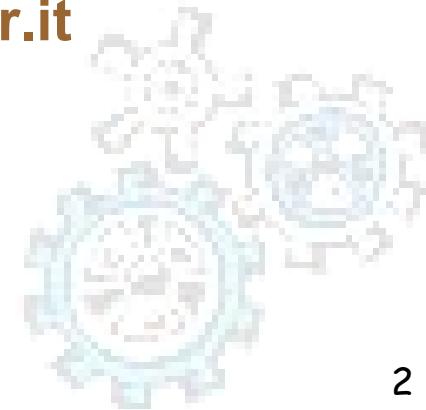


Data Mining a.a. 2009/10

Introduzione

Data Mining

- **Acronimo: DM**
- **Orario:**
 - **Martedì 14-16 aula B1**
 - **Giovedì 11-13 aula B1**
- **Docente:**
 - **Mirco Nanni, ISTI-CNR, mirco.nanni@isti.cnr.it**
- **Pagina web (wiki):**
 - **<http://www.cli.di.unipi.it/doku/doku.php/dm/>**



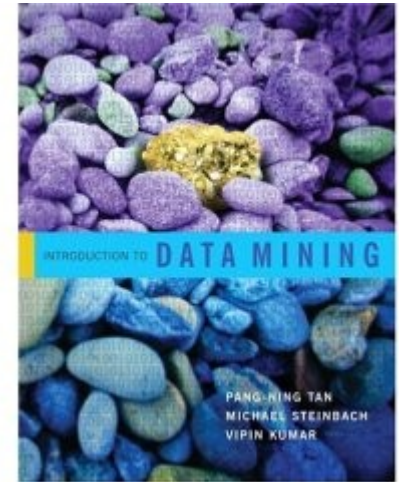
Data Mining

■ Testo di riferimento

Pang-Ning Tan, Michael Steinbach, Vipin Kumar

Introduction to DATA MINING

Addison Wesley, ISBN 0-321-32136-7, 2006



■ Altri riferimenti

- | Jiawei Han, Micheline Kamber. [Data Mining: Concepts and Techniques](#). Morgan Kaufmann Publishers, 2000
- | David J. Hand, Heikki Mannila and Padhraic Smyth. [Principles of Data Mining](#). MIT Press, 2001.
- | Barry Linoff. [Data Mining Techniques for Marketing Sales and Customer Support](#). John Wiles & Sons, 2002

■ I lucidi utilizzati nelle lezioni saranno resi disponibili attraverso il wiki del corso



Censimento studenti

Laurea spec./magistrale	#
Spec. Informatica	
Magistrale “	
Spec. Tecnologie Informatiche	2
Spec. Inf. per Ec. e Azienda	3
Magistrale “	
Spec. Informatica Umanistica	7
(Erasmus → Scienze comunicazione) (Triennale)	1
Informatica (Triennale)	2

Laurea provenienza	#
Informatica	6
Informatica Umanistica	7
Informatica Applicata	

Contenuti del corso

- Una parte preliminare dove si introducono i concetti essenziali del **processo di estrazione** della conoscenza: studio e preparazione dei dati, forme dei dati, misure e similarità dei dati
- Una parte centrale dove si introducono le principali **tecniche di datamining** (regole associative, classificazione e clustering). Di queste tecniche si studieranno gli aspetti formali e implementativi;
- Una parte più metodologica dove si visiteranno alcuni **casi di studio** nell'ambito del marketing, del supporto alla gestione clienti e dell'evasione fiscale
- L'ultima parte del corso ha l'obiettivo di introdurre **aspetti avanzati**, quali tecniche di data mining su dati complessi ed aspetti di privacy



Contenuti del corso in dettaglio

■ **Introduzione e Concetti Basici**

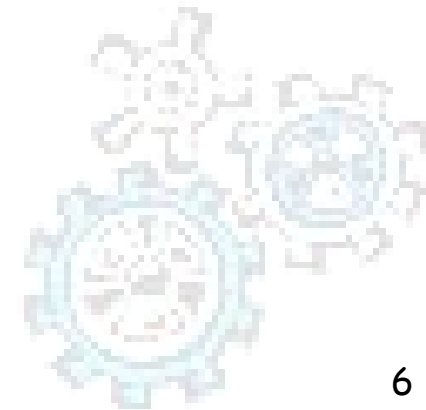
- **Il processo di knowledge discovery**
- **Esempi di applicazioni (Evasione fiscale, Business Intelligence)**

■ **Il processo di estrazione della conoscenza**

- **Le fasi iniziali: preparazione e pulizia dei dati**

■ **Introduzione alle tecniche di base**

- **Regole Associative**
- **Alberi di decisione**
- **Clustering**



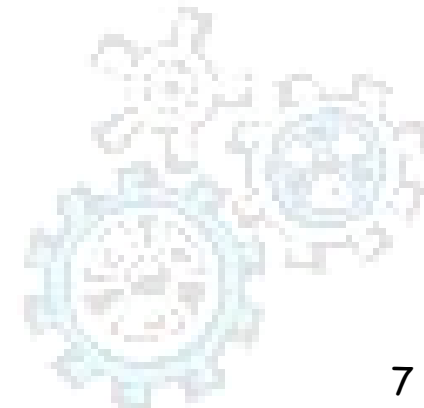
Contenuti del corso in dettaglio

■ **Algoritmi di Base**

- **Regole associative: algoritmo Apriori e varianti**
- **Alberi di Decisione: C4.5**
- **Clustering: K-Means, Hierarchical & Density-based**

■ **Argomenti avanzati**

- **Metodi e modelli alternativi**
- **Dati complessi (serie temporali, ecc.)**
- **Problemi di privacy**



Why Mine Data?

Commercial Viewpoint

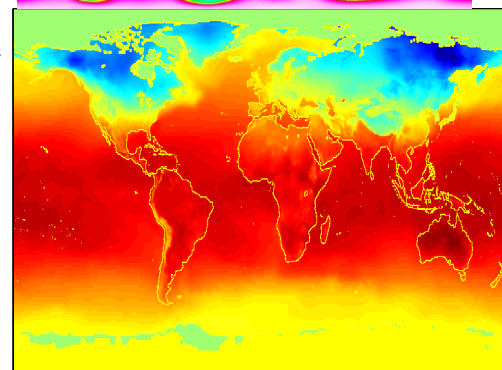
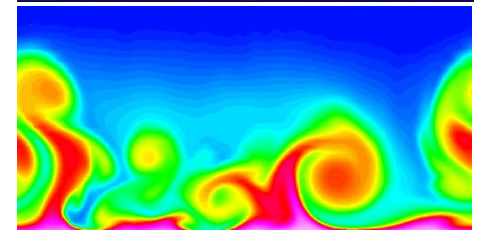
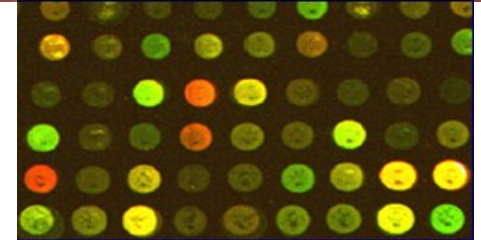
- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - purchases at department/grocery stores
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)



Why Mine Data?

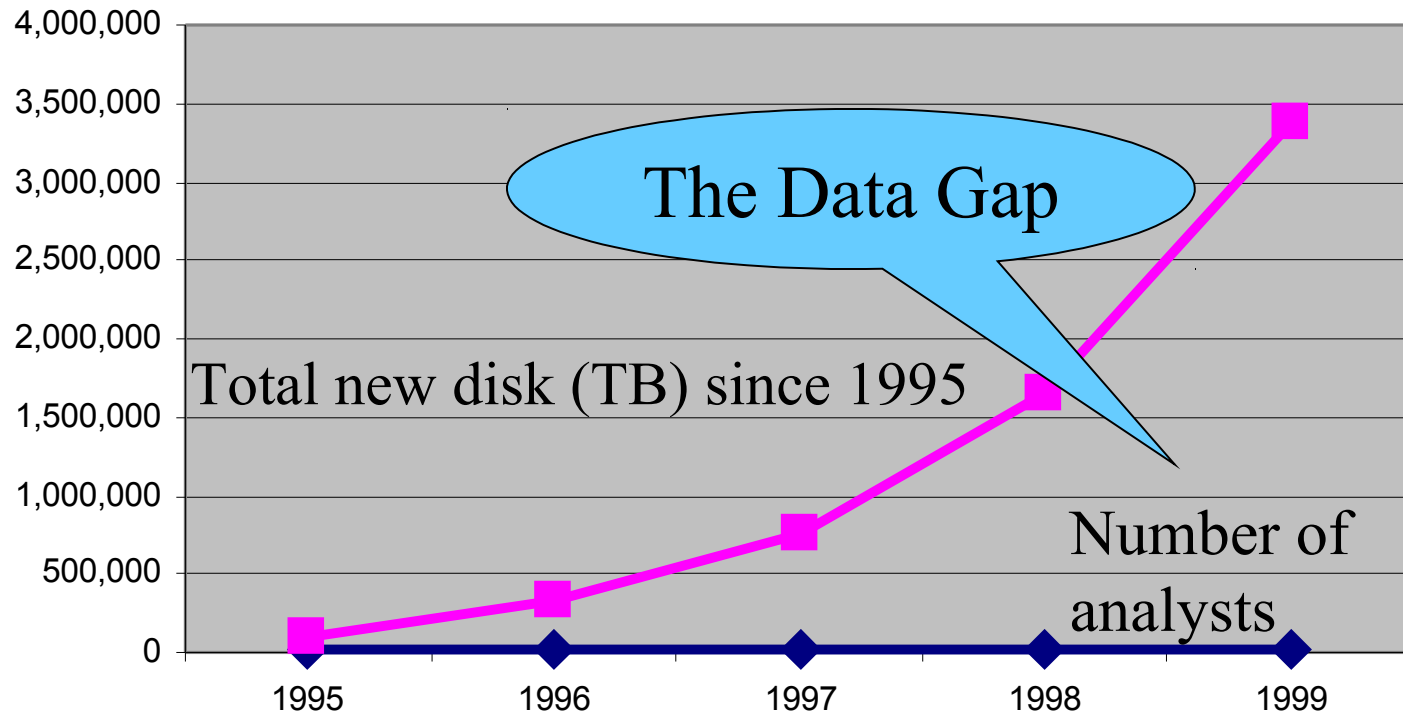
Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
 - remote sensors on a satellite
 - telescopes scanning the skies
 - microarrays generating gene expression data
 - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
 - in classifying and segmenting data
 - in Hypothesis Formation



Mining Large Data Sets - Motivation

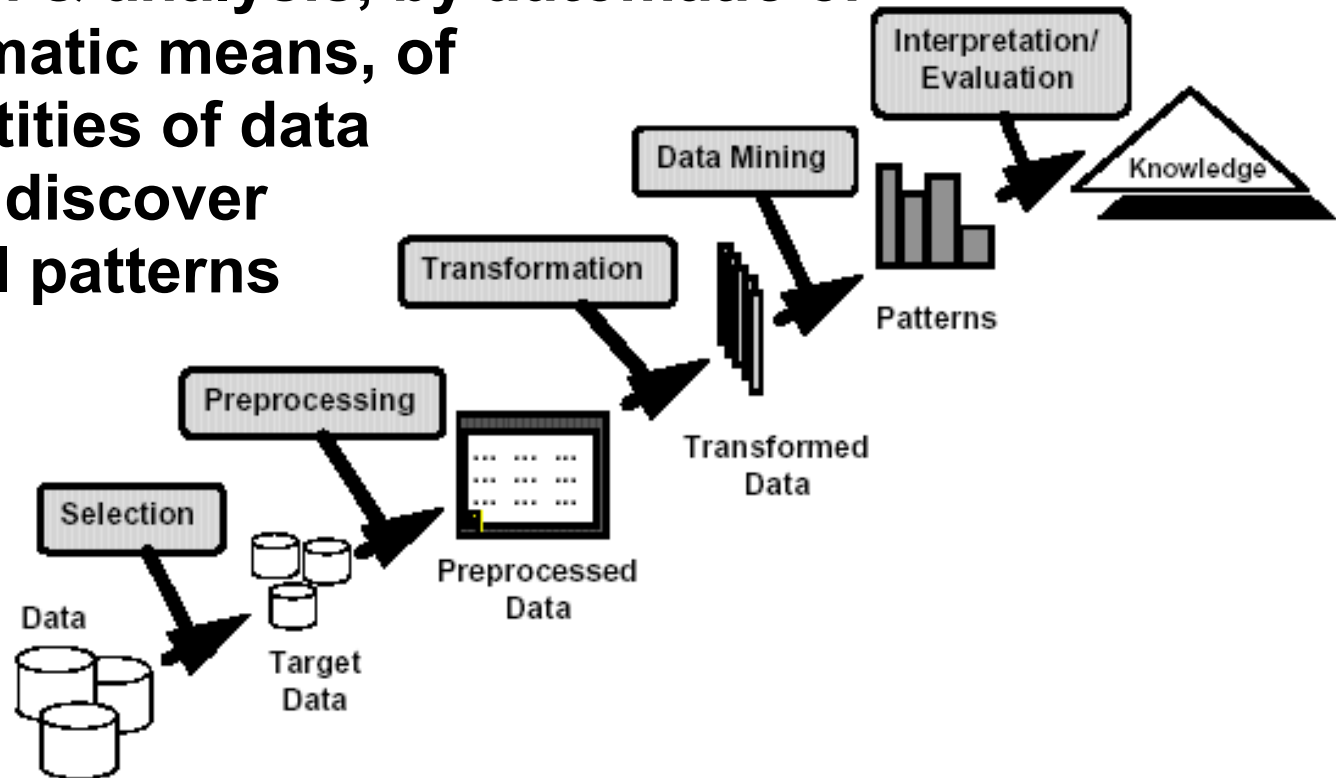
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



What is Data Mining?

● Many Definitions

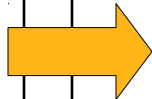
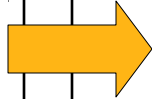
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



What is (not) Data Mining?

● What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

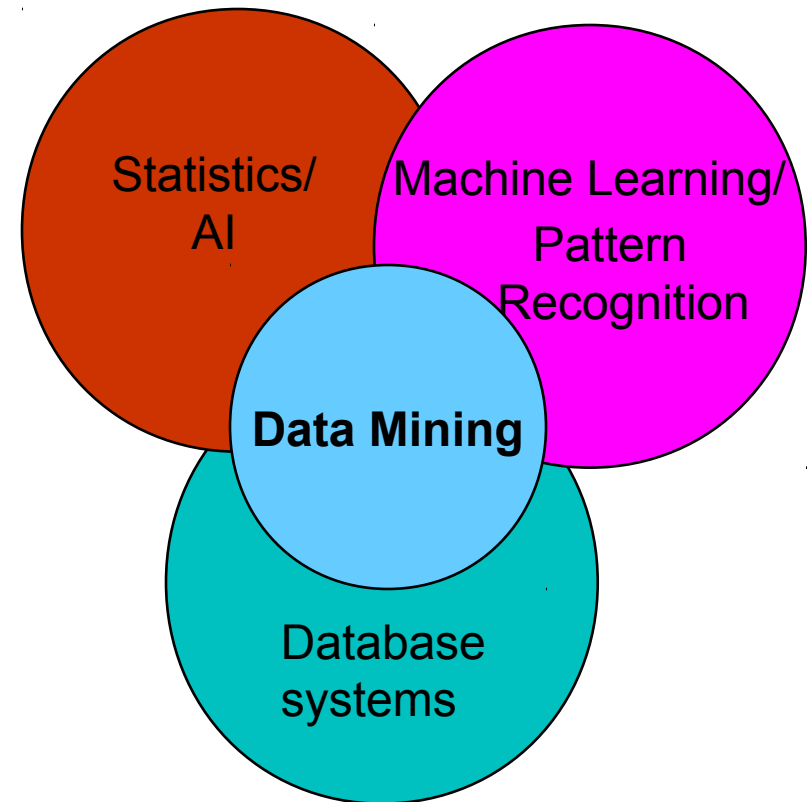


● What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



Data Mining Tasks

- Prediction Methods
 - Use some variables to predict unknown or future values of other variables.
- Description Methods
 - Find human-interpretable patterns that describe the data.

Data Mining Tasks...

- Predictive
 - Classification
 - Regression
 - Deviation Detection
- Descriptive
 - Association Rule Discovery
 - Sequential Pattern Discovery
 - Clustering

Classification: Definition

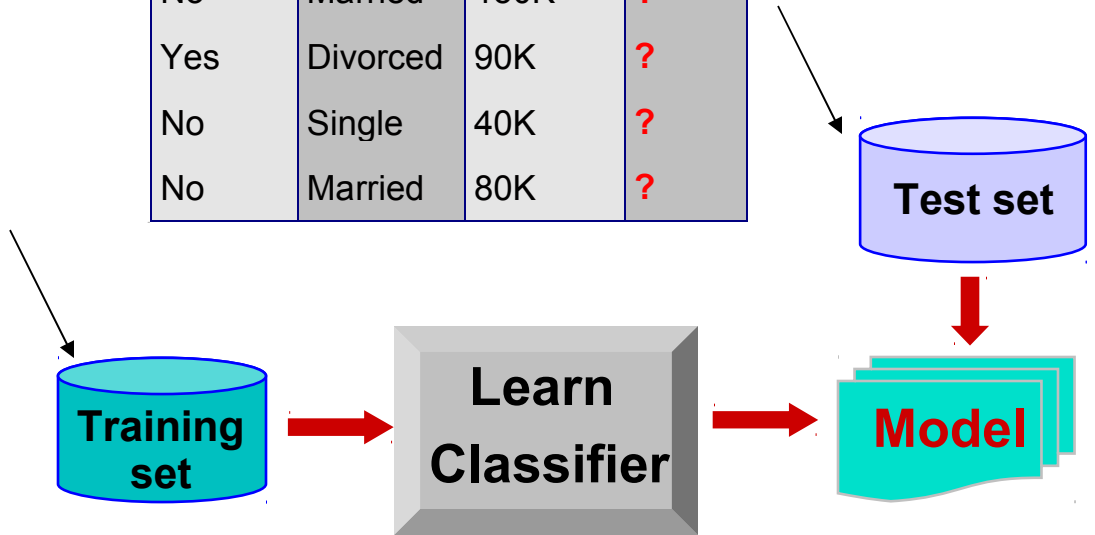
- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical
categorical
continuous
class

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- Direct Marketing
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - ◆ Use the data for a similar product introduced before.
 - ◆ We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
 - ◆ Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - ◆ Use this information as input attributes to learn a classifier model.

Classification: Application 2

- Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.

- Approach:

- ◆ Use credit card transactions and the information on its account-holder as attributes.

- When does a customer buy, what does he buy, how often he pays on time, etc

- ◆ Label past transactions as **fraud** or **fair** transactions. This forms the class attribute.

- ◆ Learn a model for the class of the transactions.

- ◆ Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - ◆ Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - ◆ Label the customers as **loyal** or **disloyal**.
 - ◆ Find a model for loyalty.

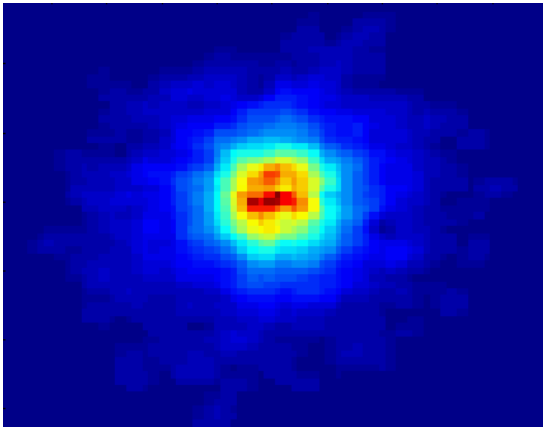
Classification: Application 4

- Sky Survey Cataloging
 - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
 - 3000 images with 23,040 x 23,040 pixels per image.
 - Approach:
 - ◆ Segment the image.
 - ◆ Measure image attributes (features) - 40 of them per object.
 - ◆ Model the class based on these features.
 - ◆ Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

Courtesy: <http://aps.umn.edu>

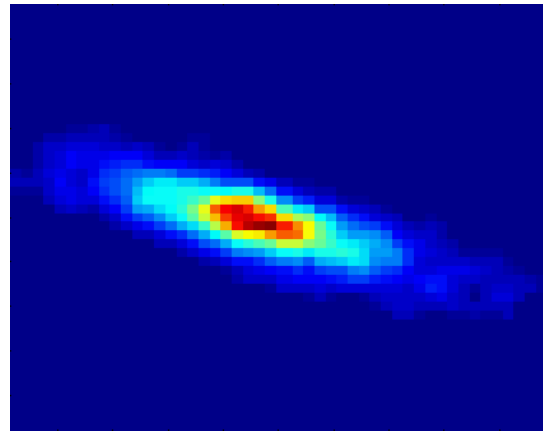
Early



Class:

- Stages of Formation

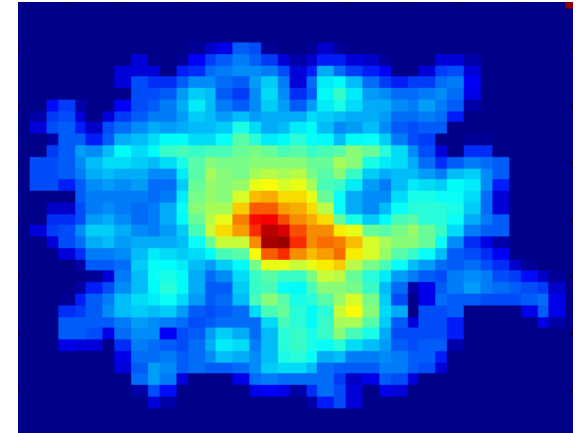
Intermediate



Attributes:

- Image features,
- Characteristics of light waves received, etc.

Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Clustering Definition

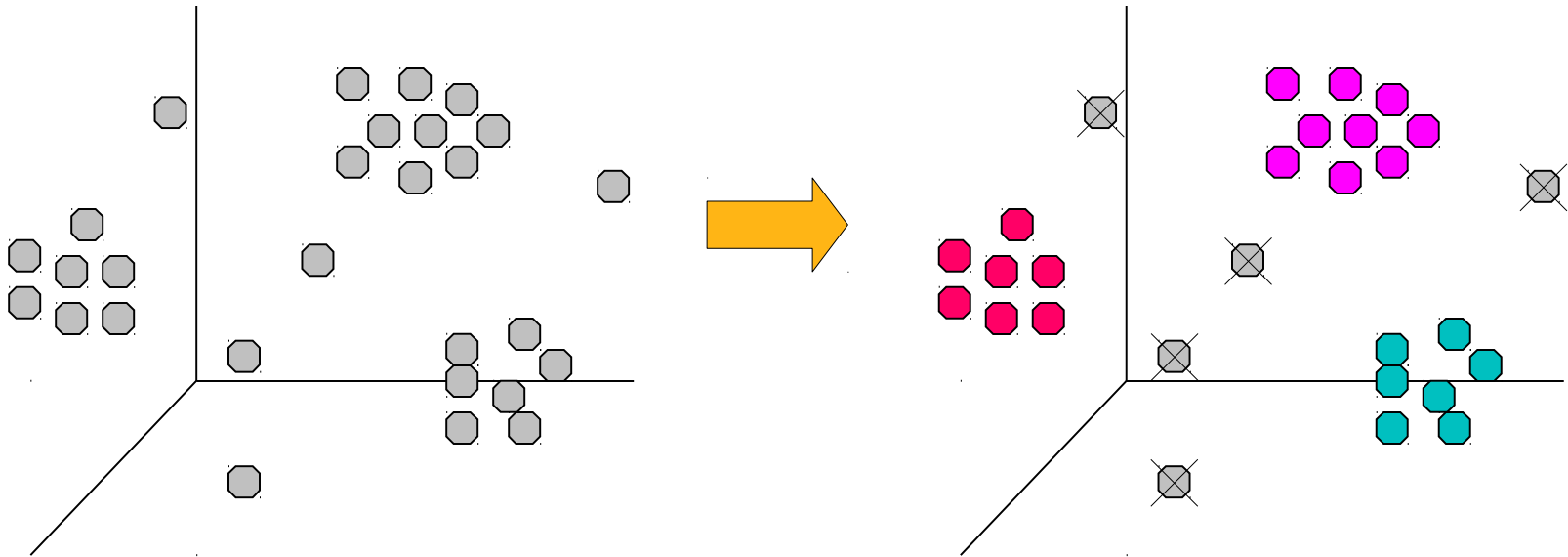
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
 - Data points in one cluster are more similar to one another.
 - Data points in separate clusters are less similar to one another.
- Similarity Measures:
 - Euclidean Distance if attributes are continuous.
 - Other Problem-specific Measures.

Illustrating Clustering

Euclidean Distance-based Clustering in 3-D space.

Intracluster distances
are minimized

Intercluster distances
are maximized



Clustering: Application 1

- Market Segmentation
 - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - ◆ Collect different attributes of customers based on their geographical and lifestyle related information.
 - ◆ Find clusters of similar customers.
 - ◆ Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering

- Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
- Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

<i>Category</i>	<i>Total Articles</i>	<i>Correctly Placed</i>
<i>Financial</i>	555	364
<i>Foreign</i>	341	260
<i>National</i>	273	36
<i>Metro</i>	943	746
<i>Sports</i>	738	573
<i>Entertainment</i>	354	278

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock- $\{UP/DOWN\}$
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOWN, Bay-Network-Down, 3-COM-DOWN, Cabletron-Sys-DOWN, CISCO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-Down, Tellabs-Inc-Down, Natl-Semiconduct-DOWN, Oracl-DOWN, SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andrew-Corp-DOWN, Computer-Assoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN, Fed-Home-Loan-DOWN, MBNA-Corp-DOWN, Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Unocal-UP, Schlumberger-UP	Oil-UP

Association Rule Discovery: Definition

- Given a set of records each of which contains some number of items from a given collection;
 - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
 - Let the rule discovered be
 $\{Bagels, \dots\} \rightarrow \{Potato\ Chips\}$
 - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
 - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
 - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
 - Goal: To identify items that are bought together by sufficiently many customers.
 - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
 - A classic rule --
 - ◆ If a customer buys diaper and milk, then he is very likely to buy beer.
 - ◆ So, don't be surprised if you find six-packs stacked next to diapers!

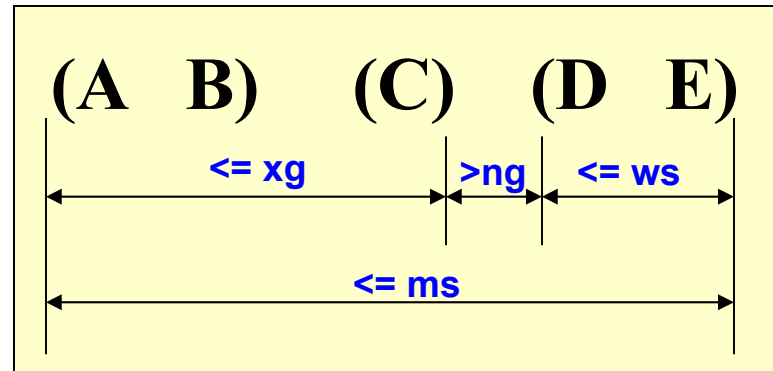
Association Rule Discovery: Application 3

- Inventory Management:
 - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
 - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery: Definition

- Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different event
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.

(A B) (C) → (D E)



Sequential Pattern Discovery: Examples

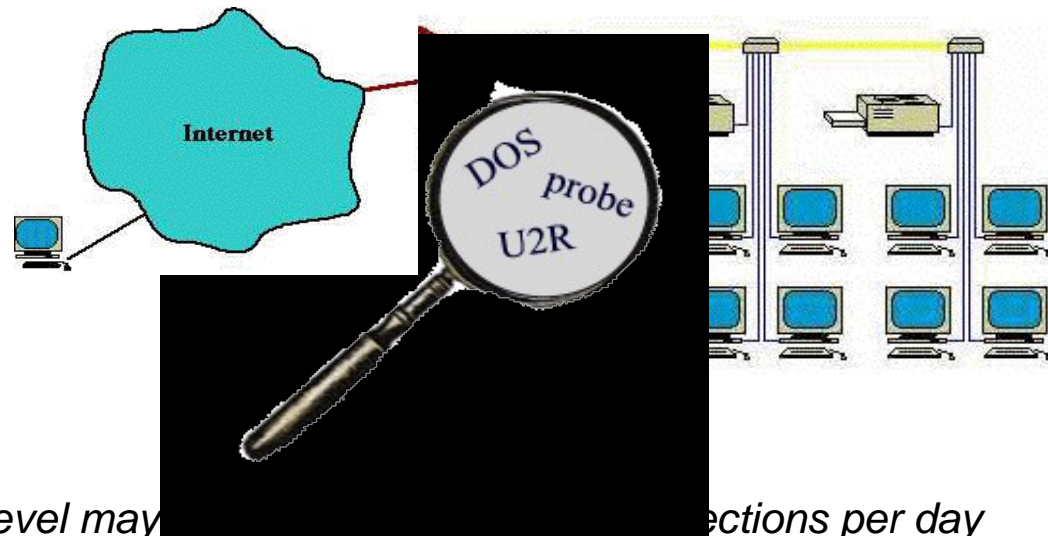
- In telecommunications alarm logs,
 - (Inverter_Problem Excessive_Line_Current)
(Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
 - Computer Bookstore:
(Intro_To_Visual_C) (C++_Primer) -->
(Perl_for_dummies,Tcl_Tk)
 - Athletic Apparel Store:
(Shoes) (Racket, Racketball) --> (Sports_Jacket)

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
 - Credit Card Fraud Detection
 - Network Intrusion Detection



Typical network traffic at University level may have 100,000 connections per day

Challenges of Data Mining

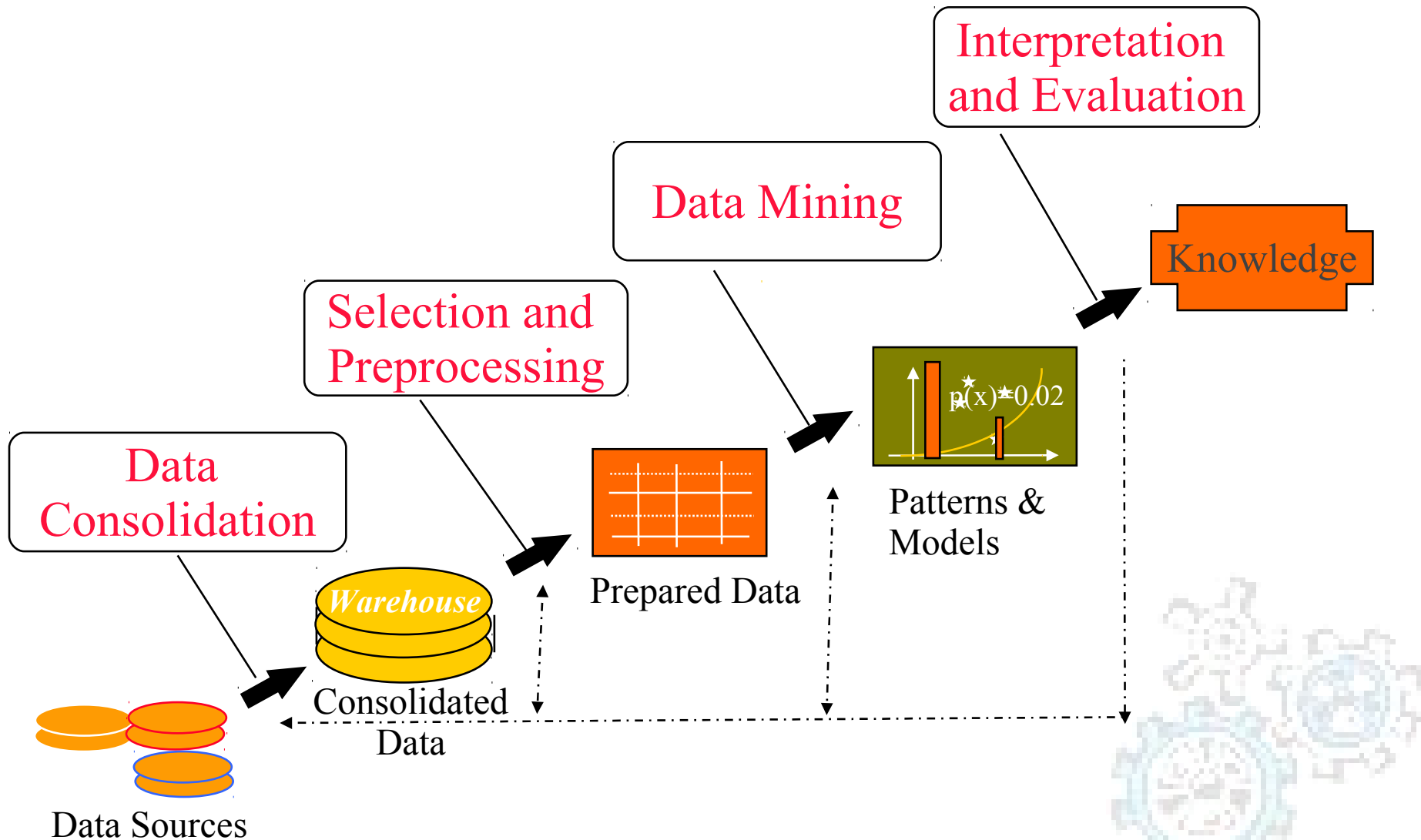
- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

The KDD process (Knowledge Discovery in Databases)

- The selection and processing of data for:
 - the identification of **novel**, accurate, and **useful** patterns, and
 - the modeling of real-world phenomena.
- **Data mining** is a major component of the KDD process - automated discovery of patterns and the development of predictive and explanatory models.

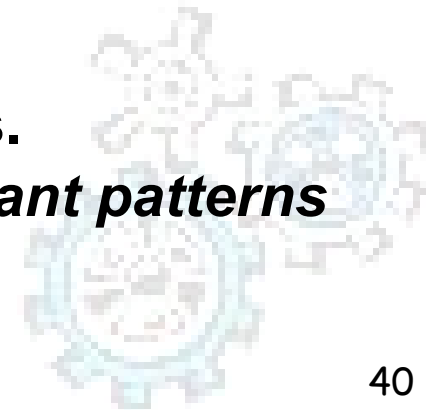


The KDD process



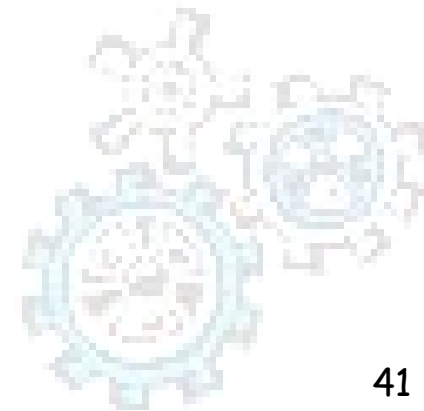
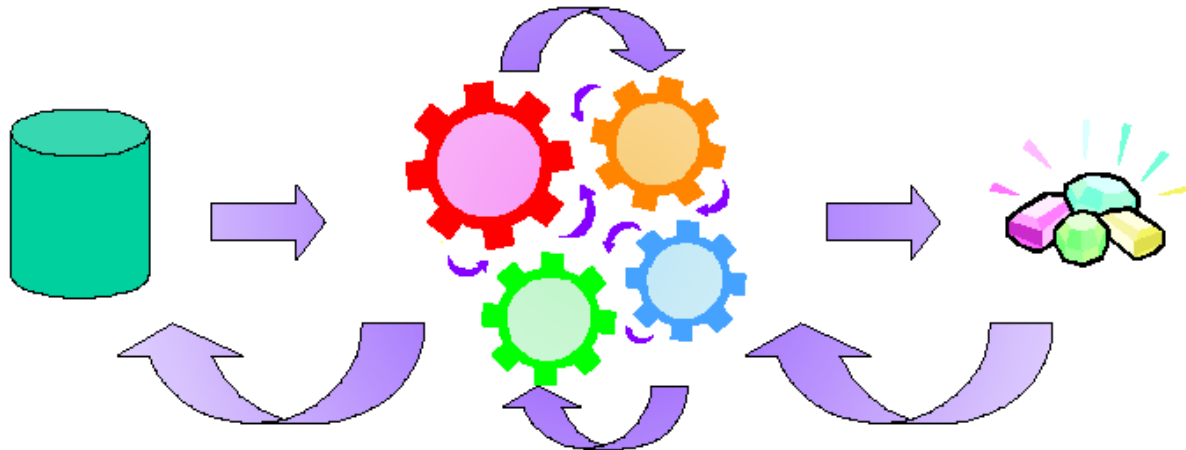
The steps of the KDD process

- Learning the application domain:
 - relevant prior knowledge and goals of application
- **Data consolidation:** Creating a target data set
- **Selection and Preprocessing**
 - *Data cleaning* : (may take 60% of effort!)
 - *Data reduction and projection*:
 - find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
 - classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- **Data mining:** search for patterns of interest
- **Interpretation and evaluation:** analysis of results.
 - *visualization, transformation, remove redundant patterns*
- Use of discovered knowledge



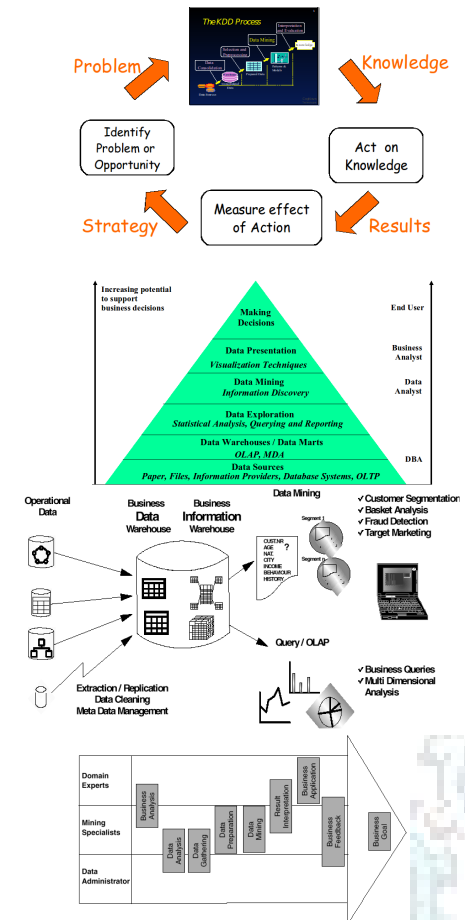
The KDD Process in Practice

- **KDD is an Iterative Process**
 - art + engineering rather than science



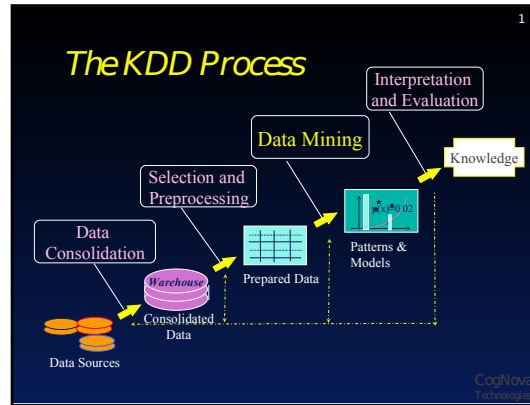
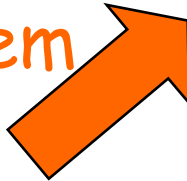
The KDD Process in real applications

- KDD as part of a virtuous cycle
- KDD/DM in the Business intelligence process
- Roles in the KDD process
- An environment for KDD and BI



The *virtuous cycle*

Problem



Knowledge



Identify
Problem or
Opportunity

Act on
Knowledge

Strategy



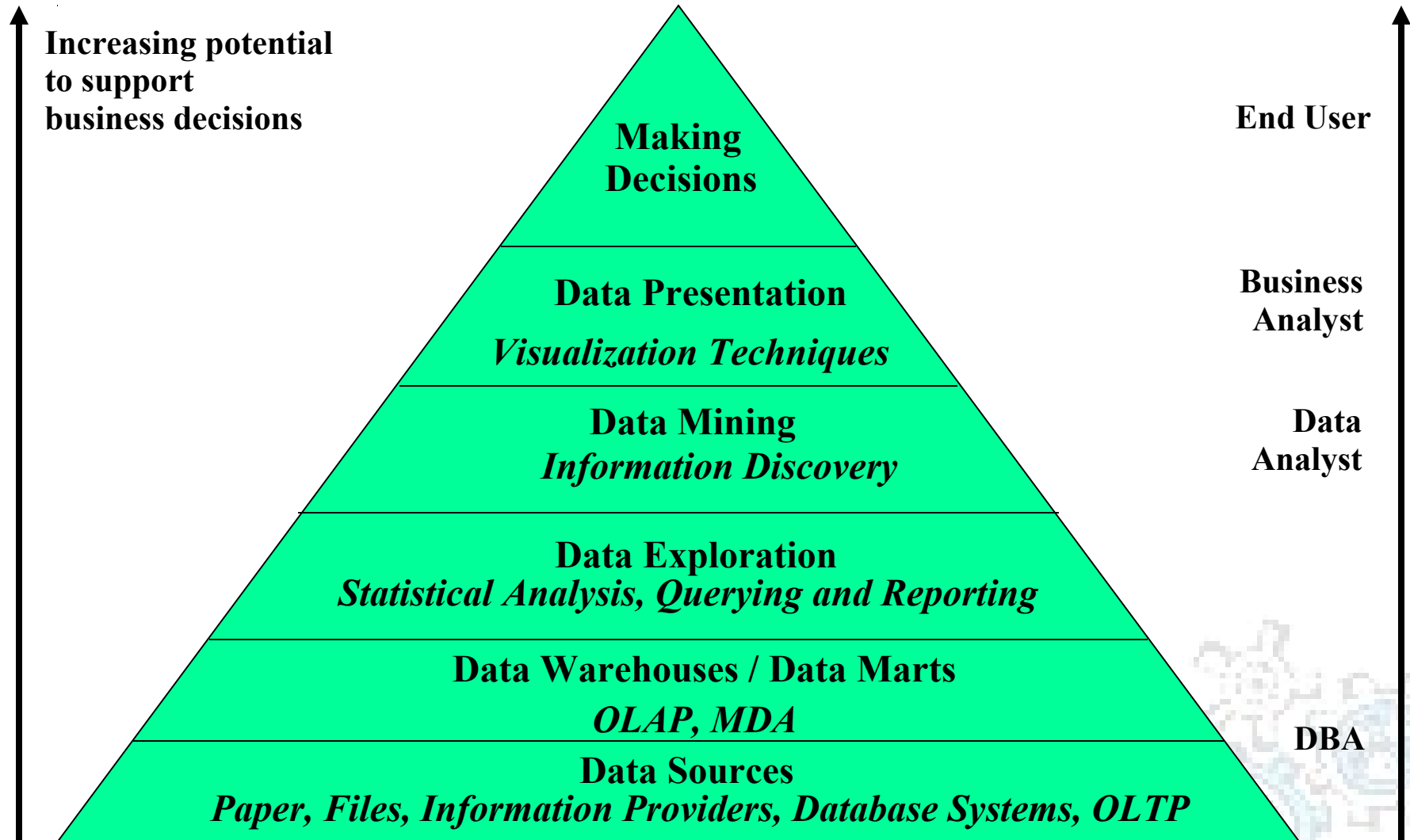
Measure effect
of Action



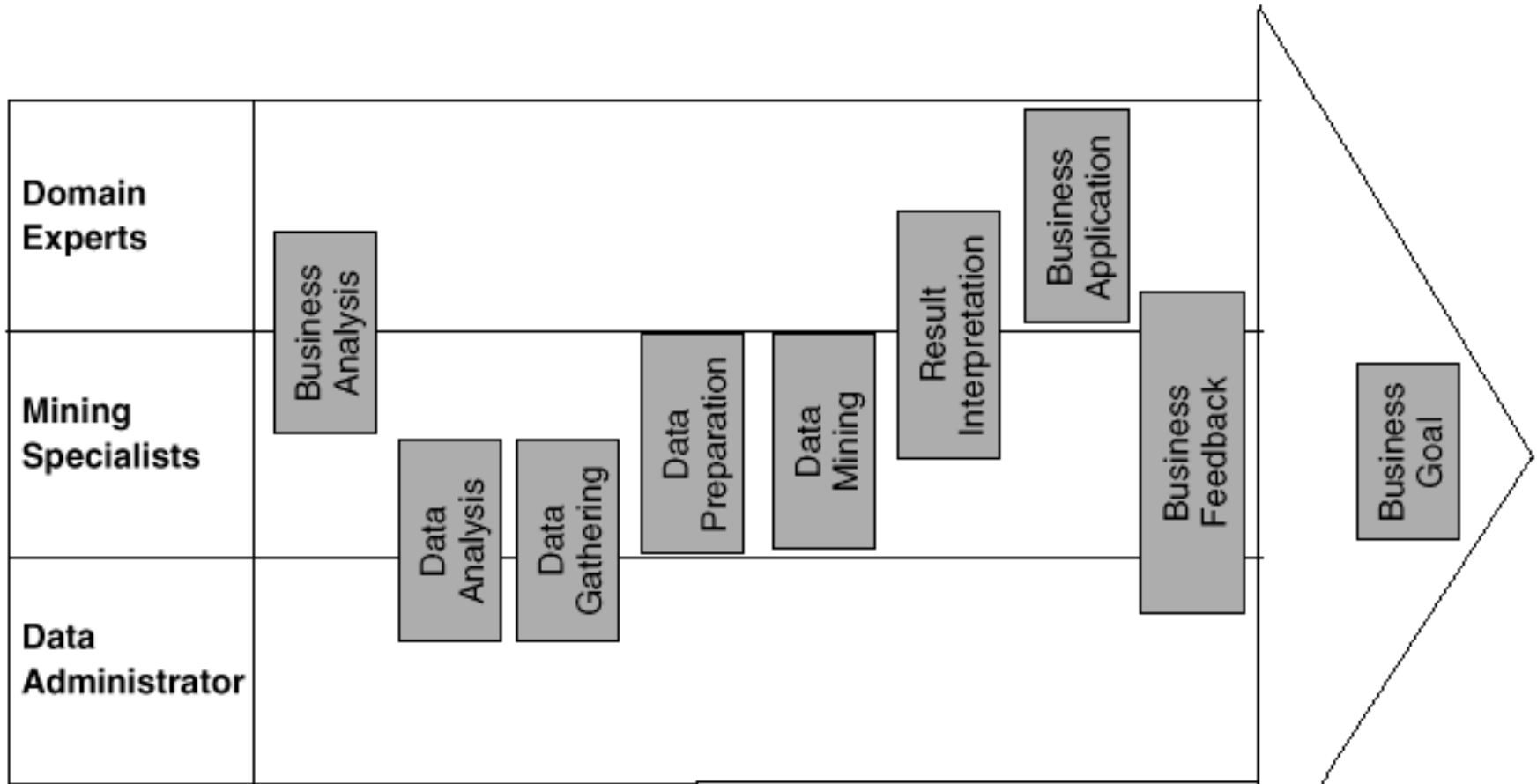
Results



Data mining and business intelligence



Roles in the KDD process



A business intelligence environment

