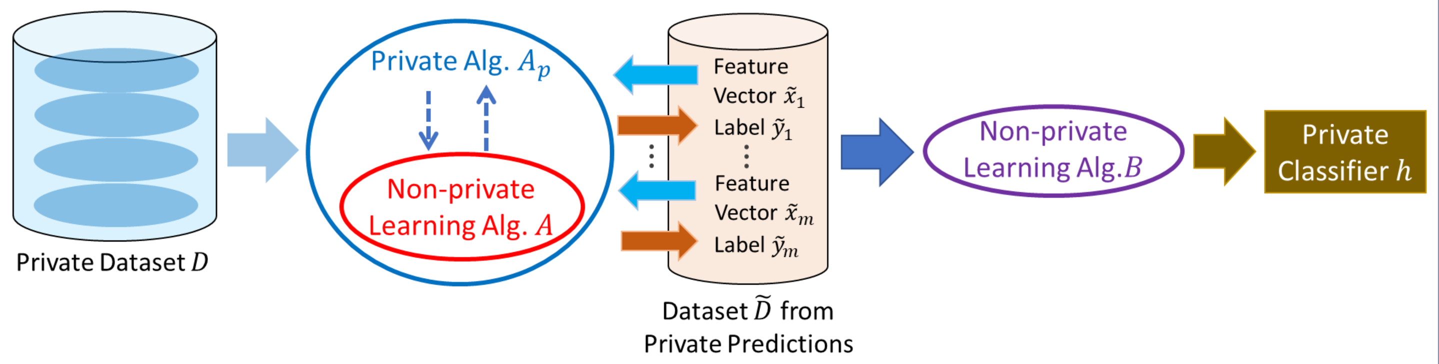
Model-Agnostic Private Learning

Raef Bassily[1], **Om Thakkar[2]**, Abhradeep Thakurta[3]

[1]The Ohio State University, [2]**Boston University**, [3]University of California-Santa Cruz

INTRODUCTION

This work provides a framework, using black-box transformations of non-private learners, for obtaining: 1) Privacy-preserving predictions, and 2) A private classifier from private predictions



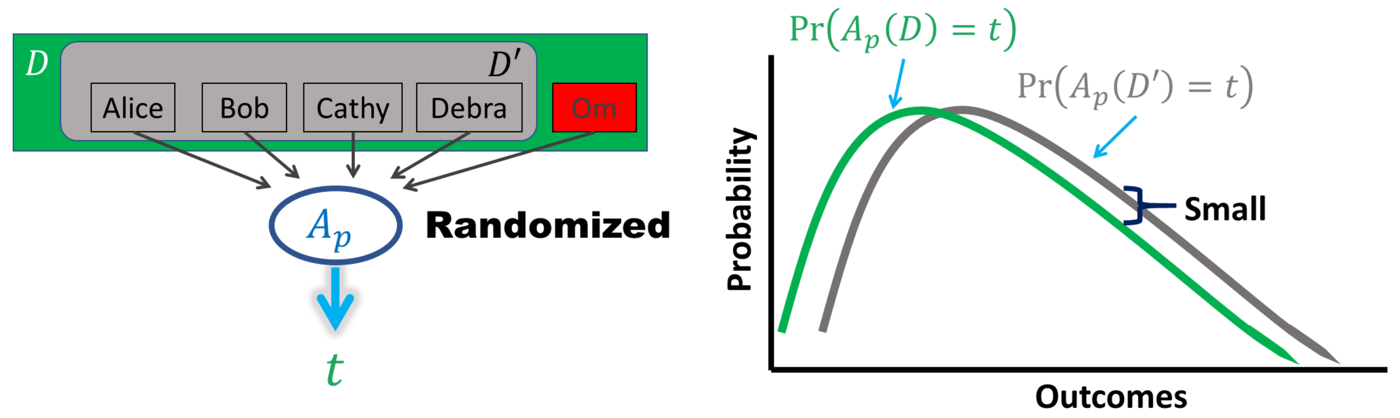
Requirement: Private alg. A_p should not reveal *too much* info. about the sensitive dataset D.

PRELIMINARIES

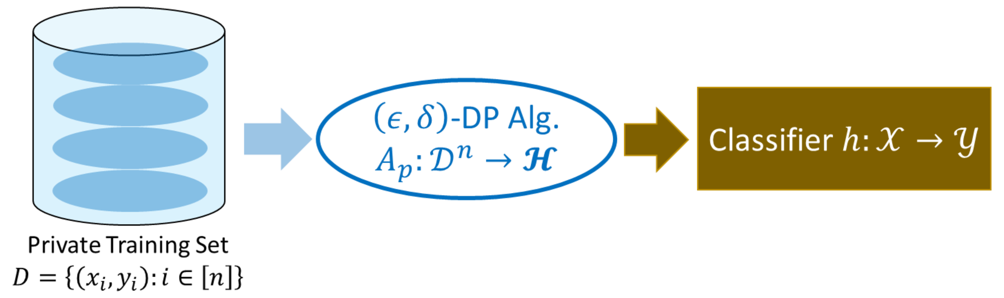
(α, β, n)-PAC learner: Alg. A is (α, β, n)-(agnostic) PAC learner for hypothesis class $\mathcal{H}$ if, given input dataset $D \sim \mathcal{D}^n$, w.p. $1-\beta$ outputs a hypothesis $h_D \in \mathcal{H}$ with $err(h_D; \mathcal{D}) \leq \gamma + \alpha$, where $err(h; \mathcal{D})$ denotes the misclassification rate of h on $\mathcal{D}$, and $\gamma \coloneqq \min_{h \in \mathcal{H}} err(h; \mathcal{D})$.

We call it the realizable case if $\gamma = 0$, else we call it the agnostic case.

(ϵ, δ)-Differentially Privacy (DP) [DMNS'06]: A randomized algorithm $A_p: \mathcal{D}^n \to \mathcal{T}$ is (ϵ, δ)-DP, if for all neighboring datasets $D, D' \in \mathcal{D}^n$, i.e., $|D \triangle D'| = 1$, and for all sets of outcomes $T \subseteq \mathcal{T}$, we have $\Pr(A_p(D) \in T) \leq e^{\epsilon} \Pr(A_p(D') \in T) + \delta$



DP LEARNING: STANDARD APPROACH



Main Issues with the Standard Approach for DP Learning:

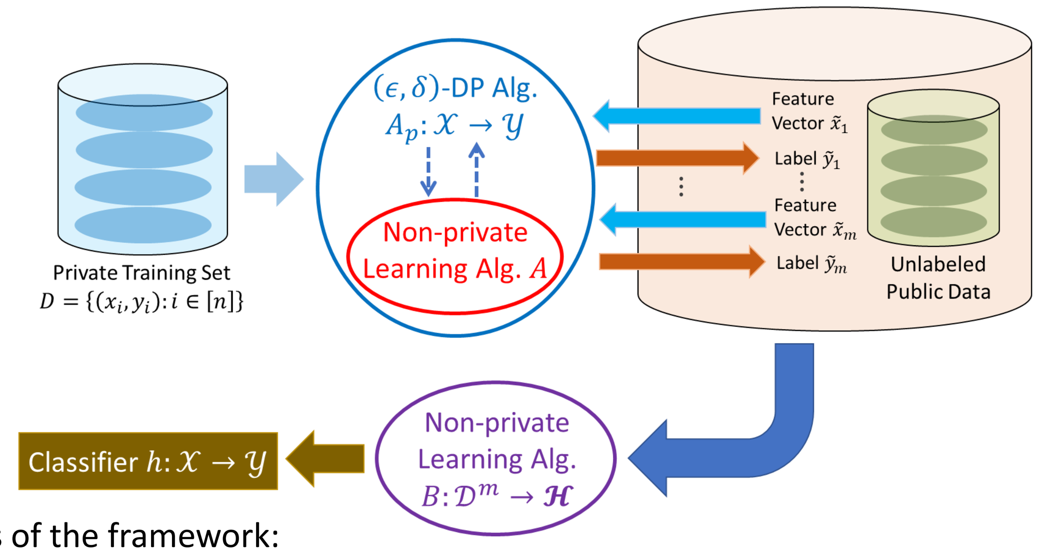
- Requires white-box modification of *non-private* learners
- Often requires knowledge about the structure of $\mathcal{H}$
- Often yields error with dependence on size of $\mathcal{H}$, even for simple model classes, e.g., learning thresholds [BNSV'15]

DP LEARNING: VIA BLACK-BOX TRANSFORMATIONS

We design a framework that uses black-box transformations of non-private learners to first privately label some **unlabeled public data**, which is then used for obtaining a classifier.

Advantages of black-box transformations:

- No change to the existing algorithmic infrastructure
- Inherit computational advantages of existing non-private algorithms



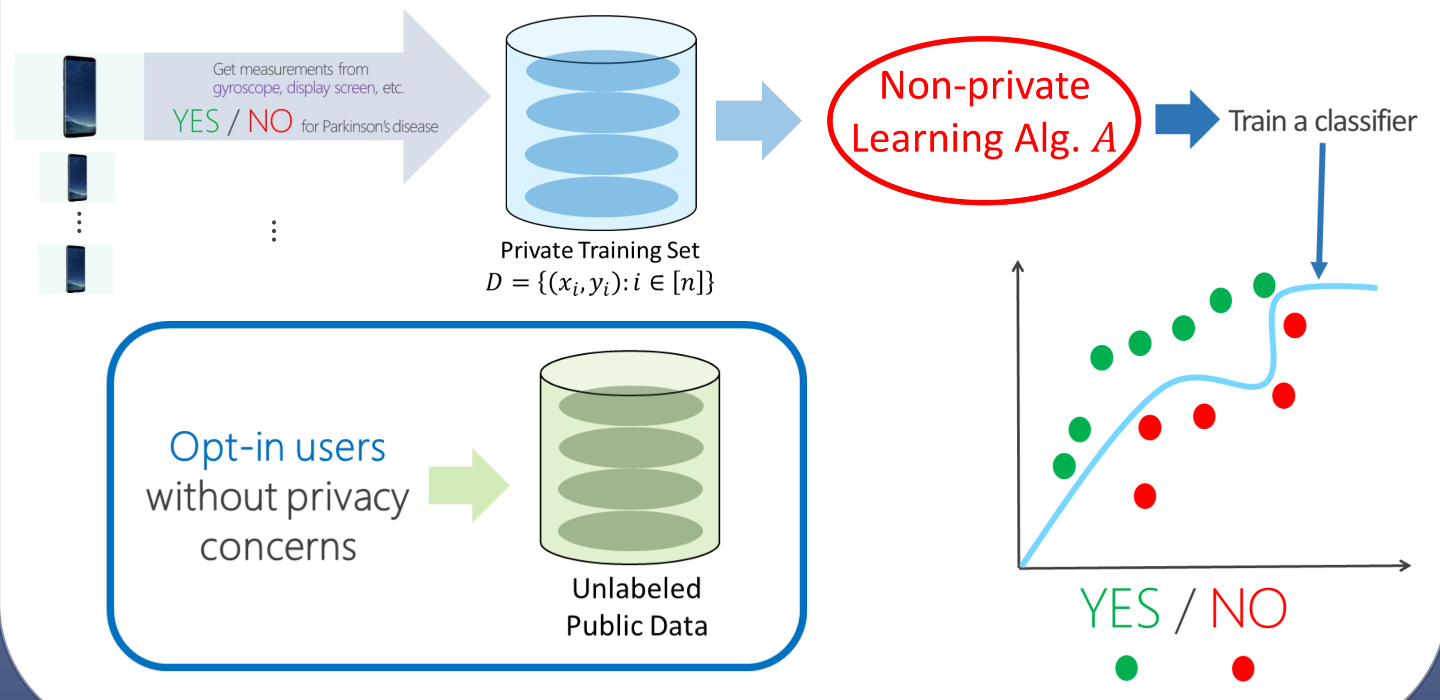
Advantages of the framework:

- Makes black-box use of non-private learners
- Conservatively uses privacy budget to label *lots* of public features [PAE+'17, This work]
- Allows knowledge transfer using public features and output labels to train a private classifier
- Results in transferrable utility guarantees [This work]:
  Non-private learner accuracy → Private learner accuracy
  - Note: [PAE+'17] used a black-box transformation, but didn't provide formal accuracy guarantees. However, the experiments in [PAE+'17] indirectly corroborate our intuition and theory on *stability*.

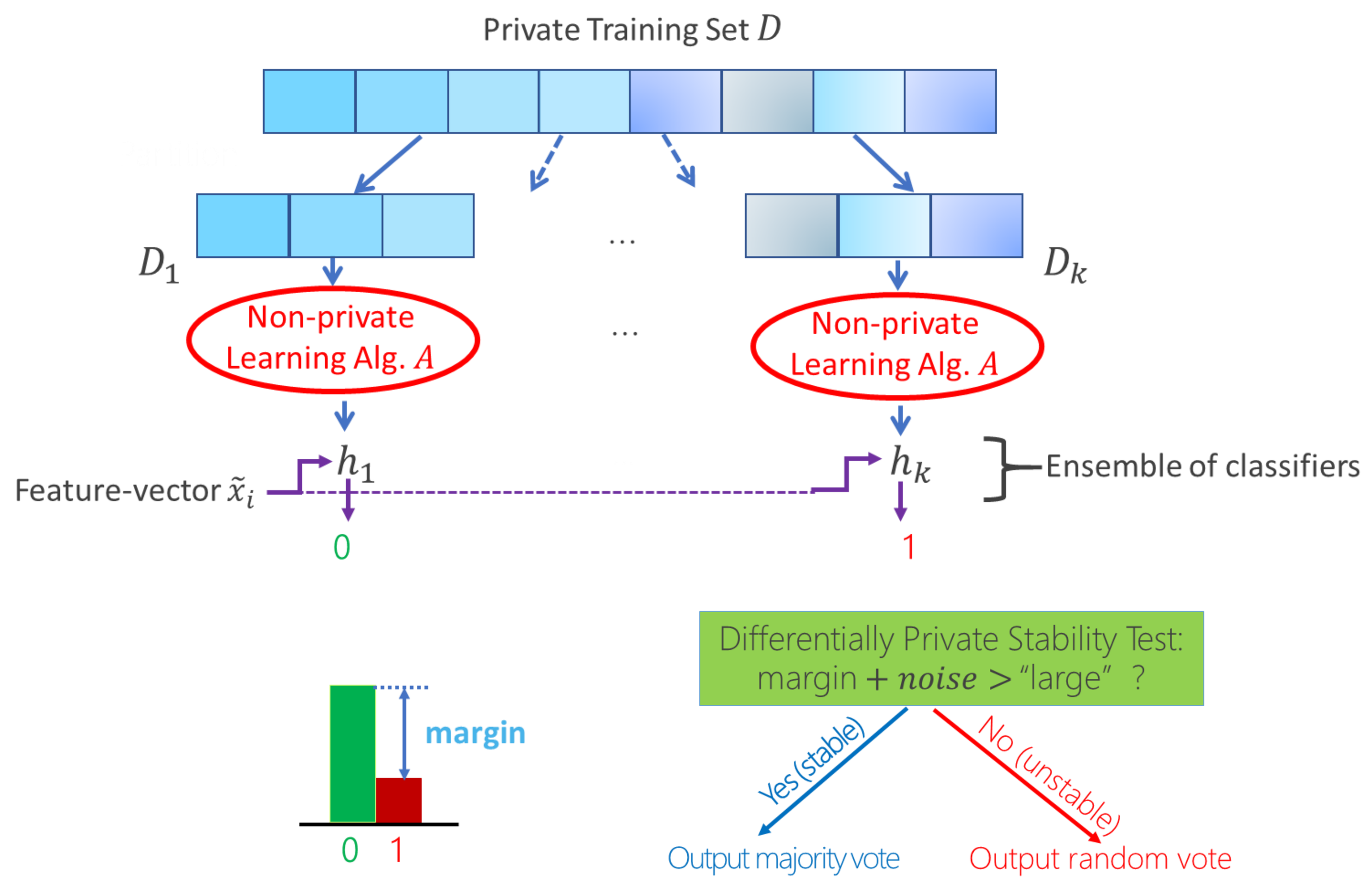
BUILDING BLOCKS

Goal: A model for binary classification, e.g., predicting Parkinson's disease

Inputs: 1) Private Training Set D (n samples),
2) Non-private Learning Algorithm A, and
3) Unlabeled Public Data (m samples)



1. PRIVATE ALGORITHM FOR CLASSIFICATION QUERIES



For each query, if margin is *sufficiently* large, w.h.p. the output is the majority vote.

Privacy: To ensure (ϵ, δ)-DP, private alg. A_p halts after $\approx k^2 \epsilon^2 / \log(1/\delta)$ unstable predictions. The privacy budget is only consumed by **unstable** predictions.

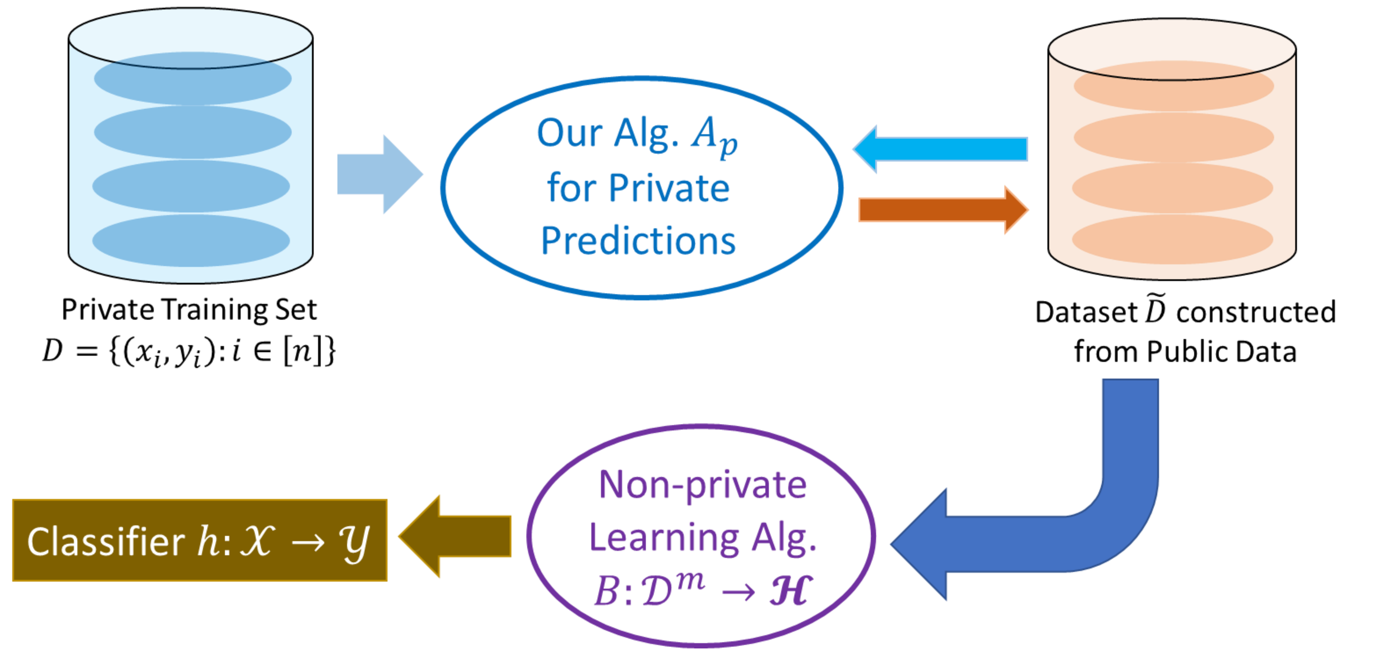
Generic Transformation of Misclassification Rate:
Misclassification rate of $A \leq \alpha \Rightarrow$ # misclassifications by $A_p \leq 3m\alpha$
(for a specific setting of k)

Proof Idea: If each classifier $h_1, \ldots, h_k$ has misclassification rate $\leq \alpha$, then by a counting argument, at least $2k/3$ classifiers will agree on the correct label, except for $\approx 3m\alpha$ queries.

- No. of ✗ in each row $\approx m\alpha$
- Total no. of ✗ $\approx km\alpha$
- No. of columns with $>\approx \frac{k}{3}$ ✗ is $< 3m\alpha$

⇒ Our algorithm can answer $1/\alpha$ more queries than prior approaches based on the "composition theorem" of differential privacy.

2. FROM PRIVATE PREDICTIONS TO A PRIVATE CLASSIFIER



This construction:

- is efficient as long as the non-private learner is efficient
- transfers accuracy of non-private learner to accuracy of the final construction

Accuracy of the final classifier h:

- By the guarantees of the previous algorithm A_p, misclassification rate in $\tilde{D}$ is small
- If B is a good learner, then misclassification rate of h is close to that in $\tilde{D}$

Realizable case: If B is a PAC learner for a concept class $\mathcal{H}$, and α is the desired bound on misclassification rate, then given $m \approx \frac{VC(\mathcal{H})}{\alpha^2}$ public feature-vectors, the private sample-complexity implied by our construction is $O\left(\frac{VC(\mathcal{H})^{3/2}}{\alpha^{3/2}}\right)$.

Agnostic case: If B is an agnostic-PAC learner for a concept class $\mathcal{H}$, and $\gamma \coloneqq \min_{h \in \mathcal{H}} err(h; \mathcal{D})$, then given $m \approx \frac{VC(\mathcal{H})}{\alpha^2}$ public feature-vectors, the private sample-complexity implied by our construction is $O\left(\frac{VC(\mathcal{H})^{3/2}}{\alpha^{5/2}}\right)$ to achieve a misclassification rate bound of $\alpha + O(\gamma)$.

Multi-label classification:

- Our construction can handle multi-label classification queries
- It is easy to obtain similar guarantees for the misclassification rate of the private learner

REFERENCES

[BNSV'15] - Bun Nissim Stemmer Vadhan, *FOCS'16.*

[DMNS'06] - Dwork McSherry Nissim Smith, *TCC'06.*

[PAE+'17] - Papernot Abadi Erligsson Goodfellow Talwar, *ICLR'17.*