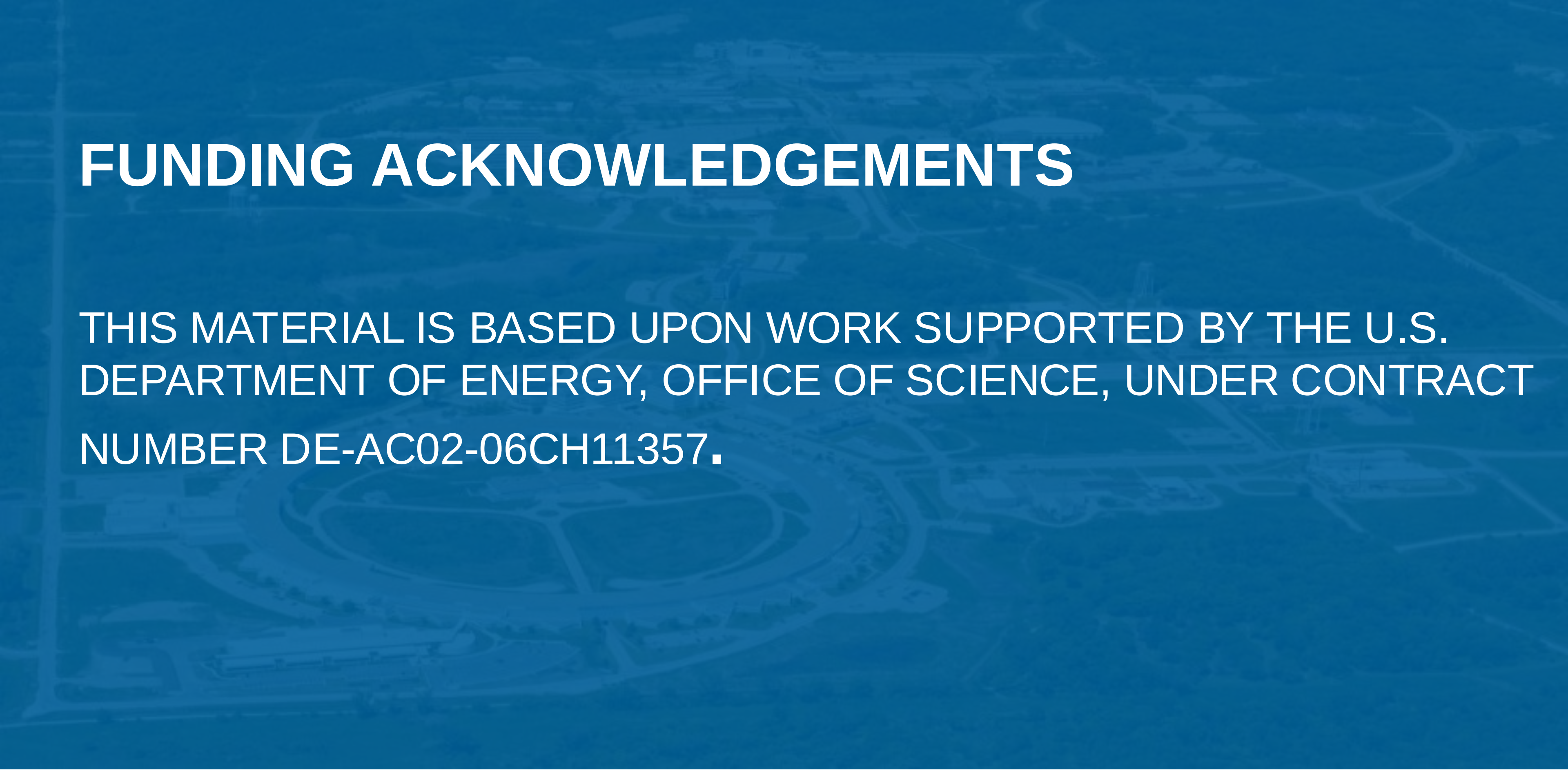
ADVANCES IN PRIVACY PRESERVING FEDERATED LEARNING TO REALIZE A TRULY LEARNING HEALTHCARE SYSTEM

RAVI MADDURI
Senior Scientist
Data Science and Learning Division
Argonne National Laboratory
madduri@anl.gov

ZILINGHAN LI
Machine Learning Research Engineer
Data Science and Learning Division
Argonne National Laboratory
zilinghan.li@anl.gov

Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Washington D.C.
Oct 29., 2024



FUNDING ACKNOWLEDGEMENTS

THIS MATERIAL IS BASED UPON WORK SUPPORTED BY THE U.S. DEPARTMENT OF ENERGY, OFFICE OF SCIENCE, UNDER CONTRACT NUMBER DE-AC02-06CH11357.

KEY FEATURES OF A LEARNING HEALTH SYSTEM (LHS)

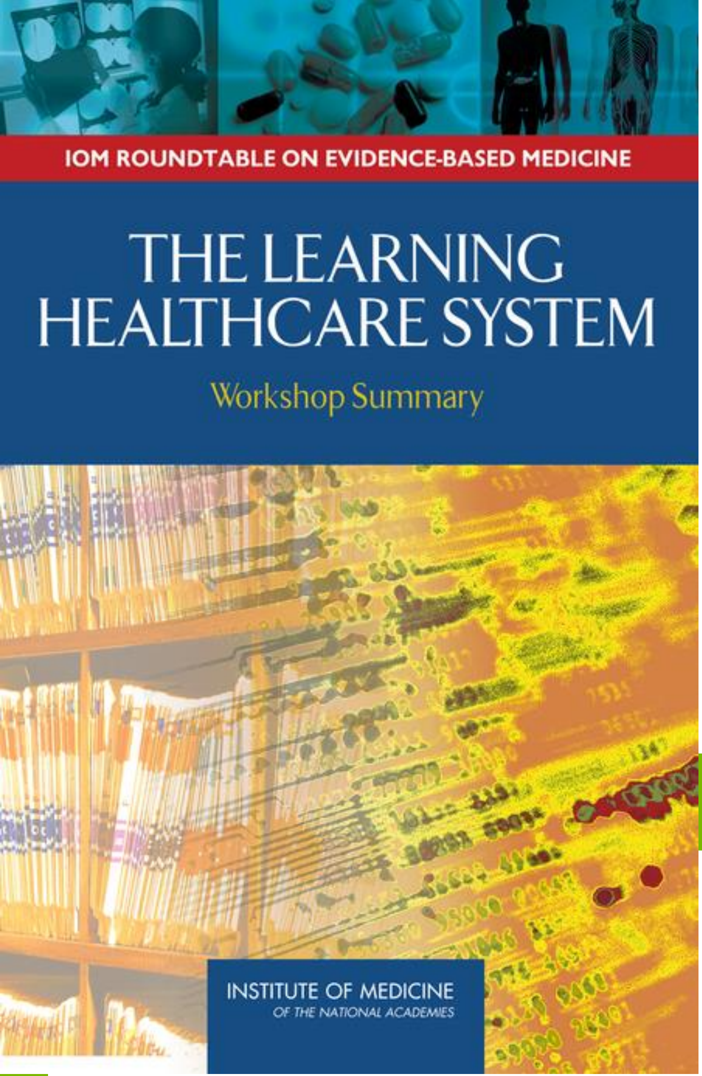
From Institute of Medicine Report

- Data Integration and Interoperability
  - Diverse Data Sources: Combines EHRs, genomics, imaging, and patient-reported outcomes.
- Collaborative Culture
  - Stakeholder Collaboration: Involves clinicians, patients, researchers, and policymakers.
  - Shared Goals: Focuses on improving outcomes collectively.
- Ethical Data Use and Privacy
  - Data Security: Implements strong protections for patient information.
  - Governance: Follows ethical guidelines for data access and use.

KEY FEATURES OF A LEARNING HEALTH SYSTEM (LHS)

From Institute of Medicine Report

- Continuous Learning Cycle
  - Data-Driven Improvement: Routinely collects and analyzes clinical data.
  - Rapid Feedback: Insights are quickly integrated back into practice.
- Feedback Mechanisms
  - Clinician Support: Offers decision support tools and alerts.
  - Patient Feedback: Incorporates patient experiences to refine care delivery.
- Evidence Generation at Point of Care
  - Embedded Research: Research activities are part of clinical workflows.
  - Real-Time Analytics: Provides immediate evidence to inform decisions.
- Patient-Centered Care
  - Active Engagement: Patients participate in their own care and data sharing.
  - Personalization: Care plans reflect individual preferences and needs.



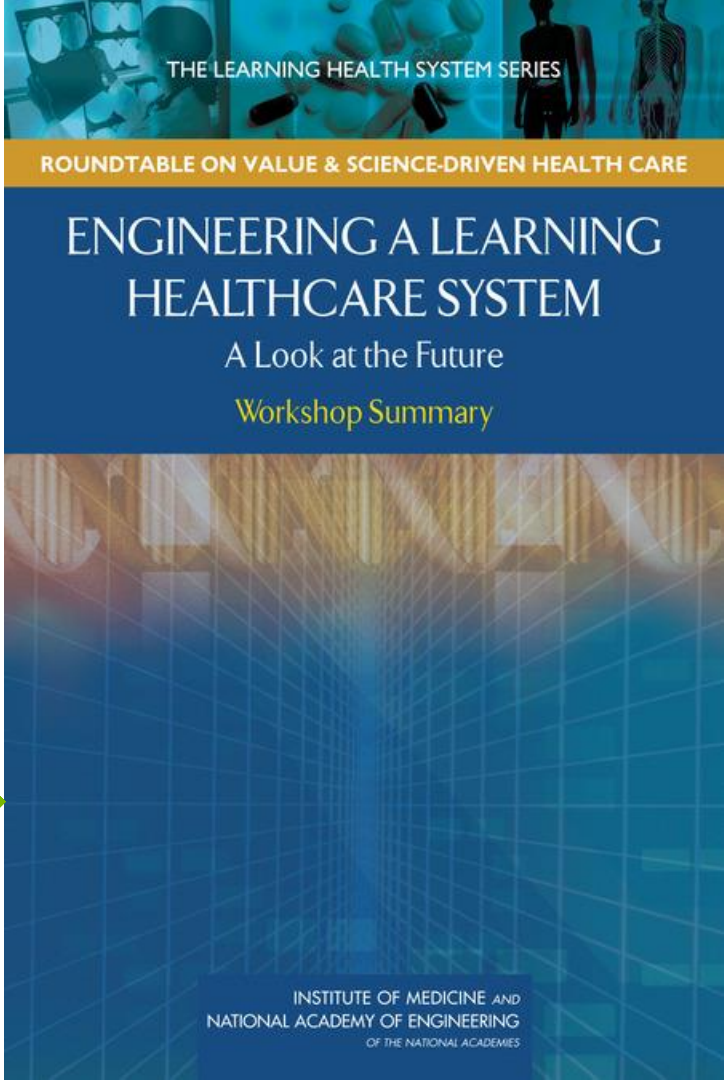
IOM ROUNDTABLE ON EVIDENCE-BASED MEDICINE

THE LEARNING HEALTHCARE SYSTEM

Workshop Summary

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Multimodal machine learning with continuous learning, enhanced collaboration, and patient privacy preservation capabilities can enable LHS.



THE LEARNING HEALTH SYSTEM SERIES

ROUNDTABLE ON VALUE & SCIENCE-DRIVEN HEALTH CARE

ENGINEERING A LEARNING HEALTHCARE SYSTEM

A Look at the Future

Workshop Summary

INSTITUTE OF MEDICINE AND
NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES

CHALLENGES TOWARDS LHS

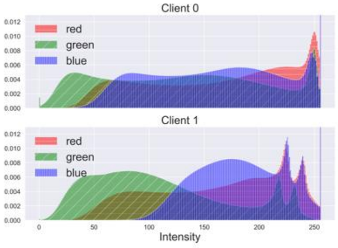
Multimodal and Heterogeneous Medical Data in Distributed Silos

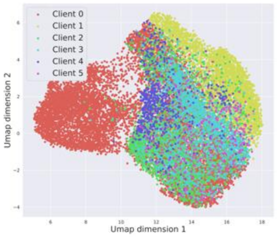
Biomedical Health Data are Multimodal

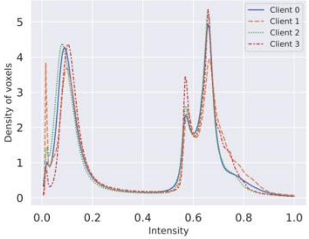
Biomedical health data are multimodal as they include different types like images, texts, signals, and structured data from various sources.

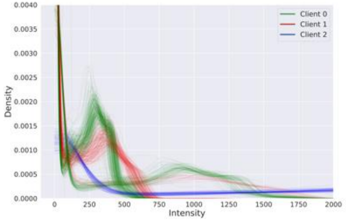
Biomedical Health Data are Heterogeneously Distributed

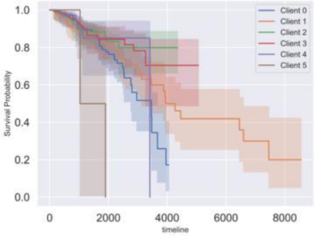
Biomedical Health data are heterogeneous across distributed data silos due to variations in patient populations, medical practices, and data collection processes among hospitals.

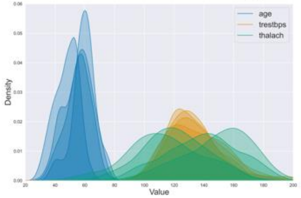












CHALLENGES TOWARDS LHS

Training Multimodal Medical AI models is Essential But Challenging

Training a multimodal medical AI model is essential

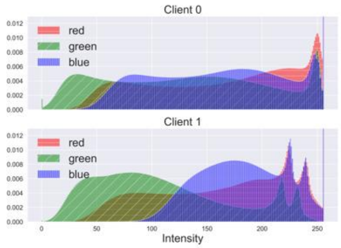
Training multimodal medical AI models is essential because it can leverage diverse data types to provide more comprehensive and accurate insights, improving diagnosis and treatment outcomes.

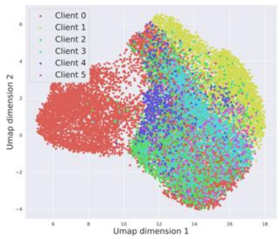
Training a robust multimodal AI model within single data silo is difficult

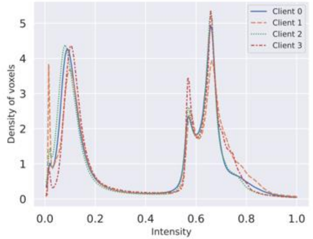
Training robust multimodal AI models within a single data silo is difficult due to the homogeneous patient population and limited data modalities, hindering generalization to real-world scenarios.

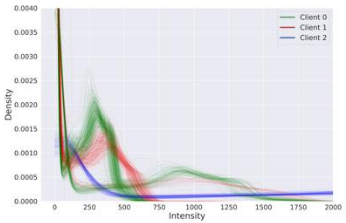
Collecting medical data centrally is challenging

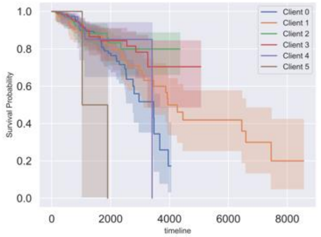
Collecting the distributed medical data centrally is challenging due to privacy concerns and varying regulations (e.g. HIPAA).

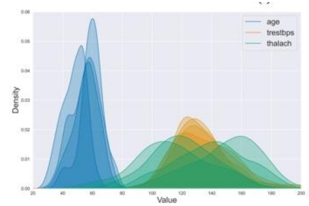












MULTIMODAL FL APPROACHES TO LHS

Federated Learning Offers a Viable Solution

- Federated learning (FL) is a distributed learning paradigm with multiple data silos as clients and a central server for orchestration.

- Each FL client has its own computing facilities and trains model using their ***private*** local data.

- Each FL client only shares the ***model*** trained on their private local data to the FL server for aggregation.

- A global model is obtained by aggregating models from different FL clients, thus implicitly leveraging private data from various data silos.

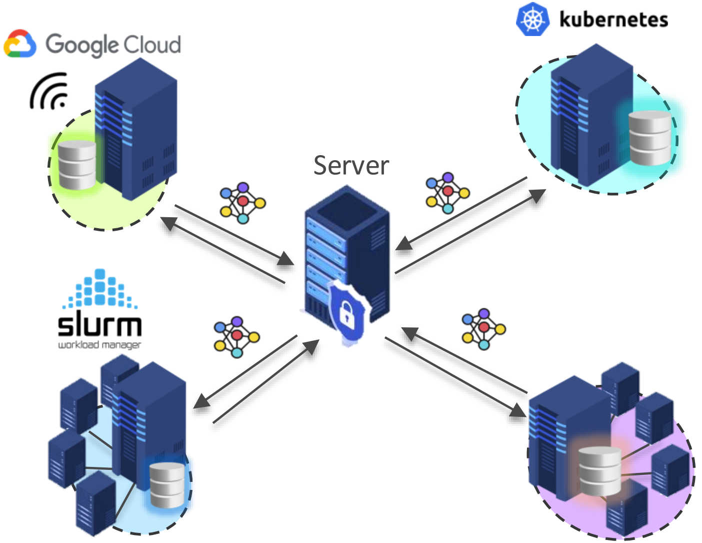


Fig. Federated Learning Illustration.

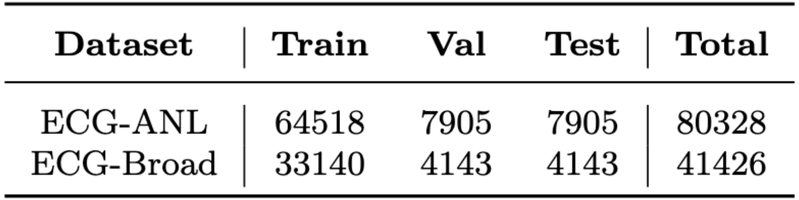
CASE STUDIES IN FEDERATED LEARNING FOR BIOMEDICINE

Federated Learning Shows Effectiveness in Uni-modal Medical Models

Challenge: Predicting age from ECG signals.

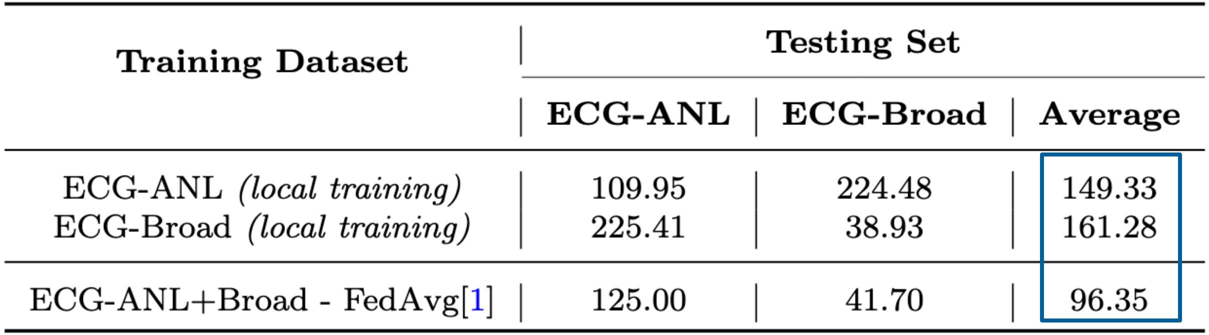
Two Clients

- ECG-ANL: the dataset is aggregated from multiple open-source datasets.
- ECG-Broad: private dataset collected by Broad Institute.



Dataset	**Train**	**Val**	**Test**	**Total**
ECG-ANL	64518	7905	7905	80328
ECG-Broad	33140	4143	4143	41426

FL can obtain a model that performs relatively well on both datasets.



Training Dataset	**Testing Set**		
	ECG-ANL	**ECG-Broad**	**Average**
ECG-ANL *(local training)*	109.95	224.48	149.33
ECG-Broad *(local training)*	225.41	38.93	161.28
ECG-ANL+Broad - FedAvg[1]	125.00	41.70	96.35

Mean Squared Error (the lower, the better)

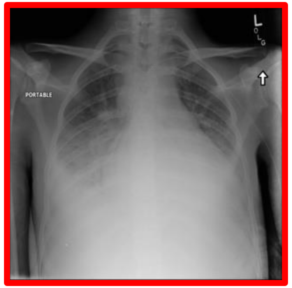
CASE STUDIES IN FEDERATED LEARNING FOR BIOMEDICINE

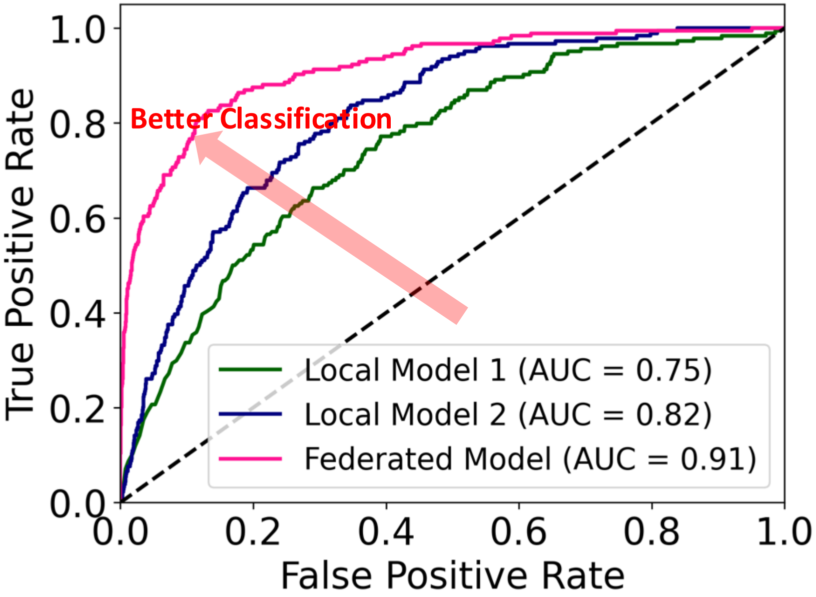
Federated Learning Shows Effectiveness in Uni-modal Medical Models

Challenge: COVID-19 prediction from Chest X-Rays.

Two Clients

- ANL-COVID: the dataset is aggregated from multiple open-source datasets.
- UChicago-COVID: private dataset collected by UChicago.





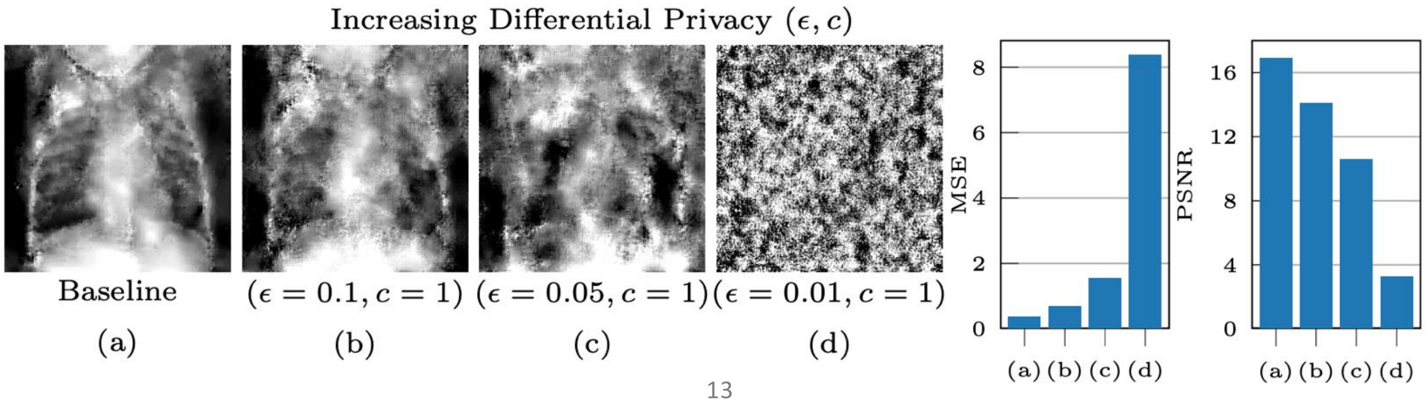
CASE STUDIES IN FEDERATED LEARNING FOR BIOMEDICINE

Federated Learning Shows Effectiveness in Uni-modal Medical Models

Challenge: Preventing the reconstruction of Chest X-Rays from model gradients

- Federated learning itself is not privacy preserving – as the training data can be reversely constructed from model gradients.

- Differential privacy (DP) techniques, can significantly increase the difficulty of reconstruction by adding noises to model parameters



VISION

From Uni-modal to Multimodal Privacy-Preserving Federated Learning

The examples only train uni-modal biomedical models using privacy-preserving federated learning, which limits the range of applications.

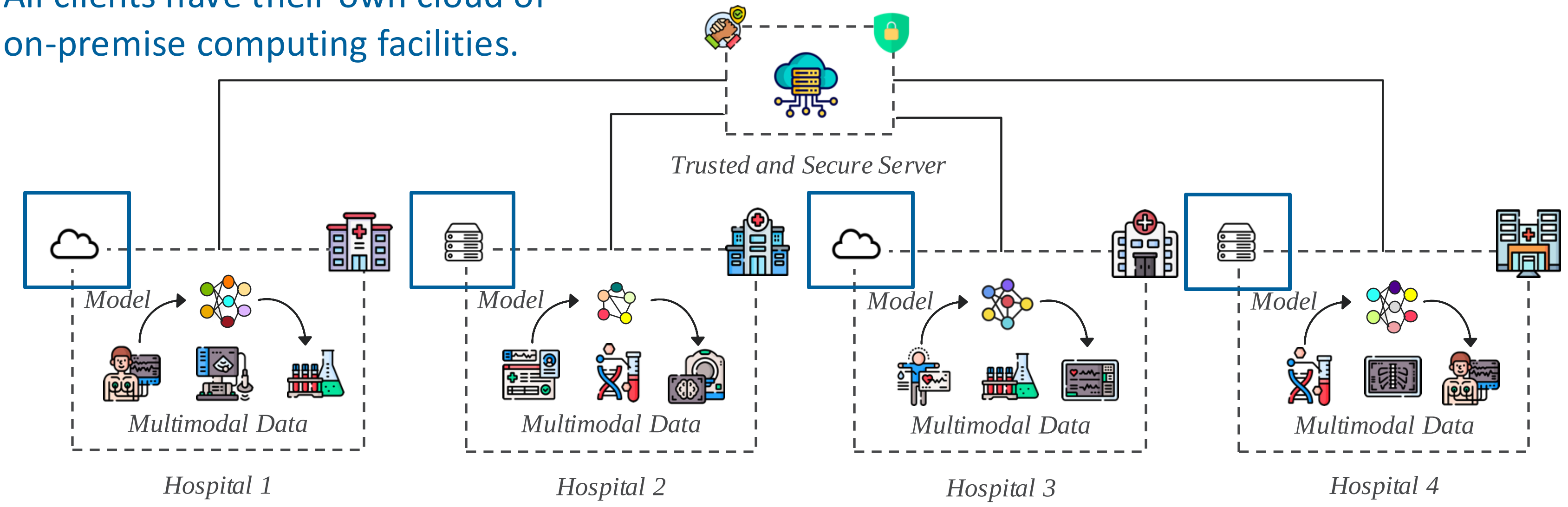
➔ Here we envision an ideal federated learning framework to:

- ❖ Enable secure and privacy-preserved training of multimodal biomedical models.
- ❖ Continuously update the trained models as new data accumulate.
- ❖ Help realize a truly learning healthcare system.

VISION

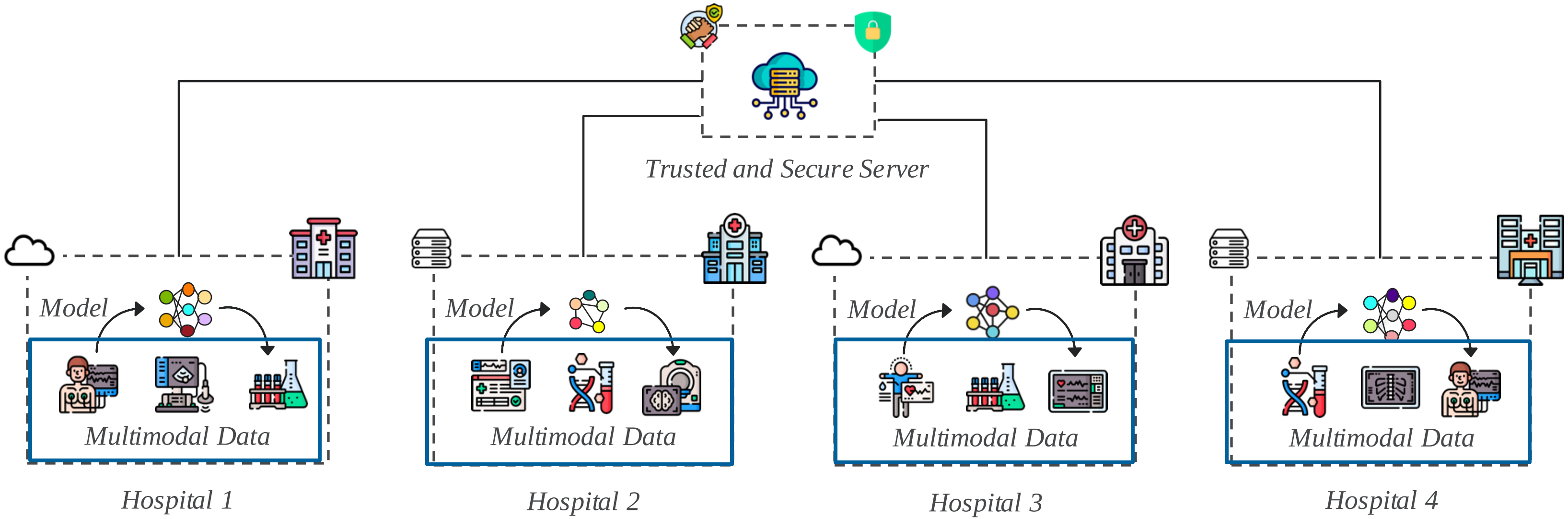
From Uni-modal to Multimodal Privacy-Preserving Federated Learning

All clients have their own cloud or on-premise computing facilities.



VISION

From Uni-modal to Multimodal Privacy-Preserving Federated Learning

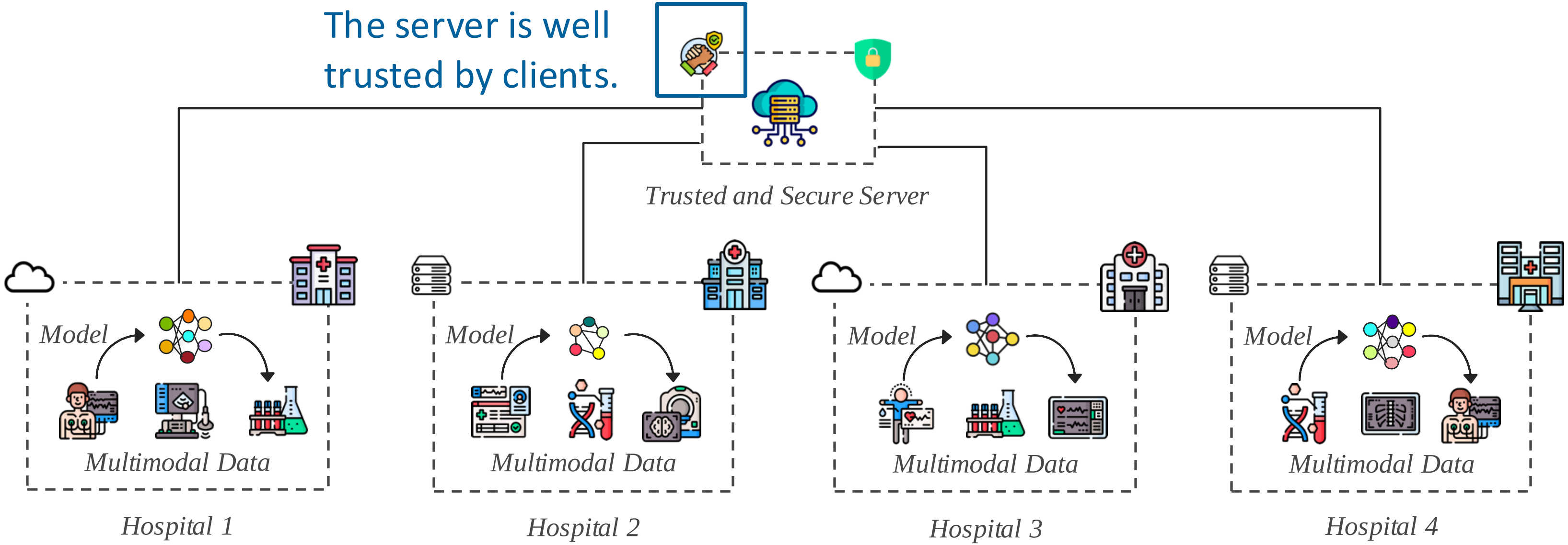


Different clients have various data modalities for model training.

VISION

From Uni-modal to Multimodal Privacy-Preserving Federated Learning

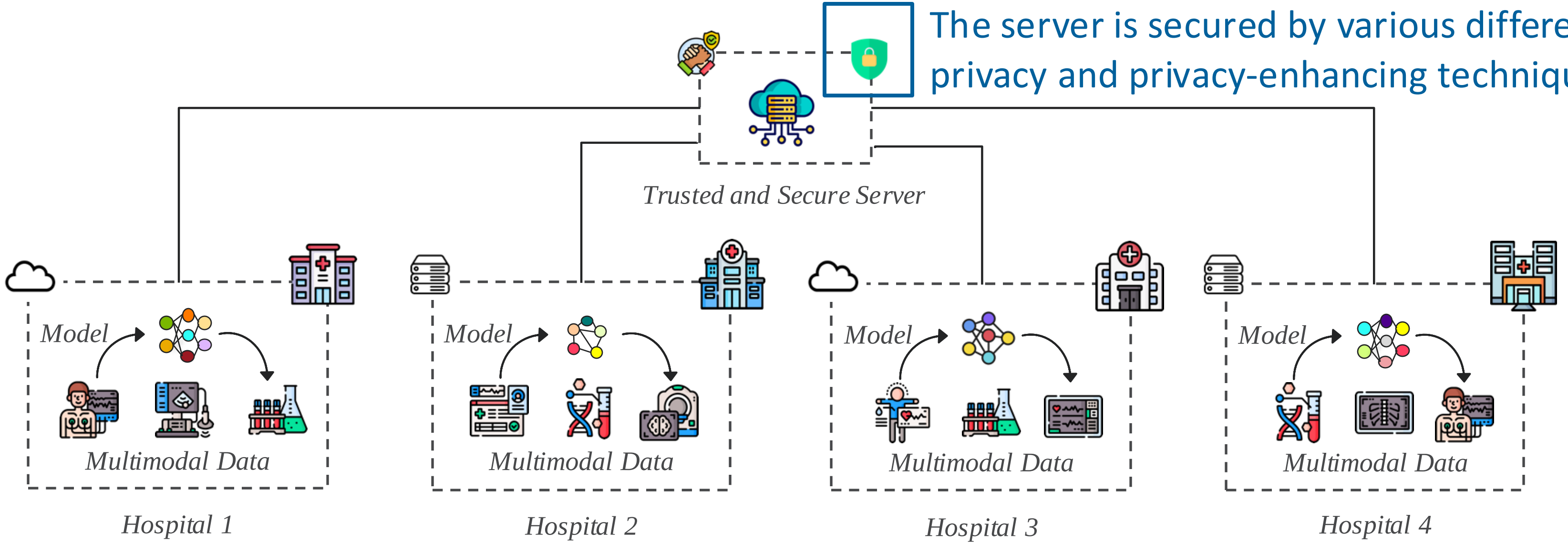
The server is well trusted by clients.



VISION

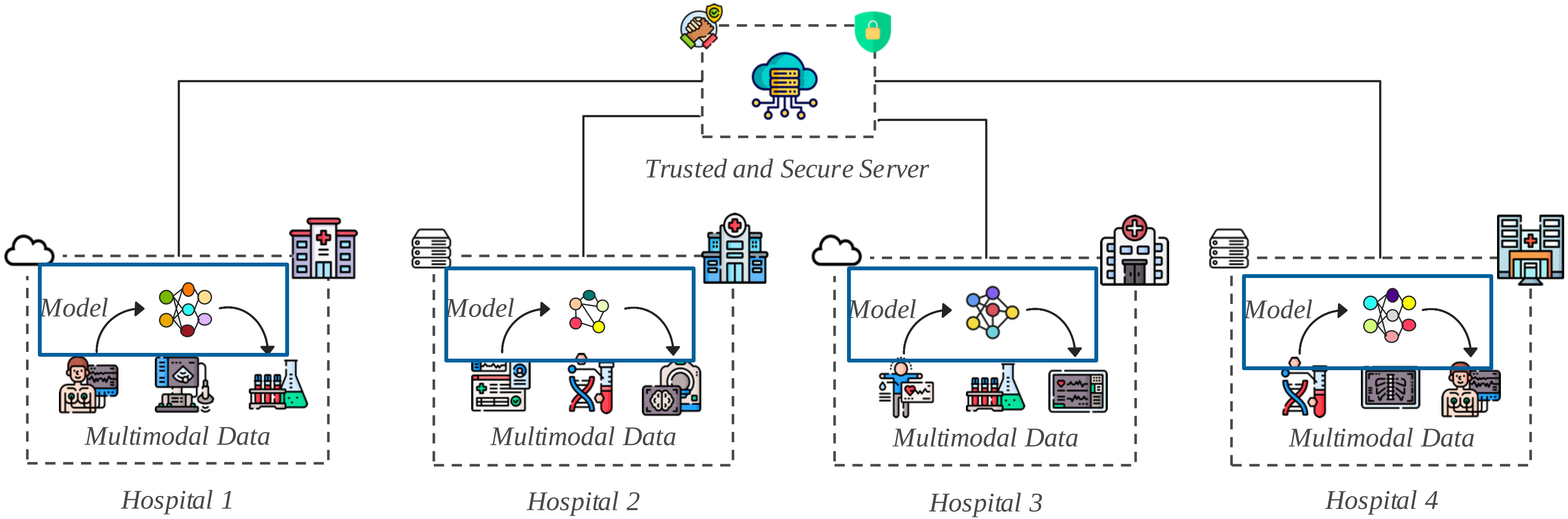
From Uni-modal to Multimodal Privacy-Preserving Federated Learning

The server is secured by various differential privacy and privacy-enhancing techniques.



VISION

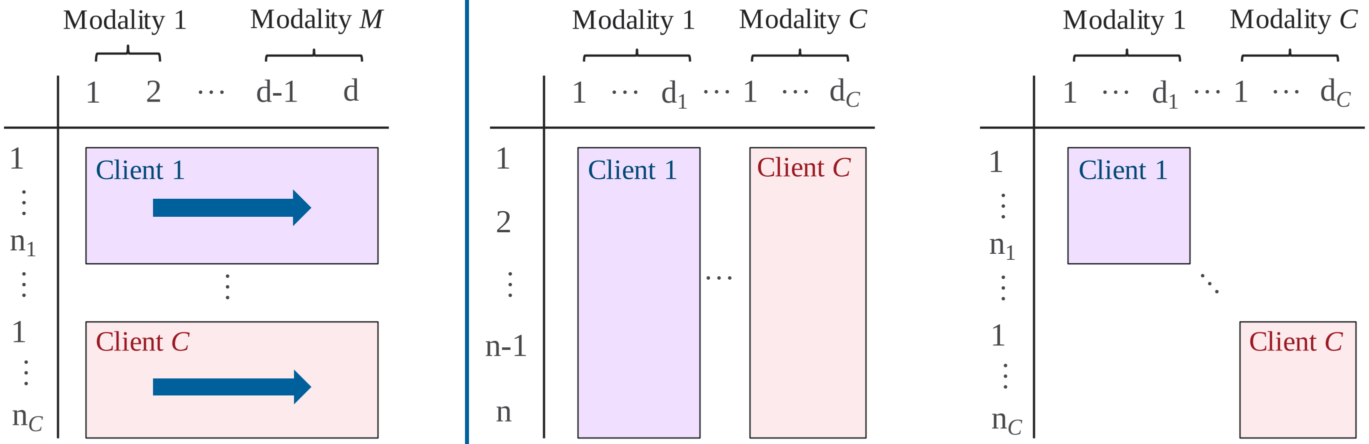
From Uni-modal to Multimodal Privacy-Preserving Federated Learning



The models can be continuously updated to adapt to any shifts in data distributions, availability of new data, and evolving health trends in real-time.

BUILDING BLOCKS

Multimodal Federated Learning



(a) Same multimodal feature space and different sample spaces

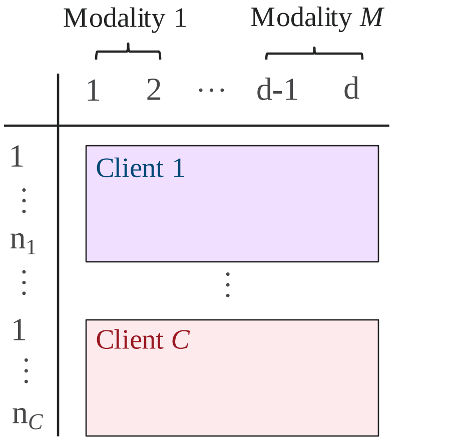
(b) Same sample space and different multimodal feature spaces

(c) Different sample spaces and multimodal feature spaces

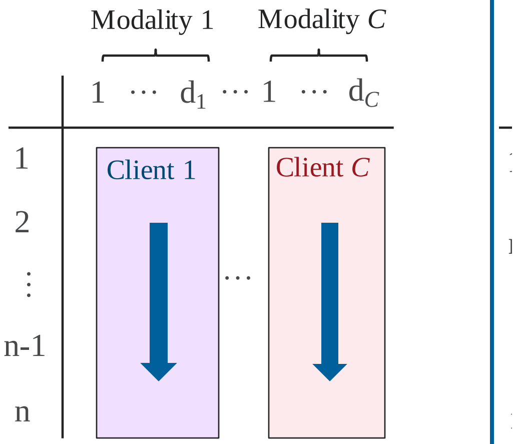
All clients have the same data modalities: All clients train the same multimodal modal using traditional federated learning. ➔ Horizontal Federated Learning.

BUILDING BLOCKS

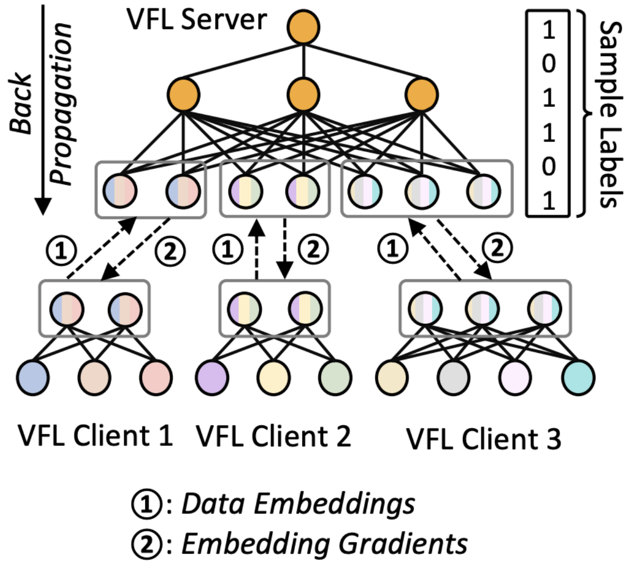
Multimodal Federated Learning



(a) Same multimodal feature space and different sample spaces



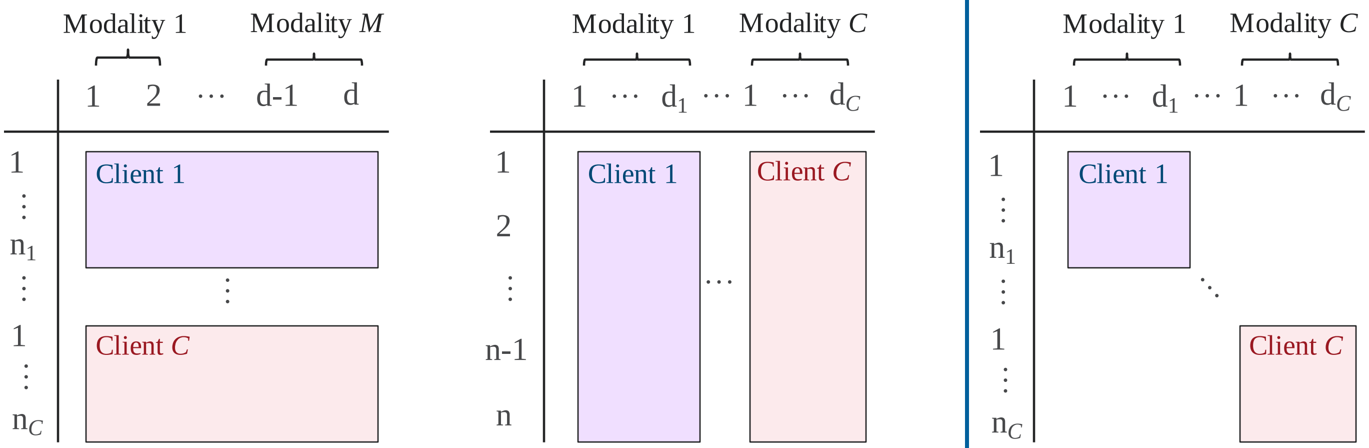
(b) Same sample space and different multimodal feature spaces



All clients have different modalities but the same sample space (rare in the real-world): Different clients train different embedding models for different modalities by sharing the data embeddings to the server and updating the local embedding models using embedding gradients sent back from the server.→ Vertical Federated Learning.

BUILDING BLOCKS

Multimodal Federated Learning



(a) Same multimodal feature space and different sample spaces

(b) Same sample space and different multimodal feature spaces

(c) Different sample spaces and multimodal feature spaces

All clients have different modalities and different sample spaces: Common in real-world scenarios, but very challenging ➔ Potential solutions involve training small embedding models locally, and leveraging synthetic multimodal data or publicly available multimodal data for sample alignment and further tuning.

BUILDING BLOCKS

Hierarchical Federated Learning

As training data are vulnerable to data reconstruction, especially without privacy-enhancing techniques:

- Some clients (e.g. small clinics) might not be willing to share their local models with the central server.

- Those small clinics might only be willing to share the local model with some large trusted local medical institutions/university ➔ this prevents the creation of a large federation.

- Hierarchical federated learning enables collaborative training beyond the trust boundary.

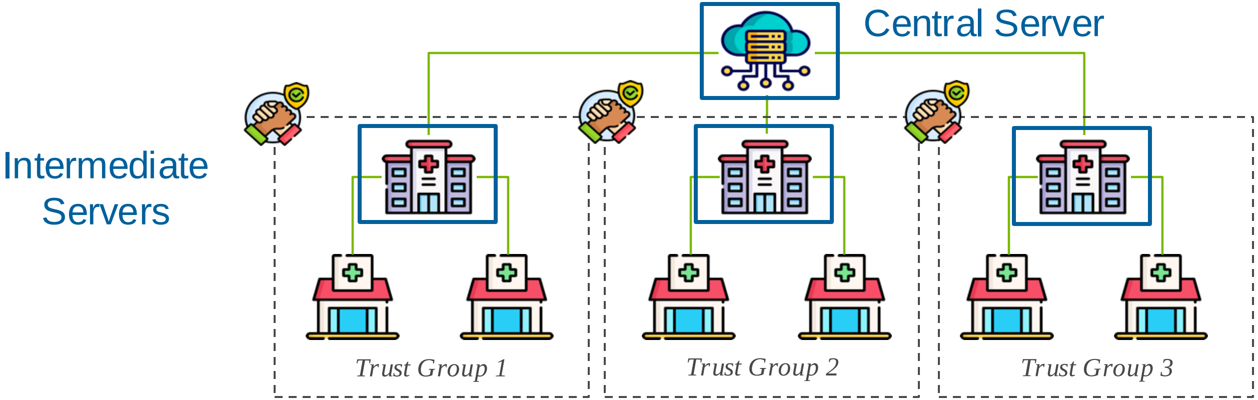
BUILDING BLOCKS

Hierarchical Federated Learning

- Each client first sends its local model to a trusted intermediate server.

- The intermediate server then sends a model aggregated from several clients to the central server.

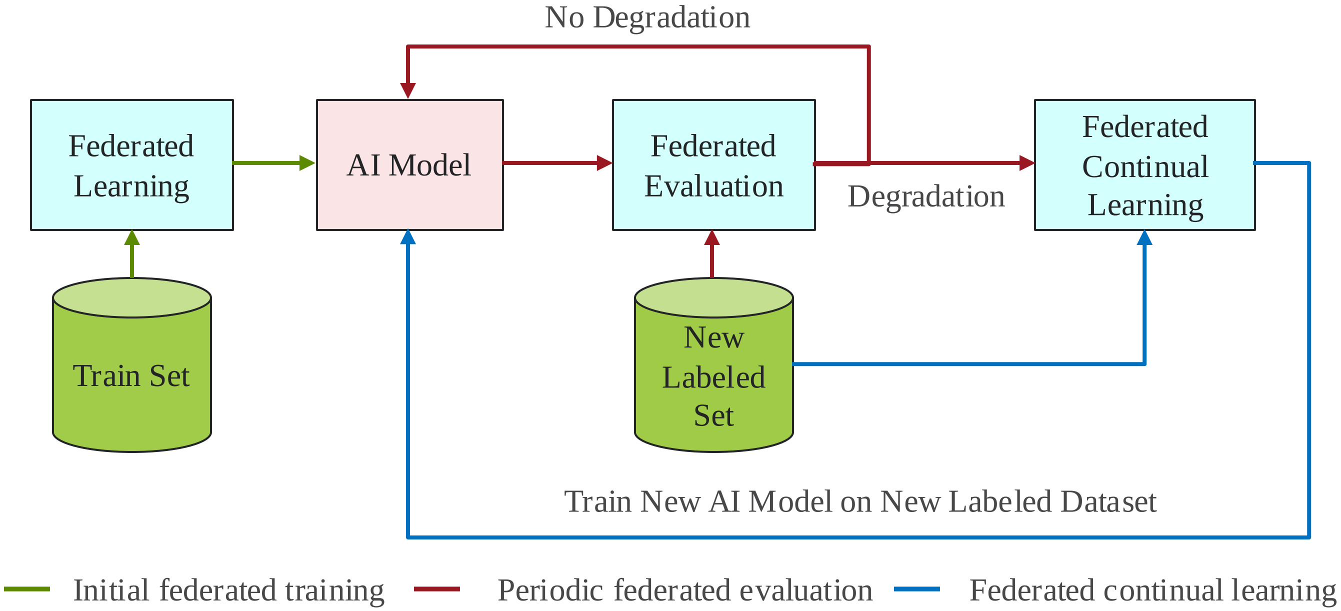
- As the aggregated model contains the information of several clients, it is hard to reconstruct the training data of any single client.



BUILDING BLOCKS

Federated Continuous Learning

Federated continuous learning ensures that the models evolve to reflect the diversity of healthcare settings and populations, reducing biases and enabling better generalization.



No Degradation

Federated Learning

AI Model

Federated Evaluation

Degradation

Federated Continual Learning

Train Set

New Labeled Set

Train New AI Model on New Labeled Dataset

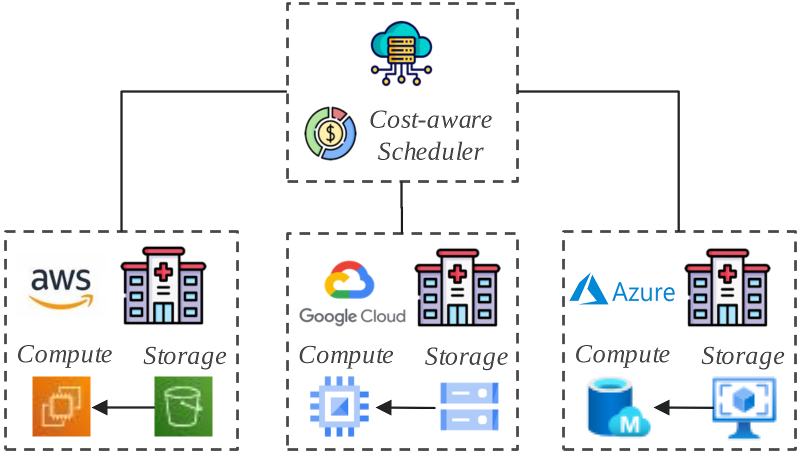
Initial federated training

Periodic federated evaluation

Federated continual learning

BUILDING BLOCKS

Cost-Aware FL on the Cloud



- Many hospitals have their private data on Cloud Storage (S3, Globus Cloud Storage, etc.) and have their computing on the Cloud as well.

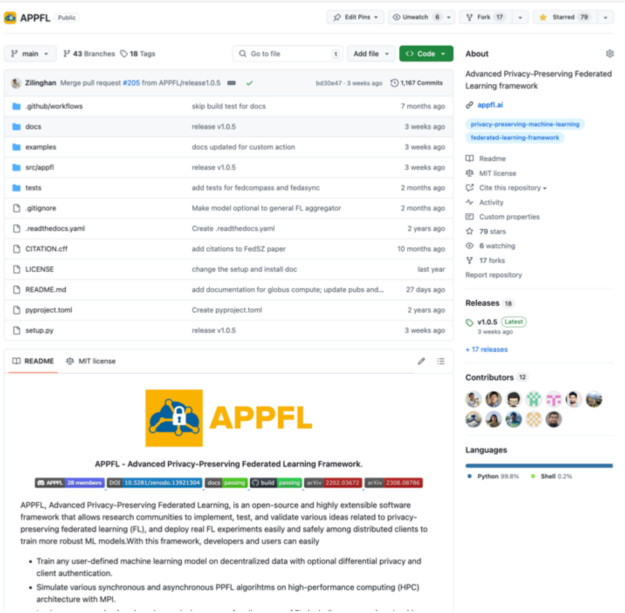
- Training on GPU cloud instances can be costly.

- AWS, Google, and Azure all have "spot computing" – AWS Spot Instances, Google Cloud Preemptable VMs, and Azure Spot VMs, which provide a low-cost computing option, but can be killed at any time with short notice.

- It would be greatly beneficial to have a cost-aware scheduler for the FL server to reduce the cost for FL experiments among heterogeneous cloud computing providers using their spot instances, while maintaining robust and tolerant to potential client failures.

REFERENCES

- Hoang, Trung-Hieu, Jordan Fuhrman, Ravi Madduri, Miao Li, Pranshu Chaturvedi, Zilinghan Li, Kibaek Kim et al. "Enabling end-to-end secure federated learning in biomedical research on heterogeneous computing environments with APPFLx." *arXiv preprint arXiv:2312.08701* (2023).
- Li, Zilinghan, Shilan He, Ze Yang, Minseok Ryu, Kibaek Kim, and Ravi Madduri. "Advances in APPFL: A Comprehensive and Extensible Federated Learning Framework." *arXiv preprint arXiv:2409.11585* (2024).
- Madduri, Ravi, Zilinghan Li, Tarak Nandi, Kibaek Kim, Minseok Ryu, and Alex Rodriguez. "Advances in Privacy Preserving Federated Learning to Realize a Truly Learning Healthcare System." *arXiv preprint arXiv:2409.19756* (2024).
- https://github.com/APPFL/APPFL
- https://appfl.ai



Zilinghan Merge pull request #205 from APPFL/release1.0.5 ✓ bd30e47 · 3 weeks ago 1,167 Commits

.github/workflows	skip build test for docs	7 months ago
docs	release v1.0.5	3 weeks ago
examples	docs updated for custom action	3 weeks ago
src/appfl	release v1.0.5	3 weeks ago
tests	add tests for fedcompass and fedasync	2 months ago
.gitignore	Make model optional to general FL aggregator	2 months ago
.readthedocs.yaml	Create .readthedocs.yaml	2 years ago
CITATION.cff	add citations to FedSZ paper	10 months ago
LICENSE	change the setup and install doc	last year
README.md	add documentation for globus compute; update pubs and...	27 days ago
pyproject.toml	Create pyproject.toml	2 years ago
setup.py	release v1.0.5	3 weeks ago

About

Advanced Privacy-Preserving Federated Learning framework

appfl.ai

privacy-preserving-machine-learning
federated-learning-framework

Readme
MIT license
Cite this repository
Activity
Custom properties
79 stars
6 watching
17 forks
Report repository

Releases 18

v1.0.5 Latest
3 weeks ago

+ 17 releases

Contributors 12

Languages

Python 99.8% Shell 0.2%

APPFL

APPFL - Advanced Privacy-Preserving Federated Learning Framework.

APPFL 28 members | DOI 10.5281/zenodo.13921304 | docs passing | build passing | arXiv 2202.03672 | arXiv 2308.08786

APPFL, Advanced Privacy-Preserving Federated Learning, is an open-source and highly extensible software framework that allows research communities to implement, test, and validate various ideas related to privacy-preserving federated learning (FL), and deploy real FL experiments easily and safely among distributed clients to train more robust ML models.With this framework, developers and users can easily

- Train any user-defined machine learning model on decentralized data with optional differential privacy and client authentication.
- Simulate various synchronous and asynchronous PPFL algorihtms on high-performance computing (HPC) architecture with MPI.

THANK YOU!