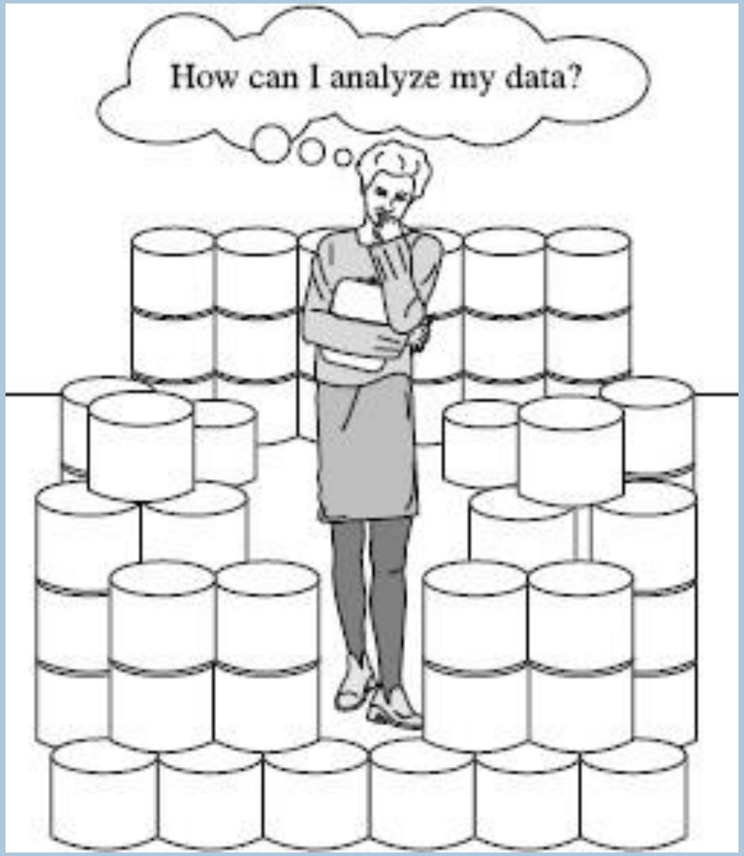
DATA MINING
Data sets
Pre-processing
Classification
Database
Statistics
Analytics
Evaluation

GJU
الجامعة الألمانية الأردنية
German Jordanian University

What Is Data Mining?

- We live in a world where vast amounts of data are generated constantly and rapidly
- **Data mining** is the process of discovering interesting patterns, models and other kinds of knowledge in large data sets
  - “Data mining”: a misnomer? It should be “knowledge mining from data”
  - Other terms: *Knowledge mining from data*, *KDD* (*Knowledge Discovery from Data*), *pattern discovery*, *knowledge extraction*, *data analytics*, *information harvesting*
- Data mining is a young, dynamic, and promising field
- Example: Data mining turns a large collection of data into knowledge
  - Google’s *Flu Trends* found a close relationship between the number of people who search for flu-related info. and the number of people who have flu symptoms
    - It can estimate flu activity up to two weeks faster than traditional systems

Why do we need Data Mining?



We are data rich, but information poor

- Huge volumes of data are accumulated in databases and data warehouses.
- Huge volumes of data also come from WWW and data streams
  - (video surveillance, telecommunication, and sensor networks)
- Effective and efficient analysis of data in different forms becomes a challenging task.
- Fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools
- Decision-making is done according to **information** (not data)

What is Data Mining?

- Many Definitions for Data Mining:
  - **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**
  - **Extracting or “mining” knowledge from large amounts of data.**

- Alternative Names for Data Mining
  - **Knowledge Discovery from Data (KDD)**
  - Knowledge Discovery (mining) in Databases, knowledge extraction, data/pattern analysis, data archeology, information harvesting, business intelligence, etc

What is (not) Data Mining?

- Is everything “data mining”?
- What is not Data Mining?
  - Simple search and query processing
  - (Deductive) expert systems
  - Look up phone number in phone directory
  - Query a Web search engine for information about “GJU"
- What is Data Mining?
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Which items are bought together in a market?

What is Data Mining: Tasks 1

Discuss whether or not each of the following activities is a data mining task?

- **Dividing the customers of a company according to their gender.**
  - **If their genders are recorded in data.**
  - **If their genders are NOT recorded in data.**

- **Computing the total sales of a company.**

- **Sorting a student database based on student identification numbers.**

What is Data Mining?: Tasks 1

Discuss whether or not each of the following activities is a data mining task?

- **Dividing the customers of a company according to their gender.**
  - **If their genders are recorded in data.** **NO**
  - **If their genders are NOT recorded in data.** **YES**

- **Computing the total sales of a company.** **NO**

- **Sorting a student database based on student identification numbers.** **NO**

What is Data Mining?: Tasks 2

Discuss whether or not each of the following activities is a data mining task?

- **Predicting the outcomes of tossing a fair coin.**

- **Predicting the future stock price of a company using historical records.**

- **Monitoring the heart rate of a patient for abnormalities.**

What is Data Mining?: Tasks 2

Discuss whether or not each of the following activities is a data mining task?

- **Predicting the outcomes of tossing a fair coin.** **NO**
- **Predicting the future stock price of a company using historical records.** **YES**
- **Monitoring the heart rate of a patient for abnormalities.** **YES**

What is Data Mining?: Tasks 3

Discuss whether or not each of the following activities is a data mining task?

- **Finding the category (economy, sport, …) of a newspaper article**
- **Deciding whether an email is a spam or not.**
- **Deciding whether an image contains an apple or not.**

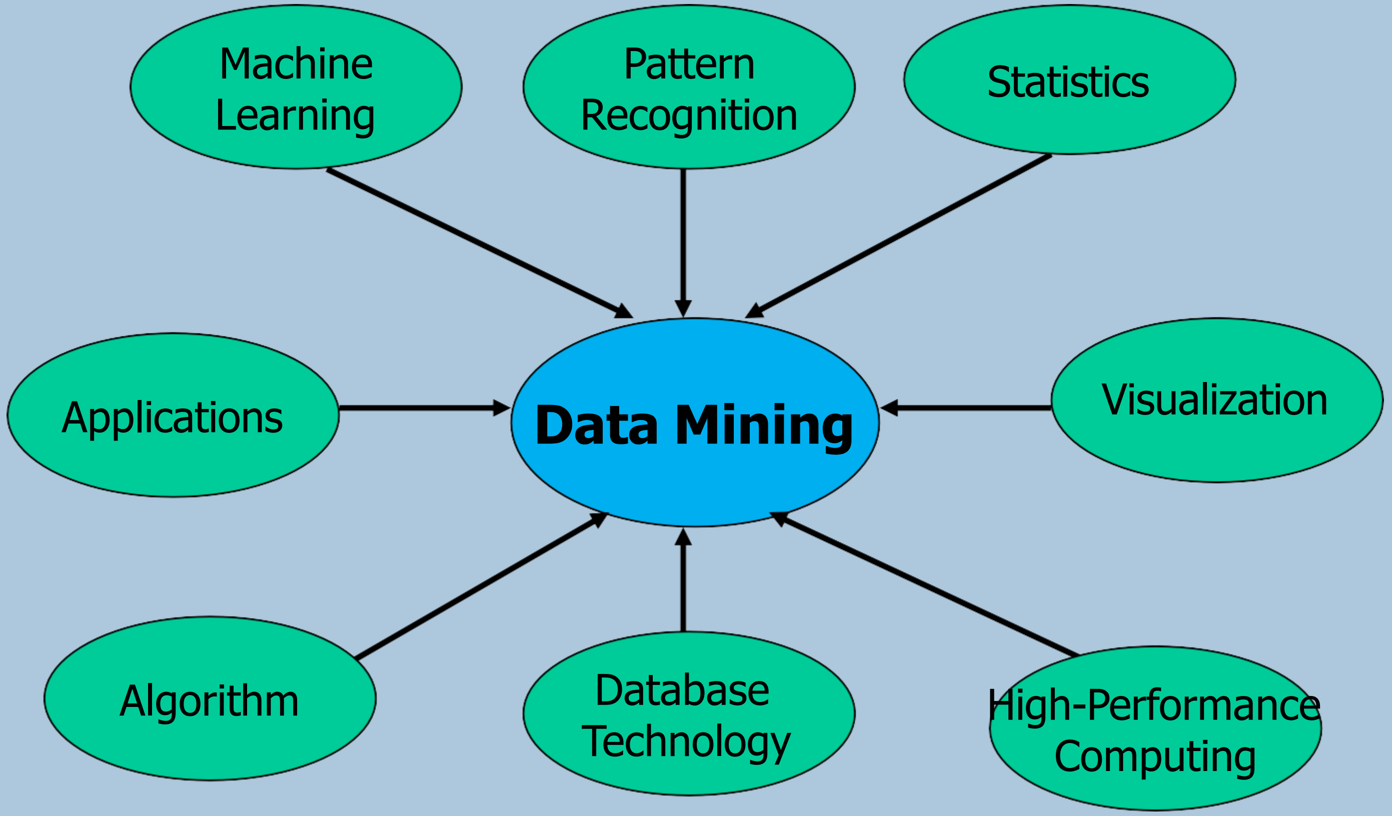
What is Data Mining?: Tasks 3

Discuss whether or not each of the following activities is a data mining task?

- **Finding the category (economy, sport, …) of a newspaper article.** **YES**
- **Deciding whether an email is a spam or not.** **YES**
- **Deciding whether an image contains an apple or not.** **YES**

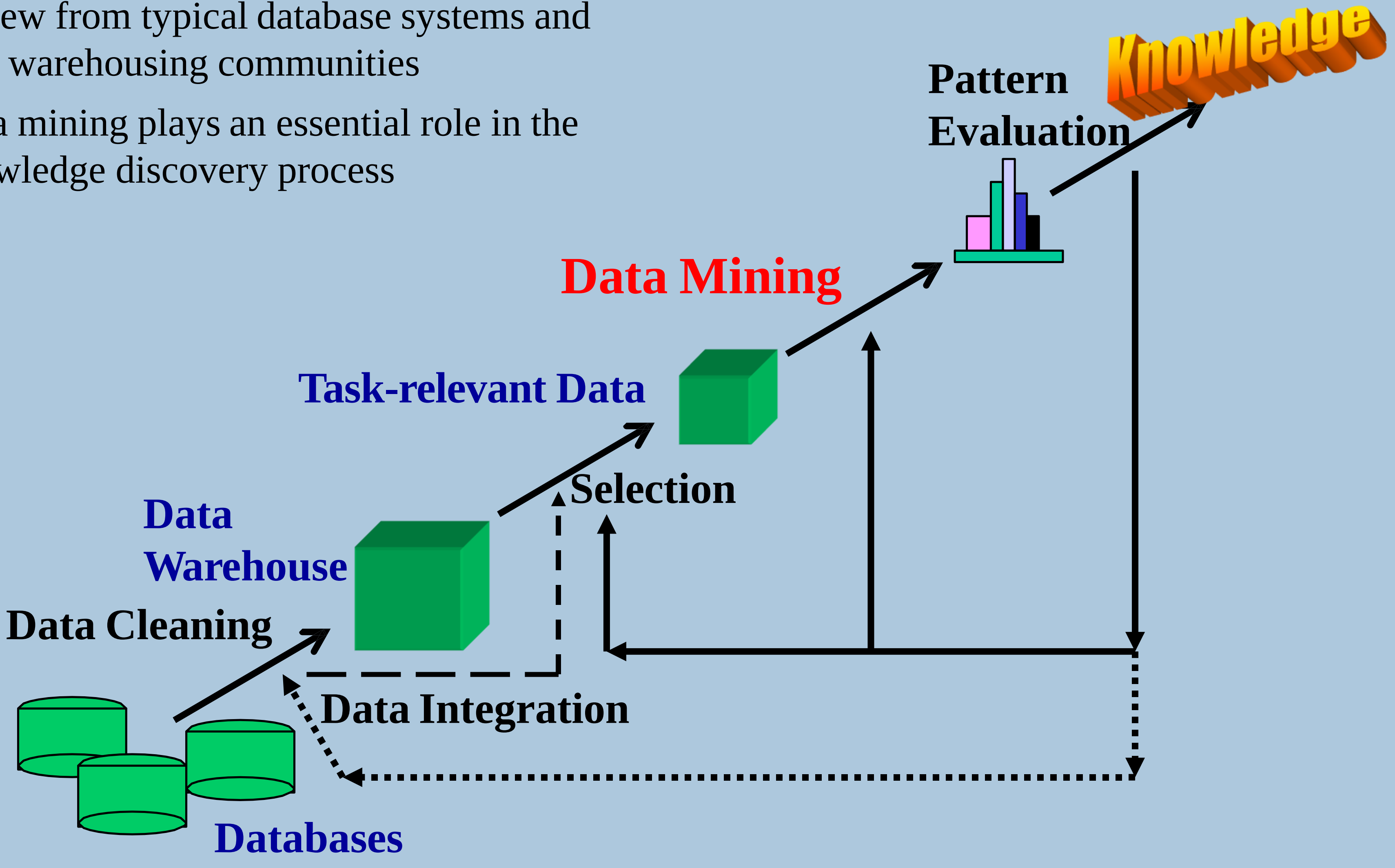
Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems, …

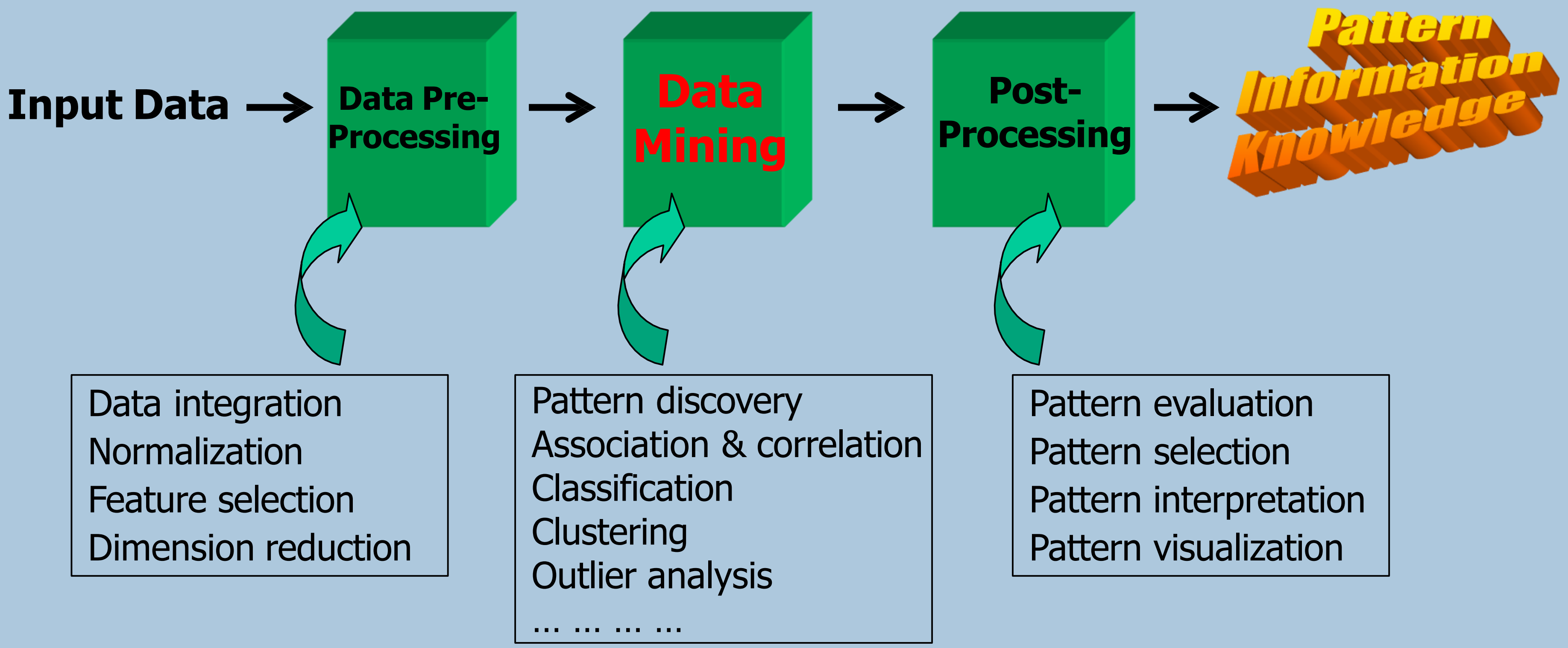


Knowledge Discovery (KDD) Process

- A view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



KDD Process:
A Typical View from ML and Statistics



Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

Diversity of Data Types for Data Mining (I)

- **Structured vs. unstructured data**
  - *Structured*: uniform, record- or table-like structures, defined by data dictionaries, with a fixed set of attributes, each with a fixed set of value ranges and semantic meaning
    - Ex. Data stored in *relational databases*, *data cubes*, *data matrices*, and many *data warehouses*
  - *Semi-structured*: allow a data object to contain a set value, a small set of heterogeneous typed values, or nested structures, or to allow the structure of objects or sub-objects to be defined flexibly and dynamically
  - Data having *certain structures* with clearly defined semantic meaning, such as *transactional data set, sequence data set* (e.g., time-series data, gene or protein data, or Weblog data)
  - *Graph or network data:* A more sophisticated type of semi-structured data set
  - *Unstructured data*: text data and multimedia (*e.g.*, audio, image, video) data
- The real-world data can often be a mixture of structured, semi-structured data and unstructured data

Diversity of Data Types for Data Mining (II)

- **Data associated with different applications**
  - Different applications: different data sets and require different data analysis methods
    - Sequence data: *Biological sequences* vs. *shopping transaction sequences*
    - *Time-series:* ordered set of numerical values with equal time interval
    - *Spatial, temporal and spatiotemporal data*
    - Graph and network data: Social networks, computer communication networks, biological networks, and information networks may carry rather different semantics

- **Stored vs. streaming data**
  - Stored data: Finite, stored in various kinds of large data repositories
  - Streaming data (e.g., video surveillance or remote sensing): Dynamic, constantly coming, infinite, real-time response—posing challenges on effective data mining

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
  - Object-relational databases, Heterogeneous databases
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and information networks
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

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**

– Classification	[Predictive]
– Clustering	[Descriptive]
– Association Rule Discovery	[Descriptive]
– Sequential Pattern Discovery	[Descriptive]
– Regression	[Predictive]
– Deviation Detection	[Predictive]

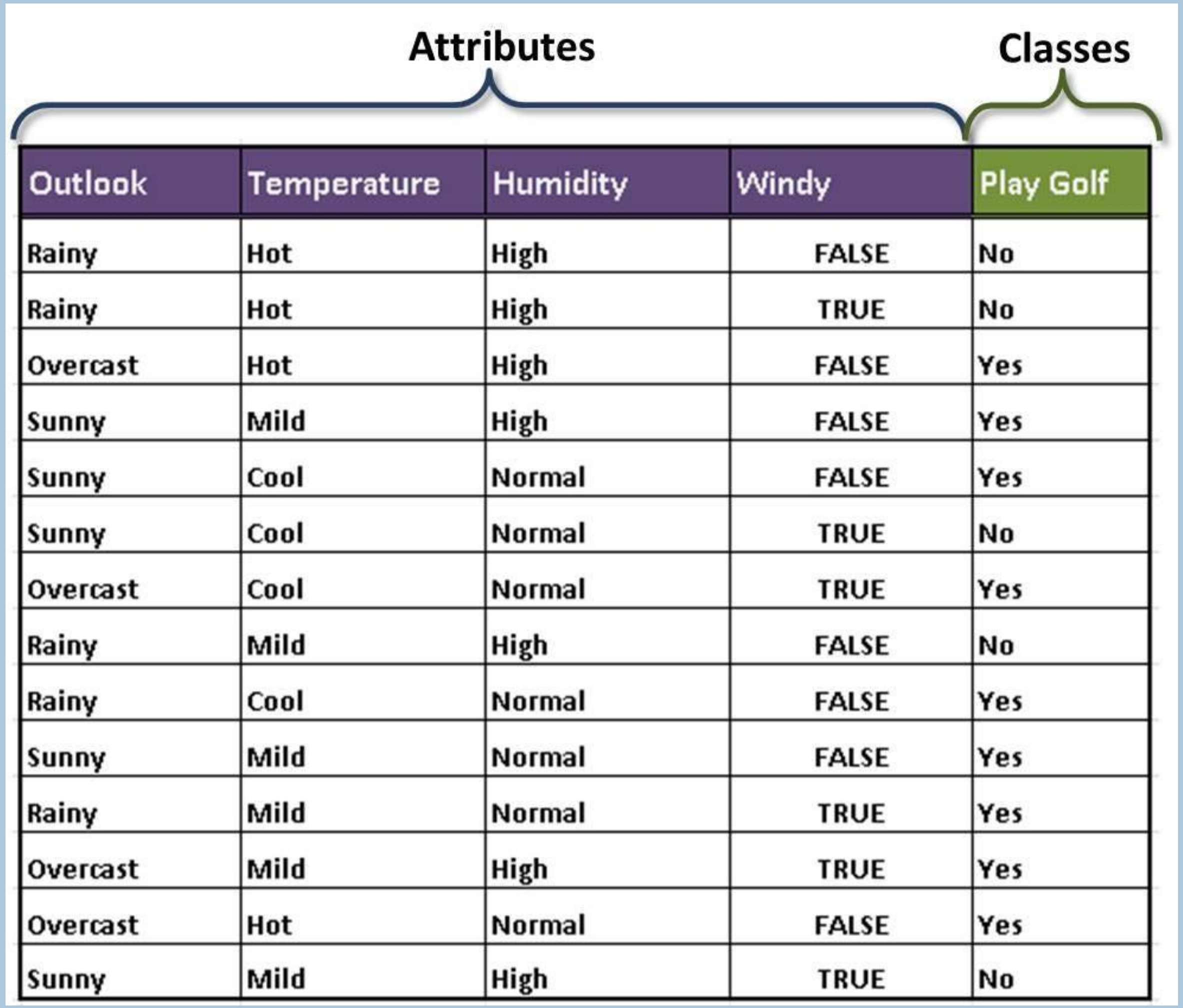
Classification

- 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.

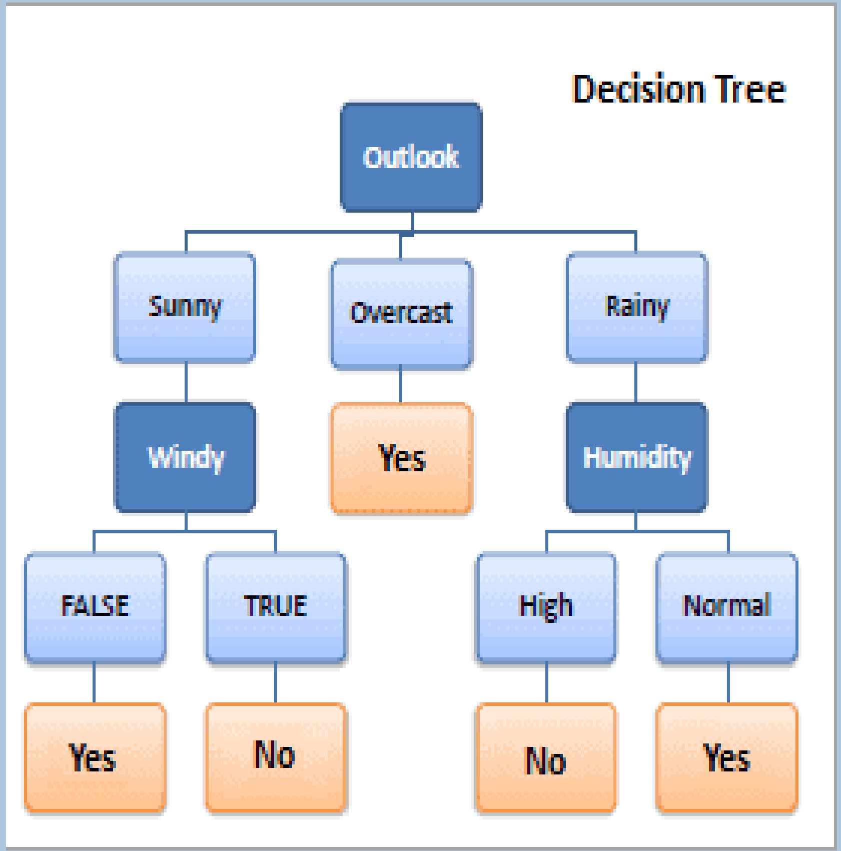
Classification: Typical Methods & Applications

- **Typical methods:**
  - Decision trees,
  - Naïve Bayesian classification,
  - support vector machines,
  - neural networks,
  - rule-based classification,
  - pattern-based classification, …
- **Typical applications:**
  - Credit card fraud detection, direct marketing,
  - Classifying diseases, web-pages,

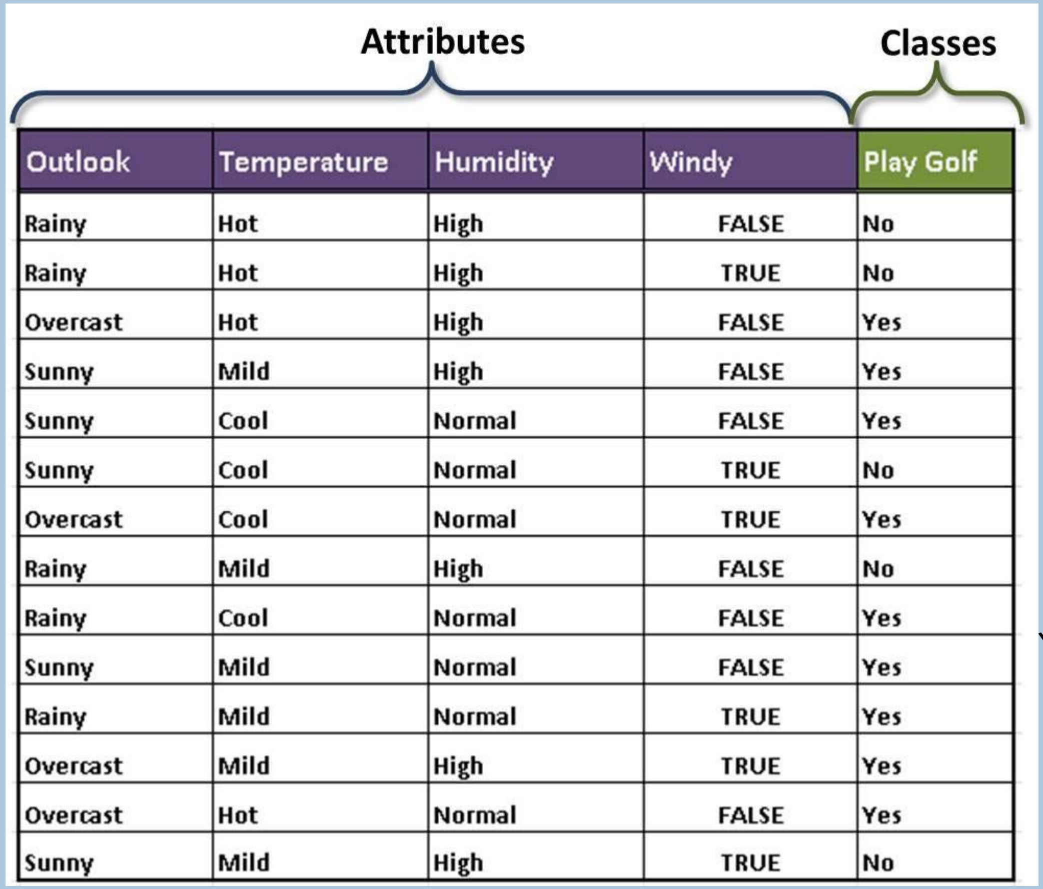
Classification: Typical Methods



Attributes				Classes
Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No



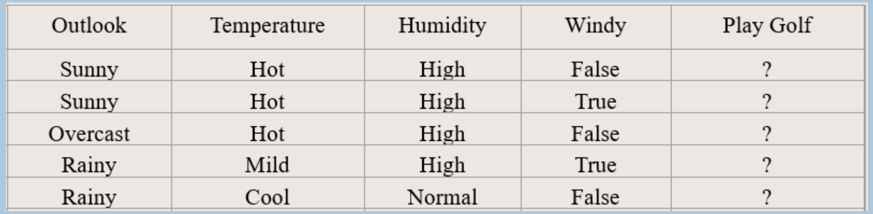
Classification: Example



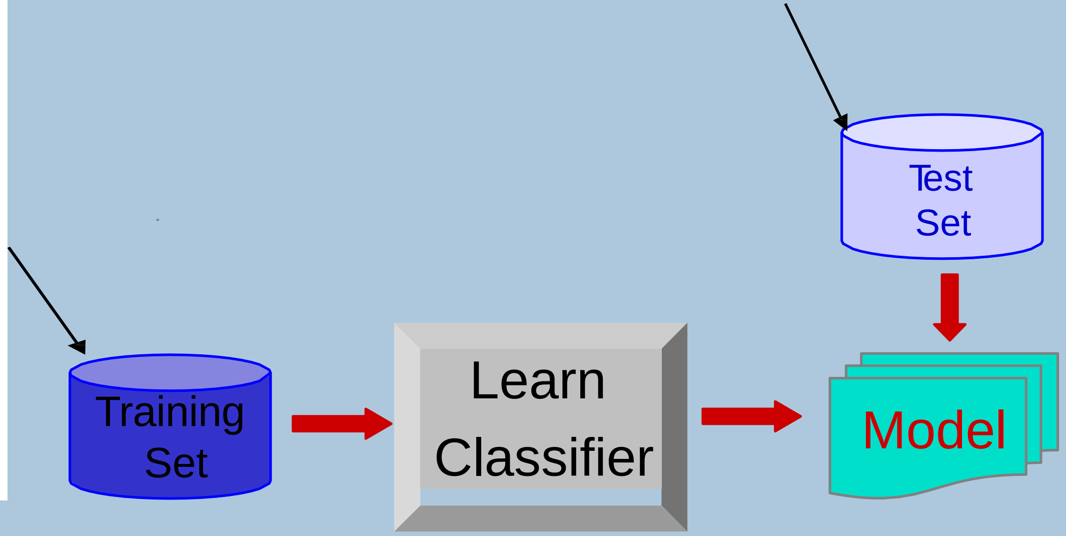
Attributes

Classes

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No



Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Hot	High	False	?
Sunny	Hot	High	True	?
Overcast	Hot	High	False	?
Rainy	Mild	High	True	?
Rainy	Cool	Normal	False	?



Classification: Application 1

Direct Marketing

Goal:

- Reduce the 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 – Example

Text Categorization

- How can we classify a given newspaper text as an ***economy***, ***sport*** or ***health*** article?

Classification – Example

Text Categorization

- How can we classify a given newspaper text as an ***economy***, ***sport*** or ***health*** article?
- **Create a training set.**
  - **Collect articles whose categories are known as economy, sport and health. For example, 100 articles from each category.**
- **Create a test set.**
  - **Similarly, collect articles whose categories are known as economy, sport and health to create a test set.**

Classification – Example

Text Categorization

- How can we classify a given newspaper text as an ***economy***, ***sport*** or ***health*** article?

- Create a training set.
  - Collect articles whose categories are known as economy, sport and health. For example, 100 articles from each category.
- Create a test set.
  - Similarly, collect articles whose categories are known as economy, sport and health to create a test set.

- **Determine the attributes:**
  - **Possible attributes are words appearing in texts (maybe not all words).**
  - **Words appearing more frequently in one category are good candidates as attributes.**

Classification – Example

Text Categorization

- How can we classify a given newspaper text as an ***economy***, ***sport*** or ***health*** article?

- Create a training set.
  - Collect articles whose categories are known as economy, sport and health. For example, 100 articles from each category.
- Create a test set.
  - Similarly, collect articles whose categories are known as economy, sport and health to create a test set.
- Determine the attributes:
  - Possible attributes are words appearing in texts (maybe not all words).
  - Words appearing more frequently in one category are good candidates as attributes.

- **Decide the classification method: decision tree, naïve Bayes, svm, …**

- **Create the model using the selected classification method.**

Classification – Example

Text Categorization

- How can we classify a given newspaper text as an ***economy***, ***sport*** or ***health*** article?
- Create a training set.
  - Collect articles whose categories are known as economy, sport and health. For example, 100 articles from each category.
- Create a test set.
  - Similarly, collect articles whose categories are known as economy, sport and health to create a test set.
- Determine the attributes:
  - Possible attributes are words appearing in texts (maybe not all words).
  - Words appearing more frequently in one category are good candidates as attributes.
- Decide the classification method: decision tree, naïve Bayes, svm, …
- Create the model using the selected classification method.
- **Test the accuracy of the created model.**

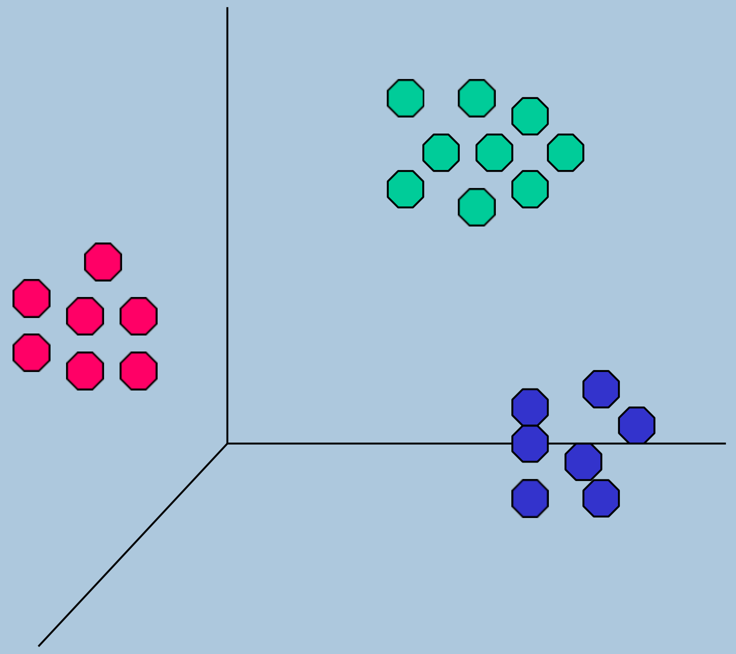
Clustering

- **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 each 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 Similarity Measures: Problem-specific measures.
- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters),
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Euclidean Distance Based Clustering in 3-D space

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application

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.

Association Rule Discovery

- **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.**

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in a store?
- A typical association rule
  - Bread → Peanut Butter [10%, 50%] (support, confidence)

    Support: This measures how often the rule is applicable to the dataset. For instance, if 100 out of 1000 transactions contain both bread and peanut butter, the support for the rule (Bread => Peanut Butter) is 0.1 or 10%.

    Confidence: This measures how often items in the consequent appear in transactions that contain the antecedent. If 100 out of 200 transactions that contain bread also contain peanut butter, the confidence for the rule is 0.5 or 50%.
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Association Rule Discovery: Application 1

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 bread and banana, then he is very likely to buy peanut butter.

Association Rule Discovery: Application 2

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.

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
- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception?
    - One person’s garbage could be another person’s treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis