main Usage is in Production Environments. Used by Hospitals and Biotech Companies.
License Apache

OPENFL
A Flexible, Extensible and Easily Learnable Federated Learning Tool for Data Scientists.
License MIT

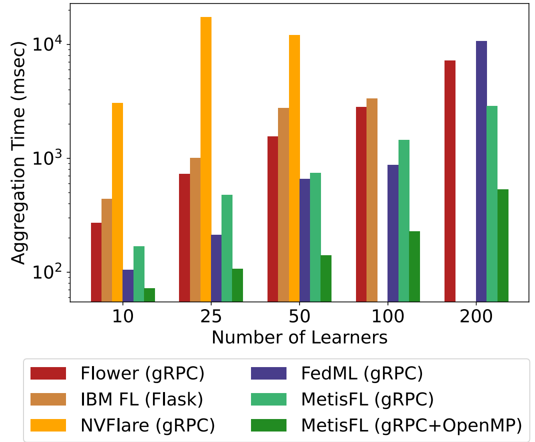
SECRETFLOW 隐语
A Unified Framework for Privacy-Preserving Data Intelligence and Machine Learning.
License Apache

FEDLAB
A Flexible Federated Learning Framework
License Apache



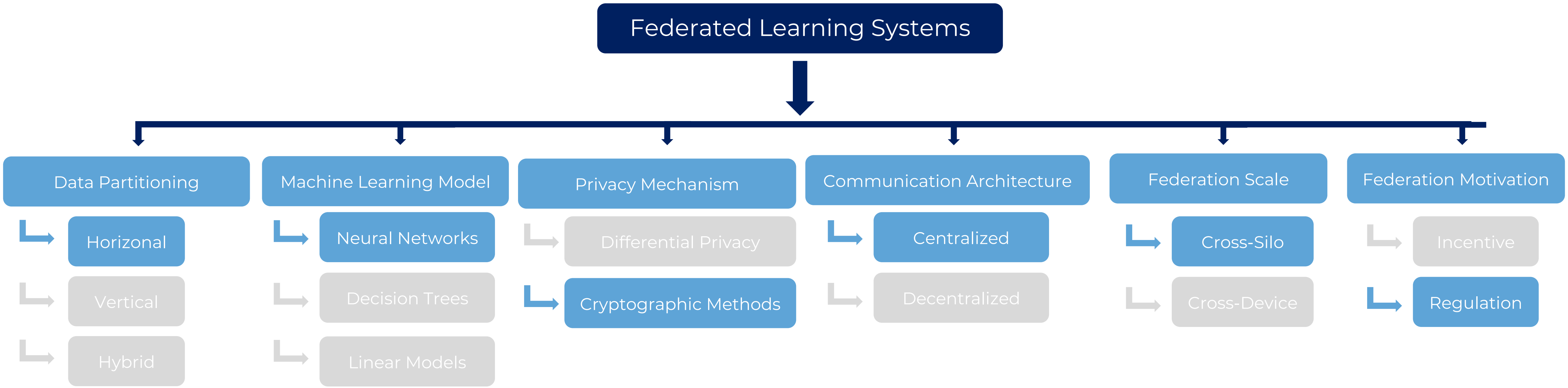
Federation Round Time
1M Parameters

Aggregation Time
10M Parameters



Towards an End-to-End Federated Learning Systems Benchmark
(Work in Progress)

Federated Learning Systems Overview



Discussed Topics.

Not Discussed Topics.

Li, Qinbin, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. "A survey on federated learning systems: Vision, hype and reality for data privacy and protection." *IEEE Transactions on Knowledge and Data Engineering* (2021).

USC INFORMATION SCIENCES INSTITUTE

Future Federated Learning
Community Directions

Raise Data Privacy and Protection **Awareness**!

Define Federated Learning Algorithms and Systems **Standards**!

Democratize Federated Learning!

Make Federated Learning the **De Facto** Distributed AI Approach!

Acknowledgements

Jose-Luis Ambite, PhD
Research Team Leader, ISI
Associate Res. Prof., CS

Muhammad Naveed, PhD
Assistant Prof., CS

Srivatsan Ravi, PhD
Research Scientist, ISI
Assistant Res. Prof., CS

Greg Ver Steeg, PhD
Research Team Leader, ISI
Associate Res. Prof., CS

Paul Thompson, PhD
Professor, Neurology, Psychiatry, Radiology, Ophthalmology, and Engineering

Hamza Saleem, BS
PhD Student,
Computer Science

Tanmay Ghai, BS
MS Student,
Computer Science

Umang Gupta, MS
PhD Student,
Computer Science

Nikhil Dhinagar, PhD
Analyst, Imaging Genetics Center

+ **Armaghan Asghar,** MS CS; **Rafa Sanchez,** MS CS; **Joel Mathew**, MS CS; **Oneeb Khan**, MS CS; **Tamoghna Chattopadhyay**, MS CS; **Patrick Toral**, PhD CS ; **Chrysovalantis Anastasiou**, PhD CS

USC INFORMATION SCIENCES INSTITUTE

A Big Applause to the
Open Source Initiative (OSI)