A General Method for Estimating Correlated Aggregates Over a Data Stream

Srikanta Tirthapura
Iowa State University
snt@iastate.edu

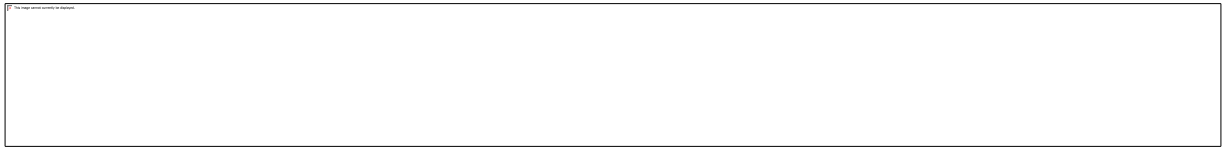
David Woodruff
IBM Research Almaden
dpwoodru@us.ibm.com

Setting: Where Would this be Useful?

- Analytics on Large Streams
  - Ex: Stream of IP Flow Records
- On-line Compression or “Sketching” of data
- Queries Encompass more than one dimension

What is a Correlated Aggregate?

Stream S = (x_1,y_1), (x_2,y_2), (x_3,y_3)……



σ = selection predicate, e.g. "σ(y) = (y < 5)"

f = an aggregation operator, say "SUM"

Problem Statement

Design a "sketch" for computing correlated aggregates on S, for various f and σ

1. Sketch size much smaller than stream size
2. Approximate answers, guarantees on accuracy
3. Sketch maintained in a single pass
4. Predicate σ not fully specified at time of observation

Why Correlated Aggregates?

Network Admin Querying a Stream of IP Flow Records

1. Median size of packet flow?
   - Answered by a Quantile Summary

2. Number of Distinct Source IPs among flows whose size greater than median flow size?

3. Number of Distinct Source IPs among flows whose size greater than 10*(median flow size)?

The Sketch allows to focus on “interesting” regions, where what is “interesting” not known during observation

Effect of Selection Predicate σ

Model assumed on σ,
parameters specified
at query time

σ completely specified
at time of stream
observation

Reduces to traditional
aggregate computation
on a stream

Nothing known
about σ till query time

Impossible, in
small space

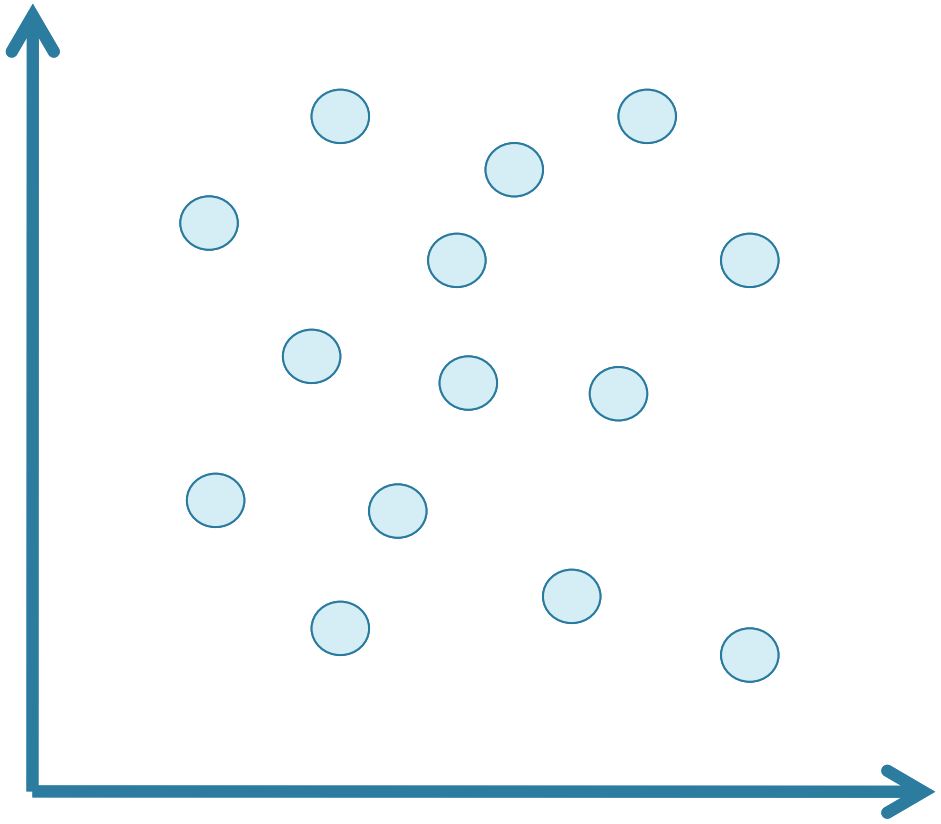
Our Model for σ

$\sigma(y) = (y \geq c)$

OR

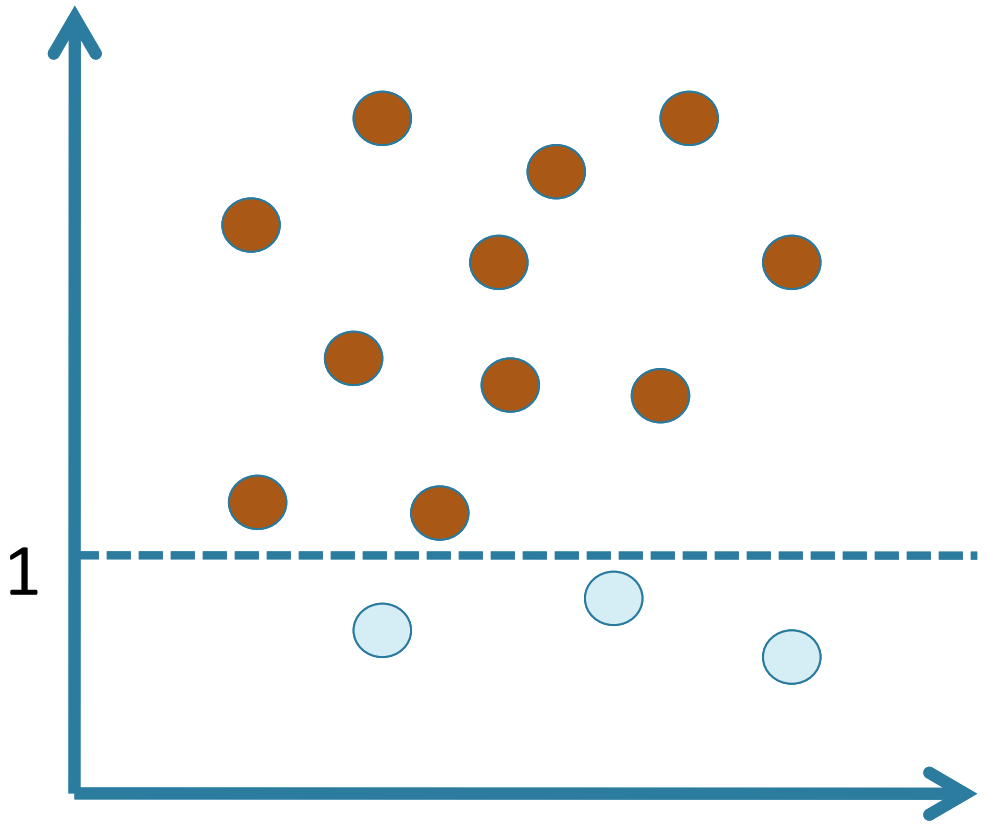
$\sigma(y) = (y \leq c)$

where c is provided at query time



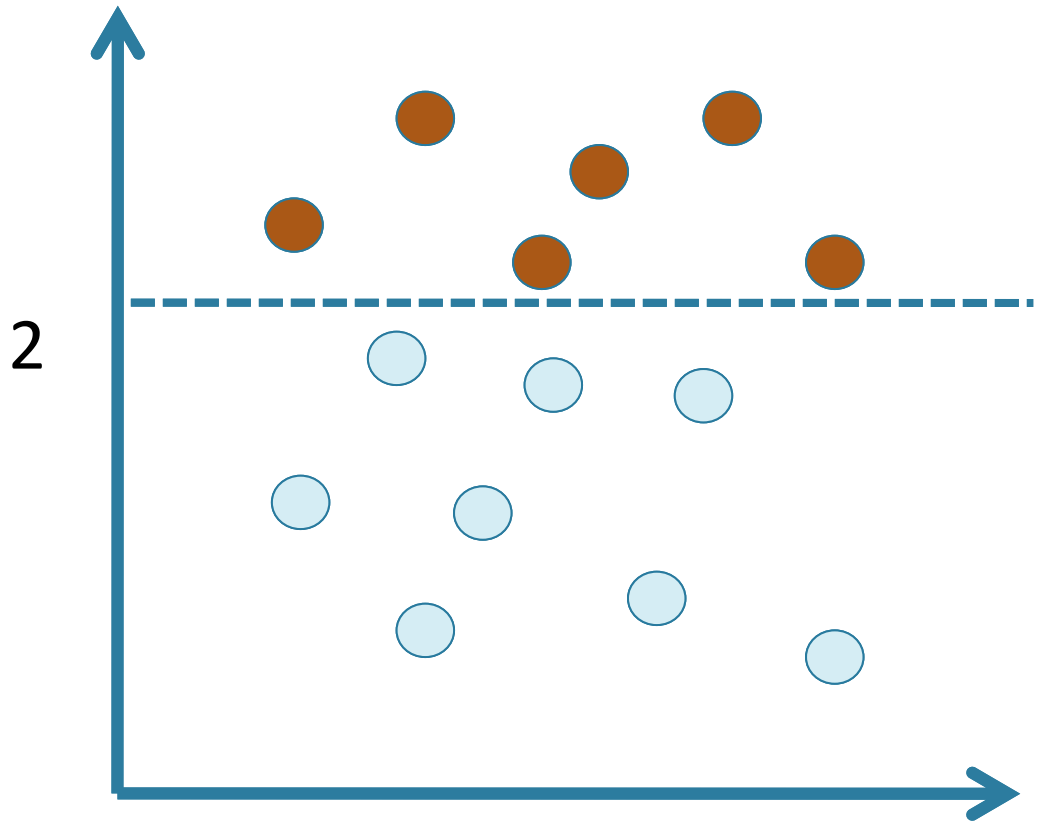
σ

y > 1

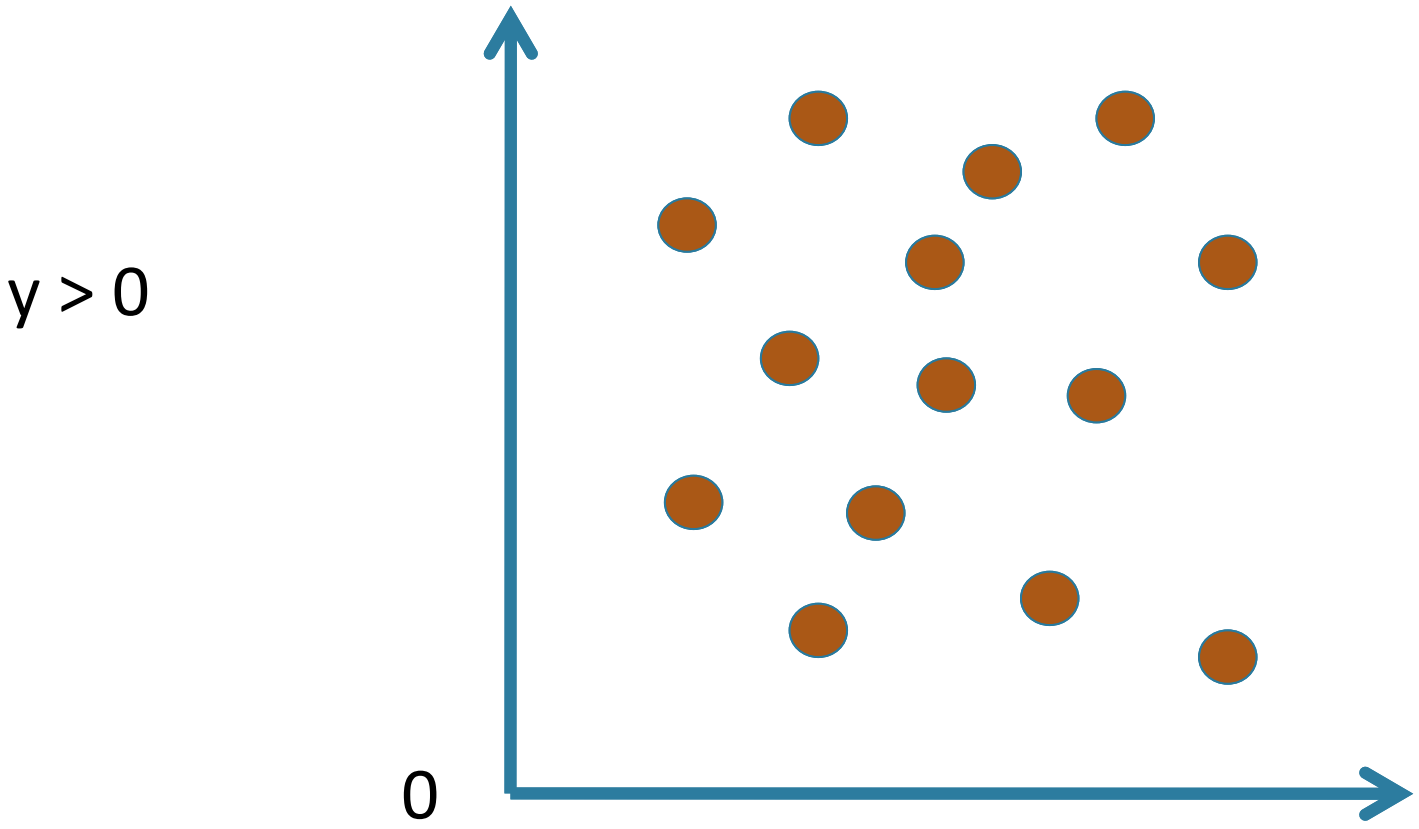


σ

y > 2

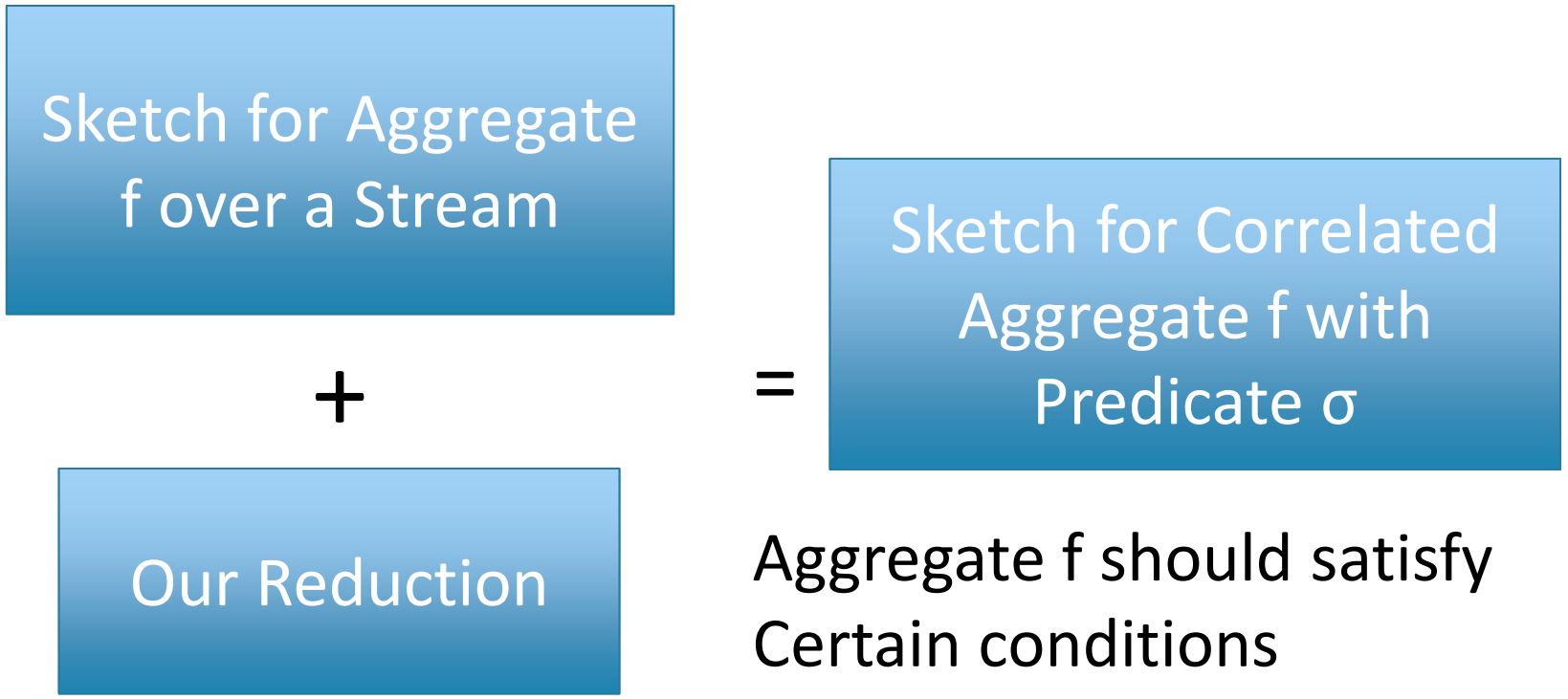


σ



Our Contributions (1)

A General Method for a Sketch for Correlated Aggregates



Aggregate f should satisfy Certain conditions

Our Contributions (2)

- First Small Space Algorithms for Estimating Correlated Frequency Moments F_k ($k \geq 0$)
  - In a multi-set of, let n_i denote the frequency of item i
    $F_k = \Sigma_i (n_i)^k$
- Memory Lower Bounds for Correlated Function Aggregation with Negative Weight Elements
- Experimental Results on F_0 (number of distinct elements), and F_2

Previous and Related Work

- Gehrke, Korn, Srivastava (SIGMOD 2001), "On Computing Correlated Aggregates over Continual Data Streams" – heuristics for correlated aggregate estimation
- Ananthakrishna et al. (TKDE 2003) – Algorithm for correlated sum with additive error guarantee
- Busch and Tirthapura (STACS 2007) – Algorithm for sum with relative error guarantee (distributed streams)
- Cormode, Korn, Tirthapura (PODS 2008) – Algorithm for Correlated Frequent Elements, Improved by Chan et al. (2009)
- Datar, Gionis, Indyk Motwani (2002): Reduction from sliding window computation to computation over infinite window

Previous and Related Work

- Estimating Aggregates on a Data Stream
- Aggregates over a Sliding Window
  - Asynchronous arrivals

Conditions on Aggregate Function f

1. f(R) bounded by polynomial in $|R|$

2. For sets R_1 and R_2, $f(R_1 \cup R_2) \geq f(R_1) + f(R_2)$

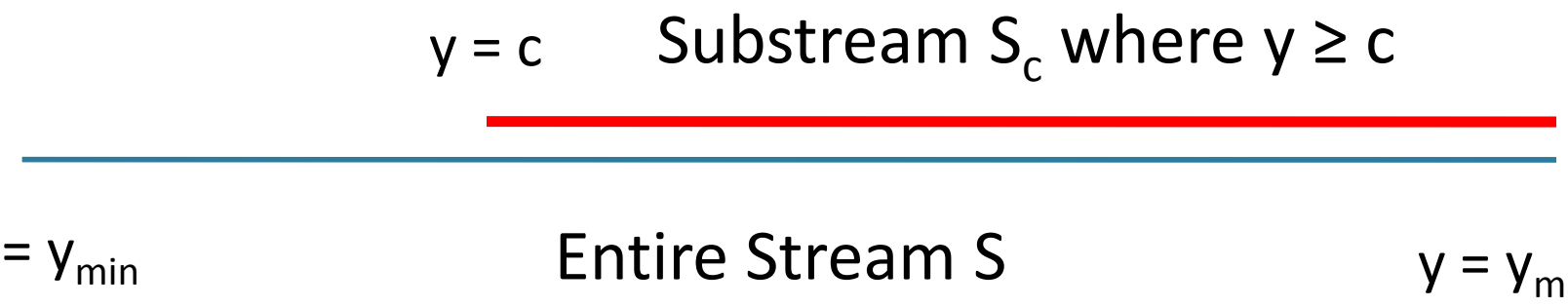
3. Smoothness 1: There exists a function c_1^f such that for sets R_1, R_2, ….. R_j , if $f(R_i) \leq \alpha$ for all i, then $f(R_1 \cup R_2 \cup \ldots..) \leq c_1^f(j).\ \alpha$

4. Smoothness 2: For $\varepsilon < 1$, there is a function c_2^f such that for two sets A and B, B subset of A, if $f(B) \leq c_2^f(\varepsilon).\ f(A)$, then $f(A-B) \geq (1-\varepsilon)\ f(A)$

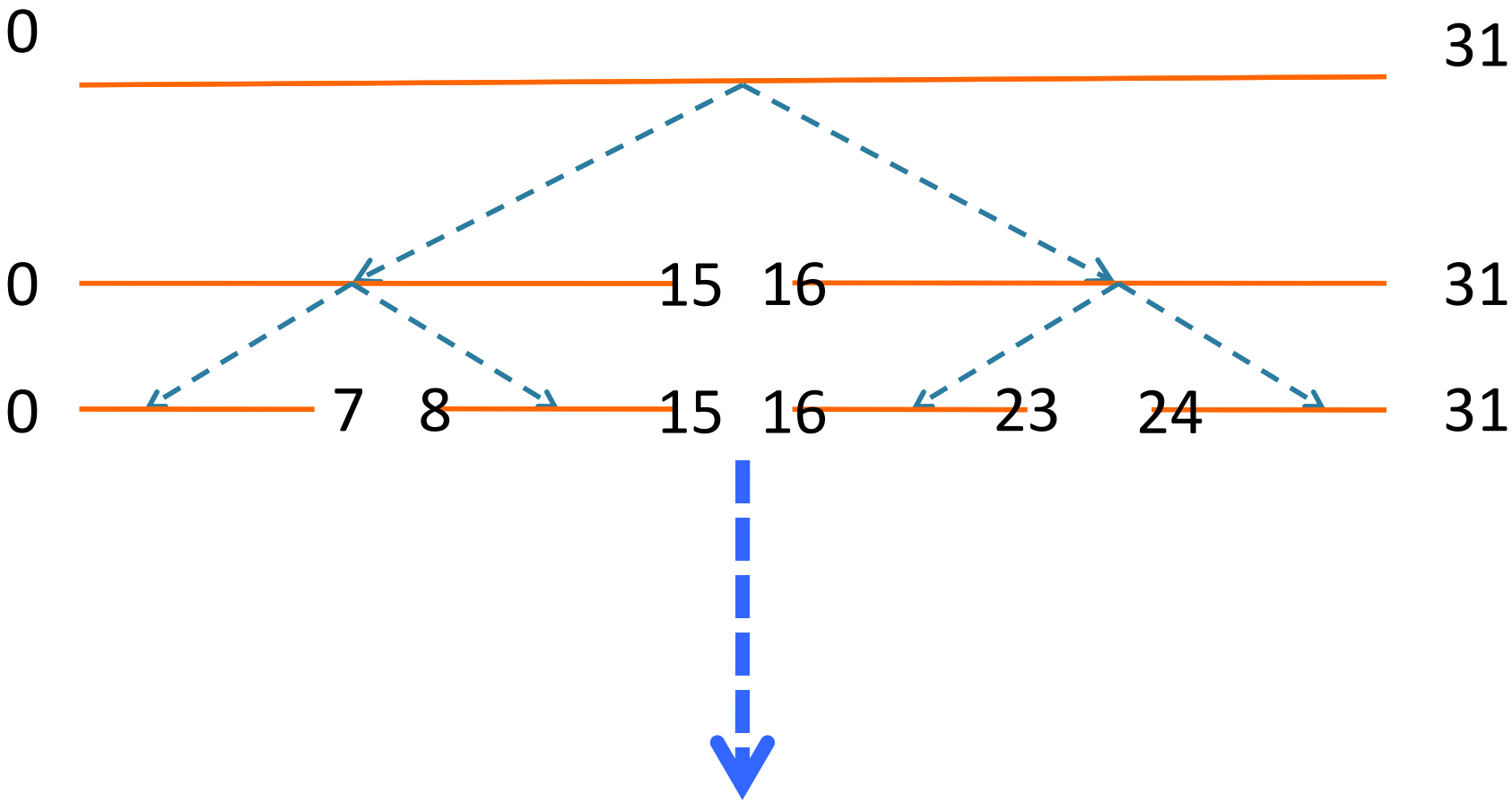
5. f can be approximated in a single pass through the stream

Intuition

Imagine the stream elements sorted according to y coordinates



Dyadic Decomposition on y universe



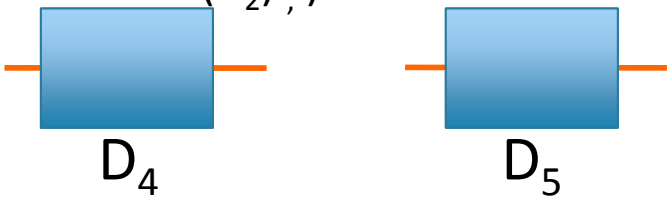
Buckets for certain nodes in dyadic decomposition

Insert all elements into Sketch D_1, until D_1 becomes too "heavy"
i.e f(D) ≥ α (α is a constant to be determined)

When D_1 is too heavy,……

Further Insertions into D_2 or into D_3, depending on y coordinate

If $f(D_2)$,) ≥ α then …

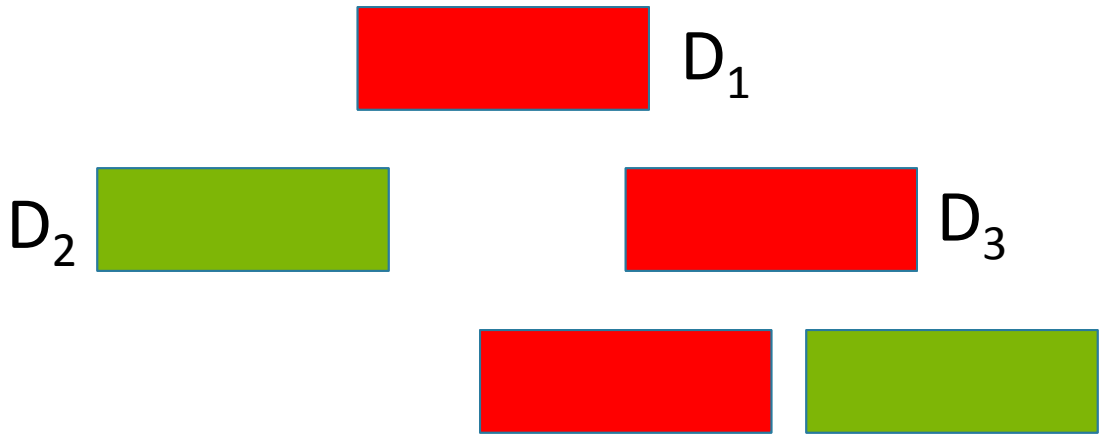


Tree of Sketch Data Structures

- Subtree of dyadic decomposition
- Depth bounded by log (y_{max})
- No control over the exact shape of tree
- Two problems:
  - Can't store the entire tree
  - Even if we did, not all intervals can be handled

Promise: $F_k \leq k.\alpha$

- Only store O(k) buckets with largest right endpoints
- We have all buckets that contain relevant data
  - $f(D_i) \leq \alpha$, for all i from 1 to k
- Some red buckets intersect the query region ($y \geq c$)
- No more than $\log(y_{max})$ buckets can be red



Error Guarantees

- Use smoothness guarantees of aggregation function f to bound error
  - Volume of uncertain portion due to union
  - Contribution of “uncertain” portion to correlated aggregate
- Removing Need for the Promise: Maintain Different Trees for $\alpha = 1, 2, 4, 8, 16, \ldots, f_{max}$

Frequency Moments

Theorem: There is a sketch that yields an (ε, δ)-estimator for correlated F_k and uses space $O(n^{1-2/k} \text{ poly}(1/\varepsilon \log(n/\delta)))$

For F_2, space is $O((\log^3 y_{max})(\log^2 f_{max}) / \varepsilon^4)$

Randomized Approximation: For $0 < \varepsilon < 1$ and $0 < \delta < 1$, an (ε, δ)-estimator of a quantity V is a random variable X such that

$$\Pr[\ |X-V| > \varepsilon V] < \delta$$

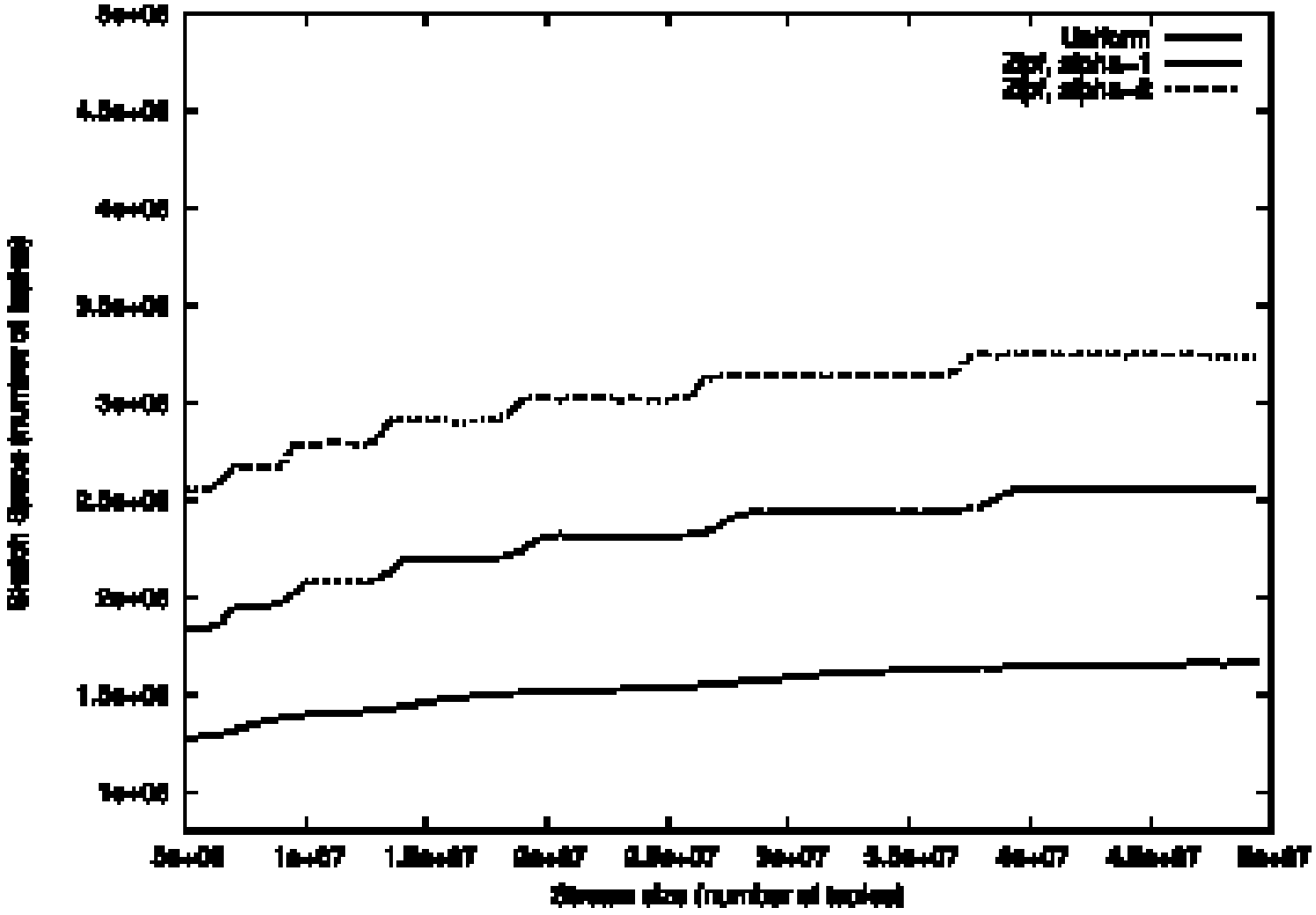
Deletions in a Stream

- Suppose stream elements took the form (x, y, +1) or (x, y, -1)

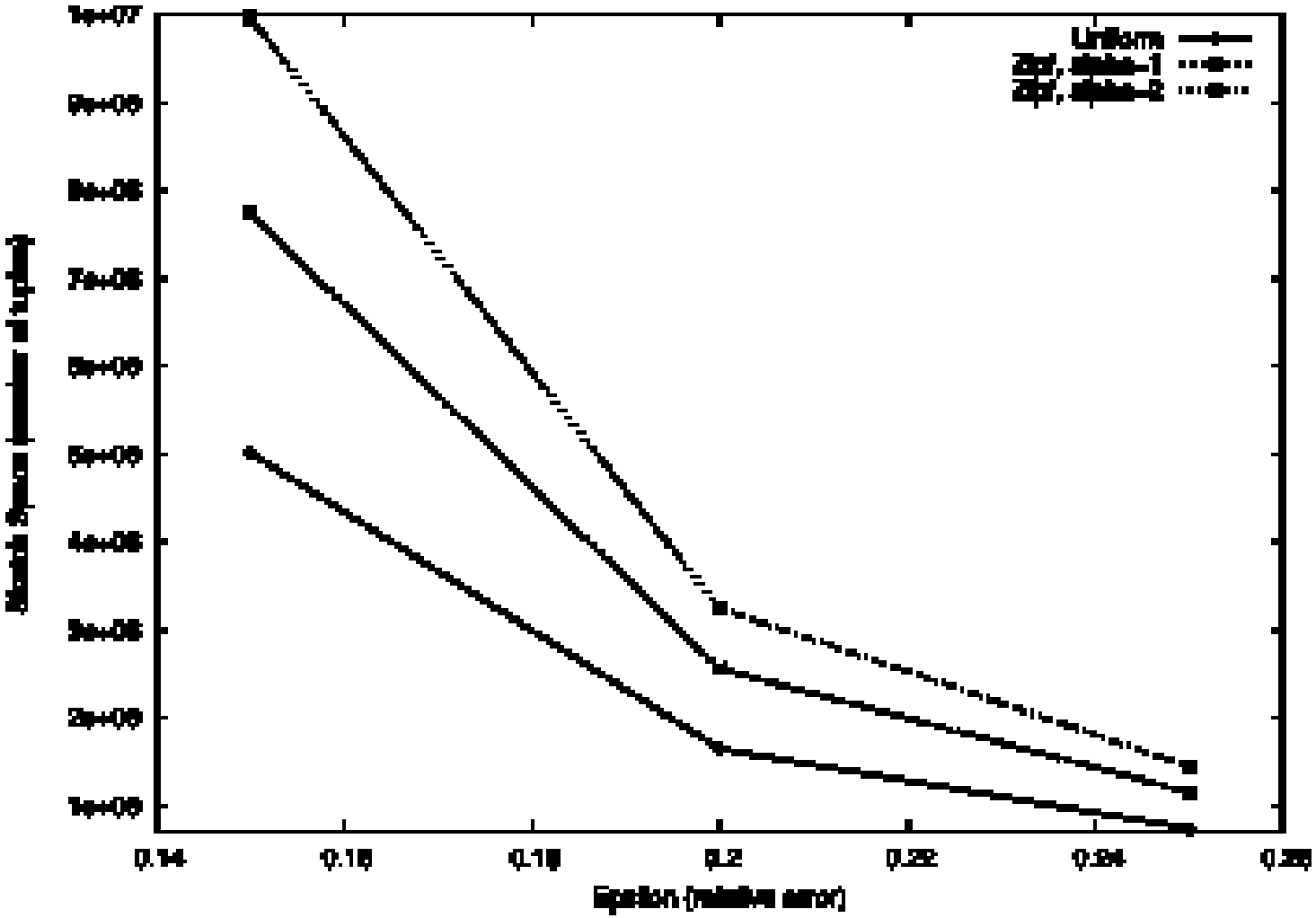
- Lower Bound Theorem: Any sketch constructed using t passes and can estimate $F_k\{x_i \mid y_i \leq c\}$ where c is given at query time must use $(y_{max})^{1/t} / \log(y_{max})$ memory in the worst case

- Contrast with streaming estimation of F_k using sub-linear space

Correlated F_2 , $\varepsilon = 0.2$, $\delta = 0.1$



Correlated F_2, space versus ε, 40 M elements



Conclusions

- Small space sketch for correlated aggregate queries over a large data stream
- Two types of selection predicates: $\sigma(y) = (y \geq c)$, $\sigma(y) = (y \leq c)$
- For aggregate function f with a "smoothness property", correlated estimation of f can be reduced to estimation of f over the entire stream
- Frequency Moments F_k, $k \geq 0$:
  - Space upper bounds for insert-only streams
  - Space lower bounds for insert-delete streams
  - Experiments

Questions

- Better Algorithms for correlated F_k
- Other selection predicates
  - Left and right hand side bounds for y
  - y belongs in a set of ranges
  - Predicates on Frequency of y
- Aggregates involving more than two dimensions