Data Mining and Exploration

Spring 2019

Lecturer: Arno Onken

Email: aonken@inf.ed.ac.uk

Institute for Adaptive and Neural Computation

School of Informatics

THE UNIVERSITY *of* EDINBURGH

Edinburgh, 17th January 2019

Logistics (1)

- Course website: tinyurl.com/ztb675b
- Lecturer office hours: Tuesdays 14-16 IF 2.27A
- For questions and answers, please use Piazza: tinyurl.com/ycmht6xh
- TA: Benedek Rózemberczki <benedek.rozemberczki@ed.ac.uk>
- Labs:
  - Weeks 2-5
  - Appleton Tower, room 6.06
  - Group 1:
    - Wednesdays: 09:00 – 10:50
    - Demonstrator: Miruna-Adriana Clinciu
  - Group 2:
    - Wednesdays: 11:10 – 13:00
    - Demonstrator: Jennifer Williams

Logistics (2)

- Presentations:
  - Poster presentations on research papers during second half of the course
  - Potential papers listed on the course website
  - Poster printing deadline for everyone: 26 February 2019
- Mini-project:
  - Apply data mining methods to a real dataset
  - List of potential datasets on the course website
  - Project report will be assessed
- Course grade:
  - 50% exam
  - 35% mini-project
  - 15% poster presentation

Data

Definition of Data from the Oxford Dictionary:

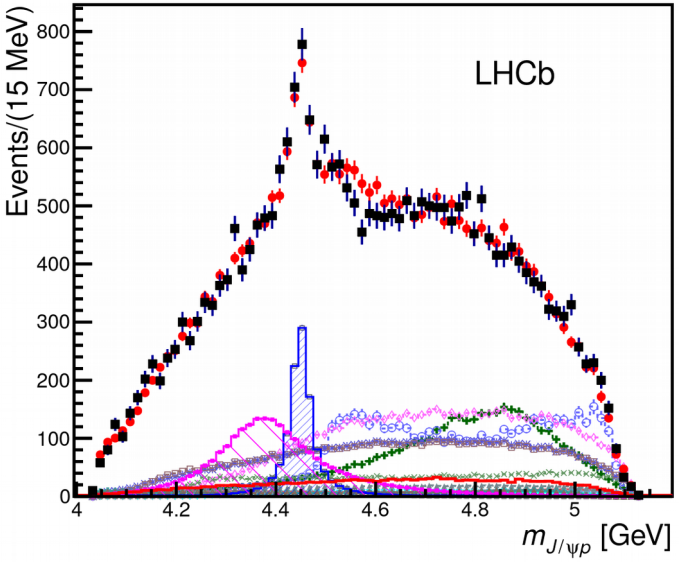
- Facts and statistics collected together for reference or analysis
  - The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media
  - Things known or assumed as facts, making the basis of reasoning or calculation.

Source: https://commons.wikimedia.org/wiki/File:DARPA_Big_Data.jpg

Source: https://commons.wikimedia.org/wiki/File:BigData_2267x1146_white.png

Data Analysis - Data Mining

Data Analysis:
Inspect, transform and model data to discover useful information



Source: https://commons.wikimedia.org/wiki/File:J-psi_p_pentaquark_mass_spectrum.svg

Server Farm at CERN

Source: https://commons.wikimedia.org/wiki/File:CERN_Server_03.jpg

Data Mining: Particular data analysis technique; extraction of patterns and knowledge from large amounts of data for predictive rather than descriptive purposes

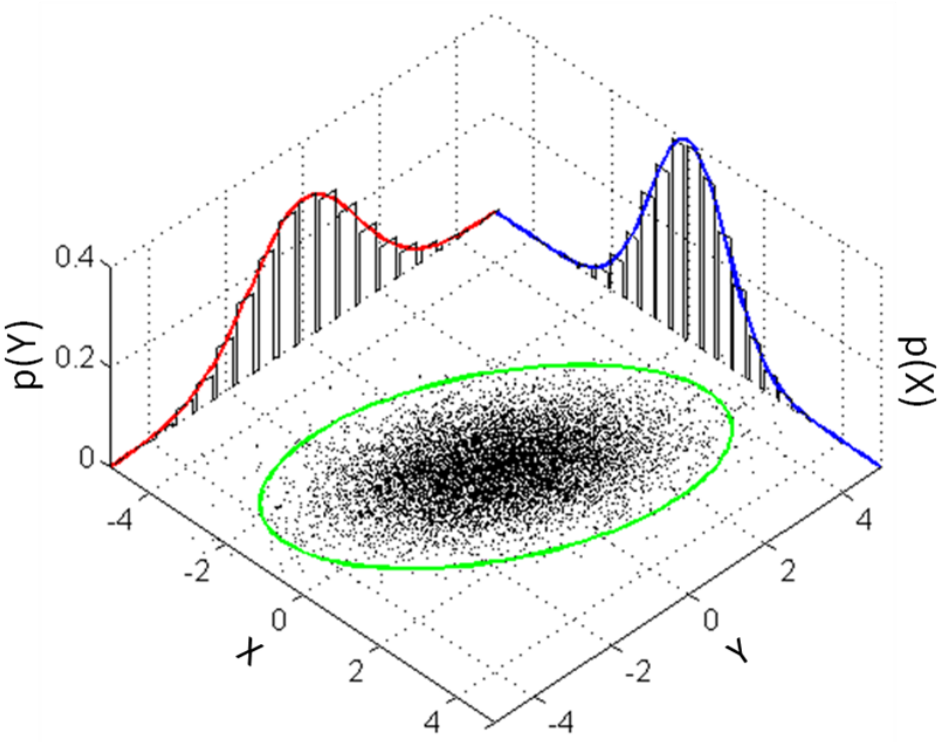
Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a tradition of data analysis to avoid wrong interpretations of suggestive results

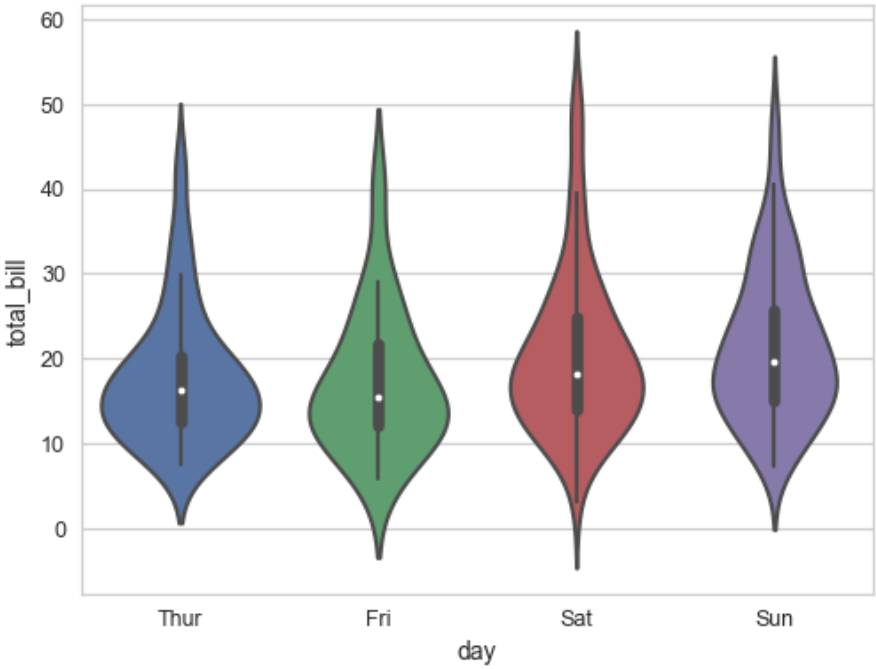
EDA emphasises:

- Graphic representation of the data
- Understanding of the data structure
- Robust measures, re-expression and subset analysis
- Tentative model building in an iterative process of model specification and evaluation
- General scepticism and flexibility with respect to the choice of methods

EDA: Graphic Representation of the Data



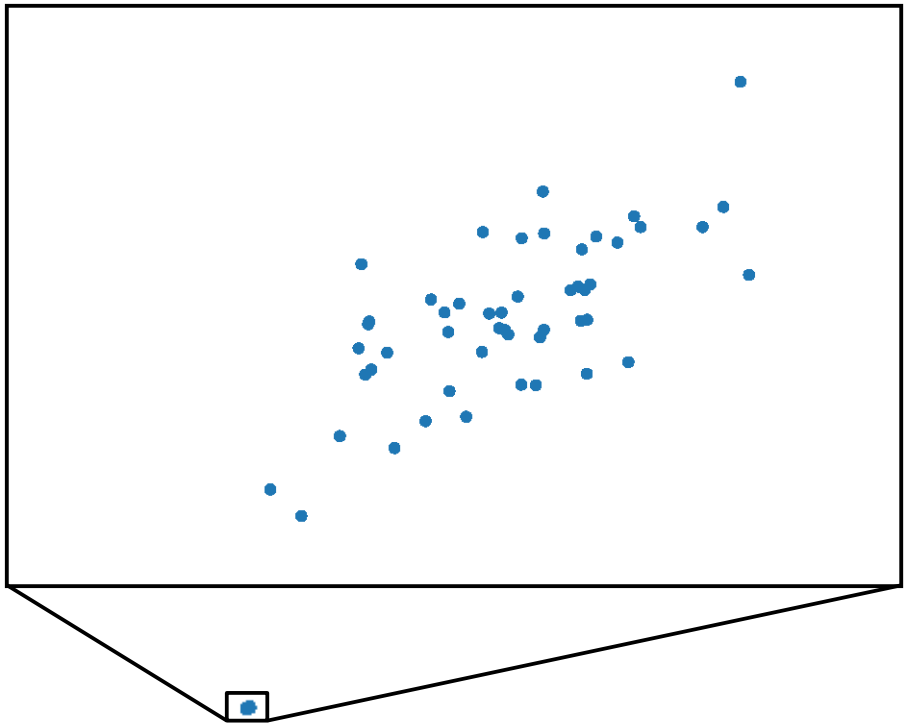
Source: https://commons.wikimedia.org/wiki/File:MultivariateNormal.png



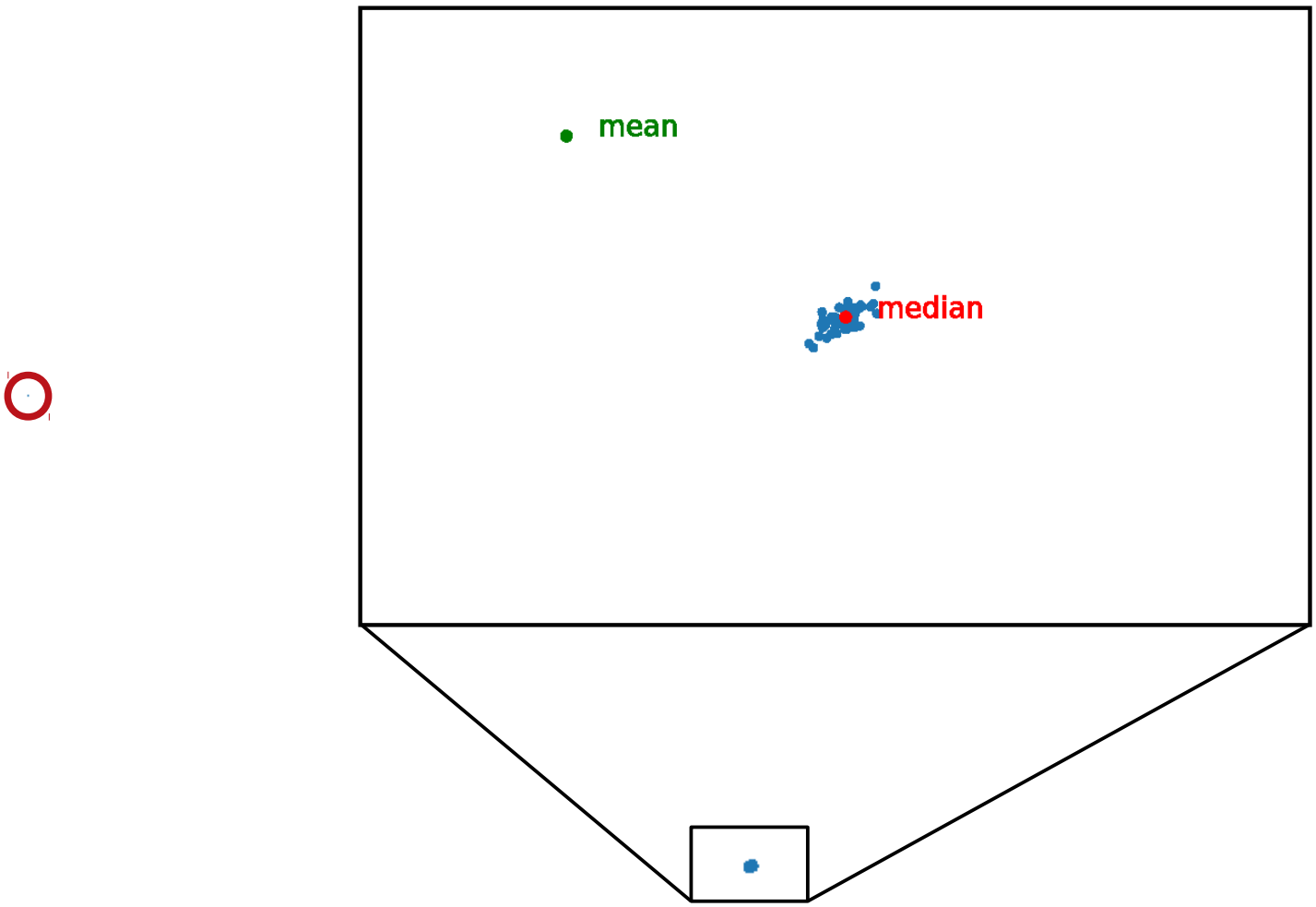
Source: https://seaborn.pydata.org/_images/seaborn-violinplot-2.png

EDA: Understanding of the Data Structure

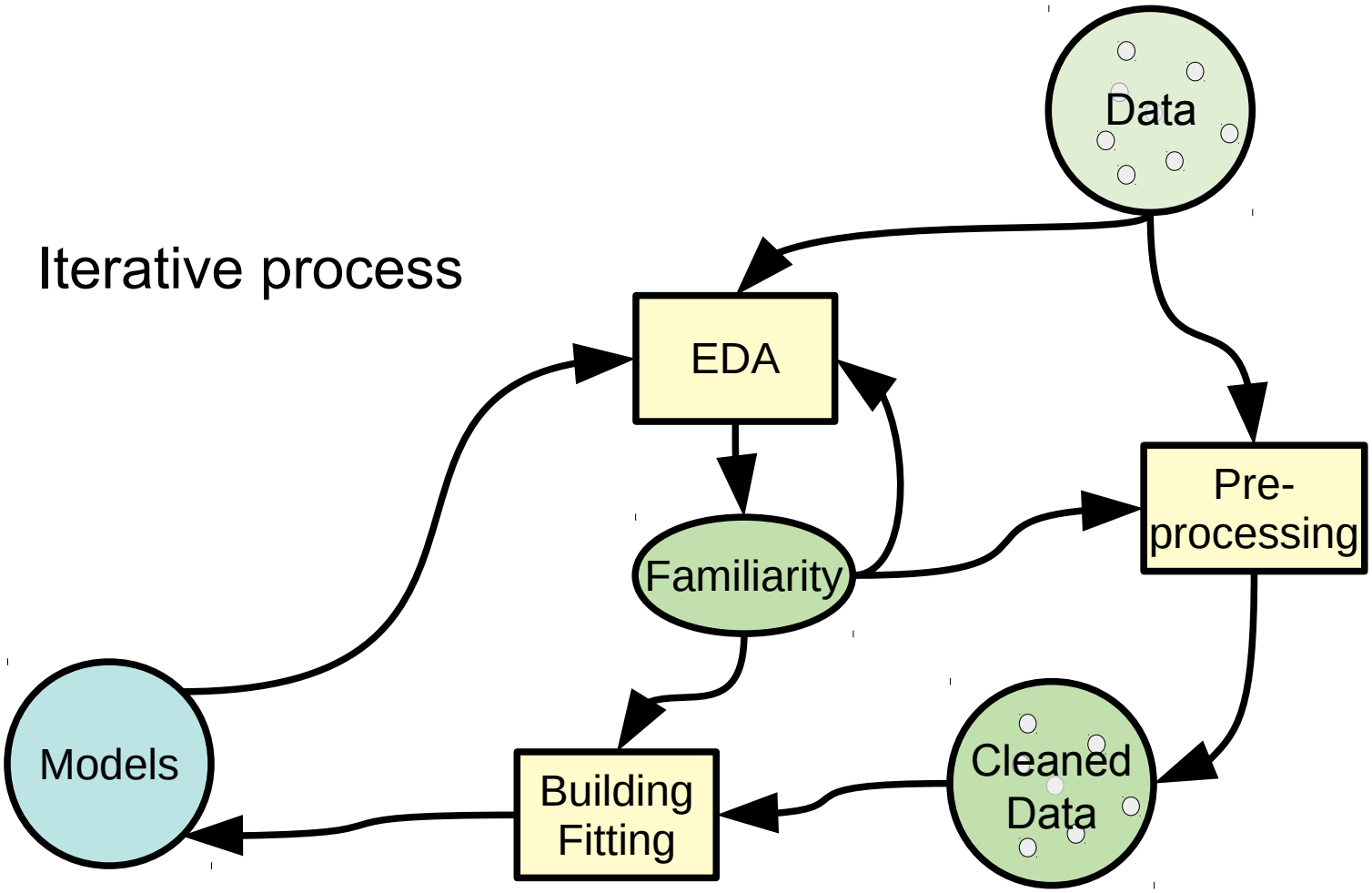
single outlier



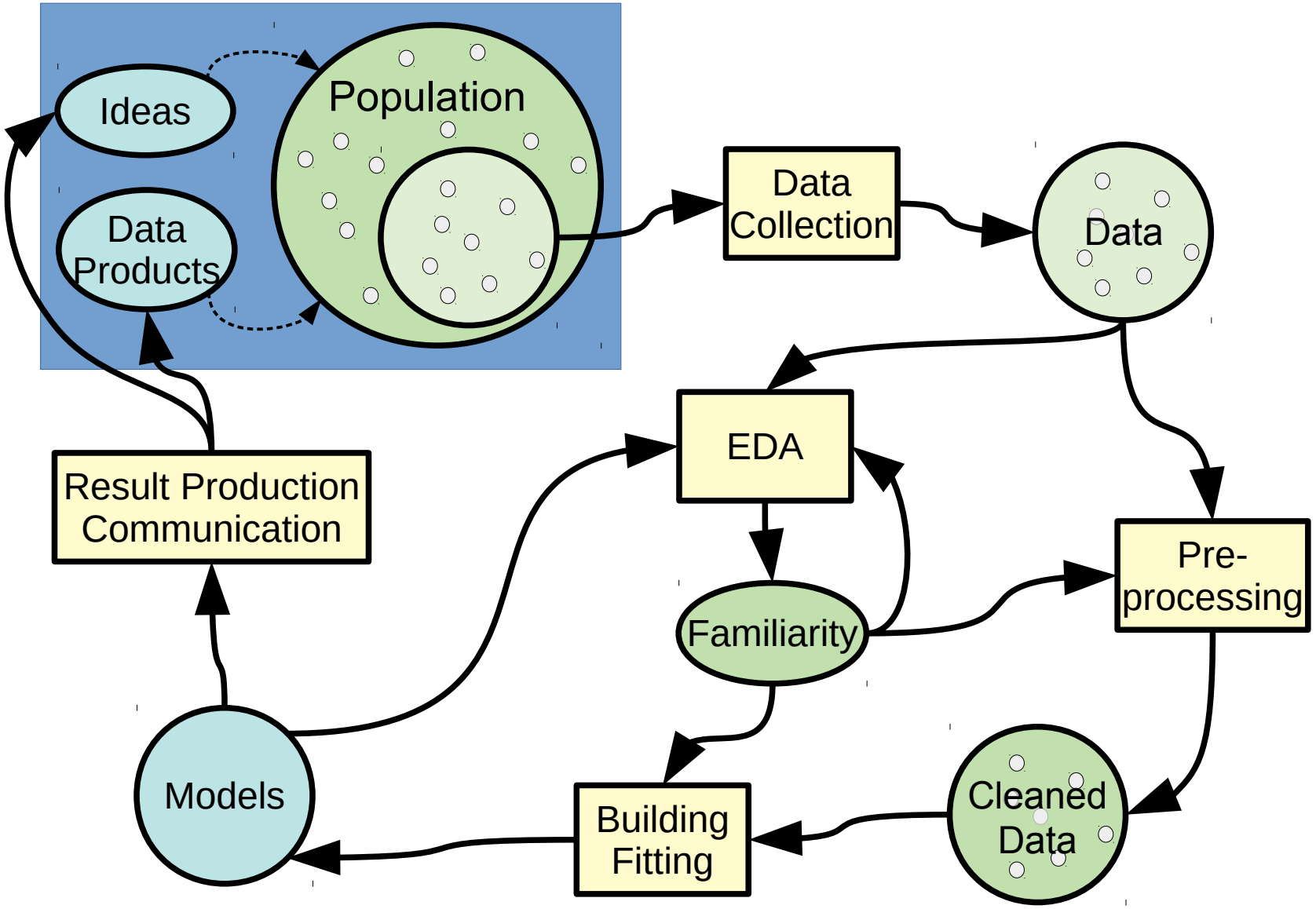
EDA: Robust Measures



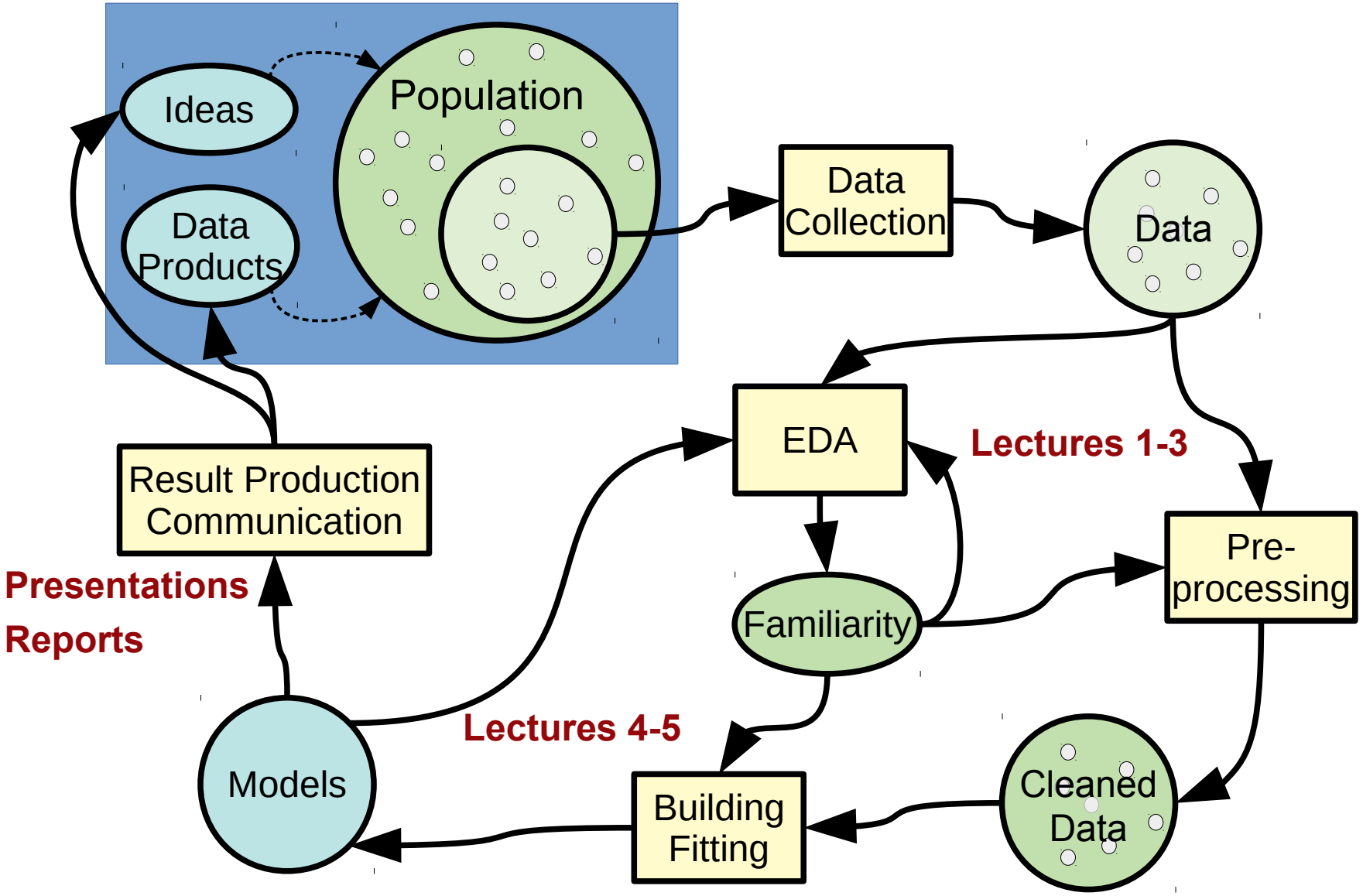
EDA: Tentative Model Building



Data Analysis Process



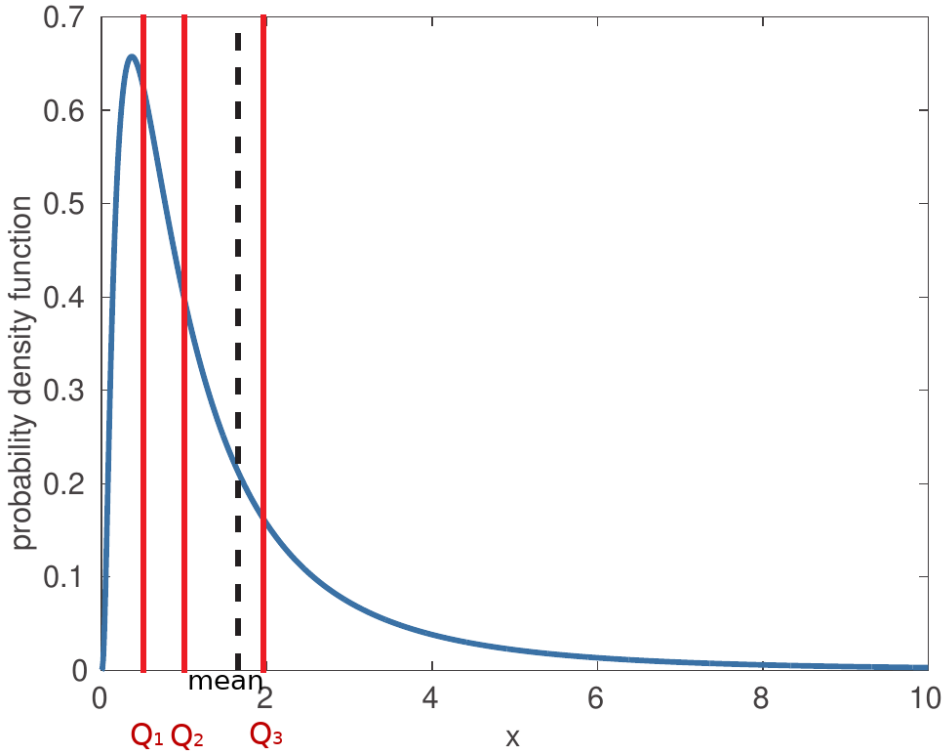
Course Content

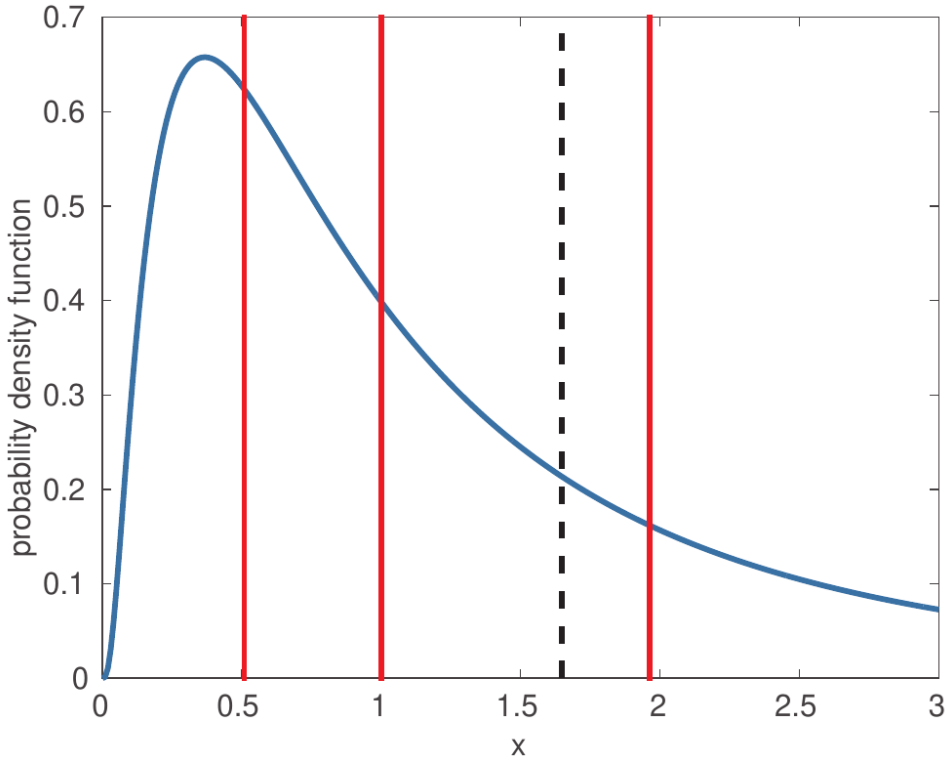


Purpose of Particular Course Elements

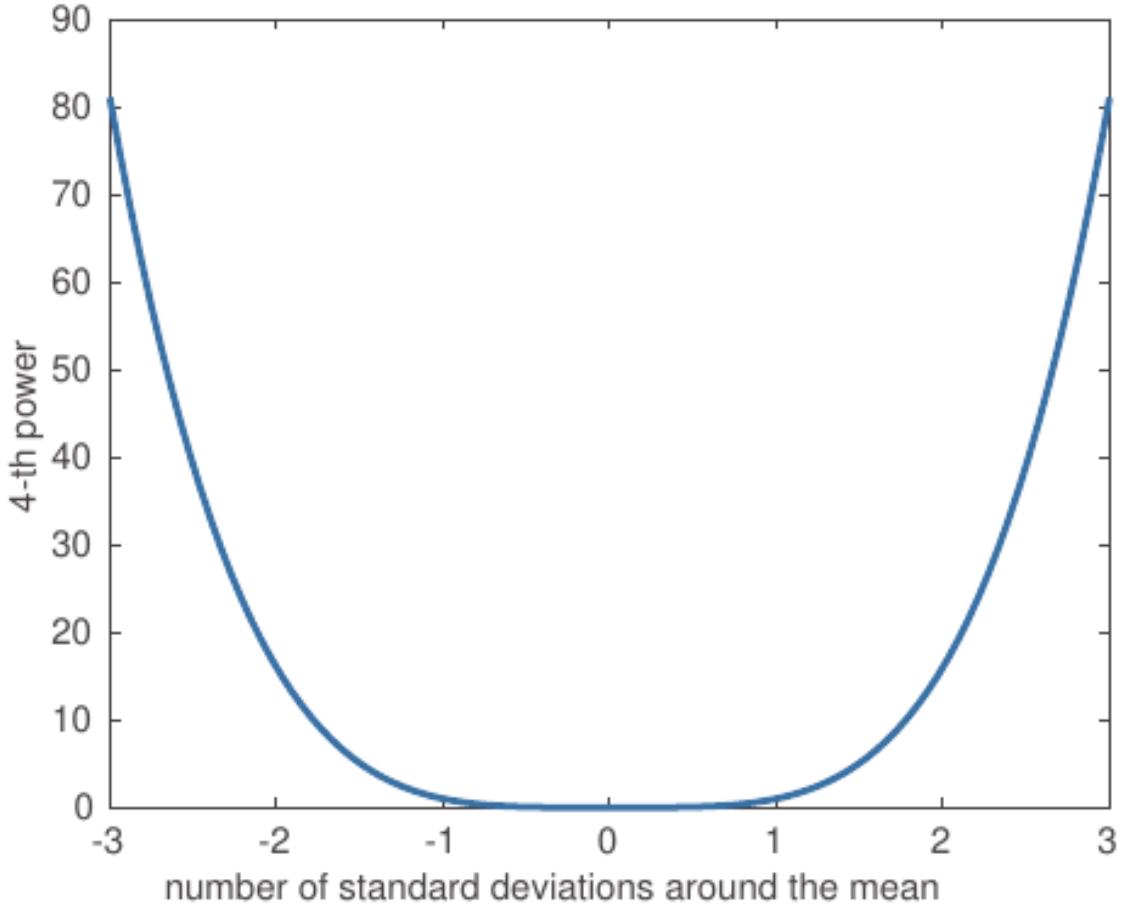
- **Lecture material and computer labs**
  - Numerical data descriptions and pre-processing (today)
    - Establish common language
    - Highlight importance of simple measures
  - In depth Principal Component Analysis (lectures 2-3)
    - Describe important method in all its aspects
  - Dimensionality reduction (lectures 3-4)
    - Closely related techniques
  - Predictive modelling and generalization (lecture 5)
    - Round off data analysis process
- **Poster sessions**
  - Train presentation of research results in the style of an academic conference
  - Exposure to wide range of topics
- **Mini-projects**
  - Full data analysis process

Positive Skewness

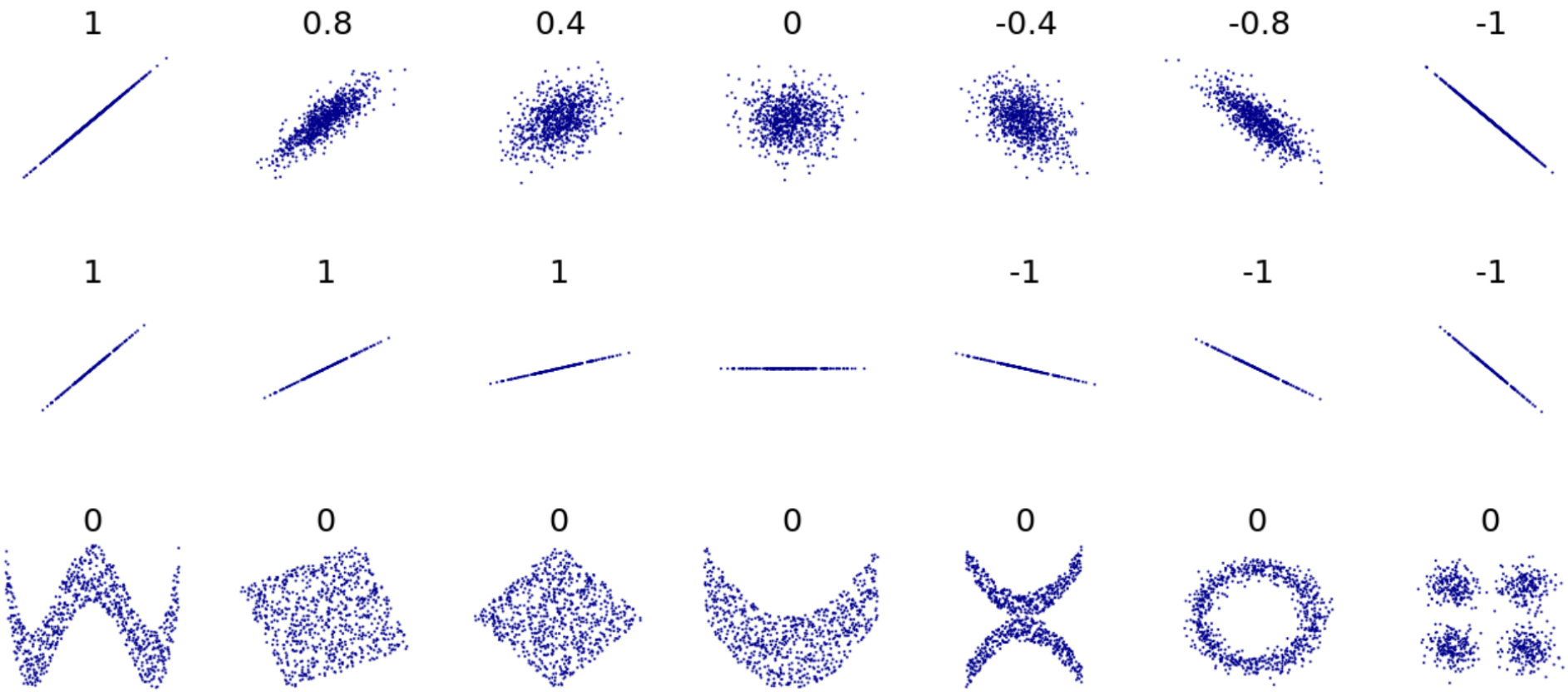




Fourth Power

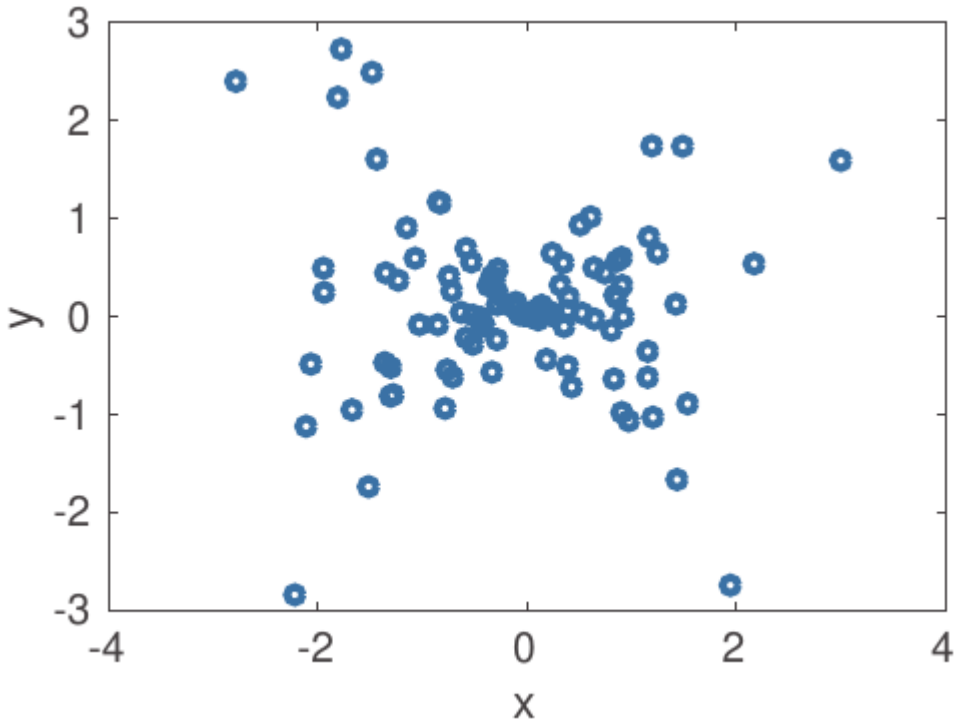


Uncorrelated and Dependent

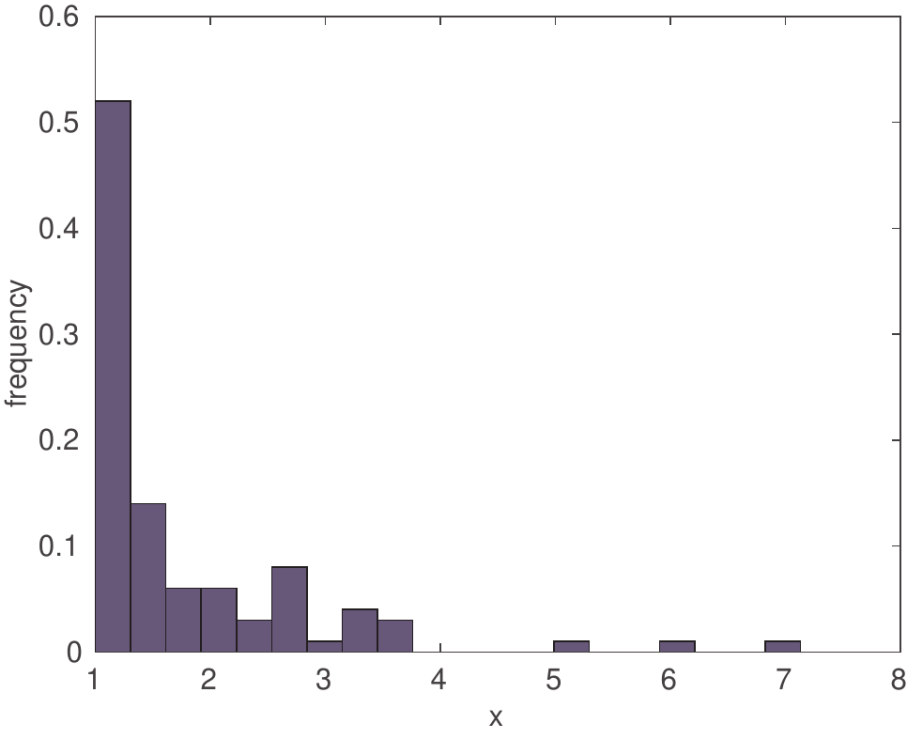


Source: https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

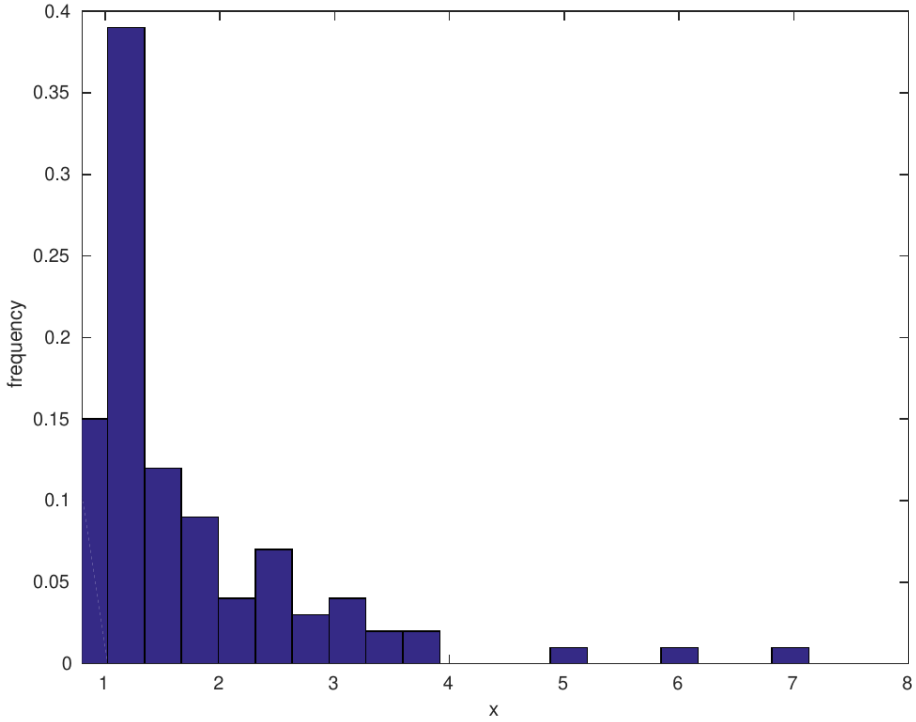
Scatter Plot



Histogram

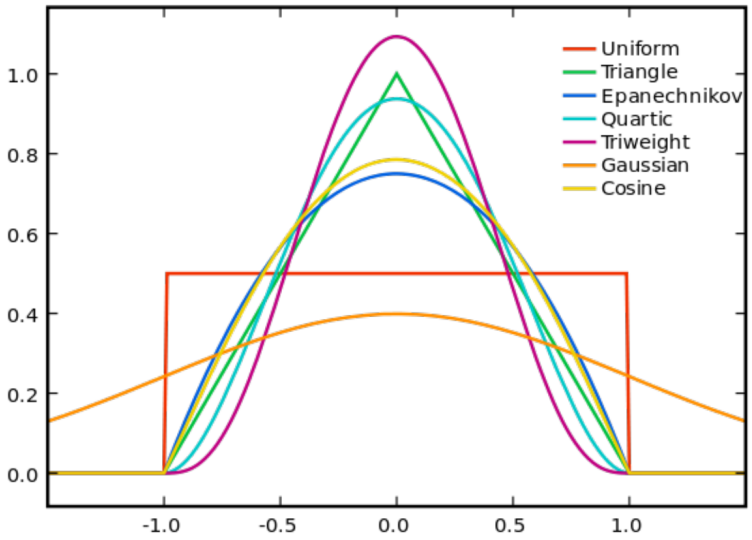


(a) $L = \min_i(x_i)$

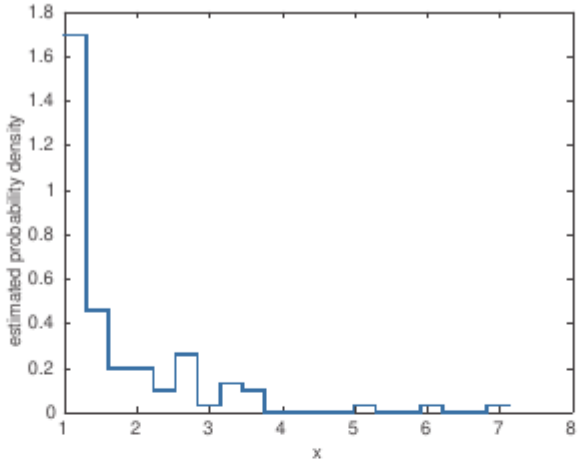


(b) $L = \min_i(x_i) - 0.3$

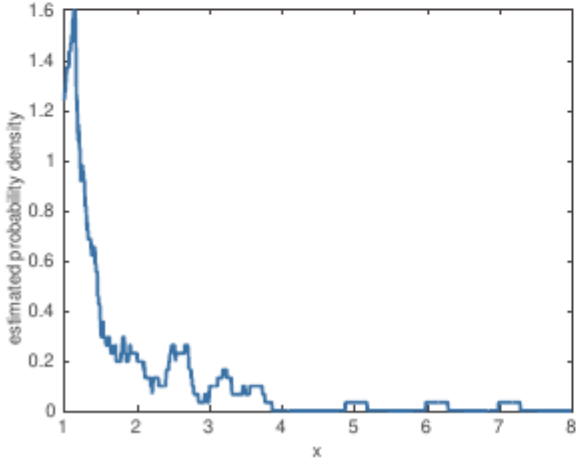
Kernel Density Plots



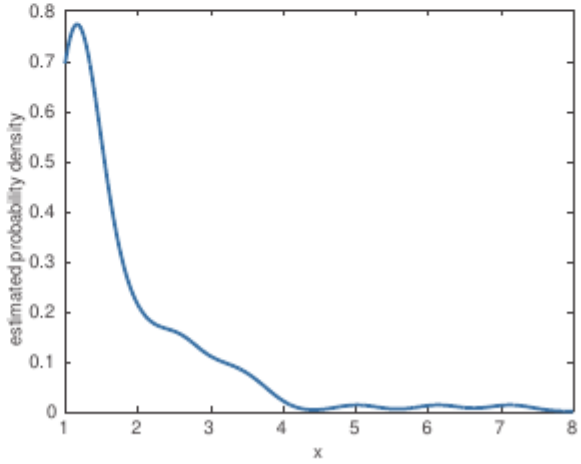
Source: https://en.wikipedia.org/wiki/Kernel_(statistics)



(a) Scaled histogram

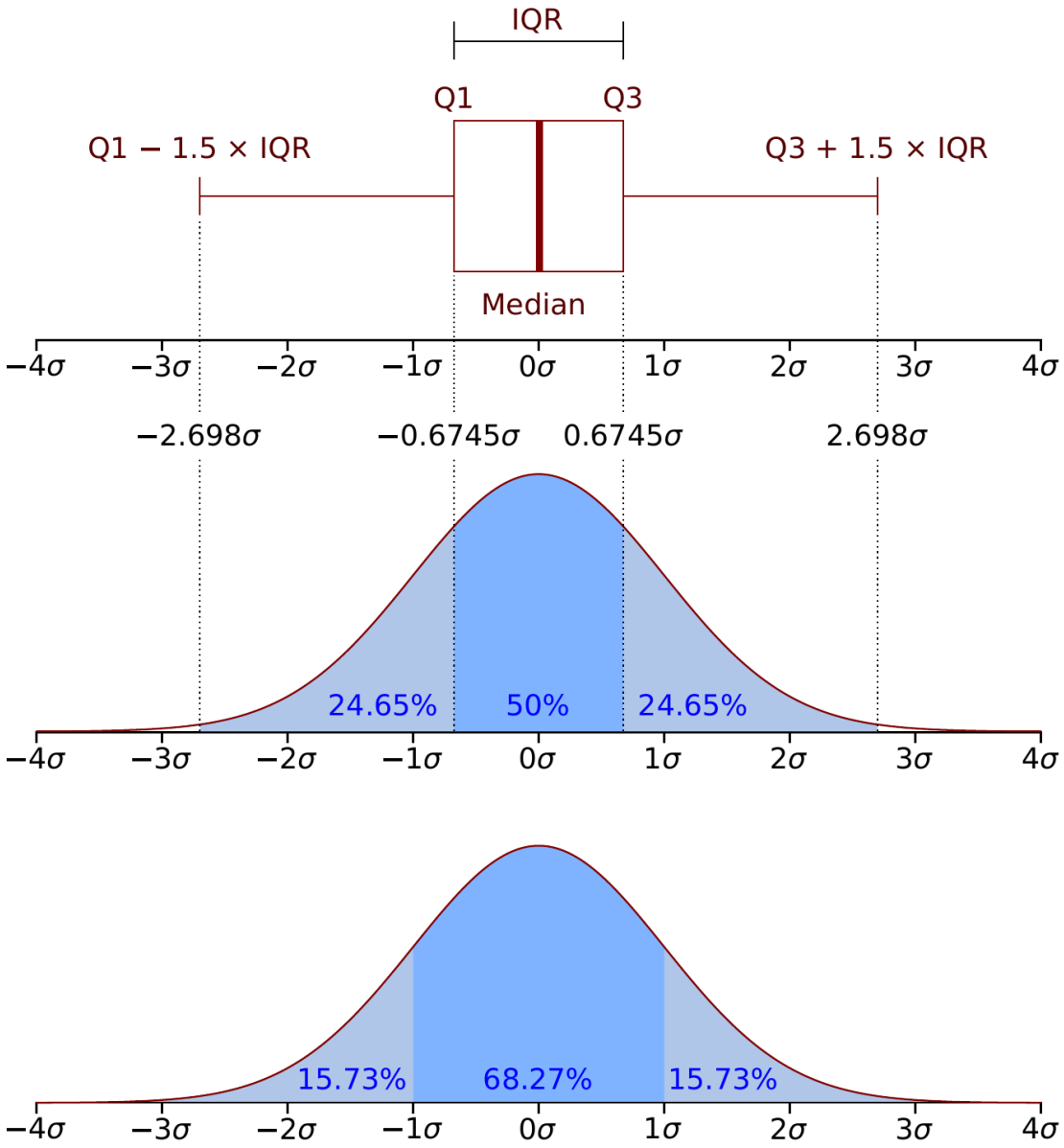


(b) Boxcar kernel



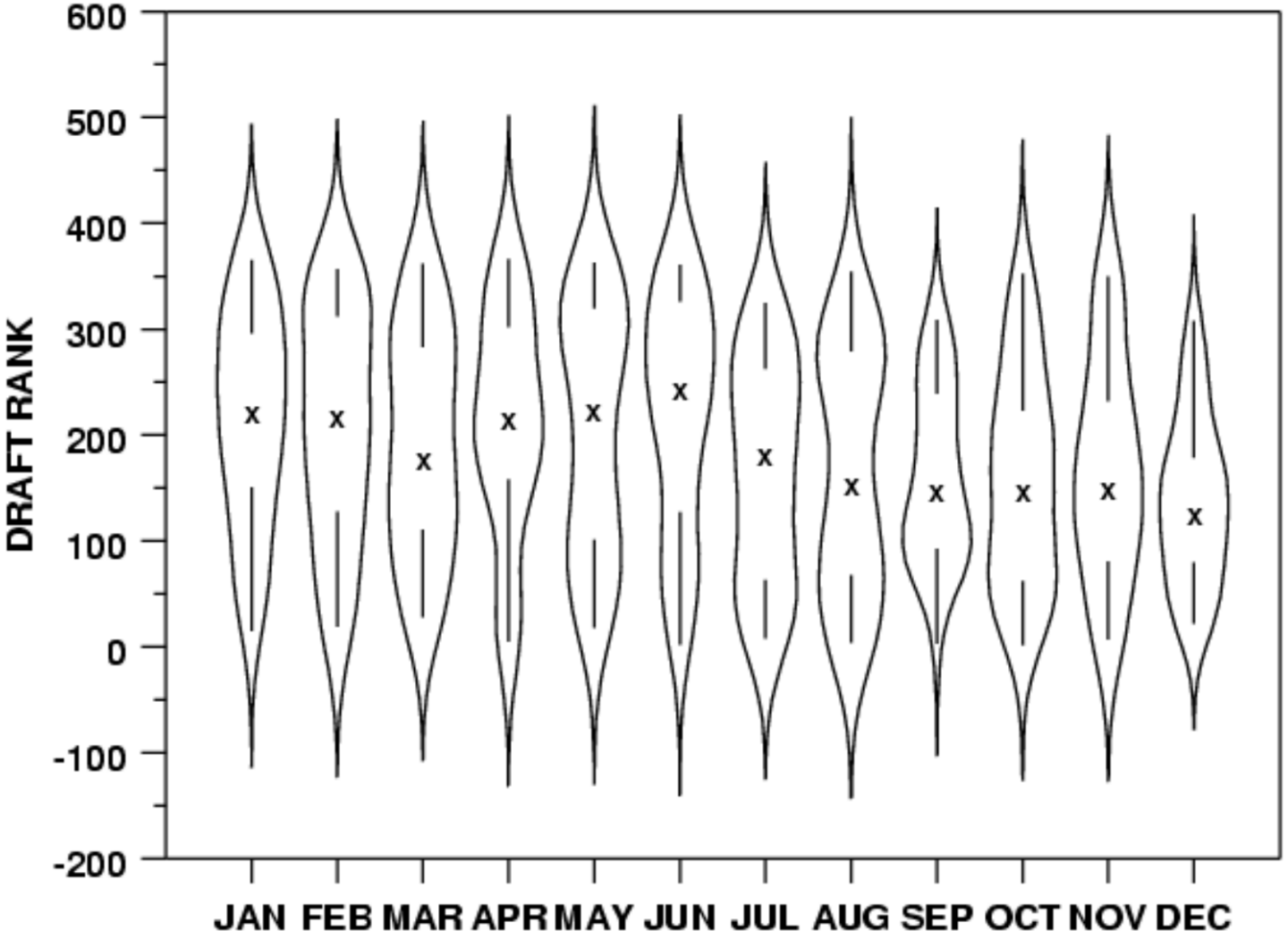
(c) Gaussian kernel

Box Plot



Source: https://en.wikipedia.org/wiki/Box_plot

Violin Plot



Source: https://en.wikipedia.org/wiki/violin_plot