Georgia Tech

CREATING THE NEXT

Privacy-Preserving Federated Data Sharing

Ferdinando Fioretto
Georgia Tech

Pascal Van Hentenryck
Georgia Tech

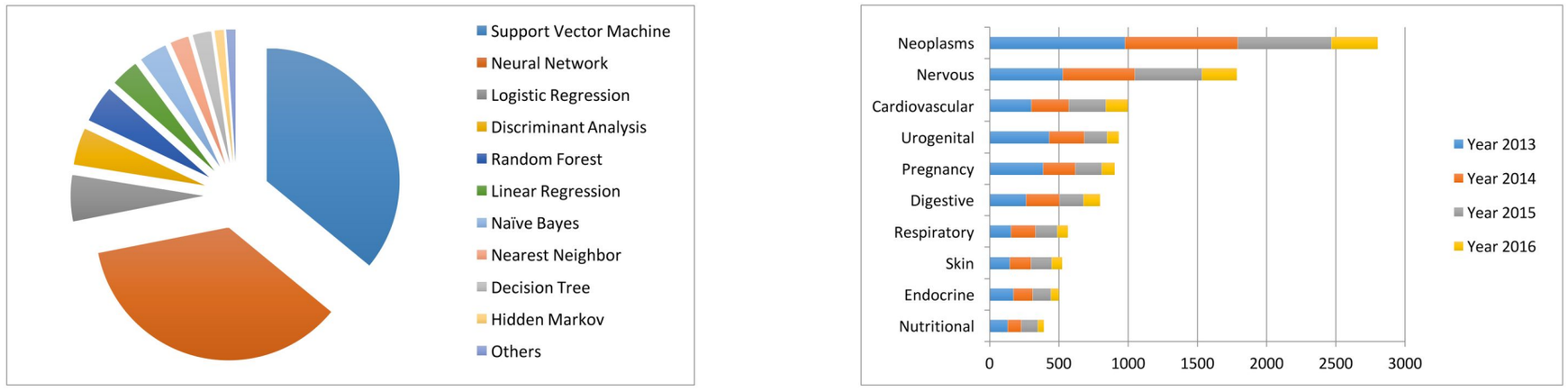
May 15, 2019

We live in a data-driven world

Georgia Tech

Motivations

- The availability of personal data is responsible for rapid advancements of AI and ML applications.
- Many data-driven ML models to support e.g., screening, diagnosis, and treatment assignment in healthcare
- Transition from proprietary data acquisition and processing to distributed data ecosystems.
- Personal data is very sensitive in many domains (e.g., healthcare)



Top ML models adopted (left) and leading disease types (right) considered in the ML & Healthcare literature [1]

The Problem

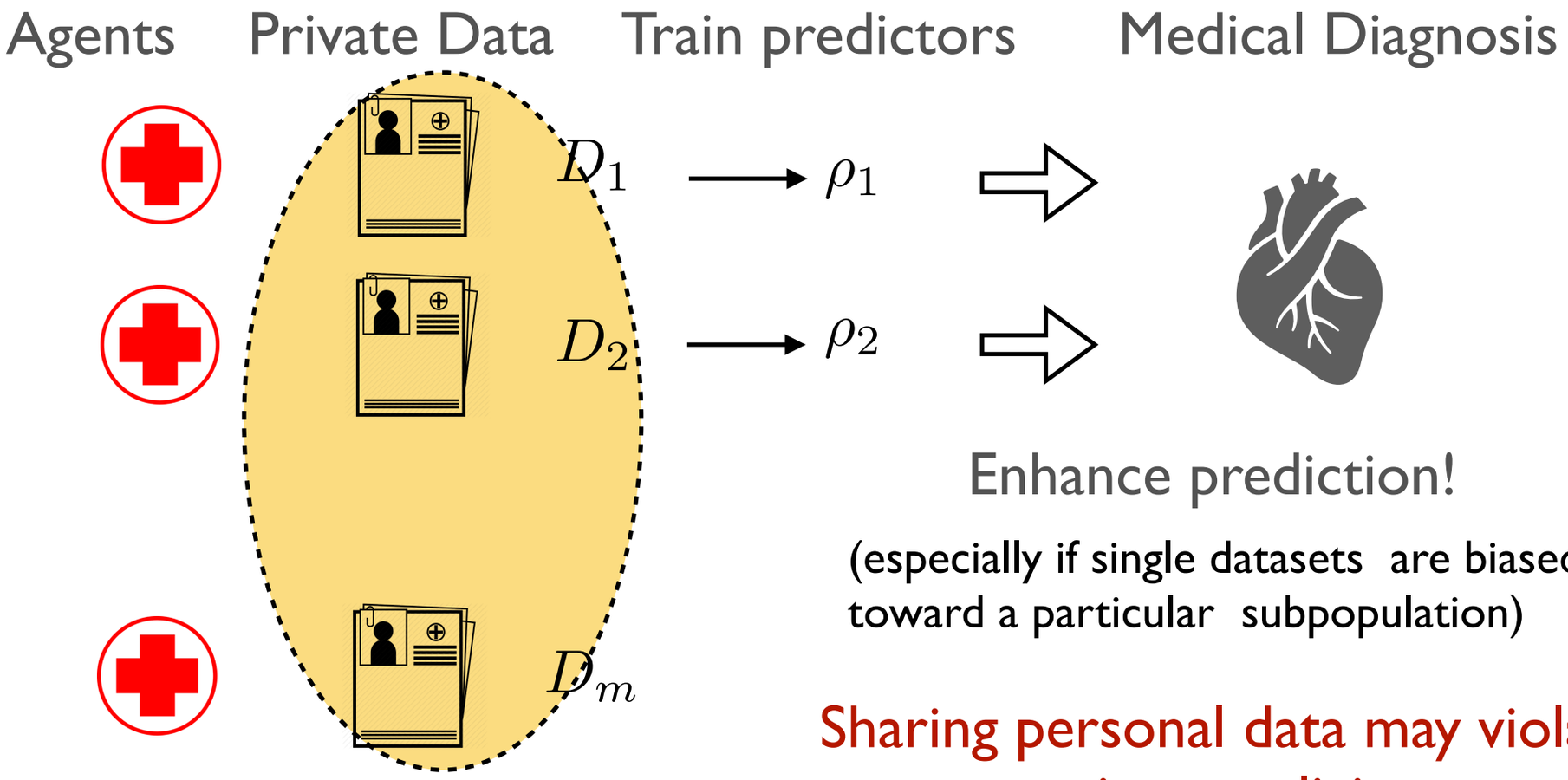
Agents Private Data Train predictors Medical Diagnosis

$D_1 \longrightarrow \rho_1$

$D_2 \longrightarrow \rho_2$

$D_m \longrightarrow \rho_m$

The Problem



Enhance prediction!

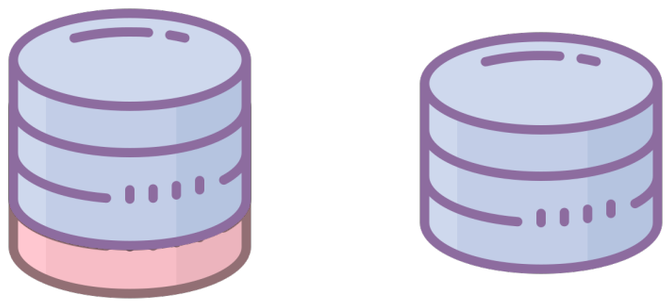
(especially if single datasets are biased toward a particular subpopulation)

Sharing personal data may violate privacy policies

How to build a Federated Data Sharing protocol that preserves the individuals data Privacy and it is suitable for prediction tasks

Differential Privacy

For every pair of inputs that differs in one row



D_1 D_2

For every output O

A randomized algorithm $\mathcal{A}$ is **ε**-differentially private if:

$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$

Intuition: adversary should not be able to use output O to distinguish between any D_1 and D_2

Differential Privacy
Sensitivity Method & Privacy Properties

- Consider a query Q on the dataset D. The *sensitivity method* injects noise to the output Q(D) that depends on the sensitivity of the query:

$$\Delta_Q = \max_{D_1 \sim D_2} \|Q(D_1) - Q(D_2)\|_1$$

- Laplace Mechanism [2], Exponential Mechanism [3], …
- **Differential Privacy Important Properties**:
  - Composability: Graceful privacy loss degradation when composing multiple DP algorithms.
  - Immunity to post-processing: Privacy guarantees are preserved by arbitrary, data-independent, post-processing steps.

Empirical Risk Minimization

- Labeled data D= $\{\mathbf{x}_i, y_i\}_{i\in[n]}$

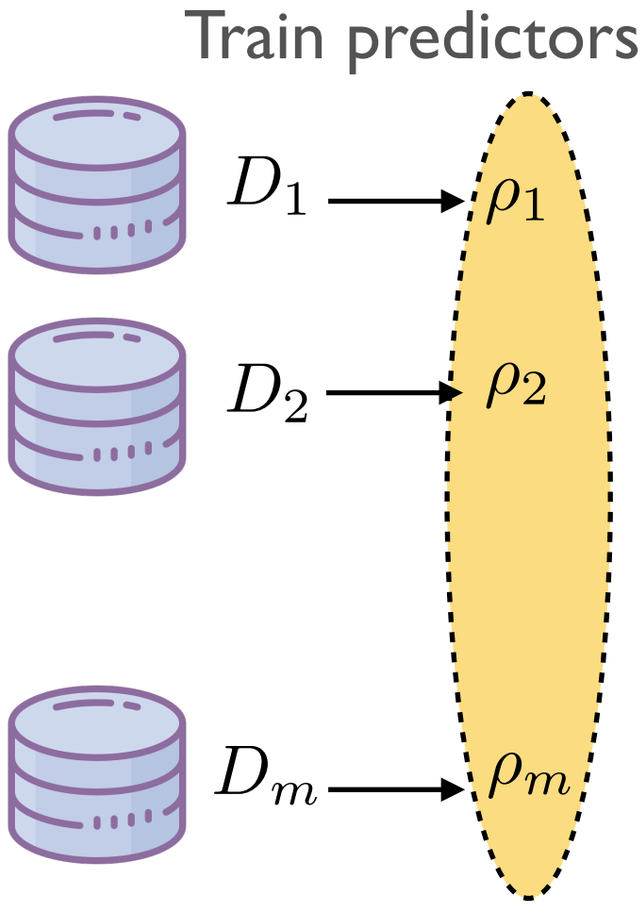
$$\overbrace{(\text{age}, \text{gender}, \ldots,}^{\mathbf{x}_i\in\mathbb{R}^d} \overbrace{\text{diabetes})}^{y_i\in[-1,1]\subseteq\mathbb{R}}$$

- ERM aims at choosing a predictor $\rho_{\mathbf{w}}$ that minimizes the (regularized) empirical loss:

$$J(\rho_{\mathbf{w}}, D) = \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(\rho_{\mathbf{w}}(\mathbf{x}_i), y_i\right)\right) + \lambda c(\mathbf{w})$$

- Different loss functions allow us to express different predictors. E.g.:
  - Logistic Regression
  - Linear Regression
  - SVM

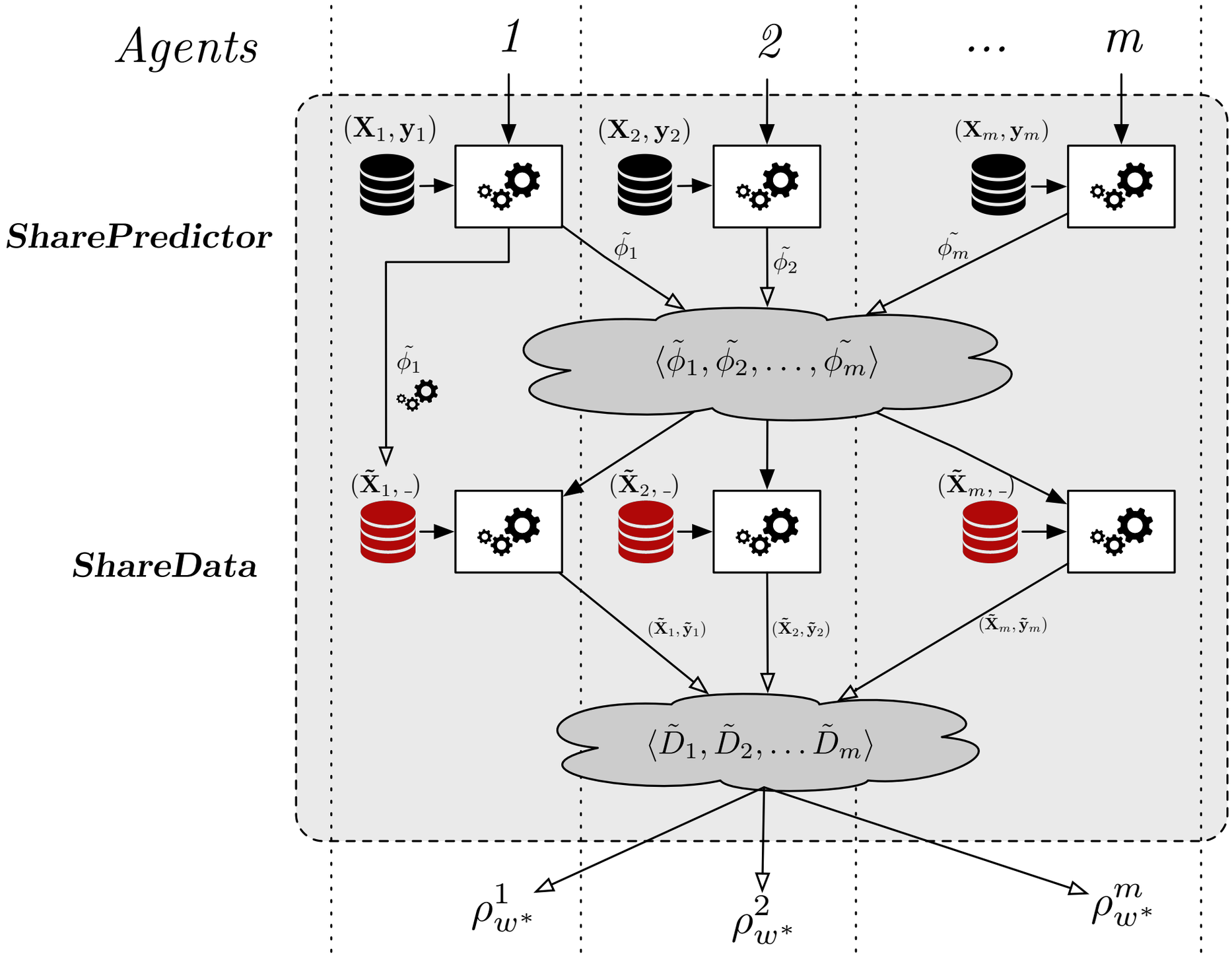
Multi-agent ERM



Sharing Data/Predictors must be privacy-preserving!

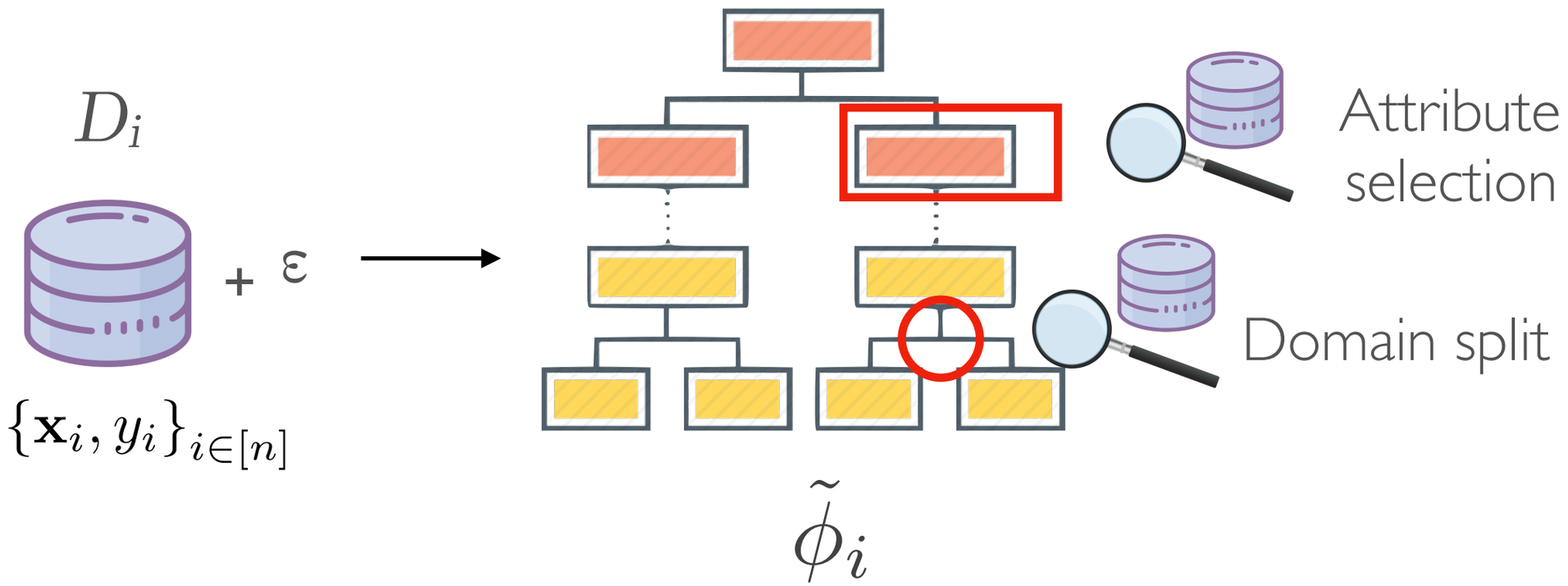
- Aggregation schemes:
  - [Data] Majority voting: Outputs the class that has been predicted more often
  - [Predictors] Committee machine: Averages the results of each predictor.
- Dataset contain private information

The PFDS Framework



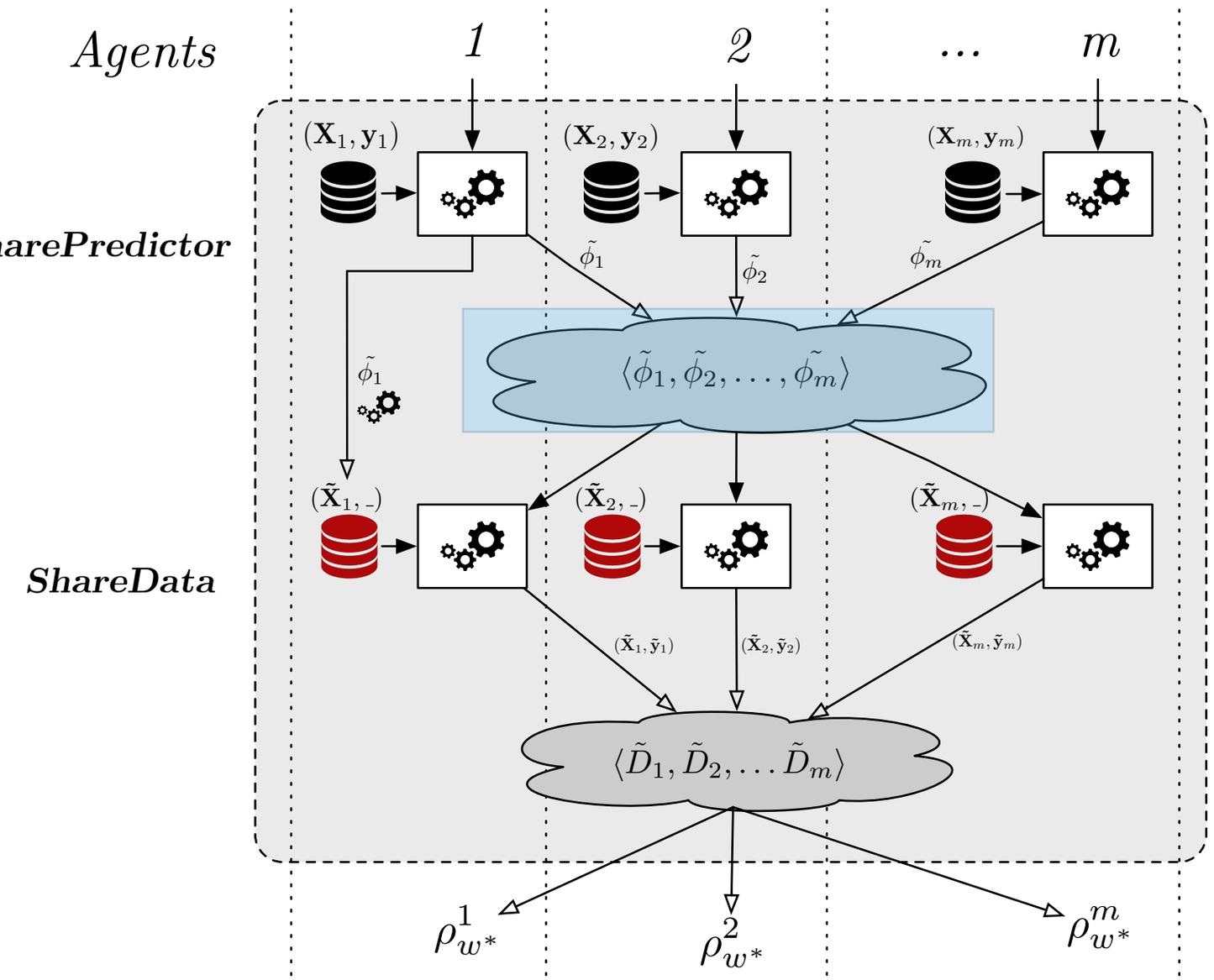
Predictor Sharing Phase

- Each agent locally builds a **ε**-differentially private decision tree $\tilde{\phi}_i$ from its own dataset Di
- Attribute selection and domain split "look at the data": require private queries
- Uses a noisy splitting function that relies on sub-sampling the domain of the select value and the Exponential Mechanism



Georgia Tech

Predictor Sharing Phase

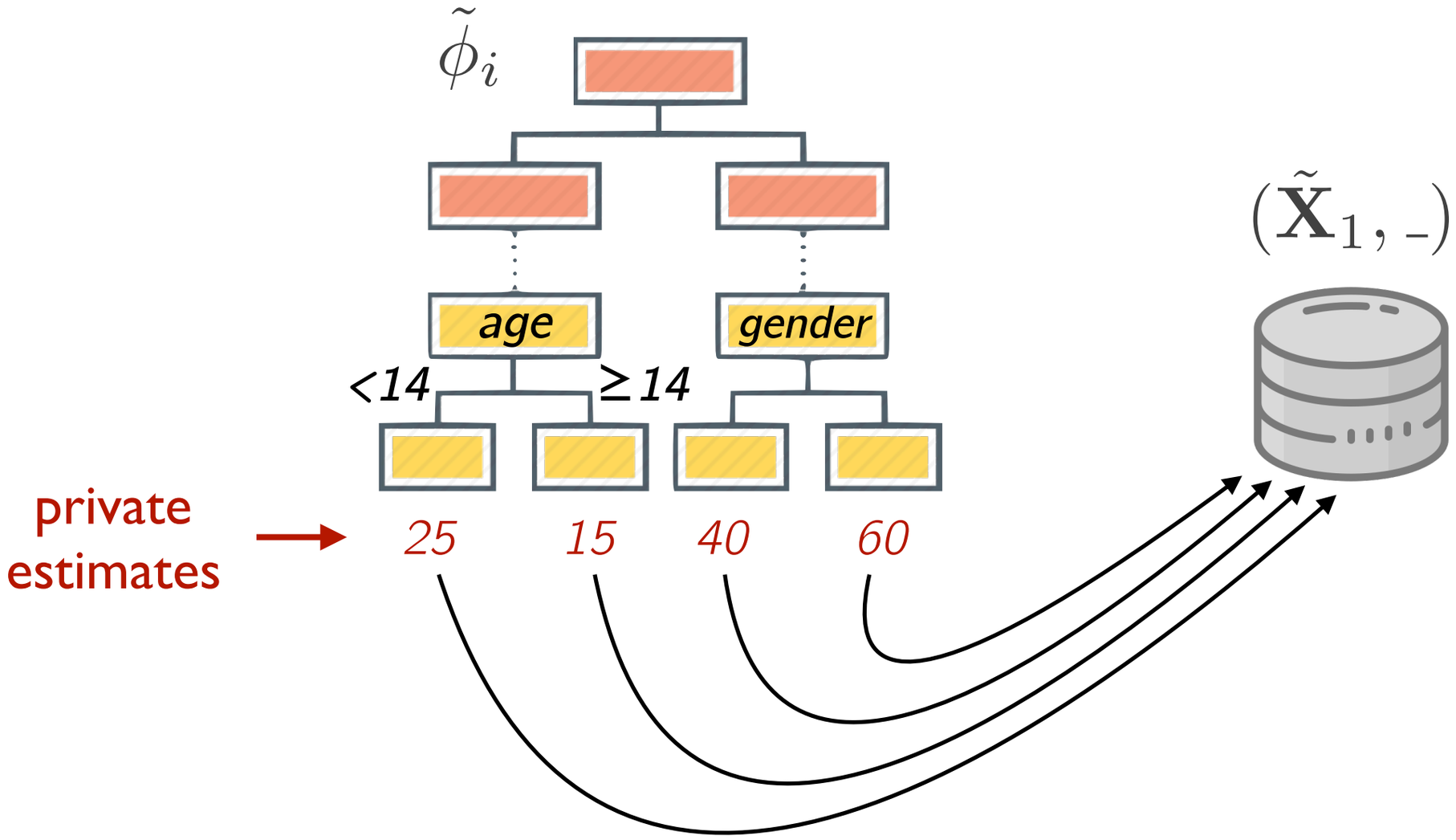


- The agents share their privacy-preserving predictors

Georgia Tech

Data Sharing Phase

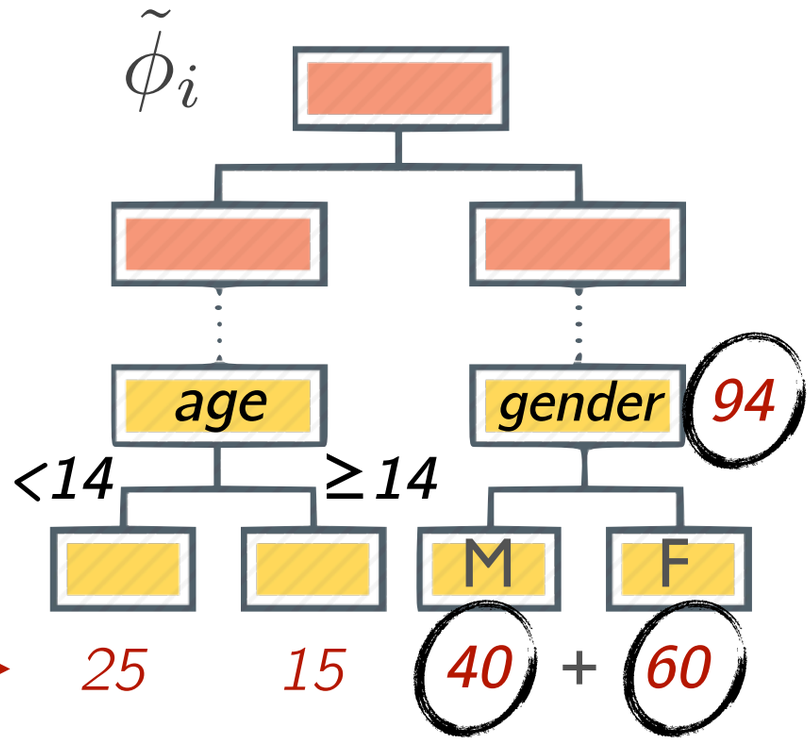
- The private predictor $\tilde{\phi}_i$ is used by the agent i to create a an unlabeled version of its dataset.
- The new dataset is generated using only private estimates (counting the number of individuals satisfying a given property in the X_1)



Georgia Tech

Data Sharing Phase

- Because noise is added independently at each node of the tree, hierarchical properties may be inconsistent



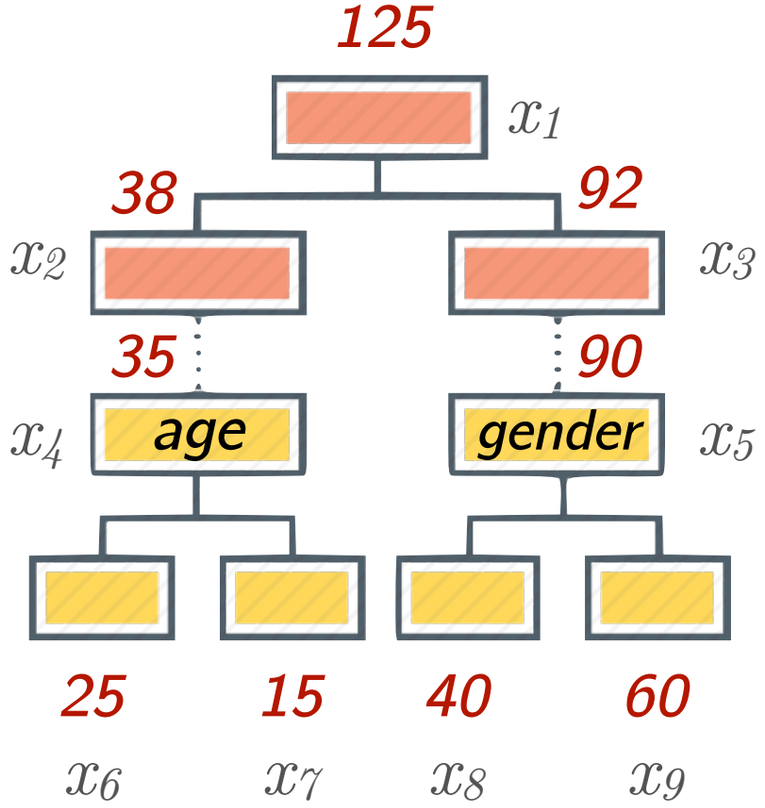
private estimates →

$(\tilde{\mathbf{X}}_1, \_)$

Georgia Tech

Data Sharing Phase
Optimization

- To restore consistency, we use convex optimization.
- It redistributed the noise to ensure count consistency.
- THM: The Optimization-based process is ε-differentially private



$$\min \sum_{x_i} (x_i - \tilde{x}_i)^2$$

$$s.t.: \quad \sum_{\tau_i \in ch(\tau_j)} x_i = x_j \quad \forall j \in [9]$$

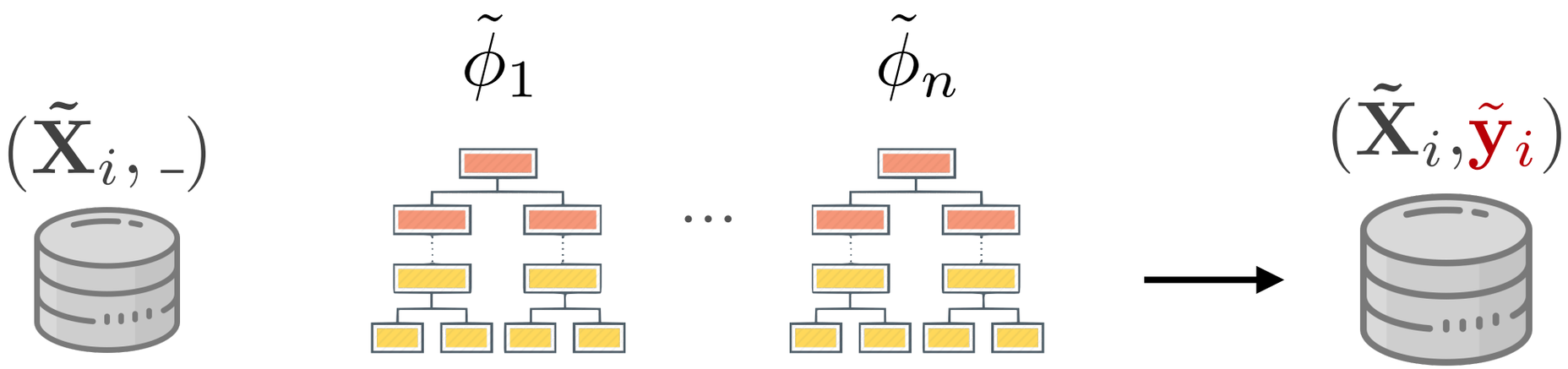
$$x_i \geq 0 \quad \forall i \in [9]$$

Data Sharing Phase
Labeling Private Data

- Each agent *i* uses all shared predictors to find the *labels* for its private dataset.
- Each dataset item $\tilde{\mathbf{x}}_i$ is labelled with

$$\tilde{\mathbf{y}}_i = \text{majority\_vote}(\tilde{\phi}_1(\tilde{\mathbf{x}}_i) \ldots \tilde{\phi}_n(\tilde{\mathbf{x}}_i))$$

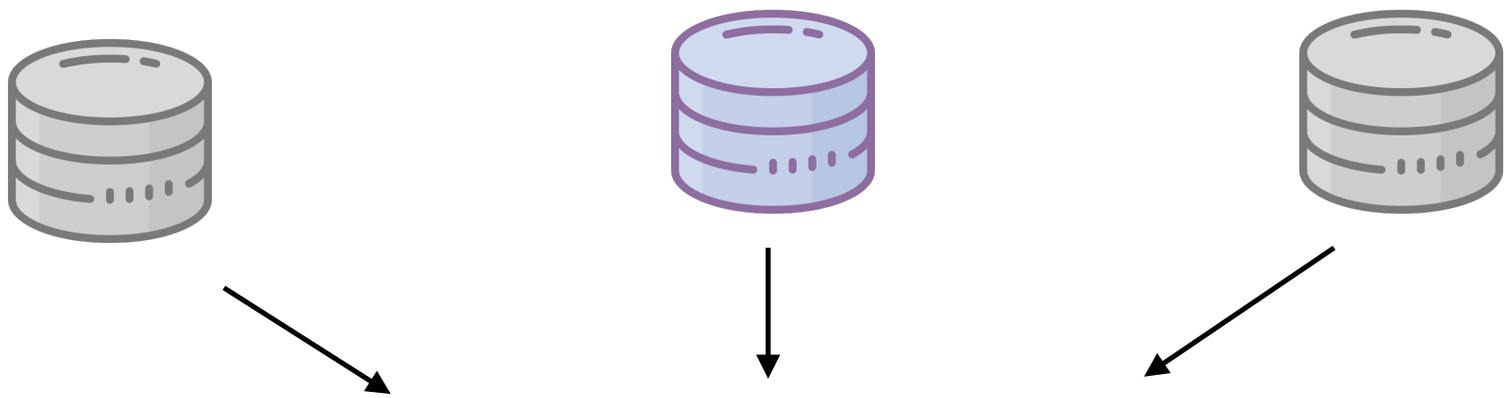
- The agent share its privacy-preserving (labeled) data with all other agents



Training on Federated Data

- Each agent uses all shared privacy-preserving dataset and its own dataset to train its predictor ρ^i_{w*}

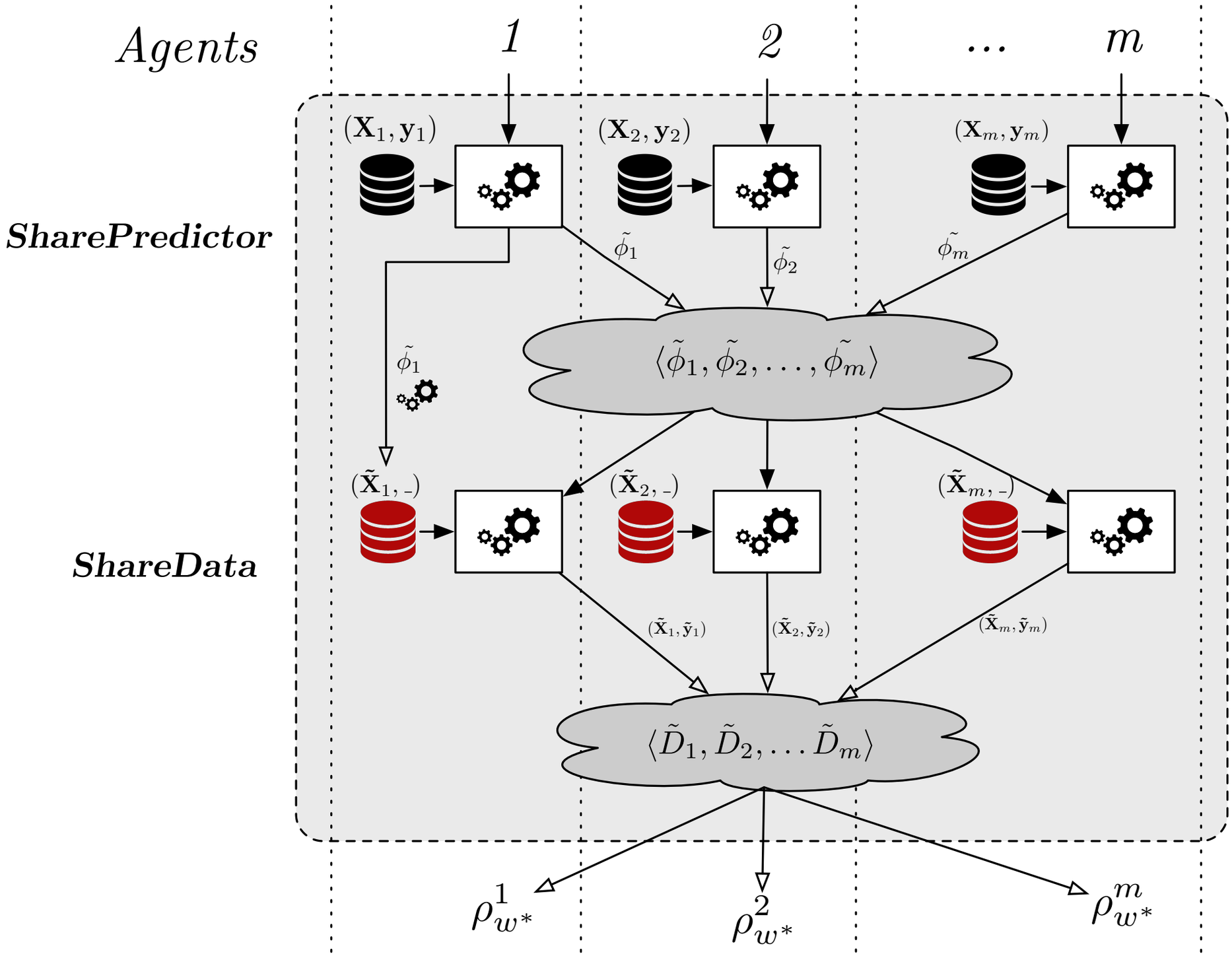
$(\tilde{\mathbf{X}}_1, \tilde{\mathbf{y}}_1)$ ... $(\mathbf{X}_i, \mathbf{y}_i)$ ... $(\tilde{\mathbf{X}}_n, \tilde{\mathbf{y}}_n)$



$$J(\rho_{\mathbf{w}}, D) = \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(\rho_{\mathbf{w}}(\mathbf{x}_i), y_i\right)\right) + \lambda c(\mathbf{w})$$

ERM training

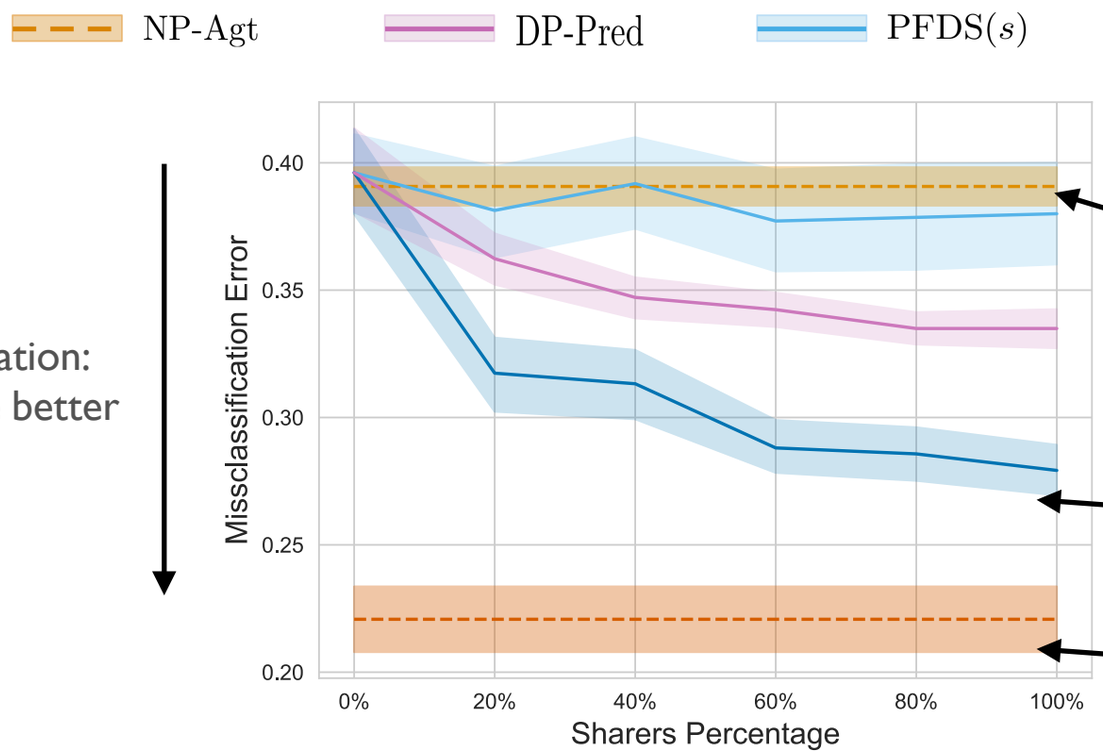
The PFDS Framework



Evaluation: Methods

- **PFDS**: method developed in this work
- **PFDS(s)**: PFDS w/out Predictor Sharing (to label data)
- **DP-Pred**: Objective perturbation[4] (log. reg. and SVM) or functional mechanism[29] (lin. reg.) w/ majority voting
- **NP-Agt**: Non-private predictor executed on the agent data
- **NP-Pred**: Non-private predictor executed on all data

Evaluation (summary)



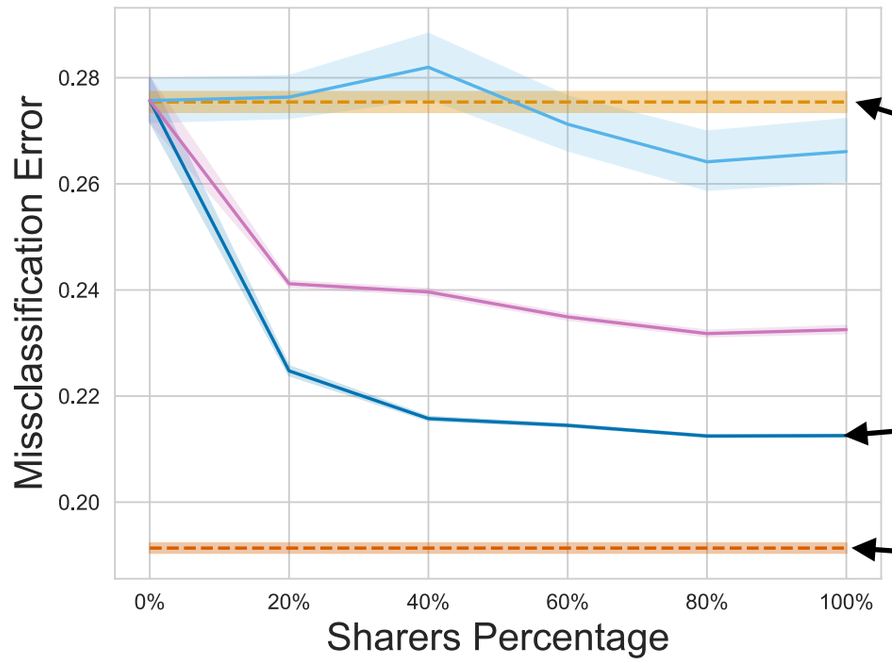
Missclassification:
The lower the better

Diabetes: *Does patient have diabetes?*

Each agent learns on its own data
(baseline)

Proposed solution

Learn with all non-private data
(best possible)



Census: *Does user gain >50k yearly?*

Each agent learns on its own data
(baseline)

Proposed solution

Learn with all non-private data
(best possible)

Georgia Tech

Conclusions

- Companies are increasingly leveraging distributed data.
- Typically, more data implies better models, however privacy regulation prevent data owners to share their data
- We investigated whether sharing private version of the agents datasets may improve the agents predictions
- We proposed a Privacy-preserving Federated Data Sharing protocol to allow agents to share their data privately.
- On Multi-agent ERM tasks, our results show that it is possible to both
    1. Guarantee privacy and
    2. Learn more accurate models than those learned by single-agent systems on non-private data.

Conclusions

- Companies are increasingly leveraging distributed data.
- Typically, more data implies better models, however privacy regulation prevent data owners to share their data
- We investigated whether sharing private version of the agents datasets may improve the agents predictions
- We proposed a Privacy-preserving Federated Data Sharing protocol to allow agents to share their data privately.
- On Multi-agent ERM tasks, our results show that it is possible to both
  1. Guarantee privacy and
  2. Learn more accurate models than those learned by single-agent systems on non-private data.

Ask us any additional question at the poster session

Thank you!

fioretto@gatech.edu
pvh@isye.gatech.edu

References

[1] Jiang, Fei, et al. "Artificial intelligence in healthcare: past, present and future." *Stroke and vascular neurology* 2.4 (2017): 230-243.