

A Study on Workload-Aware Wavelet Synopses for Point and Range Sum Queries

Michael Mathioudakis, mathiou@cs.toronto.edu

Dimitris Sacharidis, dsachar@dblab.ntua.gr

Timos Sellis, timos@dblab.ntua.gr

DOLAP 2006

Outline

- **Introduction**
- Wavelets
- Error Metrics
- Algorithms for Point Errors
- Algorithms for Range Sum Errors
- Experimental Results

Introduction

- Approximate Query Processing over **Synopses**:
An effective approach to manage **large data sets** (eg OLAP queries)
 1. **Query optimization** process - Provide highly accurate query selectivity estimates
 2. Can be used instead of the actual data - Provide quick **approximate answers** to large queries
- **Workload**-Awareness:
Take user behavior under consideration - More accuracy for important data - workload aware synopses
- **Histograms**, **Wavelet** Transformation :
Commonly Used Synopses construction techniques

Introduction - Our Contribution

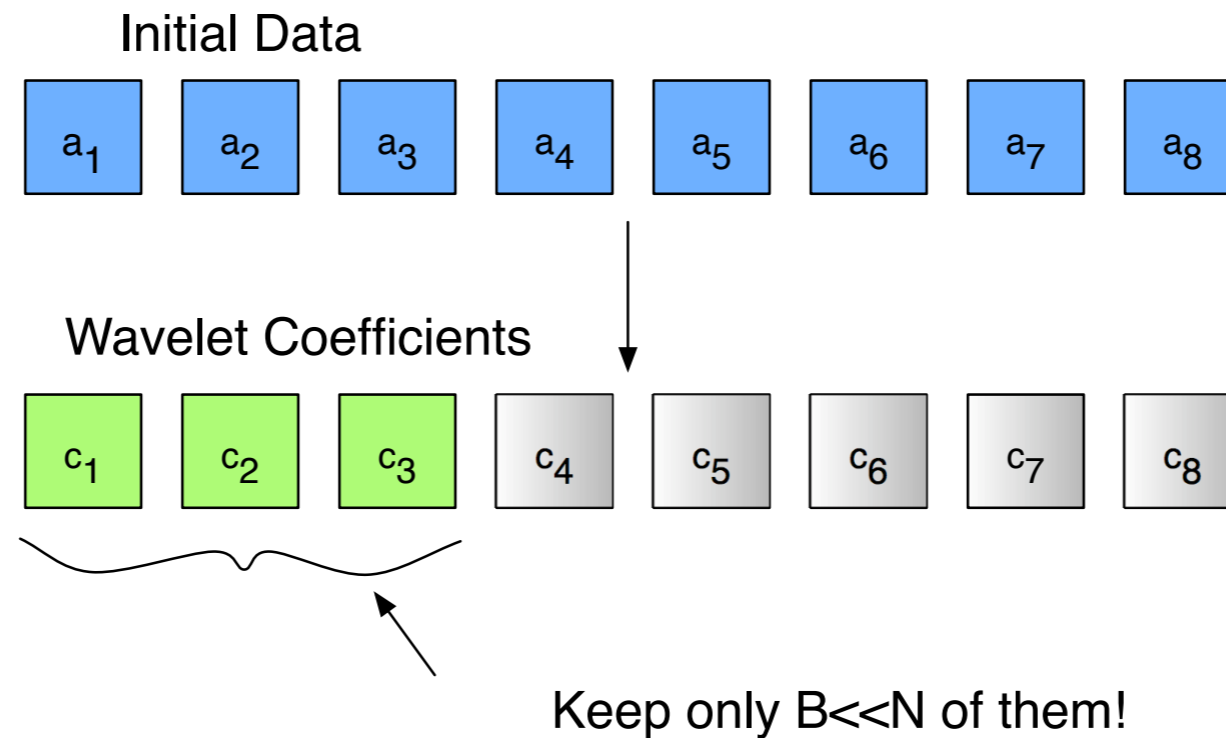
- Focus on **wavelet** synopsis construction algorithms
- **Theoretical** presentation of **existing** algorithms
- Presentation of a novel **workload-aware** algorithm for **range-sum queries**
- **Experimental** study - Accuracy vs Time Efficiency

Outline

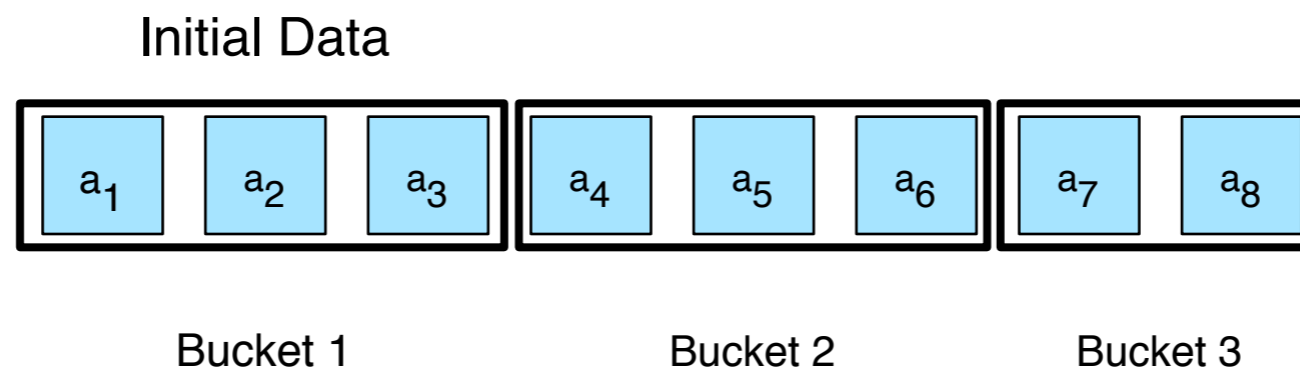
- Introduction
- **Wavelets**
- Error Metrics
- Algorithms for Point Errors
- Algorithms for Range Sum Errors
- Experimental Results

Wavelet Preliminaries

- It's a transformation!

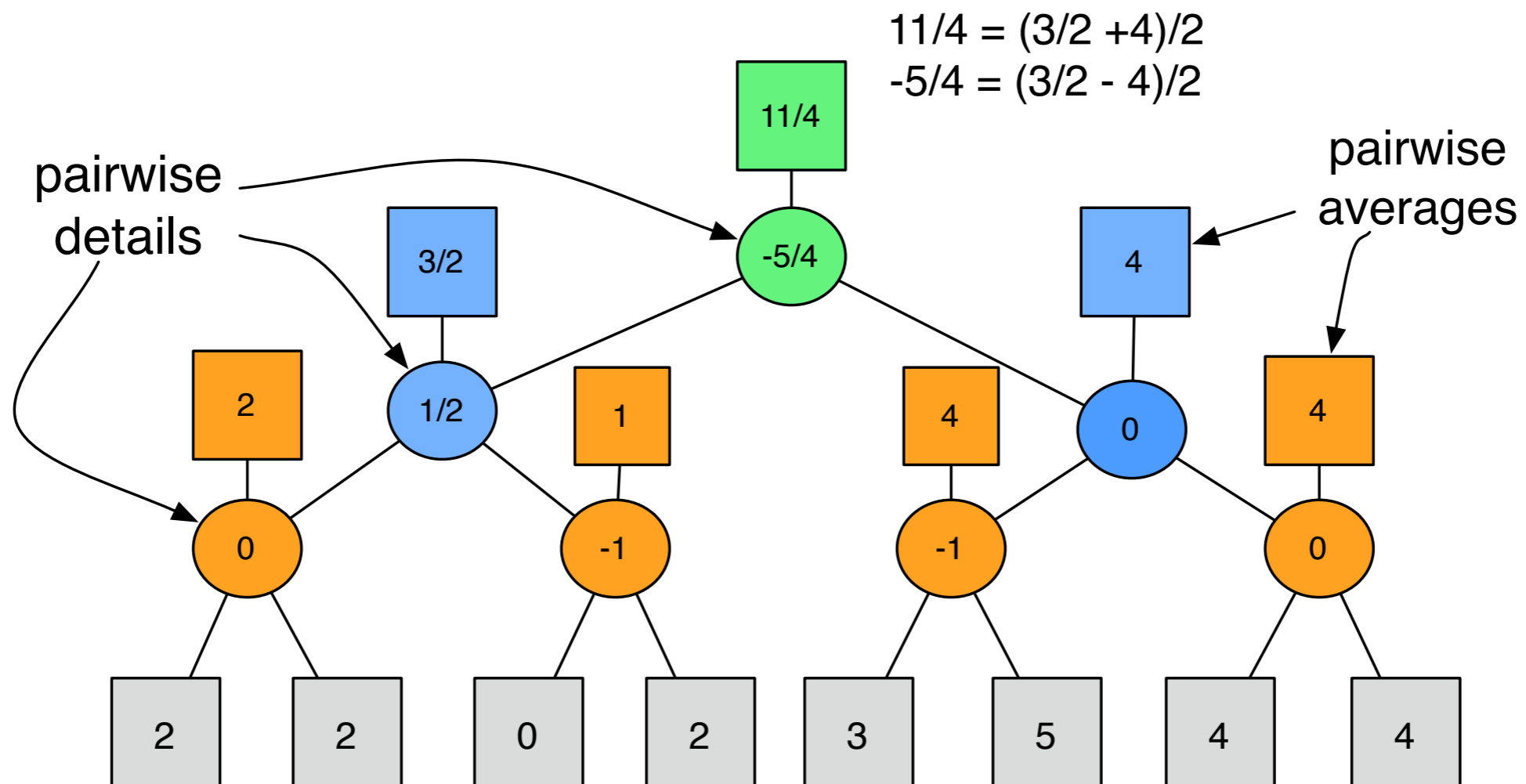


- Histograms: Construct Buckets on Initial Data - Assign one value per bucket



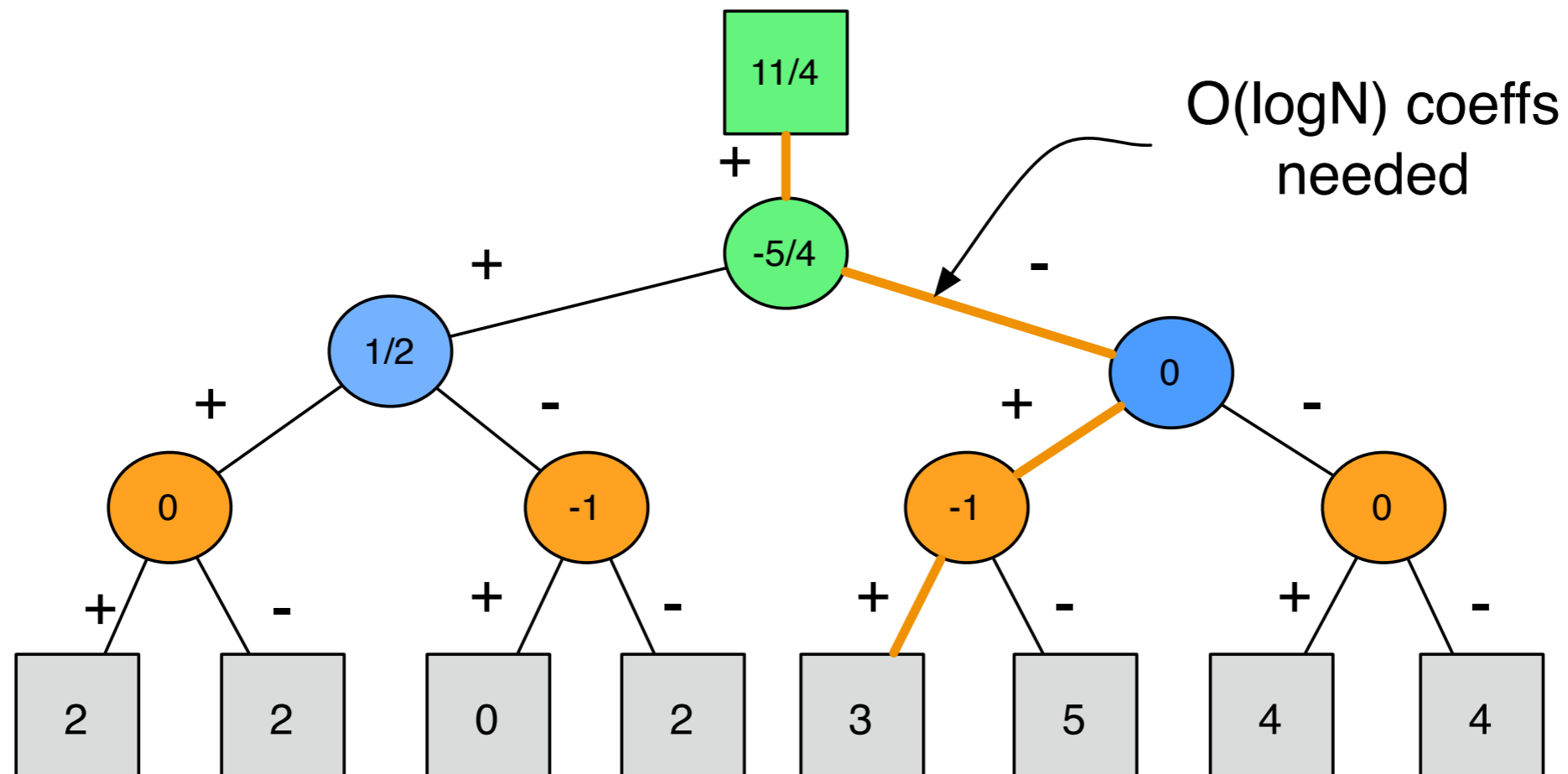
Wavelet Preliminaries

Haar W/T: recursive **pairwise** calculation of **averages** and semi-differences (**details**)



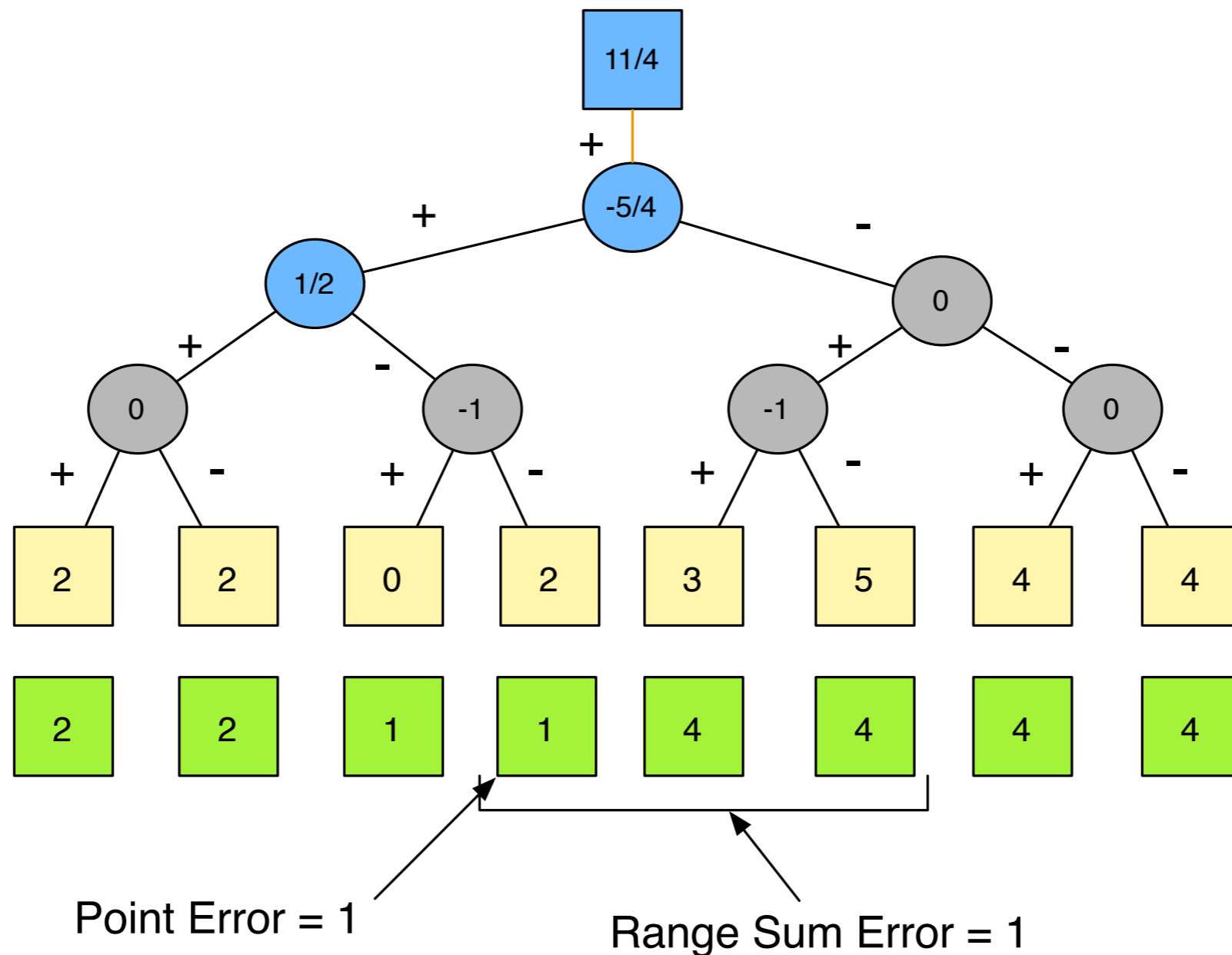
Wavelet Preliminaries

- Initial values can be reconstructed in logarithmic time
- Similar values for near data - small details
- Coefficients near the root are more important - normalization needed



Wavelet Synopses

- Keep B coefficients - Dropped coefficients are considered zero
- Error introduced to the values of our data

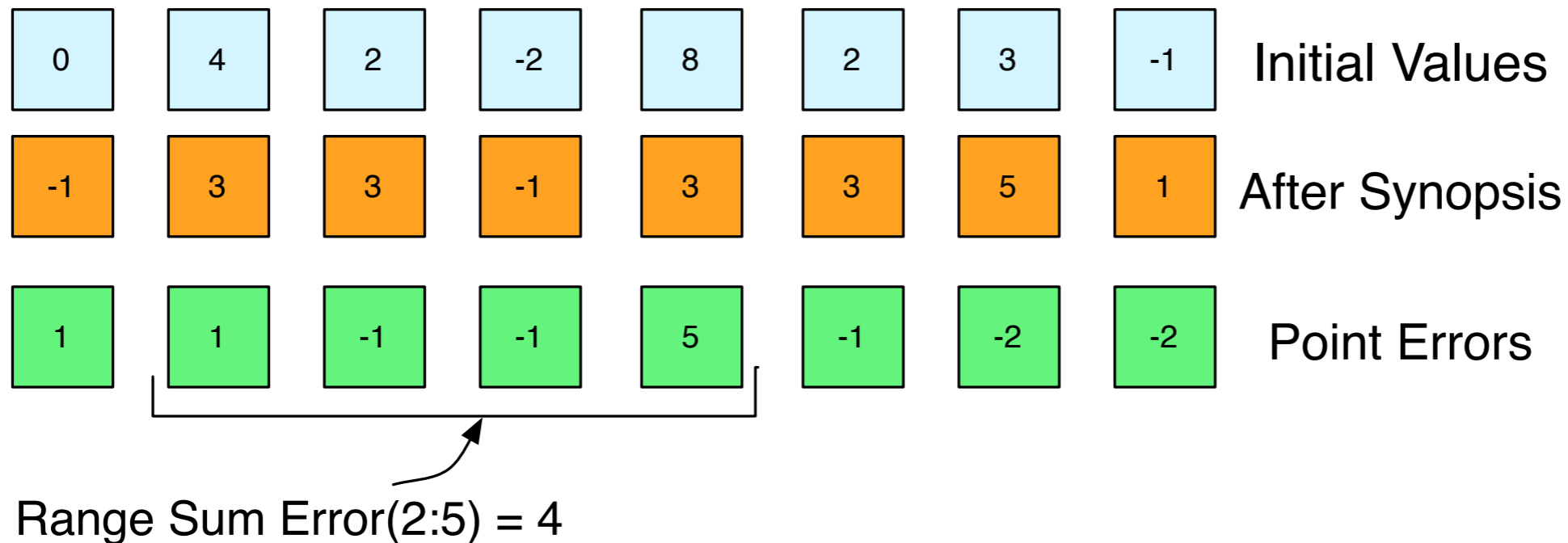


Outline

- Introduction
- Wavelets
- **Error Metrics**
- Algorithms for Point Errors
- Algorithms for Range Sum Errors
- Experimental Results

Error Metrics

- Weighted Error Metrics
- For point queries : $L_{wp} = \sum_i w[i]e[i]^p$
- For range sum queries: $L_{wp} = \sum_{i \leq j} w[i,j]e[i:j]^p$

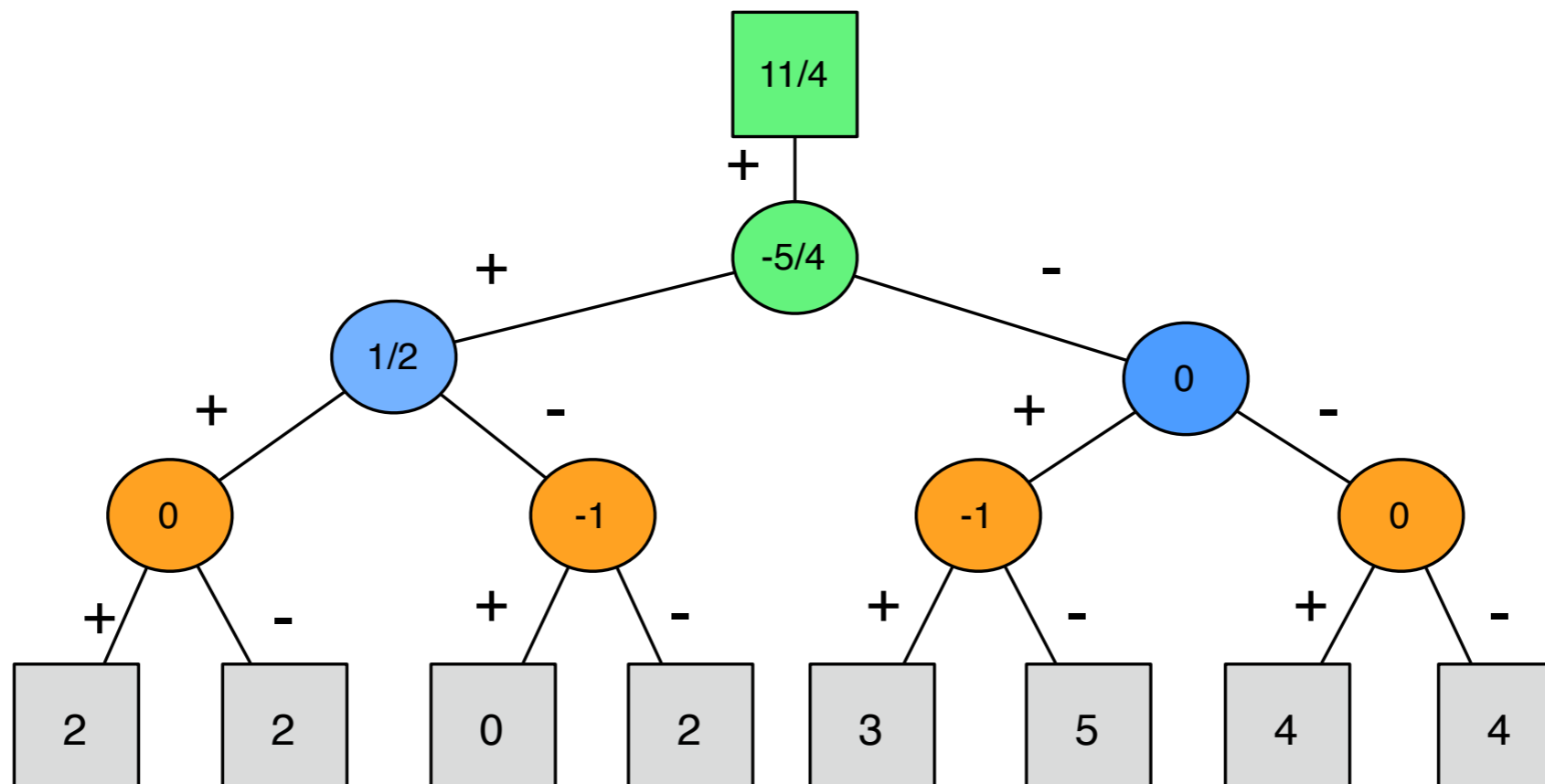


Outline

- Introduction
- Wavelets
- Error Metrics
- **Algorithms for Point Errors**
- Algorithms for Range Sum Errors
- Experimental Results

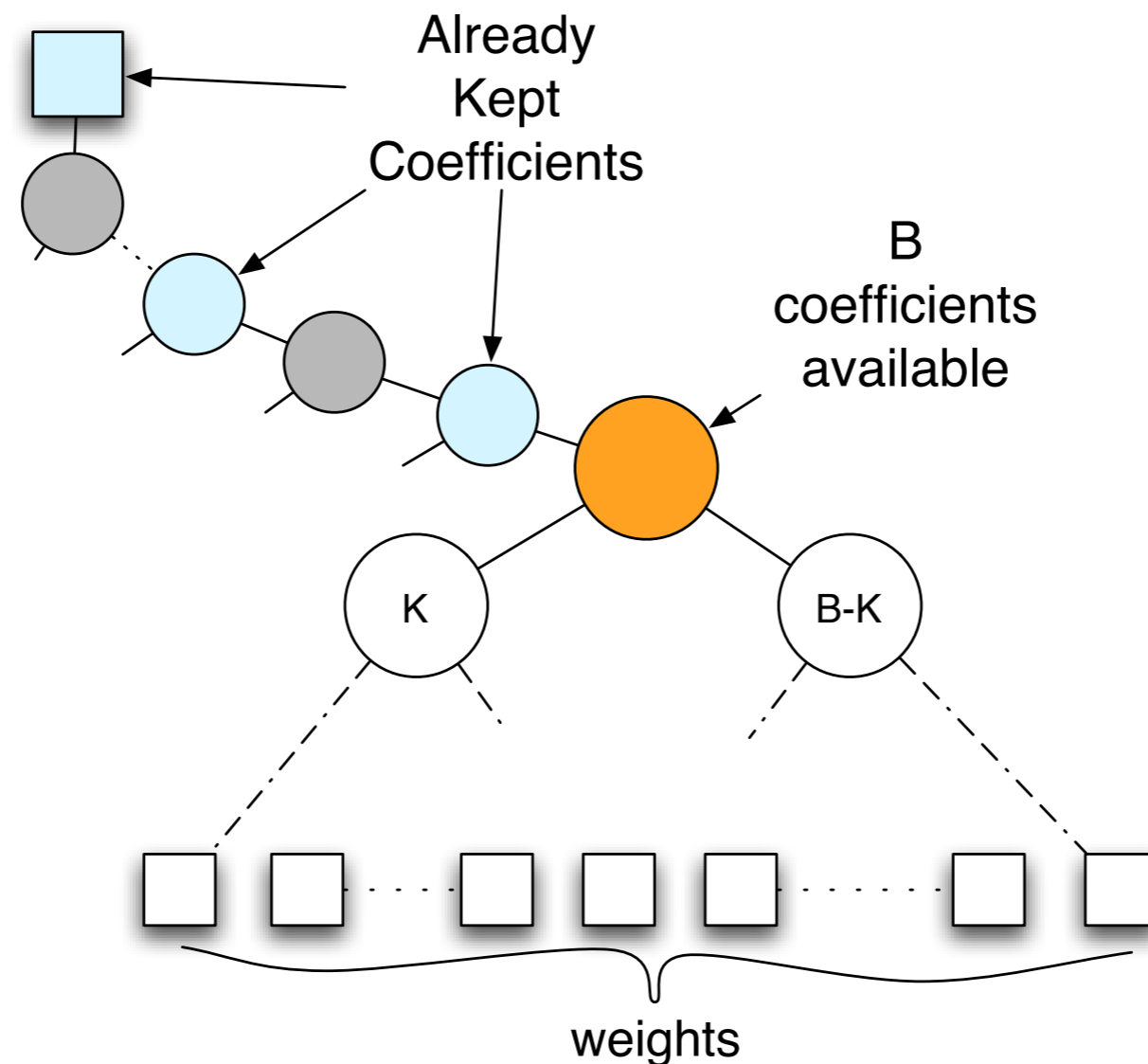
Classic Algorithm

- Minimizes L_2 of point errors
- Selects the B largest normalized coeffs, using a heap
- Complexity: $O(N)$ space, $O(N+B\log N)$ time



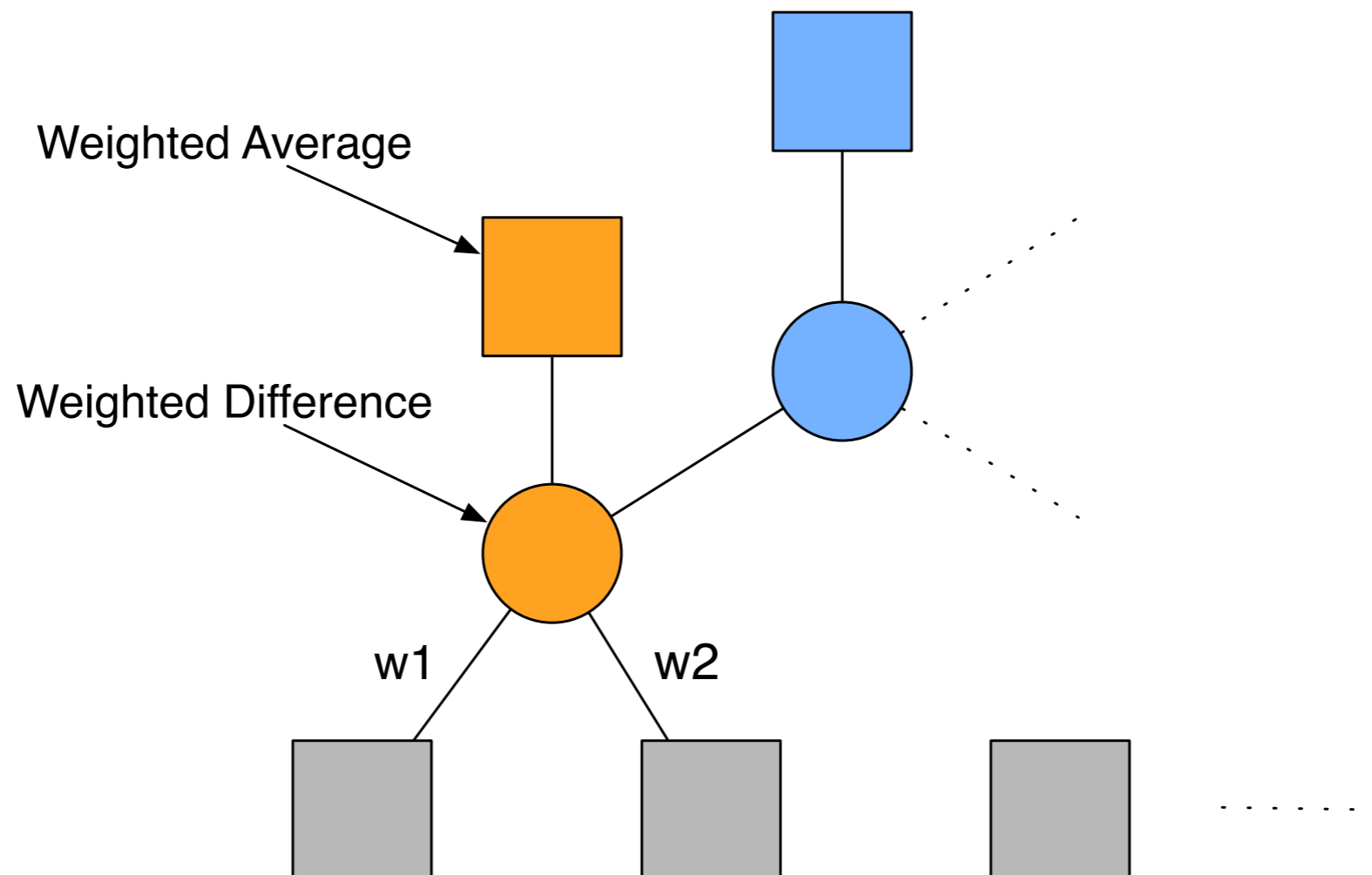
Garofalakis - Kumar

- Minimizes **Weighted** Error Metrics
- **Dynamic** Programming Algorithm on transformation's tree
- Complexity: $O(N^2)$ Space, $O(N^2 \log B)$ Time



Matias-Urieli

- Minimizes L_{w2} of point errors
- Using a **modified Haar** wavelet transformation, then apply the classic algorithm
- Complexity: $O(N)$ space, $O(N+B/\log N)$ time

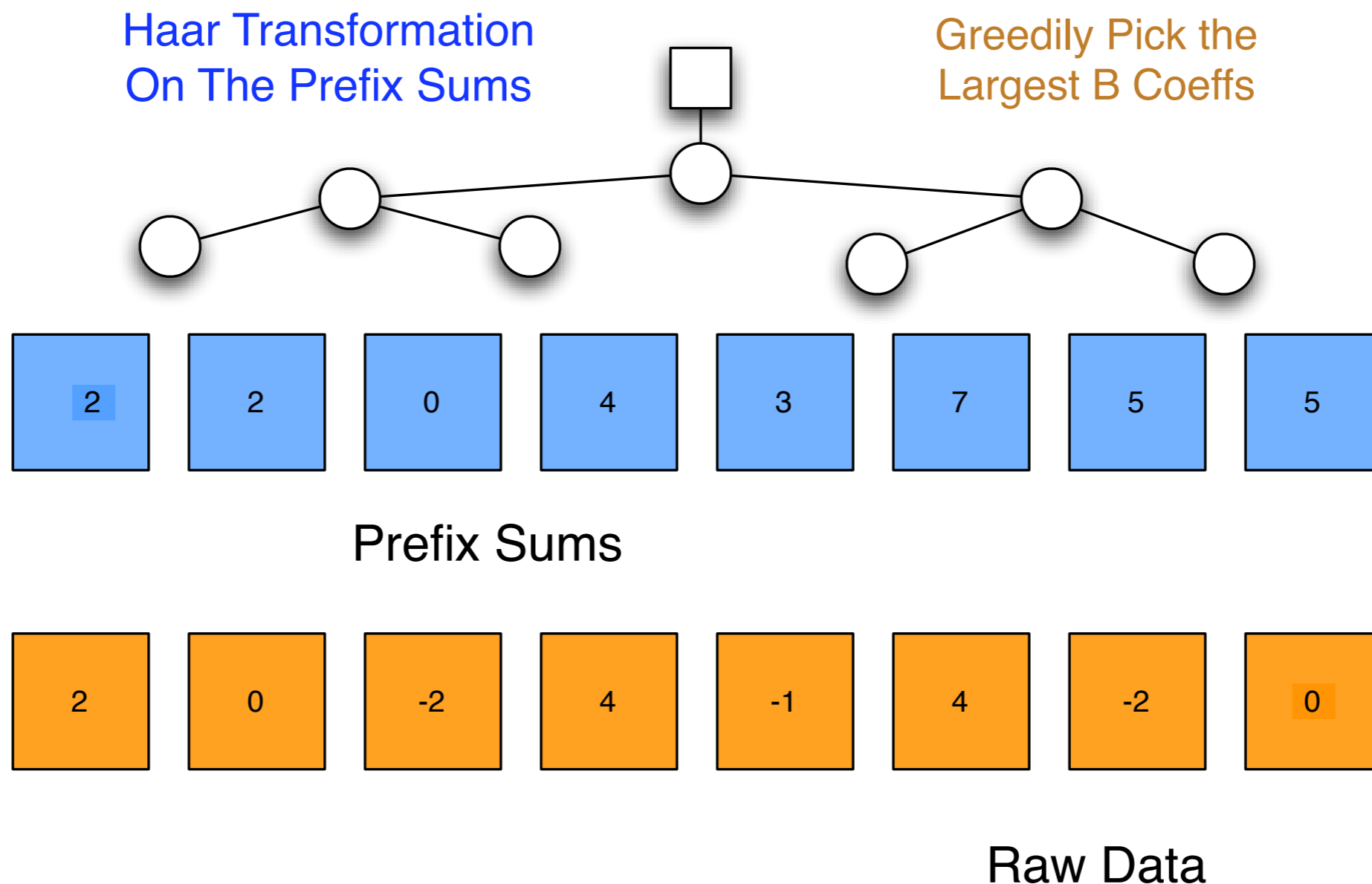


Outline

- Introduction
- Wavelets
- Error Metrics
- Algorithms for Point Errors
- **Algorithms for Range Sum Errors**
- Experimental Results

Matias - Urieli

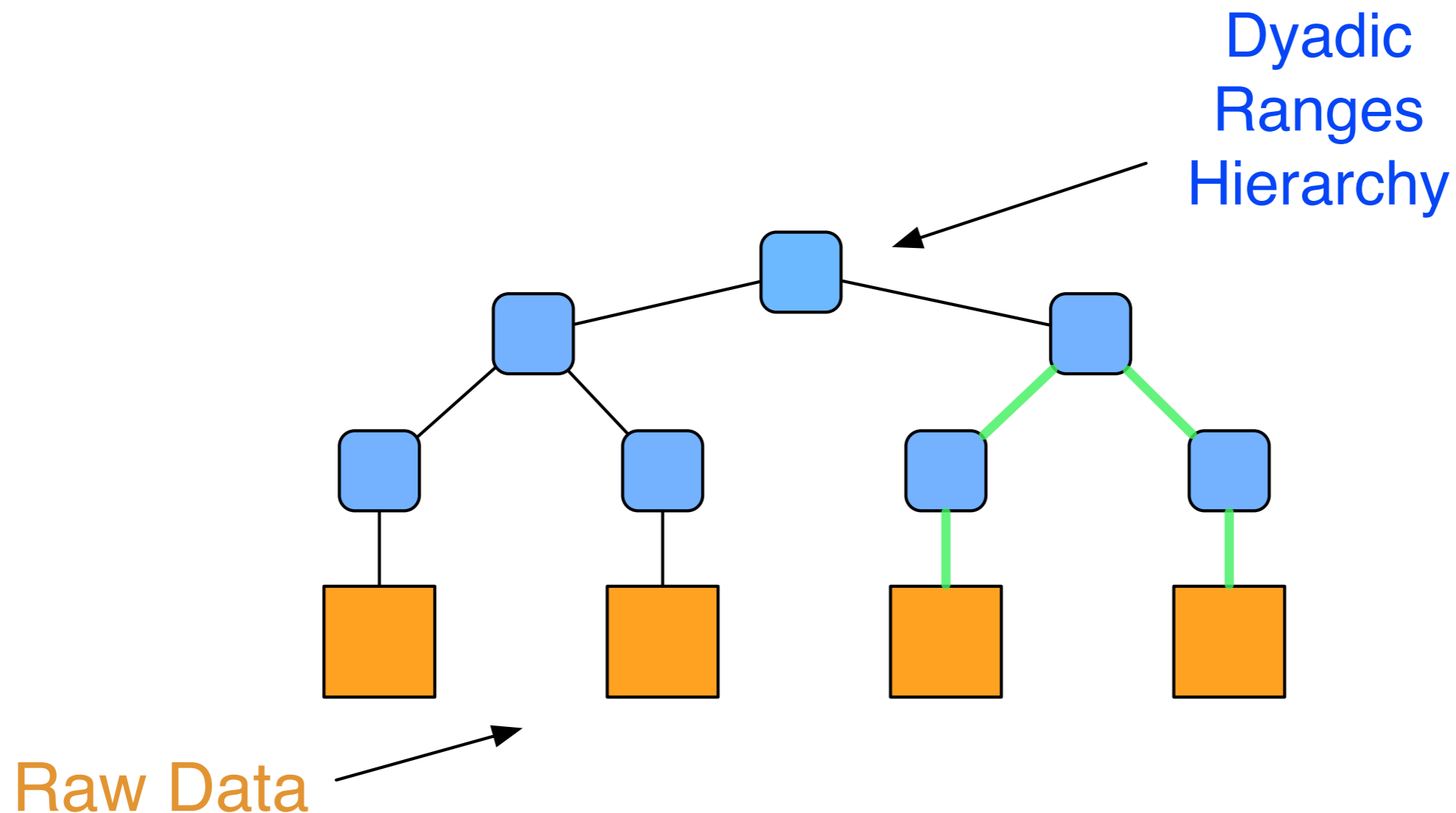
- Minimizes L_2 - Complexity: $O(N)$ space, $O(N+B\log N)$ time
- Working with prefix sums has disadvantages: sparse data become dense, difficult to update



RangeWave

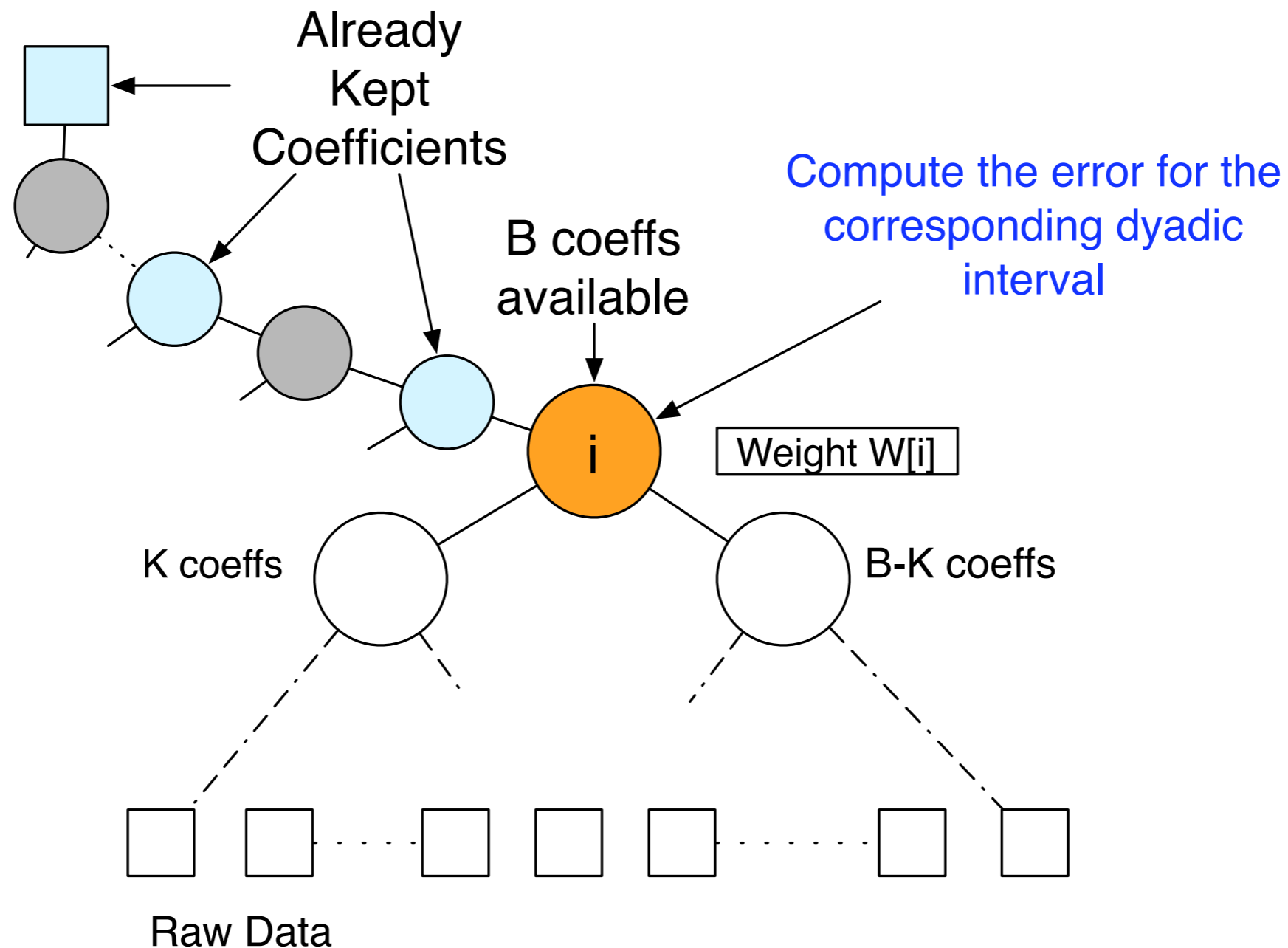
range-sum query workload

- Minimizes **Weighted- L_p** of range sum queries, that follow a dyadic hierarchy
- Workload Aware - Applies on Raw Data



RangeWave

- A **Dynamic Programming** Algorithm
- Complexity: $O(N^2 \log B)$ time, $O(N^2)$ space



Outline

- Introduction
- Wavelets
- Error Metrics
- Algorithms for Point Errors
- Algorithms for Range Sum Errors
- **Experimental Results**

Algorithms Summary

Point Query Workload

Algorithm	Time	Space	Optimal
Matias - Urieli	$N+B\log N$	N	Yes
Garofalakis - Kumar	$N^2\log B$	N^2	Yes
Classic Wavelets	$N+B\log N$	N	No
Classic Histograms	N^2B	NB	Yes

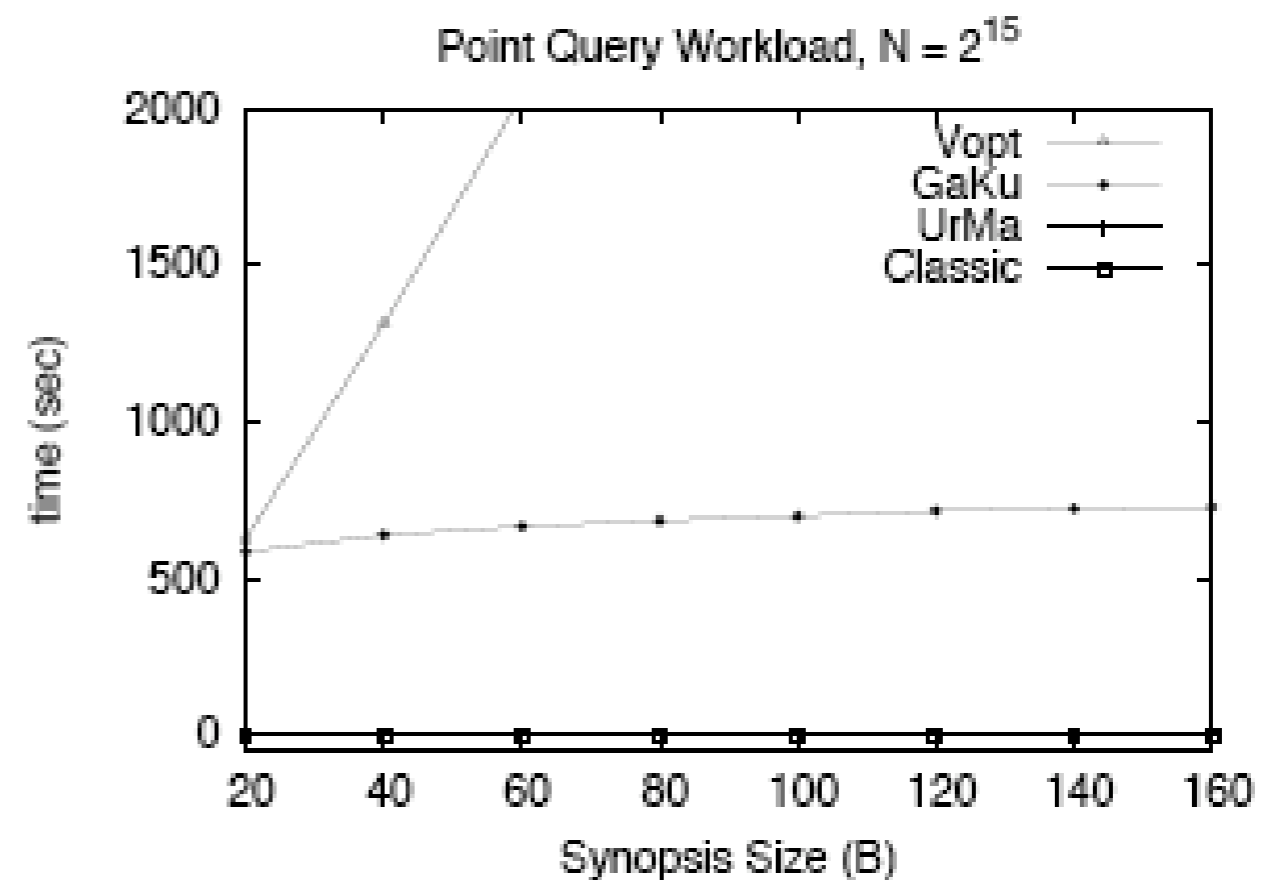
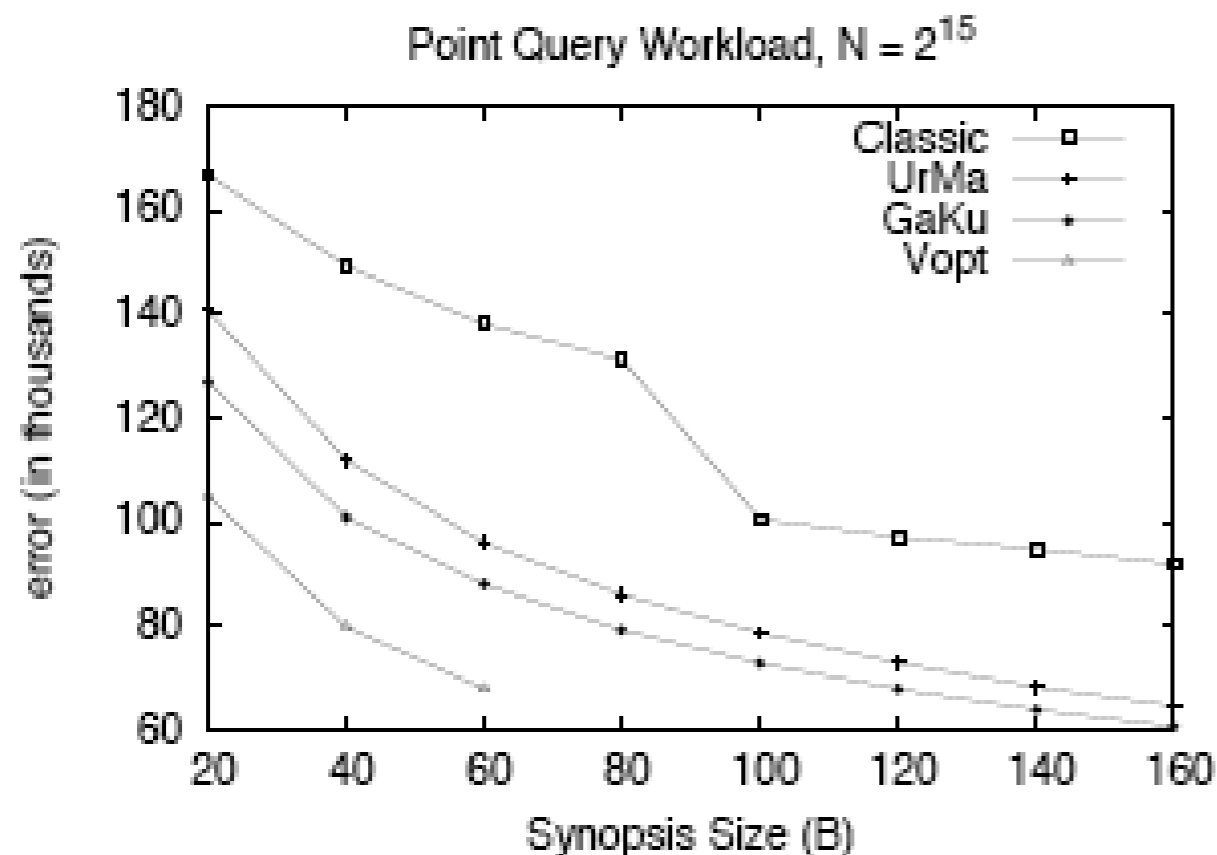
Dyadic Range Sum Query Workload

Algorithm	Time	Space	Optimal
RangeWave	$N^2\log B$	N^2	Yes
Koudas-Muthukrishnan	N^7B^2	N^5B	Yes
Matias - Urieli	$N+B\log N$	N	Only for uniform workload
Classic	$N+B\log N$	N	No

Experimental Study

Point-Query Workloads

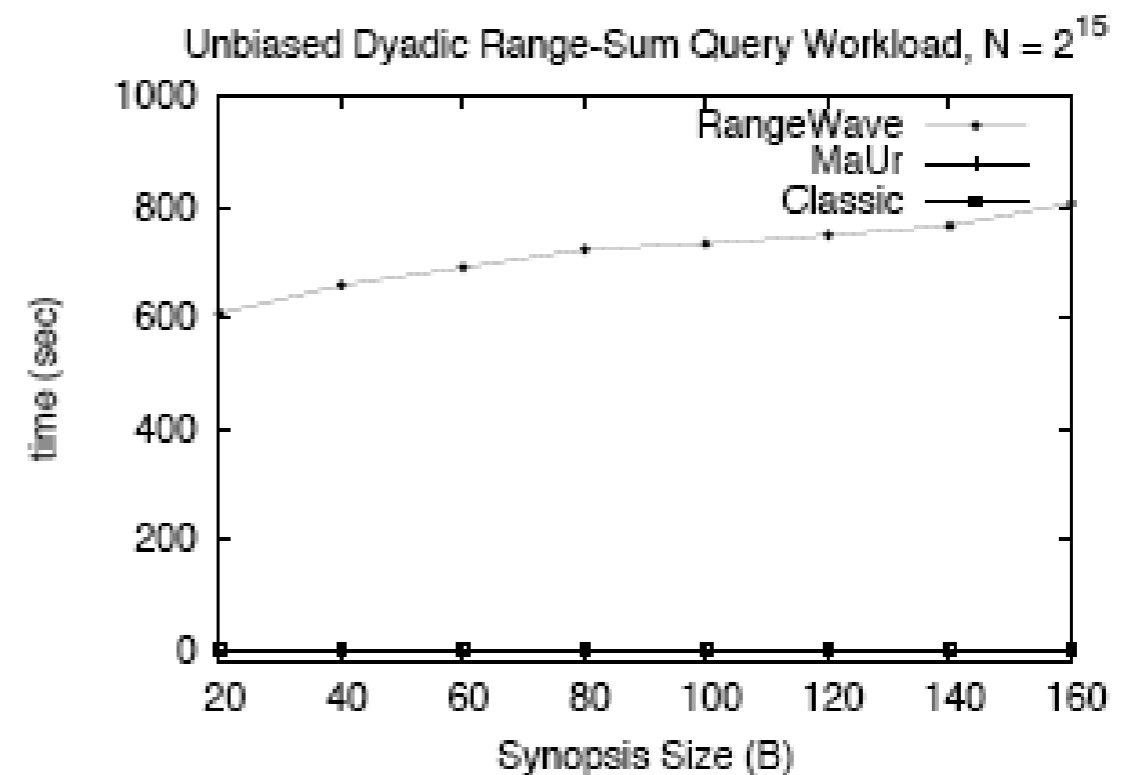
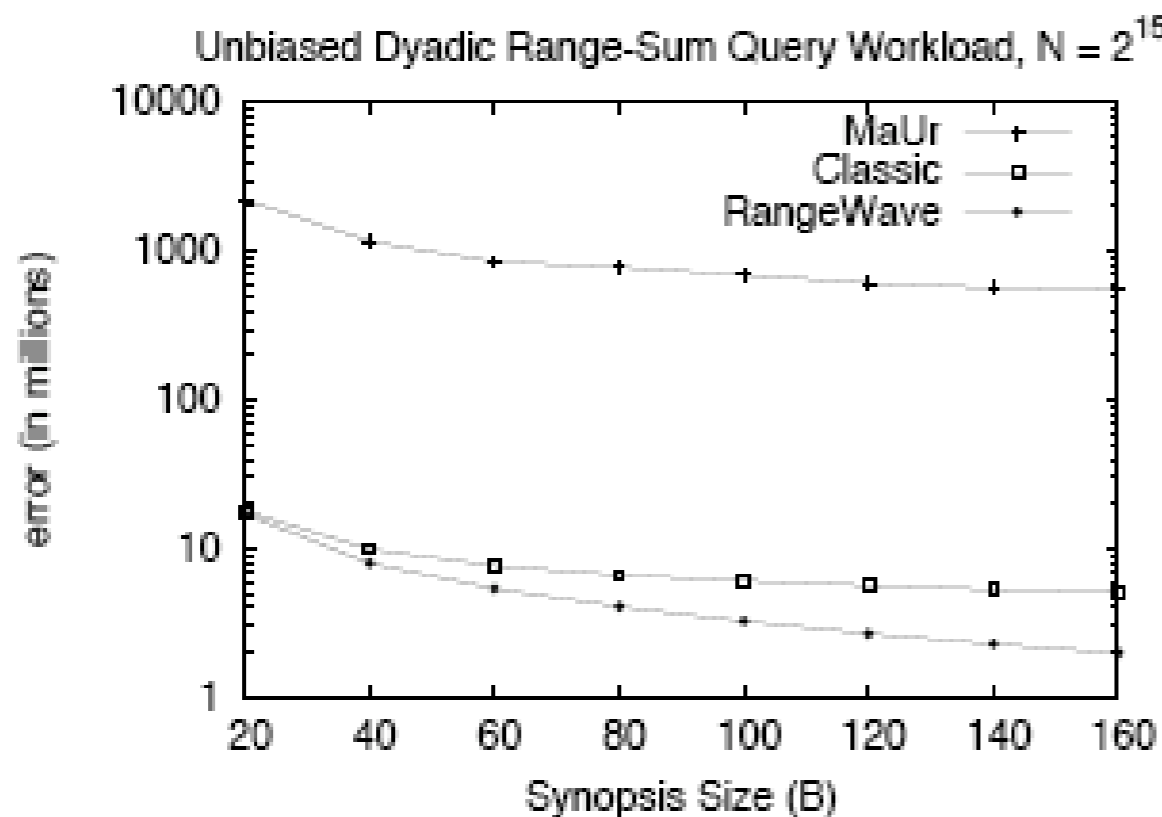
- Data and Point Workload follow Zipfian distribution
- Increasing Synopsis Size
- Urieli-Matias provides the best trade-off between accuracy (weighted L_2 error) and running time



Experimental Study

Unbiased Dyadic Range Sum Query Workload

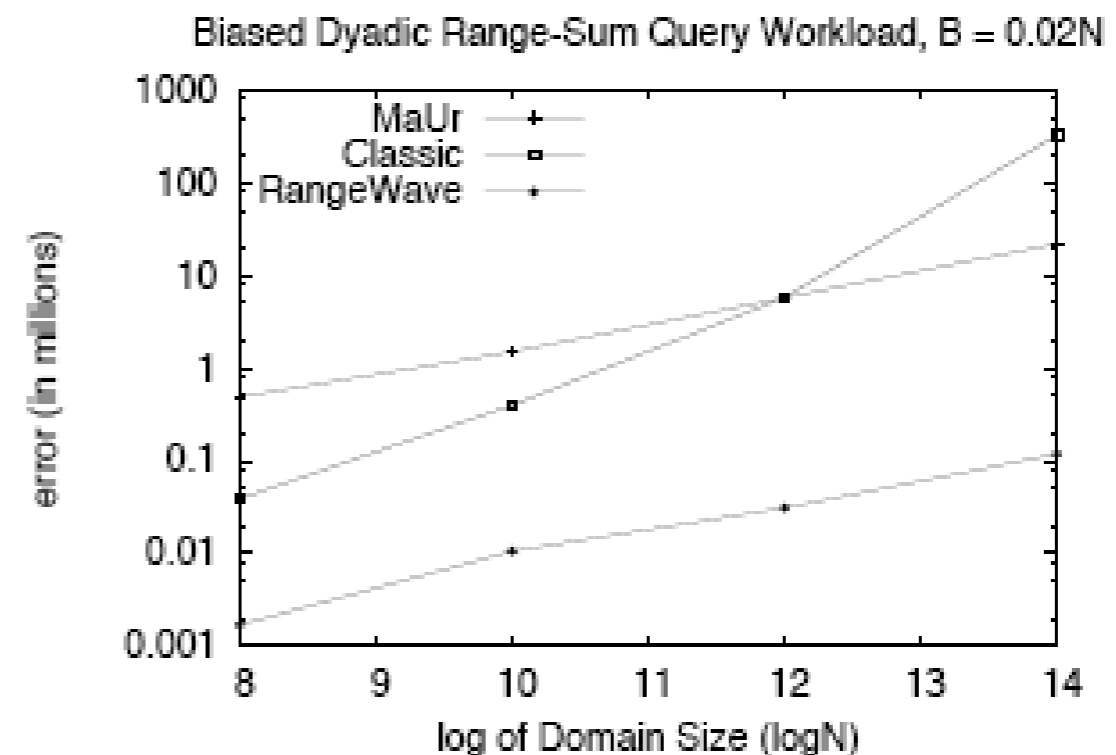
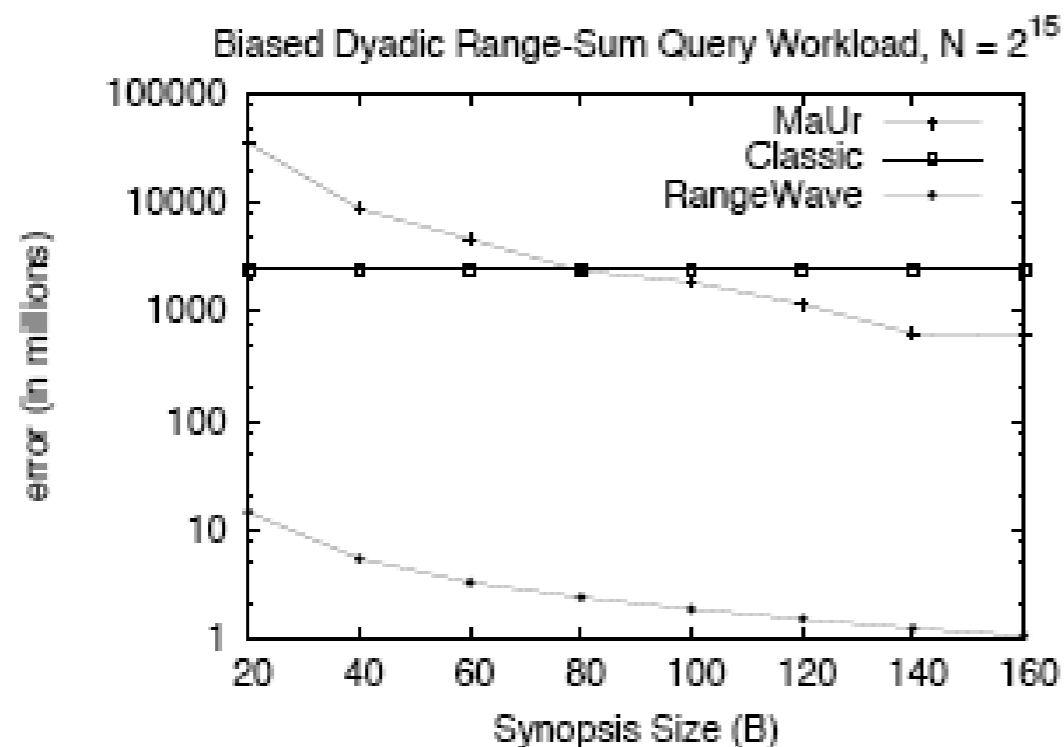
- **RangeWave** exhibits significant **accuracy gains** as the synopsis size increases for this workload
- Classic still performs well



Experimental Study

Biased Dyadic Range Sum Query Workload

- **Biased Workload** : Assigns more significance to larger range-sum queries
- The accuracy of RangeWave is orders of magnitude higher



Conclusions

- **Point** Query Workloads: You Get What You Pay

Quadratic algorithms outperform linear ones in accuracy, at a high price

- **Range Sum** Query Workloads: We can do **better**

Find a linear time algorithm for all Range Sum Queries

Extend RangeWave to general hierarchy of queries

Thank You