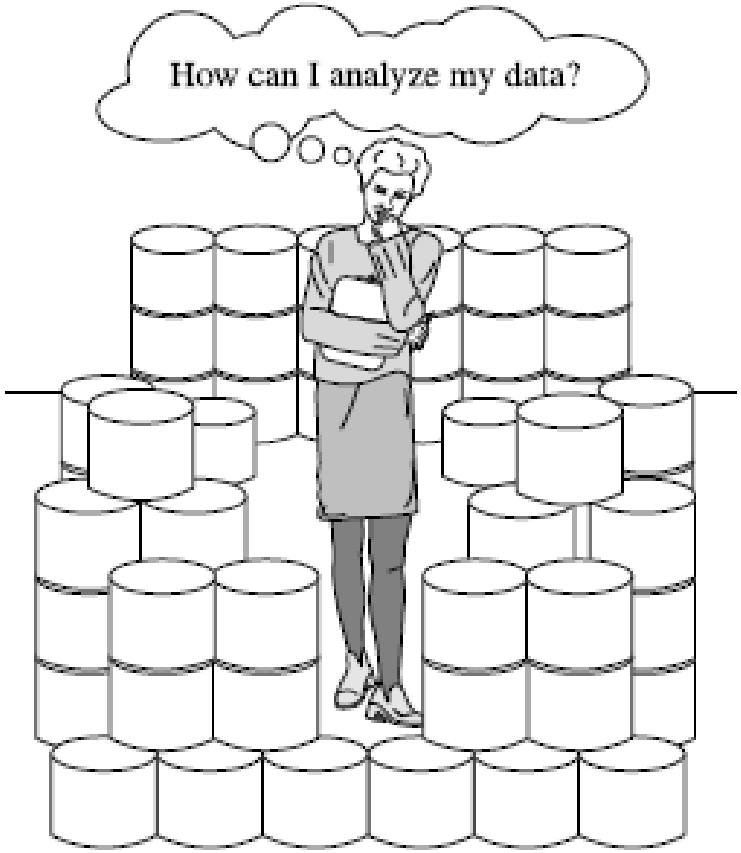
Data Mining

Introduction to Data Mining

Why do we need Data Mining?

- Explosive Growth of Data
  - Amount of data collected increasing exponentially
  - Automated data collection tools, database systems, Web, computerized society
  - *Business*: Web, e-commerce, transactions, stocks, purchases at grocery stores, Bank/Credit Card transactions
  - *Science*: remote sensors on a satellite, telescopes scanning skies, microarrays generating gene expression data,
  - *Society and everyone*: news, digital cameras, YouTube. Facebook, web
- **We have mass amount of data, but we need knowledge.**
- There is often **information (knowledge)** "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

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 “Amazon
- 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.**

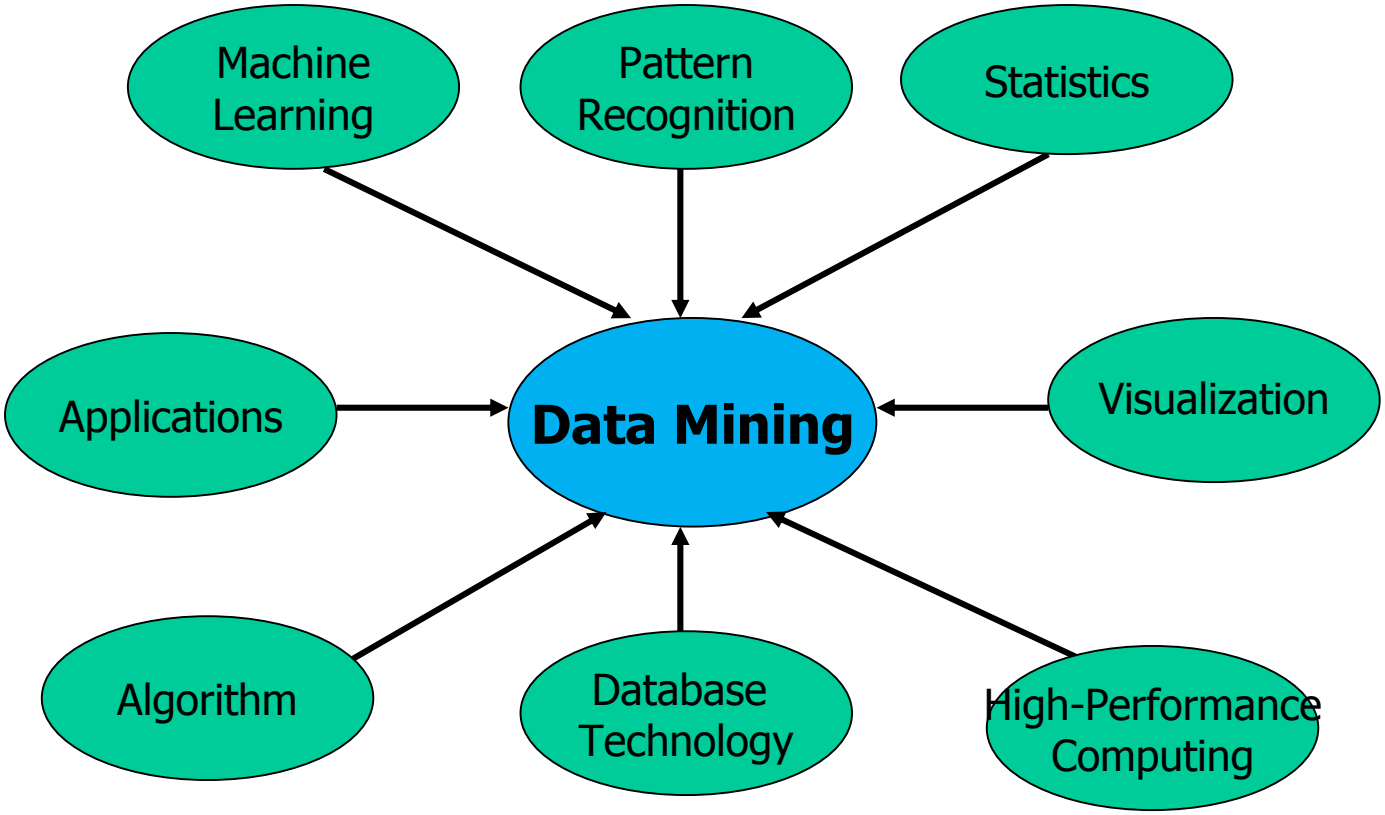
What is Data Mining?: Tasks 3

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**

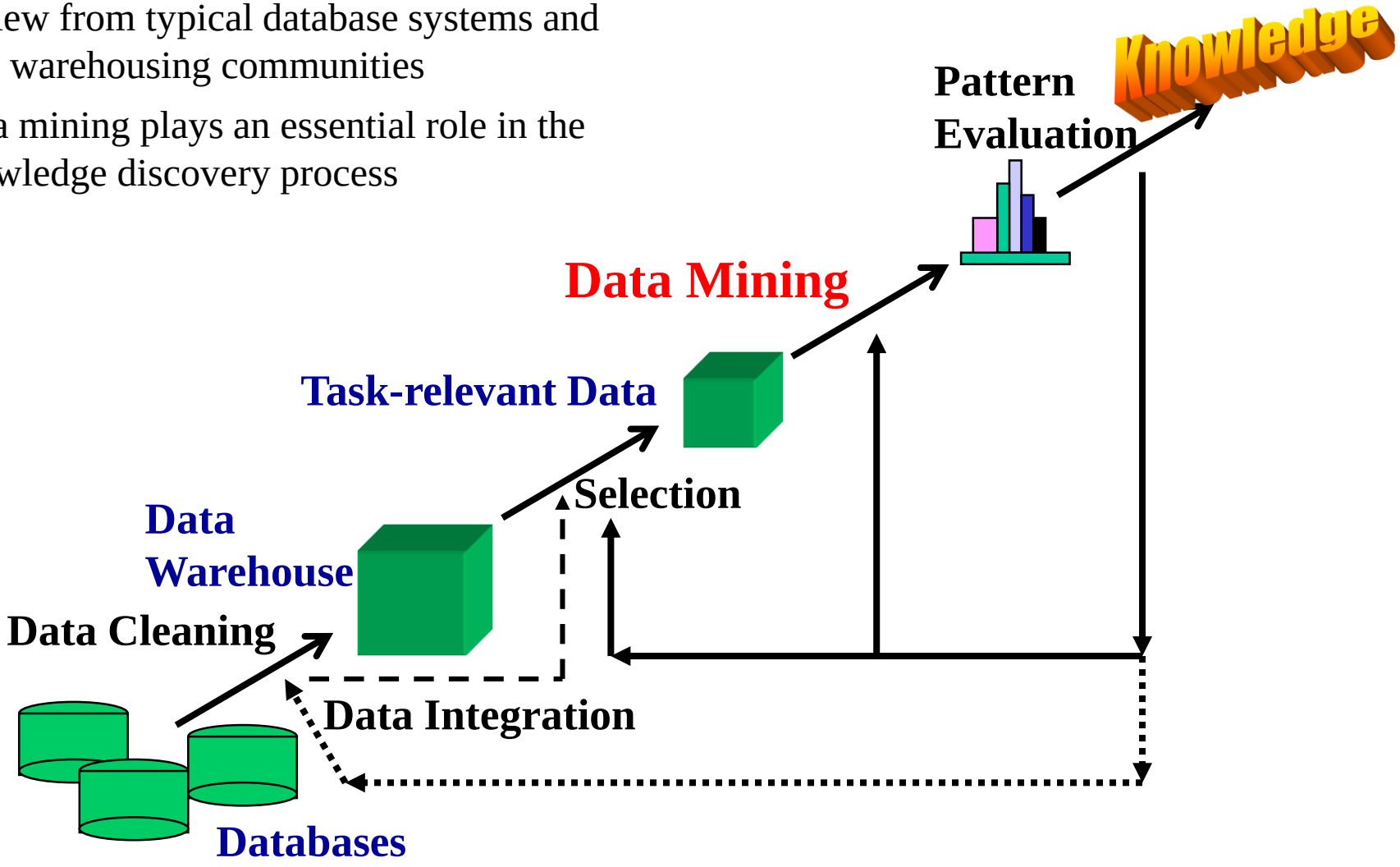
Origins of Data Mining

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

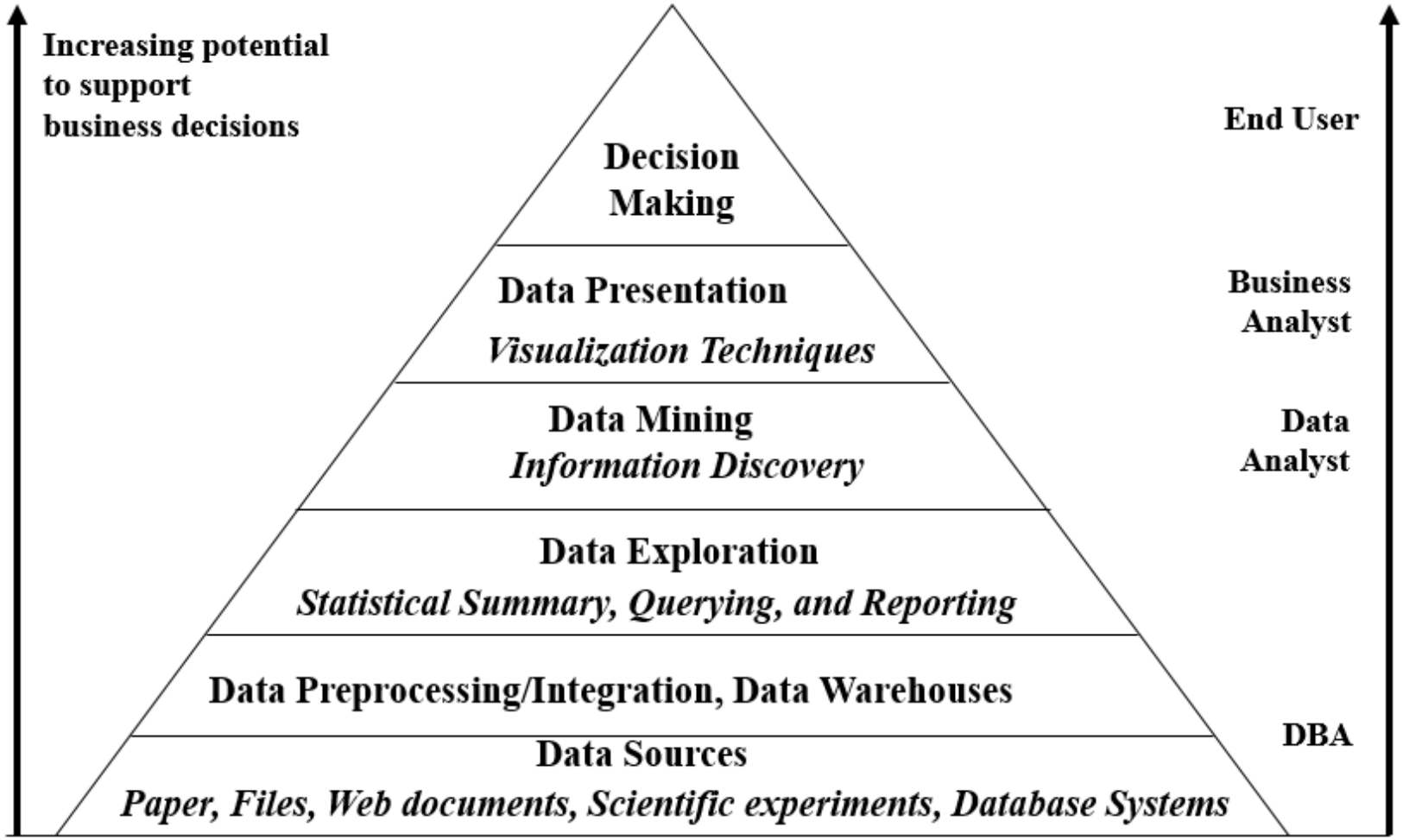


Knowledge Discovery (KDD) Process

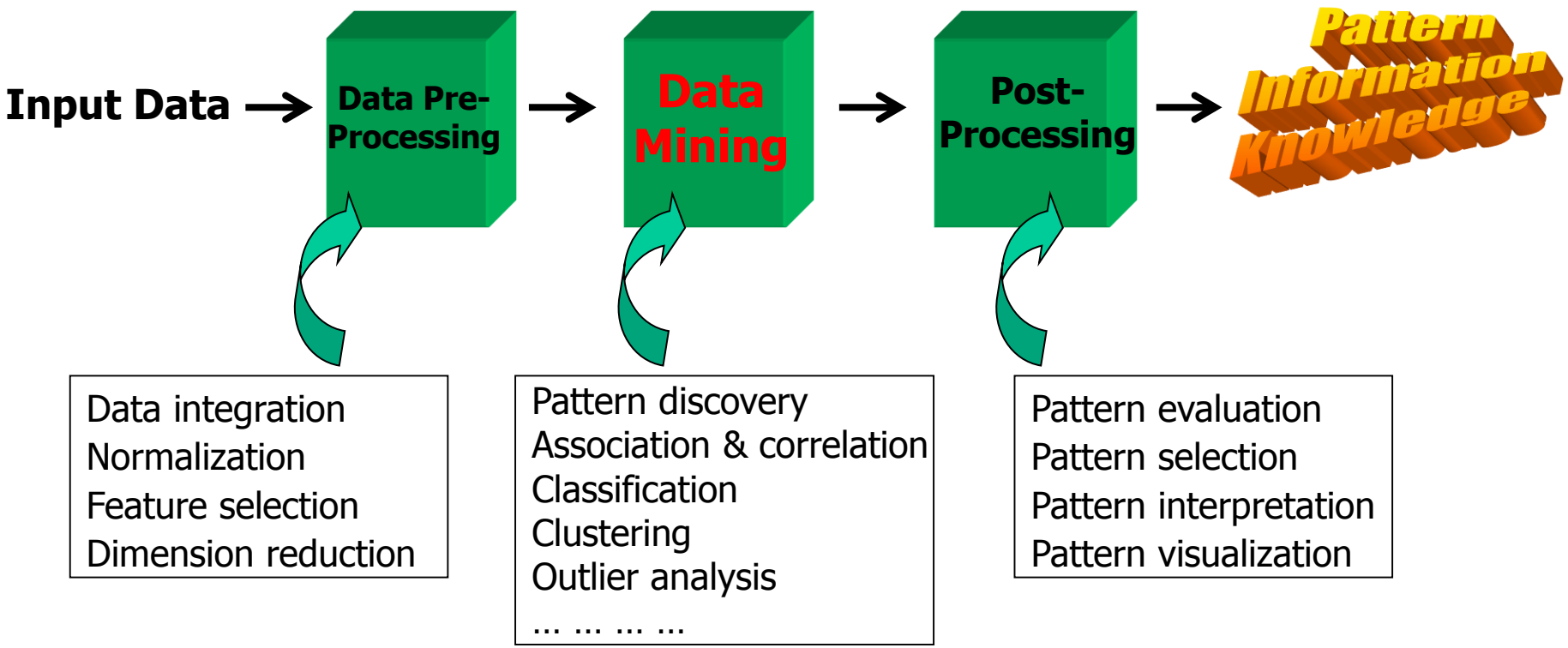
- A view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



KDD Process:
A Typical View from ML and Statistics



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.

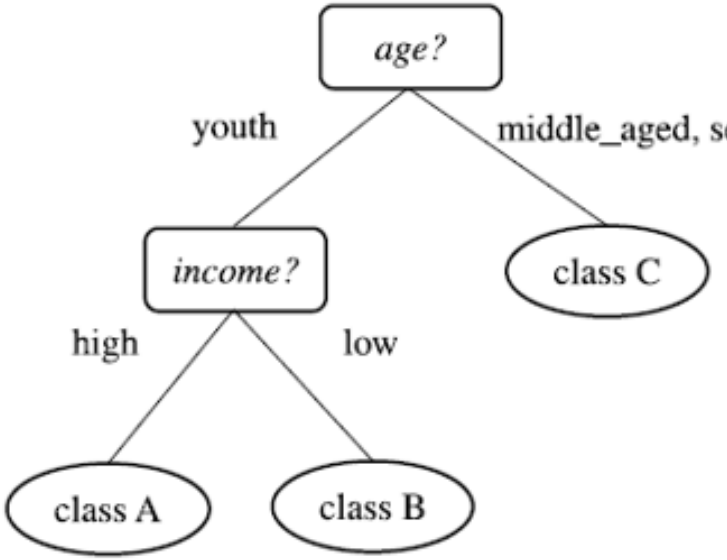
Classification: Typical Methods & Applications

- **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,

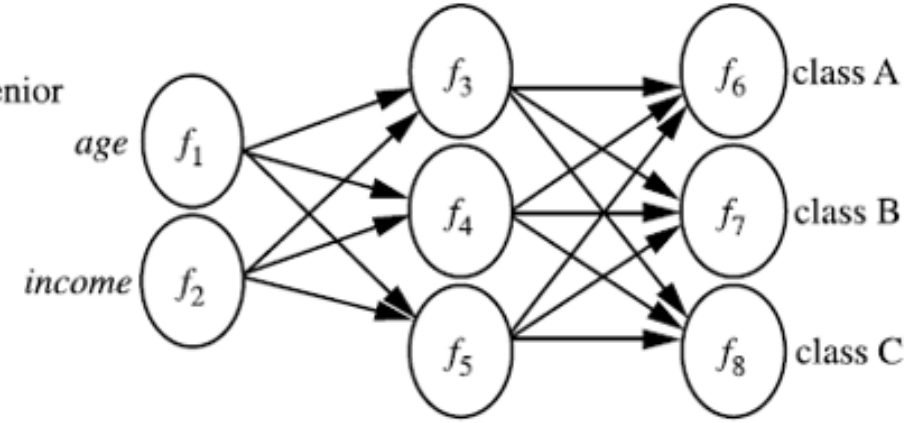
Classification: Typical Methods

age(X, "youth") AND income(X, "high") → class(X, "A")
age(X, "youth") AND income(X, "low") → class(X, "B")
age(X, "middle_aged") → class(X, "C")
age(X, "senior") → class(X, "C")

IF-THEN Rules



Decision Tree



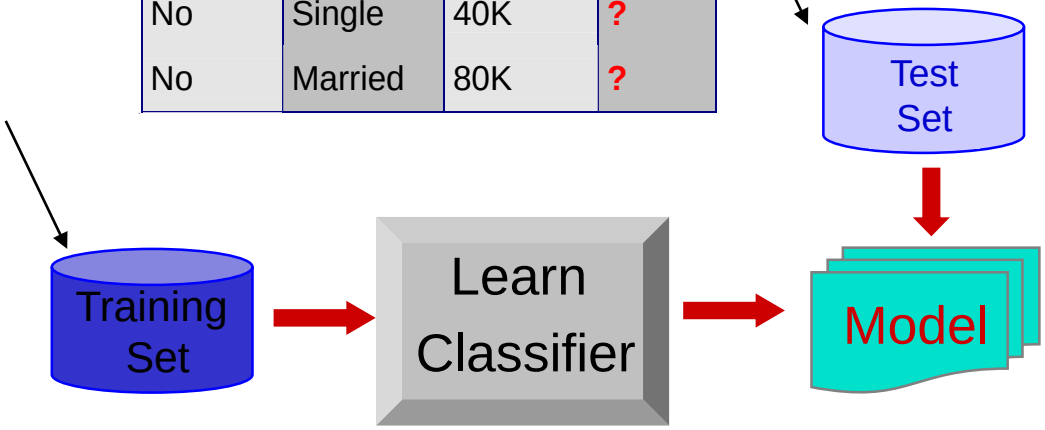
Neural Network

Classification: Example

Tid	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

categorical categorical class

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 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.**

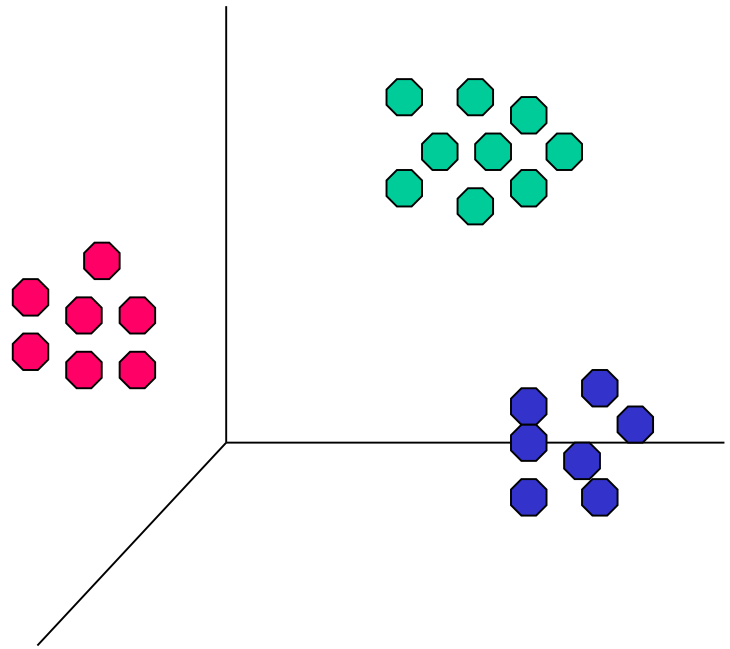
Clustering

- **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
  - Diaper → Beer [0.5%, 75%] (support, confidence)
- 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 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 2

Inventory Management

Goal:

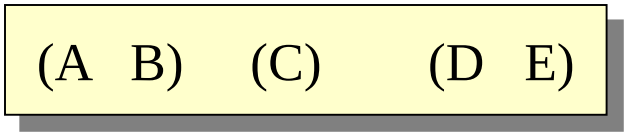
– A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.

Approach:

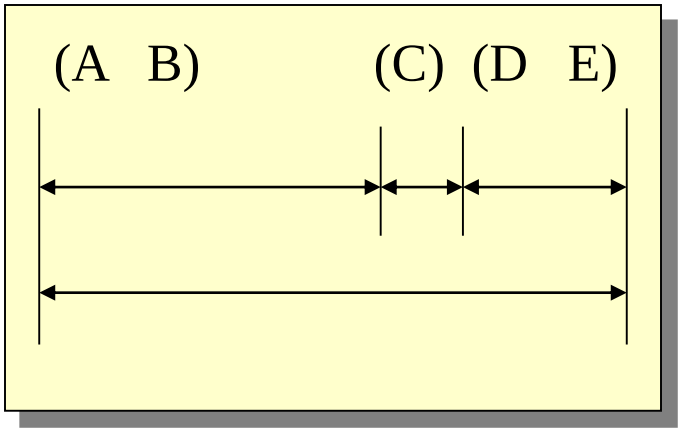
– Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery

- 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.



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



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
- 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

Structure and Network Analysis

- **Graph mining**
  - **Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)**
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates,
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

- **Are all mined knowledge interesting?**
  - One can mine tremendous amount of "patterns"
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness

Major Issues in Data Mining

- **Mining Methodology**
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- **User Interaction**
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

Major Issues in Data Mining

- **Efficiency and Scalability**
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- **Diversity of data types**
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- **Data mining and society**
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Summary – Data Mining Overview

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

Software

- The software required for this course (Weka and R/RStudio) are open source and freely available under GNU General Public License.

- Depending on your system, you can download and install Weka 3.8 which is the lasted stable version of Weka available at http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

- In addition, most recent version of R and RStudio (IDE for R) are available at https://cran.rstudio.com/ and https://www.rstudio.com/products/rstudio/download/, respectively.

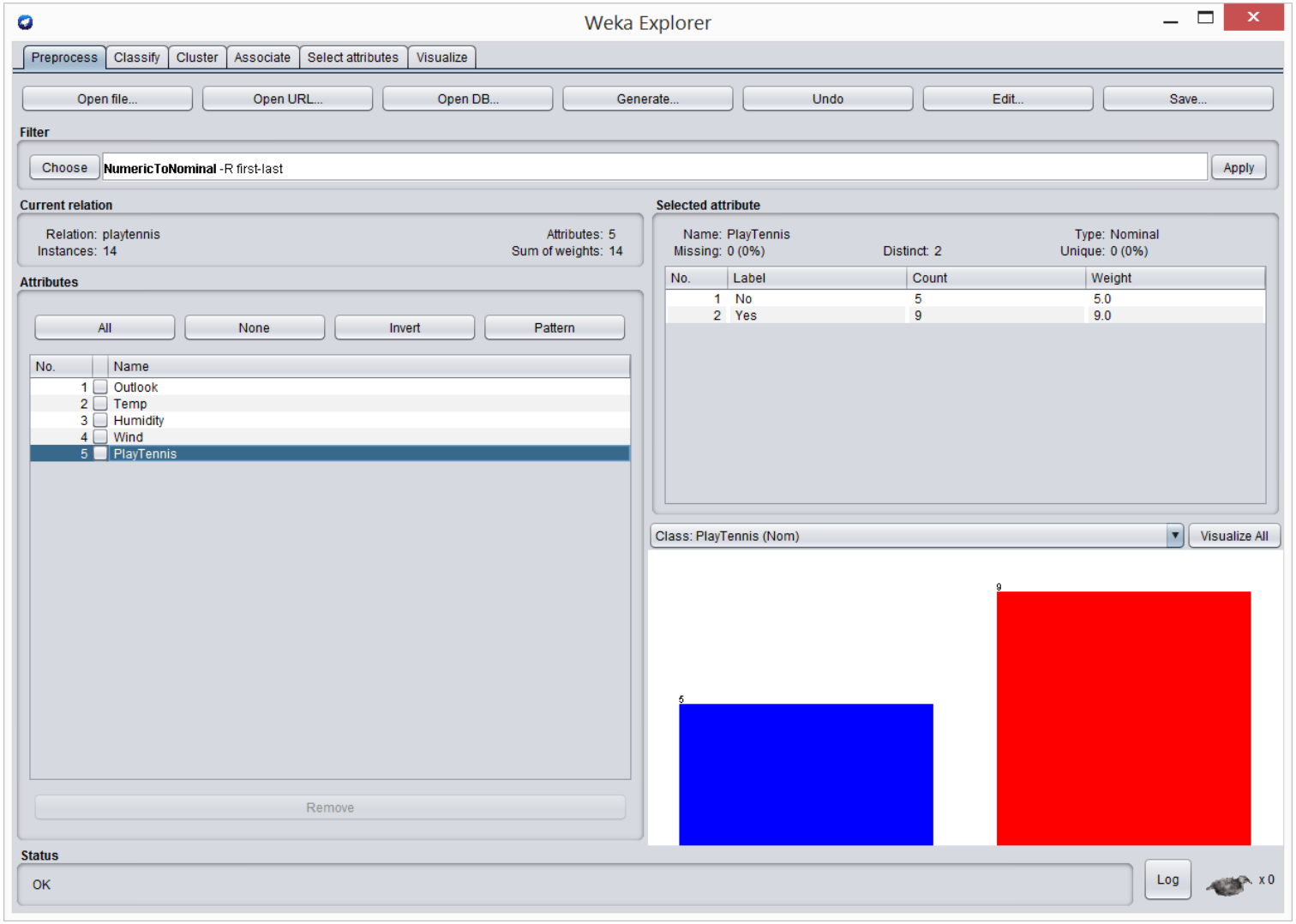
Weka

A simple training set (.csv file) holding our data.

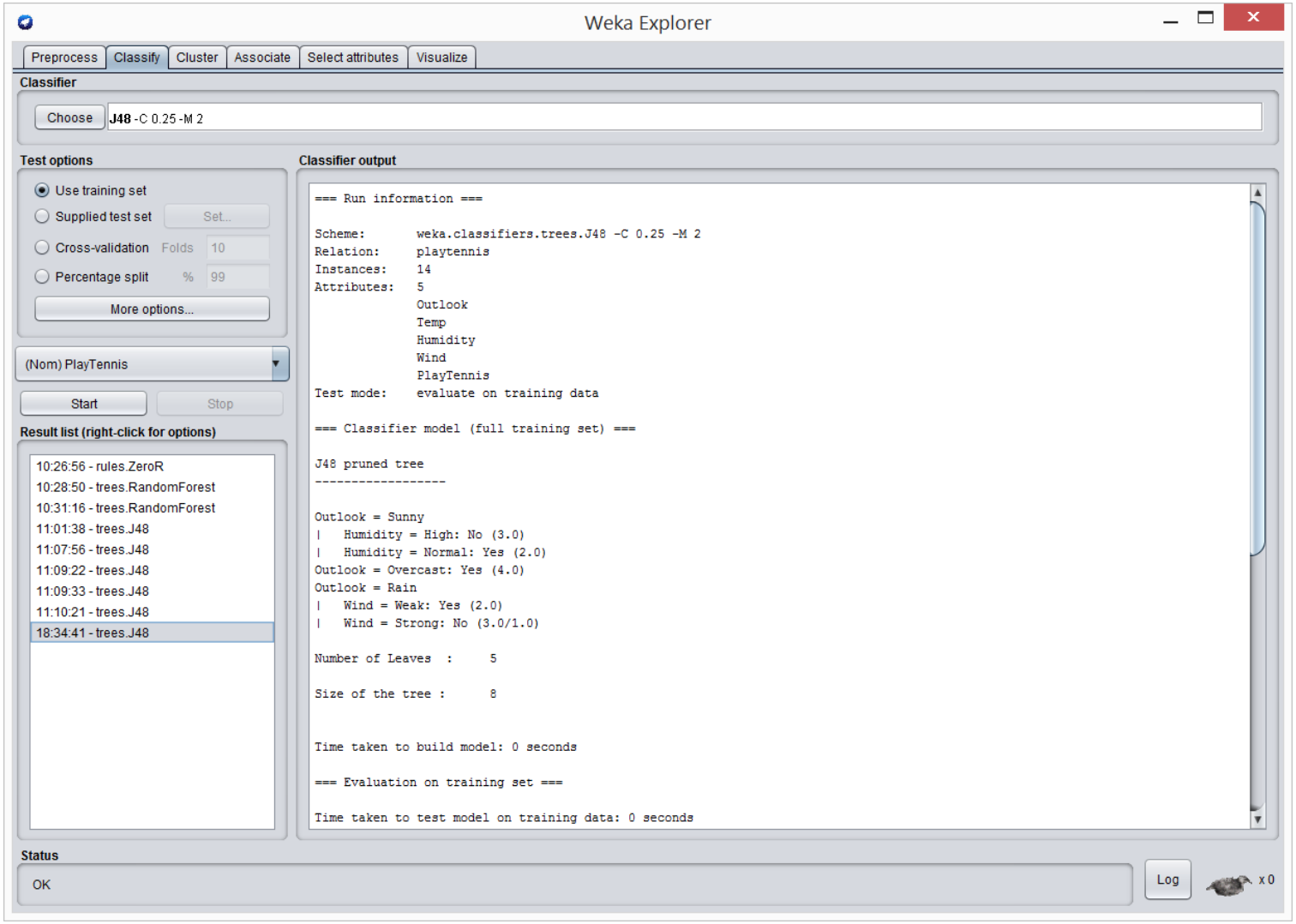
This file can be loaded by Weka.

Outlook	Temp	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Weak	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Strong	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Weka


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NumericToNominal -R first-last
Apply
Current relation
Relation: playtennis
Instances: 14
Attributes: 5
Sum of weights: 14
Selected attribute
Name: PlayTennis
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No. Label Count Weight
1 No 5 5.0
2 Yes 9 9.0
Attributes
All
None
Invert
Pattern
No. Name
1 Outlook
2 Temp
3 Humidity
4 Wind
5 PlayTennis
Remove
Class: PlayTennis (Nom)
Visualize All
9
5
Status
OK
Log
x 0

Weka


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
99
More options...
(Nom) PlayTennis
Start
Stop
Result list (right-click for options)
10:26:56 - rules.ZeroR
10:28:50 - trees.RandomForest
10:31:16 - trees.RandomForest
11:01:38 - trees.J48
11:07:56 - trees.J48
11:09:22 - trees.J48
11:09:33 - trees.J48
11:10:21 - trees.J48
18:34:41 - trees.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: playtennis
Instances: 14
Attributes: 5
Outlook
Temp
Humidity
Wind
PlayTennis
Test mode: evaluate on training data
=== Classifier model (full training set) ===
J48 pruned tree
Outlook = Sunny
| Humidity = High: No (3.0)
| Humidity = Normal: Yes (2.0)
Outlook = Overcast: Yes (4.0)
Outlook = Rain
| Wind = Weak: Yes (2.0)
| Wind = Strong: No (3.0/1.0)
Number of Leaves : 5
Size of the tree : 8
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
Status
OK
Log
x 0

Reference Books

- *Introduction to Data Mining*, PangNing Tan, Michael Steinbach, Vipin Kumar, 3rd Edition, Pearson, 2014.
- *Data Mining: Concepts and Techniques*, Jiawei Han and Micheline Kamber, Morgan Kaufmann, 2012.
- *Data Mining: Practical Machine Learning Tools and Techniques*, Ian H. Witten, Eibe Frank and Mark A. Hall, 4th Edition, Morgan Kaufmann, 2017.
- *Mining of Massive Datasets*, Anand Rajaraman & Jeffrey D. Ullman, 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009
- B. Liu, Web Data Mining, Springer 2006