AI4PH Federated Learning Workshop

JEAN-PAUL R. SOUCY

DATA SCIENCE TEAM, MCGILL UNIVERSITY HEALTH CENTRE

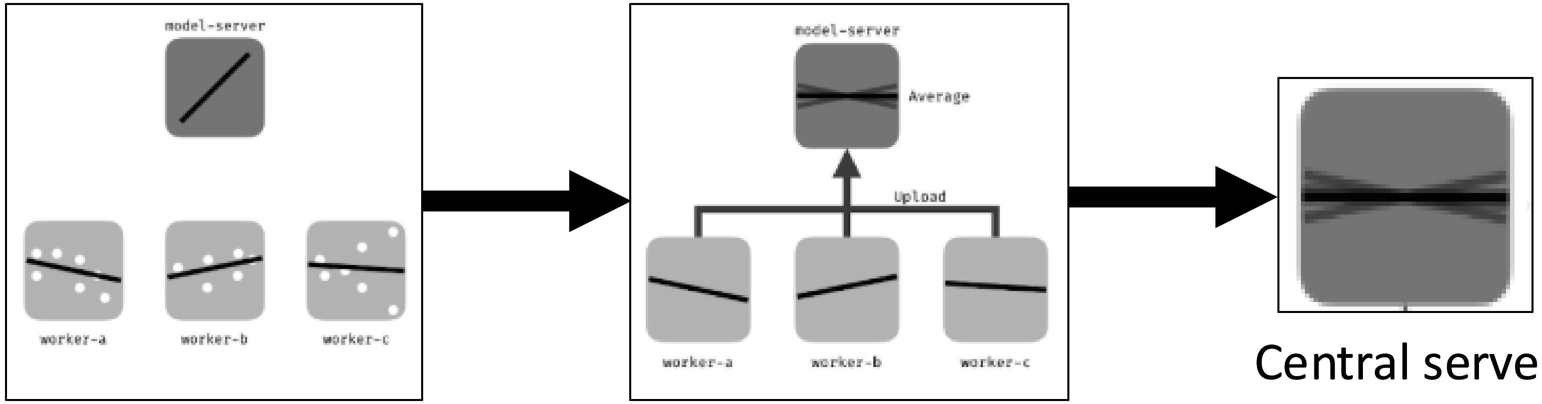
2025-07-15

Workshop objectives

- Develop an understanding of the technical foundations of federated learning
- Explore the challenges of preparing multi-site datasets for federated learning
- Compare the performance and fairness of machine learning models applied to pooled versus federated datasets
- Support participant skill development in applying machine learning models to classification problems in different contexts

What is federated learning?

- A collaborative, privacy-preserving approach to training machine learning models across decentralized data sources by sharing model updates instead of raw data



Train local models on local data

Model information shared to central server

Central server aggregates information into global model

Flower Framework

Types of federated learning

Horizontal FL

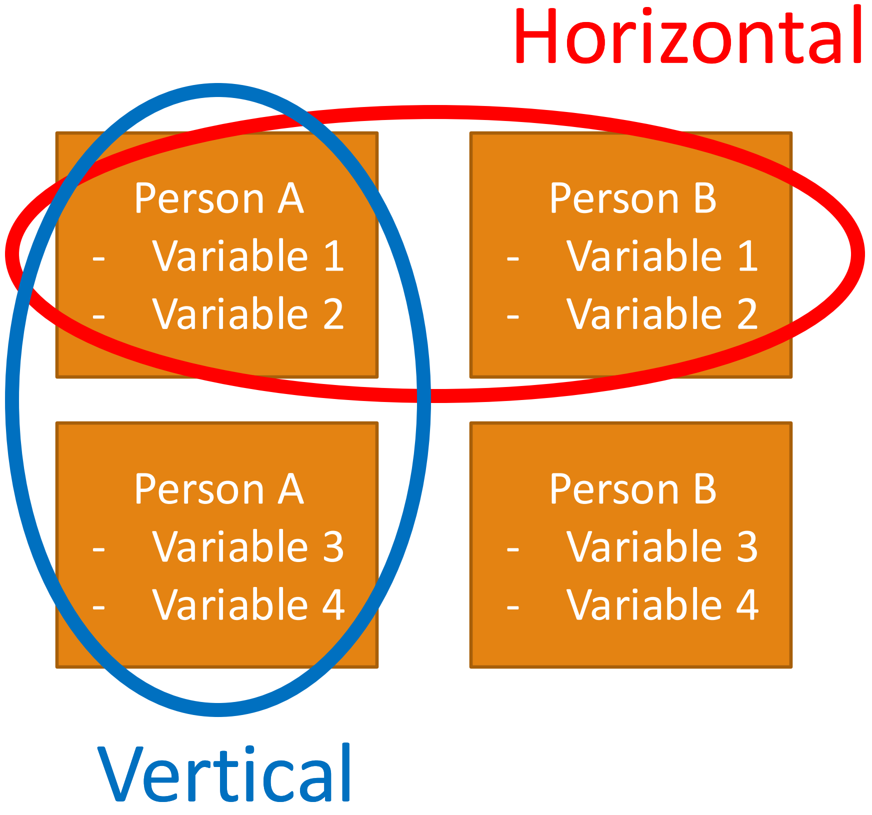
- Shared features but different samples

Vertical FL

- Shared samples but different features

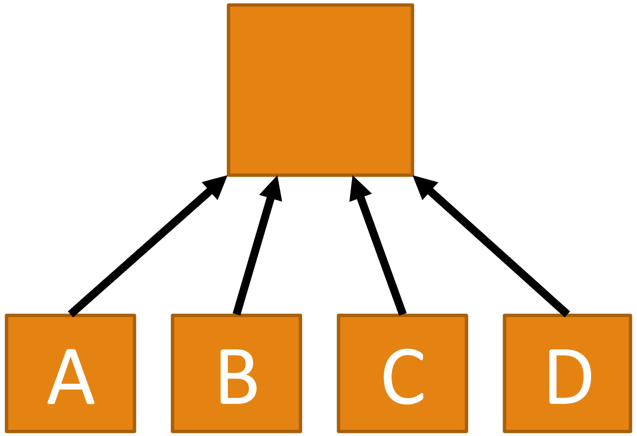
Hybrid FL

- Partially overlapping samples and features

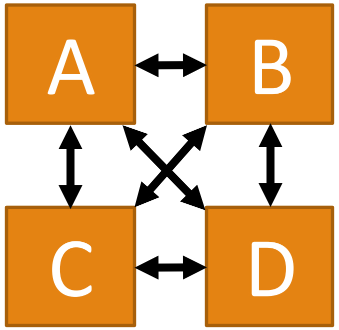


Federated learning architectures

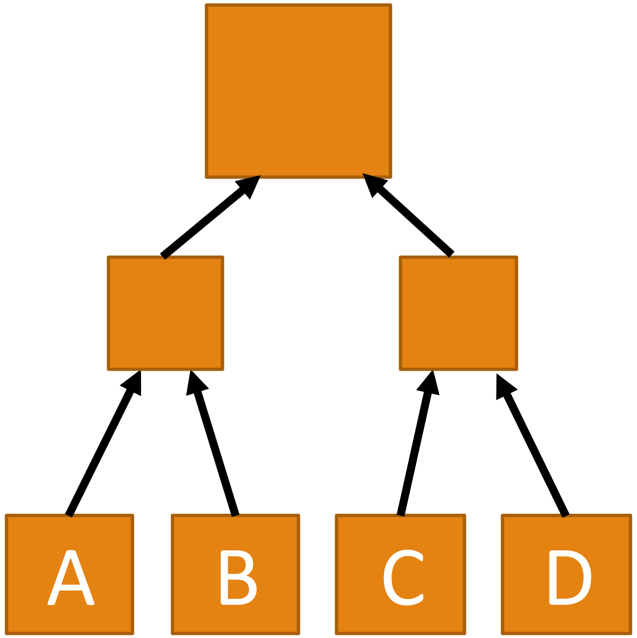
Centralized



Decentralized



Hierarchical



FedAvg

- Baseline FL algorithm (McMahan et al., 2017)
- Clients train locally on private data
- Server aggregates via weighted average of client models
- Simple, scalable, but sensitive to non-IID heterogeneity in data

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

McMahan et al., 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Heterogeneity and related challenges

- Heterogeneity (Kairouz et al., 2021):
  - Label distribution skew (prior probability shift)
  - Feature distribution skew (covariate shift)
  - Concept drift: Same label, different features
  - Concept shift: Same features, different labels
  - Sample size imbalance
- Client drift: local updates deviate in incompatible directions under skewed data

Kairouz et al., 2021. Advances and Open Problems in Federated Learning. *arXiv*. doi: 10.48550/arXiv.1912.04977

Other federated learning algorithms

- FedProx: FedAvg with a proximal term (μ)
- pFedMe: Personalized models per site w/ global alignment
- APFL: Adaptive mixture of local and global model
- Clustered FL: Clients grouped into clusters
- Distributed ensemble and stacking
- Privacy-preserving integration (e.g., MPC, HE)

Evaluation metrics

- Standard performance metrics: accuracy, precision, recall, F1 score, AUC, etc.
- Cross-client metrics: Weighted average across all clients
- System metrics: Convergence time, compute time, cost
- Privacy: Differential privacy (ε)
- Robustness: Byzantine robustness

Fairness metrics

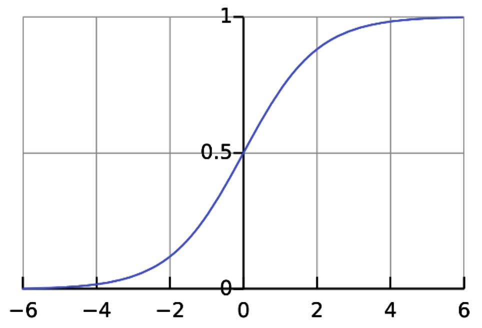
- Client-level fairness
  - Worst-client performance, cross-client variance
  - Gini index, Jain's fairness index
  - Group size ratio (GSR), predicted positive rate (PPR), false discovery rate (FDR)
- Subgroup-level fairness
  - E.g., Performance across different demographic groups

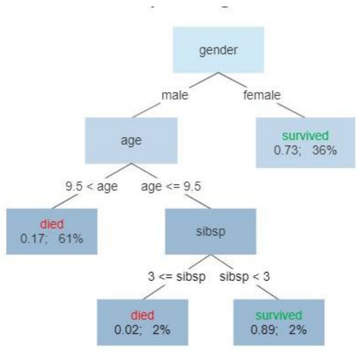
Key takeaways

- Federated learning enables collaborative ML while preserving data privacy
- Data heterogeneity (non-IID) is a core challenge, leading to client drift
- Many alternatives beyond the basic FedAvg have been proposed to address heterogeneity and other concerns
- Evaluation should consider subgroup and client fairness
- Every model has tradeoffs in performance, privacy, etc.

What you'll be doing in the data challenge

- **FedAvg**: simplified version using *glm* (logistic regression)
- **Distributed ensemble** (Random forest, XGBoost)
- Pooled models in R (any model)





FedAvg

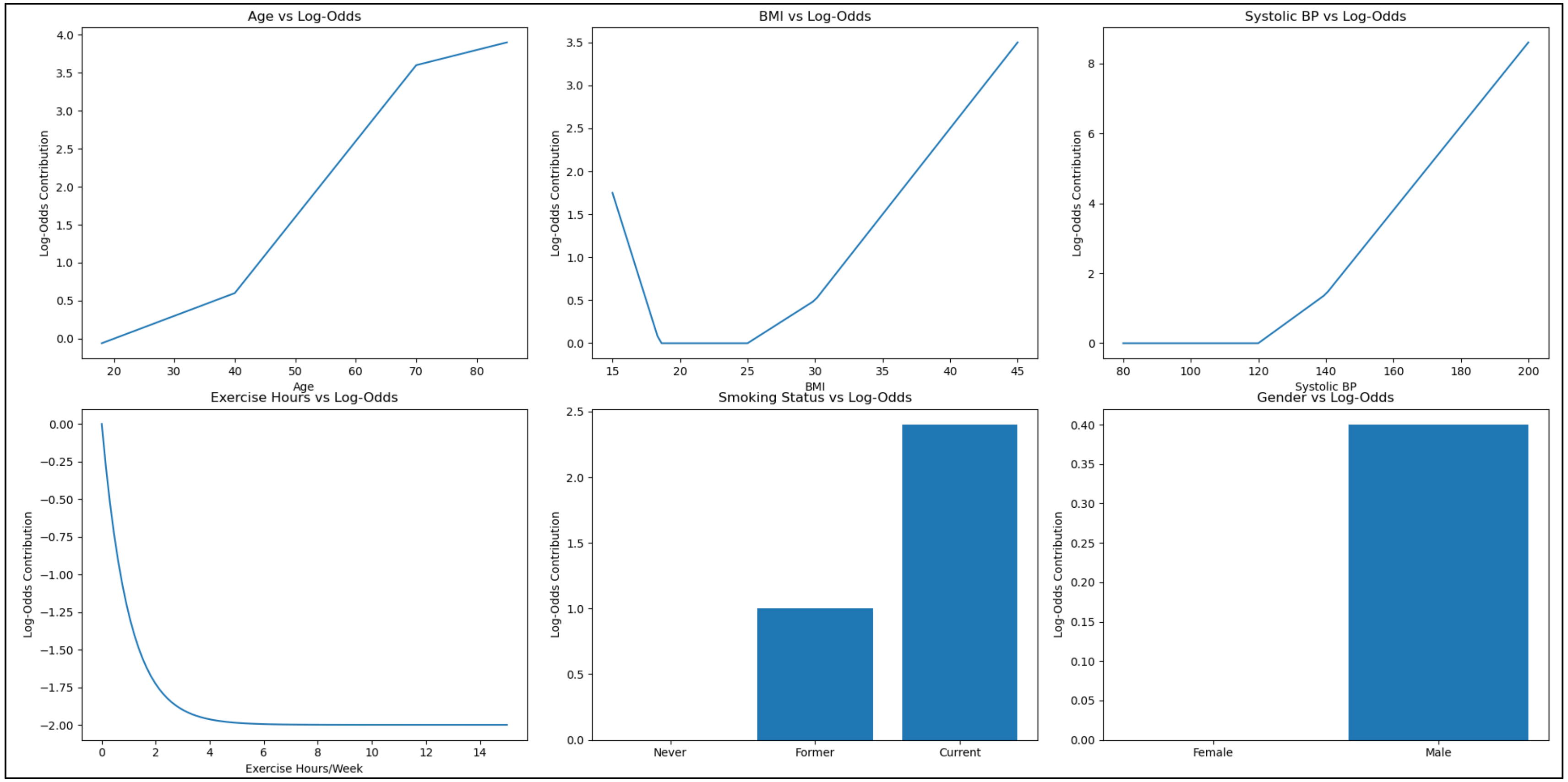
- Single round w/ all clients
- Global model is weighted average of site-specific LR coefficients

Distributed ensemble

- Single round w/ all clients
- Predicted probs. are weighted average of site-specific probs. (soft voting)

Tutorial: Simulated dataset

- Outcome: CVD
- Three sites:
  - Urban academic centre
  - Rural community clinic
  - Suburban practice
- Categorical
  - Smoking status, Gender
- Continuous:
  - Age, BMI, Systolic BP, Exercise hours



Tutorial: Simulated dataset

- Covariates are not IID across sites