“Data Mining”

COPPE/UFRJ/ntt
www.ntt.ufrj.br

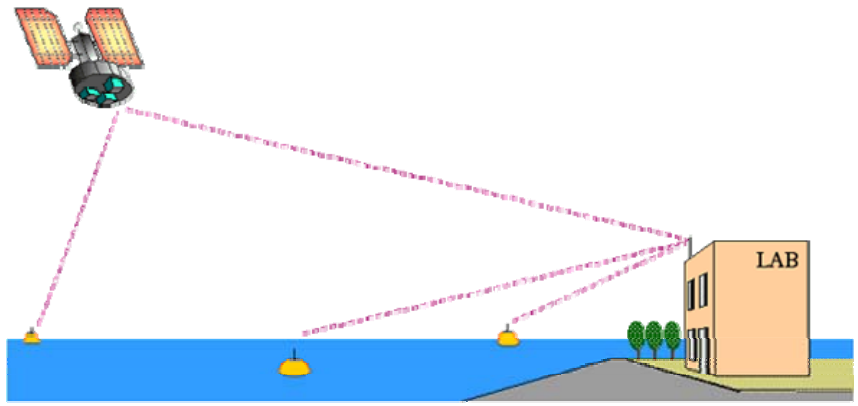
Nelson F. F. Ebecken
nelson@ntt.ufrj.br

The Evolution of Science

- **Observational Science**
  - Scientist gathers data by direct observation
  - Scientist analyzes data
- **Analytical Science**
  - Scientist builds analytical model
  - Makes predictions.
- **Computational Science**
  - Simulate analytical model
  - Validate model and makes predictions

Data Exploration Science

- Data captured by instruments
  Or data generated by simulator
- Processed by software
- Placed in a database / files
- Scientist analyzes database/files



Information Avalanche

- Both
  - better observational instruments and
  - Better simulations

  are producing a data avalanche
- Examples
  - Turbulence: 100 TB simulation
    then mine the Information
  - BaBar: Grows 1TB/day
    2/3 simulation Information
    1/3 observational Information
  - CERN: LHC will generate 1GB/s
    10 PB/y
  - VLBA (NRAO) generates 1GB/s today
  - NCBI: “only ½ TB” but doubling each year, very rich dataset.
  - Pixar: 100 TB/Movie

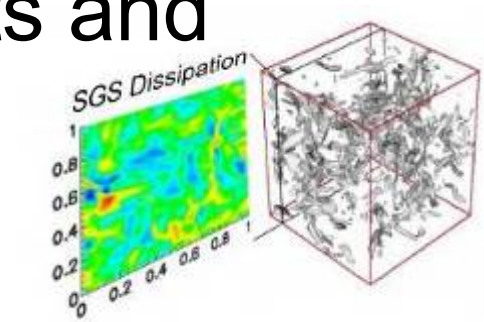


Image courtesy of C. Meneveau & A. Szalay @ JHU

The Indexable Web is More than 11.5 Billion Pages

What is the current size of the Web?
At this time (May 10, 2005)

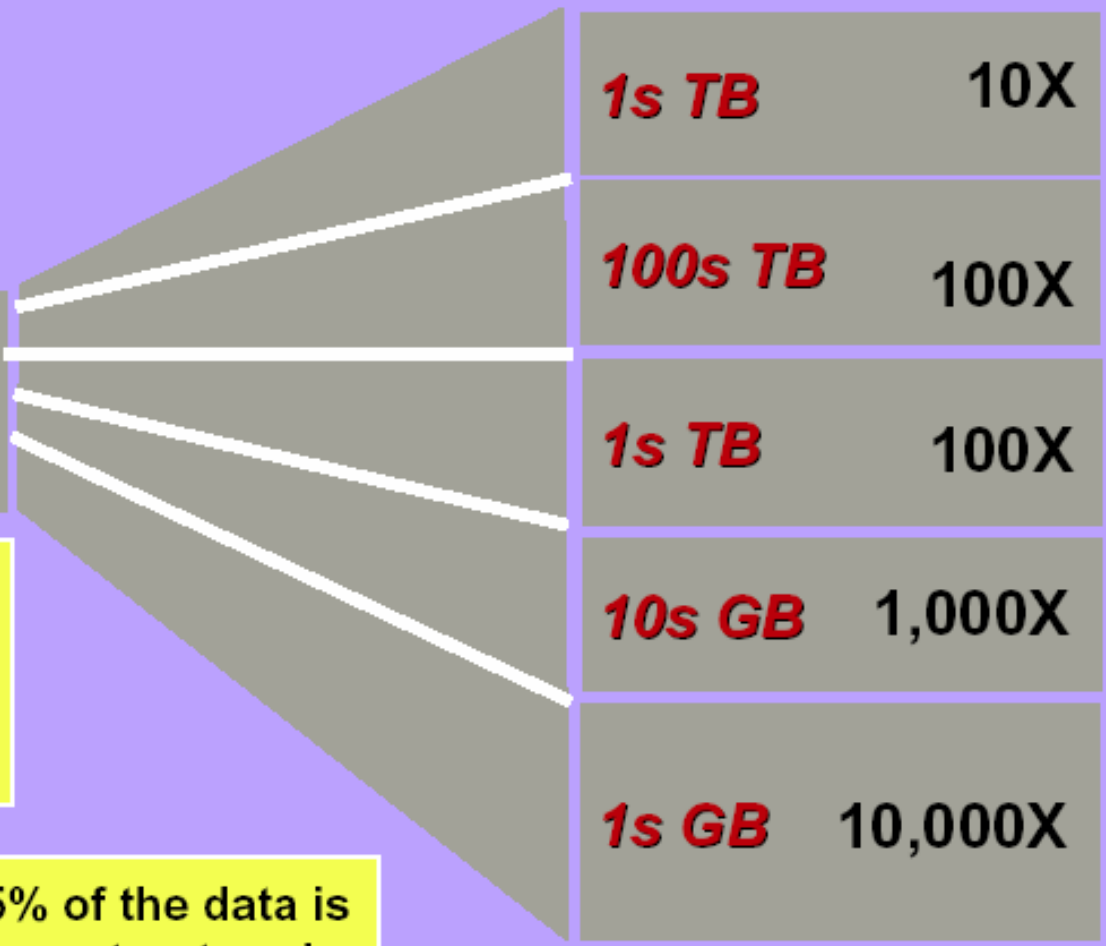
Google claims to index more than 8 billion pages,
MSN claims about 5 billion pages,
Yahoo! at least 4 billion
and Ask/Teoma more than 2 billion.

Issue: Data Volumes Exploding

2010

Common Database Sizes

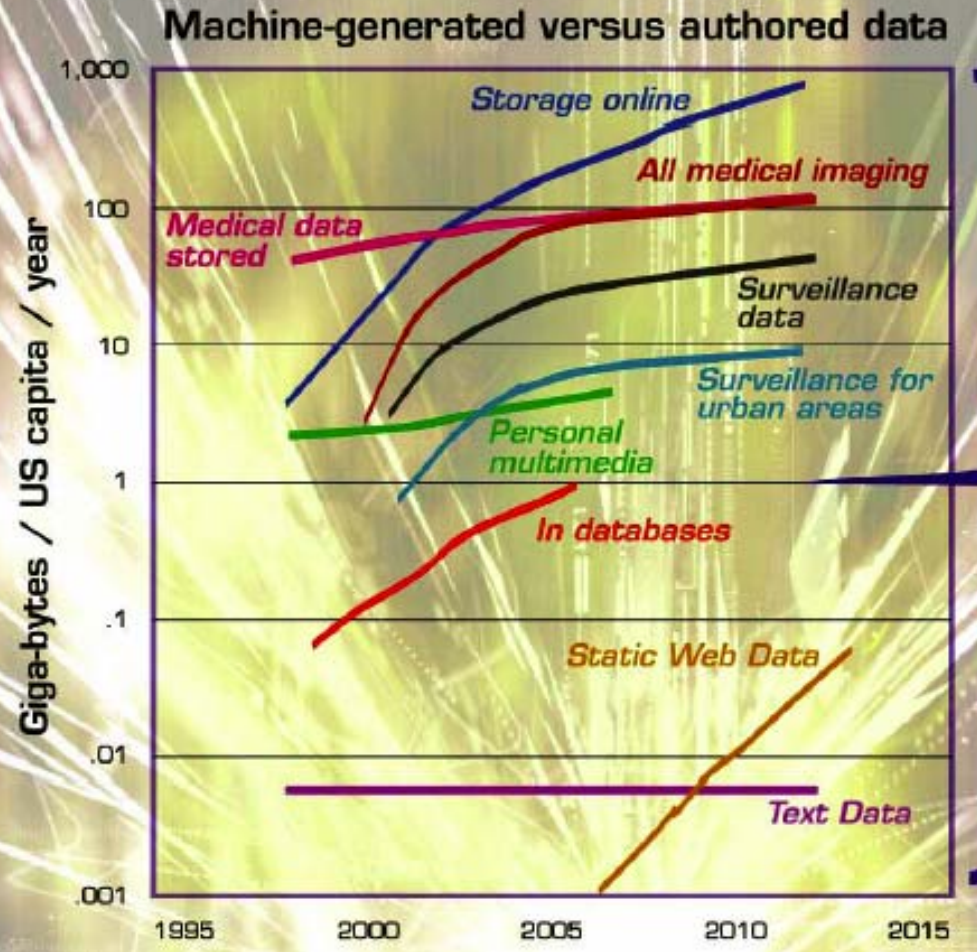
Workload	*2005*
Transactions	*100-500GB*
Warehouses	*100s GB – 10's TB*
Marts	*1 - 50 GBs*
Mobile	*100s MB*
Pervasive	*100s KB*



The world produces 250MB of information every year for every man, woman and child on earth.

85% of the data is unstructured.

Data Volume is Exploding



Machine Generated Data

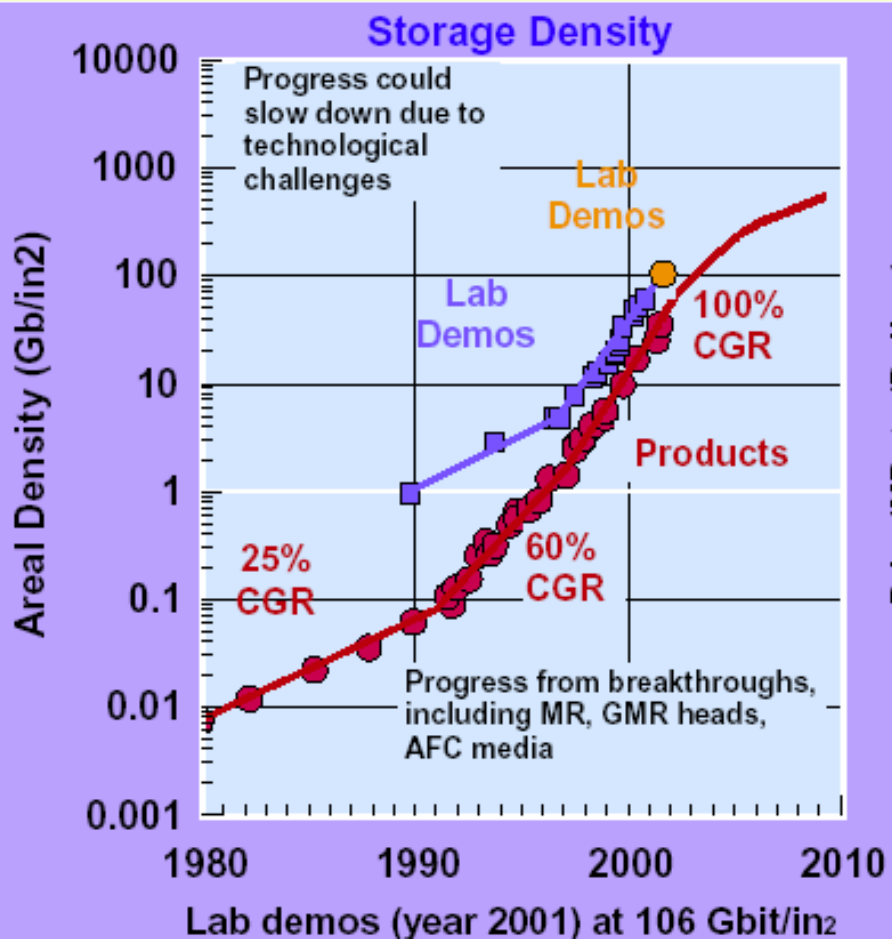
- Sensors
- High volume
- Not amenable to traditional database architectures

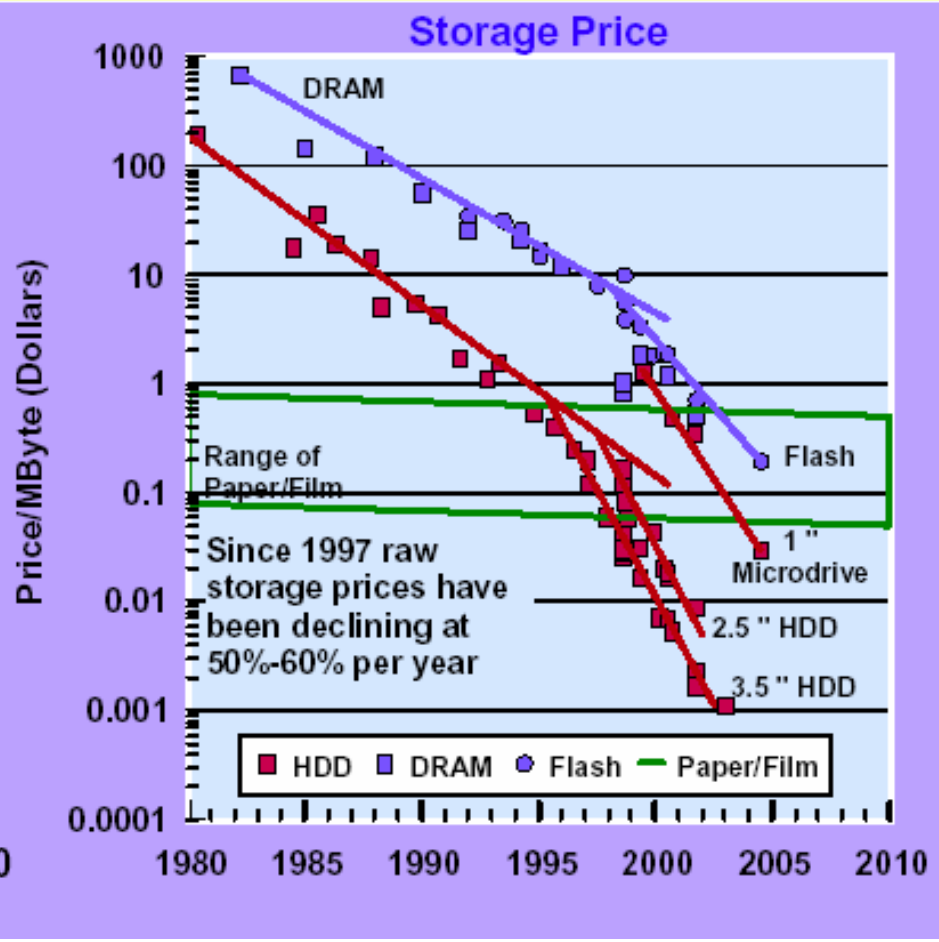
Authored Data

- Created by hand
- Low volume (but high value)

Storage Trends Aid this Data Explosion

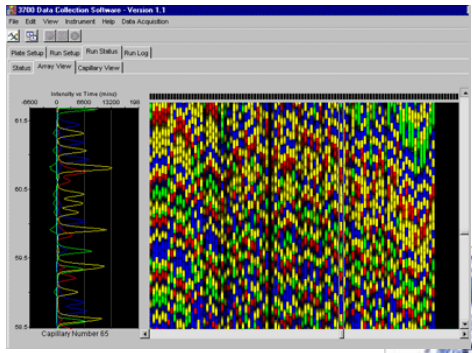
Storage aerial density CGR continues at 100% per year to >100 Gbit/in2.
The price of storage is now significantly cheaper than paper.





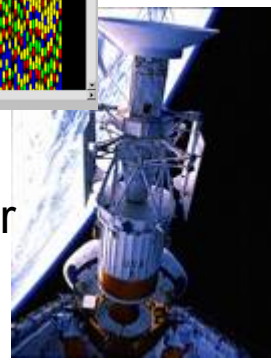
Computational Science Evolves

- Historically, Computational Science = simulation.
- New emphasis on informatics:
  - Capturing,
  - Organizing,
  - Summarizing,
  - Analyzing,
  - Visualizing
- Largely driven by observational science, but also needed by simulations.
- Too soon to say if comp-X and X-info will unify or compete.



BaBar, Stanford

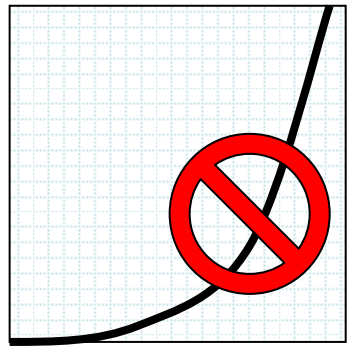
P&E
Gene Sequencer
From
http://www.genome.uci.edu/

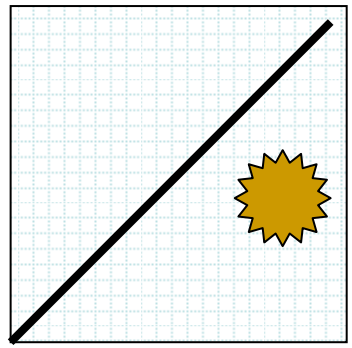


Space Telescope

Organization & Algorithms

- Use of clever data structures (trees, cubes):
  - Large speedup during the analysis
  - Tree-codes for correlations
  - Data Cubes for OLAP (all vendors)
- Fast, approximate heuristic algorithms
  - No need to be more accurate than data variance
- Take cost of computation into account
  - Controlled level of accuracy
  - Best result in a given time, given our computing resources
- Use parallelism
  - Many disks
  - Many cpus





Analysis and Databases

- Much statistical analysis deals with
  - Creating uniform samples –
  - data filtering
  - Assembling relevant subsets
  - Estimating completeness
  - censoring bad data
  - Counting and building histograms
  - Generating Monte-Carlo subsets
  - Likelihood calculations
  - Hypothesis testing

- Traditionally these are performed on files
- Most of these tasks are much better done inside DB

Information => Data => Knowledge

- KDD (Knowledge Discovery on Databases) does not exist in a vacuum
  - External forces can have more impact on the field than internal forces
- KDD is a young field with little history to guide it

  - in contrast the American Statistical Association is meeting for their 166nd annual meeting this year

- Reason for Data Model:

Data = $$

Within the scientific community:
- the data is much more dispersed
- the goal in modeling scientific problems was to find and to formulate governing laws in the form of precise mathematical terms.

- It has long been recognized that such perfect descriptions are not always possible. Incomplete and imprecise knowledge, observations that are often of a qualitative nature, the great heterogeneity of the surrounding world, boundaries and initial conditions being not completely know, all this generate the search for data models.
- To build a model that does not need complex mathematical equations, one needs sufficient and good data.

Motivation

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases
  - Oracle Data Mining, MS SQL Server, IBM DB2,…
  - Free and opened databases

What Is Data Mining?

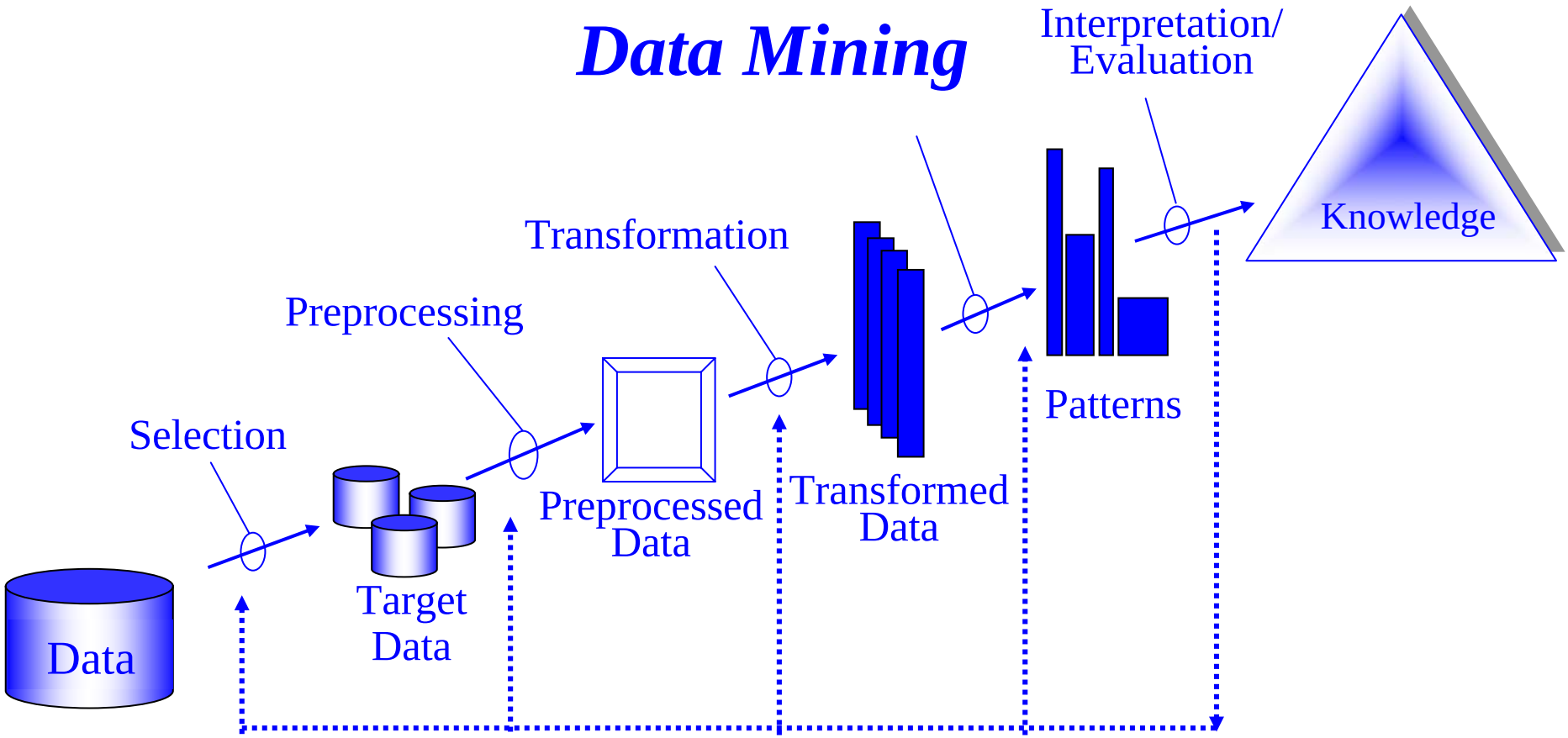
- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases

- What is not data mining?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs
  - …..
  - …..
  - Unstructured queries, OLAP(on line analytical processing) differs from SQL queries in the level of abstraction, or “open ended-ness” that can be included in the query
  - …..

Potential Applications

- Database analysis and decision support
  - Market analysis and management
  - Risk analysis and management
  - …..
- New Applications
  - Text mining (documents…)
  - Web mining

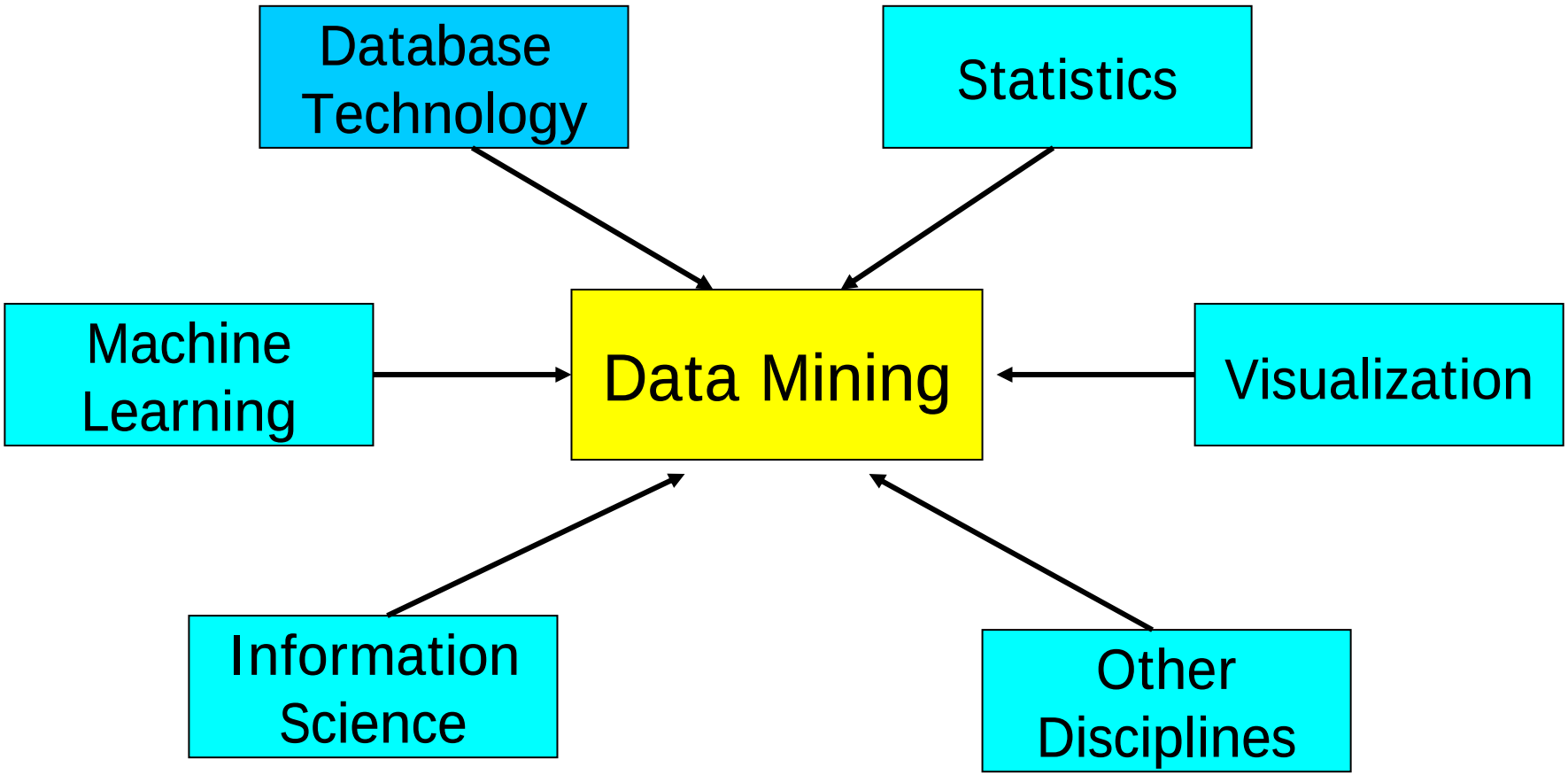
KDD - *Knowledge Discovery in Databases*



Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Multiple Disciplines



Major Tasks in Data Preprocessing

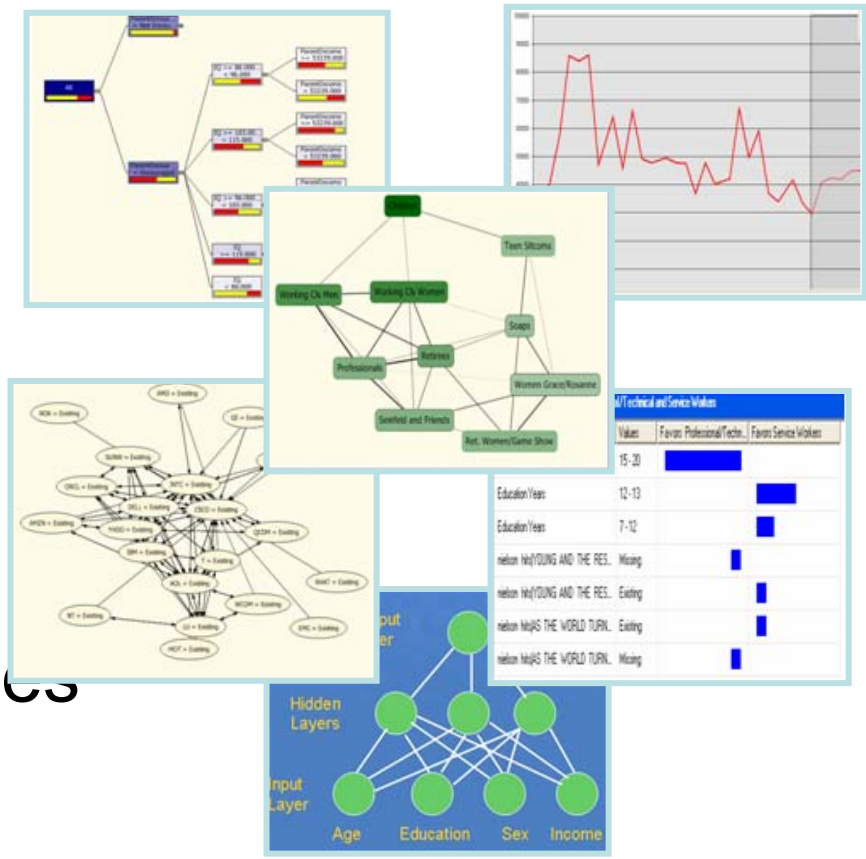
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Data Mining Tasks - Summary

- Classification
- Regression
- Segmentation
- Association Analysis
- Anomaly detection
- Sequence Analysis
- Time-series Analysis
- Text categorization
- Advanced insights discovery
- Others

Data Mining Algorithms

- Decision Trees
- Naïve Bayesian
- Clustering
- Sequence Clustering
- Association Rules
- Neural Network
- Time Series
- Support Vector Machines
- ….



Supervised and Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Semi-Supervised Learning

- To combine data with and without labels indicating the class of the observations to improve the model to be generated:
  - semi-supervised clasification: during the training explores the information obtained in large group of unlabeled data to improve the classification accuracy.
  - semi-supervised clustering: uses some small quantity of labeled data to improve the clustering process.

Classification Accuracy: Estimating Error Rates

- Partition: Training-and-testing
  - use two independent data sets, e.g., training set (2/3), test set(1/3)
  - used for data set with large number of samples
- Cross-validation
  - divide the data set into *k* subsamples
  - use *k-1* subsamples as training data and one sub-sample as test data—*k*-fold cross-validation
  - for data set with moderate size
- Bootstrapping (leave-one-out)
  - for small size data

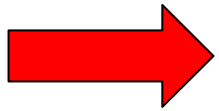
Bagging and Boosting

- General idea

Training data

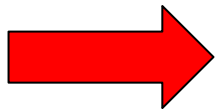
Classifier C

Altered Training data



Classifier C1

Altered Training data



Classifier C2

……..

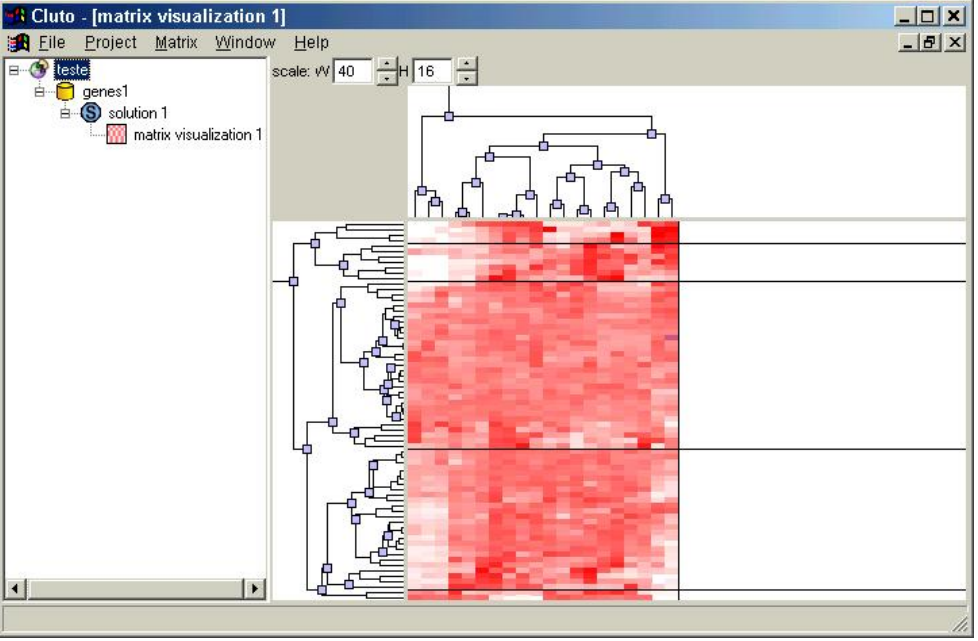
Aggregation ….

Classifier C*

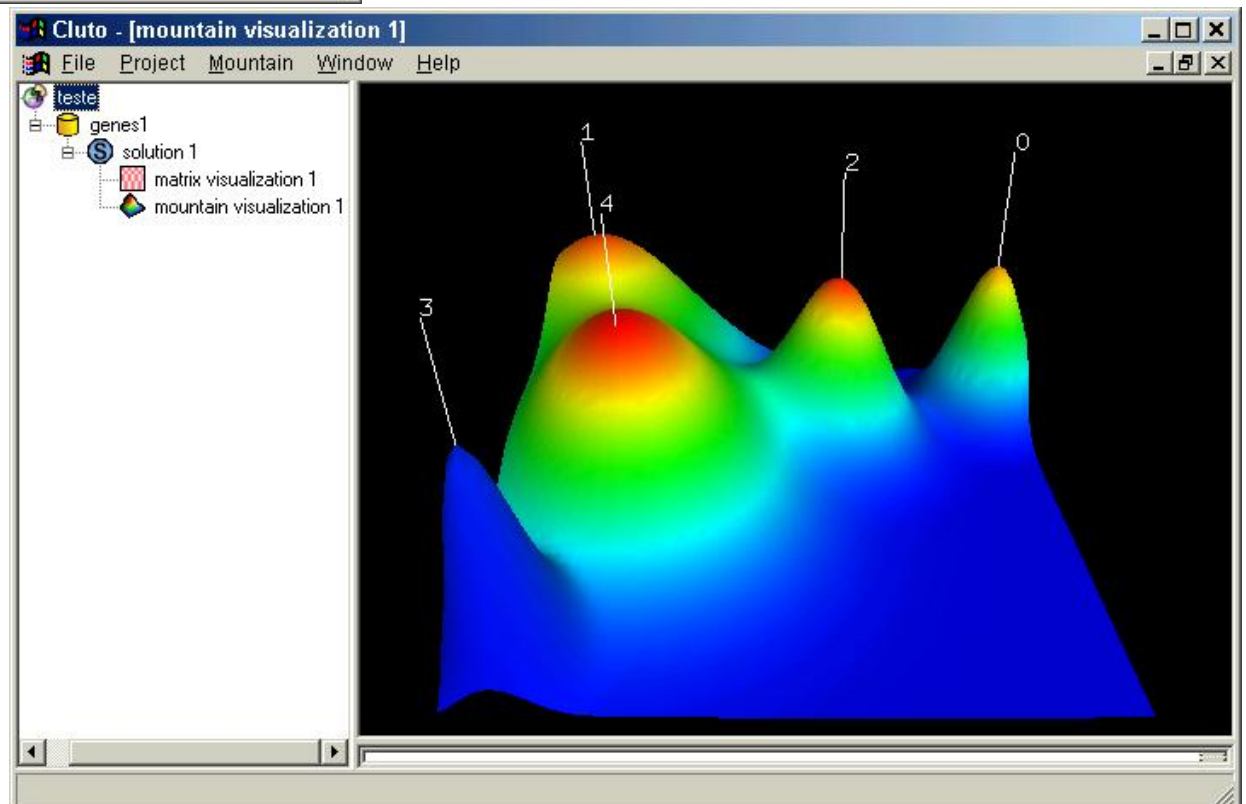
CLUTO VIZUALIZATION

www-users.cs.umn.edu/~karypis/cluto/

Matrix



Mountain



Are All the Discovered Patterns Interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**: A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures:**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

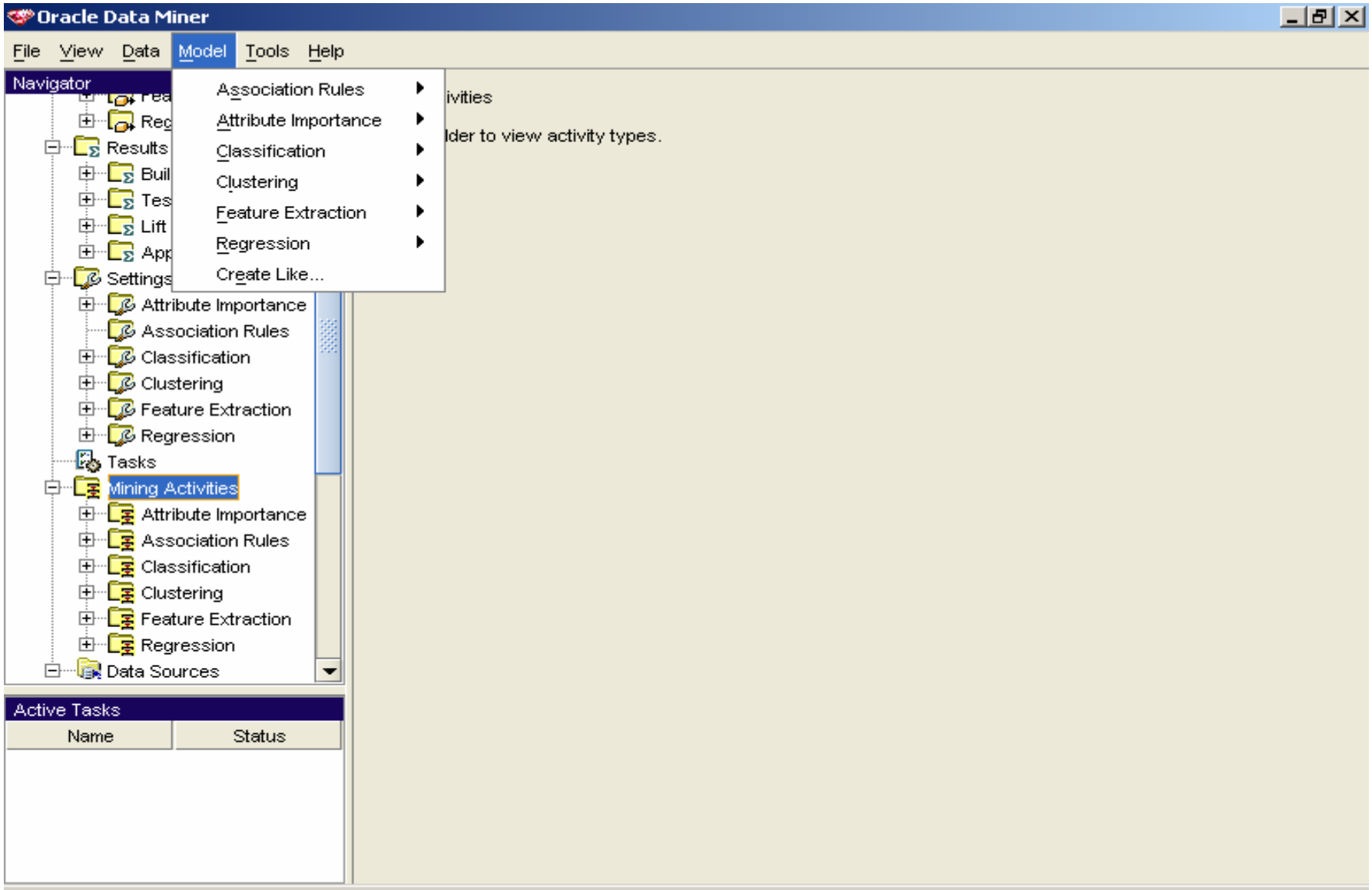
Applications

- Nowadays Data Mining is fully developed, very powerful and ready to be used. Data Mining finds applications in many areas. The more popular are:
- *Data Mining on Government*: Detection and Prevention of Fraud and Money Laundering, Criminal Patterns, Health Care Transactions, etc..
- *Data Mining for Competitive Intelligence*: CRM, New Product Ideas, Retail Marketing and Sales Pattern, Competitive Decisions, Future Trends and Competitive Opportunities, etc..
- *Data Mining on Finance*: Consumer Credit Policy, Portfolio Management, Bankruptcy Prediction, Foreign Exchange Forecasting, Derivatives Pricing, Risk Management, Price Prediction, Forecasting Macroeconomic Data, Time Series Modeling, etc..
- *Building Models from Data*: Applications of Data Mining in Science, Engineering, Medicine, Global Climate Change Modeling, Ecological Modeling, etc..

Data Mining System Architectures

- Coupling data mining system with DB/DW system
  - No coupling—flat file processing, not recommended
  - Loose coupling
    - Fetching data from DB/DW
  - Semi-tight coupling—enhanced DM performance
    - Provide efficient implement a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions
  - Tight coupling—A uniform information processing environment
    - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, query processing methods, etc.

Oracle Data Miner



MS SQL SERVER

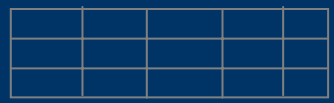
- Microsoft® (MS) Object Linking and Embedding Database for DM (OLE DB DM) technology provides an industry standard for developing DM algorithms. This technology was included in the 2000 release of the MS SQL Server™ (MSSQL). The Analysis Services (AS) component of this software includes a DM provider supporting two algorithms: one for classification by decision trees and another for clustering. The DM Aggregator feature of this component and the OLE DB DM Sample Provider made possible for developers and researchers to implement new DM algorithms.
- The MSSQL 2005 version has included more five algorithms: Naïve Bayes, Association, Sequence Clustering, Time Series and Neural Net as well as a new way to aggregate new algorithms, using a plug-in approach instead of DM providers.

Typical DM Process Using DMX

Define a model:
CREATE MINING MODEL ….

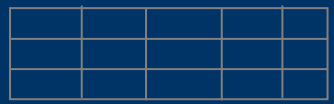
Train a model:
INSERT INTO dmm ….

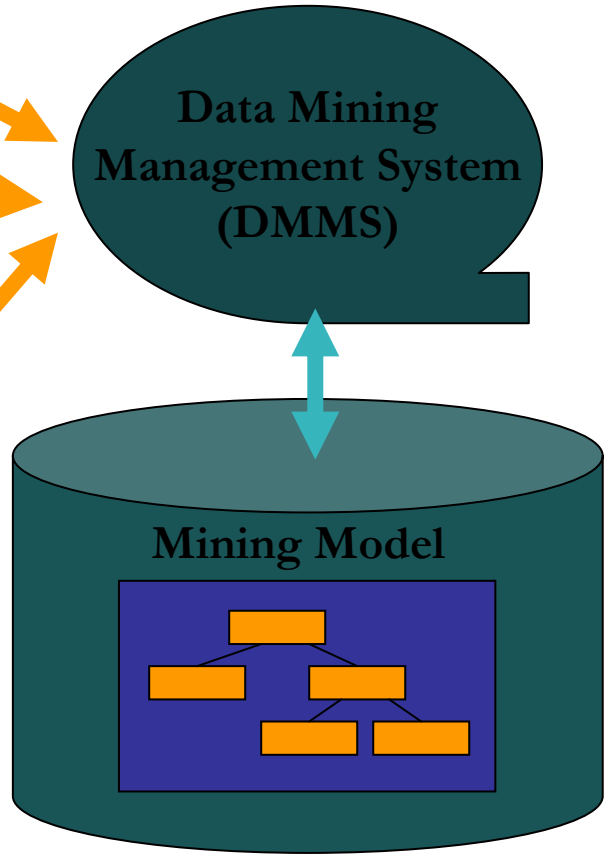
Training Data



Prediction using a model:
SELECT …
FROM dmm PREDICTION JOIN …

Prediction Input Data





Defining a DM Model

- Defines
  - Shape of “training cases” (top-level entity being modeled)
  - Input/output attributes (variables): type, distribution
  - Algorithms and parameters
- Example

```
CREATE MINING MODEL CollegePlanModel
(
   StudentID     LONG   KEY,
   Gender        TEXT   DISCRETE,
   ParentIncome  LONG   NORMAL CONTINUOUS,
   Encouragement TEXT   DISCRETE,
   CollegePlans  TEXT   DISCRETE PREDICT
) USING Microsoft_Decision_Trees
  (complexity_penalty = 0.5)
```

Nested Table Concept

- Complex types
  - Sets
  - Sequences
  - Time-series
- Sparse data
  - No need to pivot
  - Purchases, Terms

```
CREATE MINING MODEL Customer (
  ID                LONG     KEY,
  Income            DOUBLE CONTINUOUS,
  Card_Type         TEXT     DISCRETE,
  Card_ExpDate      DATE     CONTINUOUS,
  Card_IssueBank    TEXT     DISCRETE,
  Hobbies TABLE (
      Hobby         TEXT     KEY
  )
  Purchases TABLE (
      Name          TEXT     KEY,
      Quantity      LONG     CONTINUOUS
  )
  WebPageSeq TABLE (
      Page          TEXT     KEY SEQUENCE,
      Duration      LONG     CONTINUOUS
  )
) USING Microsoft_Sequence_Clustering
```

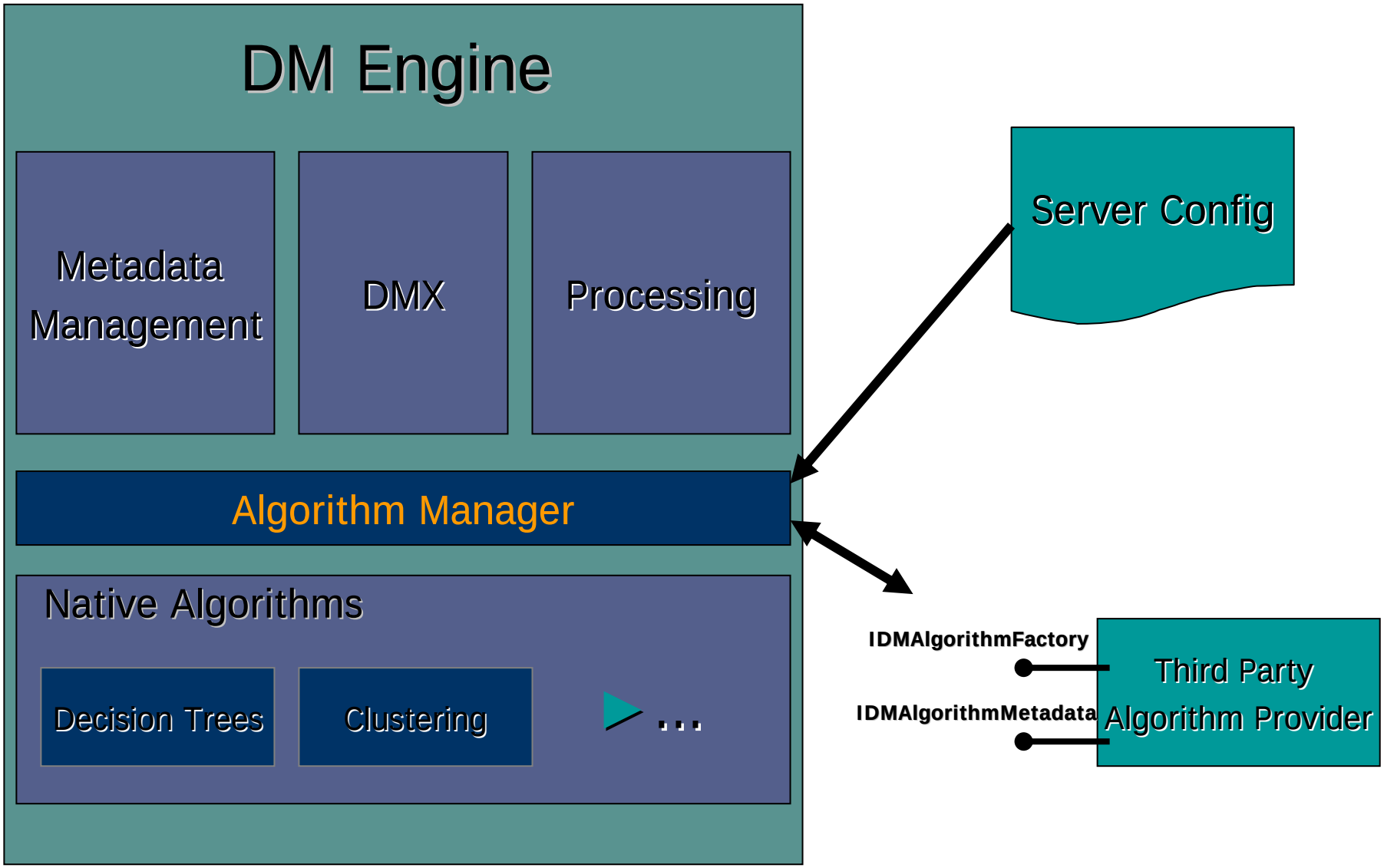
Background and Goals

- Background
  - No standard in data mining platform
    - Different ways of model development
    - Different interfaces between algorithms and other components
    - No standard language interface
    - Different application development and deployment
    - Different security models
  - Numerous algorithms for specific problems developed by academy, research and industry (in particular, niche vendors)
- Goal:
  - Allows algorithm experts from research, academy, and vendors to plug their algorithms into SQL Server so that their algorithm can take advantage of SQL Server DM functionality.
  - Positioning SQL Server DM as the leading platform for data mining

Advantages of Plug-In

- Direct use of SQL DM platform support
  - DM model development tools
  - DM mode browsers (customizable)
  - DM interfaces: XMLA, DMX, UDF
  - Application development environment: ADOMD.NET, AMO
  - Security, transaction, persistency support
  - All DM common components: case generation, transformation
  - And more…
- With little effort, the developers can make their algorithm available in mainstream of DM platform.
- Analogy: a DBMS that supports plug-in interface for an access method.

Plug-In Architecture



Incremental and Parallel Mining of Concept Description

- Incremental mining: revision based on newly added data ΔDB
  - Generalize ΔDB to the same level of abstraction in the generalized relation R to derive ΔR
  - Union R U ΔR, i.e., merge counts and other statistical information to produce a new relation R'
- Similar philosophy can be applied to data sampling, parallel and/or distributed mining, etc.

Multi-relational DM

- The common approach to solve this problem is to use a single flat table assembled by performing a relational join operation on the tables. But this approach may produce an extremely large and impractical to handle table, with lots of repeated and null data.
- In consequence, multi-relational DM (MRDM) approaches have been receiving considerable attention in the literature. These approaches rely on developing specific algorithms to deal with the relational feature of the data.
- By another way, OLE DB DM technology supports nested tables (also know as table columns). The row sets represent uniquely the tables in a nested way. There are no redundant or null data in each row set. Ex. one row per customer is all that is needed, and the nested columns of the row set contain the data pertinent to that customer.

Unique Table Approach

- Most classifiers work on a single table (attribute-value learning) with a fixed set of attributes.
- It is restrictive in DM applications with multiple tables.
- It is possible to construct a single table by performing a relational join operation on multiple tables using propositional logic.

Single Table Approach - Relational Join Operation

- For one-to-one and many-to-one relationships, we can join in the extra fields to the original relation without problems
- For one to many relationships, we have drawbacks:
  - It produces an extremely large table.
  - It produces data redundancy and lots of null values. The data duplication may introduce statistical skew.
  - loss of meaning
  - loss of information through aggregation

Multi-relational DM algorithms

- Group of several approaches:
  - Propositional learning (join of tables by propositionalization);
  - Inductive Logic Programming (ILP);
  - First Order Bayesian Networks (FOBN);
  - Multi-Relational Data Mining (MRDM).

Multi-relational DM algorithms

- Propositional learning
  - 1st step - The single table is constructed automatically using aggregate functions to deal with redundancy, loss of meaning and information;
  - 2nd step - Performed by any DM algorithm, such as Decision Trees, Naive Bayes, Bayesian Networks, Association Rules, Classification Rules, and so on.

Multi-relational DM algorithms

– Inductive Logic Programming (ILP) includes:
  - Progol;
  - First Order Inductive Logic (FOIL);
  - Top-down Induction of First-order Logical Decision Trees (TILDE);
  - Inductive Constraint Logic (ICL);
  - CrossMine.

Multi-relational DM algorithms

– First Order Bayesian Networks (FOBN):
  - Probabilistic Relational Model (PRM)
  - Probabilistic Logic Program (PLP)
  - Bayesian Logic Program (BLP)
  - Stochastic Logic Program (SLP)

Multi-relational DM algorithms

– Multi-Relational Data Mining
  - Multi-Relational Decision Tree Induction
    Multi-Relational Decision Tree (MRDT)
    Multi-Relational Naïve Bayes Classifier (Mr-SBC)
  - Multi-Relational Model Trees (Mr-SMOTI) (with support to regression)

OLE DB DM nested tables

– MS OLE DB DM:
  - Nested Data Mining Columns (nested tables);
  - Data mining models use this nested column structure for both input and output data, as the syntax used to populate a data mining model with training data allows nested columns to be represented as sub-queries.

Unstructured Information

- Text
- Chat
- Email
- Audio
- Video
- ...

Structured Information

Text Mining

- Is the art and science of extracting information and knowledge from text.
- The practice builds upon related disciplines:
  - Information Retrieval
  - Computational linguistics
  - Pattern recognition

Web Mining

- Is the use of data mining techniques to automatically discover and extract information from Web documents and Services
- Find relevant information
- Create new knowledge out the information available on the Web
- Personalization of the Information
- Learning about consumers or individual users

Extending "Mission-Critical" to Unstructured Data

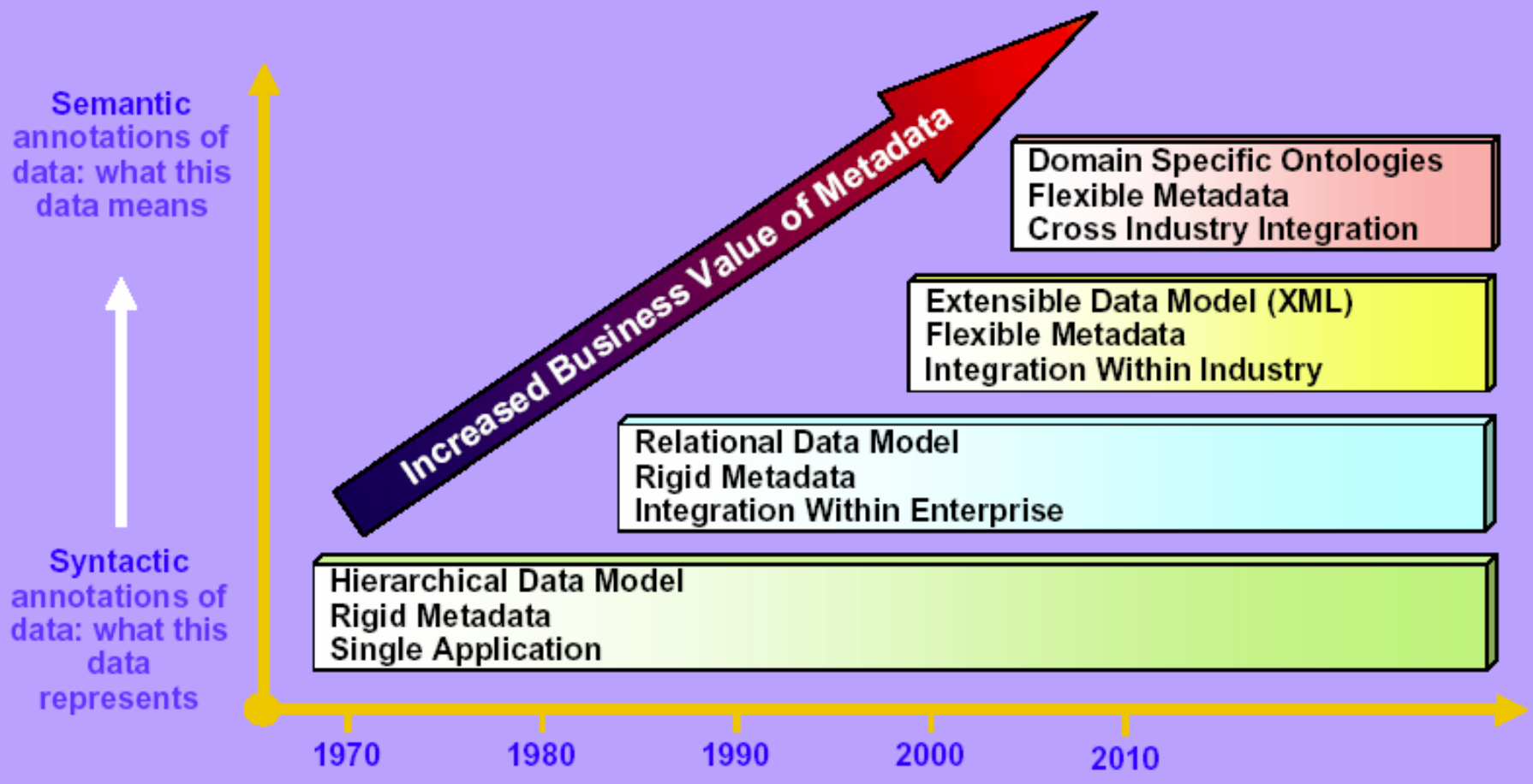
XML has become the "data interchange" format.

- ✓ ***XML View Of Relational Data***
  - SQL data viewed and updated as XML
    - Done via document shredding and composition
  - DTD and Schema Validation
- ✓ ***XML Documents As Monolithic Entities***
  - Atomic Storage And Retrieval
  - Search Capabilities

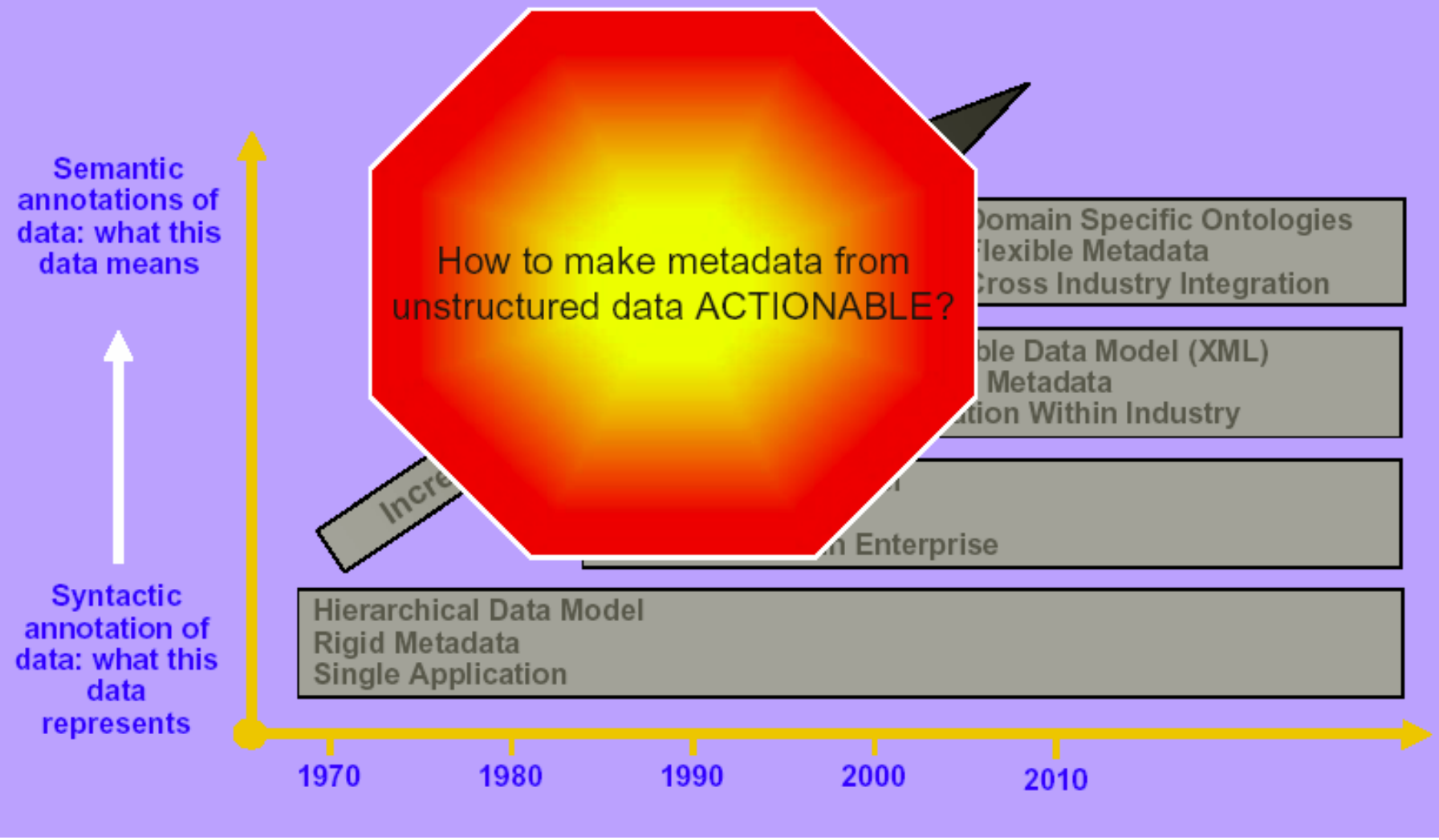
Next: *XML As A Rich Datatype*

- Full storage and indexing
- Powerful querying capabilities

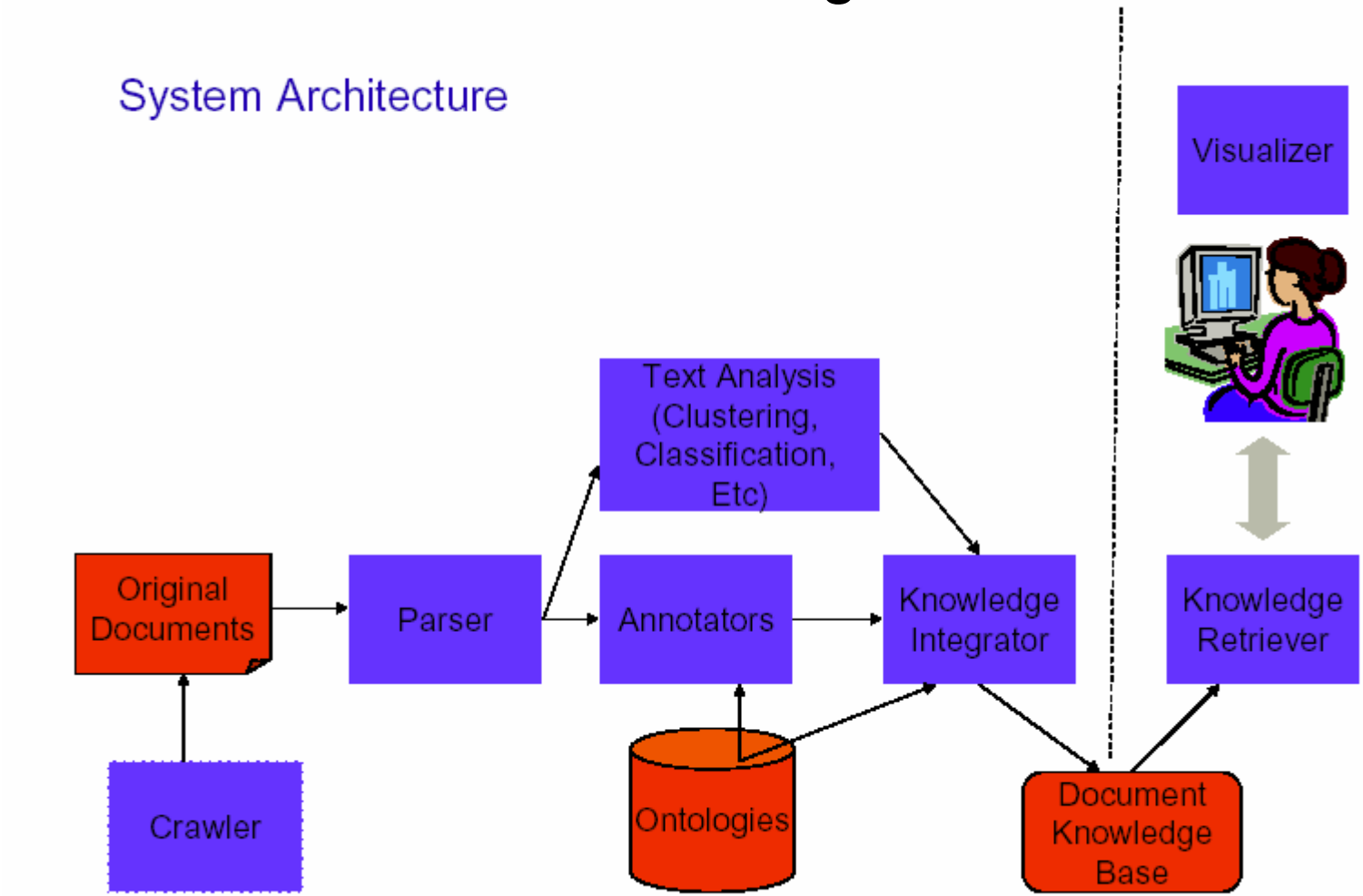
Evolution of Metadata



Metadata from UNSTRUCTURED data is growing exponentially



IBM integration



Text Mining with Oracle 10g Text

Oracle Text provides features for supervised and unsupervised classification. The ability to find, classify, cluster and visualize documents, based on their textual, content metadata or attributes, makes the Oracle database the single point of integration for all data management

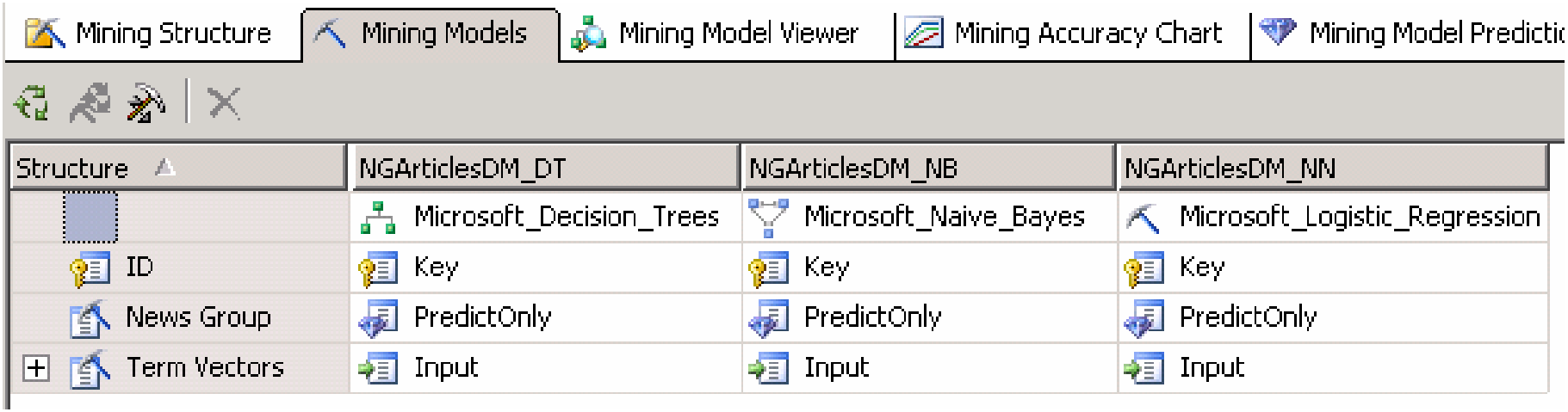
Rule Based Classification: Decision Trees and SVM

Two main clustering techniques: hierarchical (nested series of partitions) and partitional (Kmeans).

Document level visualization: Themes and Keyword

Visualization of Structures: categories and clusters

Text Classification using SQL Server 2005 Data Mining



International Conference on Data Mining

WIT Press, Southampton

- September 2-4, 1998
  Rio de Janeiro, Brazil
- July 5-7, 2000
  Cambridge, UK
- October 16-18, 2002
  Bologna,Italy
- December 1-3, 2003
  Rio de Janeiro, RJ
- September 15-17, 2004
  Malaga, Spain
- May 25-27, 2005
  Skiathos, Greece

July 11-13 2006

Prague, Czech Republic