Distributed Harmonization: Federated Clustered Batch Effect Adjustment and Generalization

Bao Hoang, Yijiang Pang, Siqi Liang, Liang Zhan, Paul Thompson, Jiayu Zhou
Illidan Lab @ University of Michigan
University of Pittsburgh
University of Southern California

Applications	Biomarkers	Disease Progression	Drug Discovery
Methodology	Transfer Learning Multi-task, Few-Shot,Adaptation	Multi-Modality Fusion	Robust Learning against Missing and Noisy Data
Infrastructure	Privacy Protection, Federated Learning (Data Heterogeneity, Resource Heterogeneity, Availability), Fairness, Distributed Optimization		

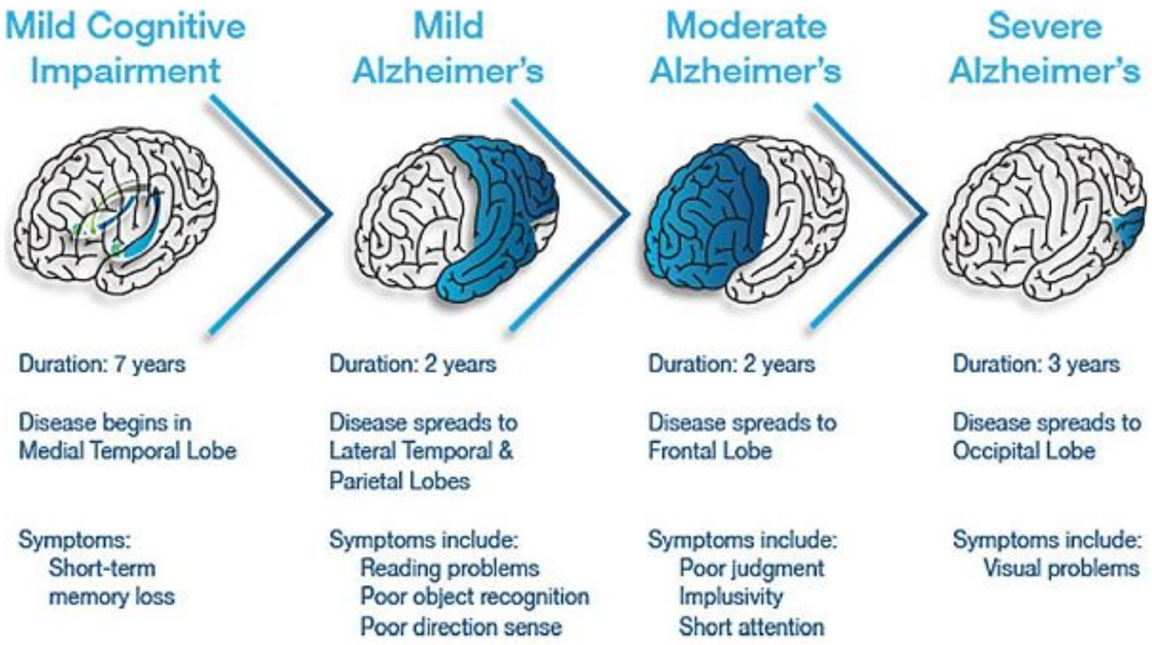
Alzheimer's disease is officially listed as the seventh leading cause of death in the United States in 2021 and 2022 with 31.0% and 28.9% death rates, respectively.

Kochanek KD, Murphy SL, Xu JQ, Arias E. Mortality in the United States, 2022. NCHS Data Brief, no 492. Hyattsville, MD: National Center for Health Statistics. 2024. DOI: https://dx.doi.org/10.15620/cdc:135850

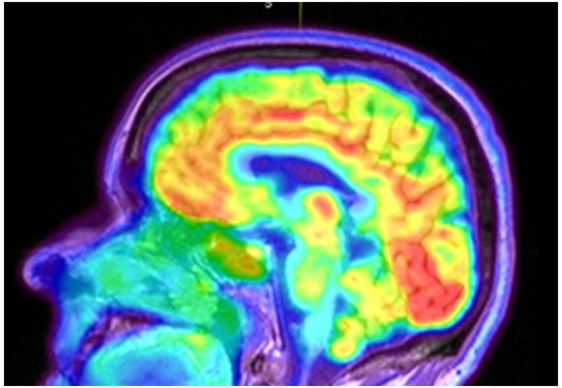
Neurodegenerative Disease - Alzheimer's

Early Stage

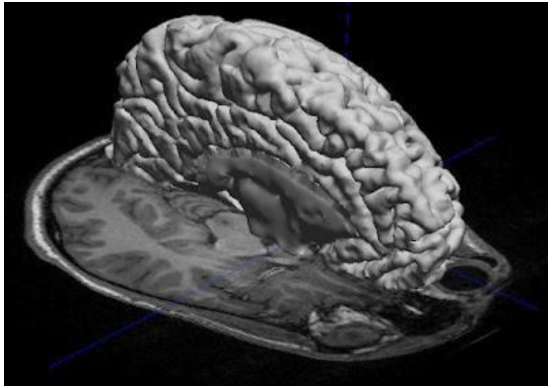
MCI



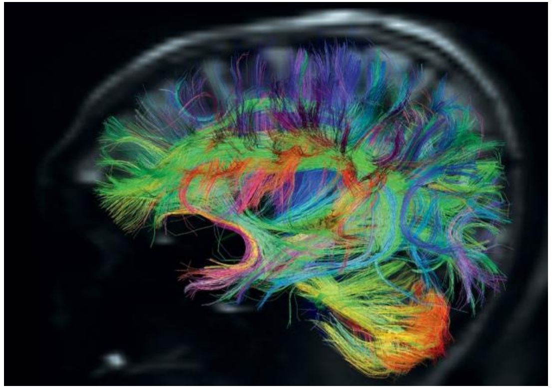
Early Diagnosis and Biomarkers



Huang, Shuai, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, Eric Reiman, and Alzheimer's Disease NeuroImaging Initiative. "*Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation*." NeuroImage 50, no. 3 (2010): 935-949.



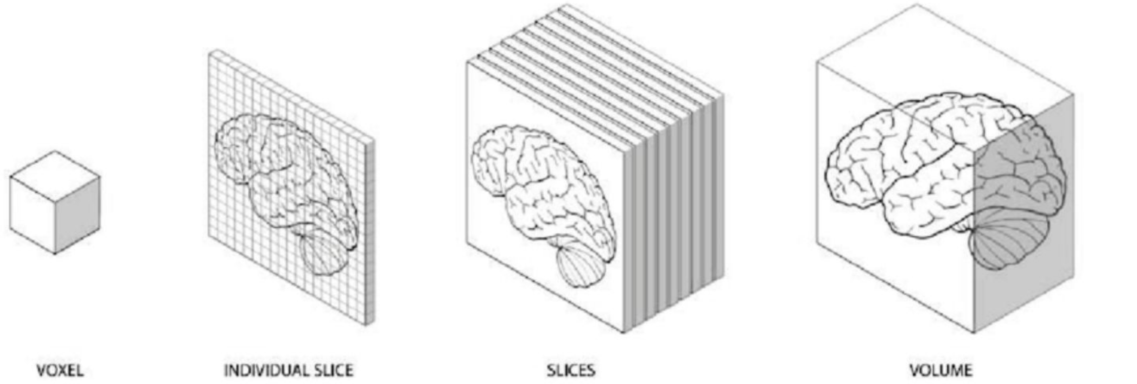
Zhou, Jiayu, Jun Liu, Vaibhav A. Narayan, Jieping Ye, and Alzheimer's Disease Neuroimaging Initiative. "*Modeling disease progression via multi-task learning*." NeuroImage 78 (2013): 233-248.

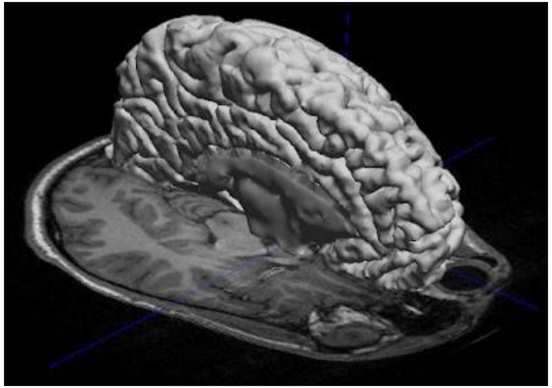


Wang, Qi, Liang Zhan, Paul M. Thompson, Hiroko H. Dodge, and Jiayu Zhou. "*Discriminative fusion of multiple brain networks for early mild cognitive impairment detection*." In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 568-572. IEEE, 2016.

Medical Imaging is High Dimensional

- Brain imaging is high dimensional with rich information
  - Brain volume 1,200,000 mm^3
  - Voxel sizes 1-3 mm in 1.5T or 3T scanners
  - ~1 Million voxels in one brain scan.





Drissi, Natasha Morales. Brain Networks and Dynamics in Narcolepsy. Linkopings Universitet (Sweden), 2018.

Medical Imaging has a Small Sample Size

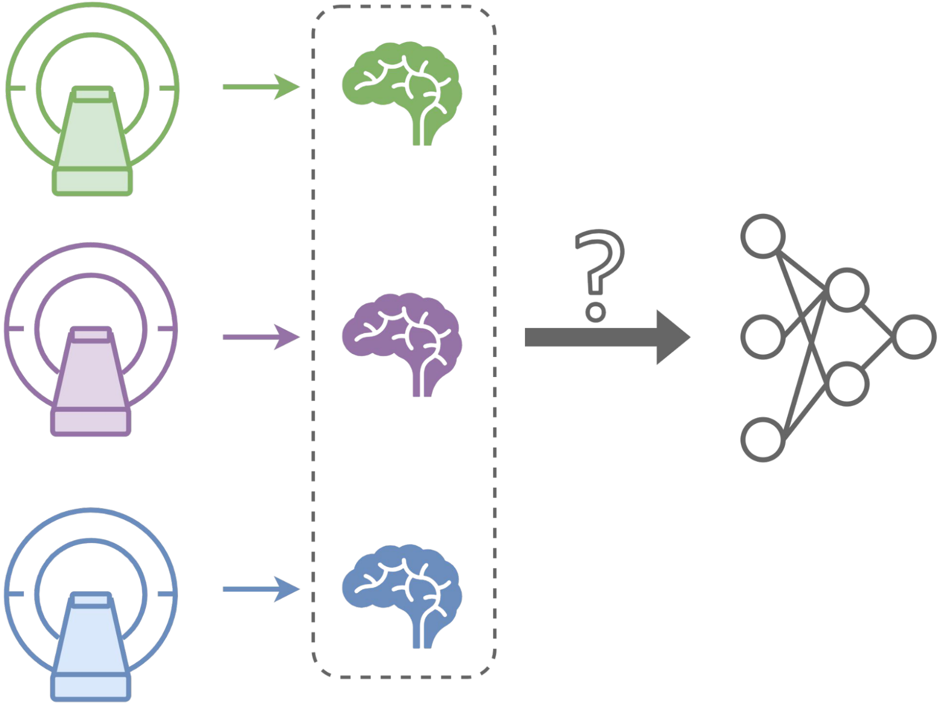
- Limited sample size due to acquisition costs. Examples:
- **ADNI**: Alzheimer's Disease Neuroimaging Initiative
  - Longitudinal, multi-center, observational study, with the goal to validate biomarkers for Alzheimer's disease (AD) clinical trials.
  - Stage I: 5 years, $60 million
  - 819 samples for machine learning studies.
- **ENIGMA**: Enhancing Neuro Imaging Genetics Through Meta Analysis
  - Brings together researchers in imaging genomics to understand brain structure, function, and disease
  - 50 working groups across the world

Jack Jr, Clifford R., et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods." Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27.4 (2008): 685-691.

Thompson, Paul M., et al. "ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries." Translational psychiatry 10.1 (2020): 100.

Batch Effect from Multi-Site Analysis

Affected by batch effect,
resulting in non-i.i.d data

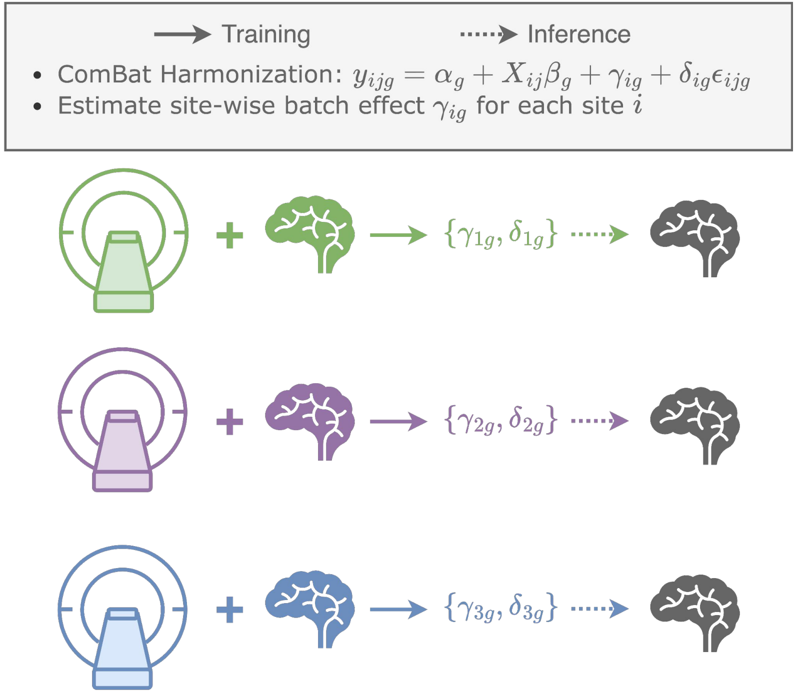


- **High Dimensionality**: Large-scale brain imaging data (and prior knowledge) is required to train an effective machine learning model
- **Sample Size**: We need to collect from multiple sites using different scanners, leading to batch effect.
- Batch effects can lead to **poor generalization** and **unstable predictions** for machine learning model [1].

[1] J. Yang, K. Zhou, Y. Li and Z. Liu, Generalized out-of-distribution detection: A survey (2021)

ComBat Harmonization

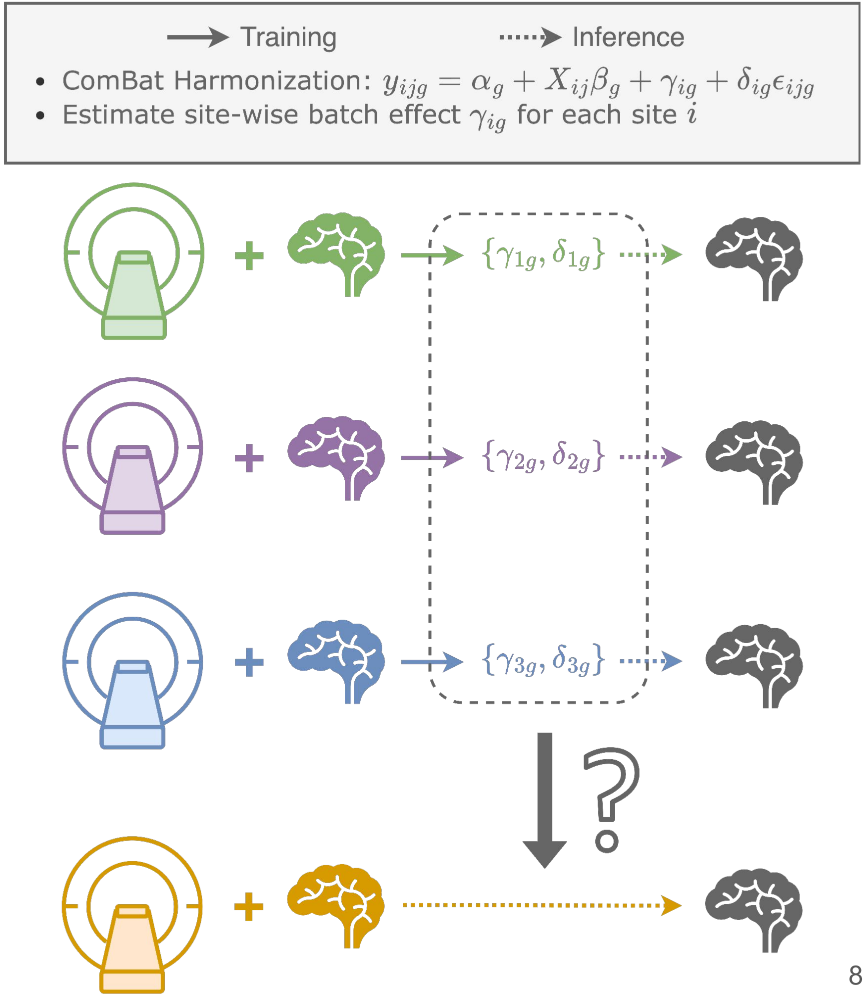
- CombBat is a well-known harmonization technique and has been shown to be helpful in mitigating the batch effect of neuroimaging data.
  - Model and estimate site-wise batch effects.
  - Remove batch effects for downstream analysis tasks.



Fortin, Jean-Philippe, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams et al. "Harmonization of cortical thickness measurements across scanners and sites." Neuroimage 167 (2018): 104-120.

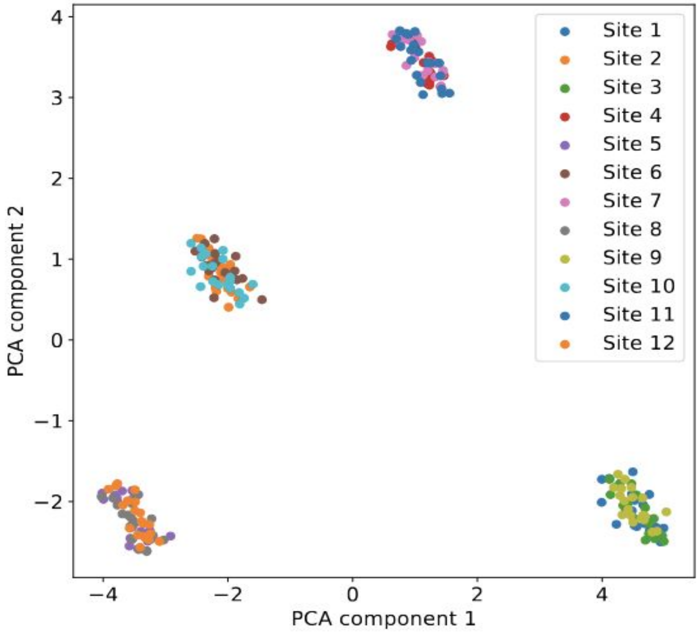
Limitation of ComBat

- However, ComBat is incompatible with harmonizing brain imaging from **unseen** sites without retraining, which introduces **significant computational cost**.



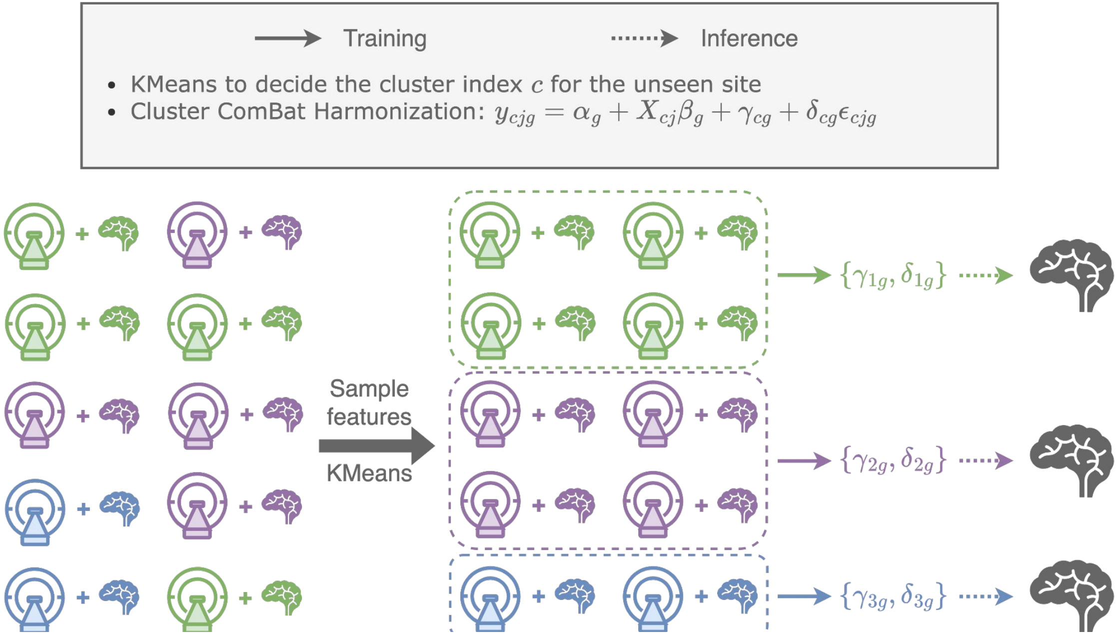
Our Approach

- We assume that some sites may exhibit clustering patterns.
    - Sites may share similarity
    - Sites in **same** cluster can **share** ComBat estimated batch effect.
- Using this assumption, we proposed Cluster ComBat to eliminate the need of retraining as original ComBat.



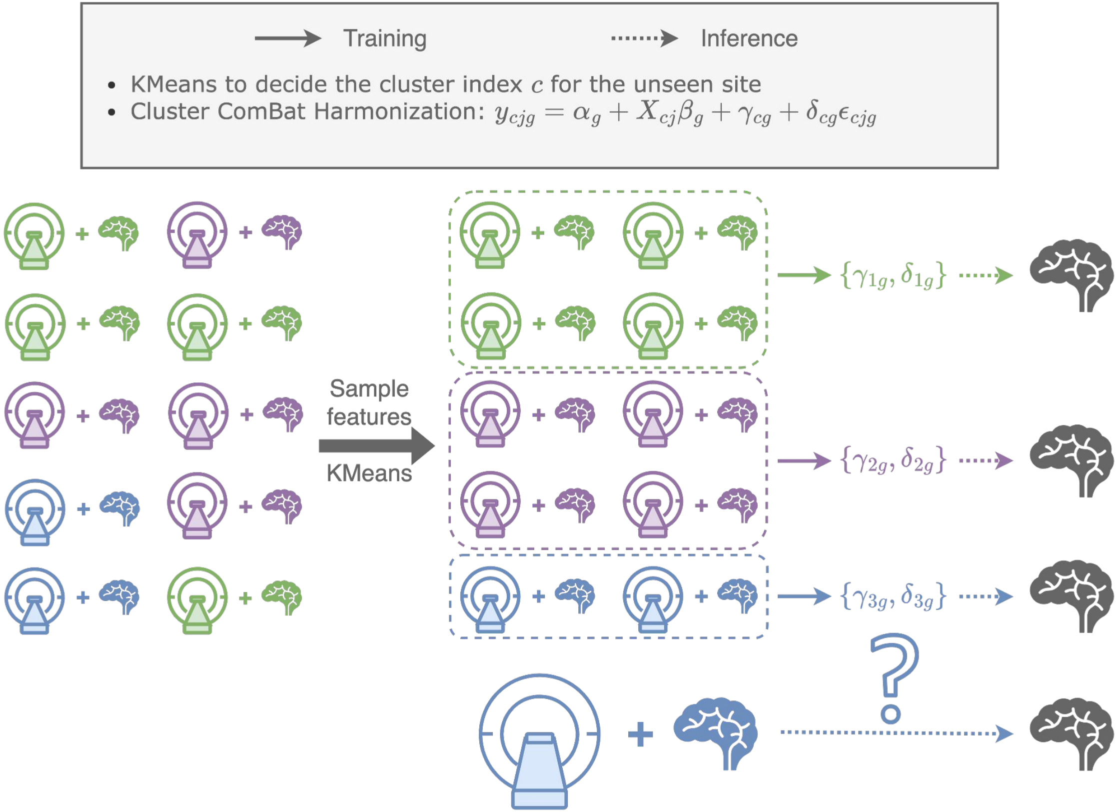
Cluster ComBat

- Instead of estimating **site-wise** batch effects, Cluster ComBat estimates **cluster-wise** batch effects.



Cluster ComBat

- The pre-estimated **cluster-wise** batch effects can be used to harmonize brain imaging from **unseen** sites from same clusters without the need of retraining.



Privacy Risk of Distributed ComBat

- Sharing data directly among multiple sites to apply harmonization poses challenges to **data security** and **patient privacy protection**.
  - Direct training on all the data is often **impractical** in the medical domain.
  - A distributed version of the original ComBat method, known as Distributed ComBat, already was proposed [1].
- Based on their framework, we have developed a distributed version of our proposed method, called **Distributed Cluster ComBat**.

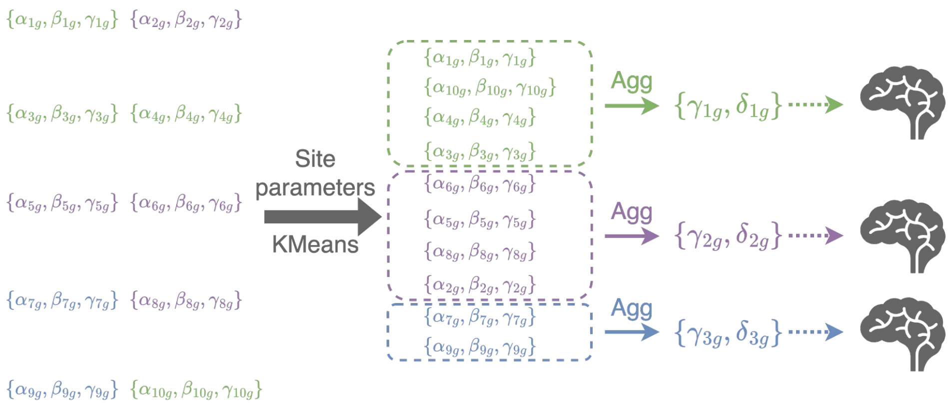
[1] Chen, A. A., Luo, C., Chen, Y., Shinohara, R. T., & Shou, H. (2022). Privacy-preserving harmonization via distributed ComBat. In NeuroImage (Vol. 248, p. 118822). Elsevier BV. https://doi.org/10.1016/j.neuroimage.2021.118822

Distributed Cluster Combat

- In decentralized settings, we **cannot** use raw brain imaging data to train K-means.
- Instead, we use **locally** estimated parameters for K-means.

→ Training ⋯→ Inference

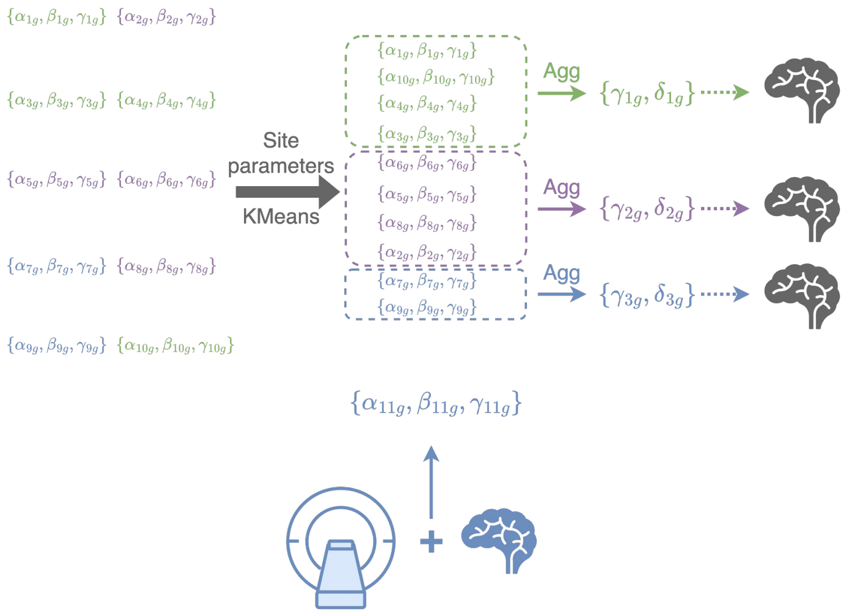
- Locally estimate the ComBat parameters first
- KMeans based on site parameters
- Parameter aggregation in each cluster
- KMeans to decide the cluster index c for the unseen site
- Cluster ComBat Harmonization: $y_{cjg} = \alpha_g + X_{cj}\beta_g + \gamma_{cg} + \delta_{cg}\epsilon_{cjg}$



Distributed Cluster Combat

Training ⋯⋯ Inference

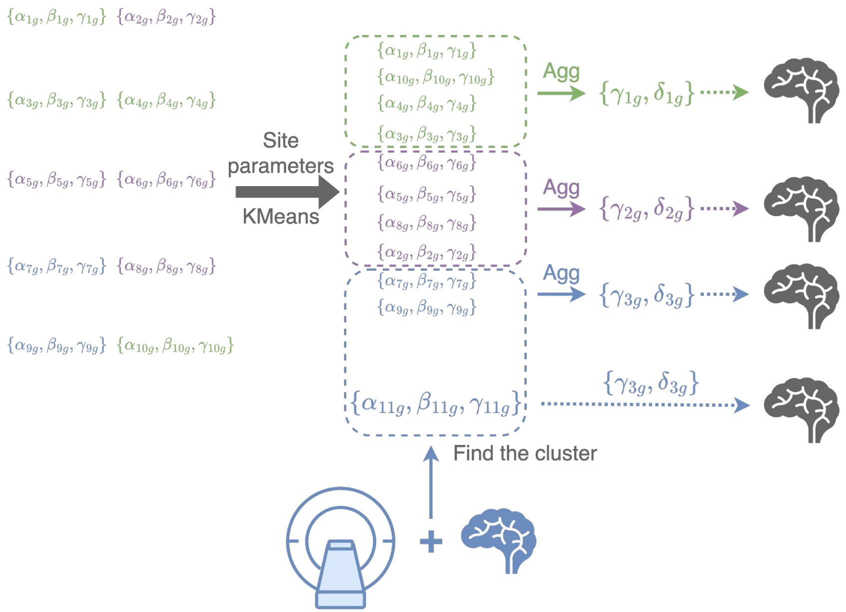
- Locally estimate the ComBat parameters first
- KMeans based on site parameters
- Parameter aggregation in each cluster
- KMeans to decide the cluster index c for the unseen site
- Cluster ComBat Harmonization: $y_{cjg} = \alpha_g + X_{cj}\beta_g + \gamma_{cg} + \delta_{cg}\epsilon_{cjg}$



Distributed Cluster Combat

→ Training ⋯→ Inference

- Locally estimate the ComBat parameters first
- KMeans based on site parameters
- Parameter aggregation in each cluster
- KMeans to decide the cluster index c for the unseen site
- Cluster ComBat Harmonization: $y_{cjg} = \alpha_g + X_{cj}\beta_g + \gamma_{cg} + \delta_{cg}\epsilon_{cjg}$



Downstream Task Performance

- ADNI Dataset
- Use Linear Regression to predict the 6 prediction tasks (MEM, MEM SLOPES, EXF, EXF SLOPES, LAN, and LAN SLOPES variables) using 228 ROI features of DTI brain imaging
    - Outperform baselines on all 6 tasks for both centralized/decentralized settings

Algorithm	MEM	MEM SLOPES	EXF	EXF SLOPES	LAN	LAN SLOPES
Centralized Setting						
Without harmonization	13.77±22.05	1.89±3.59	10.30±19.38	1.58±3.19	10.94±17.74	1.45±3.01
Generalized Linear Squares Approach [41]	1.07±0.30	0.52±0.18	0.93±0.22	0.47±0.18	0.95±0.26	0.45±0.13
ComBat[a]	**1.00±0.18**	0.16±0.04	1.03±0.18	0.13±0.04	1.04±0.20	0.13±0.03
Cluster ComBat	1.00±0.20	**0.15±0.03**	**0.91±0.12**	**0.12±0.03**	**0.87±0.15**	**0.12±0.02**
Decentralized Setting						
Distributed ComBat[a]	0.98±0.16	0.15±0.03	1.00±0.16	0.13±0.03	1.01±0.17	0.12±0.03
Distributed *Cluster ComBat*	**0.91±0.16**	**0.14±0.03**	**0.96±0.12**	**0.12±0.02**	**0.91±0.17**	**0.11±0.02**

Time Efficiency

Average time of running 100 experiments MEM regression task.

- Centralized: **2x** speedup
- Decentralized: **4x** speedup

Algorithm	Average Time (s)
Centralized Setting	
ComBat[a]	0.2427±.0.0017
Cluster ComBat	0.1127±.0.0001
Decentralized Setting	
Distributed ComBat[a]	2.5051±0.0771
Distributed *Cluster ComBat*	0.6389±.0.0027

Discussion and Future Works

- Cluster-based Combat for both centralized/decentralized settings
- Capable to generalization on unseen sites without re-training
- Design for privacy concern in medical/biomedical domains
- Integrating with the ENIGMA Consortium toolbox to further validate existing studies

Thompson, Paul M., et al. "ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries." Translational psychiatry 10.1 (2020): 100.

Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejan- dro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al . 2014. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain imaging and behavior 8 (2014), 153–182.

Thanks!

http://illidanlab.github.io

Acknowledgement: This material is based in part upon work supported by the National Science Foundation under Grant IIS-2212174, IIS-1749940, IIS 2319450, IIS 2045848, Office of Naval Research N00014-24-1-2168, and National Institute on Aging (NIA) RF1AG072449, U01AG068057, National Institute of Mental Health RF1MH125928.