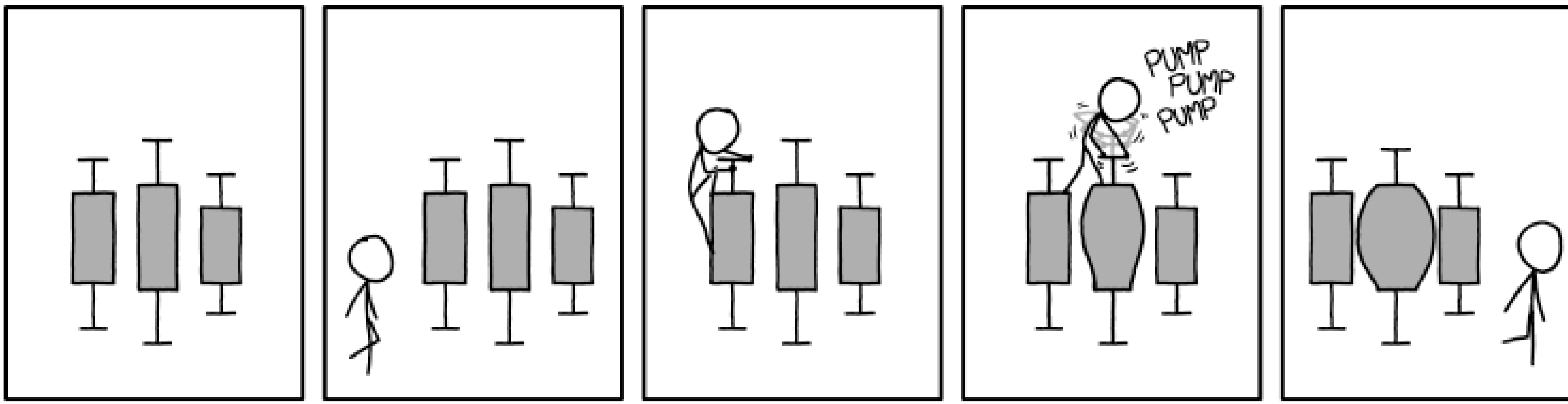
Data Analysis

Descriptive Statistics and data exploration

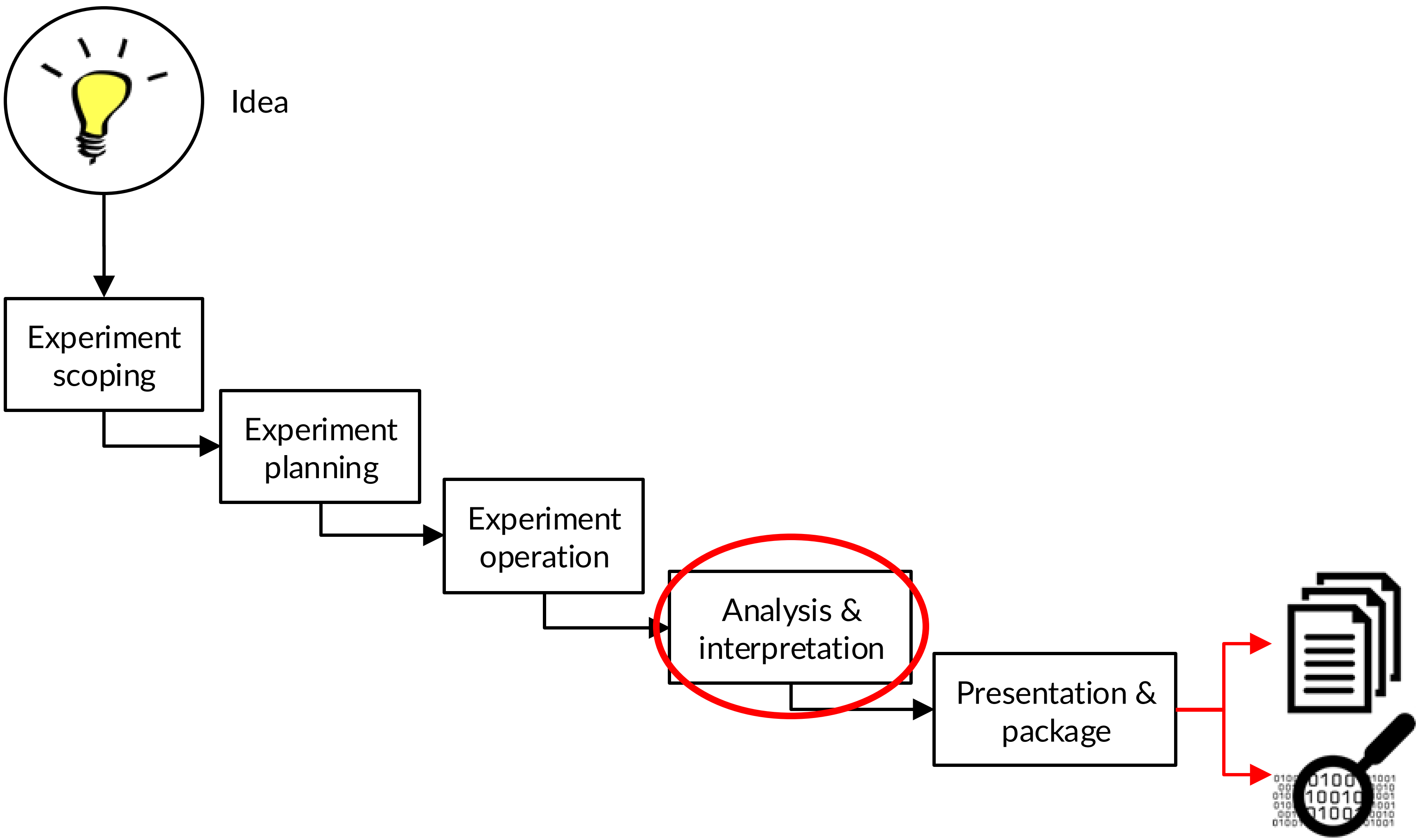
Ivano Malavolta



VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Quick Recap



Idea
Experiment scoping
Experiment planning
Experiment operation
Analysis & interpretation
Presentation & package

Analysis and Interpretation

- Understanding the data
  - descriptive statistics
  - Exploratory Data Analysis (EDA, e.g. boxplots, scatter plots)
- Data preparation (if needed)
- Data transformation (if needed)
- Hypothesis testing
- Effect size estimation
- Results interpretation

Descriptive Statistics

- Goal: get a ‘feeling’ about how data is distributed

- Properties:
  - Central tendency (e.g. mean, median)
  - Dispersion (e.g. frequency, standard deviation)
  - Dependency (e.g., correlation)

Parameter vs. statistic

- **Parameter**: feature of the population
  - μ: mean
  - σ: standard deviation
- **Statistic**: feature of the sample
  - $\bar{x}$: mean
  - s: standard deviation
- Statistics are an *estimation* of parameters

Central Tendency

- Arithmetic mean: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Geometric Mean: $$GM(x) = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- It is like the arithmetic mean, but with **multiplication**

→ used when collected data is not ”additive”, but “multiplicative”

- Less sensible to outliers
- Report it when the range of the considered values is very large

Central tendency

- Median (or 50% percentile): middle value separating the greater and lesser halves of a data set

$$\tilde{x} = x_{50\%}$$

X = [13, 18, 13, 14, 13, 16, 14, 21, 13]

X_{sort} = [13, 13, 13, 13, 14, 14, 16, 18, 21]

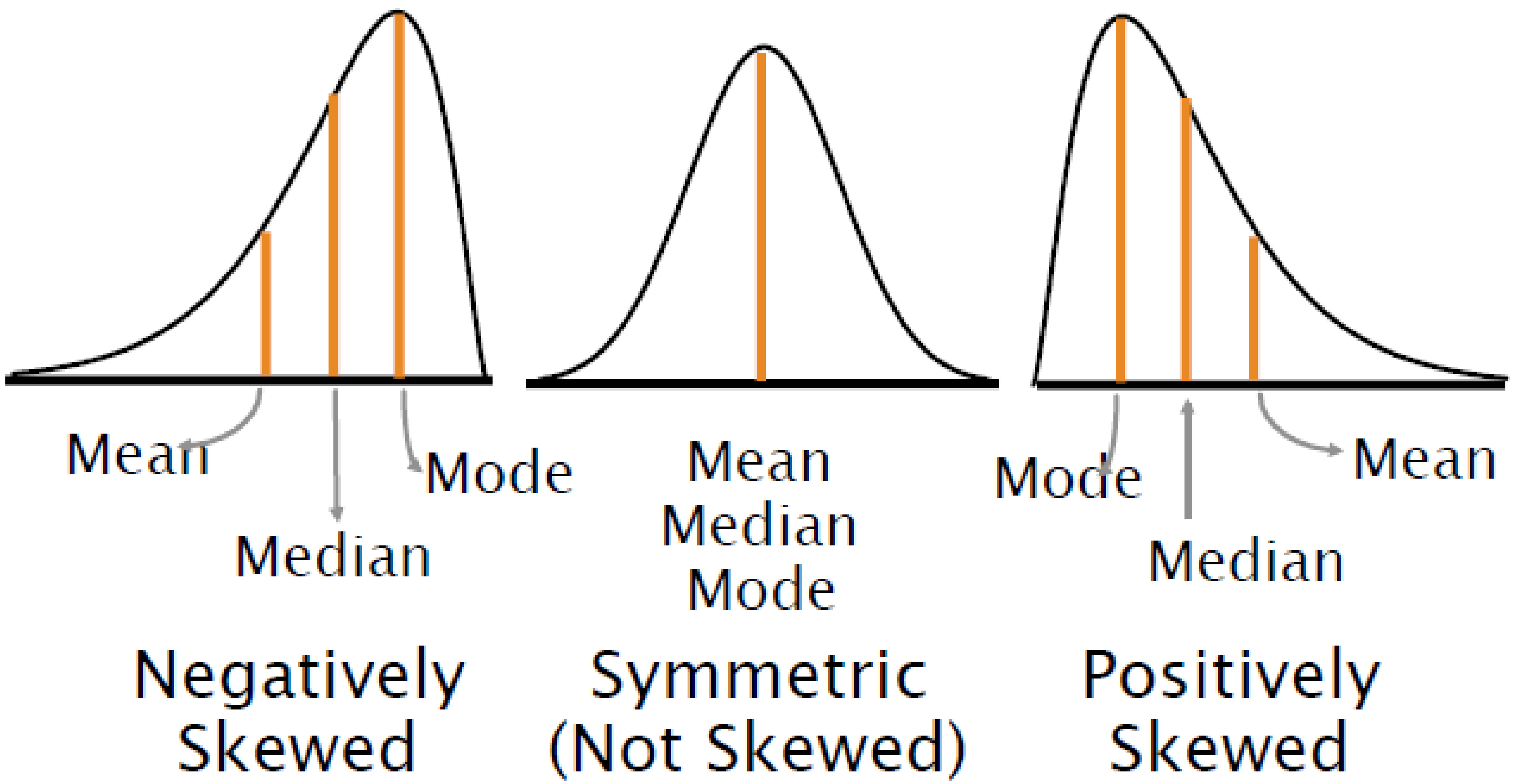
Central tendency

- Mode: most frequent value in data set

$$X = [13, 18, 13, 14, 13, 16, 14, 21, 13]$$

$$Mo_X = 13$$

Central tendency - Skewness



Dispersion

- Variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

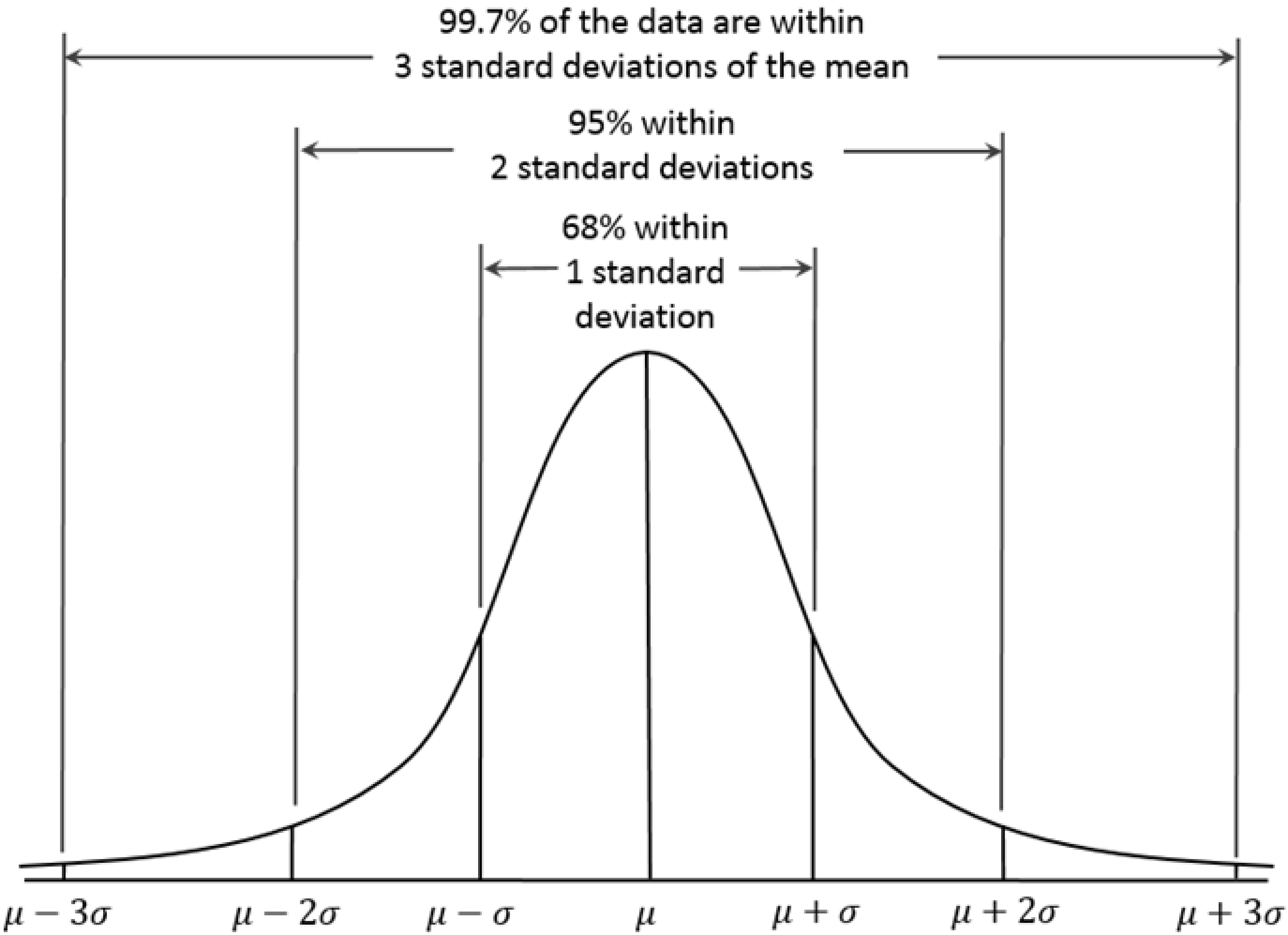
Informally: it gives an idea about how ”sparse” is data

- Standard Deviation: $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Informally: everything which is within 1 SD from the mean is “normal”

- Standard Deviation is **dimensionally equivalent** to the data

Dispersion - three-sigma-rule



Dispersion – Range and Coefficient of variation

- Range: $x_{max} - x_{min}$

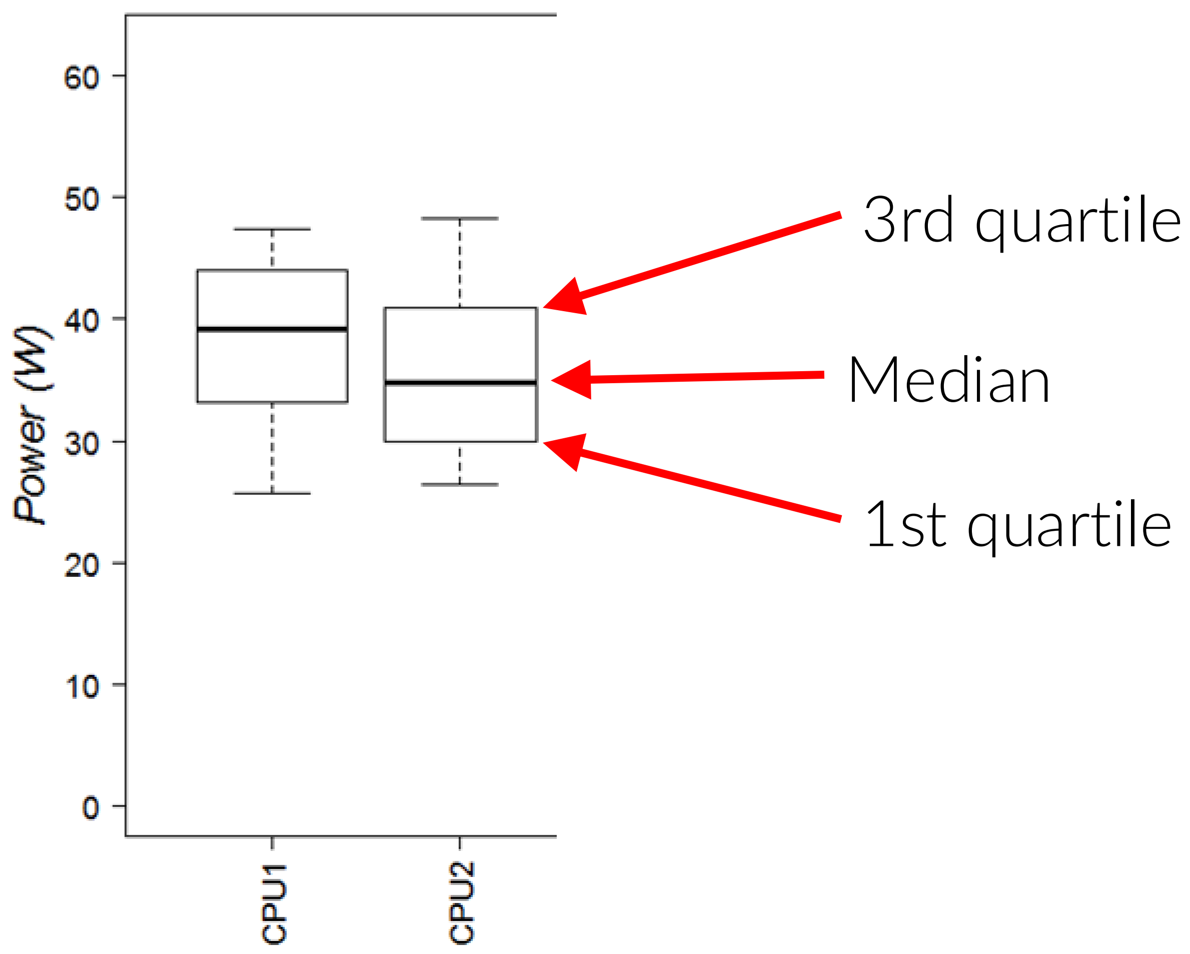
It is useful if you want to **compare the dispersion of variables with different units of measure**

- Coefficient of variation: (in percentage of mean) $CV = 100\frac{s}{x}$

- Coefficient of variation only has meaning if all values are **positive** (*ratio* scale)

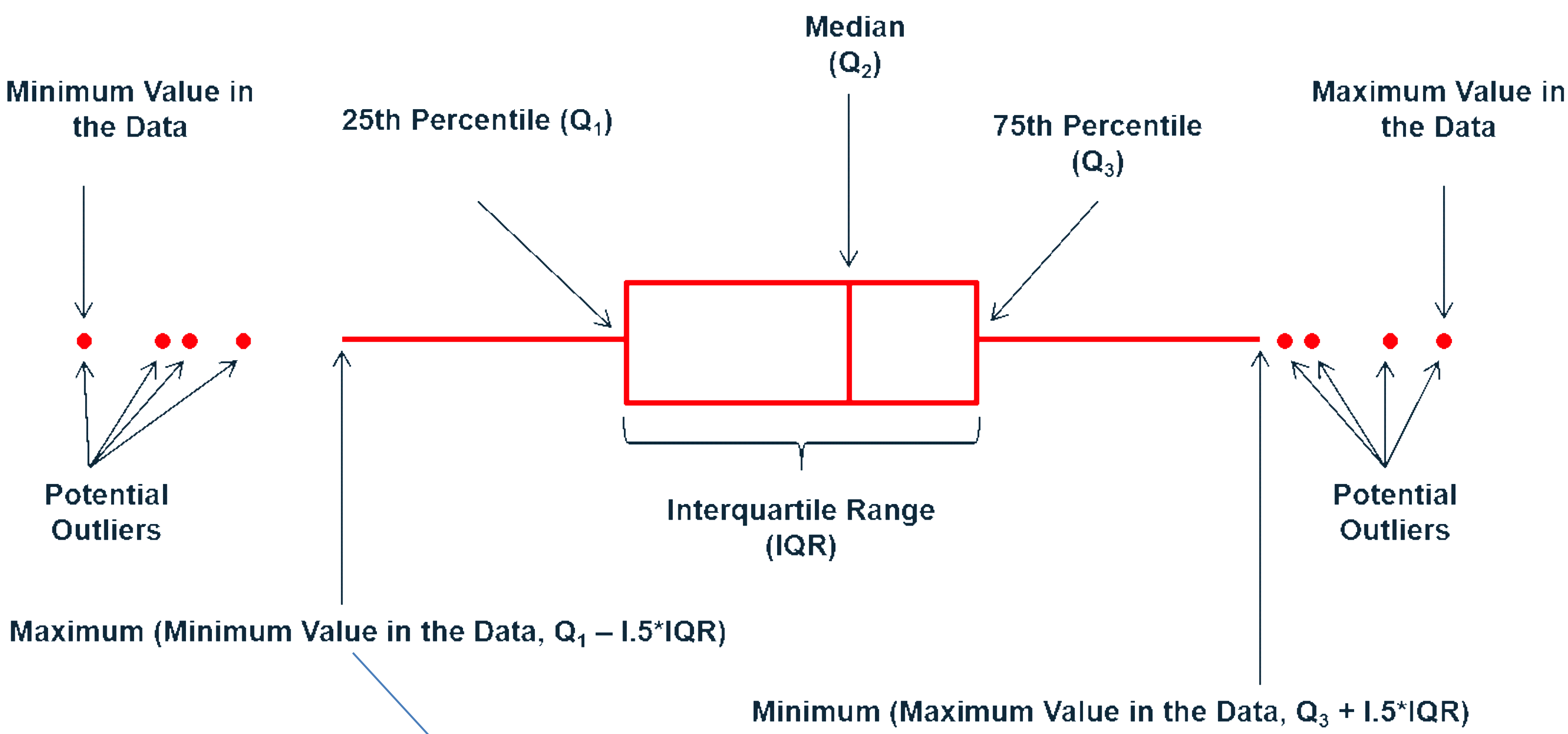
Basic visualizations

Box Plot



Basic visualizations

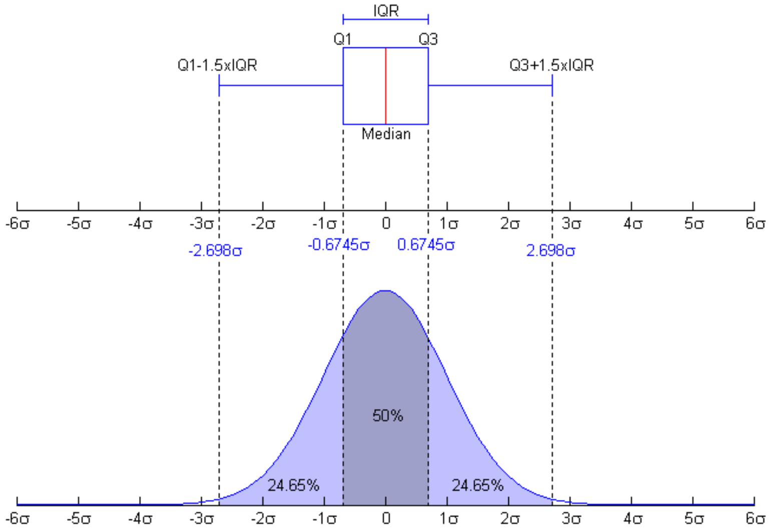
Box Plot



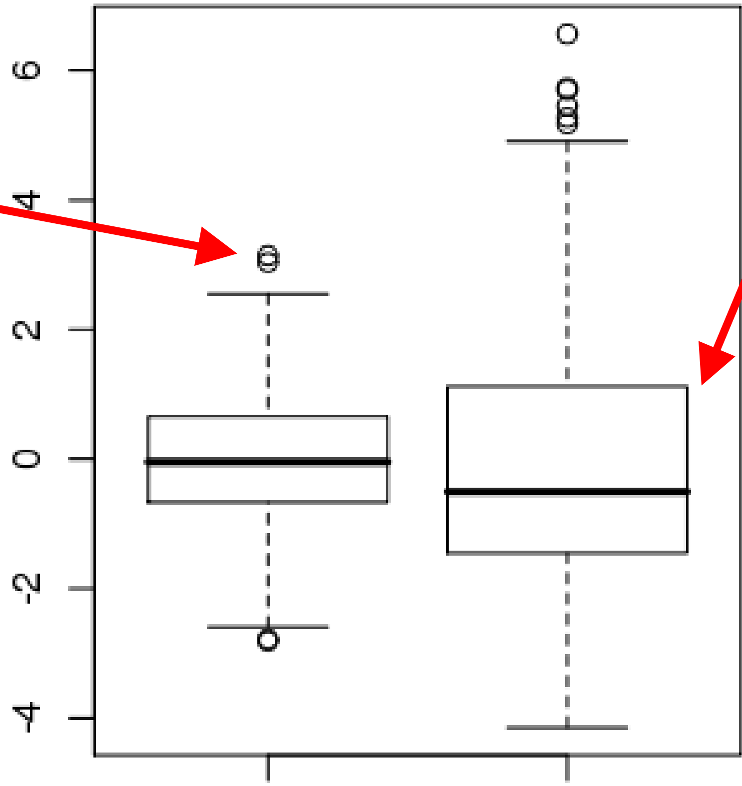
Minimum/maximum values THAT ARE NOT OUTLIERS

Basic visualizations

Box Plot



outliers



positive skewness

By Gbdivers (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Dependency: correlation

- Meaningful when comparing *paired* values/datasets

- Sample correlation coefficient (Pearson):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$$

Dependency: example

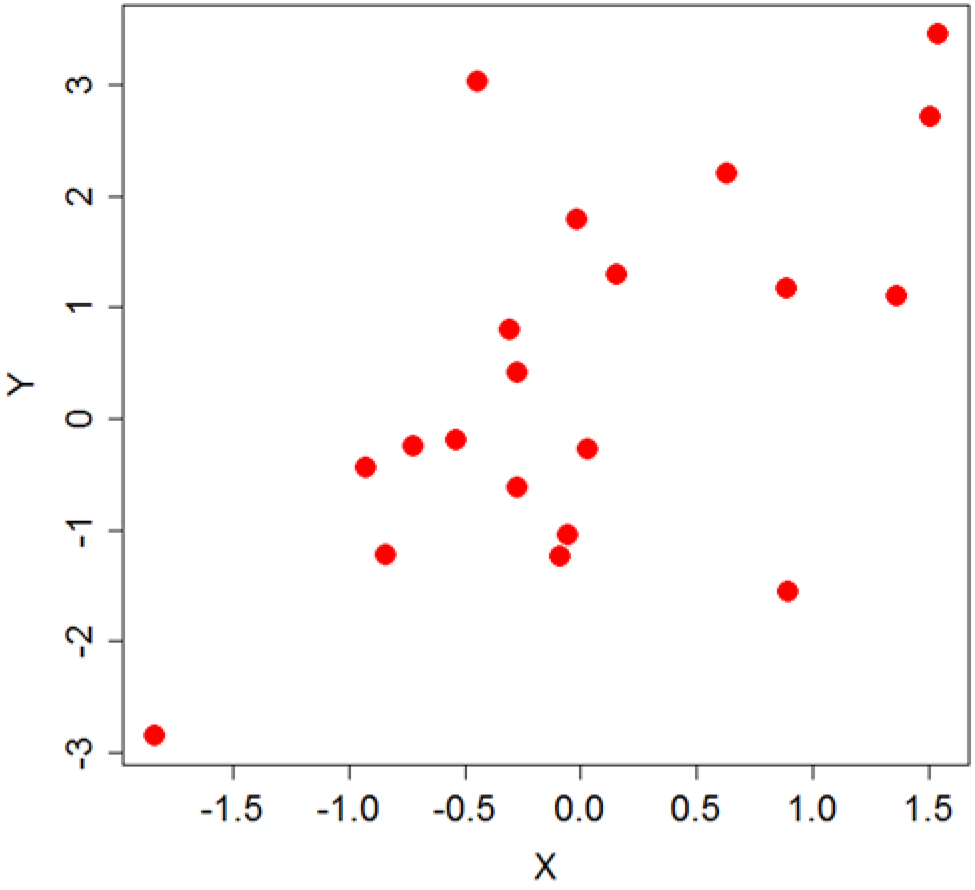
23	9.5
23	27.9
27	7.8
27	17.8
39	31.4
41	25.9
45	27.4
49	25.2
50	31.1
53	34.7
53	42.0
54	29.1
56	32.5
57	30.3
58	33.0
58	33.8
60	41.1
61	34.5

Age vs. body fat %

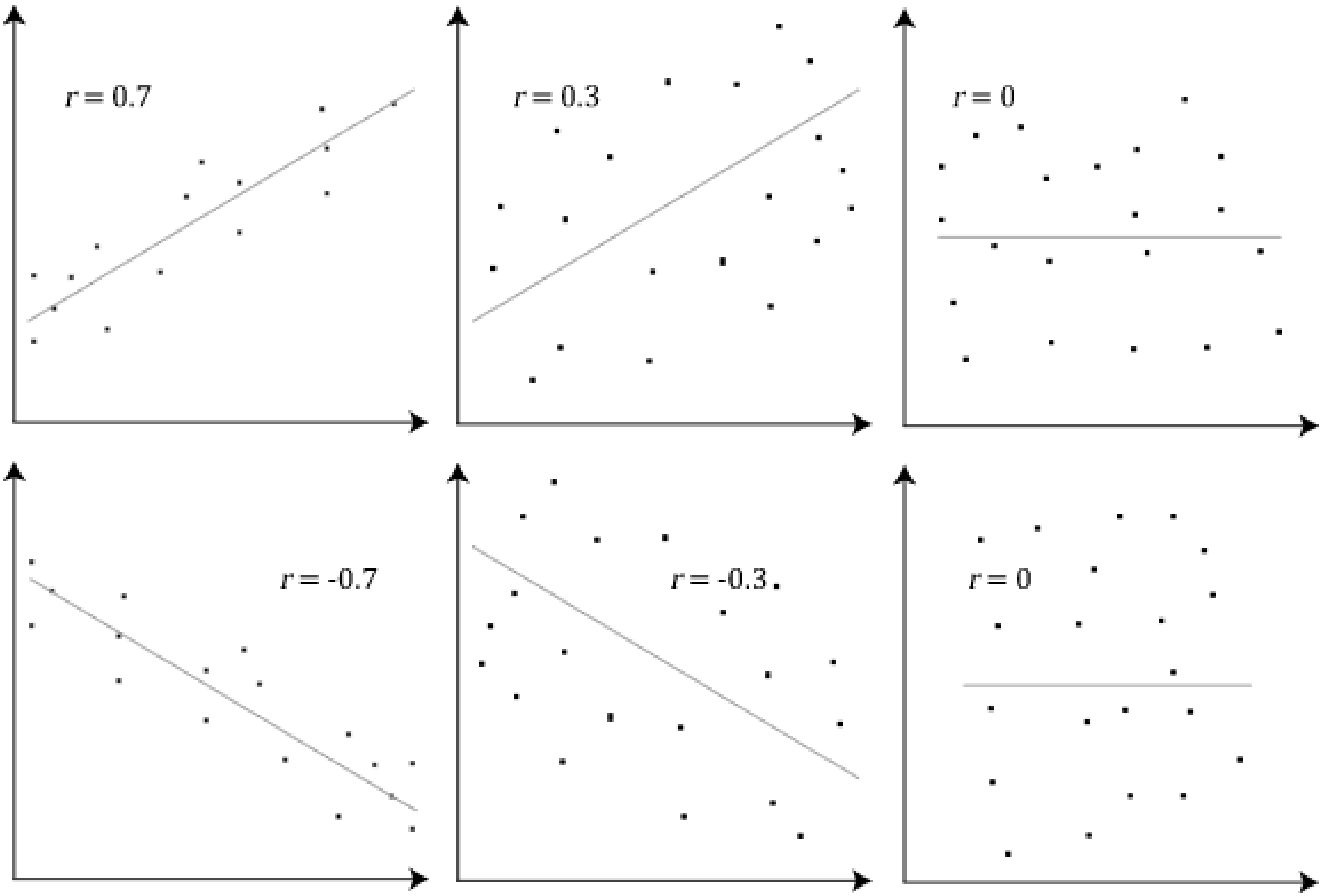
- Pearson: $r = 0.7921$
- Spearman: $\rho = 0.7539$
- Kendall: $\tau = 0.5762$

Basic Visualizations

Scatter Plot

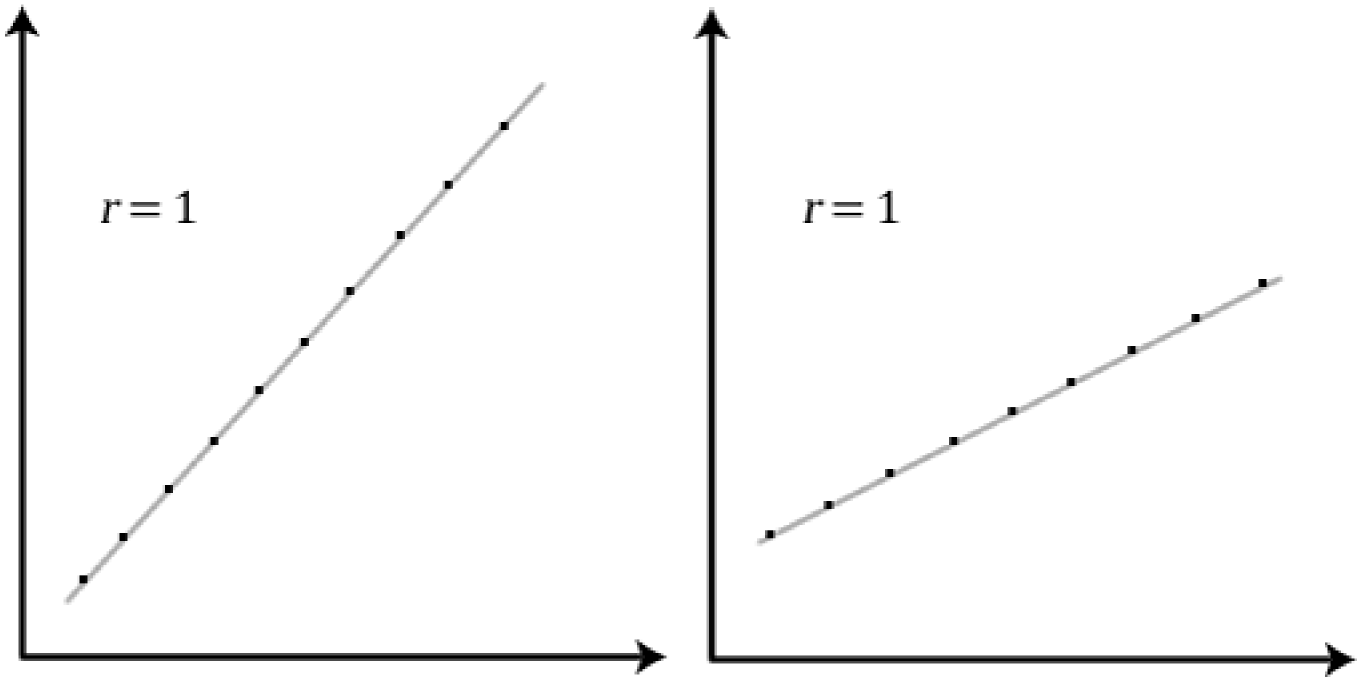


Positive VS negative correlation



https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

It does NOT indicate the slope of the line



https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Dependency: correlation

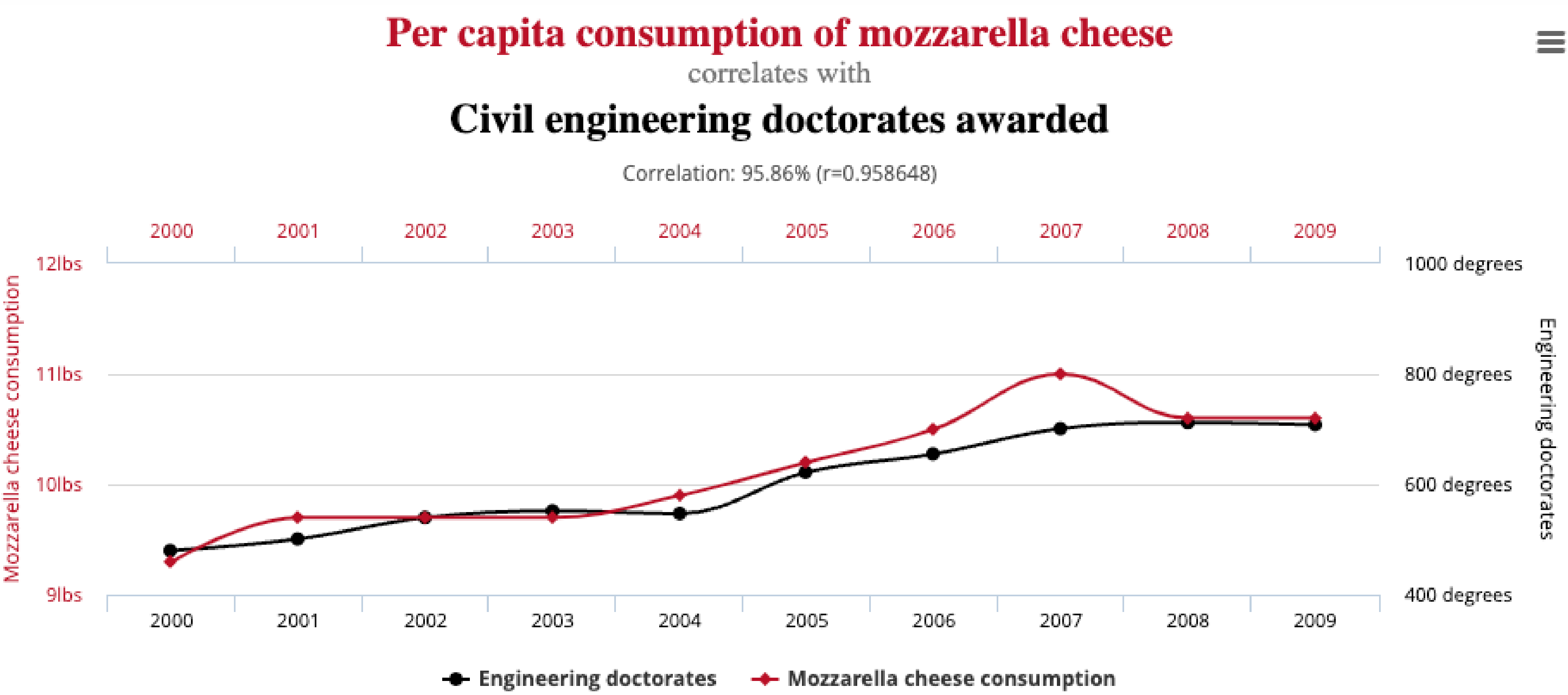
- Pearson correlation coefficient assumes normally distributed data

- Spearman's rank correlation coefficient: ρ
  - non-parametric alternative
  - also good for ordinal data

- Kendall's rank correlation coefficient: τ
  - smaller values
  - more accurate on small samples

Correlation does NOT imply causation!

- Spurious Correlations: http://tylervigen.com/



Data preparation

What if you have extreme values for a couple of runs during the experiment?

It depends on what is happening during those runs, check:

- if they make sense logically (e.g., in our EASE_2022 paper we had CPU usage going beyond 100%, and it helped us understanding that two treatments were using more than one core)
- if they all belong to the same treatments or subjects (they might indicate something interesting!)
- if other metrics behave peculiarly (e.g., cpu and duration of the run)

NOTE: there are different schools of thought about how to treat outliers in measurement-based experiments, such as:

- rerunning the runs
- keeping the data as it is
- removing the outliers
    - Example: https://ieeexplore.ieee.org/abstract/document/9830107

In your specific case, since the execution of a run does not cost a lot (thanks to automation), **it is strongly advised to redo the problematic runs**

What this module means to you?

- Now you know how to explore trends within your data
  - but you still cannot say anything about your null hypotheses
- You can have a “feeling” about
  - how disperse-correlated is your data
  - what is “standard” in your data
- You can quickly visualize interesting trends
  - box plots
  - scatterplots

Readings

Chapter 10