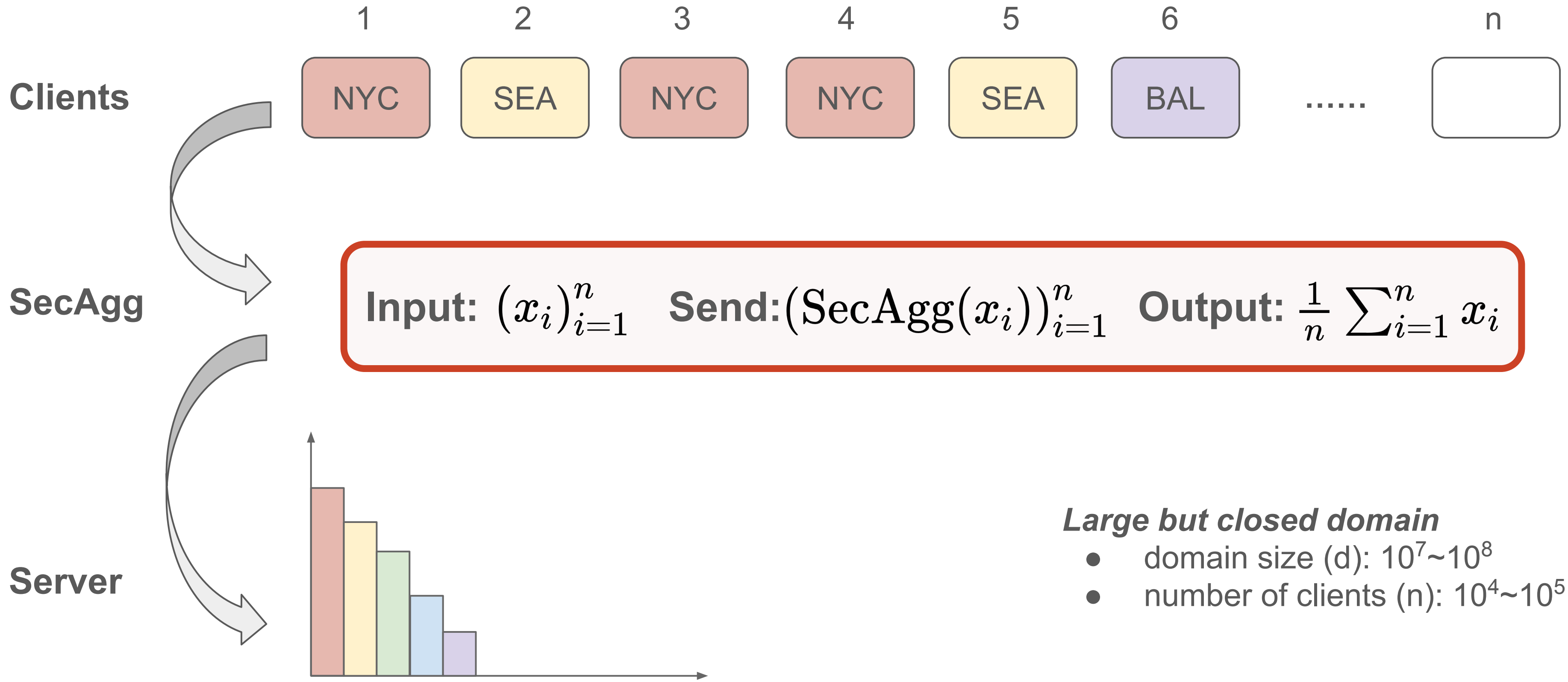
Private Federated Frequency Estimation: Adapting to the Hardness of the Instance

(NeurIPS 2023)

Jingfeng Wu

with Wennan Zhu, Peter Kairouz, and Vova Braverman

Secure Frequency Estimation



Large but closed domain

- domain size (d): $10^7 \sim 10^8$
- number of clients (n): $10^4 \sim 10^5$

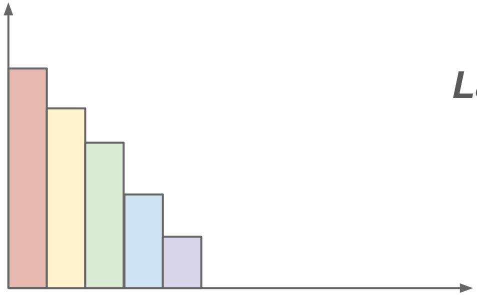
Naive approach: one-hot encoding

1	2	3		n
NYC	SEA	NYC	......	

$x_2 := (0, 1, 0, \ldots)$

Input: $x_1 \cdots x_n$ **Output:** $\frac{1}{n}\sum_{i=1}^{n} x_i$

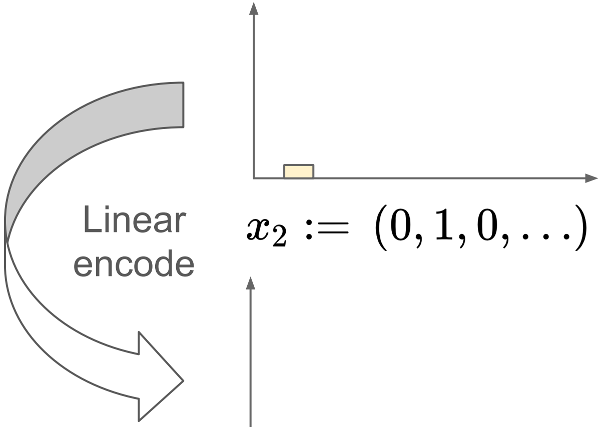
Bandwidth = Õ(d)



Large but closed domain

- domain size (d): 10^7~10^8
- number of clients (n): 10^4~10^5

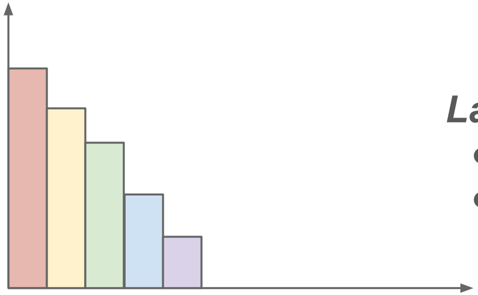
Save bandwidth with sparse encoding



$x_2 := (0, 1, 0, \ldots)$

$Sx_2 := (0, 1, 0, 0, 0)$

Input: $(Sx_i)_{i=1}^{n}$ **Output:** $\frac{1}{n} S \sum_{i=1}^{n} x_i$



Large but closed domain

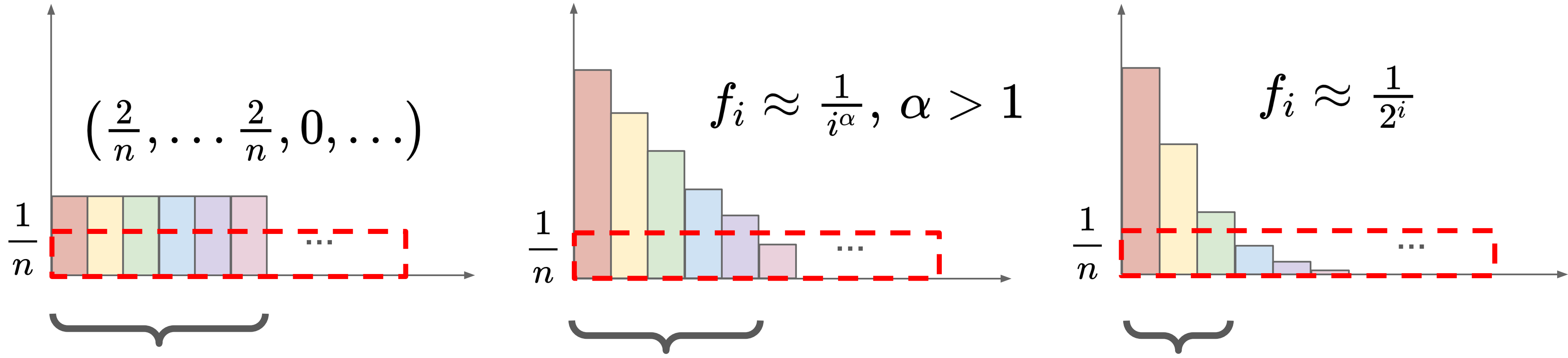
- domain size (d): $10^7 \sim 10^8$
- number of clients (n): $10^4 \sim 10^5$

At most n non-zeros

Bandwidth = Õ(n)

Price, Eric. "Sparse Recovery." (2020): 140-164.

Save more bandwidth?

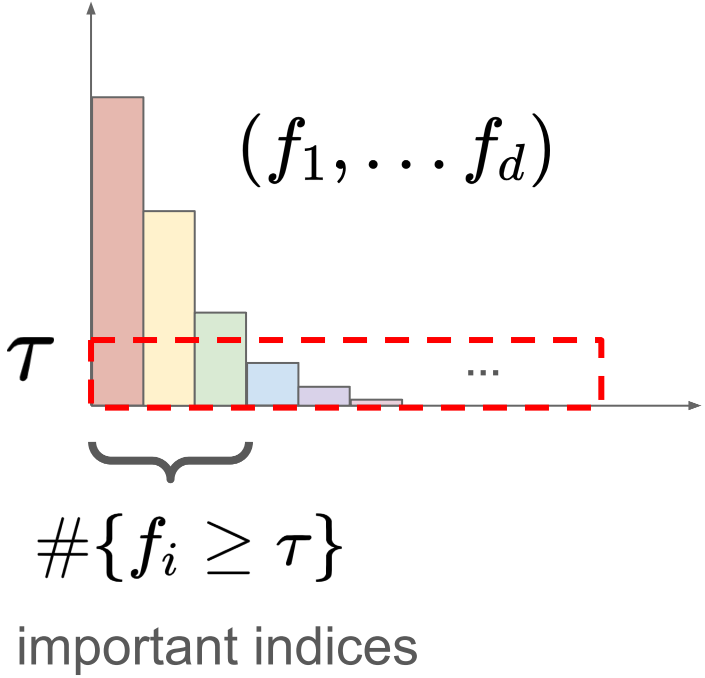


n/2
important indices

O($n^{1/\alpha}$)
important indices

O(log n)
important indices

If an oracle tells you the set of important indices

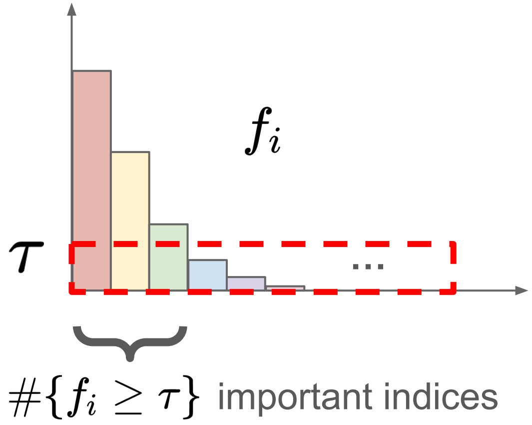


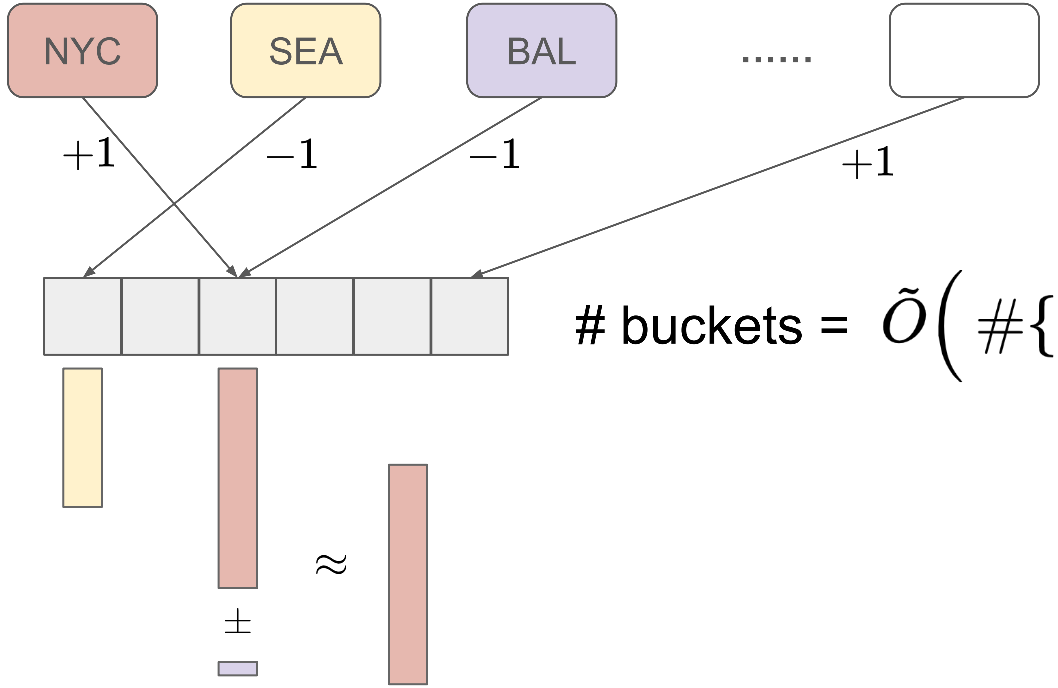
- Error parameter τ (e.g., τ = 1/n)
- Only counts over a small domain
- Bandwidth $\tilde{O}(\#\{f_i \geq \tau\})$
- This is minimum!

Without knowing the important indices,

Count-Sketch:

randomly assign the d counters to a few buckets





buckets = $\tilde{O}\left(\#\{f_i \geq \tau\} + \underbrace{\frac{1}{\tau^2} \sum_{f_i<\tau} f_i^2}_{\text{Additional buckets for dealing with collisions}} \right)$

Estimate the bandwidth $\tilde{O}\left(\#\{f_i \geq \tau\} + \frac{1}{\tau^2} \sum_{f_i < \tau} f_i^2 \right)$

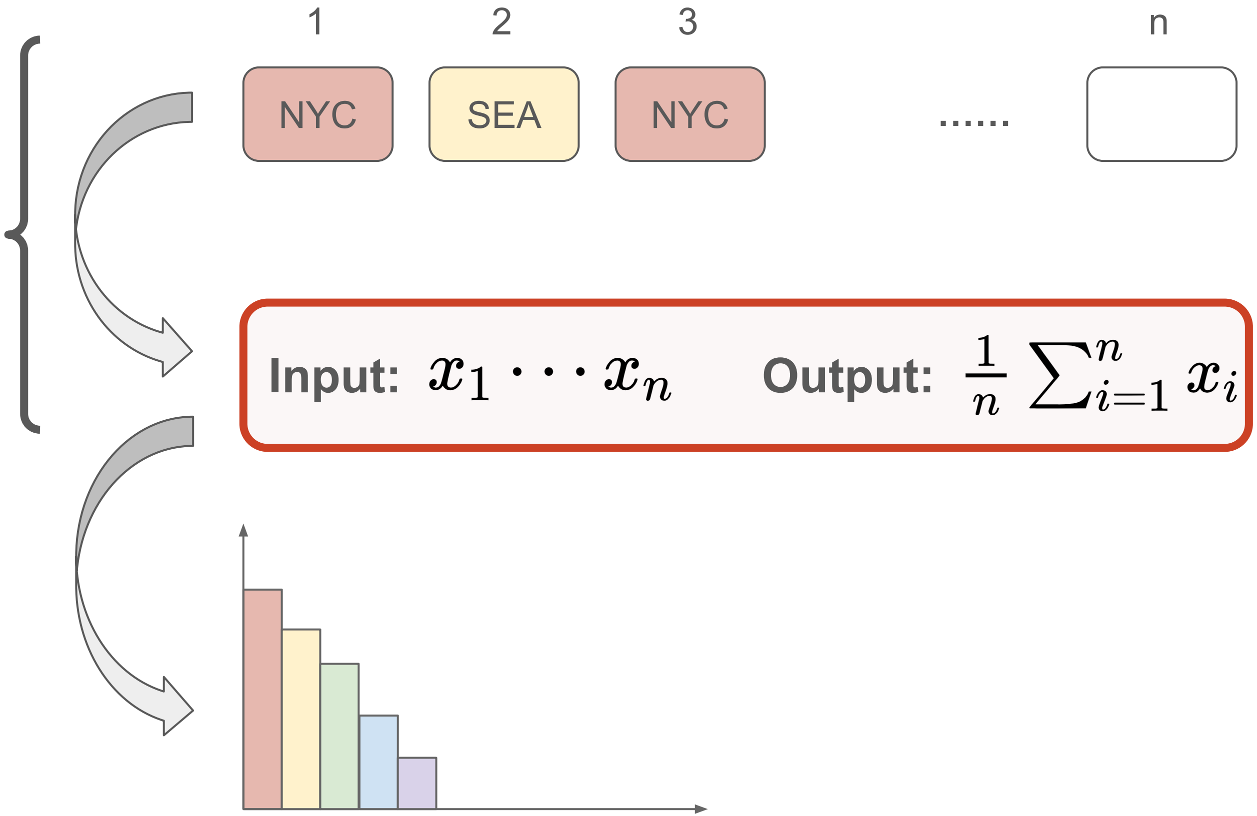
1. Assume that $f_i \approx \beta \cdot i^{-\alpha} \Leftrightarrow \log f_i \approx \log \beta - \alpha \cdot \log i$
2. **P1**. Small sketch collecting a few top counts $\left(1, \hat{f}_1\right), \ldots, \left(20, \hat{f}_{20}\right)$
3. Estimate log(β) and α by linear regression
4. $$\#\{f_i > \tau\} + \frac{1}{\tau^2} \sum_{f_i < \tau} f_i^2 \approx \left(\frac{\beta}{\tau}\right)^{\frac{1}{\alpha}} + \frac{\beta^2}{(2\alpha - 1)\tau^2} \cdot \left(\frac{\beta}{\tau}\right)^{\frac{1-2\alpha}{\alpha}}$$
5. **P2**. Set the bandwidth & run

Cevher V. Learning with compressible priors. Advances in Neural Information Processing Systems. 2009;22.

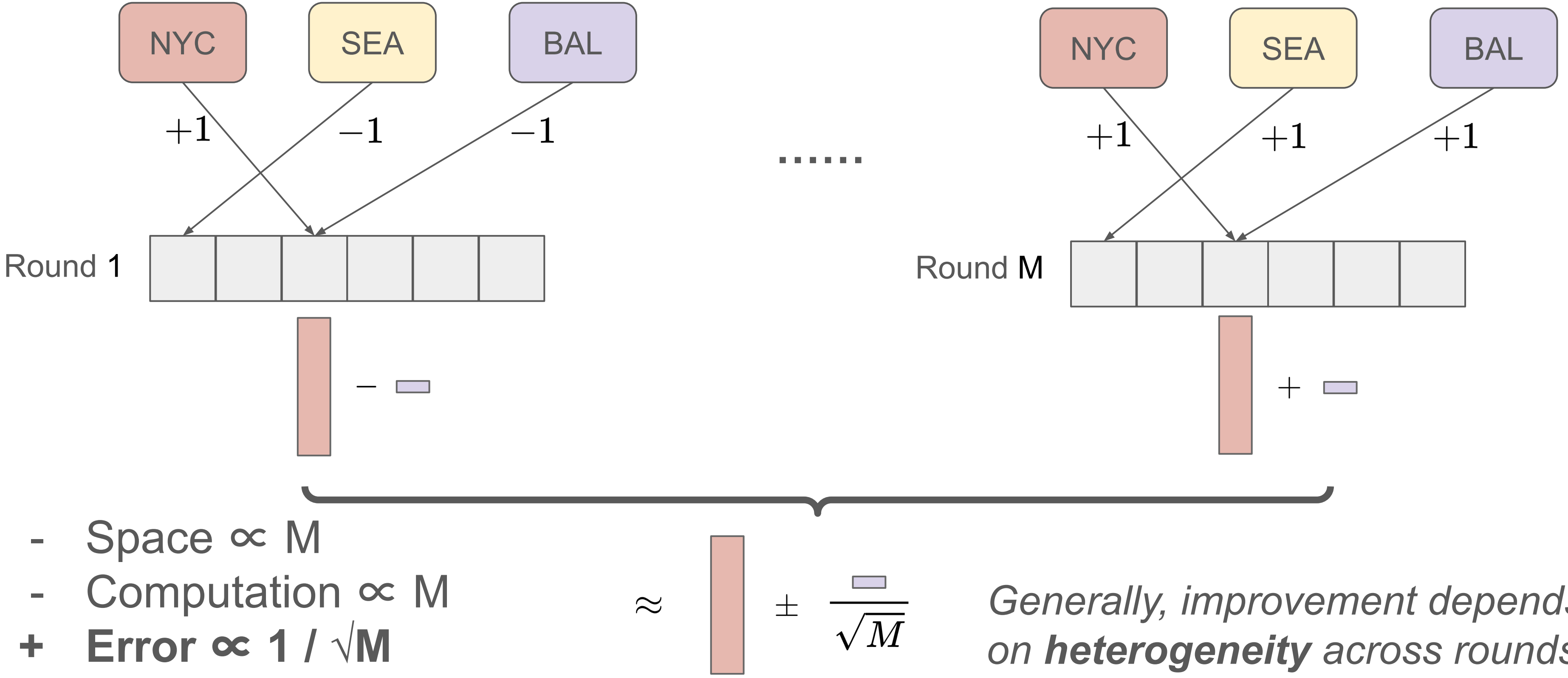
Real-world FA: from single round to multiple rounds

M rounds,
each client
participates once

Server observes
M partial sums



Hybrid sketches: shared bucket hash + fresh sign hash



- Space ∝ M
- Computation ∝ M

\+ **Error ∝ 1 / √M**

*Generally, improvement depends on **heterogeneity** across rounds*

Multi-round: DP for hybrid sketches

$\sigma \propto l_2$-sensitivity $\propto$

- 1 / #{clients *per round*}
- √#{independent sketches *per round*}

*Release **M** sketches*