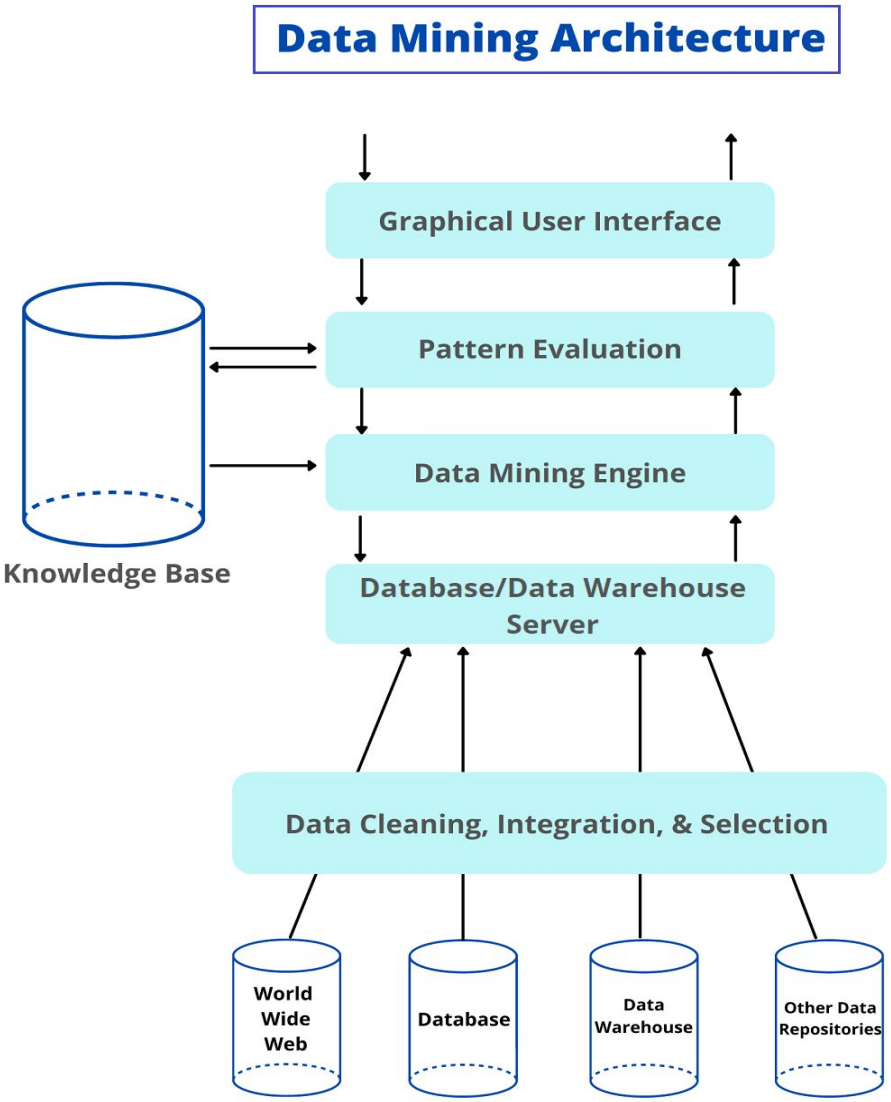
Data Mining:

Architecture and Data Types

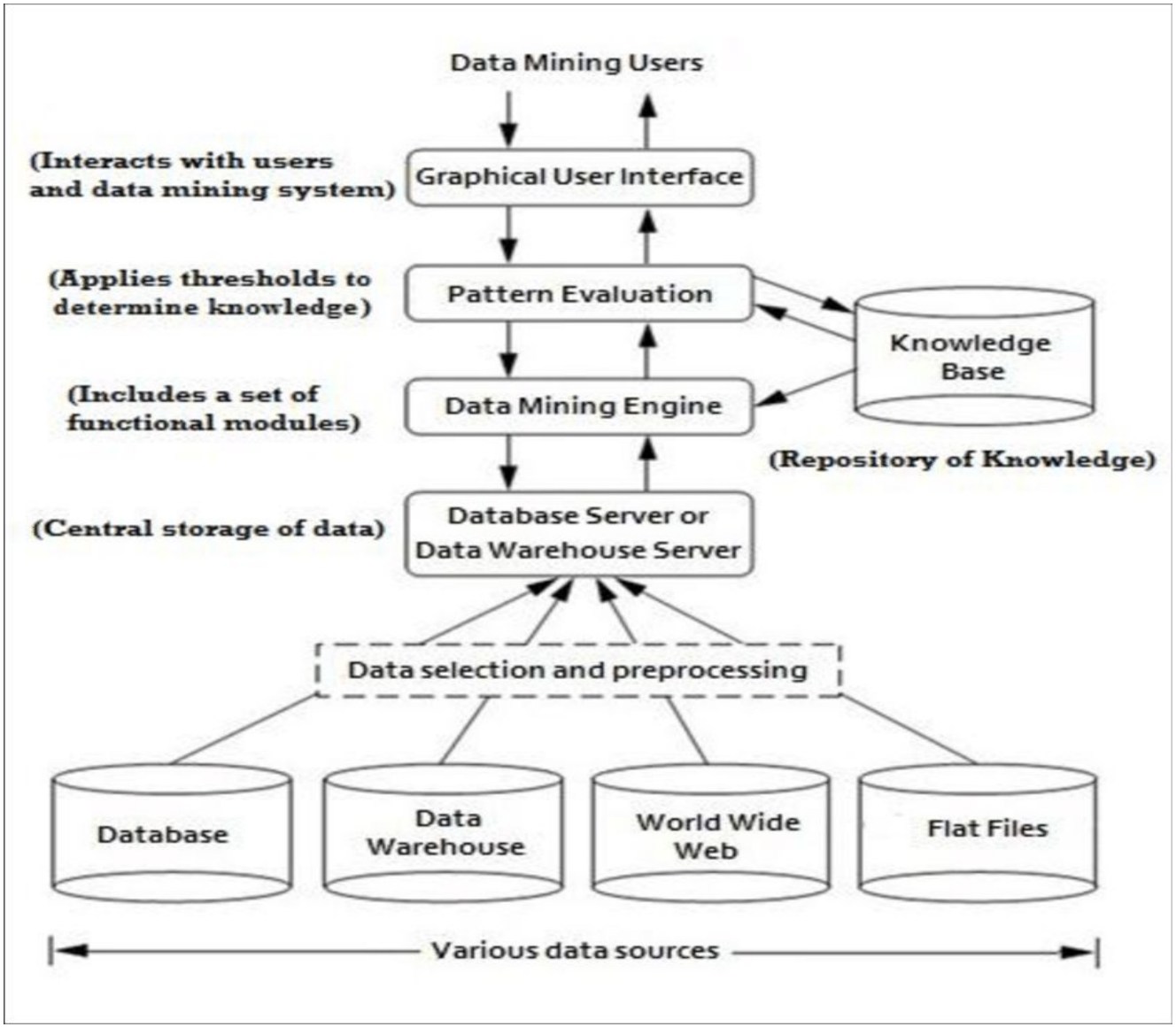
Overview

- Data Mining Architecture
- Data Objects and Attribute Types
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Data Mining Architecture



Data Mining Architecture



Data Mining Architecture

Data Mining Engine

Perform data mining tasks. That includes association, classification, characterization, clustering, prediction, etc.

Pattern Evaluation Modules

This module is mainly responsible for the measure of interestingness of the pattern.

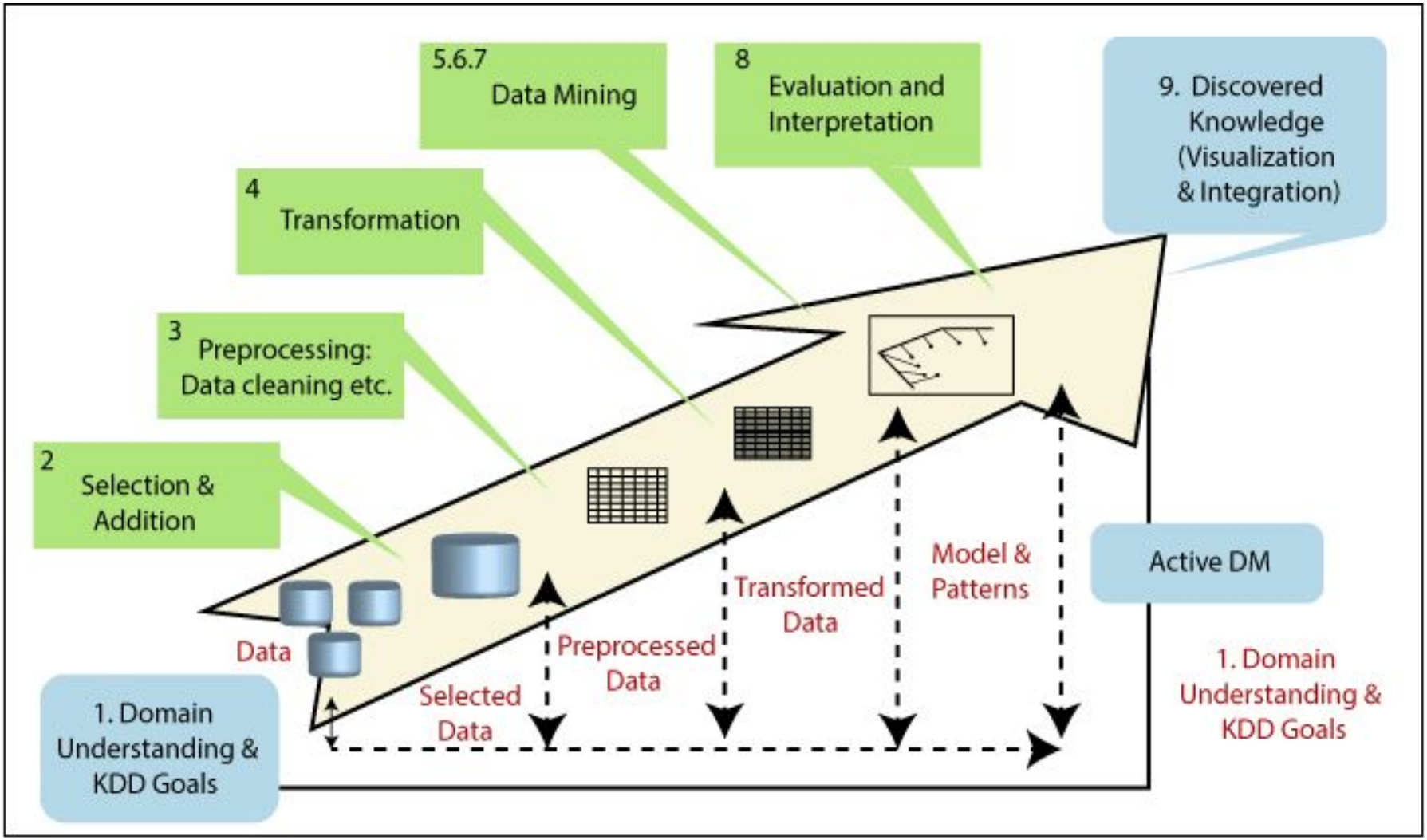
Graphical User Interface

We use this interface to communicate between the user and the data mining system. A

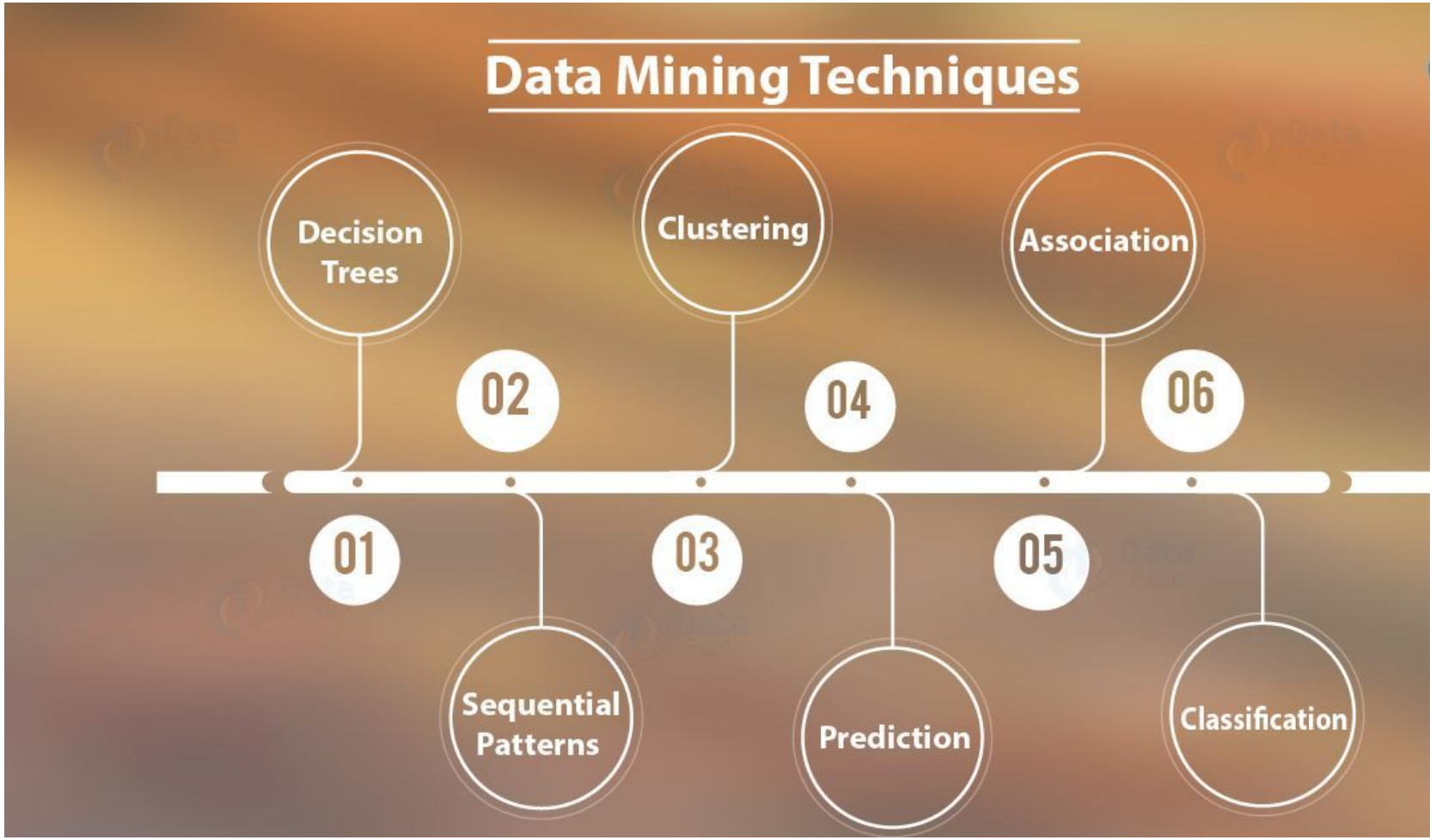
Knowledge Base

The knowledge base might even contain user beliefs and data from user experiences. That can be useful in the process of data mining.

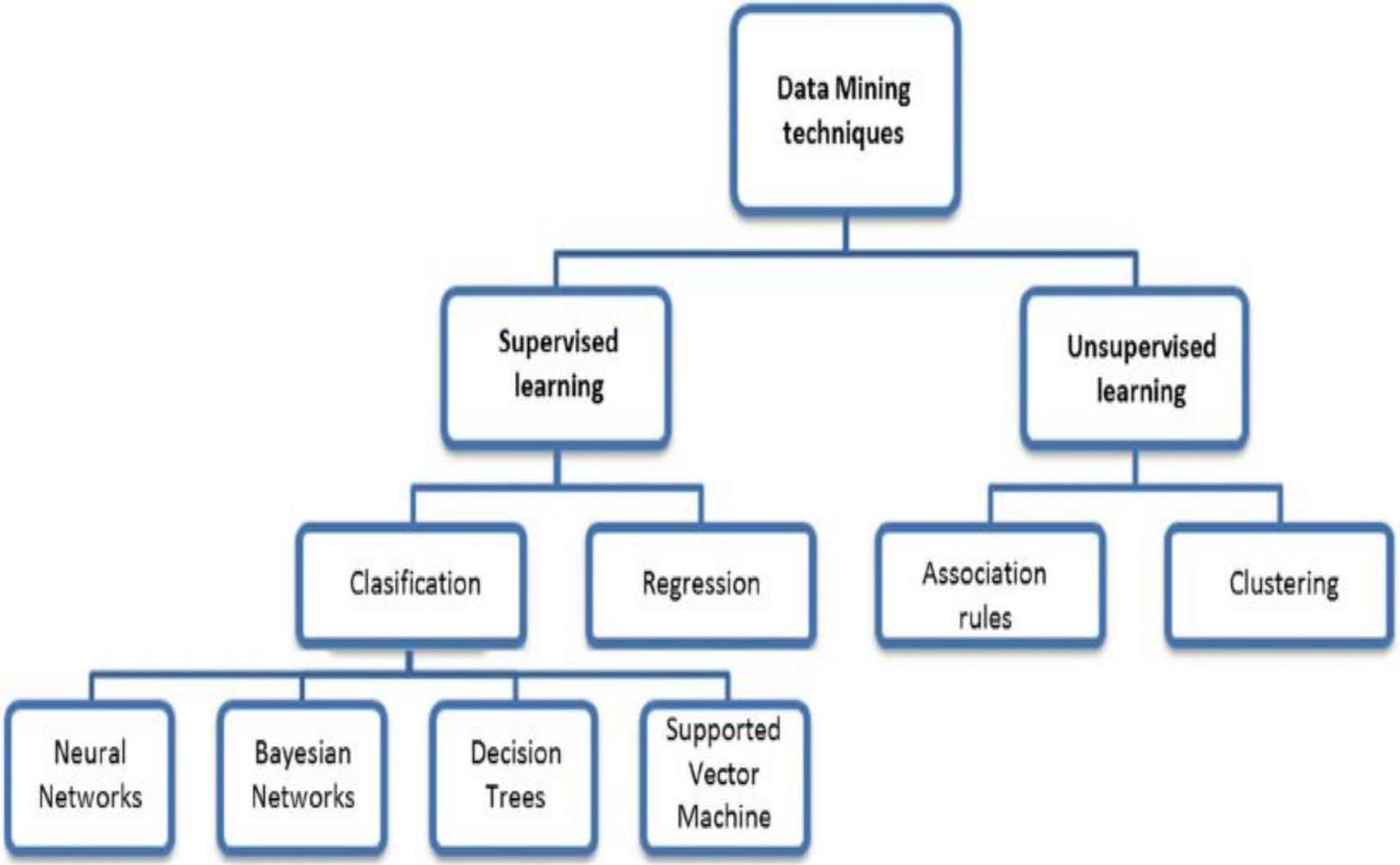
KDD PROCESS


5.6.7
Data Mining
8
Evaluation and
Interpretation
9. Discovered
Knowledge
(Visualization
& Integration)
4
Transformation
3
Preprocessing:
Data cleaning etc.
2
Selection &
Addition
Active DM
Model &
Patterns
Transformed
Data
Preprocessed
Data
Selected
Data
Data
1. Domain
Understanding &
KDD Goals
1. Domain
Understanding &
KDD Goals

DATA MINING TECHNIQUES



DATA MINING TECHNIQUES



Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes Types

- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Data Mining Patterns

A pattern means that the data (visual or not) are correlated that they have a relationship and that they are predictable.

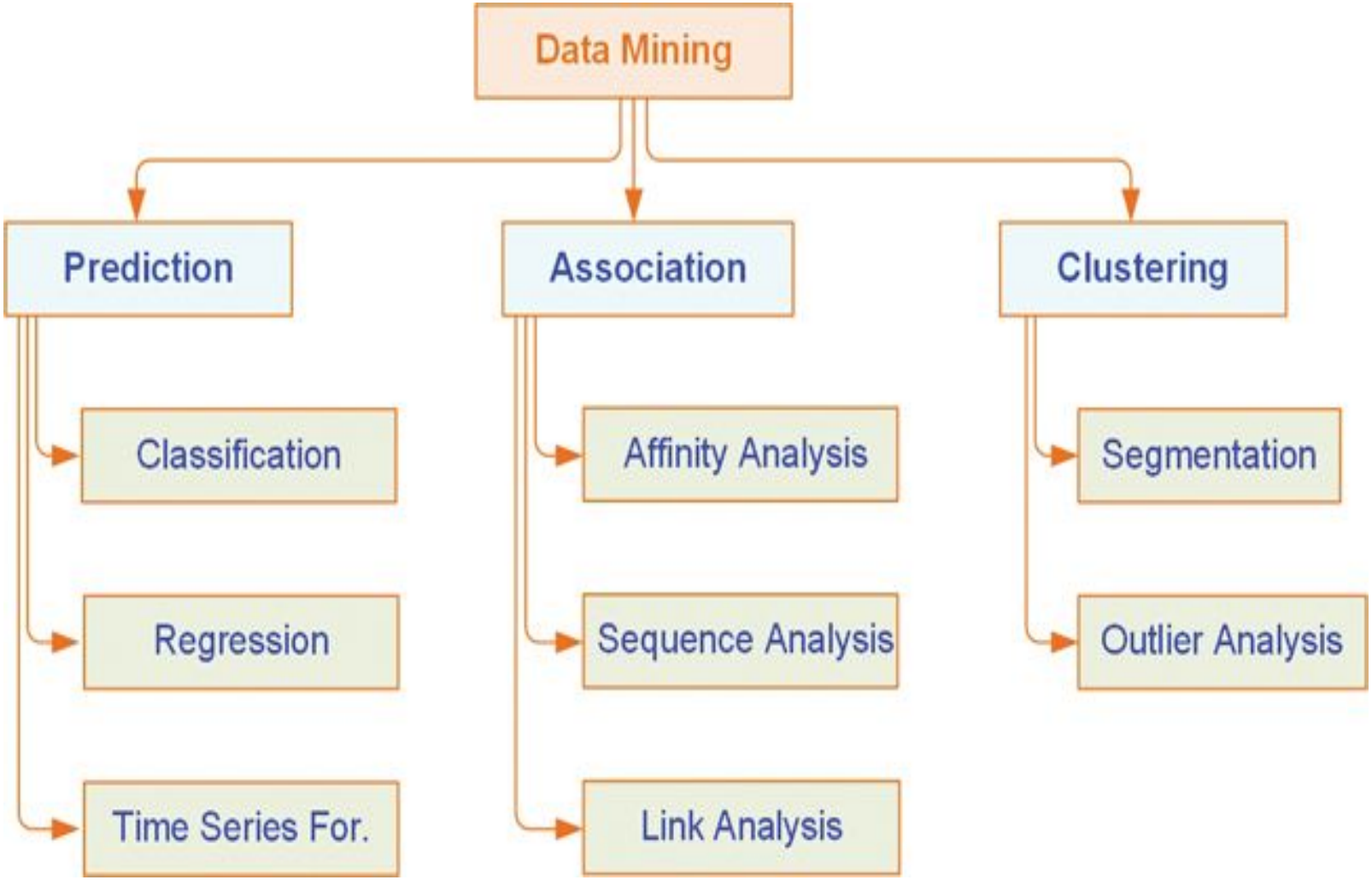
When you find a pattern, you can have a good idea when or where something will happen before it actually happens.

Data Mining Patterns

Data mining seeks to identify three major types of patterns:

- **Associations** find commonly co-occurring groupings of things, such as "bread and butter" commonly purchased and observed together in a shopping cart (i.e., market-basket analysis).
- Another type of association pattern captures the sequences of things. These sequential relationships can discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.
- **Predictions** tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the Super Bowl or forecasting the absolute temperature on a particular day.
- **Clusters** identify natural groupings of things based on their known characteristics, such as assigning customers in different segments based on their demographics and past purchase behaviors.

Data Mining Patterns



Summary

- Data Mining Architecture
- Types of Data in Data Mining
- Data Mining Techniques
- Data Mining Patterns