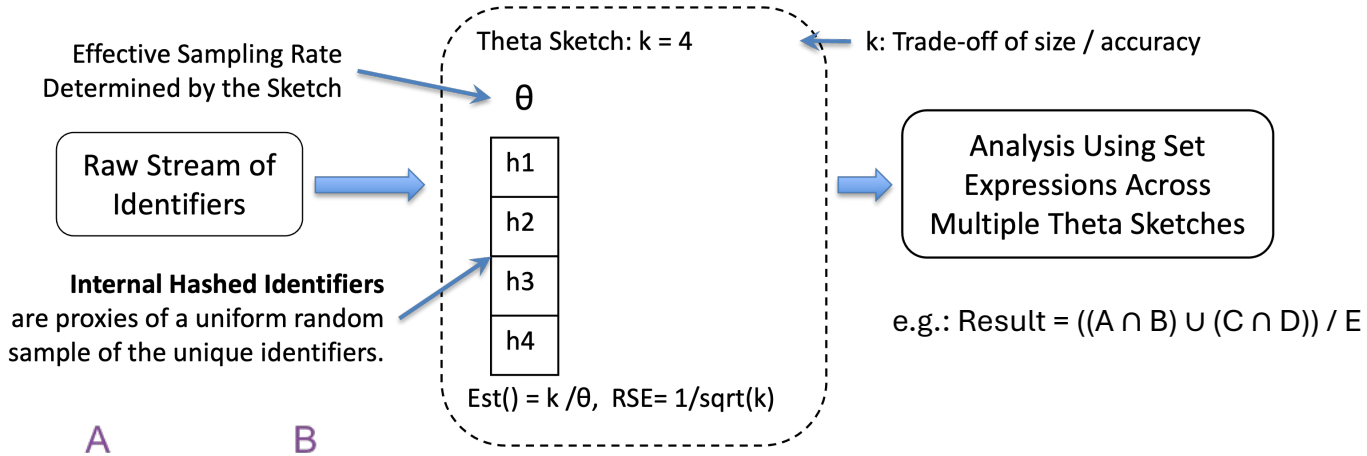
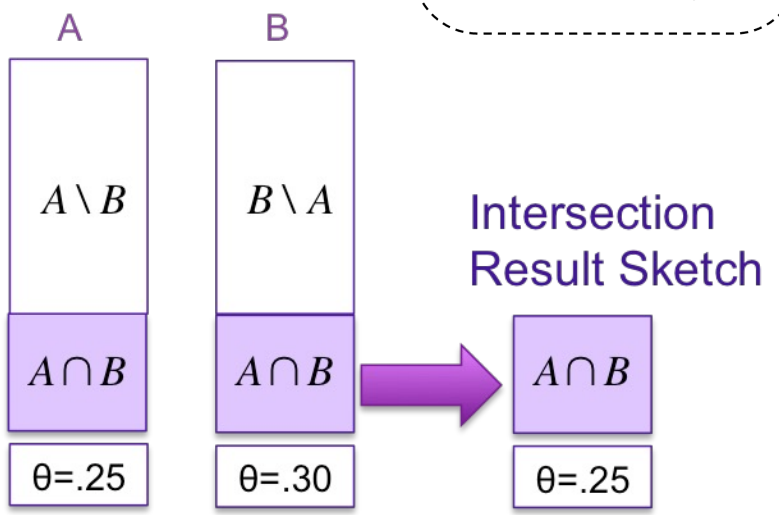
Outline

- Highlighting three powerful sketches
  - Theta Sketch
  - Tuple Sketch
  - Frequent Distinct Tuples (FDT) Sketch

Lee Rhodes
Apache DataSketches PMC Chair, ASF Member
leerho@apache.org
13 Jan 2026

The Theta Sketch With its Set Expressions

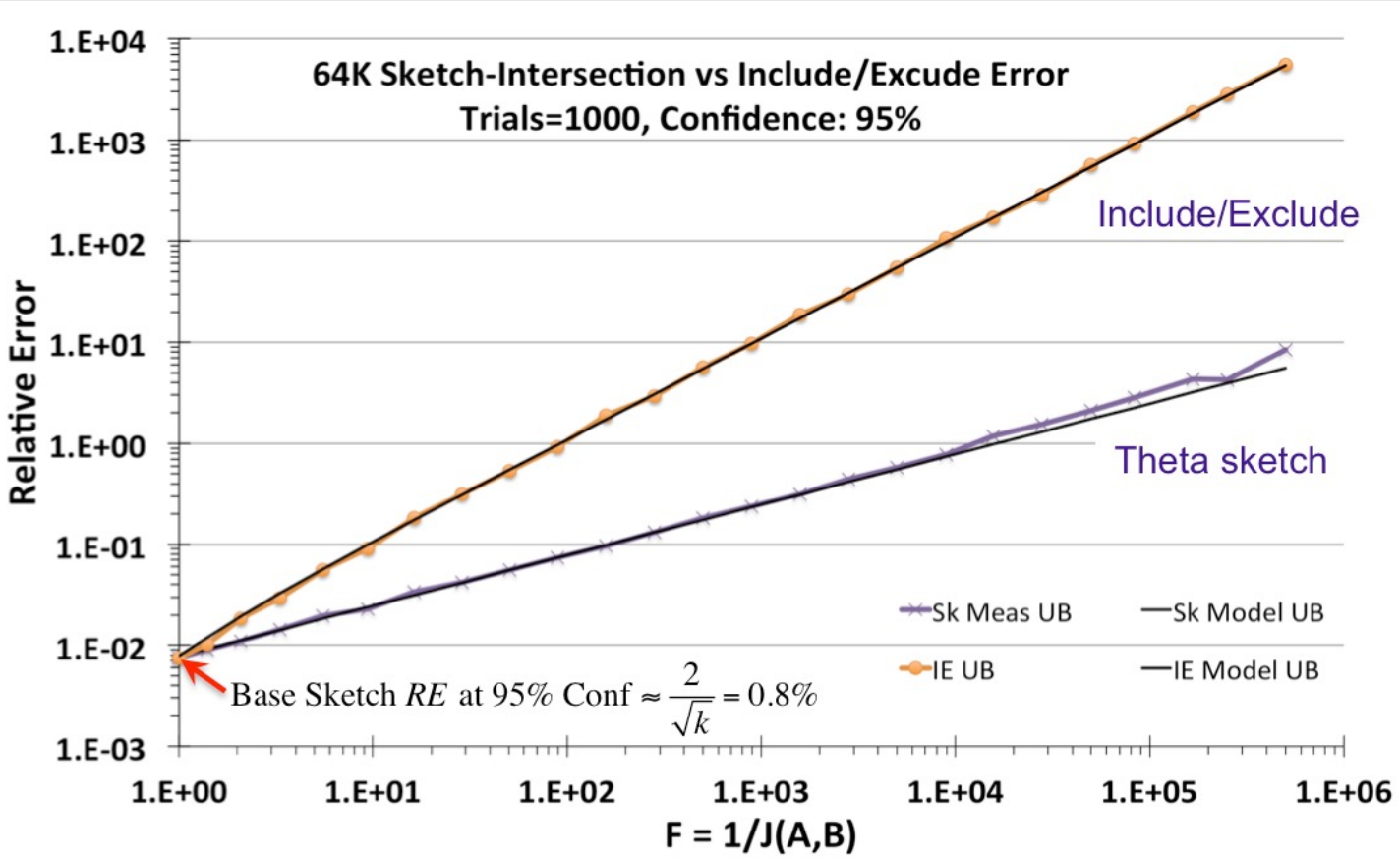




$$\Delta = \{\cup, \cap, \setminus\}; \qquad \theta_{A\Delta B} = \min(\theta_A, \theta_B); \qquad S_{A\Delta B} = \{x < \theta_{A\Delta B}; x \in (S_A \Delta S_B)\}$$

$$\text{est}(|A\Delta B|) = \frac{|S_{A\cup B}|}{\min(\theta_A, \theta_B)} \frac{|S_{A\Delta B}|}{|S_{A\cup B}|} = \frac{|S_{A\Delta B}|}{\min(\theta_A, \theta_B)}, \text{ Using "Broder Rule"}$$

Intersections have worse error, but intersection error of the Theta/Tuple Family can be orders-of-magnitude better than Include/Exclude methods.

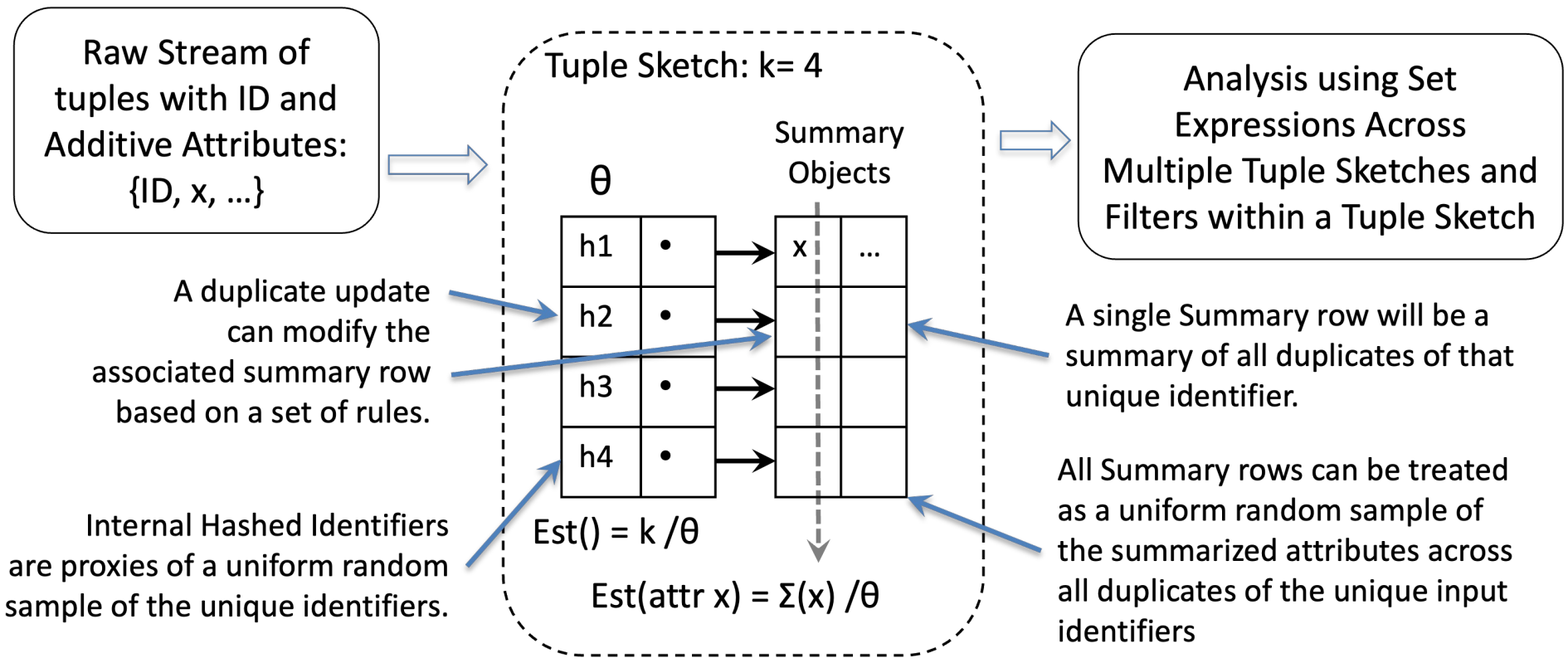


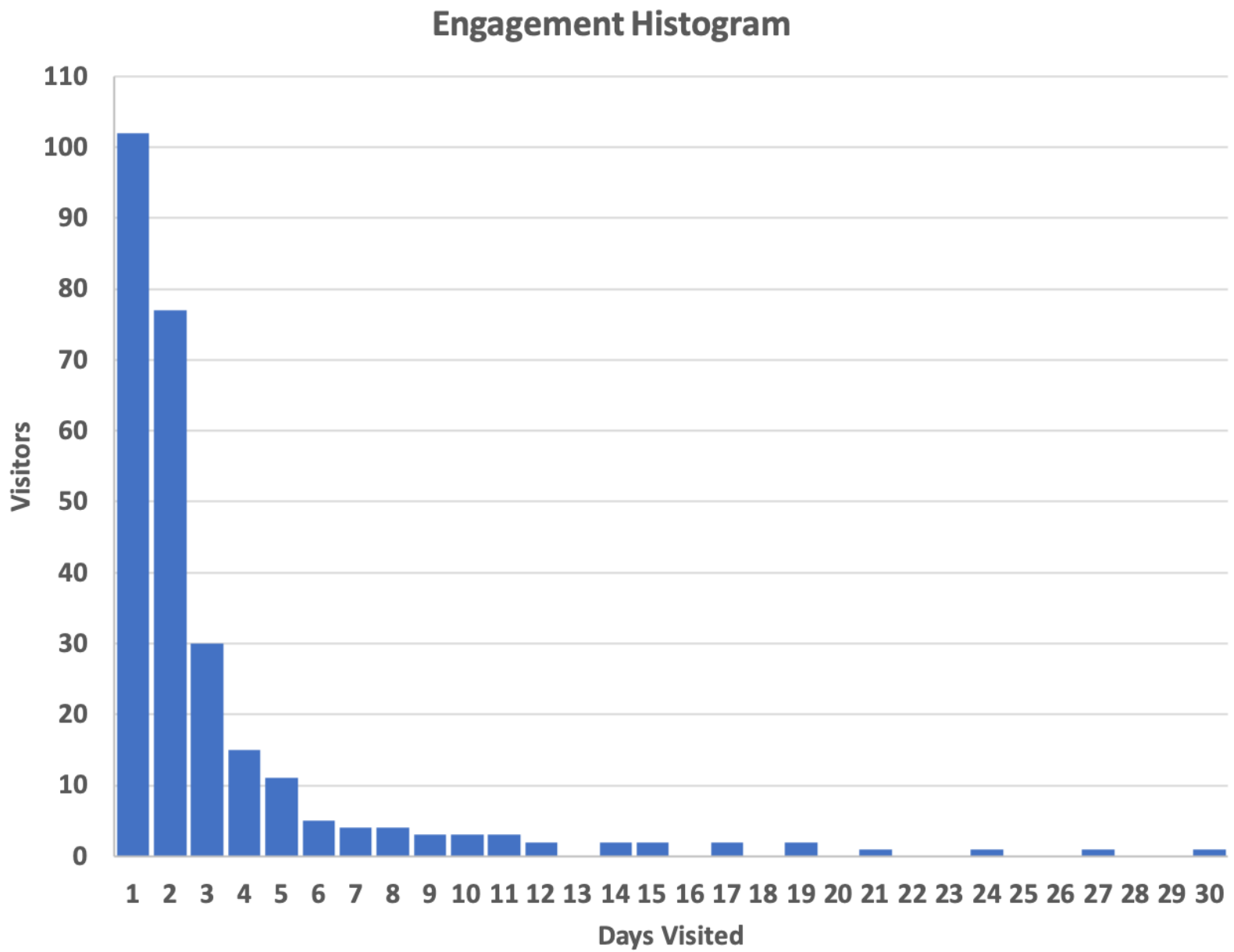
$RE_{A\cap B}$ = Relative Error for Theta Sketch Intersection $\approx \sqrt{F}\frac{1}{\sqrt{k}}$

RE_{IE} = Relative Error for Include / Exclude formula $\approx F\frac{1}{\sqrt{k}}$

$RE = \frac{Measured}{Truth} - 1 \qquad F = \frac{A \cup B}{A \cap B} = \frac{1}{J(A,B)} \qquad \frac{RE_{IE}}{RE_{A\cap B}} = \sqrt{F}$

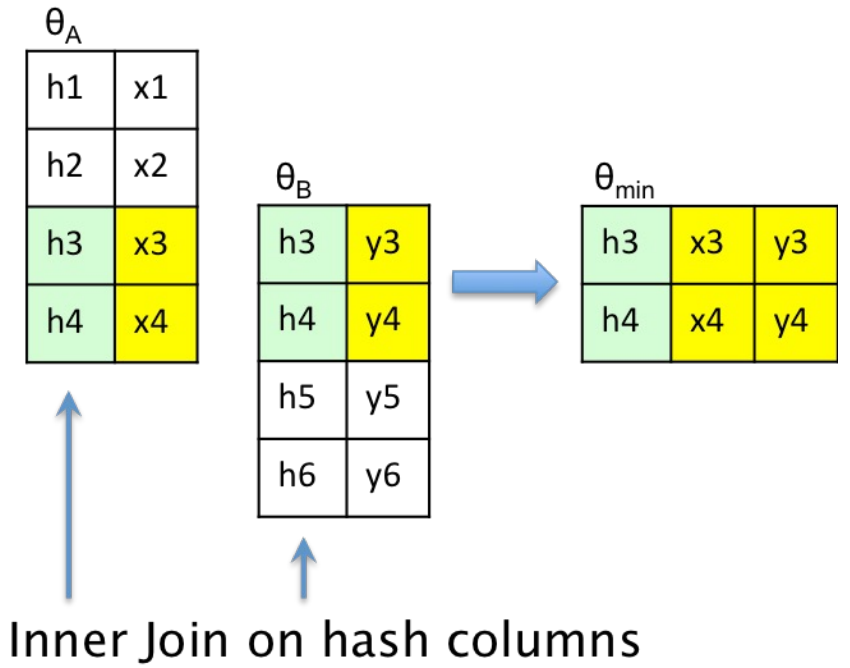
The Tuple Sketch adds an Associative Table of Data




Engagement Histogram
Visitors
110
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Days Visited

The Tuple Sketch can be used:

- To speed up and simplify certain complex analysis in a single pass
- It also includes the same Set Expression capabilities of Theta



The Frequent Distinct Tuples (FDT Sketch) Builds on the Tuple Sketch

A two-dimensional example:

Task: We want to identify the IP addresses that have the most distinct User IDs. Or conversely, we would like to identify the User IDs that have the most distinct IP addresses.

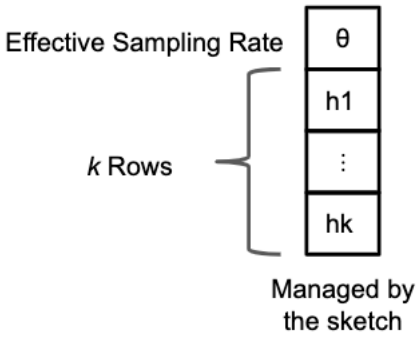
- Create a Tuple Sketch with a summary row: {IP, UID} using the appropriate types.
- Create a "plug-in" class that defines the rules for set operations and duplicates.
- Load the sketch with tuples of the form {key=<IP, UID>, IP, UID}.
- Load the summary rows into a DB table.
- For IP addresses that have the most distinct User IDs, sort the table by IP, #UIDs/IP.
- For UIDs that have the most IPs: sort the table by UIDs, #IPs/UID.
- Estimate of size of each group in the input data set just divide by *theta*.

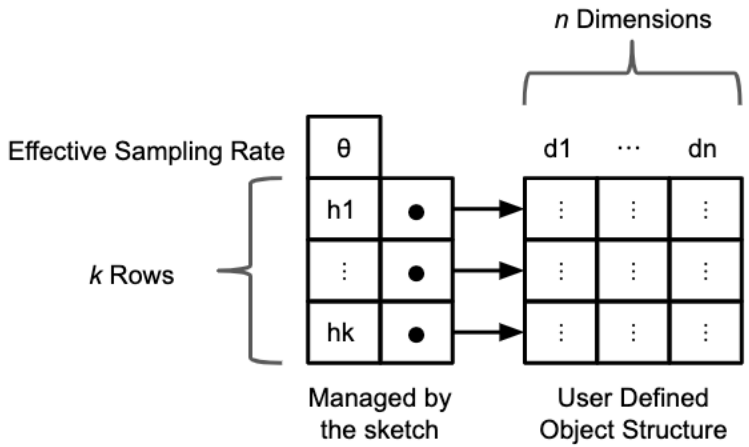
This is a multidimensional sketch!

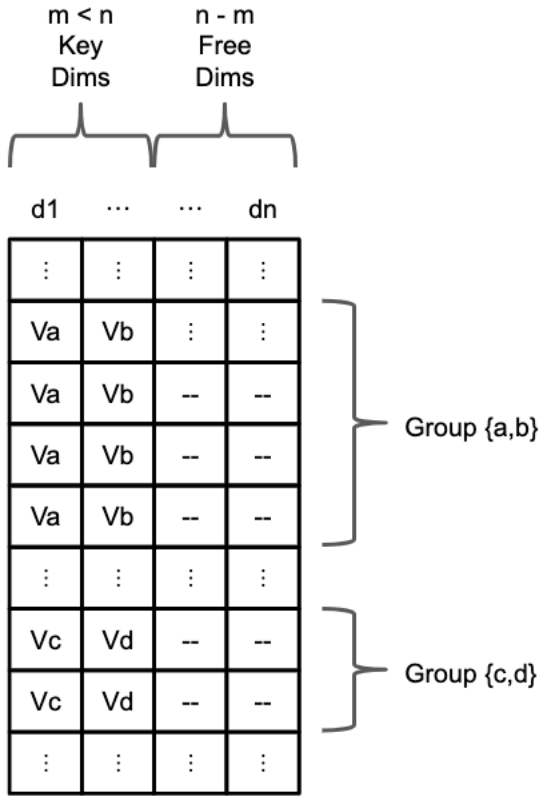
- More generally, given a multiset of tuples with *N* dimensions *{d1,d2, d3, …, dN}*
- You load the sketch by feeding it tuples of the form *{Key, v1,v2,v3,…, vN}*.
- You then choose a primary subset of dimensions $M < N$.
- Our task is to identify the combinations of *M* subset dimensions that have the most frequent number of distinct combinations of the *N-M* non-primary dimensions.

- The input data only needs to be scanned once! Then you can query the sketch multiple times with different combinations of the subset M dimensions.

Evolution in Abstraction & Power:
Theta -> Tuple -> FDT



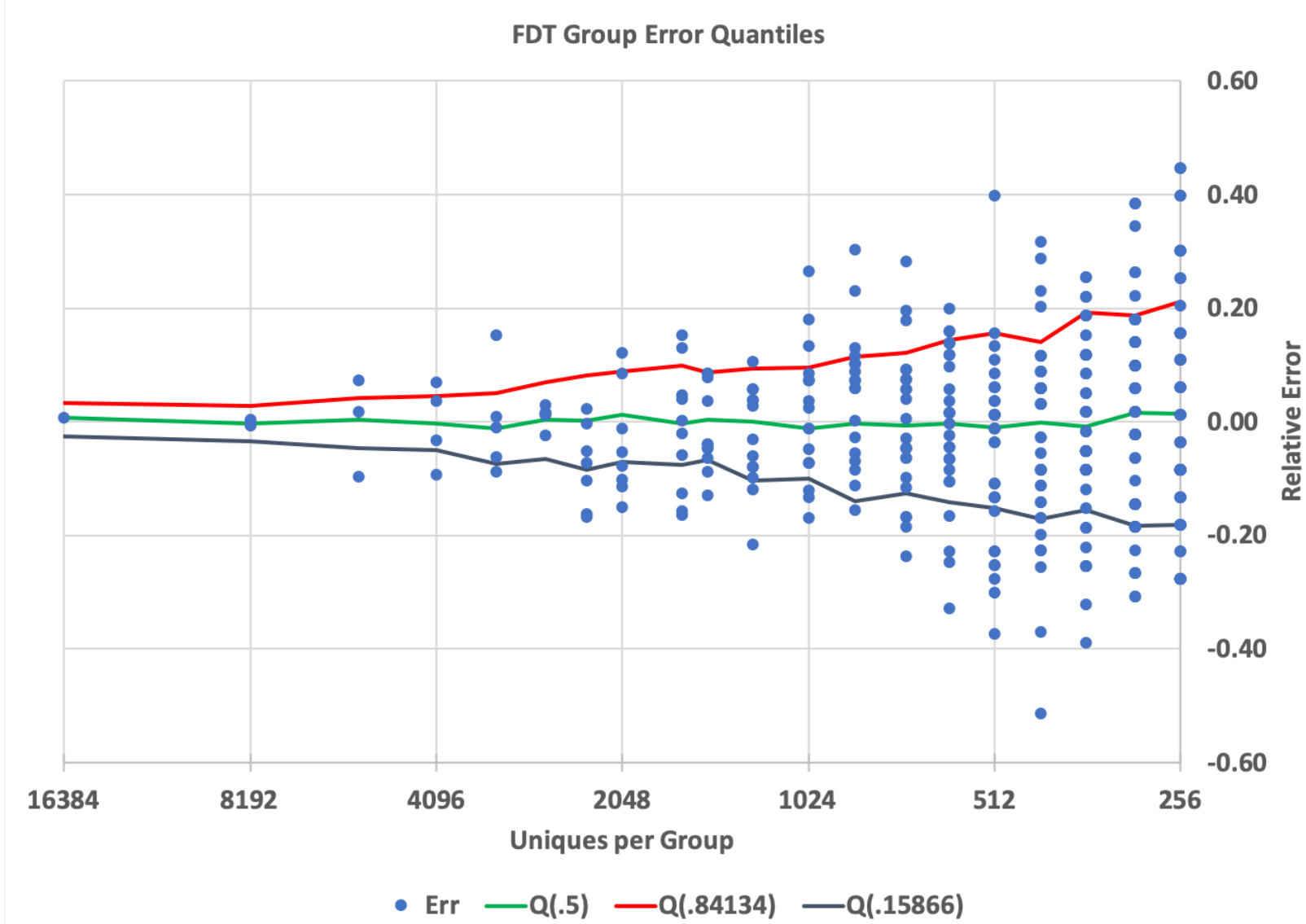




Theta

Tuple

FDT


FDT Group Error Quantiles
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
Relative Error
16384
8192
4096
2048
1024
512
256
Uniques per Group
Err
Q(.5)
Q(.84134)
Q(.15866)

Thank You!

Open Invitation for Collaboration

https://datasketches.apache.org