EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

Privacy Preserving Federated Unsupervised Domain Adaptation with Application to Age Prediction from DNA Methylation Data

Cem Ata Baykara

Ali Burak Ünal

Nico Pfeifer

Mete Akgün

Medical Data Privacy and Privacy Preserving Machine Learning

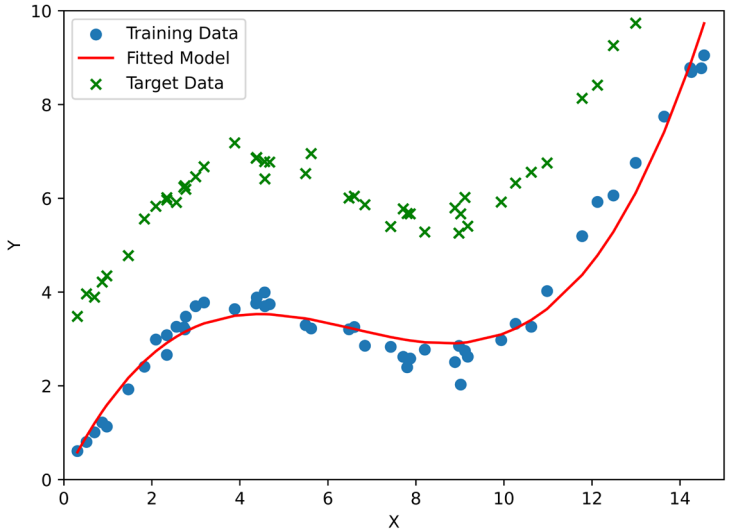
Short Bio

- B.Sc. in Computer Science at Bilkent University in Turkey, 2015
- M.Sc. in Computer Science at Bilkent University in Turkey, 2017
  - Identification of cancer patient subtypes using graph kernels and multi-view kernel clustering on patient omics data
- Ph.D. in Computer Science at the University of Tübingen in Germany, 2022
  - Development of privacy preserving machine learning algorithms for medical applications
- **Currently** postdoctoral researcher in Medical Data Privacy and Privacy Preserving Machine Learning (MDPPML) group at the University of Tübingen
  - Deep neural networks in federated learning settings

Heterogeneous Data

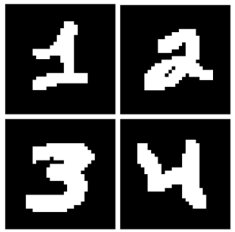
Standard machine learning applications assume training (source) data has the same joint probability distribution as (target) data on which the model is later applied.

This is often not the case in real life!



Domain Adaptation

The domain shift problem is categorized into two types, depending on the availability of labels from the target domain.

Source Domain	Target Domain	Target Domain
		
Always Labelled	Labelled	Unlabelled
{ 1, 2, 3, 4 }	{ 1, 2, 3, 4 }	{ }

Generally, quite challenging!

Unsupervised Domain Adaptation

- Domain adaptation is well studied in computer vision
  - mostly on image classification problems
  - mainly through deep neural networks
- Not well studied in
  - regression problems
  - setting where # features ≫ # samples

} Relevant in medical data

MNIST Dataset

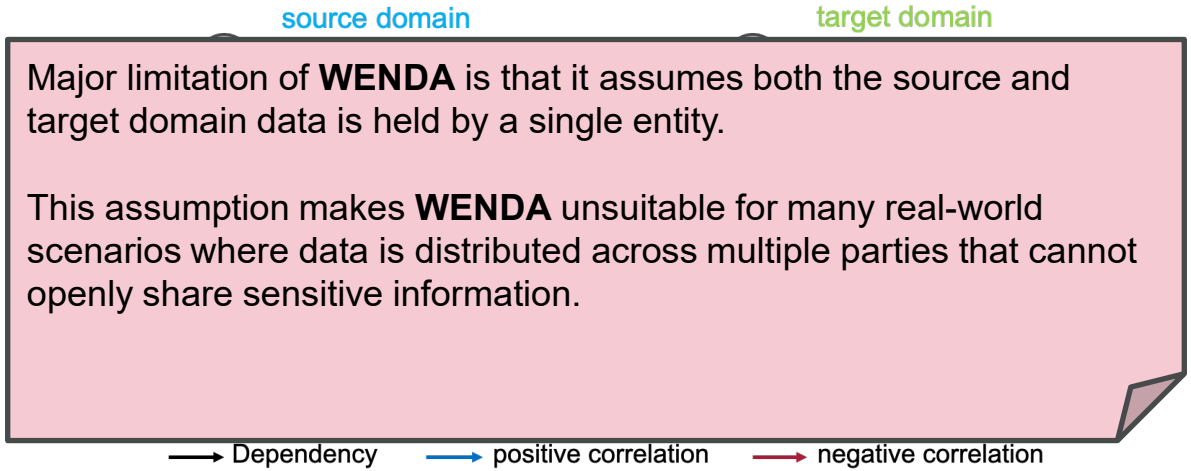
SVHN Dataset

Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data

State-of-the-art solution

WENDA (**W**eighted **e**lastic **n**et for unsupervised **d**omain **a**daptation)

Handl, L., Jalali, A., Scherer, M., Eggeling, R., & Pfeifer, N. (2019). Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics*, *35*(14), i154-i163.

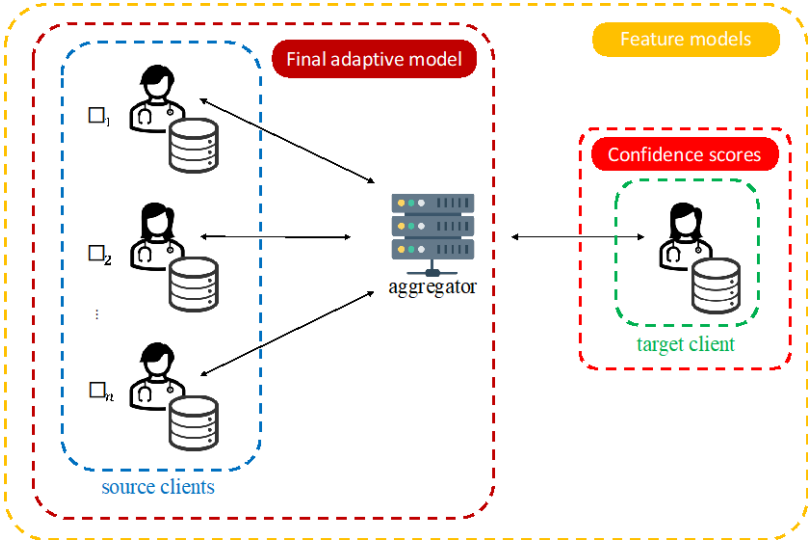


- Compares input dependency structure in the source and target domains.
- Enforces penalization on features behaving differently.
- Forces the model to rely more on inputs that behave similar.

Privacy Preserving Federated Unsupervised Domain Adaptation

Our solution

FREDA (**F**eder**a**te**d** Unsupervised **D**omain **A**daptation)



Motivation

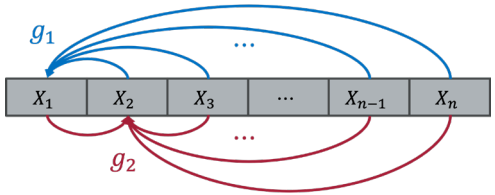
Allow multiple data owners to perform unsupervised domain adaptation without sharing their private data.

By using

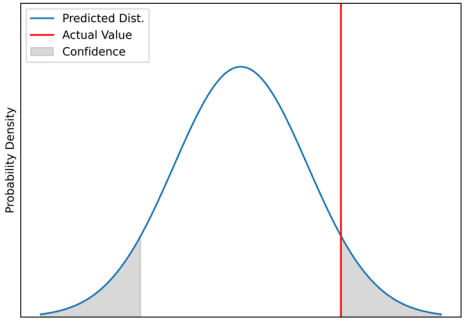
Randomized Encoding, Secure Aggregation, Federated Learning

What are the challenges?

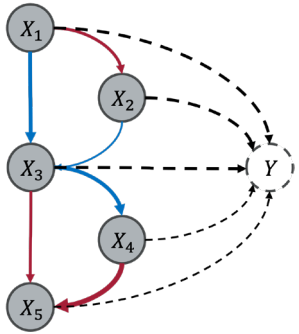
FREDA is a complex method which consists of 3 components



Feature Models



Confidence Scores



Final Adaptive Model

Challenge of Feature Models

Given training (source) data $x = [x_1, \dots, x_n]$ with targets $y = [y_1, \dots, y_n]$, with noise in each point $\mathcal{N}_{0,\sigma}$ and new data (target) points $x_* = [x_1^*, \dots, x_m^*]$ for which we want to predict y_*, Gaussian Process regressor is trained as follows:

$$\mathcal{N}(K_* K^{-1} y, K_{**} - K_* K^{-1} K_*^T)$$

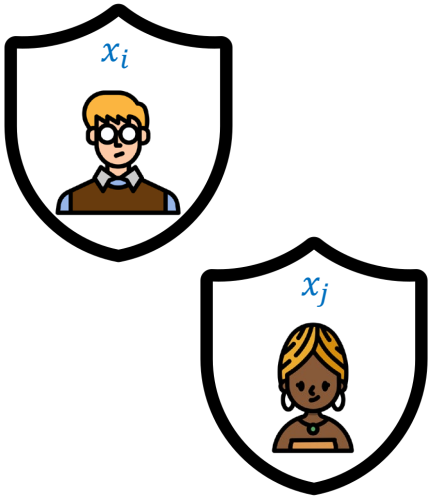
where

$$K = \left(k(x_i, x_j)\right) + \sigma^2 * \mathbb{1}_n$$

$$K_* = k(x_i^*, x_j)$$

$$K_{**} = k(x_i^*, x_j^*)$$

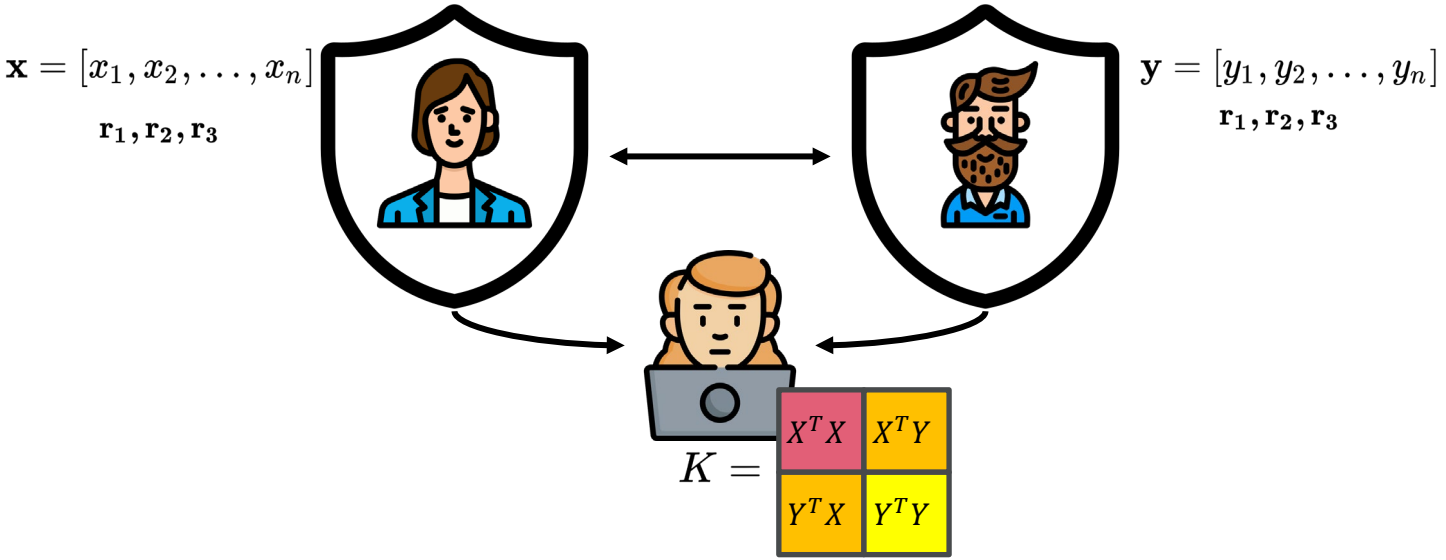
$$K(x, y) = \sum_{i=0}^{input\ dim} \alpha_i^2 x_i y_i$$



A Privacy-Preserving Framework for Collaborative Machine Learning with Kernel Methods

FLAKE (**F**ramework for **L**earning with **A**nonymized **Ke**rnels)

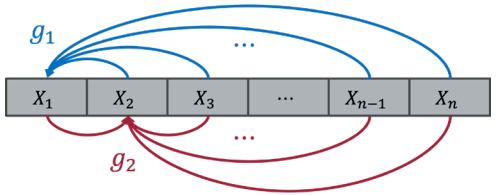
Hannemann, A., Ünal, A. B., Swaminathan, A., Buchmann, E., & Akgün, M. (2023, November). A Privacy-Preserving Framework for Collaborative Machine Learning with Kernel Methods. In 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) (pp. 82-90). IEEE.



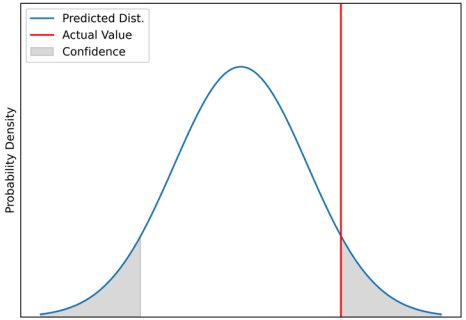
- Proposes a specialized randomized encoding-based masking method.
- Allows learning kernel-based methods on horizontally distributed data.

What are the challenges?

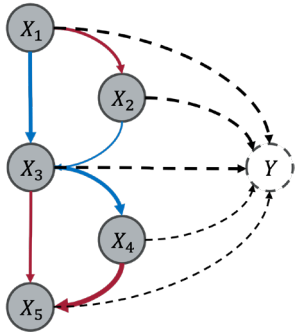
FREDA is a complex method which consists of 3 components



Feature Models



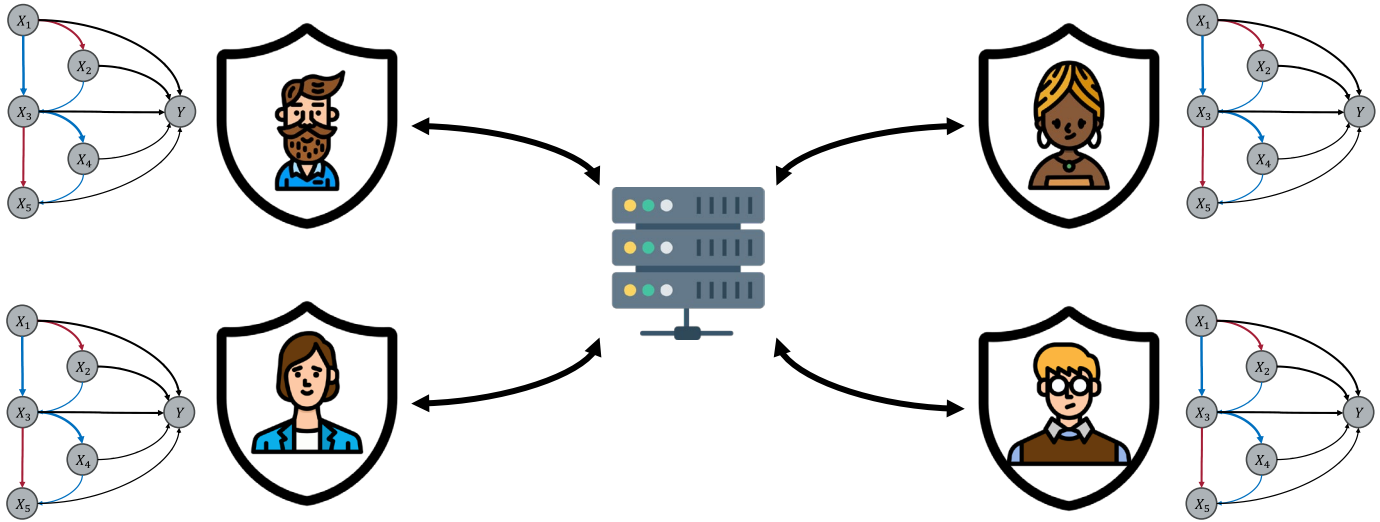
Confidence Scores



Final Adaptive Model

- Randomized Encoding
- Secure Aggregation

Challenge of Training the Final Adaptive Model

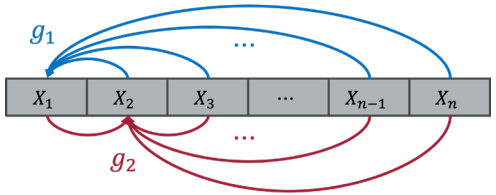


Federated Learning

- Allows multiple entities to collaboratively train a machine learning model.
- Allows entities to keep their data private during training.

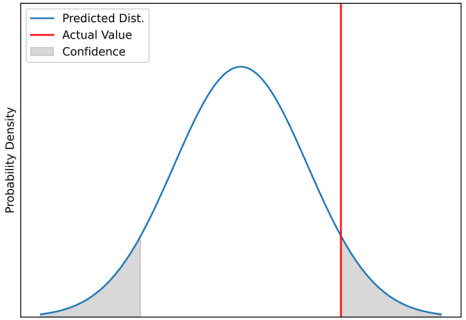
What are the challenges?

FREDA is a complex method which consists of 3 components



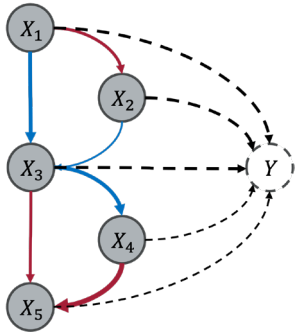
Feature Models

- Randomized Encoding
- Secure Aggregation



Confidence Scores

Can be performed locally by the target domain owner!



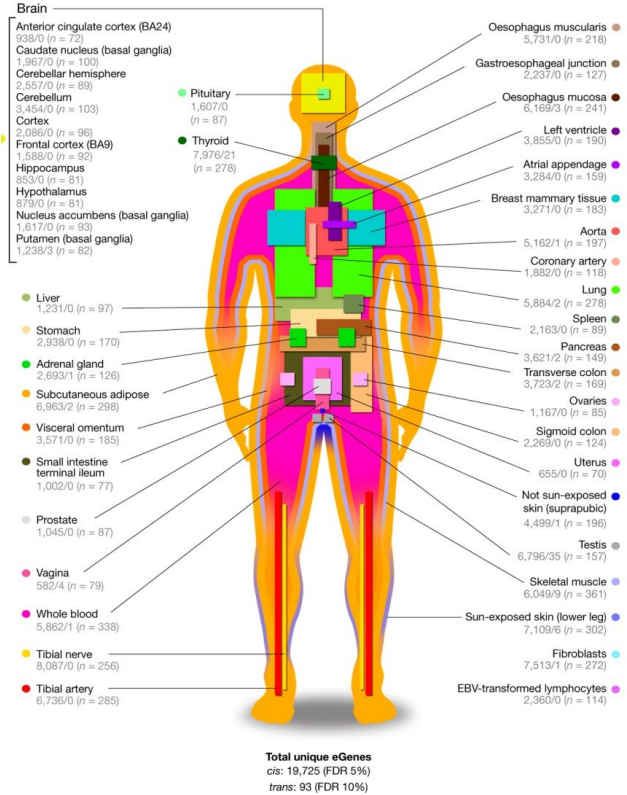
Final Adaptive Model

- Federated Learning
- Secure Aggregation

Age Prediction from DNA Methylation Data

- Healthy tissue samples from GEO and TCGA
  - Has ~13,000 features (High dimensionality)
- Training set: 1866 & Test set : 1001 samples
- We consider 10 tissues in total
  1. Brain CRBM
  2. Brain Frontal
  3. Brain Hippocampus
  4. Brain Mid-Brain
  5. Brain Occipital
  6. Brain Temporal
  7. CD4+ cells
  8. Blood
  9. Saliva`
  10. vaginal swab

- Biologically very different even from other brain tissues.
- They are not represented in the training data.



GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). https://doi.org/10.1038/nature24277

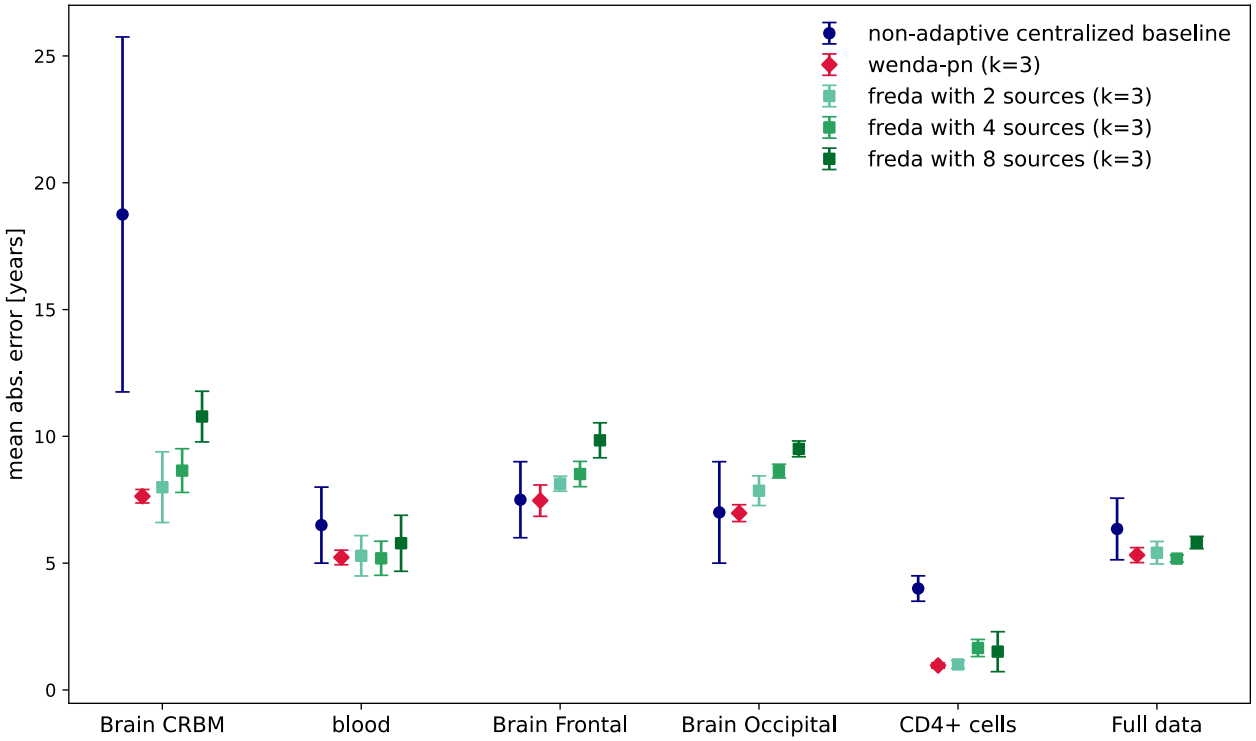
Age Prediction from DNA Methylation Data

Age Prediction from DNA Methylation Data

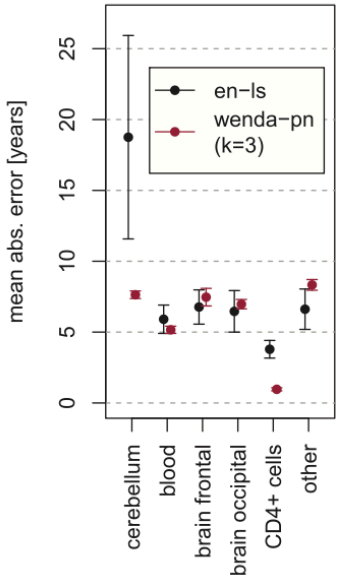
- DNA methylation changes characteristically with age.
- DNA methylation patterns are tissue-specific, meaning that each tissue can be considered its own domain.

Predicting age across tissues is a domain adaptation problem

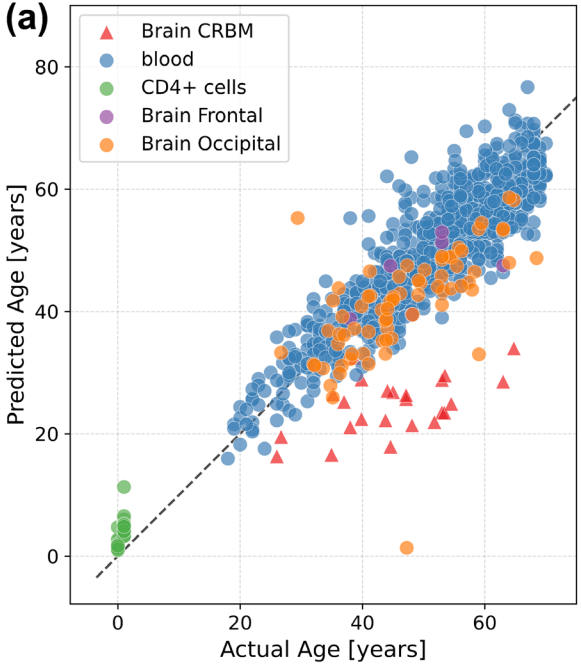
Experiments with FREDA – Comparison



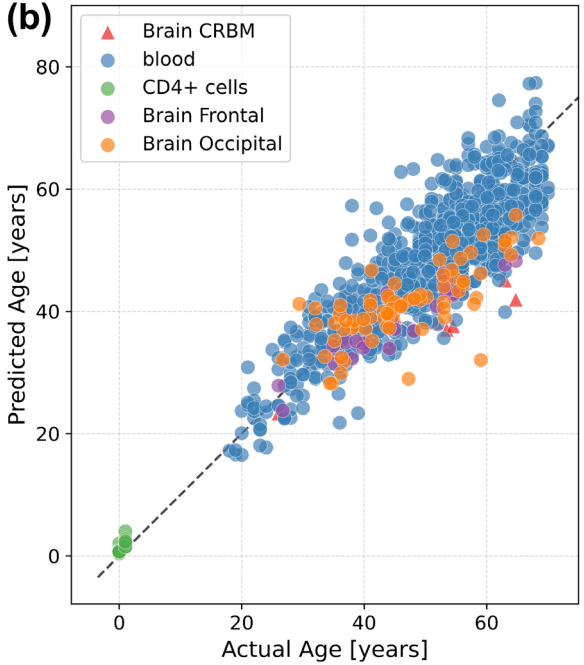
Experiments with FREDA – Baselines



Mean abs. error per test tissue.



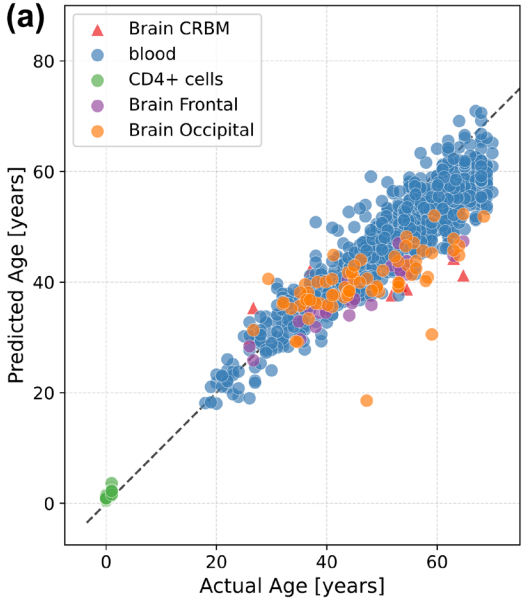
centralized non-adaptive
en-ls



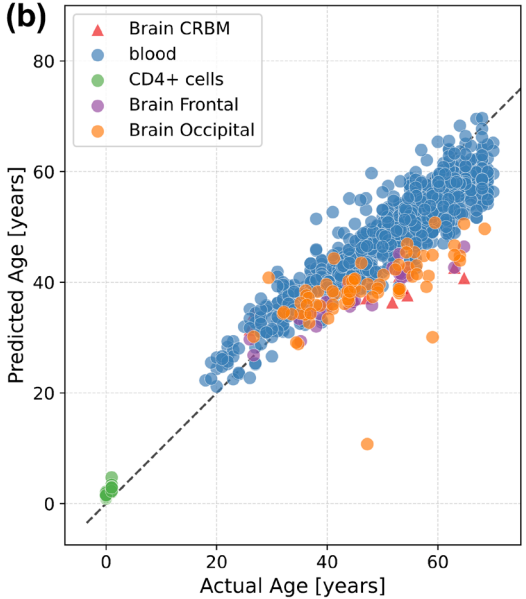
WENDA

Experiments with FREDA – Predictive Performance

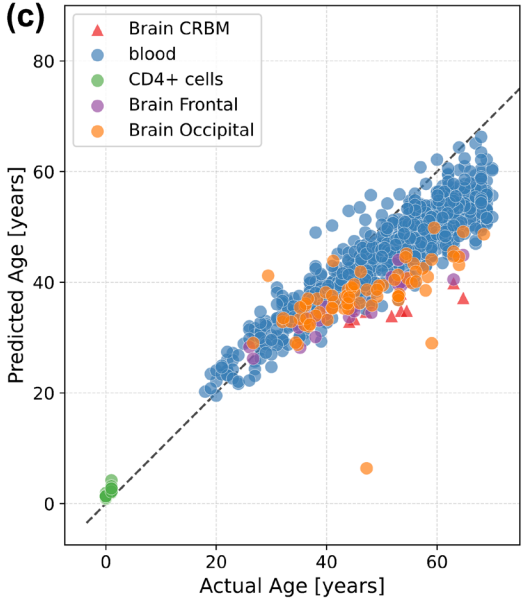
FREDA (2 source clients)



FREDA (4 source clients)



FREDA (8 source clients)



Summary

- We propose FREDA a privacy preserving federated unsupervised domain adaptation framework.
- By using various privacy enhancing technologies, FREDA allows entities in a distributed setting to collaboratively perform unsupervised domain adaptation on high dimensional tabular datasets without sacrificing the privacy of their data.
  - Randomized Encoding
  - Secure Aggregation
  - Federated Learning
- We test the performance of our framework on real data focusing on age prediction using DNA methylation data across multiple tissues.
- Our results show that FREDA achieves almost the same level of performance on the entire target domain data as WENDA.
- For the cerebellum samples, our FREDA effectively addresses the problem of distribution shift across domains, yielding performance comparable to WENDA even in a distributed setting with the source domain data distributed across up to 8 different parties.

Thanks for listening!

Any questions?