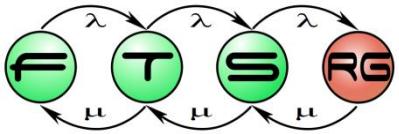
Visual Data Analysis

GUTA Gábor,

SALÁNKI Ágnes

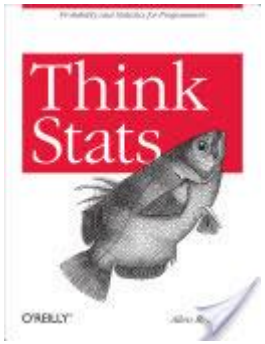
Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



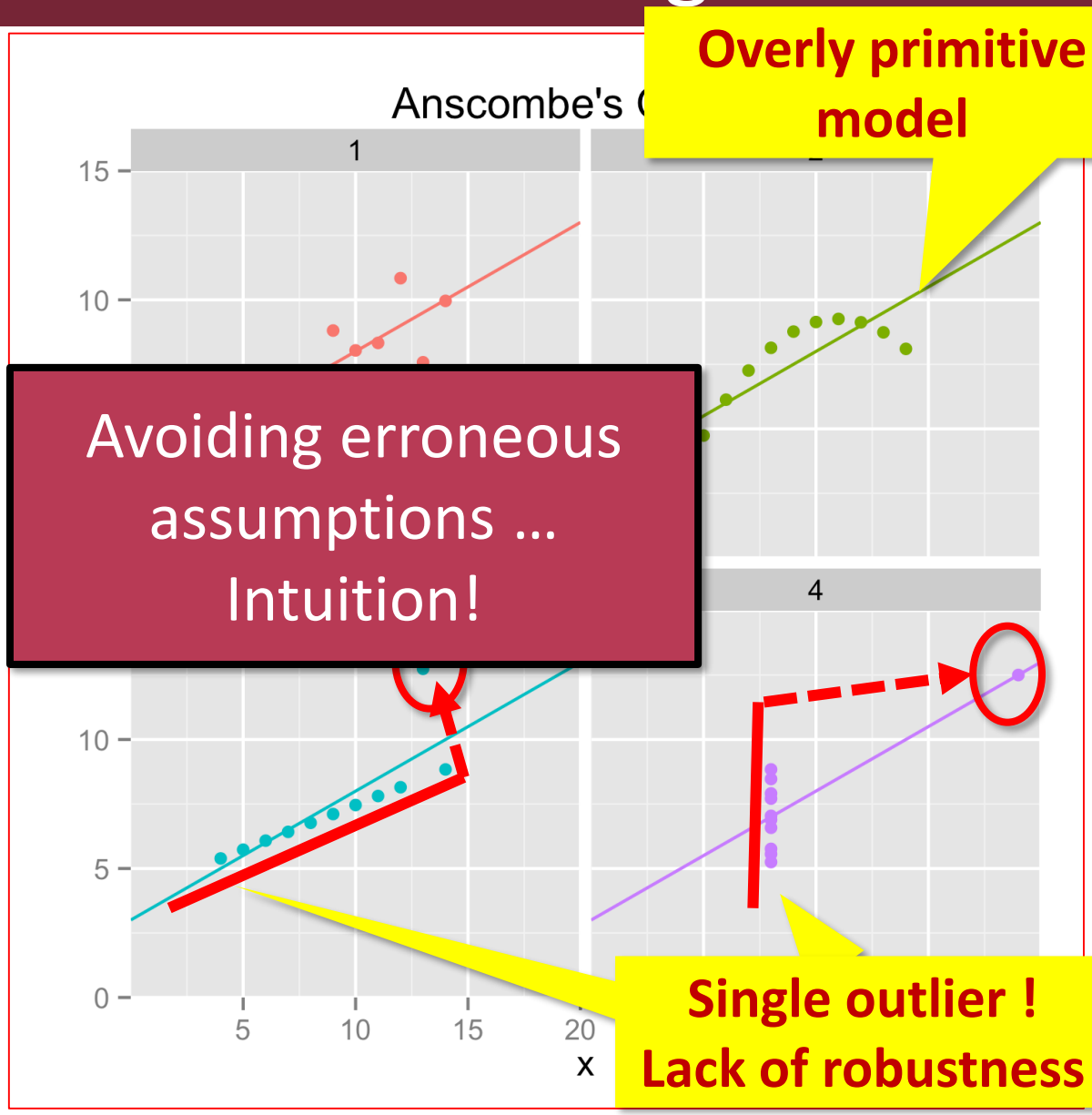
Repetition: Basic statistics

- Min: smallest value
- Max: gratest value
- Mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
- Variance: $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

- Think Stats: Probability and Statistics for Programmers



Reviewing the Calculations



- For all cases:

$Means$:

$M[x] = 9$

$M[y] \sim 7.5$

$Variance$:

$\sigma[x] = 11$

$\sigma[y] \sim 4.12$

$Correlation$:

$C(x, y) \sim 0.816$

$Regression$:

$y \sim 3 + 0.5x$

Content

Visualisation – Why?

Visualisation – What?

Visualisation – How?

Content

Visualisation – Why?

Visualisation – What?

Visualisation – How?

Visualisation in Everyday Life

Analog Display

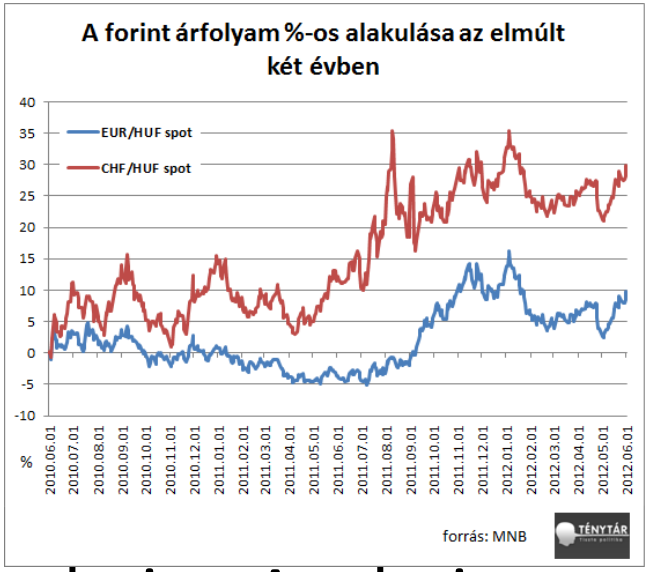
Digital Display

Analog + Coordinat

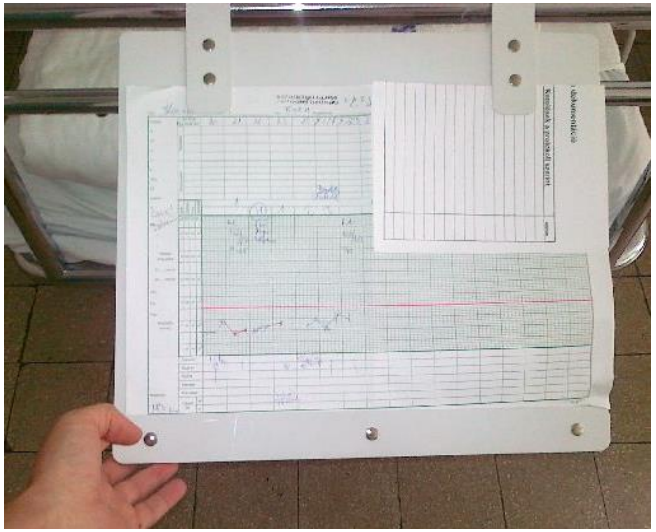
rid Display

Visualisation in Everyday Life

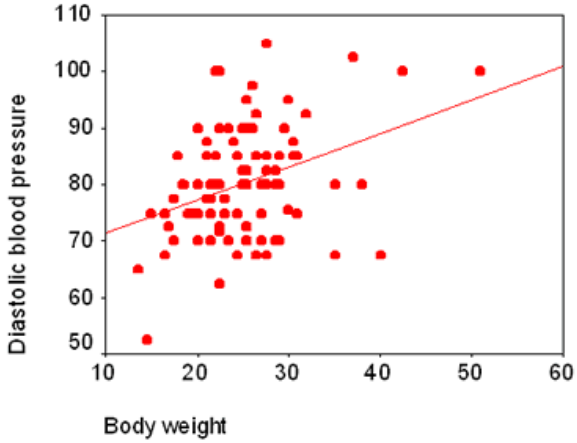
Trend Analysis and Forcast



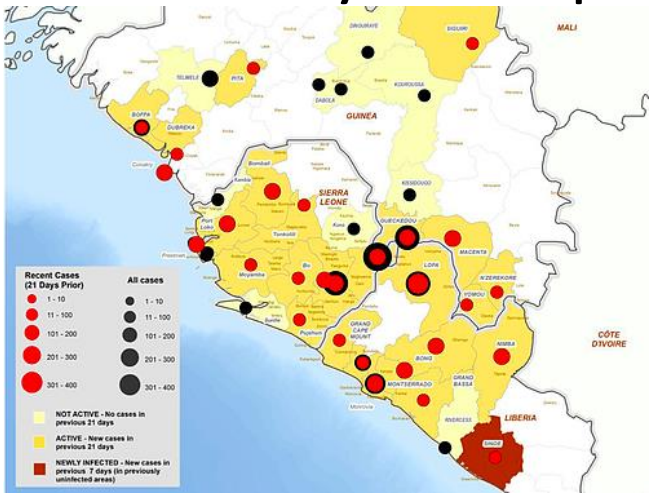
Time Series Analysis



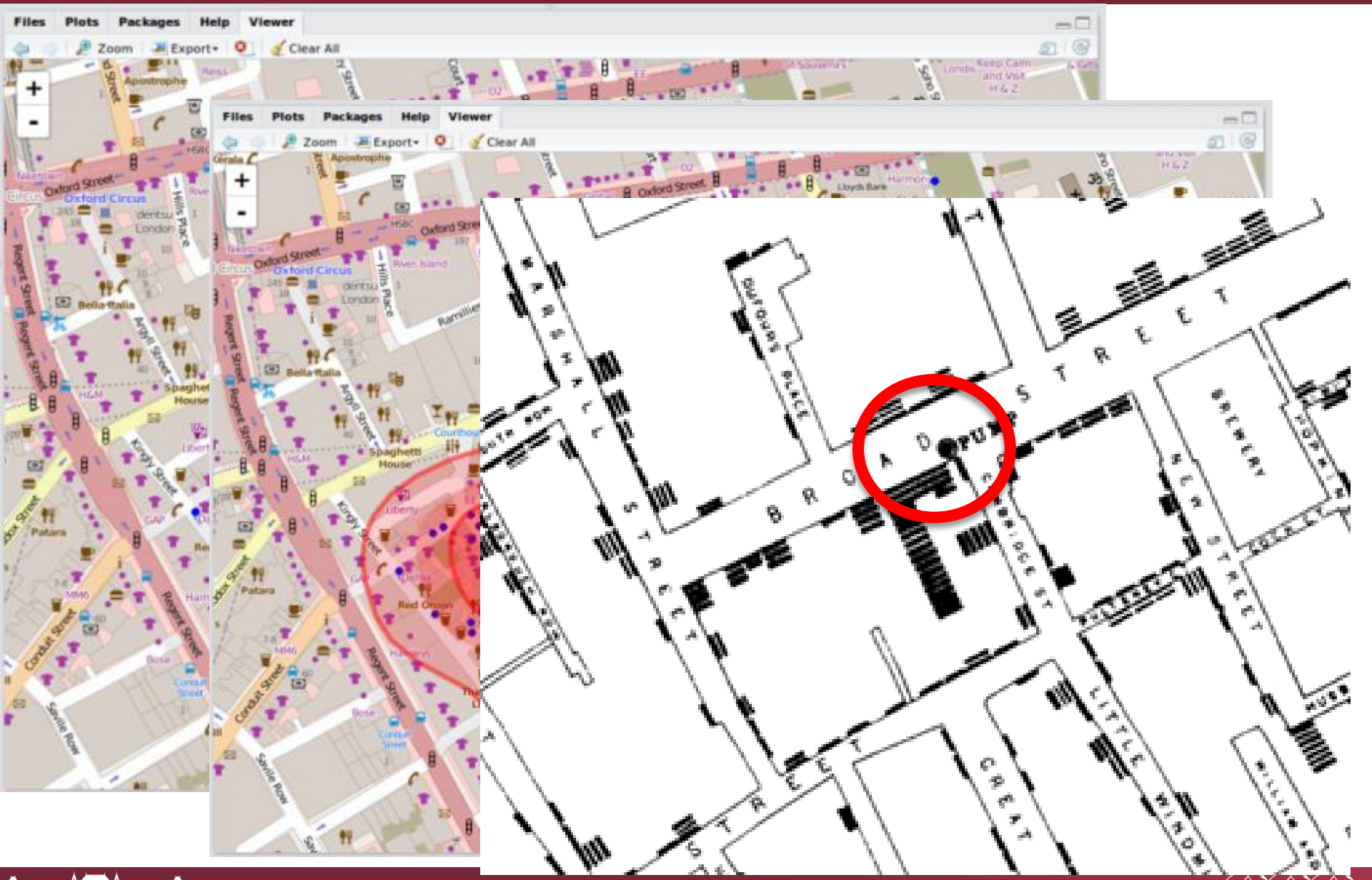
Correlation Analysis



Analysis of Spatial Data



Opening Up Relations



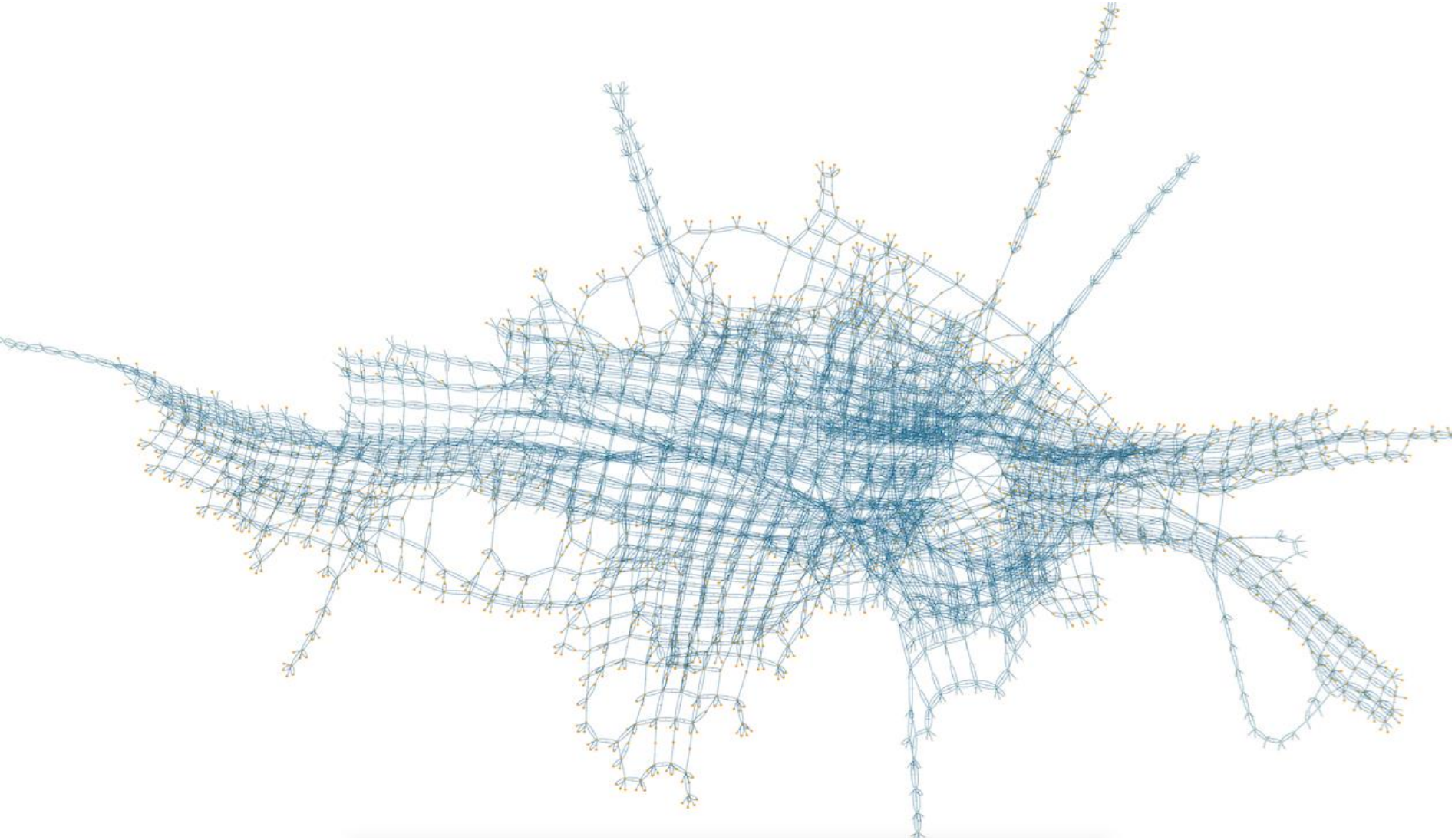
All Eyes on the Data!

„Massive Parallel” Processing

- 120.000.000 Sensors
- 10^{10} P[illegible]

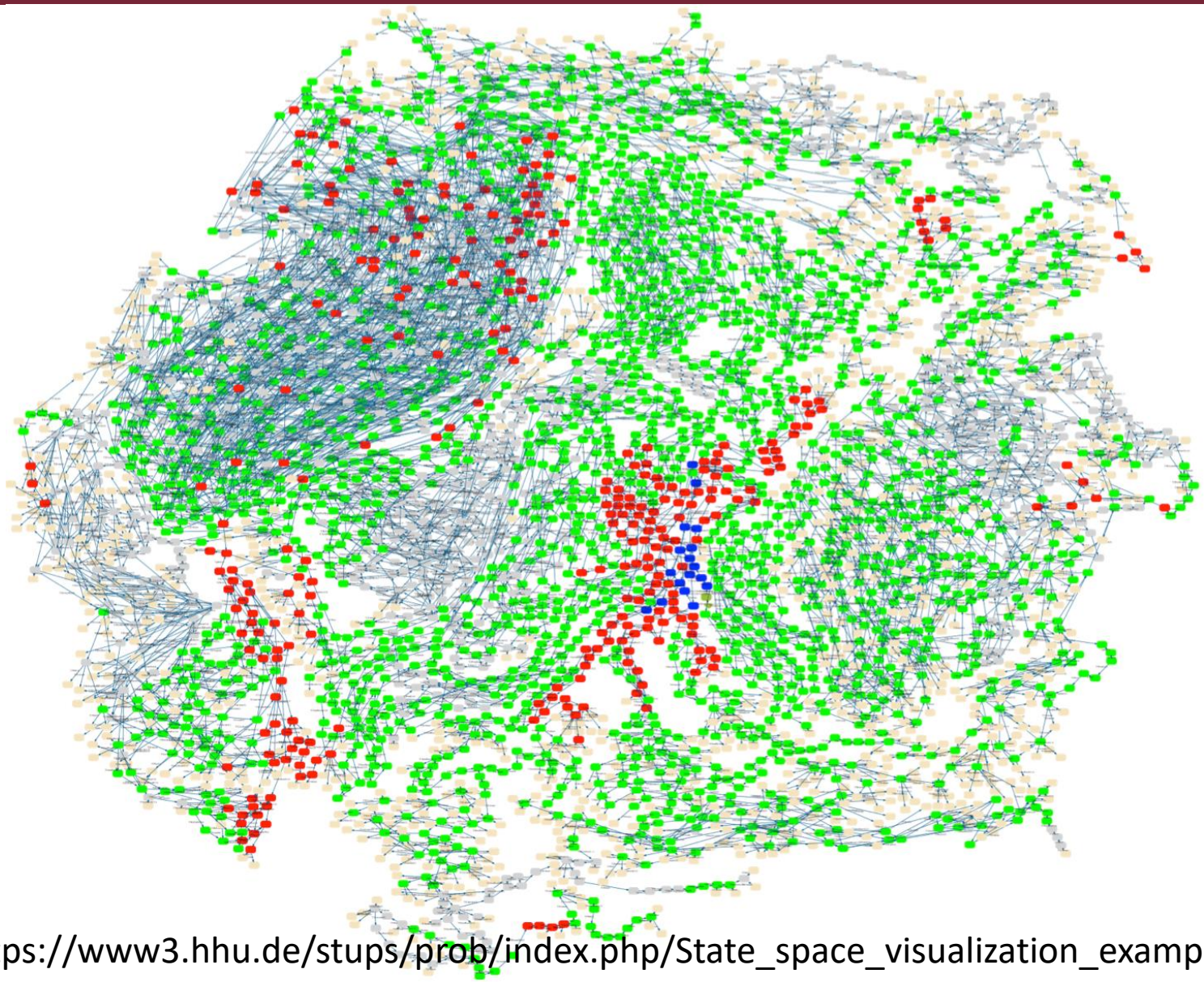
5. Visual selection and manipulation
4. Interpretation, correlation with other models, evaluation

Example: Visualisation of State Spaces



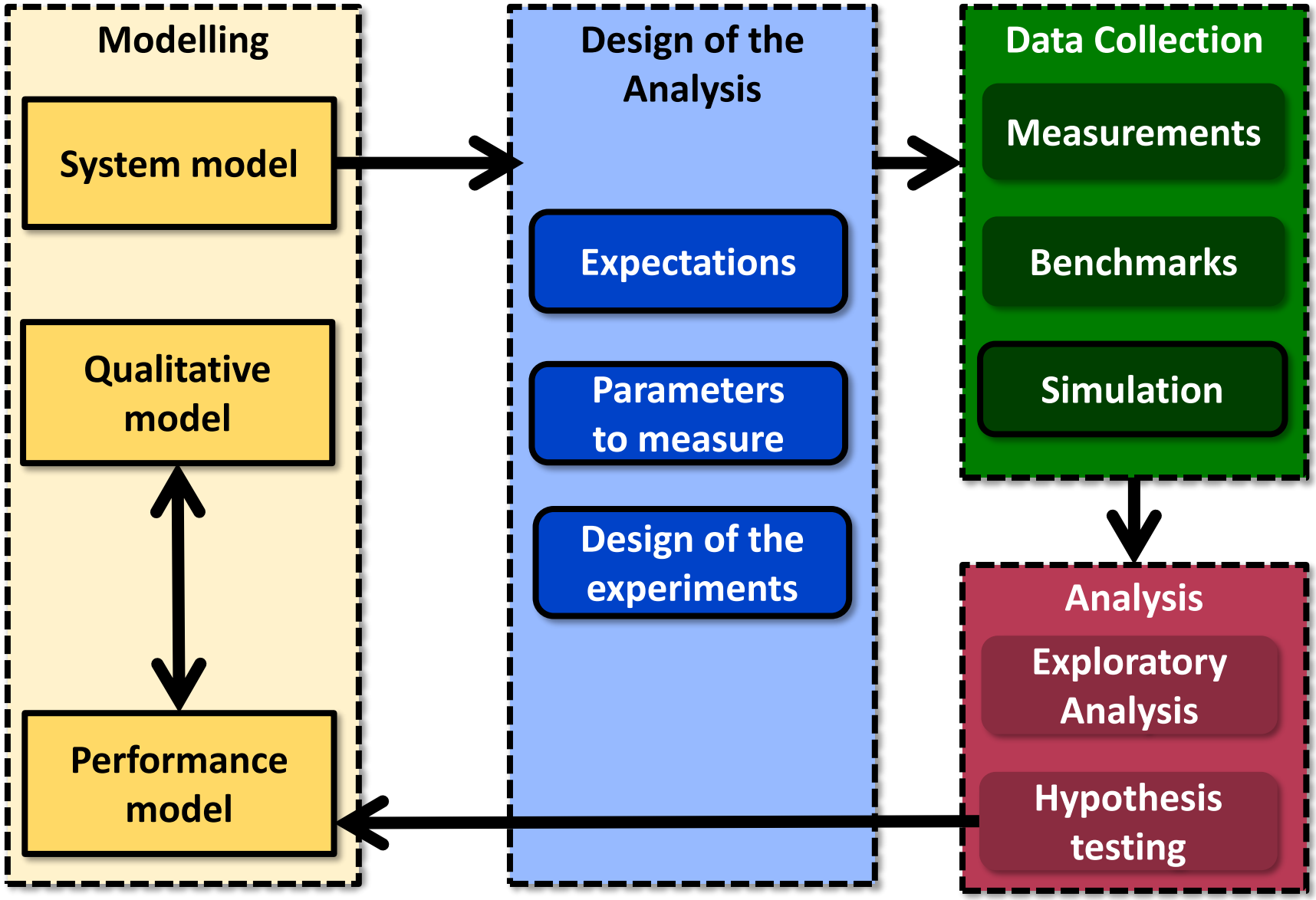
https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples

Example: State Space of the CAN Bus



https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples

Example: System Model → Performance Model



Content

Visualisation – Why?

Visualisation – What?

Visualisation – How?

Reminder: Tabular Representation

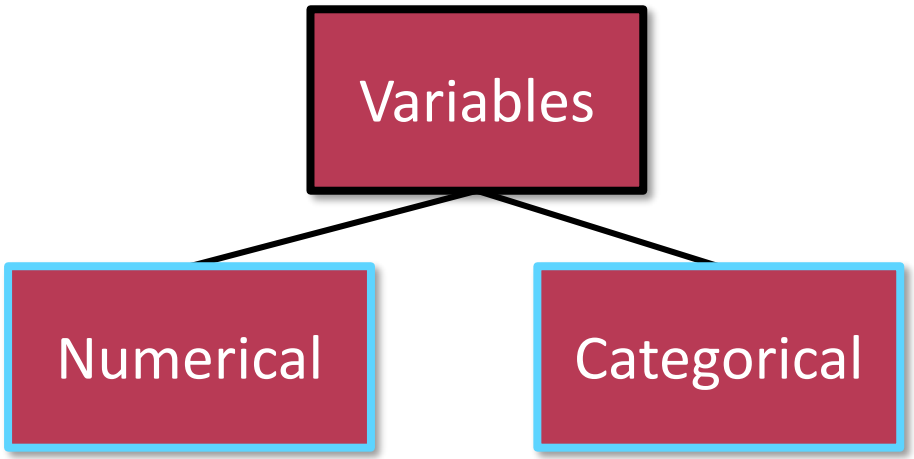
- **Rows of the table** = Model elements
- **Columns of the table** = Properties

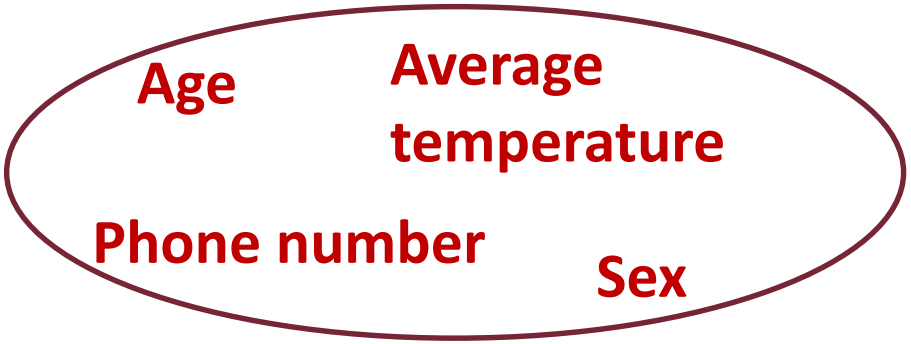
Name	Type	Size (kB)	Last modified
Documents	directory		2016.02.02
Contracts.pdf	file	569	2015.11.09
Pictures	directory		2016.02.02
Logo.png	file	92	2015.03.06
Groundplot.jpg	file	1226	2016.02.02

- Data analysis languages (e.g. R, Python): **dataframe**
  - One row: one measurement/observation
  - Columns have their own **Types**

Numerical and Categorical Variables

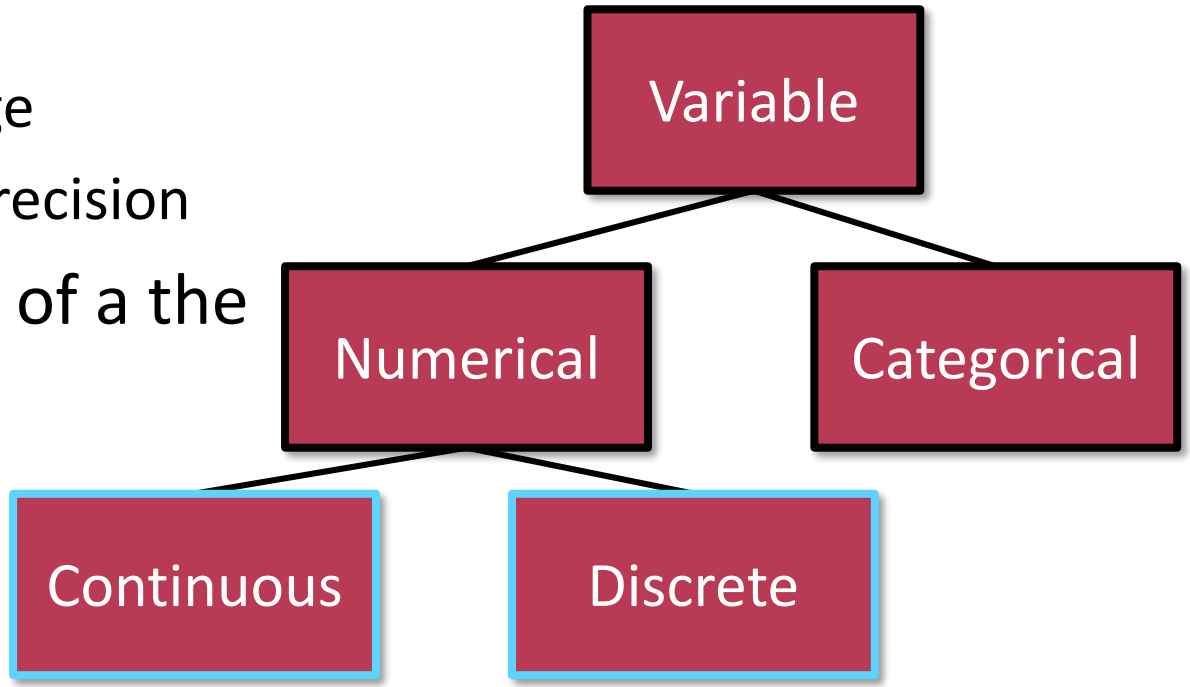
- Numerical
  - Arithmetic operations are interpreted meaningful (average, sum, inc, dec, …)
- Categorical
  - No operation between the values





Numerical Variables

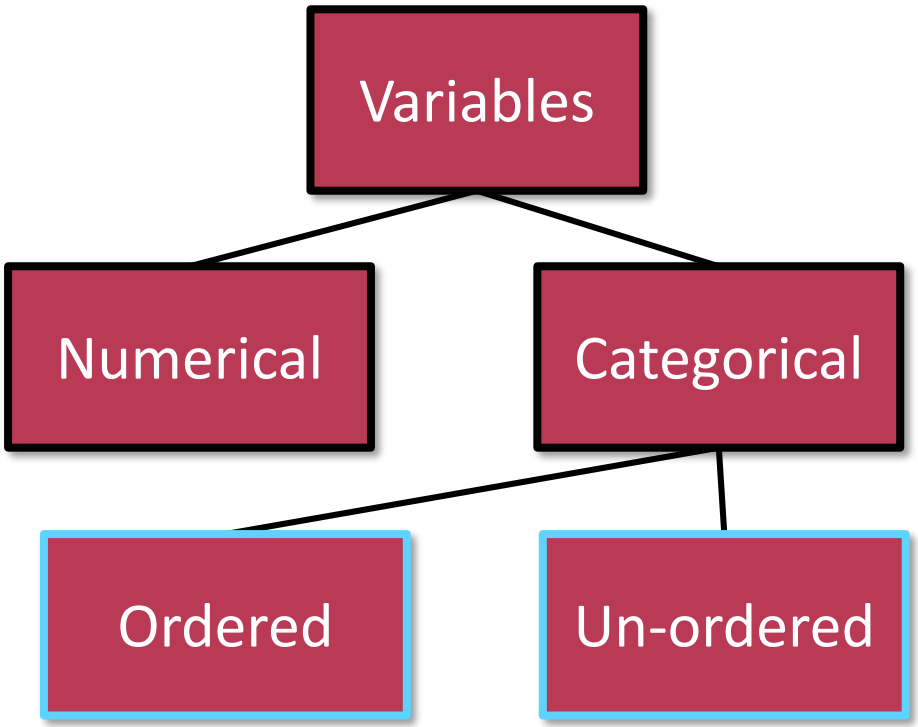
- Continuous
  - Measured
    - in a specific range
    - with a specific precision
  - e.g. temperature of a the server room



- Integer
  - Counted – finite number of values in a specific range
  - e.g. number of disks

Categorical variables

- Ordered (ordinal)
  - Excelent, good, fair, poor
  - Value range is ordered
  - Fully ordered?
- Un-ordered (nominal)
  - Types



10. Would you urge others to attend these classes regularly?

- ○ I would convince everybody to come
- ○ I would urge them to come
- ○ Maybe I would urge them to come
- ○ I would rather discourage them from coming
- ○ I would definitely discourage them from coming
- ○ I do not want to answer

Content

Visualisation – Why?

Visualisation – What?

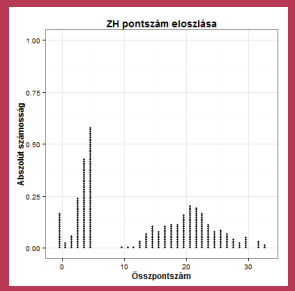
Visualisation – How?

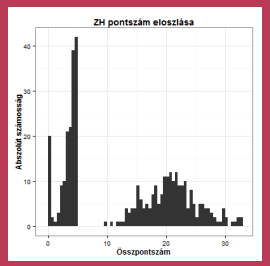
Single variable

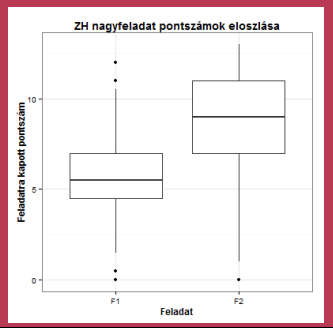
Variables

Numerical

Test results: [28, 28, 30, …]

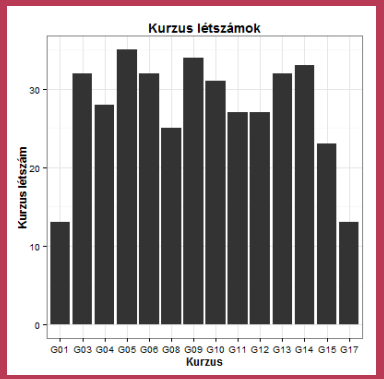






Categorical

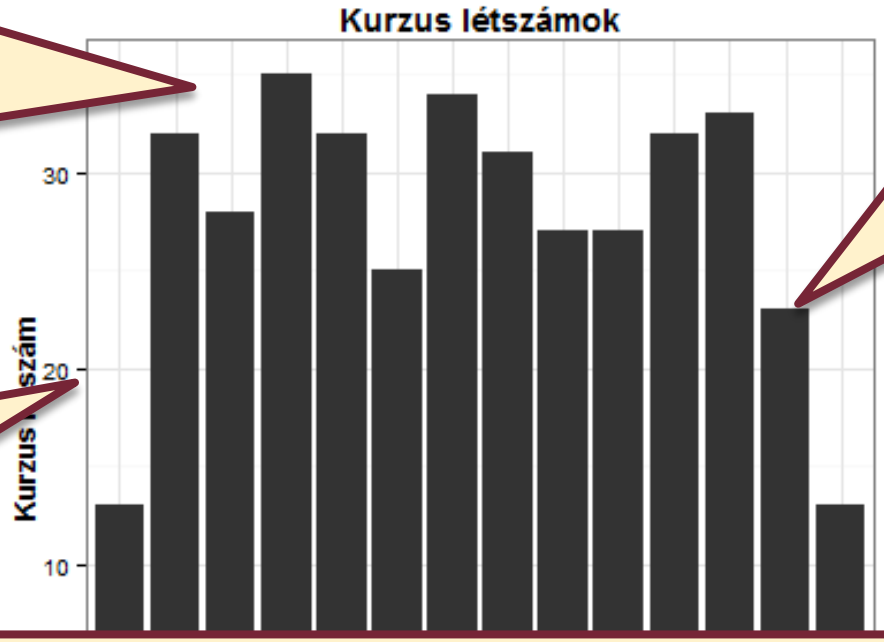
Training groups: [G01, G02, G03, …]



Column Charts / Bar Charts

- Input variable: Course codes
- Question: How many students have subscribed?

Are there popular time slots / trainers?



Height of bar: Frequency of the given value

Absolute Frequency!

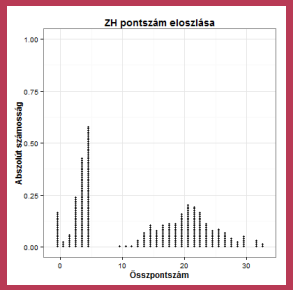
Design decision: Splitting of the value range
e.g.: Tuesday-Thursday-Friday?

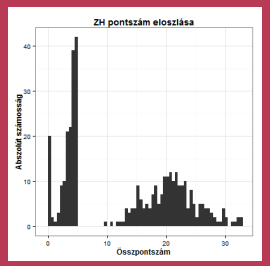
Single variable

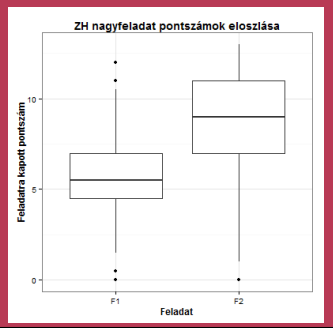
Variables

Numerical

Test results: [28, 28, 30, …]

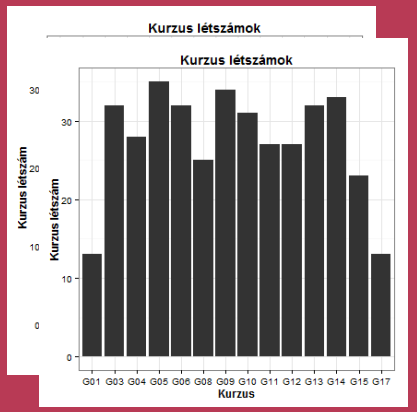






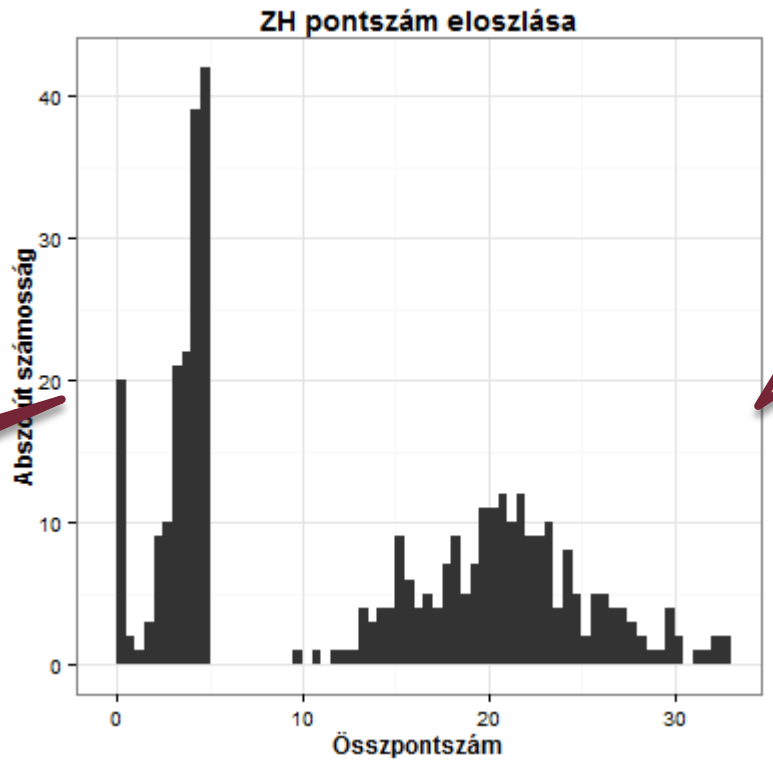
Categorical

Training groups: [G01, G02, G03, …]



Histogram

- Input variable: Test results
- Question: What results were born?



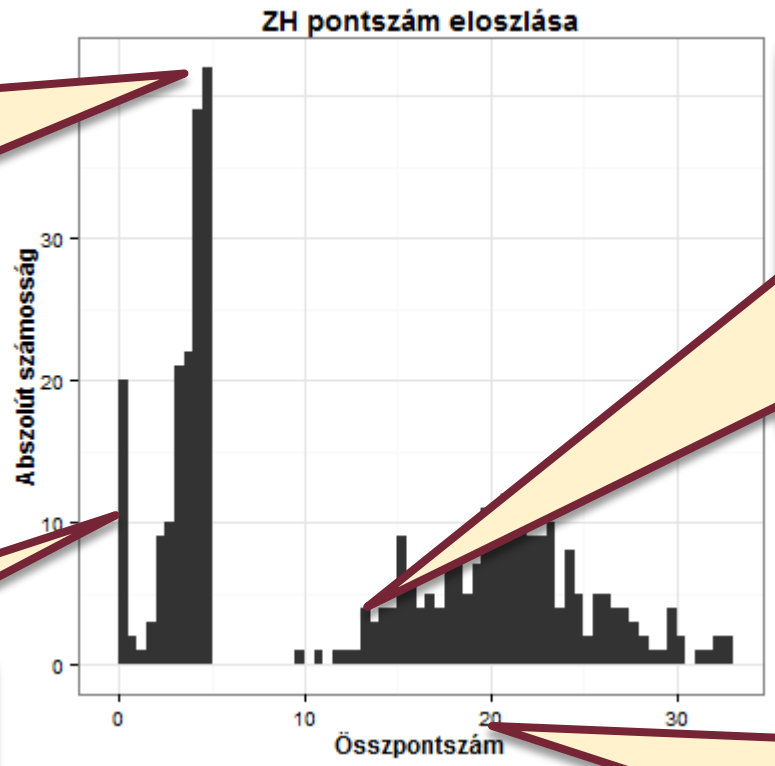
Height of bar : Frequency of the given **interval** of values

Absolute Frequency!

Design decision : Choosing length of the intervals
e.g.: 1-point-resolution vs. 0,5-point-resolution?

Histogram

- Input variable: Test results
- Question: What results were born?



Many have the entry test almost made.

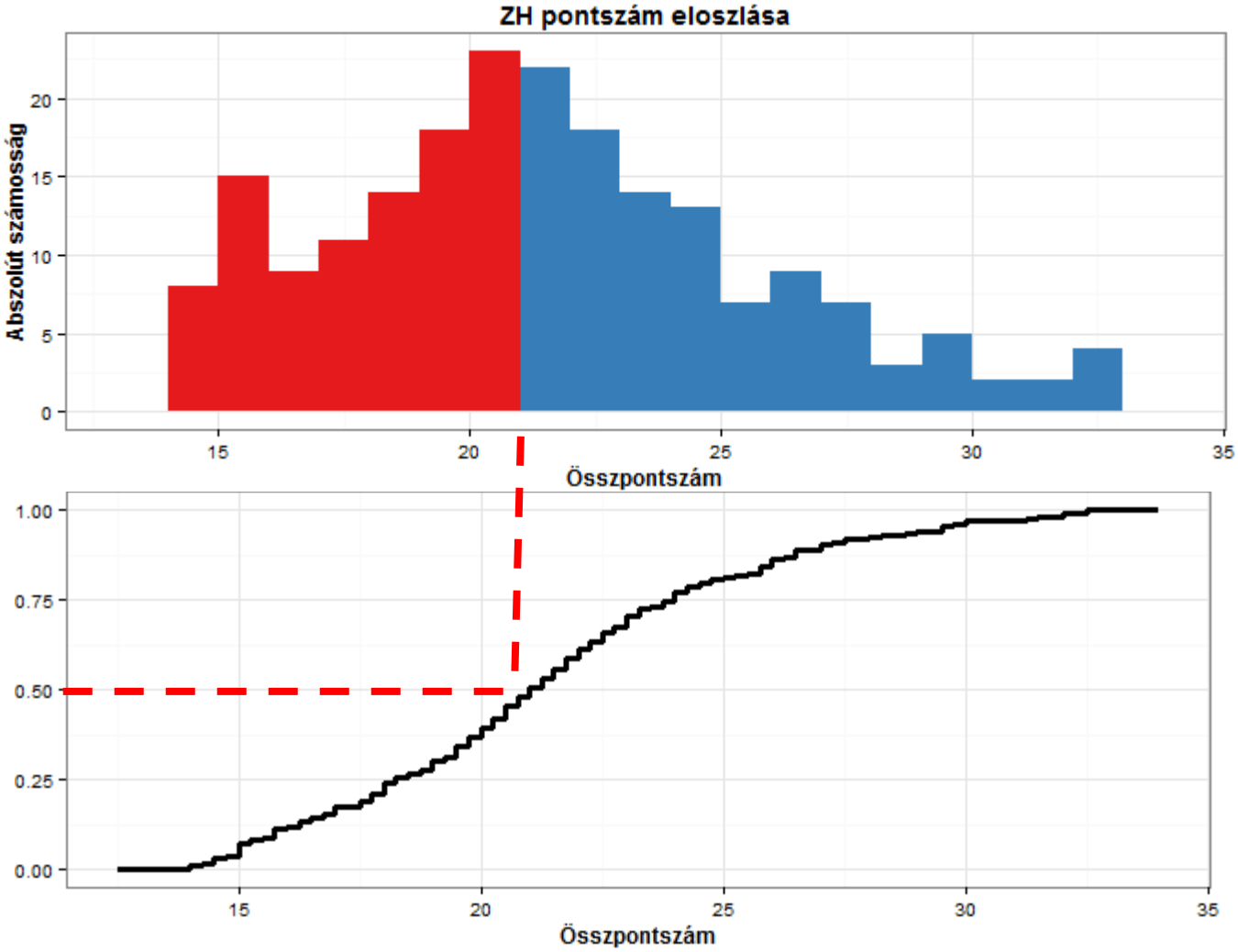
The most of those who made the entry test, made the whole test also.

Those not appeared at all.

The average/ median was at 20 points.

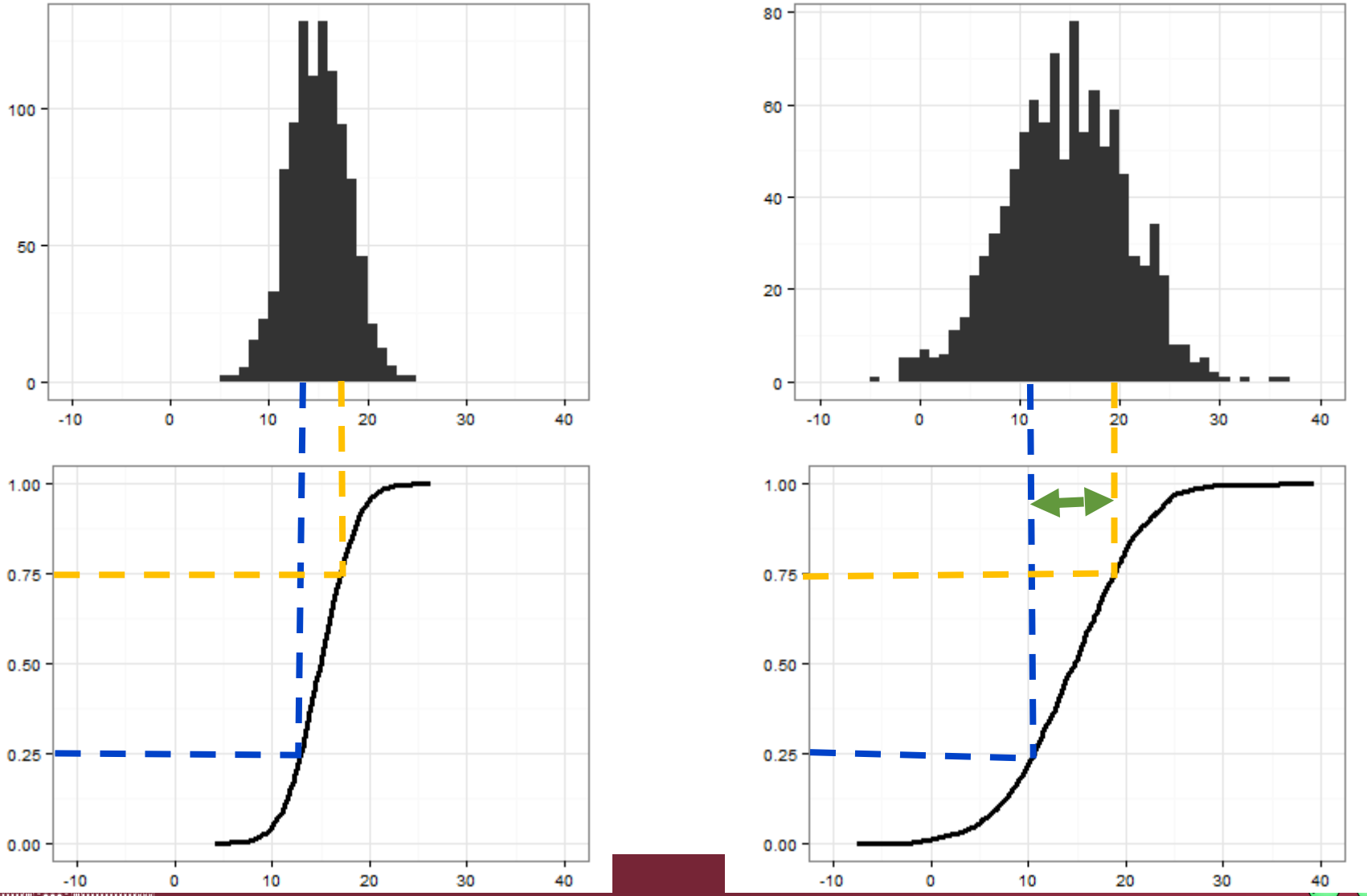
Simple Statistical Description

- Where is „the middle” of the values?



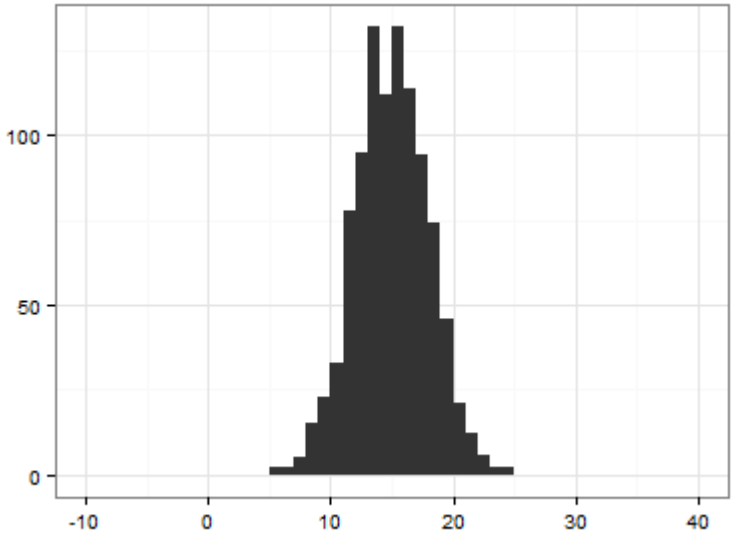
Simple Statistical Description

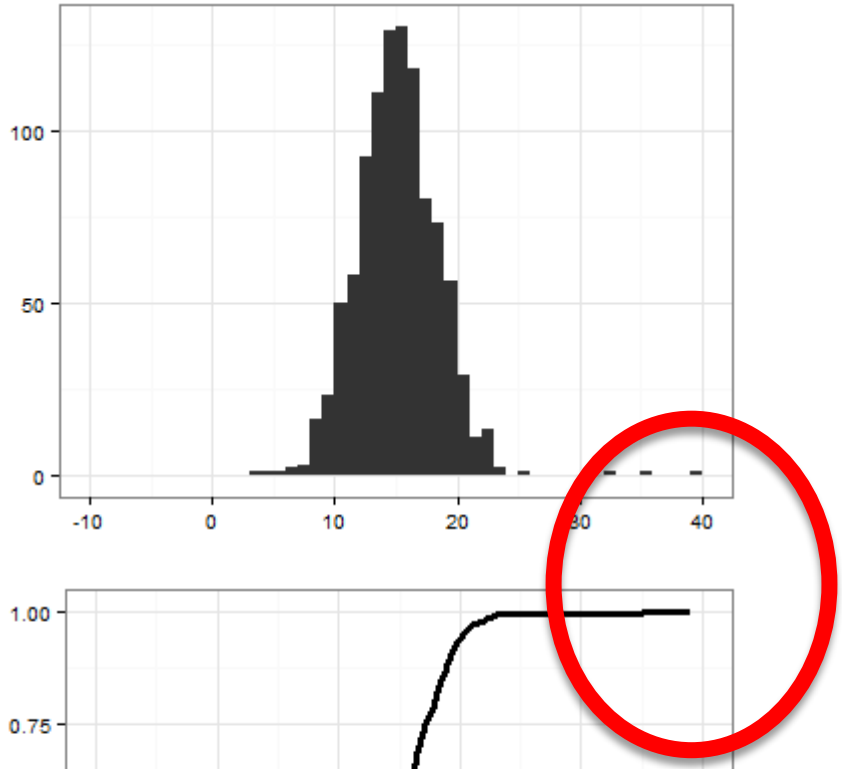
- How far are the values „scattered”?

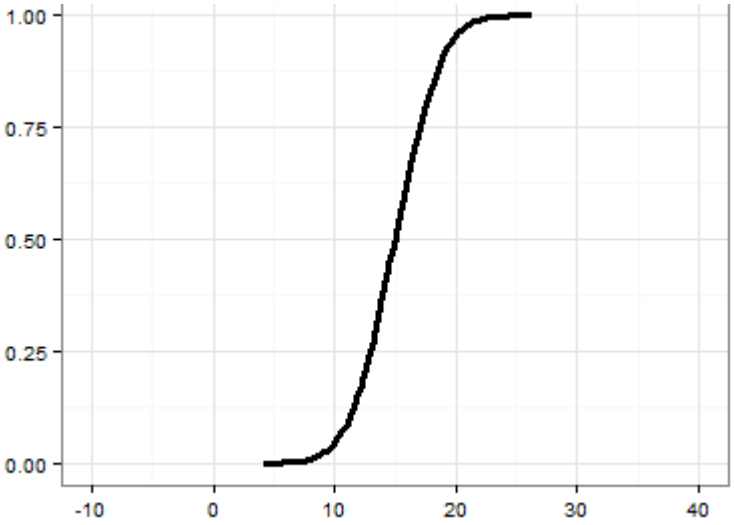


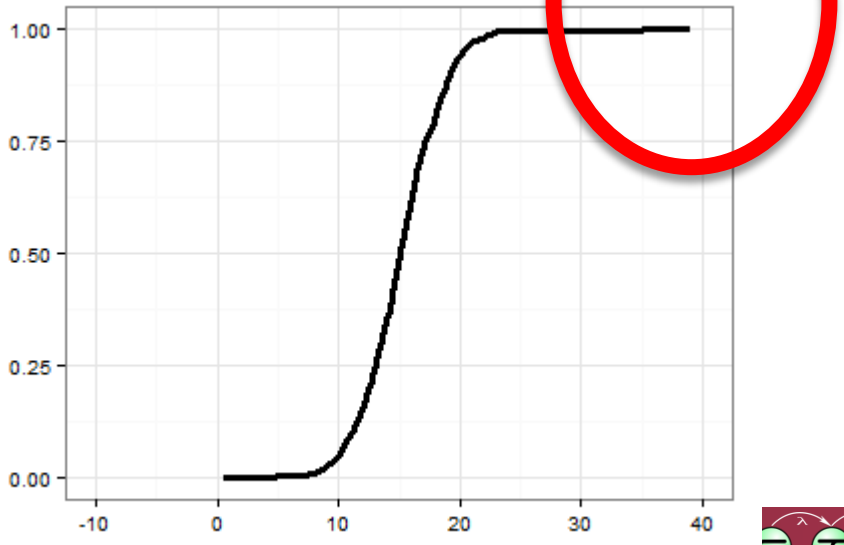
Simple Statistical Description

- Are there outliers?



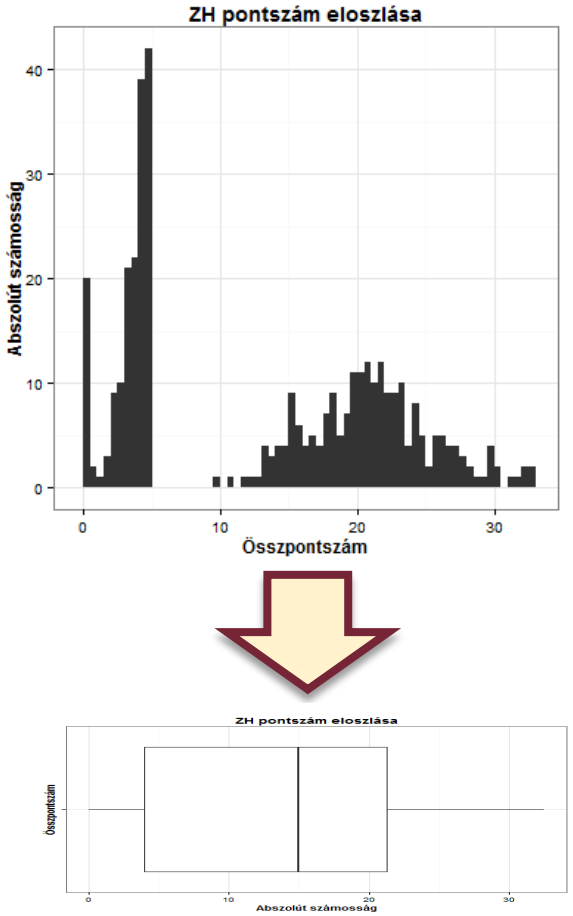






Box plots

- Input variable: Test results
- Question: What results were born **approximately**?



An art of abstraction:
just take some intervals,
the exact values are not that important

Description of (Continuous) Observations

- Description of the „middle”
  1. Average – arithmetic mean
  2. **Median** – the element separating the upper half from the lower half (ordered data sets!)
  3. Mode – the most frequent element
  - Example: {3, 4, 4, 5, 5, 6, 10, 20}
    - Mean: ~ 7.125
    - Median: 5
    - Mode: 4 and 5 (often as 4.5)
- Description of the „spread”
  - Percentiles (frequency for categorical types)

Describing (cont.) Observations

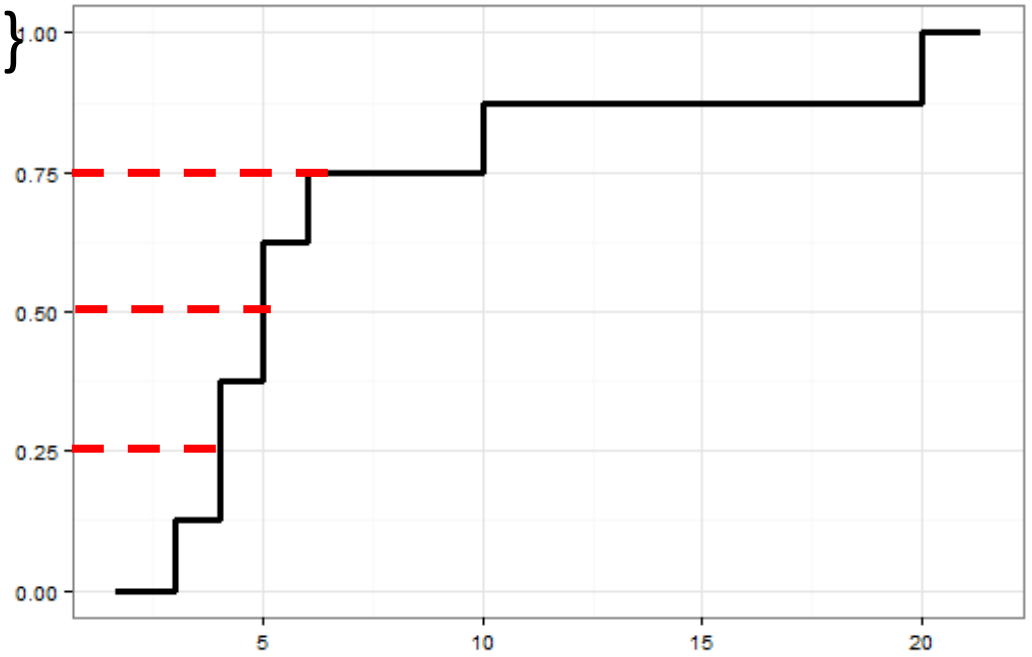
If the elements of a data set are ordered, the middle element is **the median** of the data set. In the case if there is no middle element (an even number of elements), **the median** is the average of the two middle elements.

The mode is the most frequent element (the most frequent elements) of the data set. If there is no unique *most frequent element*, the data set has multiple Modes.

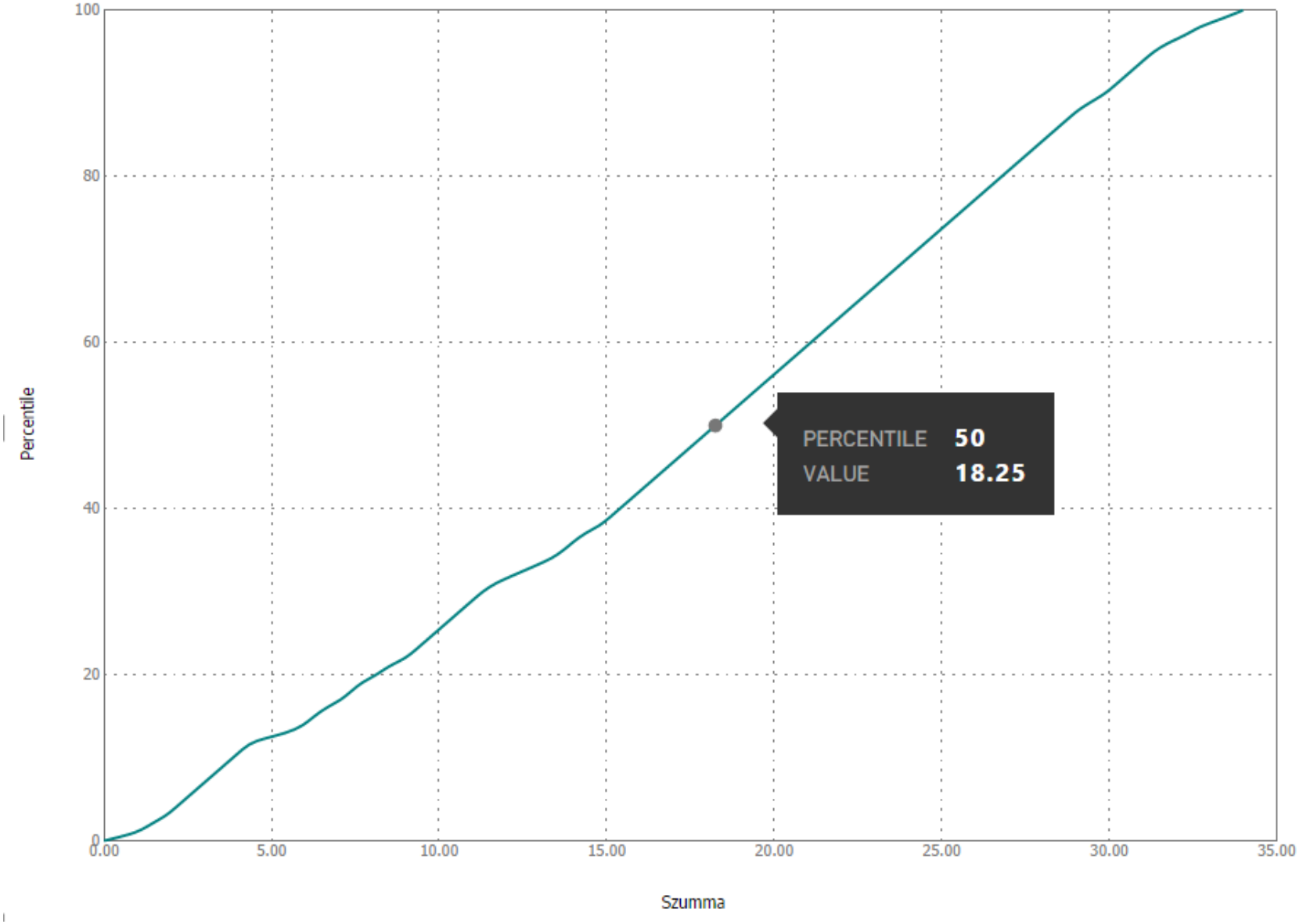
Percentiles

- Percentile
  - $n\%$ of the values are weaker than the n^{th} **percentile**
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - 50. percentile: 5
    - 25. percentile: 4
    - 75. percentile: 10
- Quartiles
  - Q1: 25. percentile
  - Q3: 75. percentile
  - **Q2: Median**

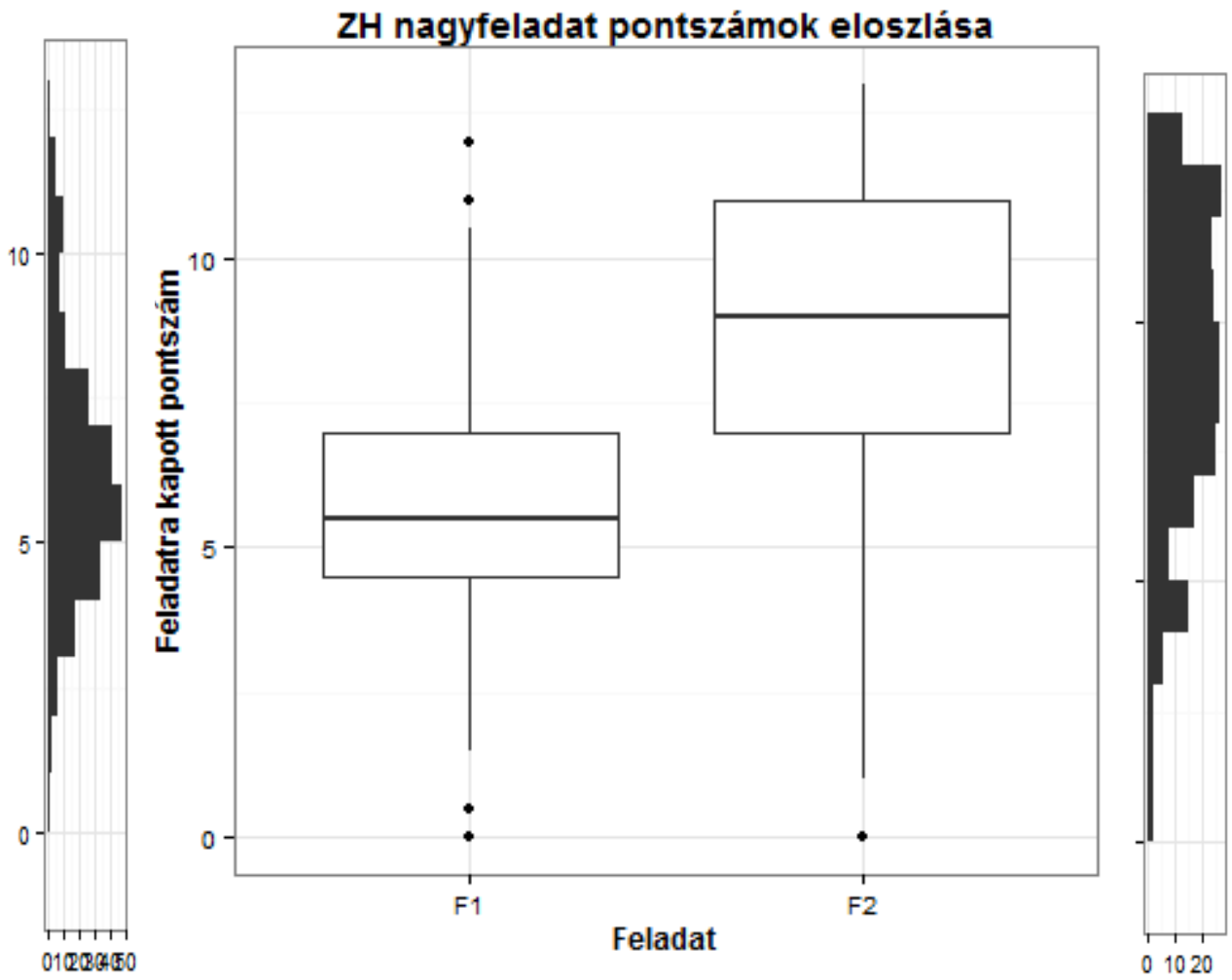
Frequency: $n\%$ of the values lie in the given categorie(s)



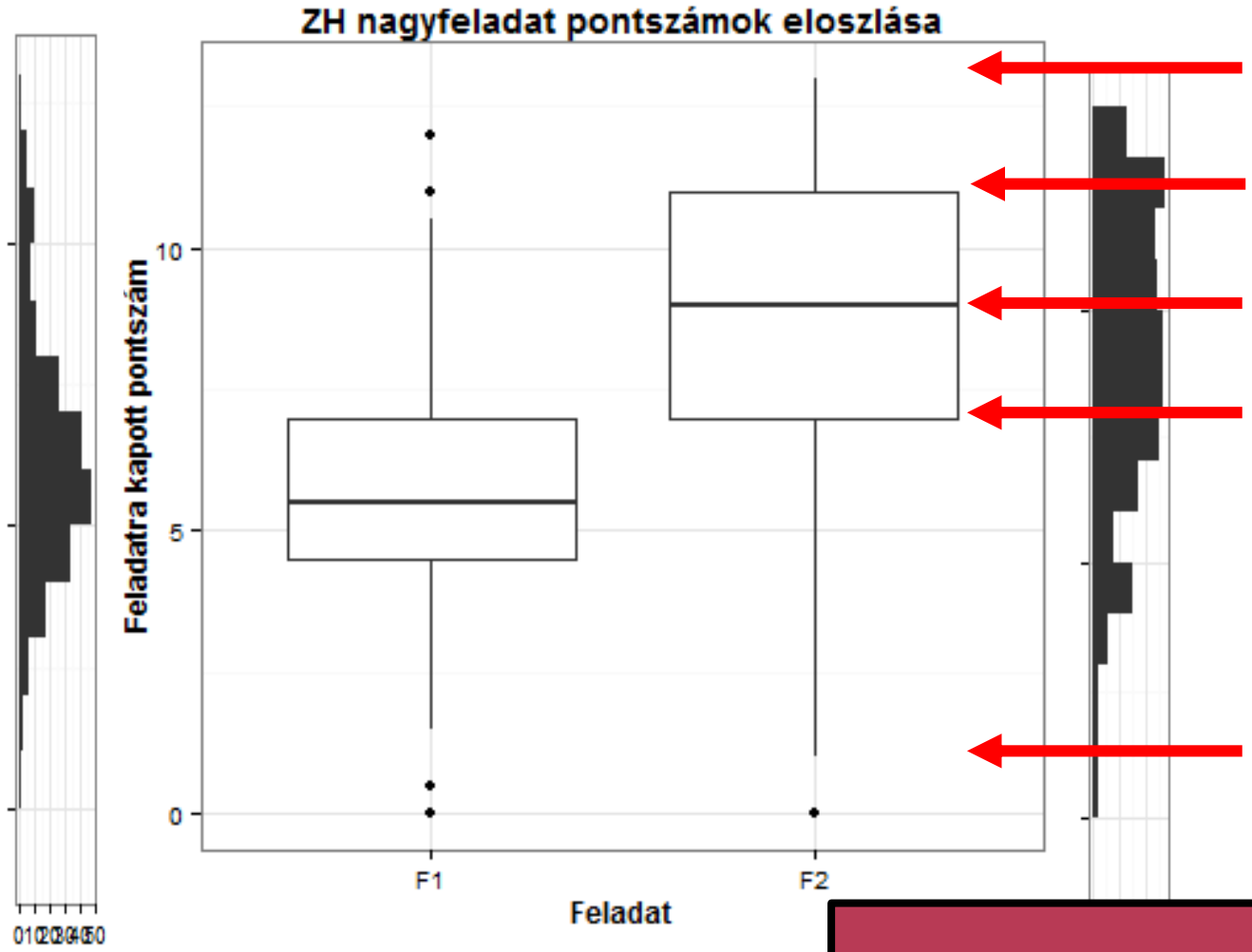
Example: Representation of Percentiles



Box and whisker plot



Box and whisker plot



Limit of the upper outliers

Q3

Median

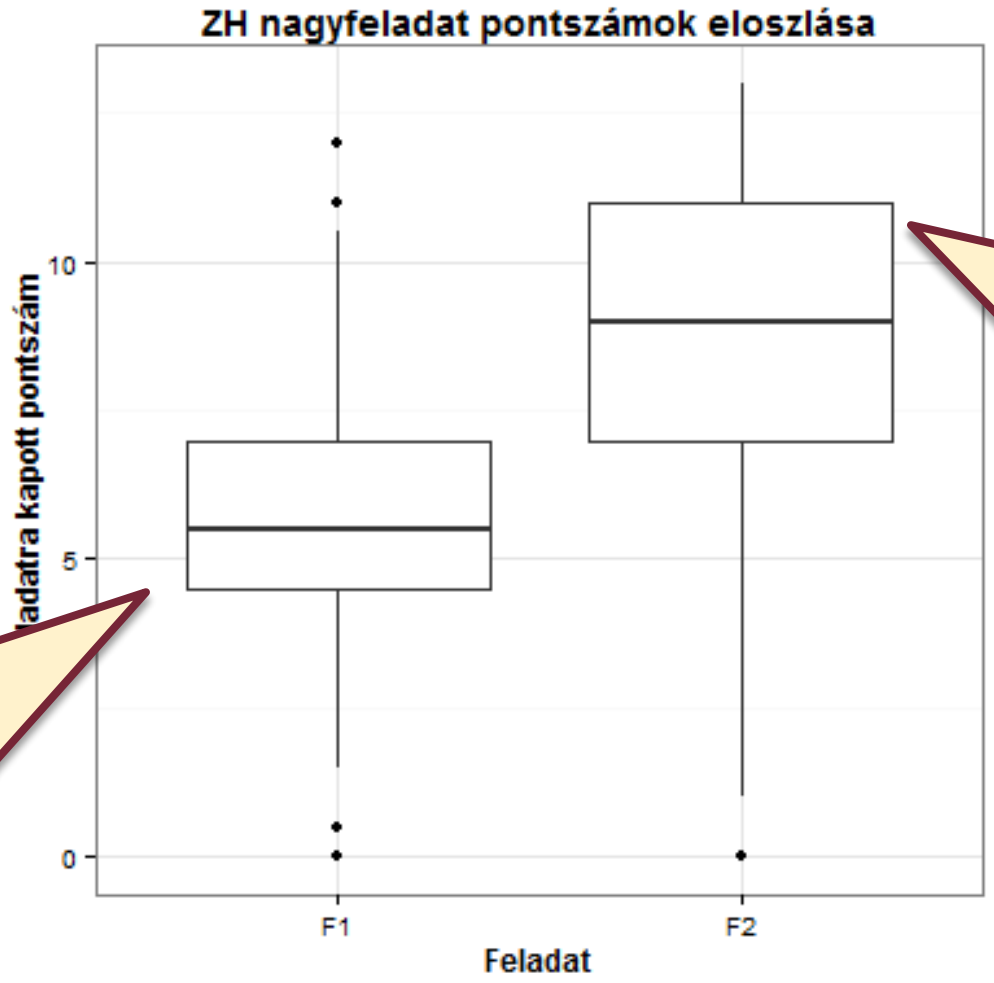
Q1

Limit of the lower outliers

How to do it in Excel?

(www.youtube.com/watch?v=ucWmfmXb1kk)

Box and whisker plot

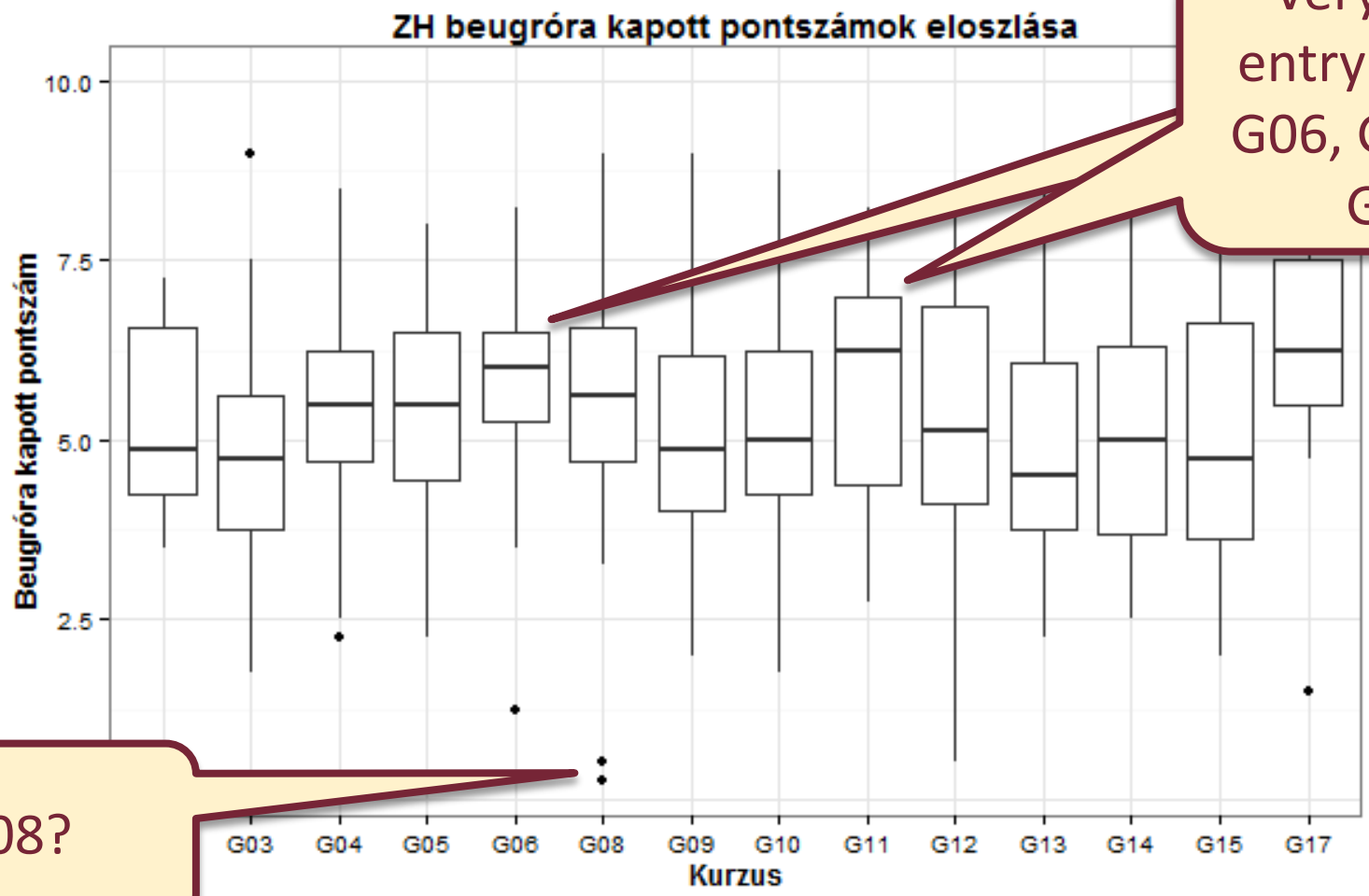


Task2 has produced (in average) better results than Task1.

50% of the results of Task1 lie between 4.5 and 7.5.

Box and whisker plot

- How were the results per training groups?

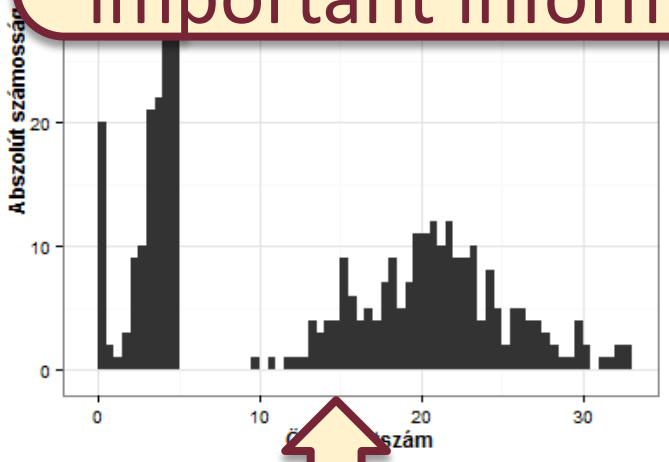


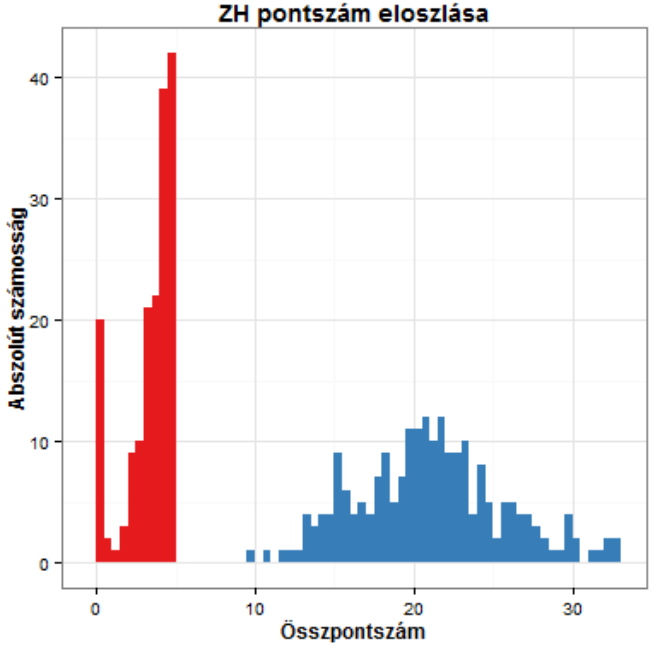
Very good entry tests in G06, G11 and G17

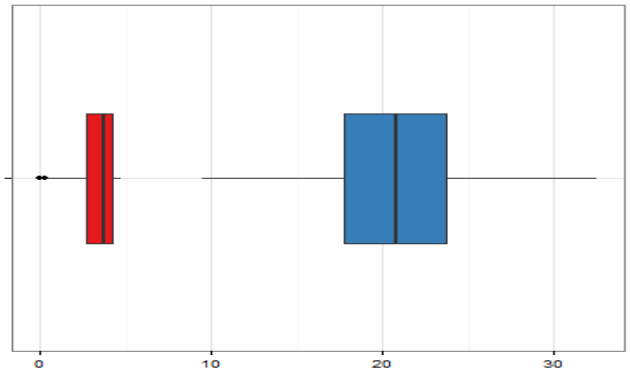
G08?

Box and whisker plot

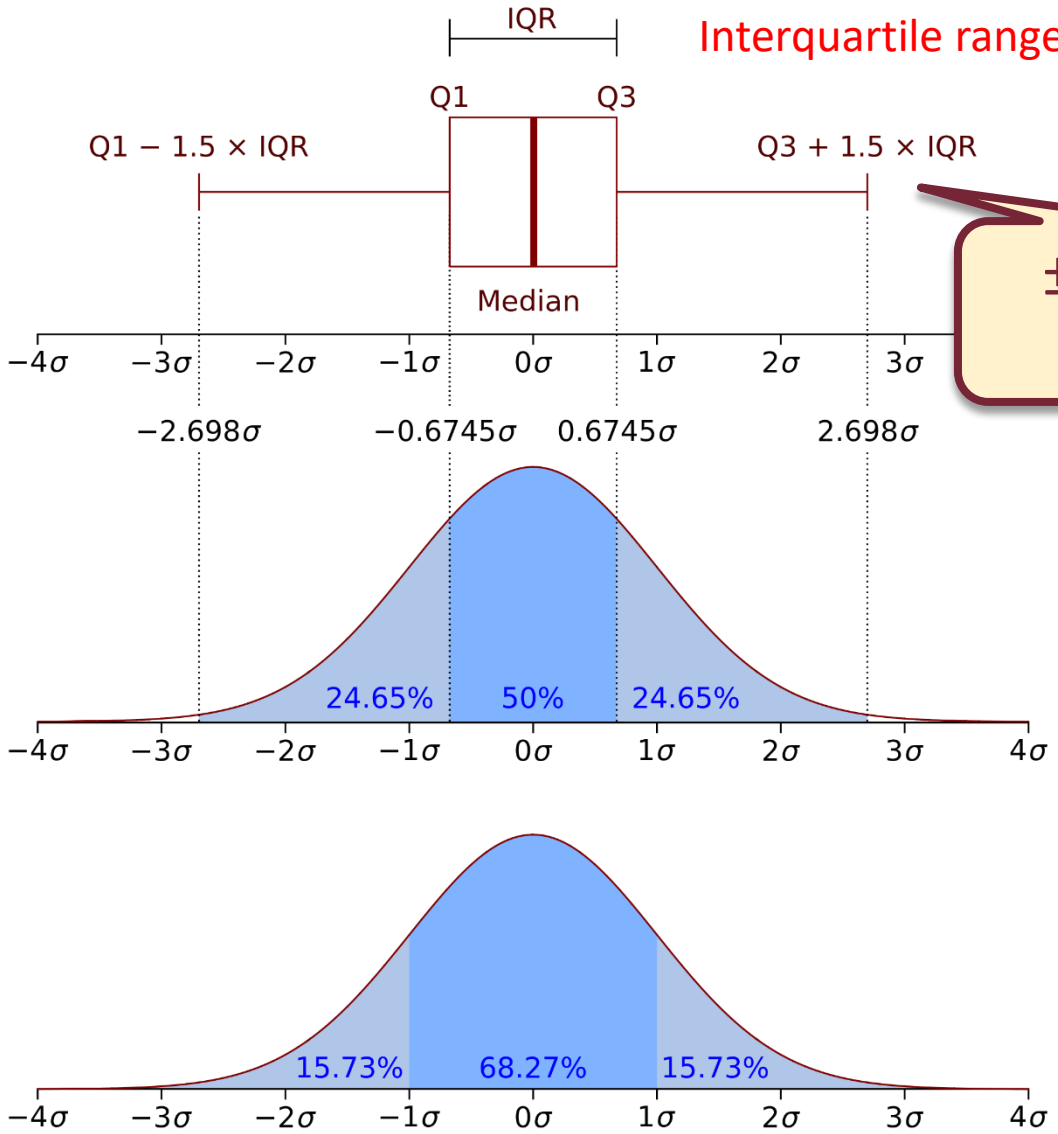
Abstraction: With box plot we can miss important information.





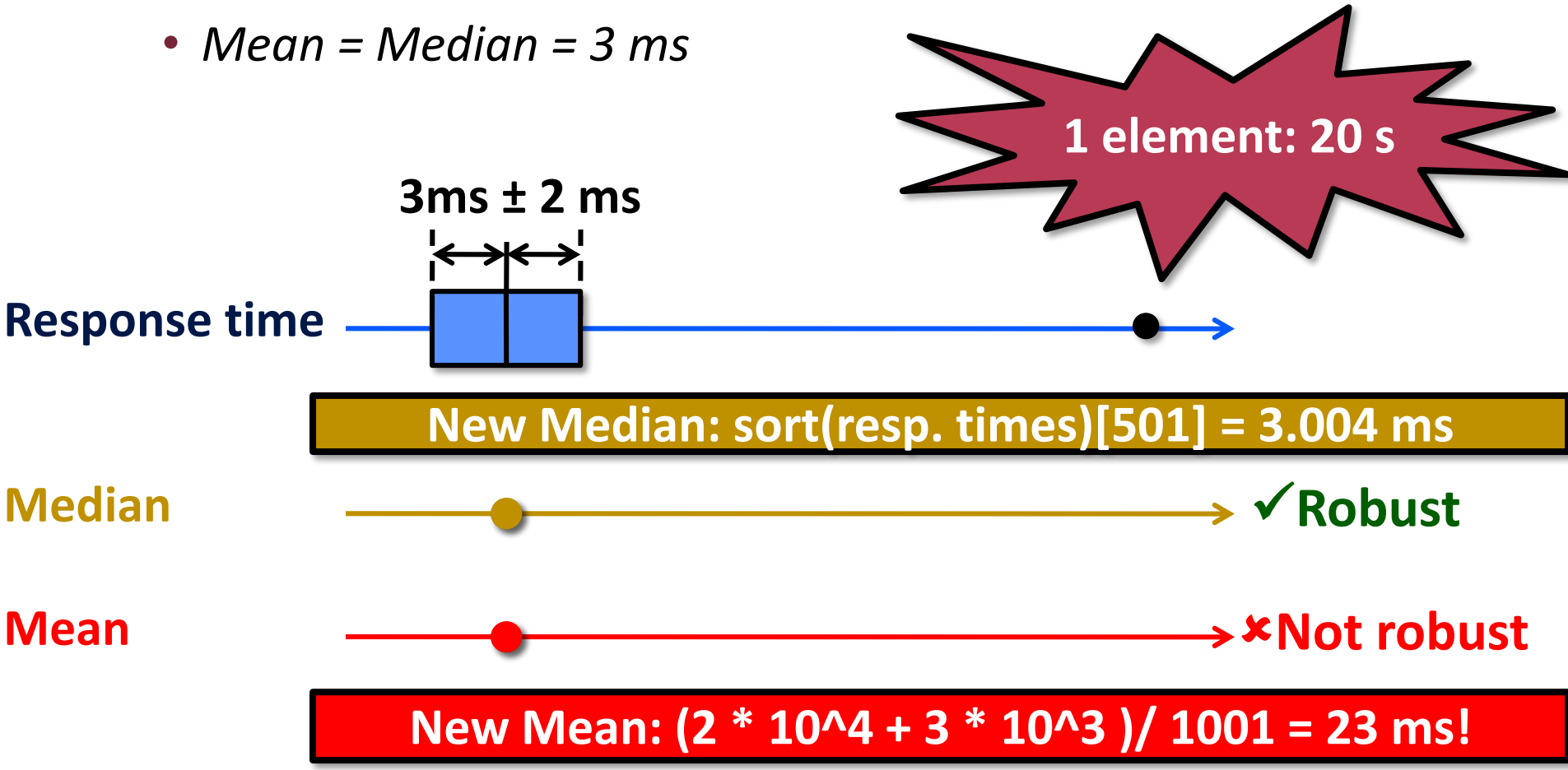


Boxplot (Box and whisker plot)

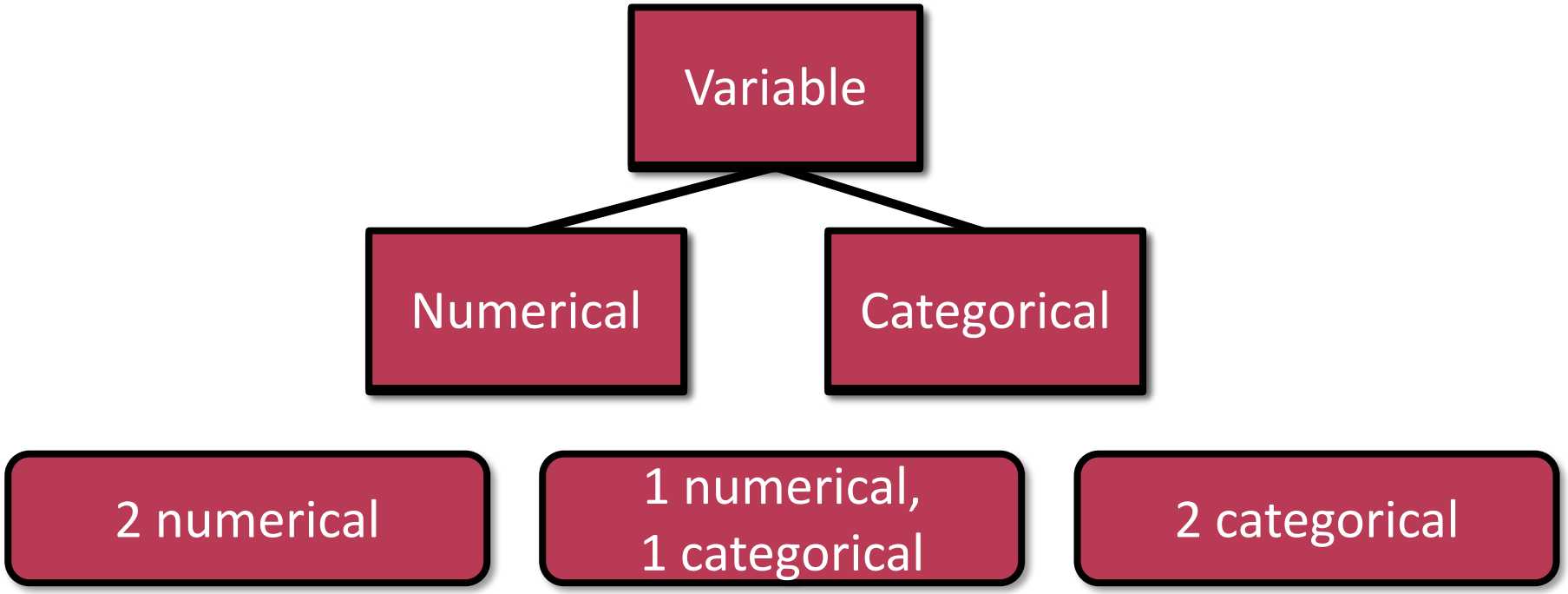


Median instead of Mean – Why?

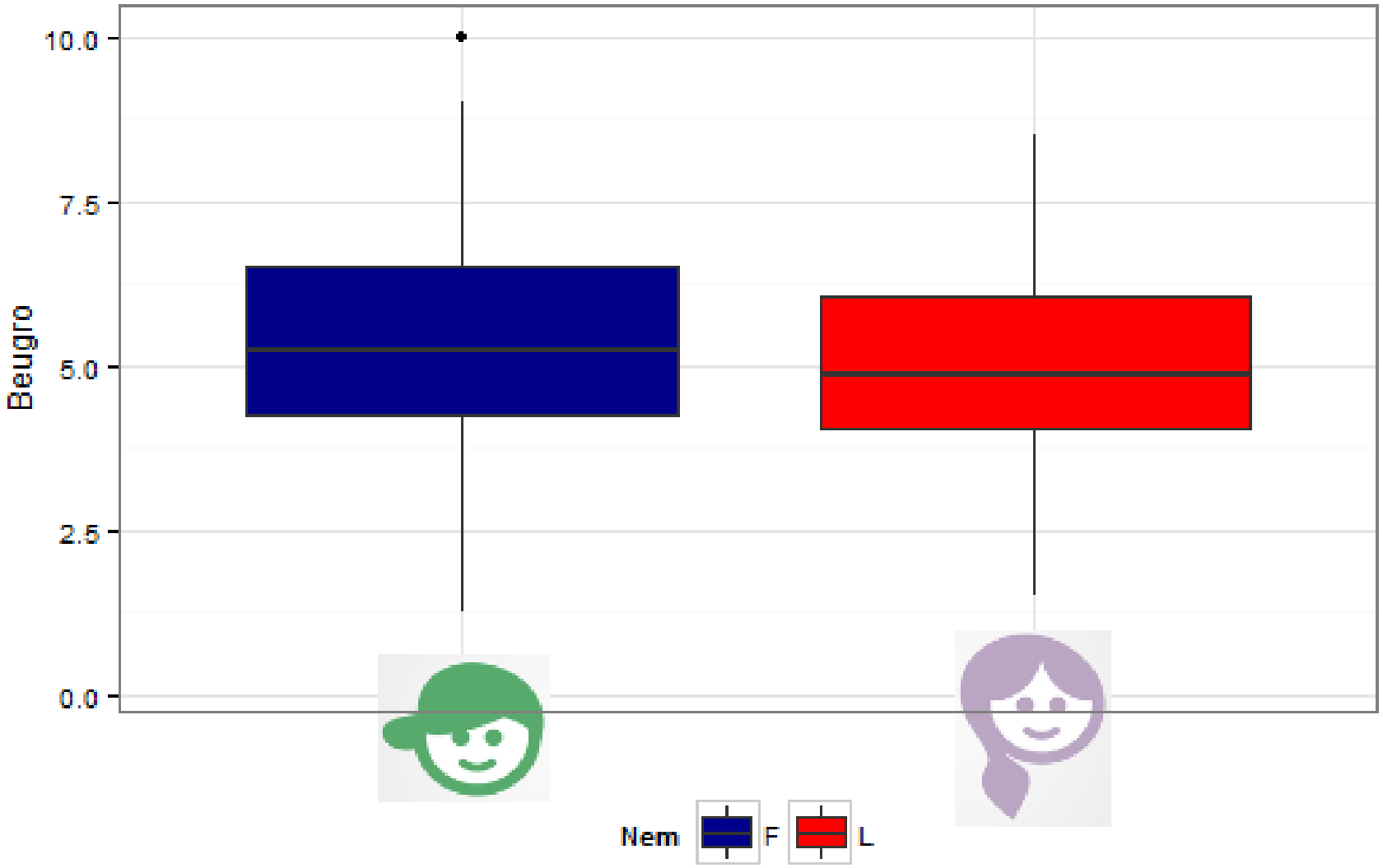
- Data set: 1000 Points in (1, 5) with uniform distribution
  - *Mean = Median = 3 ms*



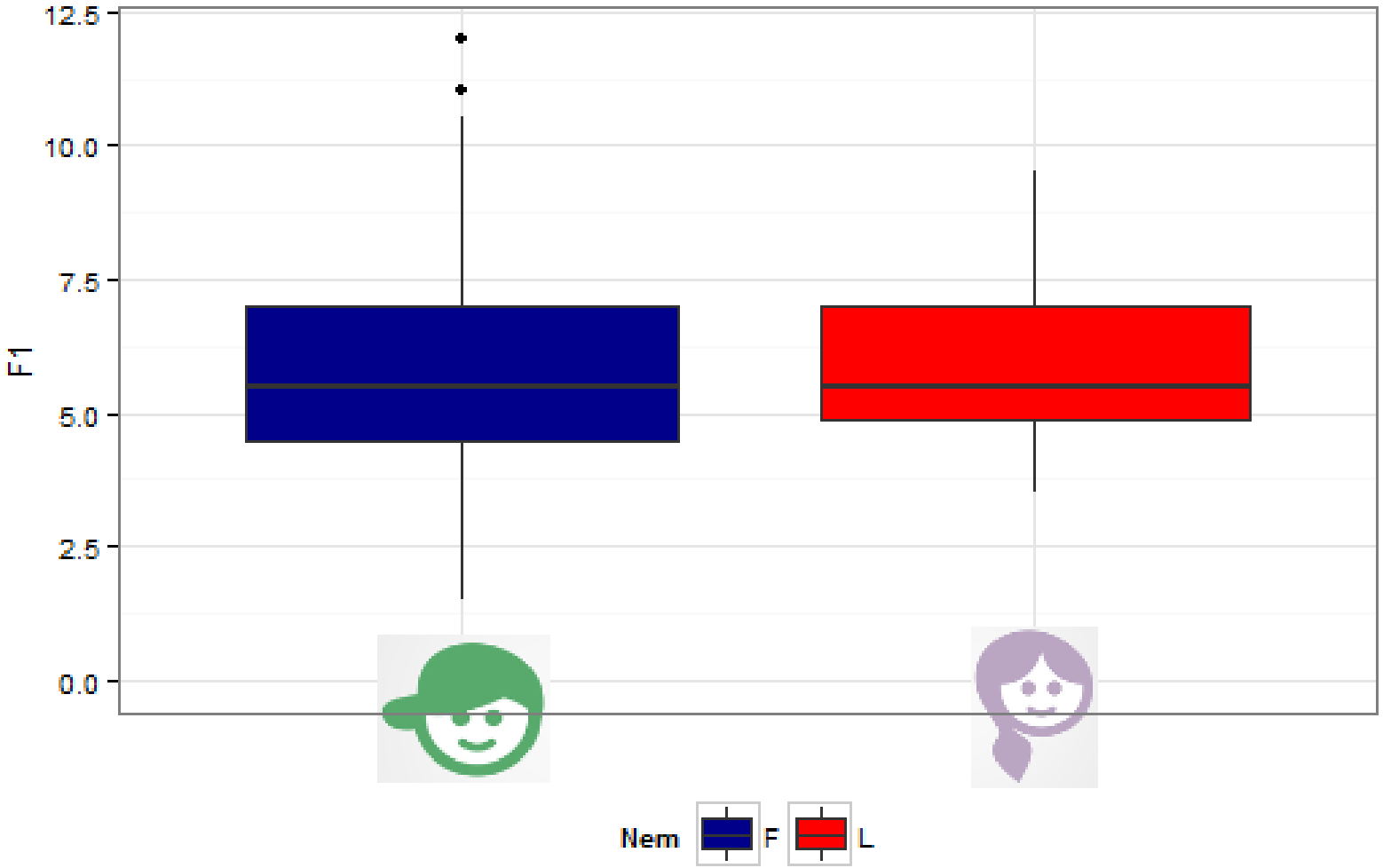
Relation between two Variables



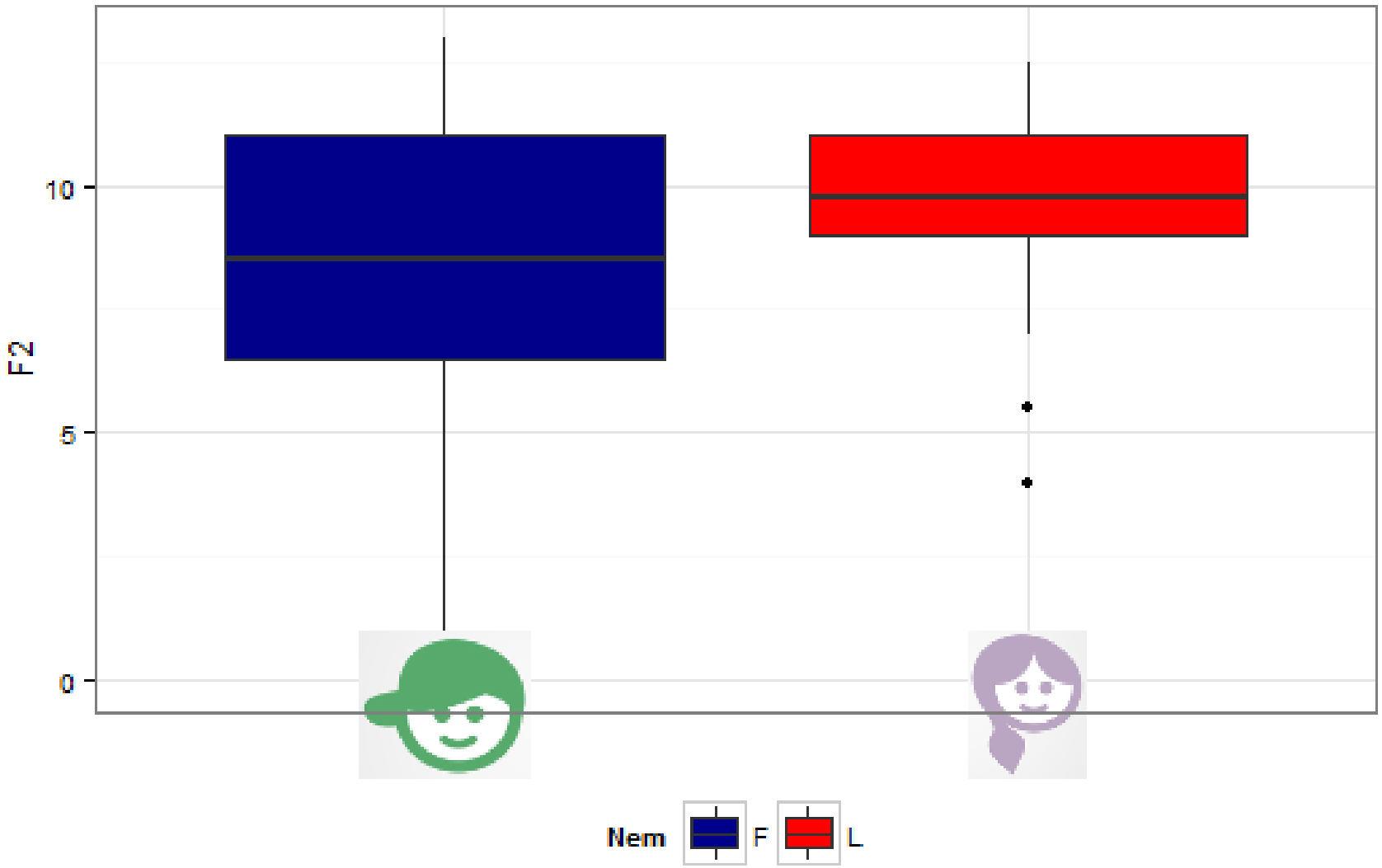
Numerical, per Category



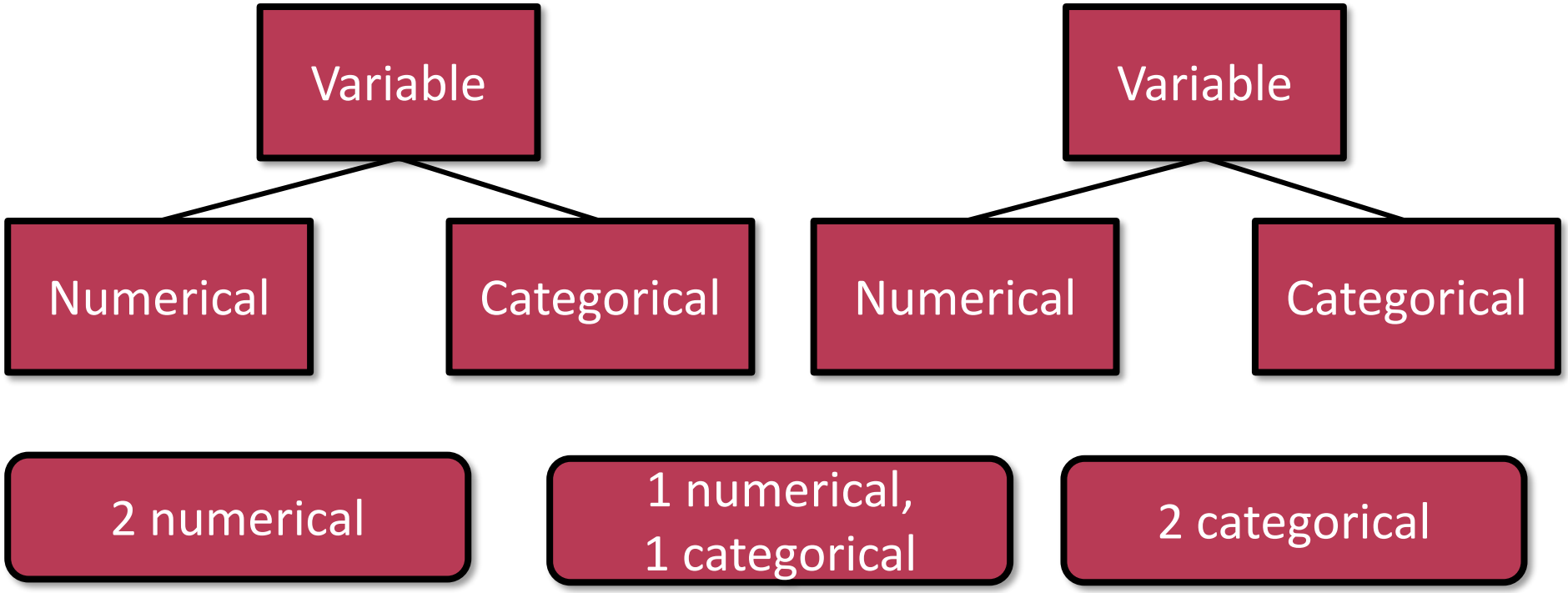
Numerical, per Category



Numerical, per Category



Relation between two Variables

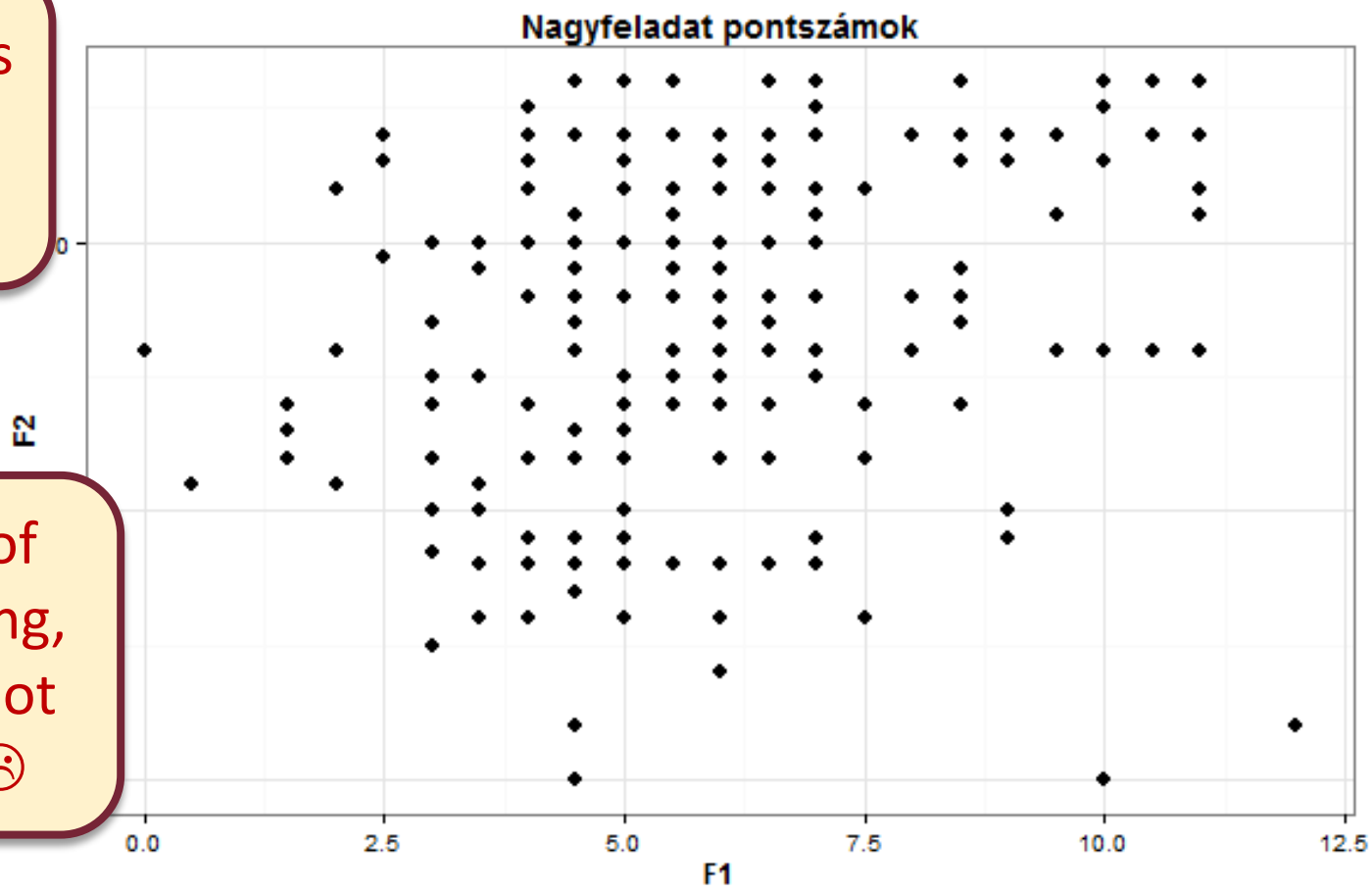


Scatter Plot

- Input variable: Results of the two main test tasks
- Question: Is there any correlation?

Pairs of results are displayed (visualised).

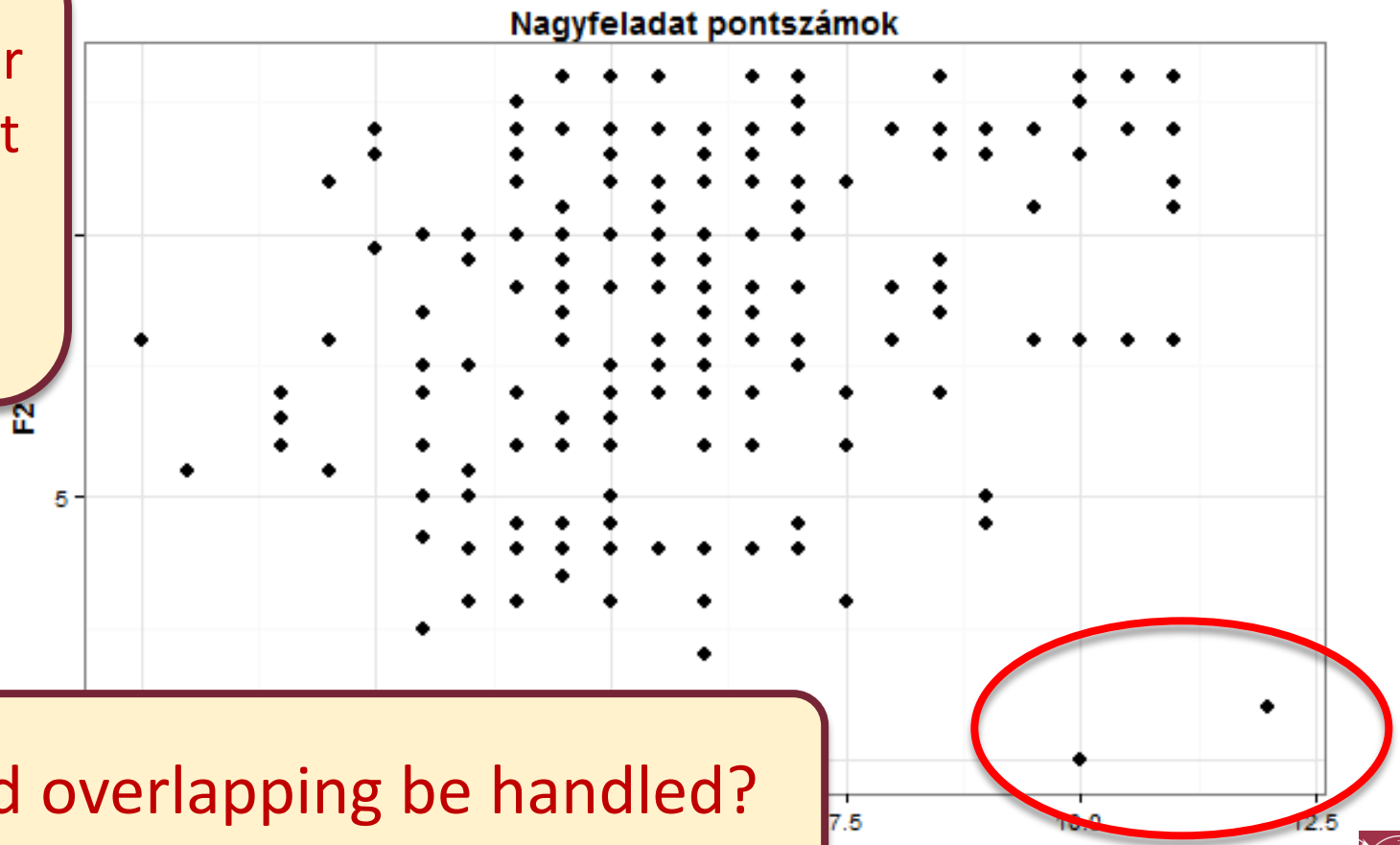
Where one of them is missing, the pair cannot be shown. ☹



Scatter Plot

- Input variable: Results of the two main test tasks
- Question: Is there any correlation?

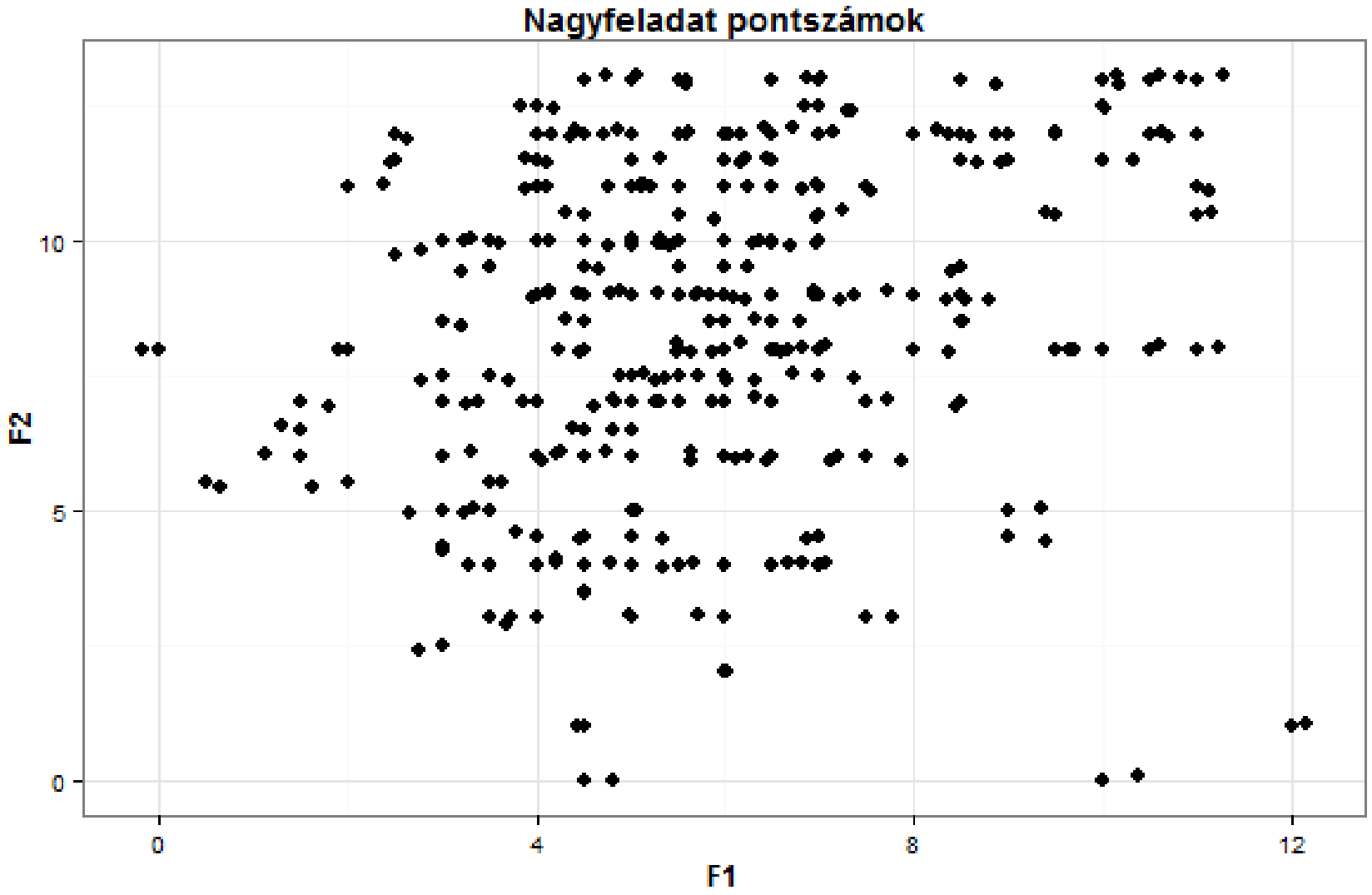
Who had a good result for Task1, had not necessarily a good one for Task2.



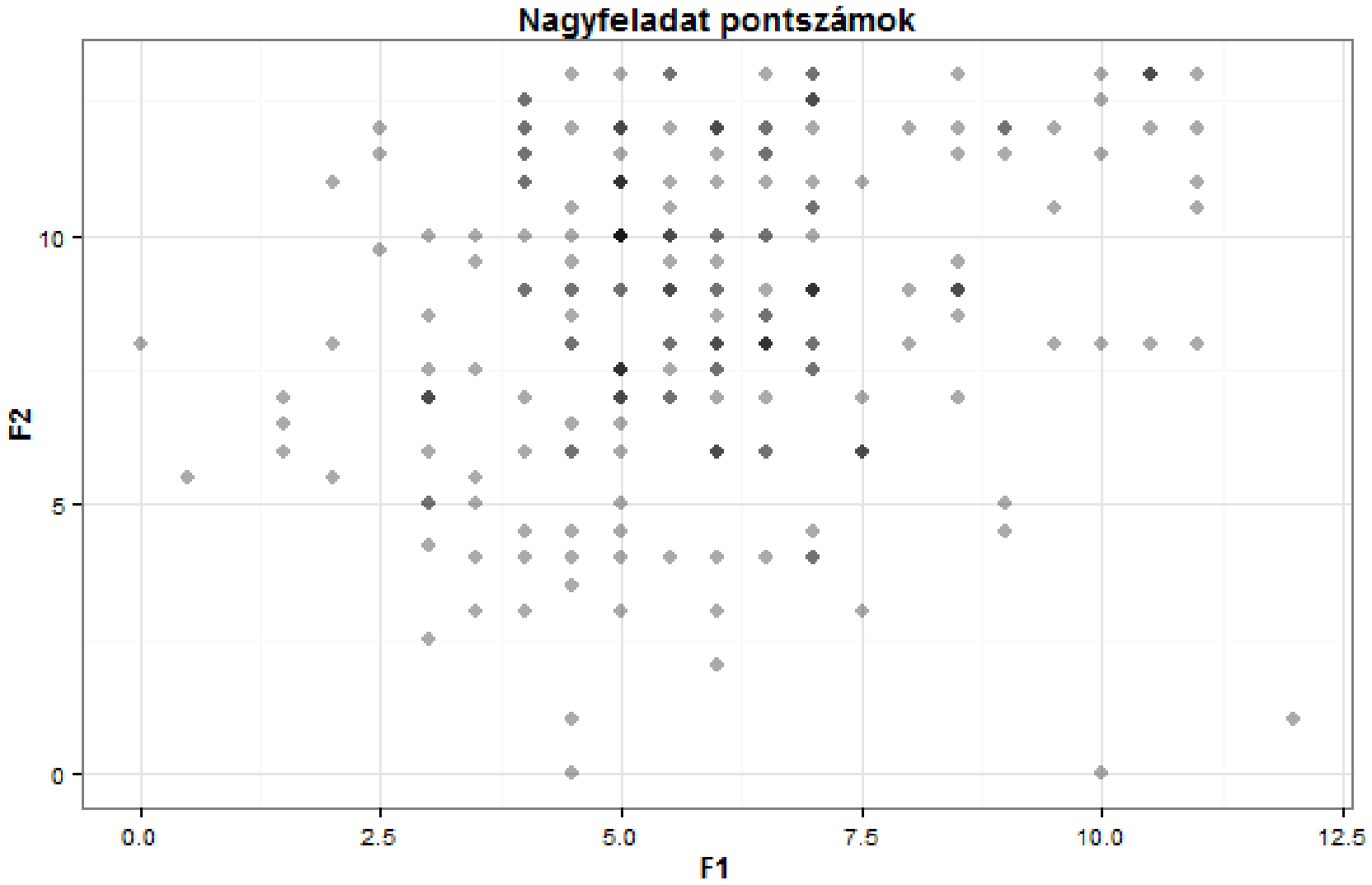
How should overlapping be handled?

Overplotting

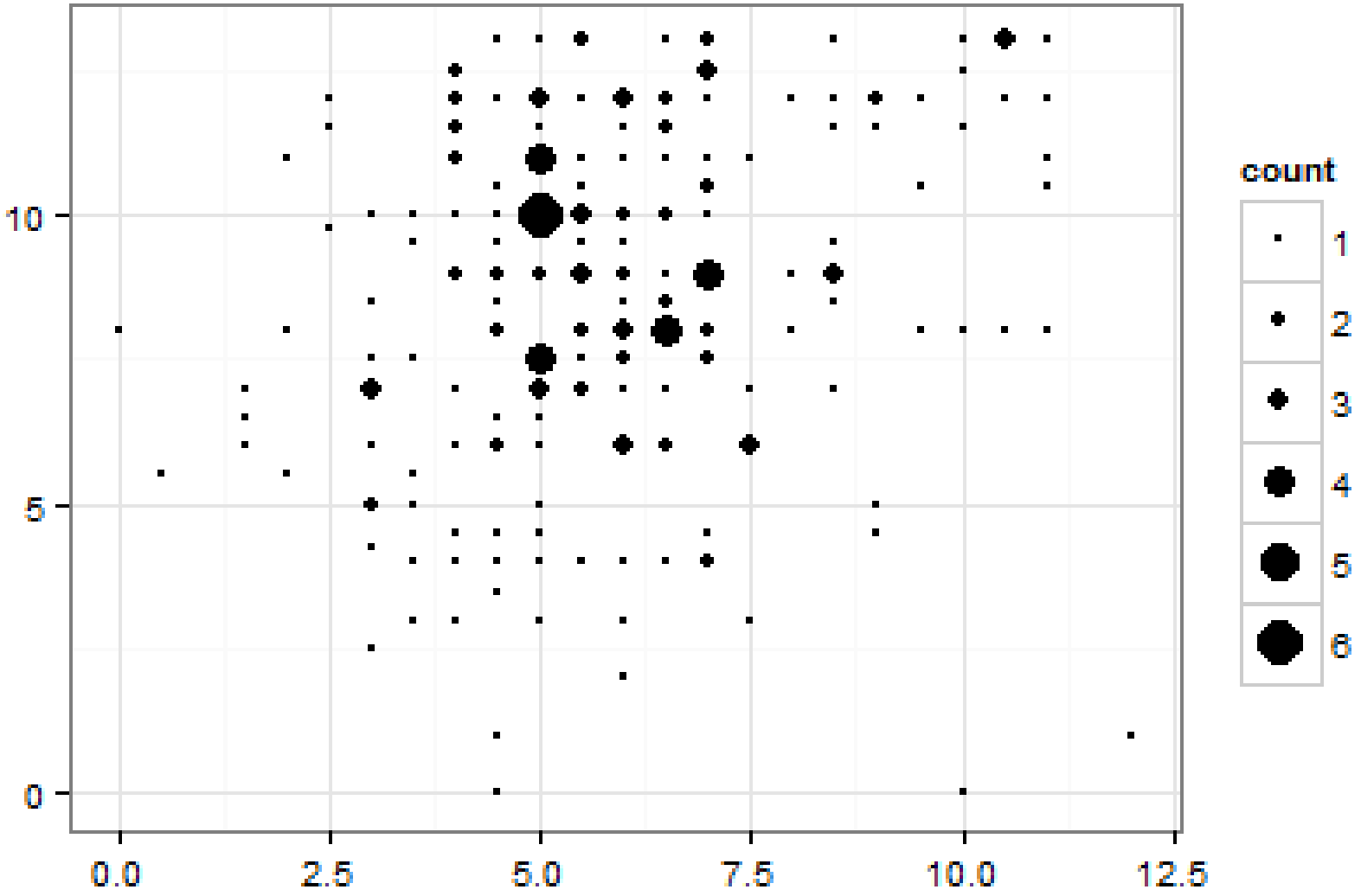
Overplotting – Solution 1: Jitter



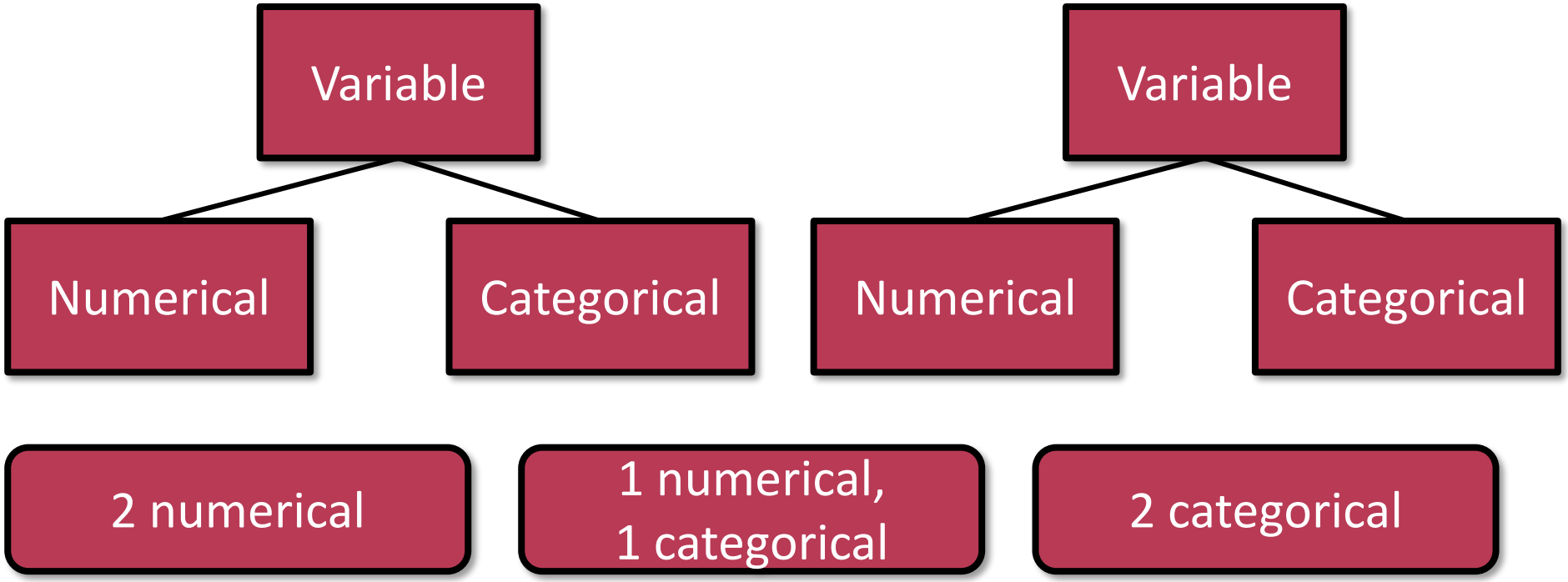
Overplotting – Solution 2: Transparency



Overplotting – Solution 3: Size

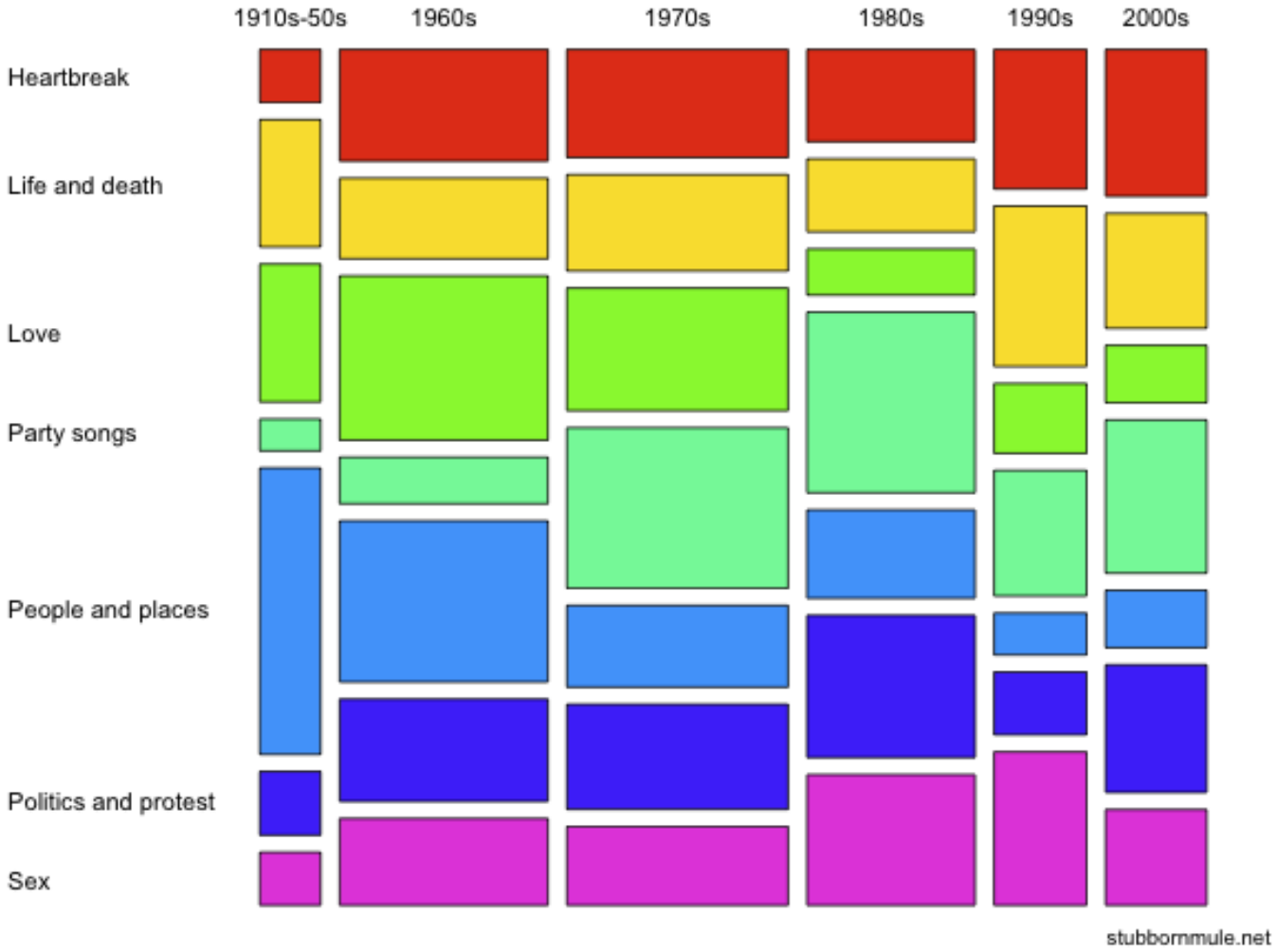


Relation between two Variables



Mosaic Plot

- Relation between 2 or more categorical variables



Guardian's list of "1000 songs to hear before you die"

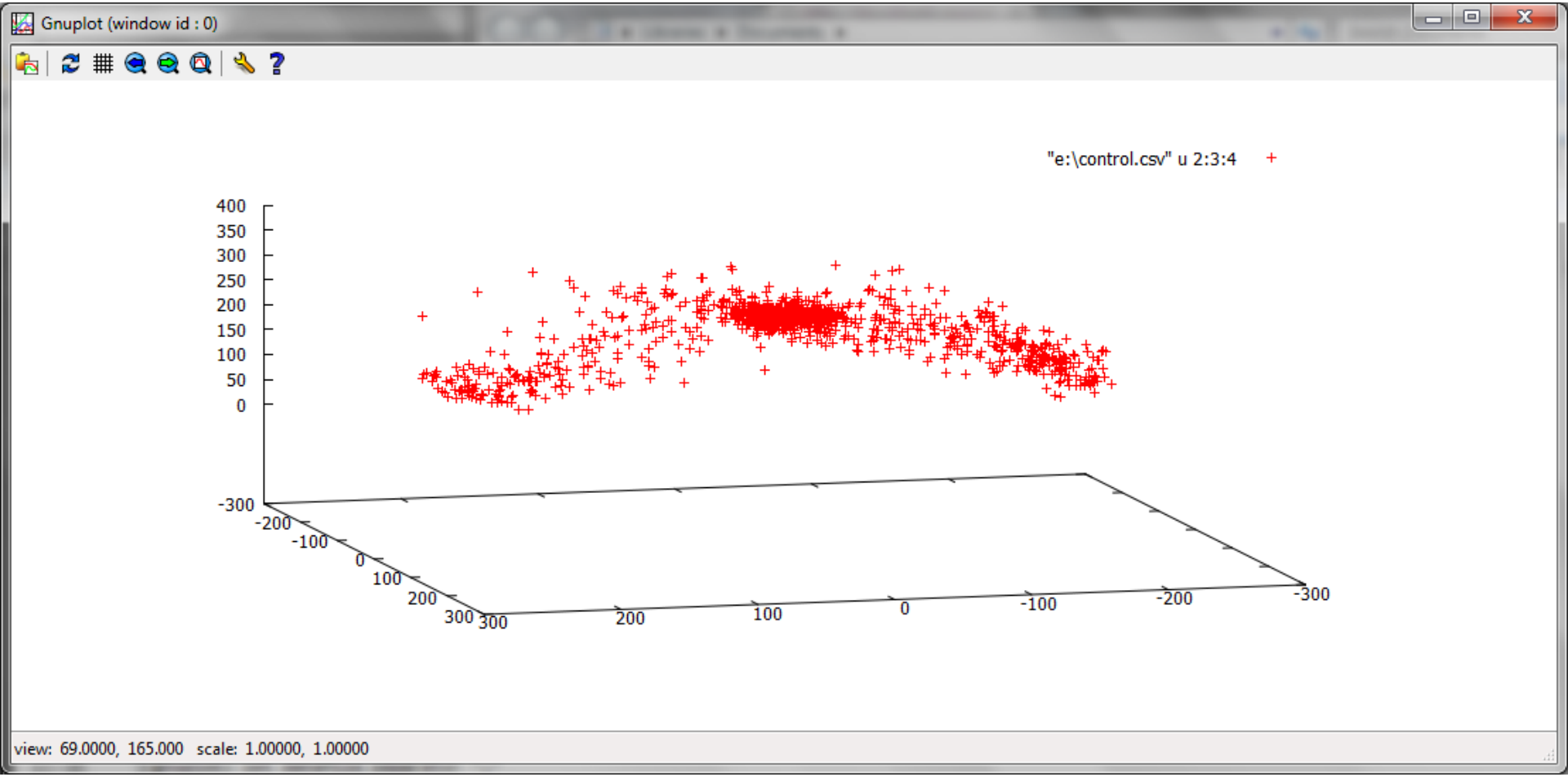
MULTIPLE VARIABLES

More than Two Variables

- Changing the properties of the graphical elements
  - Color
  - Size
  - Texture
  - Place (non-trivial way, but look at tree maps, there the place has a direct meaning)

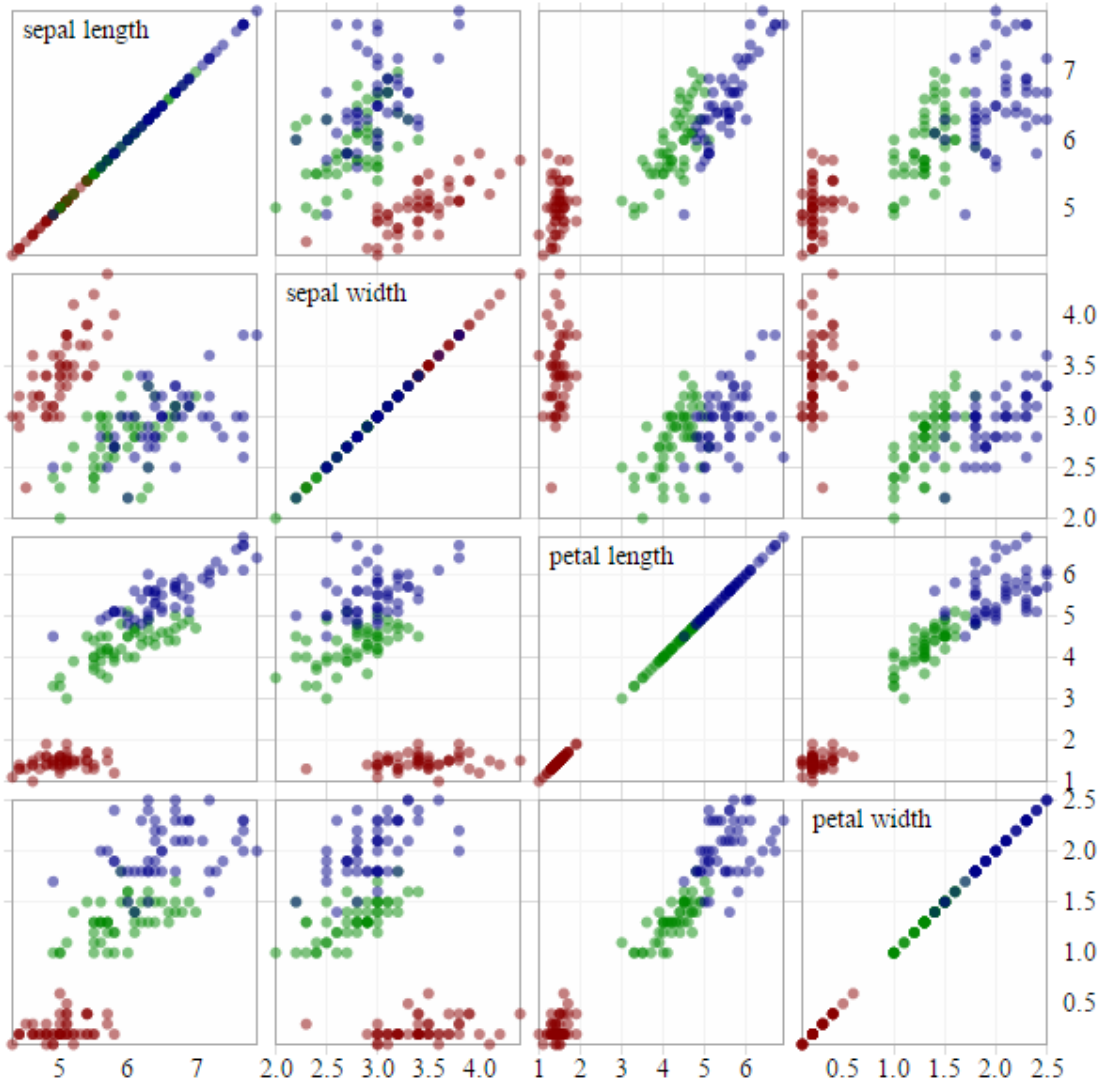
- E.g. bubble chart, heatmap, treemap

3D Plot

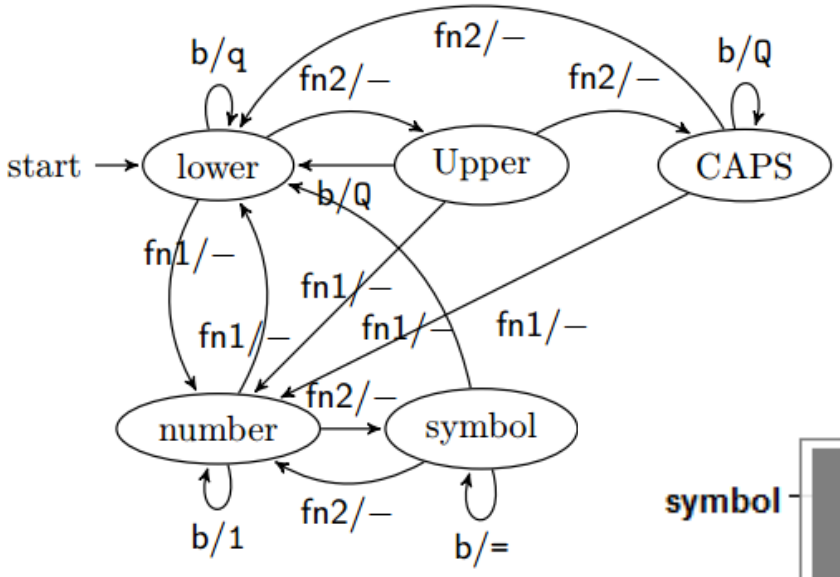

Gnuplot (window id : 0)
"e:\control.csv" u 2:3:4
400
350
300
250
200
150
100
50
0
-300
-200
-100
0
100
200
300
300
200
100
0
-100
-200
-300
view: 69.0000, 165.000 scale: 1.00000, 1.00000

Bubble Chart: Average Age by Regions

Scatterplot matrix

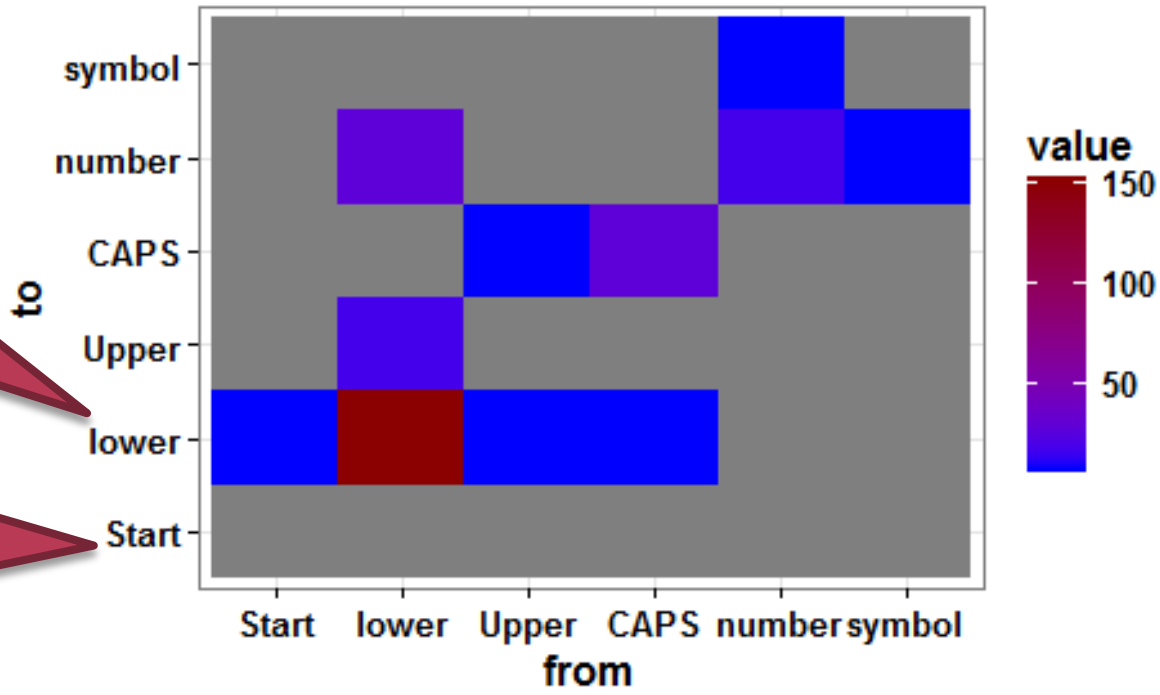


Heatmap: Operations Statistics

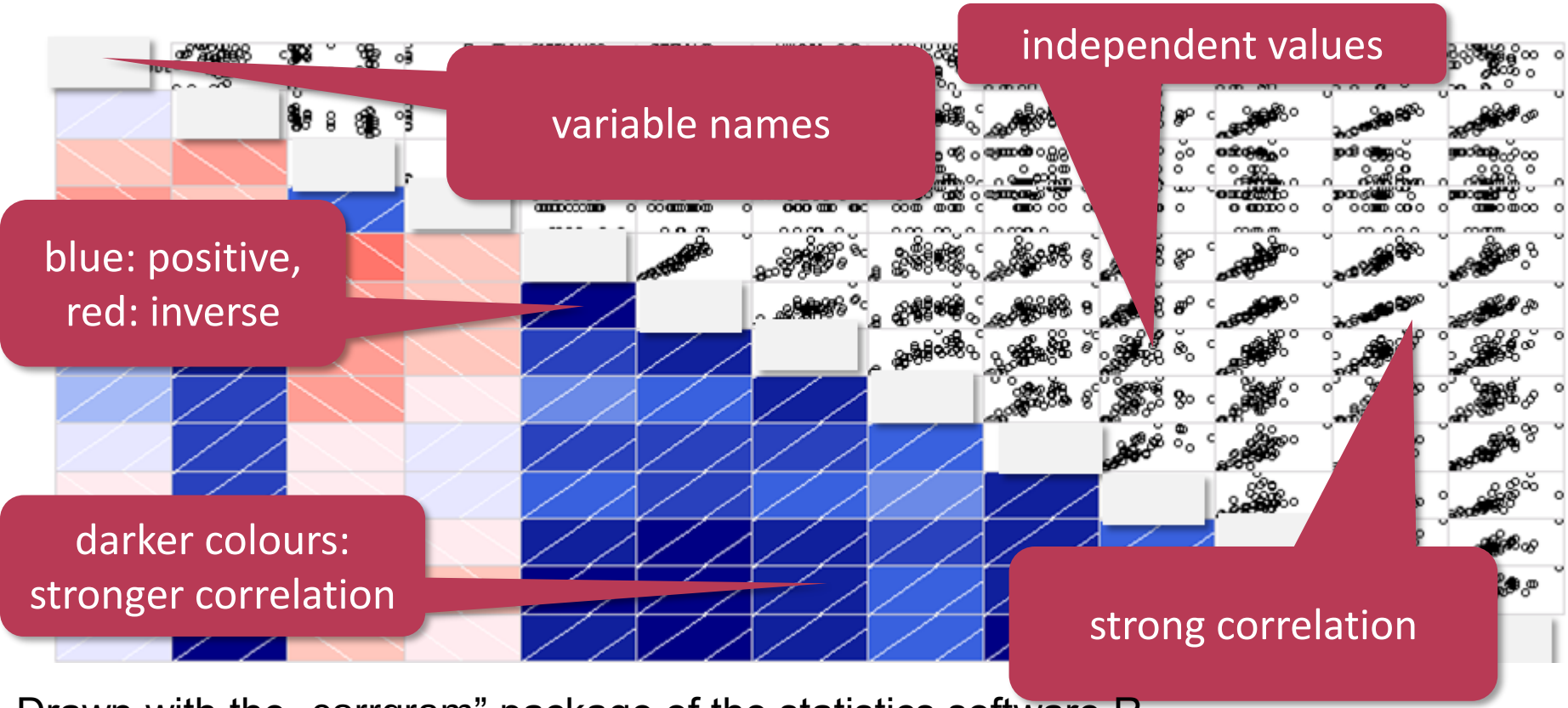


We mostly write simple texts.

In Start we only start, but we never return.



Outlook: Pairwise Correlation of Multiple Values



Drawn with the „corrgram" package of the statistics software R.

Correlation (see Probability Theory):

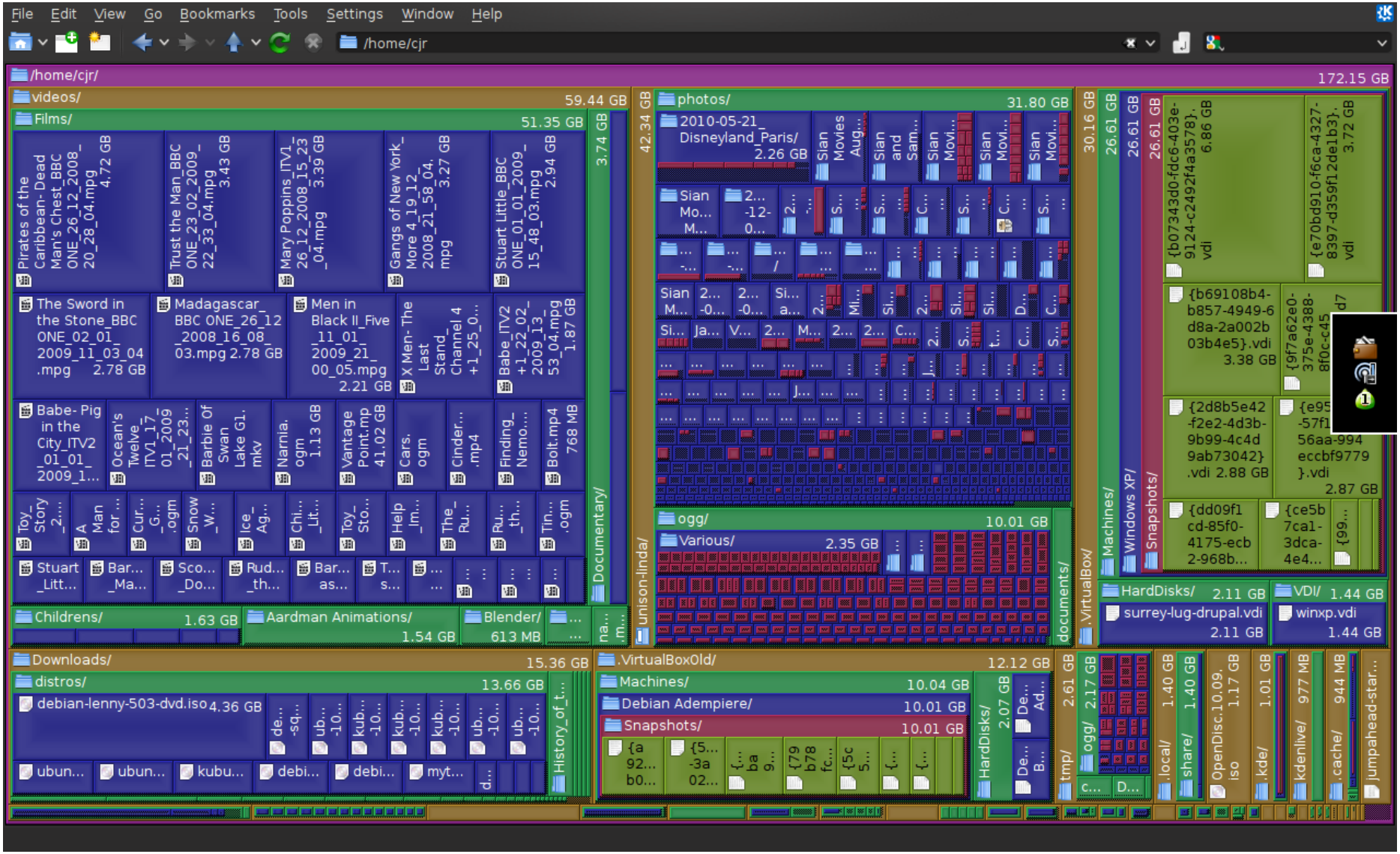
Strength and direction of the linear dependency between two variables

Over the diagonal: **scatterplot matrix**

Goal: Filtering out the related variables, identification of the **outliers**.

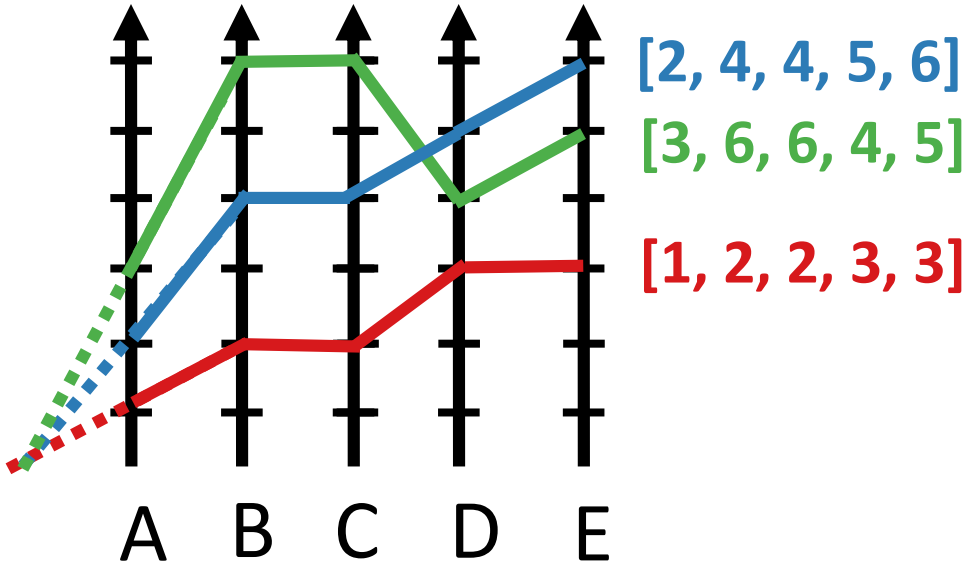
→ Which variables are important for the prediction of the load?

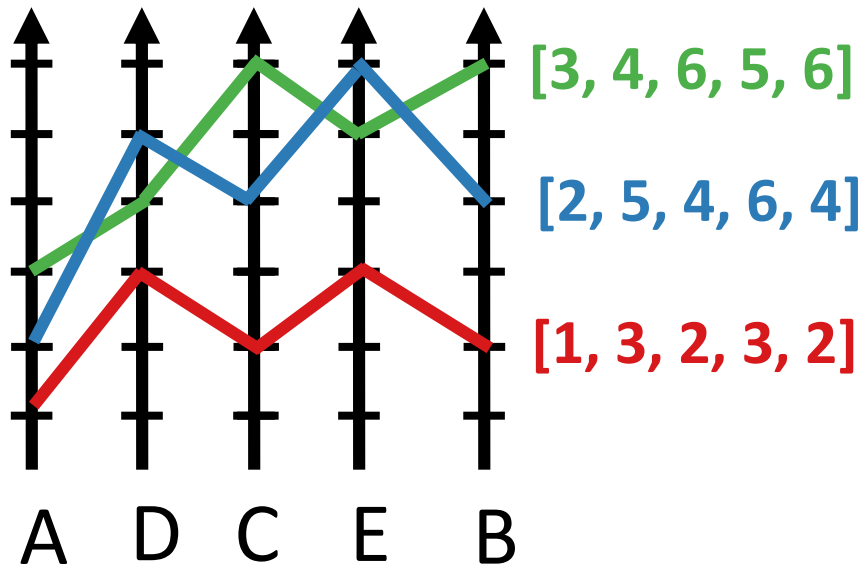
Tree Map: e.g. File System



Parallel Coordinates

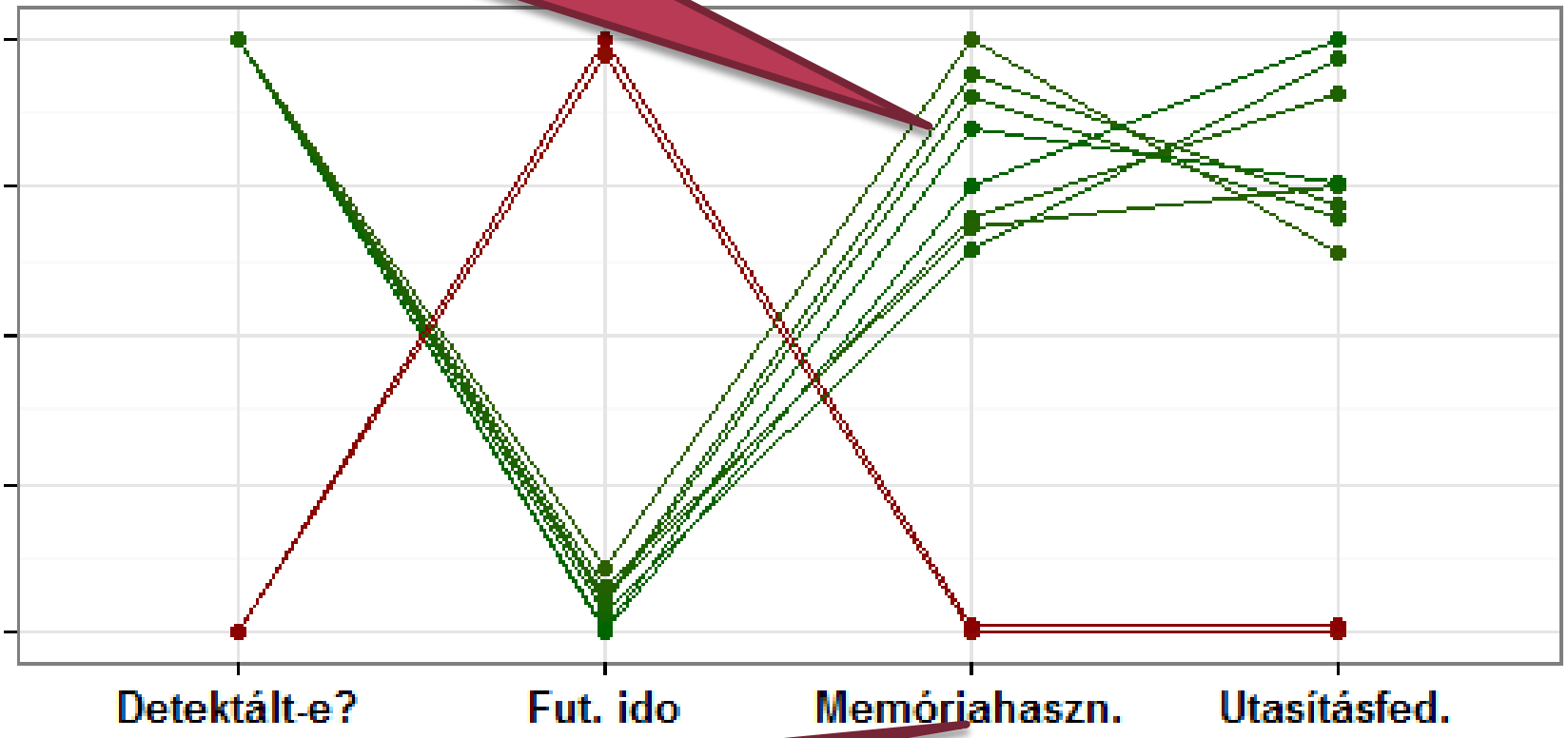
- Multi-dimensional visualization
- Compact, scalable
- Axis order?





Parallel Coordinates: Analysis of the Test Cases

1 test case: 1 broken line

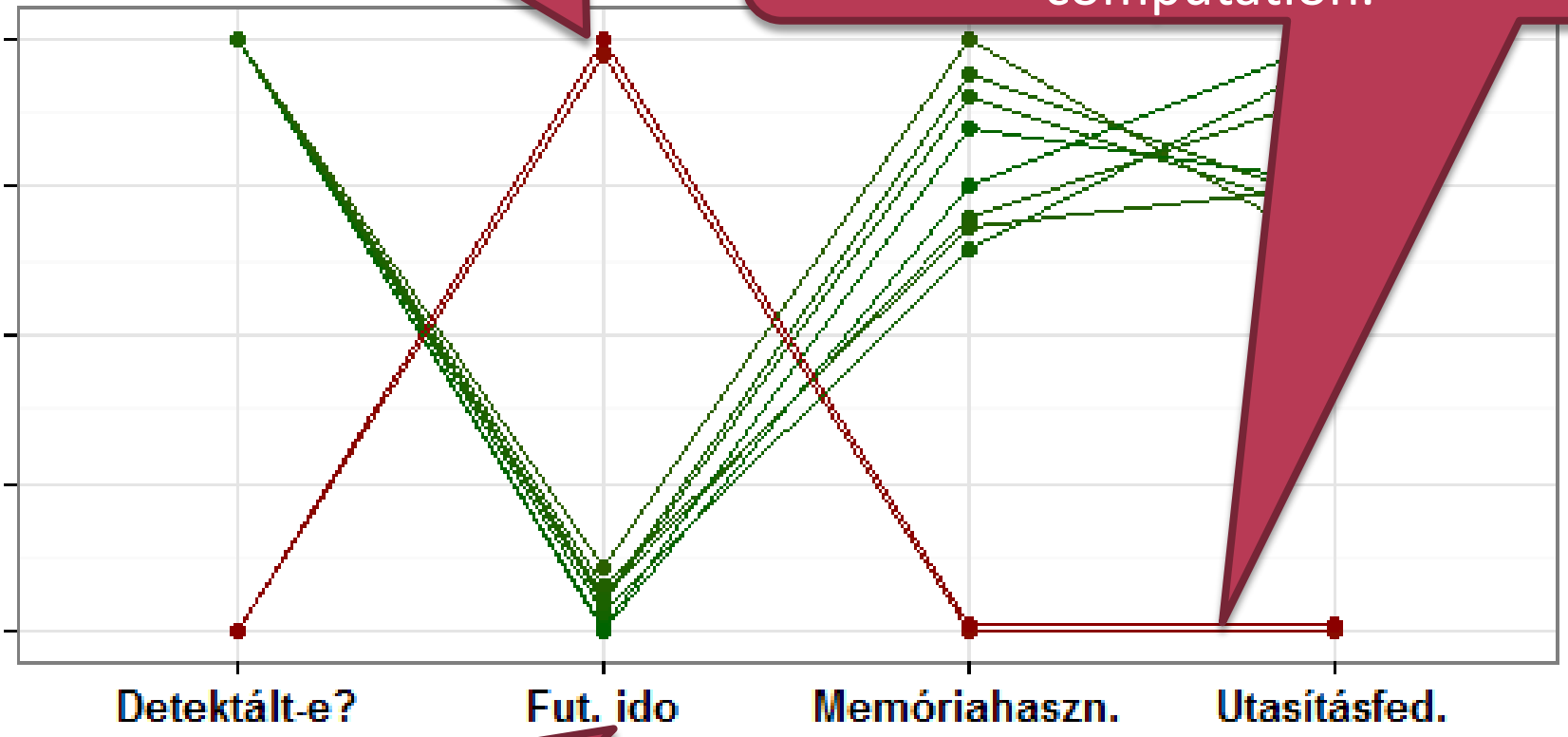


The variables appear on the x-axis

Parallel Coordinates: Analysis of the Test Cases

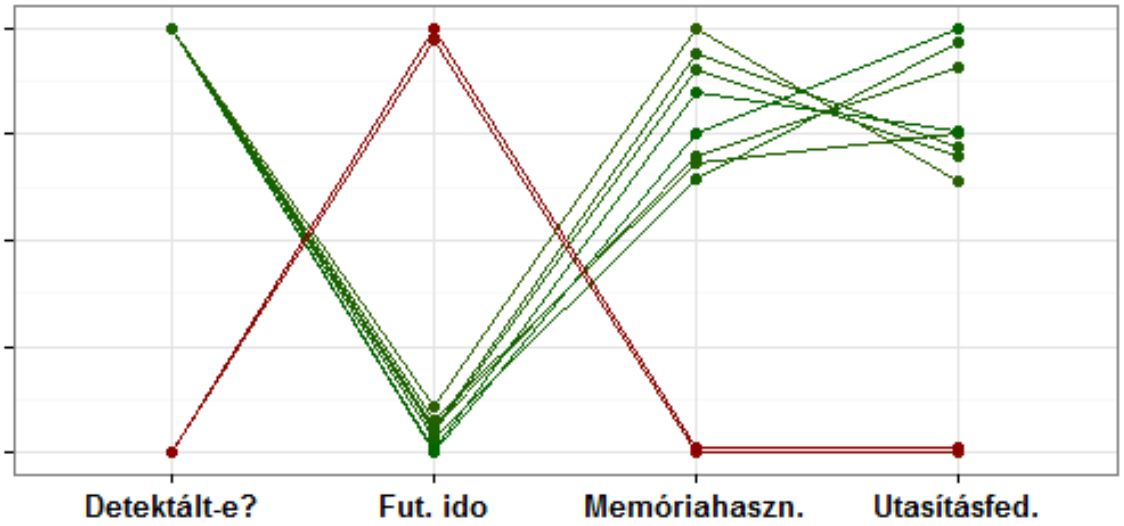
Timeout?

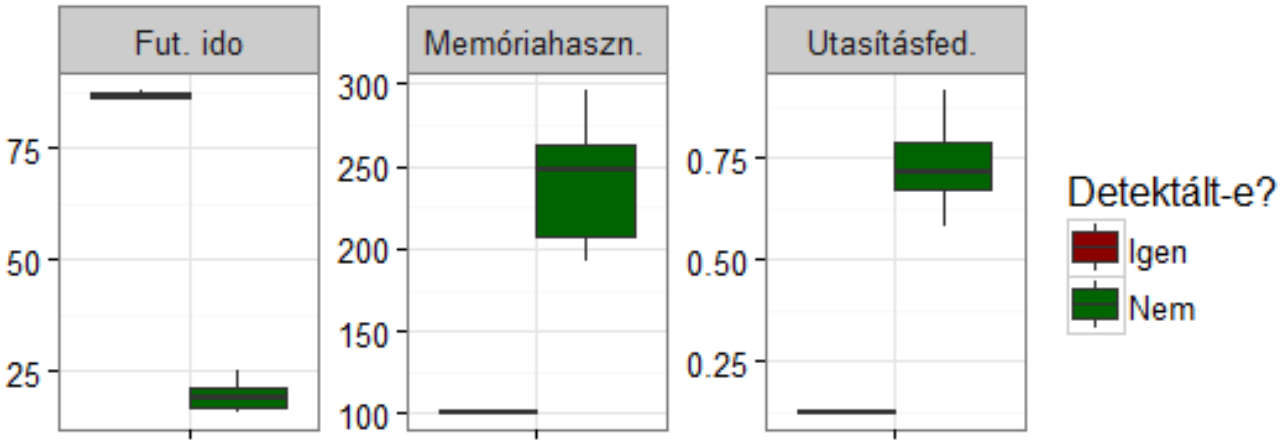
The ones detecting an error did not even came to the actual computation.



Run time and memory usage seem to be in a positive relation (if the test is successful)

Parallel Coordinates: the Alternatives





Radar Chart: An Extension of Parallel Coord.