UNIVERSITÀ DI PISA

Dottorato di Ricerca in Ingegneria dell'Informazione

# Data Mining and Soft Computing

## Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI$^2$S)

Dept. of Computer Science and A.I.

University of Granada, Spain

Email: herrera@decsai.ugr.es

http://sci2s.ugr.es

http://decsai.ugr.es/~herrera

# Summary

1. **Introduction to Data Mining and Knowledge Discovery**
2. **Data Preparation**
3. **Introduction to Prediction, Classification, Clustering and Association**
4. **Data Mining - From the Top 10 Algorithms to the New Challenges**
5. **Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation**
6. **Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning**
7. **Genetic Fuzzy Systems: State of the Art and New Trends**
8. **Some Advanced Topics I: Classification with Imbalanced Data Sets**
9. **Some Advanced Topics II: Subgroup Discovery**
10. **Some advanced Topics III: Data Complexity**
11. **Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.**

# Slides used for preparing this talk:

CS490D:
Introduction to Data Mining
*Prof. Chris Clifton*

Association Analysis: Basic Concepts and Algorithms

Lecture Notes for Chapter 6

Introduction to Data Mining

by Tan, Steinbach, Kumar

*DATA MINING*
*Introductory and Advanced Topics*
Margaret H. Dunham

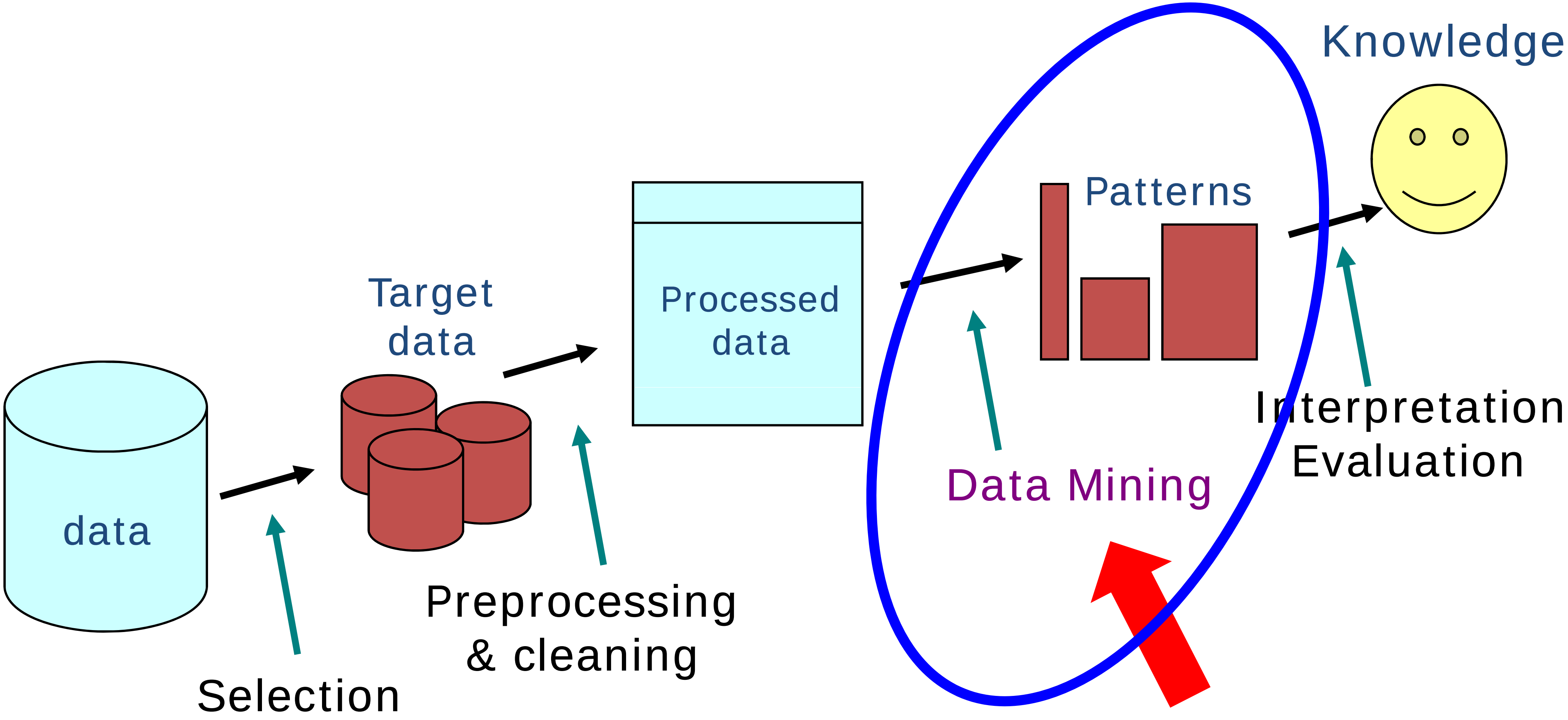
# Outline

- ✓Introduction
- ✓Classification
- ✓Prediction
- ✓ Clustering
- ✓Association
- ✓ Data Mining Systems / Data Set Repositories
- ✓Concluding Remarks

# Outline

- ✓ Introduction
- ✓ Classification
- ✓ Prediction
- ✓ Clustering
- ✓ Association
- ✓ Data Mining Systems / Data Set Repositories
- ✓ Concluding Remarks

# Introduction



# What Is The Input?

- Concepts
- Instances/Examples
- Attributes
  nominal v.s. numeric attributes
- Preparing inputs

# What to do in data mining

- **Classification**
  Find the class a new instance belong to
  e.g. whether a cell is a normal cell or a cancerous cell

- **Numeric prediction**
  Variation of classification where the output is numeric classes
  e.g. frequency of cancerous cell found

# What to do (contd.)

- **Clustering**

  Process to cluster/group the instances into classes ➔ before existence of any classes
  e.g. deriving/classify a new disease into different possible types/groups

- **Association**

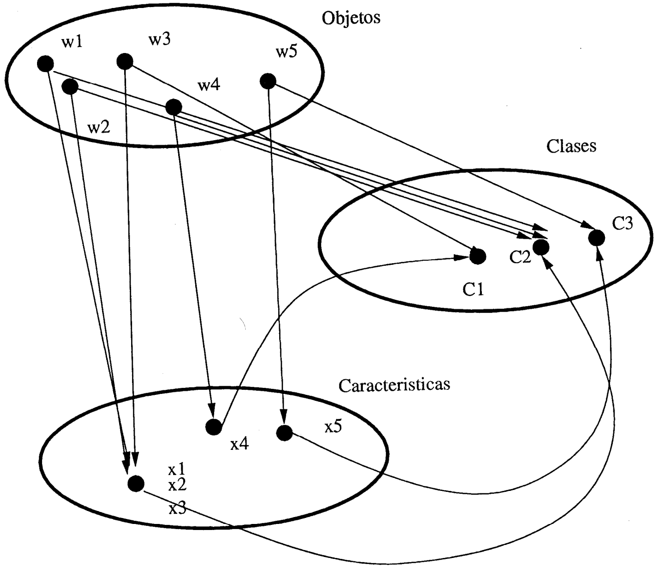
Finding rules/conclusions among attributes
e.g. a high-blood-pressure patient is most likely to have heart-attack disease

# Outline

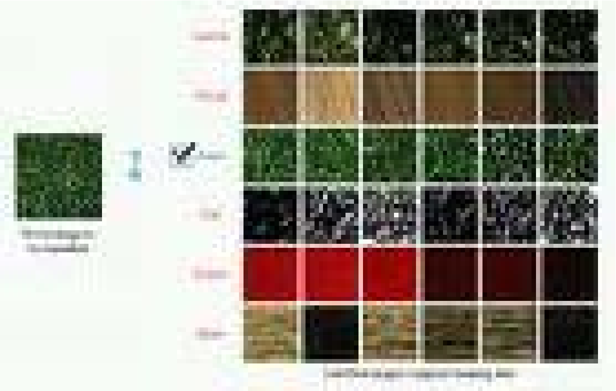
- ✓Introduction
- ✓Classification
- ✓Prediction
- ✓ Clustering
- ✓Association
- ✓ Data Mining Systems / Data Set Repositories
- ✓Concluding Remarks

# Classification Problem

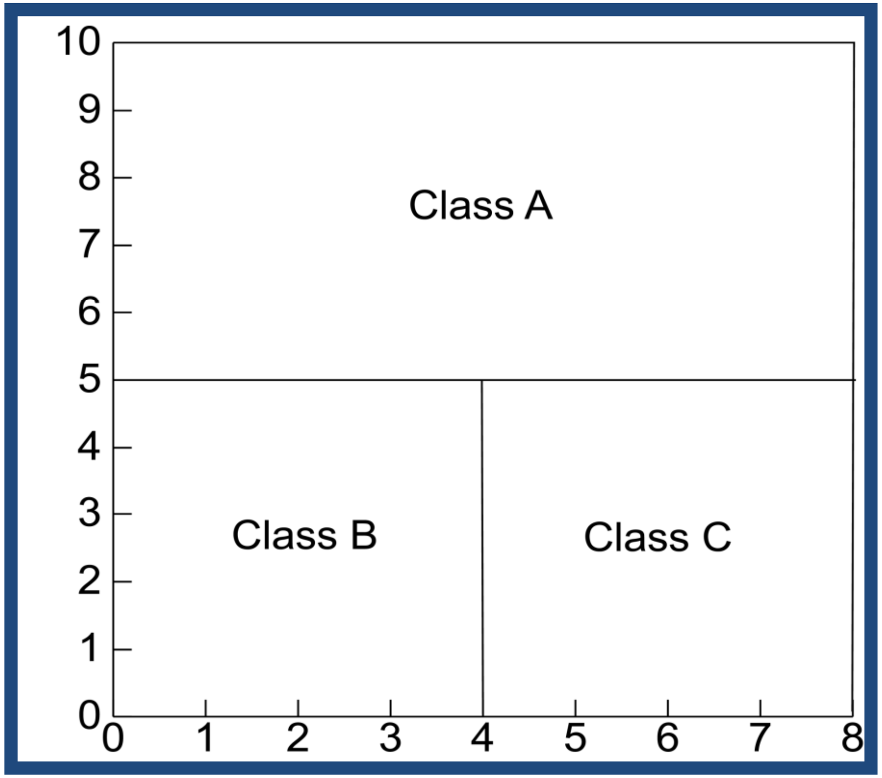
- Given a database $D=\{t_1,t_2,\ldots,t_n\}$ and a set of classes $C=\{C_1,\ldots,C_m\}$, the ***Classification Problem*** is to define a mapping $f:D \rightarrow C$ where each $t_i$ is assigned to one class.



- ***Prediction*** is similar, but may be viewed as having infinite number of classes.

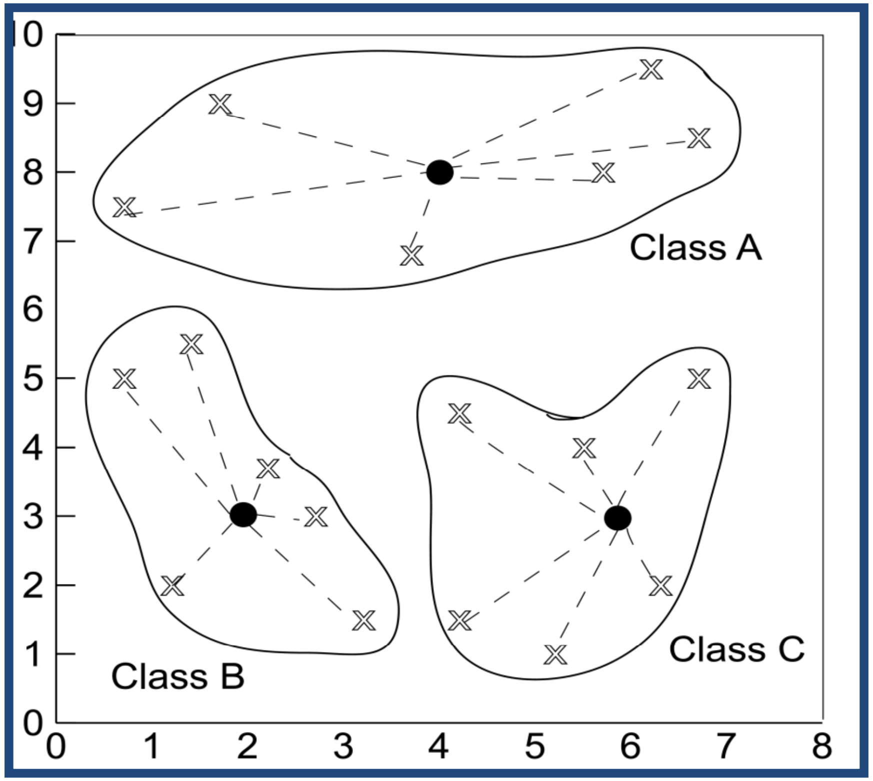


# Defining Classes



Partitioning Based

Distance Based



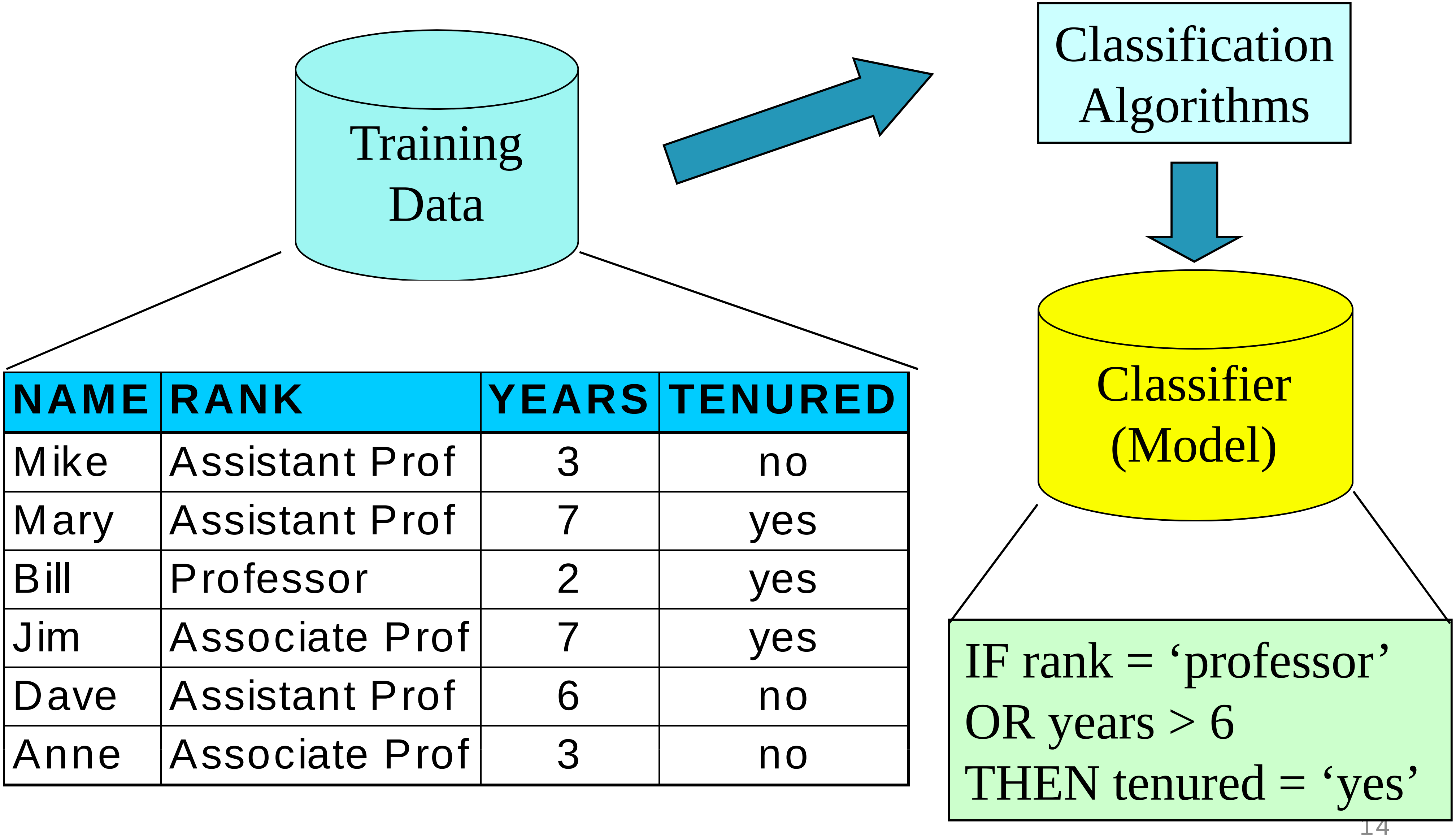
# Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae

# Classification—A Two-Step Process

- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

# Classification Process (1): Model Construction



Training Data

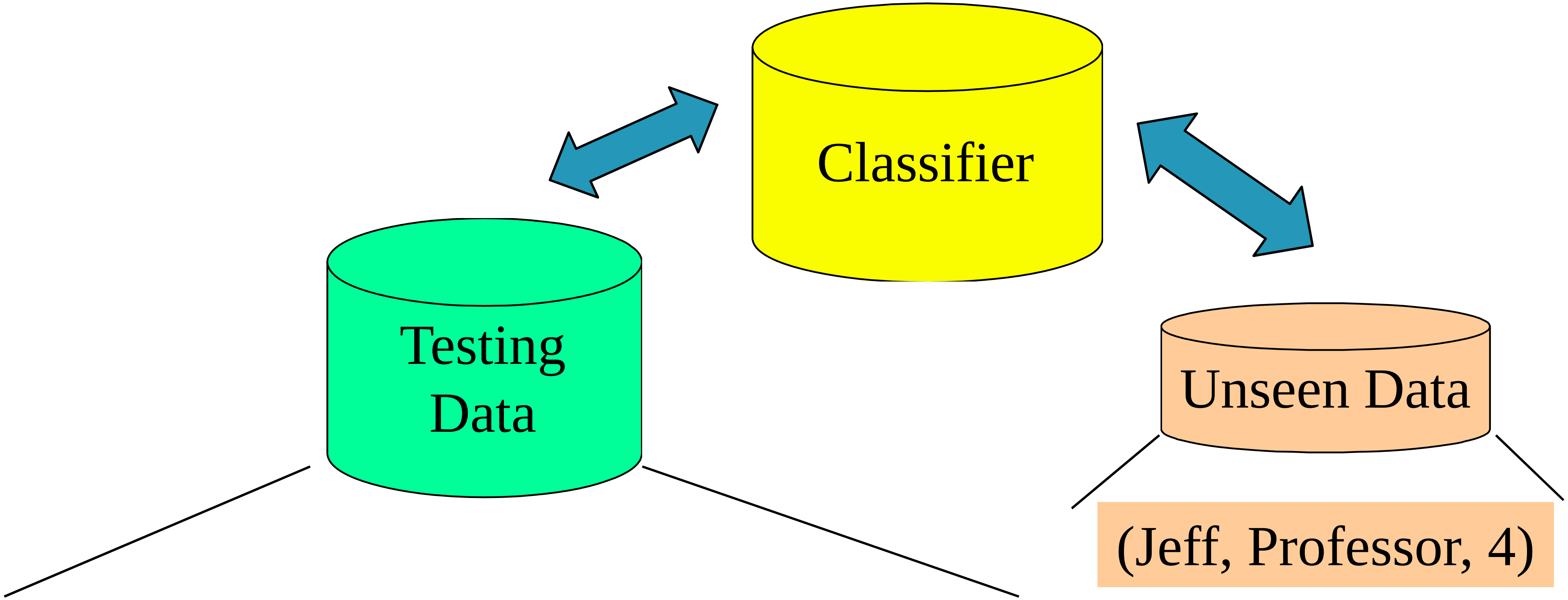
Classification Algorithms

Classifier (Model)

| NAME | RANK | YEARS | TENURED |
|---|---|---|---|
| Mike | Assistant Prof | 3 | no |
| Mary | Assistant Prof | 7 | yes |
| Bill | Professor | 2 | yes |
| Jim | Associate Prof | 7 | yes |
| Dave | Assistant Prof | 6 | no |
| Anne | Associate Prof | 3 | no |

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

# Classification Process (2): Use the Model in Prediction



| NAME | RANK | YEARS | TENURED |
|---|---|---|---|
| Tom | Assistant Prof | 2 | no |
| Merlisa | Associate Prof | 7 | no |
| George | Professor | 5 | yes |
| Joseph | Assistant Prof | 7 | yes |

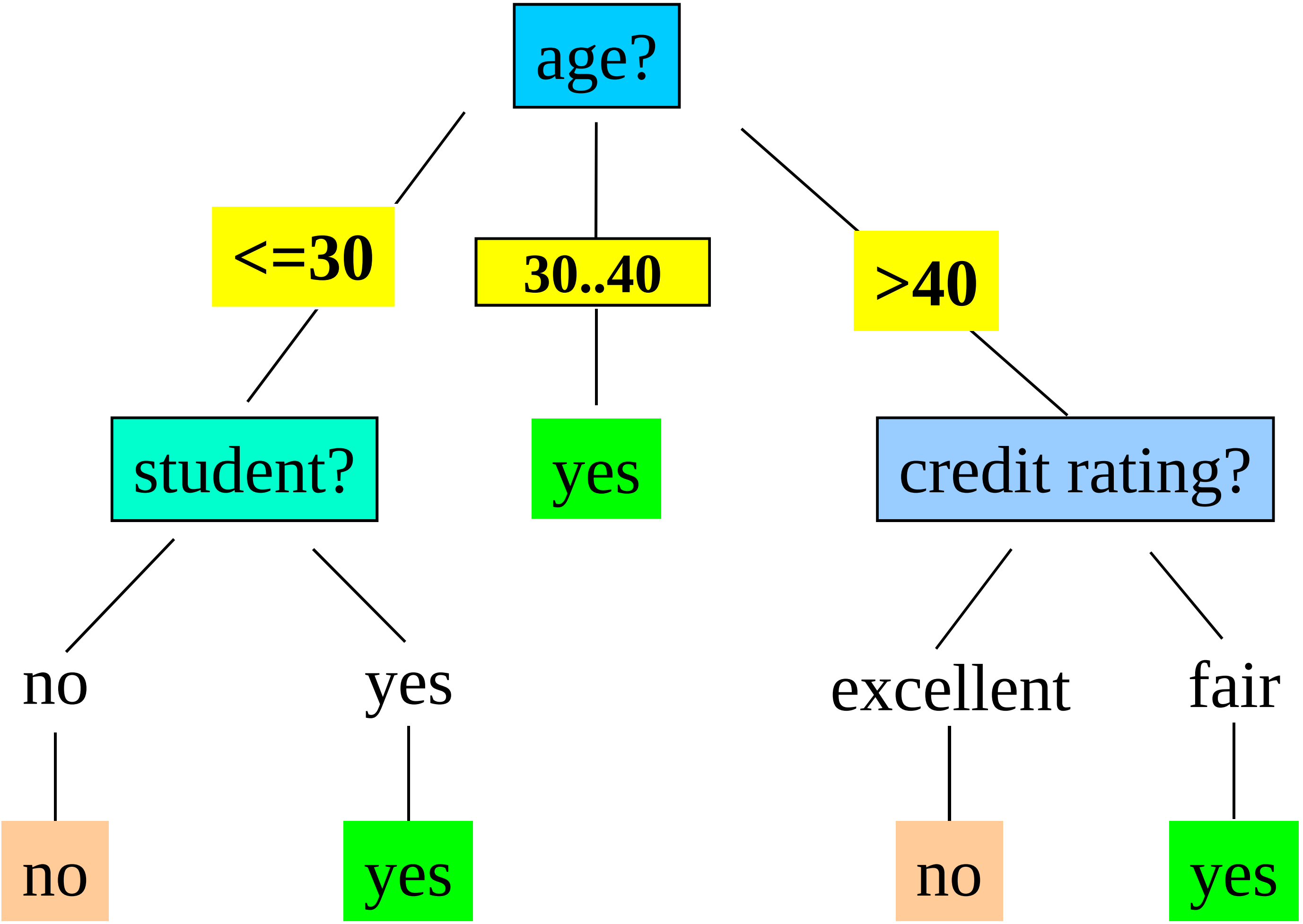
Tenured?

Yes

# Dataset

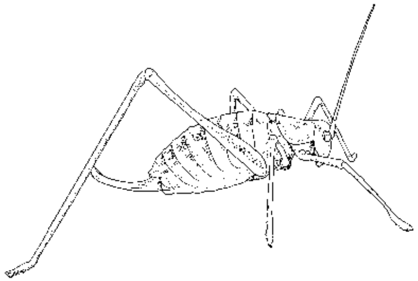
| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | |
| 31…40 | high | yes | fair | |
| >40 | medium | no | excellent | |

# A Decision Tree for *“buys_computer”*

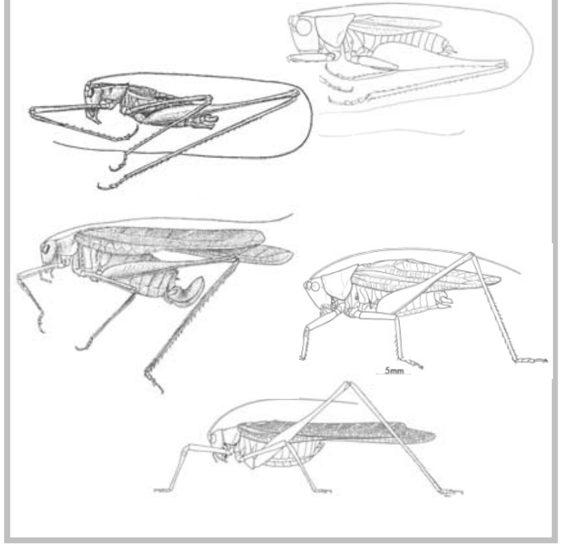


# Classification example

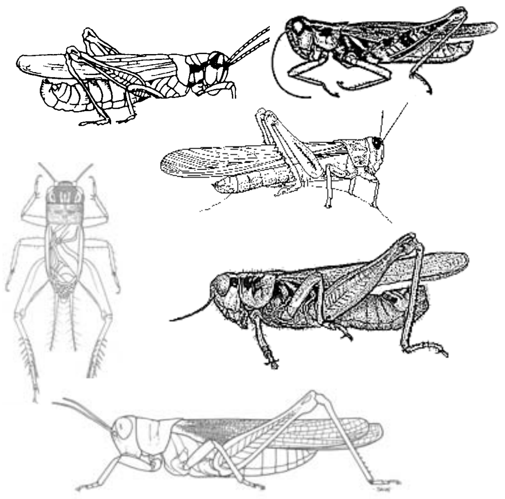
Given a collection of annotated data. (in this case 5 instances of **Katydids** and five of **Grasshoppers)**, decide what type of insect the unlabeled example is.



**Katydids**



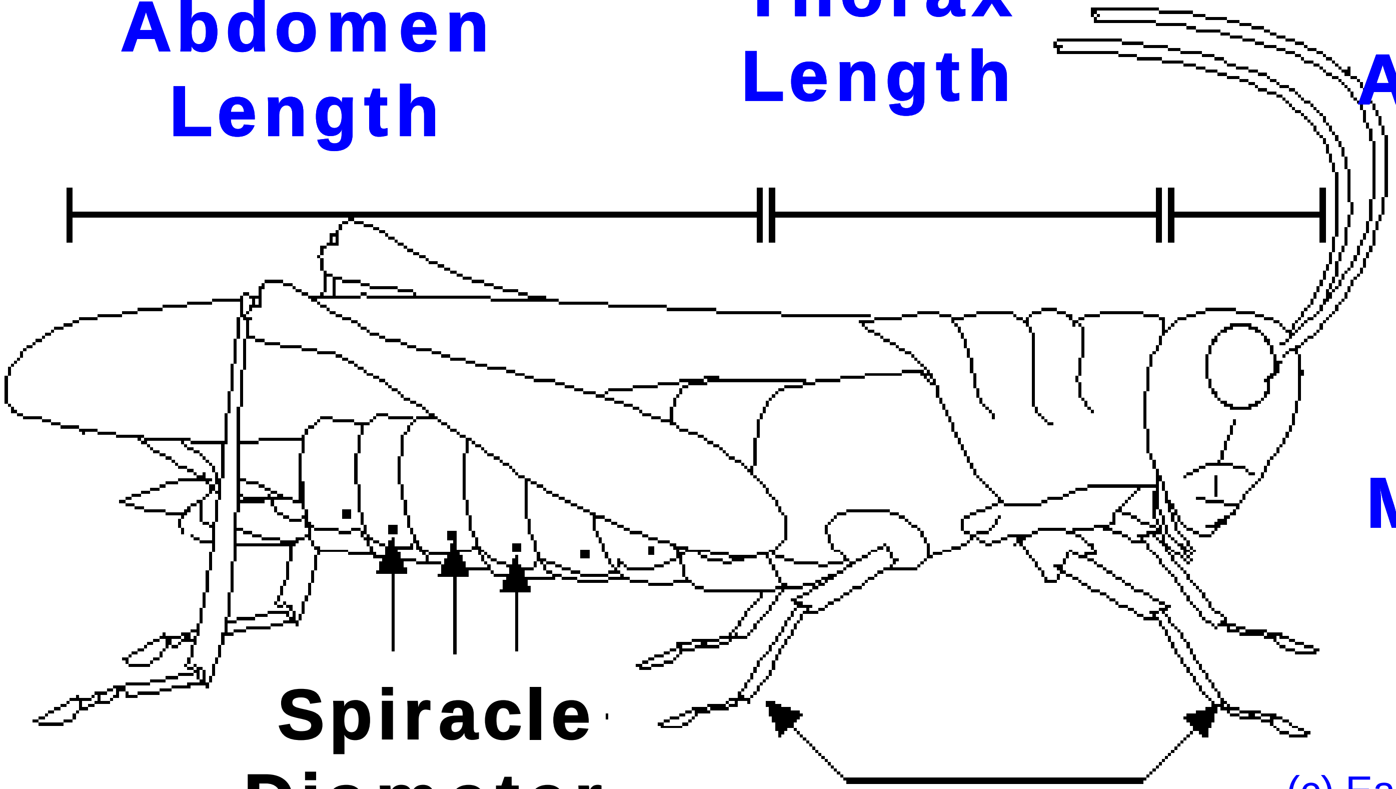
**Grasshoppers**



(c) Eamonn Keogh, eamonn@cs.ucr.edu

Spanish: Grillo - saltamontes

# Classification example



Has Wings?

Color {Green, Brown, Gray, Other}

Abdomen Length

Thorax Length

Antennae Length

Mandible Size

Spiracle Diameter

Leg Length

# Classification example

The classification problem can now be expressed as:

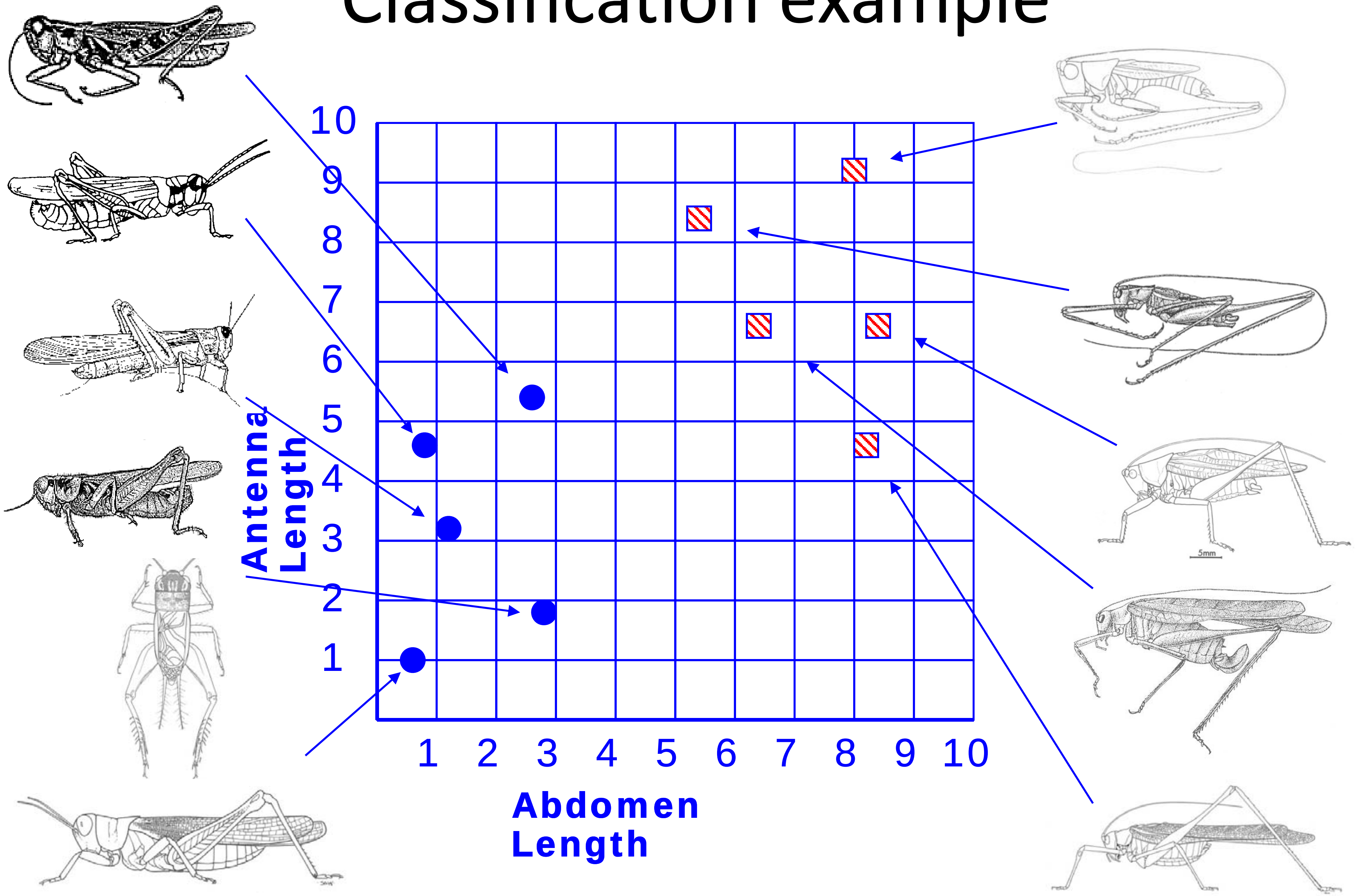
Given a training database predict the **class** label of a previously unseen instance

| Insect ID | Abdomen Length | Antennae Length | Insect Class |
|---|---|---|---|
| *1* | 2.7 | 5.5 | **Grasshopper** |
| *2* | 8.0 | 9.1 | **Katydid** |
| *3* | 0.9 | 4.7 | **Grasshopper** |
| *4* | 1.1 | 3.1 | **Grasshopper** |
| *5* | 5.4 | 8.5 | **Katydid** |
| *6* | 2.9 | 1.9 | **Grasshopper** |
| *7* | 6.1 | 6.6 | **Katydid** |
| *8* | 0.5 | 1.0 | **Grasshopper** |
| *9* | 8.3 | 6.6 | **Katydid** |
| *10* | 8.1 | 4.7 | **Katydid** |

| previously unseen instance = | 5.1 | 7.0 | **???????** |
|---|---|---|---|

(c) Eamonn Keogh, eamonn@cs.ucr.edu

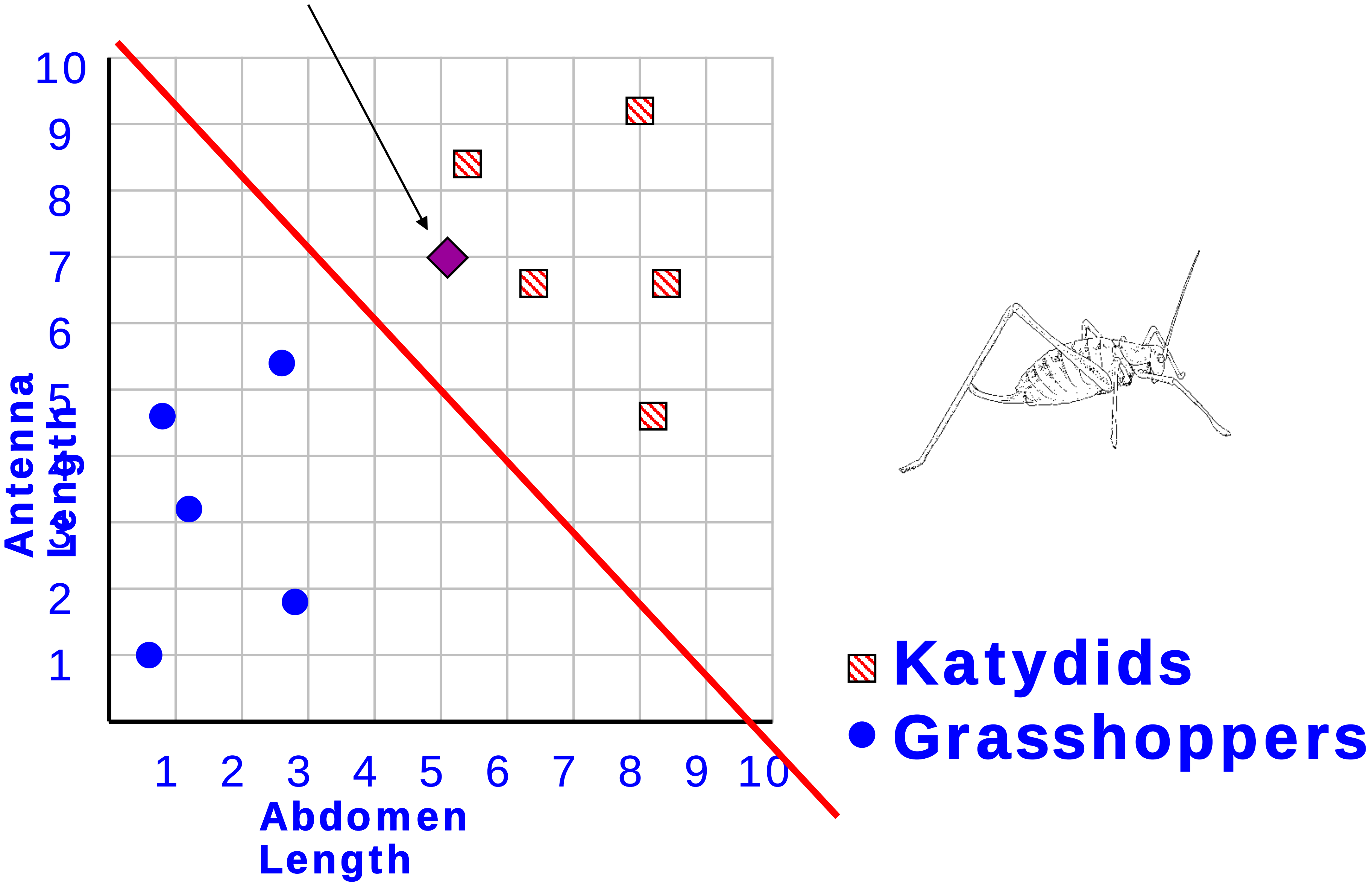
# Classification example



**Grasshoppers**

**Katydids**

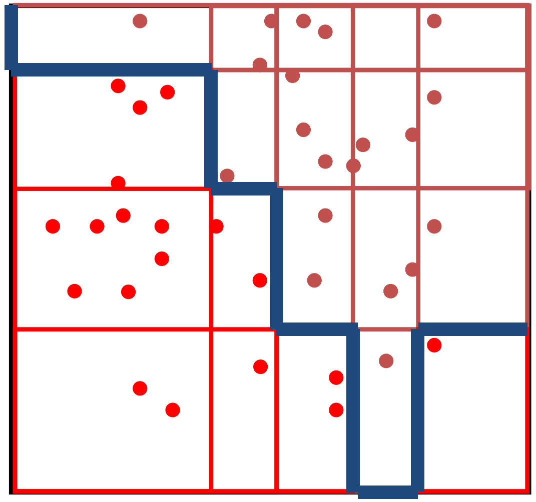
# Classification example



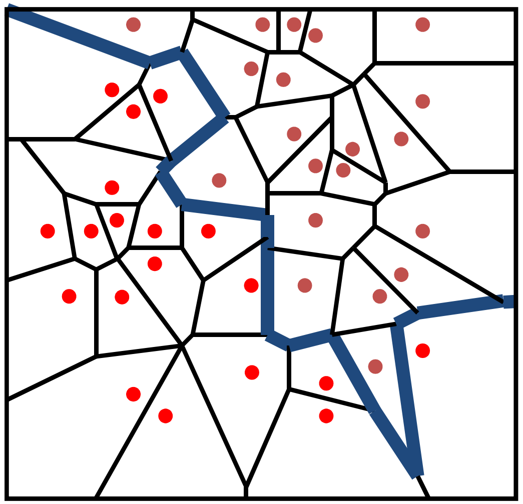
Linear classifier

# Classification models

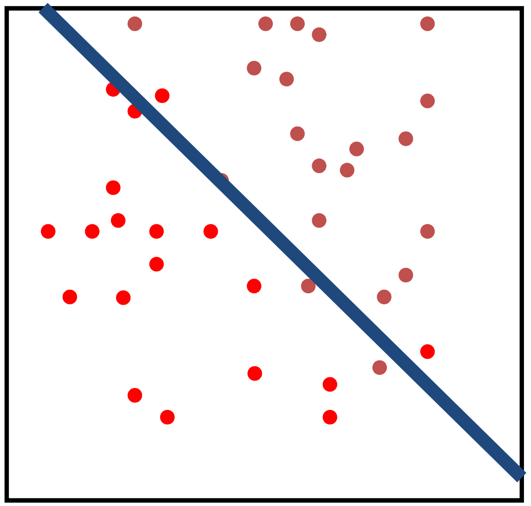
- Interval rules based classifier



- Instance based classifier



- Linear classifier



# Classification Accuracy: Estimating Error Rates

- Partition: Training-and-testing
  - use two independent data sets, e.g., training set (2/3), test set(1/3)
  - used for data set with large number of samples
- Cross-validation
  - divide the data set into *k* subsamples
  - use *k-1* subsamples as training data and one sub-sample as test data—*k*-fold cross-validation
  - for data set with moderate size
- Bootstrapping (leave-one-out)
  - for small size data

# Outline

✓Introduction

✓Classification

✓Prediction

✓ Clustering

✓Association

✓ Data Mining Systems / Data Set Repositories

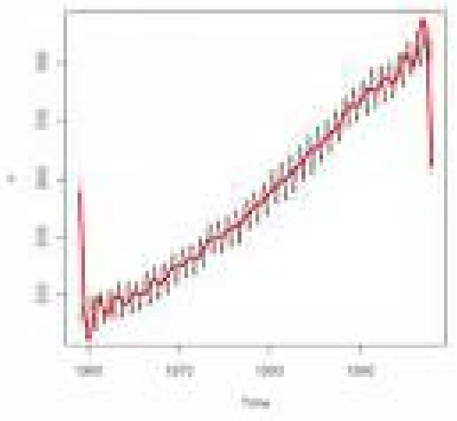
✓Concluding Remarks

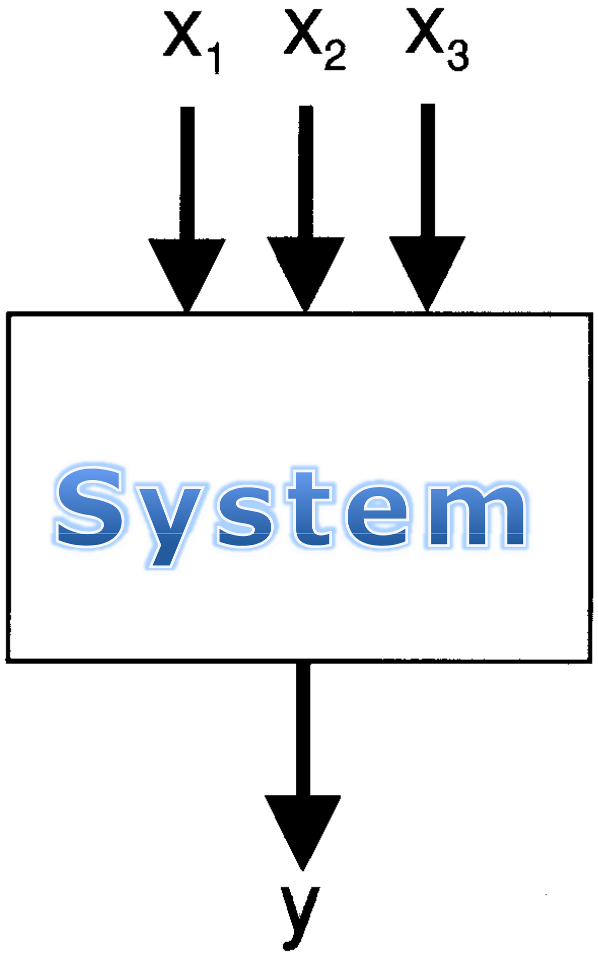
# Prediction Problem

Prediction is different from classification

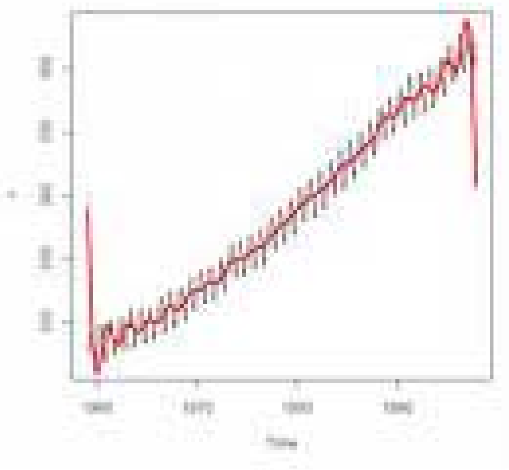
Classification refers to predict categorical class label

Prediction models continuous-valued functions





# How to work?



- Prediction work is similar to classification
  - First, construct a model
  - Second, use model to predict unknown value
    - Major method for prediction is regression
      - Linear and multiple regression
      - Non-linear regression

# Regression Analysis in Prediction

- Linear regression: Y = $\alpha$ + $\beta$ X
  - Two parameters , $\alpha$ and $\beta$ specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of Y1, Y2, …, X1, X2, ….
- Multiple regression: Y = b0 + b1 X1 + b2 X2.
  - Many nonlinear functions can be transformed into the above.
- Neural networks, fuzzy rule based systems, ….

# Introduction to Prediction, Clustering, Classification and Association

## Outline

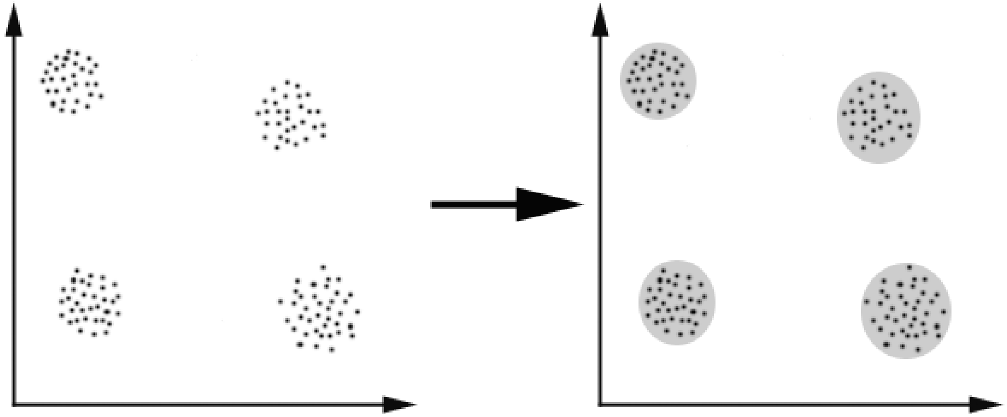
- ✓Introduction
- ✓Classification
- ✓Prediction
- ✓ Clustering
- ✓Association
- ✓ Data Mining Systems / Data Set Repositories
- ✓Concluding Remarks

# Clustering Problem

- Given a database D={$t_1$,$t_2$,…,$t_n$} of tuples and an integer value k, the ***Clustering Problem*** is to define a mapping f:D→{1,..,k} where each $t_i$ is assigned to one cluster $K_j$, 1<=j<=k.

- A ***Cluster***, $K_j$, contains precisely those tuples mapped to it.

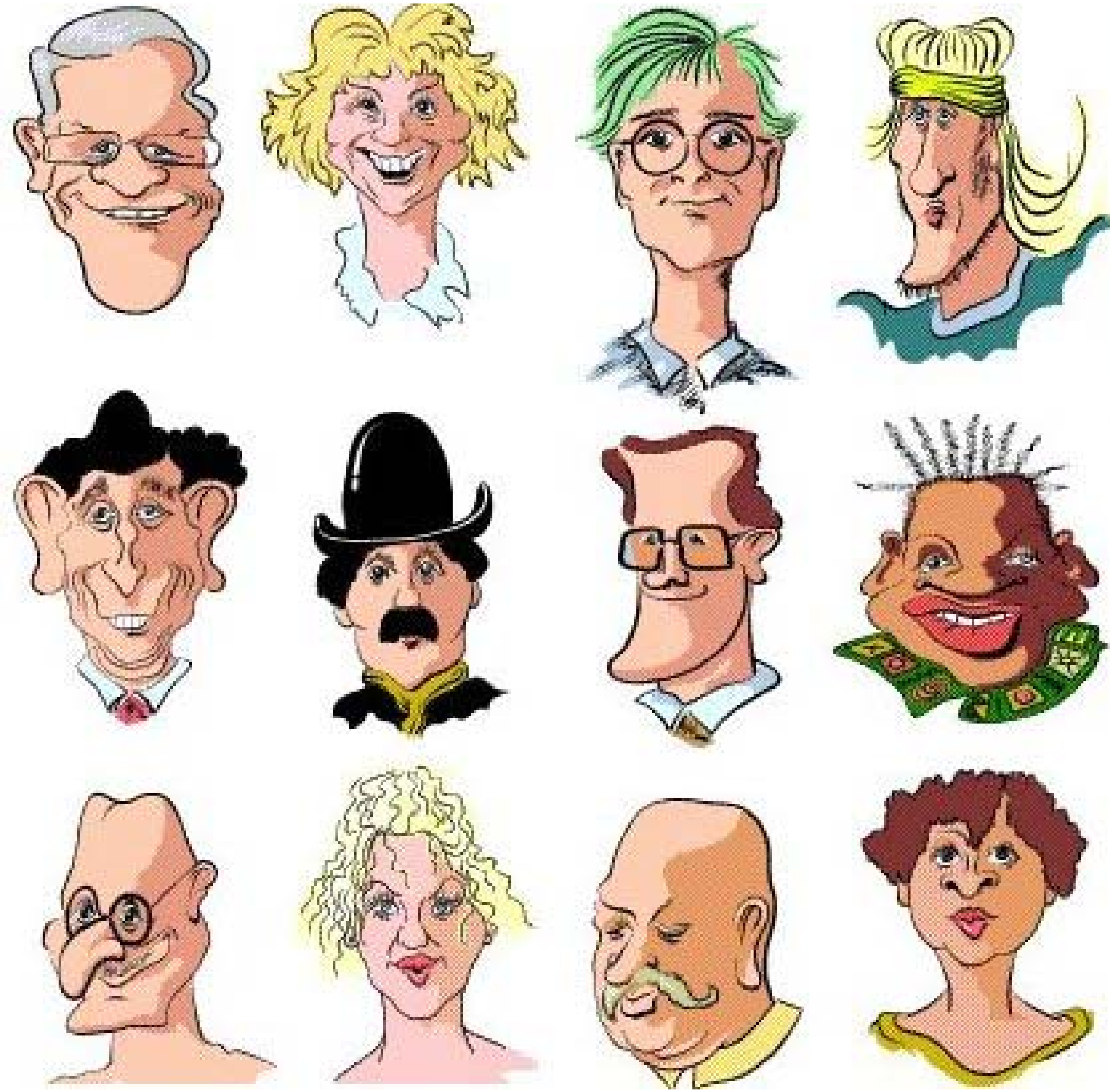
- Unlike classification
- problem, clusters are
- not known a priori.



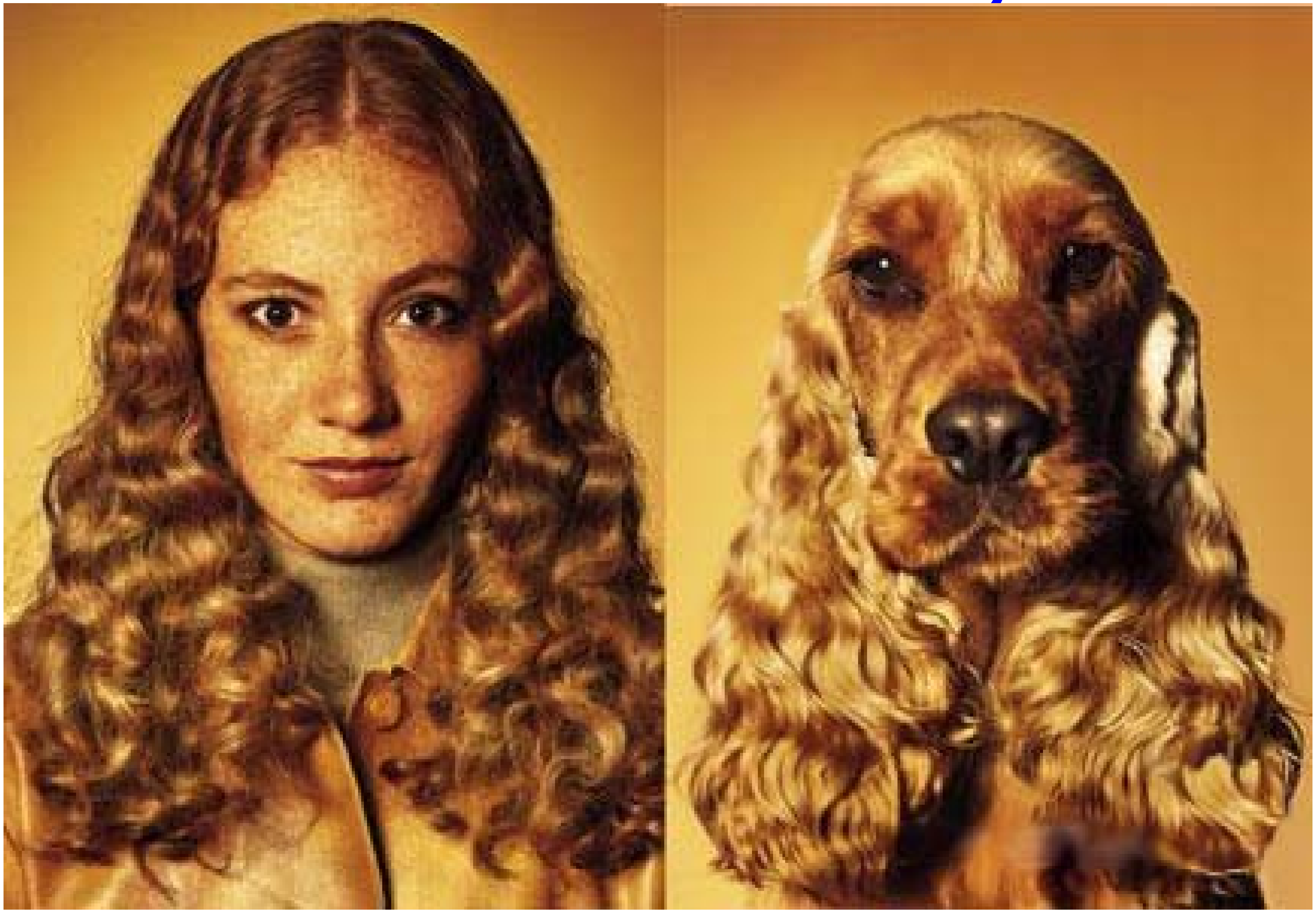
# Clustering Examples

- ***Segment*** customer database based on similar buying patterns.
- Group houses in a town into neighborhoods based on similar features.
- Identify new plant species
- Identify similar Web usage patterns

# Clustering Problem



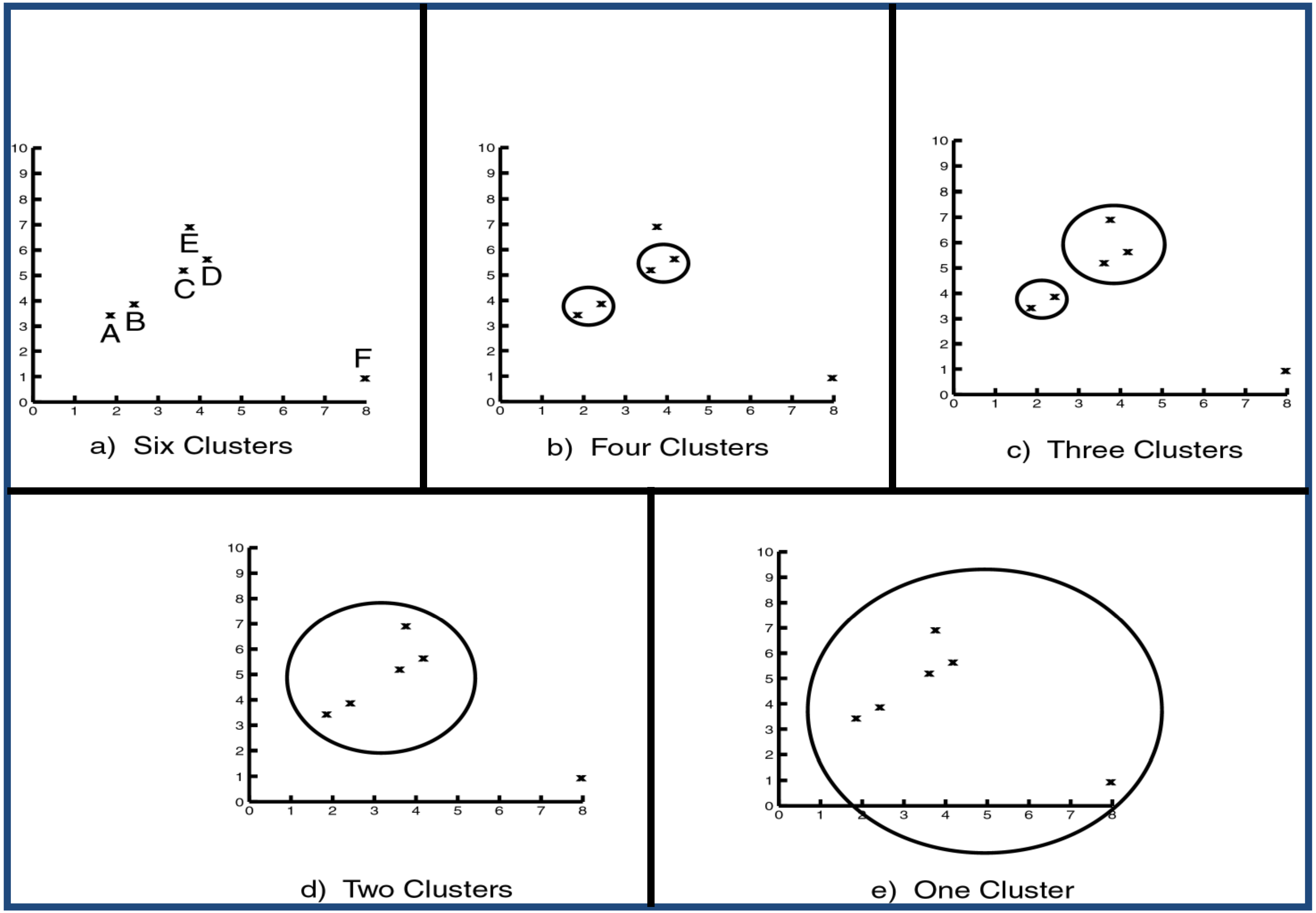
# What is Similarity?



# Clustering vs. Classification

- No prior knowledge
  - Number of clusters
  - Meaning of clusters
- Unsupervised learning

# Levels of Clustering



a) Six Clusters

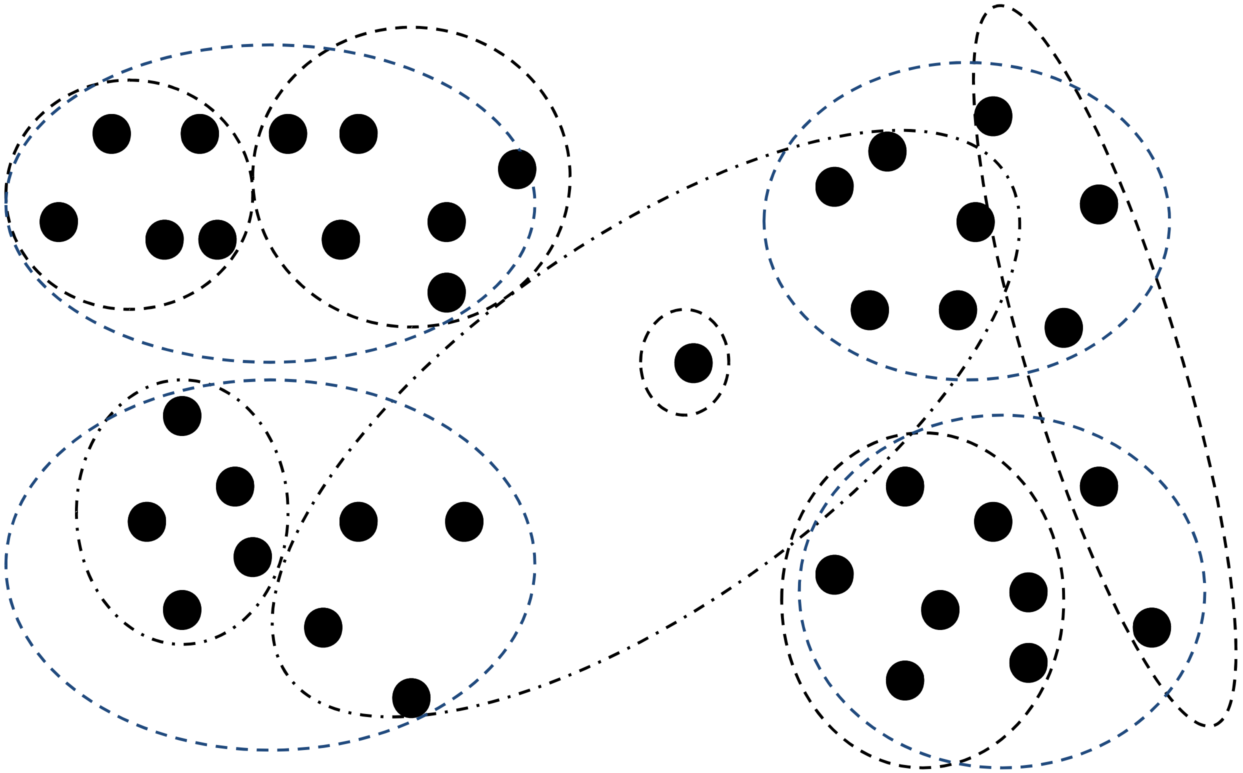
b) Four Clusters

c) Three Clusters

d) Two Clusters

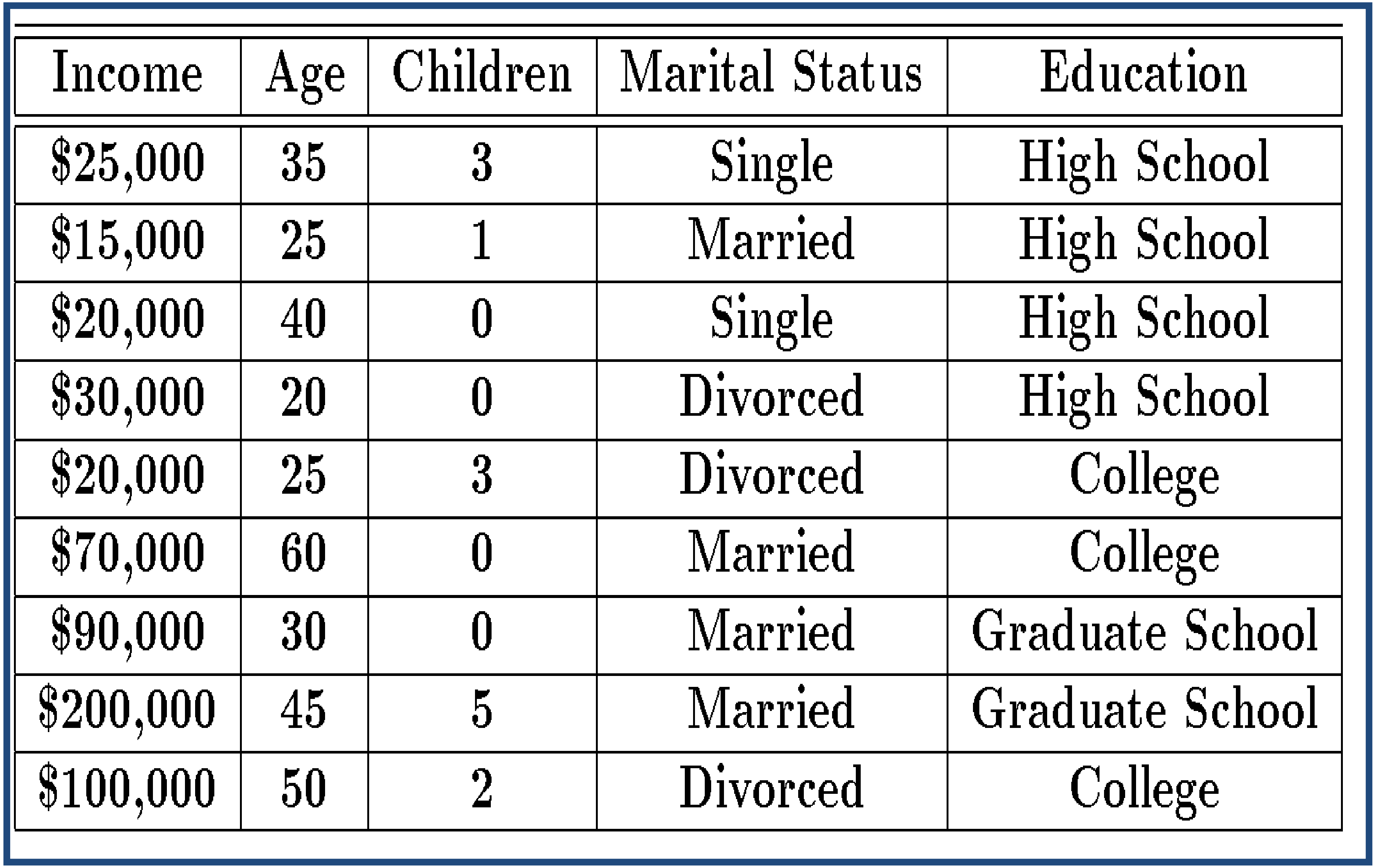
e) One Cluster

# Levels of Clustering



Size Based

# Clustering Example



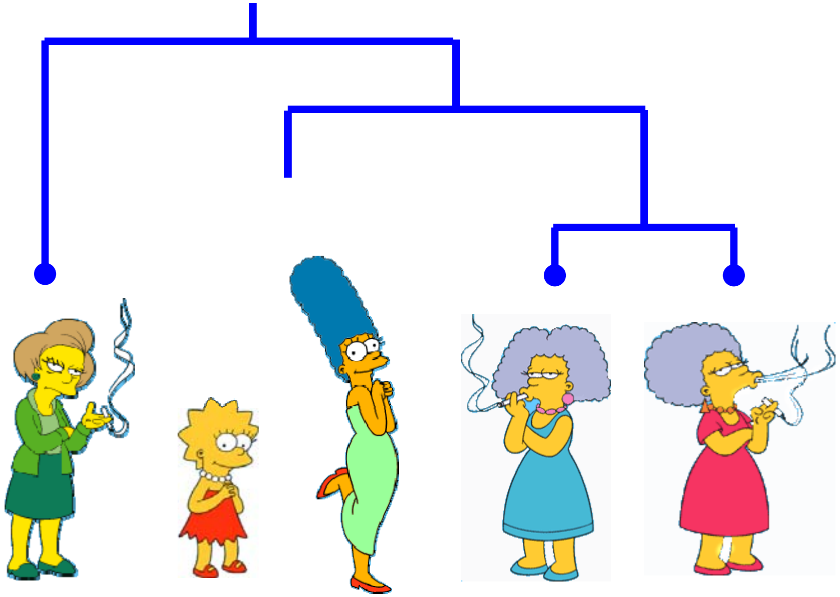
| Income | Age | Children | Marital Status | Education |
|---|---|---|---|---|
| $25,000 | 35 | 3 | Single | High School |
| $15,000 | 25 | 1 | Married | High School |
| $20,000 | 40 | 0 | Single | High School |
| $30,000 | 20 | 0 | Divorced | High School |
| $20,000 | 25 | 3 | Divorced | College |
| $70,000 | 60 | 0 | Married | College |
| $90,000 | 30 | 0 | Married | Graduate School |
| $200,000 | 45 | 5 | Married | Graduate School |
| $100,000 | 50 | 2 | Divorced | College |

# Types of Clustering

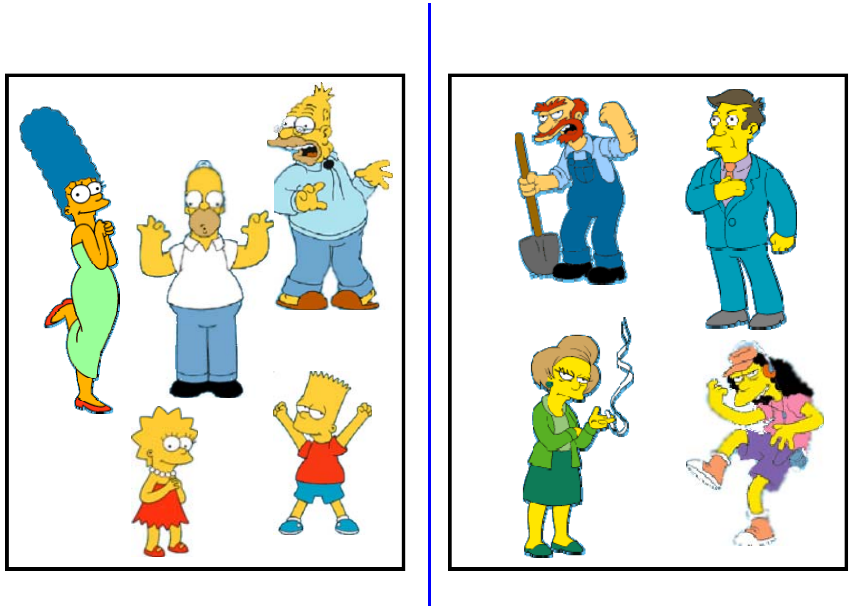
- ***Hierarchical*** – Nested set of clusters created.
- ***Partitional*** – One set of clusters created.
- ***Incremental*** – Each element handled one at a time.
- ***Simultaneous*** – All elements handled together.
- ***Overlapping/Non-overlapping***

# Types of Clustering

**Hierarchical**



**Partitional**



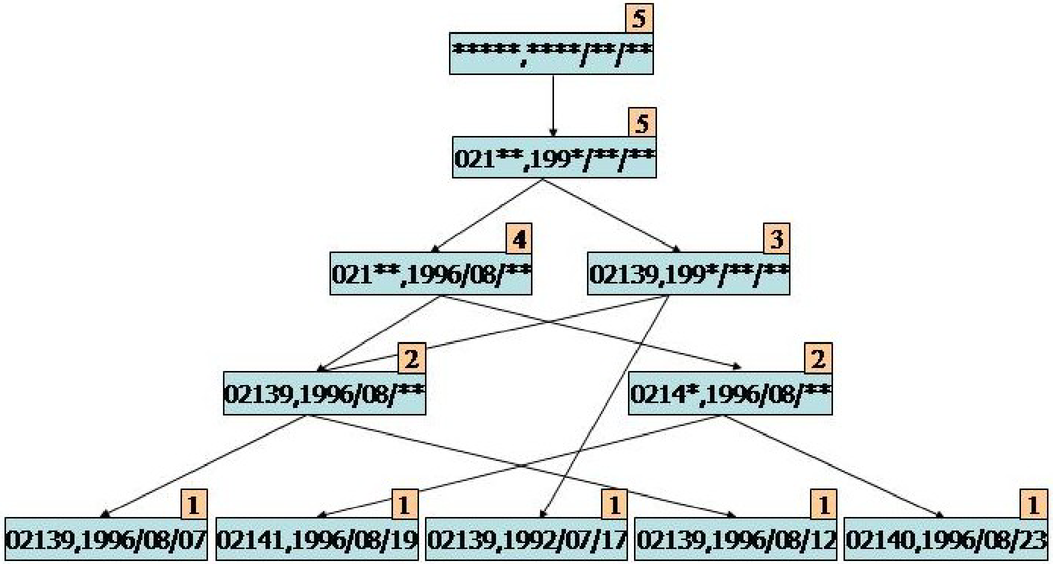
(c) Eamonn Keogh, eamonn@cs.ucr.edu

# Outline

- ✓ Introduction
- ✓ Classification
- ✓ Prediction
- ✓ Clustering
- ✓ Association
- ✓ Data Mining Systems / Data Set Repositories
- ✓ Concluding Remarks

# Association Rule Problem

- Given a set of items $I=\{I_1,I_2,\ldots,I_m\}$ and a database of transactions $D=\{t_1,t_2, \ldots, t_n\}$ where $t_i=\{I_{i1},I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, the ***Association Rule Problem*** is to identify all association rules $X \Rightarrow Y$ with a minimum support and confidence.
- Link Analysis
- ***NOTE:*** Support of $X \Rightarrow Y$ is same as support of $X \cup Y$.



# Example: Market Basket Data

- Items frequently purchased together:

  **Bread ⇒PeanutButter**

- Uses:
  - Placement
  - Advertising
  - Sales
  - Coupons
- Objective: increase sales and reduce costs

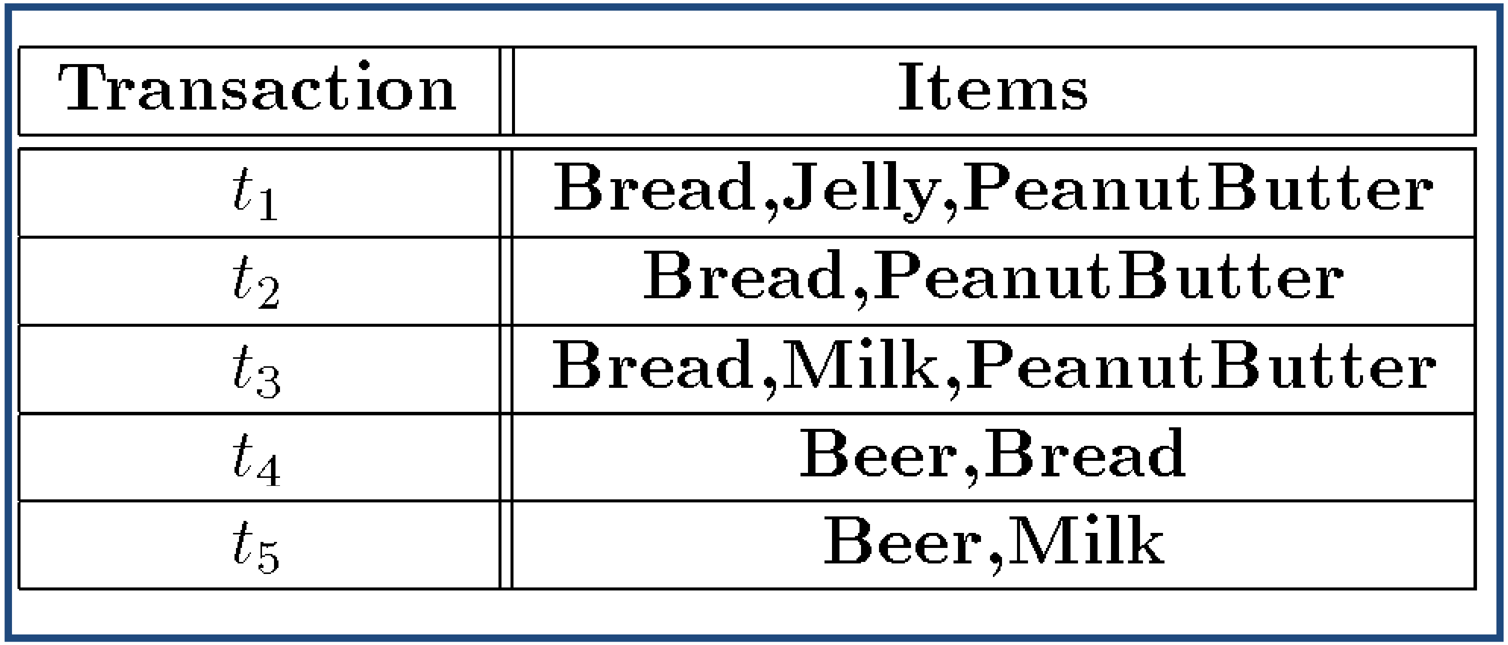
# Association Rule Definitions

- ***Set of items:*** $I=\{I_1,I_2,\ldots,I_m\}$
- ***Transactions:*** $D=\{t_1,t_2, \ldots, t_n\}$, $t_j \subseteq I$
- ***Itemset:*** $\{I_{i1},I_{i2}, \ldots, I_{ik}\} \subseteq I$
- ***Support of an itemset:*** Percentage of transactions which contain that itemset.
- ***Large (Frequent) itemset:*** Itemset whose number of occurrences is above a threshold.

# Association Rule Definitions

- ***Association Rule (AR):*** implication $X \Rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y$ = ;
- ***Support of AR (s) $X \Rightarrow Y$***: Percentage of transactions that contain $X \cup Y$
- ***Confidence of AR ($\alpha$) $X \Rightarrow Y$:*** Ratio of number of transactions that contain $X \cup Y$ to the number that contain X

# Association Rules Example

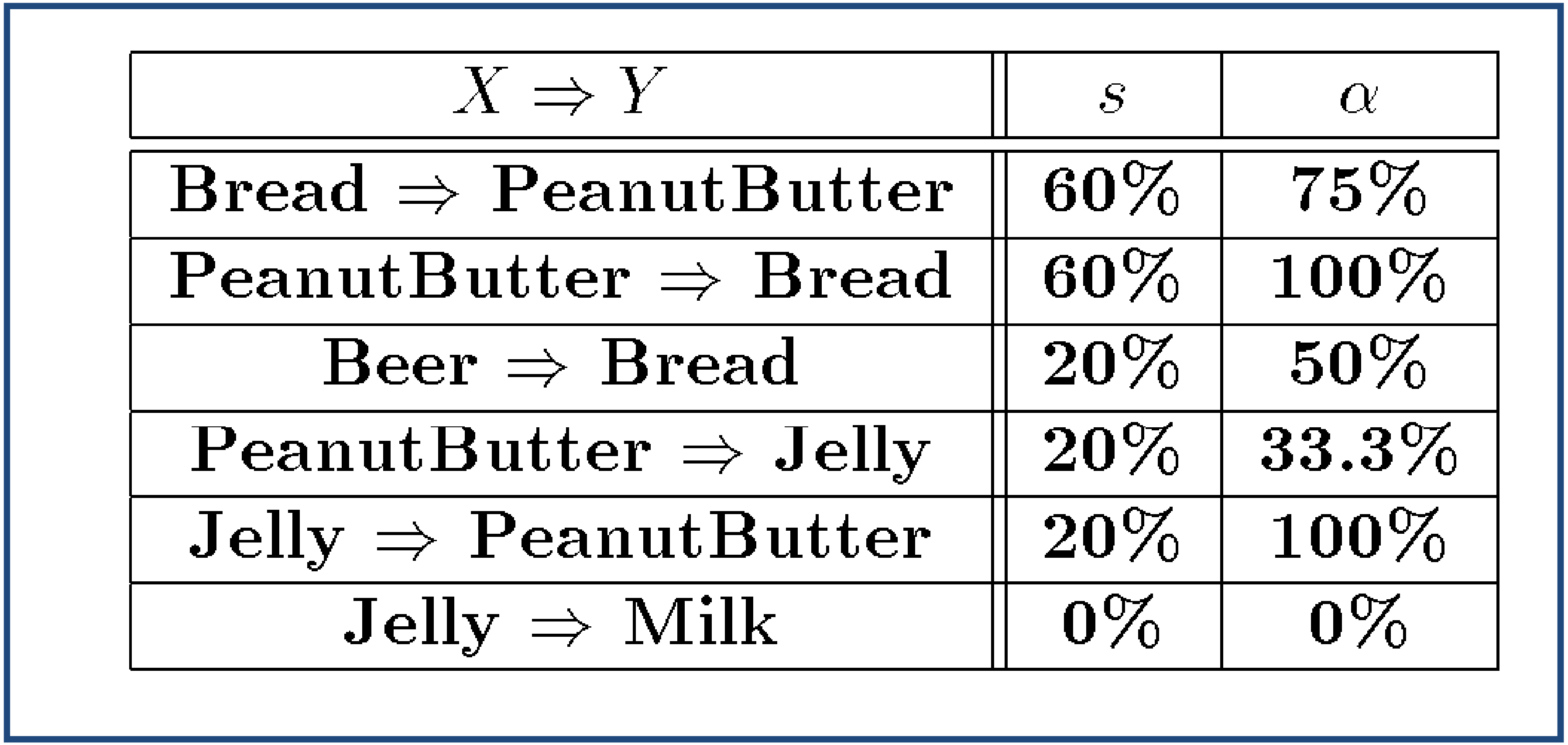


| Transaction | Items |
|---|---|
| $t_1$ | Bread,Jelly,PeanutButter |
| $t_2$ | Bread,PeanutButter |
| $t_3$ | Bread,Milk,PeanutButter |
| $t_4$ | Beer,Bread |
| $t_5$ | Beer,Milk |

I = { Beer, Bread, Jelly, Milk, PeanutButter}

Support of {Bread,PeanutButter} is 60%

# Association Rules Ex (cont’d)



| $X \Rightarrow Y$ | $s$ | $\alpha$ |
|---|---|---|
| **Bread ⇒ PeanutButter** | **60%** | **75%** |
| **PeanutButter ⇒ Bread** | **60%** | **100%** |
| **Beer ⇒ Bread** | **20%** | **50%** |
| **PeanutButter ⇒ Jelly** | **20%** | **33.3%** |
| **Jelly ⇒ PeanutButter** | **20%** | **100%** |
| **Jelly ⇒ Milk** | **0%** | **0%** |

# Association Rule Techniques

1. Find Large Itemsets.
2. Generate rules from frequent itemsets.

**Apriori (1993): Apriori** is a classic algorithm for learning association rules

- ***Large Itemset Property:***

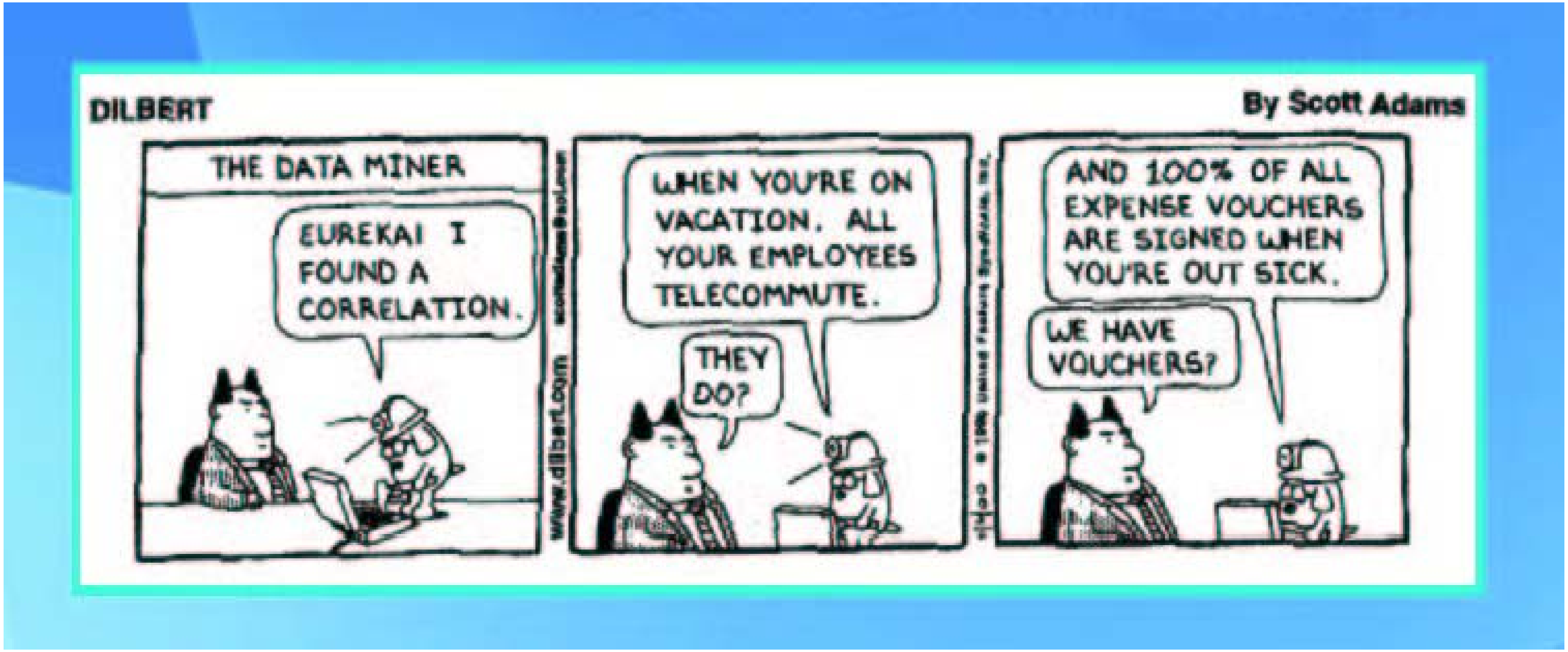
  *Any subset of a large itemset is large.*

- Contrapositive:

  ***If an itemset is not large,***
  ***none of its supersets are large.***

# Measuring Quality of Rules

- Support
- Confidence
- Interest
- Conviction
- Chi Squared Test


DILBERT
By Scott Adams
THE DATA MINER
EUREKA! I FOUND A CORRELATION.
www.dilbert.com
scottadams@aol.com
WHEN YOU'RE ON VACATION, ALL YOUR EMPLOYEES TELECOMMUTE.
THEY DO?
AND 100% OF ALL EXPENSE VOUCHERS ARE SIGNED WHEN YOU'RE OUT SICK.
WE HAVE VOUCHERS?

*Data Mining Introductory and Advanced Topics*, by Margaret H. Dunham, Prentice Hall, 2003.

DILBERT reprinted by permission of United Feature Syndicate, Inc.

# Outline

✓Introduction

✓Classification

✓Prediction

✓ Clustering

✓Association

✓ Data Mining Systems / Data Set Repositories

✓Concluding Remarks

# Data Mining System

Some data mining systems …..

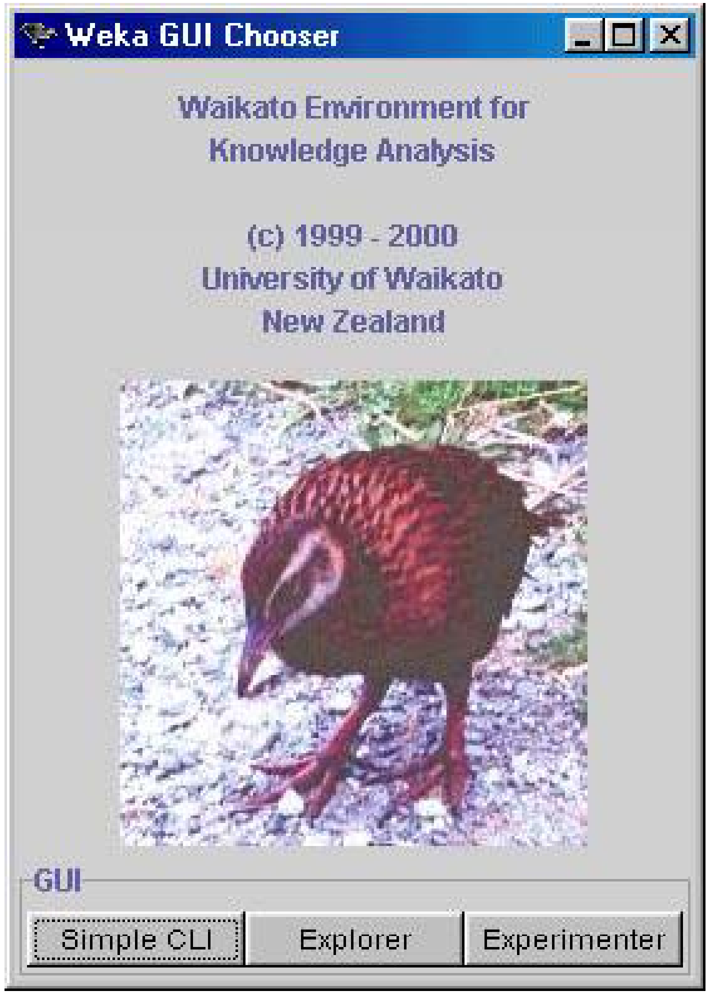
Weka

KEEL

Rapid Miner

# Weka

Data Mining System



Weka GUI Chooser

Waikato Environment for Knowledge Analysis

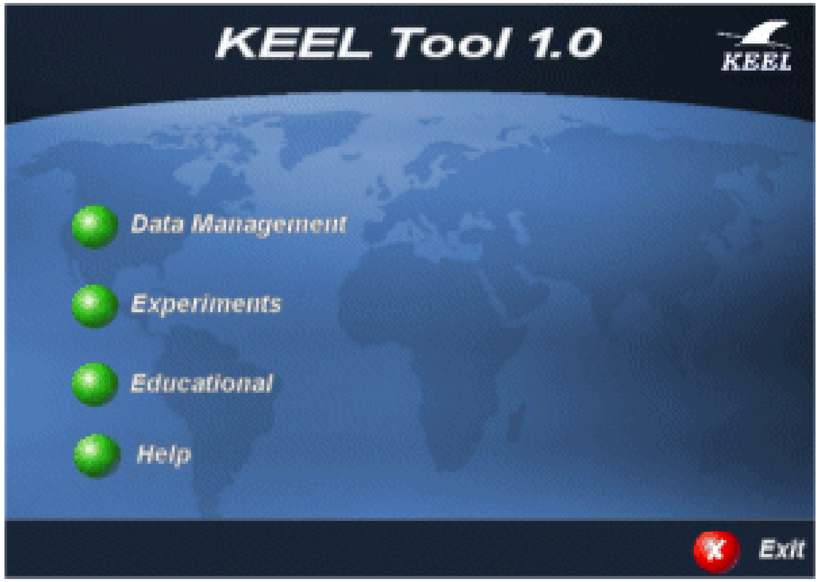
(c) 1999 - 2000
University of Waikato
New Zealand

- The University of Waikato, New Zealand
- Machine learning software in Java implementation

http://www.cs.waikato.ac.nz/ml/weka/

# KEEL

Data Mining System



- Machine learning software in Java implementation

http://www.keel.es/

# Rapid Miner

Data Mining System

- Rapid Miner YALE: Yet Another Learning Environment

http://rapid-i.com/

# Data Mining Repositories

Most of the commercial datasets used by companies for data mining area not available for others to use.

However there area a number of “libraries” of datasets that are readily available for downloading from the World Wide Web free of charge by any one.

The best known of these is the “Repository” of datasets maintained by the University of California at Irvine, generally known as the “UCI Repository”. The URL for the Repository is: **http://archive.ics.uci.edu/ml**

# Data Mining Repositories

It contains approximately 120 datasets

on topics as diverse as credit risks,

patients classification, sensor data

of a mobile robot, …

Datasets with missing values and noise are included.

A recent development is the creation of the UCI “Knowledge Discovery in Data Bases Archive” at [http://kdd.ics.uci.edu/](http://kdd.ics.uci.edu/).

This contains a range of large and complex datasets

as a challenge to the data mining

research community to scale up

its algorithms as the size of sotred datasets.

# Data Mining Repositories

It contains approximately 120 datasets on topics as diverse as credit risks, patients classification, sensor data of a mobile robot, …

Datasets with missing values and noise are included.

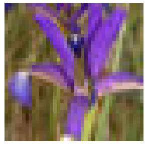
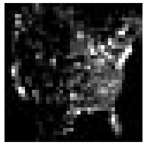
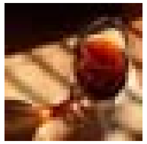
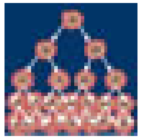
A recent development is the creation of the UCI “Knowledge Discovery in Data Bases Archive” at [http://kdd.ics.uci.edu/](http://kdd.ics.uci.edu/).

This contains a range of large and complex datasets as a challenge to the data mining research community to scale up its algorithms as the size of sotred datasets.

# Data Mining Repositories

UCI

Most popular data sets

| | | |
|---|---|---|
| 41057: |  | Iris |
| 33055: |  | Adult |
| 27764: |  | Wine |
| 24353: |  | Breast Cancer Wisconsin (Diagnostic) |
| 19211: | | Poker Hand |
| 19161: | | Abalone |

# Data Mining Repositories

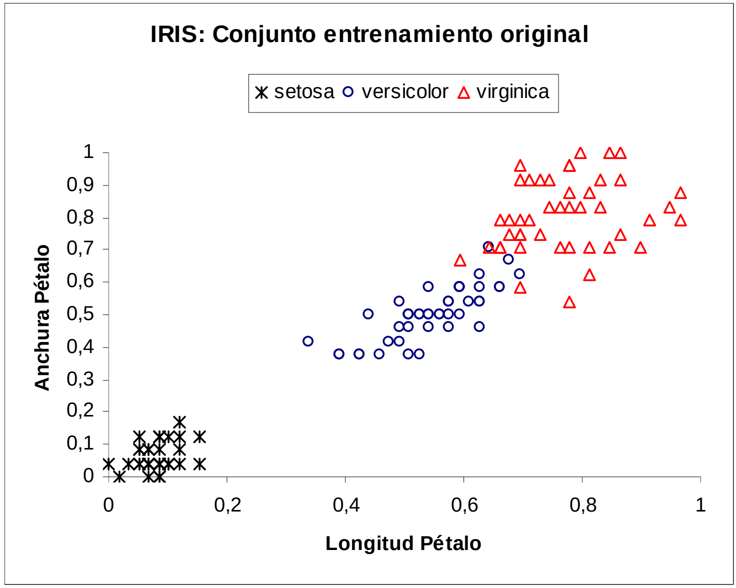
## Iris Data Set

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 150 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 4 | **Date Donated** | 1988-07-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 41063 |

**Attribute Information:**

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa<br>
-- Iris Versicolour<br>
-- Iris Virginica



# Data Mining Systems/ Repositories

Other links to Data Mining Systems and Repositories at: http://sci2s.ugr.es/keel/links.php

# Outline

✓Introduction

✓Classification

✓Prediction

✓ Clustering

✓Association

✓ Data Mining Systems / Data Set Repositories

✓Concluding Remarks

# Concluding Remarks

Some data mining tasks:

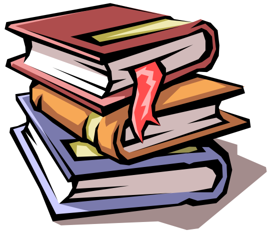
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

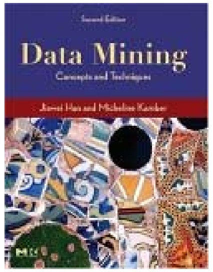
    (classification, regression)

- Description Methods
  - Find human-interpretable patterns that describe the data.

    (clustering, association, ..)

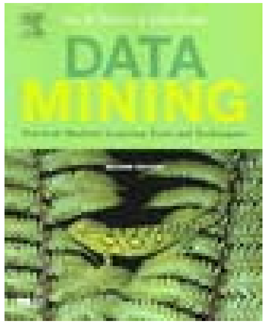
# Bibliography

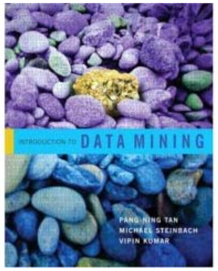




J. Han, M. Kamber.
Data Mining. Concepts and Techniques
Morgan Kaufmann, 2006 (Second Edition)
http://www.cs.sfu.ca/~han/dmbook

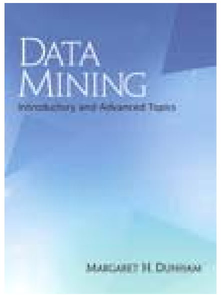
I.H. Witten, E. Frank.
Data Mining: Practical Machine Learning Tools and Techniques,
Second Edition,Morgan Kaufmann, 2005.
http://www.cs.waikato.ac.nz/~ml/weka/book.html





Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
Introduction to Data Mining (First Edition)
Addison Wesley, (May 2, 2005)
http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Margaret H. Dunham
Data Mining: Introductory and Advanced Topics
Prentice Hall, 2003
http://lyle.smu.edu/~mhd/book



# Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.