COMP 6838 Data MIning

LECTURE 1: Introduction

Dr. Edgar Acuna
Departmento de Matematicas

Universidad de Puerto Rico- Mayaguez

math.uprm.edu/~edgar

Course’s Objectives

- Understand the basic concepts to carry out data mining and knowledge discovery in databases.
- Implement on real world datasets the most well known data mining algorithms.

- Course’s Schedule: Tuesday and Thursday from 2.00pm till 3.15 pm in M118.
- Prerequisites: Two courses including statistical and probability concepts. Some knowledge of matrix algebra, databases and programming.

- Office: M314
- Office’s Hours: Monday 7.30-9am, Tuesday: 7.30-8.30am and Thursday 9.30-10.30am.
- Extension x3287
- E-mail: edgar.acuna@upr.edu, eacunaf@yahoo.com
- TA: Roxana Aparicio (M 309, M108)

References

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, 2005.
- Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
- Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, 2005.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Verlag, 2001.
- Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, Wiley-IEEE Press, 2002.
- Michael Berry & Gordon Linoff, Mastering Data Mining, John Wiley & Sons, 2000.
- Graham Williams, Data Mining Desktop Survival Guide, on-line book (PDF).
- David J. Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining , MIT Press, 2000.

Software

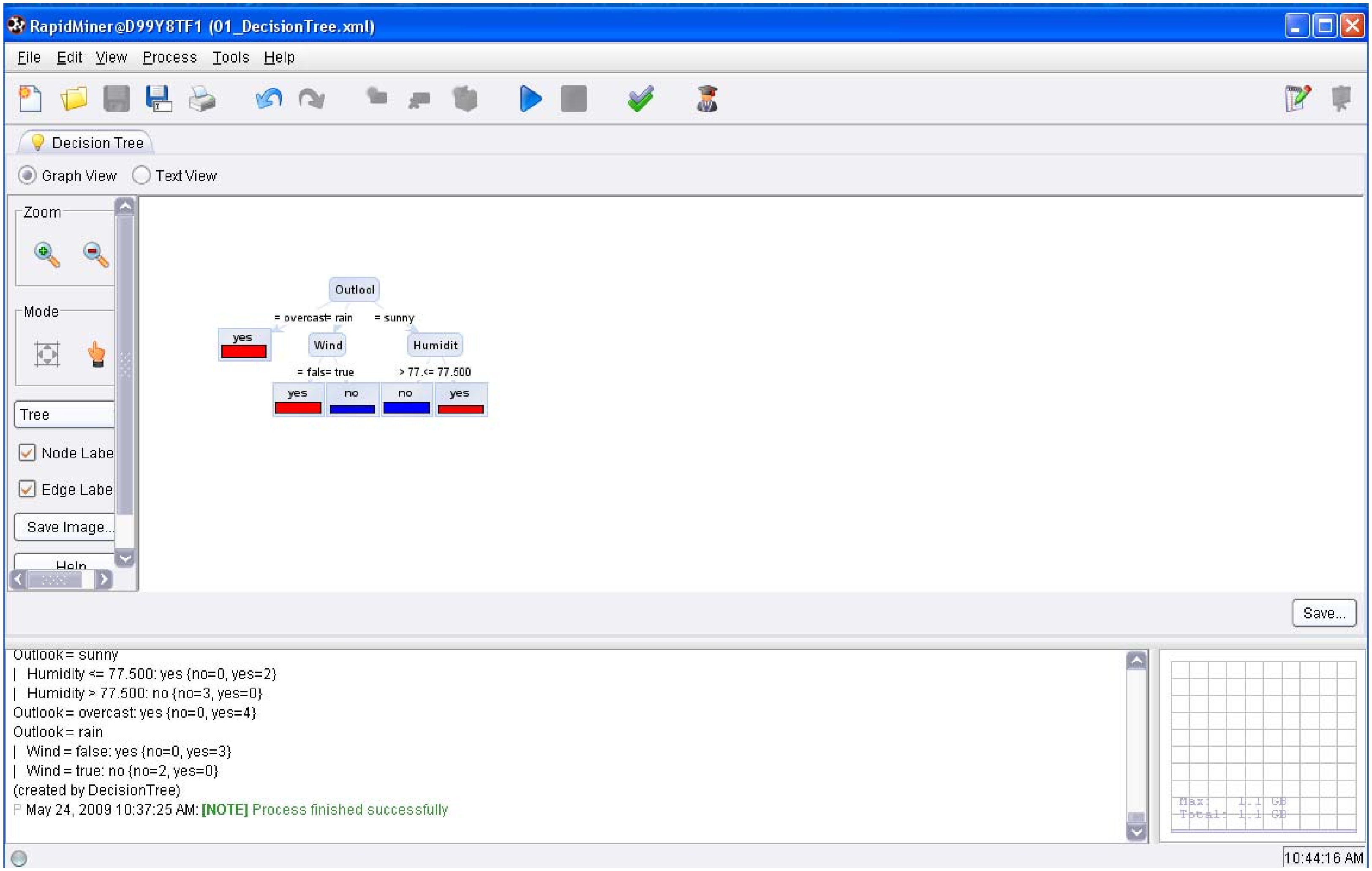
Free:

- R (cran.r-project.org). Statistical oriented.
- Weka (http://www.cs.waikato.ac.nz/ml/weka/): written in Java, manual in spanish.There is an R interface to Weka (RWeka)
- RapidMiner (YALE) (http://rapid-i.com). It has more features than Weka.
- Orange (http://www.ailab.si/orange). It requires Python and other programs.

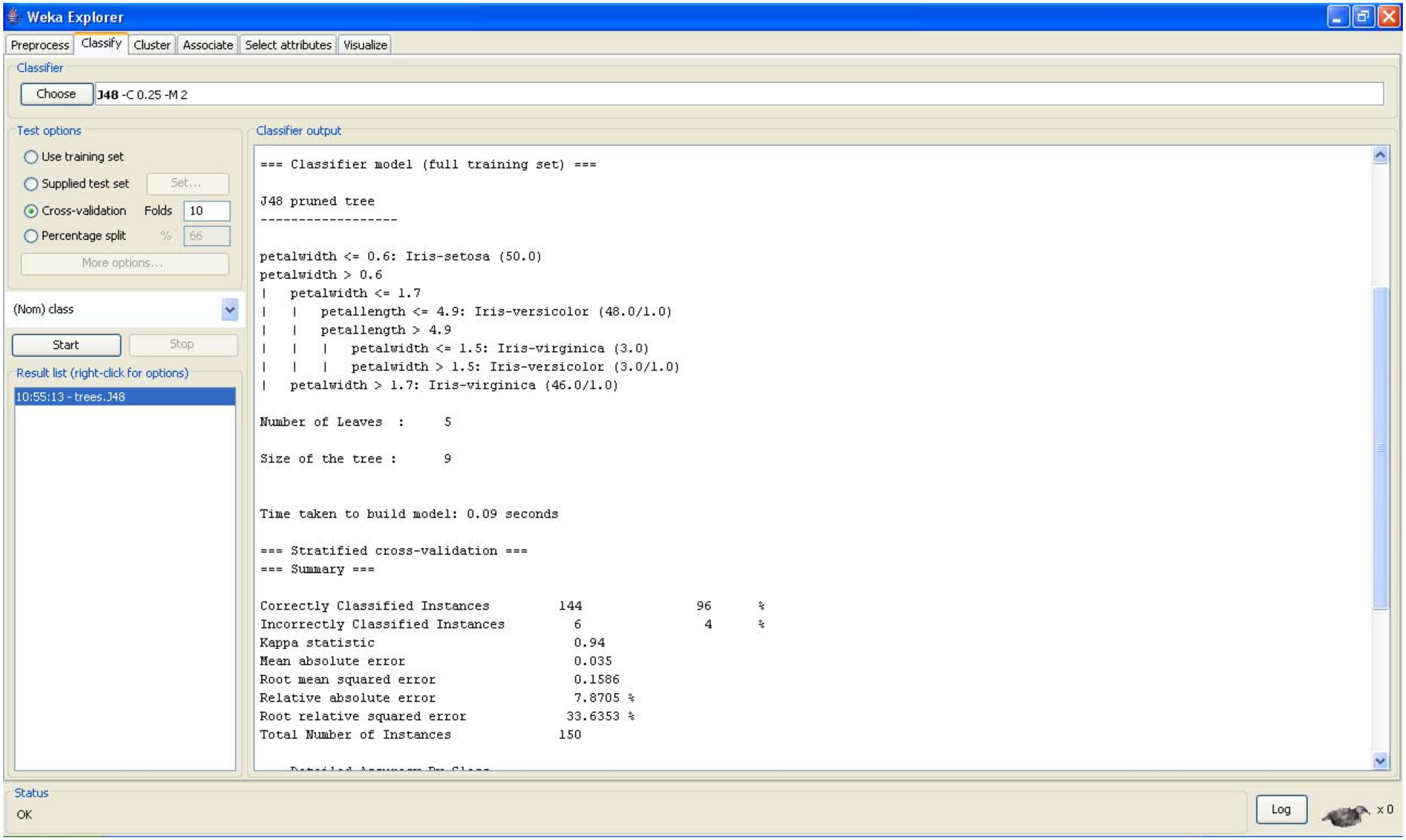
Software

Comercials:

- Microsoft SQL 2008: Analysis Services. Incluye 9 data mining procedures, 6 of them to be discussed in this course.
- Oracle,
- Statistica Miner,
- SAS Enterprise Miner,
- SPSS Clementine.
- XL Miner, written in Excel.
- Also specialized software to perform a specific data mining task.



```
Outlook = sunny
|   Humidity <= 77.500: yes {no=0, yes=2}
|   Humidity > 77.500: no {no=3, yes=0}
Outlook = overcast: yes {no=0, yes=4}
Outlook = rain
|   Wind = false: yes {no=0, yes=3}
|   Wind = true: no {no=2, yes=0}
(created by DecisionTree)
P May 24, 2009 10:37:25 AM: [NOTE] Process finished successfully
```



Classifier: J48 -C 0.25 -M 2

Test options:

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list: 10:55:13 - trees.J48

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

    === Detailed Accuracy By Class ===
```

Status: OK

Evaluation

- Homeworks (4) ………… 40%
- Partial exam………..30%
- Project. ……………. 30%

Course’s Content

- Introduction to Data Mining: 3 hrs.
- Data Preprocessing: 15 hrs.
- Visualization: 5 hrs.
- Outlier Detection 5 hrs
- Supervised Classification: 9 hrs.
- Clustering: 7 hrs

Motivation

The mechanisms for automatic recollection of data and the development of databases technology has made possible that a large amount of data can be available in databases, data warehouses and other repositories of information. Nowdays, there is the need to convert this data in knowledge and information.

“Every time the amount of data increases by a factor of ten we should totally rethink how we analyze it.” J.H.F. Friedman (1997). “Data Mining and Statistics, what is the connection”.

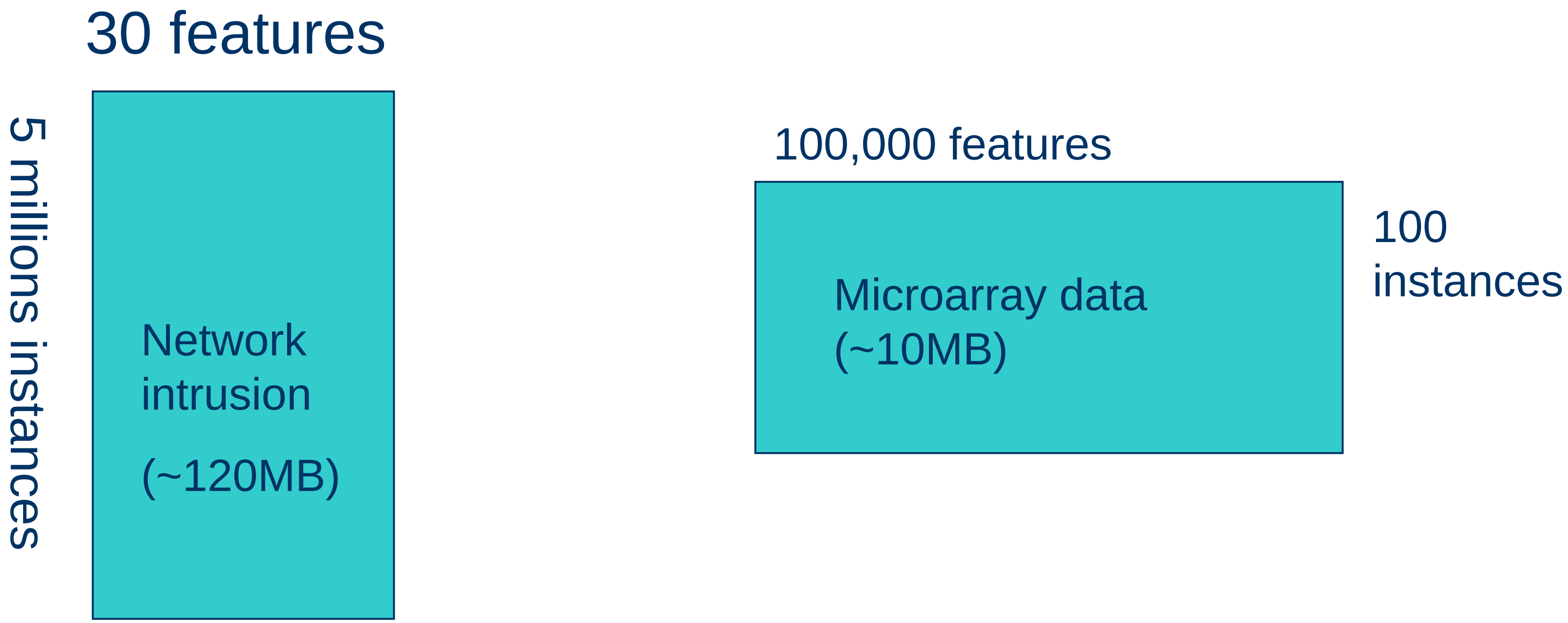
Size of datasets

Description	Size in Bytes	Mode of storage
very small	10^2	Piece of paper
Small	10^4	Several sheets of paper
Medium	10^6 (megabyte)	Floppy Disks
Large	10^9(gigabite)	A TV Movie
Massive	10^{12}(Terabyte)	A Hard Disk
Super-massive	10^{15}(Petabyte)	File of distributed data

Exabyte (10^{18} bytes), ZettaByte (10^{21} bytes), Yottabyte(10^{24} bytes)

Source: http://www.bergesch.com/bcs/storage.htm

Two different shape of datasets



30 features

5 millions instances

Network intrusion

(~120MB)

100,000 features

100 instances

Microarray data (~10MB)

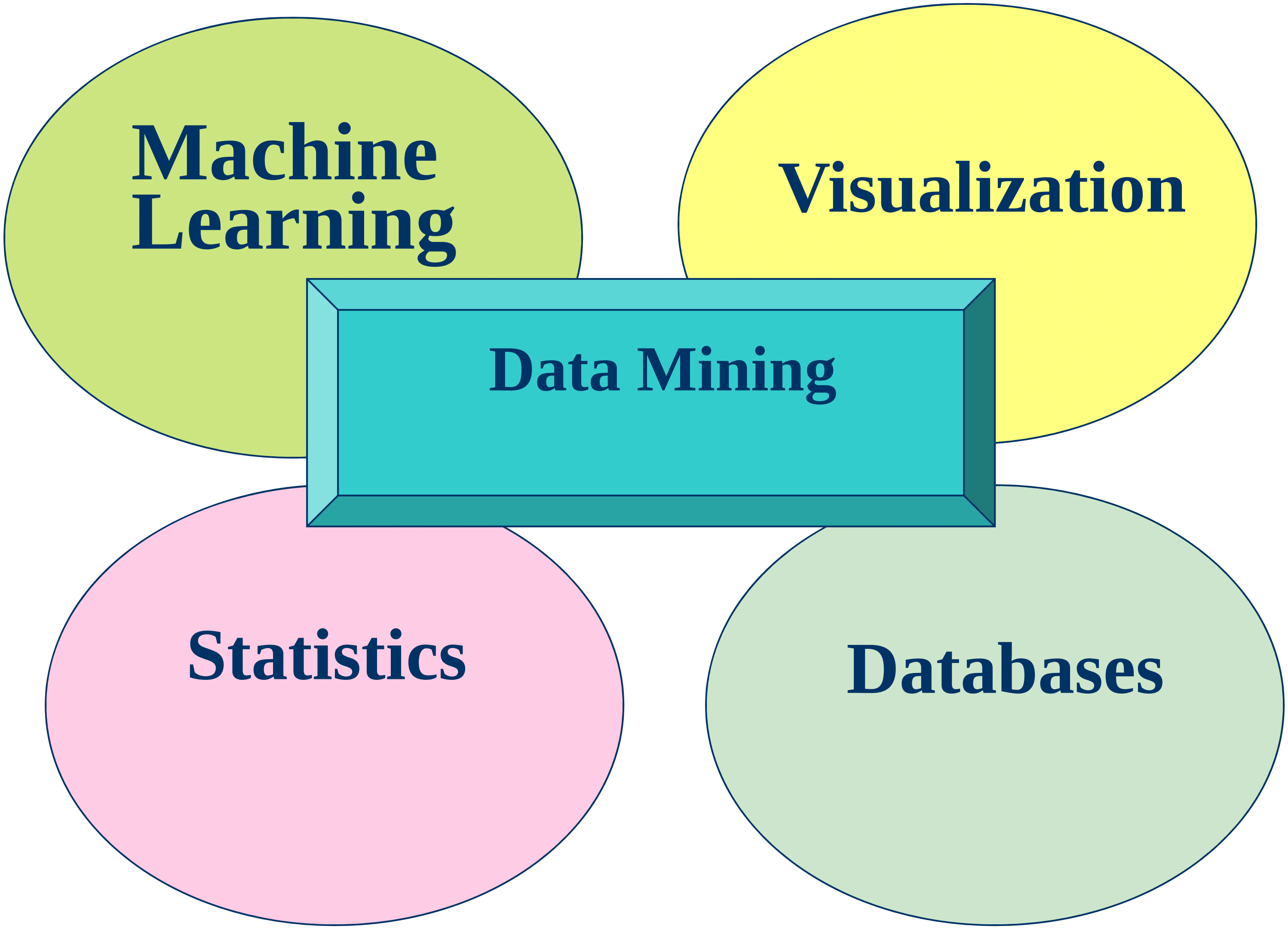
Examples of very large databases

- A telescope may generate up to 1 gigabyte of astronomical data in one second.
- ATT storages annually about 35 Terabytes of information in telephone calls (2006).
- Google searches in more than 1 trillion of internet pages representing more than 25 PetaBytes (2008).
- It is estimated that in 2002 more than 5 exabytes(5 millions of TB) of new data was generated.

What is data mining? What is KD?

- “Data mining is the process of extracting previously unknown comprehensible and actionable information from large databases and using it to make crucial business decision”. (Zekulin)
- **“Knowledge discovery** is the non-trivial extraction of implicit, unknown, and potentially useful information from data”. Fayyad et al. (1996).
- Other names: Knowledge discovery in databases (KDD), knowledge extraction, intelligent data analysis.
- **Currently: Data Mining and Knowledge Discovery are used interchangeably**

Related Areas Areas



Machine Learning

Visualization

Data Mining

Statistics

Databases

Statistics, Machine Learning

- Statistics (~40% of DM)
  - Based on theory. Assume distributional properties of the features being considered.
  - Focused in testing of hypothesis, parameter estimation and model estimation (learning process).
  - Efficient strategies for data recollection are considered.
- Machine learning (~25 % of DM)
  - Part of Artificial Intelligence.
  - More heuristic than Statistics.
  - Focused in improvement of the performance of a classifier based on prior experiences.
  - Includes: Neural Networks (Eng), decision trees (Stat), Naïve Bayes, Genetic algorithms (CS).
  - Includes other topics such as robotics that are unrelated to data mining

Visualization, databases

- Visualization (~15 % of DM)
  - The dataset is explored in a visual fashion.
  - It can be used in either pre or post processing step of the Knowledge discovery process.
- Relational Databases (~20% of DM)
  - A relational database is a set de tables and their schemas which define the structure of tables. Each table has a primary key that is used to uniquely define every record (row) in table. Foreign keys are used to define the relations between different tables in databases.
  - The goal for an RDBMS is to maintain the data (in tables) and to quickly located the requested data.
  - The most used interface between the user and the relational database is SQL(structured query language).

DM Applications

Science: Astronomy, Bioinformatics (Genomics, Proteonomics, Metabolomics), drug discovery.

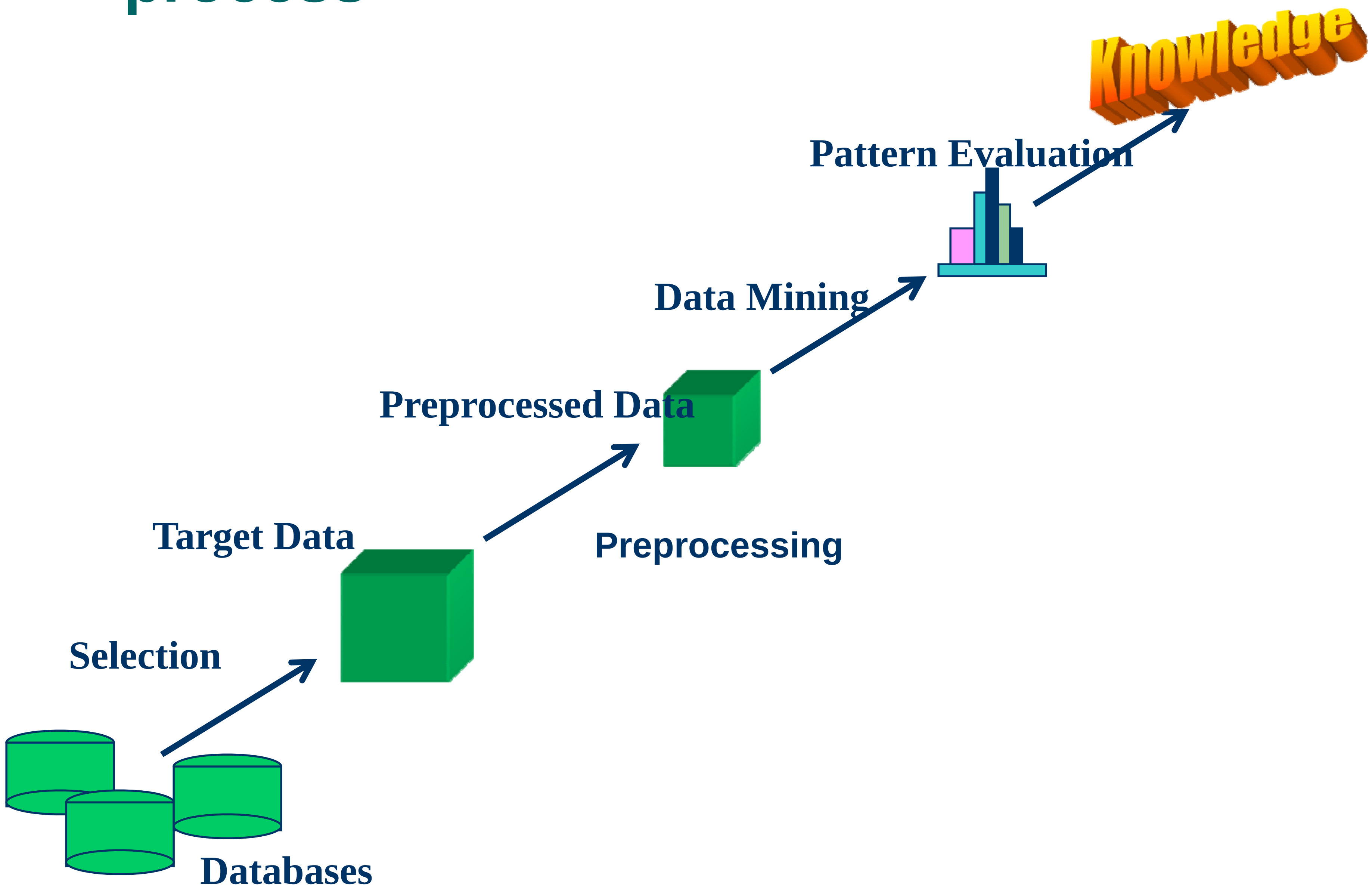
Business: Marketing, credit risk, Security and Fraud detection,

Govermment: detection of tax cheaters, anti-terrorism.

Text Mining: Discover distinct groups of potential buyers according to a user text based profile. Draw information from different written sources (e-mails).

Web mining: Identifying groups of competitors web pages. E-commerce (Amazon.com)

Data Mining as one step of the KDD process



Databases

Selection

Target Data

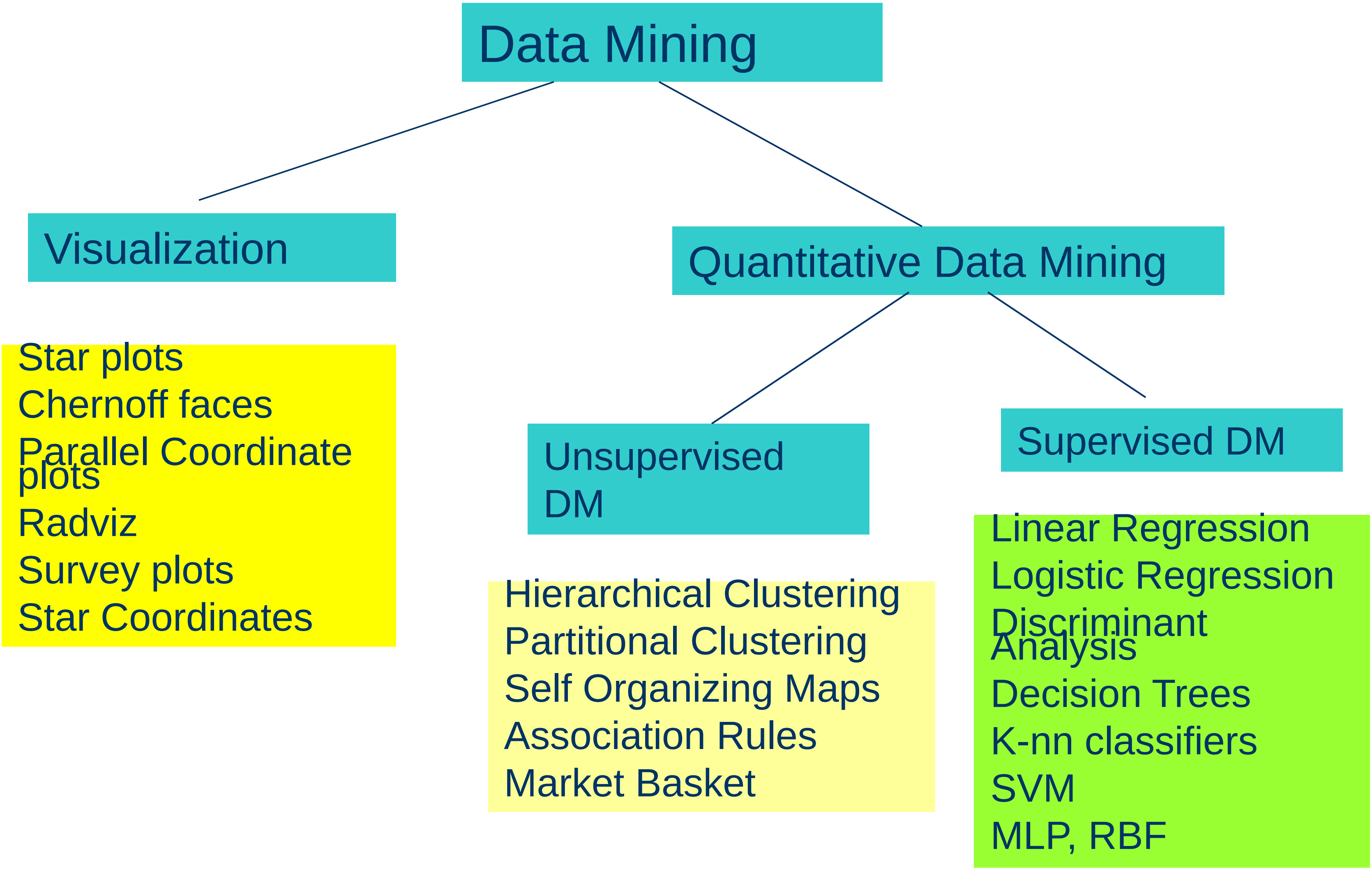
Preprocessing

Preprocessed Data

Data Mining

Pattern Evaluation

Knowledge



Data Mining

Visualization

- Star plots
- Chernoff faces
- Parallel Coordinate plots
- Radviz
- Survey plots
- Star Coordinates

Quantitative Data Mining

Unsupervised DM

- Hierarchical Clustering
- Partitional Clustering
- Self Organizing Maps
- Association Rules
- Market Basket

Supervised DM

- Linear Regression
- Logistic Regression
- Discriminant Analysis
- Decision Trees
- K-nn classifiers
- SVM
- MLP, RBF

Types of data mining tasks

- Descriptive: General properties of the database are determined. The most important features of the databases are discovered.
- Predictive: The collected data is used to train a model for making future predictions. Never is 100% accurate and the most important matter is the performance of the model when is applied to future data.

Data mining tasks

- Regression (predictive)
- Classification (predictive)
- Unsupervised Classification –Clustering (descriptive)
- Association Rules (descriptive)
- Outlier Detection (descriptive)
- Visualization (descriptive)

Regression

- The value of a continuous response variable is predicted based on the values of other variables (predictors), assuming that there is a functional relation among them.
- Statistical models, decision trees, neural networks can be used.
- Examples: car sales of dealers based on the experience of the sellers, advertisament, type of cars, etc.

Regresion[2]

- Linear Regression $Y=b_o+b_1X_1+\ldots..b_pX_p$
- Non-Linear Regression, $Y=g(X_1,\ldots,X_p)$, where g is a non-linear function. For example, $g(X_1,\ldots X_p)=X_1\ldots X_pe^{X1+\ldots Xp}$
- Non-parametric Regression $Y=g(X_1,\ldots,X_p)$, where g is estimated using the available data.

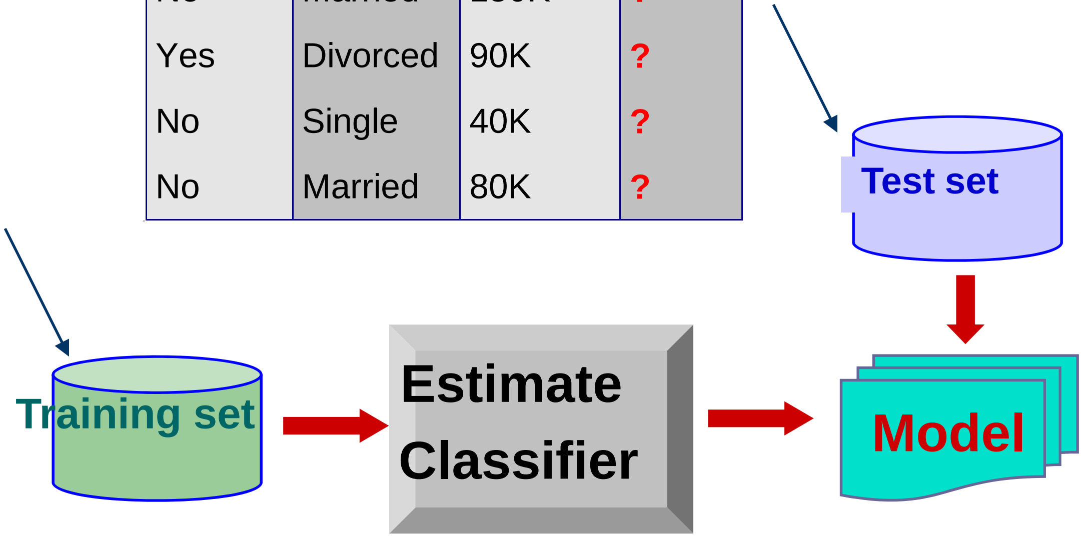
Supervised Classification

- The response variable is categorical.
- Given a set of records, called the training set (each record contains a set of attributes and usually the last one is the class), a model for the attribute class as a function of the others attributes is constructed. The model is called the classifier.
- *Goal: Assign records previously unseen (test set) to a class as accurately as possible.*
- Usually a given data set is divided in a training set and a test set. The first data set is used to construct the model and the second one is used to validate. The precision of the model is determined in the test data set.
- It is a decision process.

Example: Supervised Classification

	categorical	categorical	continuous	class
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**

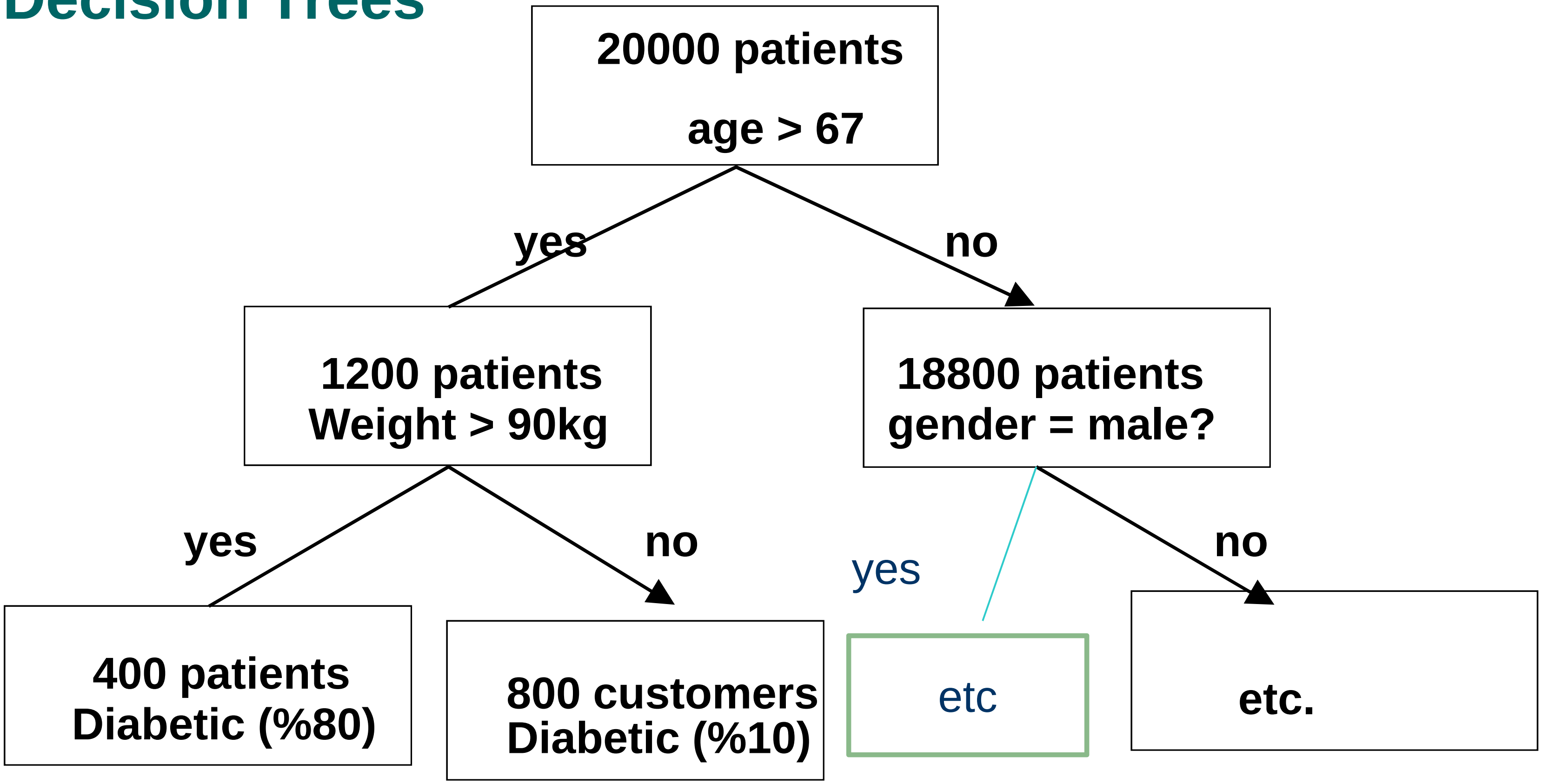
Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



Examples of Classification Techniques

- Linear Discriminant Analysis
- Naïve Bayes
- Decision trees
- K-Nearest neighbors
- Logistic regression
- Neural networks
- Support Vector Machines
- …..

Example Classification Algorithm 1
Decision Trees



20000 patients

age > 67

yes

no

1200 patients
Weight > 90kg

18800 patients
gender = male?

yes

no

yes

no

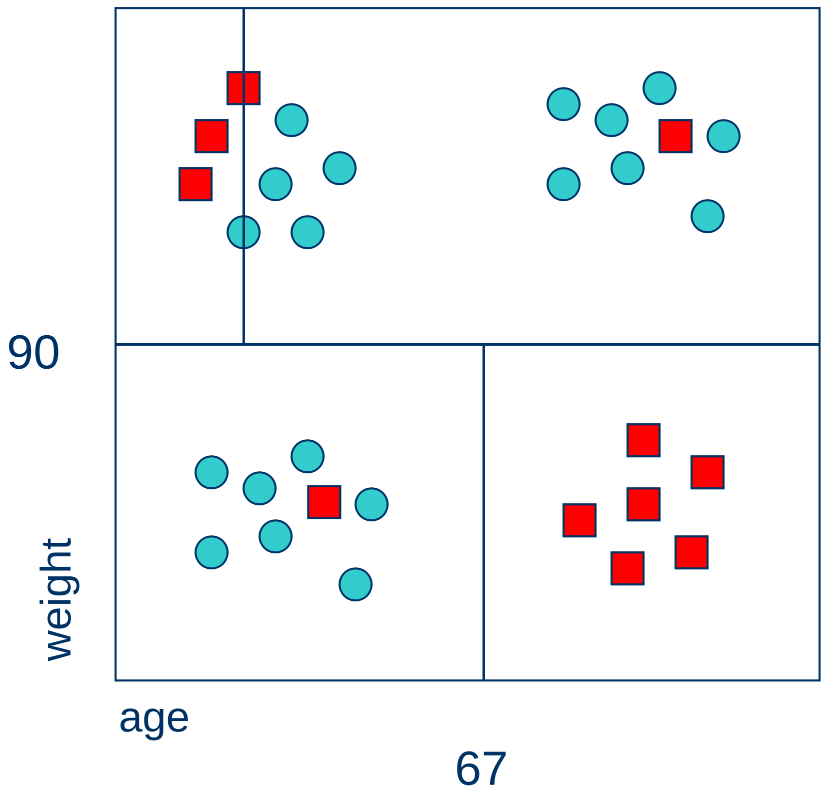
400 patients
Diabetic (%80)

800 customers
Diabetic (%10)

etc

etc.

Decision Trees in Pattern Space



The goal’s classifier is to separate classes [circle(non-diabetic), square (diabetic)] on the basis of attribute age and weight

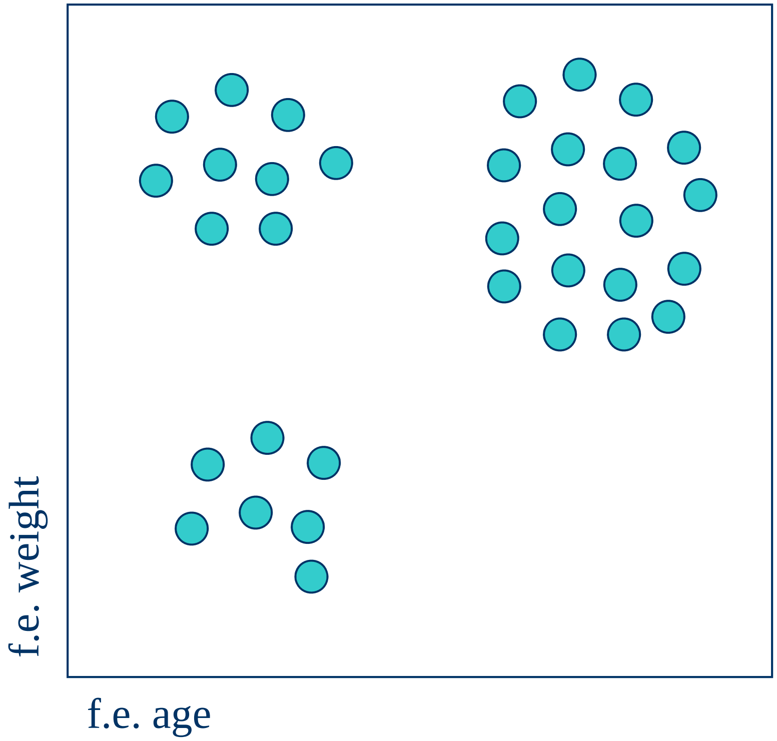
Each line corresponds to a split in the tree

Decision areas are ‘tiles’ in pattern space

Unsupervised Classification (Clustering)

- Find out groups of objects (clusters) such as the objects within the same clustering are quite similar among them whereas objects in distinct groups are not similar.
- A similarity measure is needed to establish whether two objects belong to the same cluster or to distinct cluster.
- Examples of similarity measure: Euclidean distance, Manhattan distance, correlation, Grower distance, hamming distance, etc.
- Problems: Choice of the similarity measure, choice of the number of clusters, cluster validation.

Data Mining Tasks: Clustering

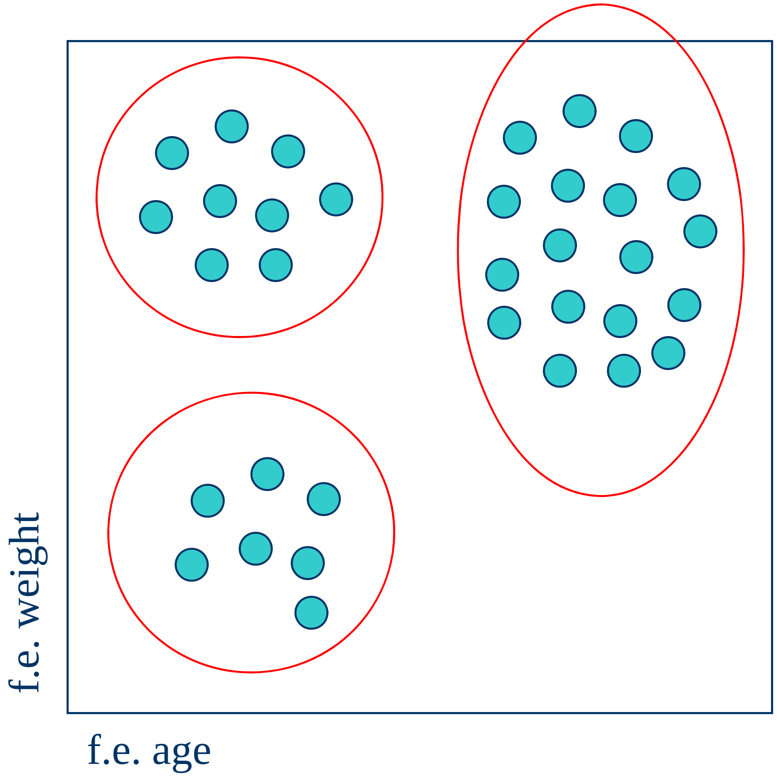


Clustering is the discovery of groups in a set of instances

Groups are different, instances in a group are similar

In 2 to 3 dimensional pattern space you could just visualise the data and leave the recognition to a human end user

Data Mining Tasks: Clustering



Clustering is the discovery of groups in a set of instances

Groups are different, instances in a group are similar

In 2 to 3 dimensional pattern space you could just visualize the data and leave the recognition to a human end user

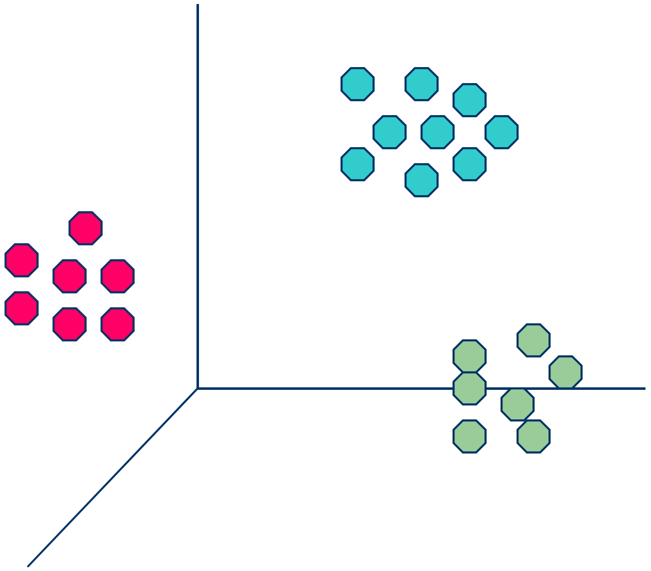
In >3 dimensions this is not possible

Clustering[2]

☒ Tri-dimensional clustering based on euclidean distance.

The Intracluster distances are minimized

The Intercluster distances are maximized



Clustering Algorithms

- Partitioning algorithms: K-means, PAM, SOM.
- Hierarchical algorithms: Agglomerative, Divisive.
- Gaussian Mixtures Models.
- ……………

Outlier Detection

- The objects that behave different or that are inconsistent with the majority of the data are called outliers.
- Outliers arise due to mechanical faults, human error, instrument error, fraudulent behavior, changes ithe system, etc . They can represent some kind of fraudulent activity.
- The goal of outlier detection is to find out the instances that do not have a normal behavior.

Outlier Detection [2]

- Methods:
  - based on Statistics.
  - based on distance.
  - based on local density.
- Application: Credit card fraud detection, Network intrusion

Association Rules discovery

- Given a set of records each of which contain some number of items from a given collection.

The aim is to find out dependency rules which will predict occurrence of an item based on occurrences of other items

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

Rules discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Reglas de Asociacion[2]

- The rules (X->Y) must satisfy a minimum support and confidence set up by the user. X is called the antecedent and Y is called the consequent.
- Support=(# records containing X and Y)/(# records)
- Confidence=(# records containing X and Y/(# de records containing X)

Example: The first rule has support .6 and the second rule has support .4.

The confidence of rule 1 is .75 and for the rule 2 is .67

Applications: Marketing and sales promotion.

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Privacy
- Streaming Data