Private Synthetic Data Generation

Steven Wu
School of Computer Science
Carnegie Mellon University

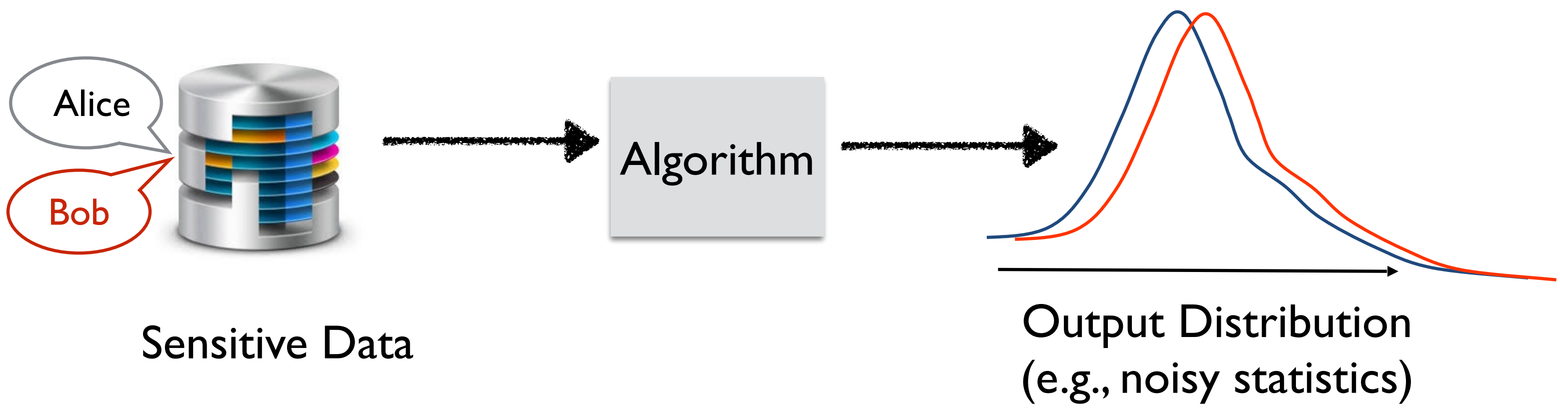
Announcement

- No recitation on Friday
- Day for Community Engagement

Today's Objectives

- Revisit PATE
- Starting Private Synthetic Data Generation

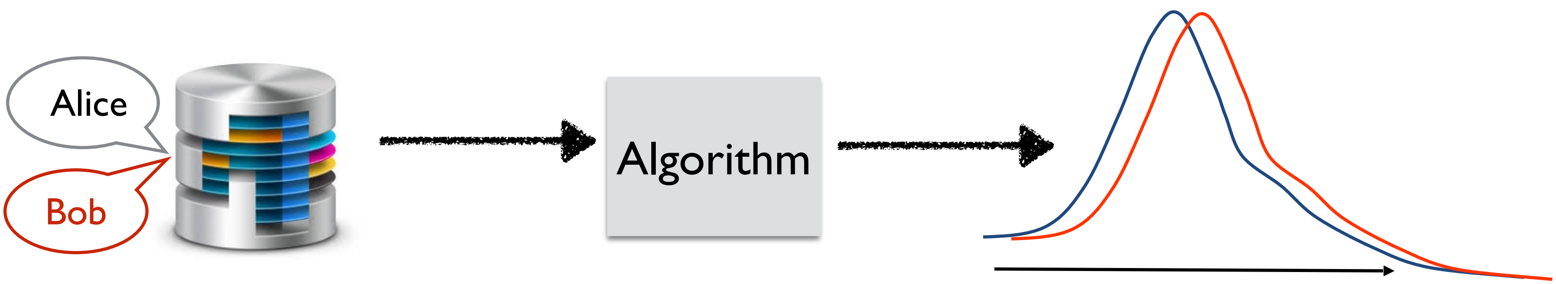
Both try to address a common challenge:
How to facilitate the use of DP for non-experts?



"An algorithm is *differentially private* if changing a single record does not alter its output distribution by much."
[DN03, DMNS06]

Definition: A (randomized) algorithm A is (ε, δ)-differentially private if for all neighbors D, D' and every $S \subseteq \mathrm{Range}(A)$

$$\Pr[A(D) \in S] \leq e^{\varepsilon} \Pr[A(D') \in S] + \delta$$



Sensitive Data Set
(e.g., medical records)

Output Distribution
(e.g., noisy statistics)

"An algorithm is *differentially private* if changing a single record does not alter its output distribution by much."
[DN03, DMNS06]

United States® Census 2020

Uber

LinkedIn

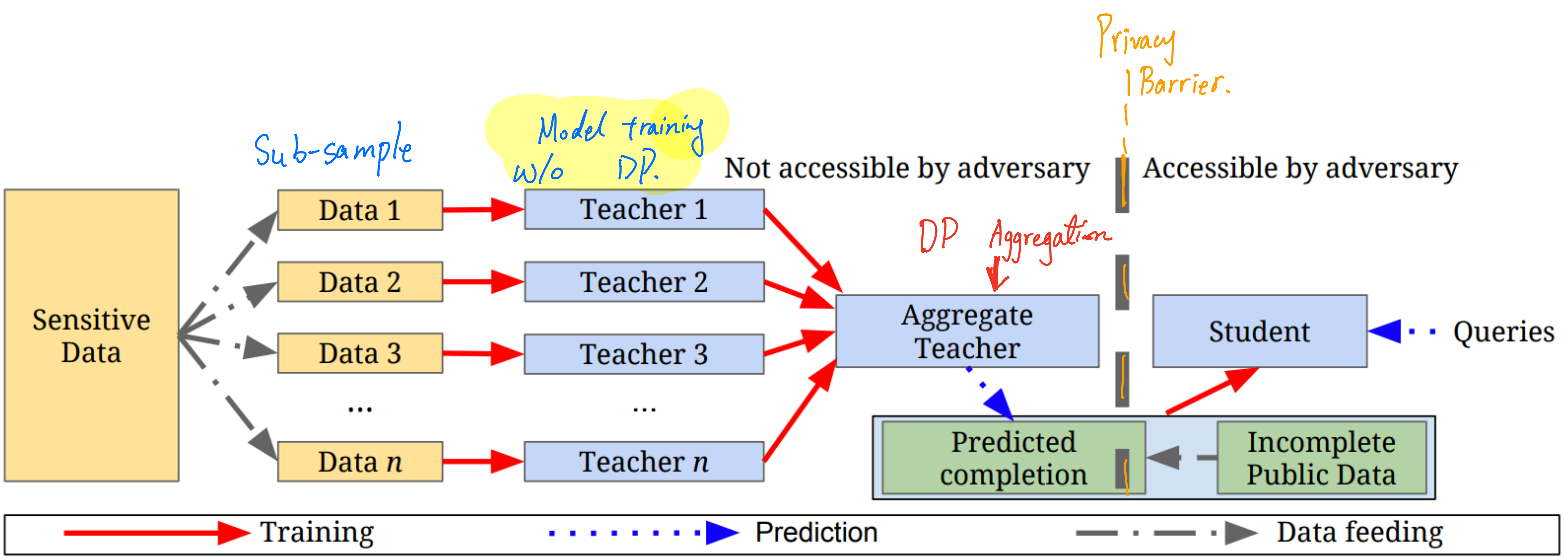
Meta

Challenge:

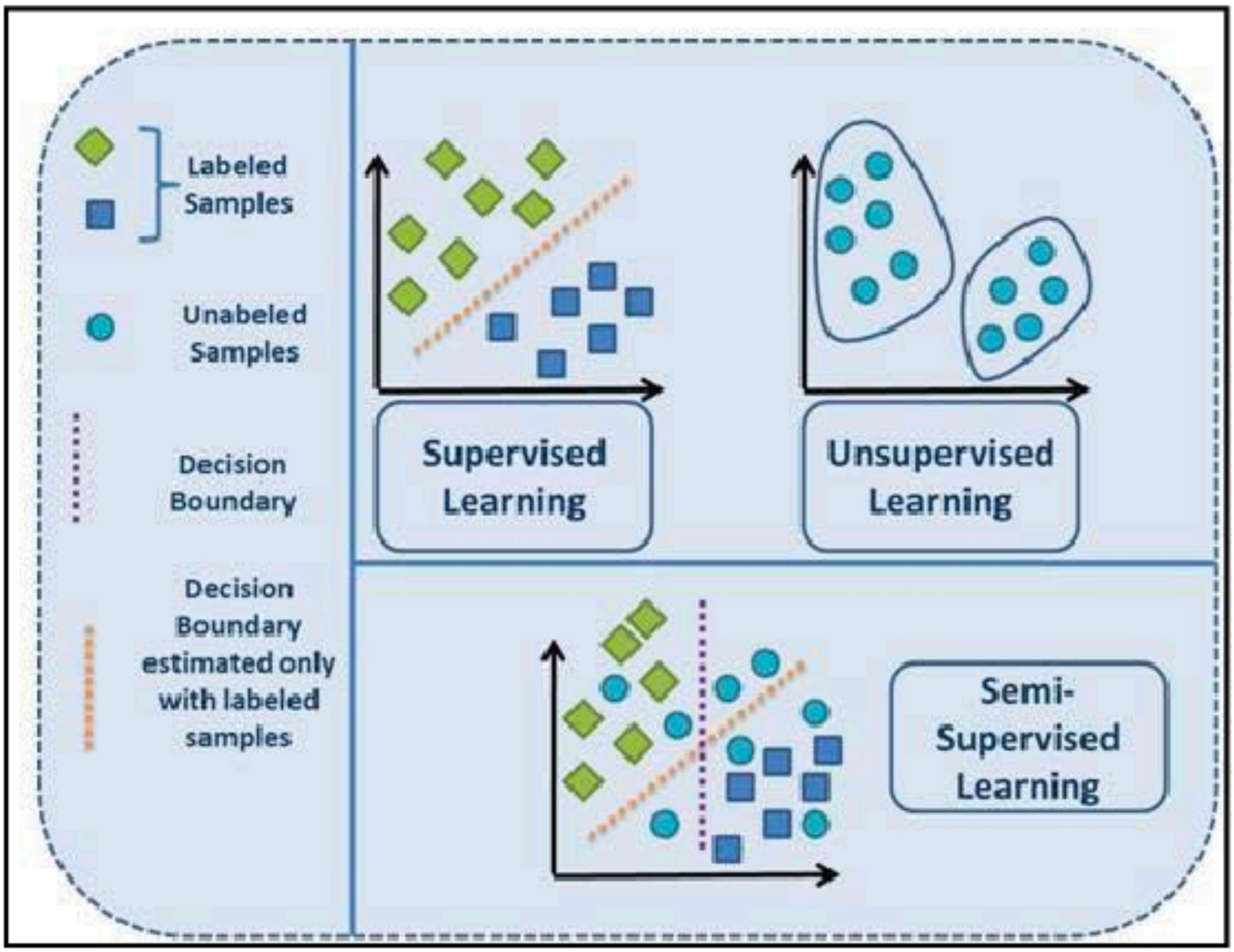
How can we enable non-experts to work with DP?

PATE

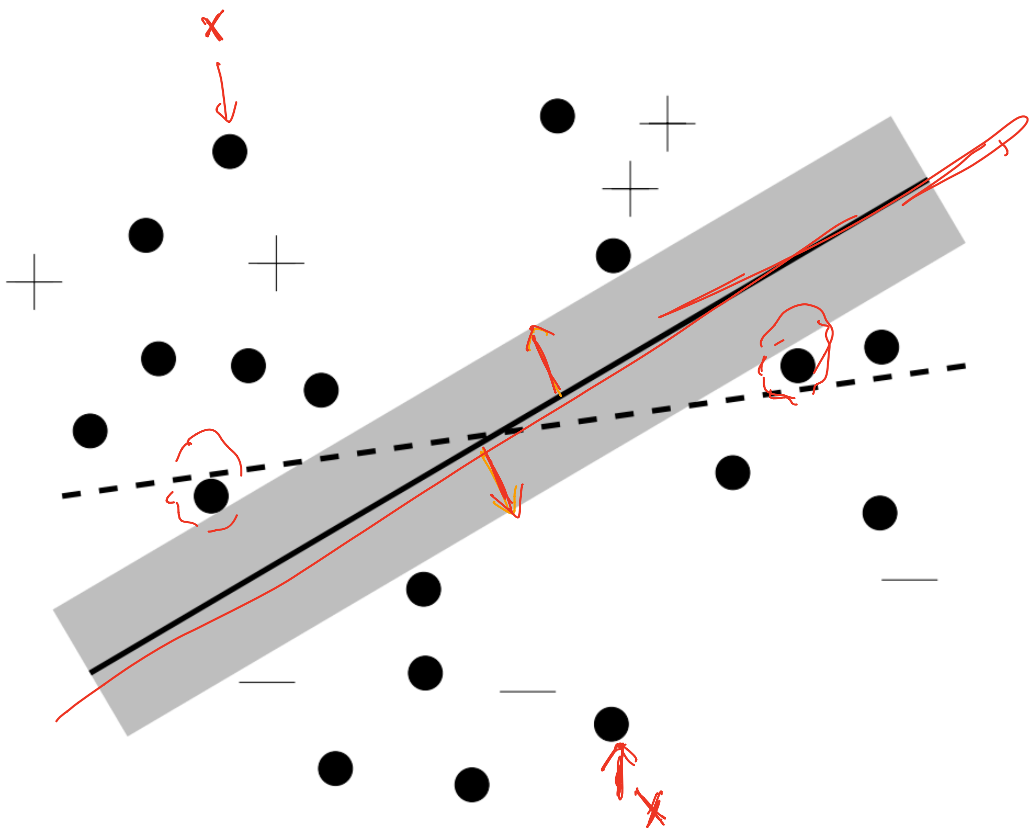
- Private Aggregation of Teacher Ensembles



Digression: Semi-Supervised Learning

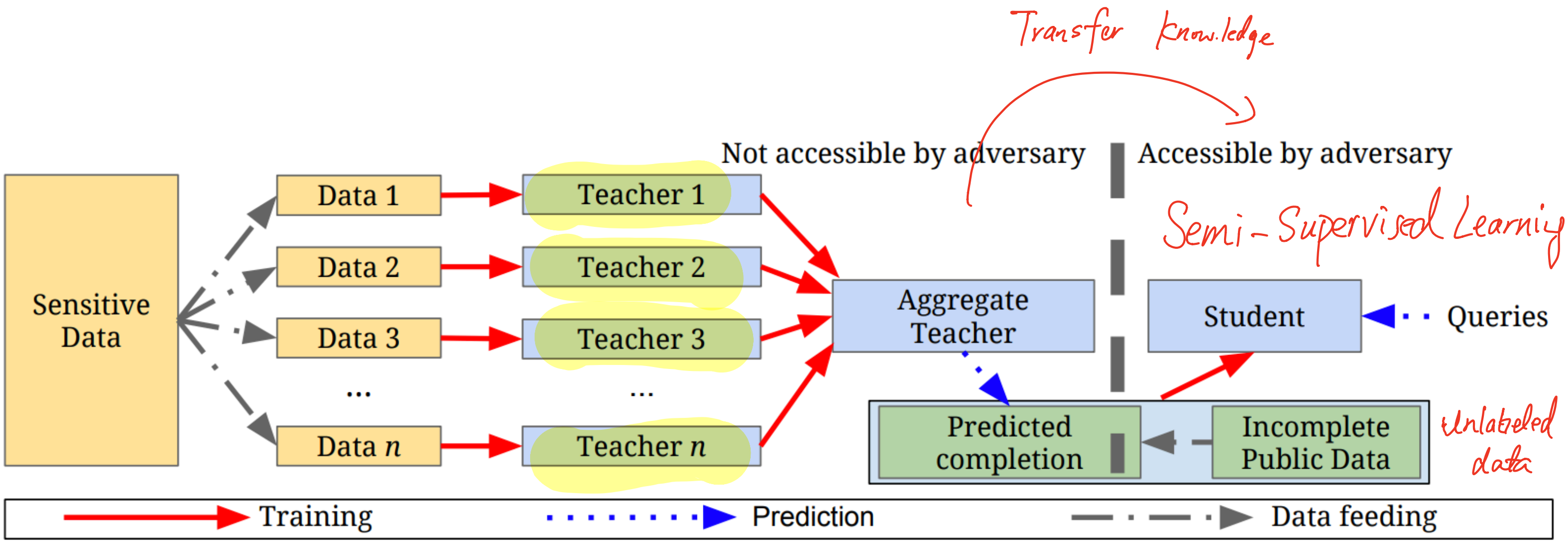


Semi-supervised Learning for SVM



PATE

- Private Aggregation of Teacher Ensembles



Unlabeled data point

$(x_1, \_)$

feature

missing label $\in \{0,1\}$

Teacher 1 $\rightarrow \hat{y} = 1$

2 $\rightarrow \hat{y} = 0$

3 $\rightarrow \hat{y} = 1$

$\vdots$

$n \rightarrow \hat{y} = 1$

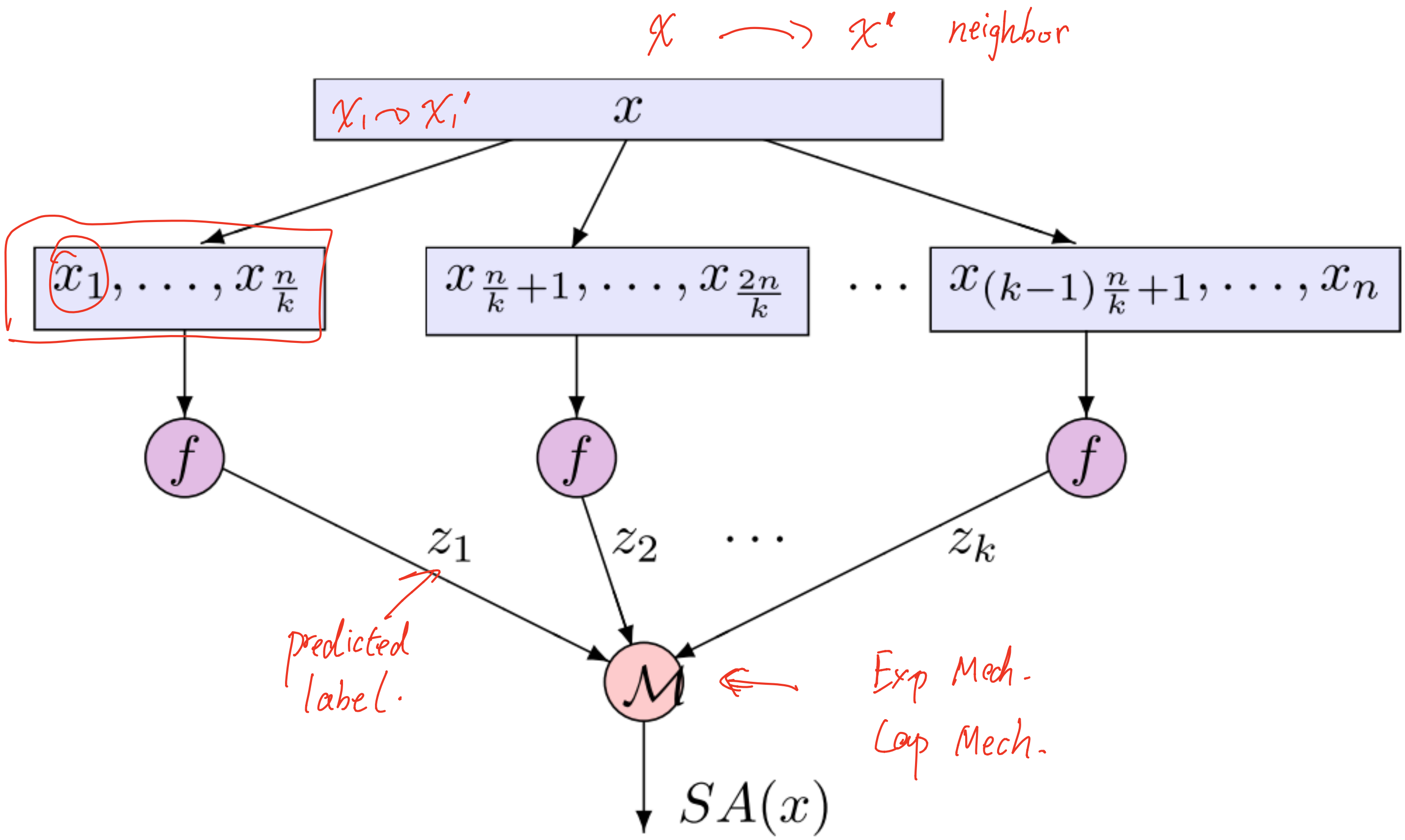
DP Aggregation

Return histogram w/ Lap noise

0 1

Return label via Exp Mech. / Report NM.

Sub-sample and Aggregate



Differentially Private Synthetic Data

Sensitive data set
(e.g. medical records)

DP
Algorithm

Synthetic data set
"*Fake*" data records that preserve important statistical properties

Allow arbitrary usage
post - processing

Data Scientist

Synthetic Data Release

1. Synthetic data for query/statistics release
   - A large collection of statistics in mind
2. General-purpose synthetic data
   - Exploratory data analysis
   - Training ML models
   - …

Synthetic Data Release

1. Synthetic data for query/statistics release
    - A large collection of statistics in mind
2. General-purpose synthetic data
    - Exploratory data analysis
    - Training ML models
    - …

Synthetic Data for Statistic/Query Release

Statistical / Counting Query Release

$$D \in (\{0, 1\}^d)^n$$

	Smoke	Lung Cancer	Diabetes	OCD	
patient_id1	1	1	1	1	$q(x) = 1$
patient_id2	1	0	0	1	$q(x) = 0$
patient_id3	1	1	0	1	$q(x) = 1$
patient_id4	0	0	1	0	$q(x) = 0$

$q(D) = 1/2$

Counting query: what is the fraction of people that satisfy some specified property q?

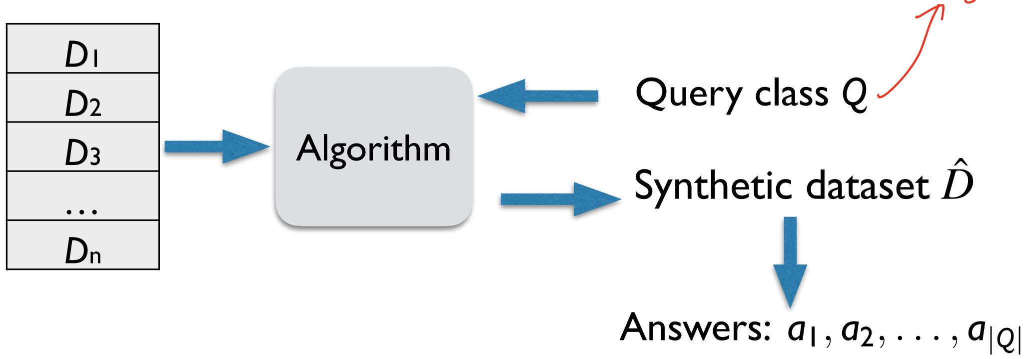
e.g. $q(x)$ = has "Smoke", "Lung Cancer" & "OCD"
(3-way Marginals)

Synthetic Data for Query Release

(used to be x)

Private data set

$\{q_1, \ldots, q_{|Q|}\}$



Goal: "max error" to be small

$$\max_{q \in Q} |a_q - q(D)| \to \text{small.}$$

Consistency:

For example,

#(smoke & lung cancer) + #(smoke & no lung cancer) = #(smoke)

Does not hold for Laplace / Gaussian

Long Line of Work

- [BLR08, RR10, HR10]
- [HLM12]
- [GGHRW14, ZCPSX14]
- [MSM19]
- [VTBSW20]
- [LVSUW21, ABKKMRS21]
- …

Theoretical Constructs

More Practical Methods

Terrance Liu, Giuseppe Vietri, Z. S. Wu
"Iterative Methods for Private Synthetic Data: Unifying Framework and New Methods"
To appear at NeurIPS 2021

Iterative Framework

w/ Adaptive Measurements

Define some loss function L that measures accuracy

For rounds $t = 1,\ldots,T$

1. <u>SELECT</u>: sample a set of queries Q_t for which the current synthetic dataset has high error

2. <u>MEASURE</u>: release noisy answers A_t for queries in Q_t

3. <u>UPDATE</u>: update the synthetic dataset to fit the noisy answers A_t according to the loss function L

Iterative Framework

w/ Adaptive Measurements

L: a loss function that measures accuracy

For rounds $t = 1, \ldots, T$

(1) Select

High-error query

$\tilde{q}_t \leftarrow Q$

(2) Measure

$\tilde{a}_t = \tilde{q}_t(\cancel{P}\,D) + z_t$

Gaussian/Lap

Synthetic Data set at iteration t.

$$D_t \leftarrow \arg\min_D \mathcal{L}(D, \{\tilde{q}_i\}_{i=1}^{t}, \{\tilde{a}_i\}_{i=1}^{t})$$

(3) Update

Adaptive Measurements

Restrict the synthetic dataset to belong to some family of distributions $\mathcal{D}$ and initialize $D_0 \in \mathcal{D}$

Adaptive Measurements

Restrict the synthetic dataset to belong to some family of distributions $\mathcal{D}$ and initialize $D_0 \in \mathcal{D}$

Define some loss function $\mathcal{L}$

Adaptive Measurements

Restrict the synthetic dataset to belong to some family of distributions $\mathcal{D}$ and initialize $D_0 \in \mathcal{D}$

Define some loss function $\mathcal{L}$

For **T** rounds (i.e., **t = 1...T**)

Adaptive Measurements

Restrict the synthetic dataset to belong to some family of distributions $\mathcal{D}$ and initialize $D_0 \in \mathcal{D}$

Define some loss function $\mathcal{L}$

For **T** rounds (i.e., **t = 1...T**)

1. **SELECT:** sample a set of queries $\tilde{Q}_t$

Adaptive Measurements

Restrict the synthetic dataset to belong to some family of distributions $\mathcal{D}$ and initialize $D_0 \in \mathcal{D}$

Define some loss function $\mathcal{L}$

For **T** rounds (i.e., **t = 1...T**)

1. **SELECT:** sample a set of queries $\tilde{Q}_t$
2. **MEASURE:** take noisy measurements of each query in $\tilde{A}_t$

Adaptive Measurements

Restrict the synthetic dataset to belong to some family of distributions $\mathcal{D}$ and initialize $D_0 \in \mathcal{D}$

Define some loss function $\mathcal{L}$

For **T** rounds (i.e., **t = 1...T**)

1. **SELECT:** sample a set of queries $\tilde{Q}_t$
2. **MEASURE:** take noisy measurements of each query in $\tilde{A}_t$
3. **UPDATE:** update the synthetic dataset to fit the noisy measurements according to the loss function $\mathcal{L}$

$$D_t \leftarrow \mathcal{L}\left(D_{t-1}, \tilde{Q}_t, \tilde{A}_t\right)$$

Adaptive Measurements

Under this framework, existing algorithms can be reduced to selections of $\mathcal{D}$ and $\mathcal{L}$

Examples:

- **MWEM** (Hardt et al., 2012)
- **DualQuery** (Gaboardi et al., 2014)
- **FEM** (Vietri et al., 2020)
- **RAP$^{\text{softmax}}$**
    - Adapted from RAP (Aydore et al., 2021)

Two-Player Zero-Sum Game

Private data D

Synthetic Data Player

Init $D^{(0)}$

Query Player.

$\xleftarrow{q^{(1)}}$

$t=1$:

$D^{(1)} \leftarrow \text{Update}\left(D^{(0)}, q^{(1)}, \tilde{a}^{(1)}\right)$

$|q^{(1)}(D^{0}) - q^{(1)}(D)|$ is Large

$\tilde{a}^{(1)} = q^{(1)}(D) + \text{Noise}$

$t=2$:

$D^{(2)} \leftarrow \text{Update}\left(D^{(1)}, \{q^{(1)}, q^{(2)}\}, \{\tilde{a}^{(1)}, \tilde{a}^{(2)}\}\right)$

$\vdots$

$\vdots$