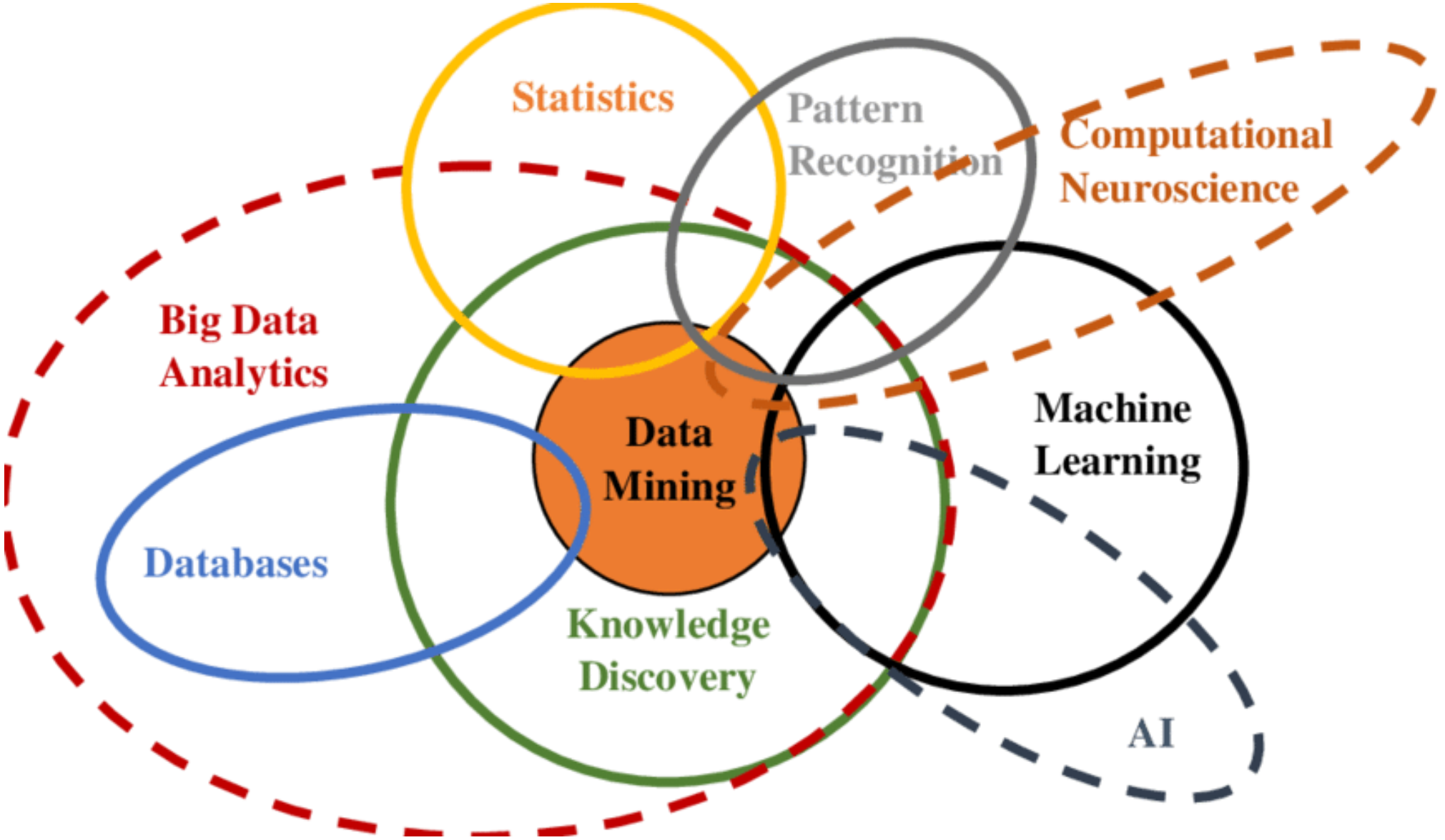
Big Data Applications and Analytics

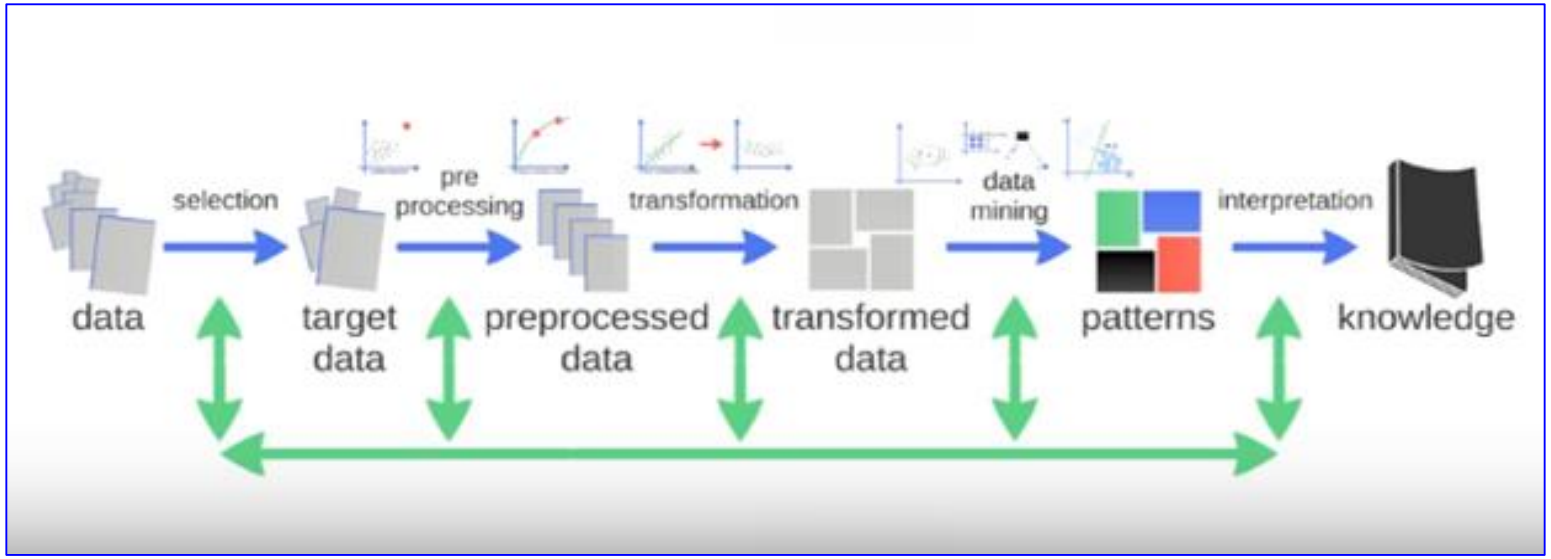

Statistics
Pattern Recognition
Computational Neuroscience
Big Data Analytics
Machine Learning
Data Mining
Databases
Knowledge Discovery
AI

Part 1: Data Mining
Part 2: Types of Data
Part 3: Big Data
Part 4: Data Analytics
Part 5: Big Data Applications

Overview

- What is Data Mining
- Data, Information, Knowledge
- How Does Data Mining Work
- Knowledge Discovery in Database (KDD)
- Data Mining Techniques
- Data Mining Tools
- Why is Data Mining useful in business?

Data, Information, and Knowledge

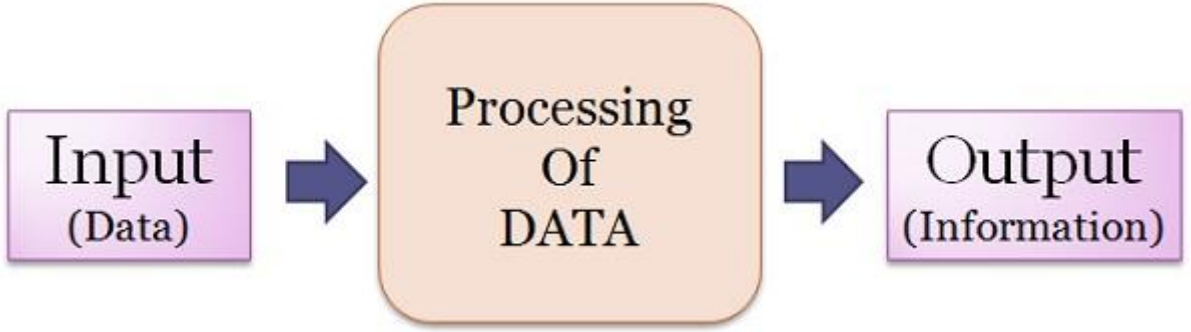


Data & Information

What is Data?
Raw, unorganized facts that need to be processed.

What is Information?
Processed, organized, structured data that is useful.

Data is plain facts that is processed, organized, structured or presented into useful information.



What is Information

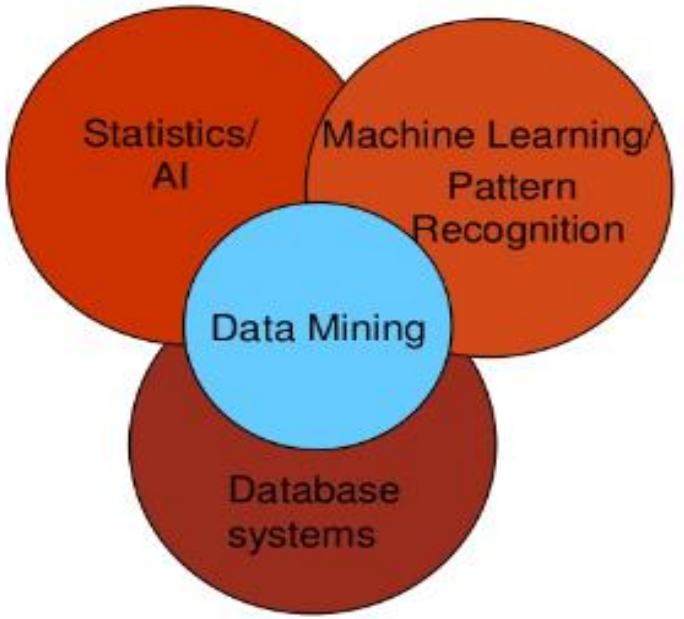
- The patterns
- associations
- relationships among all this data can provide information.
- For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Data explosion Problem

- Advance data collection tools and database technology lead to tremendous amounts of data stored in databases
- 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 using Data Mining

What Is Data Mining?

Data Mining

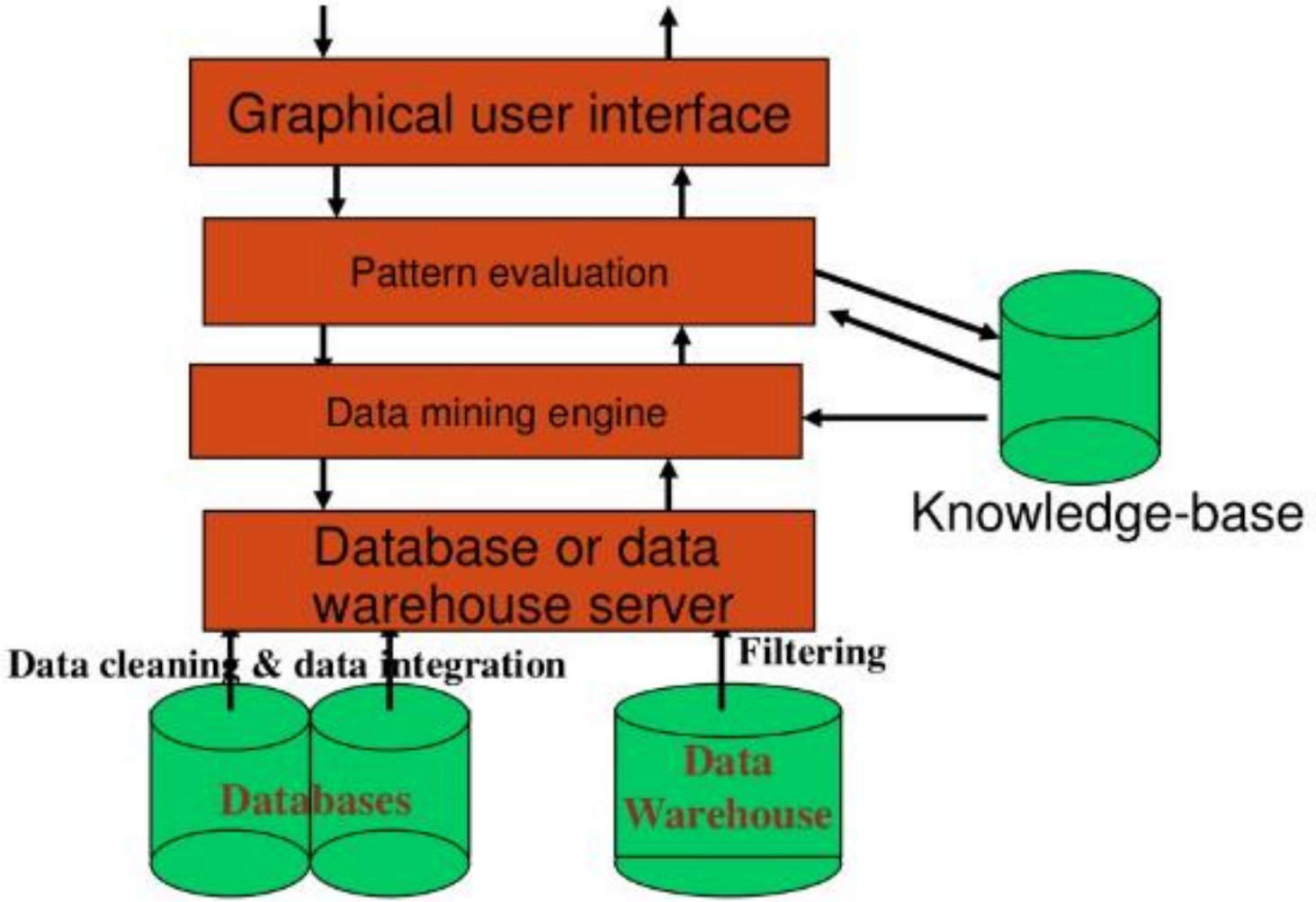


- is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- overall goal is to extract information from a data set and transform it into knowledge

Data Mining

- refers to the application of algorithms for extracting patterns from data without the additional steps of the Knowledge Discovery in Database (KDD) process.
- Data mining is considered a misnomer, because it’s the extraction of patterns and knowledge from large amounts of data not the extraction (mining) of data itself.
- is the analysis step of the “KDD” process.

Architecture of a Typical Data Mining System

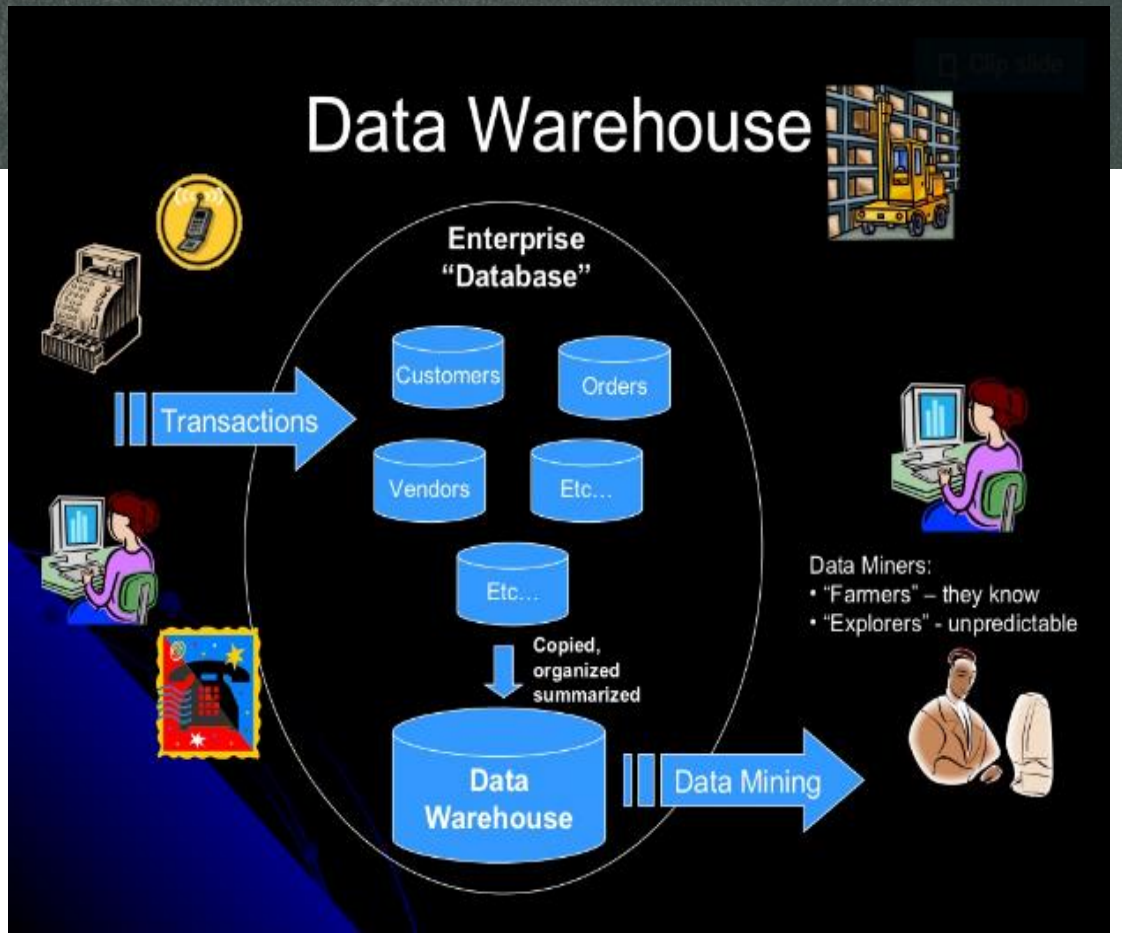


What is Knowledge

- Information can be converted into knowledge about historical patterns and future trends.
- For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

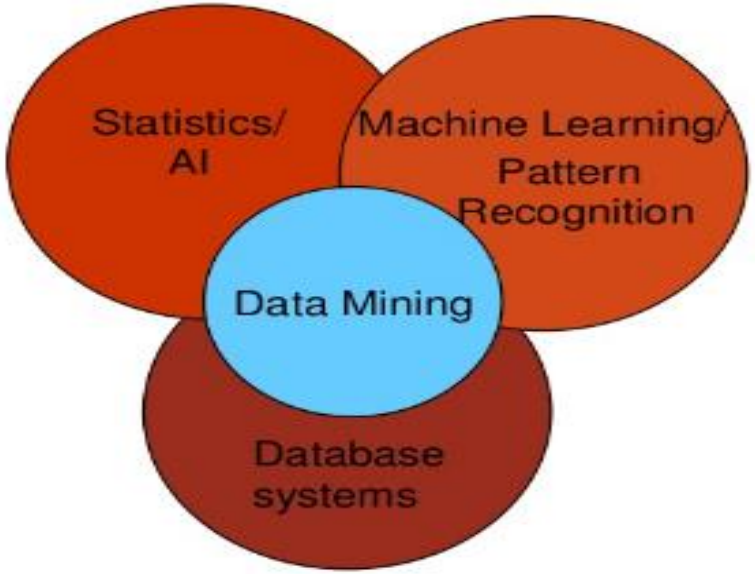
Data Warehousing

- Data Warehousing provides Enterprise with a memory
- Data Mining provides the Enterprise with intelligence
- On-Line Application Processing (OLAP):
- Few, but complex queries --- may run for hours.
- Queries do not depend on having an absolutely up-to-date database.
- Example: Analysts at Wal-Mart look for items with increasing sales in some region



How Does Data Mining Work

How Does Data Mining Work

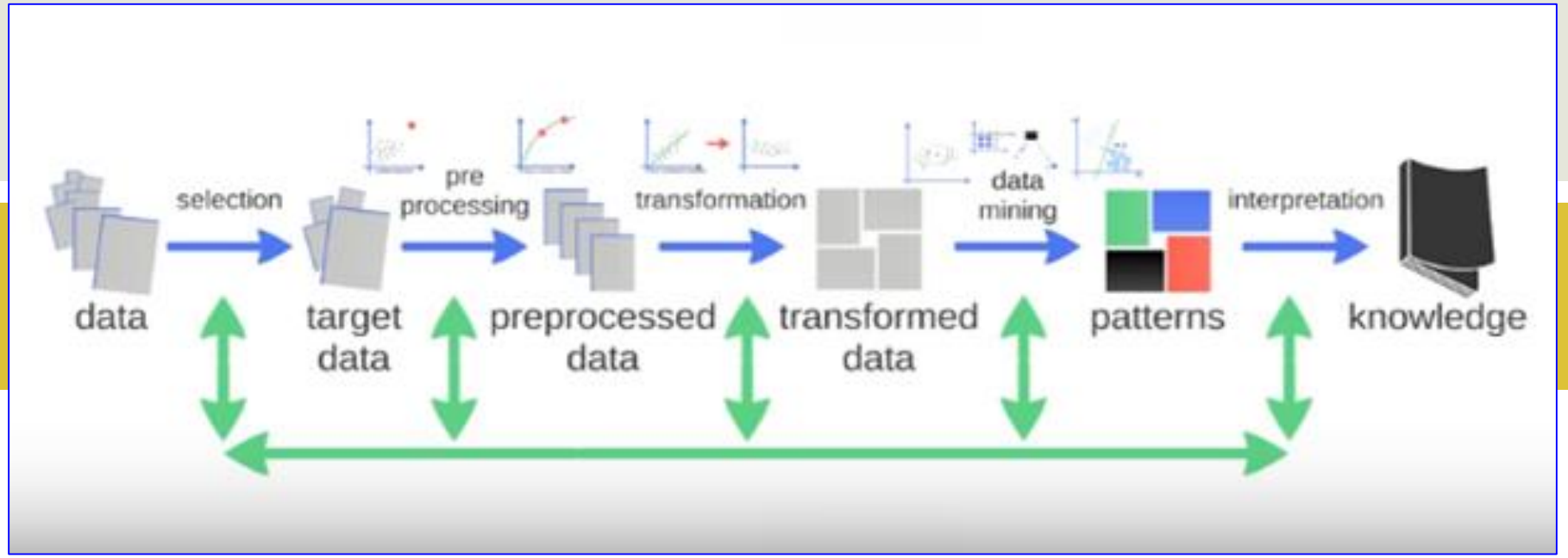


- While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two.
- Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

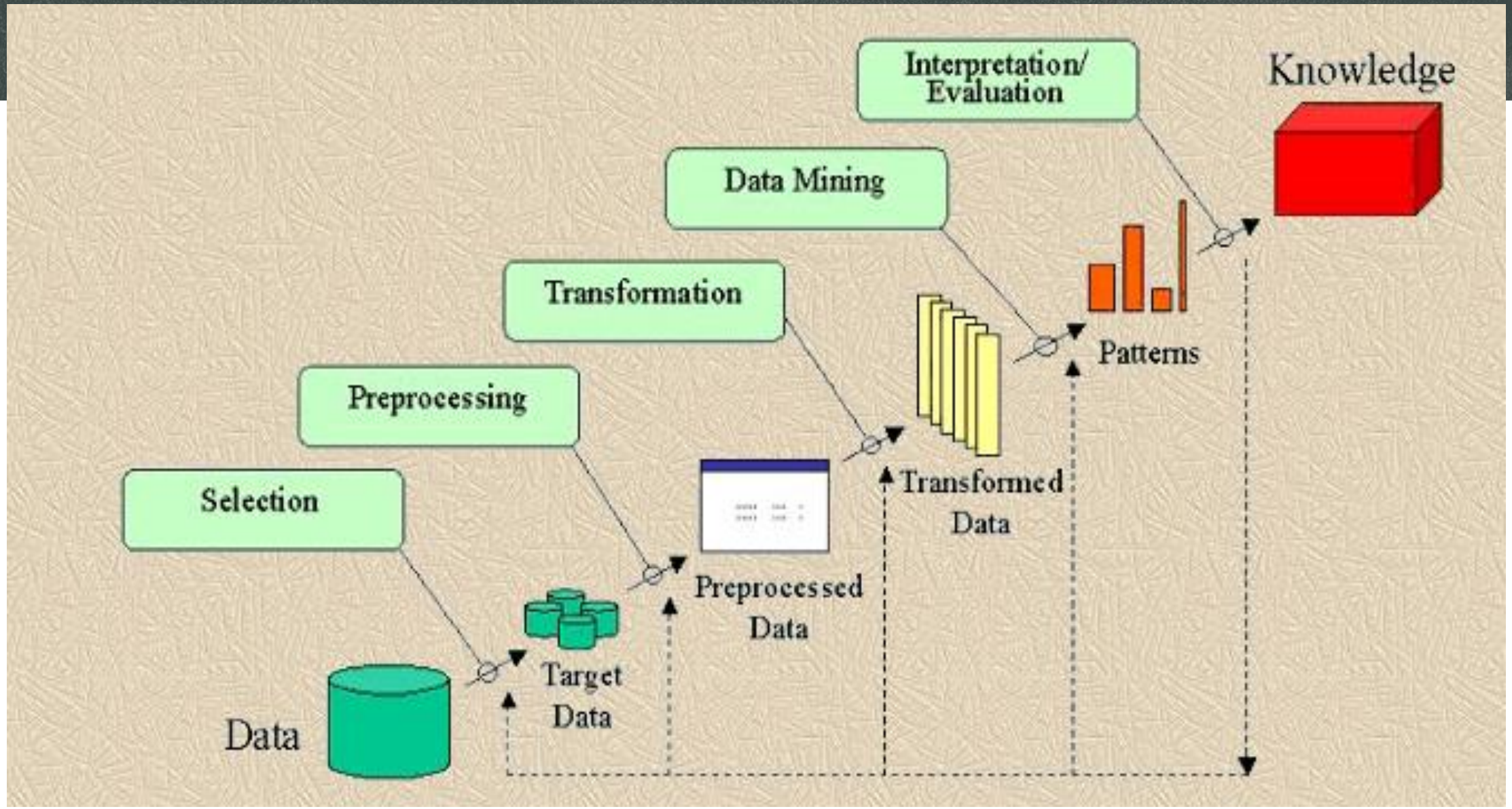
Data Mining Consists of Five Major Elements

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table

Knowledge Discovery in Database (KDD)



Knowledge Discovery in Database (KDD)

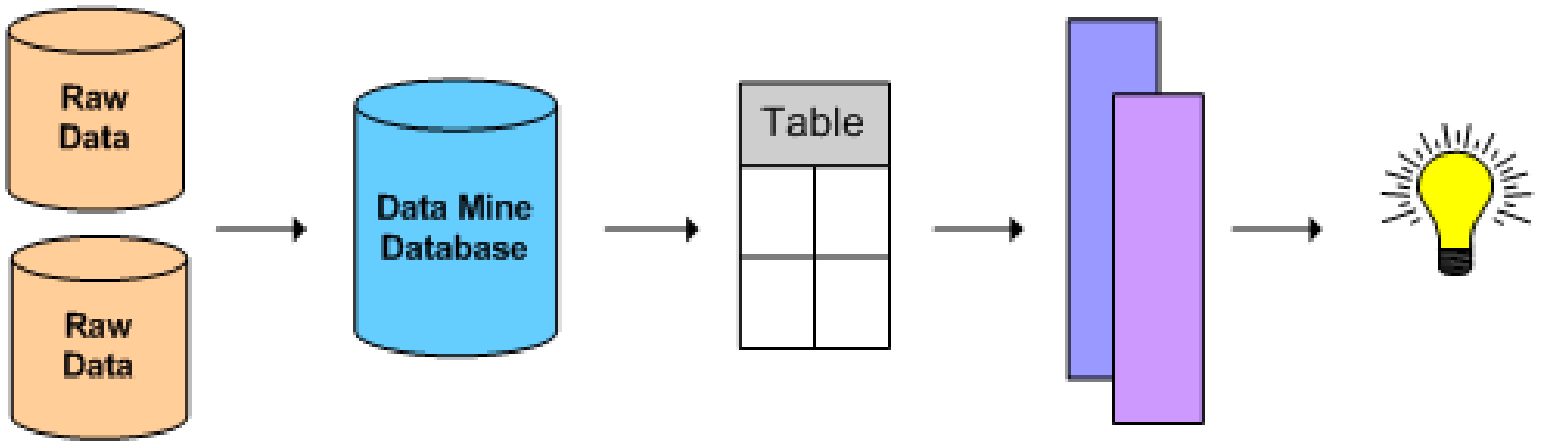


Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD)

- Refers to the overall process of discovering useful knowledge from data.
- It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge.
- It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

Basic Data Mining Process



Extraction of data from repositories

Data cleansing and loading into data mining database

Data transformation

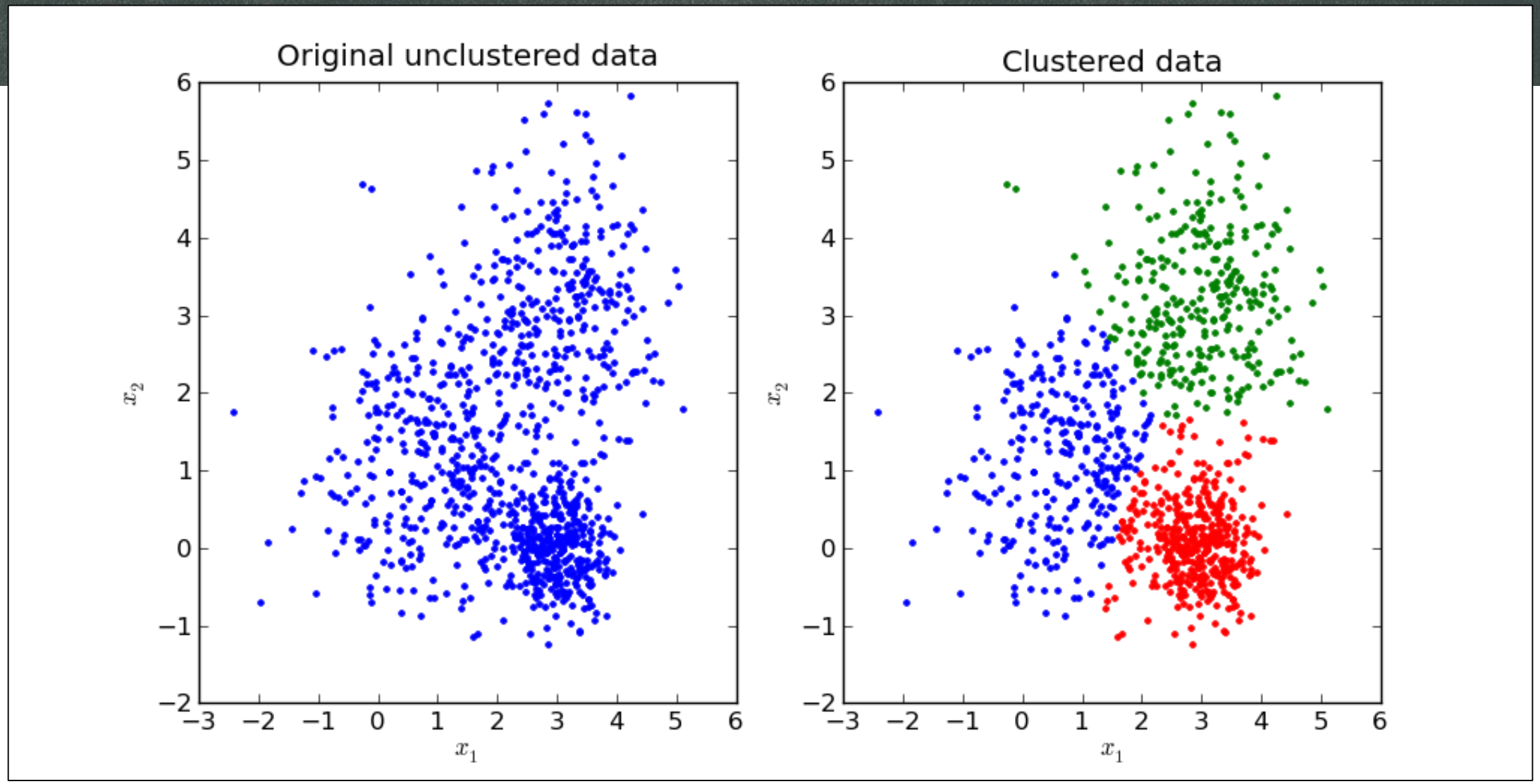
Pattern discovery using algorithms such as *clustering*, *regression* and *classification*

Data visualization and interpretation of results

Classification data example

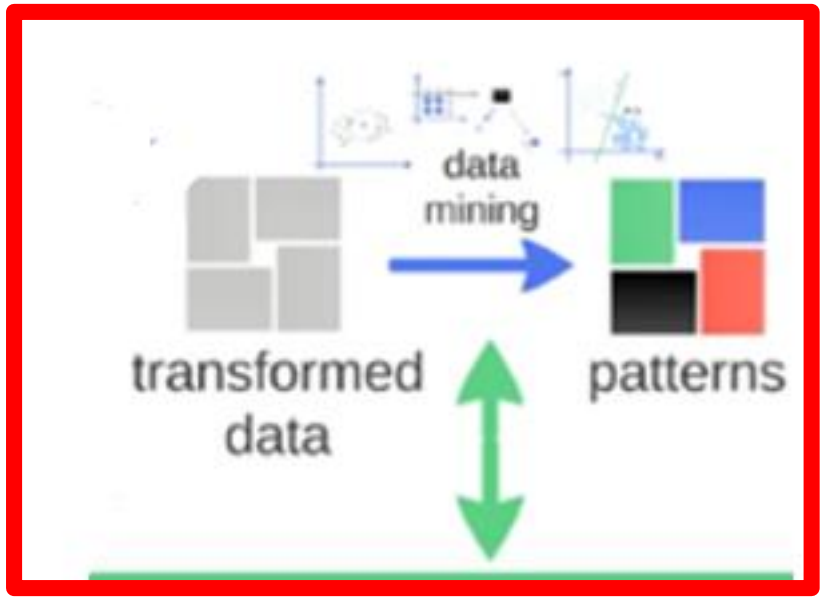
	Predictors				Response
	Outlook	**Temperature**	**Humidity**	**Wind**	**Class** **Play=Yes** **Play=No**
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

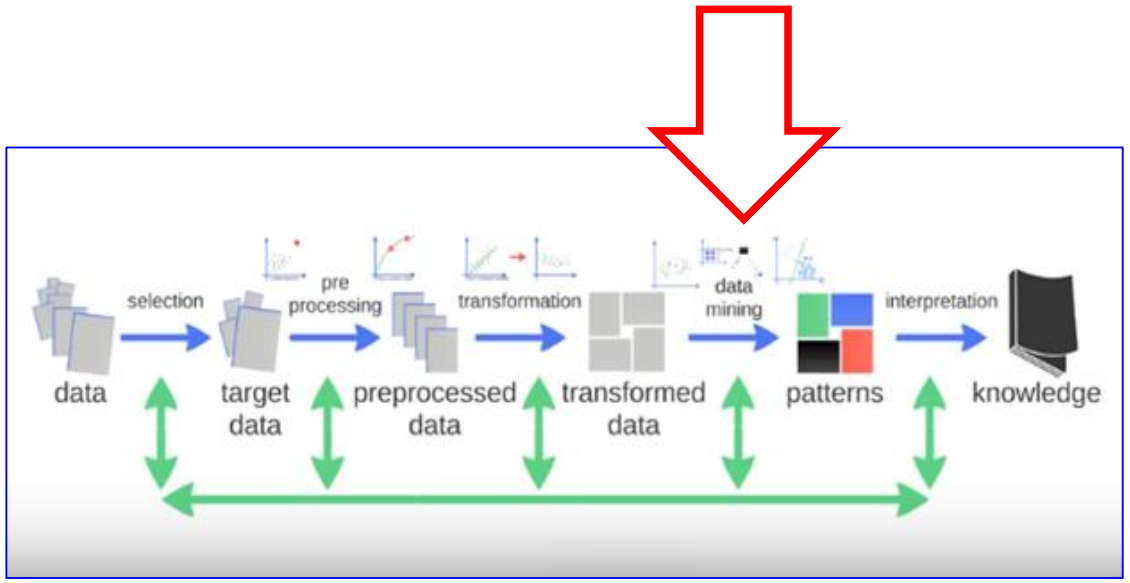
Clustered data (clustering) example



Data Mining

1. Choosing the data mining task
   Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
2. Choosing the data mining algorithms
   Selecting method(s) to be used for searching for patterns in the data.
   Deciding which models and parameters may be appropriate.
   Matching a particular data mining method with the overall criteria of the KDD process.
3. Data mining.
   Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.





Data Mining Tasks

The goals of prediction and description are achieved by using the following primary data mining tasks:

1. **Classification** is learning a function that maps (classifies) a data item into one of several predefined classes.
2. **Regression** is learning a function which maps a data item to a real-valued prediction variable.
3. **Clustering** is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data.
    - Closely related to clustering is the task of *probability density estimation* which consists of techniques for estimating, from joint multivariate probability density function of all of the variables/fields in the database.

Data Mining Tasks

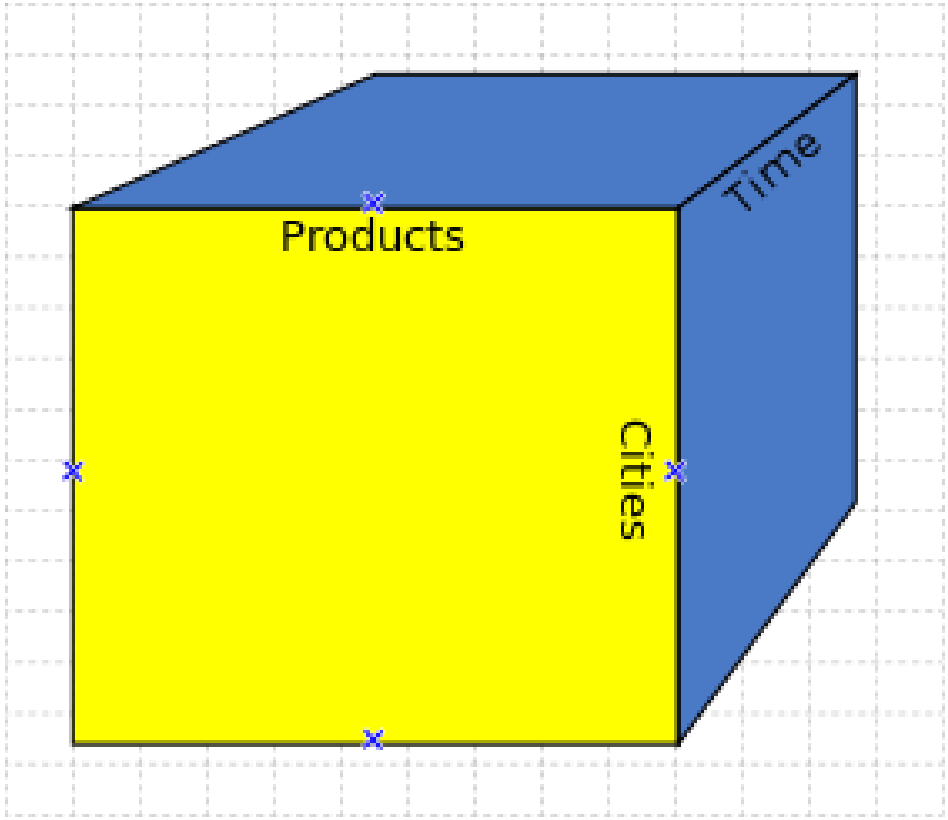
4. Summarization involves methods for finding a compact description for a subset of data.

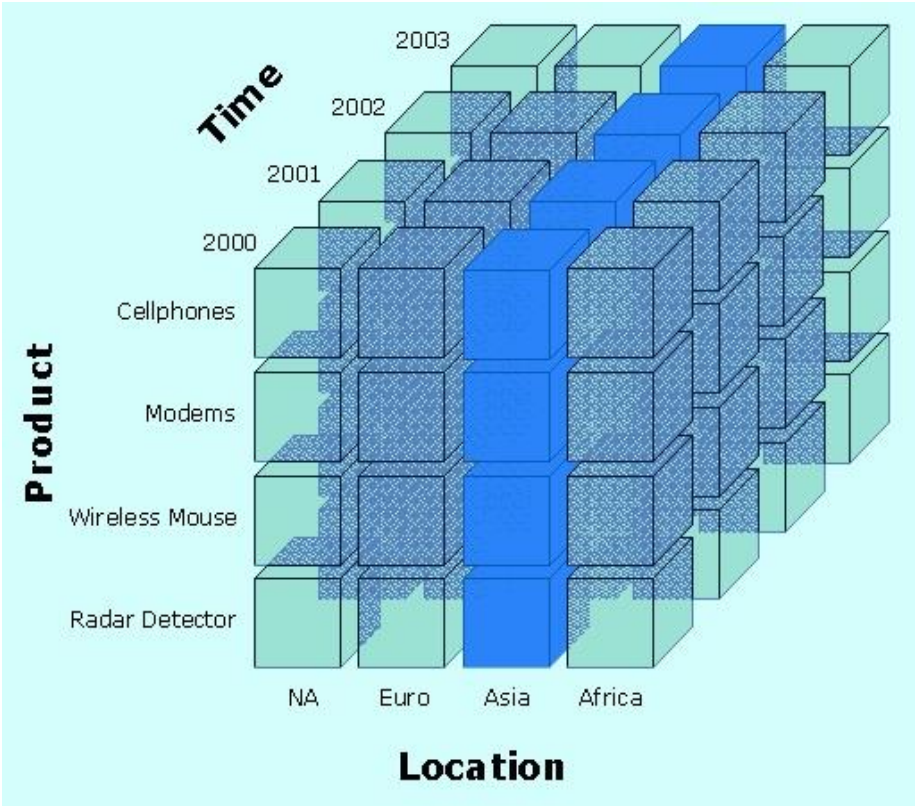
5. Dependency Modeling consists of finding a model which describes significant dependencies between variables.

- Dependency models exist at two levels:
  i. The *structural* level of the model specifies (often graphically) which variables are locally dependent on each other, and
  ii. The *quantitative* level of the model specifies the strengths of the dependencies using some numerical scale.

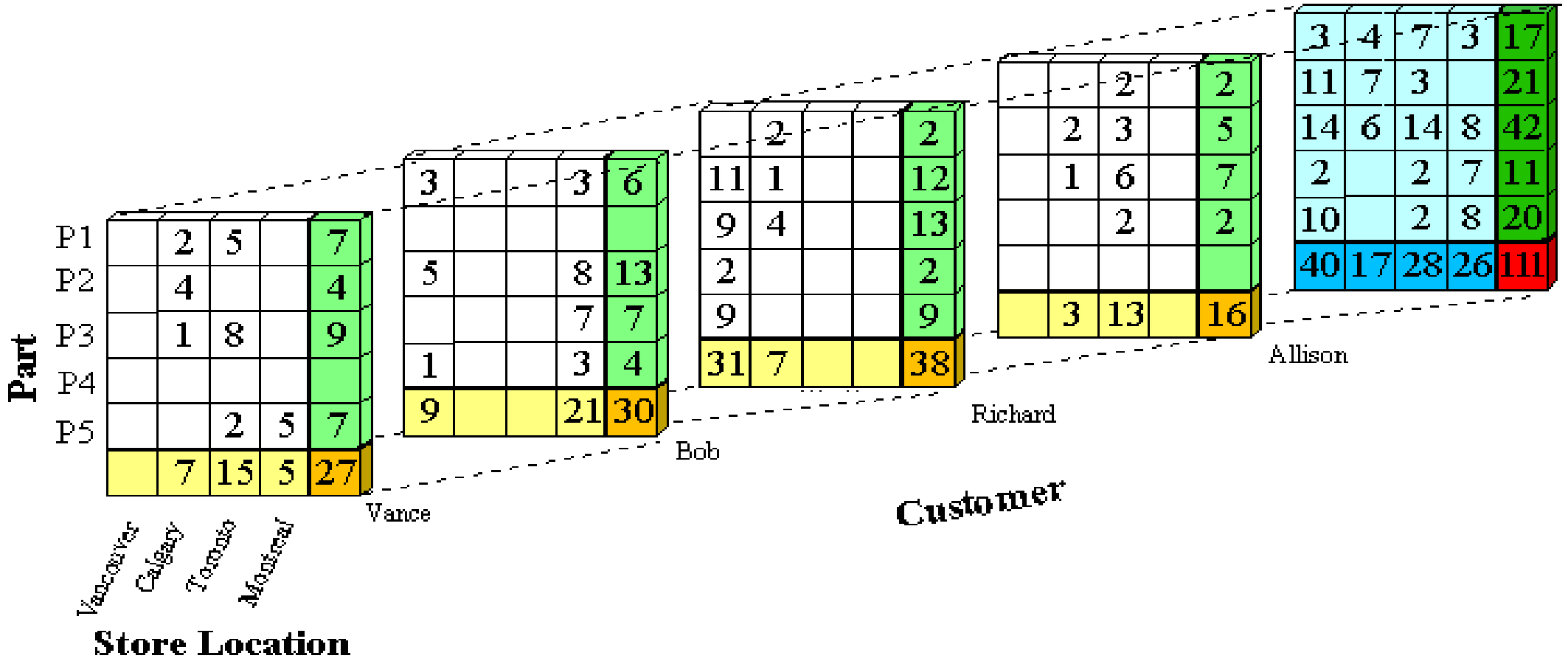
6. Change and Deviation Detection focuses on discovering the most significant changes in the data from previously measured or normative values.

Data Mining and On-Line Application Processing (OLAP)Cubes

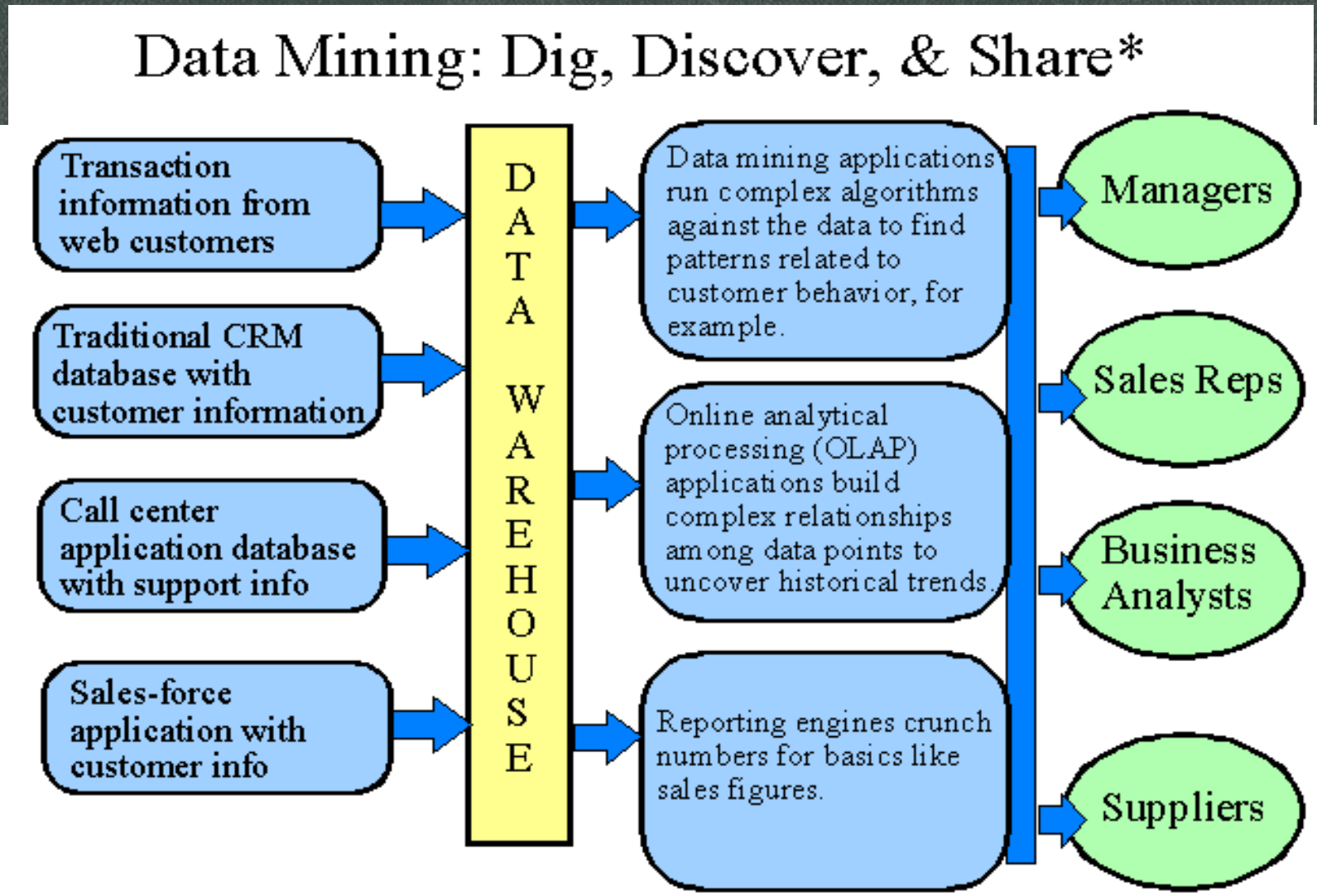




Data Mining and OLAP Cubes



Data Mining Visualization



* Adapted from Roberts-Witt's illustration, *PC Magazine*, November 19, 2002, p. ubiz 5.

Data Mining Techniques

Data Mining Techniques

Techniques

- Association rules
- Classification
- Clustering
- Decision trees
- Regression
- Neural networks

What are Data Mining Techniques Used for

- Classification and Prediction
  - Example - Focused Hiring
  - predict one or more discrete variables, based on the other attributes in the dataset.
- Cluster Analysis
  - Example - Market Segmentation
- Outlier Analysis
  - Example - Fraud Detection
- Association Analysis
  - Example - Market Basket Analysis
- Evolution Analysis
  - Example - Forecasting stock market index using Time Series Analysis

How Data Mining is Used in Business

How Data Mining is Used in Business

- It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.
- Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining is Used in Business

- Data Mining is primarily used today by companies with a strong consumer focus, to “drill down” into their transactional data.
- data mining helps determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits.
- Businesses can use this information to influence their business decisions, follow and predict trends, and improve efficiency.

How Data Mining is Used in Business

Basket Analysis

- Sometimes called “affinity analysis,” this looks at the items that a customer bought, which could help brick-and-mortar stores improve their layouts or online companies like Amazon recommend related products. The “basket” refers to what shoppers use when they are shopping.

- It’s based on the assumption that you can predict future customer behavior by past performance, including purchases and preferences. And it’s not just grocery stores that can use this data. Here are a few ways it can be applied in various industries:

How Data Mining is Used in Business

Card Marketing

- If your business involves issuing **credit cards,** you can collect the information from usage, identify customer segments and then based on information on these segments build programs that improve retention, boost acquisition, target products to develop and design prices.

- A great example of this occurred when the UN decided to issue a Visa credit card to people who traveled overseas frequently. The agency marketers segmented their database into wealthy travelers—30,000 people in high-income households.

- The agency marketers used direct mail for their appeals and generated a 3% response. That may sound small, but it actually exceeded industry standards. Large financial institutions typically see 0.5% response rates. That's how effective databases can be when marketing cards.

Q & A

Thank You