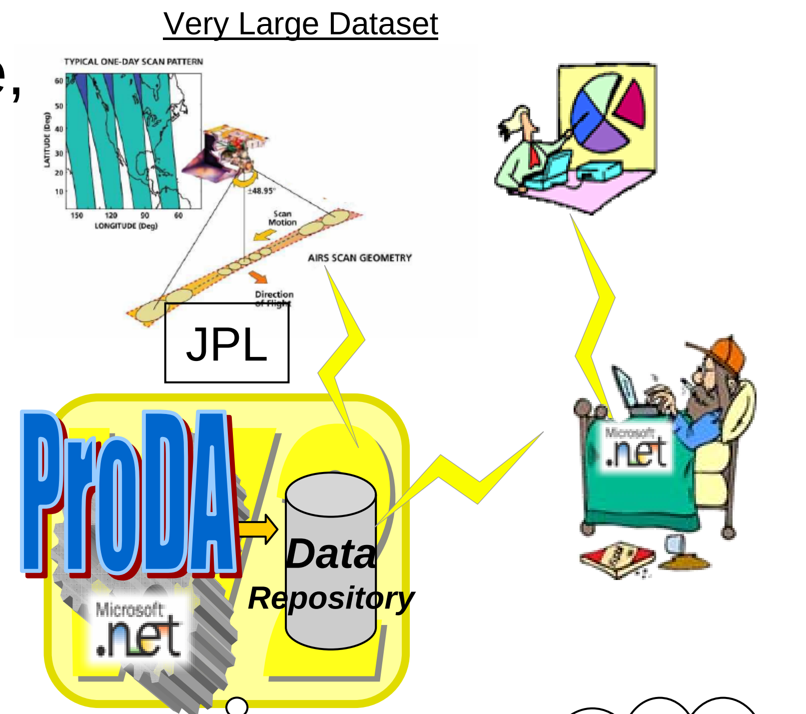
Essentials for Modern Data Analysis Systems

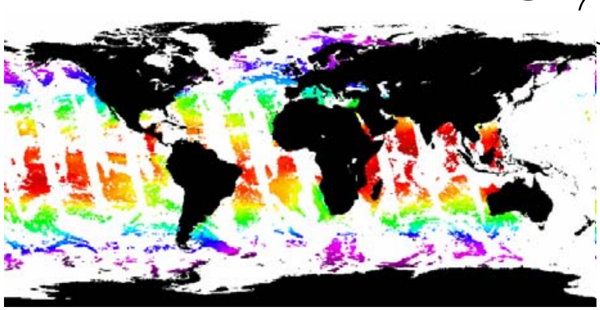
Second NASA Data Mining Workshop

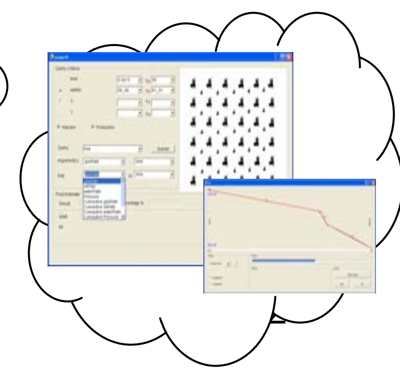
Mehrdad Jahangiri, Cyrus shahabi
University of Southern California
Los Angeles, CA 90089
{jahangir,shahabi}@usc.edu

Motivation

- Multidimensional Data
  - <latitude, longitude, altitude, time, temperature, water vapor, …>
- Massive Datasets
  - AIRS level2 data for less than a year is 320 GB
- High rate updates/increments
- Complex aggregate query
  - Variance, Correlation
- Queries are not known
- Fast approximate/exact







Motivation

- Multidimensional data
- Large data
- Aggregate queries
- Approximate answers
- Progressive answers
- Multi-resolution view

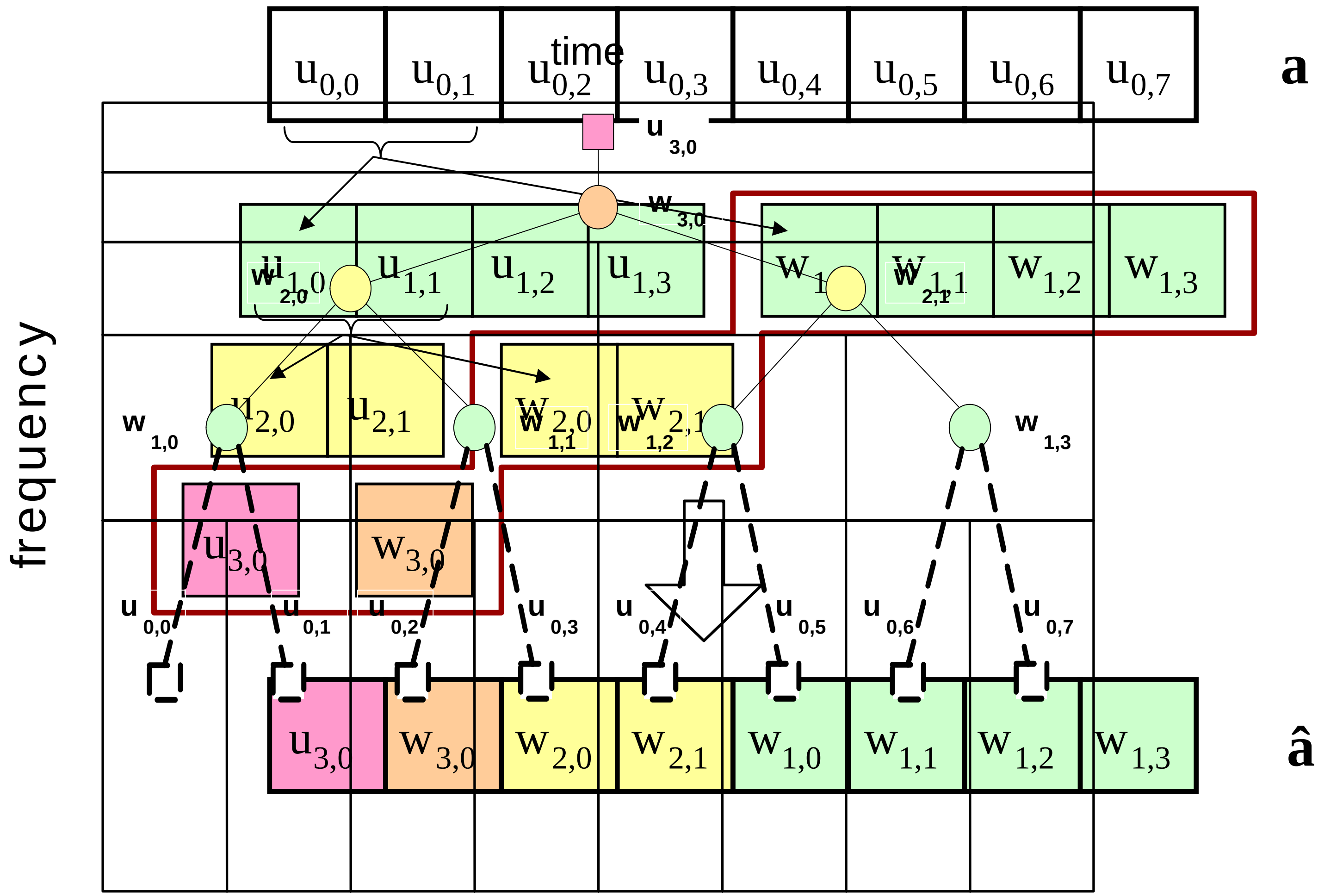
- **Wavelets!**

Outline

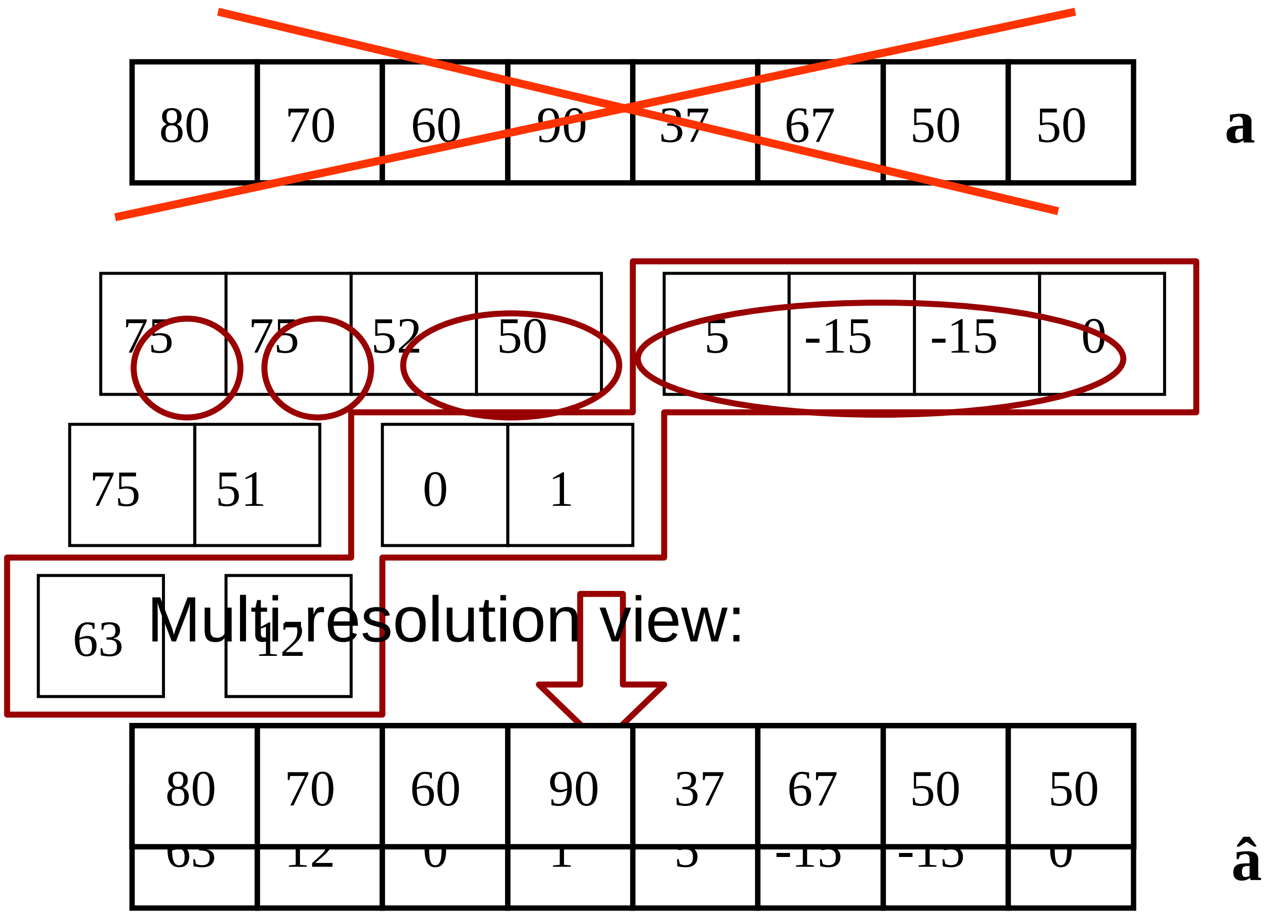
- Motivation ✓
- Introduction to Wavelets
- Related Work
- Our approach
- Essentials
- Conclusion & Future Work

Discrete Wavelet Transform

Wavelet Tree



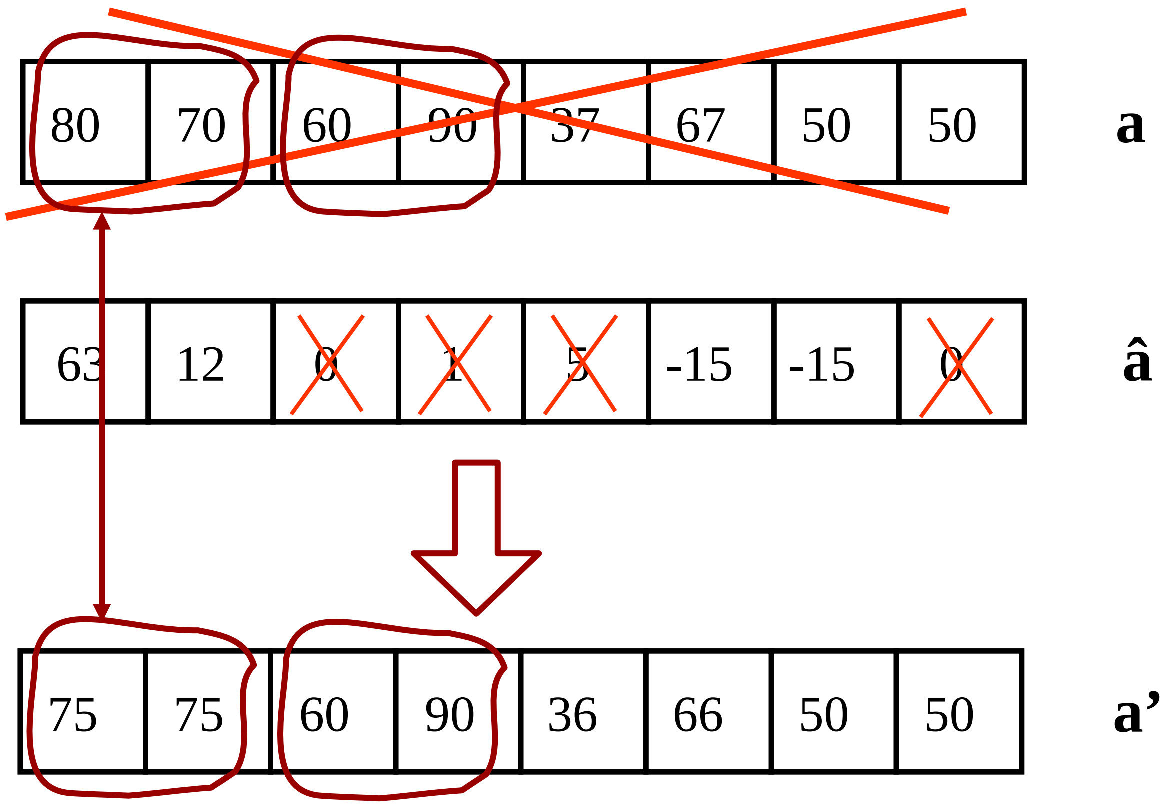
DWT Example


80 70 60 90 37 67 50 50
a
75 75 52 50
5 -15 -15 0
75 51
0 1
63
12
Multi-resolution view:
80 70 60 90 37 67 50 50
63 12 0 1 5 -15 -15 0
â

Everybody else's idea

- Let's compress data
  - Reason: save space? *(no not really!)*
  - Implicit reason: queries deal with smaller datasets and hence faster *(not always true!)*
  - More problems:
    - Only approximate results!
    - Different error rates for different queries!
      - Why? At the data population time, we don't know which coefficients are more/less important to our queries!
      - (Reminder: Different than the signal-processing objective to reconstruct the entire signal as good as possible)

Naïve use of Wavelets



Outline

- Motivation ✓
- Introduction to Wavelets ✓
- Related Work ✓
- Our approach
- Essentials
- Conclusion & Future Work

Our idea/distinction

- Storage is cheap and queries are ad-hoc; let's keep all the wavelet coefficients! (data compression is <u>not required</u>!)
- Define range-sum query as dot product of *query vector* and *data vector*
- At the query time, however, we have the knowledge of what is important to the pending query

Enabling Query in Wavelets

- Offline: Multidimensional wavelet transform of data
- At the query time: "*lazy*" wavelet transform of query vector (very fast)
- Dot product of query and data vectors in the transformed domain ➔ exact result
- Choose high-energy query coefficients only ➔ fast approximate result (90% accuracy by retrieving < 10% of data)
- Choose query coefficients in order of energy ➔ progressive result

Our query method

Original	Wavelet*

a

80	70	60	90	37	67	50	50

→ **â**

178.1	33.9	0	2	7.07	-21.2	-21.2	0

1	1	1	1	1	1	1	1

→

2.83	0	0	0	0	0	0	0

Result=504

Result=178.19*2.83=504

0	0	1	1	1	1	1	0

→

1.73	-.35	-1	.5	0	0	0	.71

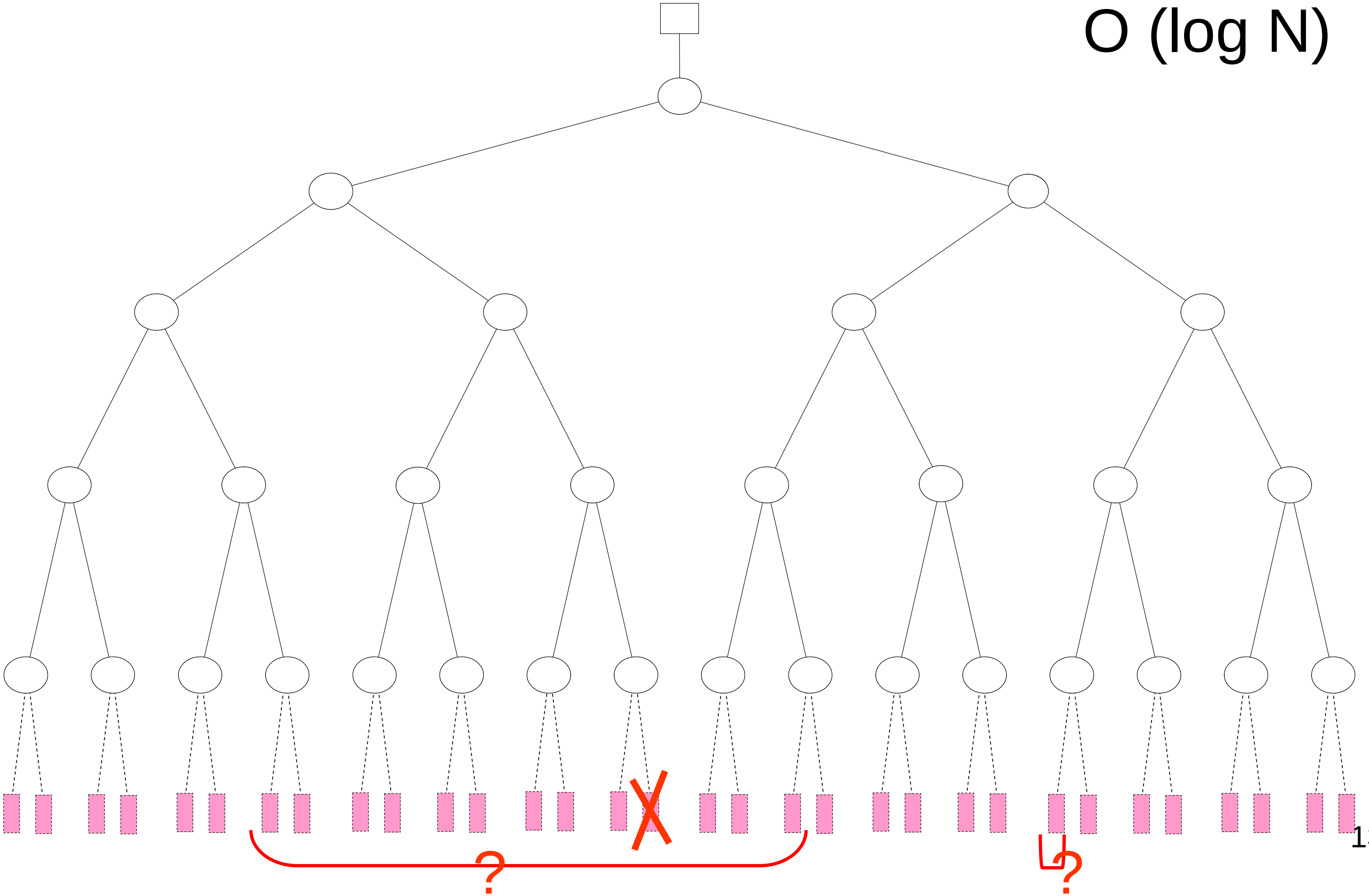
Result=304

Result=178.19*1.73+33.94*(-.35)+2*.5 =304

* Let's normalize our filters from **{1/2, 1/2}** and **{1/2, -1/2}** to **{1/√2, 1/√2}** and **{1/√2, -1/√2}**

Complexity

O (log N)



Outline

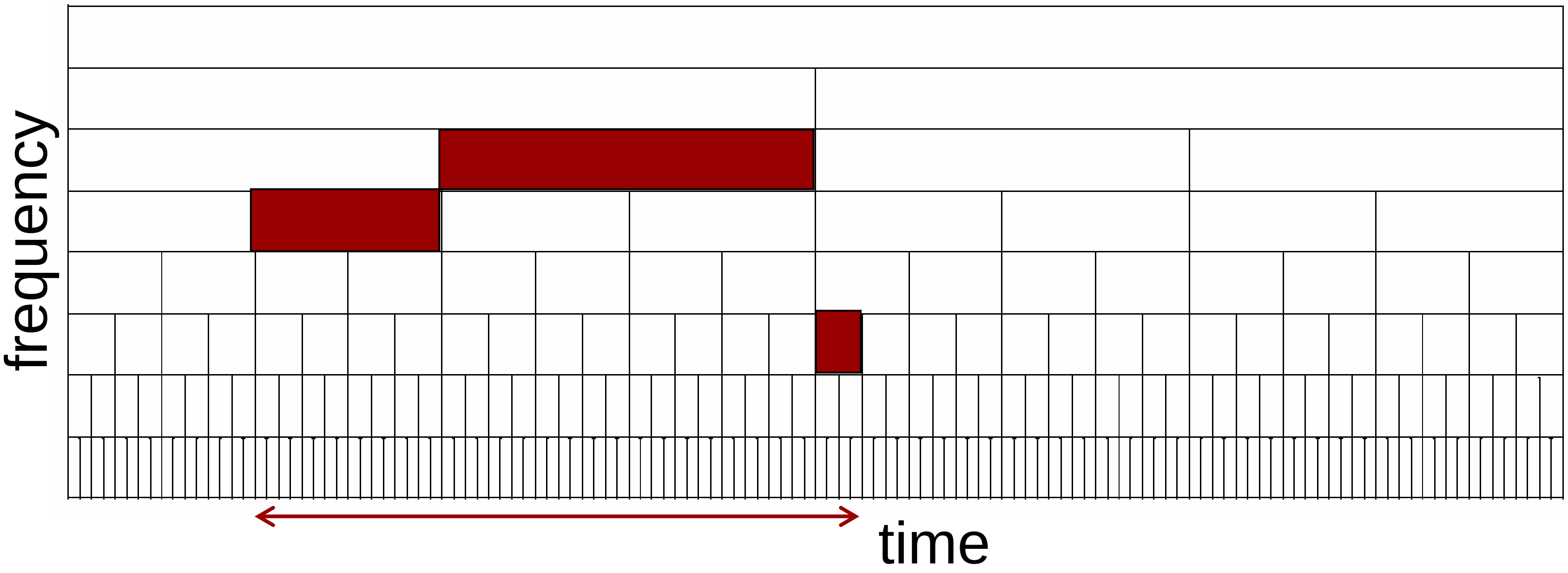
- Motivation ✓
- Introduction to Wavelets ✓
- Related Work ✓
- Our approach ✓
- Essentials
- Conclusion & Future Work

Essentials

- Multi-resolution
- Progressive
- Polynomial Queries
- Batch of Queries
- Large Multidimensional Datasets
- Archive & Synopsis
- System Architecture

Multi-resolution

- *Summarize the data at various levels of abstractions on-the-fly*
  - *e.g. daily, weekly, monthly, quarterly, yearly*
- *Access the finest resolution of the data with no extra cost*



Progressive

- Desired accuracy varies per
  - Application
  - User
  - Dataset
- Accuracy is traded-off for faster response time
- *We propose [CSMJ'05]*
  - *A class of wavelet coefficient orderings*
    - *To provide a near optimal progressive query answering.*
  - *Error forecasting*
    - *To estimate the accuracy of the generated approximate results.*

Polynomial Queries

- Predefined statistical range-aggregate queries
  - Simple: count, sum, average
  - Complex: Variance, Covariance, Correlation
- Queries can not be predicted by database designers!
- *We introduce a novel technique that supports any polynomial range-sum query [RSCS'02]*
  - *We treat all dimensions, including measure dimensions, symmetrically*
  - *Query measure can be any **polynomial** in the data dimensions.*

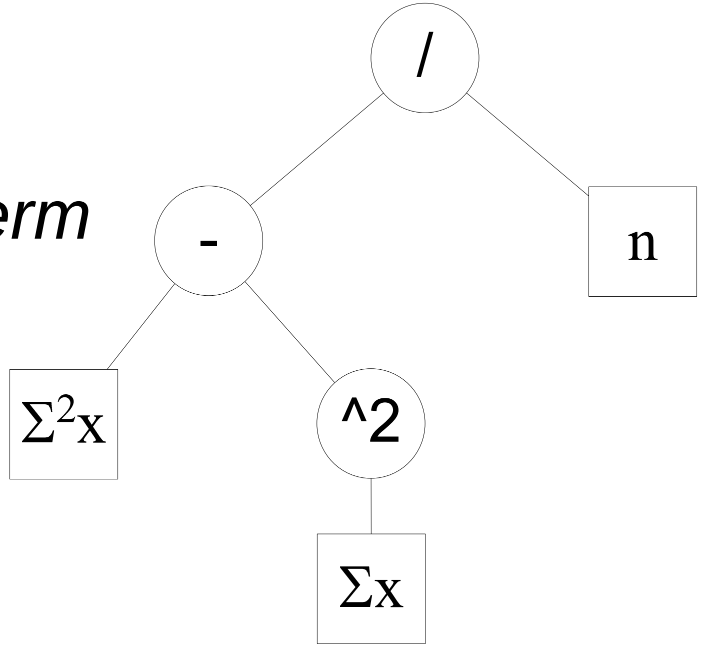
Polynomial Queries

- Consider *Var(x)*

$$Var(x) = \frac{\sum x_i^2 - (\sum x_i)^2}{n}$$

- Let us re-write *Var(x)* in post order:

$$\sum x_i^2, \sum x_i, \wedge 2, -, n, /$$



- Generate this query by *PushTerm* and *PushOperator* calls:
  - PushTerm(1,{2});
  - PushTerm(1,{1});
  - PushOperator('^2');
  - PushOperator('-');
  - PushTerm(1,{0});
  - PushOperator('/');

Batch of queries

- Most scientists typically submit batches of queries simultaneously rather than issuing individual, unrelated queries.
  - e.g. draw a grid, ask for the average measure value per grid cell
- *By exploiting I/O sharing across a query batch, we evaluate a group of queries progressively and efficiently[RSCS2'02].*
  - *We ensure that the structure of error across cells*

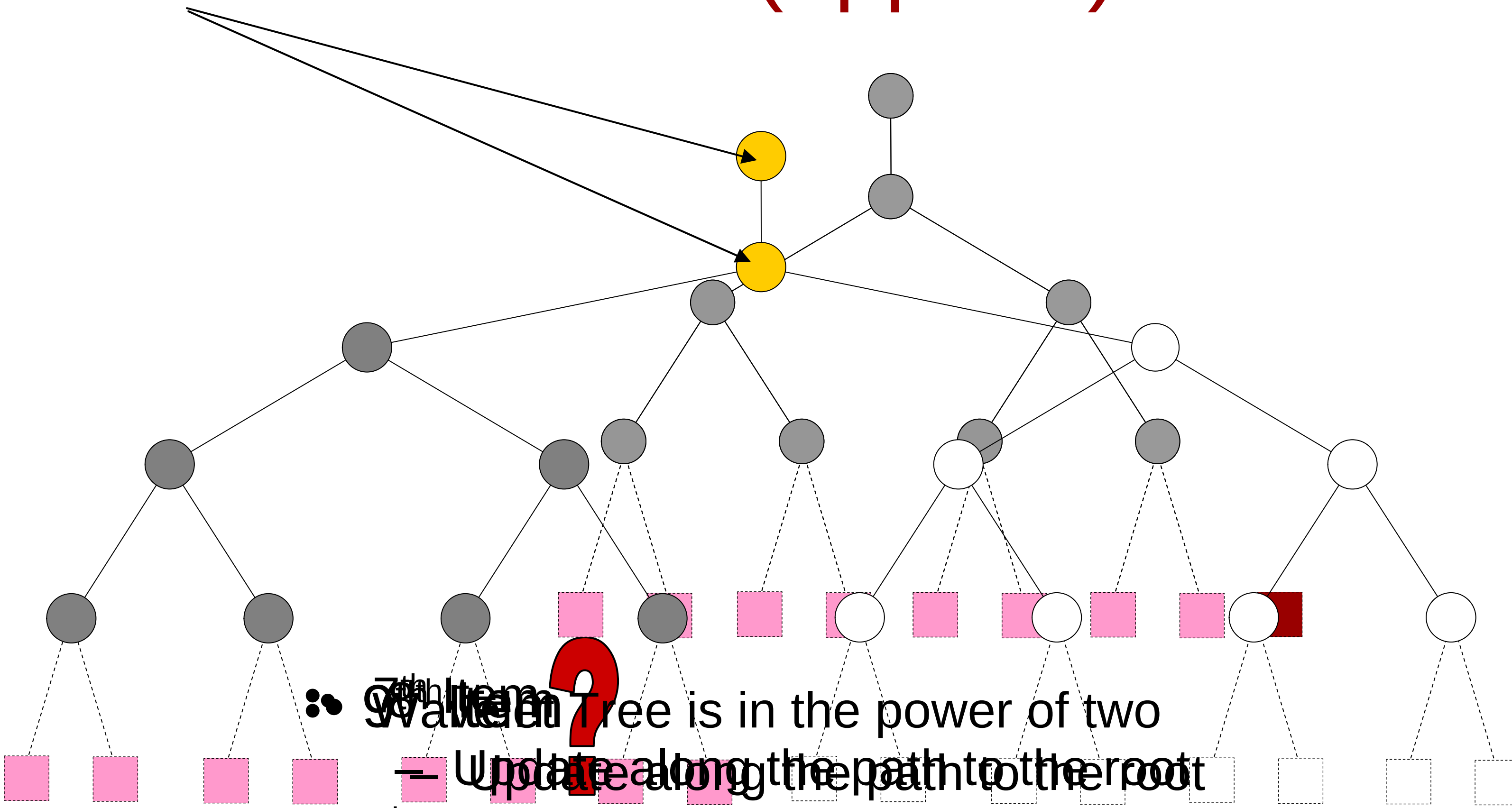
Large Multidimensional Datasets

- NASA datasets are massive (TBs of data)
- Multidimensional data
  - e.g. Longitude, Latitude, pressStd, Time, TAirStd,H2OMMStd,O3VMRStd
- *How to DWT this massive data?*
  - *Use SHIFT-SPLIT [MJDS'05]*
- *How to DWT the query on-the-fly?*
  - *Use Lazy Wavelet Transform [RSCS'02]*

Archive and Synopsis

- High rate data stream of AIRS
  - Archive
  - Synopsis
- Appending is the increase of the domain of one or more dimensions (different from update)
  - *Expanding from M to N using SHIFT-SPLIT[MJDS'05]:*
    - *Computational cost: O (M+log (N/M))*
- Maintenance of a wavelet synopsis
  - *Buffering B coefficients[MJDS'05]:*
    - *Space: O(K+log N/B+B)*
    - *Per-item computational cost: O(1/B log N/B)*

Archive (append)

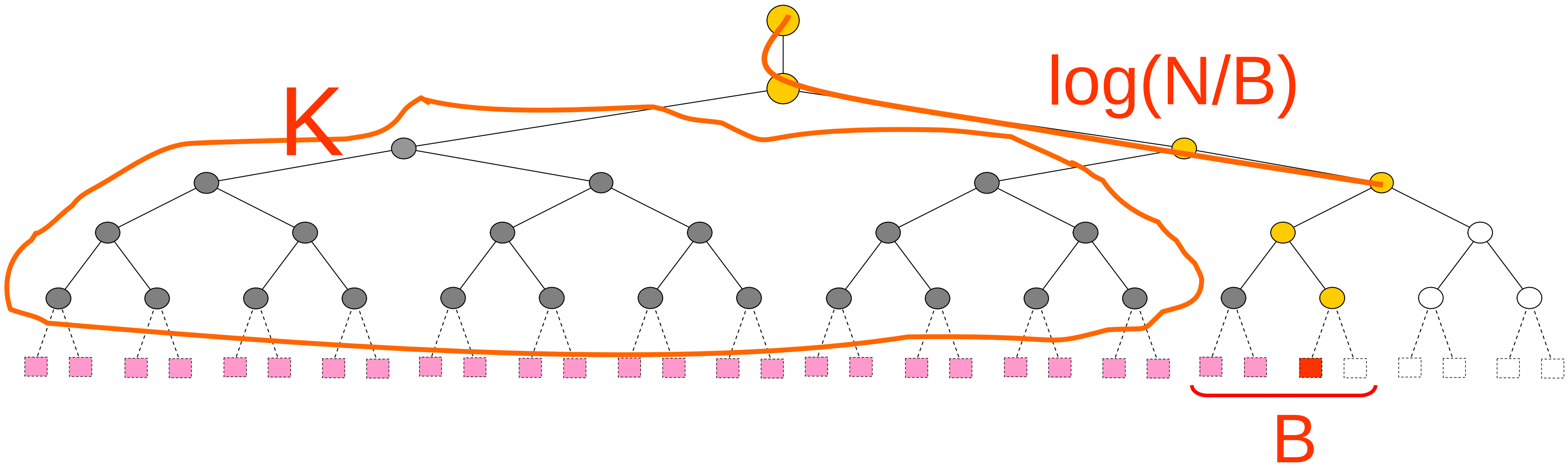


- 7th Item
- 8th Item
- Wavelet Tree is in the power of two
  - Update along the path to the root
  - Update along the path to the root
- 9^{th} Item
  - Double the tree
    - SPLIT all wavelet coefficients
    - SPLIT the average coefficient
  - Update along the path

Expanding from size M to N: $O(M + \log\frac{N}{M})$

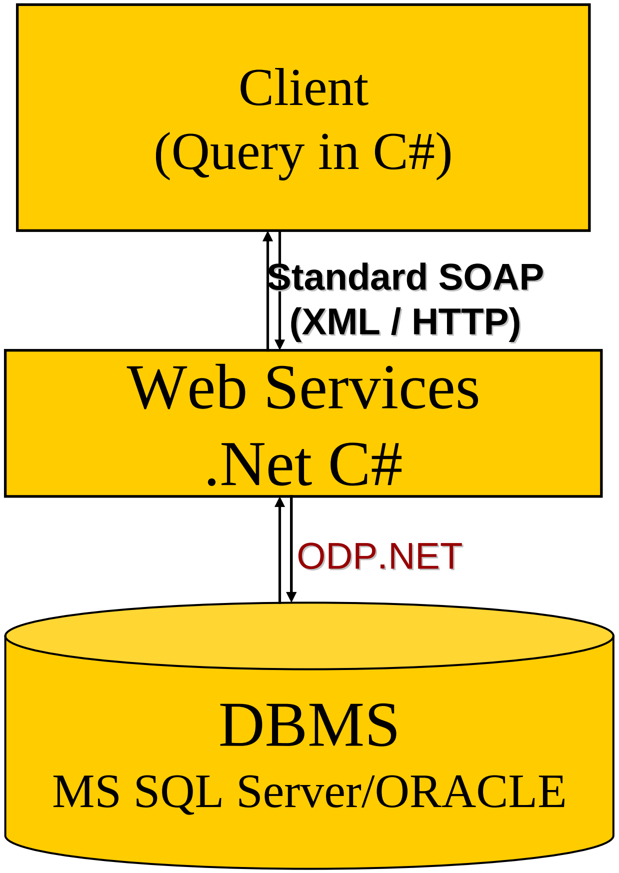
Synopsis (Data Stream Approximation)

- Buffering B coefficients
  - Space: O(K+log N/B+B)
  - Per-item computational cost: O(1/B log N/B)



System Architecture

- System requirements
  - Computations at server side
    - Process sharing
    - Result caching
    - Insignificant data transfer
  - Cross-platform (portability)
  - Network-wide accessibility
  - Code re-usability
  - Enabling rapid application development
- *A suite of web services for progressive data analysis (**ProDA**) [MJCS'05]*



3-Tier Architecture

A Sample Client for Progressive Query (in C#)

```
ProDA_WebServices pws=new ProDA_WebServices();
pws.Credentials = System.Net.CredentialCache.DefaultCredentials;
pws.CookieContainer = new CookieContainer();
```

- Creating an instance
- Storing session state

```
string []dbnames=pws.AllCubeNames();
for(int i=0;i<dbnames.Length;i++)
                Console.WriteLine(dbnames[i]);
```

- Listing all cube names

```
pws.SelectDB("…");
int n=pws.GetDimN();
for(int i=0;i<n;i++)
    Console.WriteLine("Dim"+i+"("+pws.GetDimTitle(i)+")="+pws.GetDimSize(i));
```

- Selecting a cube
- Asking for some properties

```
int []rStart={…};
int []rEnd  ={…};
pws.SetRange(rStart,rEnd);
pws.Var(1);
```

- Defining a range
- Submitting a query

```
while(pws.HasMore())
    Console.Write("Result="+pws.Advance(500)+"["+pws.GetPercentage()+"]\t\r");
pws.CloseConnection();
```

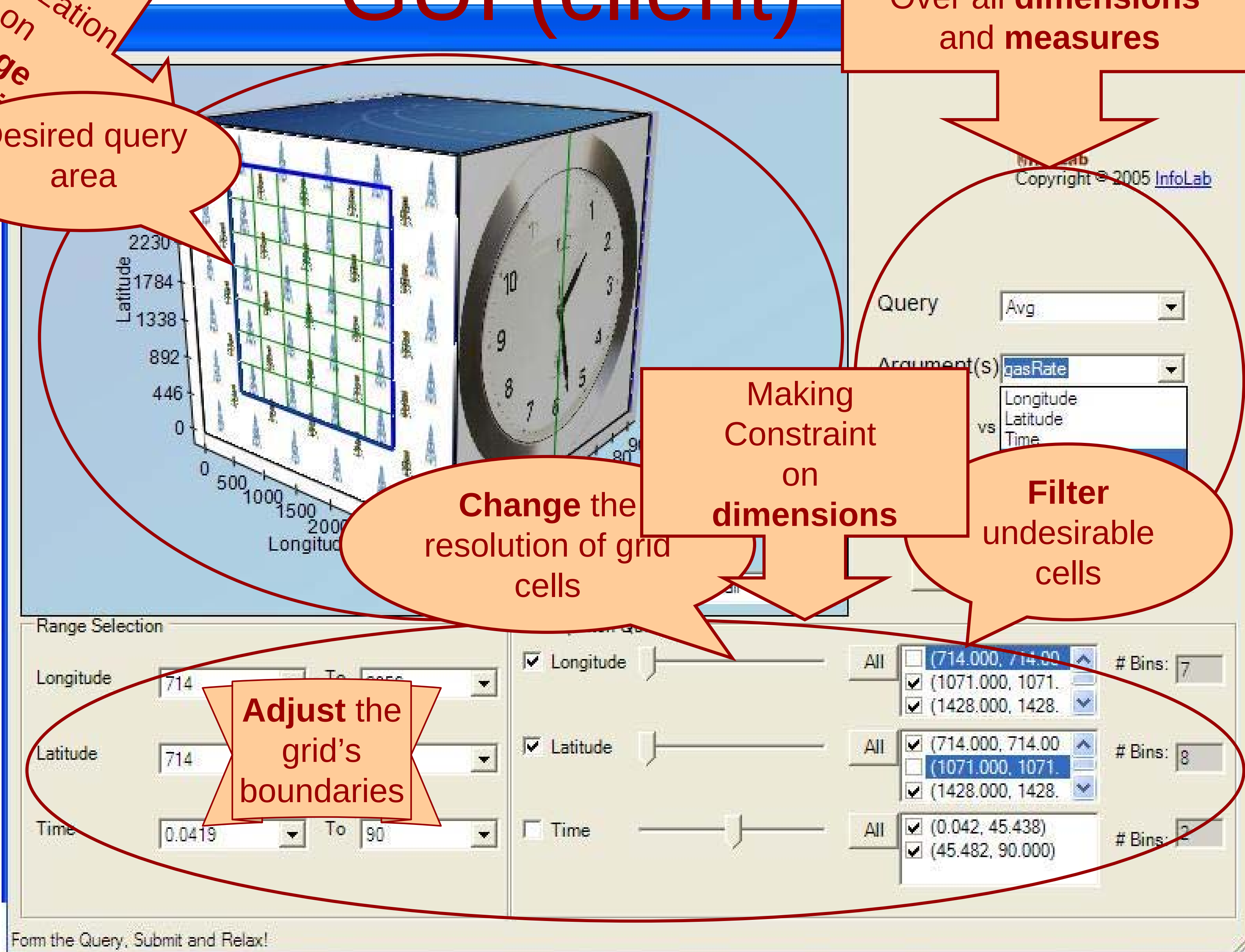
- Asking for result progressively
- Closing connection

GUI (client)

3D visualization on **range dimensions**

Different types of range aggregate queries Over all **dimensions** and **measures**

Desired query area



Change the resolution of grid cells

Making Constraint on **dimensions**

Filter undesirable cells

Adjust the grid's boundaries

GUI (client)

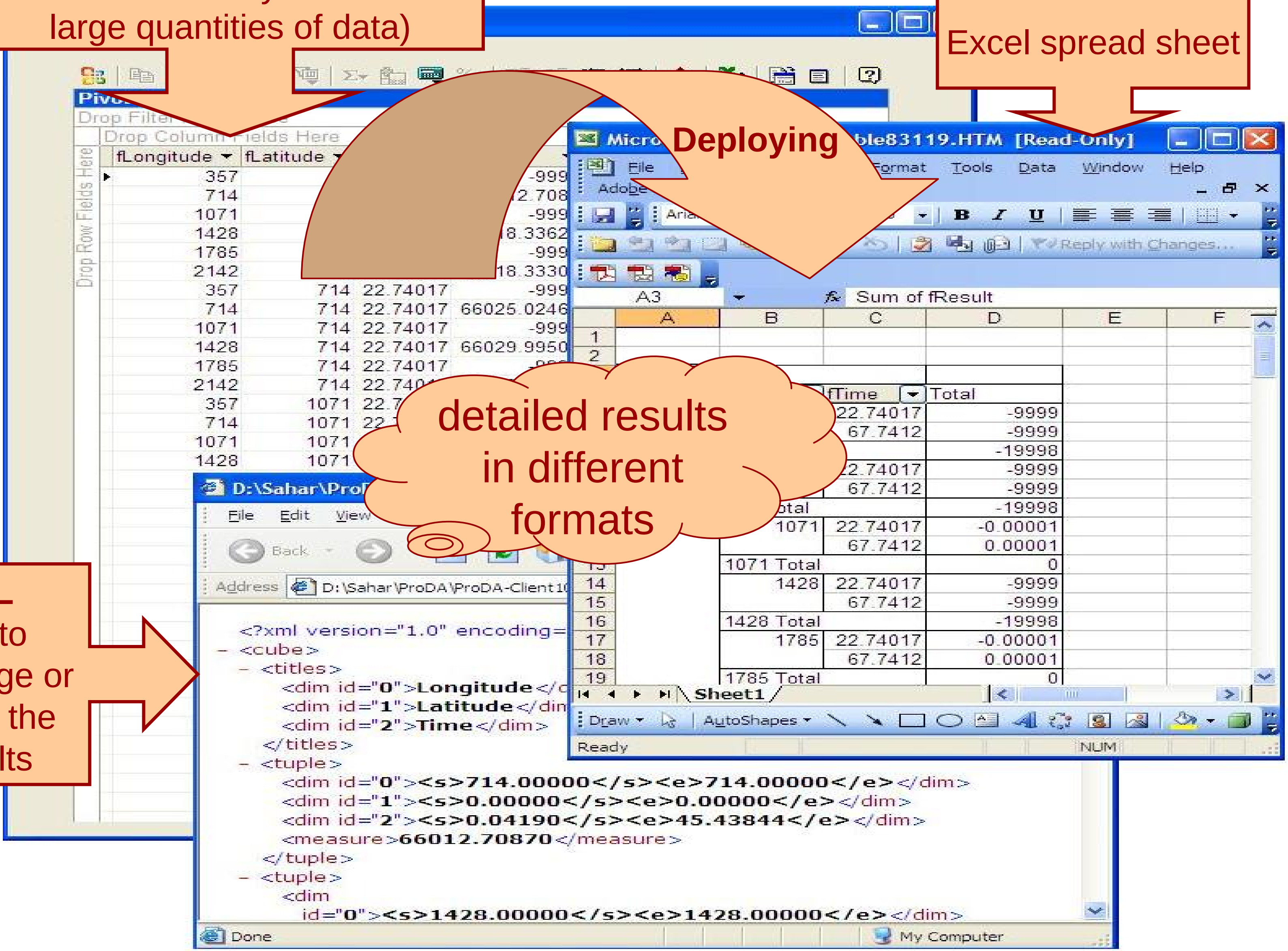
Plug in **pivot table** (powerful reporting tool for ad hoc analysis on large quantities of data)

Excel spread sheet

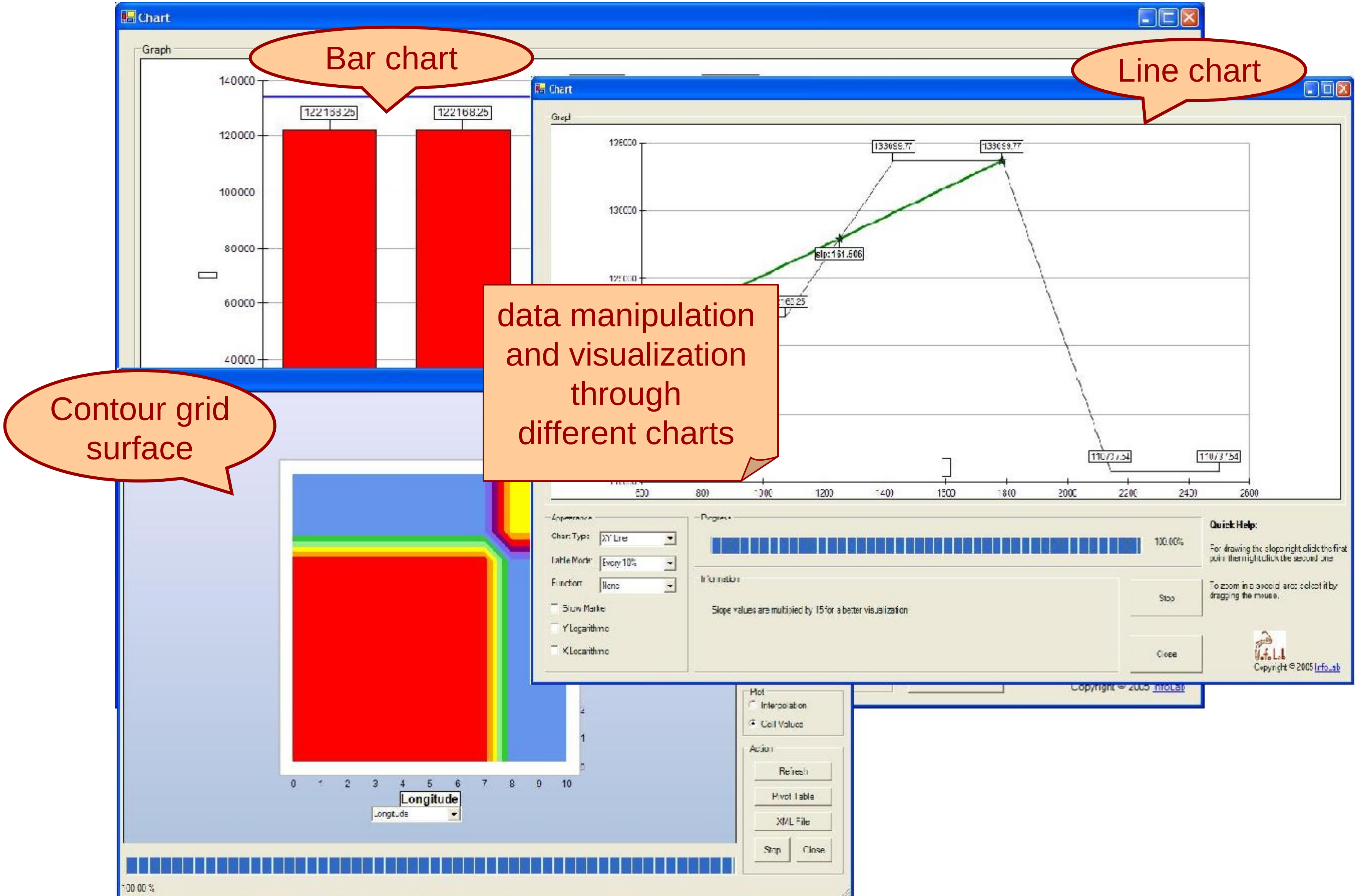
Deploying

detailed results in different formats

XML
Use to Exchange or Export the results



GUI (client)



Bar chart
Line chart
Contour grid surface
data manipulation and visualization through different charts
Chart
Graph
140000
120000
100000
80000
60000
40000
122168.25
122168.25
136000
130000
133699.77
133699.77
slp:161.506
Chart Type:
XY Line
Table Mode:
Every 10%
Function:
None
Show Marks
Y Logarithmic
X Logarithmic
100.00%
Slope values are multiplied by 15 for a better visualization
Stop
Close
Quick Help:
Longitude
Refresh
Pivot Table
XML File
Interpolation
Cell Values

Visit ProDA

ProDA Web Services:

http://mahour.usc.edu/proda/webservices.asmx

ProDA Windows Client:

http://infolab.usc.edu/projects/proda/ProDA-Chevron.zip

ProDA Website:

http://infolab.usc.edu/projects/proda/

ProDA on AIRS dataset

(complex query answering in few seconds)

- Correlation of temperature and pressure for a surface grid of 5x5
  - Let me draw the grid, download the result, and see the visualization
- Correlation of temperature and water vapor inside the golf of Mexico for all different levels of altitude
- Plot temperature vs. time for the surface of united states. Any interesting pattern?

Conclusion & Future Work

- We employ wavelets
  - exact, approximate and progressive statistical queries
  - large multidimensional datasets
  - any arbitrary polynomial query
  - operating at the server side
  - efficient maintenance
- We need to investigate:
  - Efficient storage and retrieval of sparse datasets
  - Queries across different datacubes
    - e.g. correlation of temperature between GPS and AIRS datasets

References

- [MJDS'05]M. Jahangiri, D. Sacharidis, and C. Shahabi. **SHIFTSPLIT: I/O Efficient Maintenance of Wavelet-Transformed Multidimensional Data.** In Proceedings of ACM SIGMOD, 2005.
- [MJCS'05]M. Jahangiri and C. Shahabi. **ProDA: A Suite of WebServices for Progressive Data Analysis.** In Proceedings of ACM SIGMOD (demonstration), 2005.
- [CSMJ'05]C. Shahabi, M. Jahangiri, and D. Sacharidis. **Hybrid Query and Data Ordering for Fast and Progressive Range-Aggregate Query Answering.** International Journal of Data Warehousing and Mining, 1(2):49–69,April-June 2005.
- [RSCS2'02]R. Schmidt and C. Shahabi. **How to evaluate multiple range-sum queries progressively.** In Proceedings of ACM PODS, pages 3–5.
- [RSCS'02]R. Schmidt and C. Shahabi. **Propolyne: A fast wavelet-based technique for progressive evaluation of polynomial range-sum queries.** In Proceedings of EDBT, 2002.
- [CSXT'00]C. Shahabi, Xiaoming Tian, Wugang **Zhao,TSA-tree:A Wavelet-Based Approach to Improve the Efficiency of Multi-Level Surprise and Trend Queries on Time-series Data,** SSDBM 2000
- [PRODA]ProDA: http://infolab.usc.edu/projects/proda/

Thank you

(visit http://infolab.usc.edu)