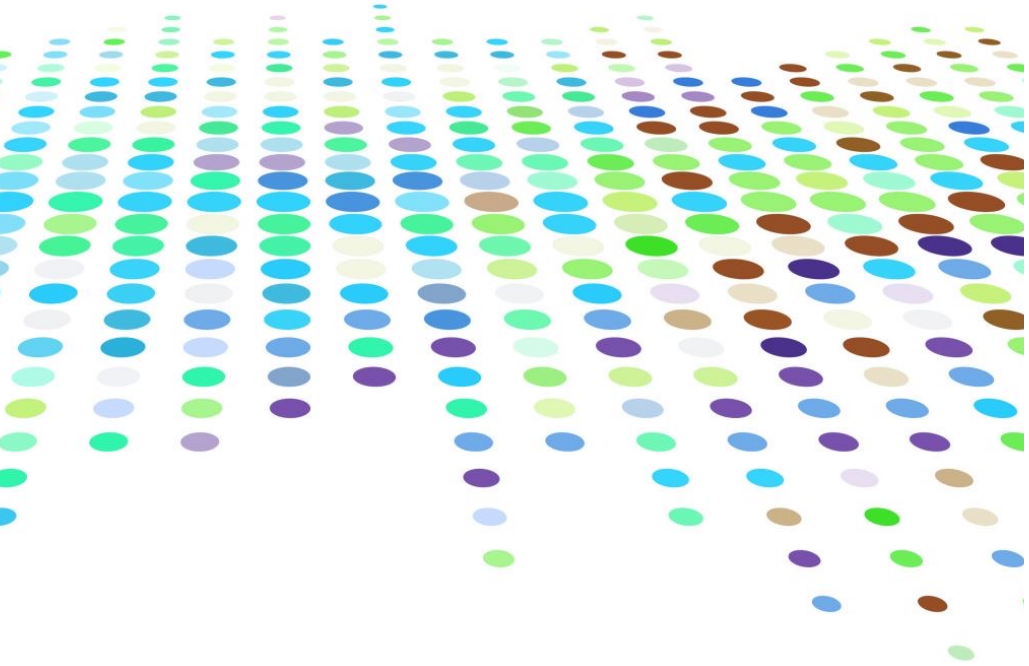


PRINCIPLES AND PRACTICES OF FEDERATED LEARNING: METHODS, CHALLENGES, AND CASE STUDIES

Rui Duan (段芮)

Harvard T.H. Chan School of Public Health

2024 International Conference on Frontiers of Data Science

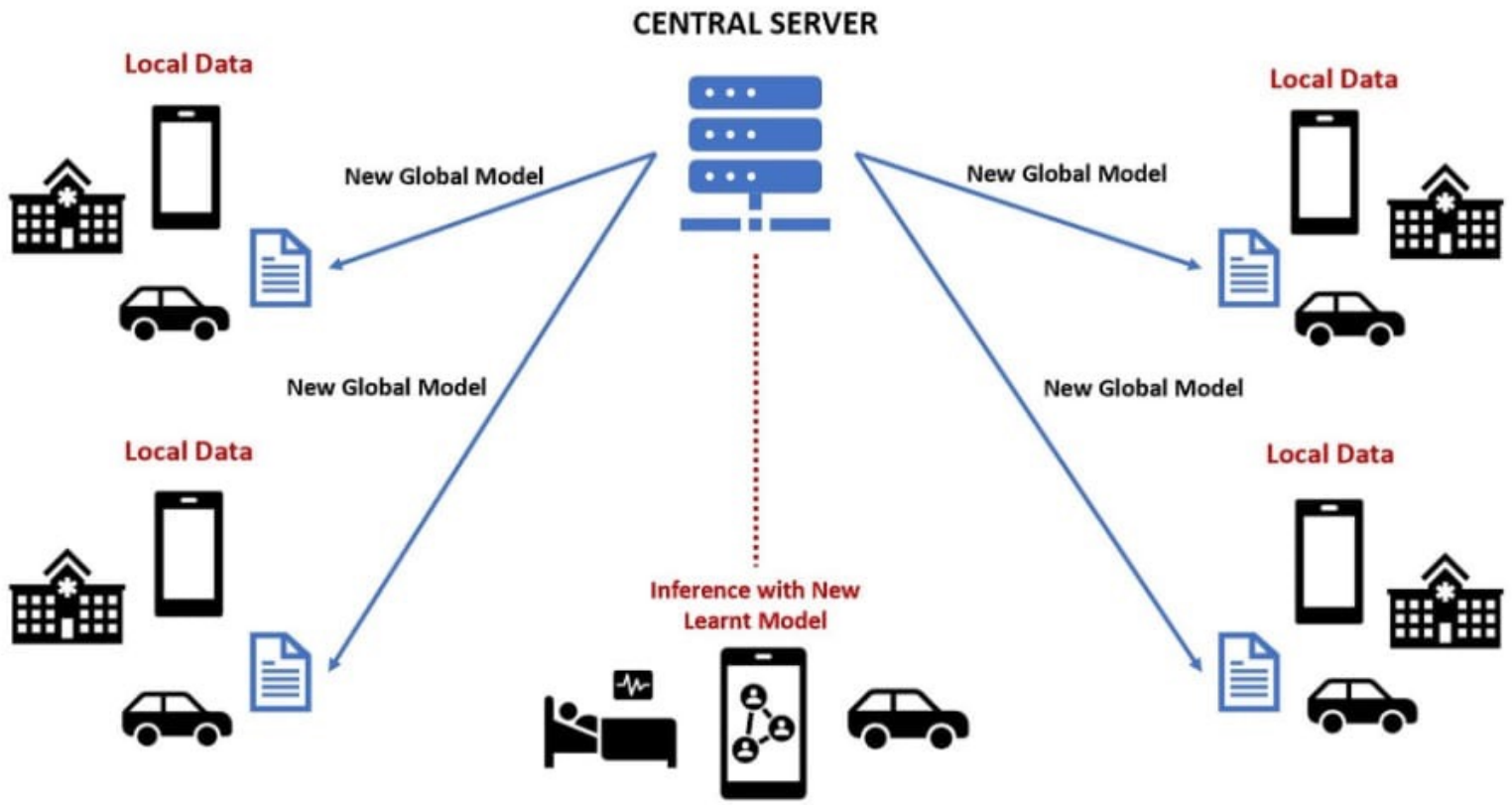
OUTLINE

- Section I — Overview of Federated Learning
  - What is Federated Learning?
  - Important applications
  - Challenges
- Section II — Federated Learning via Federated Optimization
- Section III — Federated Statistical Inference
- Discussion— Other Related Topics

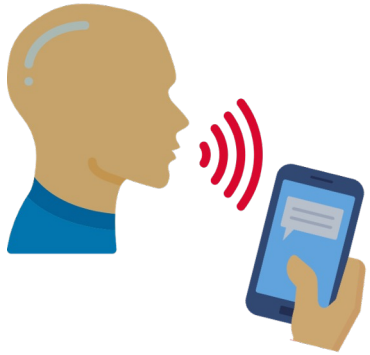
SECTION I: OVERVIEW

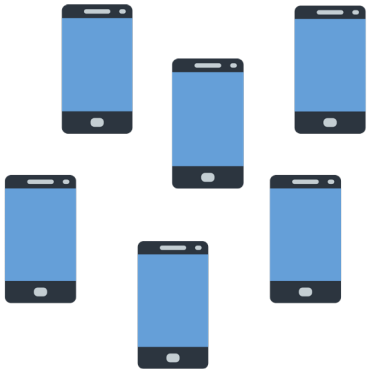
FEDERATED LEARNING

- Federated Learning is a machine learning technique that allows training models across multiple decentralized devices or servers holding local data samples without exchanging them.



Picture from internet



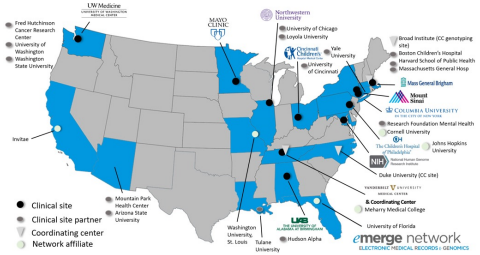


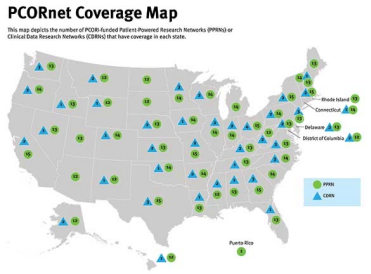
Pictures from internet

Datafication: Transforming the World into Data

DATA NETWORKS

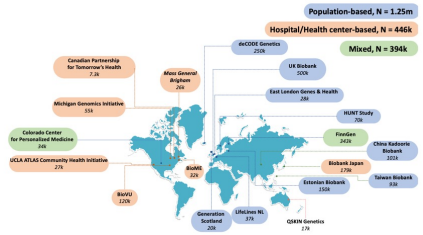
PsycheMERGE





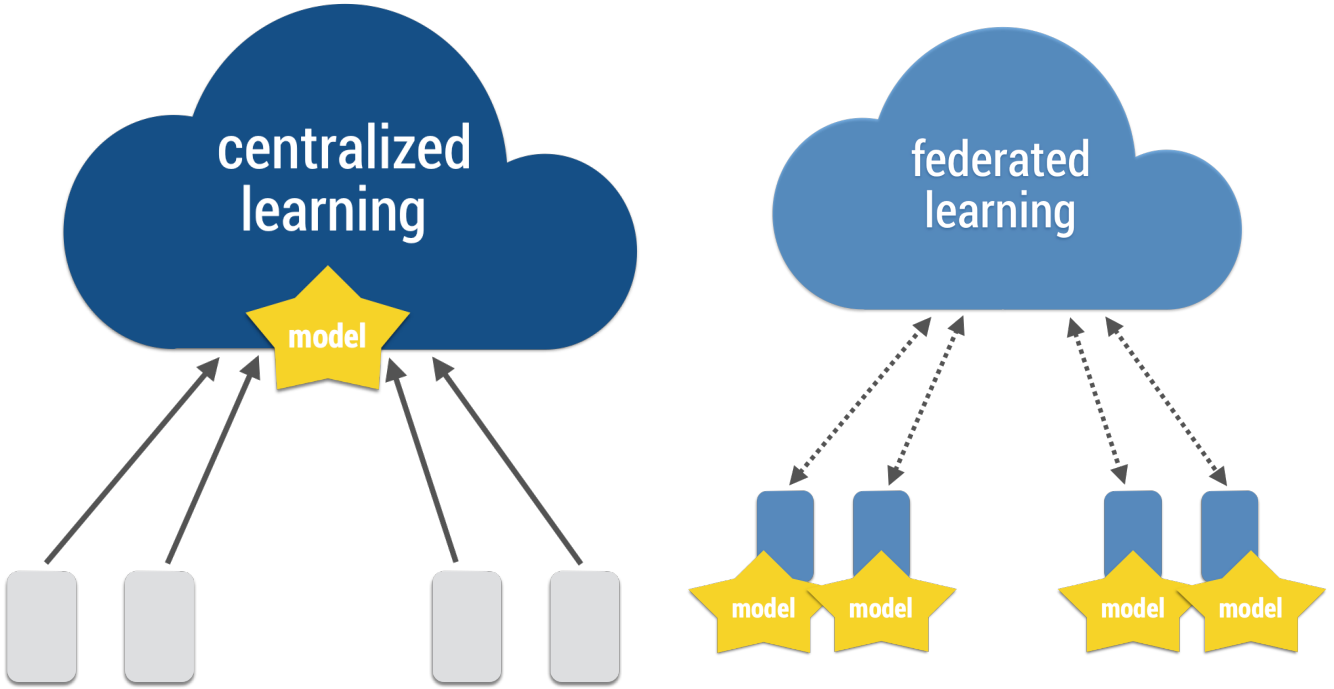
4CE

Consortium for Clinical Characterization of COVID-19 by EHR

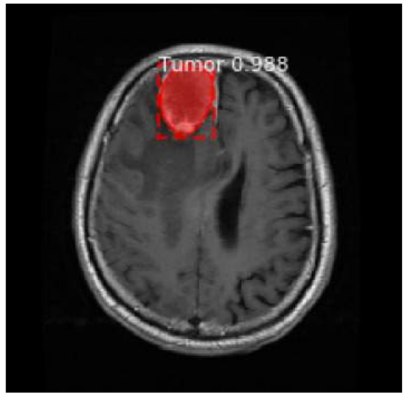


WHY FEDERATED LEARNING

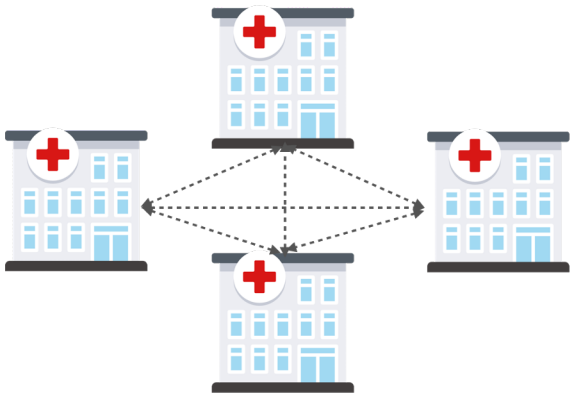
- *reduce strain on network*
- *incorporate more data*
- *improve privacy**



Picture from Federated and Collaborative Learning, Fall 2023, Virginia Smith.
https://www.cs.cmu.edu/~smithv/10719/



collaboratively learn tumor detection model without sharing patient data



NVIDIA

PLATFORMS DEVELOPERS INDUSTRIES SHOP DRIVERS SUPPORT ABOUT NVIDIA EMAIL SIGN-UP

HOME DEEP LEARNING NETWORKING DRIVING GAMING PRO GRAPHICS AUTONOMOUS MACHINES HEALTHCARE AI PODCAST

Medical Institutions Collaborate to Improve Mammogram Assessment AI with NVIDIA Clara Federated Learning

In a federated learning collaboration, the American College of Radiology, Diagnosticos da America, Partners HealthCare, Ohio State University and Stanford Medicine developed better predictive models to assess breast tissue density.

April 15, 2020 by MONA FLORES

ARTIFICIAL INTELLIGENCE, DIAGNOSTICS

UPenn, Intel partner to use federated learning AI for early brain tumor detection

The project will bring in 29 institutions from North America, Europe and India and will use privacy-preserved data to train AI models. Federated learning has been described as being born at the intersection of AI, blockchain, edge computing and the Internet of Things.

By ALARIC DEARMENT

Post a comment / May 11, 2020 at 10:03 AM

LILY HAY NEWMAN SECURITY 10.07.2020 02:19 PM

How Google's Android Keyboard Keeps 'Smart Replies' Private

The latest Gboard feature needs to know as much as possible about your digital life to work—but doesn't share that data with Google.

Artificial intelligence / Machine learning

How Apple personalizes Siri without hoovering up your data

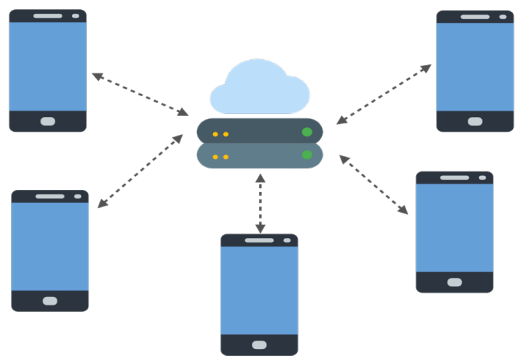
The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

by **Karen Hao** December 11, 2019

Sherpa raises $8.5M to expand from conversational AI to B2B privacy-first federated learning services

Ingrid Lunden @ingridlunden / 7:11 PM EDT • March 15, 2021 Comment

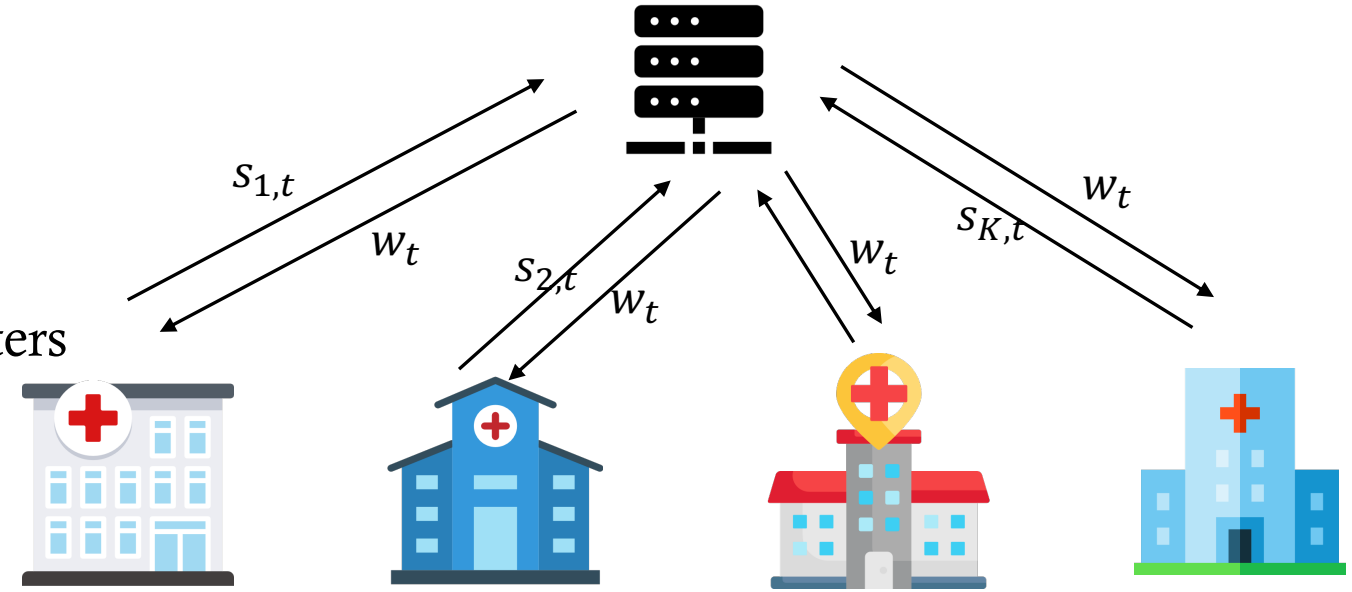
power next-word prediction in mobile keyboards without collecting texts



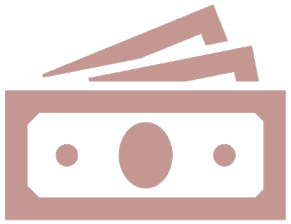
BASIC PROBLEM SET-UP

- Objective function: $L(w) = \sum_{k=1}^{K} L_k(w; D_k)$
- Optimization over all sites: $\hat{w} = \text{argmin}_w L(w)$
- Initialization
- Local sites share summary stats

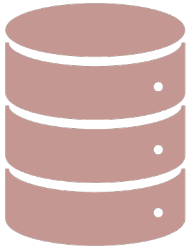
  with central server
- Central server updates model parameters



CHALLENGES



Expensive Communications



Data heterogeneity

Privacy concerns

COMMUNICATION COST

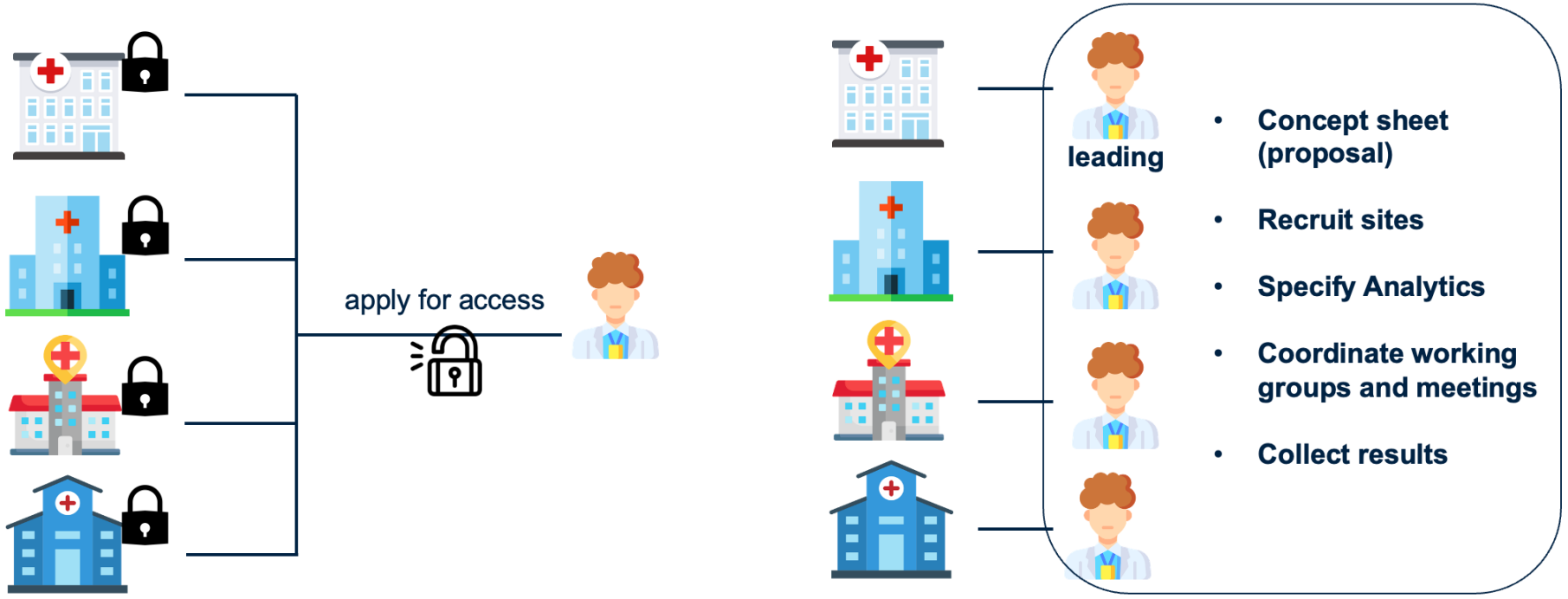
Upload and Download Costs

Limit on file sizes

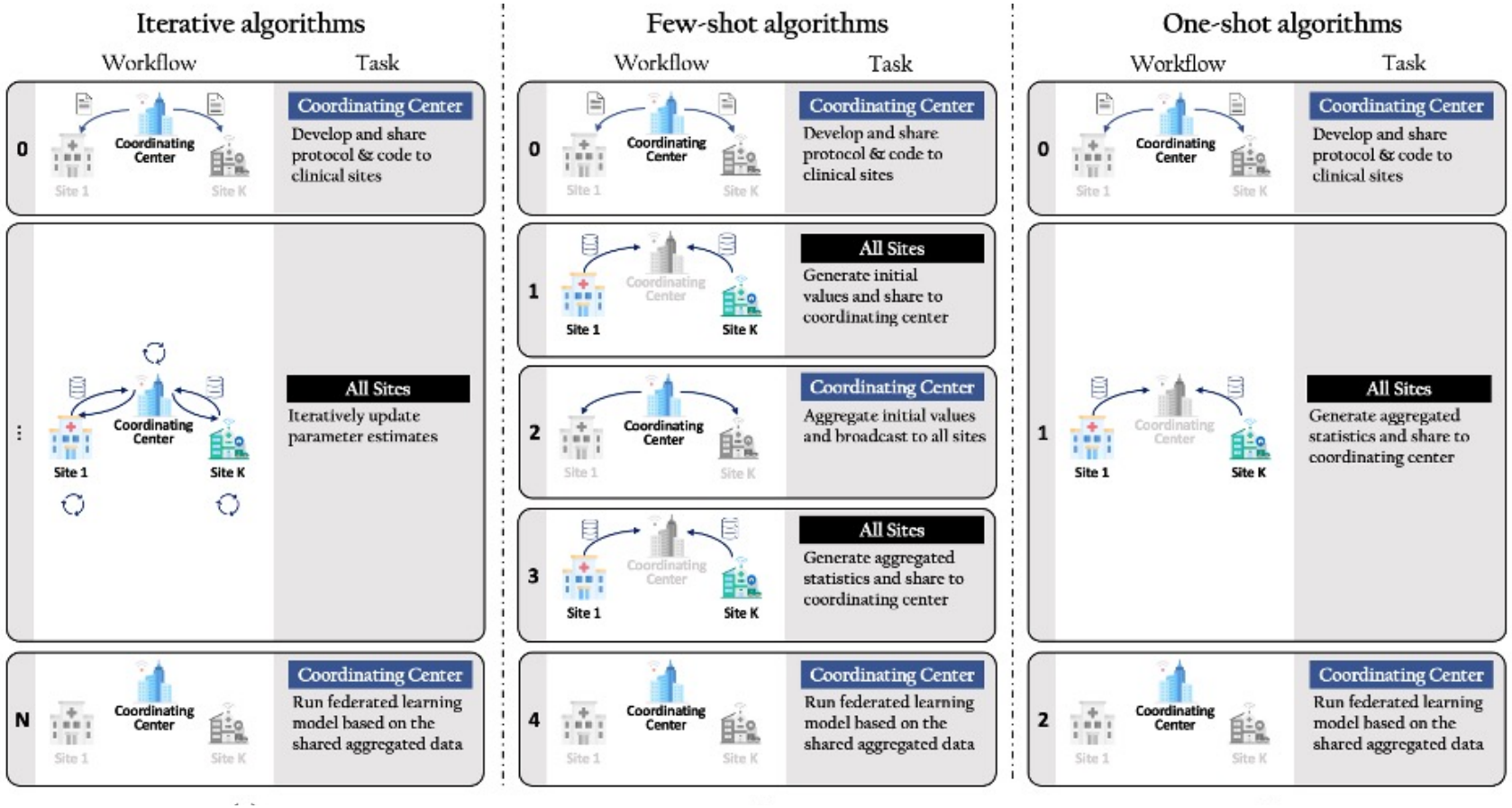
Time

Human efforts involved in the communication

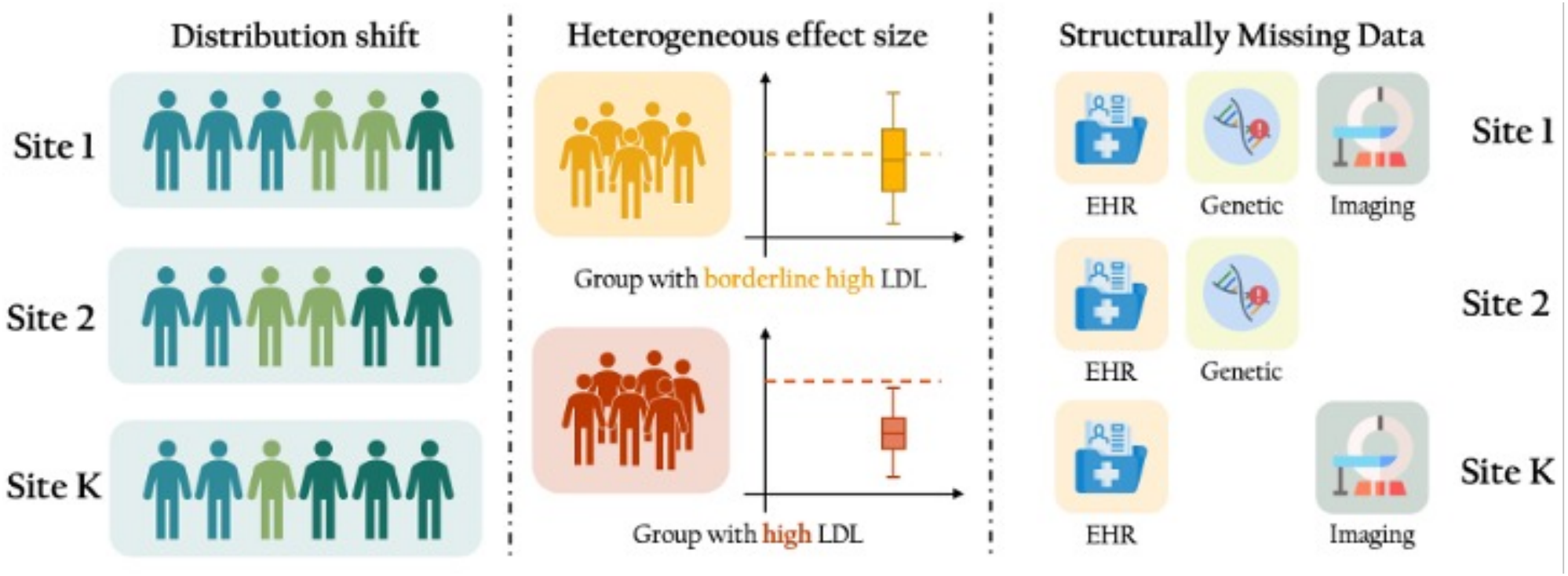
DATA SHARING IN RESEARCH NETWORK



REDUCING THE COMMUNICATION COST



DATA HETEROGENEITY

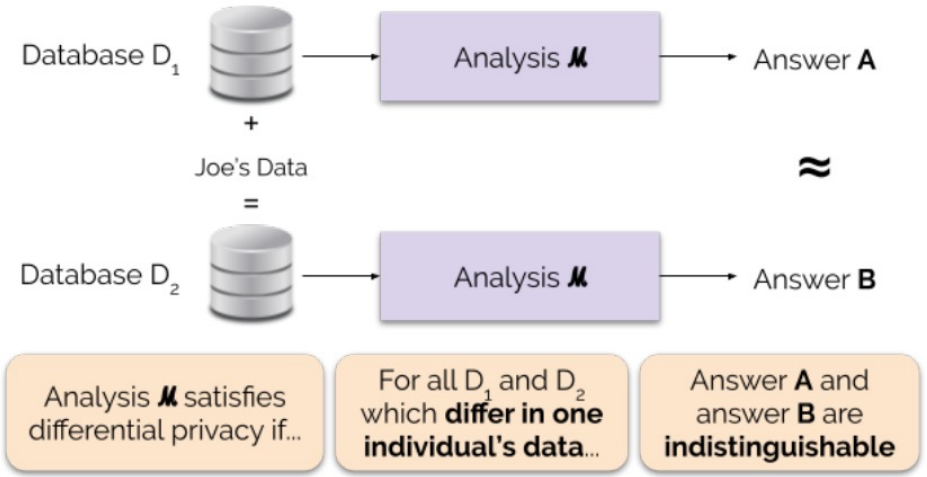


VARIOUSTASKS IN FEDERATED LEARNING

- Train a unified model for all sites
  - Better generalizability and transferability
- Jointly train site-specific models
  - Leveraging between-site similarity while accounting for site-level differences.
- Train a model for a target population of interest
  - Target population can be defined within or across sites
  - Leveraging all data within the network

PRIVACY CONCERNS

- Encryption
- Secure multiparty computation
- Differential privacy



intuition: can't determine whether or not Joe was present in the dataset, let alone the contents of his data

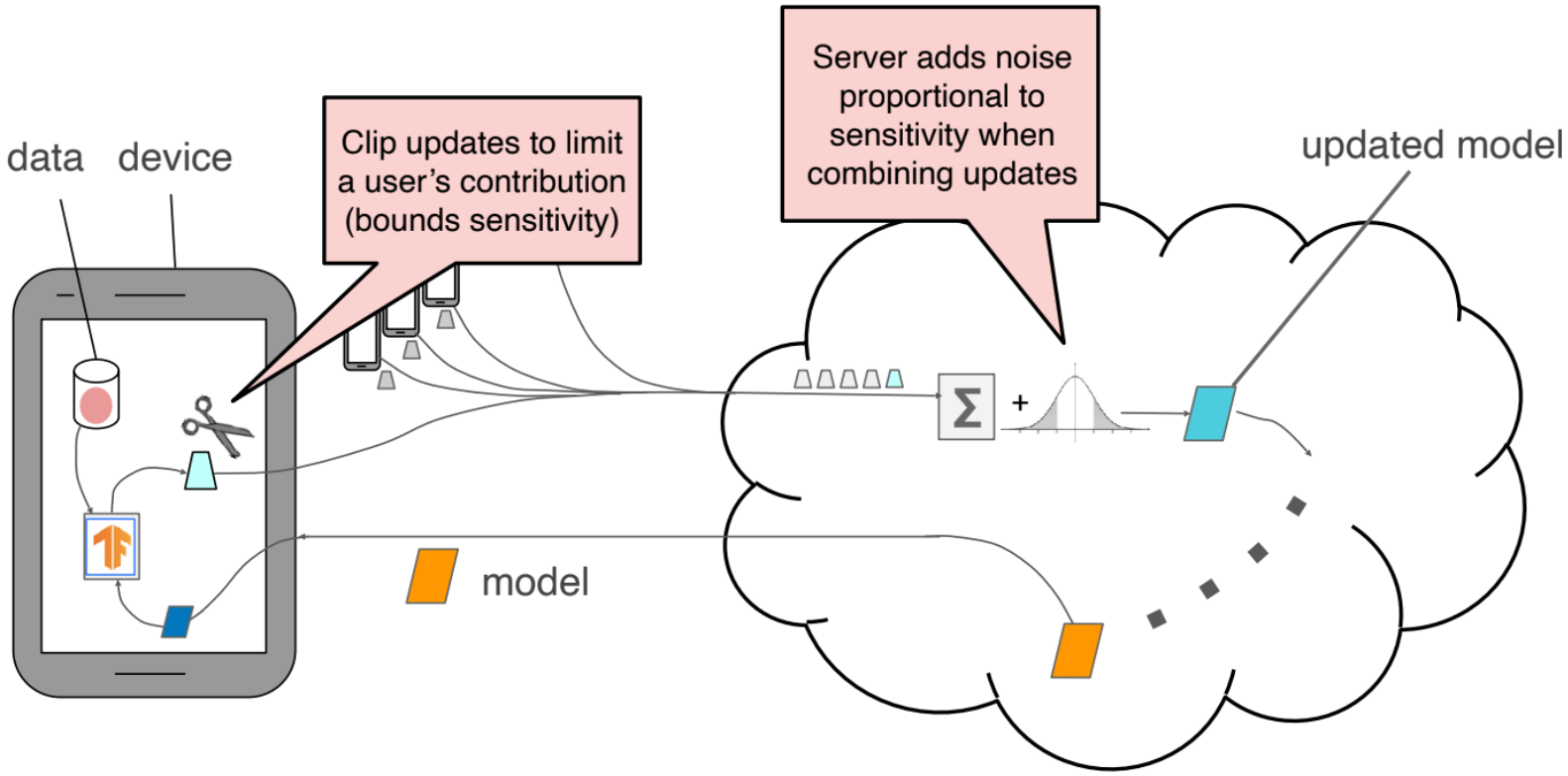
Picture from Federated and Collaborative Learning, Fall 2023, Virginia Smith. https://www.cs.cmu.edu/~smithv/10719/

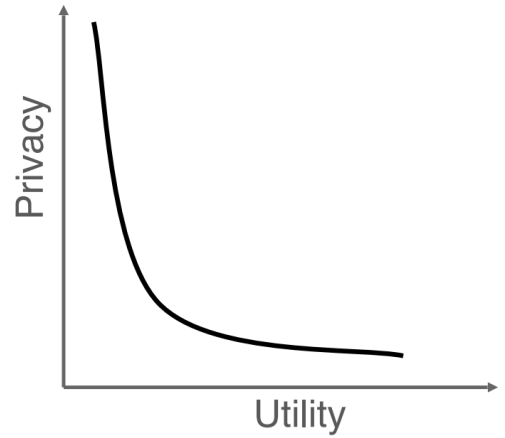
Ma, Jing, et al. "Privacy-preserving federated learning based on multi-key homomorphic encryption." *International Journal of Intelligent Systems* 37.9 (2022): 5880-5901.

Zhang, Chen, et al. "A survey on federated learning." *Knowledge-Based Systems* 216 (2021): 106775.

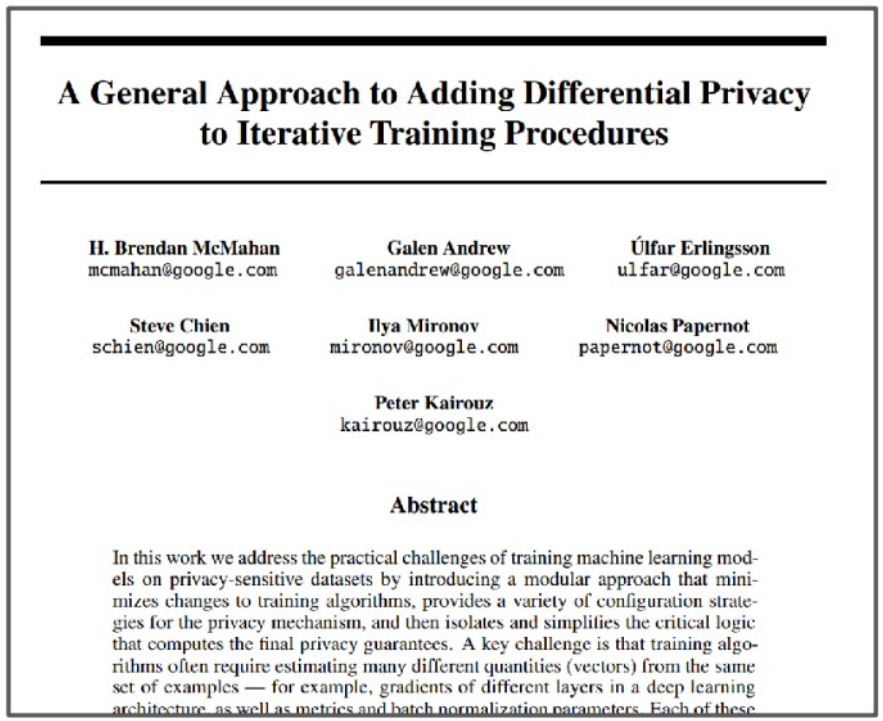
Li, Yong, et al. "Privacy-preserving federated learning framework based on chained secure multiparty computing." *IEEE Internet of Things Journal* 8.8 (2020): 6178-6186.

PRIVACY CONCERNS





[Credit: **P. Kairouz, B. McMahan**, V. Smith, FL Tutorial, NeurIPS 2020]



A General Approach to Adding Differential Privacy to Iterative Training Procedures

H. Brendan McMahan
mcmahan@google.com

Galen Andrew
galenandrew@google.com

Úlfar Erlingsson
ulfar@google.com

Steve Chien
schien@google.com

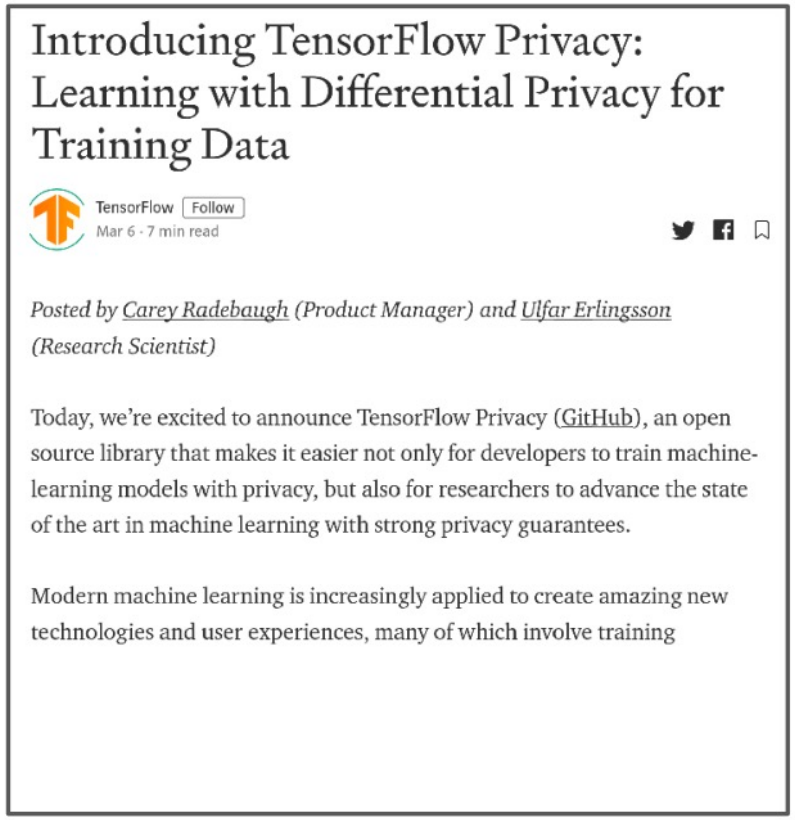
Ilya Mironov
mironov@google.com

Nicolas Papernot
papernot@google.com

Peter Kairouz
kairouz@google.com

Abstract

In this work we address the practical challenges of training machine learning models on privacy-sensitive datasets by introducing a modular approach that minimizes changes to training algorithms, provides a variety of configuration strategies for the privacy mechanism, and then isolates and simplifies the critical logic that computes the final privacy guarantees. A key challenge is that training algorithms often require estimating many different quantities (vectors) from the same set of examples — for example, gradients of different layers in a deep learning architecture, as well as metrics and batch normalization parameters. Each of these



Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data

TensorFlow Follow
Mar 6 · 7 min read

Posted by Carey Radebaugh (Product Manager) and Ulfar Erlingsson (Research Scientist)

Today, we're excited to announce TensorFlow Privacy (GitHub), an open source library that makes it easier not only for developers to train machine-learning models with privacy, but also for researchers to advance the state of the art in machine learning with strong privacy guarantees.

Modern machine learning is increasingly applied to create amazing new technologies and user experiences, many of which involve training

[Credit: **P. Kairouz, B. McMahan**, V. Smith, FL Tutorial, NeurIPS 2020]

Wei, Kang, et al. "Federated learning with differential privacy: Algorithms and performance analysis." *IEEE transactions on information forensics and security* 15 (2020): 3454-3469.

BIGGER PICTURE: COLLABORATIVE LEARNING

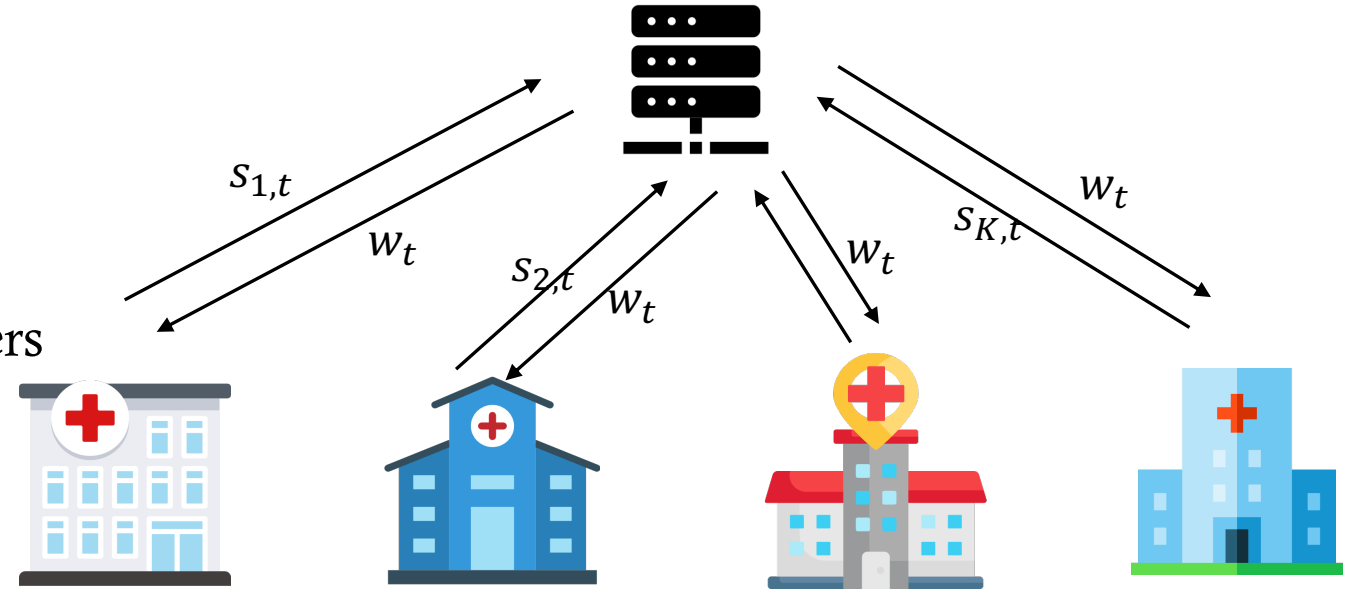
- *learning across multiple sources, stakeholders*
- ***how to incentivize participation***
- ***develop trustworthy learning schemes***
- ***effectively leveraging heterogeneous sources***

Pictures from internet

SECTION II: FEDERATED LEARNING VIA FEDERATED OPTIMIZATION

BASIC PROBLEM SET-UP

- Objective function: $L(w) = \sum_{k=1}^{K} L_k(w; D_k)$
- Optimization over all sites: $\hat{w} = \text{argmin}_w L(w)$
- Initialization
- Local sites share summary stats with central server
- Central server update model parameters



HOWTO REDUCE COMMUNICATION COST

- Leveraging properties of the loss function
- Good initialization
- Local updates
- Trade-offs between information shared per communication and the number of communications

ONE-SHOT FEDERATED LEARNING

- Linear model

$$L(\beta) = \sum_{k=1}^{K} ||y_k - X_k\beta||^2 = y_k^T y_k + \beta^T \boldsymbol{X_k^T X_K} \beta - 2\boldsymbol{y_k^T X_k}\,\beta$$

Chen et al.(2006) Regression cubes with lossless compression and aggregation.

LINEAR MIXED MODEL

- y_{ki} is the continuous outcome, $k = 1, \cdots, K, i = 1, \cdots, n_k$,
- x_{ki} is the $p-$dim covariates vector with β the fixed effects,
- z_{ki} is the $q-$dim covariates vector with u_k the random effects. z_{ki} usually is a subset of x_{ki}.
- $y_k = (y_{k1}, \cdots, y_{kn_k})^T$, $X_k = (x_{k1}, \cdots, x_{kn_k})^T$, $Z_k = (z_{k1}, \cdots, z_{kn_k})^T$,

$$y_k = X_k\beta + Z_k u_k + \epsilon_k,$$

with $u_k \sim N(0, V)$, $V = diag(\sigma_1^2, \cdots, \sigma_q^2)$, random error $\epsilon_k \sim N(0, \sigma^2 I_{n_k})$.

- Need to estimate β, and variance components $\theta = (\sigma^2, \sigma_1^2, \cdots, \sigma_q^2)$.

- Log-likelihood,

$$\ell(\beta, \theta) = -\frac{1}{2} \sum_{k=1}^{K} \{log|\Sigma_k| + (y_k - X_k\beta)^T \Sigma_k^{-1} (y_k - X_k\beta)\},$$

where $\Sigma_k = \Sigma_k(\theta) = Z_k V Z_k^T + \sigma^2 I_{n_k}$.

DLMM

- LMMs don't have closed-form solution
- Woodbury matrix identity

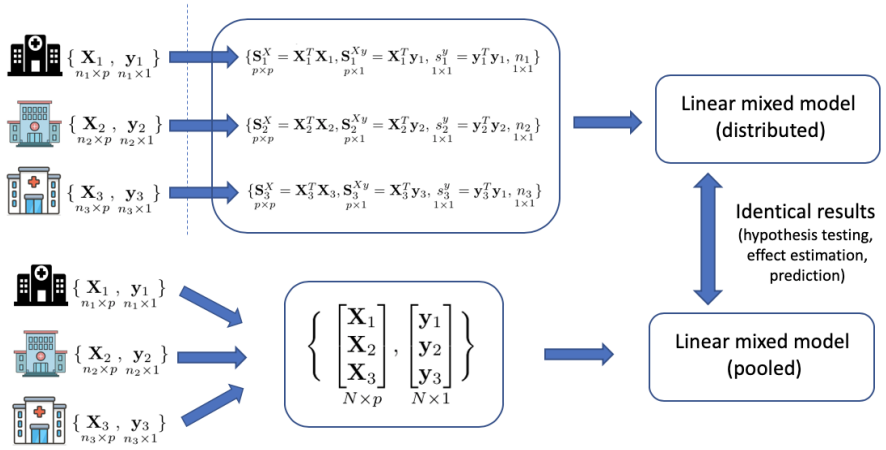
$$\Sigma_k^{-1} = \sigma^{-2}\{I_{n_k} - Z_k\left(\sigma^2 V^{-1} + Z_k^T Z_k\right)^{-1} Z_k^T\}$$

- Matrix determinant lemma

$$|\Sigma_k| = \sigma^{2n_k}|I_q + \sigma^{-2} V Z_k^T Z_k|$$

- Log-likelihood,

$$\ell(\beta, \theta) = -\frac{1}{2}\sum_{k=1}^{K}\{log|\Sigma_k| + (y_k - X_k\beta)^T \Sigma_k^{-1}(y_k - X_k\beta)\}$$



nature communications

Explore content | About the journal | Publish with us

nature > nature communications > articles > article

Article | Open Access | Published: 30 March 2022

DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models

ITERATIVE ALGORITHM: GOOD INITIALIZATION

Initialize from local site

Leveraging pre-trained models

Initialize by averaging across all local models

Nguyen, John, et al. "Where to begin? on the impact of pre-training and initialization in federated learning." *arXiv preprint arXiv:2206.15387* (2022).

ITERATIVE ALGORITHM: FEDERATED AVERAGING (FEDAVG) VERSUS DISTRIBUTED GD (SGD)

- FedAvg: Locally updating models and taking the average of all local updates to get the global update
- Distributed GD: Sequentially update model across sites.

Distributed SGD: computation on device k

for $i \in$ *mini-batch* B
  $\Delta \mathbf{w} \leftarrow \Delta \mathbf{w} - \alpha \nabla f_i(\mathbf{w})$
end
$\mathbf{w} \leftarrow \mathbf{w} + \Delta \mathbf{w}$

FedAvg: computation on device k

for $t = 1, 2, \ldots,$ *local iterations* T
  $\Delta \mathbf{w} \leftarrow \Delta \mathbf{w} - \alpha \nabla f_{i_t}(\mathbf{w})$
  $\mathbf{w} \leftarrow \mathbf{w} + \Delta \mathbf{w}$
end

Why is it useful to perform `local-updating`?

1. Can perform more local computation (i.e., more than just one mini-batch)
2. Incorporate updates more quickly (immediately apply gradient information)

✓ **Can lead to method converging in many fewer communication rounds**

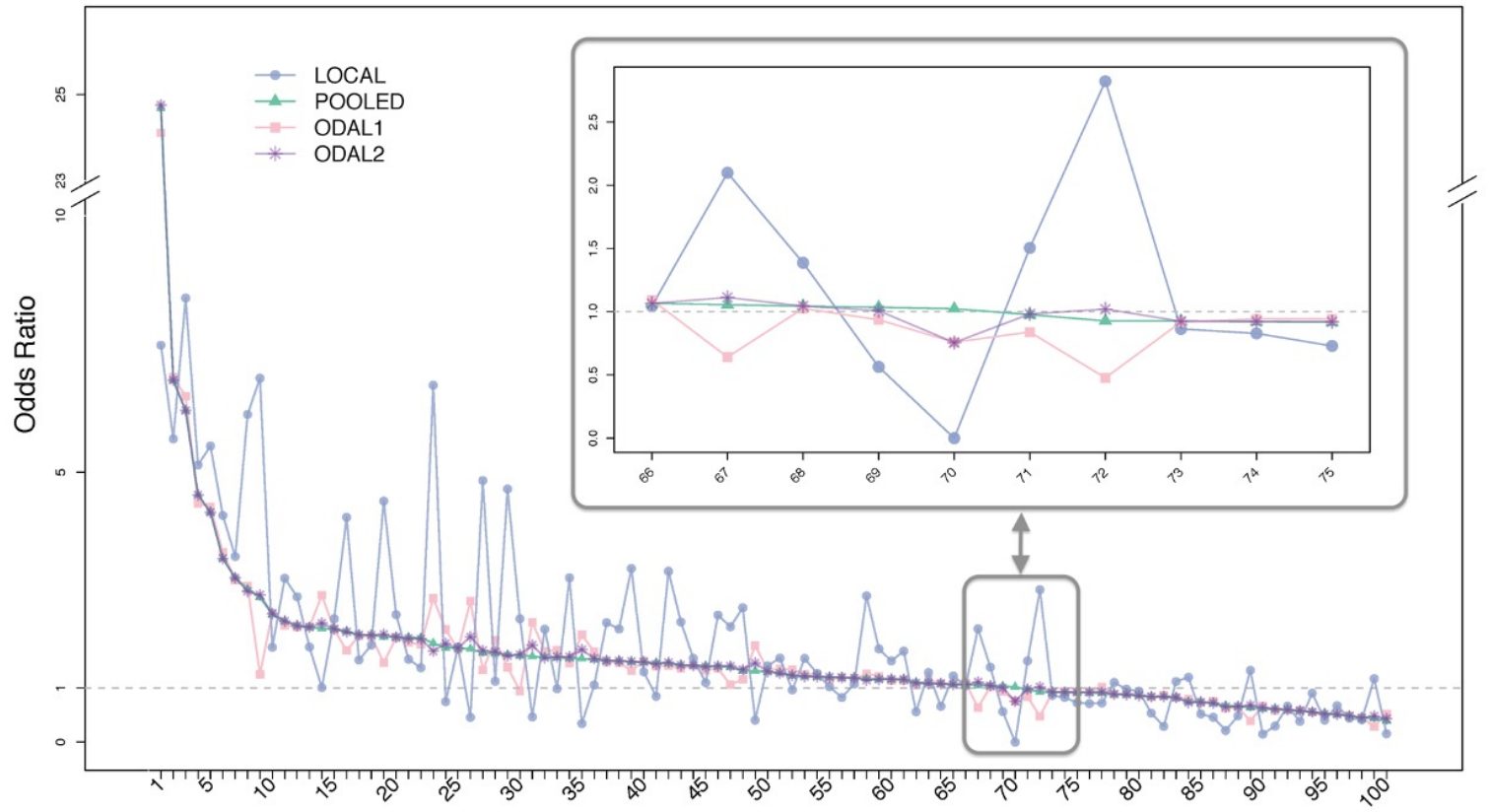
Li, Xiang, et al. "On the convergence of fedavg on non-iid data." *arXiv preprint arXiv:1907.02189* (2019).

MNIST CNN, 99% ACCURACY					
CNN	E	B	u	IID	NON-IID
FEDSGD	1	∞	1	626	483
FEDAVG	5	∞	5	179 (3.5×)	1000 (0.5×)
FEDAVG	1	50	12	65 (9.6×)	600 (0.8×)
FEDAVG	20	∞	20	234 (2.7×)	672 (0.7×)
FEDAVG	1	10	60	34 (18.4×)	350 (1.4×)
FEDAVG	5	50	60	29 (21.6×)	334 (1.4×)
FEDAVG	20	50	240	32 (19.6×)	426 (1.1×)
FEDAVG	5	10	300	20 (31.3×)	229 (2.1×)
FEDAVG	20	10	1200	18 (34.8×)	173 (2.8×)
SHAKESPEARE LSTM, 54% ACCURACY					
LSTM	E	B	u	IID	NON-IID
FEDSGD	1	∞	1.0	2488	3906
FEDAVG	1	50	1.5	1635 (1.5×)	549 (7.1×)
FEDAVG	5	∞	5.0	613 (4.1×)	597 (6.5×)
FEDAVG	1	10	7.4	460 (5.4×)	164 (23.8×)
FEDAVG	5	50	7.4	401 (6.2×)	152 (25.7×)
FEDAVG	5	10	37.1	192 (13.0×)	41 (95.3×)

Local-updating (FedAvg) can reduce communication rounds by ~100x relative to SGD

[McMahan et al, Communication-efficient learning of deep networks from decentralized data, AISTATS 2017]

HIGHER ORDER DERIVATIVES

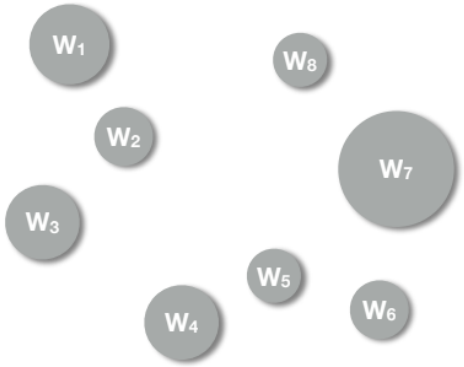


- ODAL2 needs to transfer an extra hessian matrix.
- ODAL2 provides more accurate estimation than ODAL1.

Duan, Rui, et al. "Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm." *Journal of the American Medical Informatics Association* 27.3 (2020): 376-385.

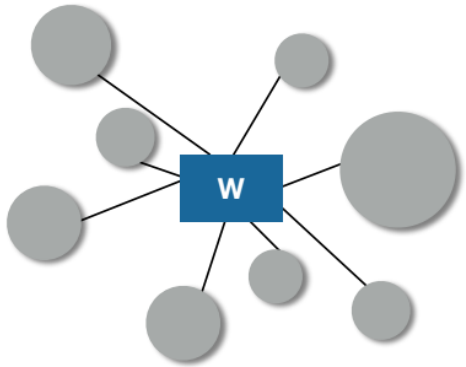
HOWTO MODEL HETEROGENOUS DATA

local



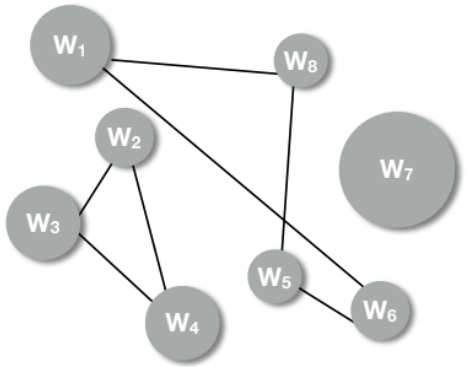
✓ *personalized models*
✗ *don't learn from peers*

global



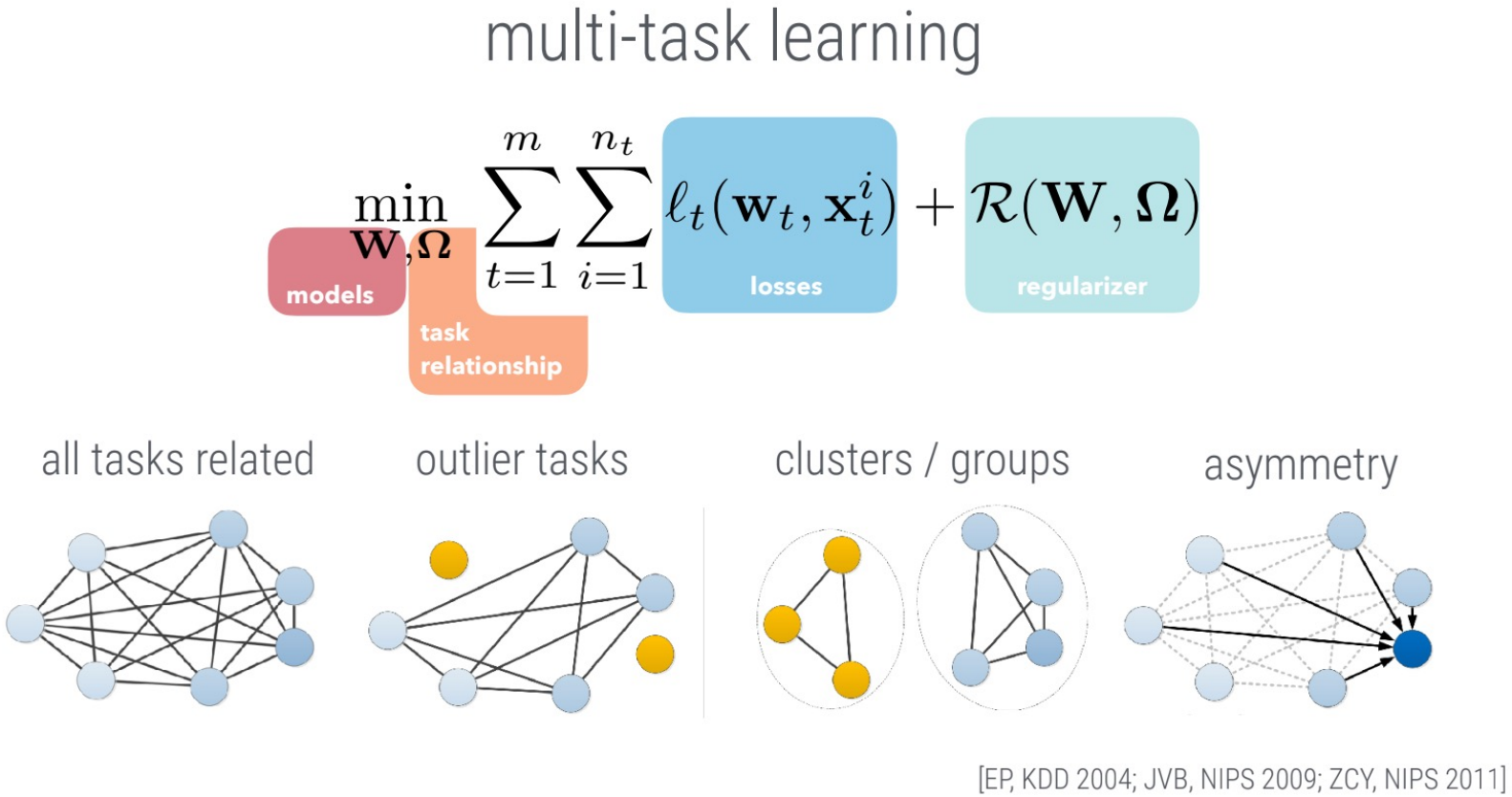
✗ *non-personalized models*
✓ *learn from peers*

??



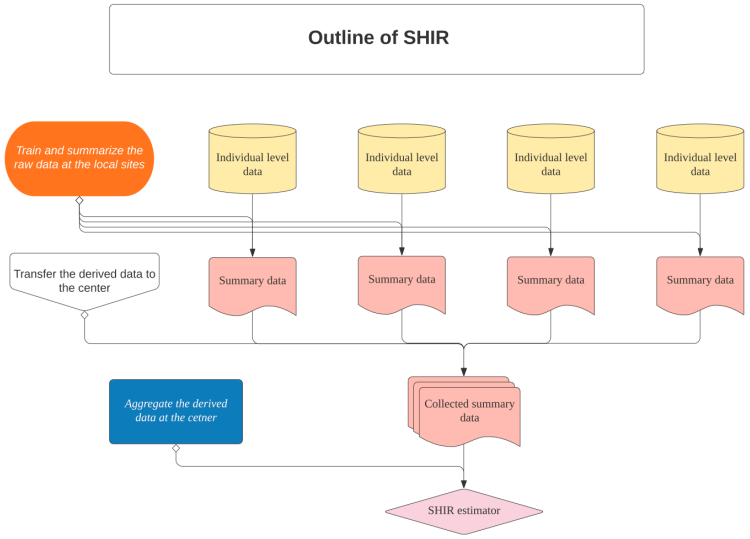
✓ *personalized models*
✓ *learn from peers*

FEDERATED MULTI-TASK LEARNING



Smith, Virginia, et al. "Federated multi-task learning." *Advances in neural information processing systems* 30 (2017).

AN EXAMPLE: SHIR



At the center node, we decompose the coefficients into mean effects and heterogeneous effects:

$$\{\mu, \alpha^{(\bullet)}\} \text{ where } \alpha^{(\bullet)} = (\alpha^{(1)}, ..., \alpha^{(M)}),\ \beta^{(m)} = \mu + \alpha^{(m)},\ \alpha^{(1)} + \cdots + \alpha^{(M)} = 0,$$

and fit the quadratic-approximated integrative regression:

$$\{\widehat{\mu}_{\text{SHIR}}, \widehat{\alpha}^{(\bullet)}_{\text{SHIR}}\} = \text{argmin}_{\mu,\alpha^{(\bullet)}} N^{-1} \sum_{m=1}^{M} n_m Q_m(\mu, \alpha^{(m)}) + \lambda(\|\mu_{-1}\|_1 + \lambda_g \|\alpha^{(\bullet)}_{-1}\|_{2,1}),$$

$$\text{where} \quad Q_m(\mu, \alpha^{(m)}) = (\mu + \alpha^{(m)})^{\mathsf{T}} \widehat{H}_m (\mu + \alpha^{(m)}) - 2\widehat{g}_m^{\mathsf{T}} (\mu + \alpha^{(m)}),$$

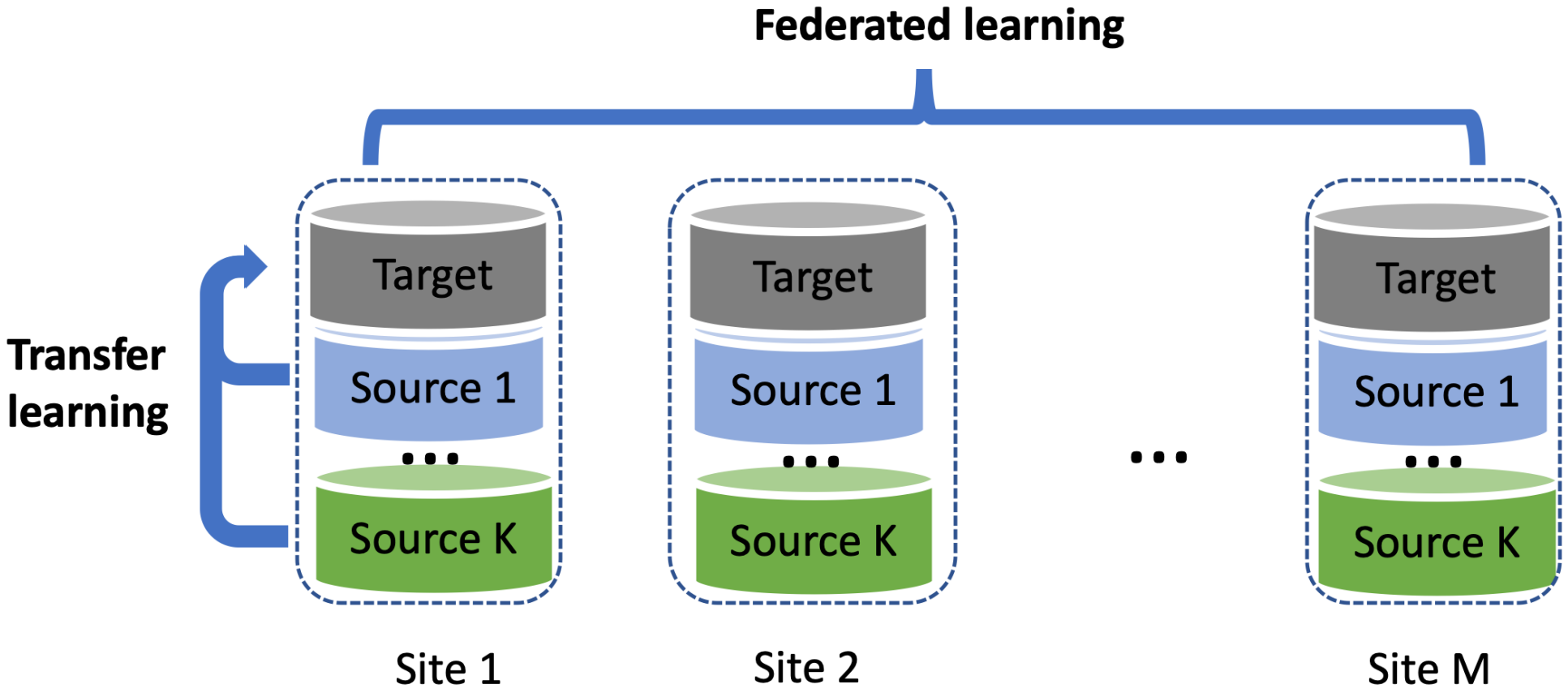
to obtain the SHIR estimator. Please see more details from the SHIR paper linked in the citation section.

https://celehs.github.io/SHIR/

Cai, Tianxi, Molei Liu, and Yin Xia. "Individual data protected integrative regression analysis of high-dimensional heterogeneous data." *Journal of the American Statistical Association* 117.540 (2022): 2105-2119.

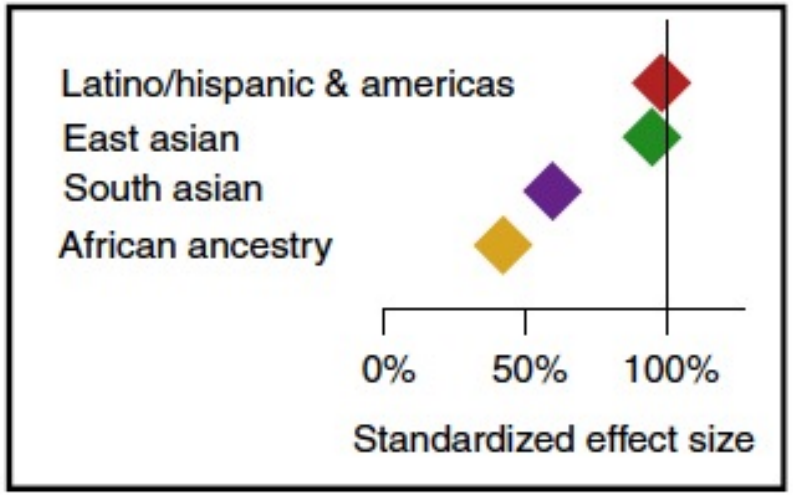
A FEDERATED TRANSFER LEARNING APPROACH

- Li S, Cai T, Duan R. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. AOAS *in press.*



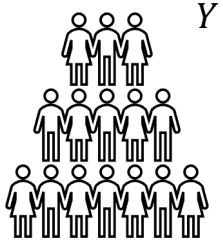
LACK OF REPRESENTATION OF BIOBANK DATA

Table 1: Sample size information of several large biobanks.					
Study	**N**	**%EA**	**%AA**	**%Hisp**	**%Asian**
Million Veteran Program (MVP)†	825K	73	19	8	0
UK Biobank (UKB)†	500K	94	1.6	0.6	2.3
eMERGE†	83K	76	18	5	1
Mass General Brigham (MGB)†	112K	82	6	5	4
Penn Medicine Biobank (PMBB)†	62K	68	21	2.5	2.4
TOPMed	144K	40	32	16	10
All of us	400K	50	18	18	7



Duncan, Laramie, et al. *Nature communications* 10.1 (2019): 1-9.

MODEL HETEROGENEITY



$$Y|X \sim f(y, \boldsymbol{x}; \boldsymbol{\beta_1})$$
$$X \sim g_1(\boldsymbol{x})$$

$$Y|X \sim f(y, \boldsymbol{x}; \boldsymbol{\beta_0})$$
$$X \sim g_0(\boldsymbol{x})$$

- Account for covariate shift (Pan & Yang 2010), i.e., $g_0(x) \neq g_1(x)$
  - Density ratio model: $w(\boldsymbol{x}) = \frac{g_0(\boldsymbol{x})}{g_1(\boldsymbol{x})}$
  - E.g., exponential tilting model (Liang & Qin, 2000, Luo & Tsai 2012)

$$w(\boldsymbol{x}; \boldsymbol{\eta}) = \exp\{\eta_0 + \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{\psi}(\boldsymbol{x})\}$$

- Outcome model shift , i.e., $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_1$.
  - Specify common and population specific parameters (Duan et al, 2022)

$$\boldsymbol{\beta}_0 = (\boldsymbol{\theta}, \boldsymbol{\gamma_0});\ \boldsymbol{\beta}_1 = (\boldsymbol{\theta}, \boldsymbol{\gamma_1})$$

  - Similarity constraints (Li et al, 2021, 2022)

$$d(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}) < h$$

CENTRALIZED SETTING: WITHOUT DATA-SHARING CONCERNS

- Fitting a density ratio model:
  - Exponential titling is equivalent as a logistic regression model with outcome being the population indicator.
  - Z = 1 (target) Z=0 (source)
  - $P(Z=1 \mid X) = \text{expit}\{\eta_0 + \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{\psi}(\boldsymbol{x})\}$

- Joint estimation of outcome model parameters

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\beta} \; -\sum_{\boldsymbol{i} \in \boldsymbol{I_T}} \log f(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{\beta_0}) - \sum_{k=1}^{K} \sum_{\boldsymbol{i} \in \boldsymbol{I_s}} \widehat{w}_i \log f(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{\beta_1}) + \lambda \mathcal{P}(\boldsymbol{\beta_0}, \boldsymbol{\beta_1})$$

  - **E.g.,** $\mathcal{P}(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}) = ||\boldsymbol{\beta_0} - \boldsymbol{\beta_1}||_q$

FEDERATED SETTING

- Estimating density ratio model when source and target data cannot be pooled together
  - For any $h(X)$, we have

$$E_T\{h(\boldsymbol{X})\} = E_S\{h(\boldsymbol{X})w(\boldsymbol{X};\boldsymbol{\eta})\}$$

  - We have the estimating equation for solving $\widehat{\boldsymbol{\eta}}$

$$\frac{1}{n_T}\sum_{i\in I_T} h\ (\boldsymbol{x_i}) = \frac{1}{n_S}\sum_{\boldsymbol{i}\in \boldsymbol{I_S}} h\ (\boldsymbol{x_i})\mathrm{w}(\boldsymbol{x_i};\boldsymbol{\eta})$$

- Approximation of the loss function: for any loss function $\boldsymbol{L}(\boldsymbol{b})$, we have the quadratic approximation with an initial value $\bar{\boldsymbol{b}}$

$$\tilde{\boldsymbol{L}}(\boldsymbol{b};\bar{\boldsymbol{b}}) \propto \nabla \boldsymbol{L}(\bar{\boldsymbol{b}})^{\mathrm{T}}(\boldsymbol{b}-\bar{\boldsymbol{b}}) + \frac{1}{2}(\boldsymbol{b}-\bar{\boldsymbol{b}})^{T}\nabla^{\boldsymbol{2}}\boldsymbol{L}(\bar{\boldsymbol{b}})(\boldsymbol{b}-\bar{\boldsymbol{b}})$$

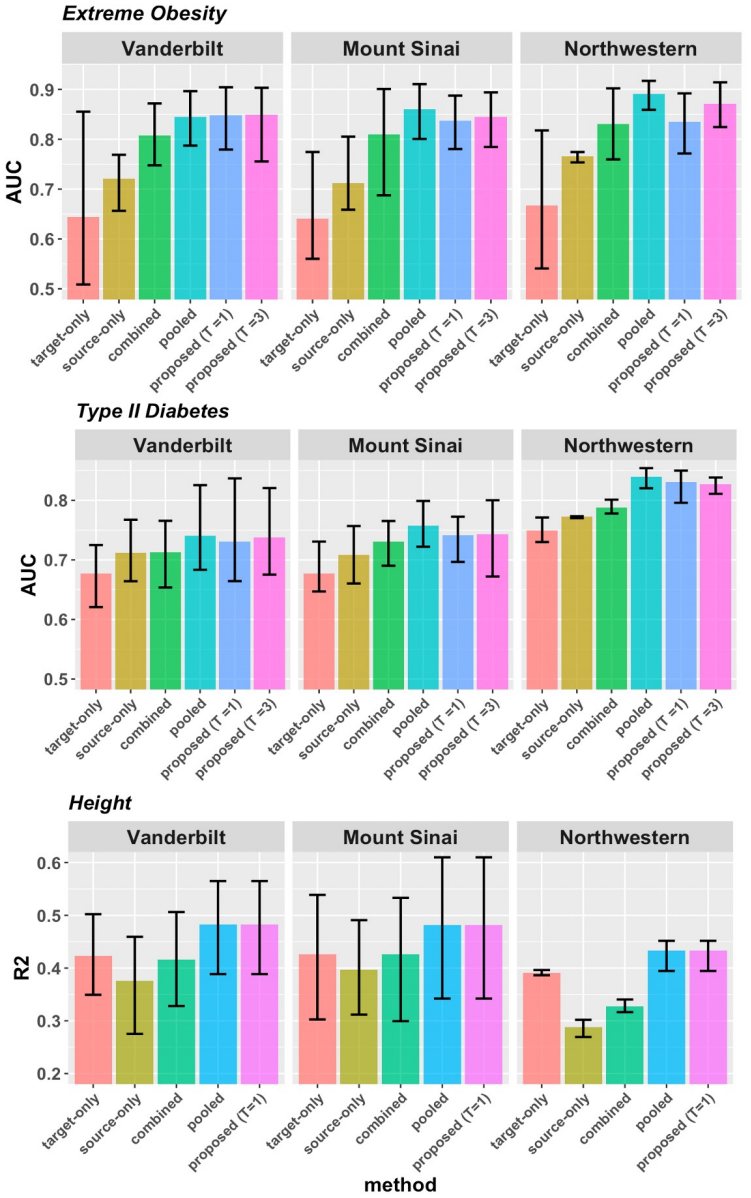
- $\frac{1}{n_T}\sum_{\boldsymbol{i}\in \boldsymbol{I_T}} h\ (\boldsymbol{x_i}), \nabla \boldsymbol{L}(\bar{\boldsymbol{b}}), \nabla^{\boldsymbol{2}}\boldsymbol{L}(\bar{\boldsymbol{b}})$ are summary statistics can be shared across sites

APPLICATION TO EMERGE

- Genetic risk models
- Integrate data from 7 sites at eMERGE
- Treat AA as target
- Three testing datasets from Vanderbilt, Mt Sinai, and Northwestern

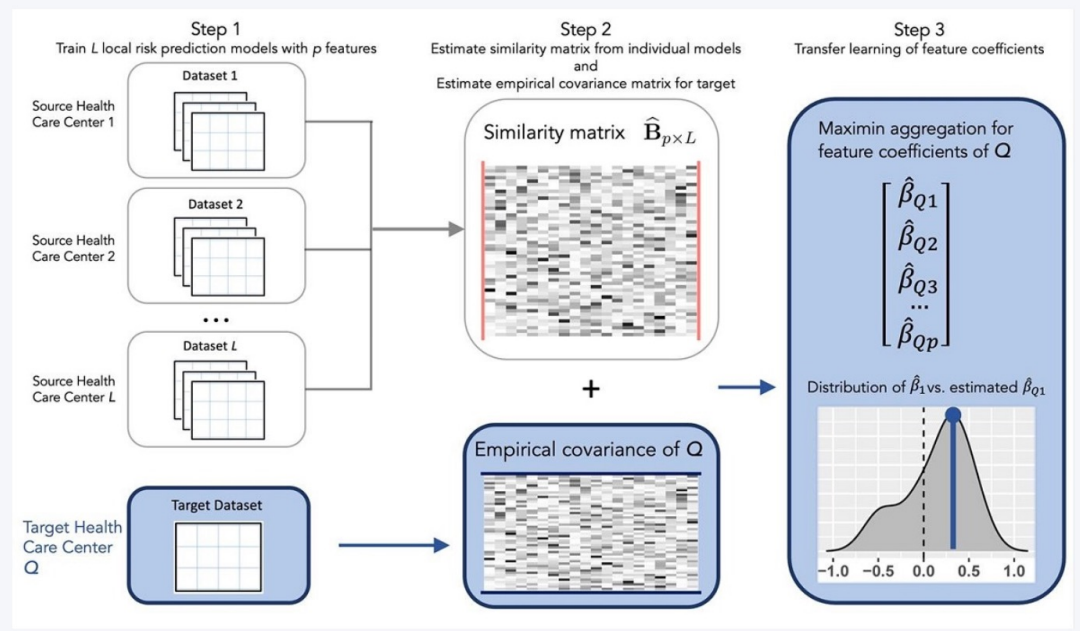
Table 1: Sample sizes of the target (AA) and the source (all other races) populations for type II diabetes, extreme obesity and height across seven sites.

	Extreme Obesity		Height		Type II Diabetes	
Site Name	source	target	source	target	source	target
Geisinger Health System	1023	4	2643	8	3081	9
Group Health Cooperative	273	13	2710	94	262	16
Marshfield Clinic	1497	0	3959	2	3977	3
Mayo Clinic	1320	3	4790	11	2880	10
Mount Sinai	534	1142	524	3402	606	3653
Northwestern University	1305	167	1178	180	1223	301
Vanderbilt University	1523	646	6598	1852	2580	1646



ZERO-SHOT LEARNING: INCORPORATING PRE-TRAINED MODELS

- Model aggregation
- Transfer learning
- Maximin Learning

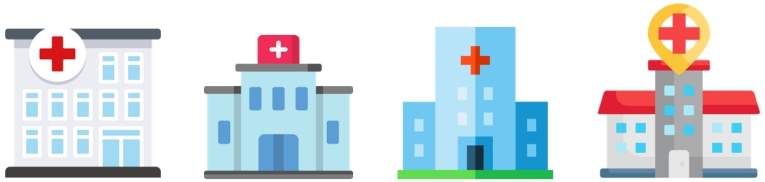


https://shiny.parse-health.org/ARRTLE/

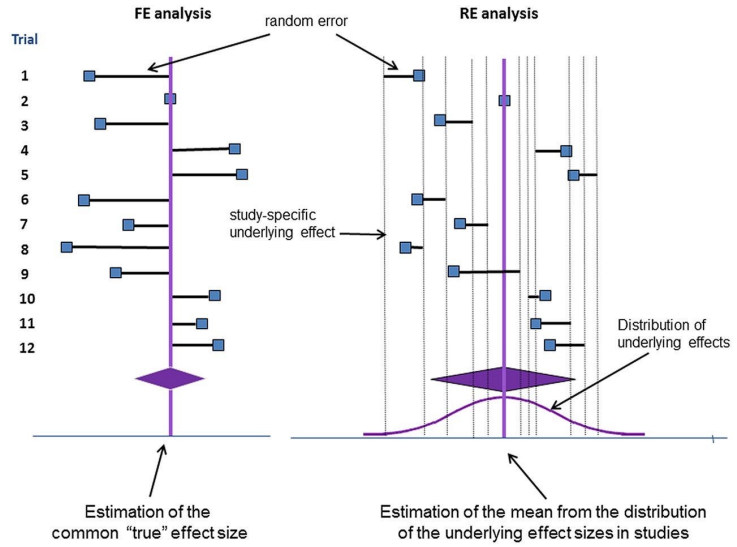
- Lecué, Guillaume, and Philippe Rigollet. "Optimal learning with Q-aggregation." (2014): 211-224.
- Parisi, Fabio, et al. "Ranking and combining multiple predictors without labeled data." *Proceedings of the National Academy of Sciences* 111.4 (2014): 1253-1258.
- Gu, Tian, Yi Han, and Rui Duan. "Robust angle-based transfer learning in high dimensions." *arXiv preprint arXiv:2210.12759* (2022).
- Wang, Zhenyu, Peter Bühlmann, and Zijian Guo. "Distributionally robust machine learning with multi-source data." *arXiv preprint arXiv:2309.02211* (2023).

SECTION III: FEDERATED STATISTICAL INFERENCE

STATISTICAL INFERENCE OF SHARED PARAMETERS ACROSS SITES

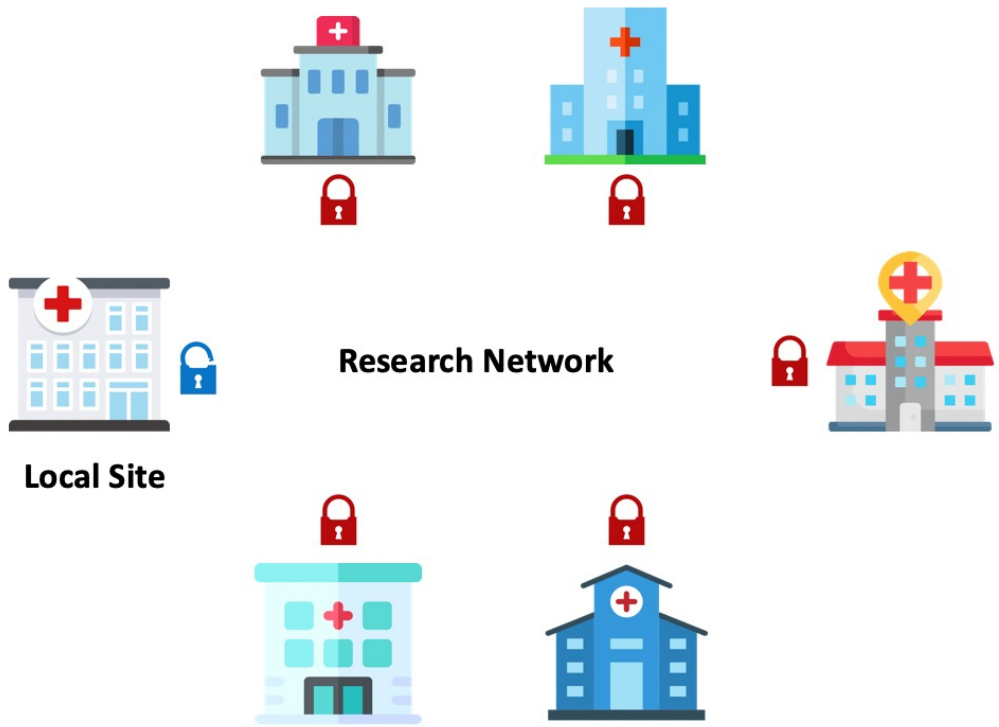


- Benchmark: Meta-analysis
- Average across local estimates



Borenstein, Michael, et al. *Introduction to meta-analysis*. John Wiley & Sons, 2021.

IID DATA: COMMUNICATION-EFFICIENT SURROGATE LIKELIHOOD



- Patient-level data from one site is accessible
- Only aggregated information can be shared between site

- **Combined likelihood function** (if data could be shared)

$$L(\beta) = \frac{1}{nK}\sum_{j=1}^{K}\log f(d_{ij};\beta)$$

- **Local likelihood function** (assume local site to be the first site, j=1)

$$L_1(\beta) = \frac{1}{n}\sum_{i=1}^{n}\log f(d_{i1};\beta)$$

- For an initial value $\bar{\beta}$,

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta-\bar{\beta}) + \sum_{t=1}^{\infty}\frac{1}{t!}\nabla^t L(\bar{\beta})(\beta-\bar{\beta})^{\otimes t}$$

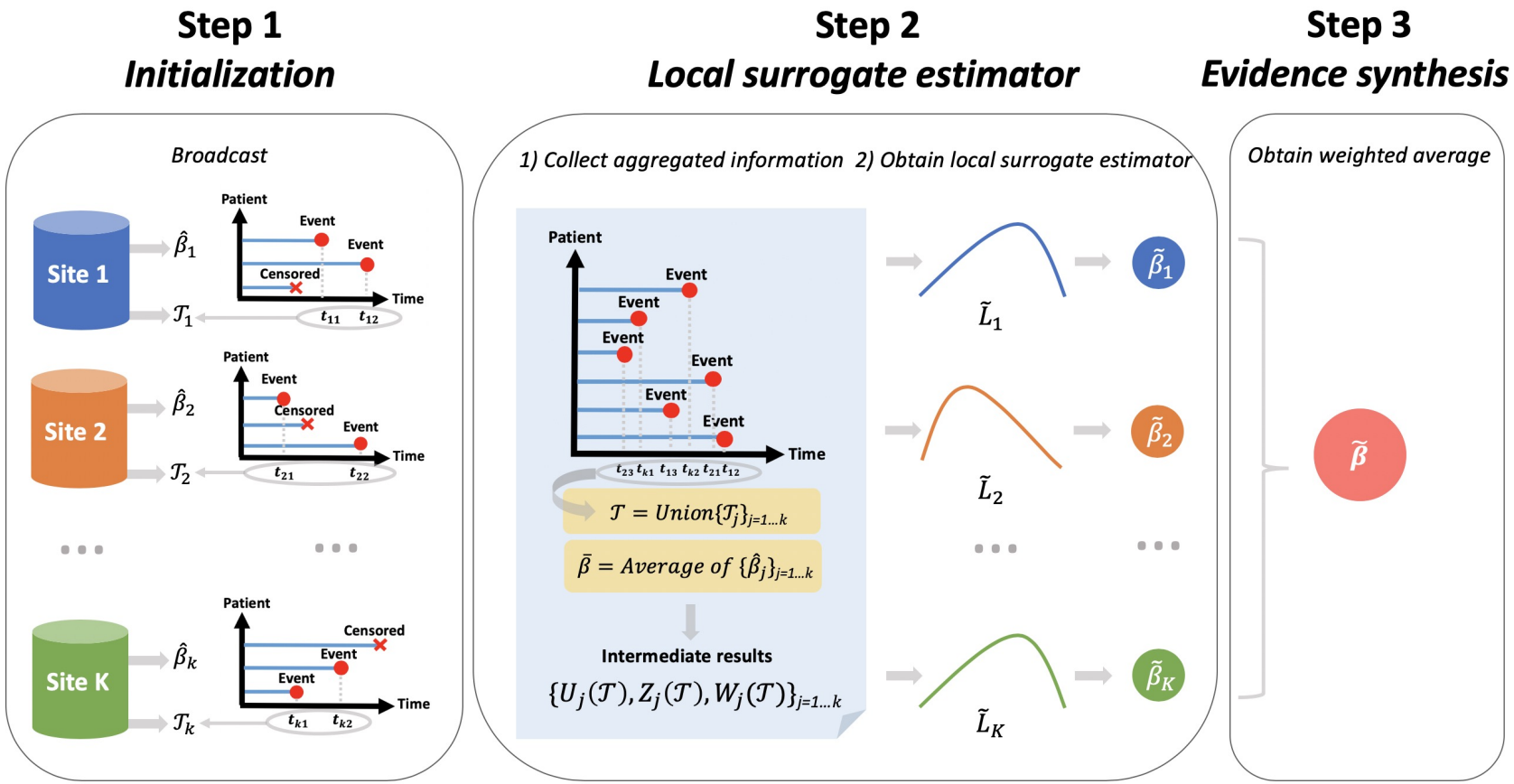
$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta-\bar{\beta}) + \sum_{t=1}^{\infty}\frac{1}{t!}\nabla^t L_1(\bar{\beta})(\beta-\bar{\beta})^{\otimes t}$$

$$\sum_{t=1}^{\infty}\frac{1}{t!}\nabla^t L_1(\bar{\beta})(\beta-\bar{\beta})^{\otimes t} = L_1(\beta) - L_1(\bar{\beta}) - \nabla L_1(\bar{\beta})^T(\beta-\bar{\beta})$$

SL function $\tilde{L}^1(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T\beta$

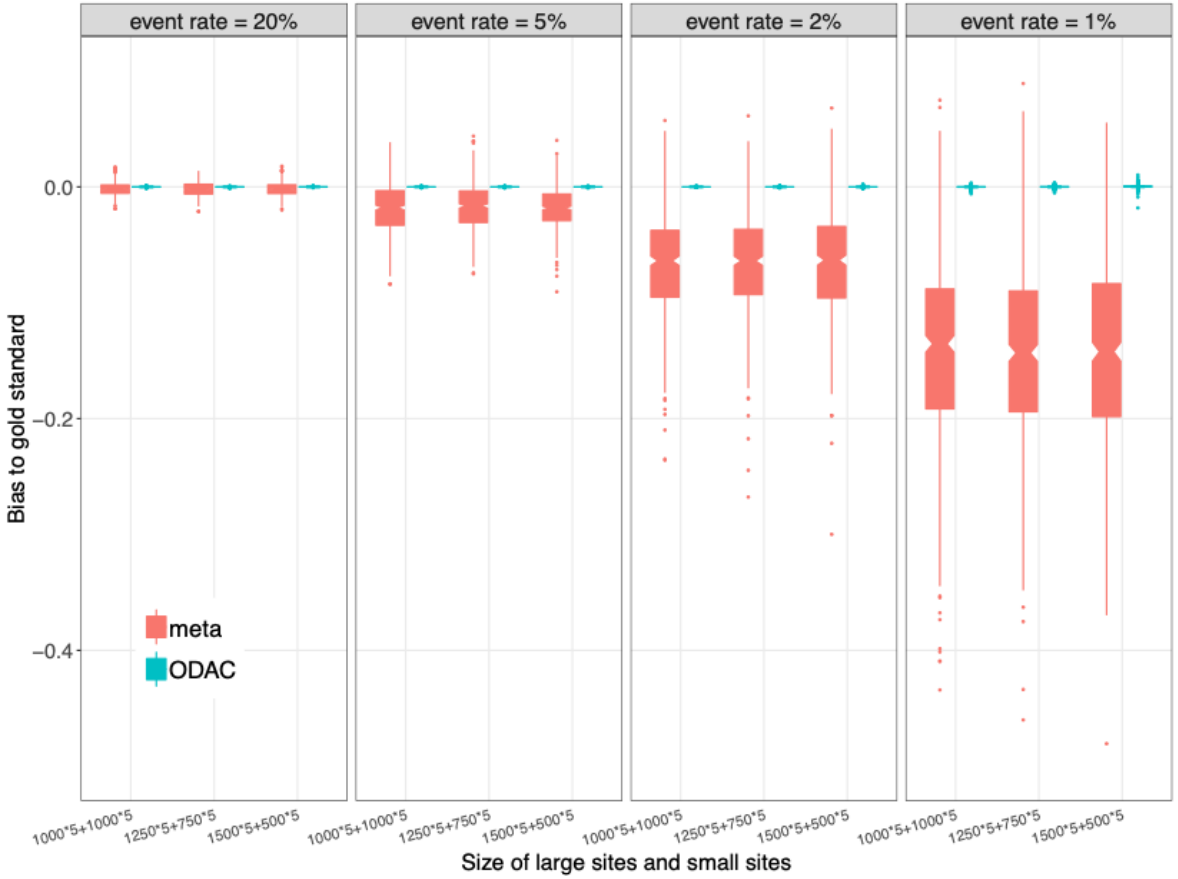
Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. Journal of the American Statistical Association. 2018 Nov 13.

ODAC



Duan, Rui, et al. "Learning from local to global: An efficient distributed algorithm for modeling time-to-event data." *Journal of the American Medical Informatics Association* 27.7 (2020): 1028-1036.

BENEFIT IN STUDYING RARE EVENTS



- Meta-analysis has increasing bias when event is rarer.
- ODAC provides estimates close to the pooled analysis.

Duan, Rui, et al. "Learning from local to global: An efficient distributed algorithm for modeling time-to-event data." *Journal of the American Medical Informatics Association* 27.7 (2020): 1028-1036.

ALLOW SITE-SPECIFIC PARAMETERS

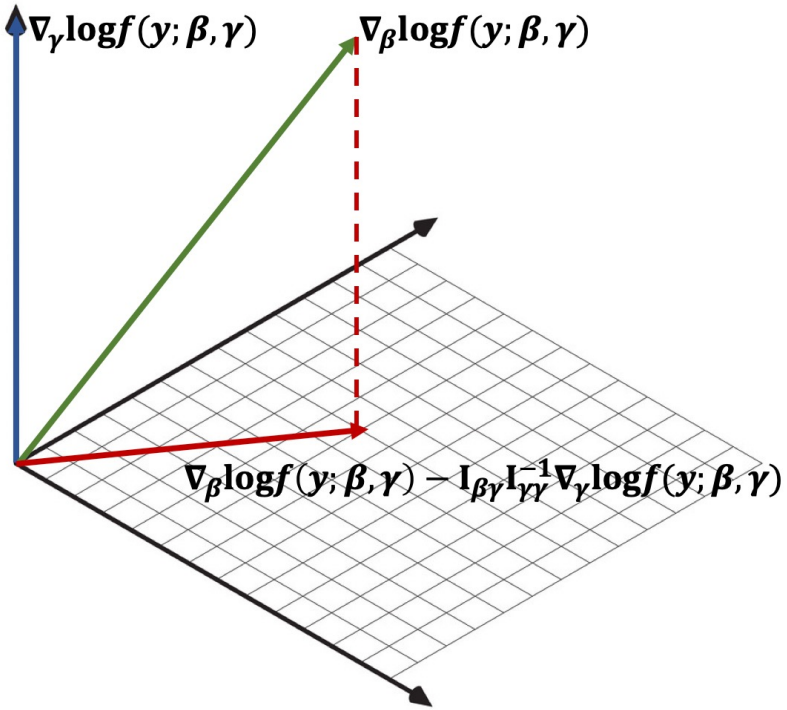
$$Y_{ij} \sim f(y; \boldsymbol{\beta}, \boldsymbol{\gamma_j})$$

- β is the *parameter of interest*
- γ_j is the *site-specific nuisance parameter--- allow site to be a covariate variable, allow interaction terms between site and other covariates.*

Change the target function to the combined efficient score function

$$L(\beta;\bar{\Gamma}) = \frac{1}{Kn}\sum_{j=1}^{K}\sum_{i=1}^{n}\log f(y_{ij};\beta,\bar{\gamma_j})$$

$$S(\beta;\bar{\Gamma}) = \frac{1}{Kn}\sum_{j=1}^{K}\sum_{i=1}^{n}\left\{\nabla_\beta \log f(y_{ij};\beta,\bar{\gamma}_j) - I^{(j)}_{\beta\gamma}{I^{(j)}_{\gamma\gamma}}^{-1}\nabla_\gamma \log f(y_{ij};\beta,\bar{\gamma}_j)\right\}$$



Duan, Rui, Yang Ning, and Yong Chen. "Heterogeneity-aware and communication-efficient distributed statistical inference." *Biometrika* 109.1 (2022): 67-83.

Find a local function and construct higher-order match

- We want to find some function $g(y;\beta)$, such that

$$\mathrm{E}_{f_1}\{\nabla^t g(Y_{i1};\beta)\} = \mathrm{E}\left\{\nabla_\beta^t S(\beta;\bar{\Gamma})\right\} = \frac{1}{K}\sum_{j=1}^{K} \mathrm{E}_{f_j}\nabla_\beta^{\mathrm{t}}\left\{\nabla_\beta \log f(Y_{ij};\beta,\bar{\gamma}_j) - I_{\beta\gamma}^{(j)} I_{\gamma\gamma}^{(j)^{-1}} \nabla_\gamma \log f(Y_{ij};\beta,\bar{\gamma}_j)\right\}$$

- Density ratio tilting
- For initial estimator $\bar{\beta}, \bar{\gamma}_j$, define

$$g(y;\beta) = \frac{1}{K}\sum_{j=1}^{K} \frac{\boldsymbol{f(y;\bar{\beta},\bar{\gamma}_j)}}{\boldsymbol{f(y;\bar{\beta},\bar{\gamma}_j)}}\left\{\nabla_\beta \log f(y;\beta,\bar{\gamma}_j) - \tilde{H}_{\beta\gamma}^{(j)} \tilde{H}_{\gamma\gamma}^{(j)^{-1}} \nabla_\gamma \log f(y;\beta,\bar{\gamma}_j)\right\}$$

Duan, Rui, Yang Ning, and Yong Chen. "Heterogeneity-aware and communication-efficient distributed statistical inference." *Biometrika* 109.1 (2022): 67-83.

- Define

Surrogate efficient score function

$$\widetilde{U}(\beta) = \boldsymbol{S}(\bar{\boldsymbol{\beta}}; \bar{\boldsymbol{\Gamma}}) + \boldsymbol{U}_1(\boldsymbol{\beta}) - \boldsymbol{U}_1(\bar{\boldsymbol{\beta}})$$

where $U_1(\beta) = \sum_{i=1}^{n} g(y_{i1}; \beta)/n$

- Obtain the proposed estimator by $\widetilde{U}(\beta) = 0$

- Each site need to provide $\bar{\gamma}_j$, and

$$S_j(\bar{\beta}; \bar{\gamma}_j) = \frac{1}{n}\sum_{i=1}^{n}\left\{\nabla_\beta \log f(Y_{ij}; \bar{\beta}, \bar{\gamma}_j) - H_{\beta\gamma}^{(j)} H_{\gamma\gamma}^{(j)^{-1}} \nabla_\gamma \log f(Y_{ij}; \bar{\beta}, \bar{\gamma}_j)\right\}$$

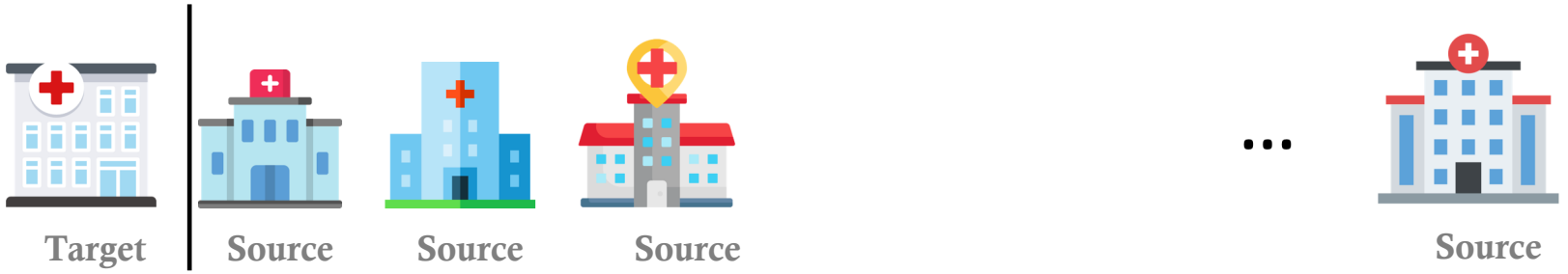
Duan, Rui, Yang Ning, and Yong Chen. "Heterogeneity-aware and communication-efficient distributed statistical inference." *Biometrika* 109.1 (2022): 67-83.

THEORETICAL PROPERTIES

Desired Properties	Meta-analysis	Proposed estimator (T=1)	Proposed estimator (T=2)
Consistency	consistent	consistent	consistent
Distance to the Gold Standard Estimator	$\frac{C}{\sqrt{Kn}}$	$\preccurlyeq \frac{C}{n}$	$\preccurlyeq \frac{C}{n\sqrt{K}} + \frac{C}{n\sqrt{n}}$
Asymptotic Normality	asymptotic normal	asymptotic normal	asymptotic normal
Asymptotic Efficient	not efficient	efficient	efficient

Duan, Rui, Yang Ning, and Yong Chen. "Heterogeneity-aware and communication-efficient distributed statistical inference." *Biometrika* 109.1 (2022): 67-83.

STATISTICAL INFERENCE IN A TARGET POPULATION



Target R = 0 | Source R = 1 | Source R = 2 | Source R = 3 | ... | Source R = K

Distributions $Q^0 = P_0^0$ P_1^0 P_2^0 ..... P_K^0

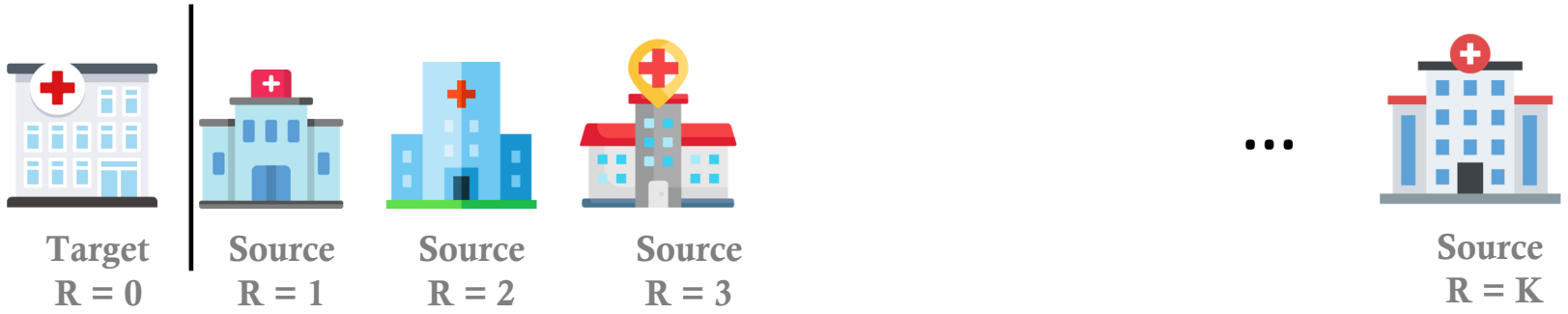
Estimand of interest is defined in the target $\psi(Q^0)$

How can all sites work collaboratively to estimate $\psi(Q^0)$?

CAUSAL INFERENCE OF TREATMENT EFFECTS

Consortium for Clinical Characterization of COVID-19 by EHR

ESTIMATE TARGET AVERAGE TREATMENT EFFECT (TATE)



- $Y^{(a)}$: potential outcome of patients under treatment $a = 0, 1$
- Estimand: $\Delta_T = \mathbf{E}(Y^{(1)} - Y^{(0)}|R = 0) = \mathbf{E}_{\mathbf{Q^0}}(Y^{(1)} - Y^{(0)})$
- At target, we assume positivity, consistency, no unmeasured confounding assumptions for the identification of TATE

$$\mathbf{E}_{\mathbf{Q^0};\mathbf{A},\mathbf{X}}(\mathbf{E}_{\mathbf{Q^0};\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{X}}(Y|A = 1, X)) - \mathbf{E}_{\mathbf{Q^0};\boldsymbol{A},\boldsymbol{X}}(\mathbf{E}_{\mathbf{Q^0};\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{X}}(Y|A = 0, X))$$

- Distribution shifts in $[Y, A, X]$ across sites

HOW CAN A HETEROGENEOUS SOURCE SITE HELP WITH ESTIMATING THE TARGET ESTIMAND?

- Distributional shifts in [A, X] can be handled by adjusting sample weights according to the density ratio and propensity.
- The conditional distribution [Y | A, X] is the key for consistently estimating the target ATE.
- Many assumes exchangeability: [Y | A, X] is the same across sites
  - Causal transportability: Bareinboim & Pearl, 2014; Rudolph & van der Laan, 2017; Dahabreh et al., 2019; Rudolph & van der Laan, 2017; Dahabreh & Hernán, 2019
  - More general data fusion: ;Athey et al., 2019; Kallus et al., 2020; Lee et al., 2023; Brantner et al., 2023

WHAT IS MORE PRACTICAL IN THE REAL-WORLD SETTINGS?

- Setting 1:
  - Among K source sites, exchangeability holds only among a subset (could be empty) of all source sites.
  - We do not know the membership of the subset

Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects

Larry Han[1,2], Jue Hou[3], Kelly Cho[4], Rui Duan[1†], Tianxi Cai[1,5†]

1 Department of Biostatistics, Harvard University
2 Department of Health Sciences, Northeastern University
3 Division of Biostatistics, University of Minnesota
4 Massachusetts Veterans Epidemiology Research and Information Center, US Department of Veteran Affairs
5 Department of Biomedical Informatics, Harvard Medical School
† Co-corresponding authors

Other recent work: Yang et al 2023-–test and pool, etc

OUTLINE OF THE METHOD

- Target site: obtain $\widehat{\Delta}_{T,0}$ from augmented inverse probability weighting (AIPW) approach.

$$\widehat{\Delta}_{T,0} = M_0 + \delta_0$$

$$M_0 = n^{-1} \sum_{i \in I_t} \{\widehat{m}(1, X_i) - \widehat{m}(0, X_i)\}$$

$$\delta_0 = n^{-1} \sum_{i \in I_T} \left\{ \frac{I(A_i=1)}{\widehat{\pi}_0(1,X_i)} - \frac{I(A_i=0)}{\widehat{\pi}_0(0,X_i)} \right\} \{Y_i - \widehat{m}(A_i, X_i)\}$$

- Source site(s): obtain $\widehat{\Delta}_{T,k}$ with the help of a density ratio model

$$w_k(\boldsymbol{X}) = \frac{f(\boldsymbol{X}|R=0)}{f(\boldsymbol{X}|R=k)}$$

$$\delta_k = n_k^{-1} \sum_{i \in I_k} \widehat{w}_k(X_i) \left\{ \frac{I(A_i=1)}{\widehat{\pi}_k(1,X_i)} - \frac{I(A_i=0)}{\widehat{\pi}_k(0,X_i)} \right\} \{Y_i - \widehat{m}(A_i, X_i)\}$$

$$\widehat{\Delta}_{T,k} = M_0 + \delta_k$$

- Combine source and target estimators through an adaptive aggregation step

ADAPTIVE AGGREGATION

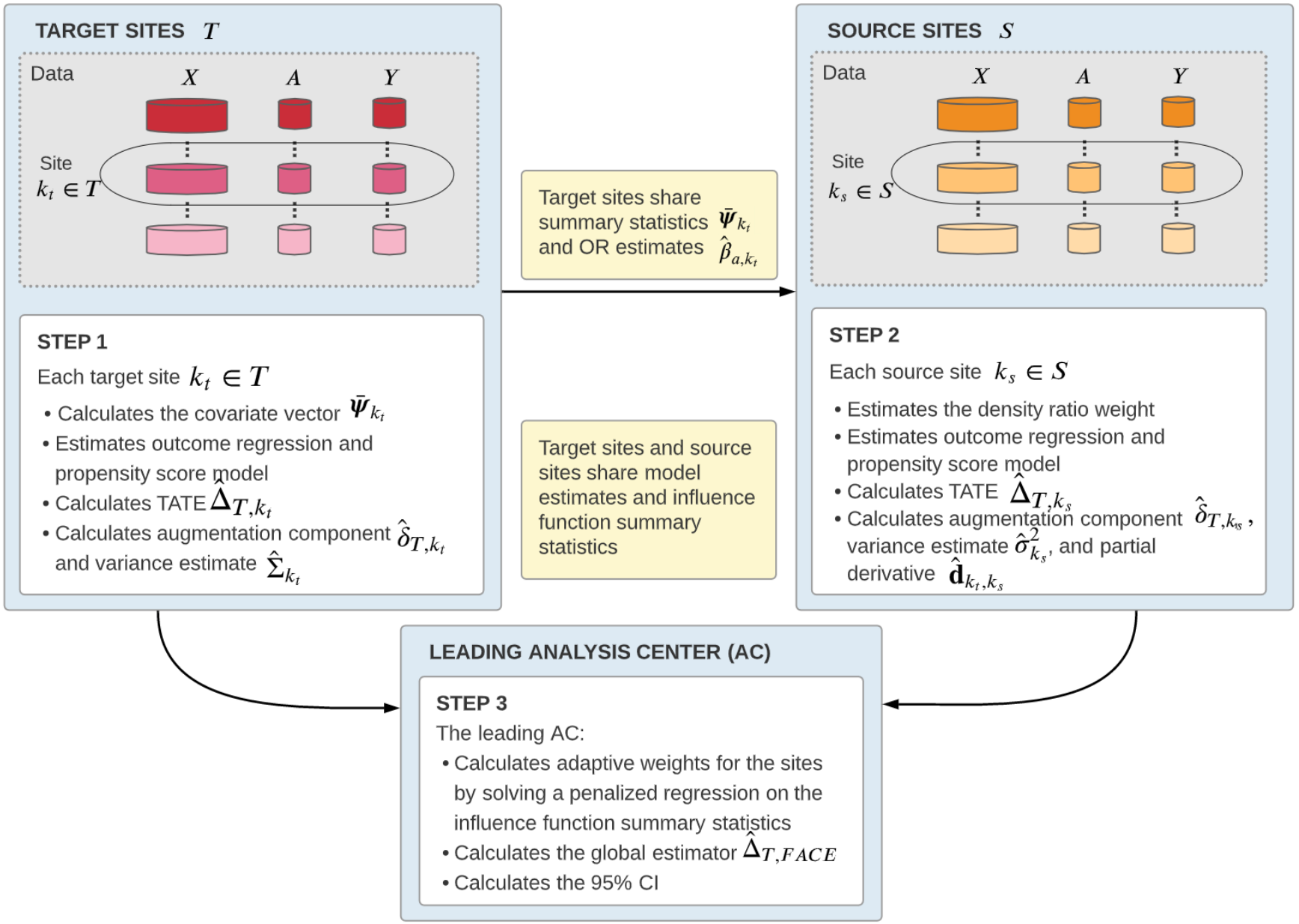
- Anchor and aggregation

$$\widehat{\Delta}_T = \widehat{\Delta}_{T,0} + \sum_{k \in S} \eta_k \{\widehat{\Delta}_{T,k} - \widehat{\Delta}_{T,0}\}$$

- Best weights learned through

$$\widehat{\boldsymbol{\eta}} = \text{argmin Var}\{\widehat{\Delta}_{T,0} + \textstyle\sum_{k \in S} \eta_k \{\widehat{\Delta}_{T,k} - \widehat{\Delta}_{T,0}\}\} + \lambda \textstyle\sum_{k \in S} |\eta_k| \left(\widehat{\Delta}_{T,k} - \widehat{\Delta}_{T,0}\right)^2$$

- Data splitting needed for adaptive aggregation
- Theoretical guarantees for asymptotical normality when target OR or PS model is correctly specified.
- Efficiency gain when exchangeability holds in at least one site


TARGET SITES T
Data
X A Y
Site
$k_t \in T$
SOURCE SITES S
Data
X A Y
Site
$k_s \in S$
Target sites share summary statistics $\bar{\Psi}_{k_t}$ and OR estimates $\hat{\beta}_{a,k_t}$
STEP 1
Each target site $k_t \in T$
• Calculates the covariate vector $\bar{\Psi}_{k_t}$
• Estimates outcome regression and propensity score model
• Calculates TATE $\hat{\Delta}_{T,k_t}$
• Calculates augmentation component $\hat{\delta}_{T,k_t}$ and variance estimate $\hat{\Sigma}_{k_t}$
Target sites and source sites share model estimates and influence function summary statistics
STEP 2
Each source site $k_s \in S$
• Estimates the density ratio weight
• Estimates outcome regression and propensity score model
• Calculates TATE $\hat{\Delta}_{T,k_s}$
• Calculates augmentation component $\hat{\delta}_{T,k_s}$, variance estimate $\hat{\sigma}^2_{k_s}$, and partial derivative $\hat{\mathbf{d}}_{k_t,k_s}$
LEADING ANALYSIS CENTER (AC)
STEP 3
The leading AC:
• Calculates adaptive weights for the sites by solving a penalized regression on the influence function summary statistics
• Calculates the global estimator $\hat{\Delta}_{T,FACE}$
• Calculates the 95% CI

APPLICATION TO COMPARATIVE EFFECTIVENESS OF COVID VACCINES

- To illustrate FACE, we study the comparative effectiveness of BNT162b2 (Pfizer) versus mRNA-1273 (Moderna) for the prevention of COVID-19 outcomes in five VA sites.
- It is of interest to understand the real-world effectiveness of these vaccines, but head-to-head comparisons have been rare.
- Inclusion criteria:
  - veteran status,
  - $\geq$ 18 years by Jan 1, 2021,
  - no previously documented COVID-19 infection, and
  - documented two-dose COVID-19 vaccination with either Pfizer or Moderna between Jan 1-Mar 24, 2021.
- All models were adjusted for age, sex, ethnicity, residence, comorbidities (CLD, CVD, HTN, T2DM, CKD, Autoimmune diseases, Obesity), and time.
- The outcomes of interest were documented SARS-CoV-2 infection either 120 or 180 days after baseline, and death with COVID-19 infection either 120 or 180 days after baseline.

APPLICATION TO COMPARATIVE EFFECTIVENESS OF VACCINES

- Han, Larry, et al. (2023+) "Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects." JASA minor revision. *arXiv:2112.09313*.

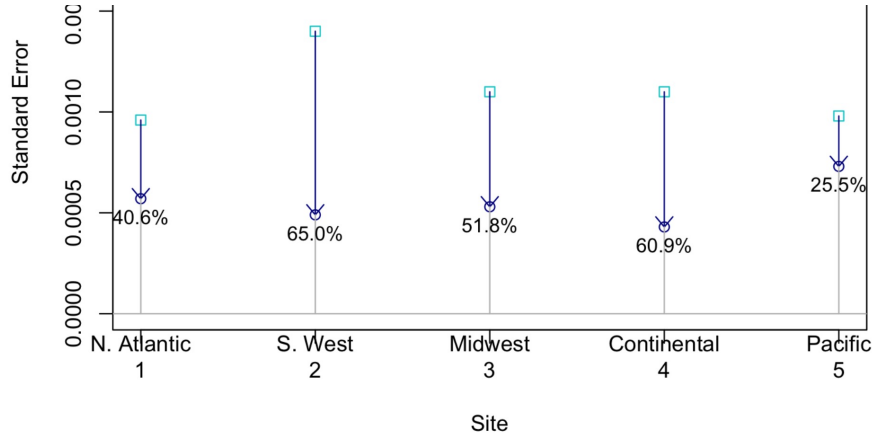
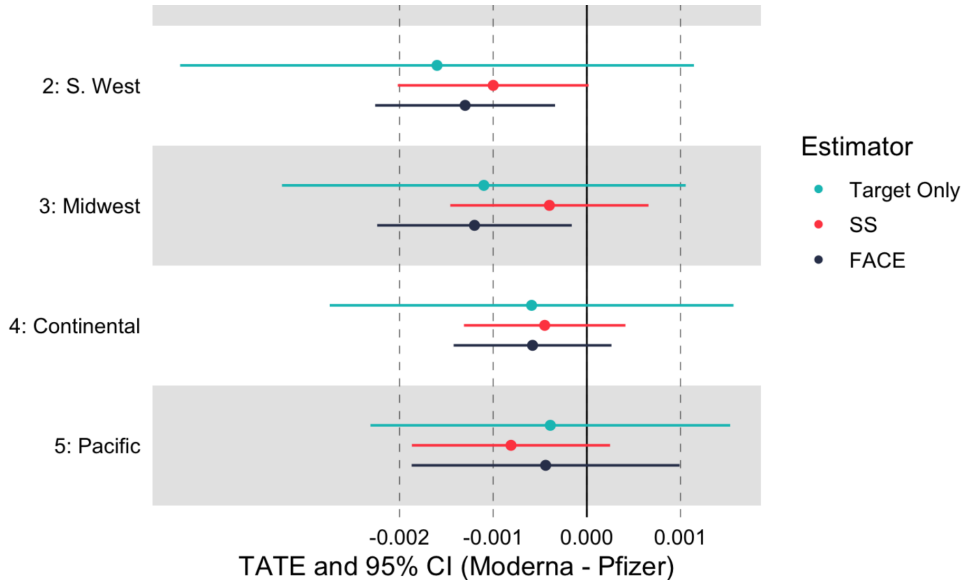


Figure: Reductions in SE are substantial using FACE



WHAT IS MORE PRACTICAL IN THE REAL-WORLD SETTINGS?

- When exchangeability does not hold in any of the source sites, the previous methods may be worse than the target only estimator.
- It does not distinguish two sites, one with a small shift in $Y|A, X$, and one with a large shift

NEW PROPOSAL

Our more recent proposal:

- A semiparametric model for the shifts in Y | A, X (selection bias model)

$$p_k^0(y|x,a) = w_k^*(y,x,a;\beta_k,Q^0)q^0(y|x,a),$$

$$w_k^*(y,x,a;\beta_k,Q^0) = \frac{w_k\ (y,x,a;\beta_k)}{E_{Q^0}(w_k\ (y,x,a;\beta_k)|X=x,A=a)}$$

- For example, the exponential tilting model $w_k\ (y,x,a;\beta_k) = \exp(\phi(z)^T\beta)$
- Efficient data integration through EIF

PROTECT AGAINST MODEL MISSPECIFICATION

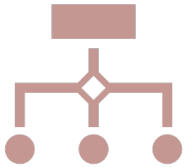
Use Domain Knowledge

Understand the meaning of the distributional shift parameters, e.g., shifts in log odds or risks in stratified groups.

Perform Goodness-of-Fit Tests and Sensitivity Analyses

Regularly test the fit of the model to ensure accuracy.

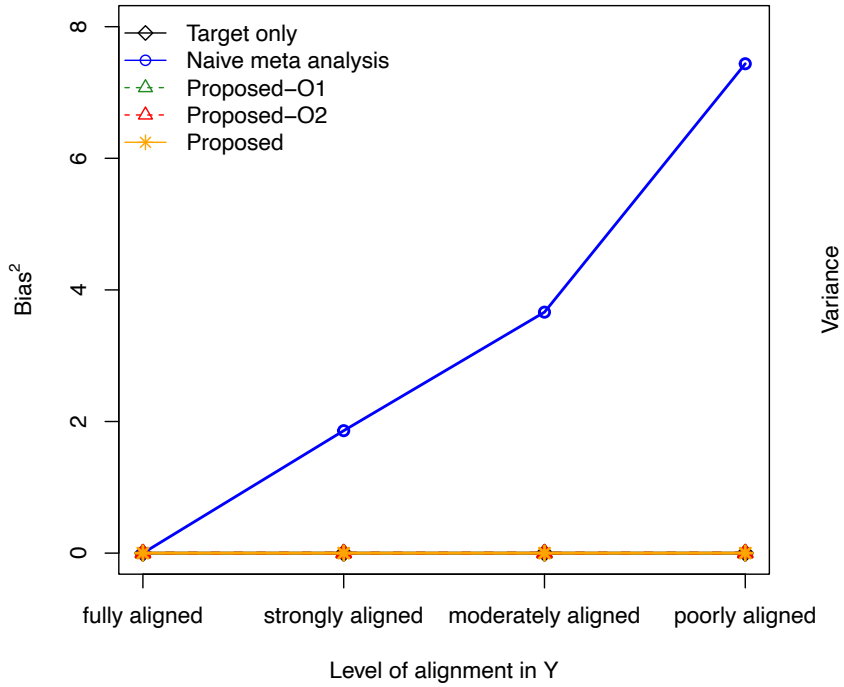
Conduct sensitivity analyses to assess the robustness of the model.

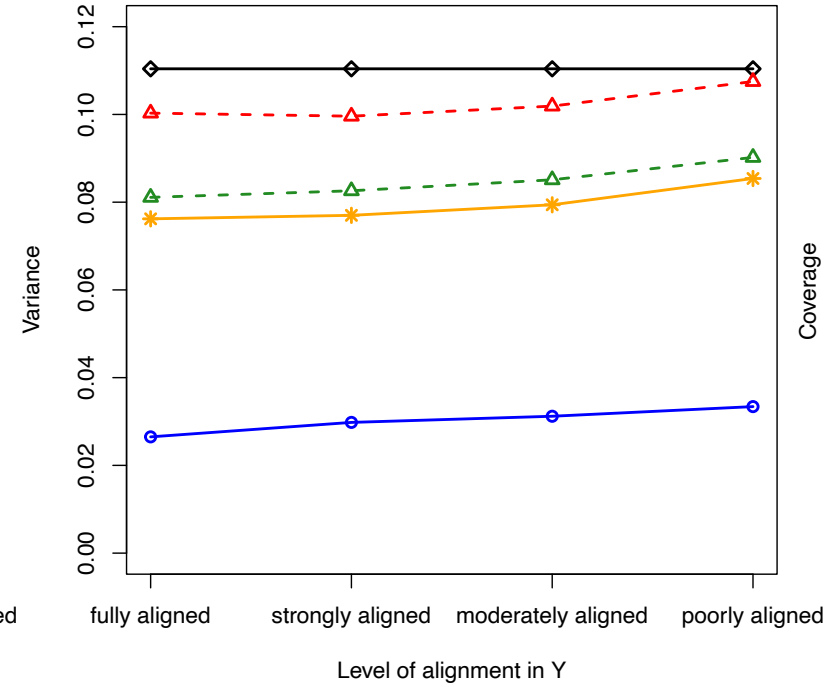


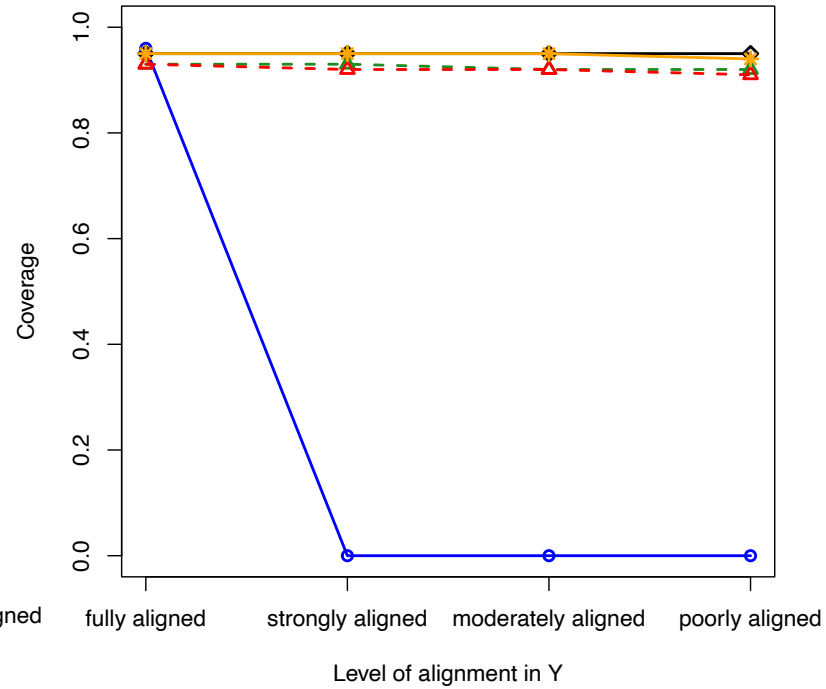
Over-parametrize the Model

Include more basis functions in the density ratio model---loss of efficiency

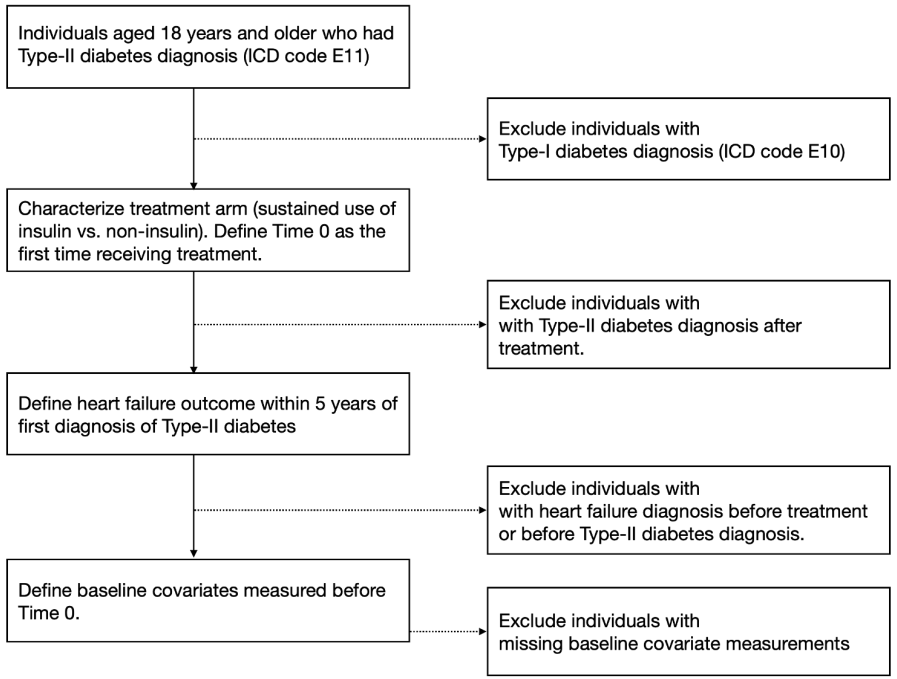
SIMULATION

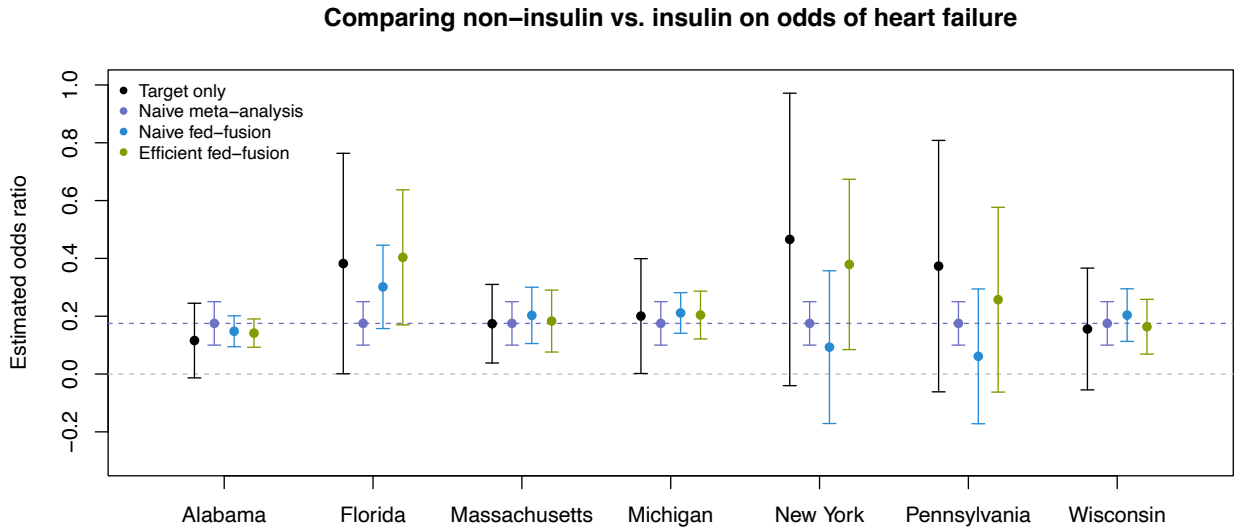

Target only
Naive meta analysis
Proposed–O1
Proposed–O2
Proposed
Bias^2
0
2
4
6
8
fully aligned
strongly aligned
moderately aligned
poorly aligned
Level of alignment in Y


Variance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
fully aligned
strongly aligned
moderately aligned
poorly aligned
Level of alignment in Y


Coverage
0.0
0.2
0.4
0.6
0.8
1.0
fully aligned
strongly aligned
moderately aligned
poorly aligned
Level of alignment in Y

REAL DATA—AOU RESEARCH HUB





Preprint coming shortly

DISCUSSION

- can we *effectively* and *efficiently* learn from heterogeneous sources?
  - communication-efficient distributed optimization, heterogeneity-aware optimization, personalization and multi-task learning, connections to pretraining

- can we provide *trustworthy* learning schemes?
  - techniques efficient/accurate/usable private and secure learning in large-scale distributed settings, threat modeling and auditing, robustness to attacks

- can we develop systems/tools that *support* collaboration?
  - game-theoretic perspectives on collaborative learning, data markets, data valuation and risk assessment, incorporating social dynamics

- Fairness: can we equalize performance across diverse networks?

REFERENCE

- Lecture Notes, Federated and Collaborative Learning, Fall 2023, Virginia Smith. https://www.cs.cmu.edu/~smithv/10719/ Access date: 07-05-2024.
- Li, Li, et al. "A review of applications in federated learning." *Computers & Industrial Engineering* 149 (2020): 106854.
- Mammen, Priyanka Mary. "Federated learning: Opportunities and challenges." *arXiv preprint arXiv:2101.05428* (2021).
- Zhang, Chen, et al. "A survey on federated learning." *Knowledge-Based Systems* 216 (2021): 106775.
- Chen, Shuxiao, Bo Zhang, and Ting Ye. "Minimax rates and adaptivity in combining experimental and observational data." *arXiv preprint arXiv:2109.10522* (2021).
- BAREINBOIM, E. & PEARL, J. (2014). Transportability from multiple environments with limited experiments: Completeness results. Advances in neural information processing systems 27, 280–288.
- BRANTNER, C. L., CHANG, T.-H., NGUYEN, T. Q., HONG, H., DI STEFANO, L. & STUART, E. A. (2023). Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. arXiv preprint arXiv:2302.13428
- DAHABREH, I. J. & HERN´AN, M. A. (2019). Extending inferences from a randomized trial to a target population. Eur. J. Epidemiol. 34, 719–722
- LI, S. & LUEDTKE, A. (2023). Efficient estimation under data fusion. Biometrika 110, 1041–1054
- RUDOLPH, K. E. & VAN DER LAAN, M. J. (2017). Robust estimation of encouragement-design intervention effects transported across sites. J. R. Stat. Soc. 79, 1509
- HAN, L., HOU, J., CHO, K., DUAN, R. & CAI, T. (2021). Federated adaptive causal estimation (face) of target treatment effects. arXiv preprint arXiv:2112.09313 .
- YANG, S., GAO, C., ZENG, D. & WANG, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. Journal of the Royal Statistical Society Series B:Statistical Methodology 85, 575–596.