Data Mining and Exploration

Spring 2018

Lecturer: Arno Onken

Email: aonken@inf.ed.ac.uk

Institute for Adaptive and Neural Computation

School of Informatics

THE UNIVERSITY *of* EDINBURGH

Edinburgh, 18th January 2018

Logistics

- 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/Demonstrator: Maria Astefanoaei
- Labs:
  - Wednesdays: 09:00 – 10:50 (only one lab per week)
  - Appleton Tower, room 6.06
- Presentations:
  - Poster presentations on research papers during second half of the course
  - Potential papers listed on the course website
- Mini-project:
  - Apply machine learning 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% presentation (10% poster presentation; 5% presentation summaries)

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

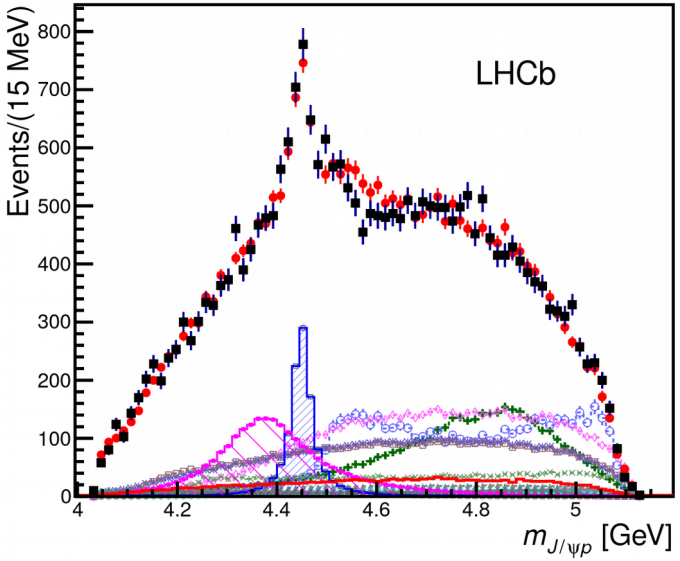
Source: https://commons.wikimedia.org/wiki/File:BigData_2267x1146_white.png

Data Analysis - Data Mining

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

Server Farm at CERN

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



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

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

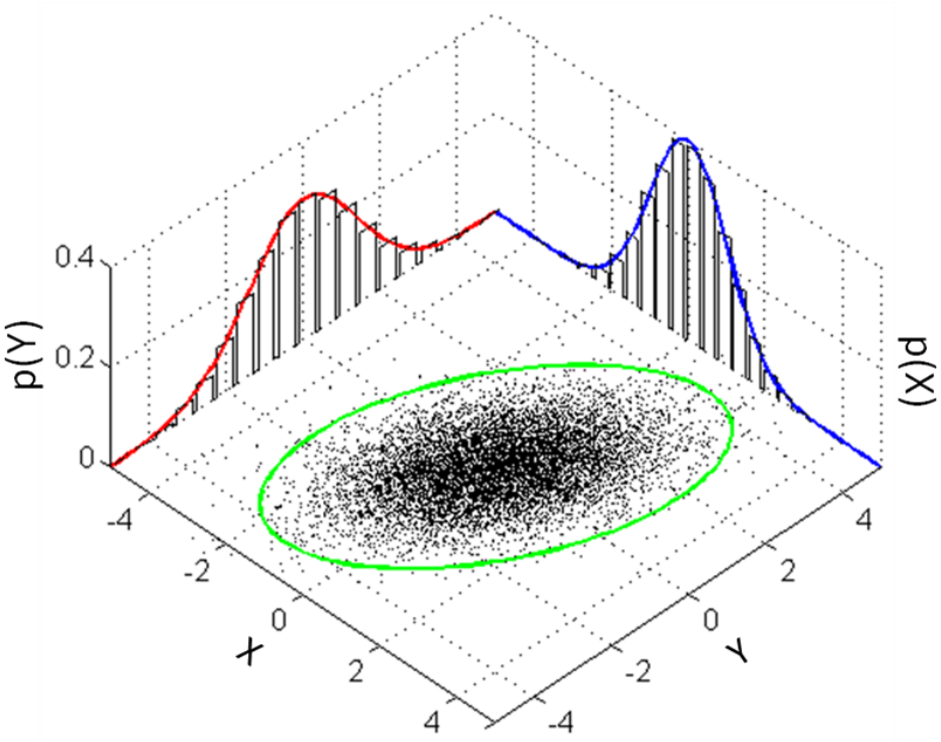
Exploratory Data Analysis

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

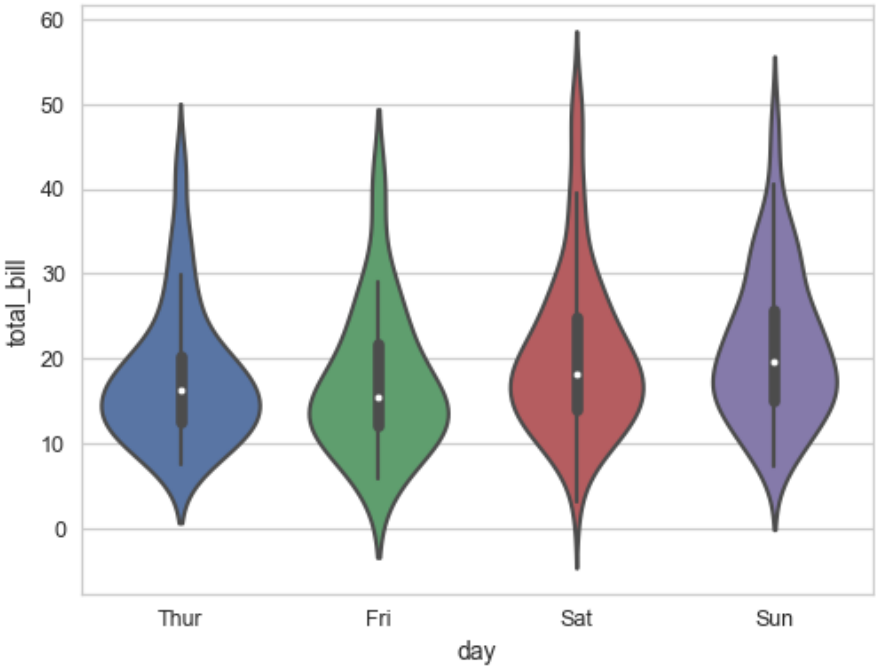
EDA emphasises:

- 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



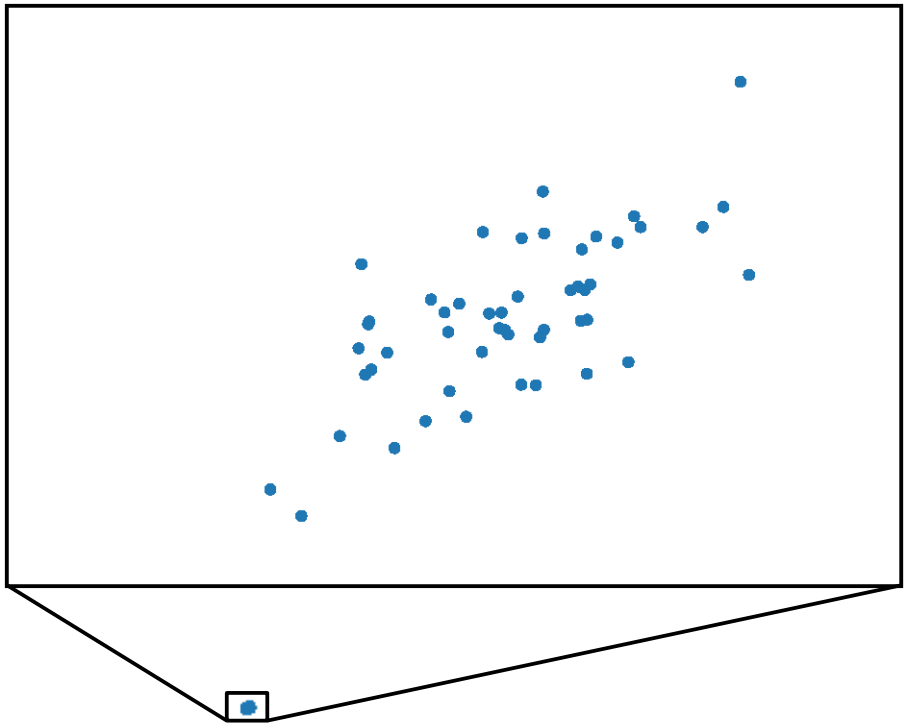
Source: https://commons.wikimedia.org/wiki/File:MultivariateNormal.png



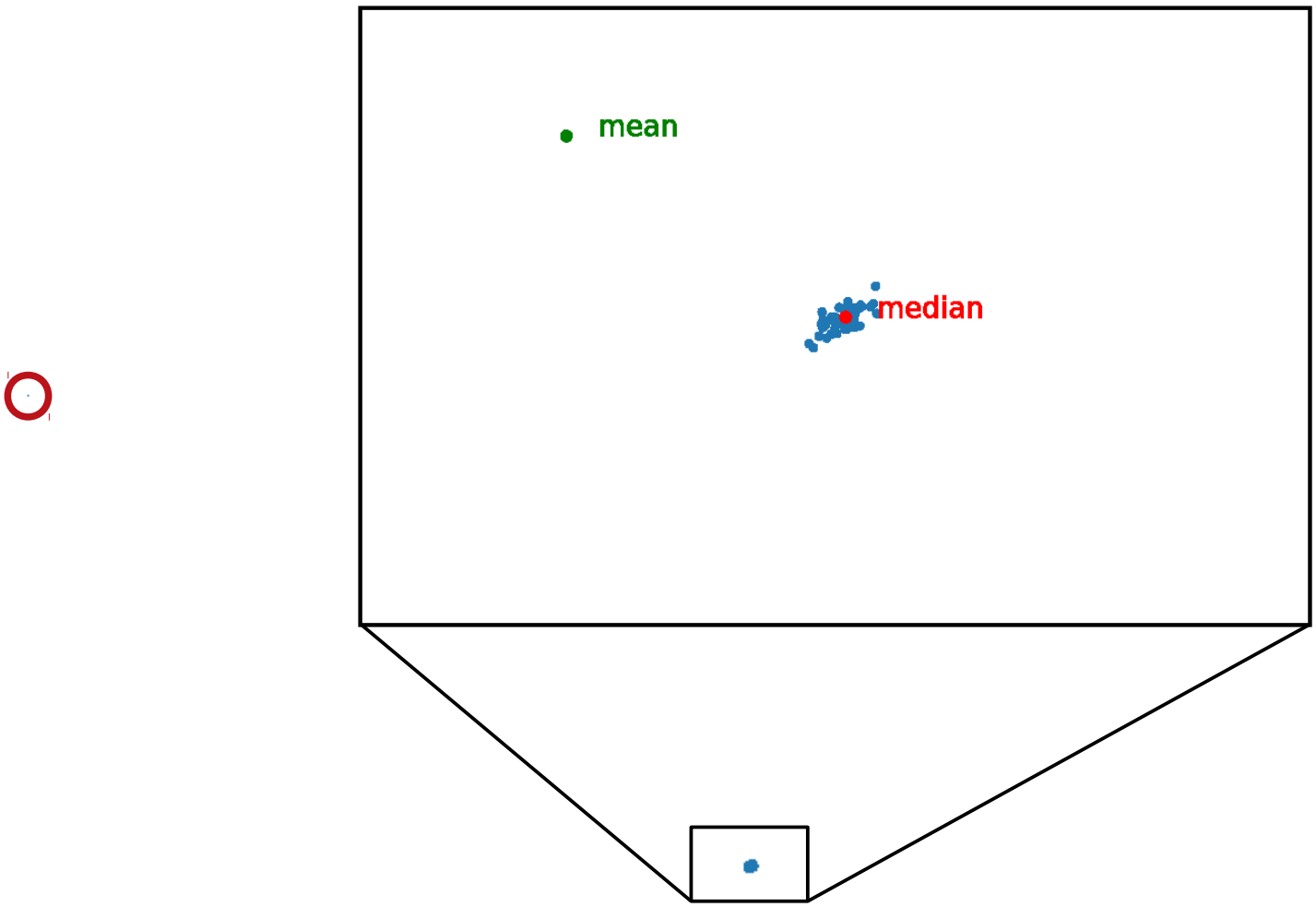
Source: https://seaborn.pydata.org/_images/seaborn-violinplot-2.png

EDA: Understanding of the Data Structure

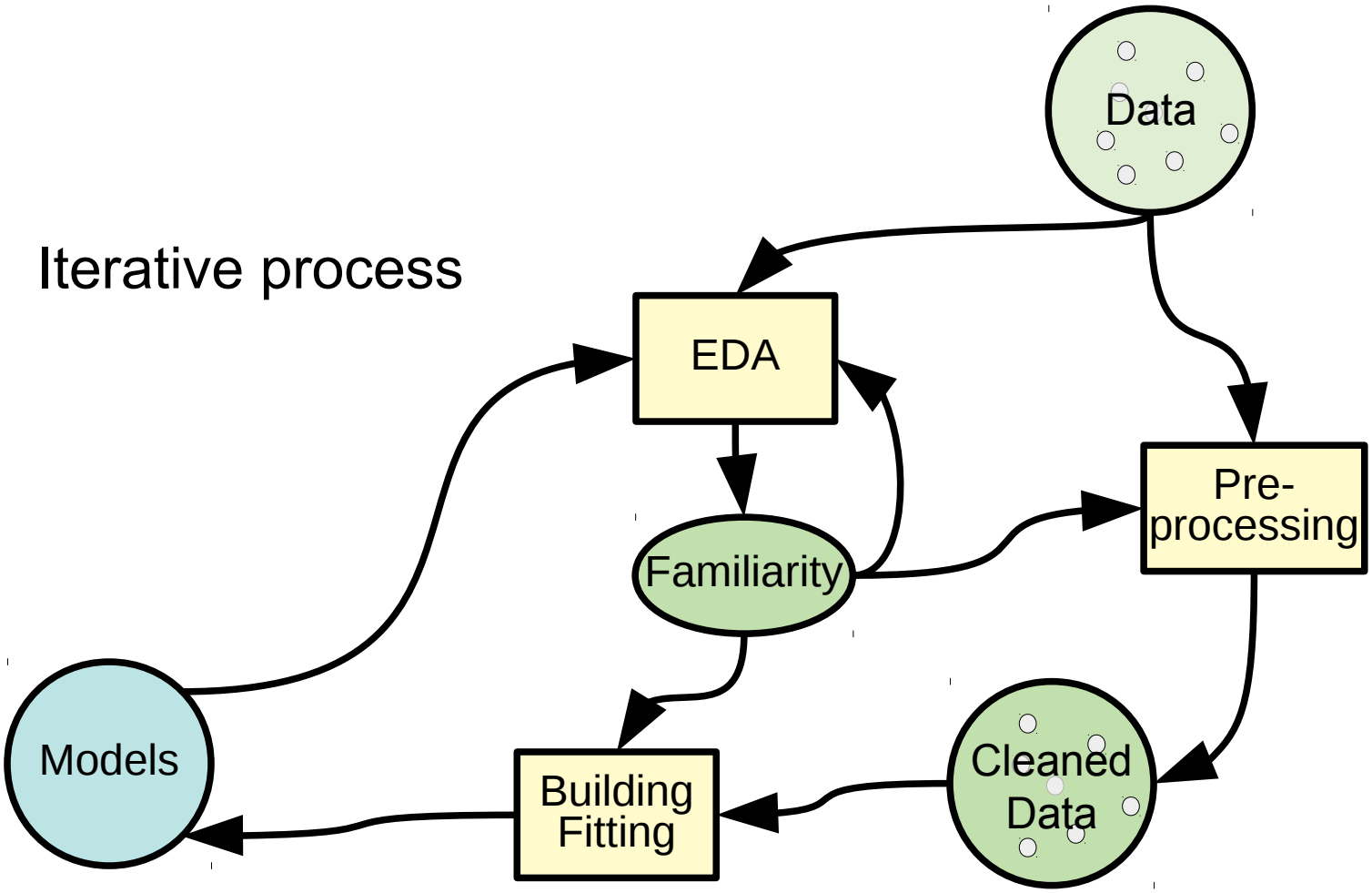
single outlier



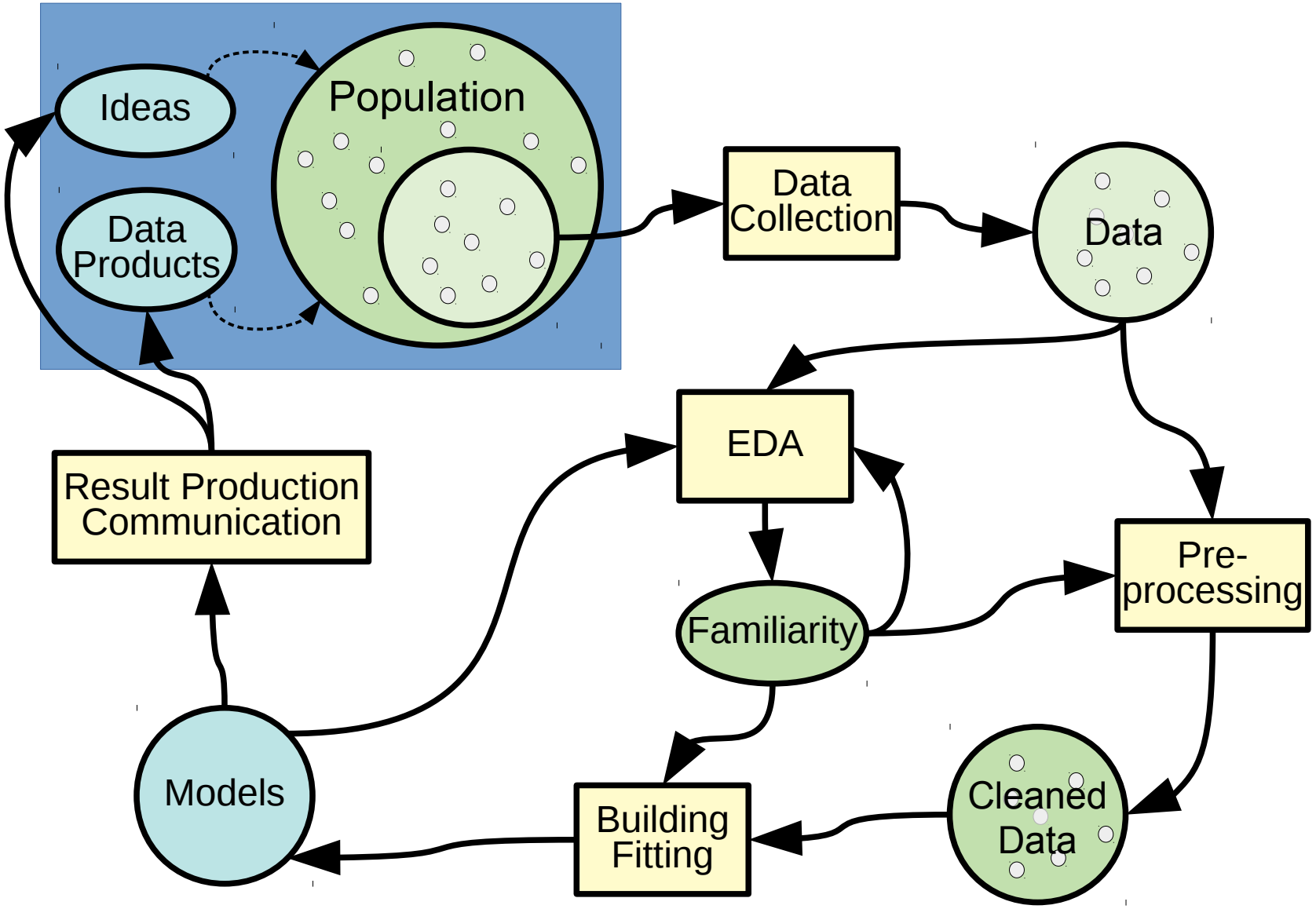
EDA: Robust Measures



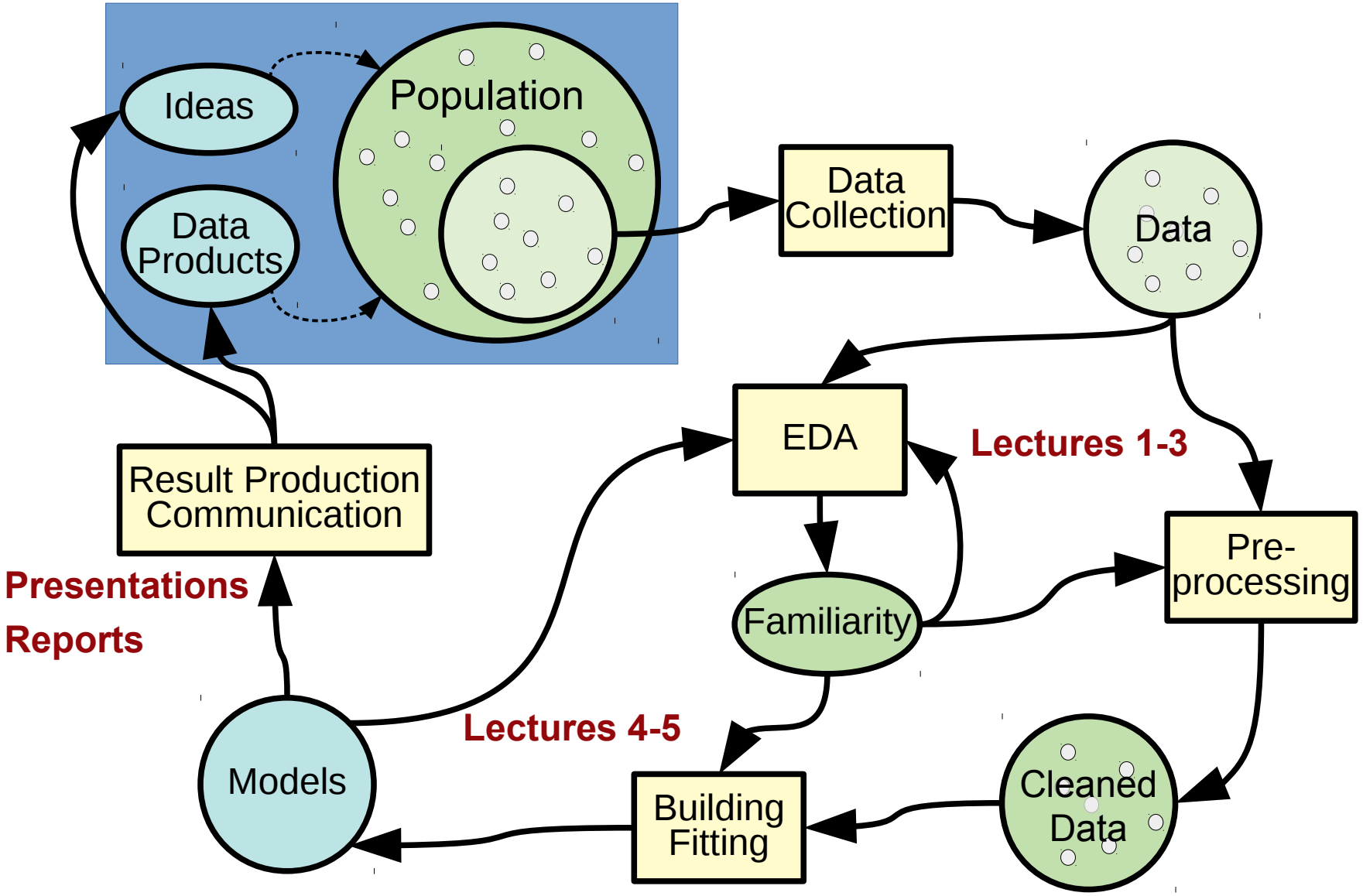
EDA: Tentative Model Building



Data Analysis Process



Course Content



Today

- Numerical data description
  - Univariate measures
    - Location: Where are the data located?
    - Scale: How spread out are the data?
    - Shape: How symmetric and extended are the data?
  - Multivariate measures:
    - Covariance and correlation – linear and nonlinear

      How strongly are variables associated?
- Data pre-processing
  - Standardisation
    - Centring matrix
    - Scaling to unit variance
  - Outlier detection and removal
- Data visualisation
  - Various plots