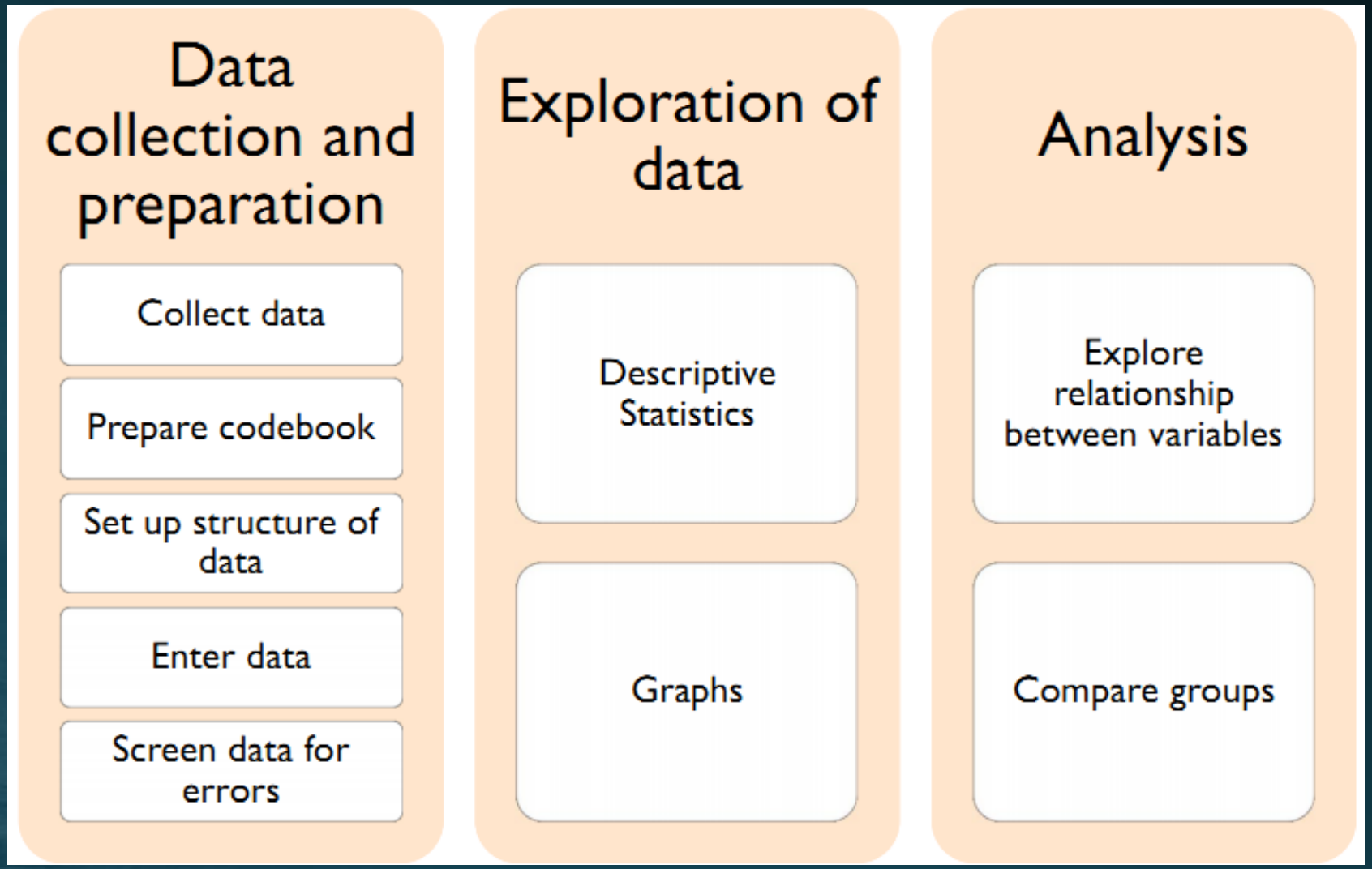
Iris data analysis example

Author: Do Thi Duyen

Overview: data analysis process



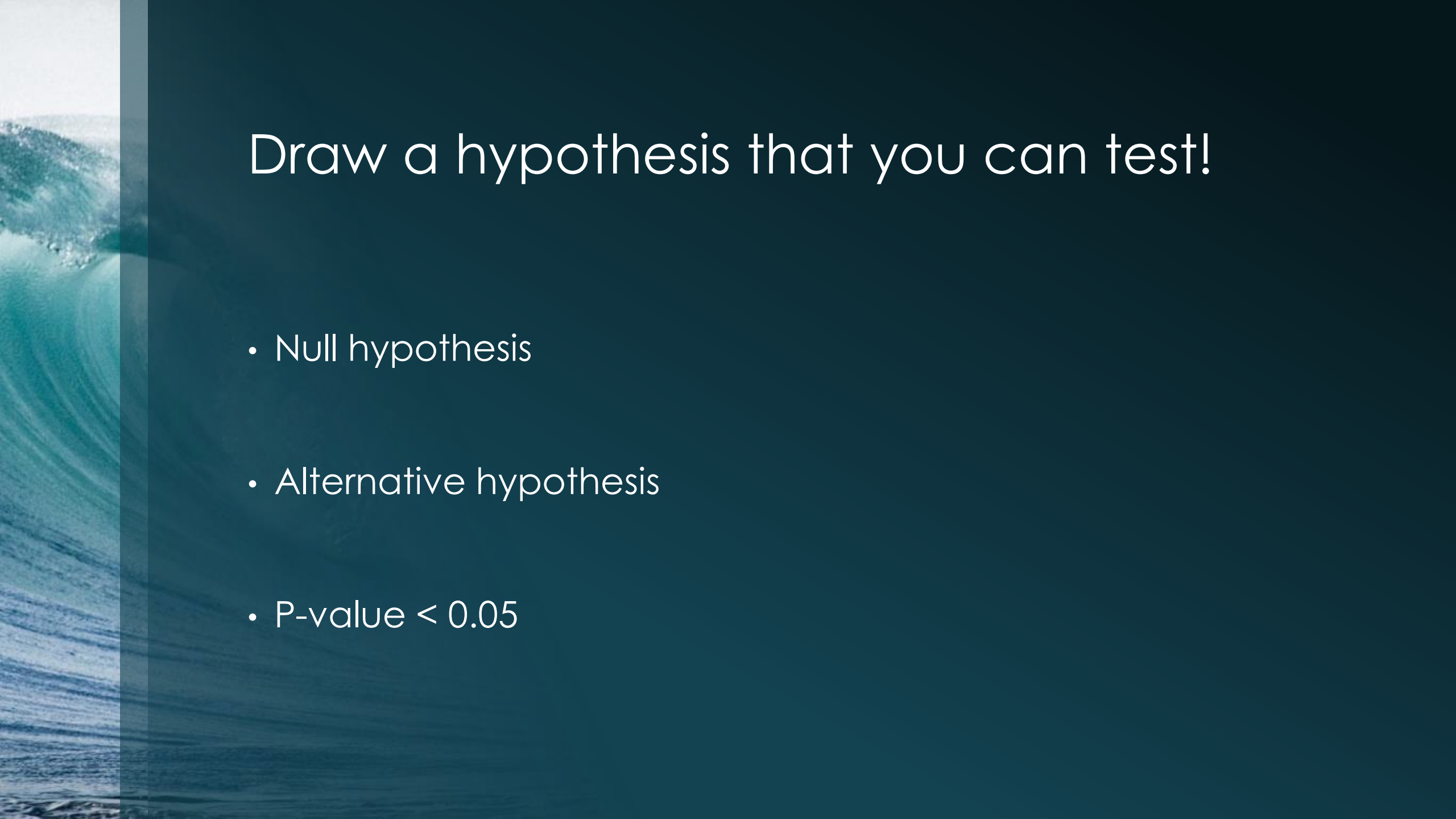
Iris setosa

Iris versicolor

Iris virginica

Iris flower data set

- Also called Fisher's Iris data set or Anderson's Iris data set
- Collected by Edgar Anderson and Gaspé Peninsula
- To quantify the morphologic variation of Iris flowers of three related species
- >iris



Draw a hypothesis that you can test!

- Null hypothesis
- Alternative hypothesis
- P-value < 0.05

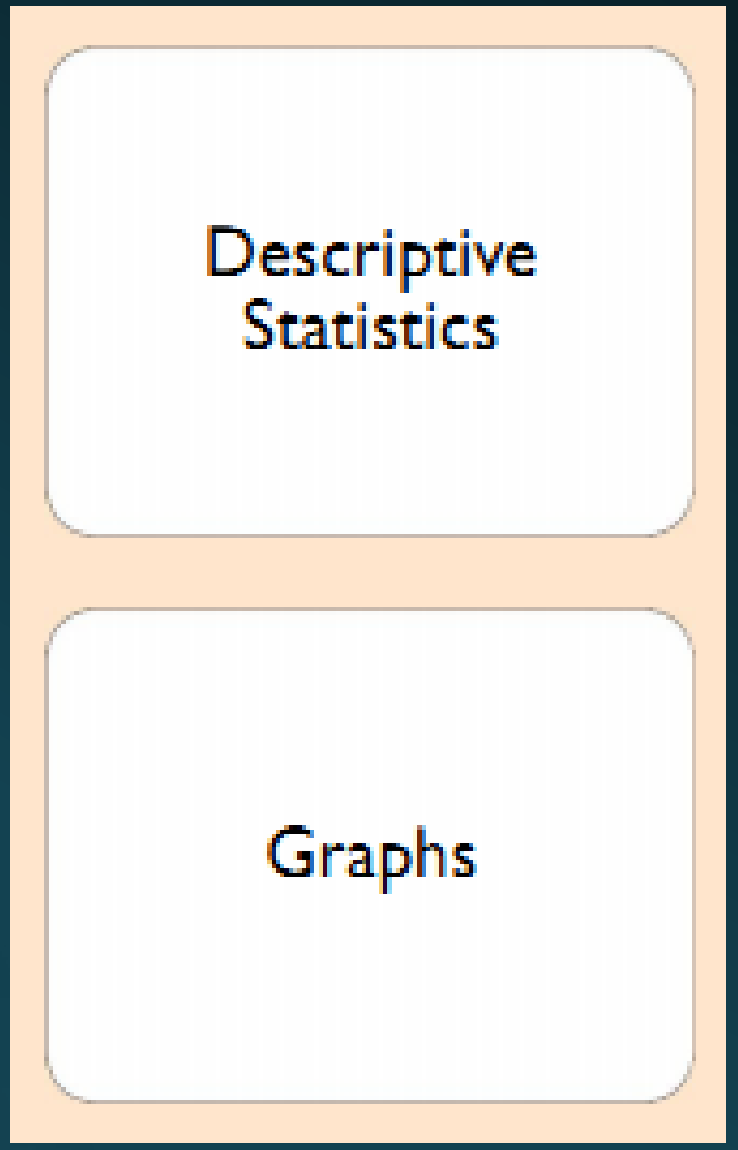
Get data!

Some ways to read data in R:

- read.table, read.csv, read.xls, data.frame,..
- edit,…
- …

=> Hint: Never modify your raw data file; always work on a copy!

Exploration of data



Some basic function in R to examine iris data:

```
>?iris
>names(iris)
>iris
>str(iris)
>iris$new_class_specis<-as.character(iris$Species)
>iris$new_class_specis<-NULL
>iris$Species <- gsub("%","",iris$Species))
>iris<-na.omit(iris)
```

Summarize and plot your data!

>summary(iris)

>plot(iris)

Descriptive statistics

Categorical:

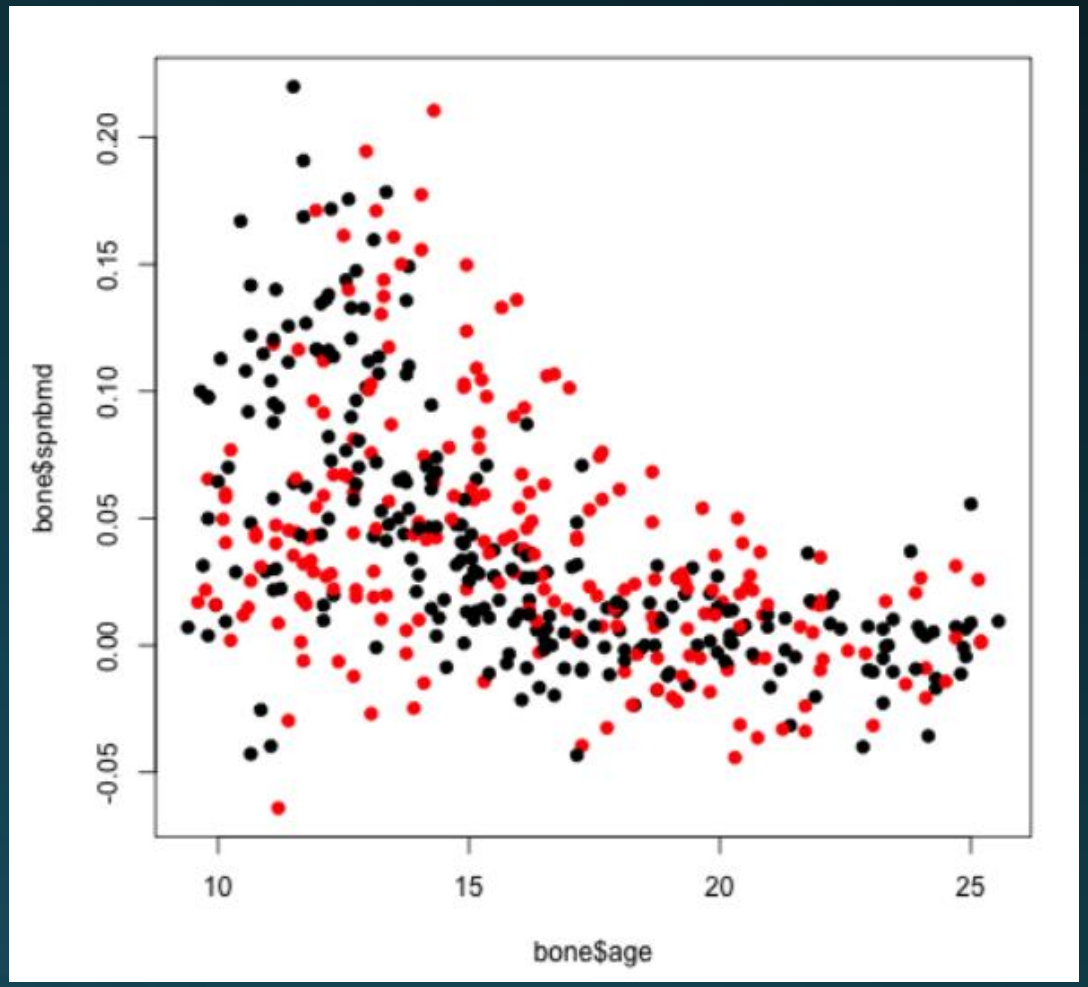
Frequencies

Numerical:

Descriptives:

- mean
- standard deviation
- minimum
- maximum
- skewness (symmetry)
- kurtosis (peakness)

Scatter plot



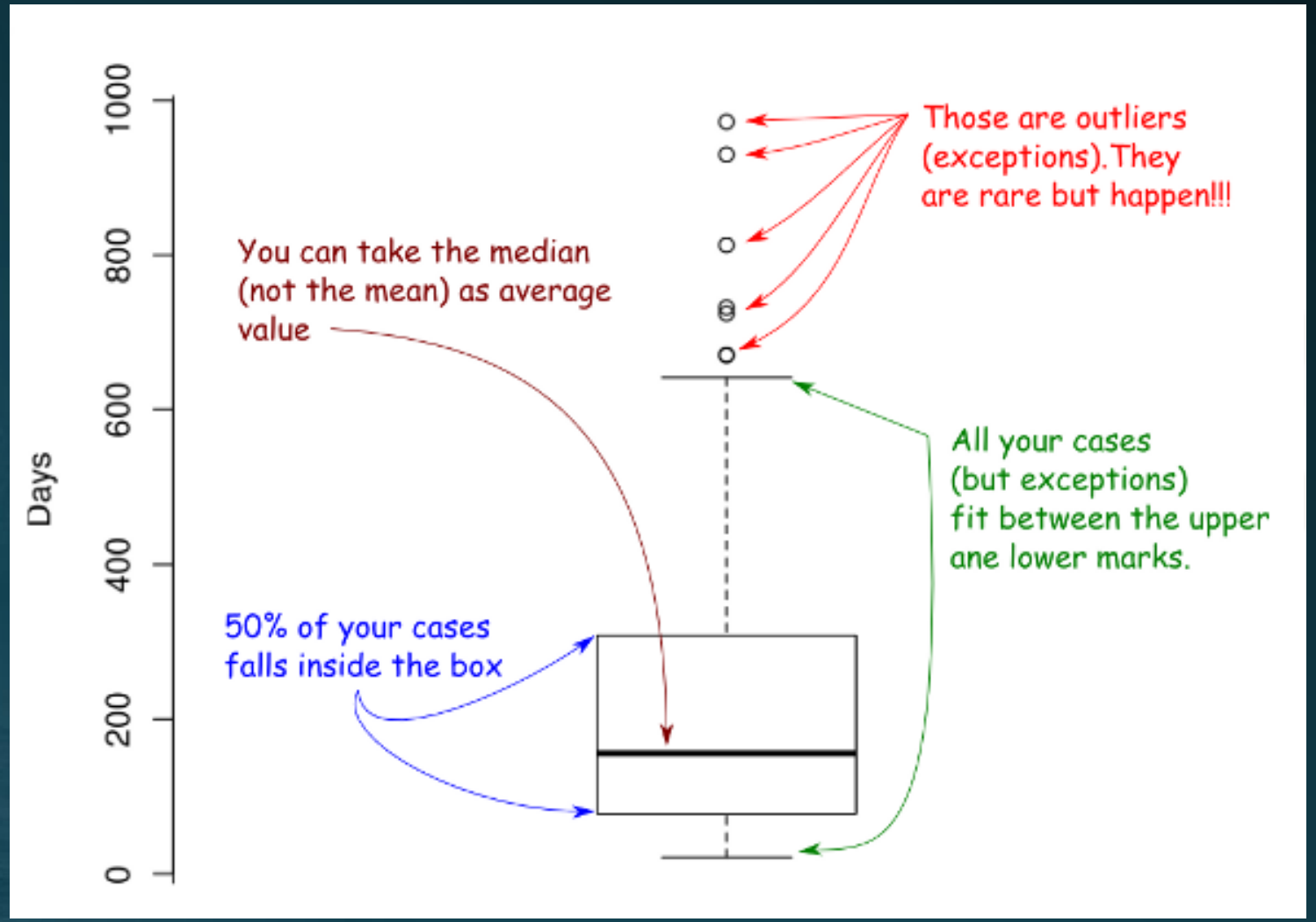
Scatter plot

```
plot(iris, col=iris$Species)

legend(7,4.3,unique(iris$Species),col=1:length(iris$Species),pch=1)
```

Lattice library

Ggplot2 library

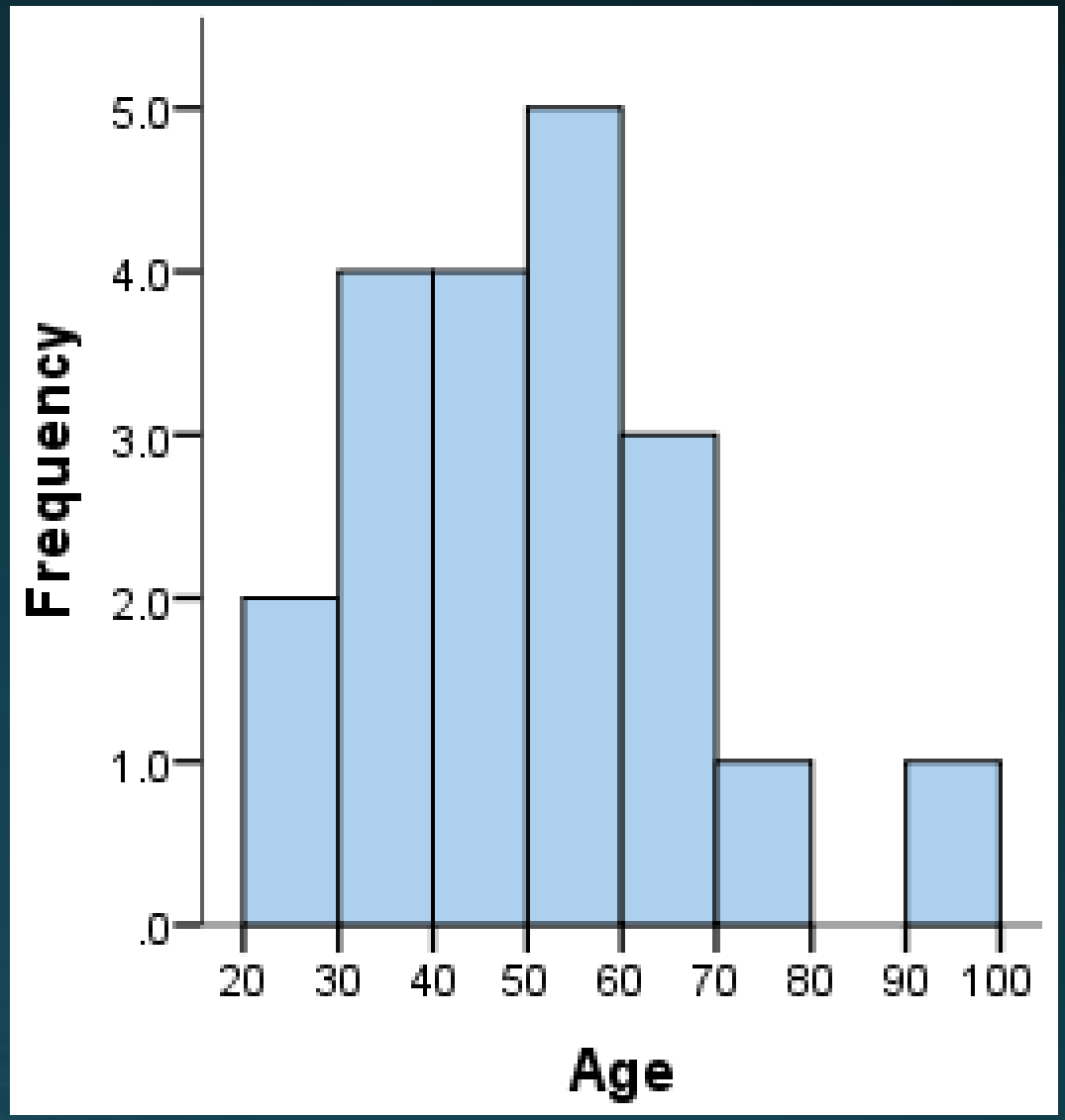
Box plot


Days
0
200
400
600
800
1000
Those are outliers (exceptions). They are rare but happen!!!
You can take the median (not the mean) as average value
All your cases (but exceptions) fit between the upper ane lower marks.
50% of your cases falls inside the box

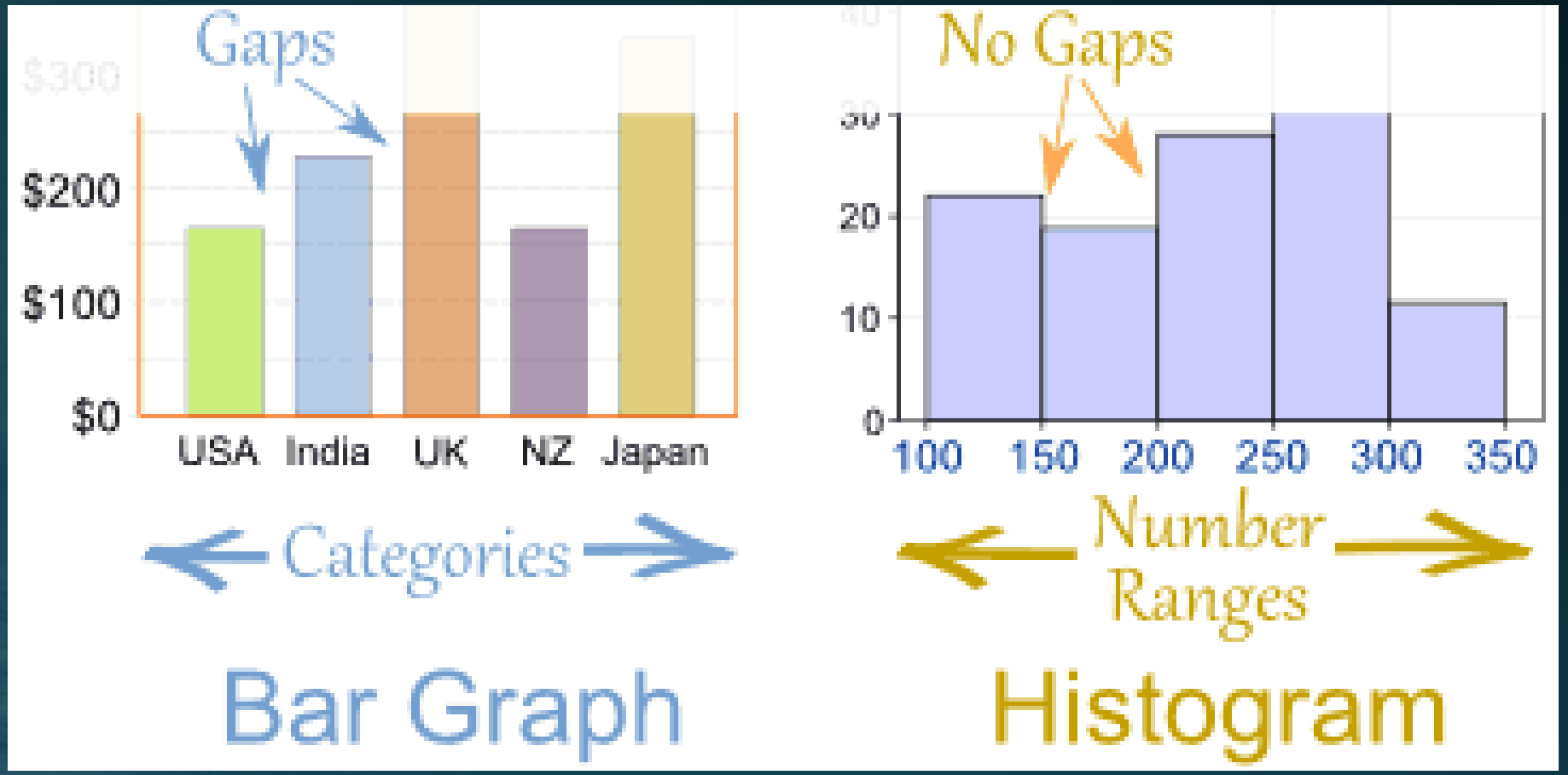
Box plot

```
>par(mfrow=c(1,2))

> plot(iris$Petal.Length)

>boxplot(iris$Petal.Length~ iris$Species)


> par(mfrow=c(2,2)) # to draw four figs in one window

> for(i in 1:4) boxplot(iris[,i] ~ Species, data=iris,
main=names(iris)[i])
```

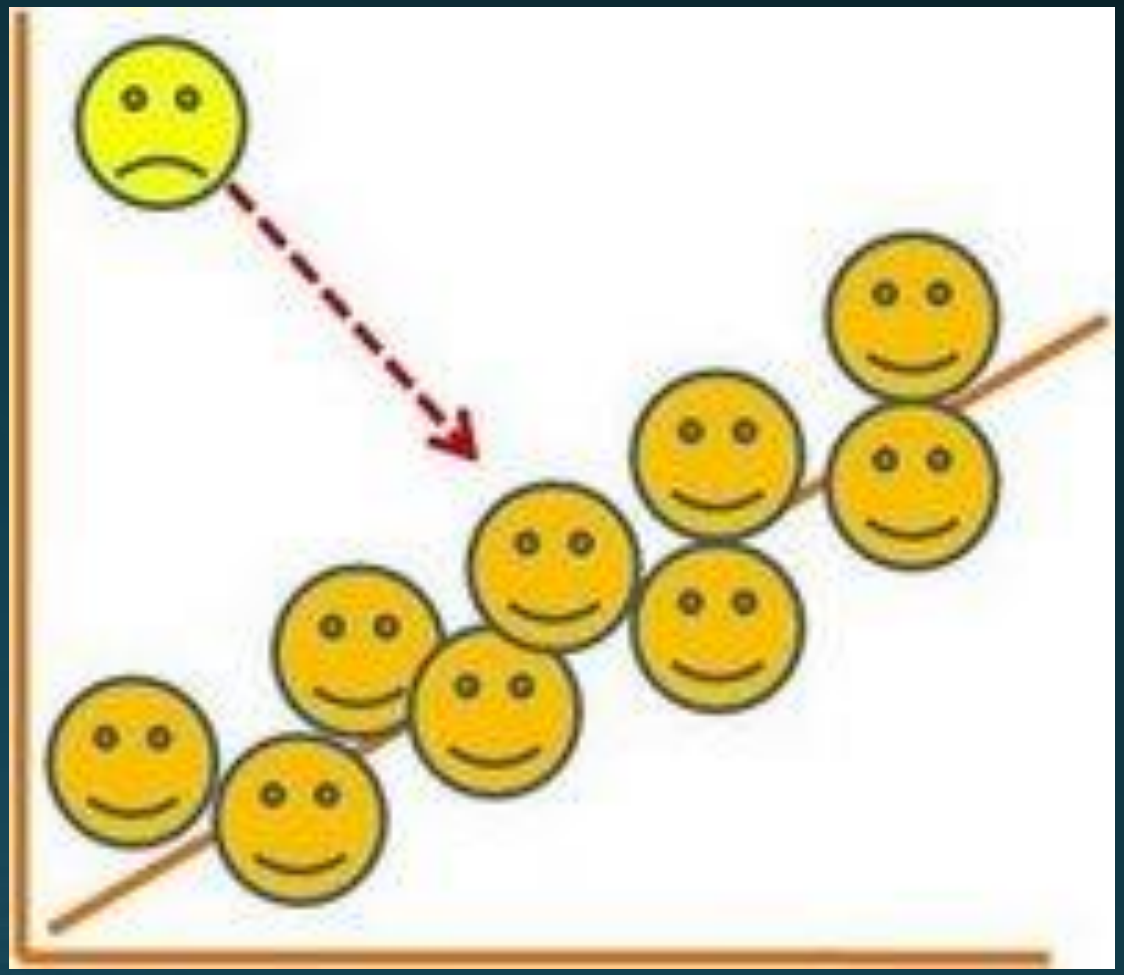
Histogram



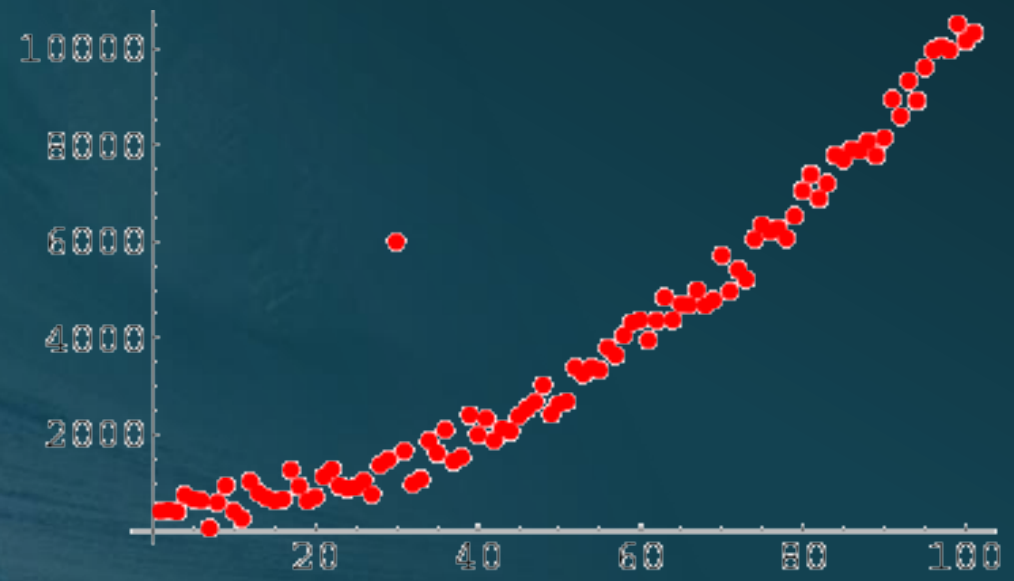
Histogram vs Bar chart

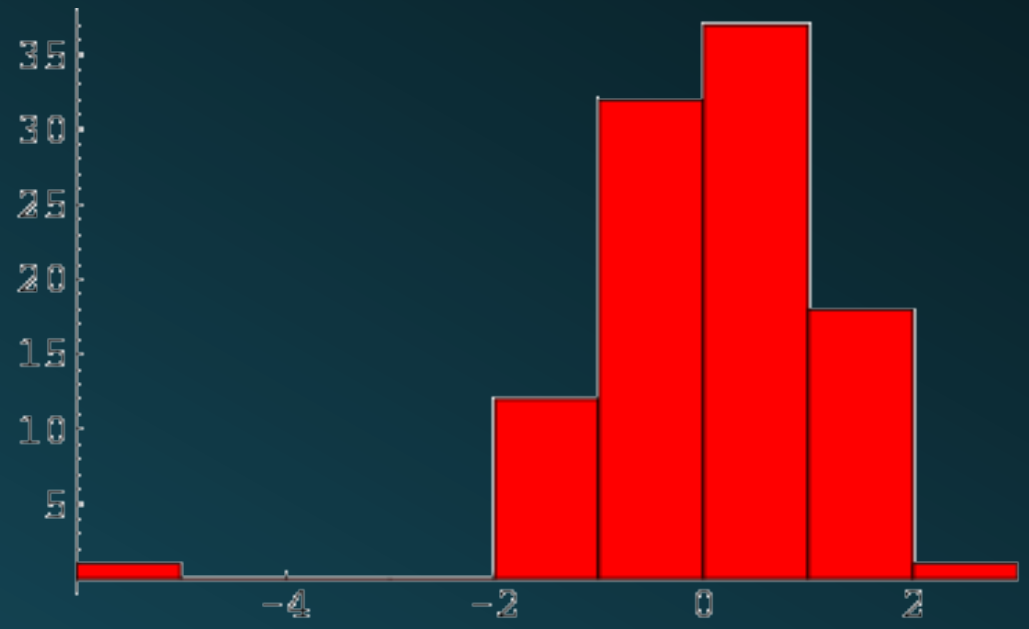


Outlier



Outlier





Histogram

```
> par(mfrow=c(1,1))
>hist(iris$Petal.Length[1:50])
```

Subsetting:

```
>iris$Sepal.Length[1:50]
>iris$Sepal.Length[-(1:50)]
```

Select by name:

```
>iris$Sepal.Length[iris$Species == "setosa"]
```

Change the order of data frame:

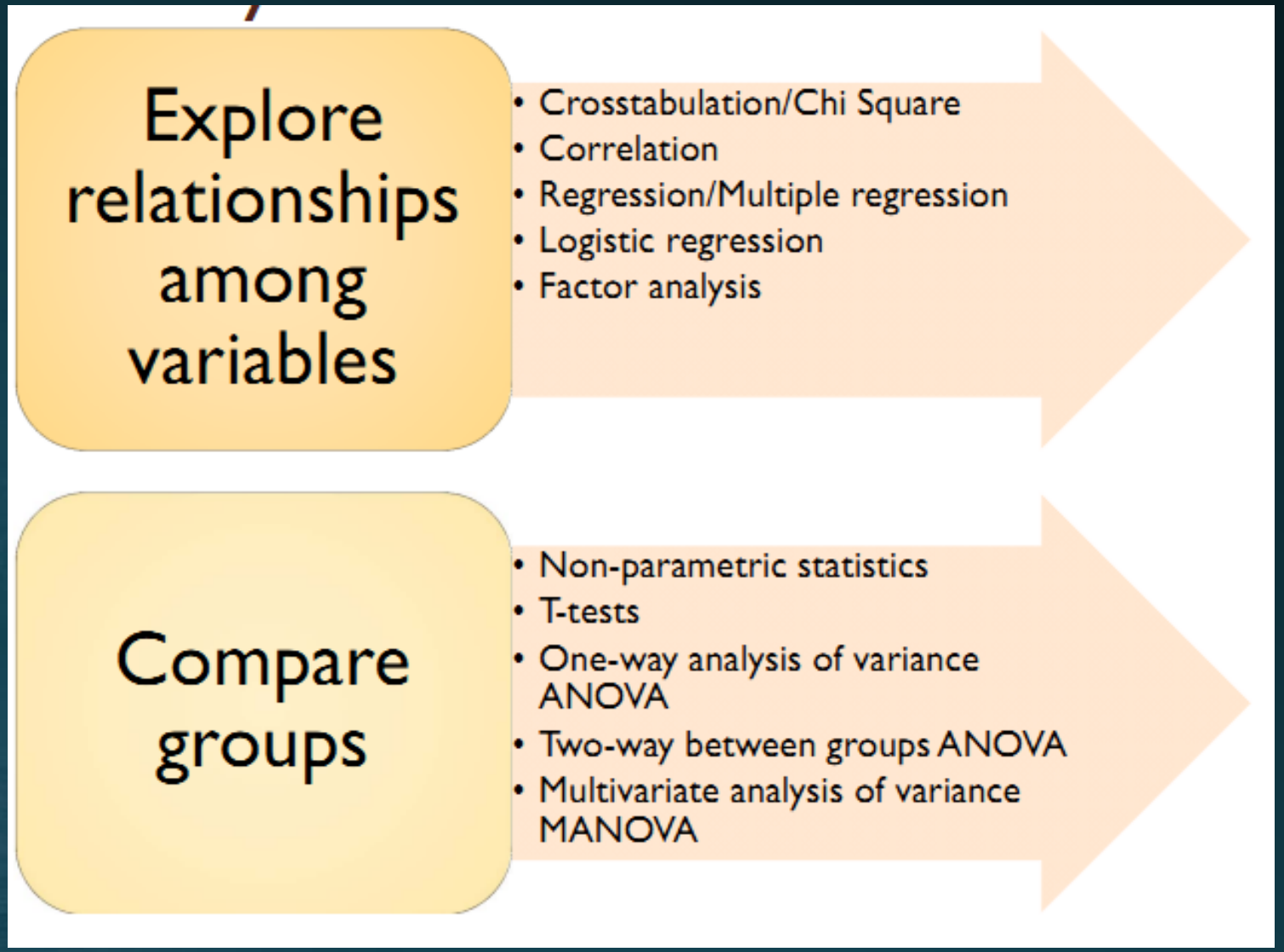
```
>iris.ordered<-iris[order(iris$Sepal.Length),]
```

Build a statistical model!

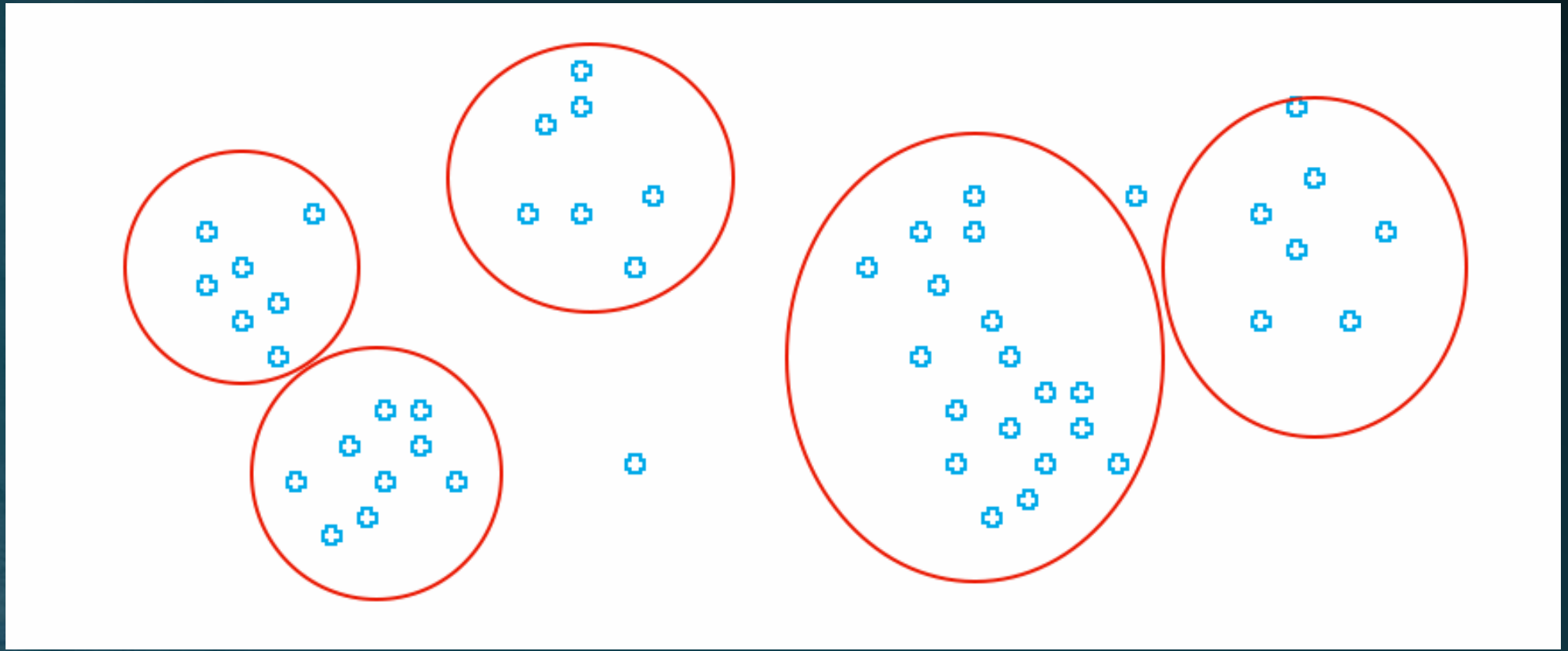
Data mining:

- Predict:
  - Classification
  - Regression
  - Deviation detection
- Descript:
  - Clustering
  - Association Rule Discovery

Analysis


Explore relationships among variables
• Crosstabulation/Chi Square
• Correlation
• Regression/Multiple regression
• Logistic regression
• Factor analysis
Compare groups
• Non-parametric statistics
• T-tests
• One-way analysis of variance ANOVA
• Two-way between groups ANOVA
• Multivariate analysis of variance MANOVA

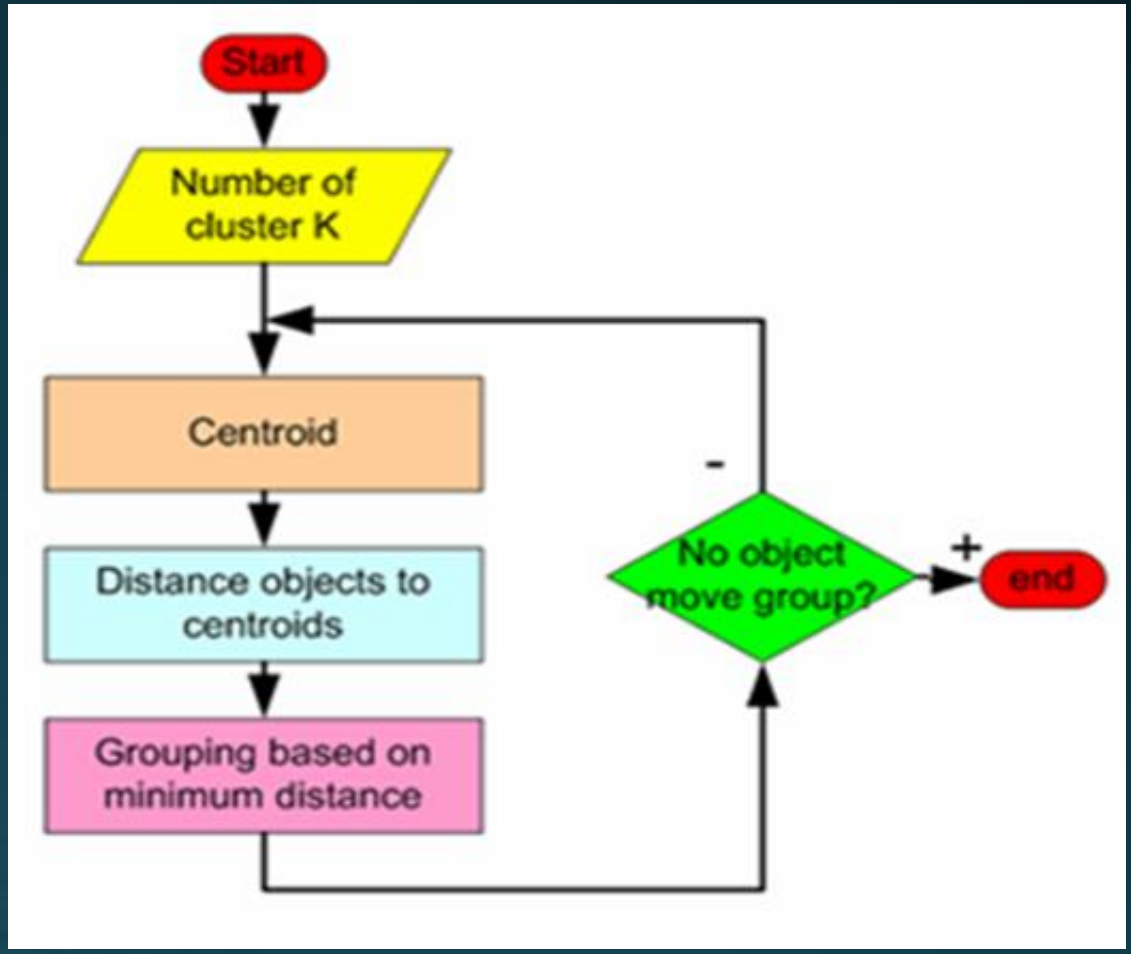
Clustering



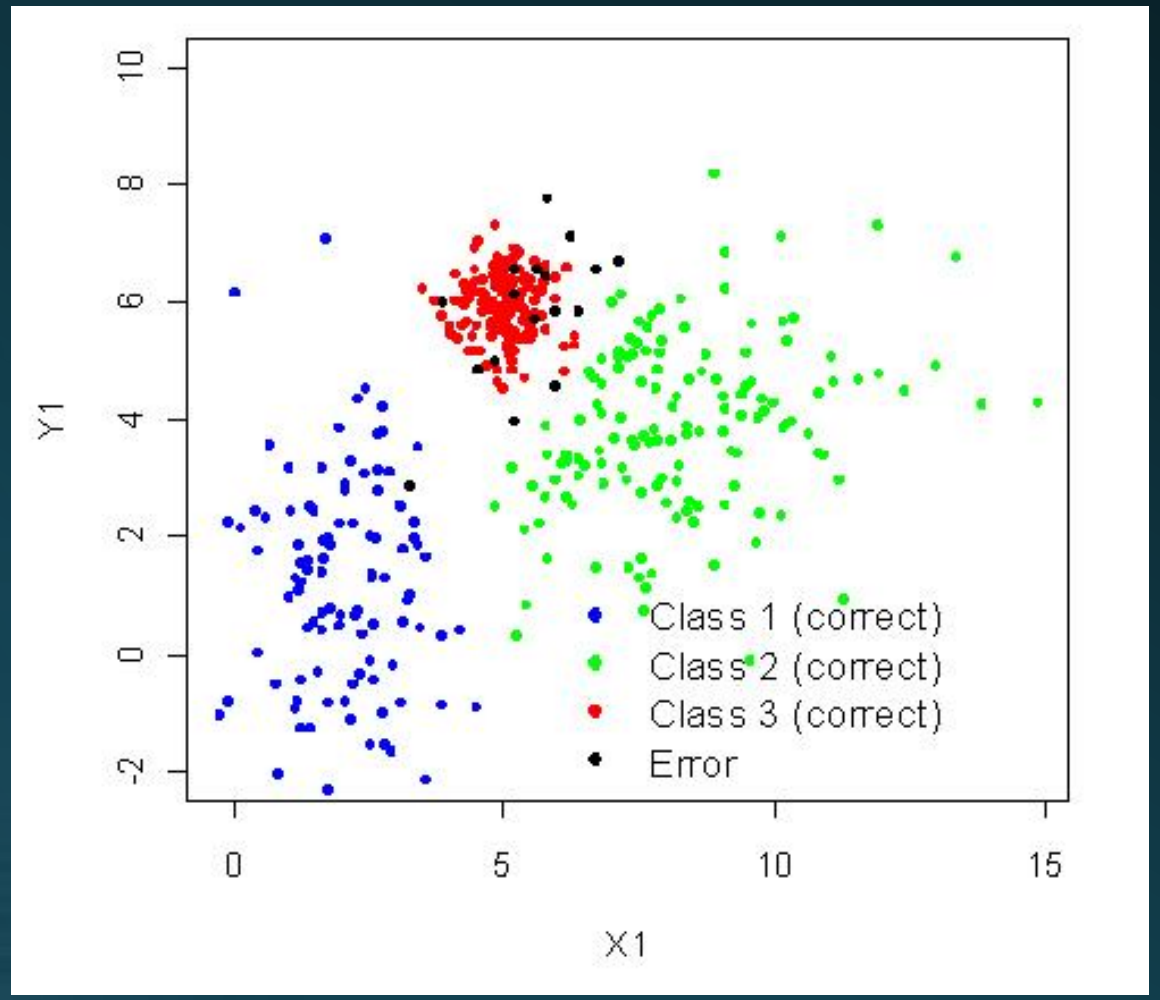
Clustering

- Principle: based on measure of distances
- Algorithms:
  - Hierarchical clustering: bottom up, top down
  - Centroid-based clustering: k-mean, PAM, CLARA, CLARANS,..
  - Distribution-based clustering: STING, WAVECluster, CLIQUE,..
  - Density-based clustering: DBSCANS, OPTICS, DENCLUE,..
  - Model-based cluatering: statistical model + Neural network
  - ...

K-Mean

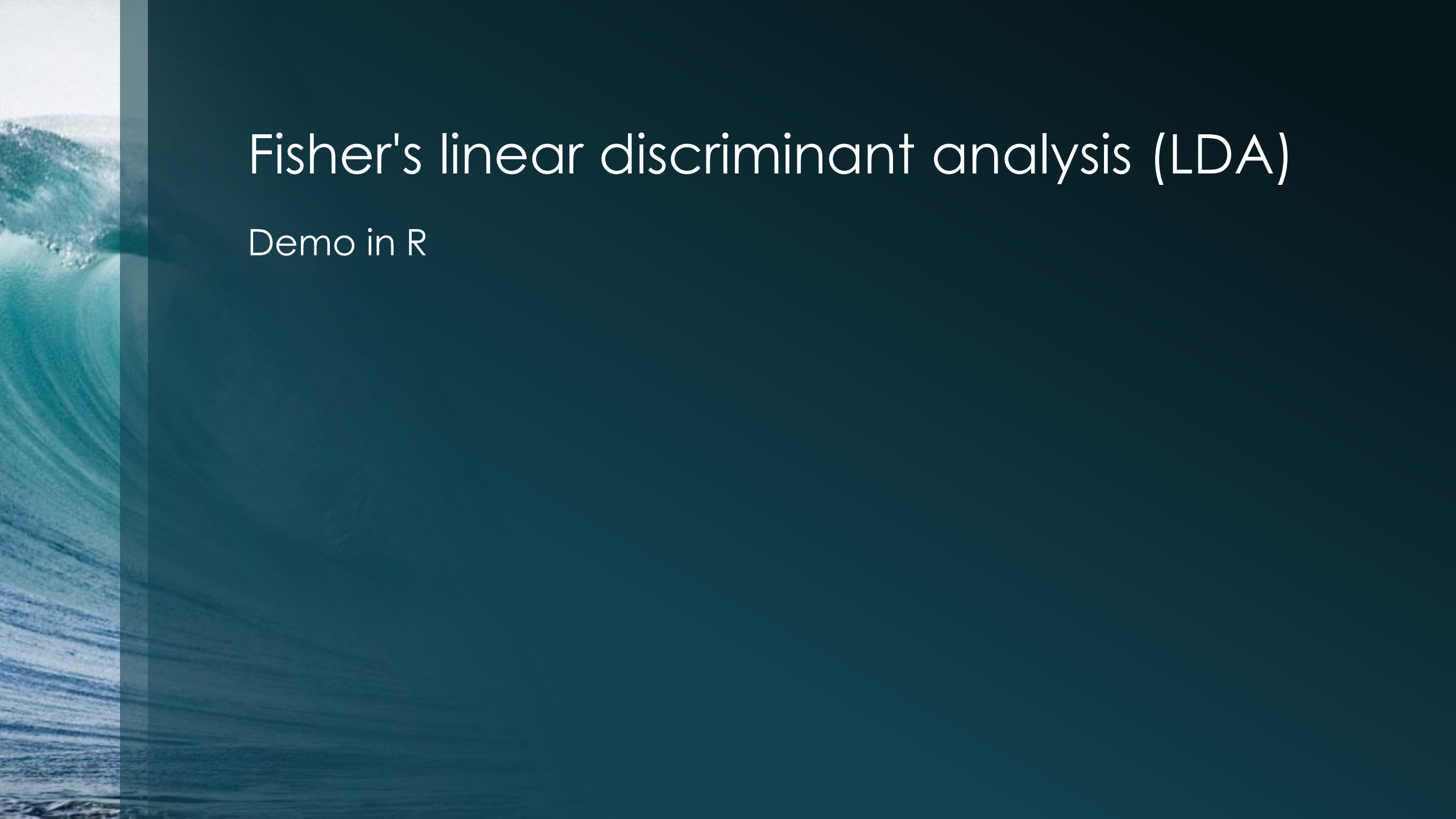

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Classification



Clasification algorithms

- Linear classifiers: Fisher's linear discriminant analysis, Naive Bayes classifier,..
- Support vector machines: Least squares support vector
- Quadratic classifiers
- Kernel estimation: k-nearest neighbor
- Boosting (meta-algorithm)
- Decision trees: Random forests
- …

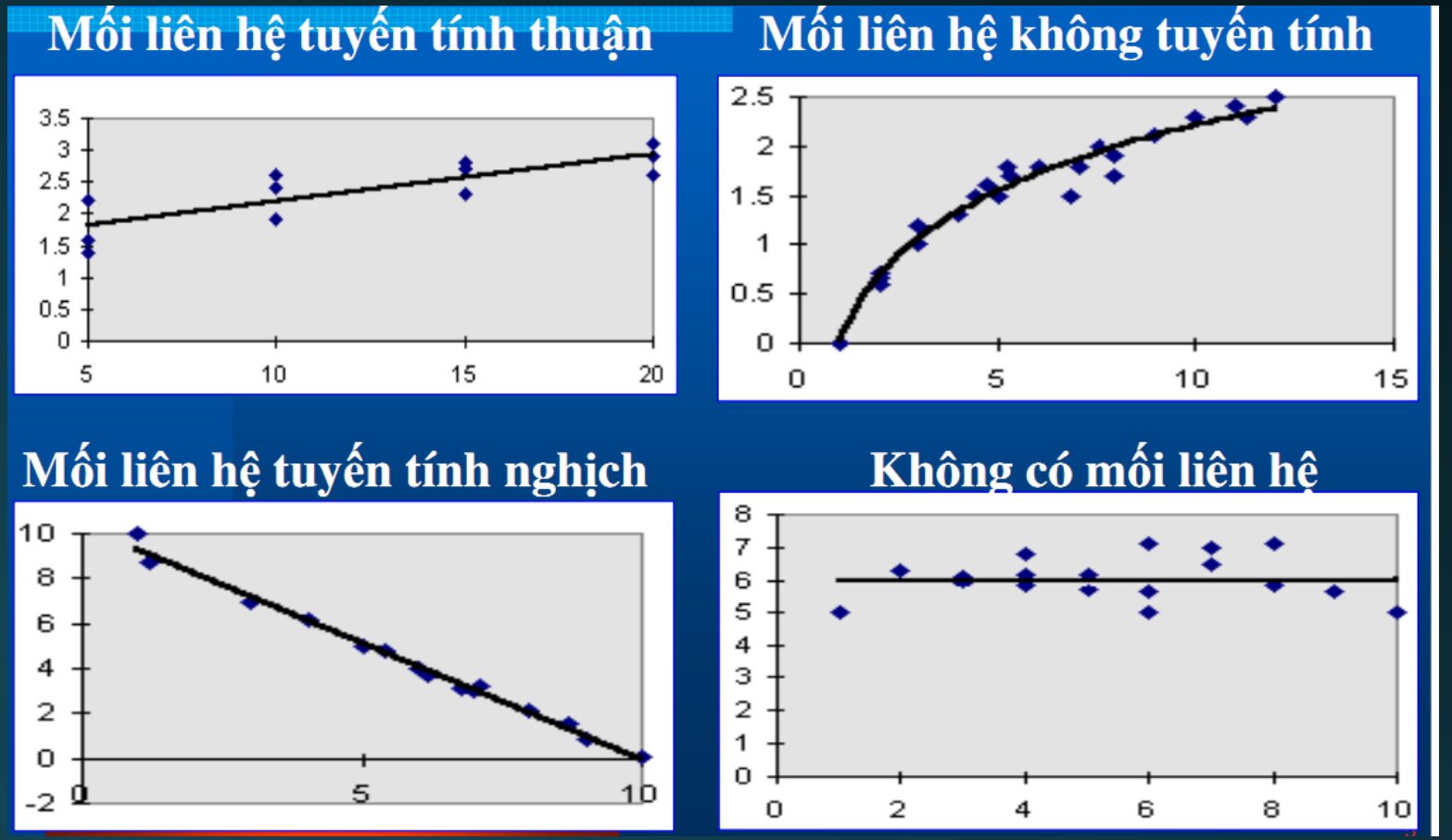


Fisher's linear discriminant analysis (LDA)

Demo in R

Regression analysis

$$Y \approx f(X, \beta)$$

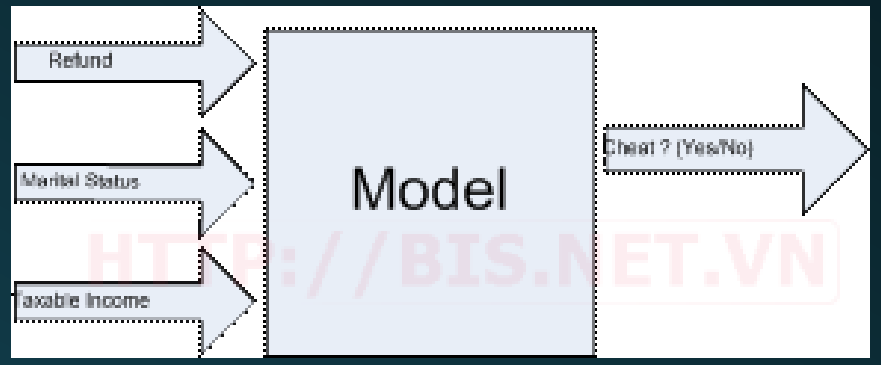


Regression analysis

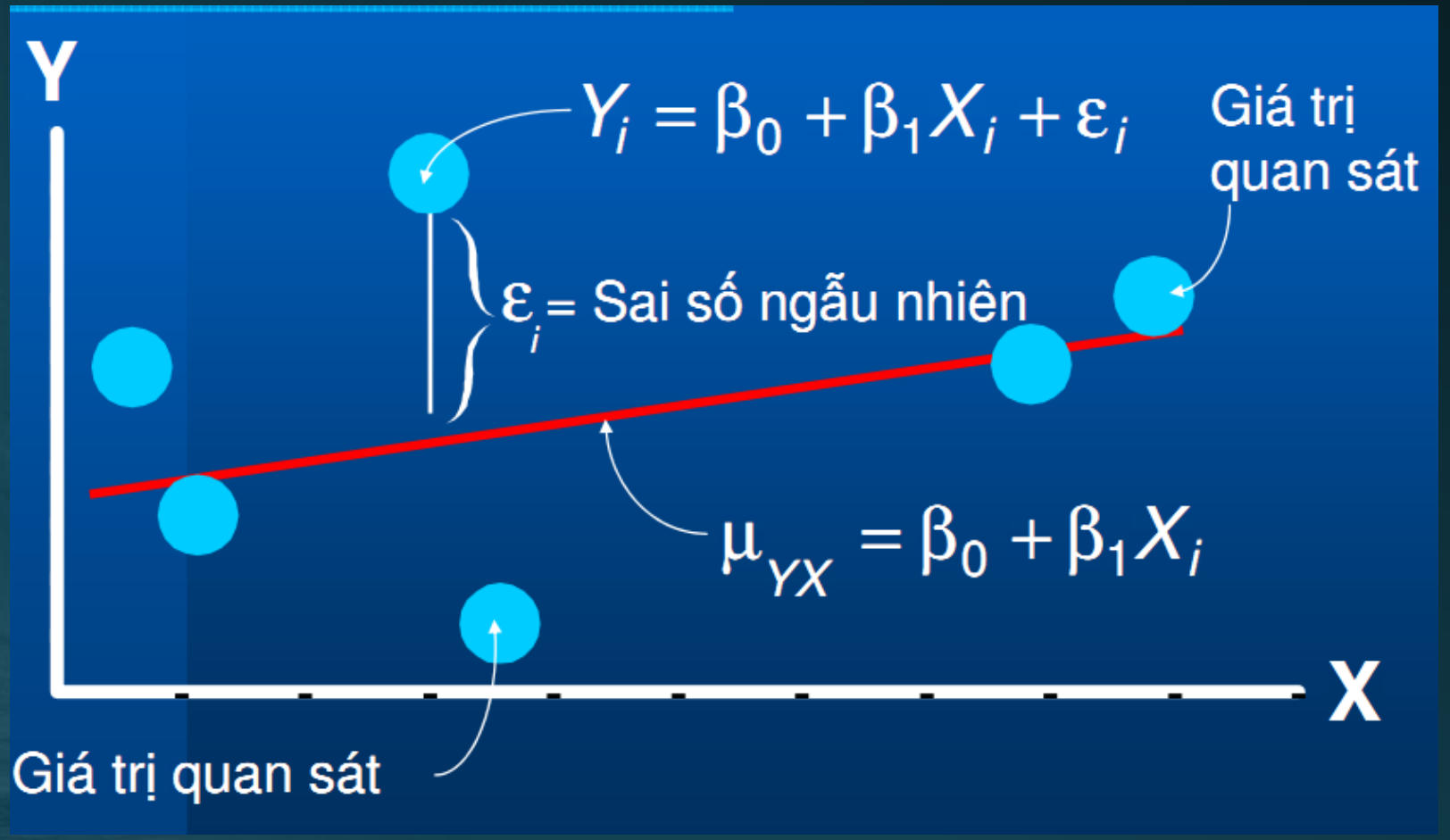
- Methods: Linear regression, Logistic regression, Poisson regression

- Regression analysis is widely used for prediction and forecasting

Predict model



Linear regression



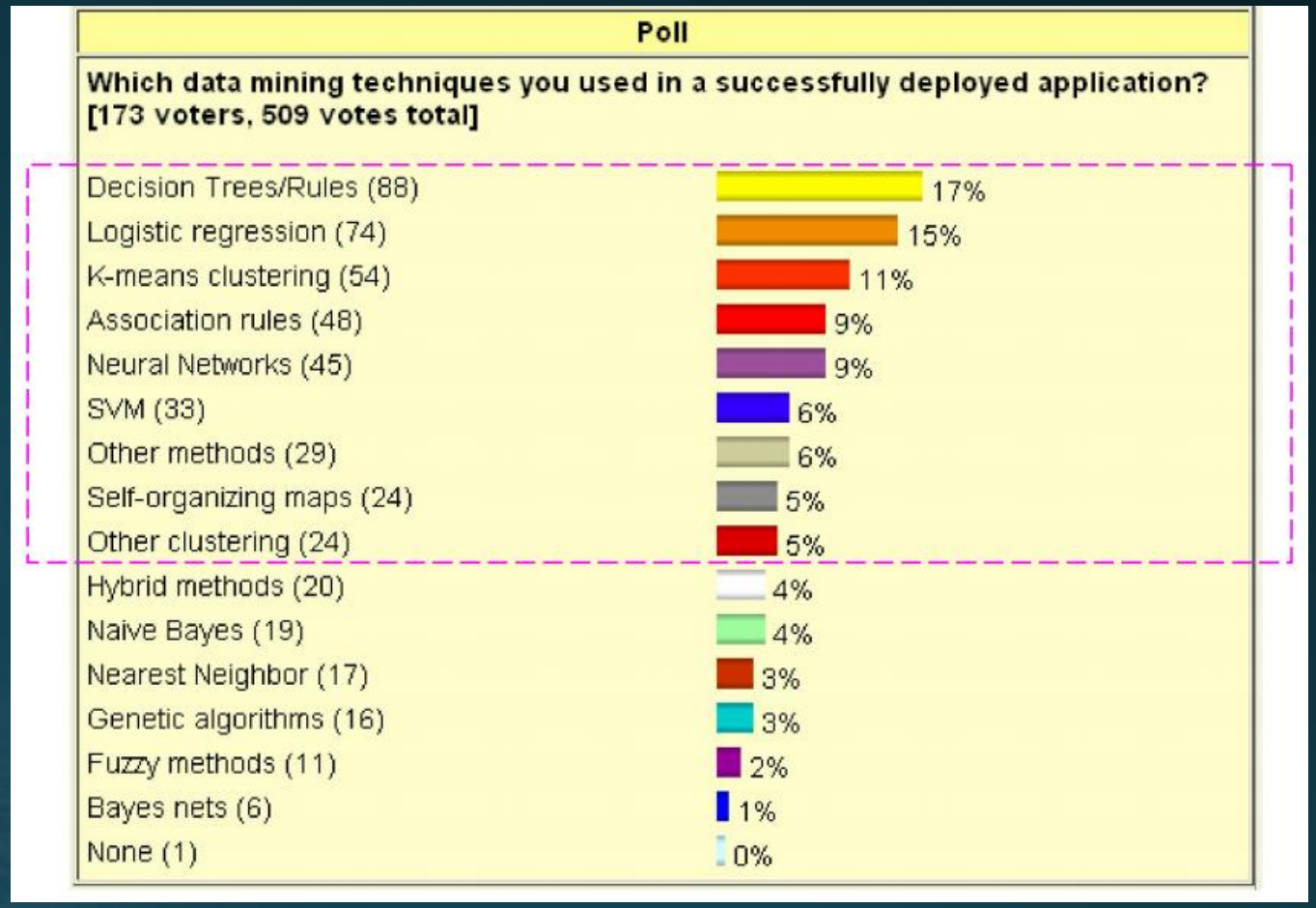
Analysis simple linear regression in R

Demo in R

Estimate

- Clustering: hard, using: user examine, similarity measure, classification algorithms, entropy, F-measure, pure,…
- Clasification: holdout, k-fold cross validation,…
- Regression: statistical hypothesis test

Report


Poll
Which data mining techniques you used in a successfully deployed application? [173 voters, 509 votes total]
Decision Trees/Rules (88) 17%
Logistic regression (74) 15%
K-means clustering (54) 11%
Association rules (48) 9%
Neural Networks (45) 9%
SVM (33) 6%
Other methods (29) 6%
Self-organizing maps (24) 5%
Other clustering (24) 5%
Hybrid methods (20) 4%
Naive Bayes (19) 4%
Nearest Neighbor (17) 3%
Genetic algorithms (16) 3%
Fuzzy methods (11) 2%
Bayes nets (6) 1%
None (1) 0%

References

- http://en.wikipedia.org/wiki/Iris_flower_data_set
- http://ykhoa.net/r/R/Chuong%2010.%20%20Phan%20tich%20hoi%20qui%20tuyen%20tinh.pdf
- http://www.statsoft.com/Textbook/Elementary-Statistics-Concepts
- http://bis.net.vn/forums/p/366/628.aspx

Q&A

Thank for listening