DOE ADVANCED SCIENTIFIC COMPUTING ADVISORY COMMITTEE MEETING

Argonne NATIONAL LABORATORY

BUILDING ON SUCCESS: ADVANCING PRIVACY-PRESERVING FEDERATED LEARNING WITH DISTRIBUTED OPTIMIZATION

KIBAEK KIM
Computational Mathematician
Mathematics and Computer Science
Argonne National Laboratory

Collaborators:
Wendy Di (ANL),
Ravi Madduri (ANL),
Minseok Ryu (former ANL; ASU)

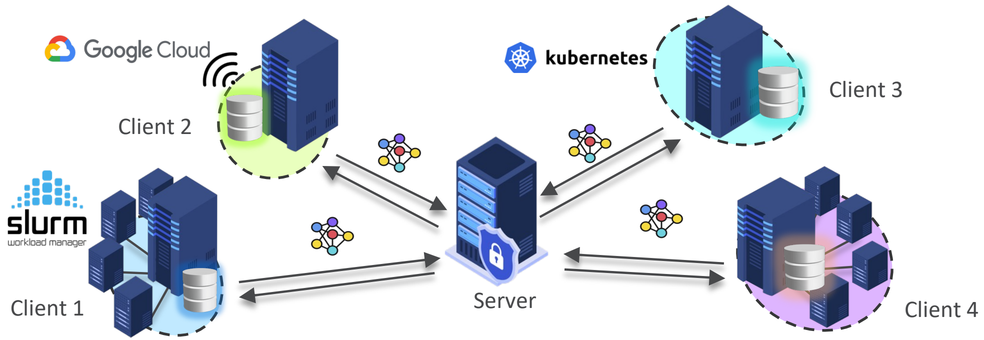
September 27, 2024

WHAT IS FEDERATED LEARNING?

Collaboratively Training Models without Sharing Data

Distributed learning approach with key benefits:

- **Privacy**: Models are trained locally.
- **Efficiency**: Only model updates are shared, reducing data transfer.
- **Scalability**: Supports large-scale applications across many computing devices.



Mathematical formulation of FL

# of clients

# of samples at client k

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\mathbf{w})$$

Training loss at client k

FedAvg:
one of the simplest and most widely-used algorithms

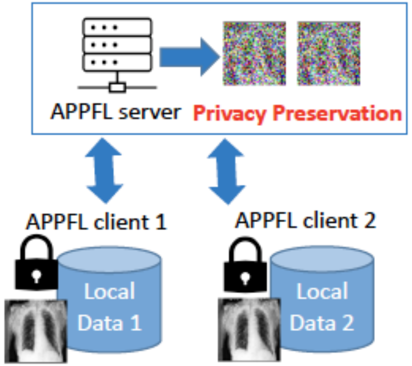
$$\mathbf{w}_t^k = \mathbf{w}_t - \eta \nabla F_k(\mathbf{w}_t)$$ *Local training at client k*

$$\mathbf{w}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbf{w}_t^k$$

PRIVACY-PRESERVING IN FEDERATED LEARNING

Ensuring Data Privacy and Secure Updates

- **Local Data Retention:** Raw data stays on client devices, but model updates alone can still leak sensitive information.
- **Potential Data Leakage:** Without privacy-preserving techniques, attackers can reconstruct raw data from gradients or model updates.
- **Differential Privacy:** Adds noise to model updates to prevent accurate data reconstruction by attackers.



> (j) The term "differential-privacy guarantee" means protections that allow information about a group to be shared while provably limiting the improper access, use, or disclosure of personal information about particular entities.
>
> ***From Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence***

Argonne NATIONAL LABORATORY

APPLICATIONS ACROSS KEY DOE DOMAINS

Use Cases of Privacy-Preserving Federated Learning for DOE

- **Scientific Experiments:**
  - Collaborative experiments using multimodal data (e.g., from DOE light source facilities) while preserving data privacy across institutions.

Argonne's APS

- **Climate Science:**
  - Secure data collaboration between research centers, allowing them to share insights from climate models and data (e.g., from the ARM facility) without sharing raw data.

ARM Facility

- **Electric Grid Data Analysis:**
  - Privacy-preserving FL for analyzing electricity consumption patterns across smart meters, enhancing prediction models while maintaining consumer data privacy.

Smart Meters and Sensors

Argonne NATIONAL LABORATORY

MATH & ALGORITHM CHALLENGES

ALGORITHMS FOR PRIVACY-PRESERVING FL

Balancing Privacy and Utility in Federated Learning

- **Key Challenge:** Managing the privacy-utility trade-off.
- **Algorithm design:** Critical to optimize both privacy and performance.
- **Noise Injection Points:**
  - **Data (input):** Perturb data before training.
  - **Model (output):** Add noise before sharing the model.
  - **Training Loss (objective):** Incorporate noise during training.
- **Goal:** Enhance training performance & maintaining privacy guarantees.

Mathematical formulation of FL

# of clients

# of samples at client k

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\mathbf{w})$$

Training loss at client k

Model randomization:

$$\tilde{\mathbf{w}}_{t+1}^{k} = \mathbf{w}_t - \eta \nabla F_k(\mathbf{w}_t) + \mathbf{z}$$

random noise

Training randomization:

$$\tilde{\mathbf{w}}_{t+1}^{k} = \mathbf{w}_t - \eta \nabla \tilde{F}_k(\mathbf{w}_t),$$

$$\text{where } \tilde{F}_k(\mathbf{w}) = F_k(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2 + \mathbf{z}^\top \mathbf{w}$$

randomizing training loss

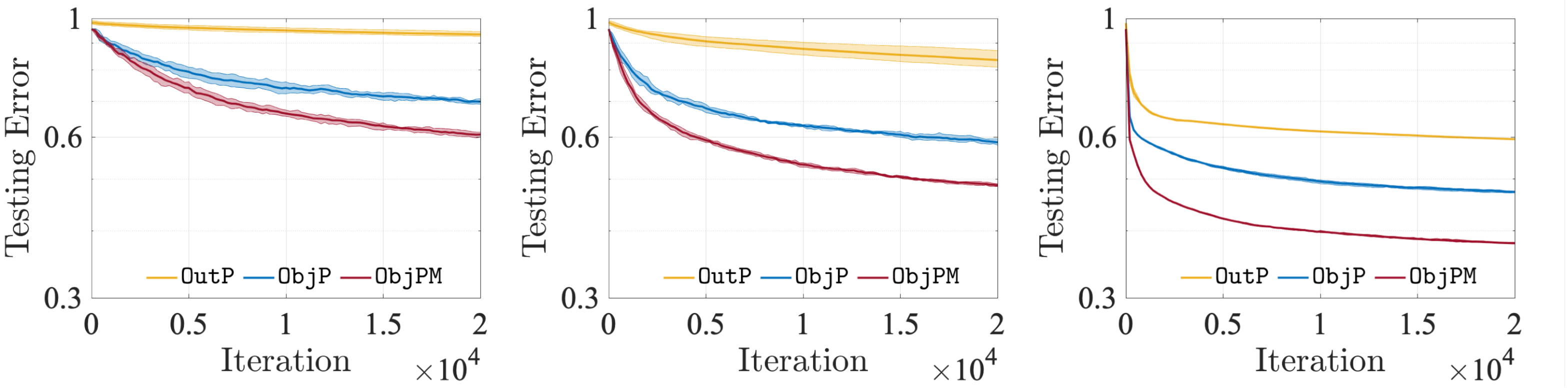
PERFORMANCE UNDER PRIVACY SETTINGS

Testing Accuracy vs. Privacy Level

- **OutP (State of the Art):** FL with noise added to the model output.
- **ObjP (APPFL):** FL with noise added during training.
- **ObjPM (APPFL):** FL with training noise and multiple local updates.
- **Results:** Our methods perform better as privacy increases, compared to current approaches.

Stronger privacy
Weaker learning

Weaker privacy
Stronger learning



Argonne NATIONAL LABORATORY

CHALLENGE OF HETEROGENEOUS COMPUTING

Stragglers and Resource Waste

- **Computing Variance:** Client machines have widely varying capabilities, causing significant differences in local training times.
- **Synchronous FL Drawback:** The server waits for all clients, leading to **resource waste** when slower clients (stragglers) delay the entire process.

Heterogeneous client computing resources.

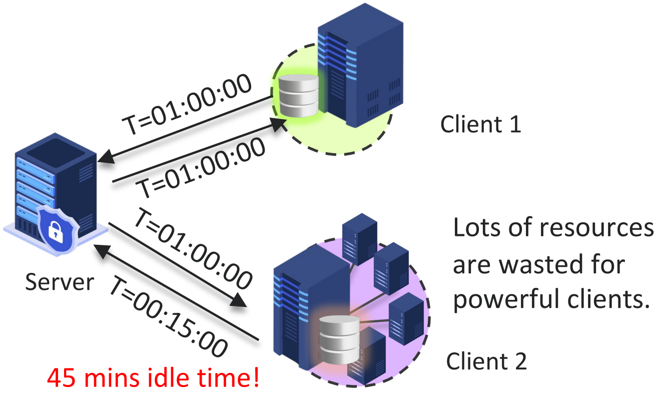
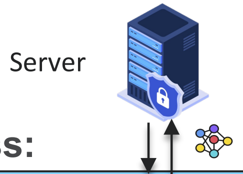


Illustration of Resource waste in synchronous FL.

ADAPTIVE ASYNC UPDATES FOR EFFICIENT FL

FedCompass: Faster Training with Higher Accuracy

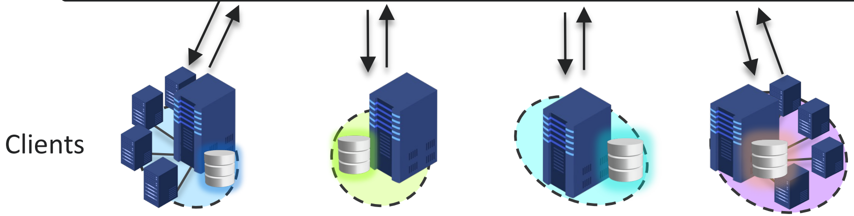


FedCompass:

Step 1: Dynamically estimate each client's computing power.

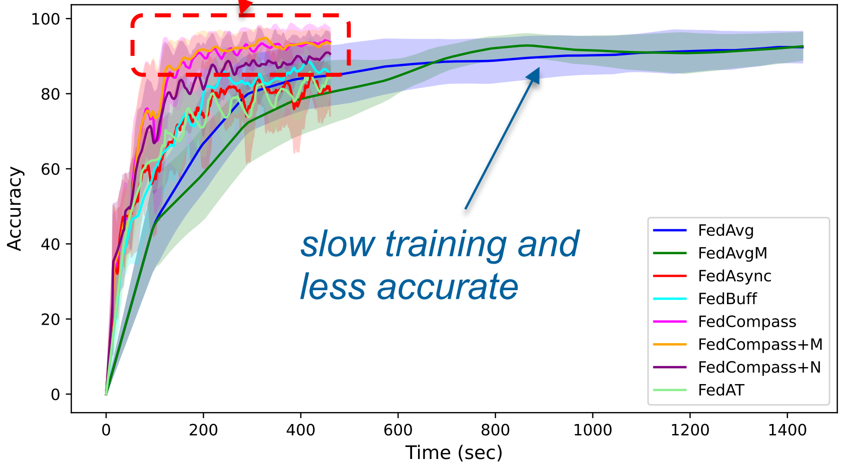
Step 2: Adjust local tasks based on client capabilities to better synchronize model updates.

Step 3: Collaborate with the server to update the global model using client results.



Federated learning with a computing power aware scheduler.

FedCompass achieves faster training and higher accuracy compared to state-of-the-art methods.

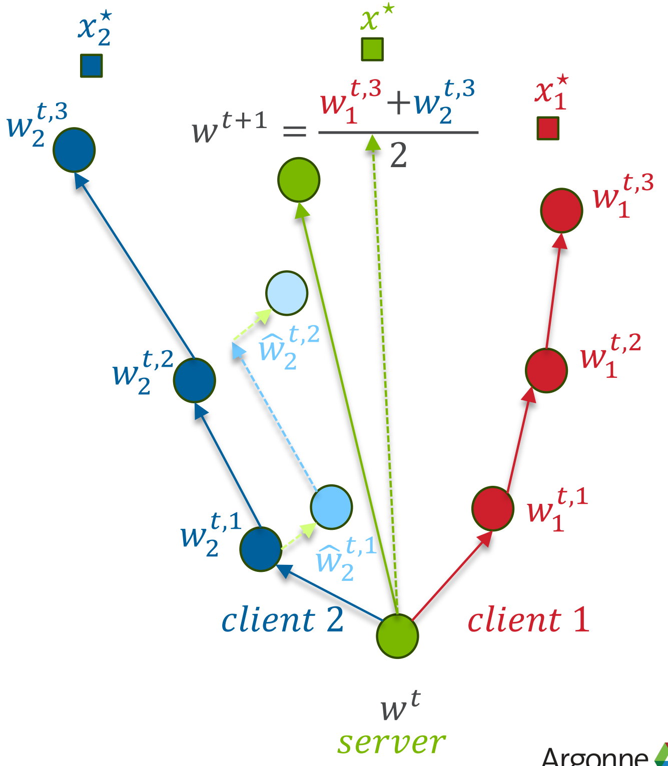


Argonne NATIONAL LABORATORY

CHALLENGE OF CLIENT DRIFT

Balancing Communication Efficiency and Modal Accuracy

- **Client Drift:** Clients run multiple updates locally, leading to misalignment with the global model, reducing overall accuracy.
- **Existing Solutions:** Drift correction methods (e.g., FedProx, SCAFFOLD, FedLin) help mitigate drift but come with trade-offs:
  - **Higher Costs:** Increased communication and storage for correction terms.
  - **Practical Limitations:** Solutions can be unstable and lack asynchronous methods, limiting scalability.

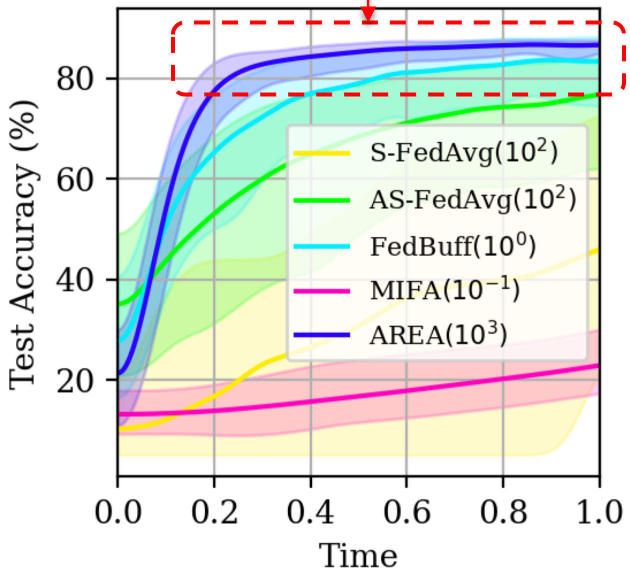


AREA: ASYNCHRONOUS EXACT AVERAGING

Asynchronous Client Drift Correction

- **Client-Side:** Clients save information from previous updates to improve future updates sent to the server.
- **Server-Side:** The server combines these improved updates to create a more accurate global model.
- **Secure:** Compatible with privacy-focused protocols, ensuring data remains secure during the process.

AREA achieves faster training and higher accuracy compared to state-of-the-art methods.



MNIST classification, 128 heterogeneous clients.

OPEN-SOURCE SOFTWARE

APPFL: Advanced Privacy-Preserving Federated Learning

Argonne NATIONAL LABORATORY

https://appfl.ai

APPFL

APPFL V1.0 (08/2024)

Building and Deploying Secure, Scalable FL Algorithms

- **First Code:** Started in 10/2021; **First Release:** 02/2022.
- **For Developers:** Design, simulate, and evaluate new privacy- and FL algorithms.
- **For Users:** Deploy secure, scalable FL experiments across distributed clients.
- **Key Features**
  - **Comprehensive:** Handles data and system heterogeneity and privacy challenges.
  - **Easy-to-use:** Simplifies transitioning from centralized to federated learning.
  - **Extensible:** Modular interface for integrating new algorithms in aggregation, training, and privacy.
  - **Scalable Deployment:** Capable of running FL across multiple HPC clusters and over DOE Energy Sciences Network (ESnet) facility for large-scale distributed experiments.

Argonne Polaris

National Center for Supercomputing Applications Delta

COMPARISON OF OPEN-SOURCE FL SOFTWARE

Key Capabilities Across FL Frameworks

TABLE I: Comparison of popular open-source federated learning frameworks. As of Aug, 2024

Framework	Data Hetero.	Sync. FL	Async. FL	Compression	Versatile Comm.	Privacy	Auth.	Real Deployment	FL Variants
LEAF [54]	✗	✓	✗	✗	✗	✗	✗	✗	✗
TFF [40]	✓	✓	✗	✗	✗	✓	✗	✗	✗
APPFL-V0 [22]	✓	✓	✗	✗	✗	✓	✗	✓	✗
FEDERATEDSCOPE [55]	✓	✓	✗	✗	✗	✓	✗	✓	*VFL*
FLARE [56]	✓	✓	✗	✗	✗	✓	✓	✓	*VFL*
OPENFL [57]	✓	✓	✗	✓	✗	✓	✓	✓	*VFL*
FEDSCALE [21]	✓	✓	✓	✓	✗	✓	✗	✓	✗
FLGO [58]	✓	✓	✓	✗	✗	✗	✗	✓	*VFL*
FEDLAB [59]	✓	✓	✓	✓	✗	✗	✗	✓	✗
FLOWER [19]	✓	✓	✗	✗	✓	✓	✓	✓	*VFL*
FEDML [20]	✓	✓	✗	✗	✓	✓	✓	✓	*VFL, HierFL, DFL*
APPFL (this work)	✓	✓	✓	✓	✓	✓	✓	✓	*VFL, HierFL, DFL*

- *APPFL v1.0 stands out with enhanced support for privacy, asynchronous algorithms, and versatile communication, advancing beyond APPFL v0 and other platforms.*

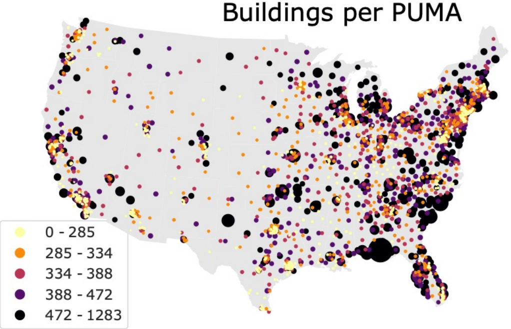
DOE USE CASES OF PRIVACY-PRESERVING FEDERATED LEARNING

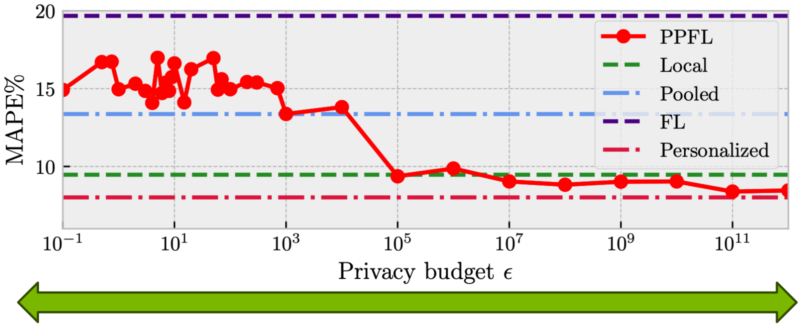
Argonne NATIONAL LABORATORY

FEDERATED LEARNING FOR LOAD FORECASTING

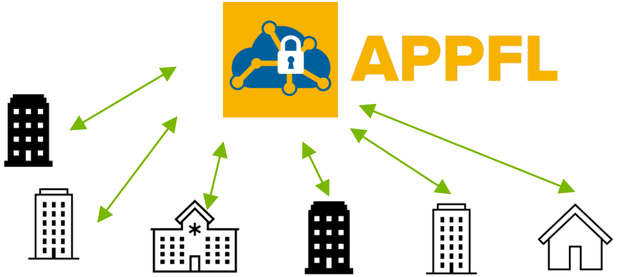
Accurate, Secure Predictions using Building Energy Data

- **Data:** Electricity consumption from 42 buildings in CA, IL, NY.
- **Challenge:** Heterogeneous patterns across buildings.
- **Model:** Attention-based LSTM (long short-term memory) neural network architecture with personalized layers.
- **Results:**
  - **Personalized FL** achieves the lowest error.
  - **PPFL** successfully integrates to ensure data privacy.





Stronger privacy *Weaker privacy*



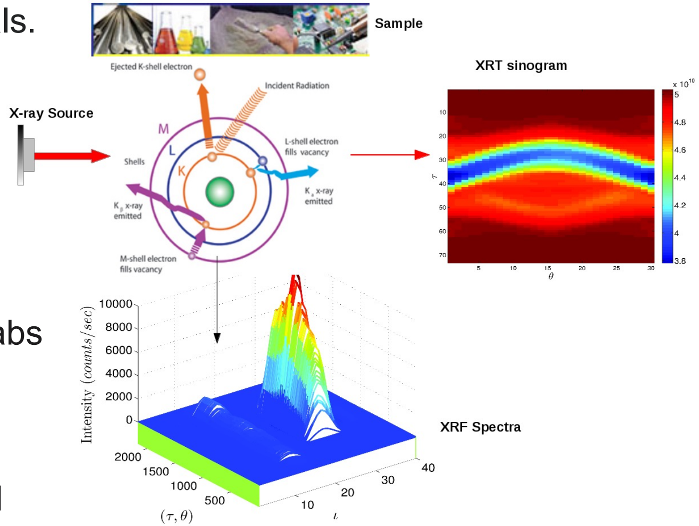
Training a forecast model without moving data + Privacy preserving technique

Argonne NATIONAL LABORATORY

XRT AND XRF AT ARGONNE'S APS FACILITY

Complimentary Data for Advanced Materials Research

- **X-Ray Tomography (XRT):** Provides **3D structural imaging** of materials.
- **X-Ray Fluorescence (XRF):** Maps **elemental composition** of materials.
- **Complimentary Nature:**
  - XRT shows **physical structure**, while XRF reveals **chemical composition**.
  - Together, they offer a **complete view** of material properties.
- **Why Federated Learning?**
  - **Scalable Collaborative Research:** Enable joint analysis across labs without sharing raw data.
  - **Data Privacy:** Keep sensitive data local, further protection with differential privacy.
  - **Better Models:** Combines data from diverse sources for improved generalization.
  - **Resource Efficiency:** Utilizes distributed computing power across multiple facilities.

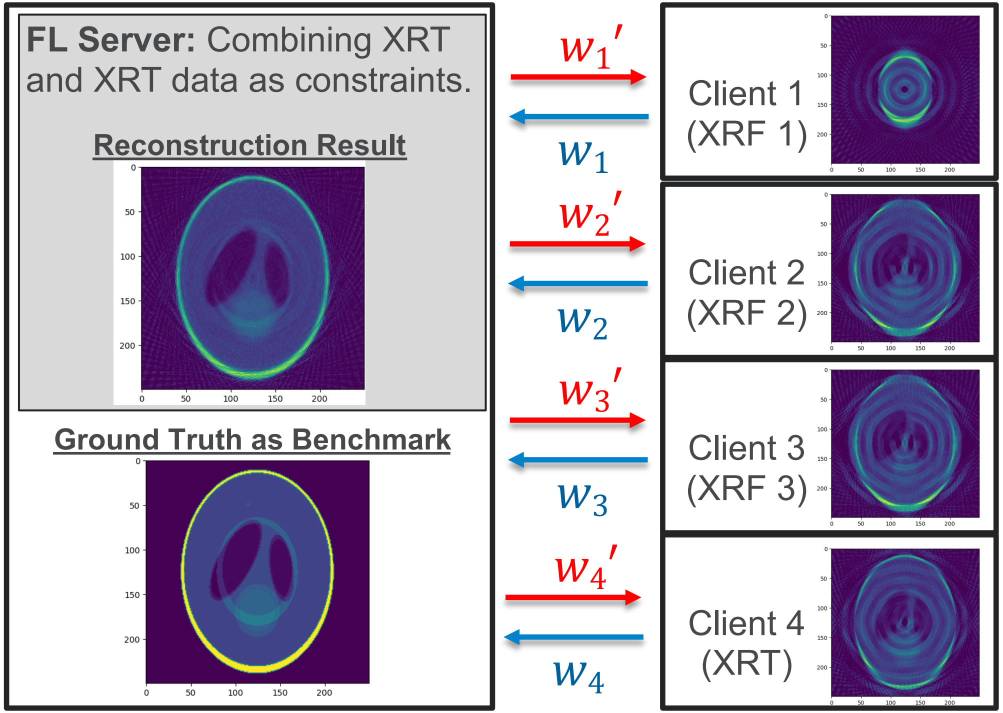


FEDERATED LEARNING ON XRT AND XRF DATA

Empirical Results and Performance Insights

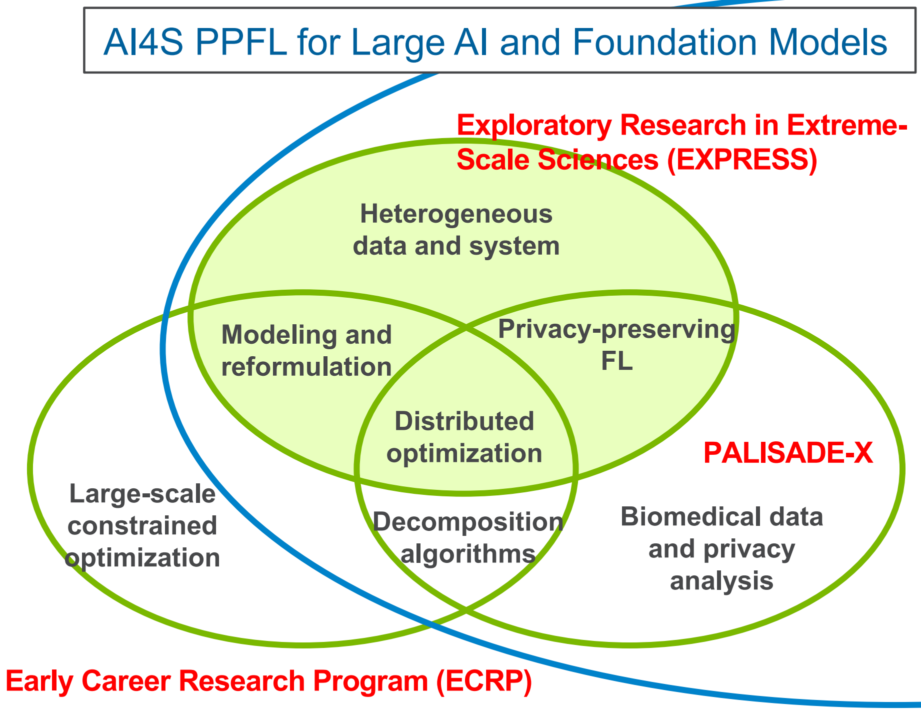
- FL integrates **distributed XRT and XRF data** for improved, **privacy-preserving** image reconstruction.
- **Key Results**
  - Combined XRT and XRF data improves reconstruction accuracy.
  - Developed communication efficient algorithm for federated reconstruction.
- **Takeaway:** Combining data and efficient algorithms boost accuracy and scalability in multimodal federated analysis.

Results of individual reconstruction



PATHWAY TO AI FOR SCIENCE (AI4S)

Success Built on ECRP, PALISADE-X, and EXPRESS



- Large AI and foundation models
- Incentive and fairness to FL clients
- Privacy preservation at scale
- Synthetic data generation
- Sustainable and robust workflows
- Interdisciplinary study: AM + CS + Facilities

Argonne NATIONAL LABORATORY
Brookhaven National Laboratory
OAK RIDGE National Laboratory
ASU
RUTGERS UNIVERSITY

ACKNOWLEDGEMENTS

- DOE ASCR Early Career Research Program (2019 - 2024)
- DOE ASCR PALISADE-X Project (2022 – 2024)
- DOE ASCR EXPRESS (2023 – 2024)
- DOE ASCR Resilient Distributed Systems (2024 – 2028)
- NSF NAIRR Pilot (2024)
- LDRD Seed (2024)
- DOE ASCR AI4S (2025 – 2027)
- **Collaborators**:

THANK YOU