GridMiner

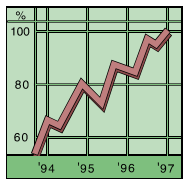
A Framework for Knowledge Discovery on the Grid – Scientific Drivers and Contributions

Peter Brezany

University of Vienna
Institute for Software Science

brezany@par.univie.ac.at

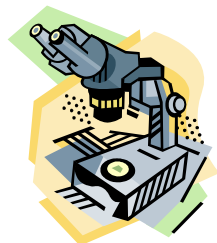
Motivation



Business

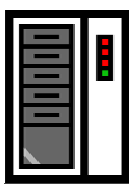
Medicine

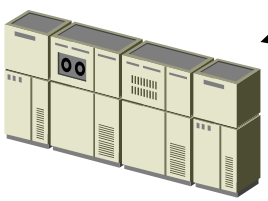
Scientific experiments



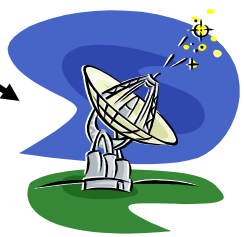
Data and data exploration cloud

Simulations





Earth observations

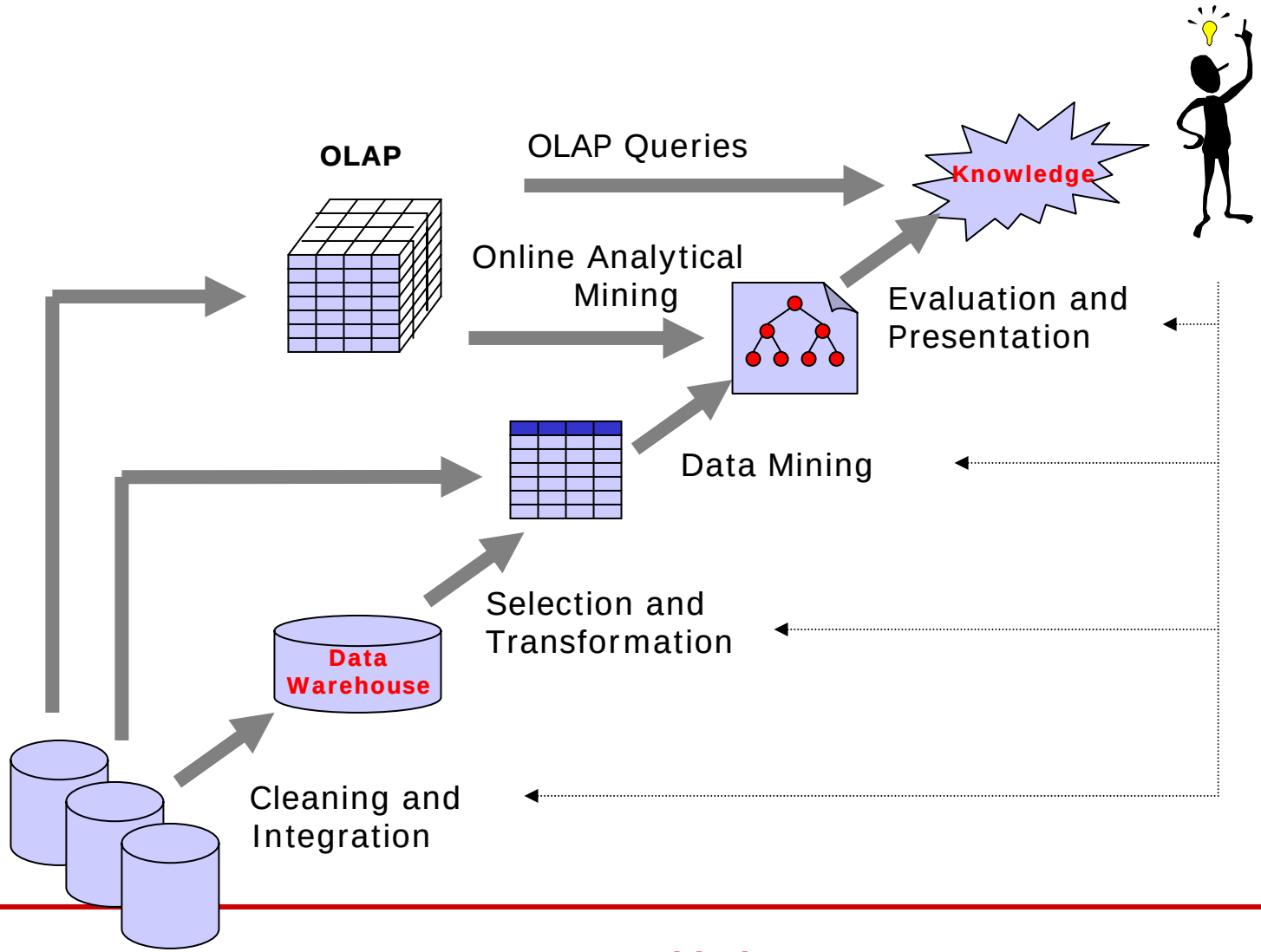


Stages of a Data Exploration Project

Based on:
Data Preparation for Data Mining, by Dorian Pyle, Morgan Kaufmann

	Time to complete (percent of total)		Importance to success (percent of total)	
1. Exploring the problem	10	20	15	80
2. Exploring the solution	9		14	
3. Implementation specification	1		51	
4. Knowledge discovery		80		20
a. Data preparation	60		15	
b. Data surveying	15		3	
c. Data modeling	5		2	

The Knowledge Discovery Process



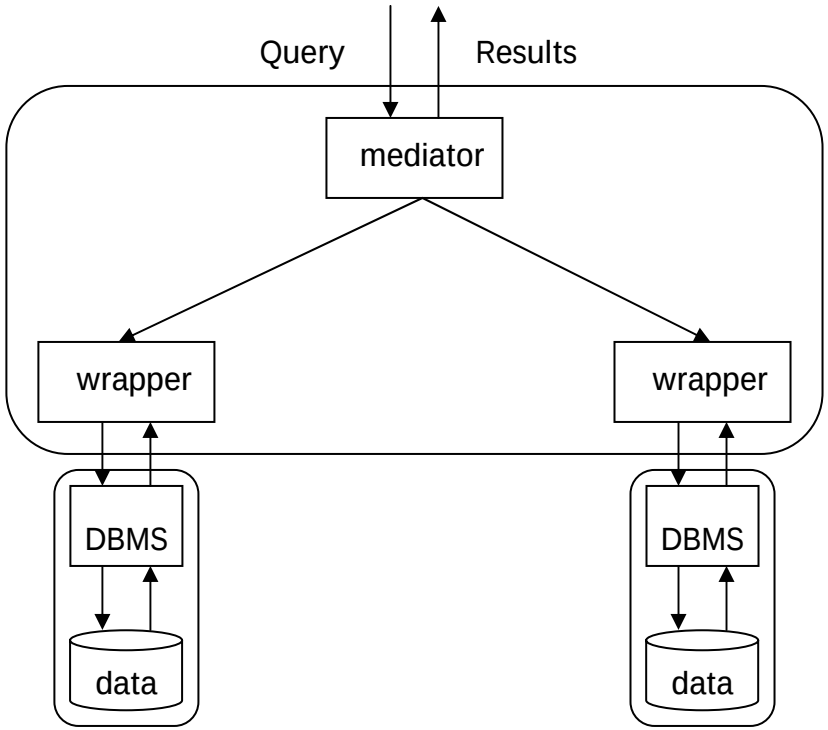
Outline

- Introduction/Motivation
- **What Does the Grid Offer to Knowledge Discovery Processes?** ←
- **Applications Addressed**
- **Novel Challenges**
- **Research Results Summary**
- **Conclusions**

The Grid offers.. (1)

- **Resource virtualisation:**
  - Dynamic Data and Computational Resource Discovery using service registries;
  - Mechanisms for dynamic resource
    - Allocation
    - Monitoring (MDS, NWS, etc.)
- **Systematic access to resources addressing:**
  - Security
  - Authentication
  - Authorization

The Grid offers.. (2)



- **Database access services**
- **Distributed query processing**
- **Data integration services** - the wrappers reconcile differences and impose a global schema.

The Grid offers.. (3)

- **Support for Job (Operation) Management**

  (important, e.g., for long-running data preprocessing)

Notification interfaces

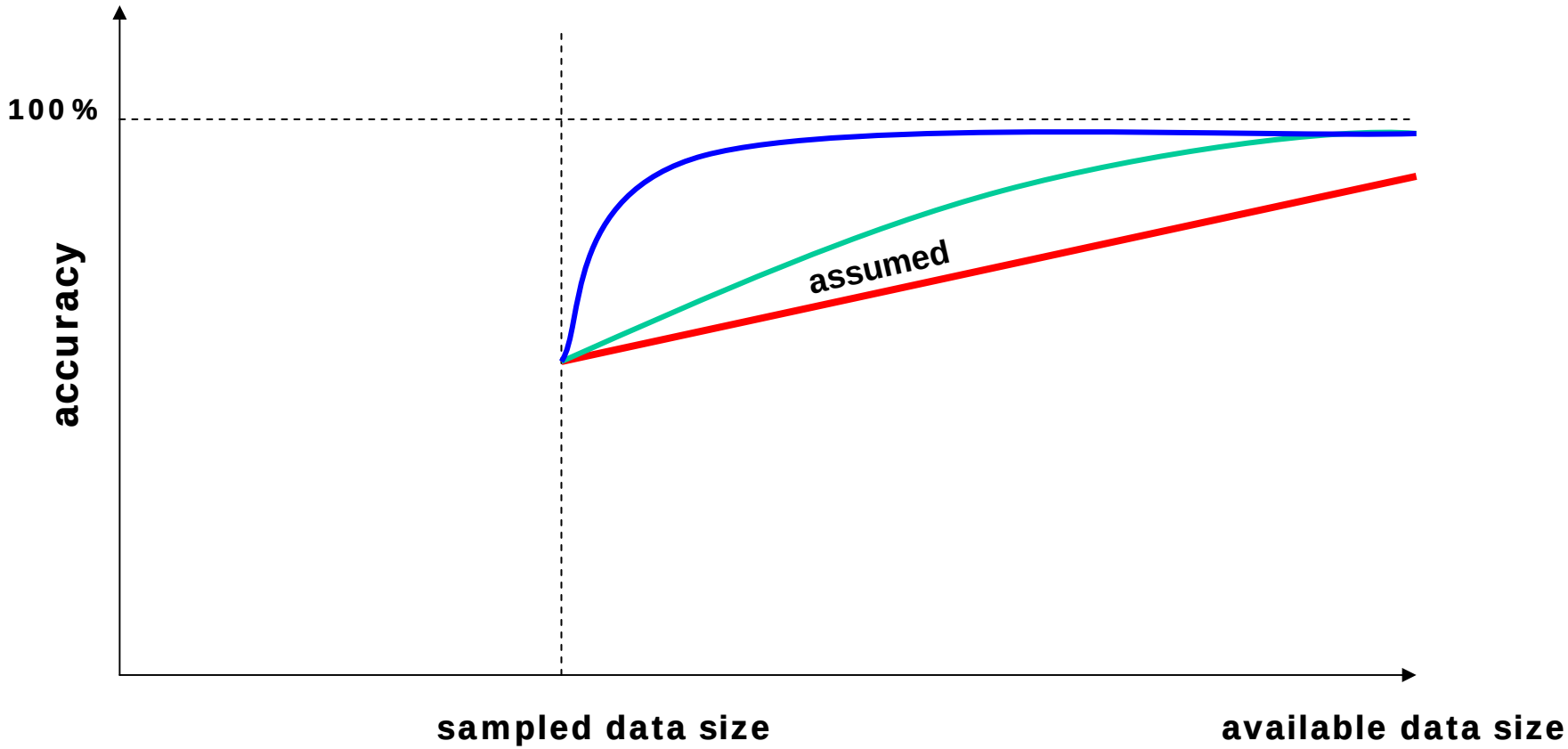
- NotificationSource for client subscription
- NotificationSink for asynchronous delivery of notification messages

The Grid offers.. (4)

Theoretically, the Grid can have unlimited size (the number of data and computational resources) **–** support for scaling up

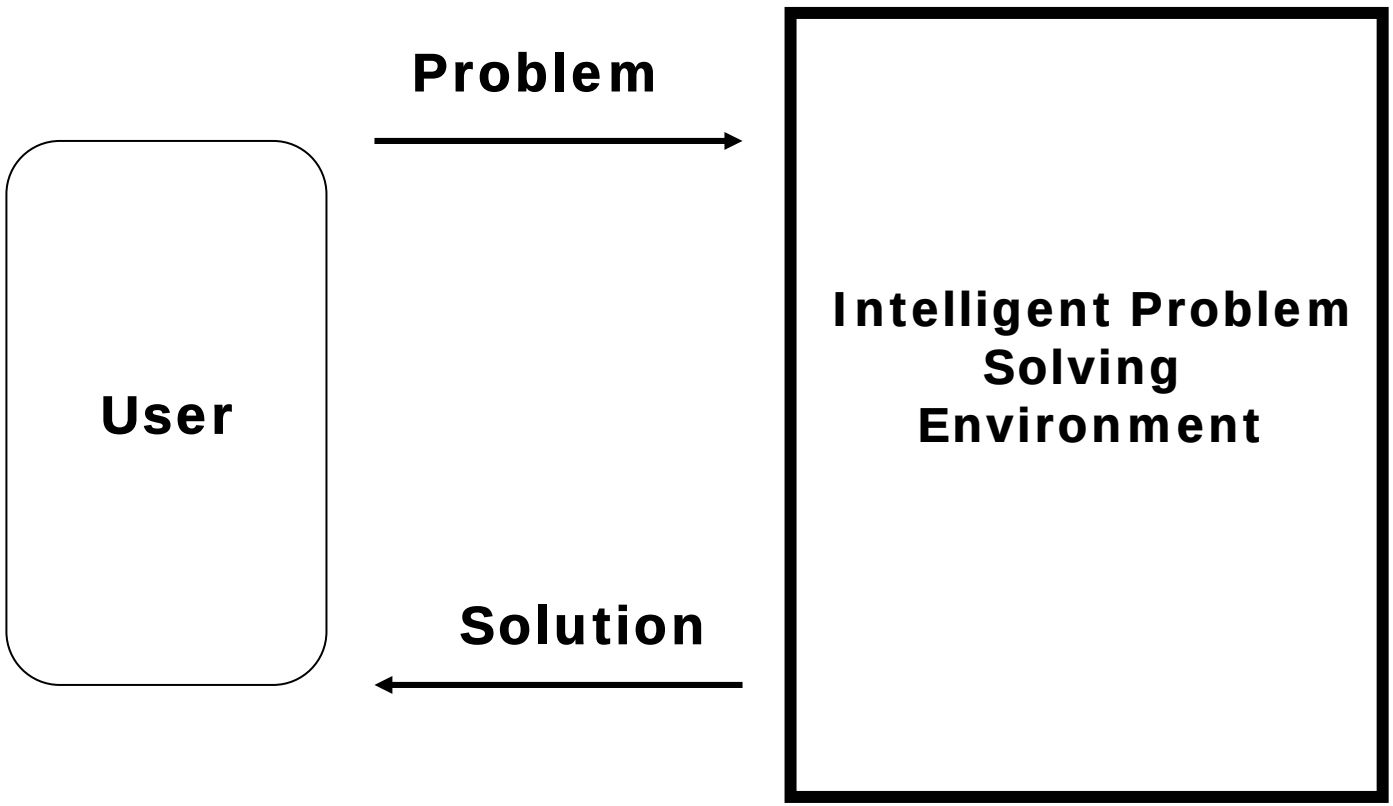
- **Questions:**
  - When is it necessary to mine huge databases, as opposed to mining a sample of the data?
  - Should not data mining algorithms be able to take advantage of all the data that is available?
- **Answers:**
  - Scaling up is desirable, because increasing the size of the training set often increases the accuracy of induced classification models.
  - Determining how much data to use is difficult, because the smallest sufficient amount depends on factors not known a priori.
  - Today‘s mining techniques can have problems when data sets exceed 100 megabytes.

Data Mining Accuracy vs. Data Size



Novel Challenges

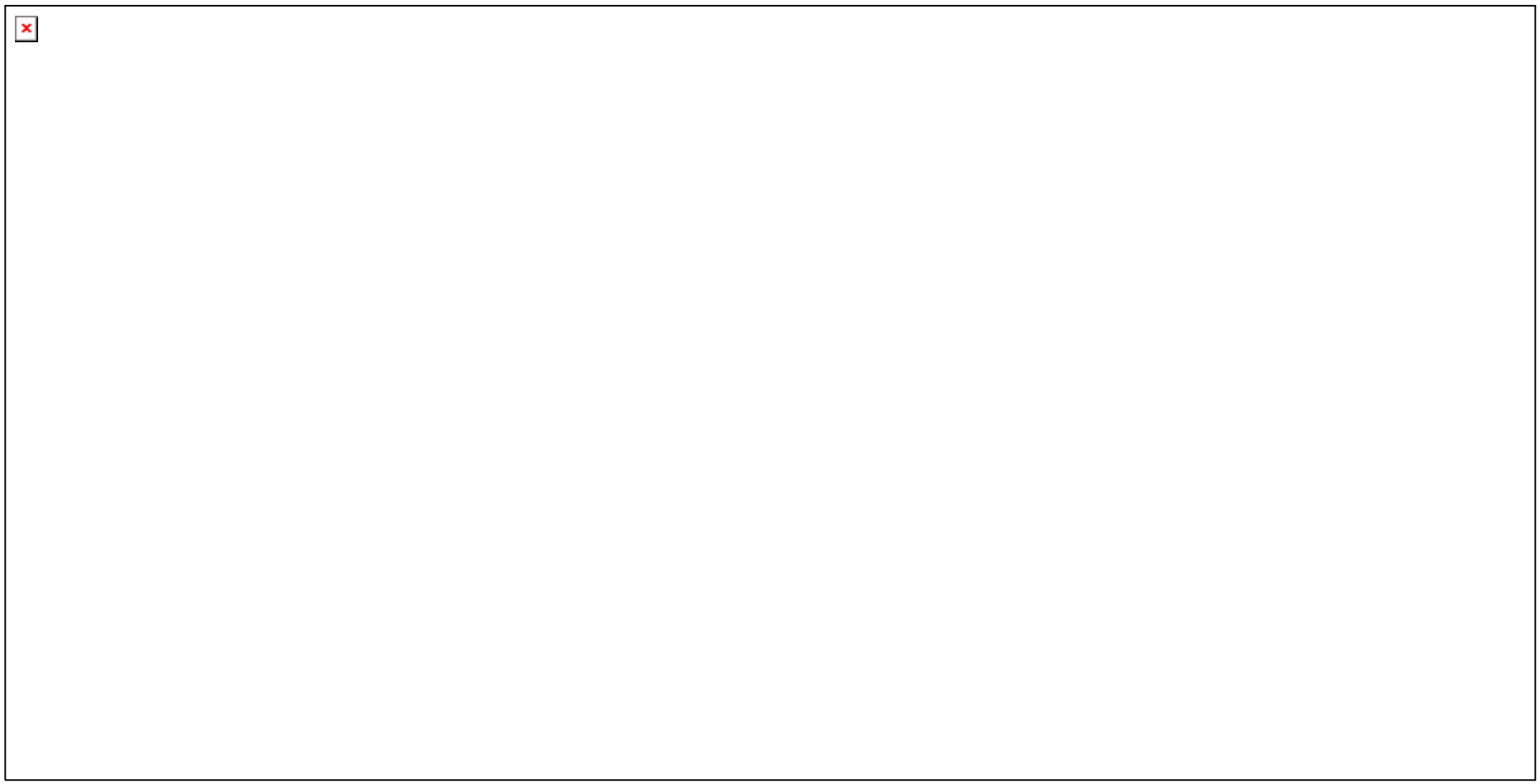
Toward Wisdom Grid / Web Infrastructures



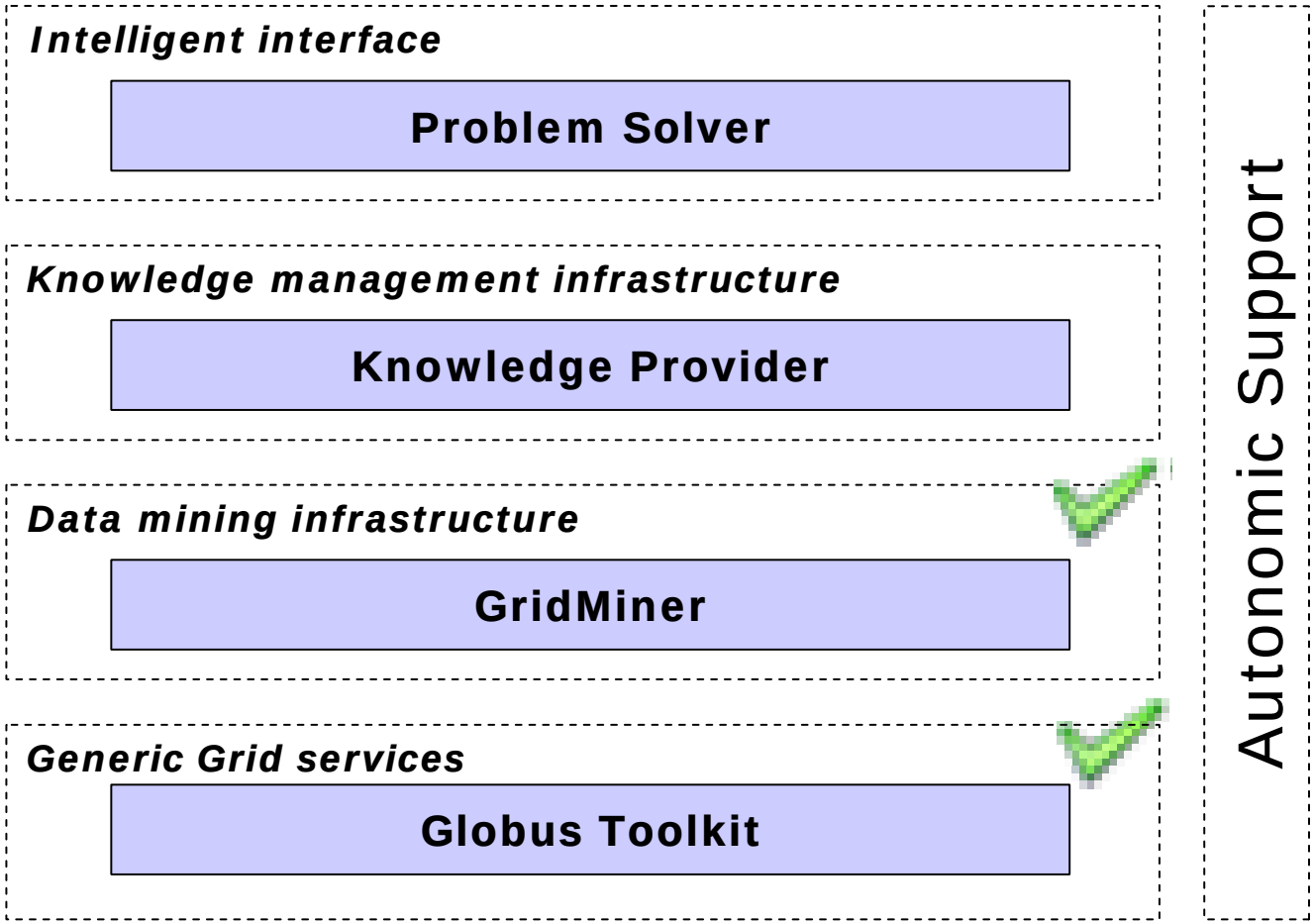
Traumatic Brain Injury Application

- Traumatic brain injuries (TBIs) typically result from accidents in which head strikes an object.
- The treatment of TBI patients is very resource intensive.
- The trajectory of the TBI patients management:
  - Trauma event
  - First aid
  - Transportation to hospital
  - Acute hospital care
  - Home care
- All the above phases are associated with data collection into databases – now managed by individual hospitals.

Scenario – Traumatic Brain Injury (TBI) Application



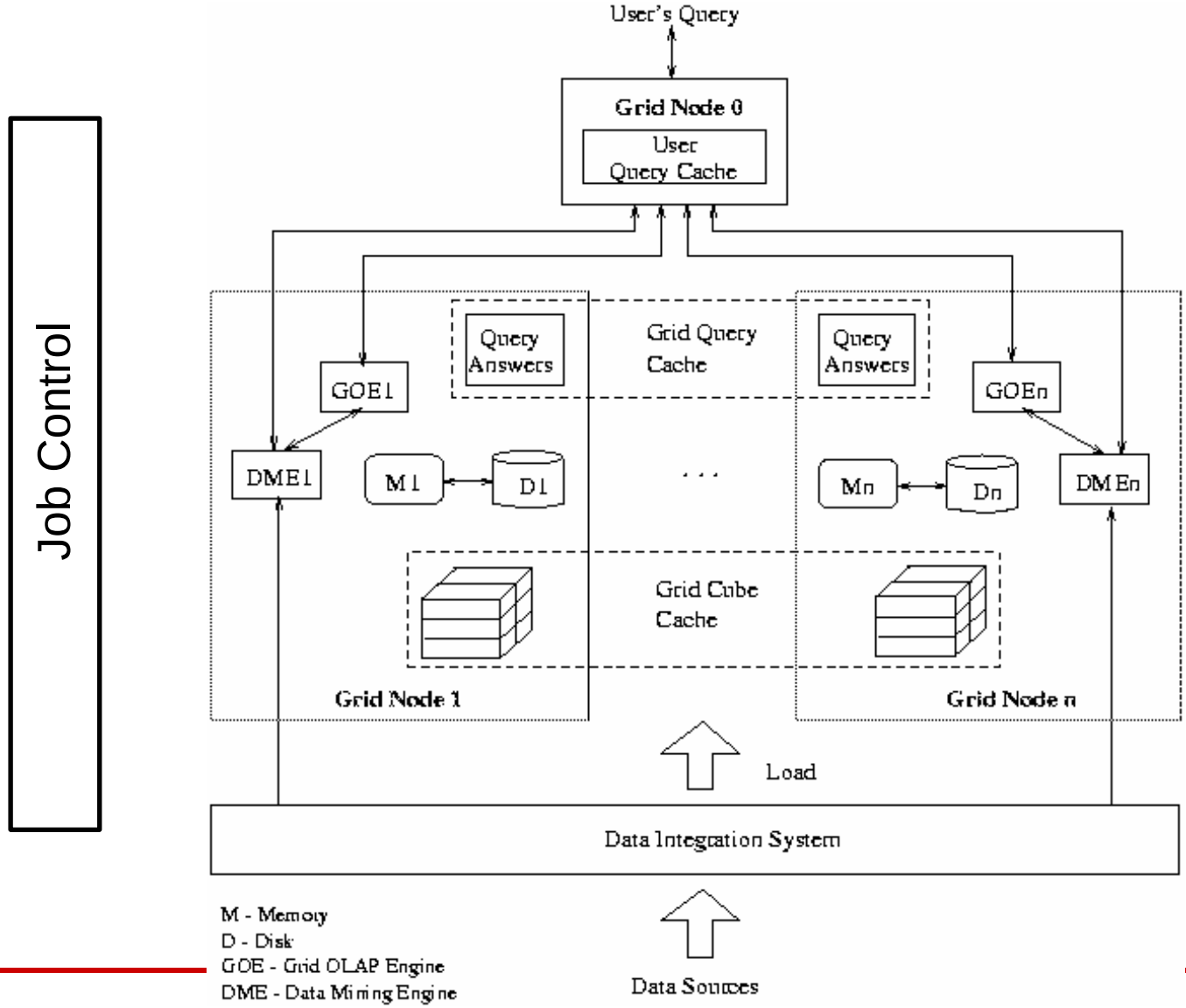
Autonomic Wisdom Grid Framework



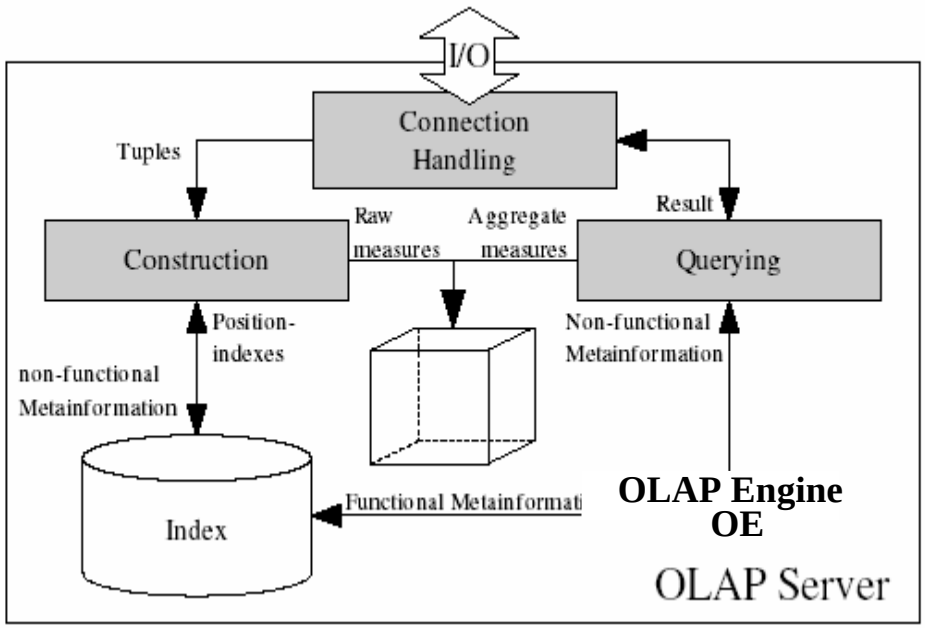
Scientific Results

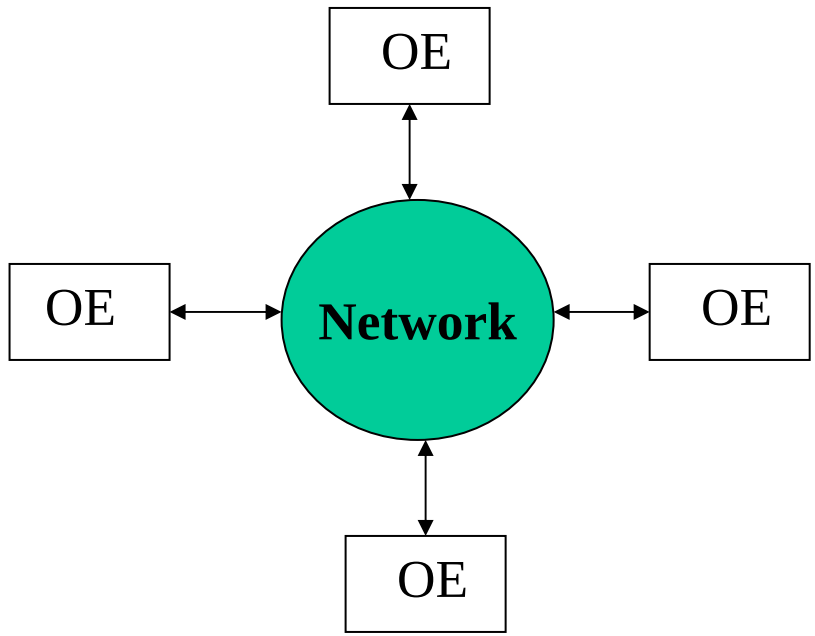
- GridMiner Architecture
- Workflow Management
- Data Mediation
- OLAP
- Data Mining

Retrospection: Once upon a time...



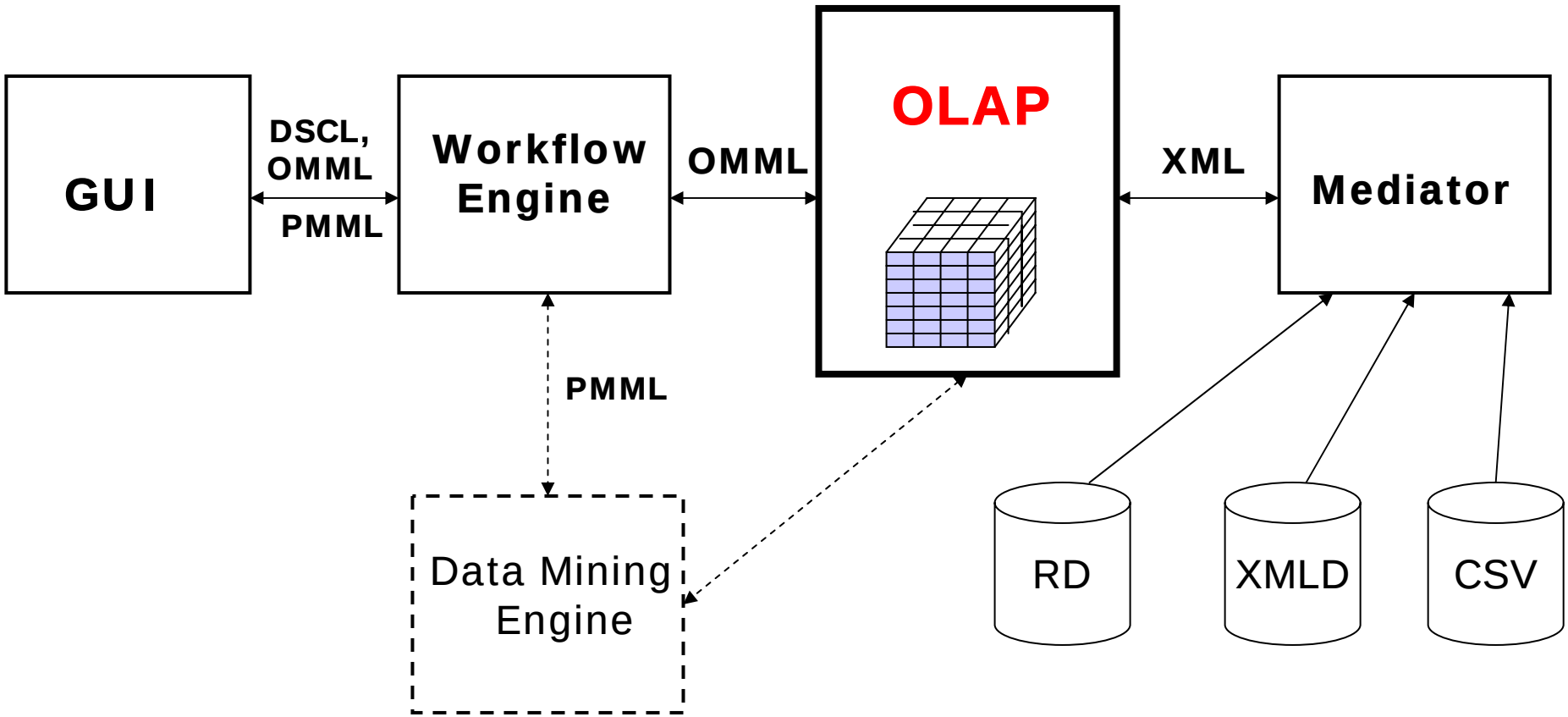
OLAP Strategy





Novel Dynamic Bit Encoding Method

Towards Centralized Service



Toward Indexing

Model				
Mini Van	6	5	4	4
Coupe	3	5	5	7
Sedan	4	3	2	3
	Red	Blue	White	Green
		Color		

The simplest method for computing a linear address from the multidimensional one:

(1) assign each possible position within one dimension an unique integer value and store these matching information in another table

(2) Bit-shift the integer assigned to the row dimension and logical OR it with the integer assigned to the column dimension.

(3) Use the combined integer as your memory address.

Model	Index(hex)
Mini Van	0x00
Coupe	0x01
Sedan	0x02

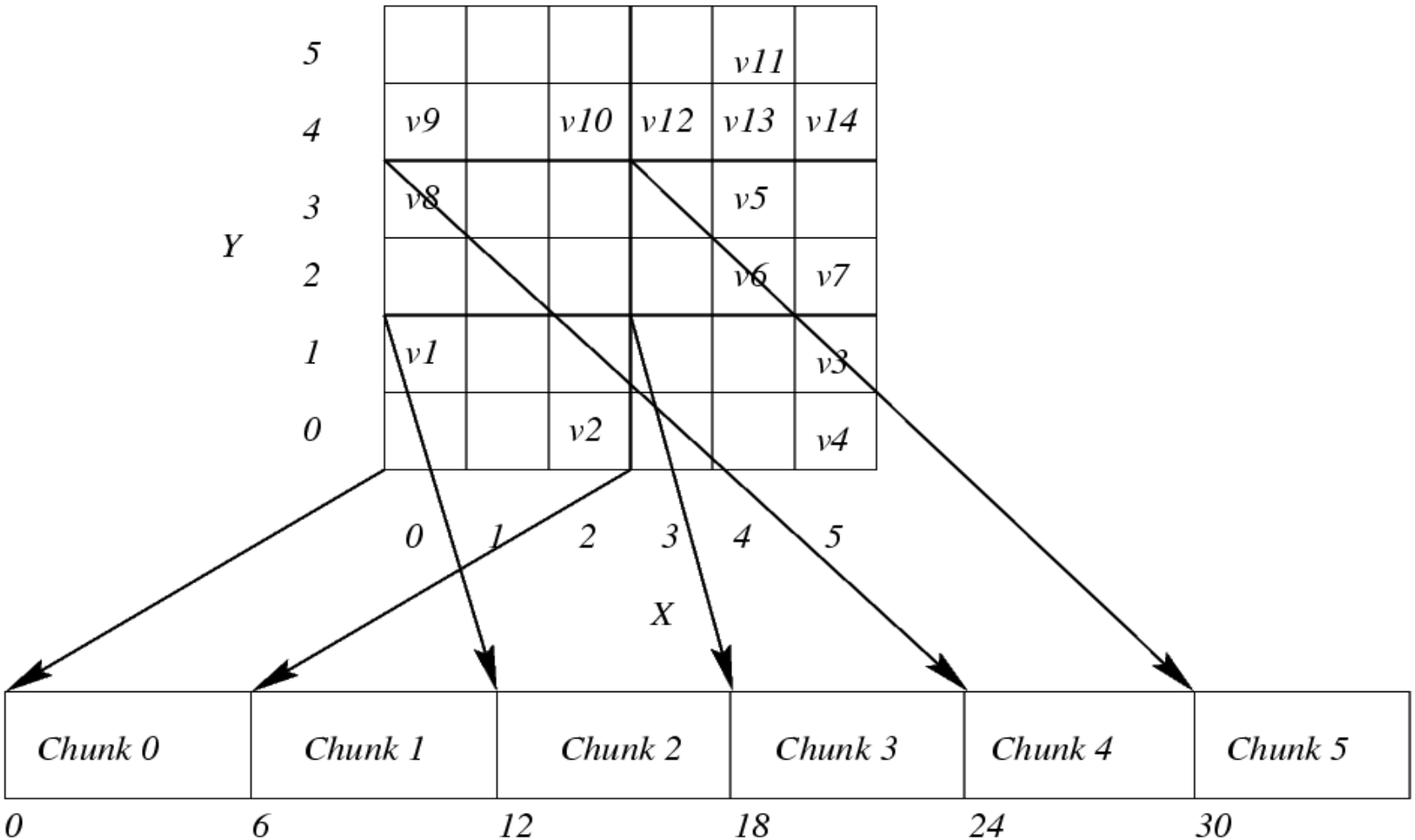
Color	Index(hex)
Red	0x00
Blue	0x01
White	0x02
Green	0x03

Drawback: We want to store 12 values, but we reserve 65534 addresses.

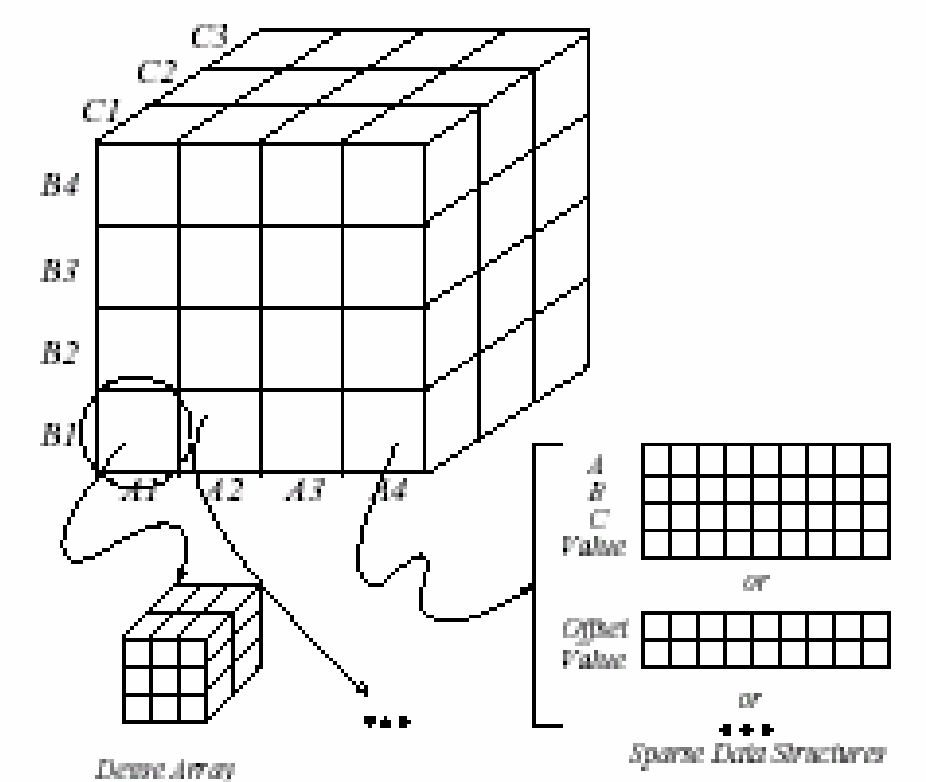
Another important issue: How to determine the position index size?

(Coupe, White) ⇒ 0x0102 (a linear address of the measure)

Chunking


5
4
3
2
1
0
Y
v11
v9
v10
v12
v13
v14
v8
v5
v6
v7
v1
v3
v2
v4
0
1
2
3
4
5
X
Chunk 0
Chunk 1
Chunk 2
Chunk 3
Chunk 4
Chunk 5
0
6
12
18
24
30

Dense and Sparse Chunk Storage



Sparsity Example

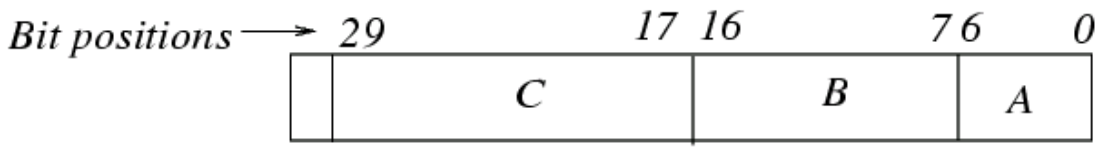
HP Application

- Web access analysis engine
- E.g., a newspaper Web site received 1.5 million hits a week
- Modeling the data using 4 dimensions
  1. ip address of the originate site (48,128 values)
  2. referring site (10,432 values)
  3. subject uri (18,085 values)
  4. hours of day (24 values)
- The resulting cube contained over 200 trillion cells!

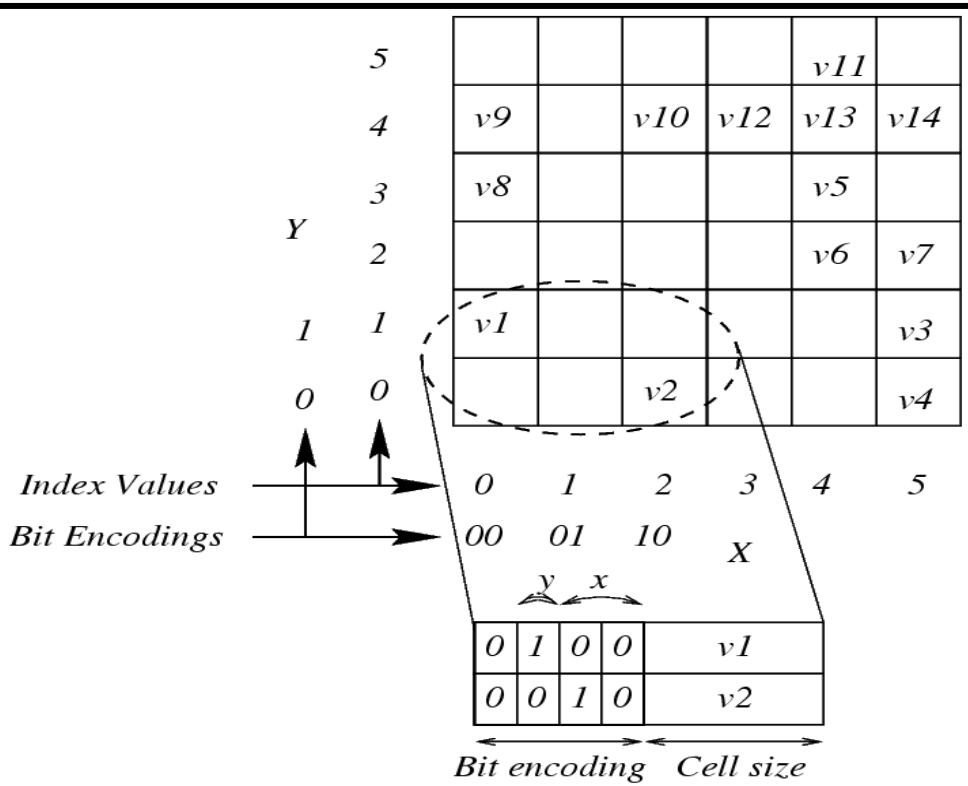
Bit Encoded Sparse Structure (BESS)

Principles:

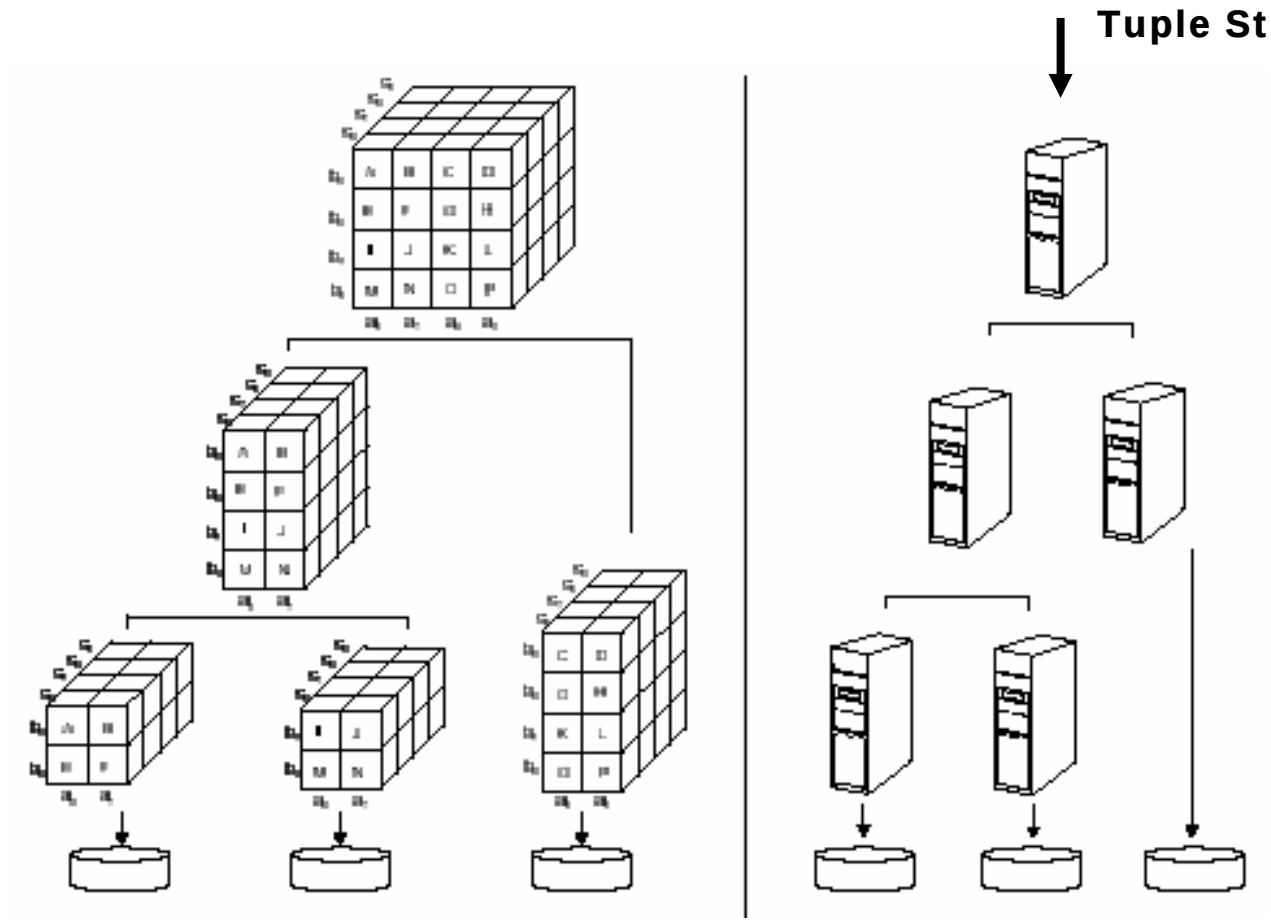
|A| = 100, |B| = 1000, |C| = 1000



Chunking:



Distributed OLAP –
Aggregation of Compute and Storage Resources vs. Federation



Federated OLAP

Motivating Example

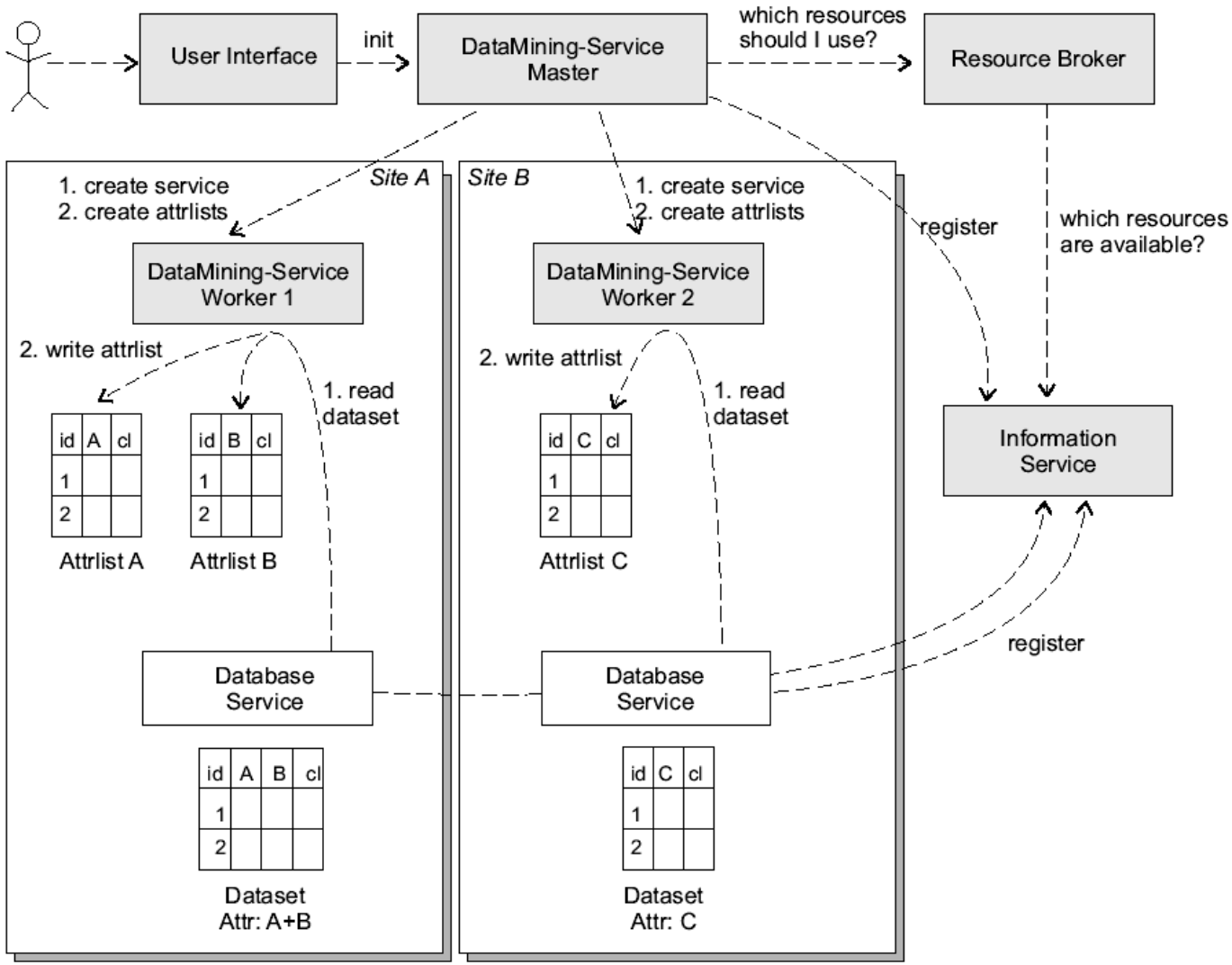
- Effective management of a network requires collecting, correlating, and analyzing a variety of network trace data.
- Analysis of flow data collecting at each router and stored in a local data warehouse „adjacent“ to the router is a challenging application.
- All flow information is conceptually part of a single relation with the following schema:

Flow (RouterId, SourceIP, SourcePort, SourceMask, SourceAS, DestIP, DestPort, DestMask, DestAS, StartTime, EndTime, NumPackets, NumBytes)

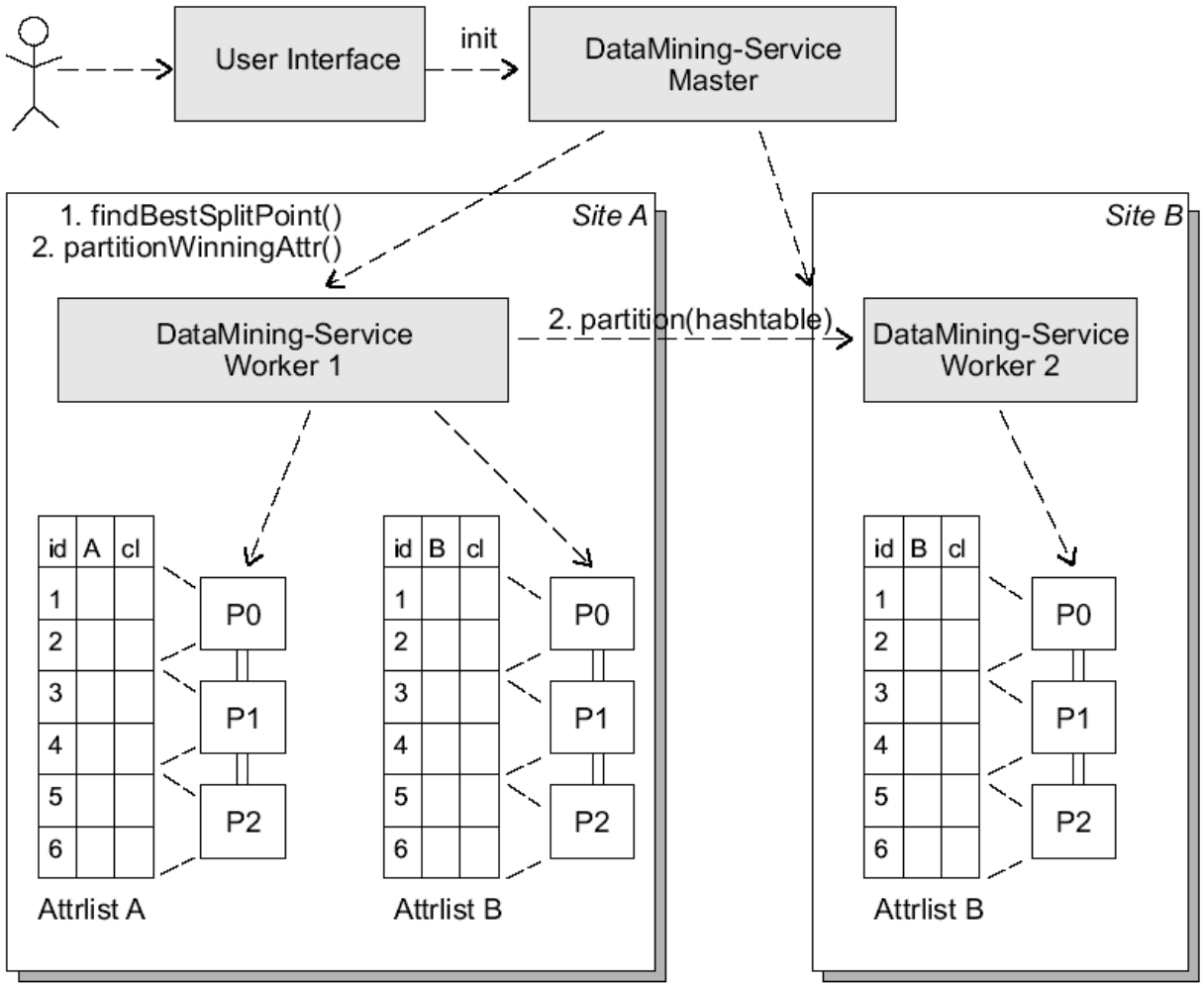
DIGIDT – **Di**stributed **G**rid-Enabled **I**nduction of **D**ecision **T**rees

DIGIDT:

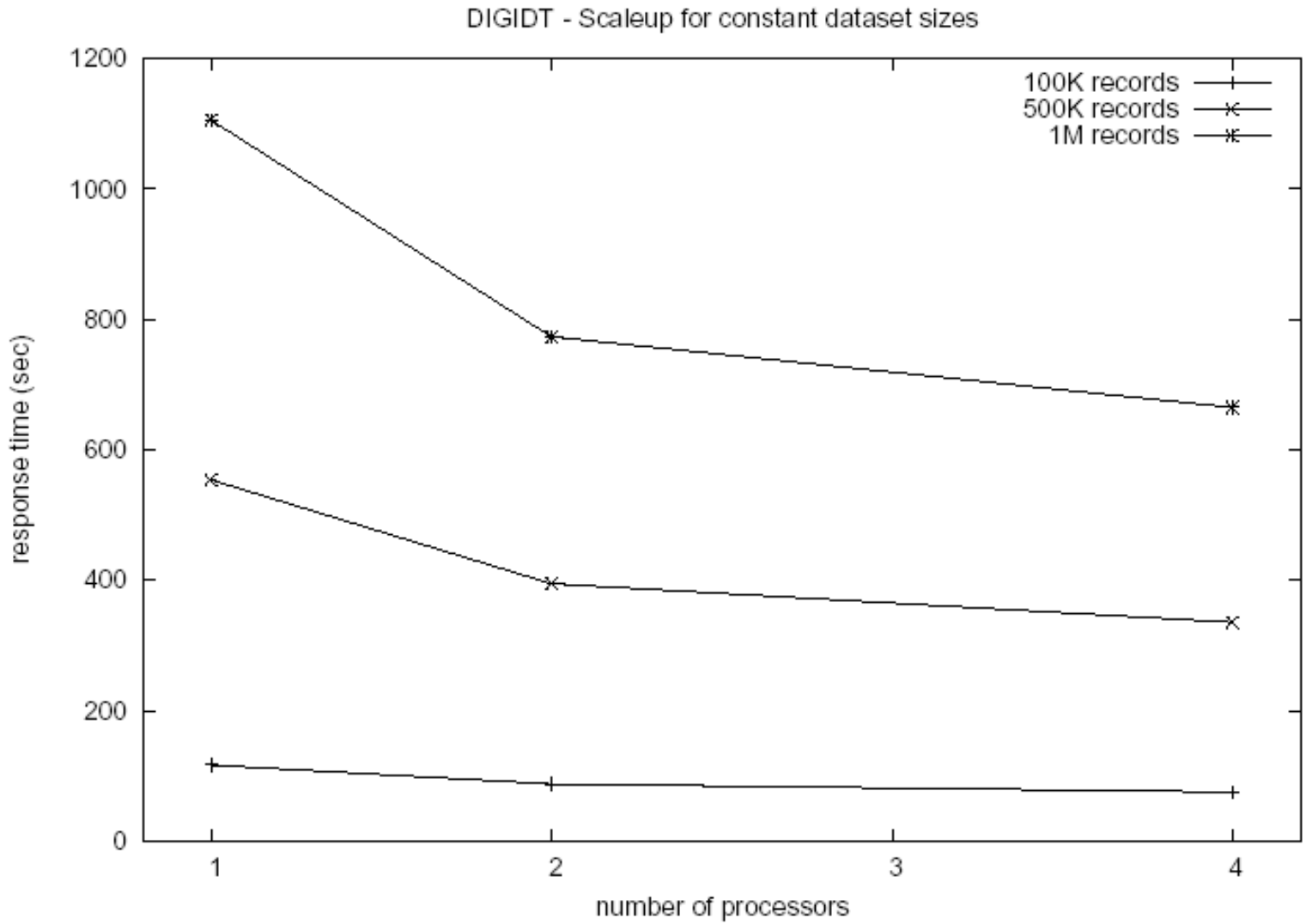
Phase 1 - Preparation



Phase 2 - Execution



DIGIDT: Experiments



Conclusions

- **Discussion of some issues driving Grid knowledge discovery research**
- **Development of the GridMiner architecture**
- **Outline of results achieved**