# 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

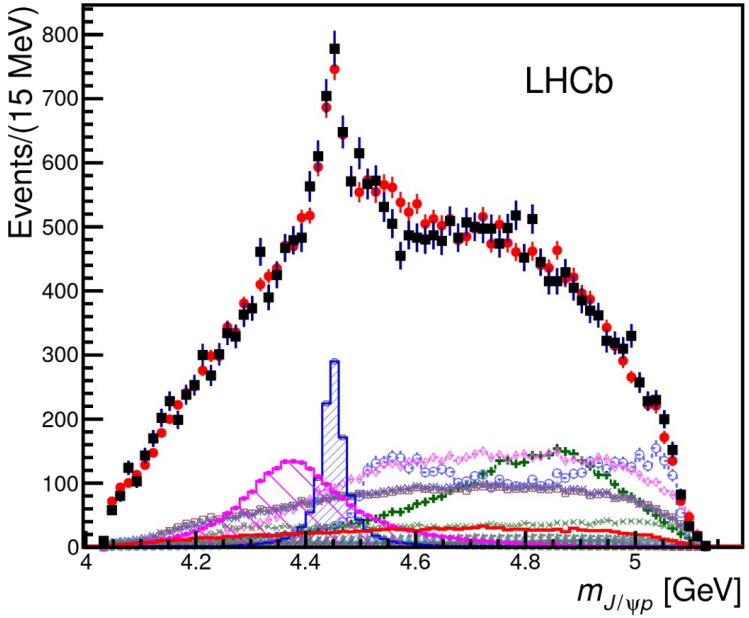

Source: https://commons.wikimedia.org/wiki/File:BigData_2267x1146_white.png

# 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

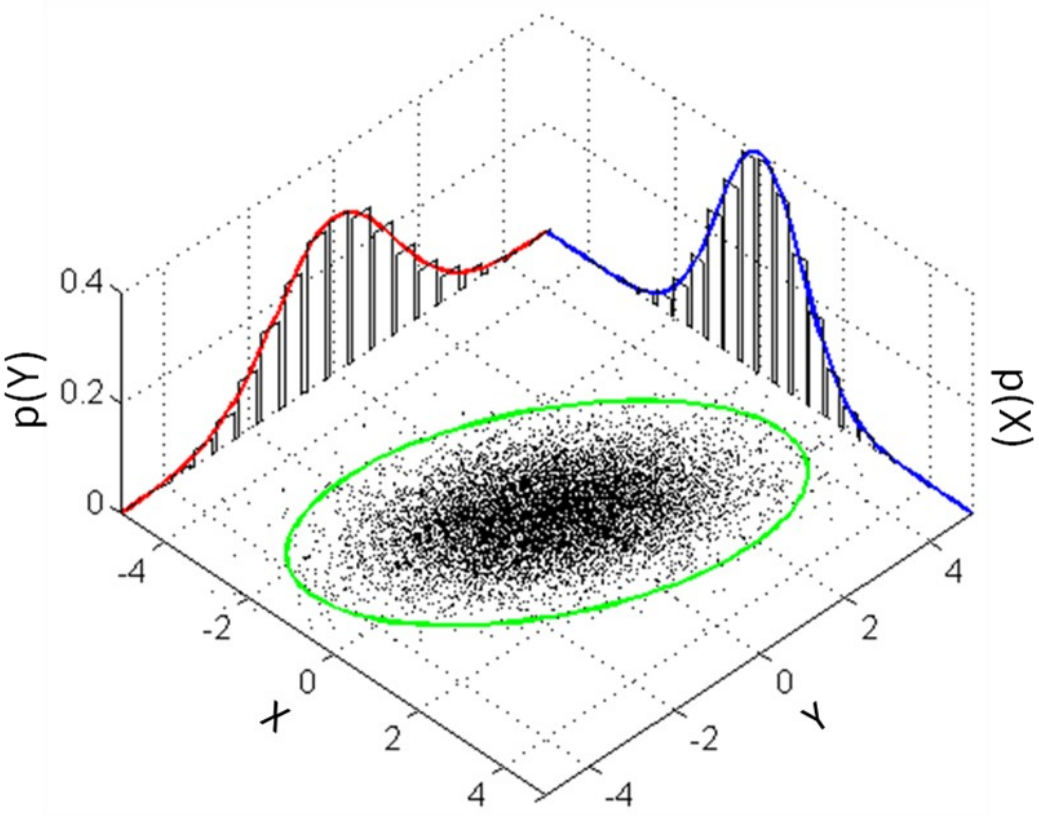
# Exploratory Data Analysis

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

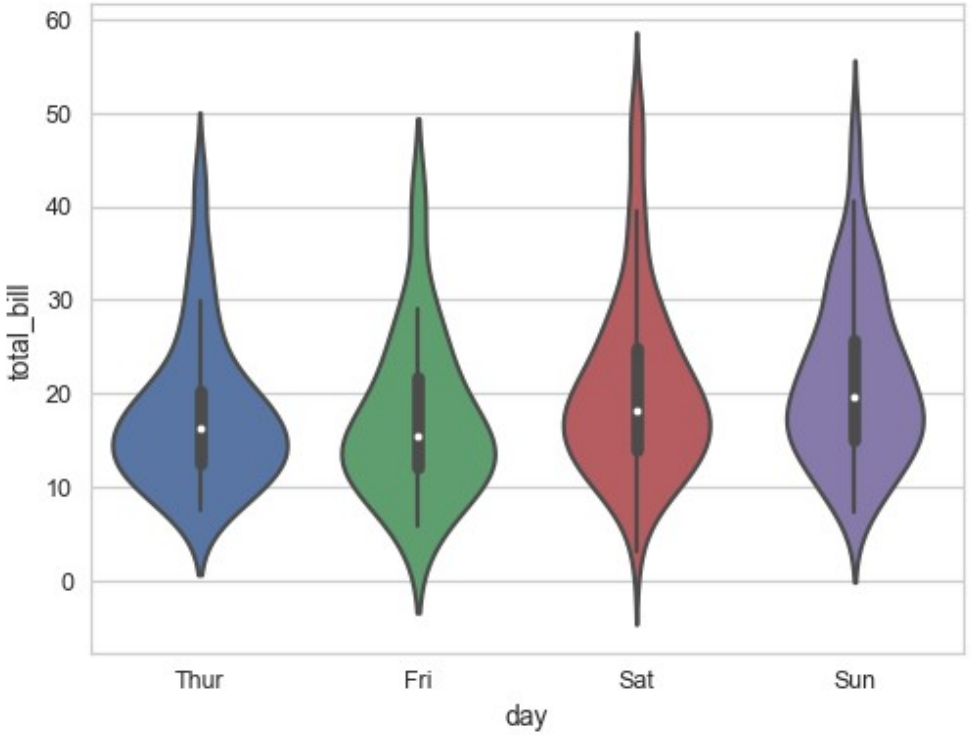
EDA emphasises:

- 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



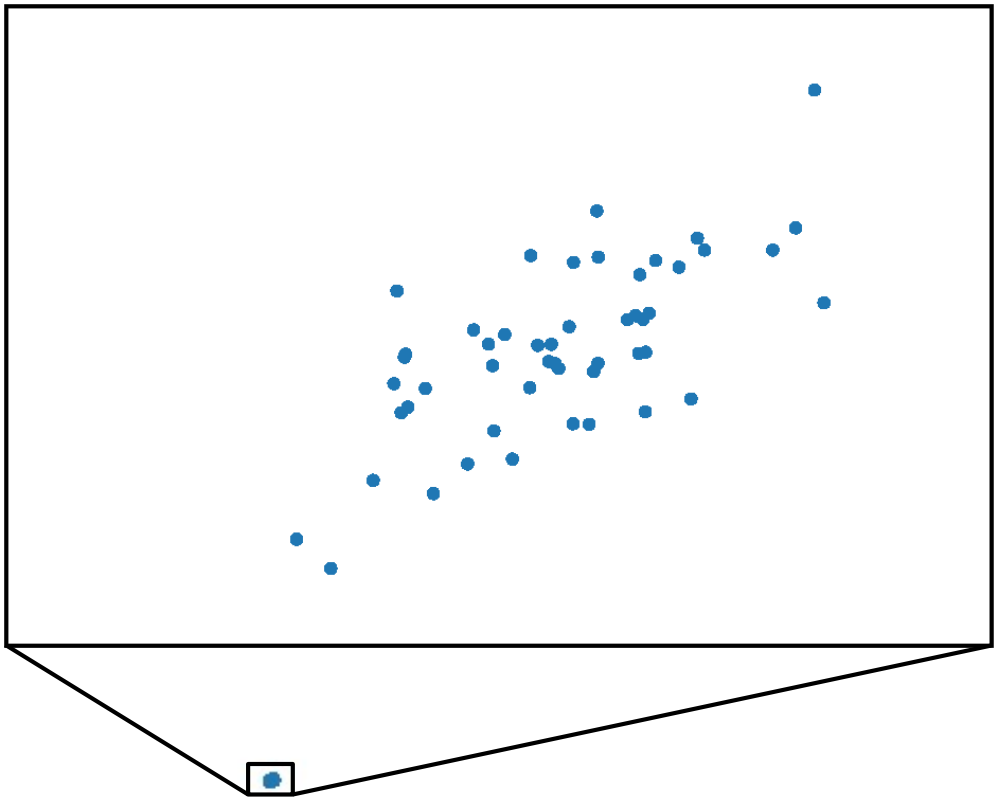

Source: https://commons.wikimedia.org/wiki/File:MultivariateNormal.png



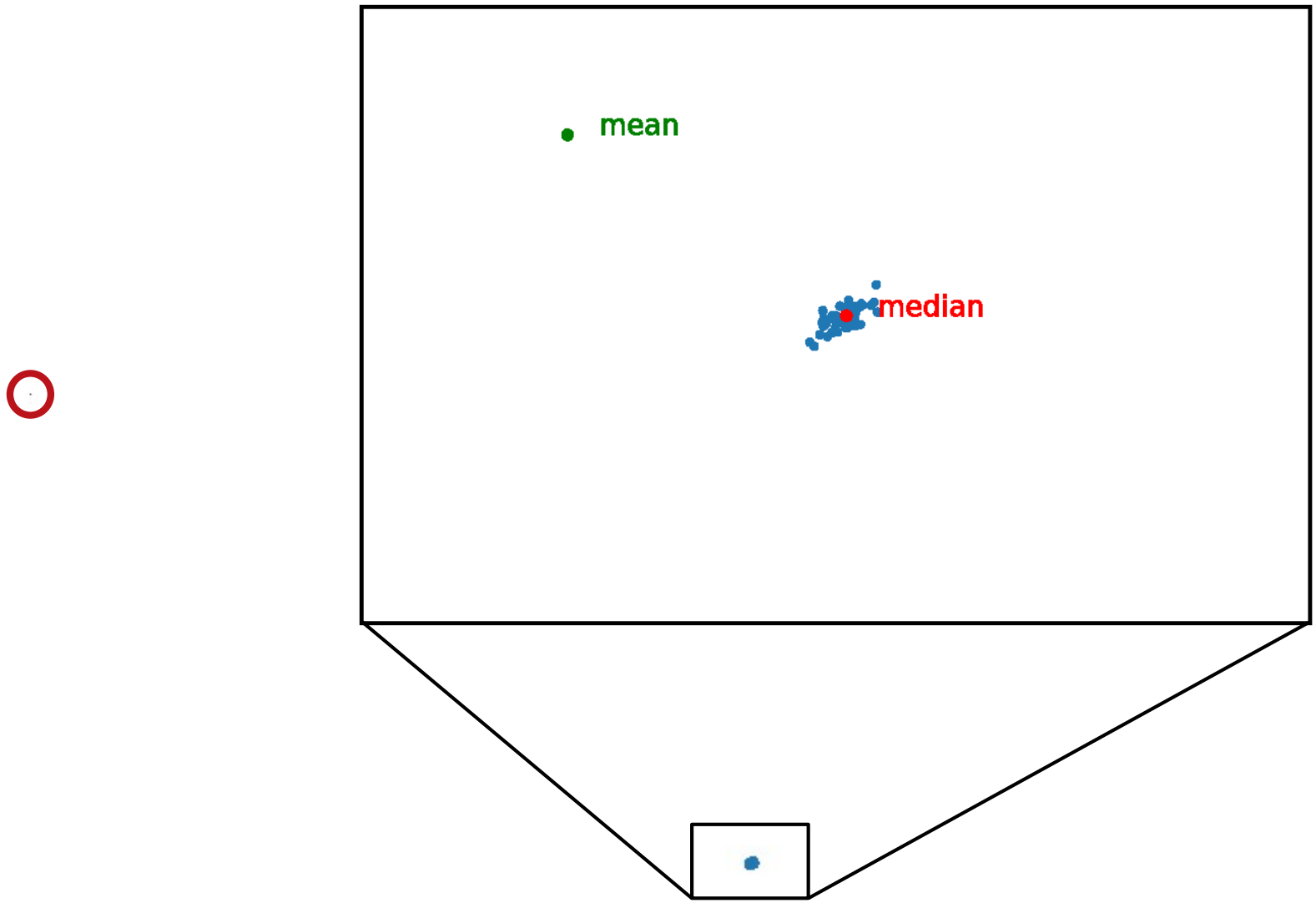

Source: https://seaborn.pydata.org/_images/seaborn-violinplot-2.png

# EDA: Understanding of the Data Structure

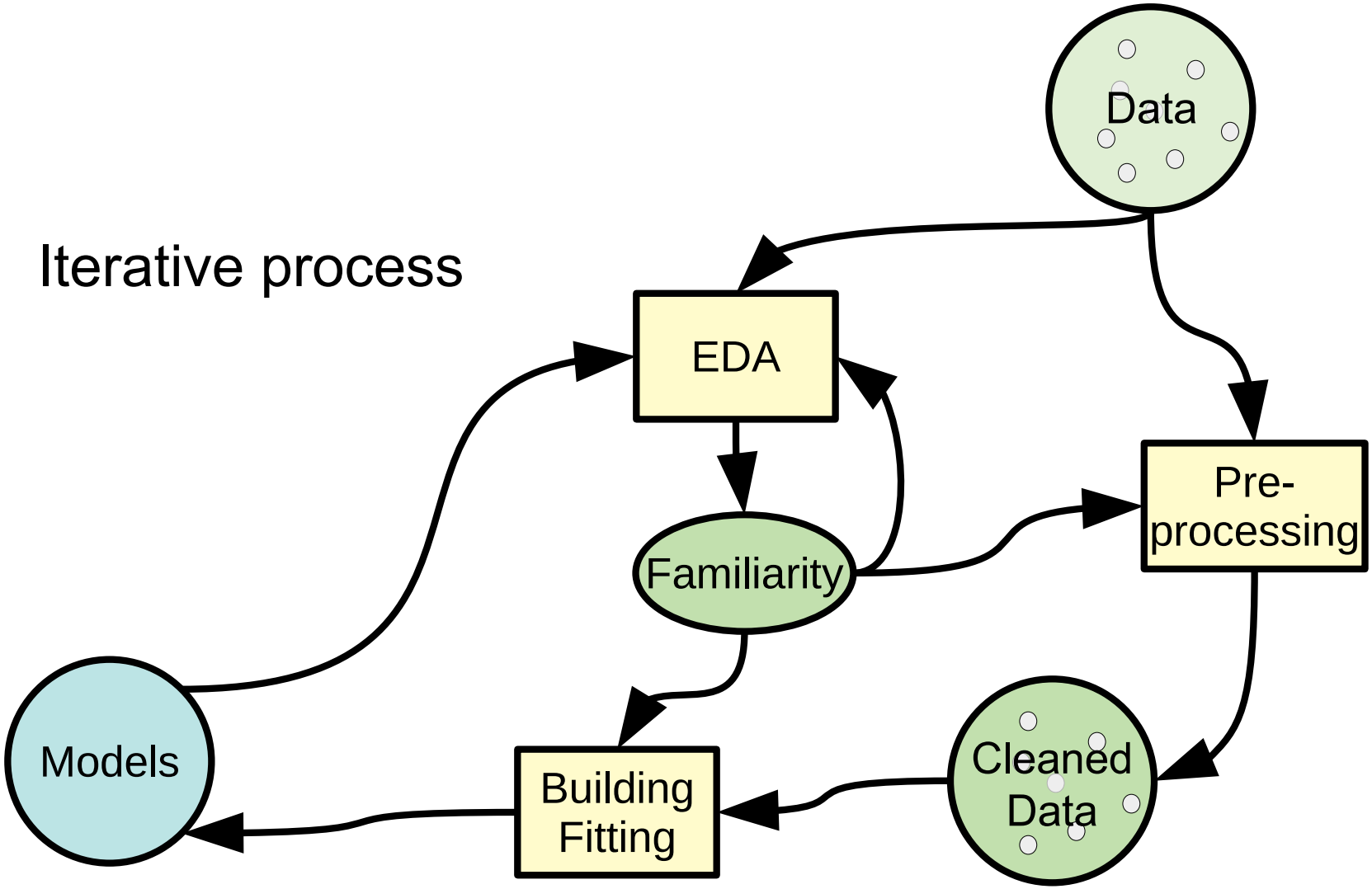
single outlier



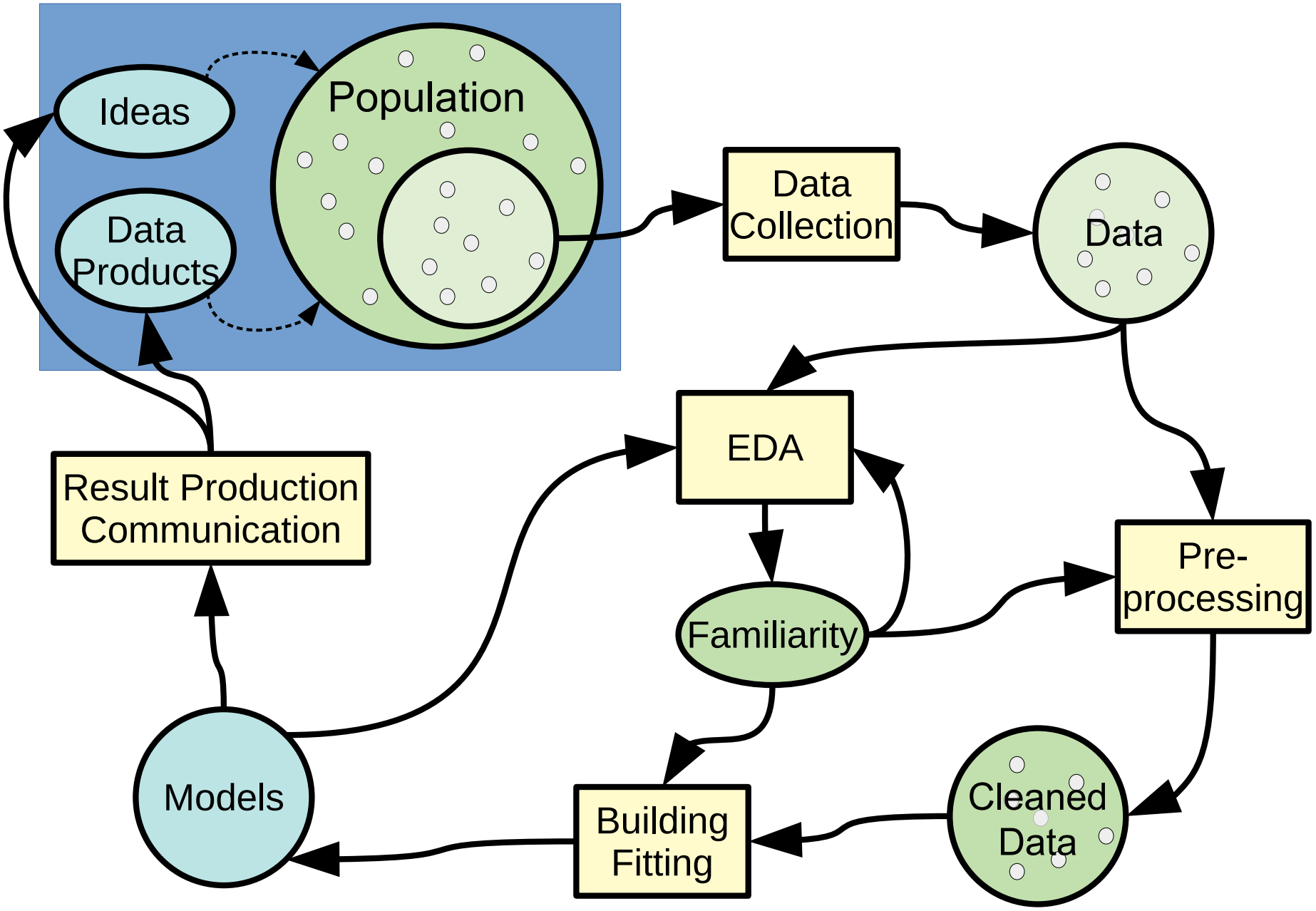
# EDA: Robust Measures



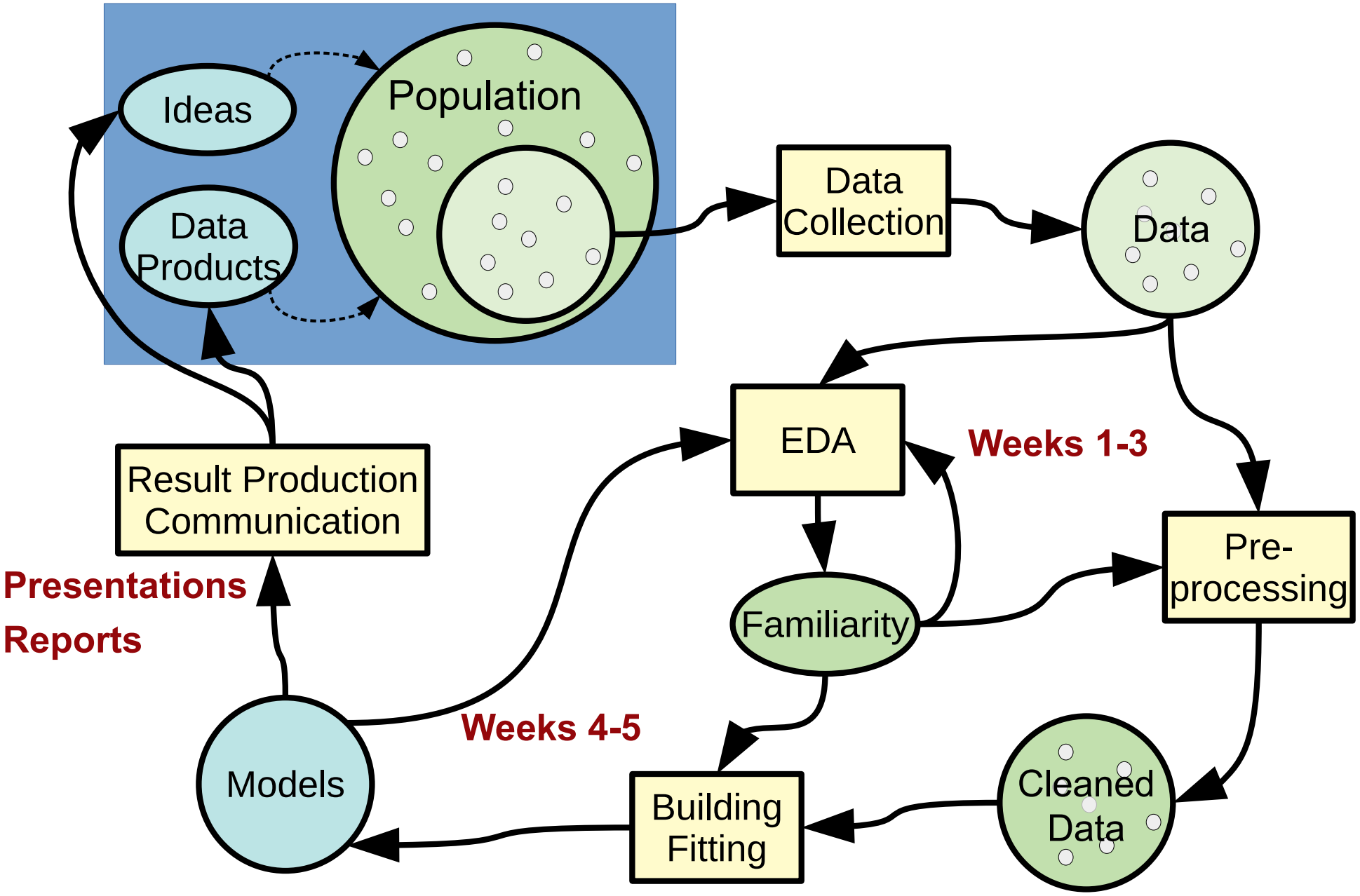
# EDA: Tentative Model Building



# Data Analysis Process



# Course Content



# Purpose of Particular Course Elements

- **Lecture material and computer labs**
  - Numerical data descriptions and pre-processing (Week 1)
    - Establish common language
    - Highlight importance of simple measures
  - In depth Principal Component Analysis (Week 2)
    - Describe important method in all its aspects
  - Dimensionality reduction (Weeks 3-4)
    - Closely related techniques
  - Predictive modelling and generalization (Week 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

# Numerical data description

Observations: $x = (x_1, x_2, \dots, x_n)$

Location measures:

- Sample mean: not robust

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is estimate of expected value:

$$\mathbb{E}[X] = \int y \, p(y) \, dy$$

For multivariate data: elementwise

- Median: robust

Splits samples into two chunks

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ denotes sorted data

$$\text{median}(x) = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{if } n \text{ is even} \end{cases}$$

Example:

$x = (0, 1, 1, 1, 2, 3, 4, 4, 5, 9)$

$\bar{x} = 3$, median $(x) = 2.5$

$y = (0, 1, 1, 1, 2, 3, 4, 4, 5, 9000)$

$\bar{y} = 902.1$, median $(y) = 2.5$

- Mode: robust

  Value that occurs most frequently

  • not necessarily unique

  • applicable even for unordered categorical data

- $\alpha$-th sample quantile $q_\alpha$:

  Proportion $\alpha$ of ordered data $x_{(i)}$ to the left of $q_\alpha$

  $q_0 = \min(x)$, $q_1 = \max(x)$ $q_{0.5} = \text{median}(x)$

- Quartiles:

  Split data into four partitions

  $Q_1 = q_{0.25}$ $Q_2 = q_{0.5} = \text{median}(x)$ $Q_3 = q_{0.75}$

Example:

$x = (0, 1, 1, 1, 2, 3, 4, 4, 5, 9)$

$Q_1$ $Q_2$ $Q_3$

# Scale measures

- Sample variance: not robust (even less robust than $\bar{x}$)

$$var(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is estimate of the variance

$$Var[X] = \int (\xi - \mathbb{E}[X])^2 \, p(\xi) \, d\xi$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- Sample standard deviation: not robust

$$std(x) = \sqrt{var(x)}$$

- Median absolute deviation: robust

$$MAD(x) = median(|x_i - median(x)|)$$

- Interquartile range: robust

$$IQR = Q_3 - Q_1$$

Useful for outlier detection

Example:

$x = (0, 1, 1, 1, 2, 3, 4, 4, 5, 9)$

$Q_1$ ... $Q_3$, $IQR = 4 - 1 = 3$

$std(x) = \sqrt{6.4} \approx 2.53$ $\quad std(y) = 284.5.10$

$MAD(x) = 1.5$

2021111