

The Terabyte Challenge

Discovering Information in Distributed and Massive Data

Robert Grossman

National Center for Data Mining

University of Illinois at Chicago

&

Magnify, Inc.



Part 1.

Introduction



What is Data Mining?

- Data mining is the automatic discovery of patterns, associations, changes, & anomalies, in large data sets.
- Emphasis on discovery vs validation (of patterns)
- Most data is distributed and was never meant to be correlated.
 - A fundamental challenge is to mine distributed data.
 - This will enable a new paradigm for scientific discovery.

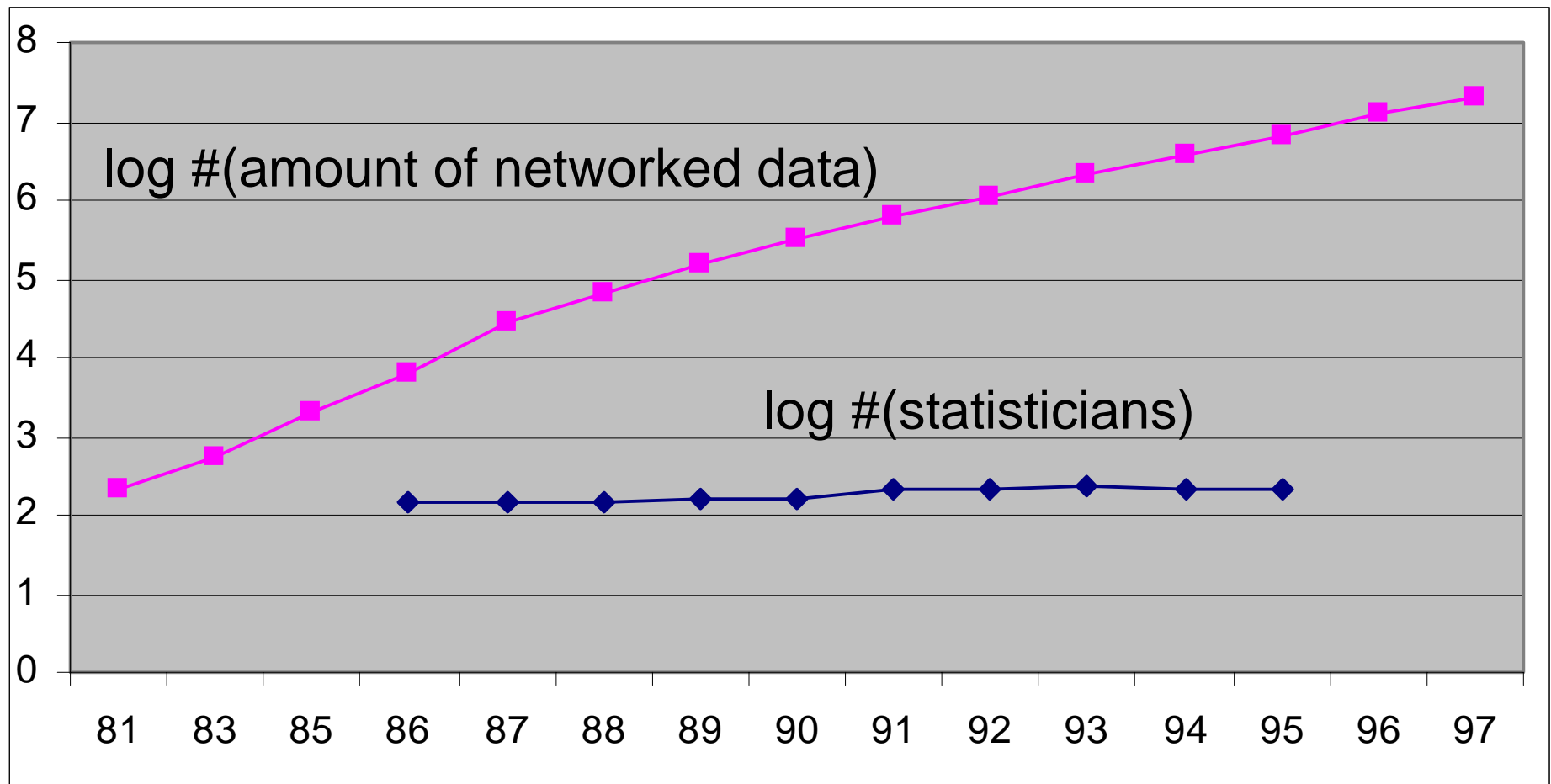


What are some Data Mining Applications?

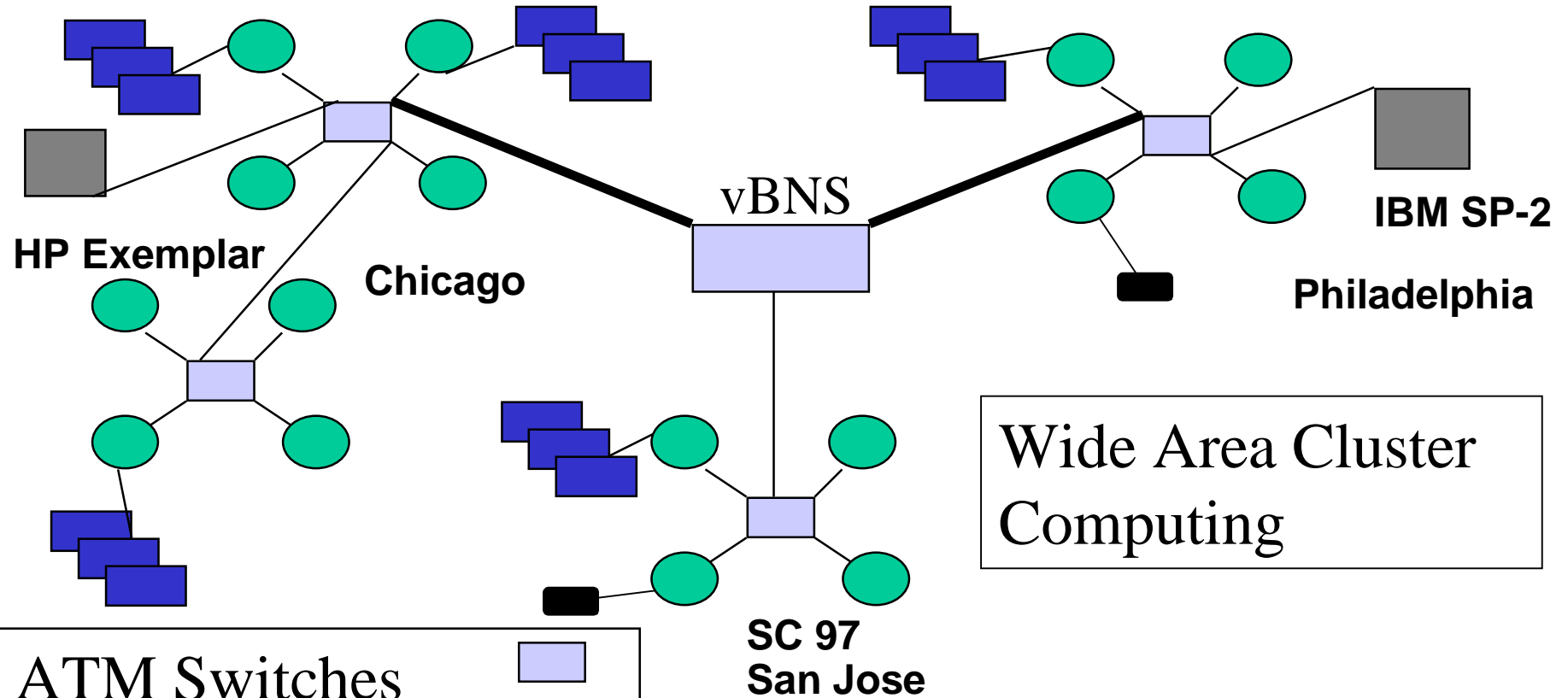
- Science
 - providing the infrastructure for searching large data sets for new stars, new elementary particles, new drugs
- Health Care
 - identifying patterns in health care data to improve diagnosis and treatment
- Business
 - uncovering fraud and other anomalies







Why is the Emergence of Data Mining Inevitable?



NSCP Meta-Cluster Facility

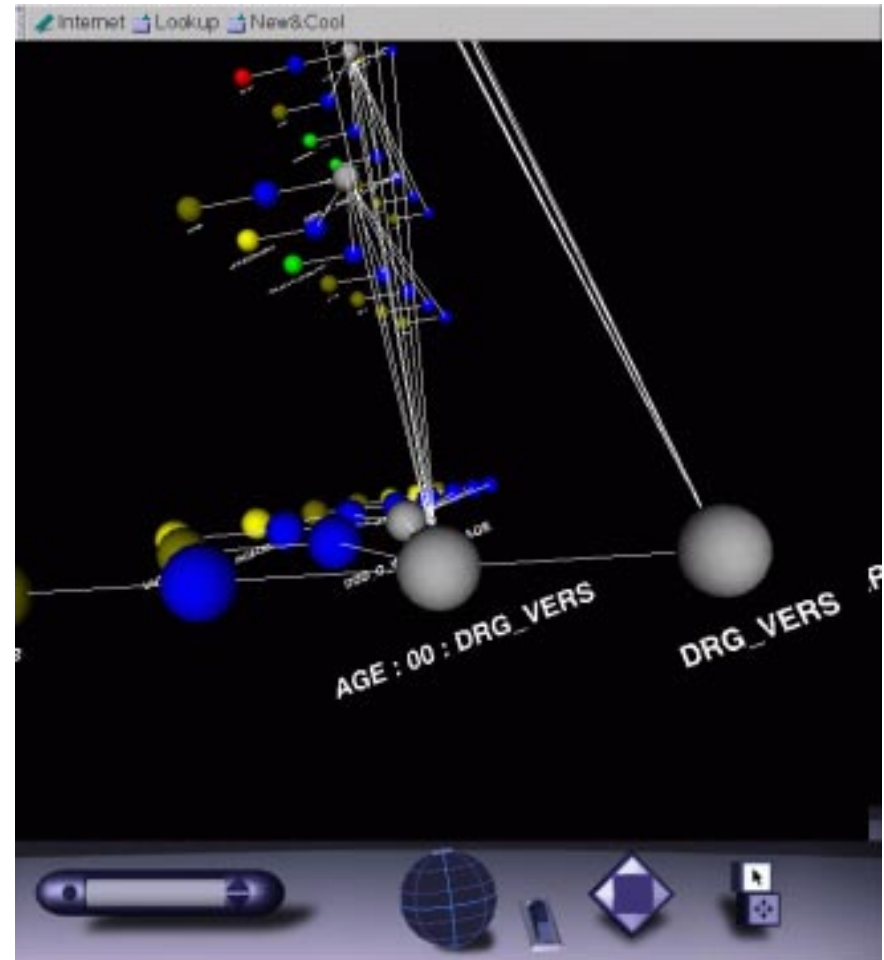


- | | |
|-------------------|-------------------------------------------------------------------------------------|
| ATM Switches |  |
| Nodes (150 nodes) |  |
| Disks (2 TB) |  |
| Tape (6 TBs) |  |

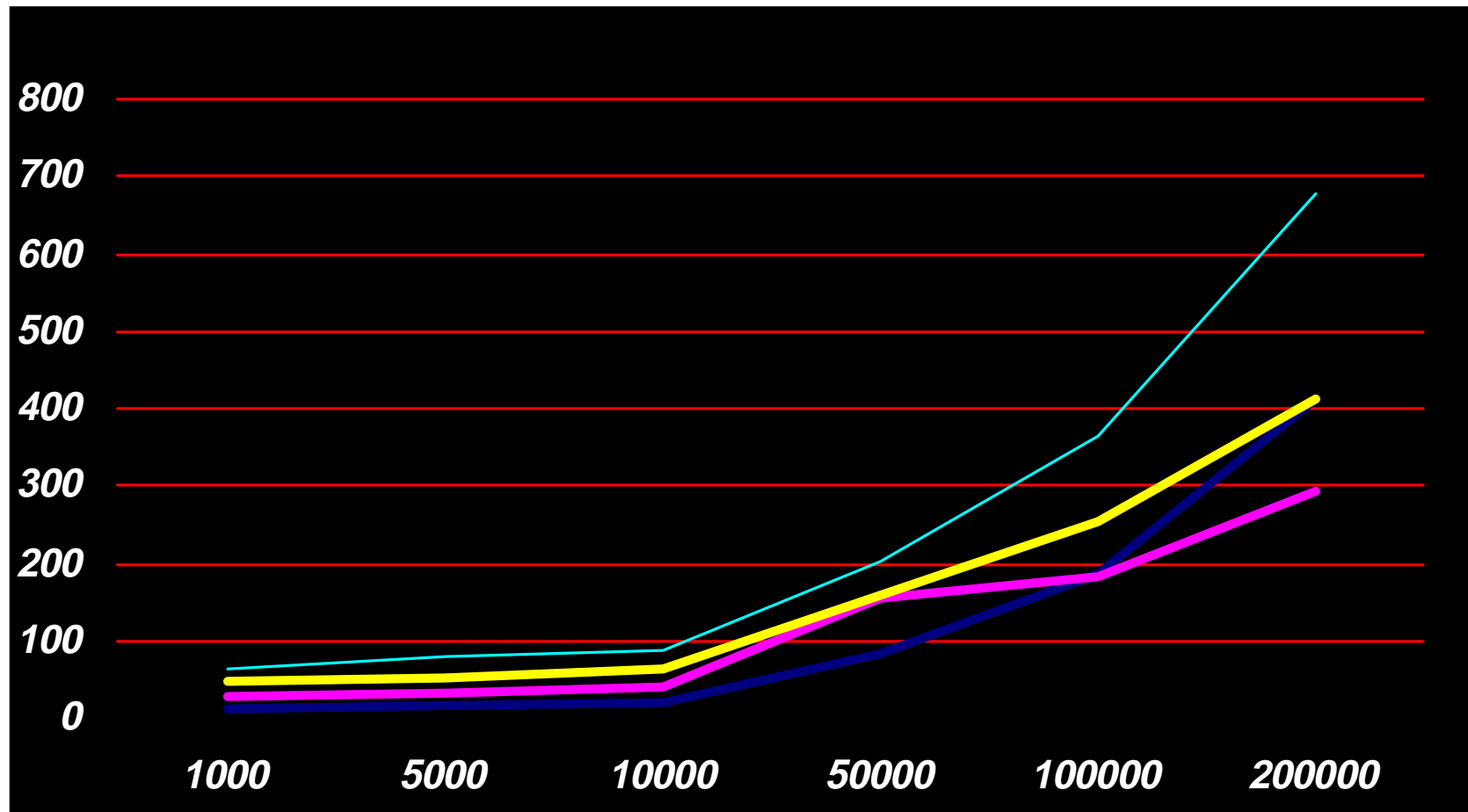
Clusters of query nodes, compute nodes, & i/o nodes

Example 1: Distributed Data Mining of Health Care Data

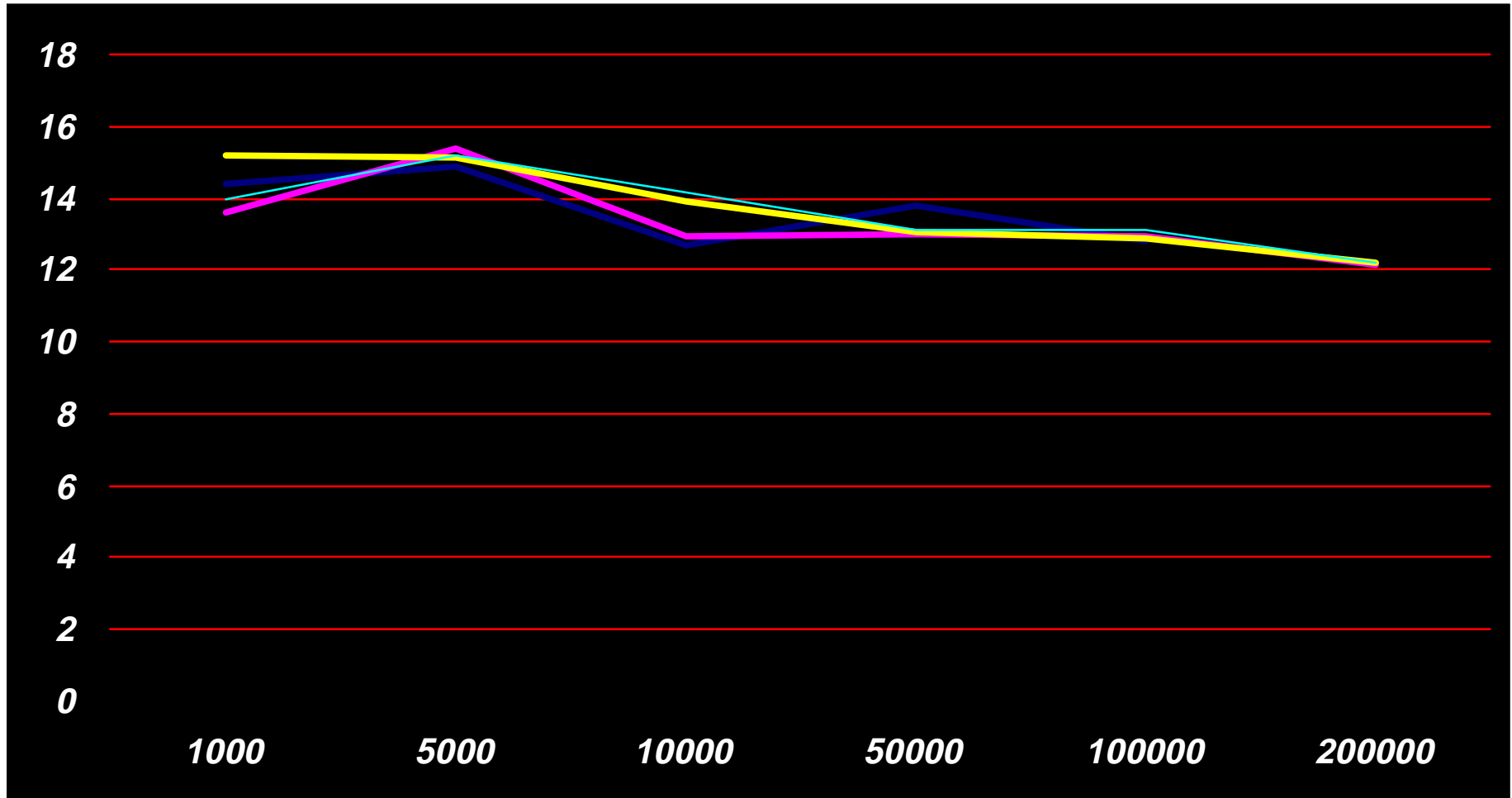
- **Data from University Healthcare Consortium (UHC)**
- **Data distributed by hospital**
- **Eg. length of stay (LOS) differs between hospitals for MI**
- **What patterns learned from billing data characterize benchmark hospitals (= top quartile)?**



Time to Compute 5, 9, 15, 21 Distributed Models



Error for 1, 5, 9, & 21 Models

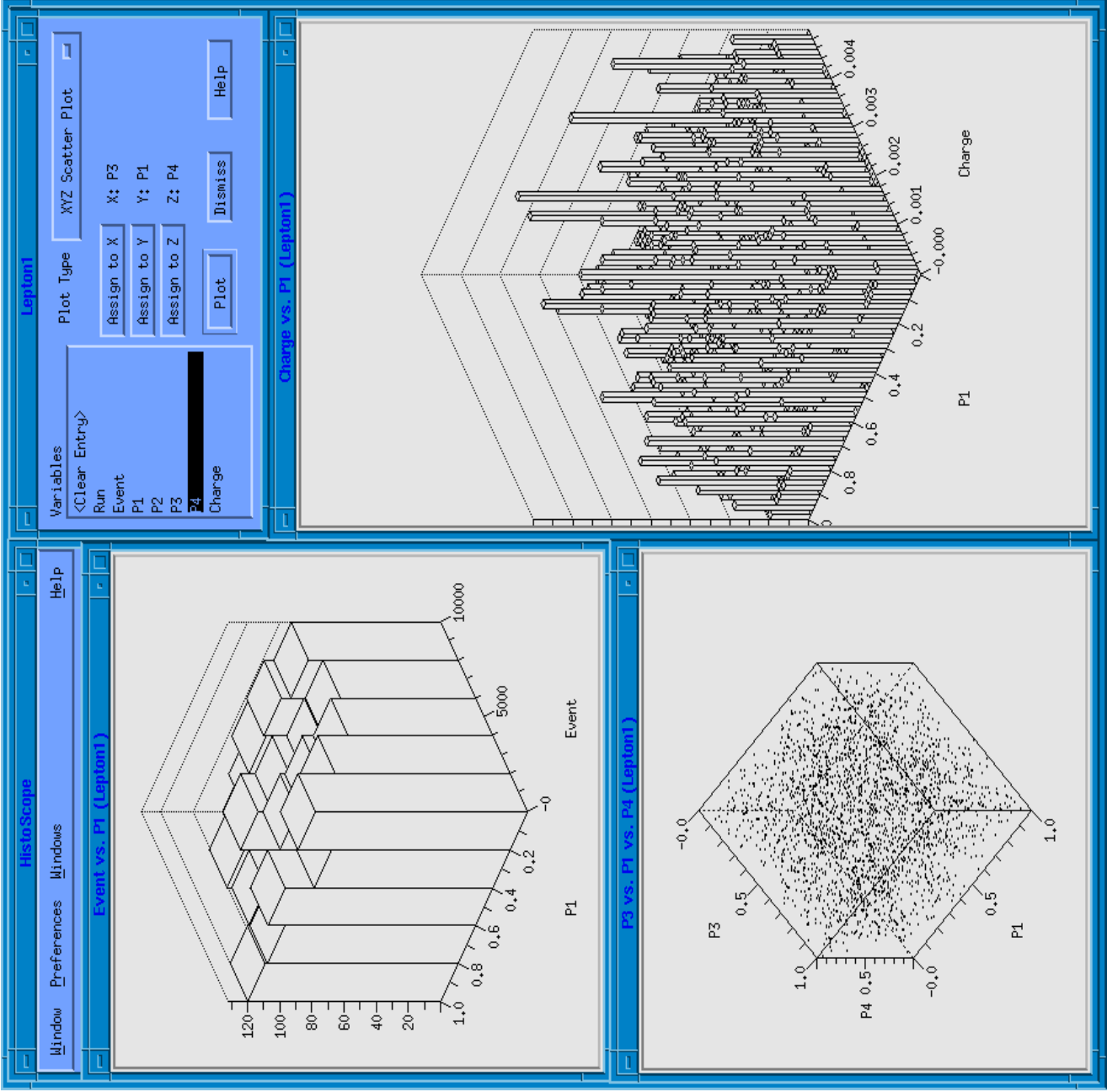


Example 2:

High Energy Physics Top Quark

- Experiment produced
 - 10^9 events before trigger
 - 10^6 events after trigger
 - 6.3 events predicted (if no top quark)
 - 15 events observed (demonstrating top quark exists)
- Hundreds of scientists around the world analyze the data locally and interactively
- How can this be done in a collaborative fashion?





Example 3: Detecting Credit Card Fraud

[0] Account number:	4500089050201002
[1] Transaction Amount:	45.06
[2] Timestamp:	95214013238
[3] Acquiring Bank:	840
[4] Issuing Bank:	124
[5] Store Type:	5310

[6] Velocity 1:	-0.795767
[7] Velocity 2:	-0.609230

plus 75 additional attributes

Basic data attributes via
data cleaning

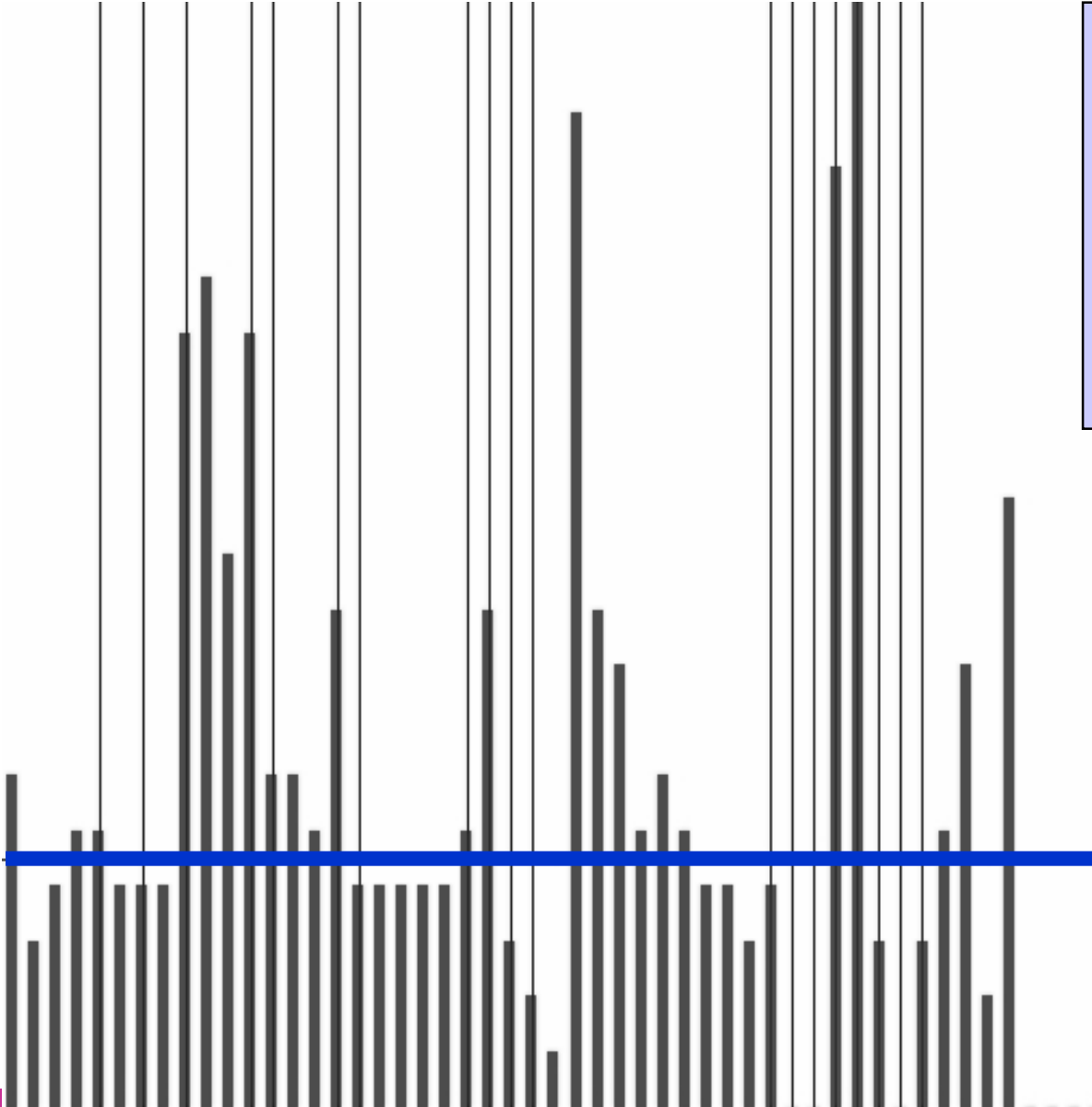
Derived attributes via data
transformations

Fraud Alert

Predictive attributes via
data mining



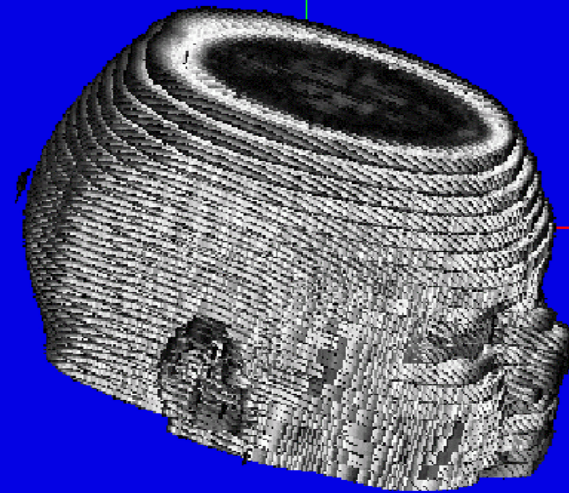
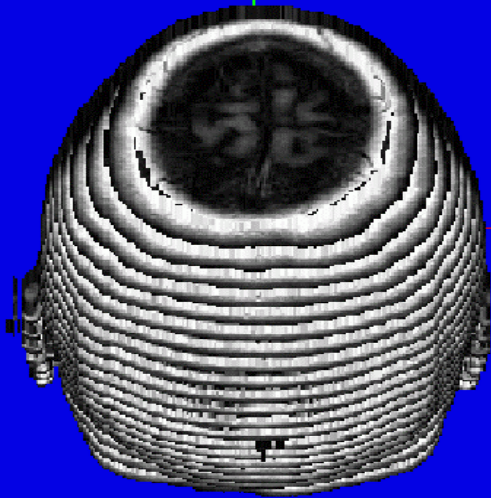
Fraud Alerts



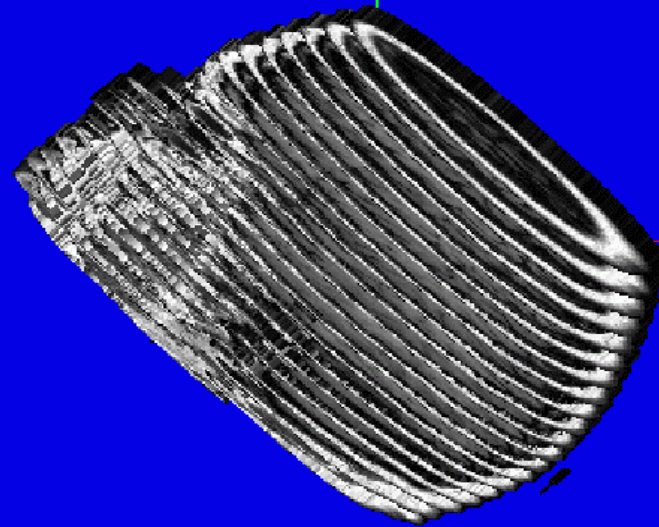
Height of bar indicates likelihood of fraud. Bars above blue line sound alert.

Use QOS to enable real time decision support.

Example 4: Remote Diagnosis



- Feature extraction and volume rendering from data warehouses of medical images (University of Illinois at Chicago).
- Utilizes the widely distributed data warehouses, compute resources, and visualization resources of the NSCP.



Part 2.

Third Generation Data Mining Systems: Distributed, High Performance Data Mining

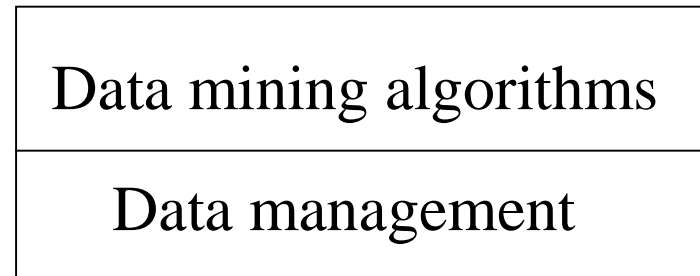


Three Generations of DM Systems

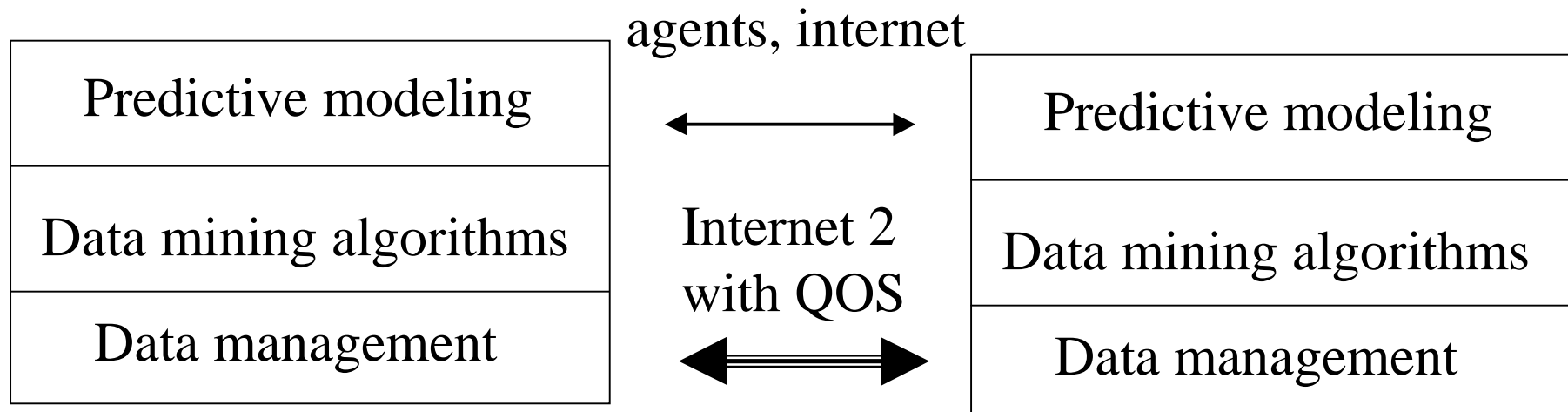
First Generation



Second Generation



Third Generation

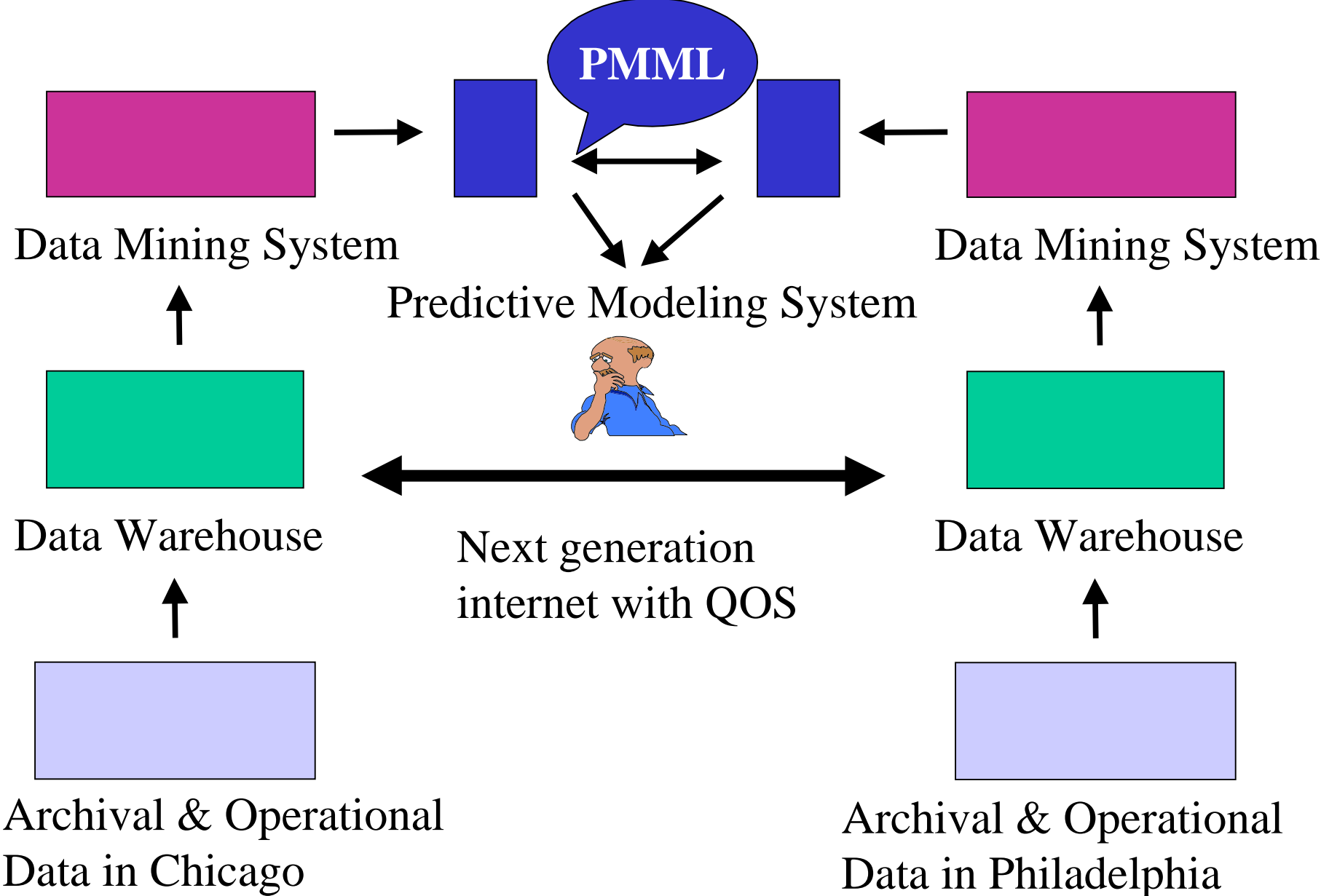


Fundamental Question

- Distributed databases
 - do we move the data or the query?
- Distributed data mining
 - do we move the data, the query (or computation), or the predictive model?
- Third generation data mining
 - enables new type of discovery



Distributed Data Mining



Predictive Model Markup Language (PMML)

<data-attributes>

<attribute-descriptor attribute-number=1 attribute-name=
“doc-id” attribute-label=“Source”>

< attribute-descriptor type=derived attribute-number=8 attribute-
name=“velocity-three-hours”>

</data-attributes>

<logistic-regression weight = 0.3>

<model-attribute-number=1 coefficient=0.239494>

<model-attribute-number=2 coefficient=0.495858>

</logistic-regression>

<cart-tree weight = 0.7>

<tree-node node-attribute=8 threshold= 0.459507 left-child=12
right-child=18 node-label=“velocity three hours”>



Quality of Service Enables: Interactive Exploration of Data

- How do we enable interactive exploration and discovery of distributed data?
 - For the first time we can correlate data that has never been correlated before.
 - Example: sun spot data with global change data.
 - Enables “casual” exploration of distributed data.
 - Requires quality of service to be effective.



Quality of Service Enables: Decision Support & Data Mining

- Decision support
 - is this transaction fraudulent?
 - is this patient at risk?
- QOS allows us to remove data source if it is a “spoiler”
- Moving to a world where there are many data sources and we are interesting in correlating as many as “feasible” for decision support.

Quality of Service Enables: Incorporating Visualization and Continuous Media Data into Data Mining

- Observe user profiles interacting with distributed digital libraries of video data
 - improving caching, prefetching
 - improving precision and recall
- Incorporating visualization into the data mining process



Summary

- Data mining is a key enabling technology for a variety of scientific, engineering, health care, and business applications.
- Today data mining is by and large concerned with mining centralized data.
- Wide area data mining utilizing Internet 2 allows the analysis and mining of geographically distributed data.
- QOS is critical.



For More Information

Robert Grossman

National Center for Data Mining

University of Illinois at Chicago

851 S. Morgan Street

Chicago, IL 60606

312 413 2176

312 355 0373 fax

<http://www.ncdm.uic.edu>

grossman@uic.edu

