

# Knowledge Discovery and Data Mining: Applications, Techniques, and Performance Issues

**Liaquat Majeed Sheikh**

**[liaquat.majeed@nu.edu.ok](mailto:liaquat.majeed@nu.edu.ok)**

# Trends leading to Data Flood

- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific Data: astronomy, biology, etc
  - Web, text, and e-commerce
- More data is captured:
  - Storage technology faster and cheaper
  - DBMS capable of handling bigger DB

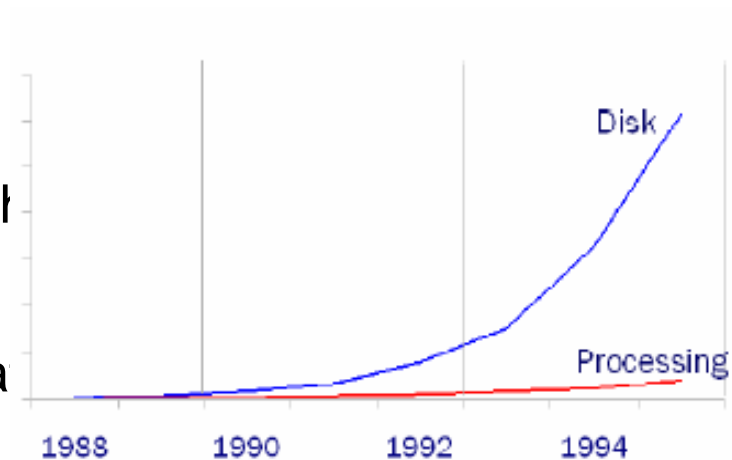


# Examples

- Walmart reported to have 24 Tera-byte DB
- AT&T handles billions of calls per day
  - data cannot be stored -- analysis is done on the fly

# Growth Trends

- Moore's law
  - Computer Speed doubles every 18 months
- Storage law
  - total storage doubles every 9 months
- Consequence
  - very little data will ever be looked at by a human



- Knowledge Discovery is **NEEDED** to make sense and use of data.

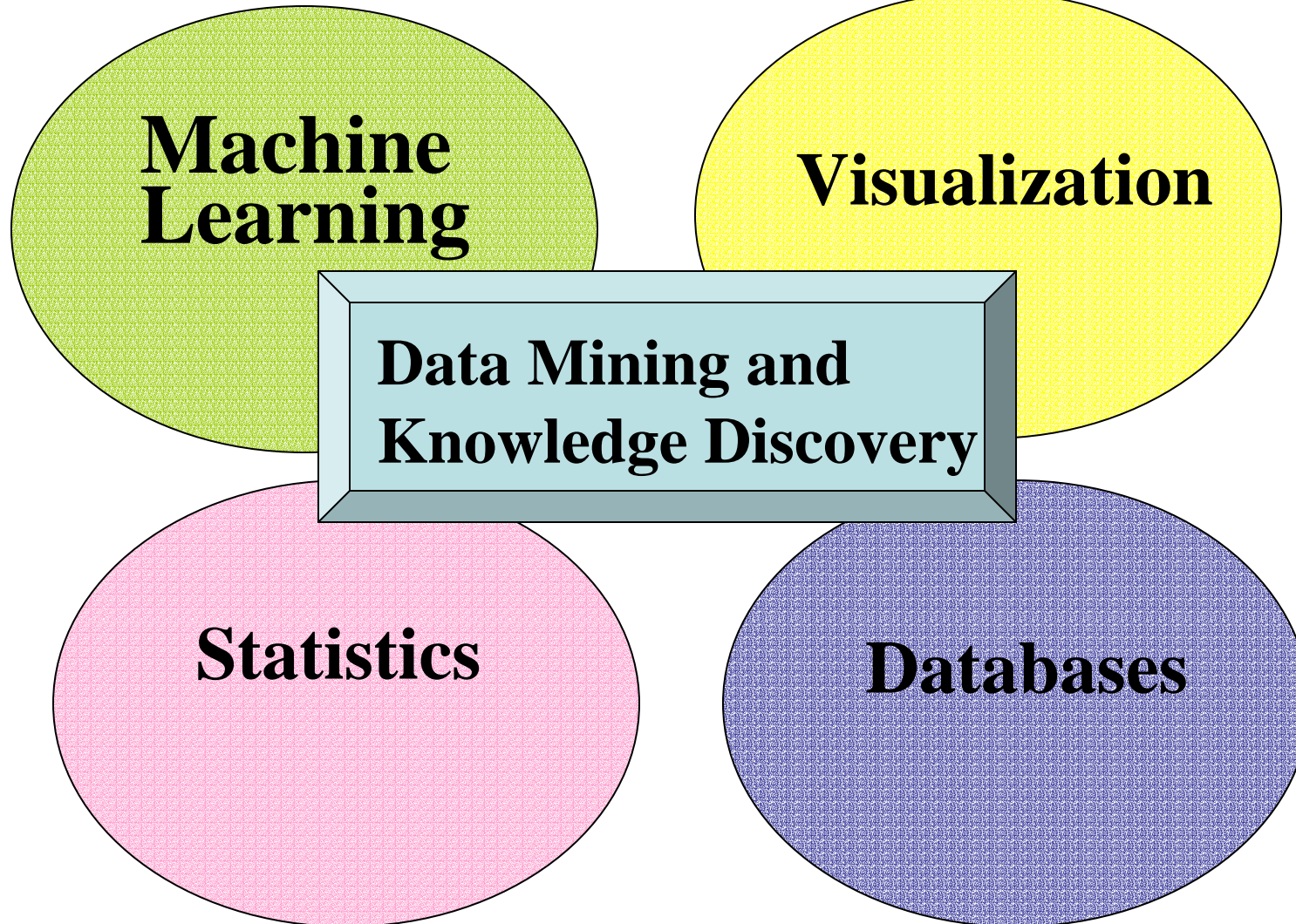
# Knowledge Discovery Definition

Knowledge Discovery in Data is the *non-trivial* process of identifying

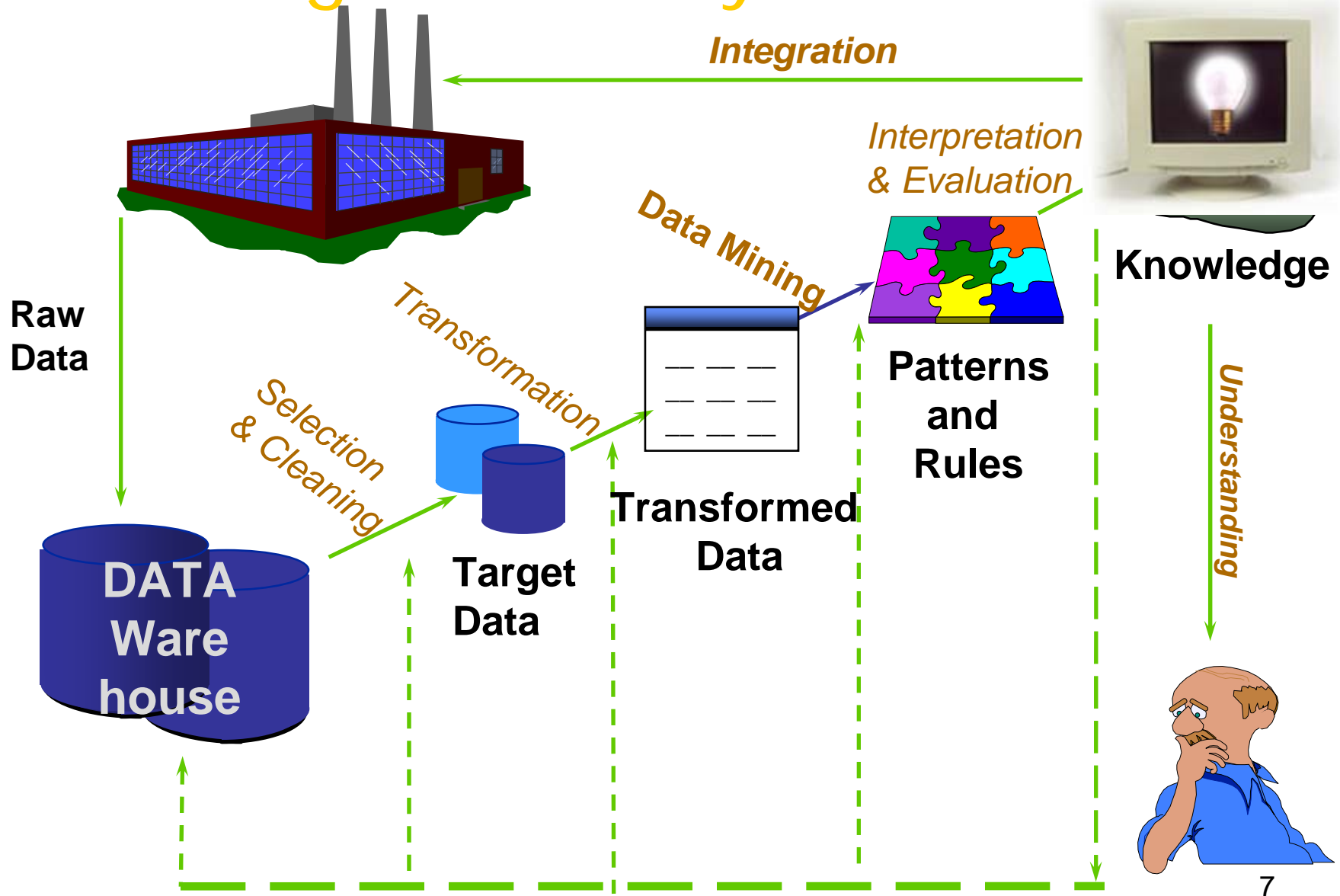
- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining*,  
Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy,  
(Chapter 1), AAAI/MIT Press 1996

# Related Fields

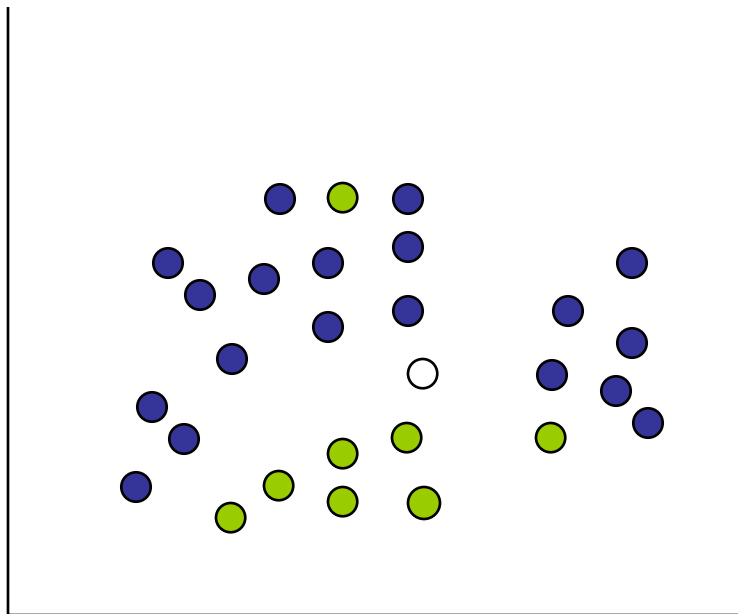


# Knowledge Discovery Process



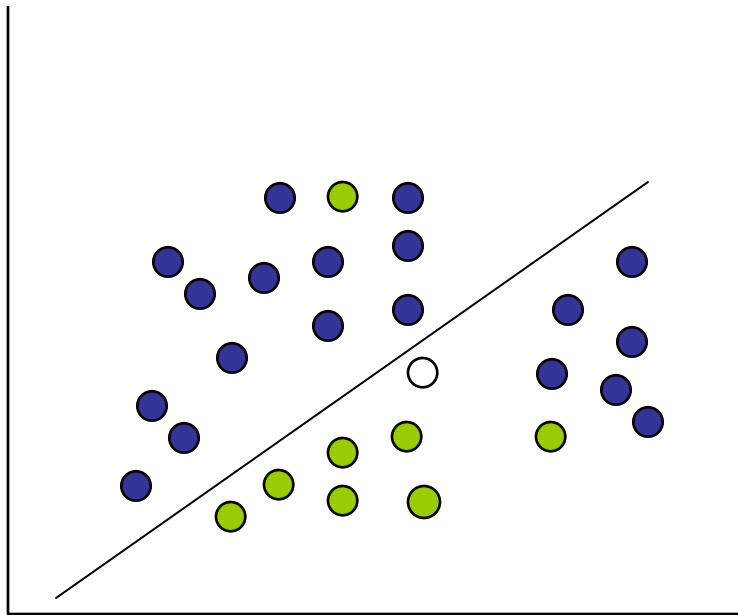
# Data Mining Tasks: Classification

**Learn a method for predicting the instance class  
from pre-labeled (classified) instances**



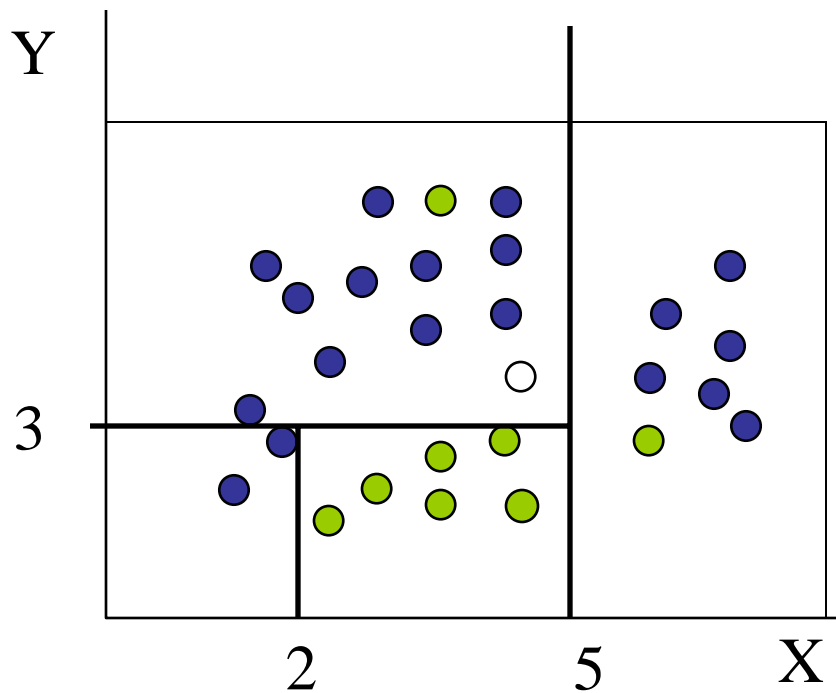
Many approaches:  
Statistics,  
Decision Trees,  
Neural Networks,  
...

# Classification: Linear Regression



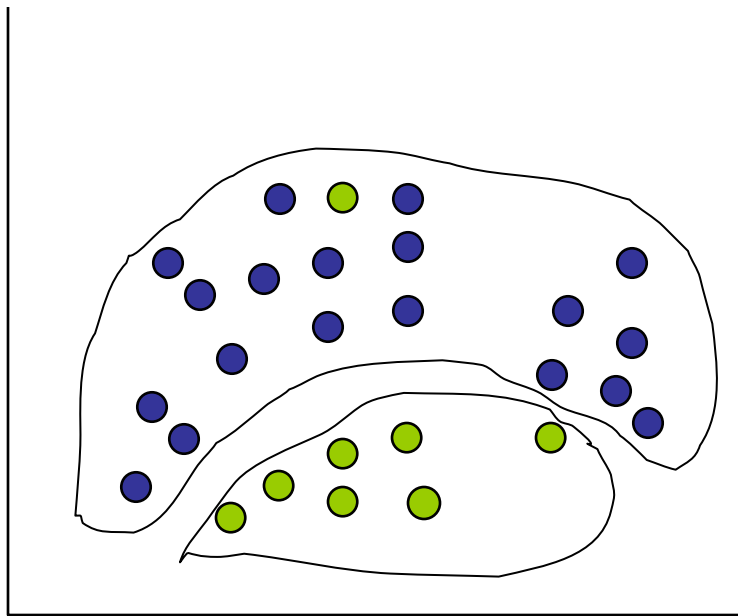
- Linear Regression  
 $w_0 + w_1 x + w_2 y \geq 0$
- Regression  
computes  $w_i$  from  
data to minimize  
squared error to  
'fit' the data
- Not flexible enough

# Classification: Decision Trees



if  $X > 5$  then blue  
else if  $Y > 3$  then blue  
else if  $X > 2$  then green  
else blue

# Classification: Neural Nets



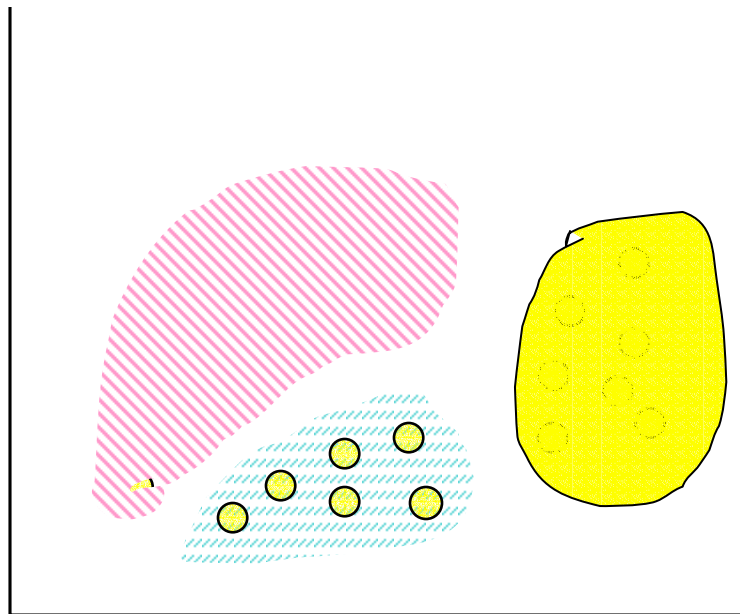
- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

# Data Mining Central Quest

Find true patterns  
and avoid *overfitting*  
(false patterns due  
to randomness)

# Data Mining Tasks: Clustering

Find “natural” grouping of instances given un-labeled data



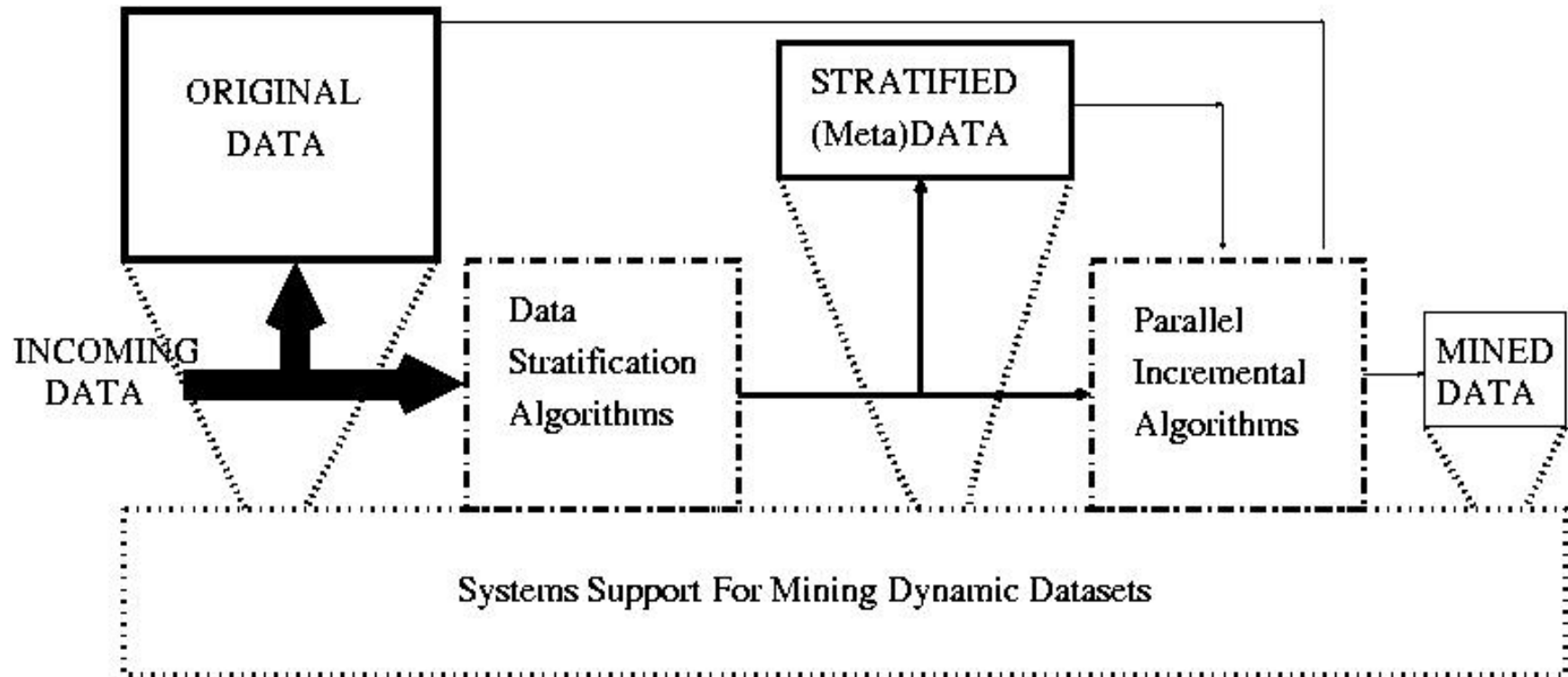
# Major Data Mining Tasks

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Estimation:** predicting a continuous value
- **Deviation Detection:** finding changes
- **Link Analysis:** finding relationships
- ...

# Active Mining Problem

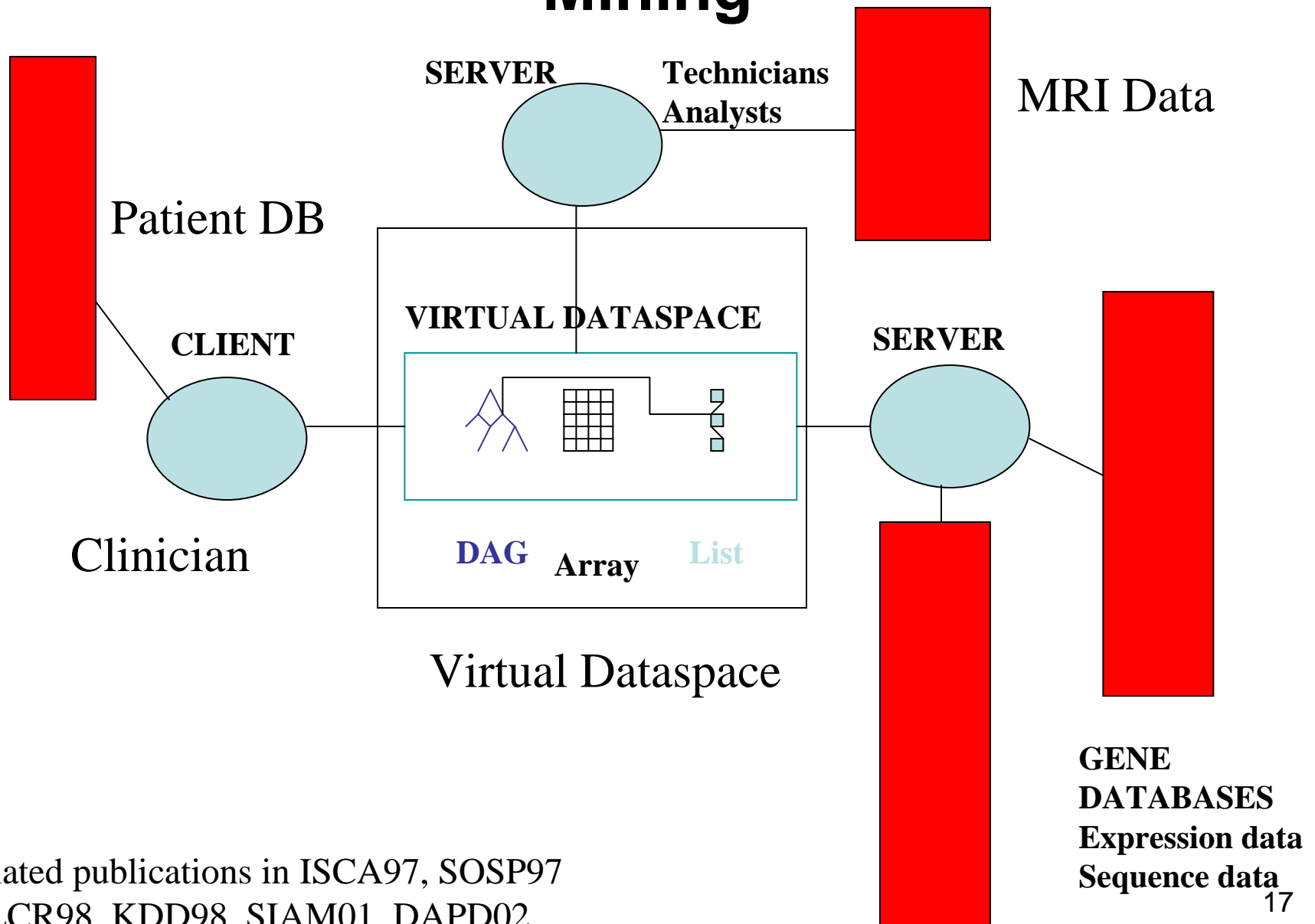
- Databases are constantly changing
- Mining is an interactive & iterative process
- Set of patterns  $P = f(\text{data}, \text{user parameters})$ 
  - Data update or User Interaction =>  $P_{\text{new}}$
- Most current approaches simply re-execute the algorithm to generate  $P_{\text{new}}$ 
  - High computational and I/O requirements
  - Poor response time to an interactive problem
  - Poor idea for time-sensitive mining!
- Can we do better? Yes!
  - Minimize access to previously processed data [CIKM99, SIAM02]
    - New data + summary → New summary
  - Approximate Results based on trends [IPDPS02, PKDD02]

# Active Mining on Streaming Data



- **Crucial Issue: Data Influx Rate Exceeds Processing Rate**
  - Problem for time-critical applications (e.g. Network Intrusion Detection)
- Data Stratification: Reduce rows (sampling), Reduce columns (PCA)
- Process incrementally, minimize access to original data.
- Systems support
  - Memory placement, Compression, Disk-filters, Data manipulation.

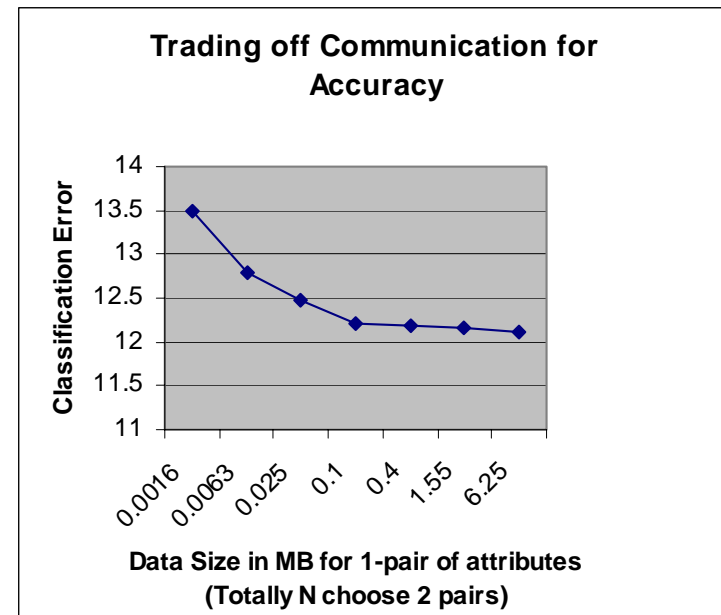
# InterAct: Middleware for Active Collaborative Mining



Related publications in ISCA97, SOSP97  
LCR98, KDD98, SIAM01, DAPD02

# Resource-Aware Data Mining

- *Distributed Data Mining in a Shared Environment*
- Problem
  - Dynamic Contention for Resources
    - E.g. TCP/IP contention-avoidance → poor comm. performance
    - Leads to unpredictable & poor performance
- Potential Solution:
  - Monitor resources, identify constraints
  - Have application adapt to resource constraints
  - Police resource allocation → predictable performance
- Can DM applications adapt?
  - Yes, see graph below for example
- How should the API look?
- Scheduling Policies?
- Resource sharing?

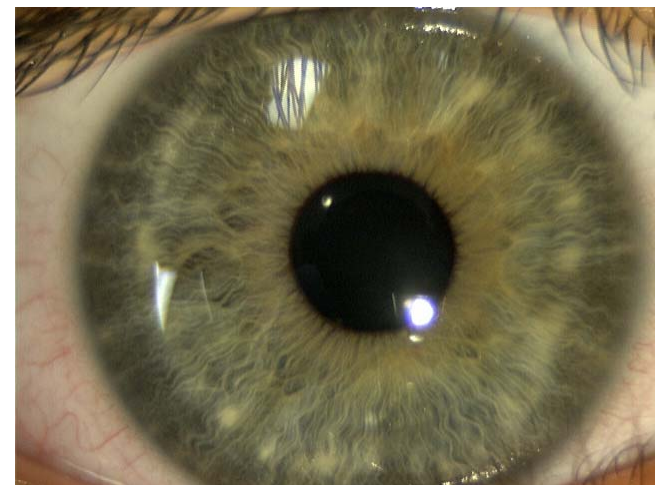
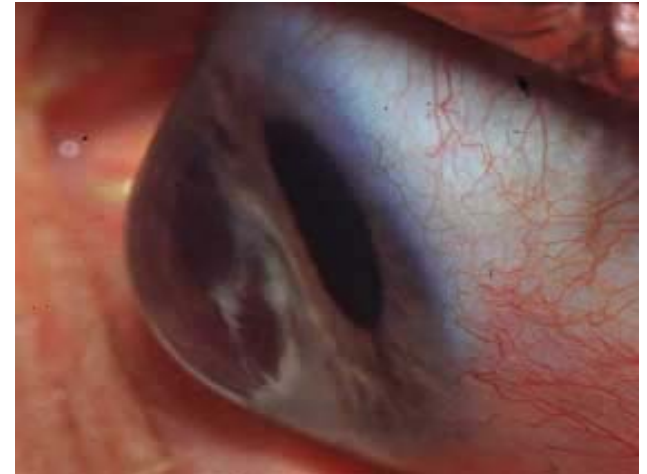


# Applications: Mining Scientific Data

- Challenges
  - Complex data with many structural relationships
    - Function-structure relationships abound
    - Modeling such relationships are crucial
  - Understanding and embedding the science in mining process is crucial
    - For performance and quality
  - The interface and resulting model needs to be in a form familiar to domain specialists
    - Otherwise the scientists are never going to use it!
  - Finally, in many cases there is a need for (soft) real-time analysis
    - Sometimes the data is streaming! E.g. simulation data

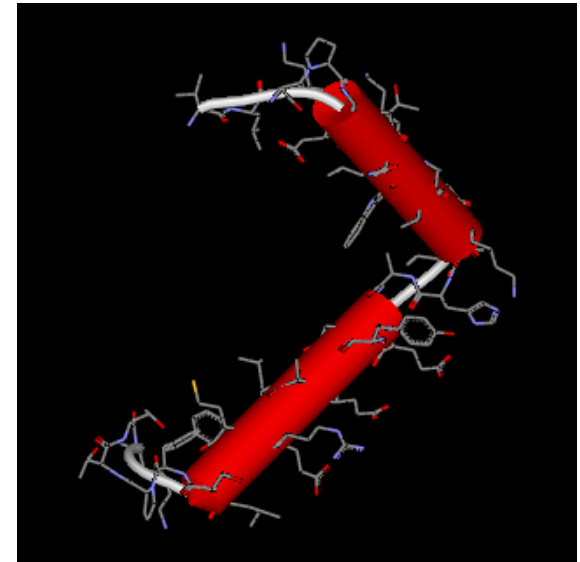
# Case Study: Keratoconus

- Problem (joint with M.Twa et al)
  - Classification of normal vs. keratoconic patients
  - Patient data + examination data
- Solution
  - Embedding the science
    - Need to model the shape and structure of cornea
    - Zernike representation
    - Emperically determine polynomial order
  - Apply easy to interpret classification model
    - Decision tree model
  - Results
    - 90-95% accuracy, easy to interpret, visual representation possible!

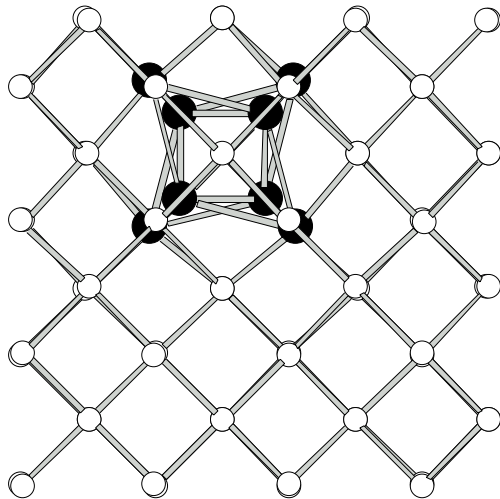
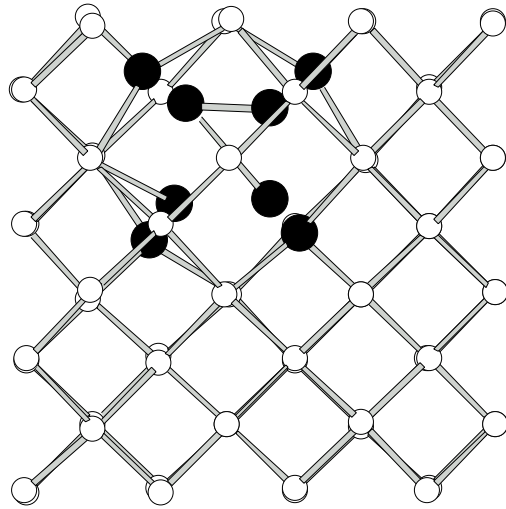


# Mining Protein Structure

- Protein Substructure Detection [ICDM02]
- Embedding the science
  - Need to model structure-activity relationship
  - Lots of self repeating structures → functional role?
- Goal: Find self repeating structures
  - Distance based structure representation
  - Frequency based pattern identification
  - Fuzzy Hashing for handling noisy data.
  - Key results so far
    - Can detect multi-level substructures
      - Backbone elements
      - Alpha Helix-related substructures (figure)
  - Same idea may be applicable to other scientific domains
    - MD simulation data
    - Structural similarity in Drugs



# Mining MD simulation data



- Utopian Objective
  - ***Real time analysis of dataset produced by MD simulation***
  - To understand the mechanics of defect creation, propagation in materials
  - Joint with R.Machiraju & J.Wilkins
- Techniques
  - **Feature Detection**
  - **Categorization of Defect Structure**
  - Feature Tracking
  - Spatio-Temporal Mining of Defects
- Embedding the science
  - Bond angles/Bond lengths/Atom types



# Other Projects

- Mining Text (email/newsgroup/citation) Data
- Mining Gene Expression Data
- Mining file system access
- Mining Performance/Simulation Trace Data
- Mining Lipid Phase Behavior
  - Joint with D. Kerr and M. Caffrey
- Coherence Issues in Distributed Federated Databases
  - Joint with M. Lauria & J. Saltz

# Conclusions

- **KDD**: A rich, promising, young field with broad applications and many challenging research issues.
- **Data mining tasks**: characterization, association, classification, clustering, prediction, sequence and pattern analysis, etc.
- **Data mining domains**: relational, transactional, text, spatial, time-series, multimedia, active DBs, data warehouses, and WWW.
- **Data mining methods**: Data-intensive, statistics, visualization, information science, and other disciplines.
- **Vision** : Scalable methods and high performance systems that support them.