

Data Mining & Machine Learning

- *Dino Pedreschi & Anna Monreale*
- *Dipartimento di Informatica*
- *Tutor: Riccardo Guidotti, Dipartimento di Informatica*

DIPARTIMENTO DI INFORMATICA - Università di Pisa

Data Mining

■ Riferimenti bibliografici

- Berthold et. al. Guide to Intelligent Data Analysis
- Pyle, D. Business Modeling and Data Mining. Morgan Kaufmann, (2003)
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, [Introduction to DATA MINING](#), Addison Wesley, ISBN 0-321-32136-7, 2006
- Jiawei Han, Micheline Kamber, [Data Mining: Concepts and Techniques](#), Morgan Kaufmann Publishers, 2000 http://www.mkp.com/books_catalog/catalog.asp?ISBN=1-55860-489-8
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (editors). Advances in Knowledge discovery and data mining, MIT Press, 1996.
- Provost, F., Fawcett, T. Data Science for Business (2012)
- Barry Linoff Data Mining Techniques for Marketing Sales and Customer Support, John Wiles & Sons, 2002



Contenuti del corso in dettaglio

- ❑ Introduction & Basic notions
- ❑ The process for the knowledge extraction
- ❑ Introduction to the DM technologies
 - **Classification:**
 - **Clustering**
 - **Pattern Mining**
- ❑ *Overview of BigData Analytics*
 - **Social Network analysis**
 - **Mobility Data Analysis**
 - **Social Media Analysis & Privacy**



Evolution of Database Technology: from data management to data analysis

- 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:
 - Data mining and data warehousing, multimedia databases, and Web technology.



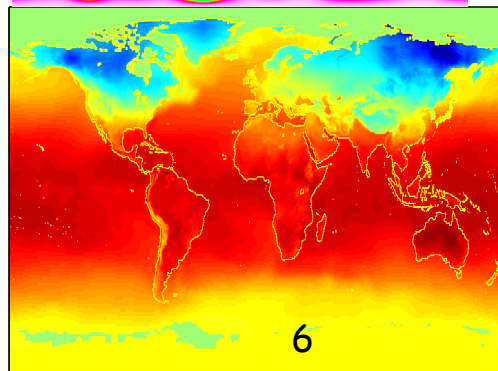
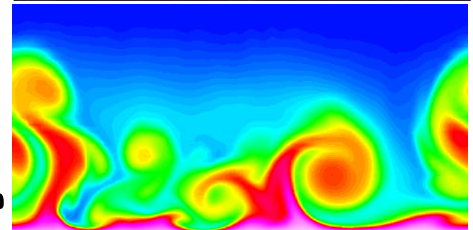
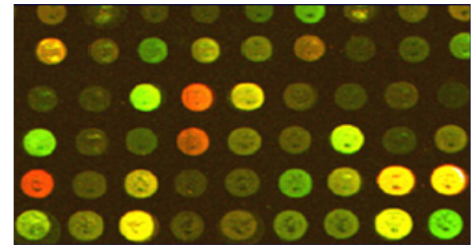
Why Mine Data? Commercial Viewpoint

- 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 data
- Data mining may help scientists
 - in classifying and segmenting data
 - in Hypothesis Formation



Why Data Mining

■ Increased Availability of Huge Amounts of Data

- | point-of-sale customer data
- | digitization of text, images, video, voice, etc.
- | World Wide Web and Online collections

■ Data Too Large or Complex for Classical or Manual Analysis

- | number of records in millions or billions
- | high dimensional data (too many fields/features/attributes)
- | often too sparse for rudimentary observations
- | high rate of growth (e.g., through logging or automatic data collection)
- | heterogeneous data sources

■ Business Necessity

- | e-commerce
- | high degree of competition
- | personalization, customer loyalty, market segmentation



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,)

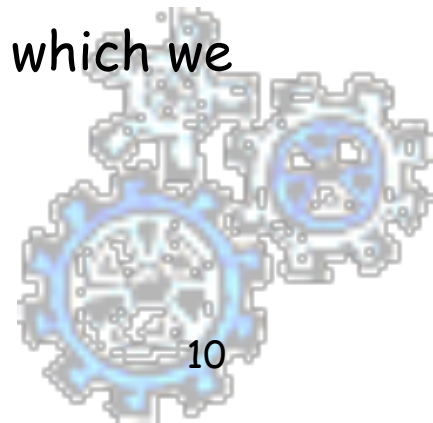
Sources of Data

■ Business Transactions

- widespread use of bar codes => storage of millions of transactions daily (e.g., Walmart: 2000 stores => 20M transactions per day)
- most important problem: effective use of the data in a reasonable time frame for competitive decision-making
- e-commerce data

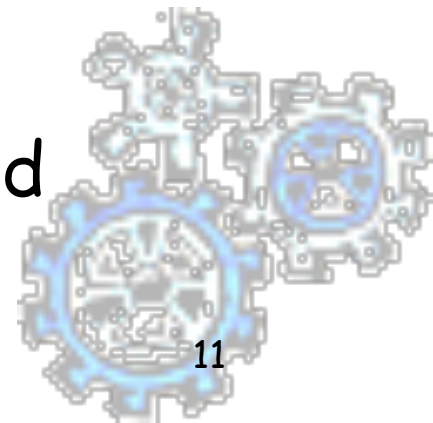
■ Scientific Data

- data generated through multitude of experiments and observations
- examples, geological data, satellite imaging data, NASA earth observations
- rate of data collection far exceeds the speed by which we analyze the data



Sources of Data

- Financial Data
 - company information
 - economic data (GNP, price indexes, etc.)
 - stock markets
- Personal / Statistical Data
 - government census
 - medical histories
 - customer profiles
 - demographic data
 - data and statistics about sports and athletes



Sources of Data

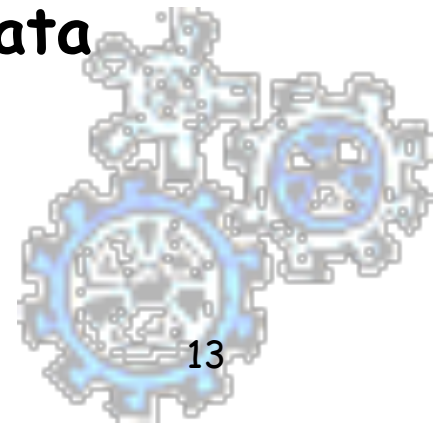
■ World Wide Web and Online Repositories

- email, news, messages
- Web documents, images, video, etc.
- link structure of of the hypertext from millions of Web sites
- Web usage data (from server logs, network traffic, and user registrations)
- online databases, and digital libraries



Classes of applications

- **Database analysis and decision support**
 - **Market analysis**
 - target marketing, customer relation management, market basket analysis, cross selling, market segmentation.
 - **Risk analysis**
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis.
 - **Fraud detection**
- **New Applications from New sources of data**
 - **Text (news group, email, documents)**
 - **Web analysis and intelligent search**



Market Analysis

- **Where are the data sources for analysis?**
 - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies.
- **Target marketing**
 - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- **Determine customer purchasing patterns over time**
 - Conversion of single to a joint bank account: marriage, etc.
- **Cross-market analysis**
 - Associations/co-relations between product sales
 - Prediction based on the association information.



Market Analysis (2)

■ Customer profiling

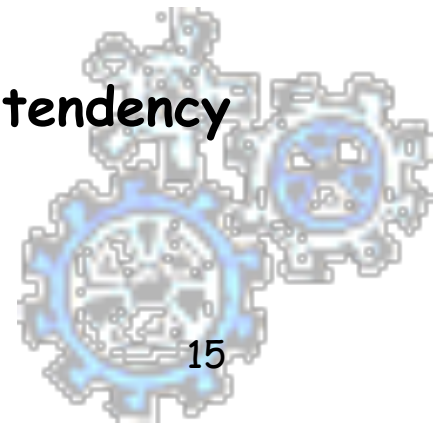
- data mining can tell you what types of customers buy what products (clustering or classification).

■ Identifying customer requirements

- identifying the best products for different customers
- use prediction to find what factors will attract new customers

■ Summary information

- various multidimensional summary reports;
- statistical summary information (data central tendency and variation)



Risk Analysis

- Finance planning and asset evaluation:
 - cash flow analysis and prediction
 - contingent claim analysis to evaluate assets
 - trend analysis
- Resource planning:
 - summarize and compare the resources and spending
- Competition:
 - monitor competitors and market directions (CI: *competitive intelligence*).
 - group customers into classes and class-based pricing procedures
 - set pricing strategy in a highly competitive market



Fraud Detection

■ Applications:

- widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.

■ Approach:

- use historical data to build models of fraudulent behavior and use data mining to help identify similar instances.

■ Examples:

- auto insurance: detect a group of people who stage accidents to collect on insurance
- money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
- medical insurance: detect professional patients and ring of doctors and ring of references

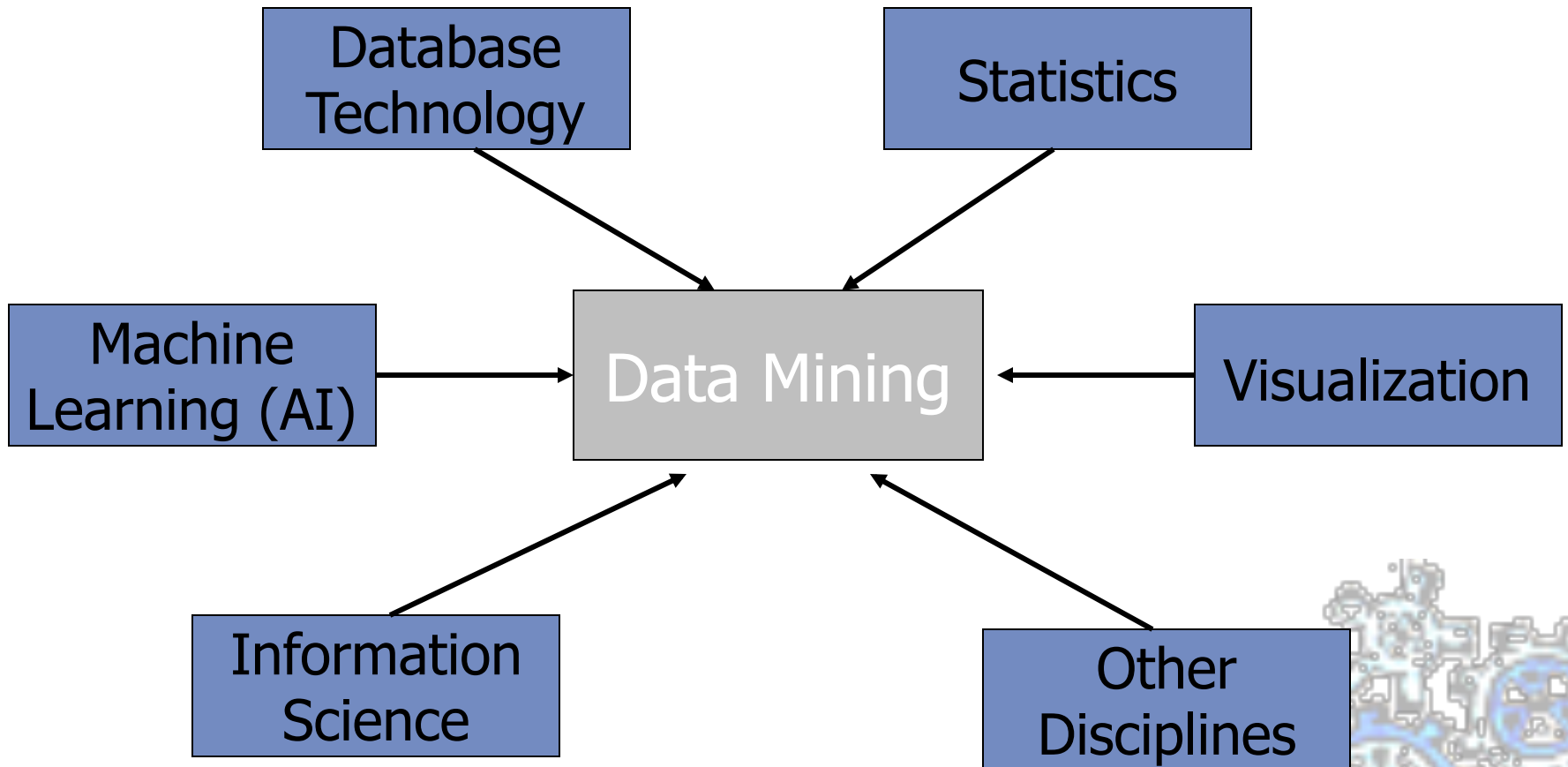


What is Knowledge Discovery in Databases (KDD)? A process!

- 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.

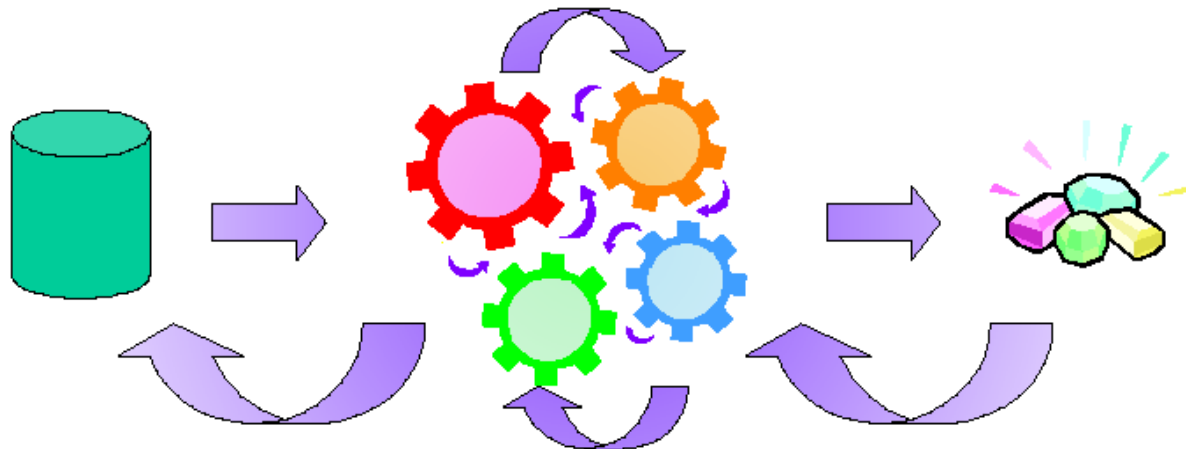


Data Mining: Confluence of Multiple Disciplines

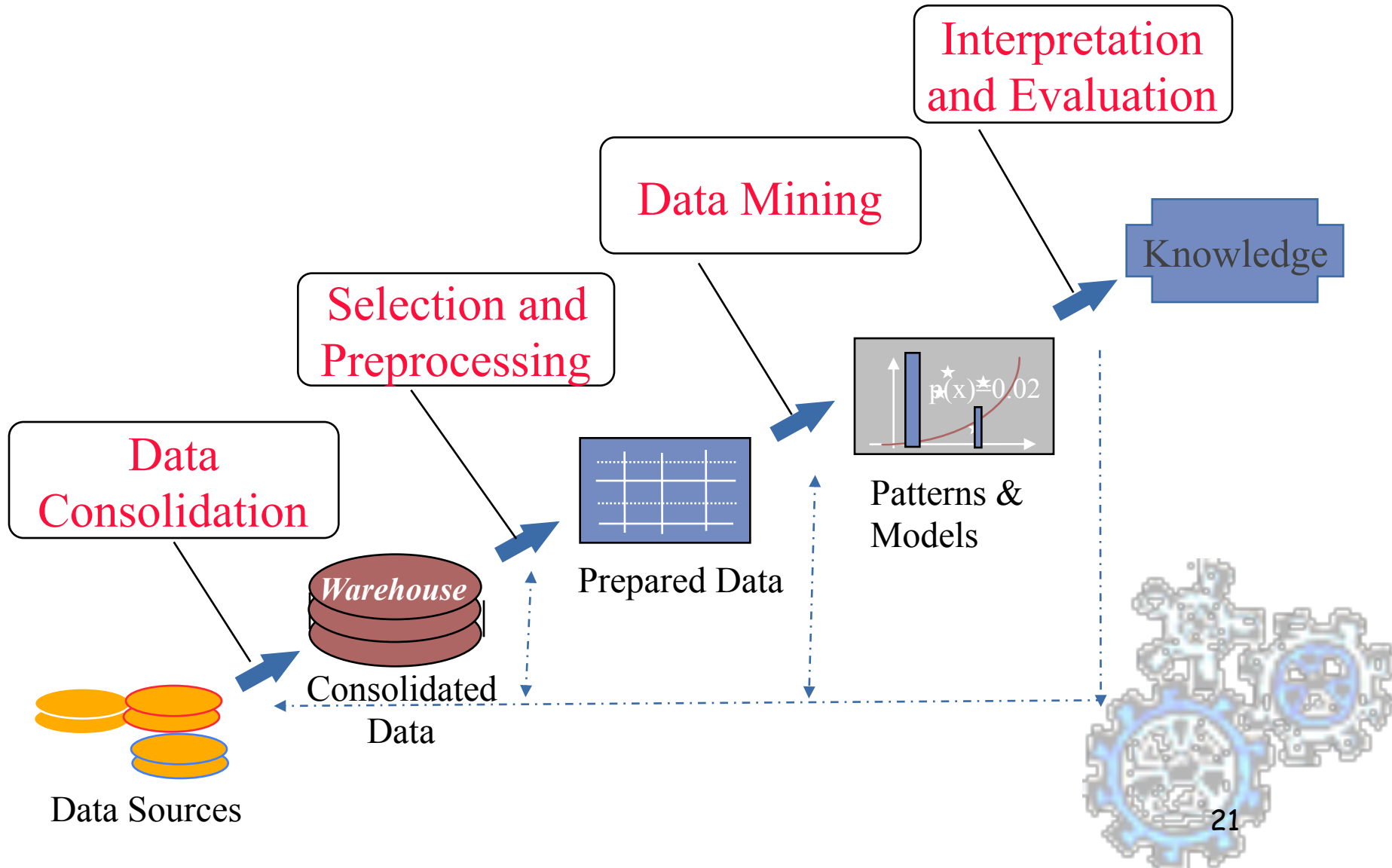


The KDD Process in Practice

- KDD is an Iterative Process
 - art + engineering ...and science



The KDD process

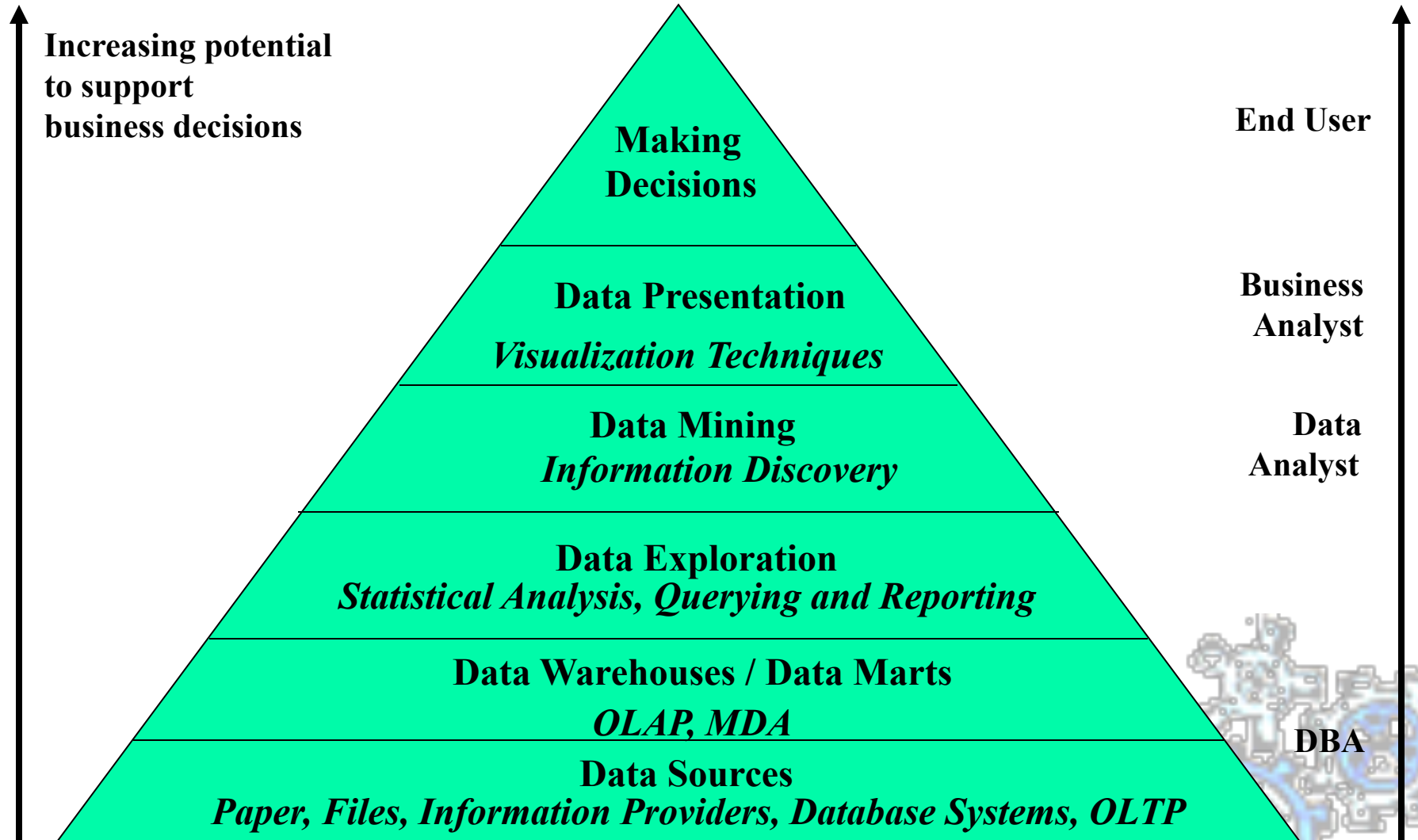


The steps of the KDD process

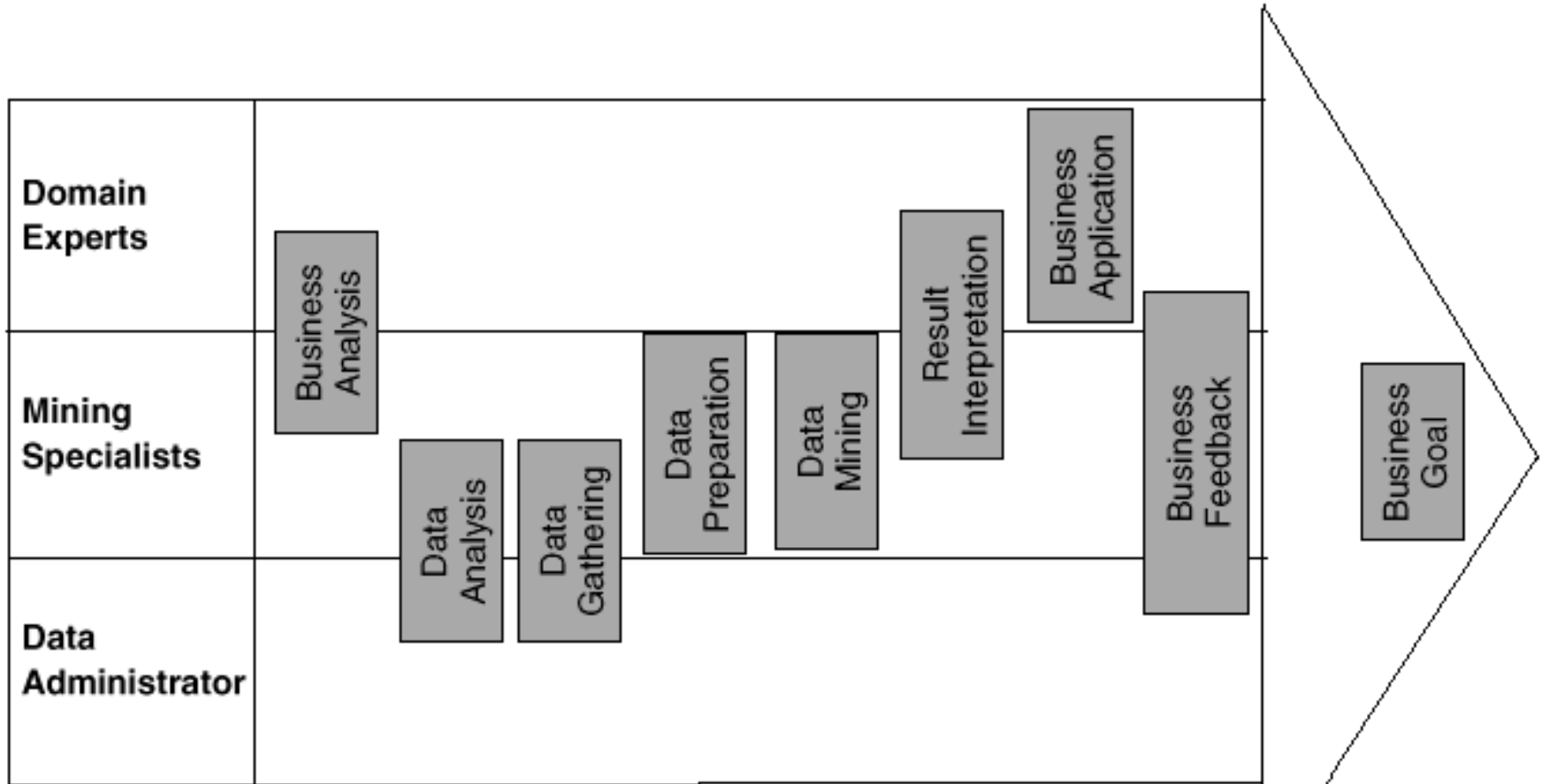
- Learning the application domain:
 - relevant prior knowledge and goals of application
- **Data consolidation:** Creating a target data set
- **Selection and Preprocessing**
 - *Data cleaning* : (may take 60% of effort!)
 - *Data reduction and projection:*
 - find useful features, dimensionality/variable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



Data mining and business intelligence



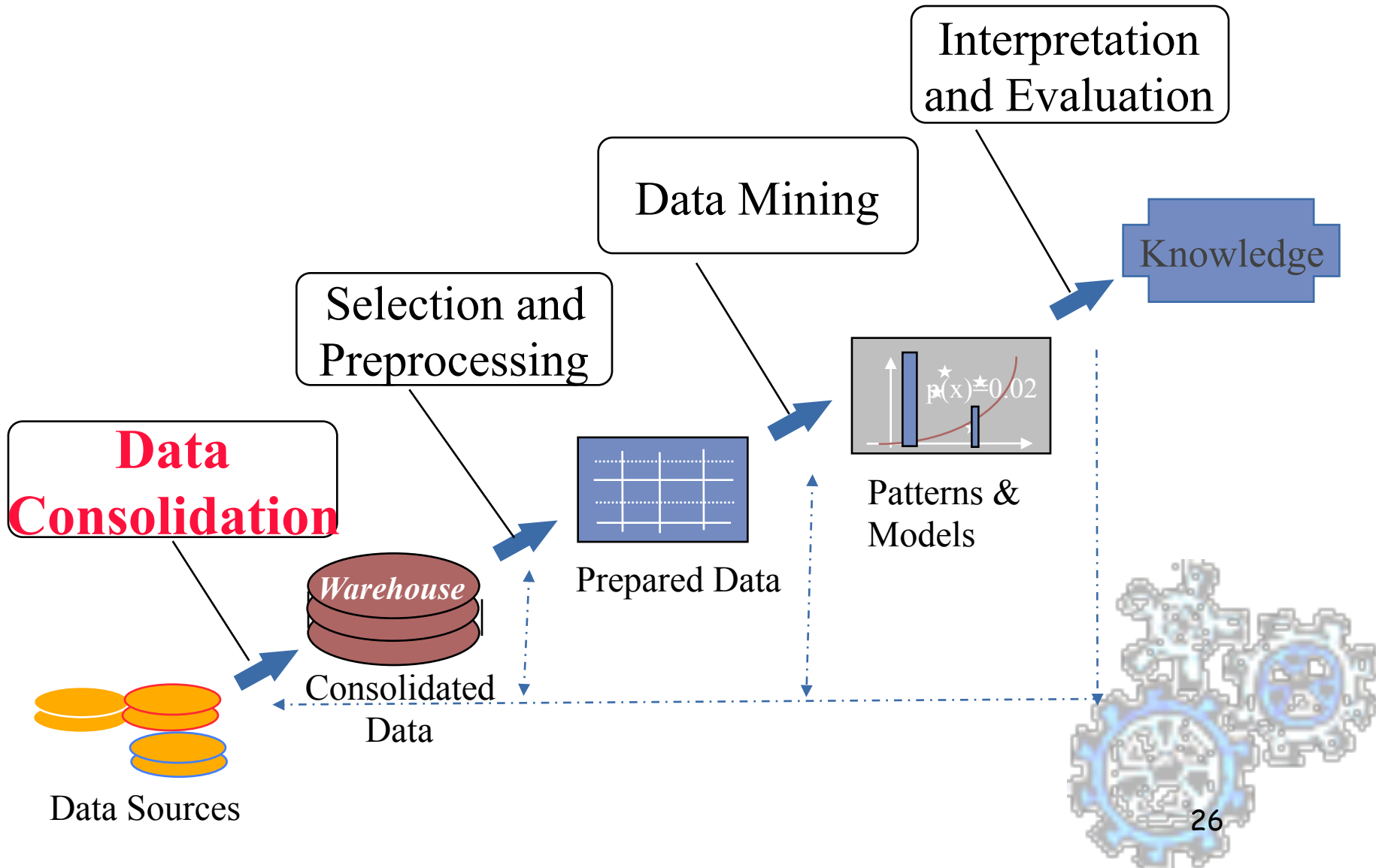
Roles in the KDD process



THE KDD PROCESS



The KDD process



Data consolidation and preparation

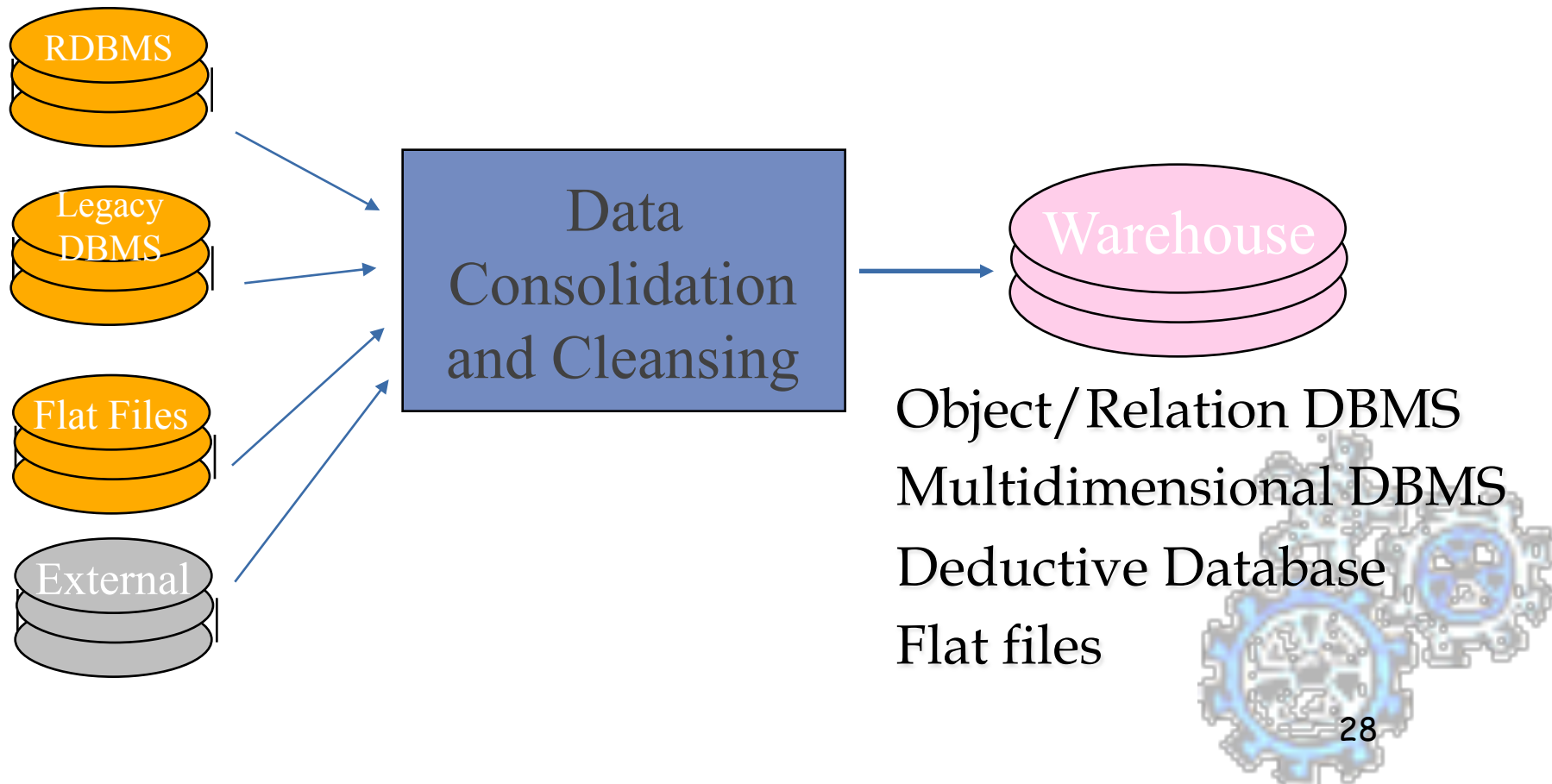
Garbage in → Garbage out

- The quality of results relates directly to quality of the data
- 50%-70% of KDD process effort is spent on data consolidation and preparation
- Major justification for a corporate data warehouse



Data consolidation

From data sources to consolidated data repository

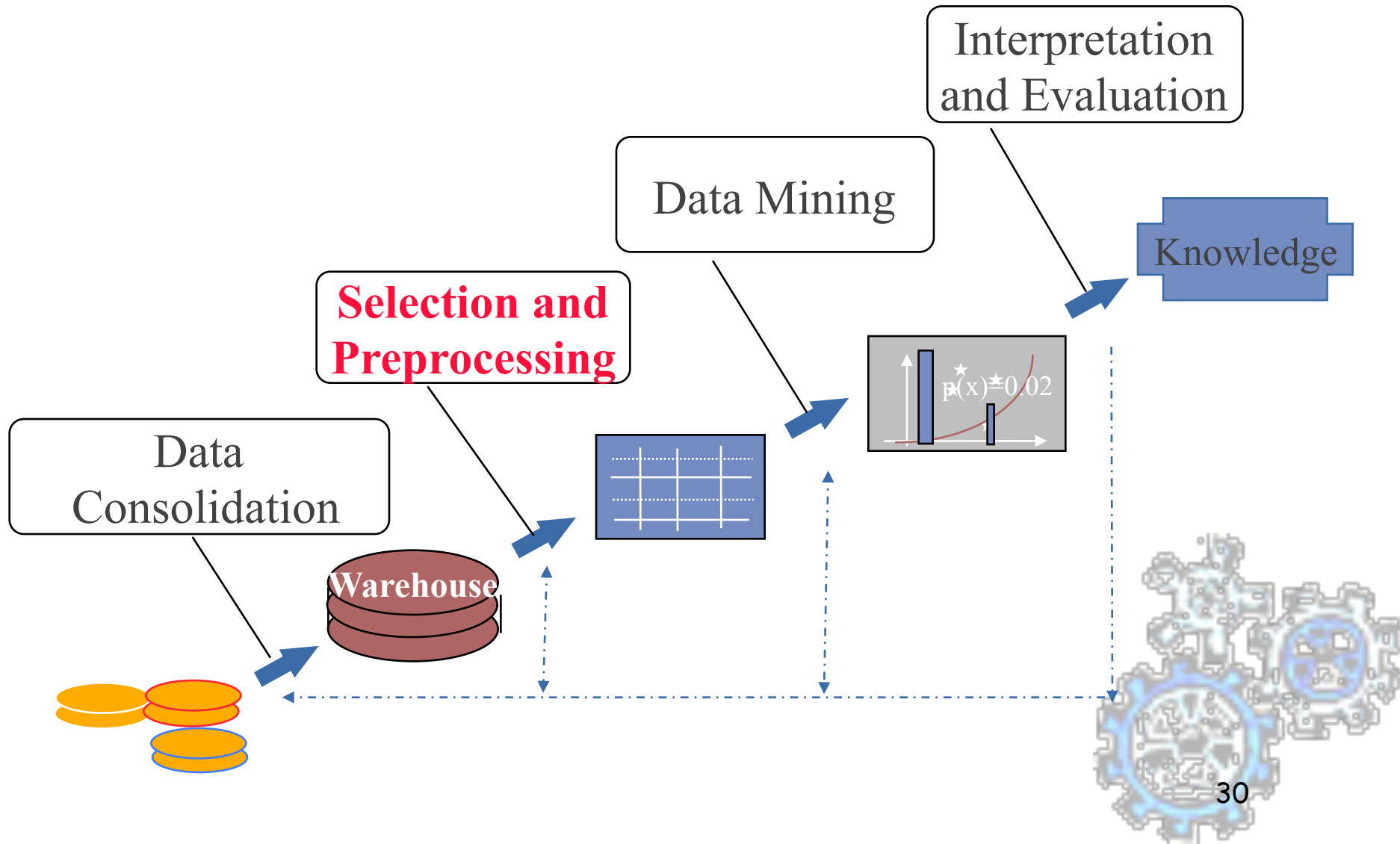


Data consolidation

- Determine preliminary list of attributes
- Consolidate data into working database
 - Internal and External sources
- Eliminate or estimate missing values
- Remove *outliers* (obvious exceptions)
- Determine prior probabilities of categories and deal with *volume bias*



The KDD process

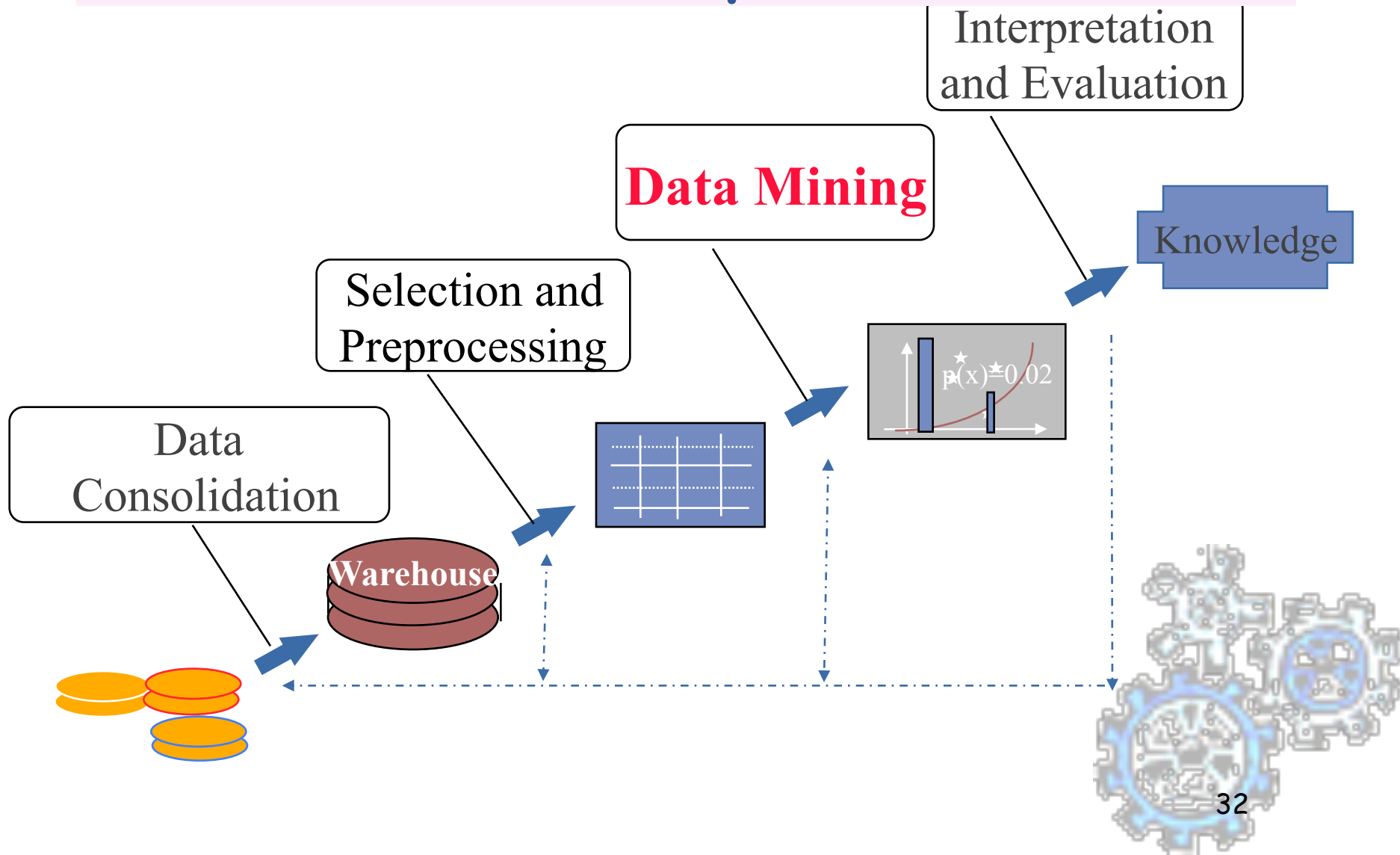


Data selection and preprocessing

- **Generate a set of examples**
 - ▮ choose sampling method
 - ▮ consider sample complexity
 - ▮ deal with volume bias issues
- **Reduce attribute dimensionality**
 - ▮ remove redundant and/or correlating attributes
 - ▮ combine attributes (sum, multiply, difference)
- **Reduce attribute value ranges**
 - ▮ group symbolic discrete values
 - ▮ quantify continuous numeric values
- **Transform data**
 - ▮ de-correlate and normalize values
 - ▮ map time-series data to static representation
- **OLAP and visualization tools play key role**



The KDD process



Data mining tasks and methods

- **Supervised (Directed) Knowledge Discovery**
 - **Purpose:** Explain value of some field in terms of all the others (goal-oriented)
 - **Method:** select the target field based on some hypothesis about the data; ask the algorithm to tell us how to predict or classify new instances
 - **Examples:**
 - what products show increased sale when cream cheese is discounted
 - which banner ad to use on a web page for a given user coming to the site



Data mining tasks and methods

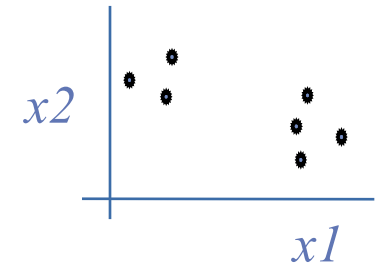
- **Unsupervised (Undirected) Knowledge Discovery (Explorative Methods)**
 - **Purpose:** Find patterns in the data that may be interesting (no target specified)
 - **Task:** clustering, association rules (affinity grouping)
 - **Examples:**
 - | which products in the catalog often sell together
 - | market segmentation (groups of customers/users with similar characteristics)



Data Mining Tasks

■ Automated Exploration/Discovery

- | *e.g.*.. **discovering new market segments**
- | clustering analysis

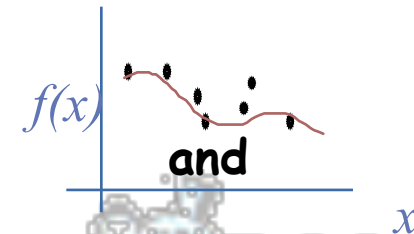


■ Prediction/Classification

- | *e.g.*.. **forecasting gross sales given current factors**
- | regression, neural networks, genetic algorithms, decision trees

■ Explanation/Description

- | *e.g.*.. **characterizing customers by demographics**
- | purchase history
- | decision trees, association rules



if age > 35
and income < \$35k
then ...

Data Mining Tasks

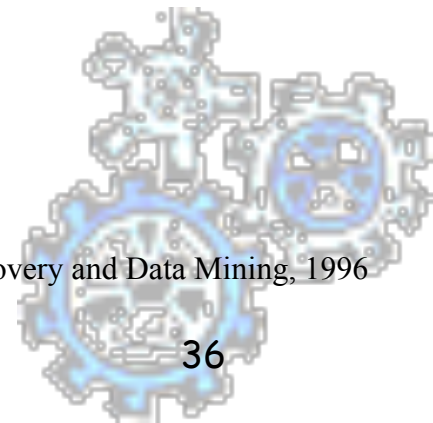
■ Prediction Methods

- Use some variables to predict unknown or future values of other variables.

■ Description Methods

- Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



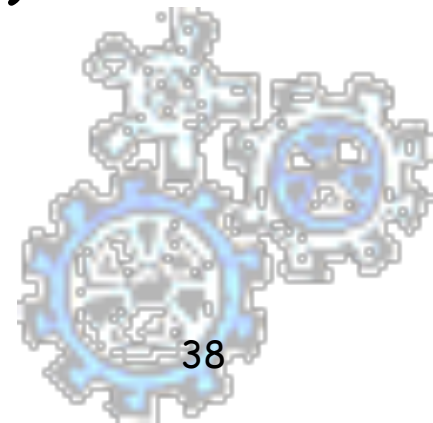
Data Mining Tasks...

- **Classification** [Predictive]
- **Clustering** [Descriptive]
- **Association Rule Discovery** [Descriptive]
- **Sequential Pattern Discovery** [Descriptive]
- **Regression** [Predictive]



Prediction and classification

- **Learning** a predictive model
- **Classification** of a new case/sample
- **Many methods:**
 - Artificial neural networks
 - Inductive decision tree and rule systems
 - Genetic algorithms
 - Nearest neighbor clustering algorithms
 - Statistical (parametric, and non-parametric)



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.

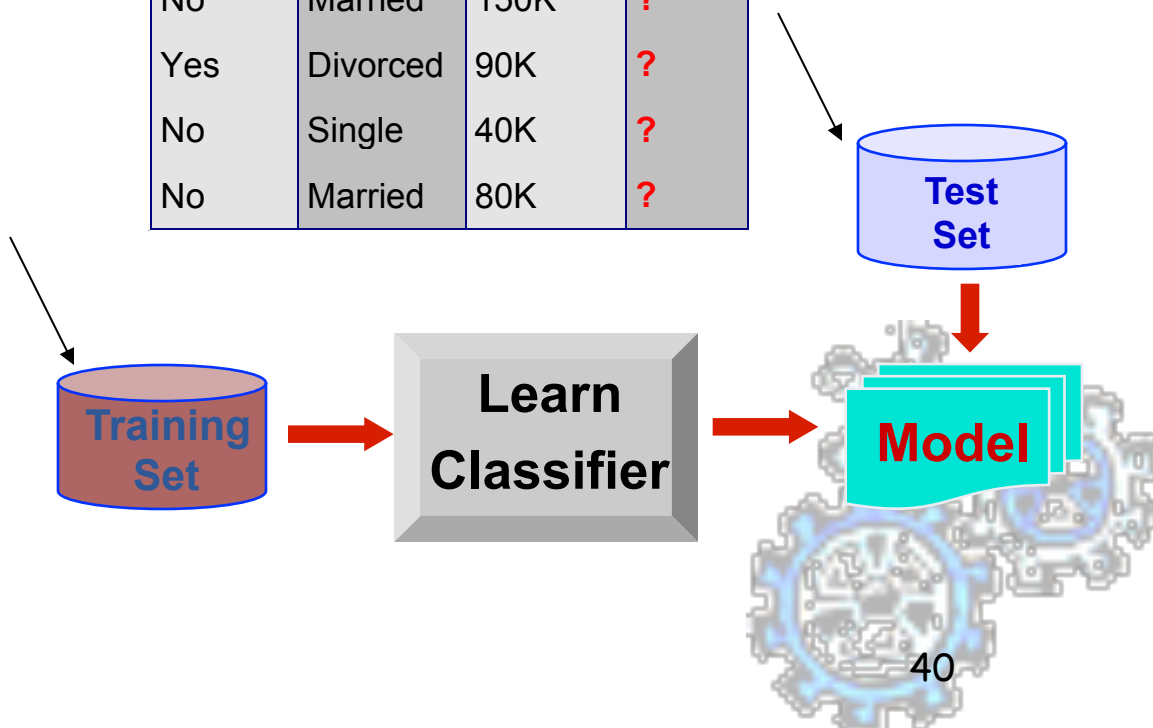


Classification Example

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.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

■ Fraud Detection

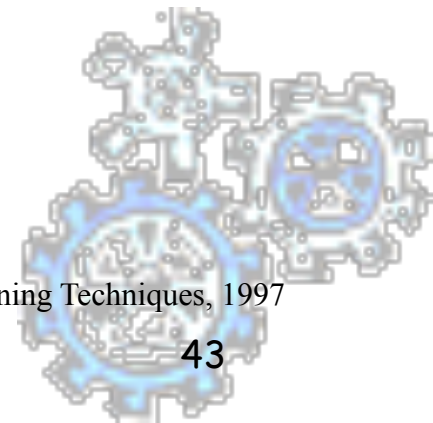
- | **Goal:** Predict fraudulent cases in credit card transactions.
- | **Approach:**
 - | Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - | Label past transactions as fraud or fair 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

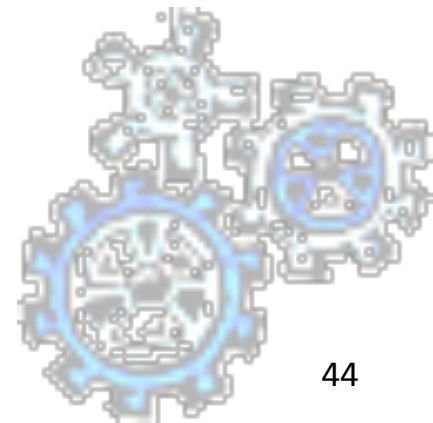
■ Customer Attrition/Churn:

- 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.



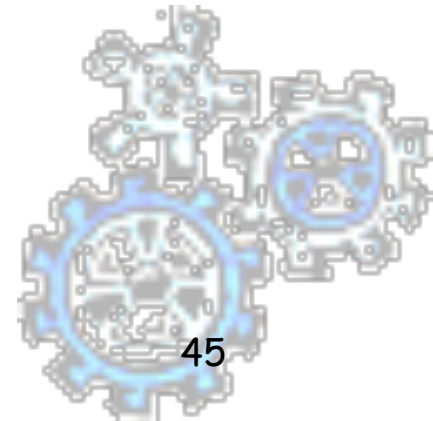
Regression

- 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.



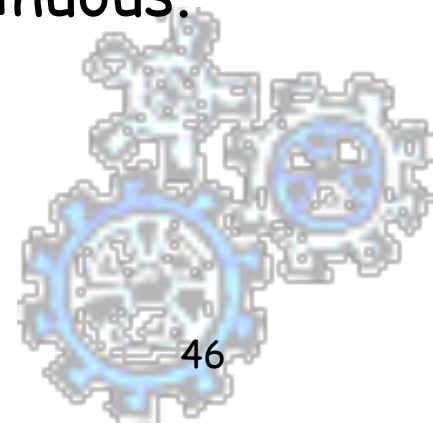
Automated exploration and discovery

- **Clustering:** partitioning a set of data into a set of classes, called *clusters*, whose members share some interesting common properties.
- **Distance-based numerical clustering**
 - ▮ metric grouping of examples (K-NN)
 - ▮ graphical visualization can be used
- **Bayesian clustering**
 - ▮ search for the number of classes which result in best fit of a probability distribution to the data
 - ▮ AutoClass (NASA) one of best examples



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.

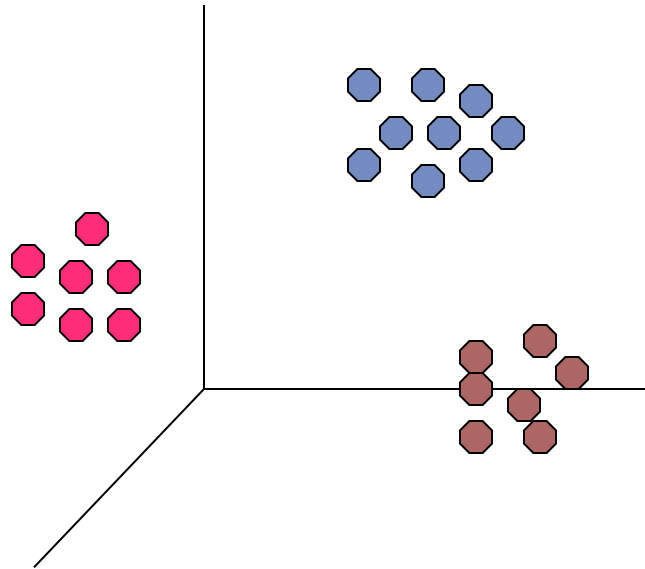


Illustrating Clustering

- I Euclidean Distance Based Clustering in 3-D space.

Intracluster distances
are minimized

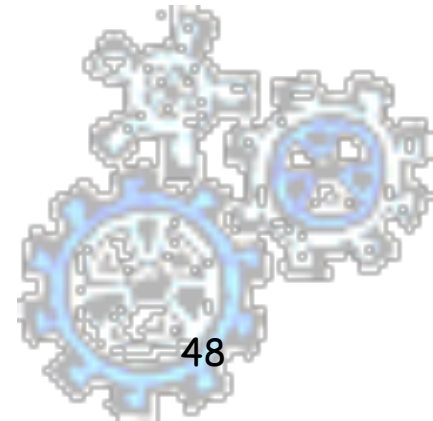
Intercluster distances
are maximized



Clustering: Application 1

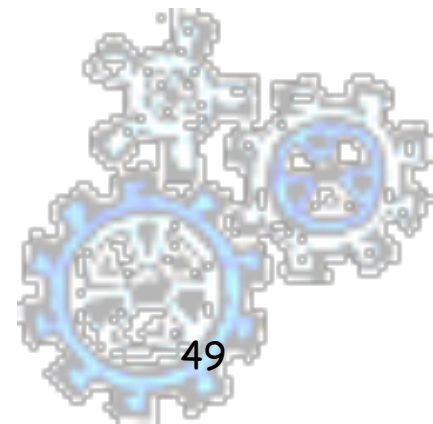
■ Market Segmentation:

- 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.



Pattern Mining

- **Determine what items often go together (usually in transactional databases)**
- **Often Referred to as Market Basket Analysis**
 - used in retail for planning arrangement on shelves
 - used for identifying cross-selling opportunities
 - “should” be used to determine best link structure for a Web site
- **Examples**
 - people who buy milk and beer also tend to buy diapers
 - people who access pages A and B are likely to place an online order
- **Suitable data mining tools**
 - association rule discovery
 - clustering
 - Nearest Neighbor analysis (memory-based reasoning)



Association Rule Discovery: Definition

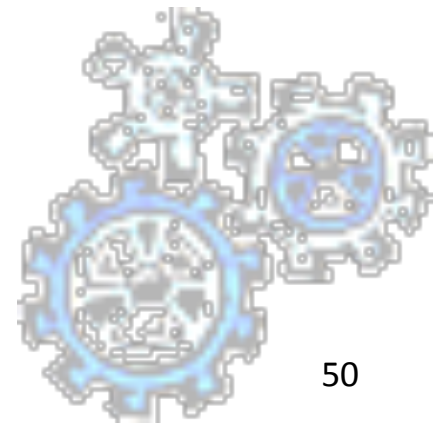
- Given a set of records each of which contain some number of items from a given collection;
 - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<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, ... } --> {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!

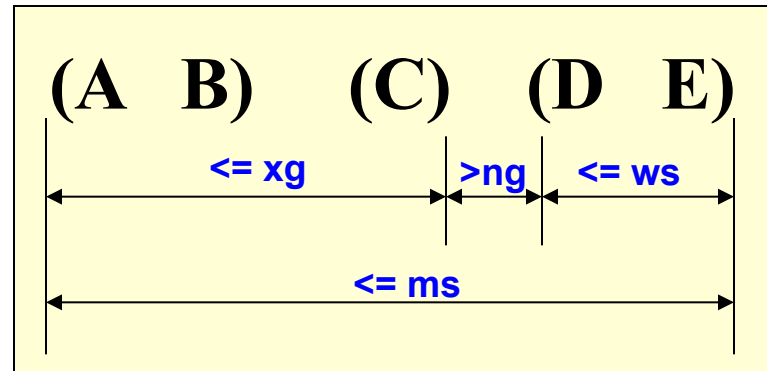


Sequential Pattern Discovery: Definition

- 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 events.

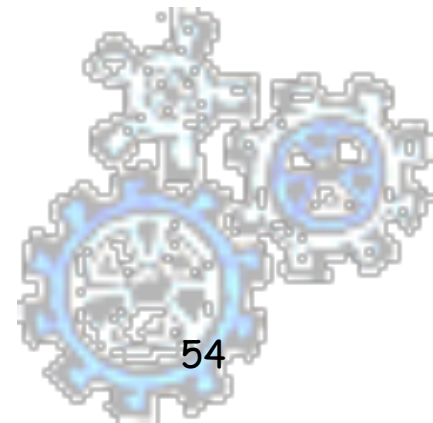
(A B) (C) \longrightarrow (D E)

- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Exception/deviation detection

- **Generate a model of normal activity**
- **Deviation from model causes alert**
- **Methods:**
 - Artificial neural networks
 - Inductive decision tree and rule systems
 - Statistical methods
 - Visualization tools

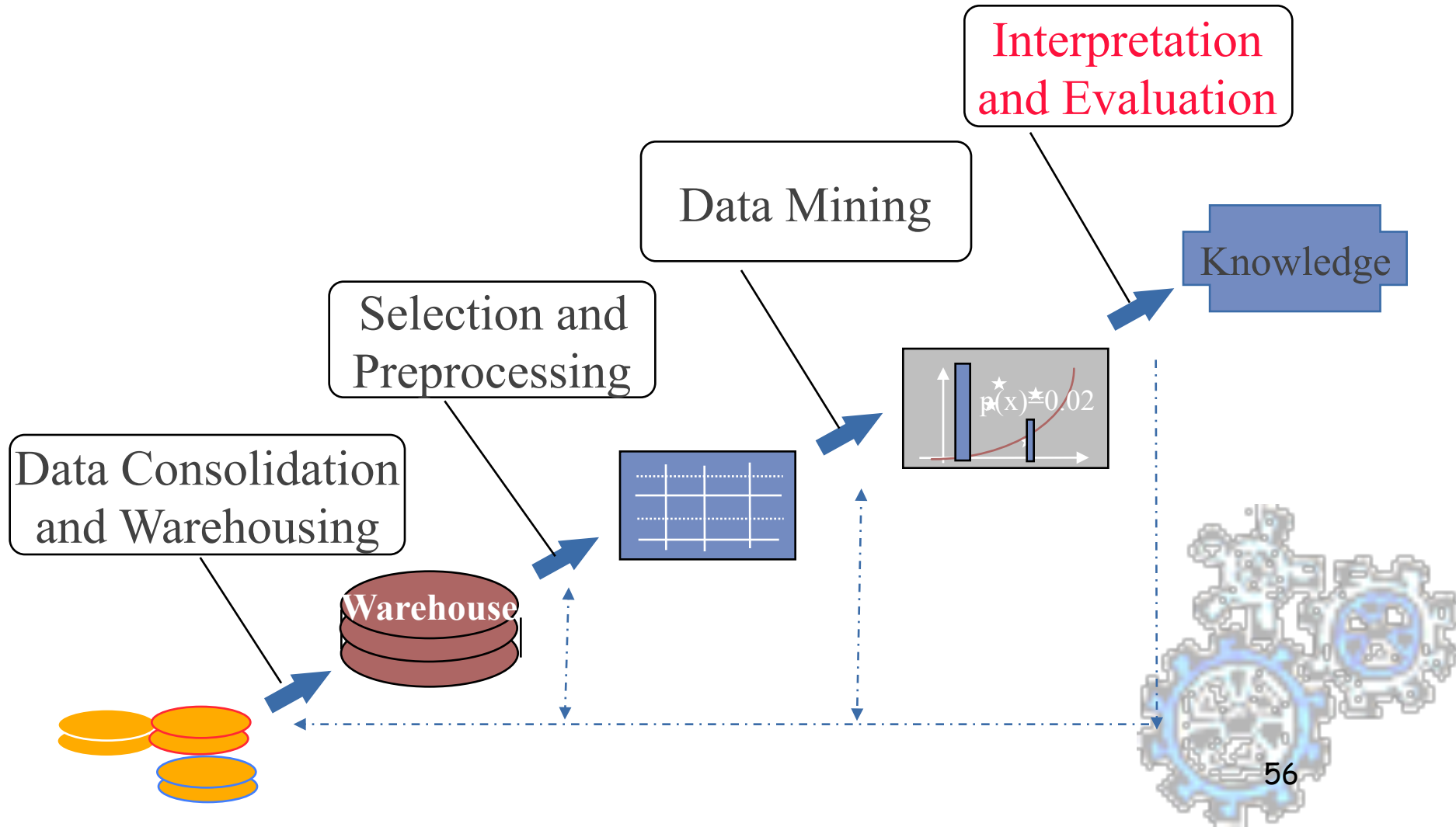


Outlier and exception data analysis

- Time-series analysis (trend and deviation):
 - Trend and deviation analysis: regression, sequential pattern, similar sequences, trend and deviation, e.g., stock analysis.
 - Similarity-based pattern-directed analysis
 - Full vs. partial periodicity analysis
- Other pattern-directed or statistical analysis



The KDD process



Are all the discovered pattern interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
- **Interestingness measures:**
 - 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



Interpretation and evaluation

Evaluation

- Statistical validation and significance testing
- Qualitative review by experts in the field
- Pilot surveys to evaluate model accuracy

Interpretation

- Inductive tree and rule models can be read directly
- Clustering results can be graphed and tabled
- Code can be automatically generated by some systems (IDTs, Regression models)

