Data Science 2

TILBURG UNIVERSITY

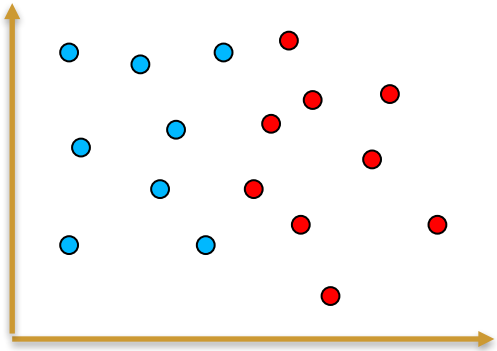
Understanding Society

Overview

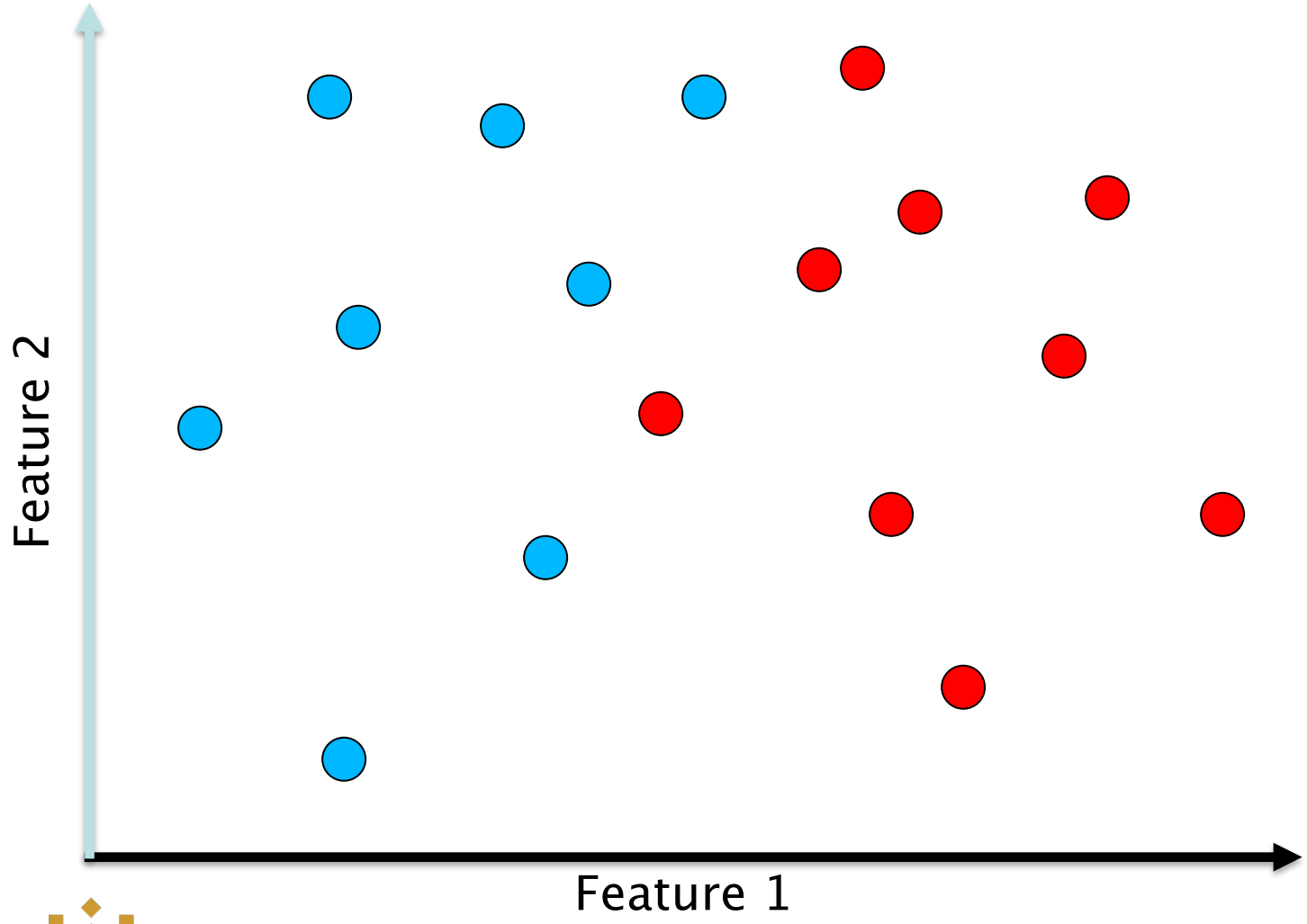
- Classification
  - Decision boundaries
- Exploratory Data Analysis
  - Univariate (single feature)
  - Multivariate (multiple features)
- EDA in Orange
  - iris.tab
  - glass.tab

Classification

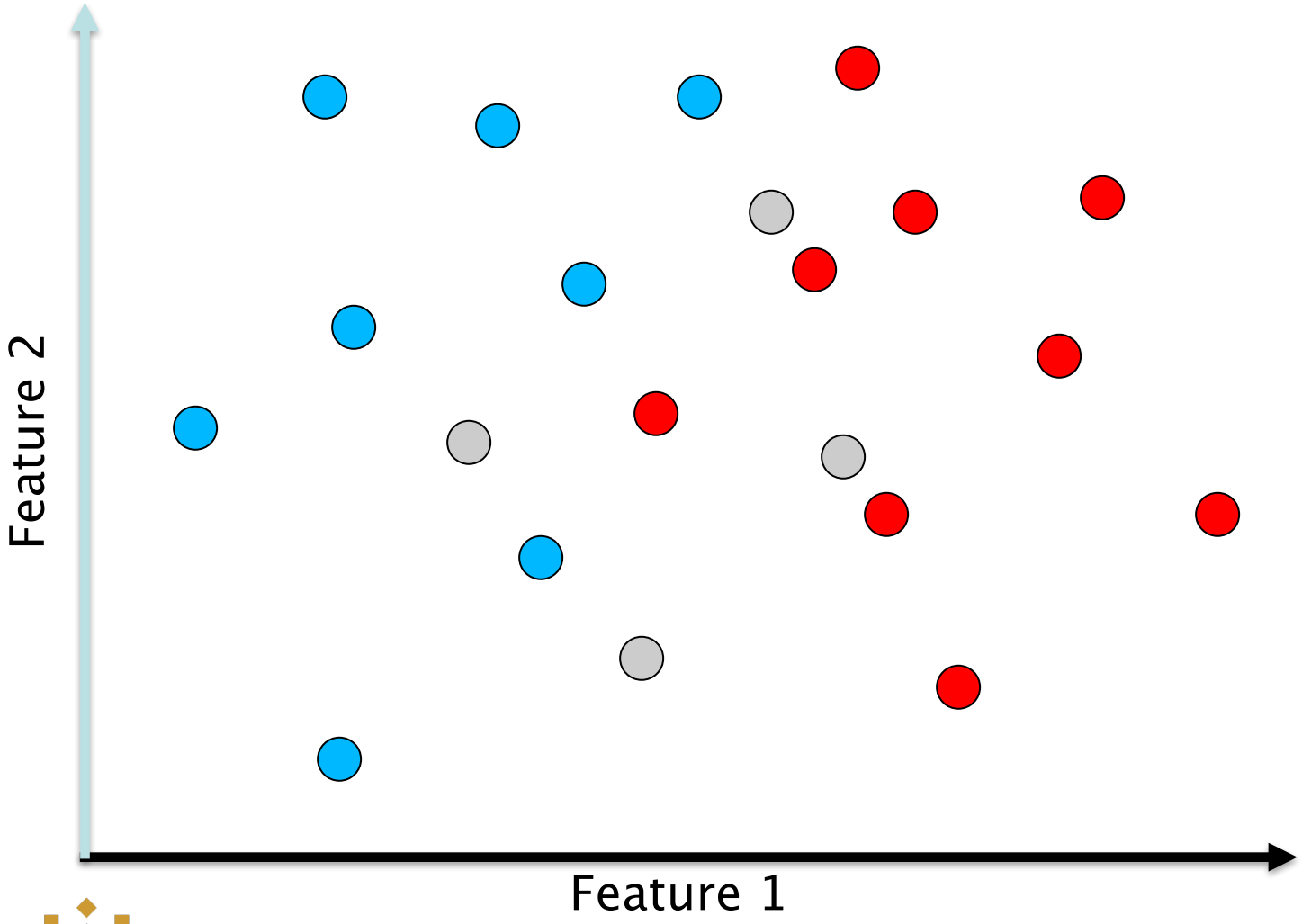
- Things are represented by feature vectors (points)
- Each thing (point) has a class label
- Which we represent by colours
- Examples of classification tasks:
  - Stock Market features —> BUY/SELL?
  - BLOGpost features —> MALE/FEMALE?
  - Fruit features —> ORANGE/APPLE/KIWI?
  - Image features —> INDOOR/OUTDOOR?



2 classes (blue and red) defined by 2 features



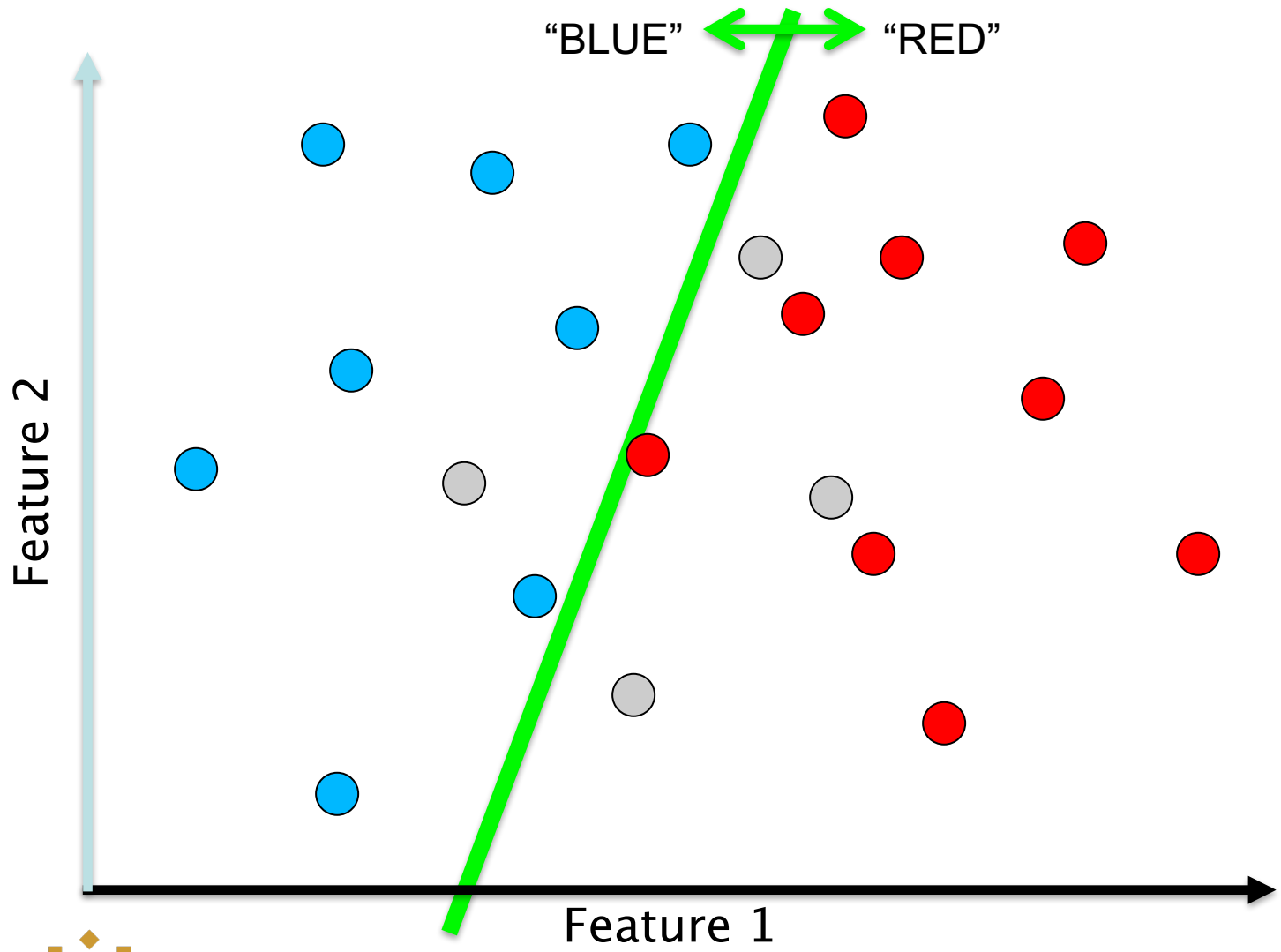
What are the class labels assigned to the grey instances?



Decision Boundaries

- Classifiers are trained on the dataset (labelled data points) and automatically “draw” a decision boundary between the two classes
- The decision boundary can be a straight line (“stiff”) or a wiggly line (“flexible”)
- The decision boundary is considered to be a model of the separation between the two classes
- The model is induced from the data
- The complexity of the model is proportional to the wiggly-ness of the decision boundary

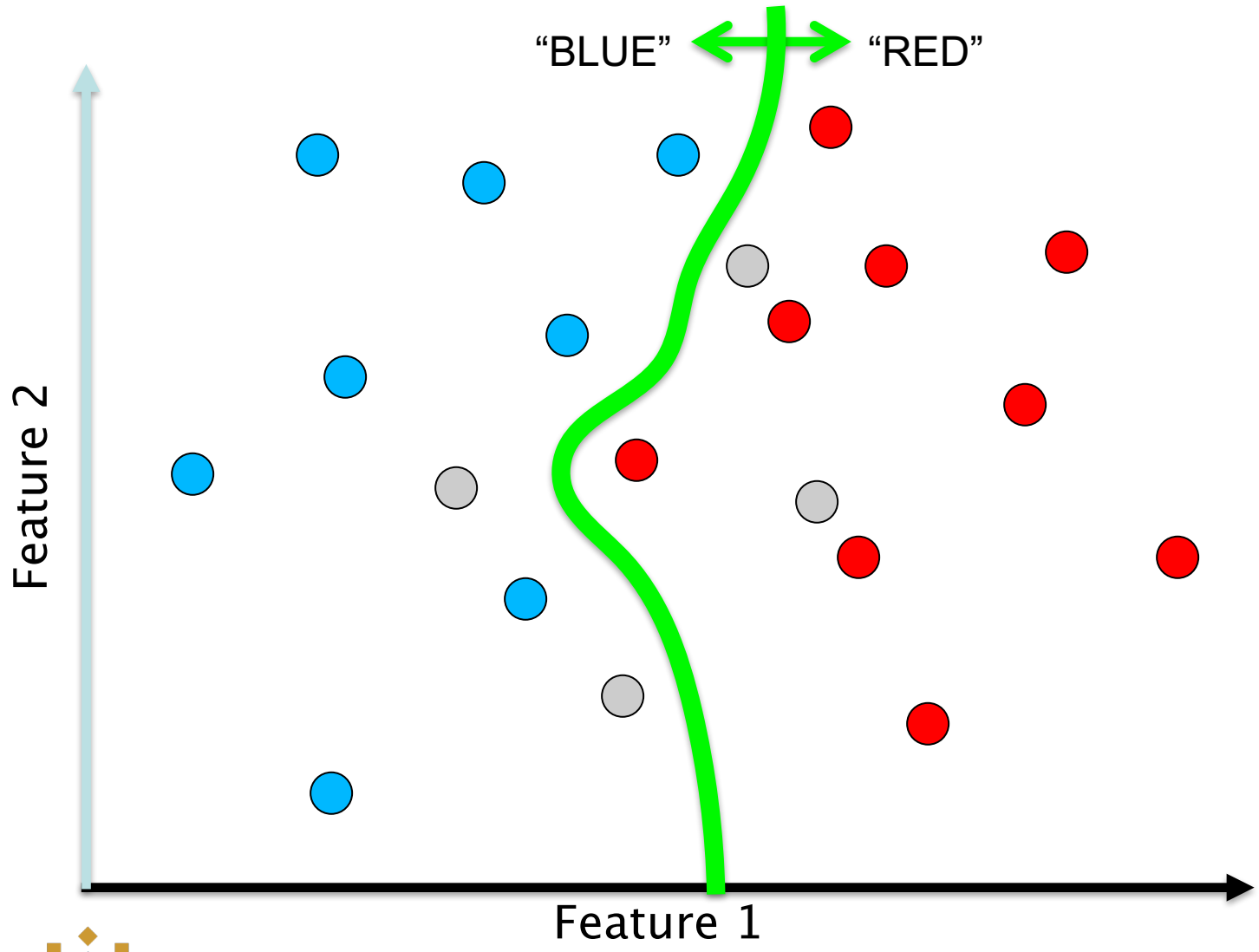
Decision Boundaries: linear decision boundary



TILBURG UNIVERSITY

simple model

Decision Boundaries: nonlinear decision boundary

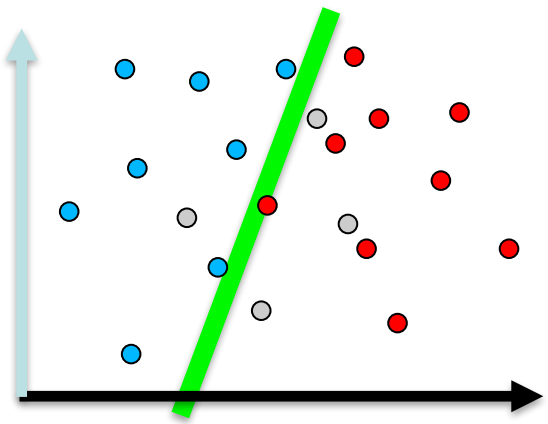


TILBURG UNIVERSITY

complex model

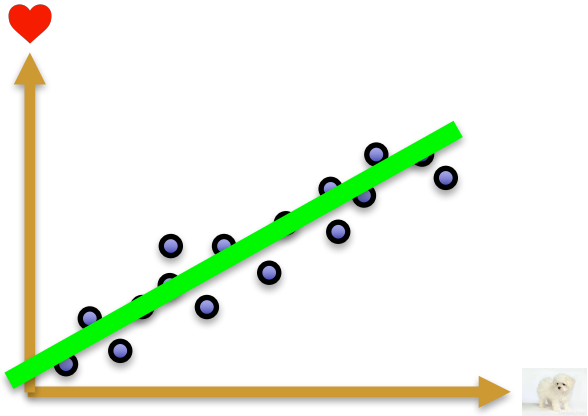
PLEASE NOTE!
Classification versus Regression

- In classification, the model induced from the data defines a decision boundary that **separates** the data described by 2 features into 2 classes (e.g., *cats* versus *dogs*)



separates the data

- In regression, the model induced from the data **fits** the data to describe the relation between 2 features or between a feature (e.g., *furriness*) and the label (e.g., *cuteness*)



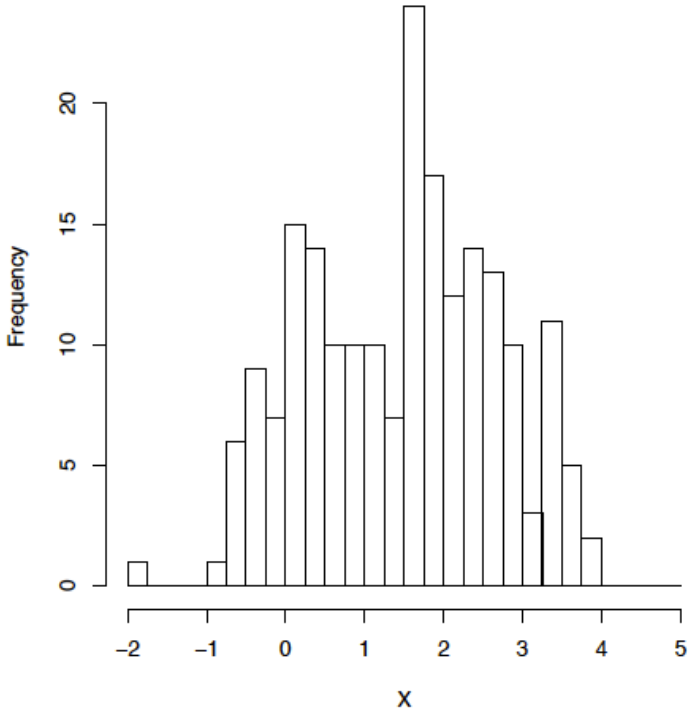
fits the data

Exploratory Data Analysis

- Getting to know your data is at the heart of Data Science
- Do not treat data as a “black box” that you can simply throw into your data science software! (“Trash in, trash out”)
- Data Science is mainly about making sense of the raw data and about defining good features
- Exploratory Data Analysis (EDA) refers to the use of statistical and visualisation tools to make sense of the structure of the data
- It should be complemented by domain knowledge (*what does the data represent and why is that important for the task at hand?*)

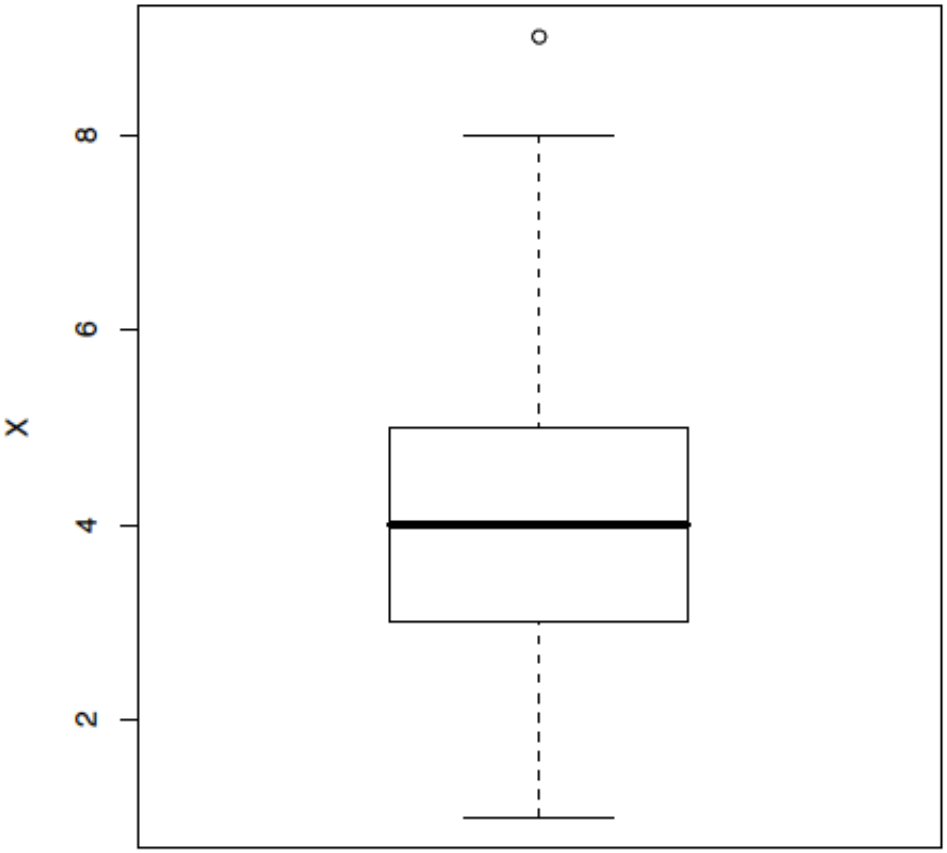
EDA: Univariate Analysis

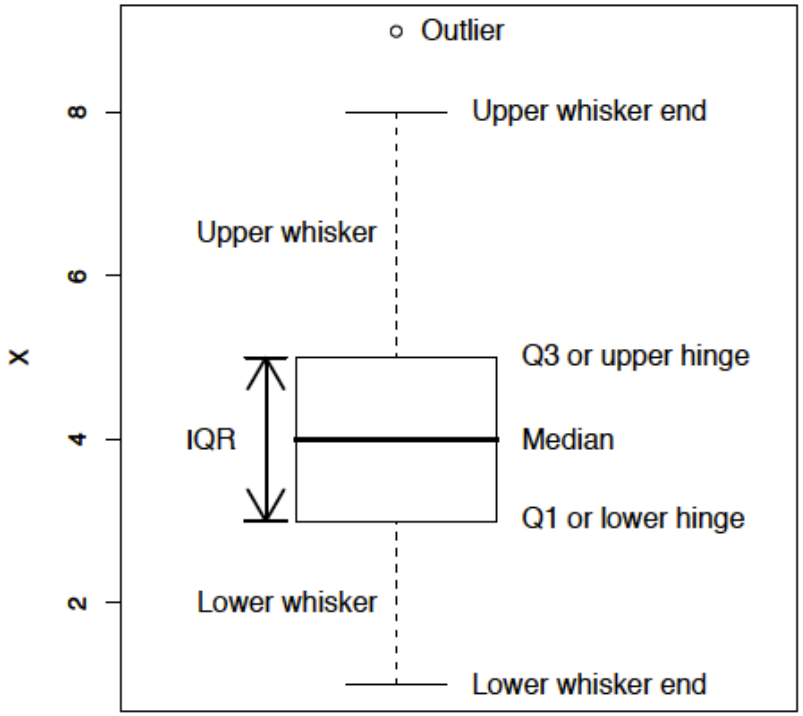
- Statistical descriptors: mean, standard deviation/variance, median, mode, …
- Distributional features: histogram (visualisation)



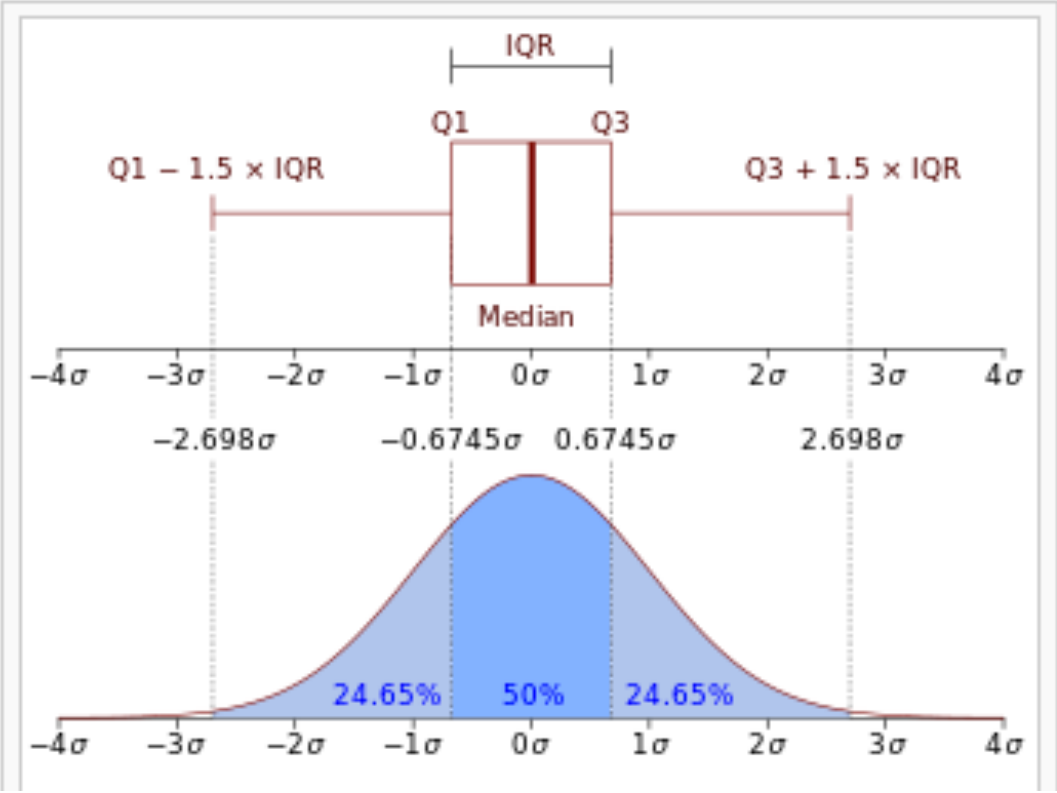
EDA: Univariate Analysis

- Box Plots



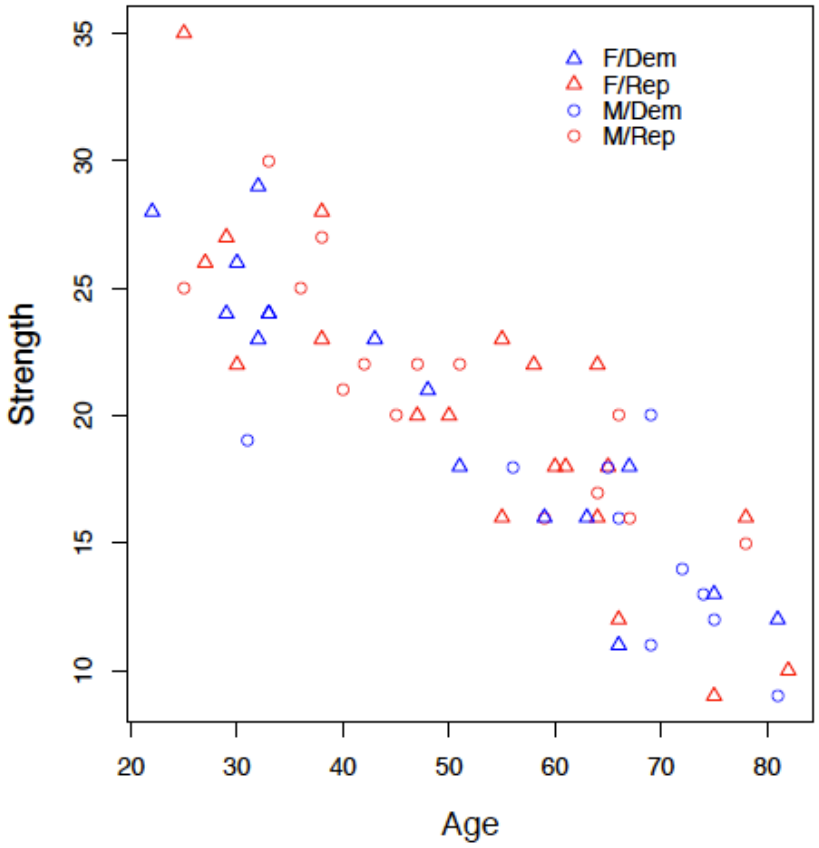


Box plot versus histogram



- Box plot and histogram are similar representations (reproduced from: https://en.wikipedia.org/wiki/Quartile)

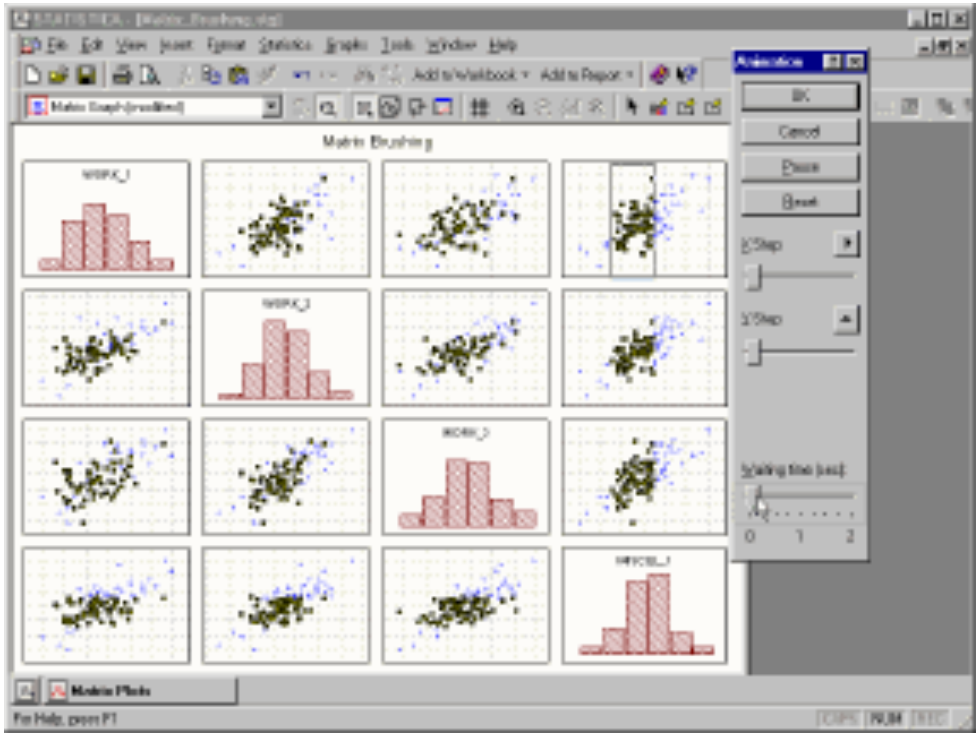
EDA: multivariate analysis (2 features)



- In a scatterplot, the locations of points shows the relation between two features (variables) and their colours represent their classes

EDA: multivariate analysis (more than 2 features)

pair-wide scatterplots



or: use dimensionality reduction methods, e.g., PCA (to be discussed)

Visual Analytics

- Extending EDA with Data Science tools yields Visual Analytics

The basic idea of visual analytics is to visually represent the information, allowing the human to directly interact with the information, to gain insight, to draw conclusions, and to ultimately make better decisions. The visual representation of the information reduces complex cognitive work needed to perform certain tasks. People may use visual analytics tools and techniques to synthesize information and derive insight from massive, dynamic, and often conflicting data by providing timely, defensible, and understandable assessments.
(Keim, Mansmann, Schneidewind,Thomas, & Ziegler, 2008)

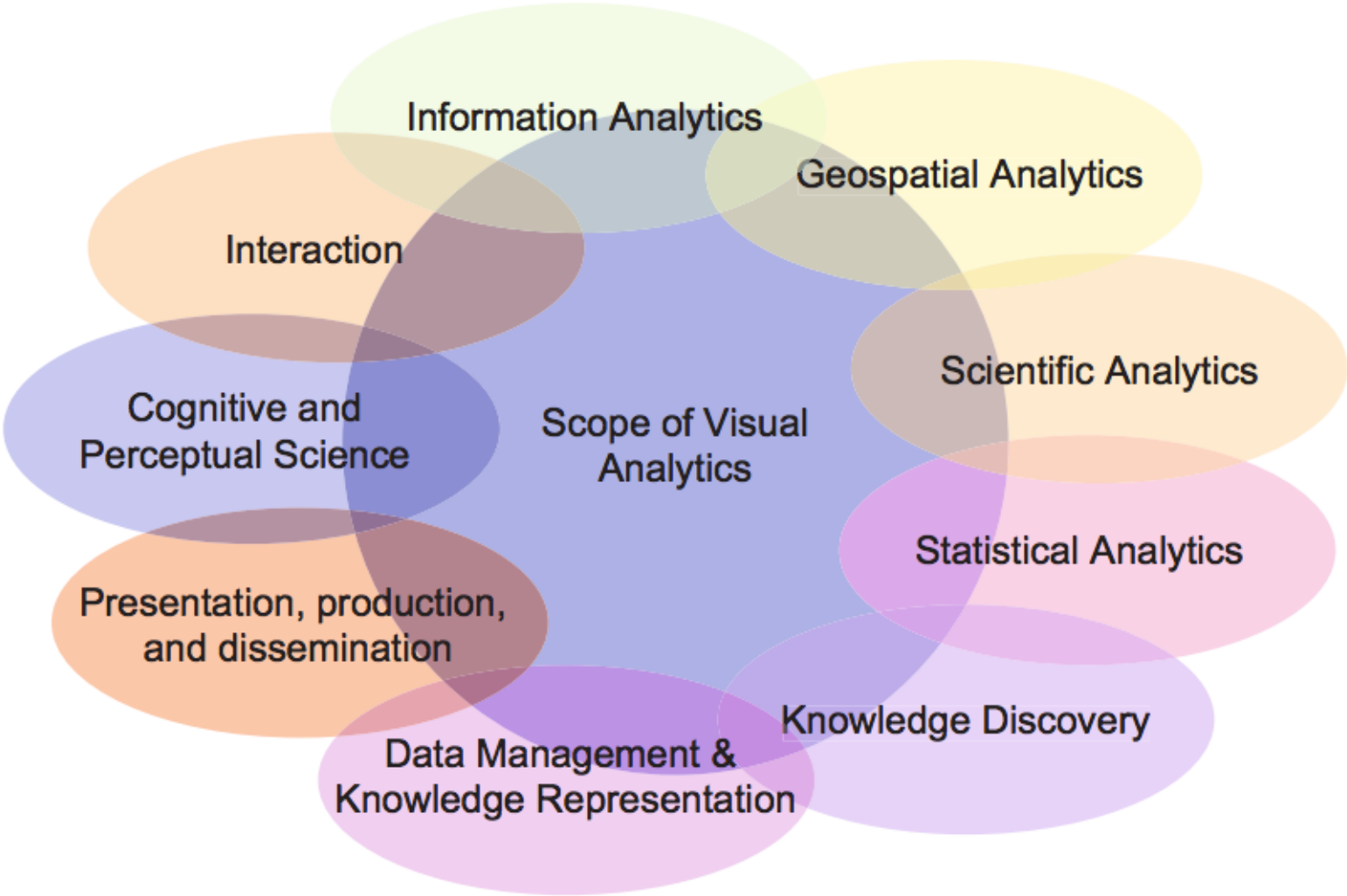
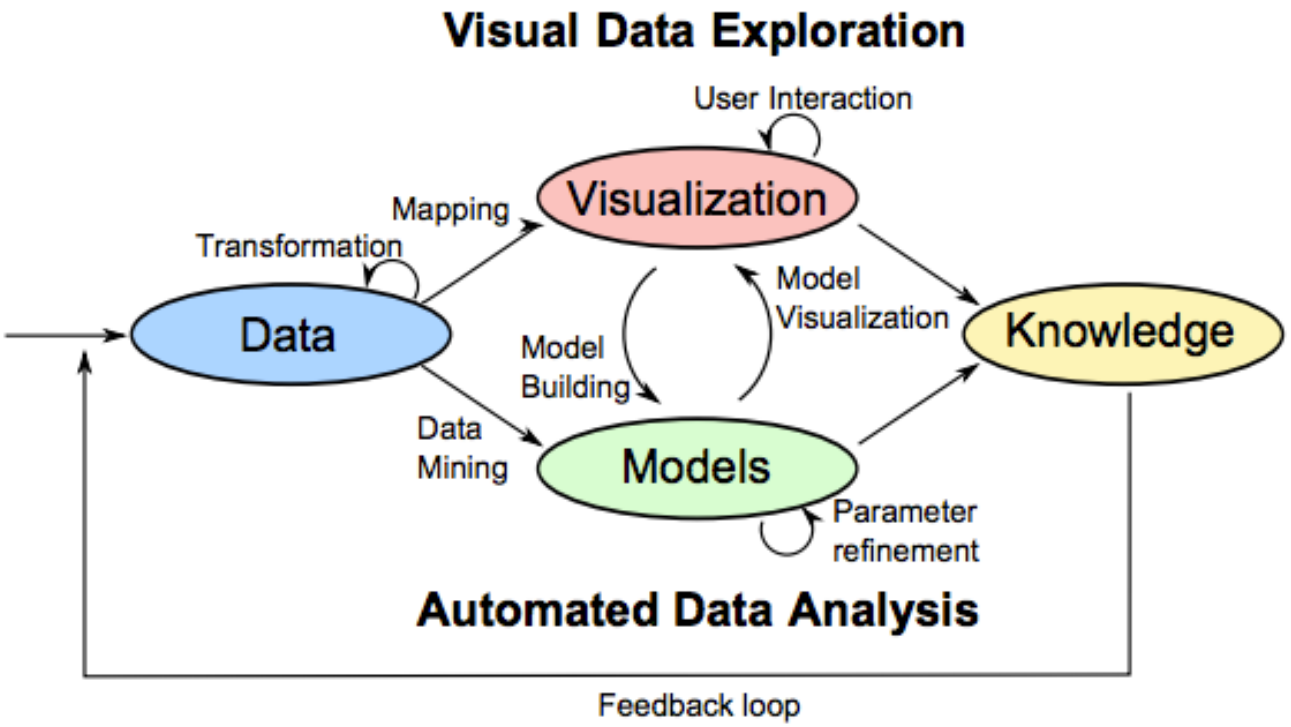


Fig. 1. The Scope of Visual Analytics

(Keim, Mansmann, Schneidewind,Thomas, & Ziegler, 2008)

Visual Analytics Process is interactive and iterative



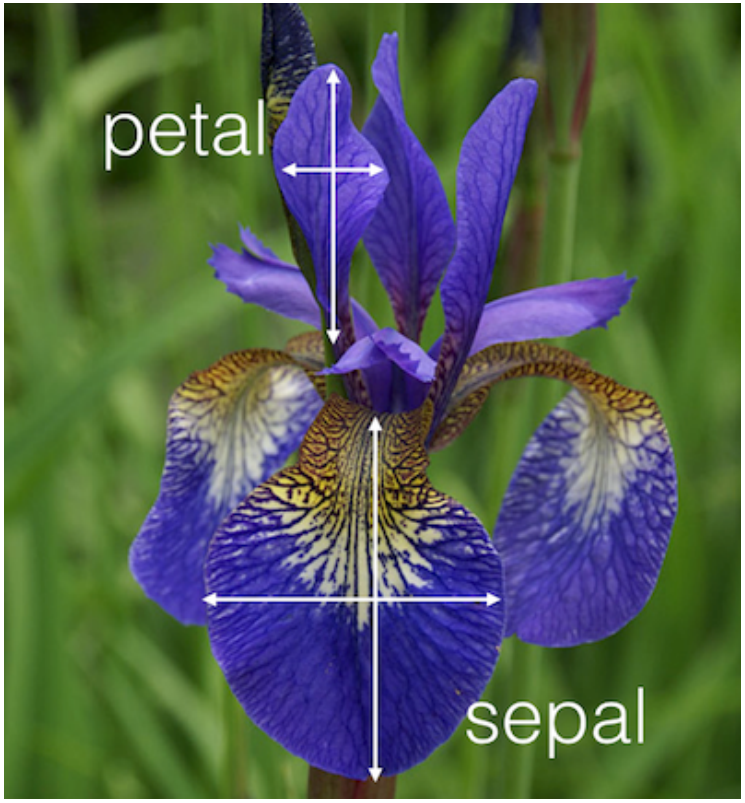
Iris Data Set: classification task (3 classes)

Iris Setosa

Iris Versicolor

Iris VIrginica

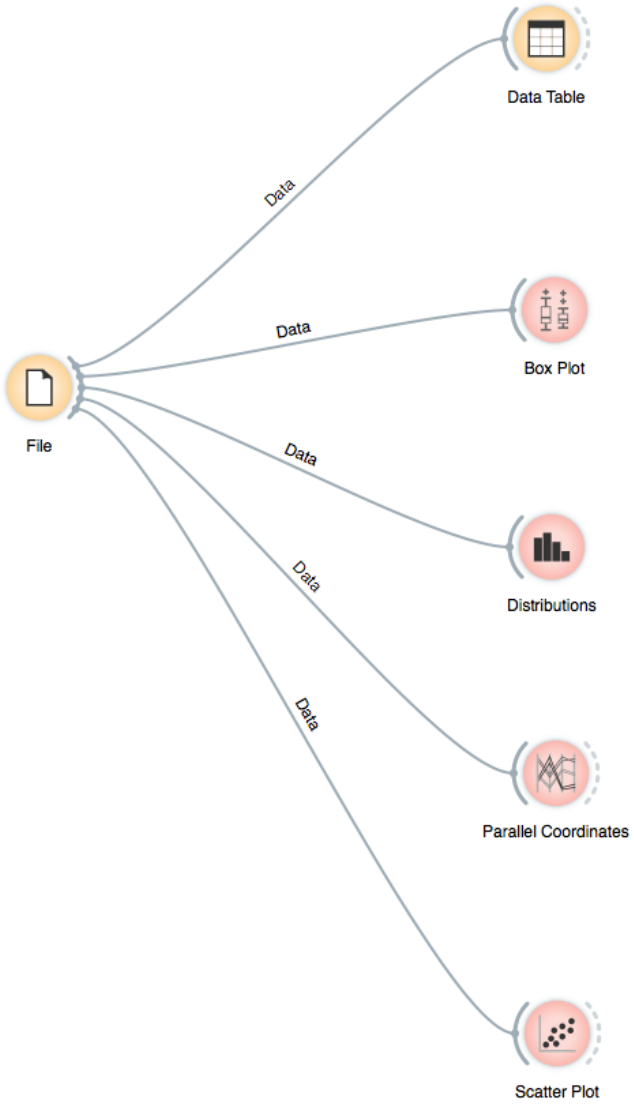
Iris Data Set: 4 features



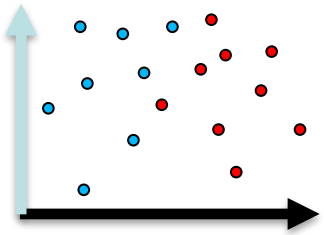
sepal length	sepal width	petal length	petal width	iris

4 continuous values per row + class label

Data Visualisation with Orange



- Read the iris.tab dataset
- 4D features, 1 output
- Put the following widgets on the canvas
  - Data Table
  - Box Plot
  - Histogram
  - Parallel Coordinates
  - Scatter Plot
- Explore the separation of the three classes
  - find separations between each pair of classes



GLASS.tab dataset

1. Na: Sodium (unit measurement: weight percent in corresponding oxide)
2. Mg: Magnesium
3. Al: Aluminum
4. Si: Silicon
5. K: Potassium
6. Ca: Calcium
7. Ba: Barium
8. Fe: Iron
9. Type of glass: (class attribute)
    -- 1 building_windows_float_processed
    -- 2 building_windows_non_float_processed
    -- 3 vehicle_windows_float_processed
    -- 4 vehicle_windows_non_float_processed (none in this database)
    -- 5 containers
    -- 6 tableware
    -- 7 headlamps

Literature

- Seltman, H.J. (2015). Chapter 4. Exploratory Data Analysis. Experimental Design and Analysis, an online book. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

- Keim, D.A. Mansmann, F. Oelke, D., & Ziegler, H. (2008). Visual Analytics: Combining Automated Discovery with Interactive Visualizations. Proceedings of the 11th International Conference on Discovery Science (DS 2008), Springer-Verlag, pages 2-14, 2008. http://bib.dbvis.de/uploadedFiles/324.pdf

- Wang, L., Wang, G., & Alexander, C.A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, Vol. 1, No. 1, 33-38. http://pubs.sciepub.com/dt/1/1/7