Efficient Federated Learning in Heterogeneous and Corrupted Environments by Secure Performance Weighting

Jose-Luis Ambite, Muhammad Naveed, Paul Thompson,
Dimitri Stripelis, Hamza Saleem, Nikhil Dhinagar
University of Southern California

Team

Jose-Luis Ambite, PhD
Research Team Leader, ISI
Associate Research Professor, CS

Muhammad Naveed, PhD
Assistant Professor, CS

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

Dimitris Stripelis, MS
PhD Student,
Computer Science

Hamza Saleem, BS
PhD Student,
Computer Science

Nikhil Dhinagar, PhD
Analyst,
Imaging Genetics Center

Can we learn from distributed data respecting privacy?

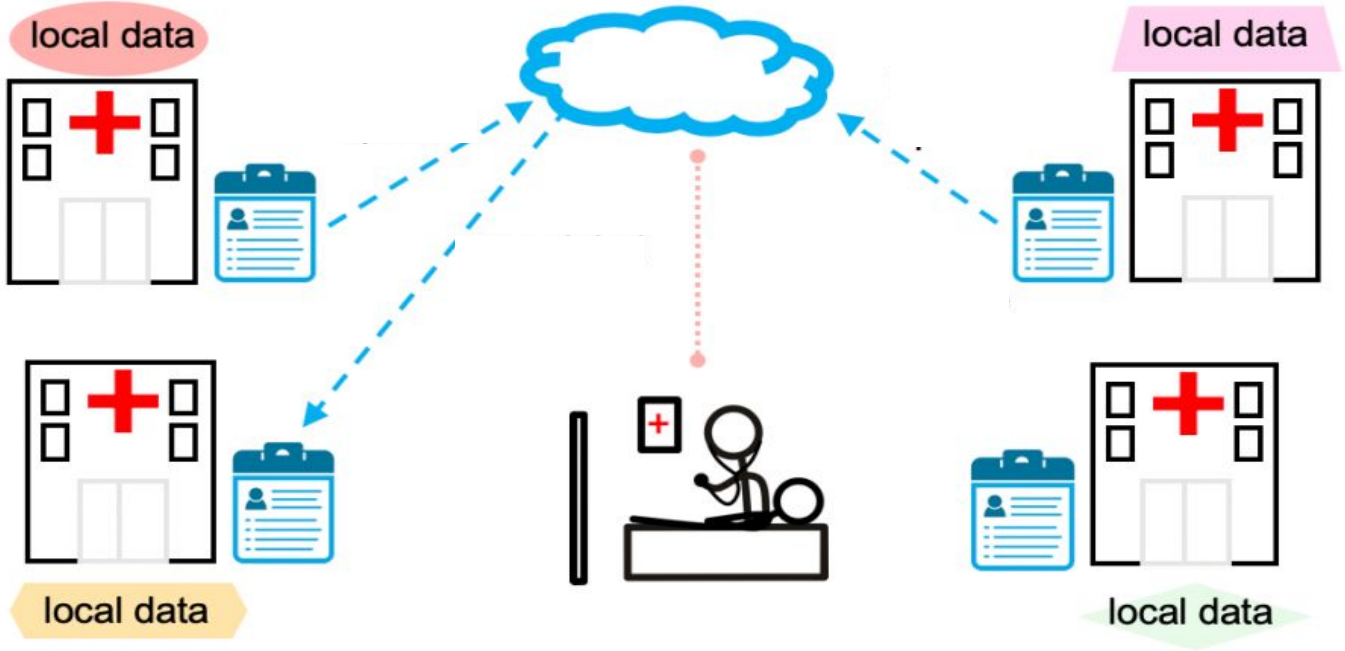
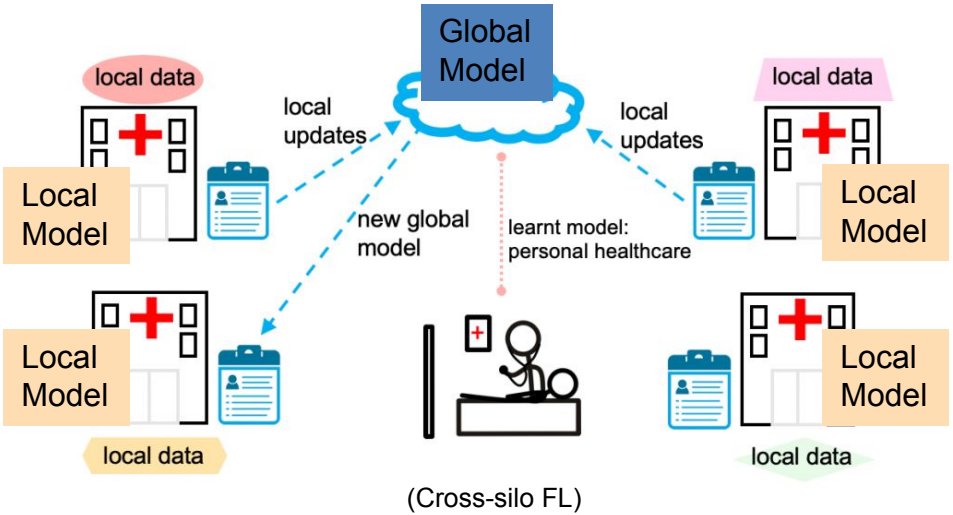


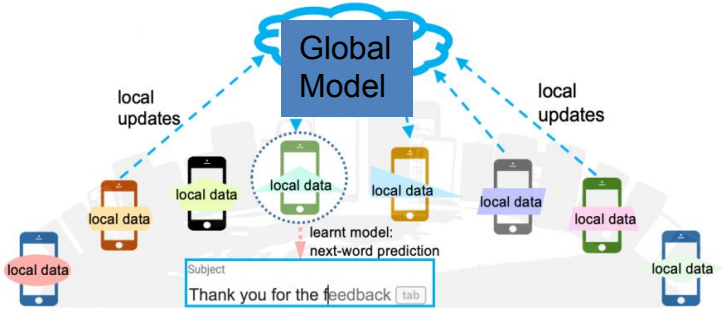
image courtesy: https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

Federated Learning
Learning a Joint Model from Distributed Data Silos

- No data shared among data sources
- Training is pushed down to data sources
- Sources share parameters (e.g., gradients)



(Cross-silo FL)



(Cross-device FL)

Challenges

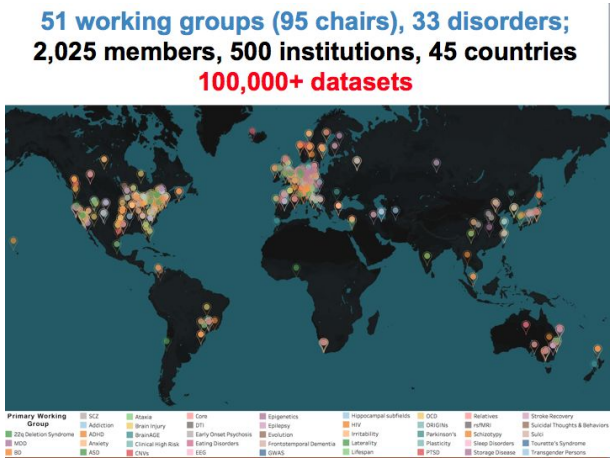
- Efficient learning
- Privacy/Security
- Heterogeneous environments in
  - Computational power
  - Data distributions (size, target classes)

Neuroimaging Datasets

- ADNI is a longitudinal natural history study to design therapeutic trials in Alzheimer's disease
- Data: clinical, PET, biofluid and genetic information, MRI data
- 1,828 subjects
- ADNI I: 400 with mild cognitive impairment (MCI), 200 with early AD, 200 elderly control subjects.
- ADNI II: 150 elderly controls, 100 Early MCI subjects, 150 late mild cognitive impairment (LMCI), and 150 mild AD patients, 107 Significant Memory Concern (SMC)
- ADNI-3 will enroll 133 normal, 151 amnestic MCI, 87 AD subjects

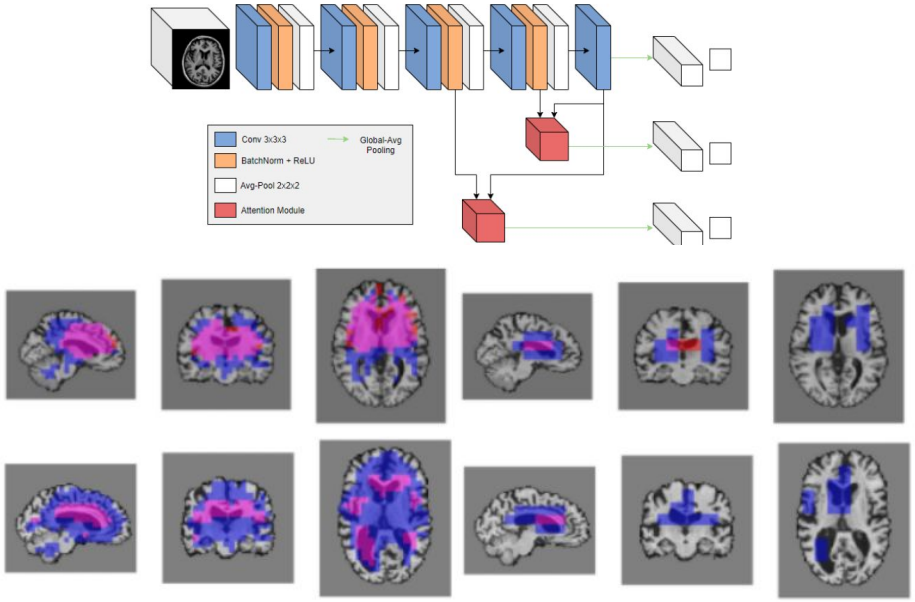
- **UK Biobank** comprises 500,000 individuals with genetic data and extensive phenotyping
- Includes 40,000 individuals with MRI and other imaging

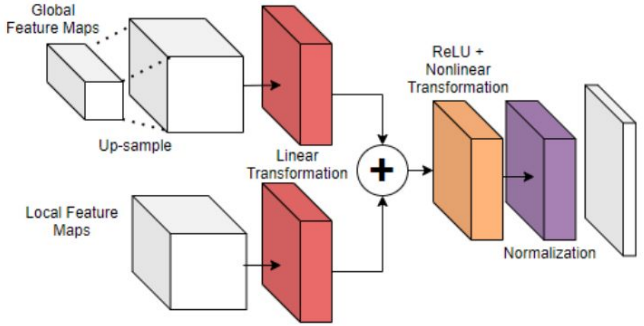
ENIGMA (Enhancing NeuroImaging Genetics through MetaAnalysis).



Neural Architectures for NeuroScience

Examples we have benchmarked: 3D CNNs, 2D Slice RNNs, CNNs with attention mechanisms (multi-task: predict age and diagnosis)





arXiv:2011.09115 (q-bio)

[Submitted on 18 Nov 2020]

3D Grid-Attention Networks for Interpretable Age and Alzheimer's Disease Prediction from Structural MRI

Pradeep Lam, Alyssa H. Zhu, Iyad Ba Gari, Neda Jahanshad, Paul M. Thompson

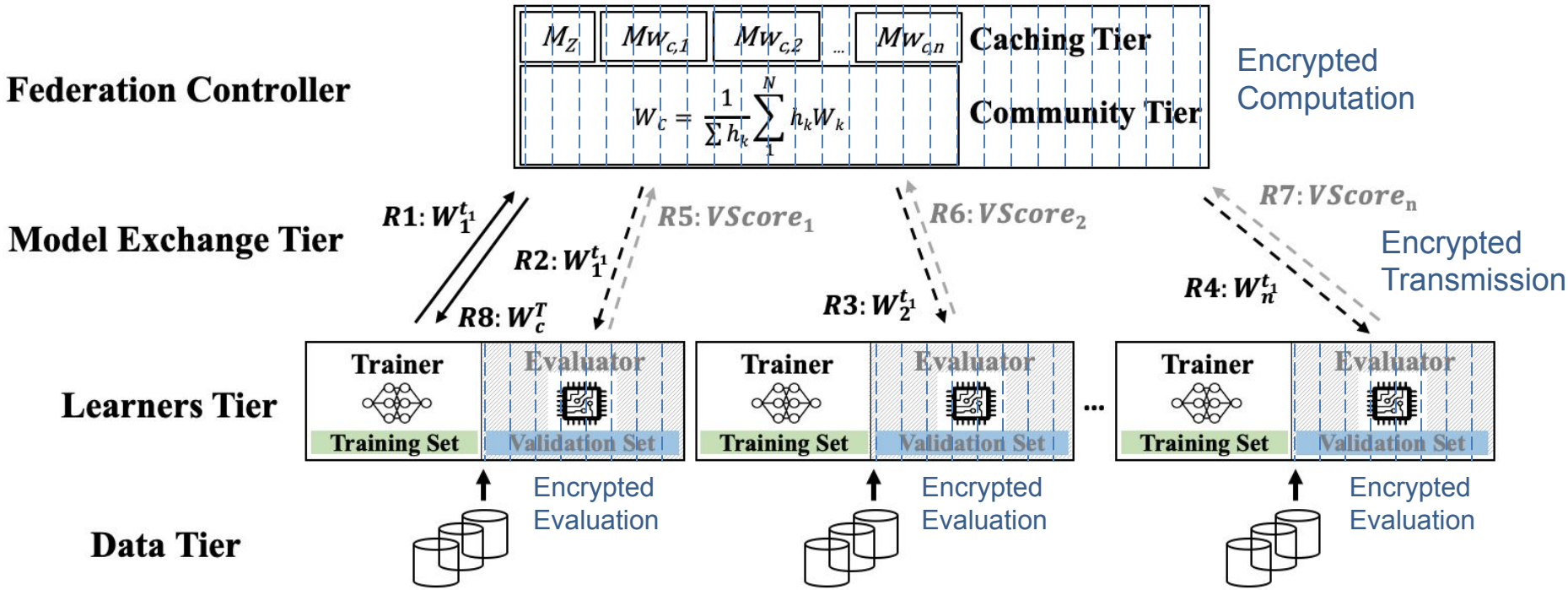
Neuroscience Tasks

By increasing difficulty:

- **Brain Age**: predict age of a subject based on brain MRI
- **AD Prediction**: predict Alzheimer's Disease status based on brain MRI; relatively easier to predict, relatively more data
- **PD Prediction**: predict Parkinson's Disease status based on brain MRI; harder to predict, low data regime

- **Imaging Genetics**: association between genetic variation and brain imaging. Can use MRI to predict genetic variant (SNP, gene)? Are effects of genes/SNPs localized on different brain regions?

METIS Federated Learning Architecture



Dimitris Stripelis, Jose Luis Ambite. Accelerating Federated Learning in Heterogeneous Data and Computational Environments. arXiv:2008.11281.

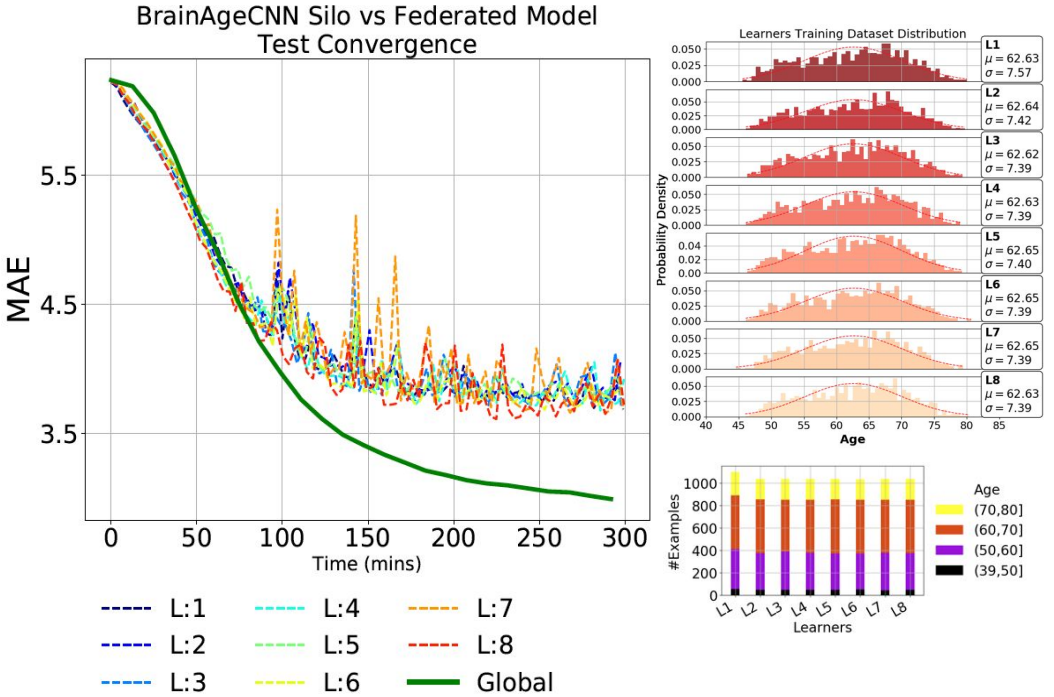
Single Silo vs Federation Performance: Uniform Data, IID

Specifications

- L_x: Centralized training on silo
- **Global:** Federated Execution

Federated Learning Environment

- **Model:** 5-CNN
- **Domain:** BrainAge
- **Data Amount:** Uniform, equal number of samples
- **Data Distribution:** IID



Community Model significantly outperforms any local model

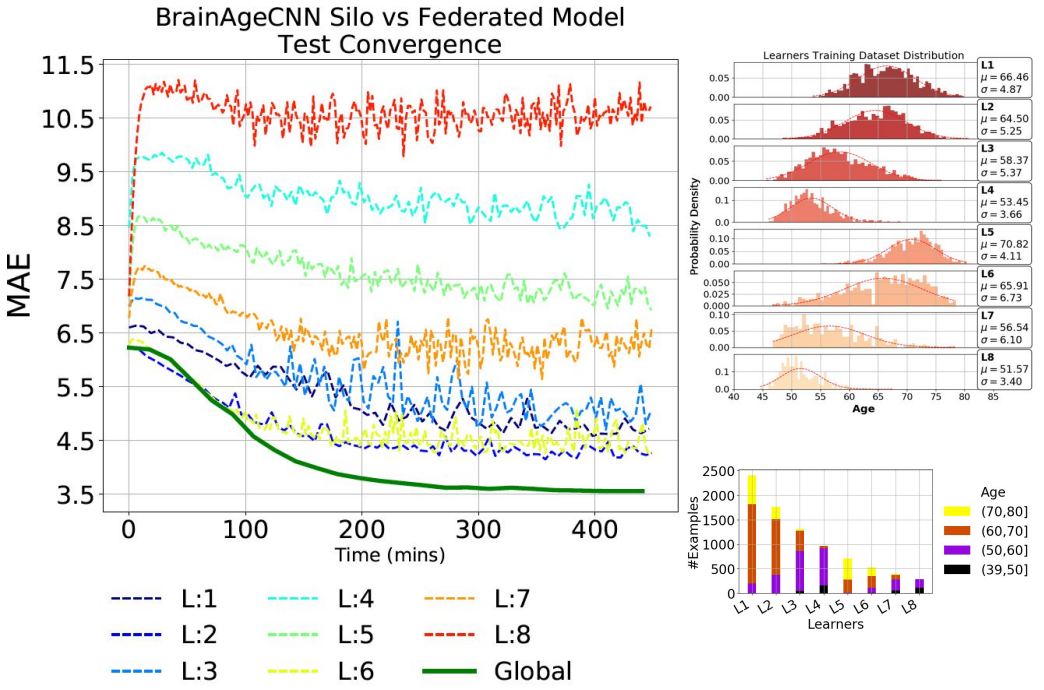
Single Silo vs Federation Performance: Skewed Data, Non-IID

Specifications

- L_x: Centralized training on silo
- **Global:** Federated Execution

Federated Learning Environment

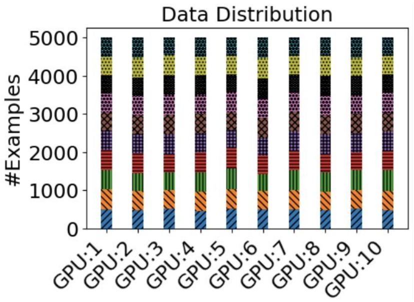
- **Model:** 5-CNN
- **Domain:** BrainAge
- **Data Amount:** Skewed, rightly skewed samples assignment
- **Data Distribution:** Non-IID



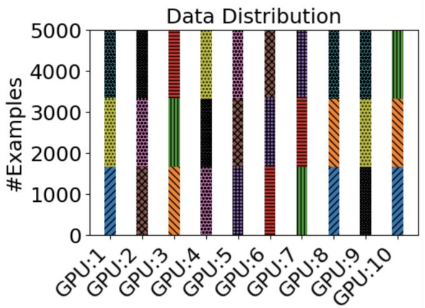
Community Model significantly outperforms any local model

Data Heterogeneity

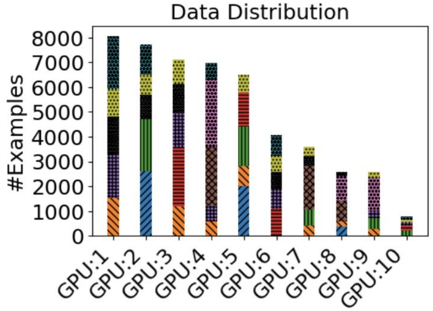
- Different amounts of data per site (uniform, skewed, powerlaw)
- Different classes/distribution per site (IID vs Non-IID)



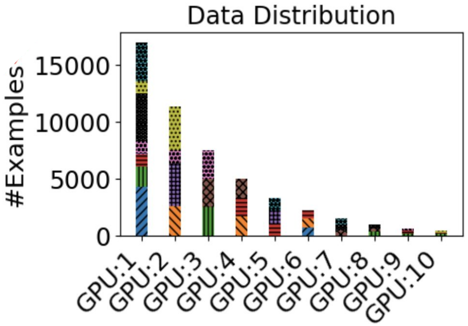
Uniform & IID
(10 target classes per learner)



Uniform & Non-IID(3)
(3 target classes out of 10 per learner)



Skewed & Non-IID(5)
(5 target classes out of 10 per learner)



PowerLaw & Non-IID(3)
(8,4,3x8 target classes per learner)

Easier ← - - - - - - - - → ***Harder***

Learning Difficulty

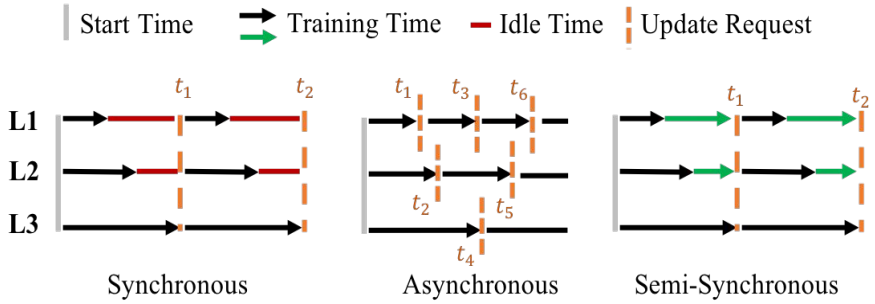
Training Policies: Model Mixing and Communication

Communication Protocols

Synchronous: new community model when all learners complete local training

Asynchronous: new community model whenever a single learner completes its local training

Semi-Synchronous: like sync, but temporally defined synchronization points based on learners' computational power

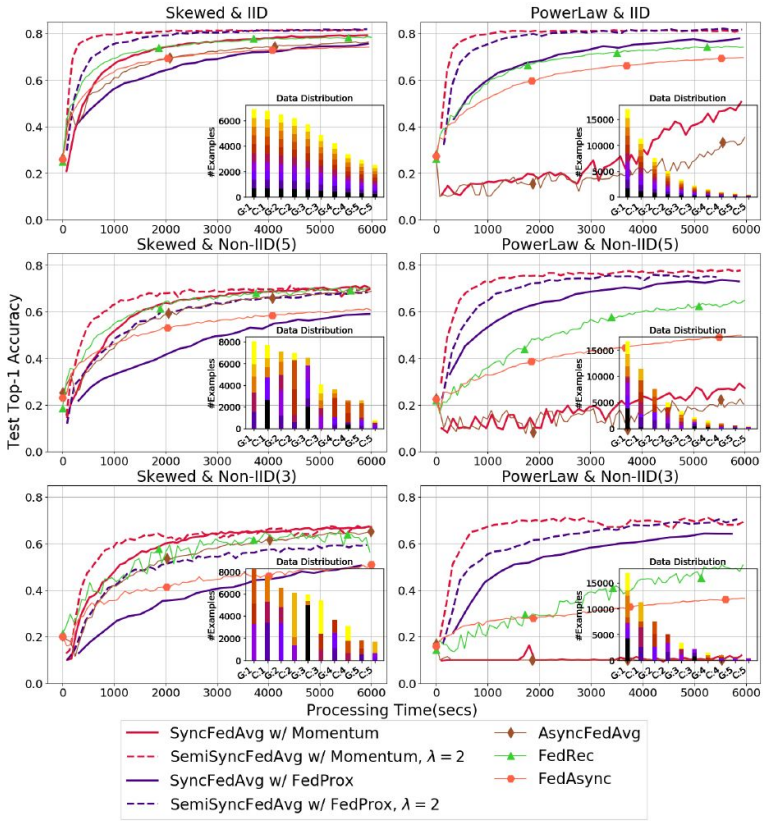


Local Models Mixing Strategies

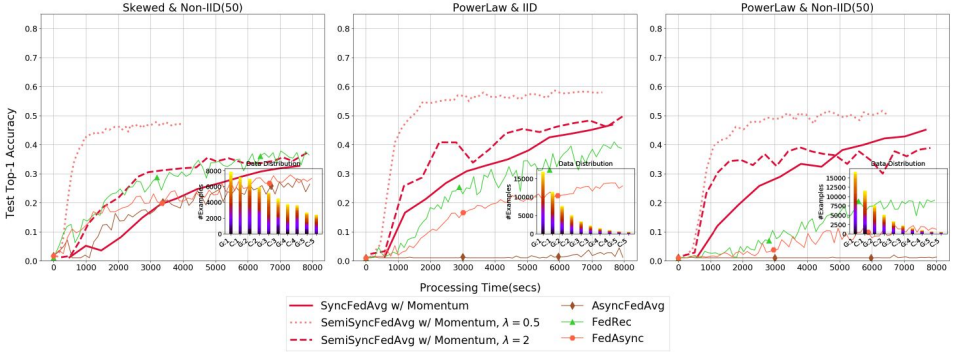
- **Federated Average (FedAvg)**
  - Weighted average based on local data size
- **Proximity (FedProx)**
  - Size-based weighted average (as FedAvg), with regularization term that penalizes deviations from community model to prevent learners from overfitting their local datasets
- **Staleness (FedAsync, FedRec)***
  - Weight models based on time elapsed between last local model update and current time
- **Distributed Validation Weighting (DVW)**
  - Weight local models based on their task performance on a distributed validation dataset

**only applicable to asynchronous protocols*

CIFAR-10 (10 classes)

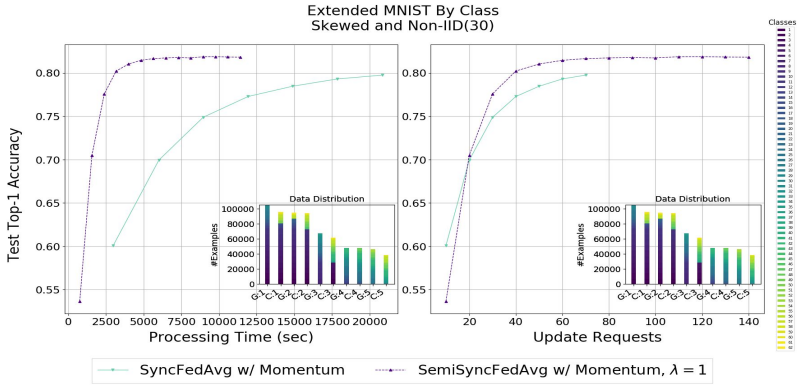


CIFAR-100 (100 classes)



Extended MNIST (62 classes)

Heterogeneous Data and Computational environment
[5 fast (GPU), 5 slow (CPU) learners]

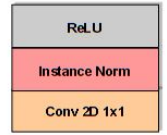


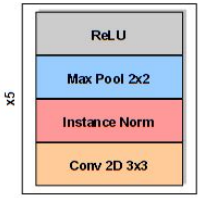
Information Sciences Ins

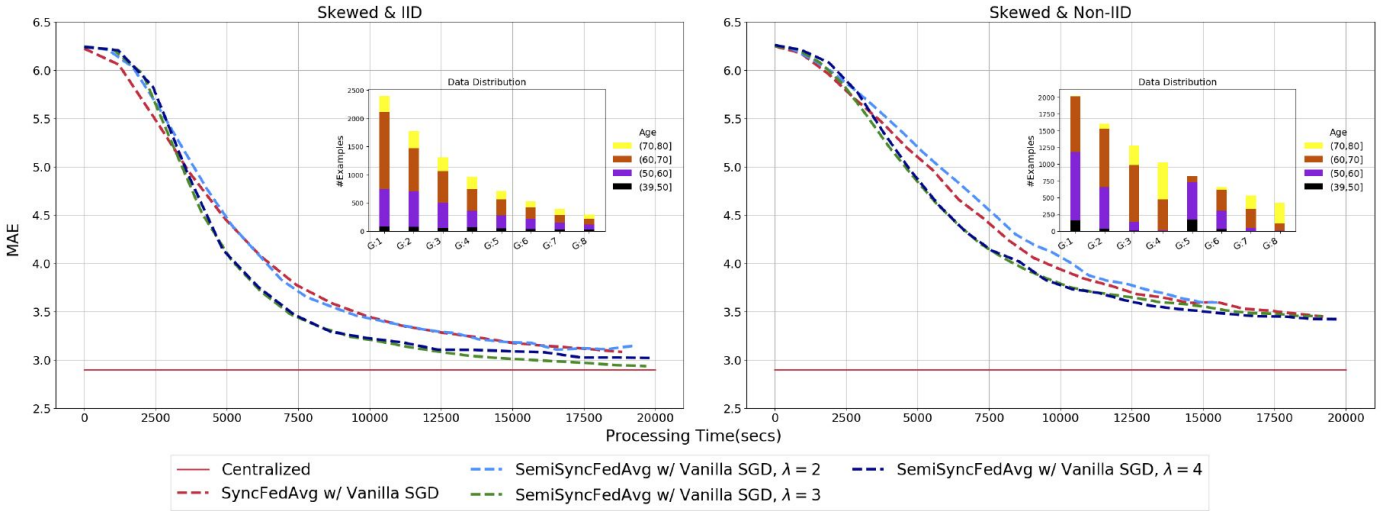
Stripelis, Thompson, Ambite. Semi-synchronous federated learning for energy-efficient training and accelerated convergence in cross-silo settings. 2021. Under Review

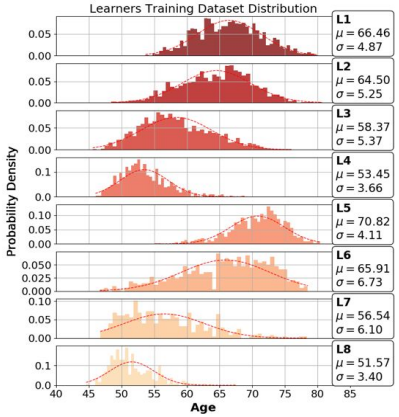
Predicting **Brain Age** from MRI Scans

- **Dataset:** UK Biobank
- **Data Distribution:** Non-IID (diverse age ranges across learners)
- **Computational Environment:** 8xGPU
- **Model:** 5-CNN









Information Sciences In

Dimitris Stripelis, José Luis Ambite, Pradeep Lam, Paul Thompson. Scaling Neuroscience Research Using Federated Learning. International Symposium on Biomedical Imaging (ISBI) 2021

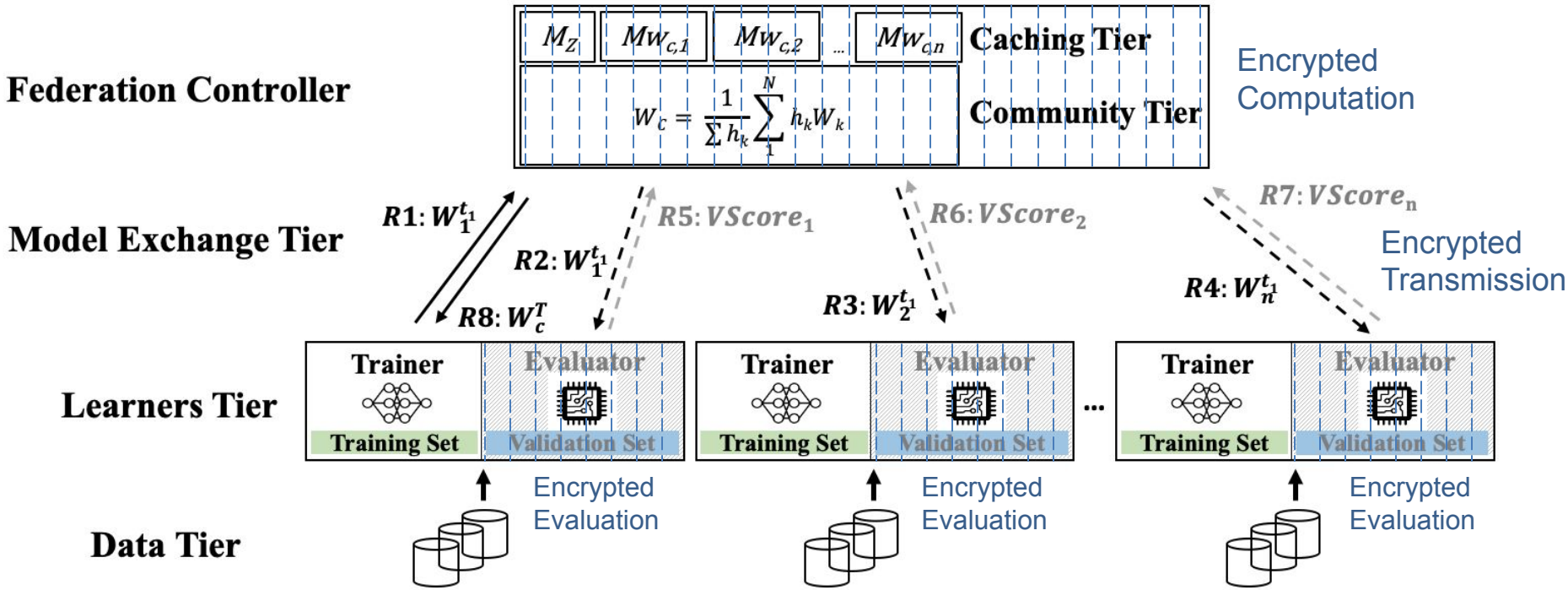
CIFAR-10 *Heterogeneous* Cluster: Energy Efficiency

Semi-Sync is 3 to 9 times more energy efficient than baseline (Synchronous FedAvg with Vanilla SGD)

Energy cost based on processing time on each type of processor weighted by its Thermal Design Power (TDP) [arXiv 20]
GPU's energy cost is two times CPU's

Exp.	Policy	Fast (GPU) Iter.	Fast (GPU) CT(s)	Fast (GPU) EC	Slow (CPU) Iter.	Slow (CPU) CT(s)	Slow (CPU) EC	Total Iter.	Total PT(s)	Total CT(s)	Total CC	Total EC(EF)
Uniform & IID 0.75	Sync w/ Vanilla	24000	578	1157	24000	16129	16129	48000	3225	16707	240	17286
	SemiSync ($\lambda = 2$) w/ Vanilla	45250	1018	2036	4750	3045	3045	50000	631	4064	100	5082 (3.4x)
	Sync w/ Momentum	11000	185	371	11000	2700	2700	22000	540	2885	110	3071 (5.6x)
	SemiSync ($\lambda = 2$) w/ Momentum	20250	335	671	2250	1217	1217	22500	**269**	**1553**	**50**	**1889 (9.1x)**
	Sync w/ FedProx	24000	548	1097	24000	22015	22015	48000	4403	22564	240	23113 (0.7x)
	SemiSync ($\lambda = 2$) w/ FedProx	40250	817	1634	4250	4074	4074	44500	928	4891	90	5708 (3x)
	FedRec	50800	1588	3177	1600	1449	1449	52400	732	3038	261	4627 (3.7x)
	AsyncFedAvg	69200	2439	4879	8800	2870	2870	78000	1315	5310	389	7750 (2.2x)
	FedAsync	76600	2703	5406	8000	3147	3147	84600	1406	5850	422	8554 (2x)
Skewed & Non-IID(5) 0.65	Sync w/ Vanilla	28200	1010	2020	22300	22812	22812	50500	4562	23822	250	24832
	SemiSync ($\lambda = 2$) w/ Vanilla	164082	3585	7170	16603	10006	10006	180685	2153	13591	220	17176 (1.4x)
	Sync w/ Momentum	28200	670	1341	22300	10896	10896	50500	2179	11567	250	12238 (2x)
	SemiSync ($\lambda = 2$) w/ Momentum	93882	1446	2893	9583	4434	4434	103465	**1059**	**5881**	**130**	**7328 (3.3x)**
	Sync w/ FedProx	29328	1065	2130	23192	22998	22998	52520	4599	24063	260	25129 (0.9x)
	SemiSync ($\lambda = 2$) w/ FedProx	210882	4677	9354	21283	12872	12872	232165	2884	17549	280	22226 (1.1x)
	FedRec	155731	5205	10411	4378	4448	4448	160109	2431	9654	803	14860 (1.6x)
	AsyncFedAvg	183982	6185	12370	17361	7349	7349	201343	3346	13534	1021	19719 (1.2x)
	FedAsync	*did not reach target accuracy*										
Power Law & Non-IID(3) 0.6	Sync w/ Vanilla	9664	600	1201	6496	10107	10107	16160	2021	10707	80	11308
	SemiSync ($\lambda = 2$) w/ Vanilla	80102	1831	3662	8183	4939	4939	88285	1103	6770	80	8601 (1.3x)
	Sync w/ Momentum	*did not reach target accuracy*										
	SemiSync ($\lambda = 2$) w/ Momentum	34502	573	1147	3623	1723	1723	38125	**438**	**2296**	**40**	**2870 (3.9x)**
	Sync w/ FedProx	13288	863	1727	8932	16363	16363	22220	3272	17227	110	18091 (0.6x)
	SemiSync ($\lambda = 2$) w/ FedProx	102902	2084	4168	10463	7084	7084	113365	1719	9168	100	11252 (1x)
	FedRec	90458	3671	7343	2647	2922	2922	93105	1832	6594	664	10266 (1.1x)
	AsyncFedAvg	*did not reach target accuracy*										
	FedAsync	*did not reach target accuracy*										

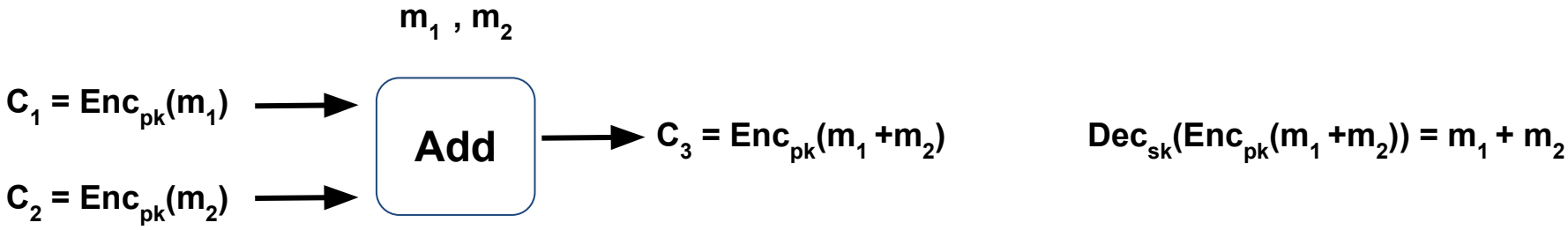
METIS Federated Learning Architecture

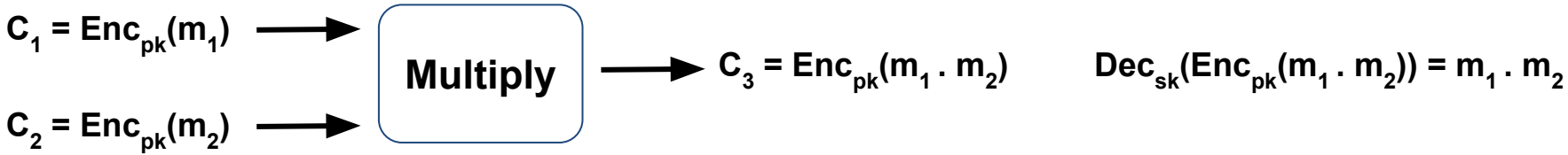


Dimitris Stripelis, Jose Luis Ambite. Accelerating Federated Learning in Heterogeneous Data and Computational Environments. arXiv:2008.11281.

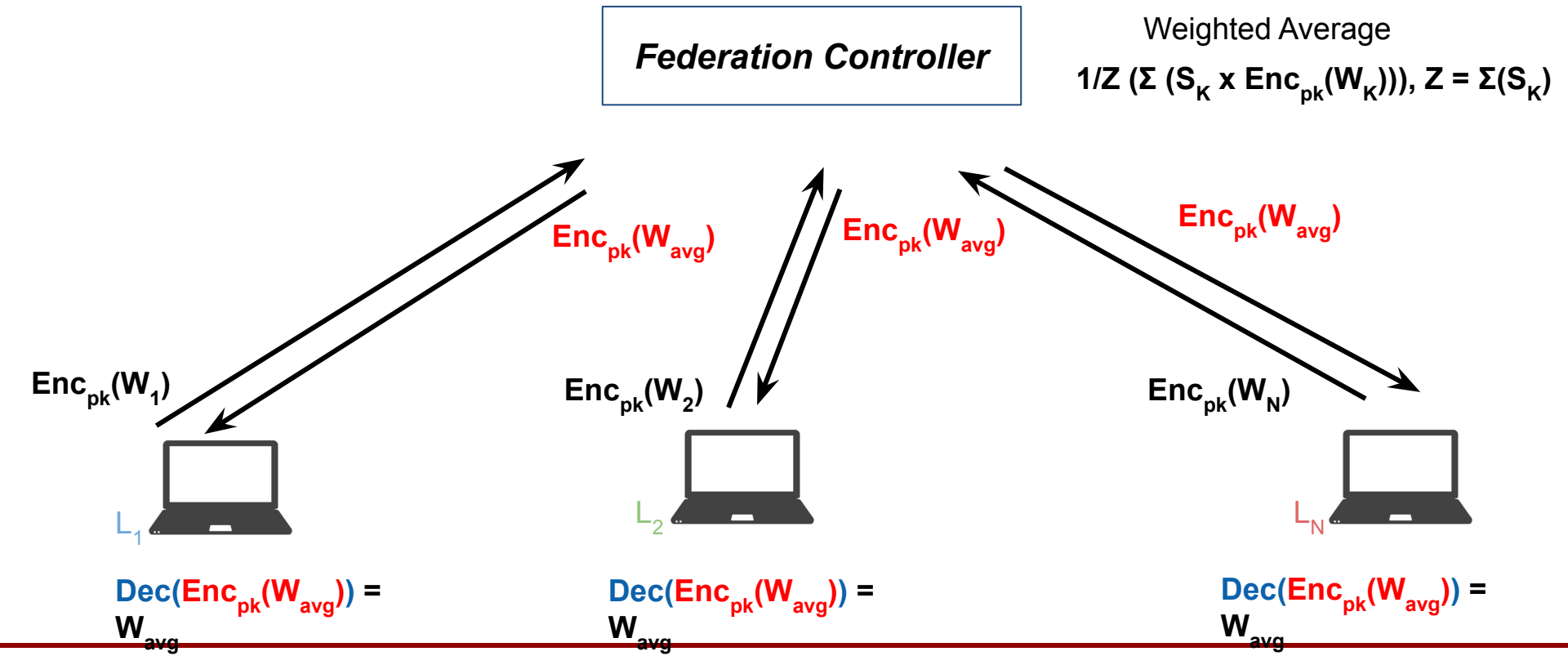
Homomorphic Encryption

- Computations on encrypted data without decrypting





Federated Average with Homomorphic Encryption



Types of Homomorphic Encryption

- Partial Homomorphic Encryption
  - Additions or Multiplications
- Somewhat Homomorphic Encryption
  - Subsets of circuits
- Levelled Fully Homomorphic Encryption
  - Circuits with bounded depth
- Fully Homomorphic Encryption
  - Unbounded Depth

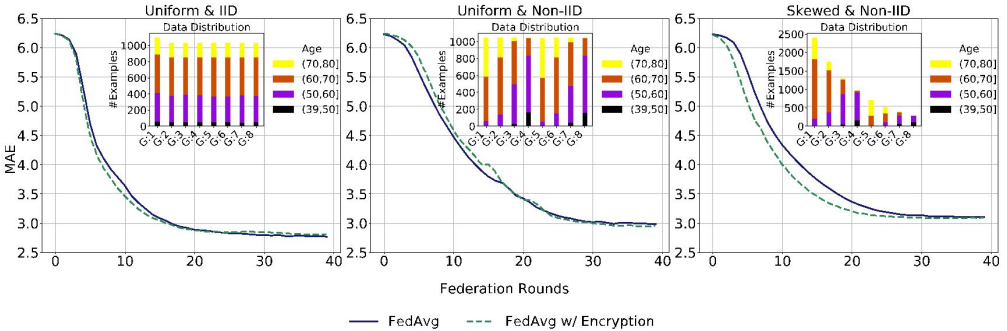
In this work we use:

- Partial Homomorphic Encryption
  - **Paillier Scheme**
- Levelled Fully Homomorphic Encryption
  - **CKKS**
  - **BGVrns**

Scheme	Scaling factors	Capabilities
Paillier	Public	Arbitrary Additions
CKKS	Private	Arbitrary Additions Arbitrary Multiplication
BGVrns	Private	Arbitrary Additions Arbitrary Multiplication

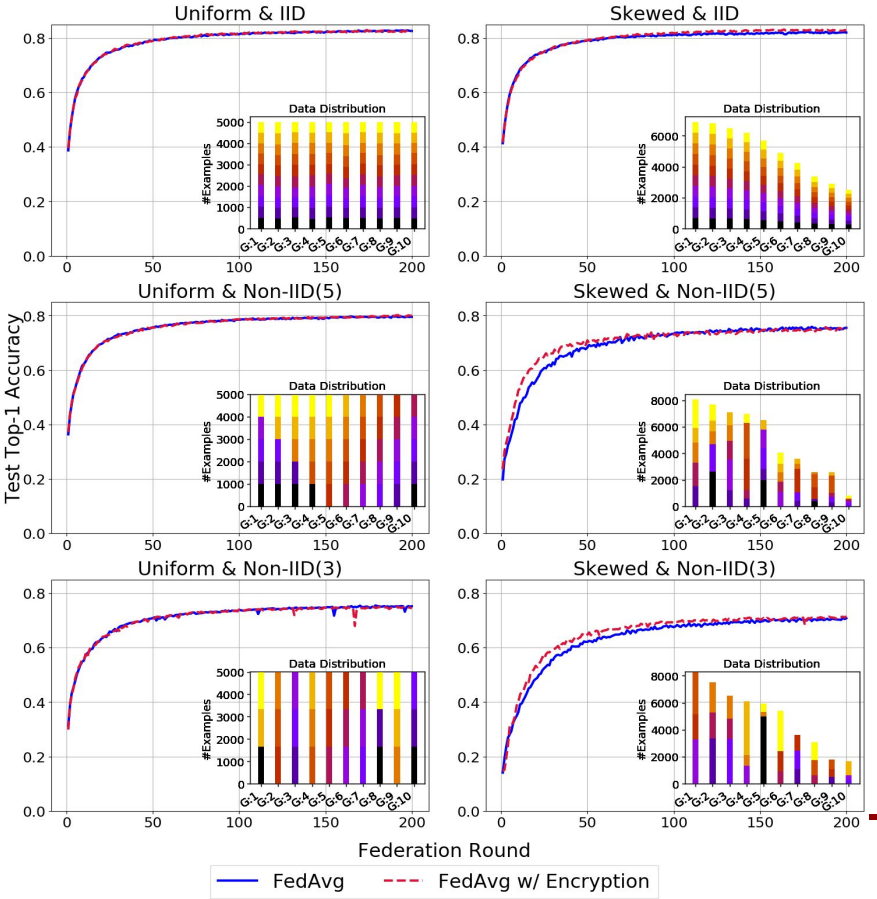
Encrypted Computation preserves federation model learning performance

BrainAge



Fully-homomorphic Encryption:
Cheon-Kim-Kim-Song (CKKS) construction

CIFAR-10



Stripelis, Saleem, Ghai, Dhinagar, Gupta, Anastasiou, Ver Steeg, Ravi, Naveed, Thompson, Ambite. Secure neuroimaging analysis using federated learning with homomorphic encryption. 17th International Symposium on Medical Information Processing and Analysis (SIPAIM), Campinas, Brazil, 2021.

Corruption Modes

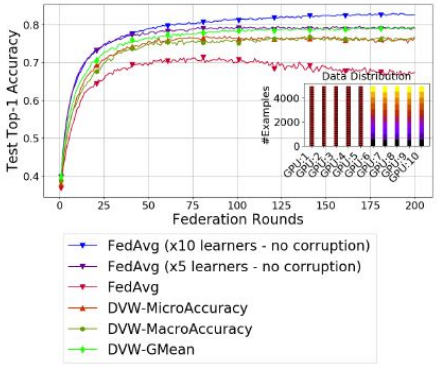
Focus primarily on errors, not on adversarial attacks*

- Random Label Shuffling
  - Labels for examples randomly assigned
- Systematic Label Flipping
  - Label planes as birds
- Image Flip
  - MRI with right/left hemispheres switched
- Attribute Coding Errors
  - Flipping code for males and females

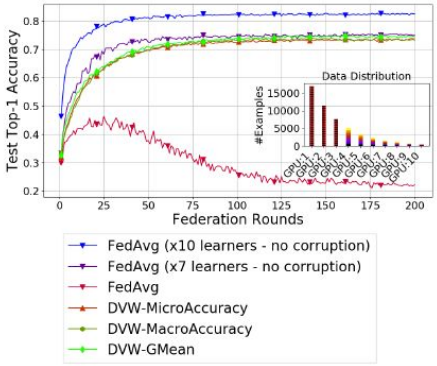
...

* It seems that FedAvg is quite resilient to realistic adversarial attacks already: Shejwalkar et al. Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Federated Learning. 2021. arXiv:2108.10241

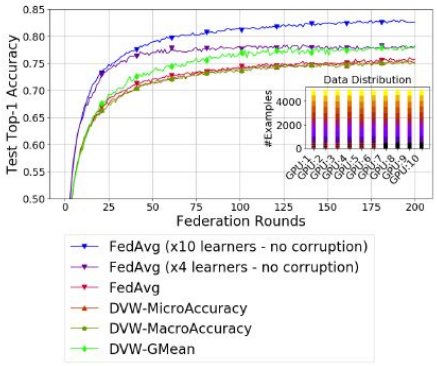
Initial Experiments on Corrupted Sources



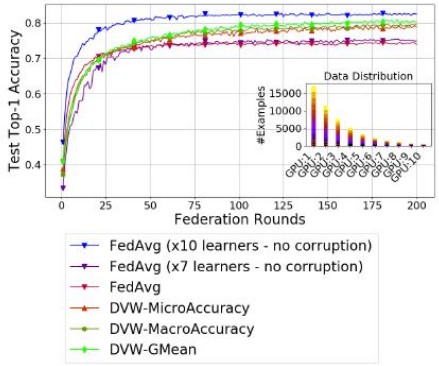
(a) Label Shuffling Uniform & IID



(b) Label Shuffling PowerLaw & IID



(c) Label Flipping Uniform & IID



(d) Label Flipping PowerLaw & IID

Convergence for Cifar10 under Label Shuffling and Label Flipping. Using our DVW performance weighting method based on the geometric mean value on the per-class accuracy for each learner, we are able to detect and minimize the effect of corrupted learners. Panels a, b, c, d have 5, 3, 6, and 3 corrupted learners in the federation, respectively.

Efficient Secure Evaluation using MPC

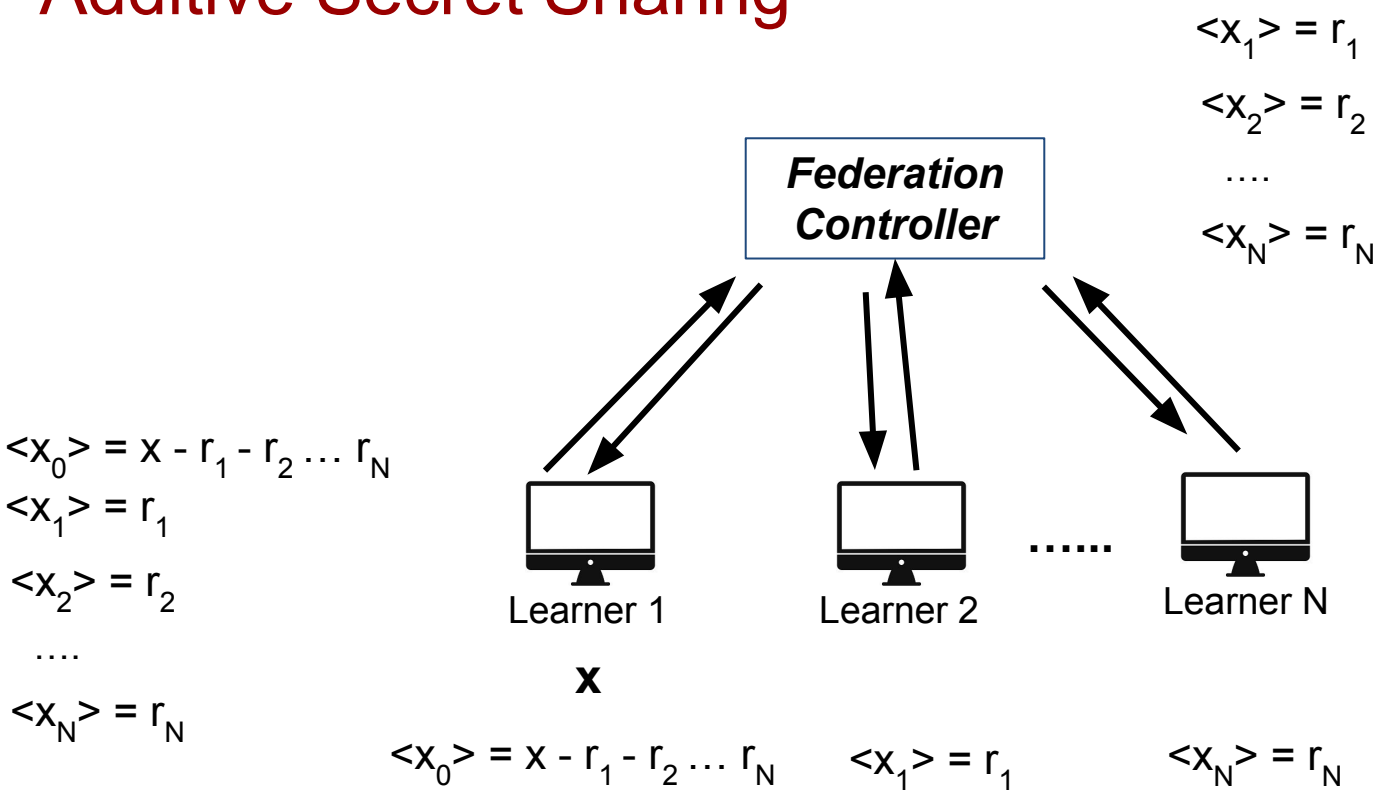
- Distributed Validation Weighing (DVW) requires the learners to test performance without disclosing their validation sets.
- We use secure multi-party computation for DVW.
- State-of-the-art* suffers from high communication overhead and ad-hoc approximations that reduce accuracy.
- We develop protocols for faster and accurate machine learning.

*

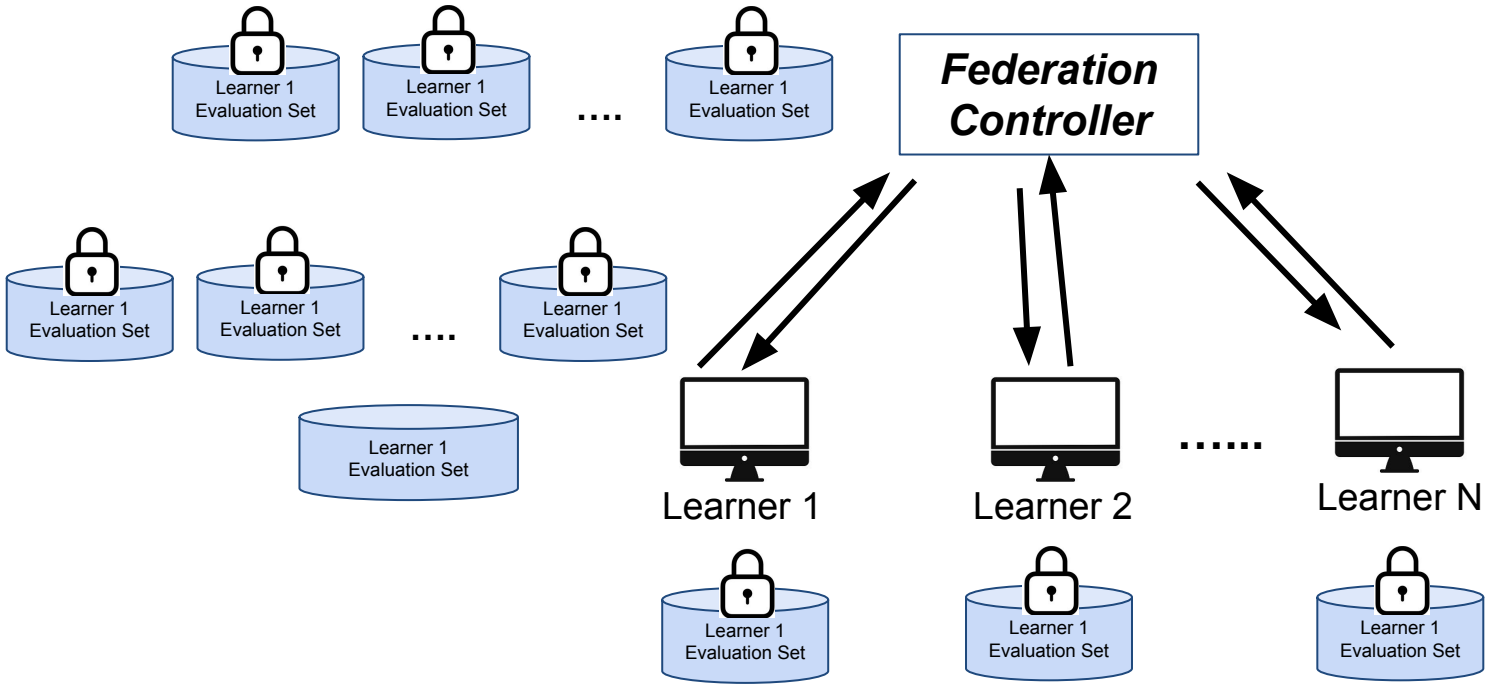
Mohassel, Payman, and Yupeng Zhang. "Secureml: A system for scalable privacy-preserving machine learning." In *IEEE symposium on security and privacy (SP 2017)*.

Wagh, Sameer, Divya Gupta, and Nishanth Chandran. "SecureNN: Efficient and Private Neural Network Training." in Privacy Enhancing Technologies Symposium. (PETS 2019)

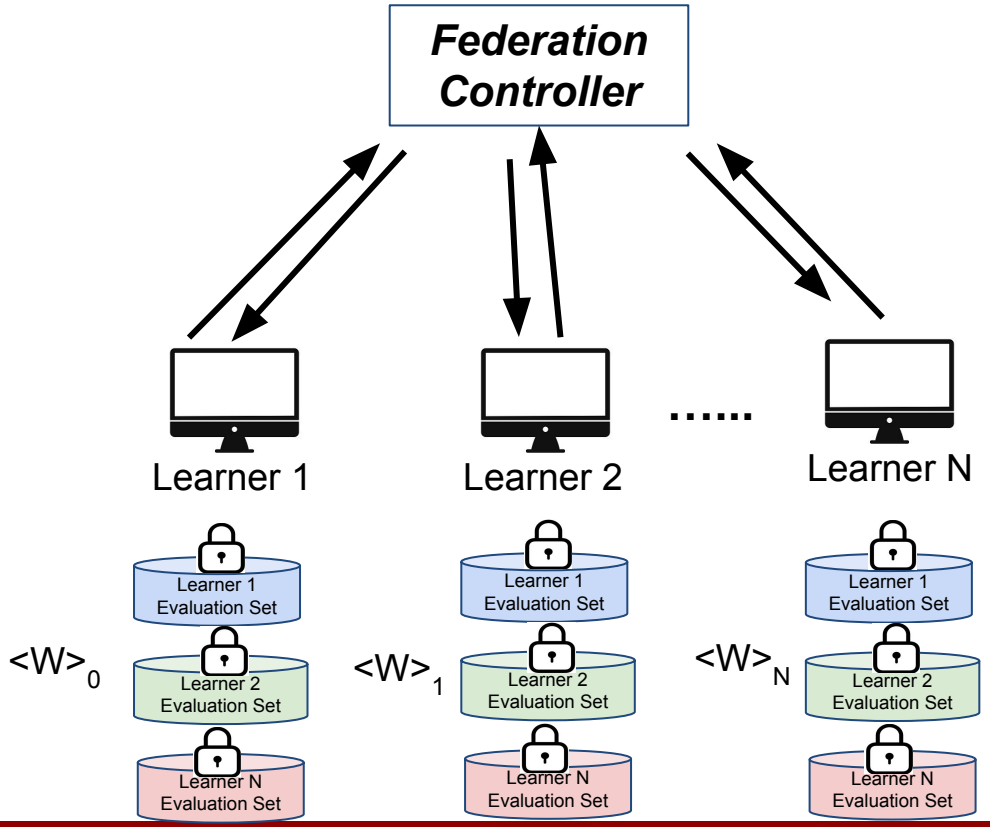
Additive Secret Sharing



Efficient Secure Evaluation: Our Proposed Technique



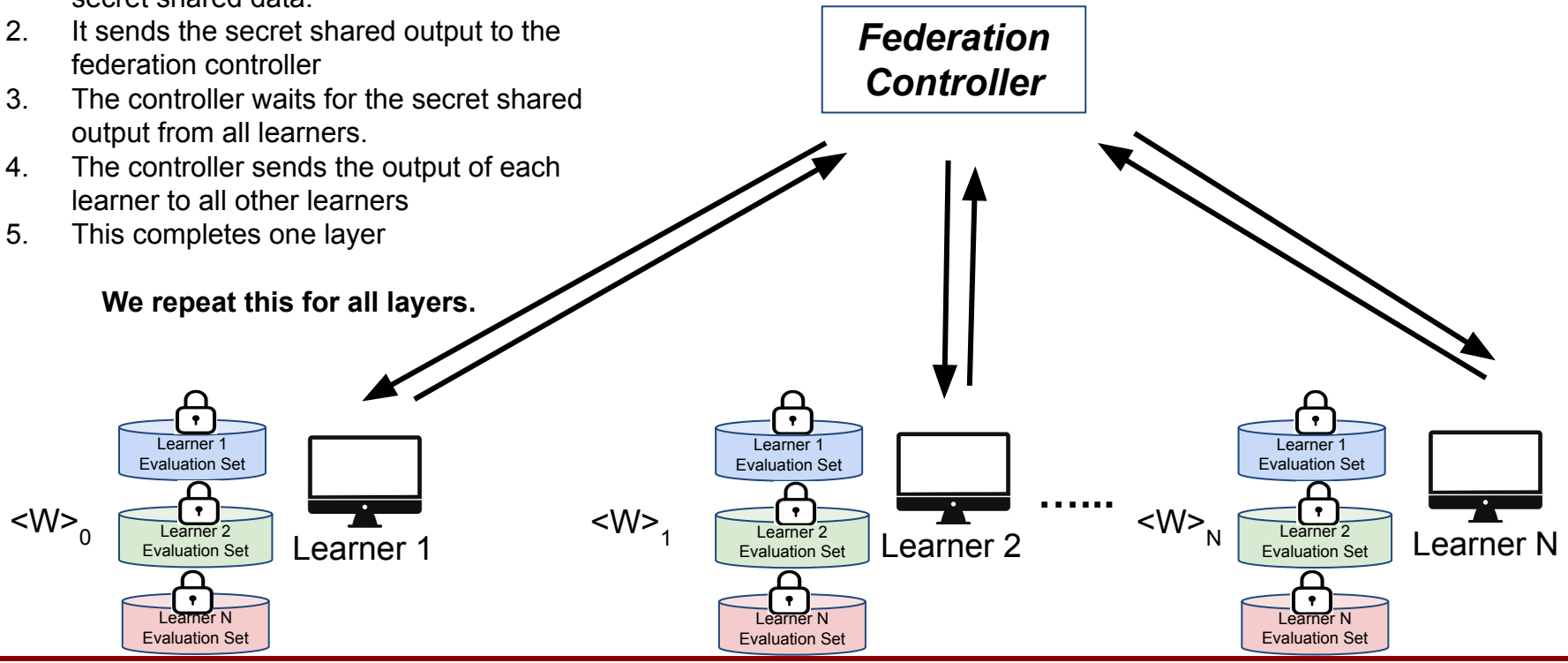
Efficient Secure Evaluation: Secret Sharing Data



Efficient Secure Evaluation: Secure Computation

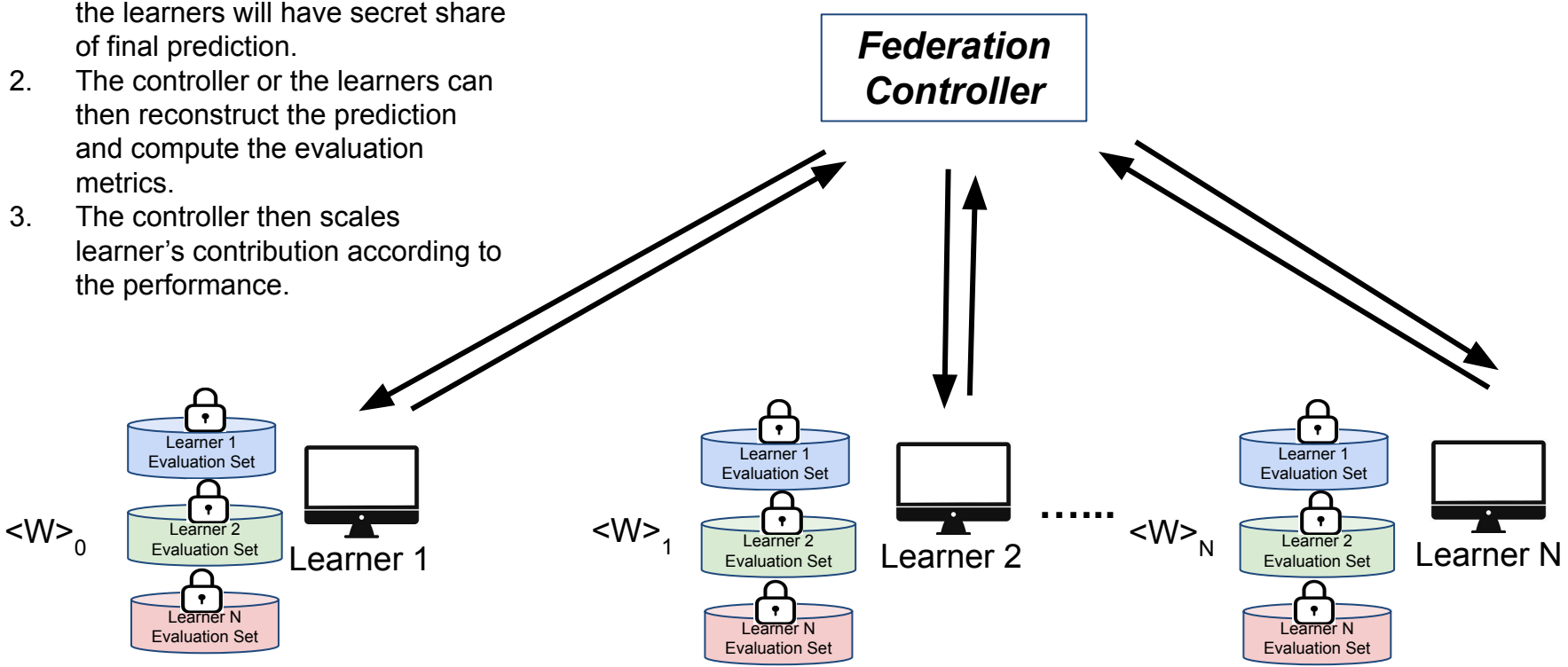
1. Every learner computes one layer on the secret shared data.
2. It sends the secret shared output to the federation controller
3. The controller waits for the secret shared output from all learners.
4. The controller sends the output of each learner to all other learners
5. This completes one layer

We repeat this for all layers.



Efficient Secure Evaluation: Secure Computation

1. After all layers are computed, all the learners will have secret share of final prediction.
2. The controller or the learners can then reconstruct the prediction and compute the evaluation metrics.
3. The controller then scales learner's contribution according to the performance.



Efficient Secure Evaluation using MPC

- Machine Learning building blocks:
  - Addition
  - Multiplication
  - Non-linear activations (sigmoid, ReLU, softmax, maxpool).
- Addition and Multiplication
  - can be performed using secret-sharing based protocols in rounds linear in the number of learners.
- Non-Linear Functions
  - Secure computation protocols require functions to be represented as circuits
  - No efficient way to represent activation function as a circuit, so they use approximations
- Our Proposed Approach: Secure Lookup Tables for computing nonlinear functions in MPC.

Efficient Secure Evaluation - Generating SLTs

$PRF_{K2}(x_0)$	$<y_0>_0$
...	...
$PRF_{K2}(x_i)$	$<y_i>_0$
...	...
$PRF_{K2}(x_m)$	$<y_m>_0$

SLT_0

$PRF_{K1}(x_0)$	$<y_0>_1$
...	...
$PRF_{K1}(x_i)$	$<y_i>_1$
...	...
$PRF_{K1}(x_m)$	$<y_m>_1$

SLT_1

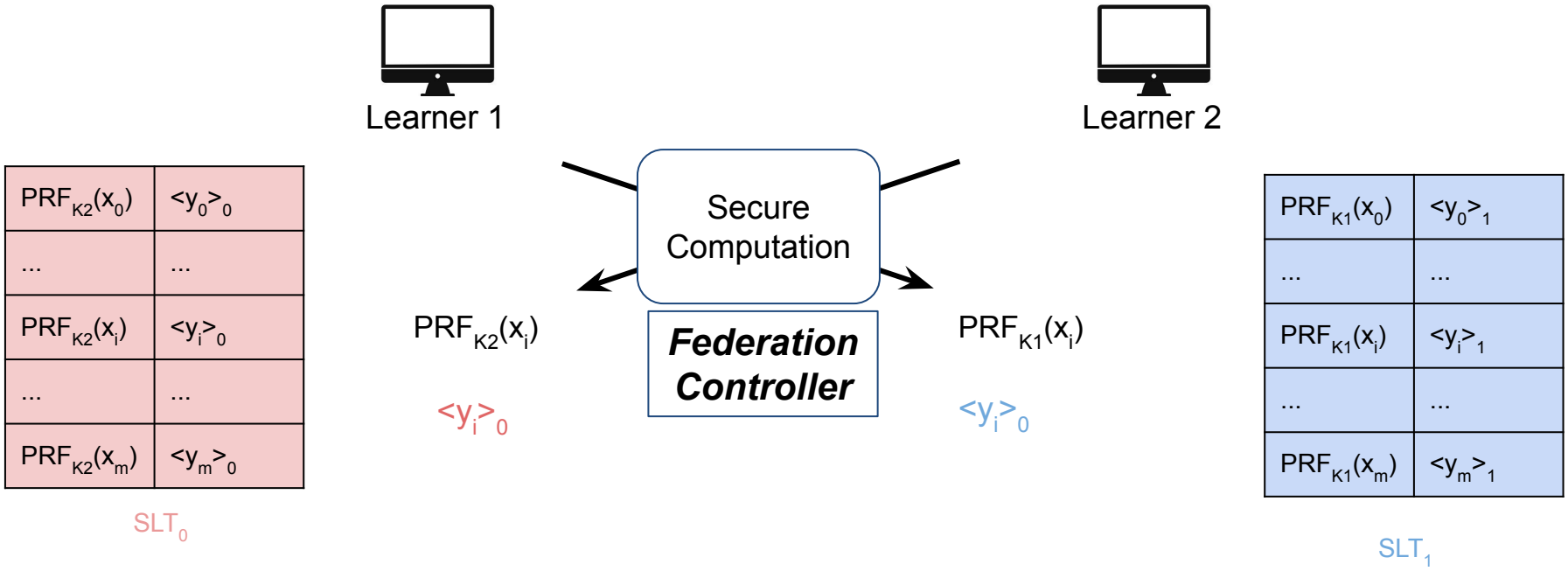
x_0	y_0
...	...
x_i	y_i
...	…
x_m	y_m

Learner 1

Learner 2

- Compute ***y = f(x)*** for all values of **x**.
- Generate two random keys **K1** and **K2**
- Create empty SLT_0 , SLT_1
- For every entry **(x_i , y_i)**:
  a. generate Key-pair:
  $(PRF_{K2}(x_0) , <yi>_0)$
  and add it to SLT_0
  b. generate Key-pair:
  $(PRF_{K1}(x_0) , <yi>_1)$
  and add it to SLT_1
- Randomly shuffle the **SLT**

Efficient Secure Evaluation - SLT Lookup



Efficient Secure Evaluation - Next Steps

- Evaluating the performance of the DVW scheme
- Providing resiliency against a subset of learners dropping in the federation

- Using HE for model evaluation
- Challenges:
    - Computing non-linear activation functions
    - Approximating activations

Plan

- *3 months:* Evaluate and select homomorphic encryption method for model sharing and evaluation.
- *6 months:* Develop encrypted model evaluation module and integrate in the Federated Learning architecture.
- *9 months:* Evaluate alternative performance weighting schemes under different corruption models on standard benchmarks.
- *12 months:* Evaluate performance weighting under typical corruption models present in neuroimaging datasets.

Thanks! Questions?