

The Analysis & Mining of Globally Distributed Data

Chapter 2. Data Mining

Robert Grossman
Laboratory for Advanced Computing
University of Illinois at Chicago
&
Open Data Partners

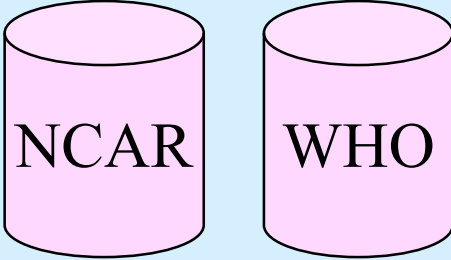
2.1 Data Mining

Extracting patterns,
changes, associations and
anomalies from data.

What is Data Mining?

- Data mining is the semi-automatic discovery of patterns, changes, associations, anomalies, and other statistical significant structures.
- Data mining is one step in the data mining process, consisting of 1) ETL, 2) data warehousing, 3) data shaping, 4) data mining algorithms, 5) deployment of models


Working with Data - End to End Viewpoint



NCAR WHO


Phase A.
Access &
EDA

45%



Phase B. Data
Analysis &
Mining

10%



Phase C.
Deployment &
Decision

45%

- Web services play an important role in Phases A and C.

Fundamental Question is Changing

- 1990s – How can I build better algorithms on my data?
- 2000's – How can I make more effective use of other's peoples data?

Data Mining/Statistical Models

Summarization Models

- clustering
- associations

- tree-based methods
- neural nets
- k-nearest neighbors

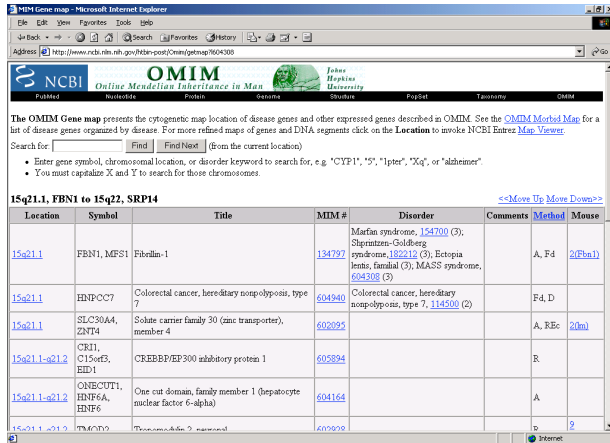
- contact chaining
- social network analysis

Predictive Models

Network/Graph

Copyright 2003 Robert L. Grossman

Nearing a Trifurcation Point



The screenshot shows the OMIM website interface. At the top, there's a search bar and navigation tabs. Below, a table lists gene-disorder associations for the 15q21.1 region. The table has columns for Location, Symbol, Title, MIM #, Disorder, Comments, Method, and Mouse.

Location	Symbol	Title	MIM #	Disorder	Comments	Method	Mouse
15q21.1	FBN1, MFS1	Fibrillin-1	134797	Marfan syndrome, 134700 (G); Shprintz-Goldberg syndrome, 182212 (G); Ectopia lentis, familial (E); MASS syndrome, 604302 (G)		A, Fd	2[Fbn1]
15q21.1	HNPCC7	Colorectal cancer, hereditary non-polyposis, type 7	604946	Colorectal cancer, hereditary non-polyposis, type 7, 114500 (C)		Fd, D	
15q21.1	SLC30A4, ZNT4	Solute carrier family 30 (zinc transporter), member 4	602095			A, REc	2[Slc30a4]
15q21.1-q21.2	CRE1, C15orf3, HD1	CREBBP/EP300 inhibitory protein 1	605894			R	
15q21.1-q21.2	ONECUT1, HNF6A, HNF6	One cut domain, family member 1 (hepatocyte nuclear factor 6-alpha)	604164			A	
15q21.1-q21.2	TM6SF2	Transmembrane 6 superfamily 2, member 2	609909				9

Web accessible
Databases

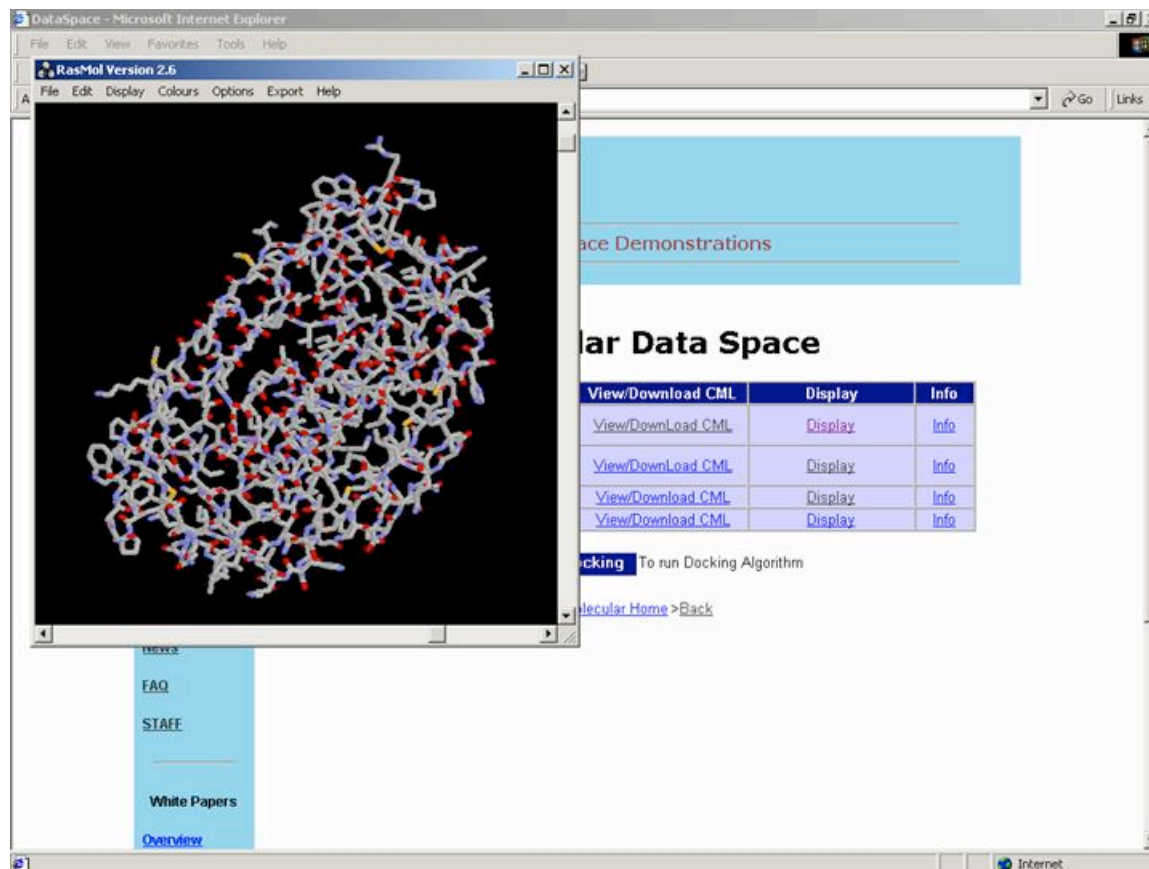
2003 - 2008

Data webs – remote
data analysis and
distributed mining

Data grids – transparent
high end computing

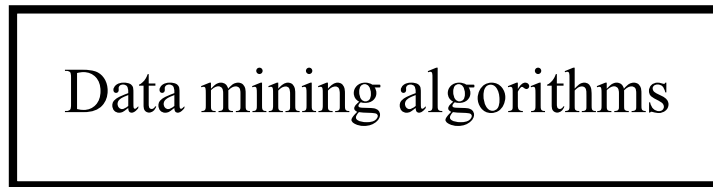
Semantic webs –
working with
knowledge

What will Future Data Mining Systems Look Like?

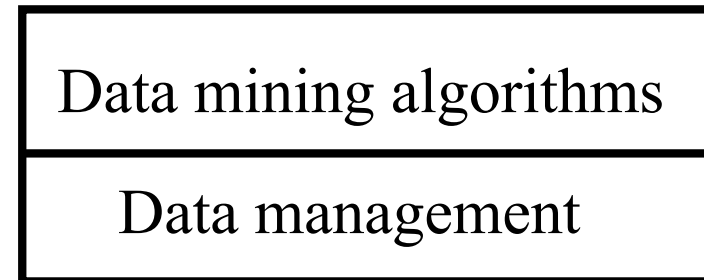


Four Generations of DM Systems

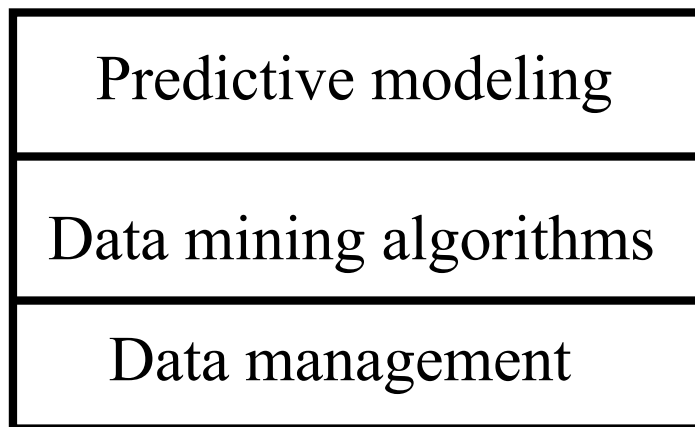
First Generation



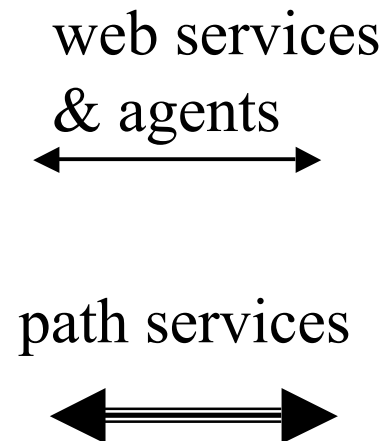
Second Generation



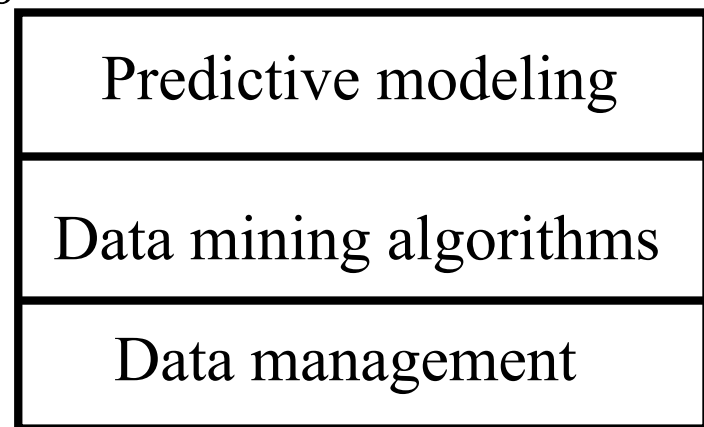
Fourth Generation



Third Generation



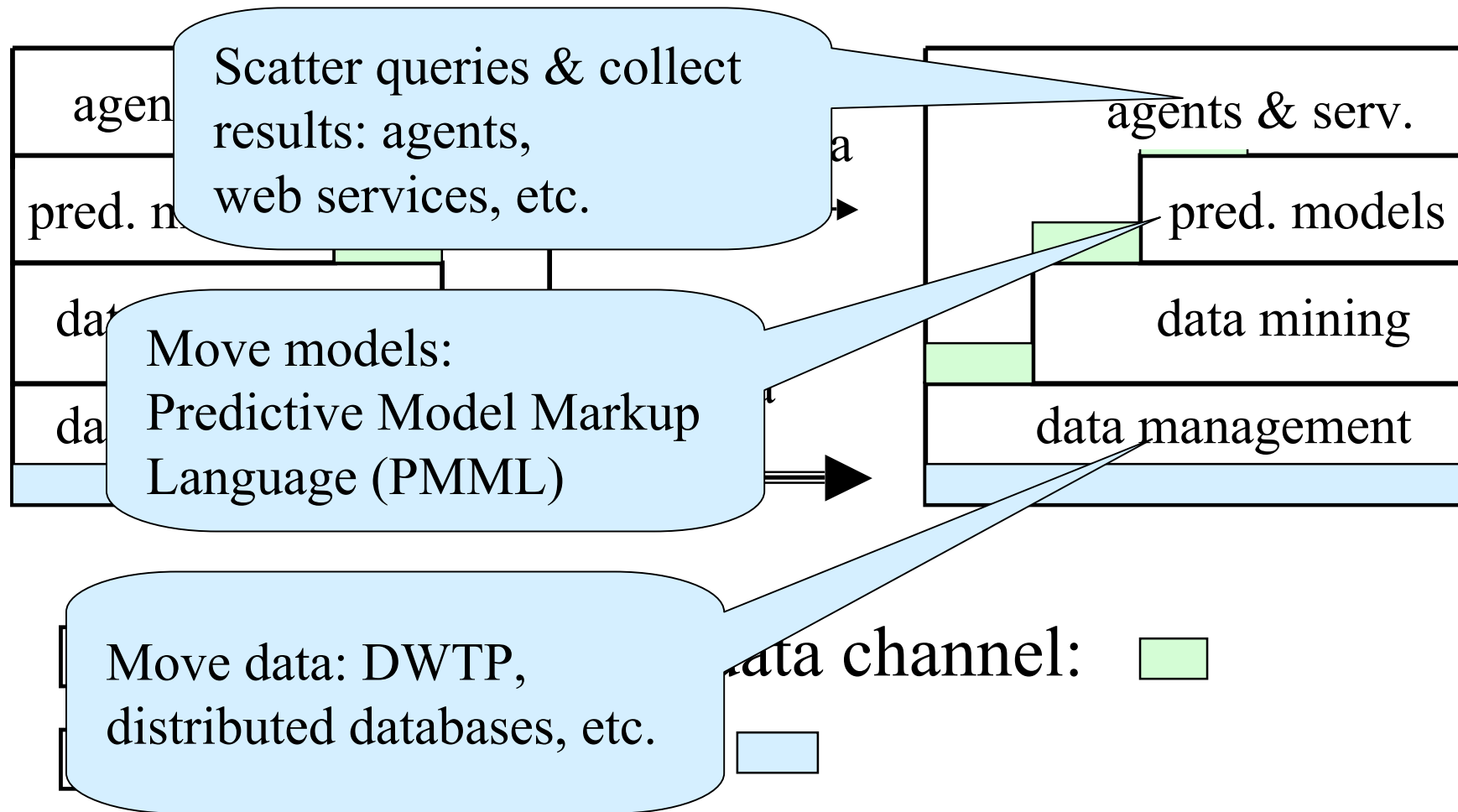
□ ubiq. data



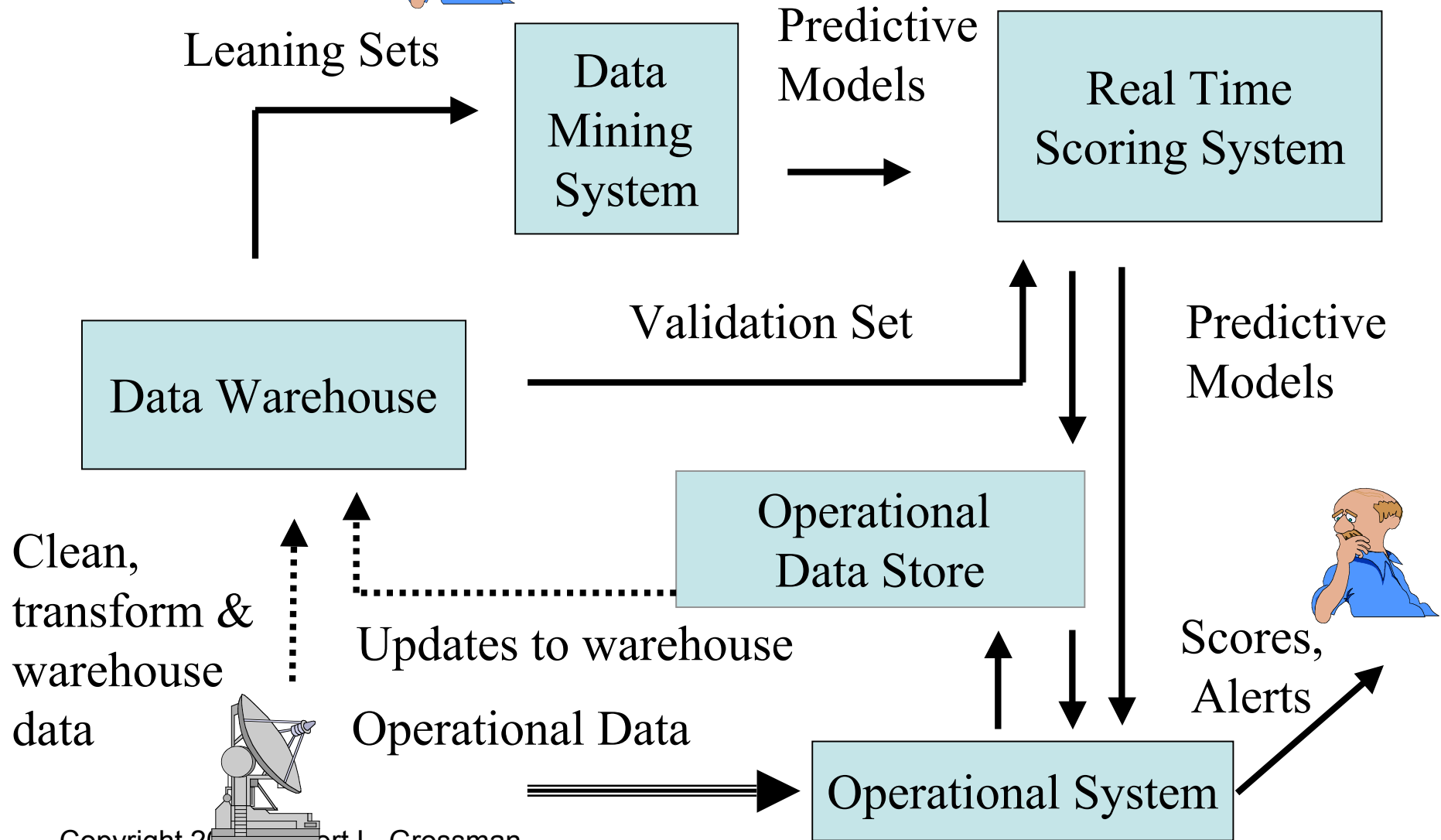
What Happened to Data Mining Systems During the Past 25 Years?

- 1980's – built statistical systems which ran on workstations (SAS, SPSS, SPlus, ...)
- 1990's – we moved algorithms to clusters of workstations and new types of data (just moving out of the labs)
- 2000's – we are beginning to build systems for exploring distributed data (still being developed)

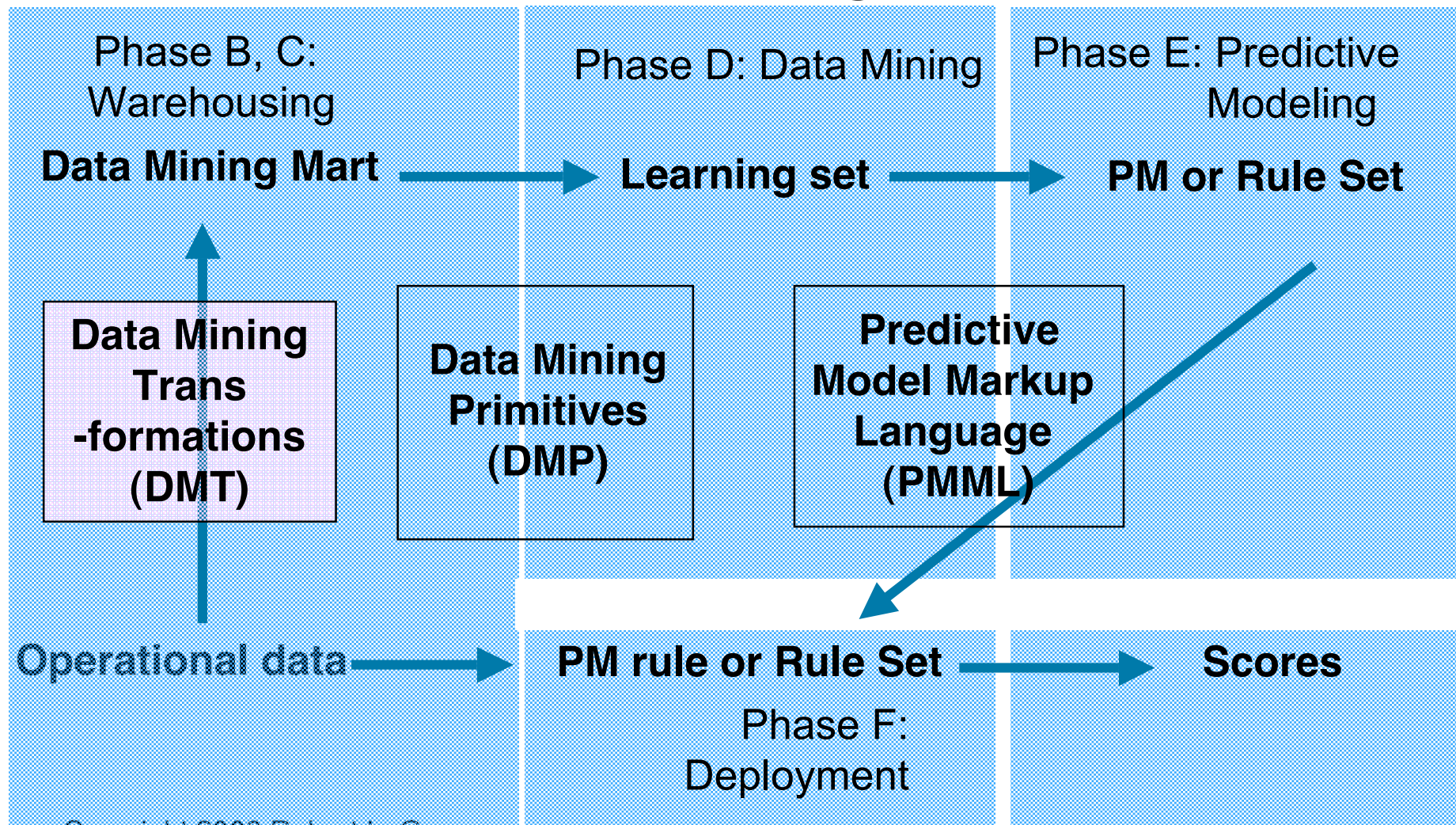
Layered Systems for DM & PM



Data Mining Process



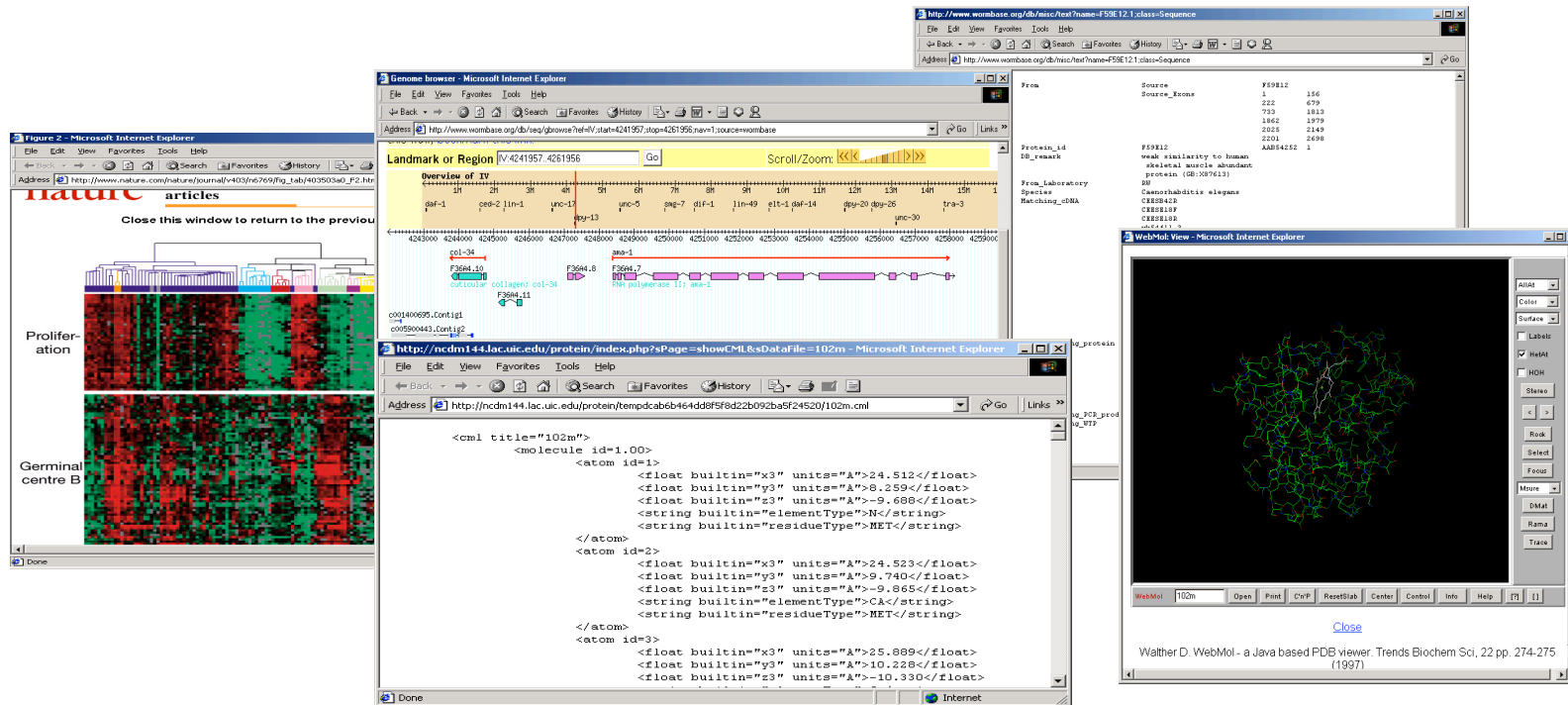
Phases in the Data Mining & Predictive Modeling Process



2.2 Distributed Data Mining

Extracting patterns from
distributed data.

Distributed Data Mining – In a Word



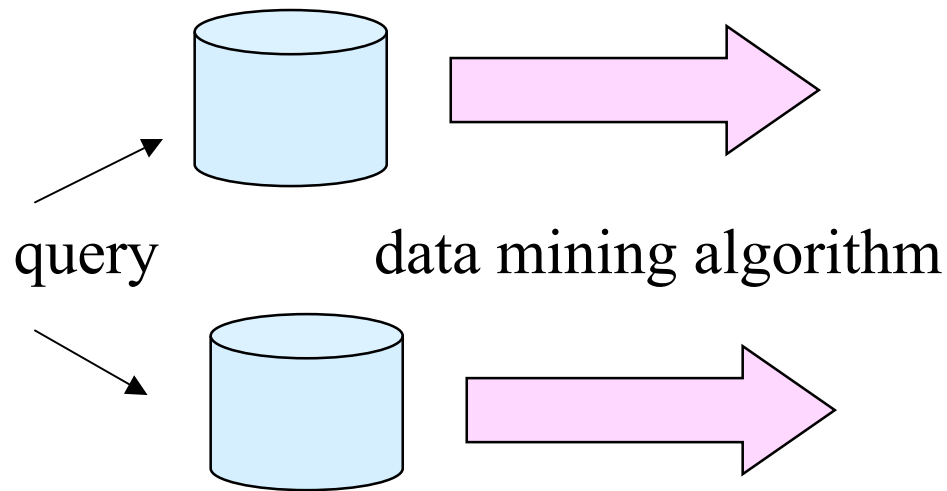
□ Predictive models are stronger by overlaying additional data.

Copyright 2003 Robert L. Grossman

Essential Distributed Data Mining Services

1. Scattering the data mining querying
2. Transporting the appropriate data
3. Performing the local and centralized data mining algorithms
4. Combining the results

1 & 2. Scatter Query and Transport Data



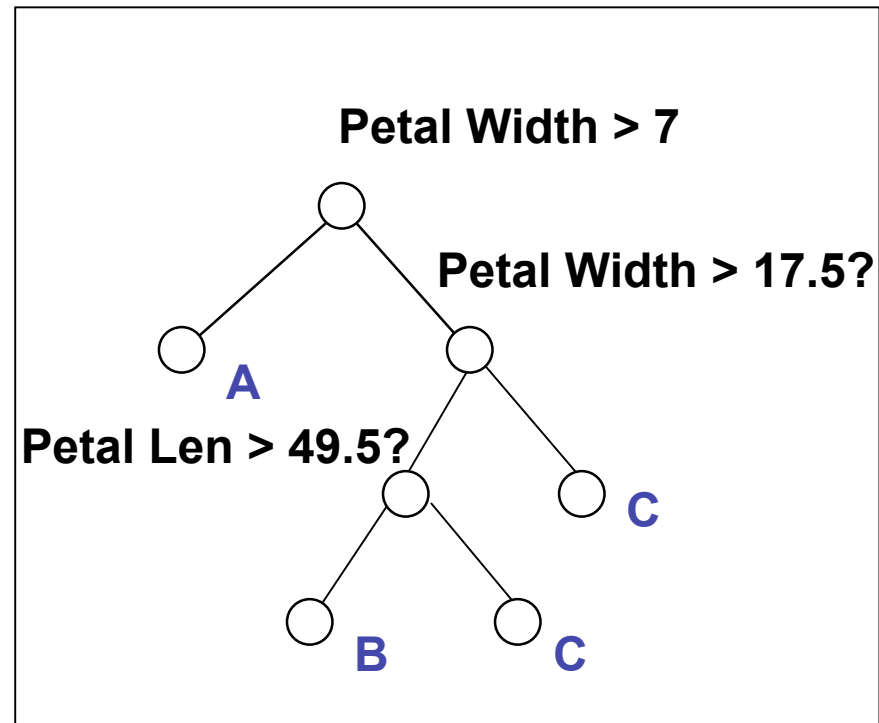
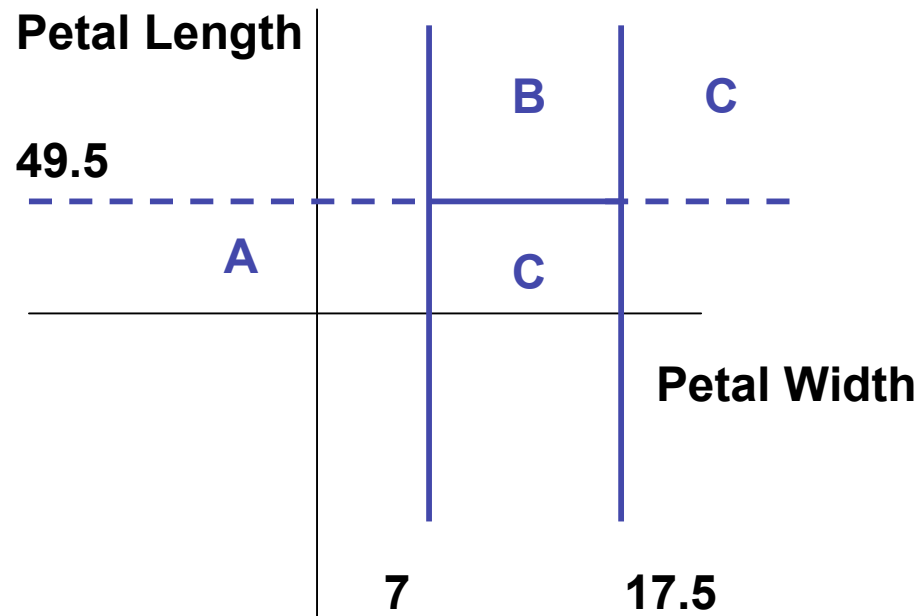
learning sets

```
<pmml>  
  
<pmml>  
<tree weight = 0.3>  
<tree-node node-id=8  
    threshold = 0.239494  
    etc. >  
</pmml>
```

statistical model

□ Data mining is the semi-automatic extraction of patterns, models, changes, associations, and anomalies from large data sets.

3. Example: Tree-Based Classifiers

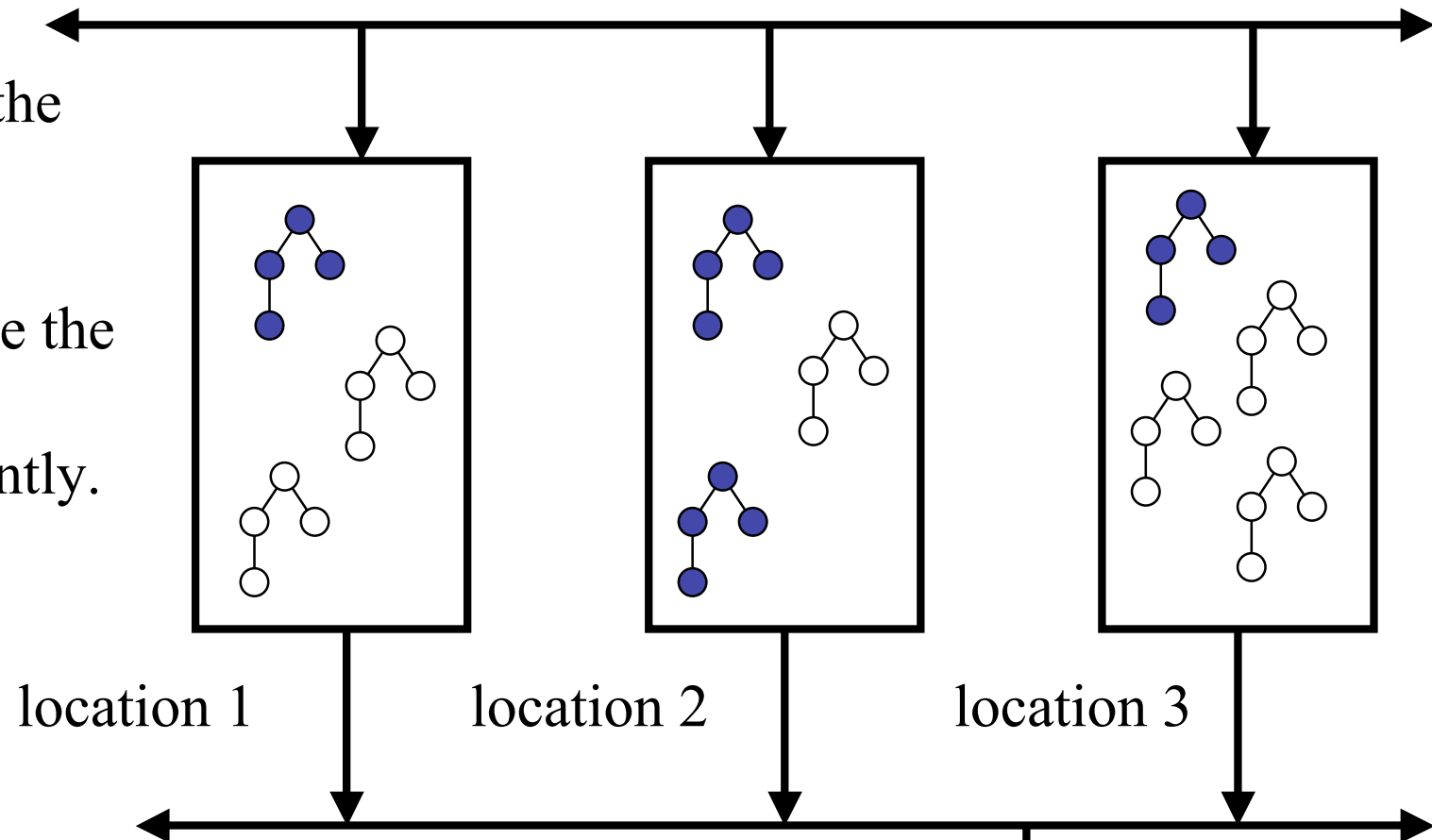


□ Trees partition the feature space into regions by asking whether an attribute is less than a threshold.

4. Combining Classifiers – Basic Idea

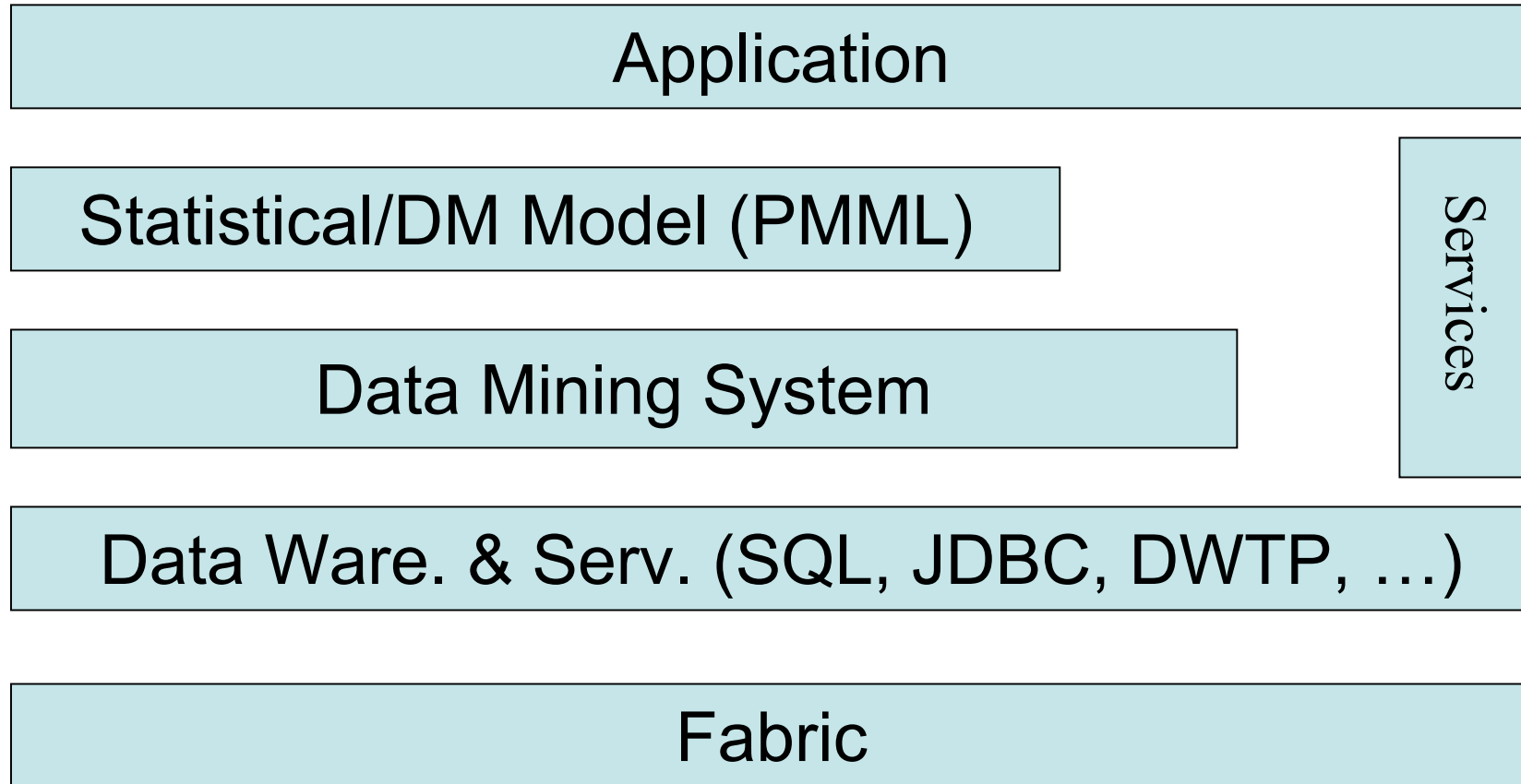
1. Scatter the query.

2. Compute the classifiers independently.



3. Gather and merge the classifiers.

Stack – Distributed Data Mining



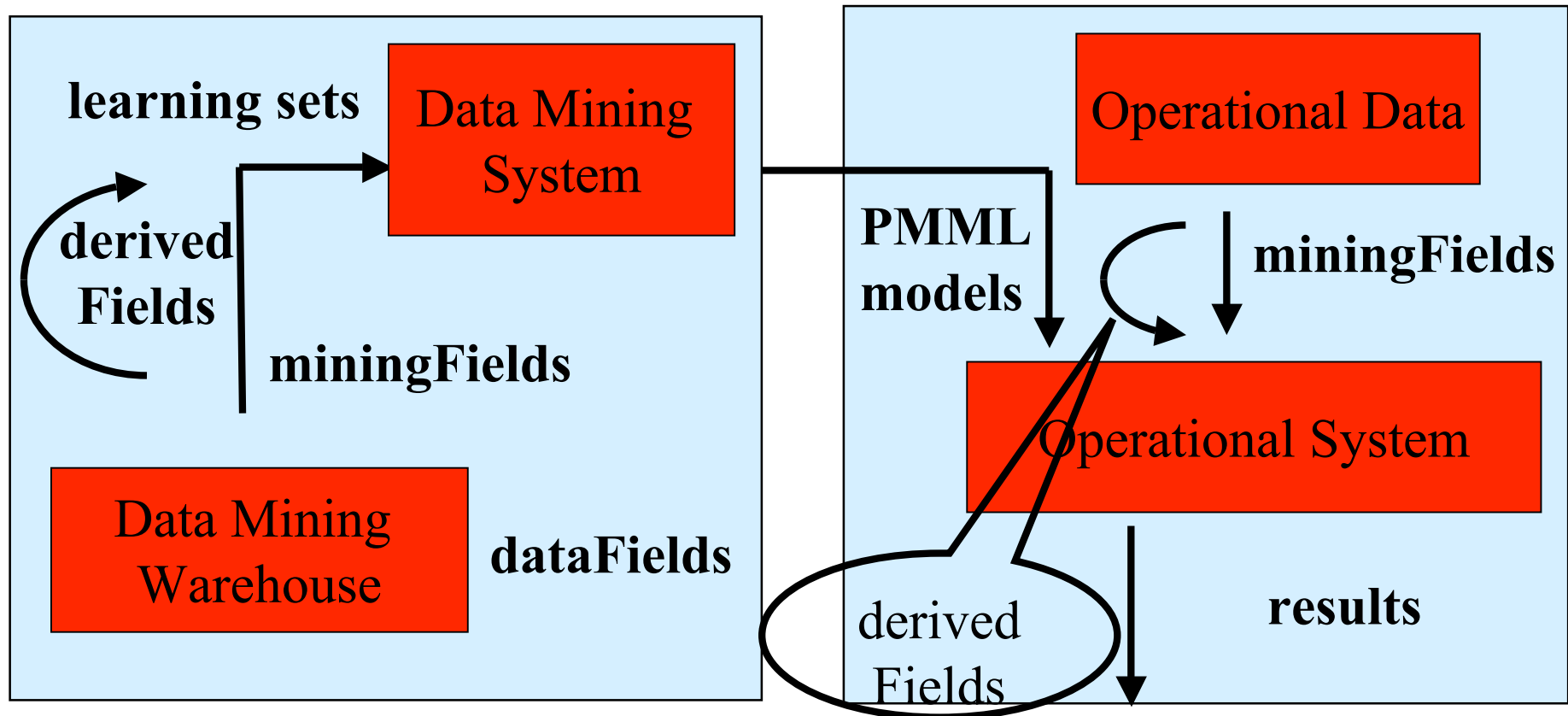
Predictive Model Markup Language (PMML)

- Based on XML
- Benefits of PMML
 - Open standard for Data Mining & Statistical Models
 - Not concerned with the process of creating a model
 - Provides independence from application, platform, and operating system
 - Simplifies use of data mining models by other applications (consumers of data mining models)

Model Architecture: Example 1

PMML Producers

PMML Consumers



Model Architecture: Example 2

