Overview of Stream Algorithms

Sudipto Guha
AT&T Research

The Data Stream Model

Data Items $x_1, x_2, \ldots, x_n, \ldots$

Small storage space : sublinear

Example $\sqrt{n}, \ldots, \log^2 n, \ldots, k$

Compute $f(x_1, x_2, \ldots, x_n, \ldots)$ access in order

Any item not explicitly stored is lost.

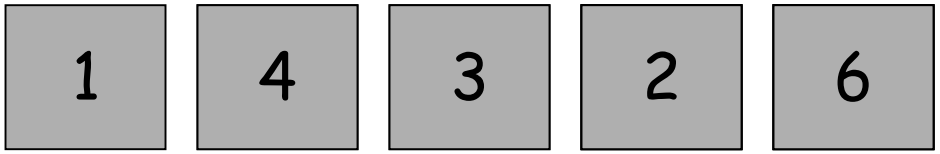
An Example

- Given a set of numbers, $x_1, x_2, \cdots, x_n$ compute their minimum

- Seems easy.
- Sum ? Sum of Powers …
- K largest elements…

Bad News

- Given a set of numbers, $x_1, x_2, \cdots, x_n$ compute their median.
- Cannot be done in less than $\frac{n}{2}$ space.
- Why:



Bend the mind

- Exact answer is hard
- What if close to actual answer

That is return a value close to median.

What is "close"?

Approximate Computation

- Typically for NP Hard optimization
- Most stream problems are Ptime
- Space restriction (also Time)

  - Harder to prove lower bounds
  - Fewer Reductions

Target problems

- Statistics: Quantiles etc
- Clustering
- Spline approximations
- Histograms: Dynamic?
- Wavelets
- Problems in representation/querying

The Online Model

- More an optimization version
- Allows arbitrary space
- Solution has an unchangeable part. Example: assigning queues etc…

- The difficulty is foreseeing future
- Streams are difficult due to space

To live in small spaces ..

- Store some "synopsis/sketch"
  - Sample data
  - Embeddings
  - Store states of computation
  - All of above
  - And more …

A Sampling Approach

- Consider finding the median
- Sample $S = O(\delta^{-1} \log^2 n)$ values
- Sort and return the median of S
- Error is $\pm \delta n$ with high prob.
- Uses Hoeffding's Inequality
- [Manku, Rajagopalan, Lindsay '98]

Sampling is strictly weaker

- Cannot find maximum or minimum Every item can be Inspected but not stored…
- One pass sampling is streaming …
- Good for robust global properties: remains unchanged by local changes
- Note: definition of local

Embeddings

- Basically Dimentionality reduction
- To compute f
- Reduce dimention to fit the space
- Operate in new space to compute g
- Invert back to get f' close to f

Linear Embeddings

- [Johnson, Lindenstrauss '84]

$$\|x\|_2 \leq \|Ax\|_2 \leq (1+\varepsilon)\|x\|_2$$

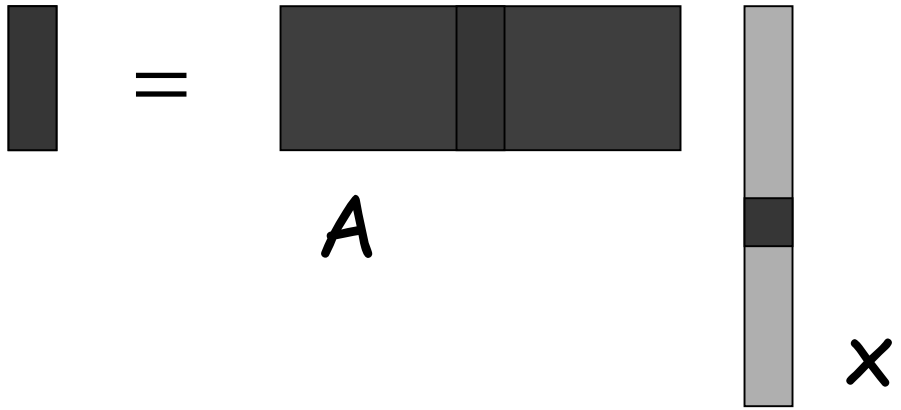
- A is a Random $(\varepsilon^{-2}\log n)\times n$ Matrix drawn from Gaussian distribution.
- Too many elements!

Use Pseudorandom Generators

[Indyk '00]

What it achieves

- Computes Norm when x_i arrive out of order.



Difference of two streams

- Two streams x, y asynchronous etc…

  Compute Ax, Ay and thus $\| Ax - Ay \|_2$

- For 1-norm use Cauchy distribution
- Return the median of $| Ax - Ay |_i$
- Motivation: Statistics of Networks

Dynamic Model

- Supports operations as
  " increase x_i by 5"
- The Cash Register Model
- Can consider a fixed domain [1..n], with insert/delete of elements where the insertions etc. are stream
- Frequency moments, Histograms …

Embeddings II

- A different approach

  [FKSV `99]

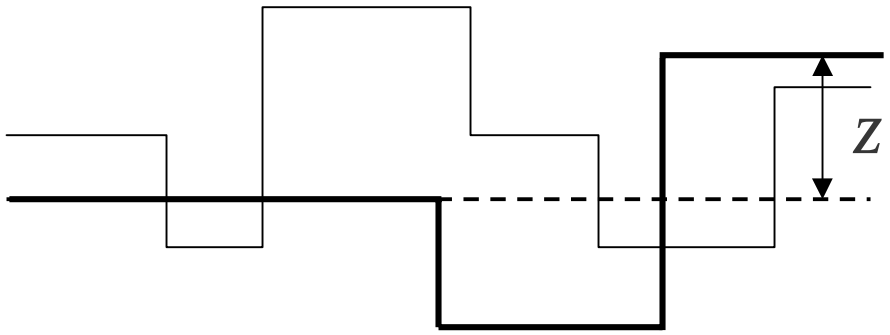
- Consider Ax where the rows of A is a 4-wise independent sequence of ± 1
- A has $O(\varepsilon^{-2})$ Rows
- Compute $\varepsilon \| Ax - Ay \|_2$
- Repeat log n times, return median

Why do we care ?

- Sum of coefficients $\sum_{j=a}^{b} A_{ij}$ can be computed quickly.
- Range summability
- Suppose we have two streams x,y
- We ask: if we replace $y_a, \ldots, y_b$ by a single value v, what happens to $\|x-y\|_2$

Dynamic Wavelets

- [GiKMS '01]
- Have signal x and an approximation y
- Consider adding a wavelet to y



$$A(y+w) = Ay + Aw = Ay + cz$$

A Greedy Approach

- Solve one variable optimization

$$\min \quad \text{median}_{i=1 \text{ to } O(\log n)} \left\{ \left\| Ax - Ay - cz \right\|_2 \right\}_i$$

- Given w, the best wt. can be found
- Pick the best w
- Repeat

Recap.

- Sample data Weak.
- Embeddings Sketches …
- Store states of computation
- All of above
- And more …

States of Computation

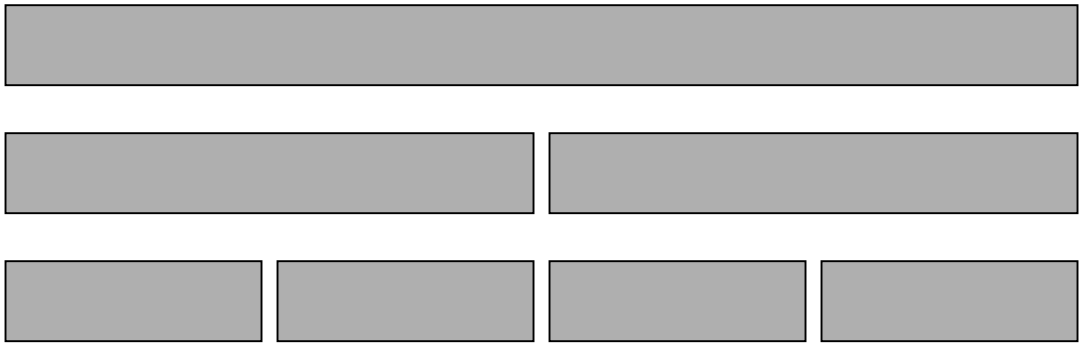
- Consider any offline computation
- Can we store the computation in succint form ?

Targets

- Divide and Conquer
- Dynamic Program

Divide and Conquer

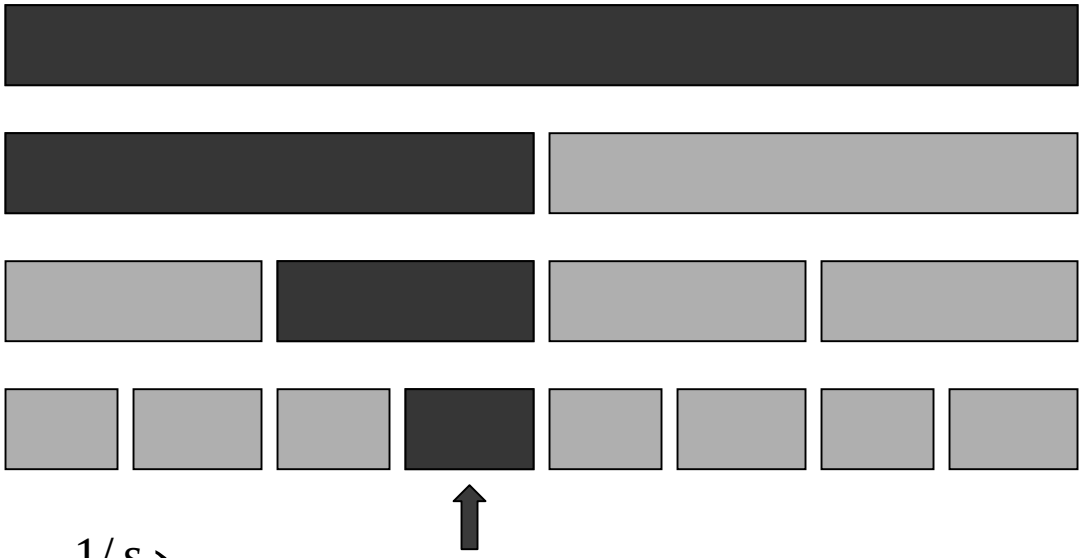
- [GMMO '00]
- Consider the following clustering



R-ary Tree. At each node cluster into K clusters and send to parent

Approximate Clustering

- For s levels, $O(2^s)$ approximation
- But space ?



$O(sn^{1/s})$

Basic Building Block

- Consider a two level process

- Prove that combined new problem has solution close to original cost
- Find (approximate) it

The Dynamic Program

- Store the table in compressed form
- Approximate entries to indicate only the large changes

- For new element, search is reduced since the table is small

Histograms

Given $x_1, x_2, \ldots, x_n$ construct $f(i)$

$$\sum_{i=1}^{n} [f(i) - x_i]^2 = \|f - x\|_2^2$$

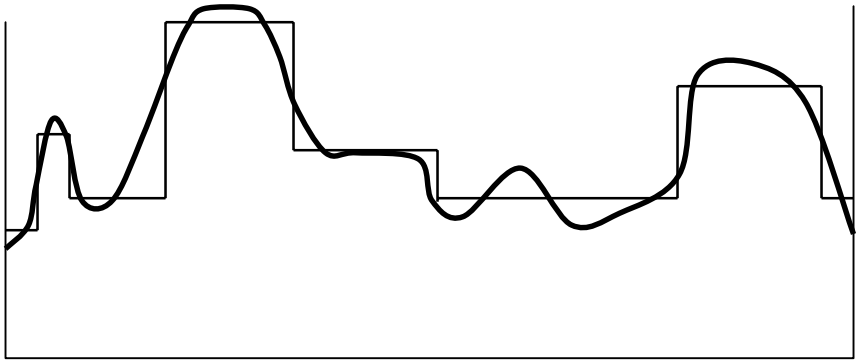
where $f(i)$ is a step fn. with k steps.

Consider x_i as the frequency item i

Basically approximating a signal.

Histogram Example

- Consider a data distribution, $k = 8$

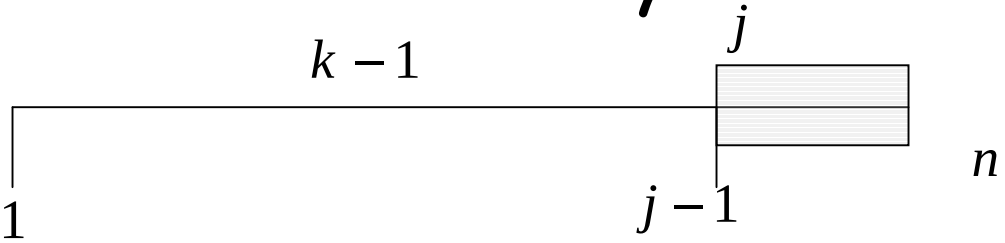


Optimal Offline Algorithm

- Within "step/bucket": Mean is best.
- Assume that the last bucket is $[j..n]$

.

What can we say about the rest $k-1$?

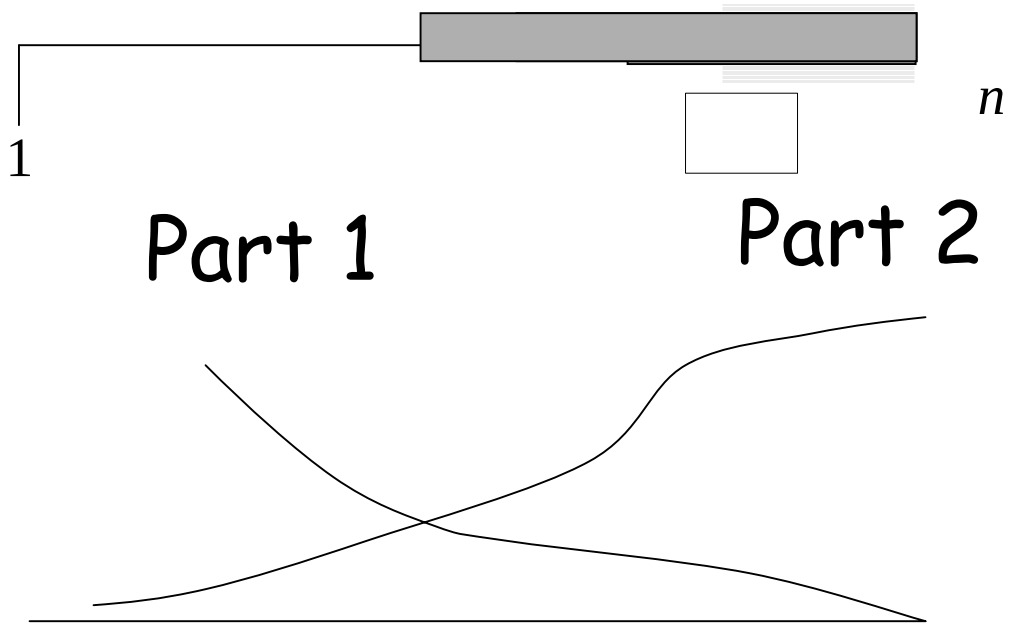


Must also be optimal for $[1..j-1]$!

Dynamic Programming !!

The Algorithm Idea

- Sum of two functions



Analysis of Optimal Alg.

- Algorithm takes $n \times k \times n = n^2 k$ time
- The space required is $n \times k$

- Biggest drawback: Back References !!
- Consider k is 2,

Reducing Search?

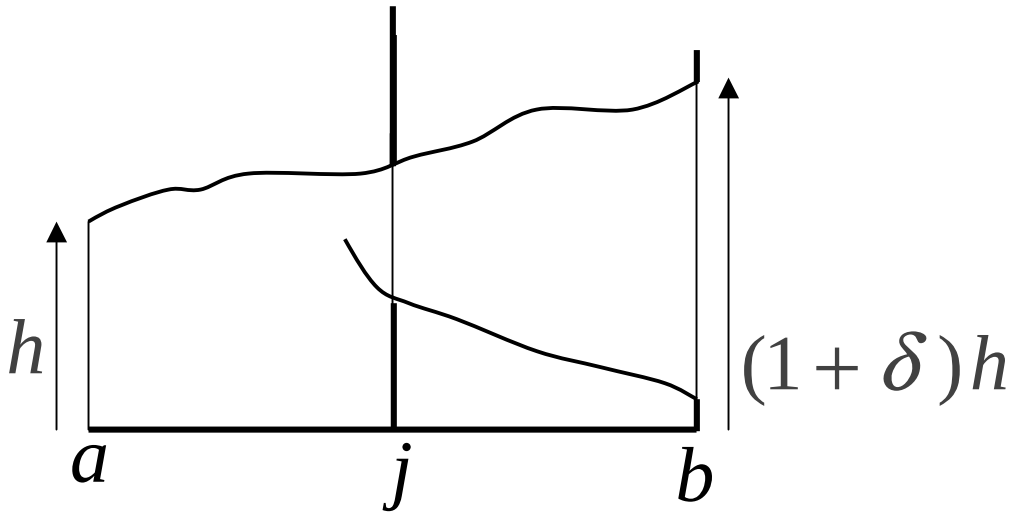
- Not Really.
- Exact Minimum:

  Finding minimum of any sequence of positive numbers can be reduced to finding min of sum of two such functions.

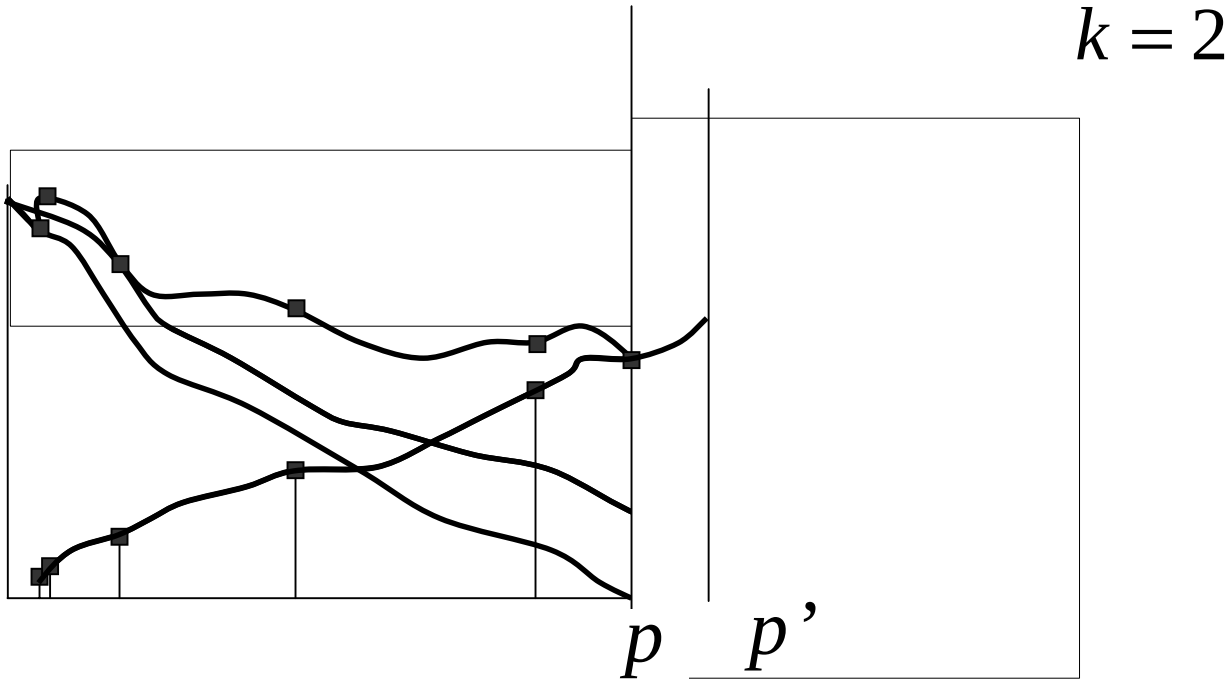
  $\Rightarrow \Omega(n)$ time.

Another Try

- [Guha Koudas `01]



Algorithm by Picture



Works for piecewise small degree poly.

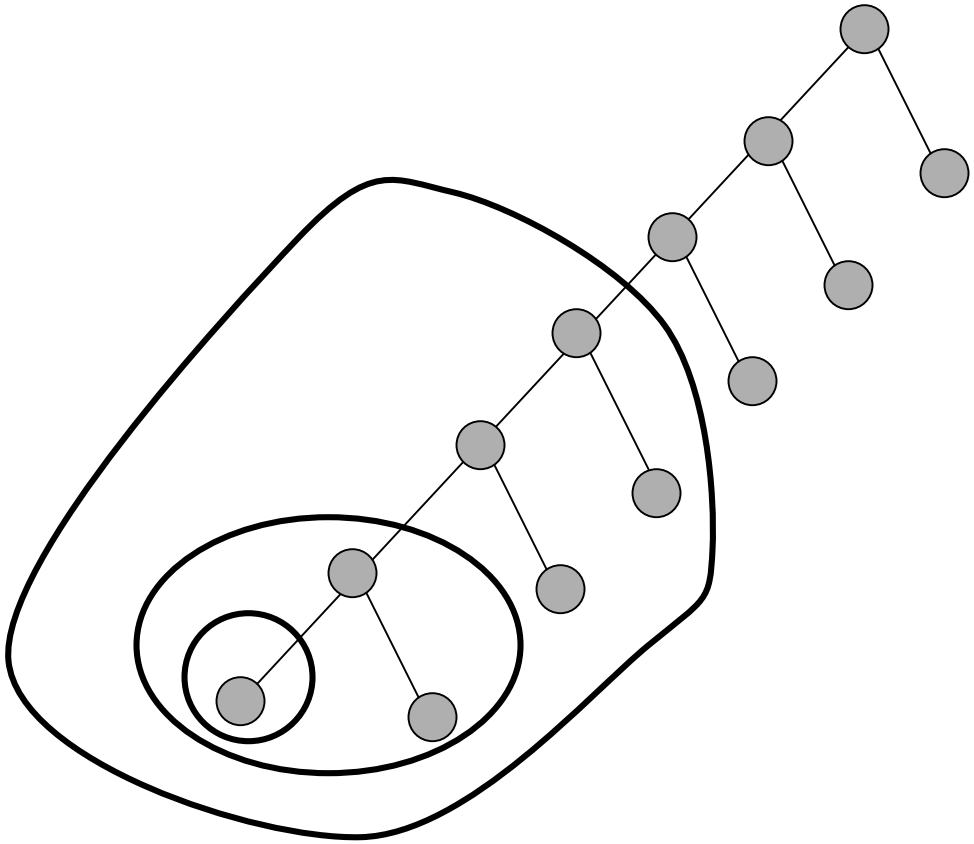
Net Gain

Space $O(k\varepsilon^{-1}\log n)$

Time $O(k^2\varepsilon^{-1}\log n)$ per item

Approximation $1+\varepsilon$

The Computation Tree

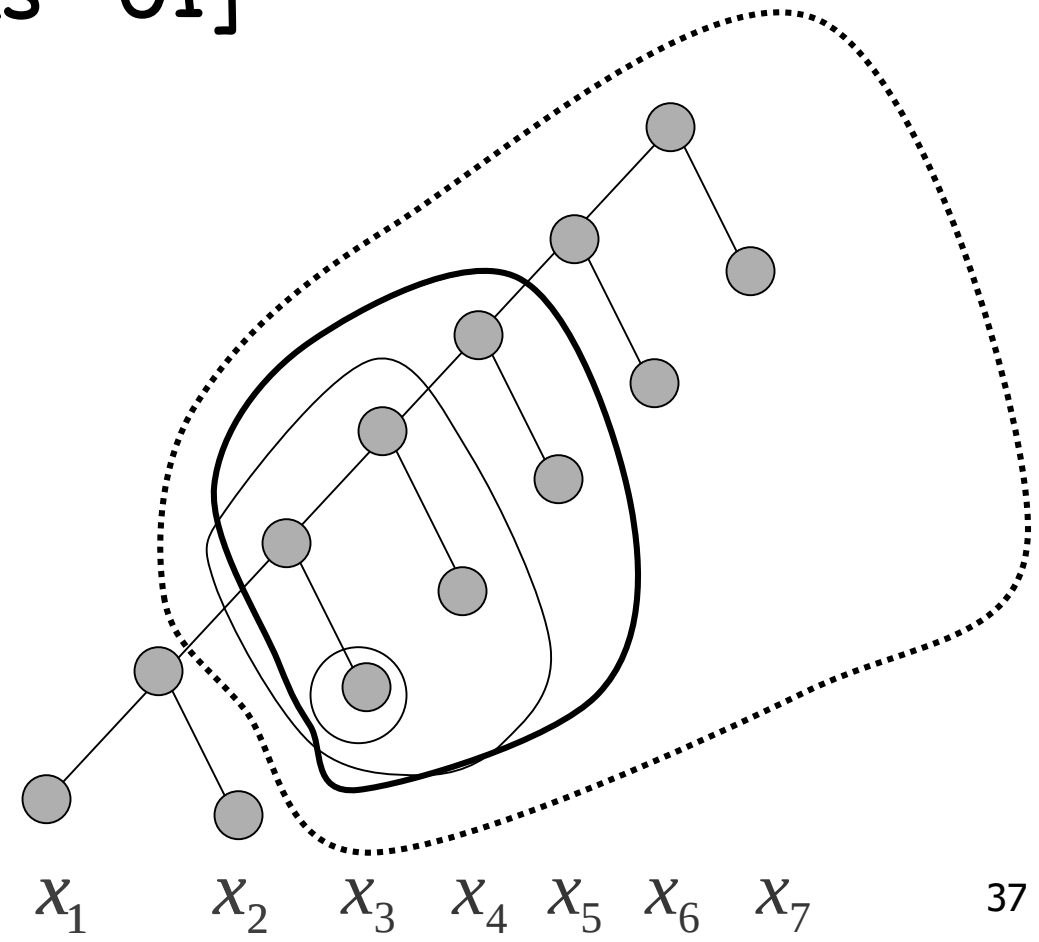


Store information for few prefixes

Sliding Window on Streams

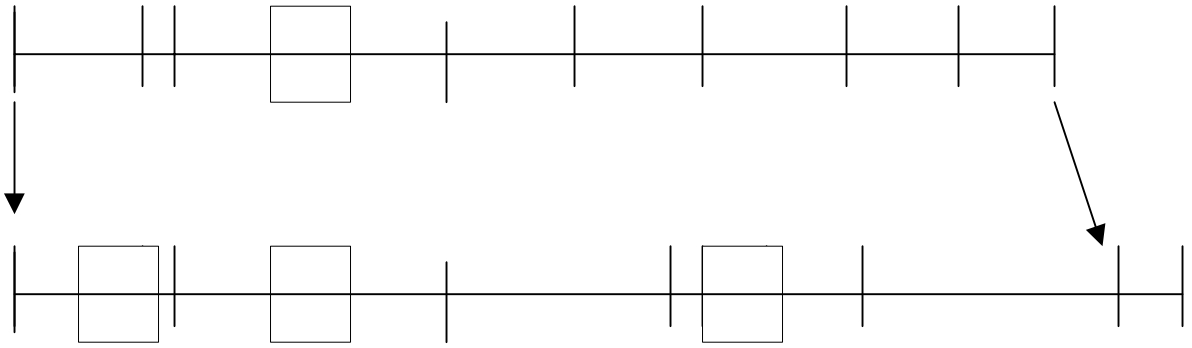
- [Guha Koudas `01]

Generate required terms by binary search



Another Median Finding Alg.

- [Greenwald, Khanna `01]



$O(\varepsilon^{-1} \log n)$ Space suffices

Still a tree, but not contiguous

Recap.

- Sample data Weak.
- Embeddings Sketches …
- Store states of computation

    Tree Merges, divide and conquer

One last example combining them…

Combining Approaches

- Dynamic Histograms
- [GiGIKMS `01]
- Get a dynamic wavelet representation with not more than $O(k \log n)$ terms
- Treat this wavelet to a Dynamic Program reducing to k buckets

Conclusions

- Several approaches
- Several models
- A feet above the tip of the iceberg

HOWTO?

- Dimentionality Reduction
- Store few states
- Compose different algorithms