VLDB, Vancouver, August 30 2023
PVLDB, Volume 16, Issue 4

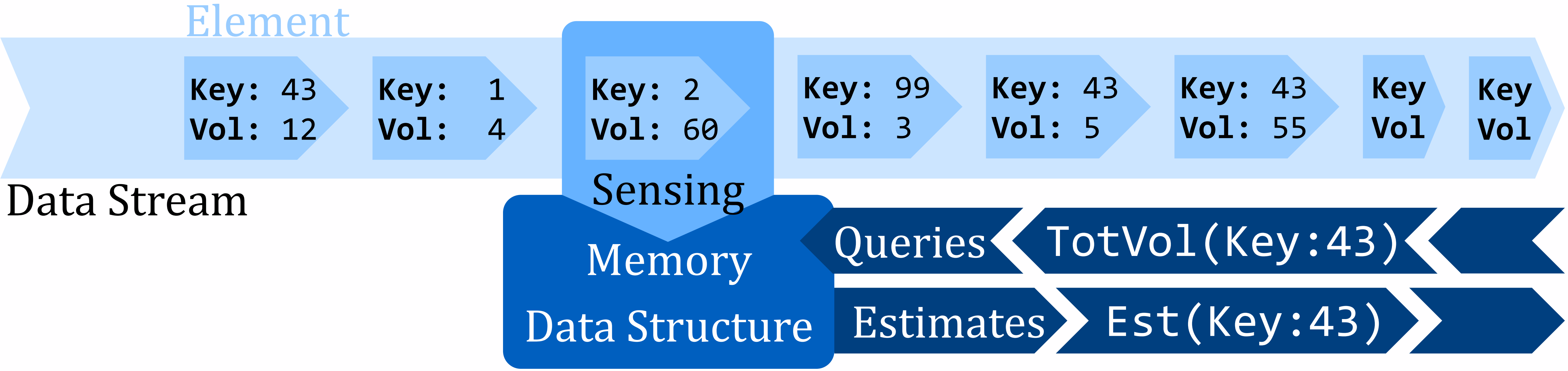
Bayesian Sketches for Volume Estimation in Data Streams

Author | Presenter : Francesco Da Dalt

Co-Authors: Simon Scherrer
Adrian Perrig

ETH zürich

Volume Estimation in Data Streams



Problem Constraints

High Throughput Data Stream

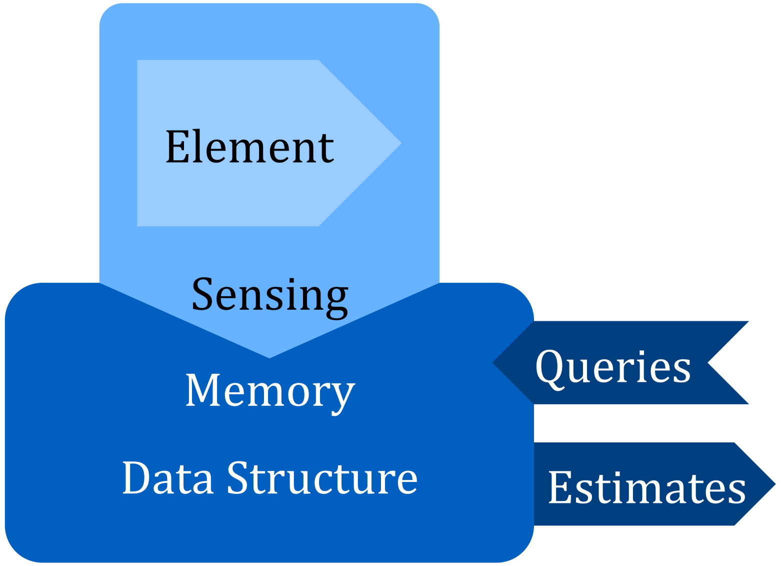
Low Overhead Sensing

Little Memory

Large Number of Keys

Network Supervision	**Key:** 192.268.123.132 **Vol:** 14 kB
Media Analytics	**Key:** #VLDB **Vol:** 12 clicks
Page Rank	**Key:** www.wikipedia.org **Vol:** +50 ranks

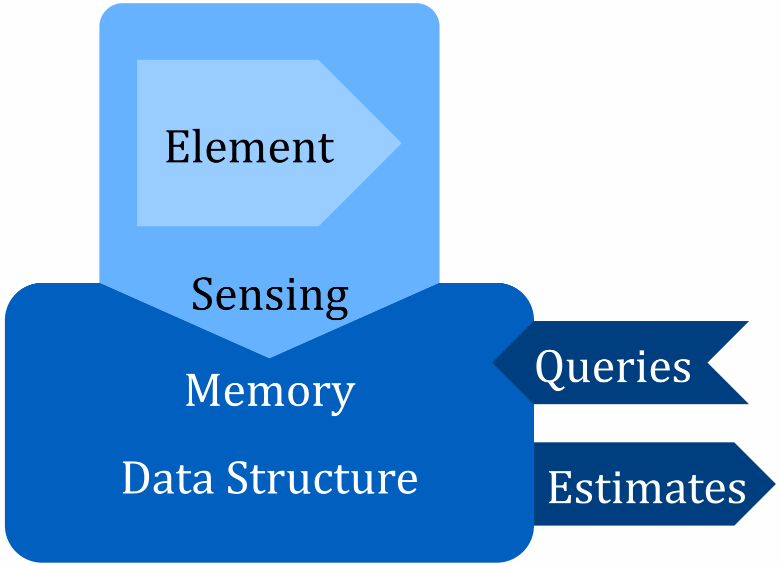
State Of The Art

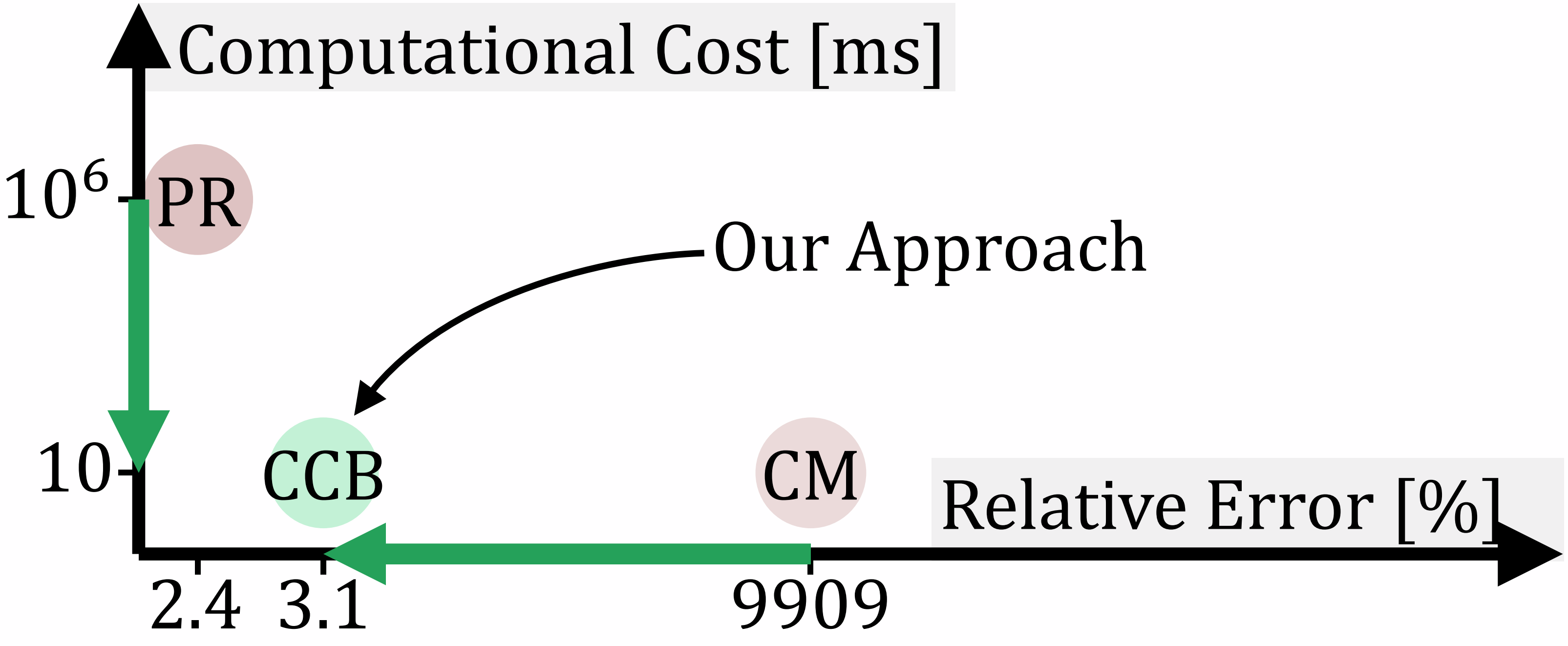


Where do we stand today ?

Two Approaches		Accuracy	Efficiency
Classical	▪ Count-Min (CM) Sketch ▪ Count (C) Sketch	−	+
Modern	▪ Seq Sketch ▪ PR Sketch	+	−

State Of The Art





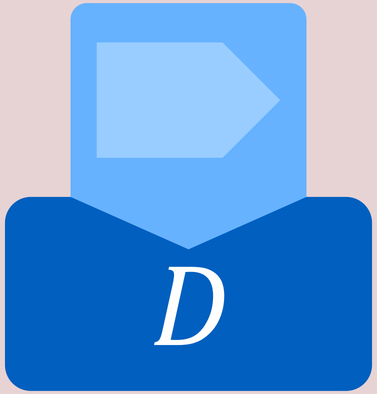
Where do we stand today ?	We developed	+	+
Classical	▪ Count-Min (CM) Sketch ▪ Count (C) Sketch	−	+
Modern	▪ Seq Sketch ▪ PR Sketch	+	−
Two Approaches		Accuracy	Efficiency

Problem Formulation

$$\begin{bmatrix} TotVol(key{:}\,1) \\ \vdots \\ TotVol(key{:}\,i) \\ \vdots \\ TotVol(key{:}\,n) \end{bmatrix} = a \qquad \hat{a} = \begin{bmatrix} Est(key{:}\,1 \mid D) \\ \vdots \\ Est(key{:}\,i \mid D) \\ \vdots \\ Est(key{:}\,n \mid D) \end{bmatrix}$$

„Ground Truth“ „Estimates“

Compress Data Stream



Compute Estimates

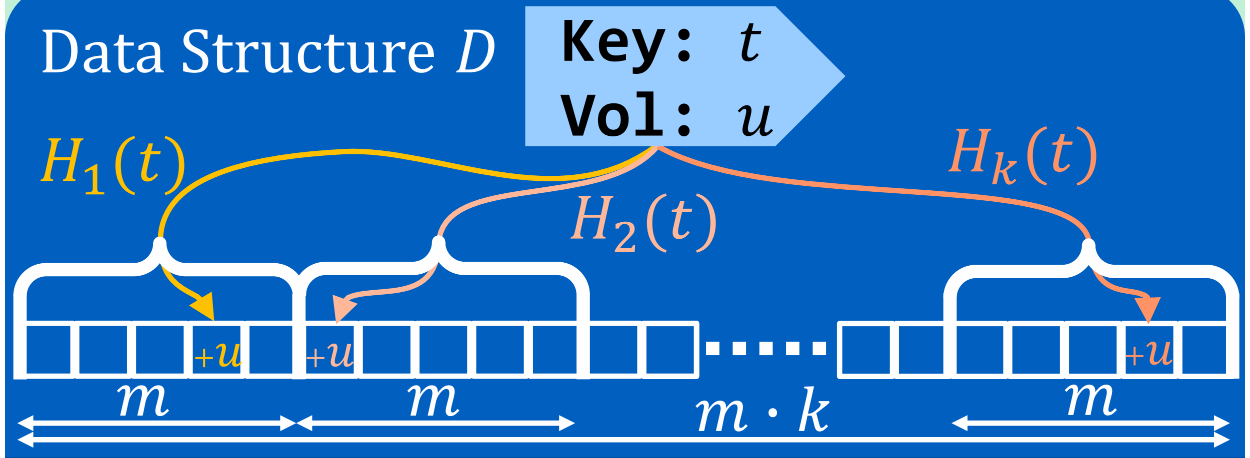
Such That $\|a - \hat{a}\|$ Is Small

New Perspective

What if $\mathbb{P}[D \mid a]$ exists ?

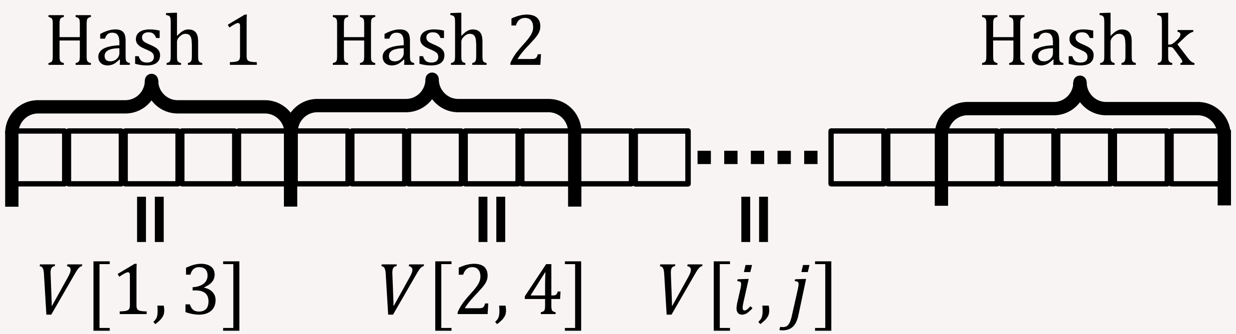
Then what is $\underset{\hat{a}}{\operatorname{argmax}}\, \mathbb{P}[\hat{a} \mid D]$ $\xrightarrow{\text{Bayes}}$ $\underset{\hat{a}}{\operatorname{argmax}}\, \mathbb{P}[D \mid \hat{a}] \cdot \mathbb{P}[\hat{a}]$

How to compute ? Keep D simple:



i-th Counter of j-th Hash: $V[\,j,\ i\,]$

Fundamental Observation



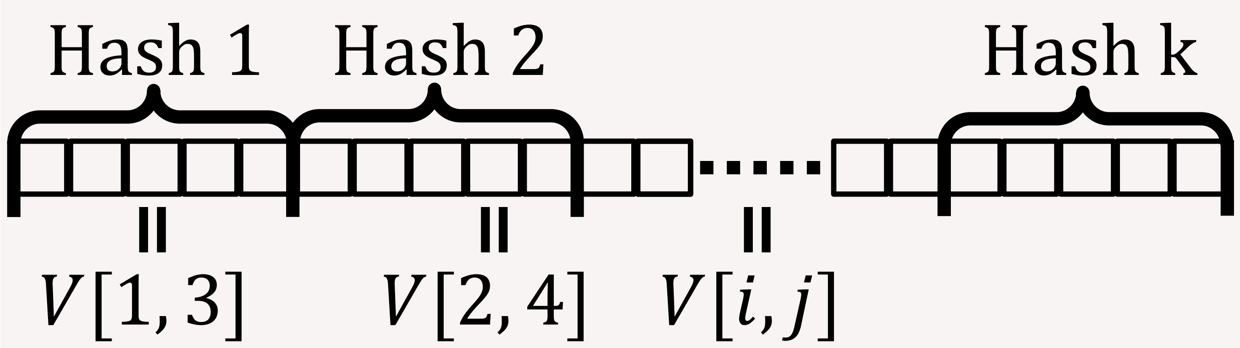
$$V[i,j] = \sum_{t=1}^{n} \underbrace{\mathbf{1}_{\{H_i(t)=j\}}}_{\text{Universal Hash Function}} \cdot a_t$$

$$V[i,j] = \sum_{t=1}^{n} Ber\left(\frac{1}{m}\right) \cdot a_t$$

$V[i,j]$ is a Sum of Independent Random Variables

ETH zürich

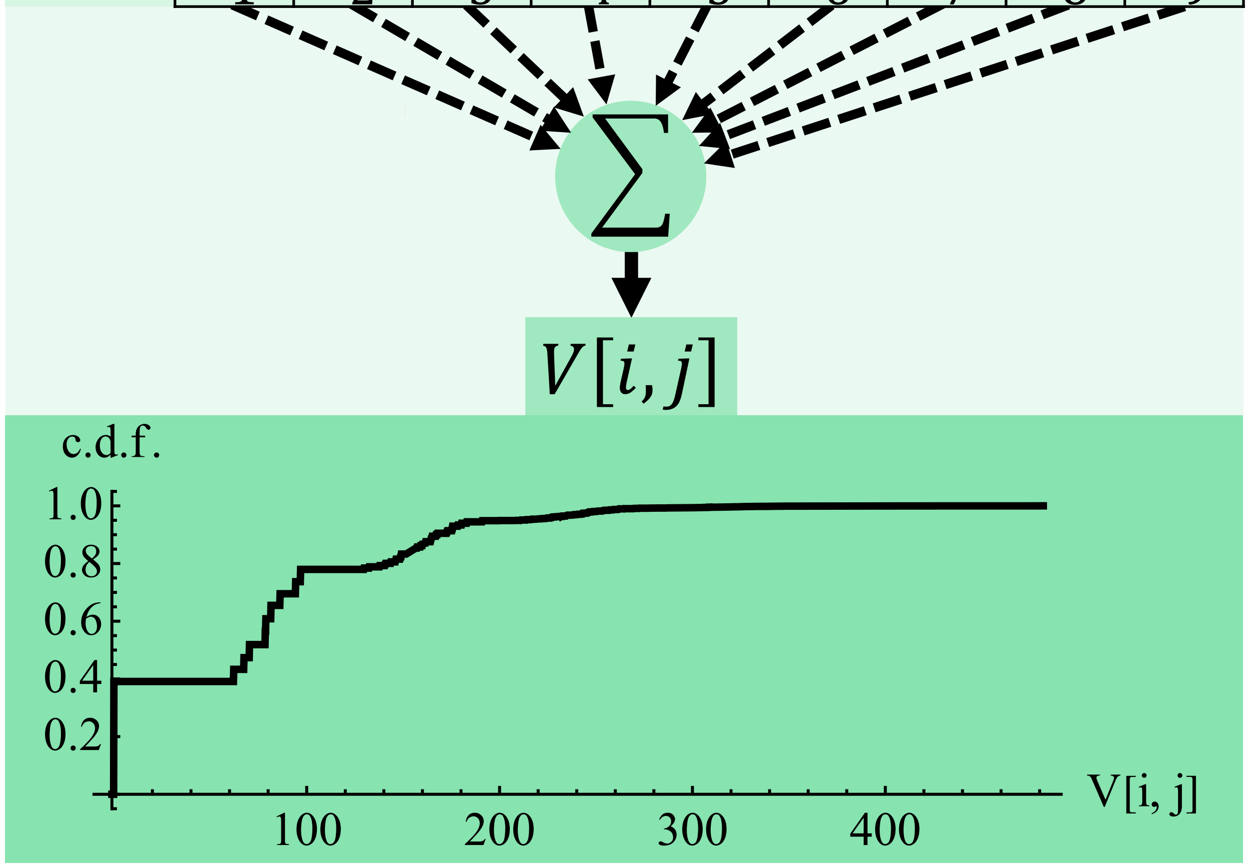
Fundamental Observation



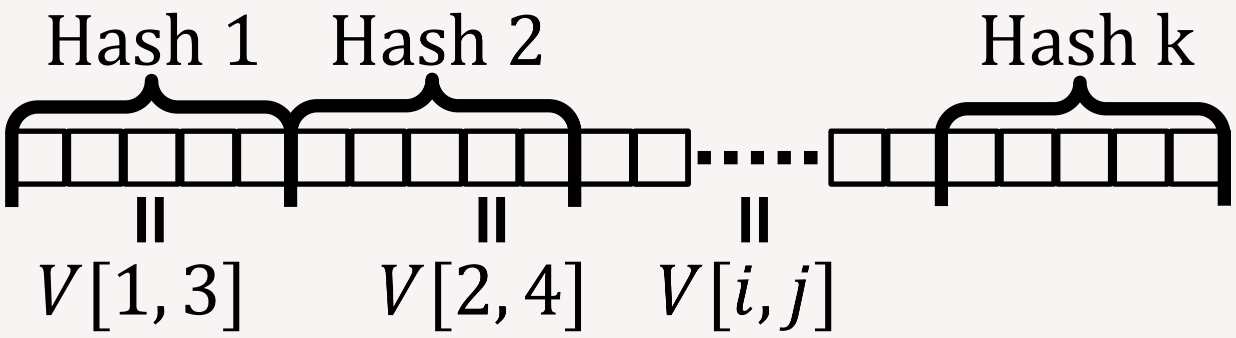
$$V[i,j] = \sum_{t=1}^{n} \underbrace{\mathbf{1}_{\{H_i(t)=j\}}}_{\text{Universal Hash Function}} \cdot a_t$$

$$\underbrace{V[i,j] = \sum_{t=1}^{n} Ber\left(\frac{1}{m}\right) \cdot a_t}_{V[i,j] \text{ is a Sum of Independent Random Variables}}$$

Key	1	2	3	4	5	6	7	8	9
Vol	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9



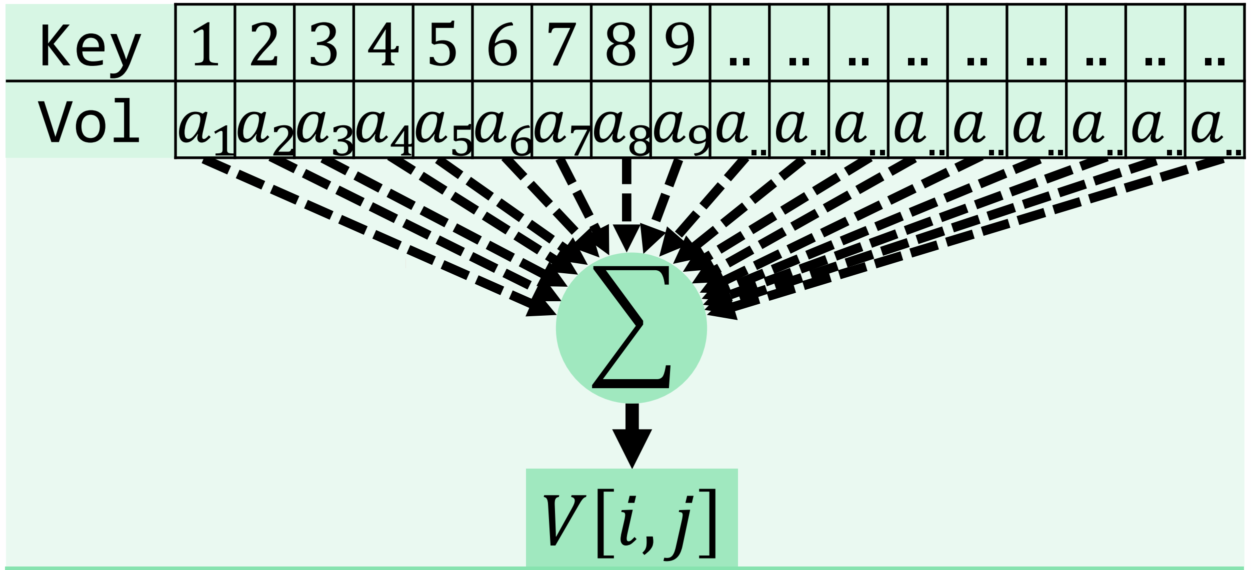
Fundamental Observation

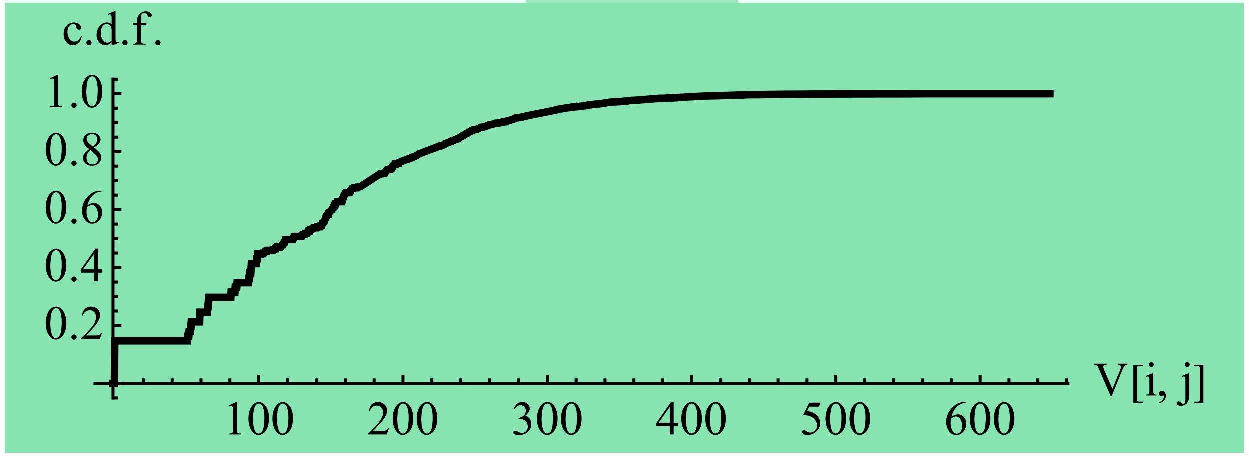


$$V[i,j] = \sum_{t=1}^{n} \underbrace{\mathbf{1}_{\{H_i(t)=j\}}}_{\text{Universal Hash Function}} \cdot a_t$$

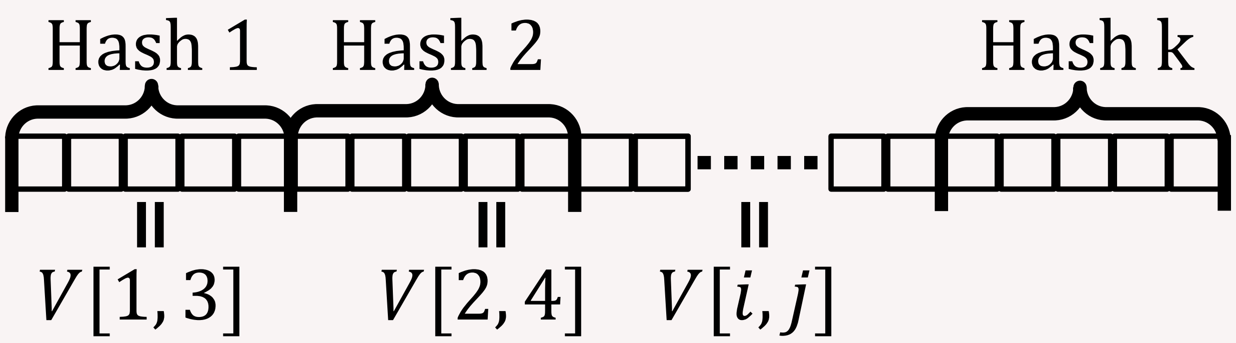
$$\underbrace{V[i,j] = \sum_{t=1}^{n} Ber\left(\frac{1}{m}\right) \cdot a_t}$$

$V[i,j]$ is a Sum of Independent Random Variables



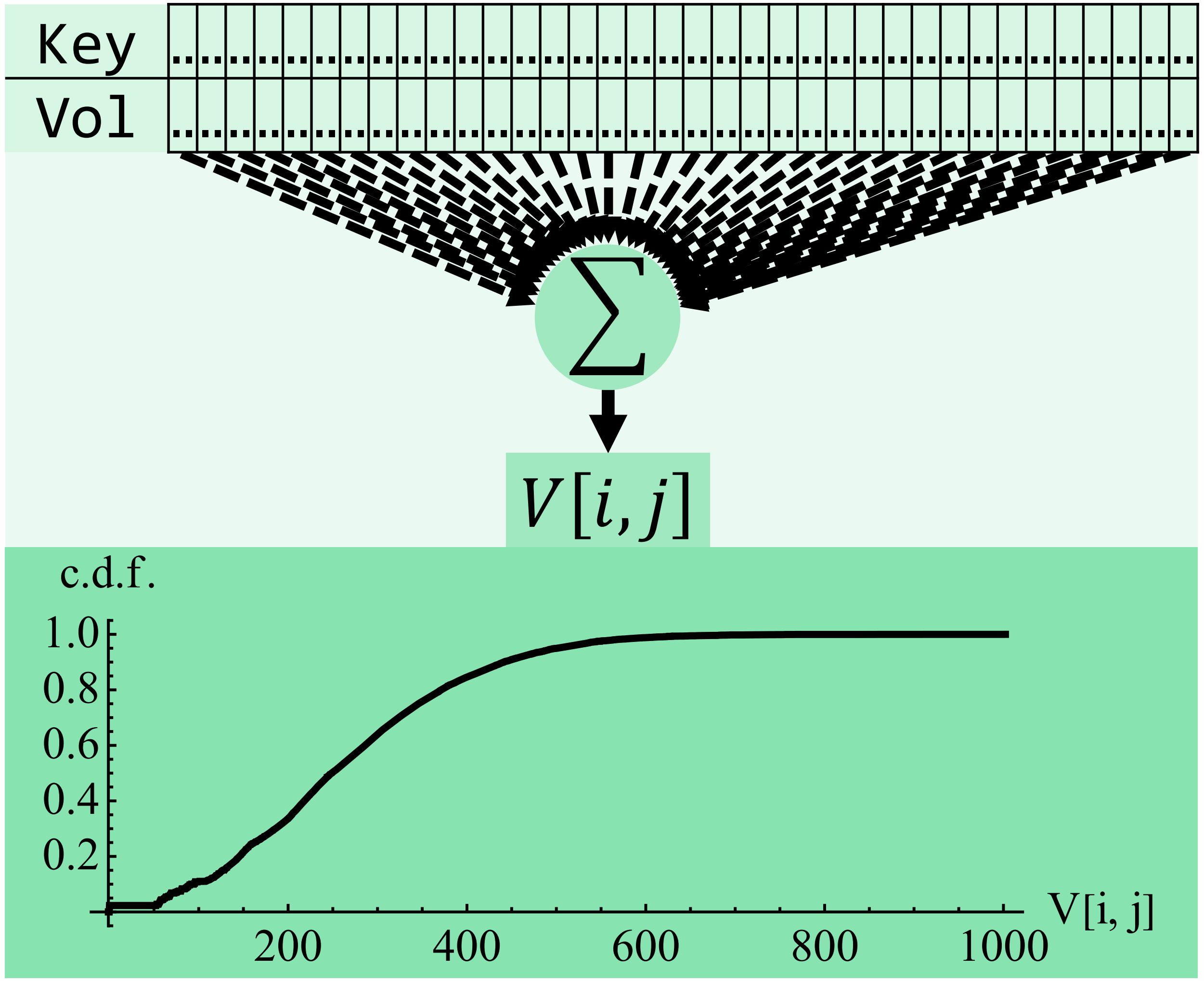


Fundamental Observation

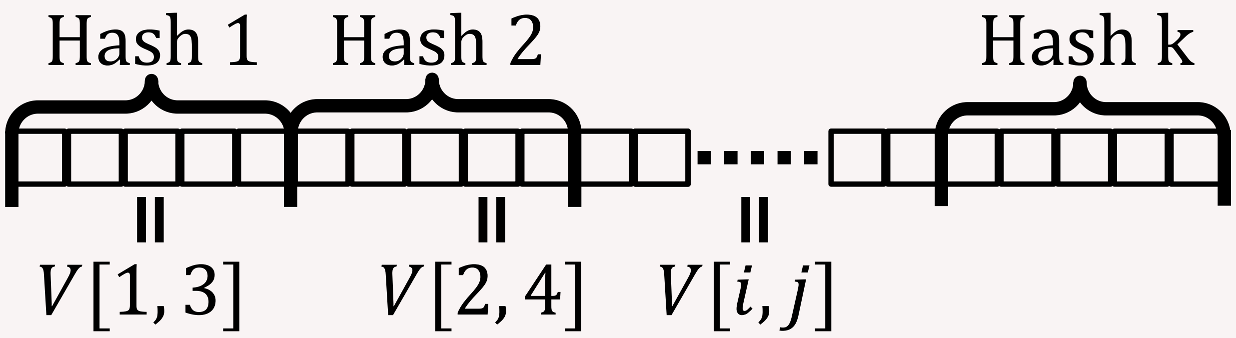


$$V[i,j] = \sum_{t=1}^{n} \underbrace{\mathbf{1}_{\{H_i(t)=j\}}}_{\text{Universal Hash Function}} \cdot a_t$$

$$\underbrace{V[i,j] = \sum_{t=1}^{n} Ber\left(\frac{1}{m}\right) \cdot a_t}_{V[i,j] \text{ is a Sum of Independent Random Variables}}$$



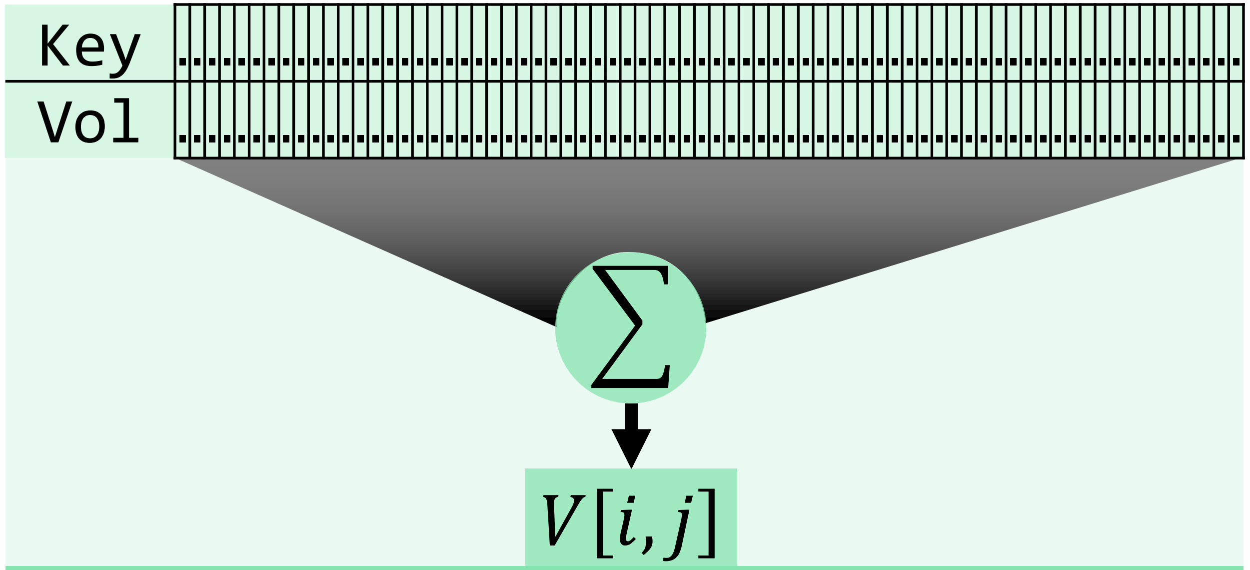
Fundamental Observation

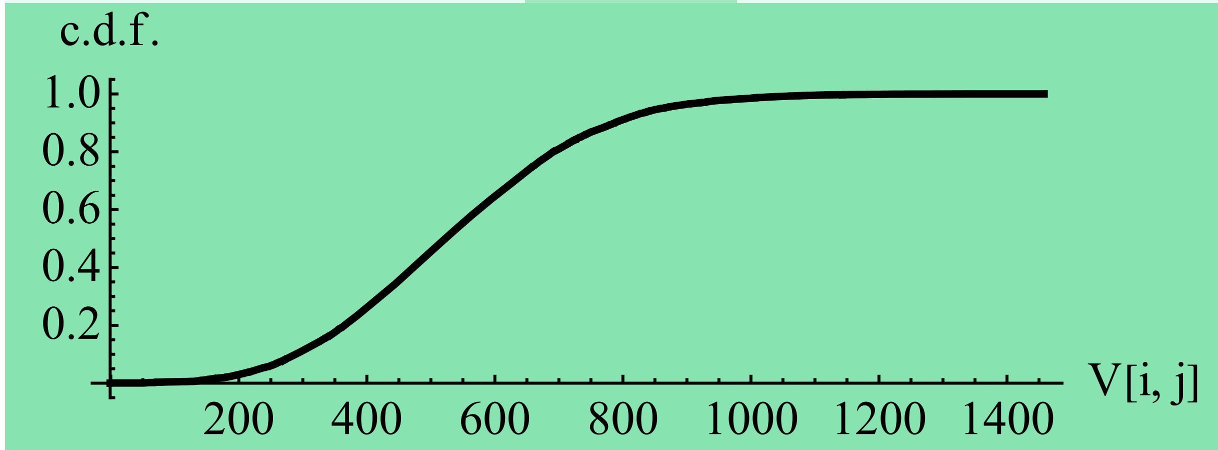


$$V[i,j] = \sum_{t=1}^{n} \underbrace{\mathbf{1}_{\{H_i(t)=j\}}}_{\text{Universal Hash Function}} \cdot a_t$$

$$\underbrace{V[i,j] = \sum_{t=1}^{n} Ber\left(\frac{1}{m}\right) \cdot a_t}_{}$$

$V[i,j]$ is a Sum of Independent Random Variables





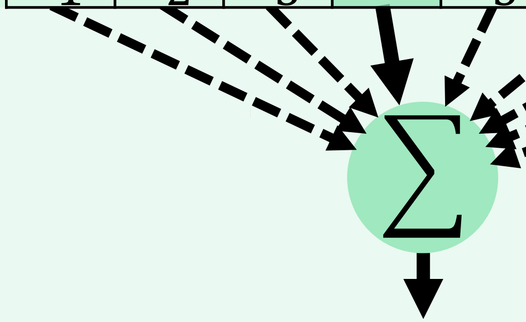
$V[i,j]$ converges to a Normal Distribution

New Sketch Algorithm

In General: $V[i,j] \sim \mathcal{N}(\mu, \sigma^2)$

What if we condition on $a_t = \alpha$?

Key	1	2	3	4	5	6	7	8	9
Vol	a_1	a_2	a_3	42	a_5	a_6	a_7	a_8	a_9



$V[i, H_i(4)] \mid a_4 = 42$

Then: $\underbrace{\overbrace{V[i, H_i(t)]}^{V_{i,t}} \mid a_t = \alpha}_{\mathcal{N}(\mu_t(\alpha), \sigma_t^2(\alpha))}$

Putting together the pieces:

$$\hat{a}_t = \underset{\alpha}{\operatorname{argmax}} \, \mathbb{P}[a_t = \alpha \mid D]$$

$$= \underset{\alpha}{\operatorname{argmax}} \left(\prod_{i=1}^{m} \mathcal{N}(V_{i,t}; \mu_t(\alpha), \sigma_t^2(\alpha)) \cdot \underbrace{\mathcal{N}(\alpha; \tilde{\mu}_t, \tilde{\sigma}_t^2)}_{\text{Prior}} \right)$$

Finding the MAP $\hat{a}_t$ is relatively straightforward:

$$\hat{a}_t = \frac{\tilde{\mu}_t \, \tilde{\chi}_t^{-2} + m \cdot \sum_{i=1}^{k} V[i, H_i(t)] - k \cdot \left(\sum_{j=1}^{n} a_j\right)}{\tilde{\chi}_t^{-2} + k \cdot (m-1)}$$

Where $\tilde{\chi}_t^2 = \tilde{\sigma}_t^2 \left(\|a\|_2^2 - \alpha^2\right)^{-1}$

Augmentation

How to improve $\hat{a}_t = \underset{\alpha}{\operatorname{argmax}}\, \mathbb{P}[a_t = \alpha |\, D]$?

Add more information to D !

$$\left.\begin{aligned} V[i,j] &= \sum_{t=1}^{n} \mathbf{1}_{\{H_i(t)=j\}} \cdot a_t \longleftarrow \text{``Volume“} \\ C[i,j] &= \sum_{t=1}^{n} \mathbf{1}_{\{H_i(t)=j\}} \longleftarrow \text{``Cardinality“} \end{aligned}\right\} D$$

$\|a\|_2^2 - {a_t}^2$ is in general not known

- MLE: $\tilde{\sigma}_t^2 = \infty \Longrightarrow \|a\|_2^2 - {a_t}^2$ not needed
- MAP:
  - Estimate $\|a\|_2^2 - {a_t}^2$ from model
  - Regress $\|a\|_2^2 - {a_t}^2$ from data

Again:

$$\hat{a}_t = \underset{\alpha}{\operatorname{argmax}}\, \mathbb{P}[a_t = \alpha |\, D]$$

Where $\tilde{\chi}_t^2 = \tilde{\sigma}_t^2\, (n-2)\left((n-1)\overbrace{(\|a\|_2^2 - \alpha^2)} - \sum_{j=1}^{n} a_j \right)^{-1}$

The MAP $\hat{a}_t$ is given by:

$$\hat{a}_t = \frac{\dfrac{\tilde{\mu}_t}{\tilde{\chi}_t^2} - k \cdot \left(\sum_{j=1}^{n} a_j\right) + (n-1) \cdot \sum_{i=1}^{k} \dfrac{V[i, H_i(t)]}{C[i, H_i(t)] - 1}}{{\tilde{\chi}_t}^{-2} + \sum_{i=1}^{k} \dfrac{n - C[i, H_i(t)]}{C[i, H_i(t)] - 1}}$$

ETH zürich

Benchmarks

Error Metrics:

- Average Relative Error: $Mean\left(\frac{|\hat{a}_t - a_t|}{a_t}\right)$
- Average Absolute Error: $Mean(|\hat{a}_t - a_t|)$

Proposed Sketches

$$\hat{a}_t^{CB} = \frac{\tilde{\mu}_t\ \tilde{\chi}_t^{-2} + m \cdot \sum_{i=1}^{k} V[i, H_i(t)] - k \cdot (\sum_{j=1}^{n} a_j)}{\tilde{\chi}_t^{-2} + k \cdot (m-1)}$$

Count-Bayes Sketch (CB)

$$\hat{a}_t^{CCB} = \frac{\frac{\tilde{\mu}_t}{\tilde{\chi}_t^2} - k(\sum_{j=1}^{n} a_j) + (n-1)\sum_{i=1}^{k} \frac{V[i, H_i(t)]}{C[i, H_i(t)] - 1}}{\tilde{\chi}_t^{-2} + \sum_{i=1}^{k} \frac{n - C[i, H_i(t)]}{C[i, H_i(t)] - 1}}$$

Cardinality-Count-Bayes Sketch (CCB)

Two versions of CB and CCB Sketch :

- Default: $\tilde{\chi}_t^2 = \infty \Longrightarrow$ no prior
- Trained: $\tilde{\chi}_t^2$ is regressed on a ``training“ trace

Performance compared to:

- Count-Min Sketch (CM) [1]
- Count Sketch (C) [2]
- Seq Sketch (Seq) [3]
- PR Sketch (PR) [4]

$$\hat{a} = \underset{a \mid V = M \cdot a}{\operatorname{argmin}} |a|$$

[1]: Cormode, G. & Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications
[2]: Moses, C. et. al. (2004). Finding frequent items in data streams
[3]: Huang, Q. et. al. (2021). Toward Nearly-Zero-Error Sketching via Compressive Sensing
[4]: Sheng, S. et. al. (2021). PR-sketch: monitoring per-key aggregation of streaming data with nearly full accuracy

Benchmarks

Experiments on various traces:

Trained on UNI1 Pt.1

Trained on first 10%
Evaluated on last 10%

Bloom filter for cardinality counting

UNI1 Pt.2 -

UNI1 Pt.5 -

UNI1 Pt.20 -

KOSARAK -

RETAIL -

Ferenc Bodon. 2010. UNIV1 dataset. https://pages.cs.wisc.edu/~tbenson/IMC10_Data.html.
Ferenc Bodon. 2003. KOSARAK dataset. http://fimi.uantwerpen.be/data/kosarak.dat.gz.
Tom Brijs. 2003. RETAIL dataset. http://fimi.uantwerpen.be/data/retail.dat.gz.

ETH zürich

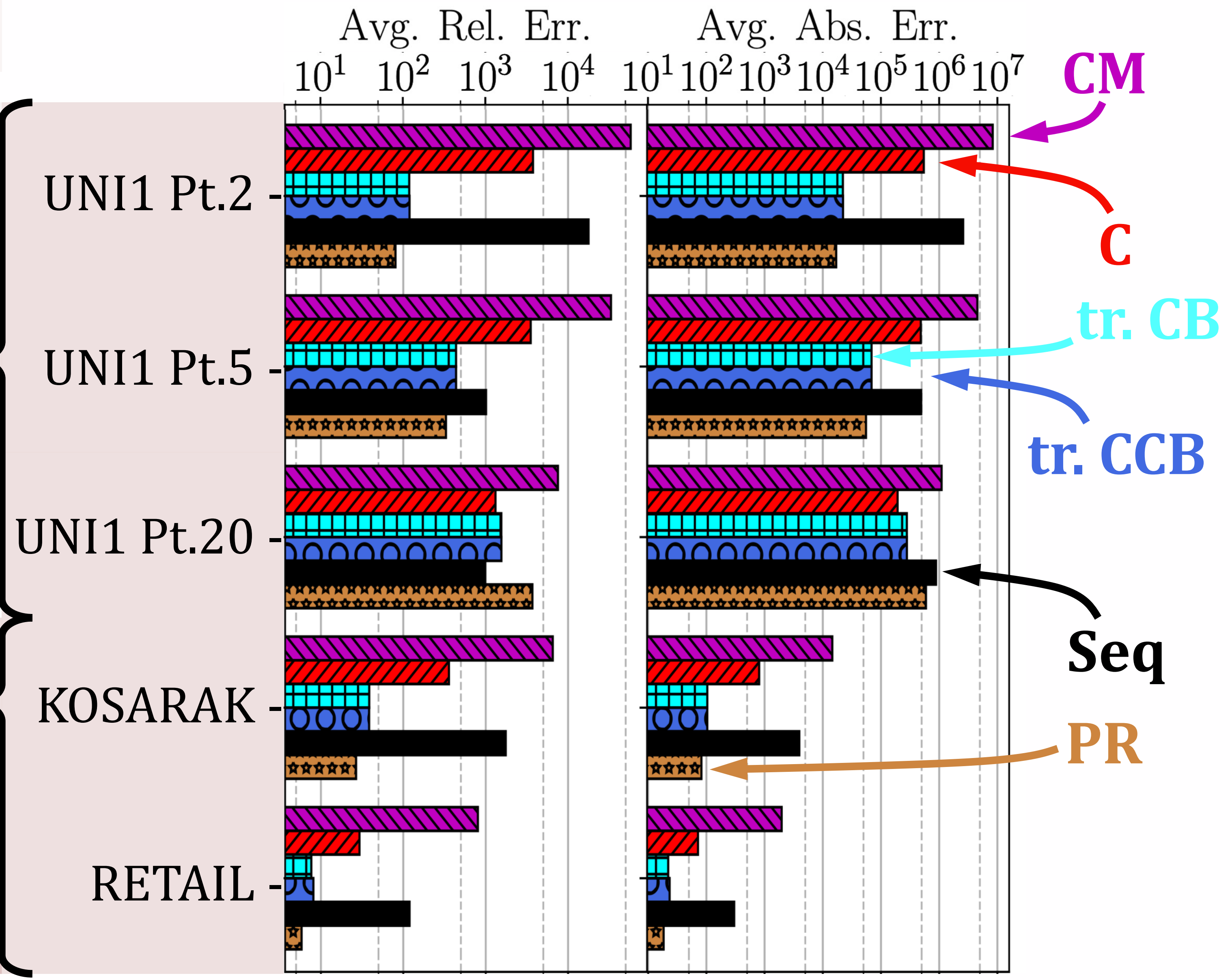
Benchmarks

Experiments on various traces:

Trained on UNI1 Pt.1

Trained on first 10%
Evaluated on last 10%

Bloom filter for cardinality counting

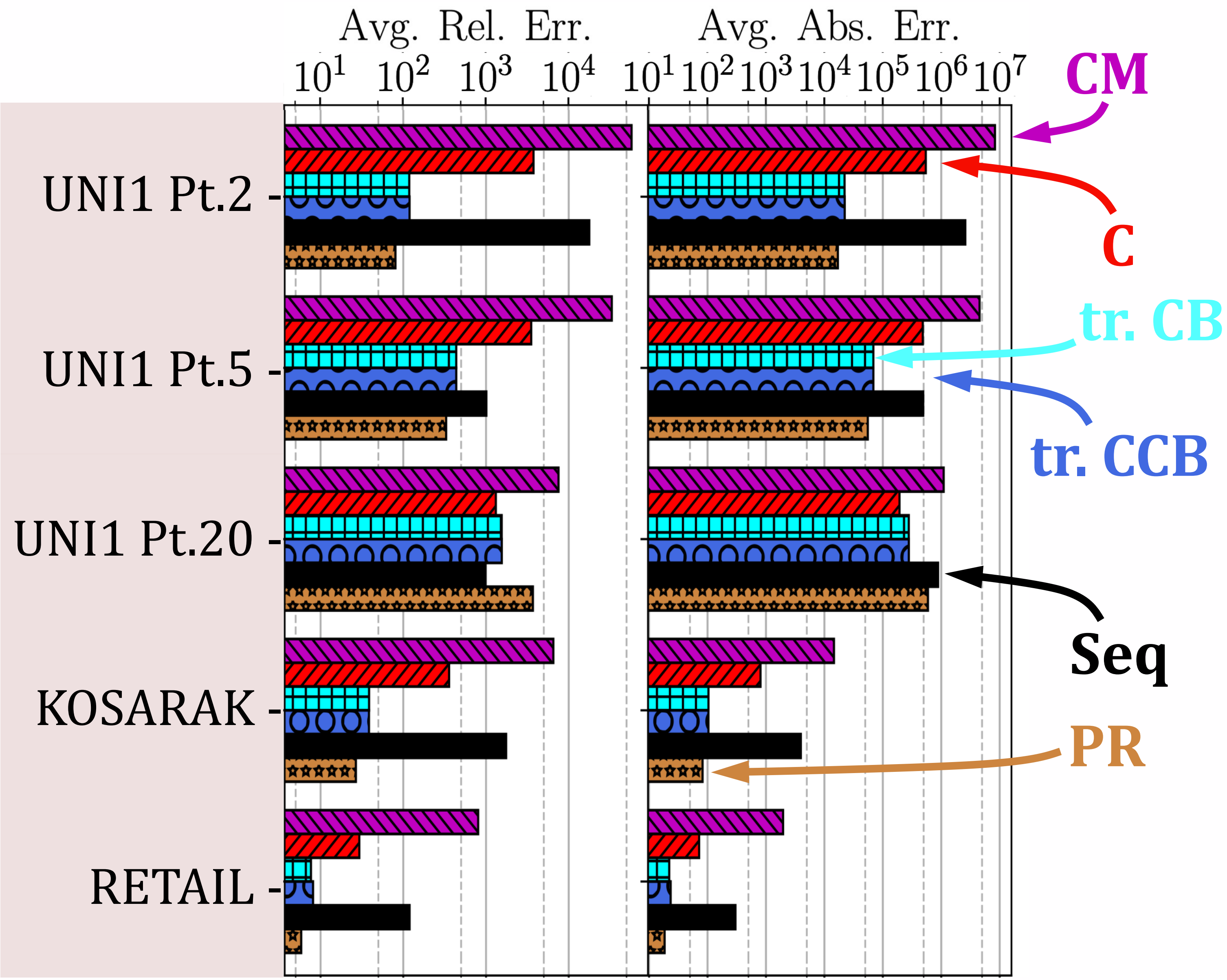


Ferenc Bodon. 2010. UNIV1 dataset. https://pages.cs.wisc.edu/~tbenson/IMC10_Data.html.

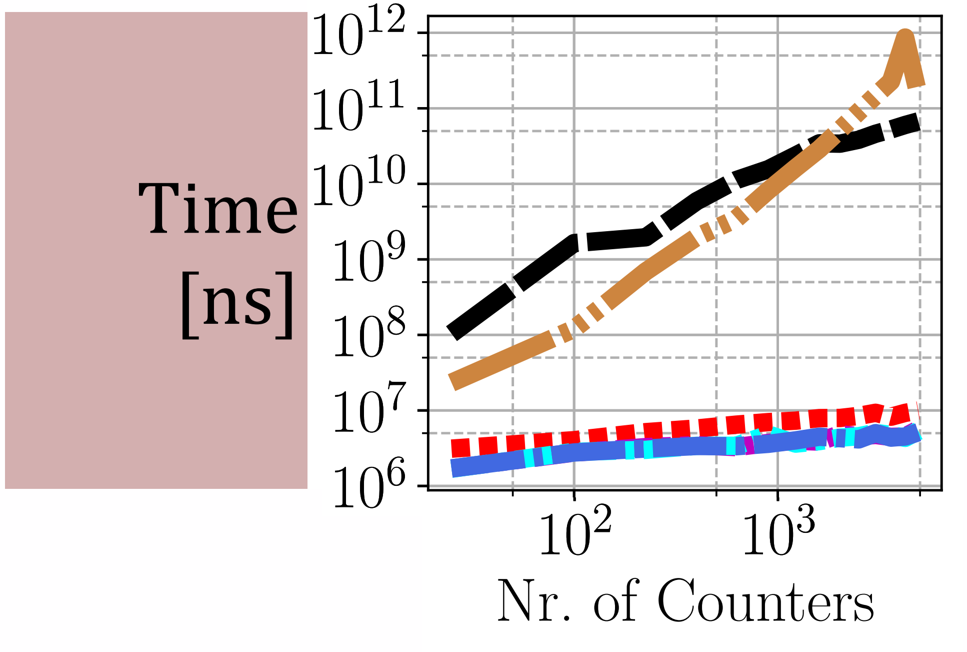
Ferenc Bodon. 2003. KOSARAK dataset. http://fimi.uantwerpen.be/data/kosarak.dat.gz.

Tom Brijs. 2003. RETAIL dataset. http://fimi.uantwerpen.be/data/retail.dat.gz.

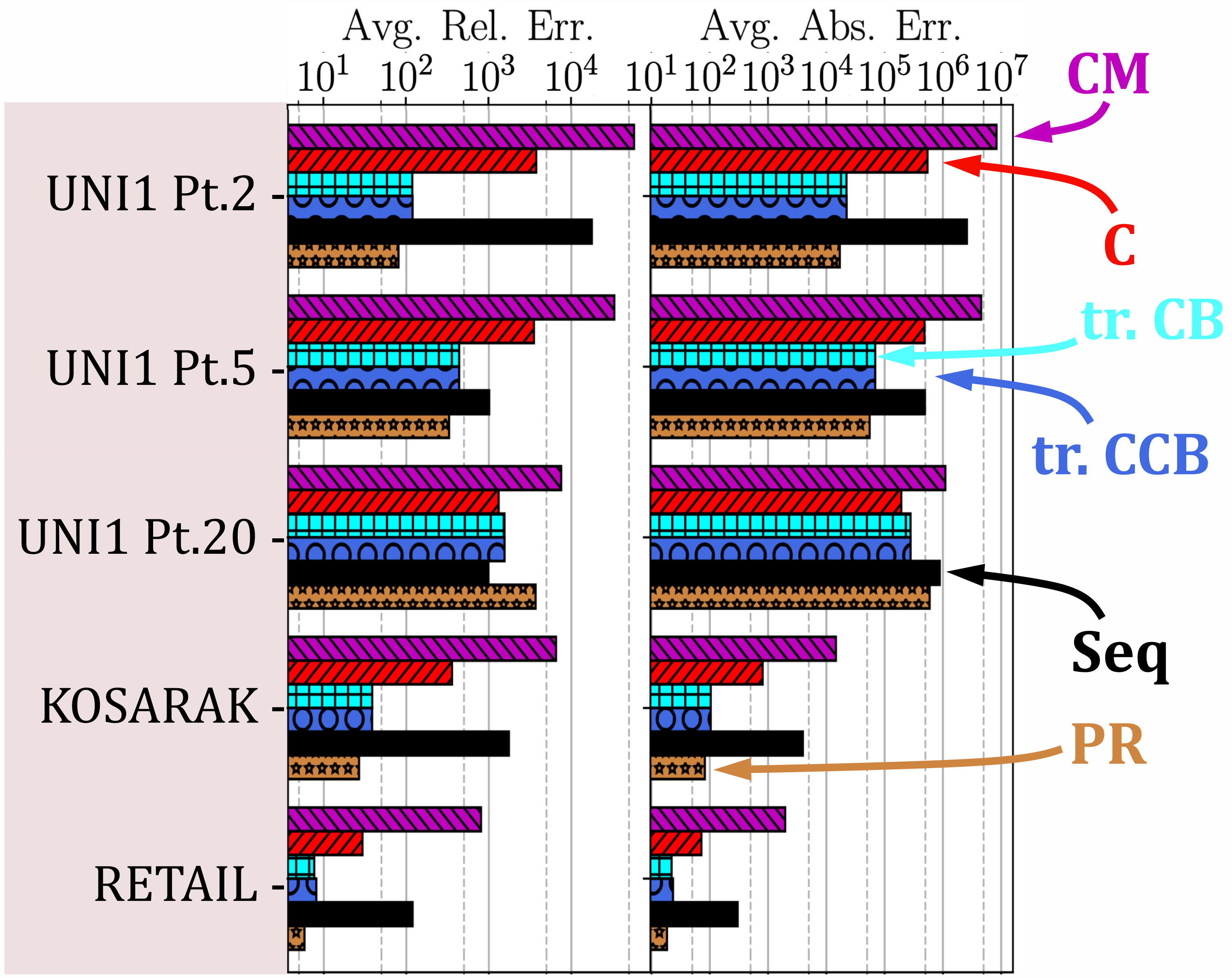
ETH zürich

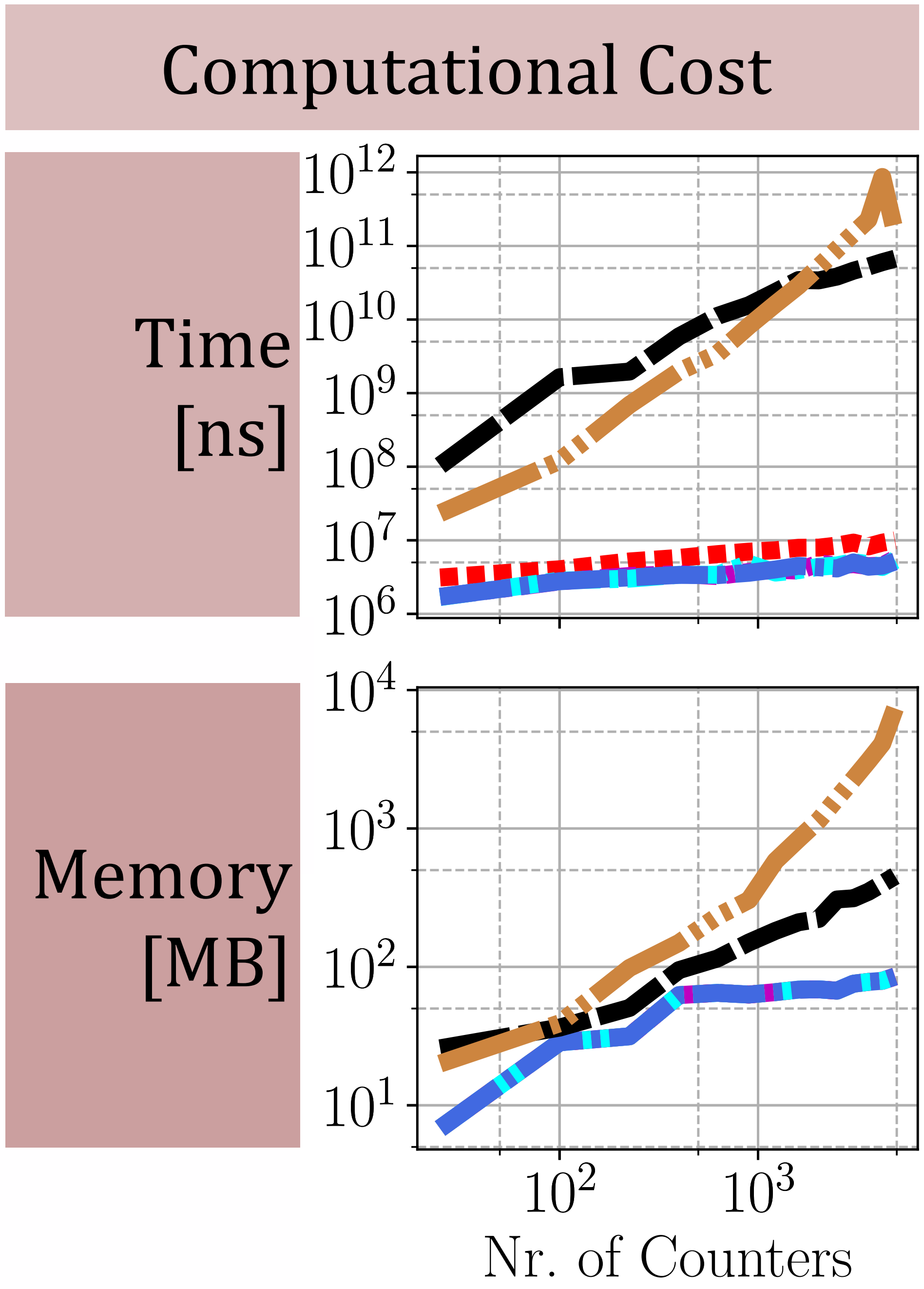


Computational Cost



Ferenc Bodon. 2010. UNIV1 dataset. https://pages.cs.wisc.edu/~tbenson/IMC10_Data.html.
Ferenc Bodon. 2003. KOSARAK dataset. http://fimi.uantwerpen.be/data/kosarak.dat.gz.
Tom Brijs. 2003. RETAIL dataset. http://fimi.uantwerpen.be/data/retail.dat.gz.



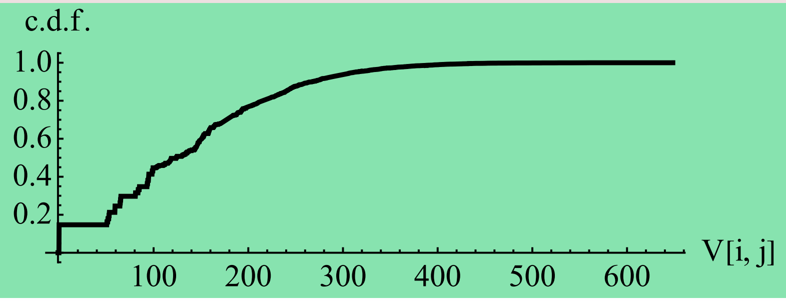


Ferenc Bodon. 2010. UNIV1 dataset. https://pages.cs.wisc.edu/~tbenson/IMC10_Data.html.
Ferenc Bodon. 2003. KOSARAK dataset. http://fimi.uantwerpen.be/data/kosarak.dat.gz.
Tom Brijs. 2003. RETAIL dataset. http://fimi.uantwerpen.be/data/retail.dat.gz.

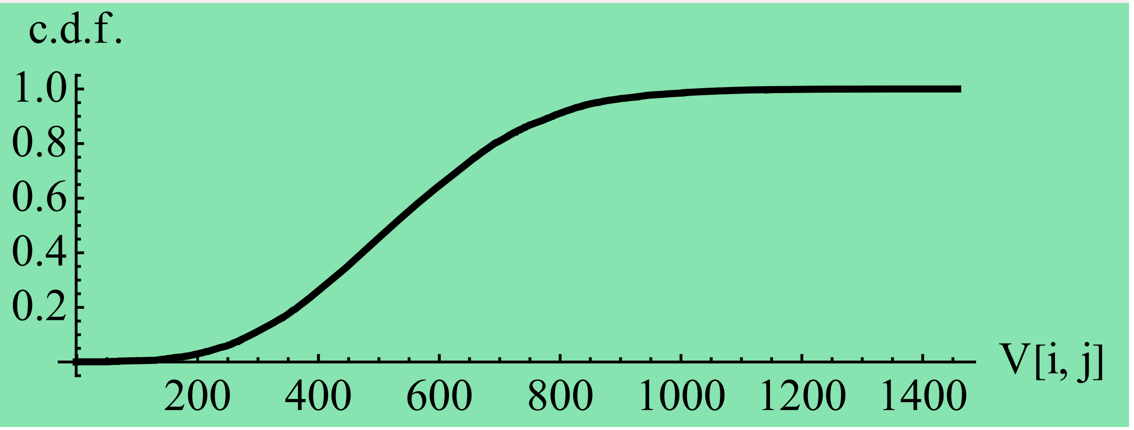
Recapitulation

New Perspective: $\mathbb{P}[D \mid a] \Longleftrightarrow argmax\, \mathbb{P}[D \mid \hat{a}] \cdot \mathbb{P}[\hat{a}]$

Random Hashing for Normal Model:

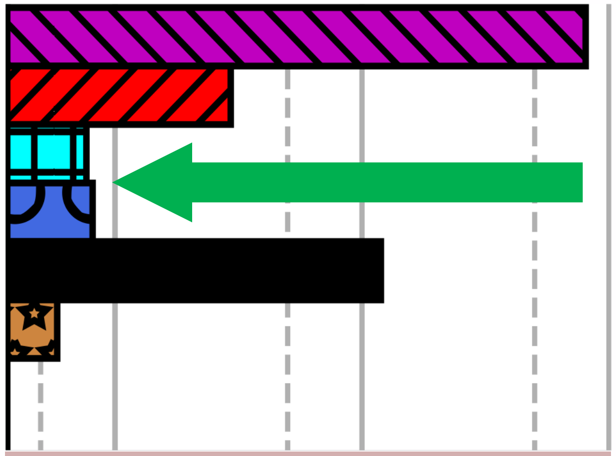


$\Longrightarrow$



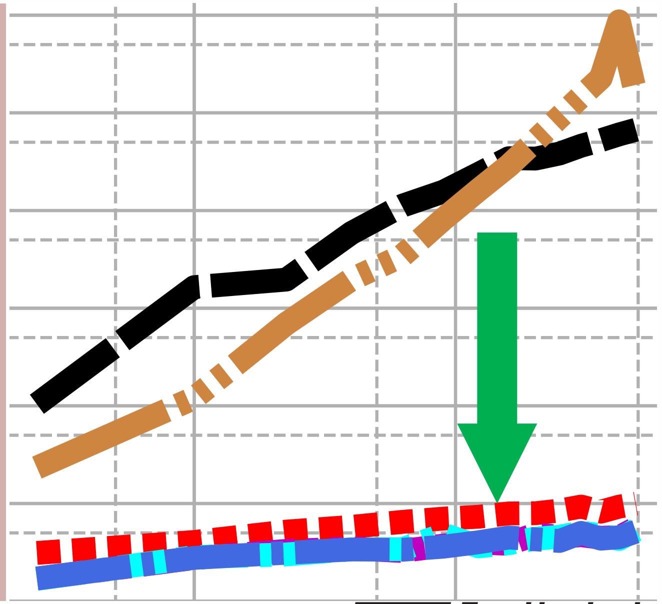
Two Novel MAP Estimates:

CB Sketch	Volume Information Only
CCB Sketch	With Cardinality Information



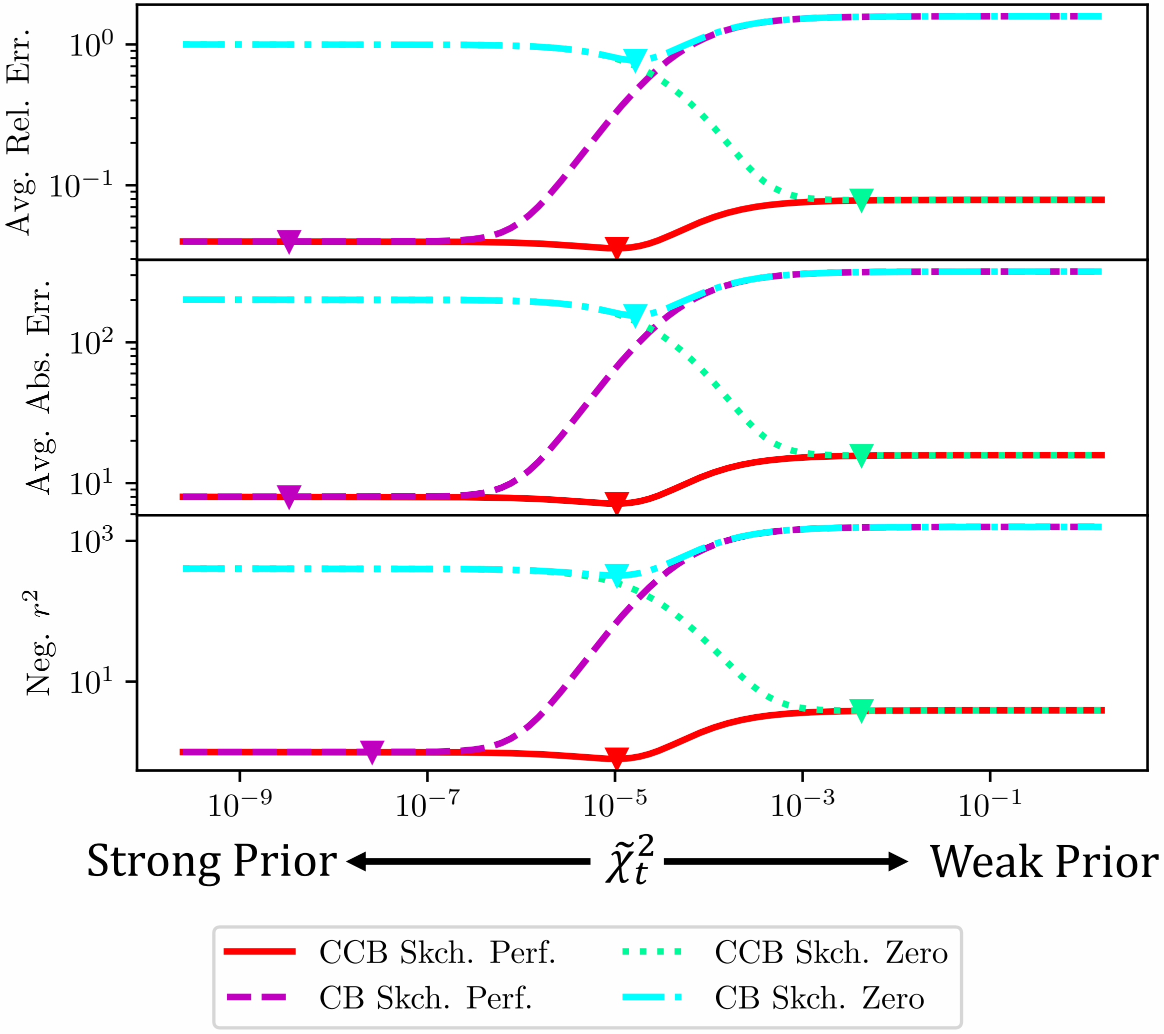
Error

Time



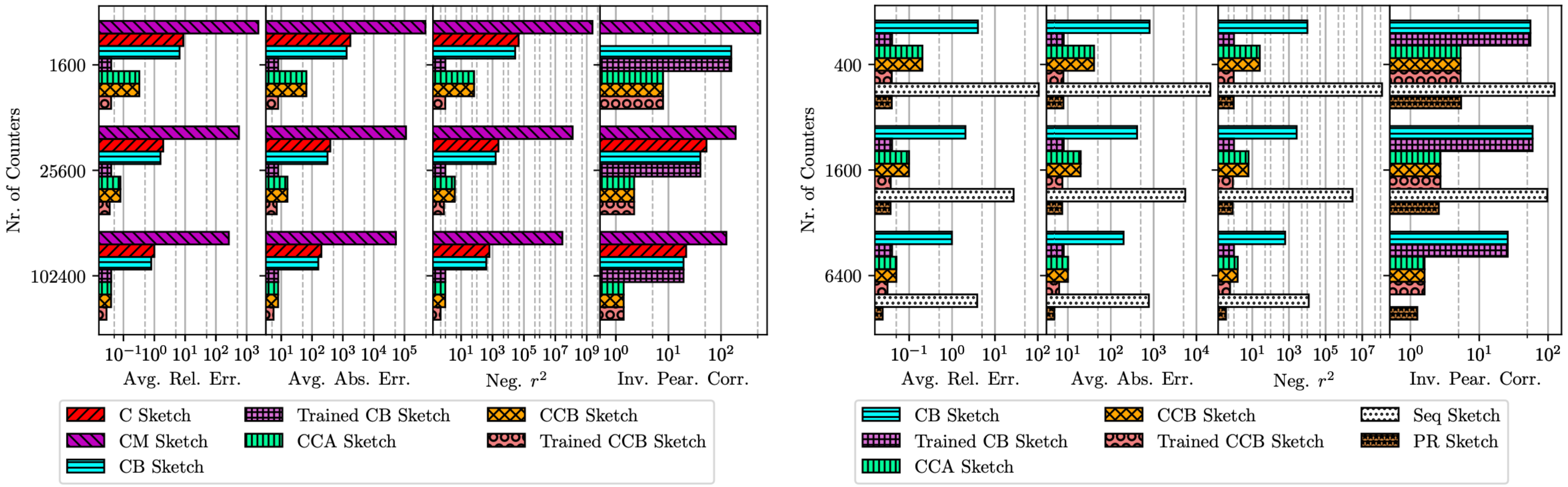
Benchmarks

The Effect of Prior Information



ETH zürich

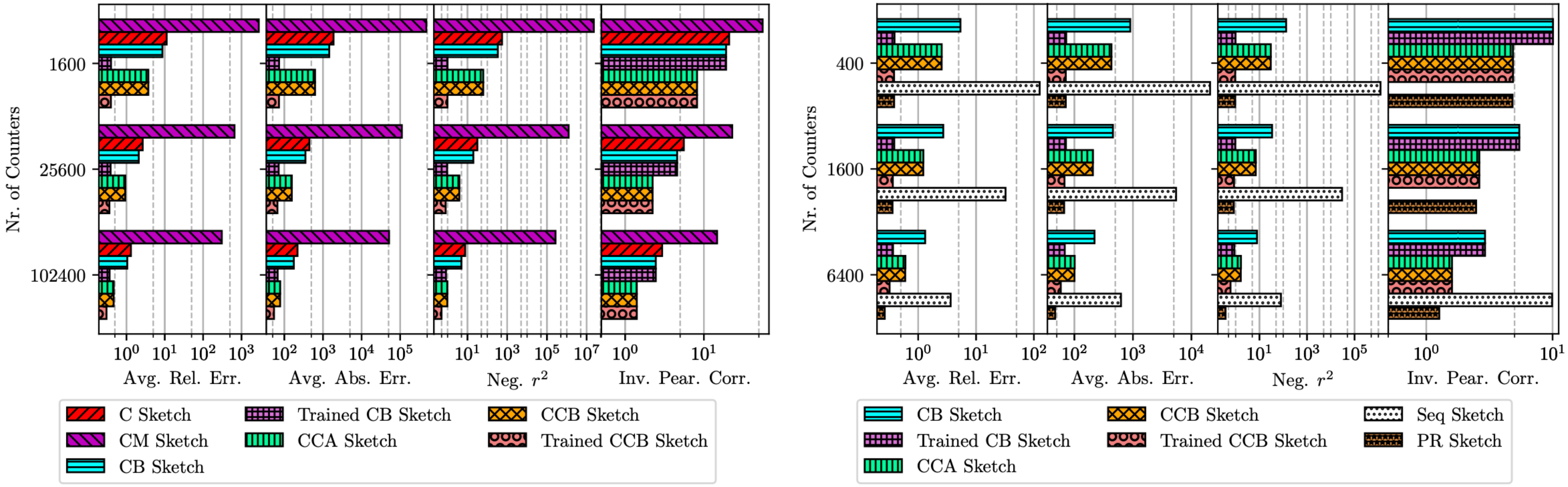
Benchmarks



(a) Comparison to local sketches (Trace of 100*k* keys). **(b) Comparison to global sketches (Trace of 10*k* keys).**

Figure 3: Comparison based on Poisson trace.

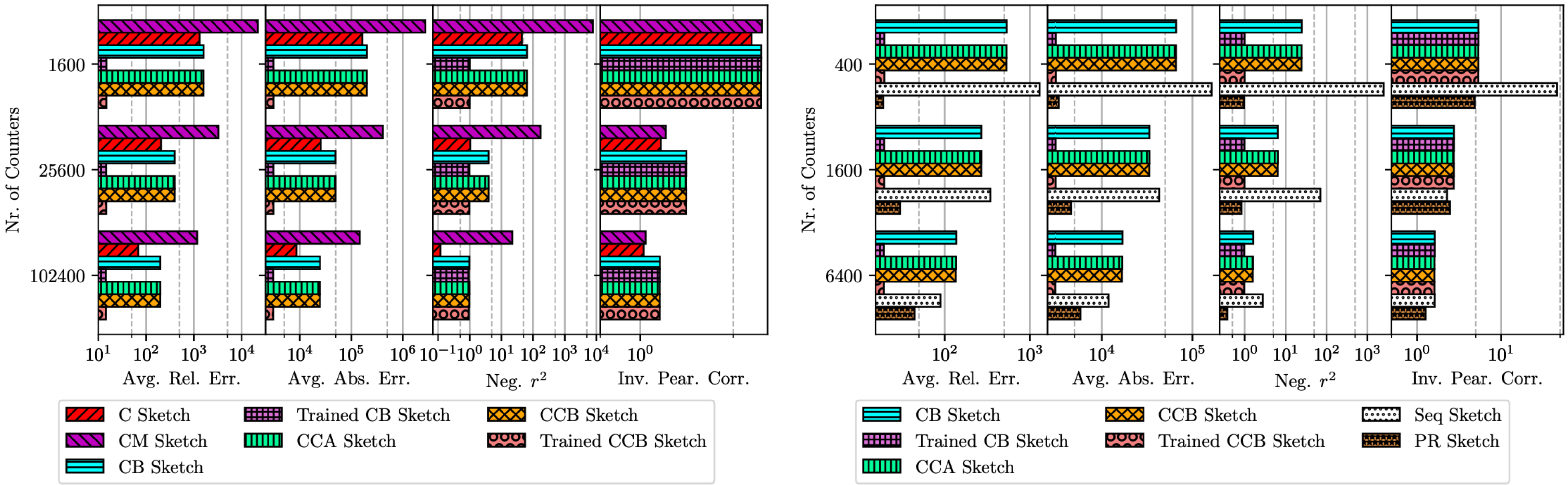
Benchmarks



(a) Comparison to local sketches (100k keys). **(b) Comparison to global sketches (10k keys).**

Figure 4: Comparison based on geometric trace.

Benchmarks



(a) Comparison to local sketches (Subsample of *100k* keys).

(b) Comparison to global sketches (Subsample of *10k* keys).

Figure 5: Comparison based on CAIDA trace.

Theory

THEOREM 1. *Let* a_f *the aggregate volume of key* $f \in \mathcal{I}$, *and* $\hat{a}_f^{CB}$ *be the estimate according to Equation 12 with a uniform prior* $\chi_p^{CB} = \infty$. *In the limit* $\ell_0 w^{-1} \to \infty$, *assuming that* $\max_{g \neq f} |a_g| \kappa_f \to 0$, *where*

$$\kappa_f = \sqrt{\left(\ell_2 - a_f^2\right)^{-1} (w-1)}$$

the accuracy of this estimate is bounded by:

$$\mathbb{P}\left[|\hat{a}_f^{CB} - a_f| \geq \epsilon \cdot \ell_1\right] \leq 2\exp\left(-\frac{\epsilon^2 \cdot \ell_1^2 \cdot d(w-1)}{2 \cdot \left(\ell_2 - a_f^2\right)}\right) \qquad (17)$$