Training Private Deep Learning Model with DPSGD

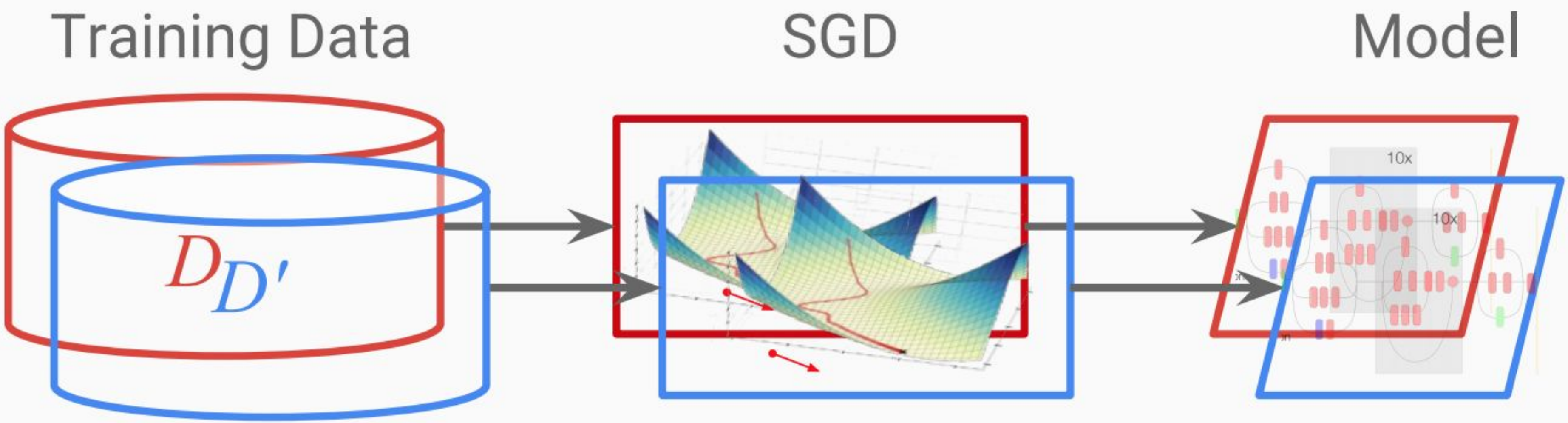
Tianhao Wang, Silin Zou, Litao Yan, Ruoxi Yang

(ε, δ)-Differential Privacy: The distribution of the output $M(D)$ on database D is (nearly) the same as $M(D')$:

$$\forall S: \quad \Pr[M(D)\in S] \le \exp(\varepsilon) \cdot \Pr[M(D')\in S]+\delta.$$

quantifies information leakage

allows for a small probability of failure



1. [1607.00133] Deep Learning with Differential Privacy

DPSGD Algorithm

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

[1607.00133] Deep Learning with Differential Privacy

Data

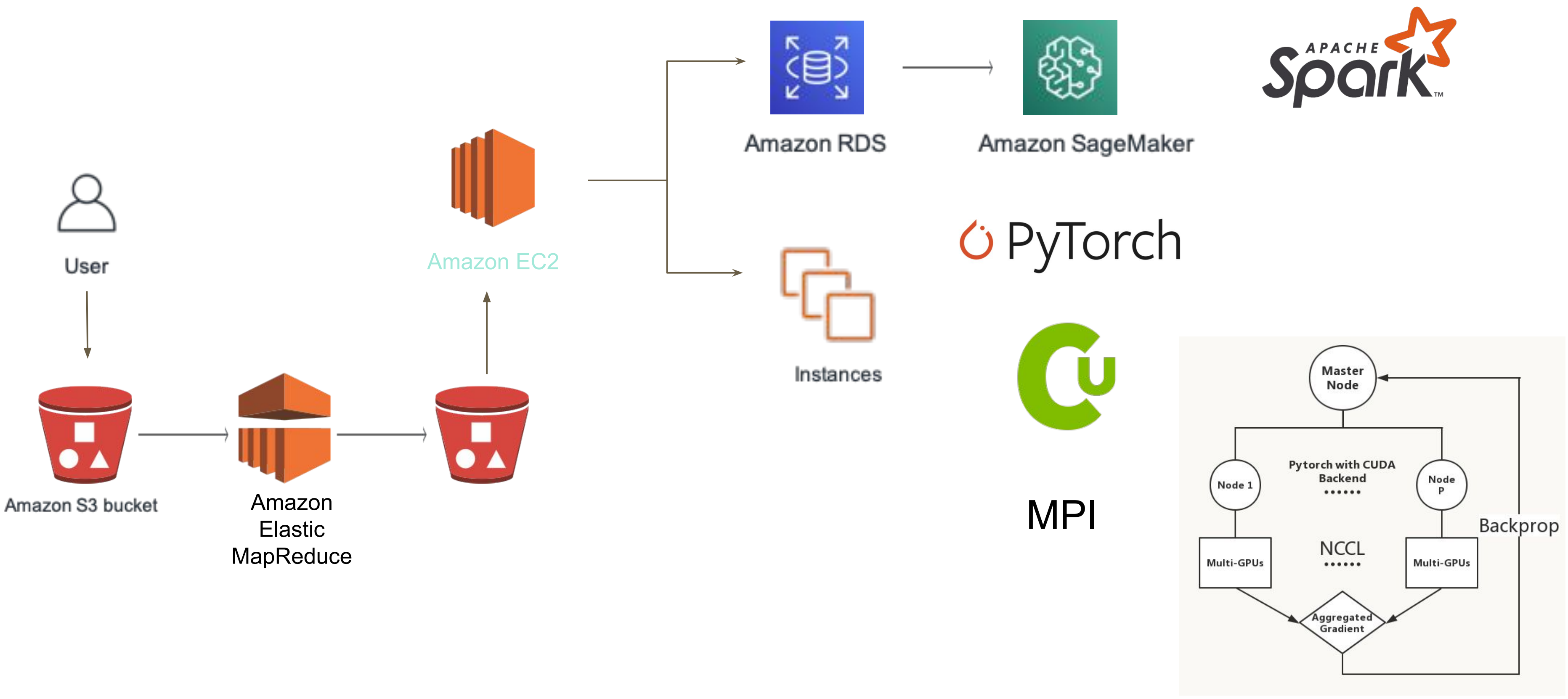
Source: Public Use Microdata Sample (PUMS)

https://www.census.gov/programs-surveys/acs/data/pums.html

Features: Sex, Married, Collegedegree, Employed, Millitary Service, US Citizen, etc.

Objective: Train a classifier to predict the unemployment rate without exposing sensitive information.

	sex	married	black	asian	collegedegree	employed	militaryservice	uscitizen	disability	englishability	blackfemale
0	1	1	0	0	0	1	0	1	0	1	0
1	1	1	0	0	0	0	0	1	0	1	0
2	0	1	0	0	0	0	1	1	0	1	0
3	0	0	0	0	0	0	0	1	1	1	0
4	0	1	0	0	0	0	0	1	0	1	0



Credit: https://sophieyanzhao.github.io/model