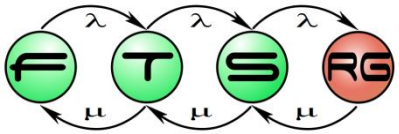
Variables and plots

Visual Analysis of Measurement Data

2018.09.13.

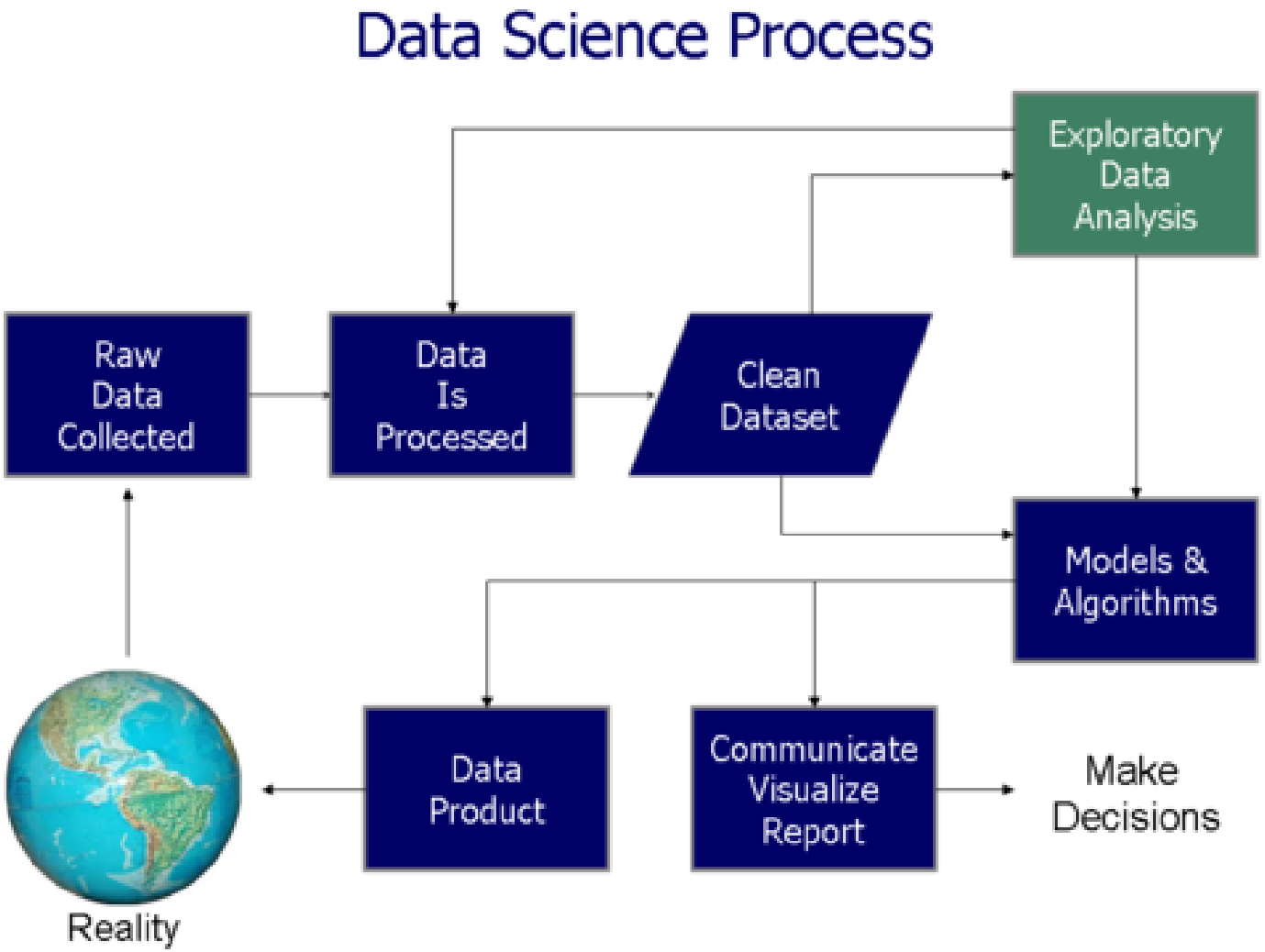
Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Outline

- Data format/representation
- Data processing
- ETL, workflow support
- Outlook: OLAP
- Case studies

Data science „process”



https://en.wikipedia.org/wiki/Data_science

DATA FORMAT

Tidy data

- 3 Simple rules to facilitate statistics and visualization
- One variable – one column
- One observation – one row
- Each type of observational unit – one table
- … seems to be trivial
- … not true in most practical cases
- … and even for staitstical tools (e.g. output of R packages)

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23. https://github.com/hadley/tidy-data

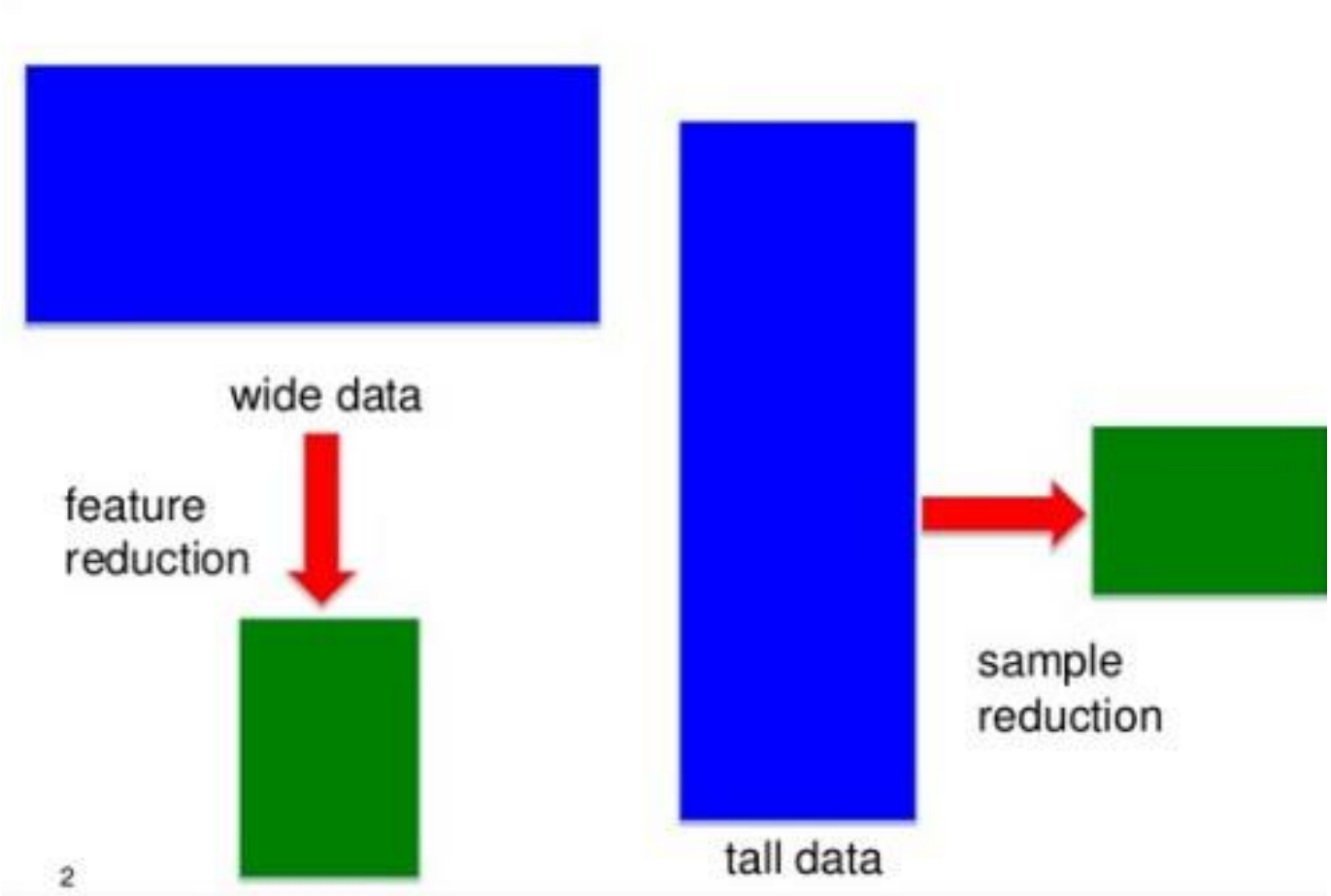
Data originally: long/wide

Person	Age	Weight
Bob	32	128
Alice	24	86
Steve	64	95

Person	Variable	Value
Bob	Age	32
Bob	Weight	128
Alice	Age	24
Alice	Weight	86
Steve	Age	64
Steve	Weight	95

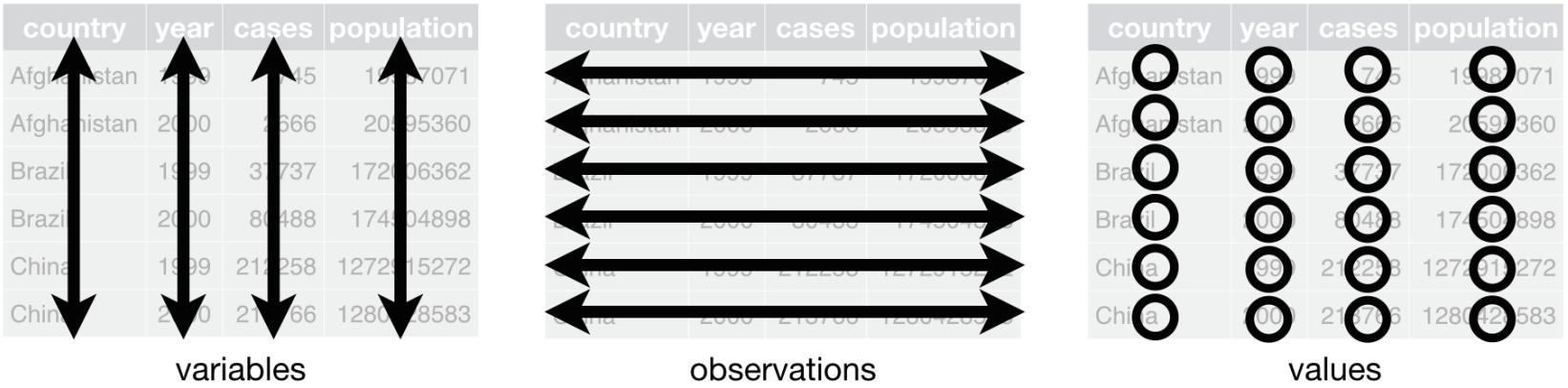
https://en.wikipedia.org/wiki/Wide_and_narrow_data

How to use these formats?

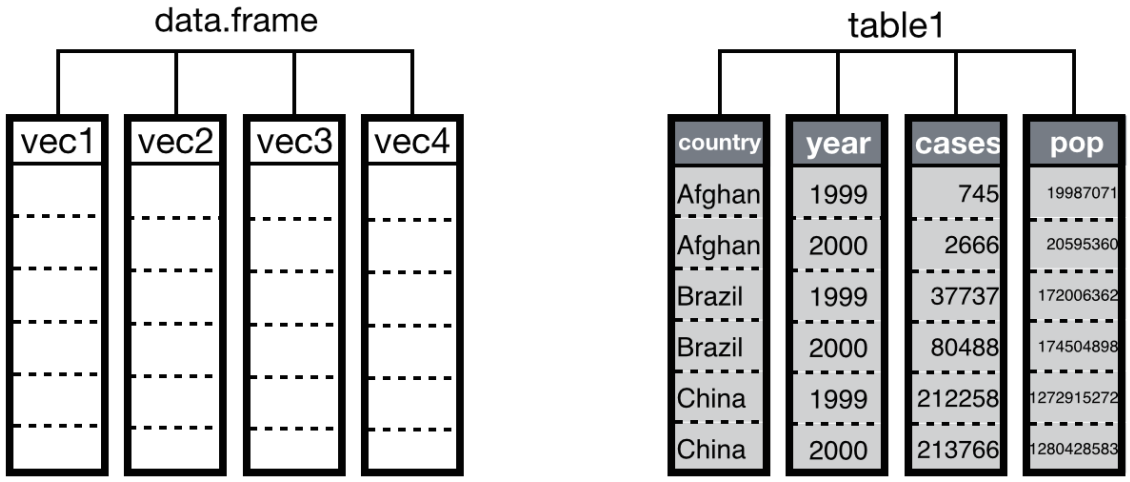


Sparse Screening for Exact Data Reduction. Jieping Ye, Arizona State University

Examples for tidy data



R dataframe representation:



http://garrettgman.github.io/tidying/

„tidying”

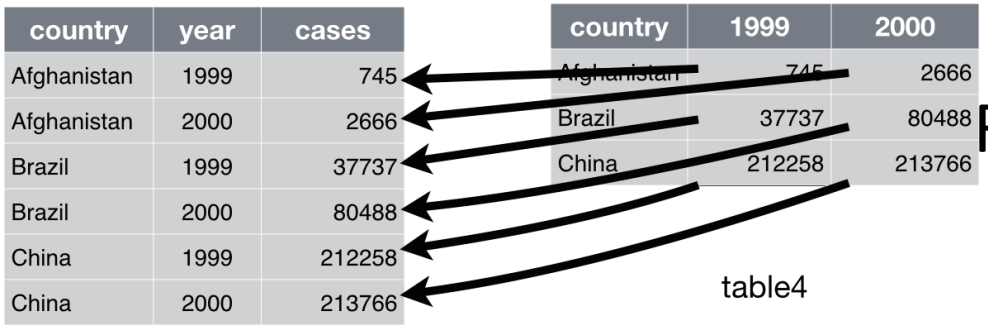
country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2

R: spread(data,key,value)

http://garrettgman.github.io/tidying/

„tidying”



country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

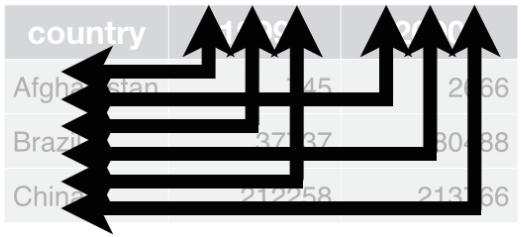
table4

R: spread(data,key,value)

Generalization?

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

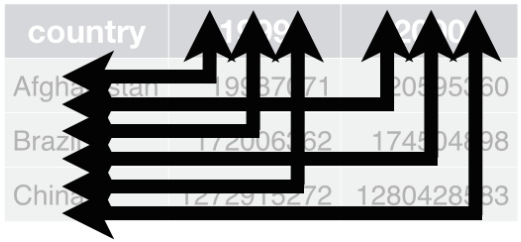
table4



country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

table5

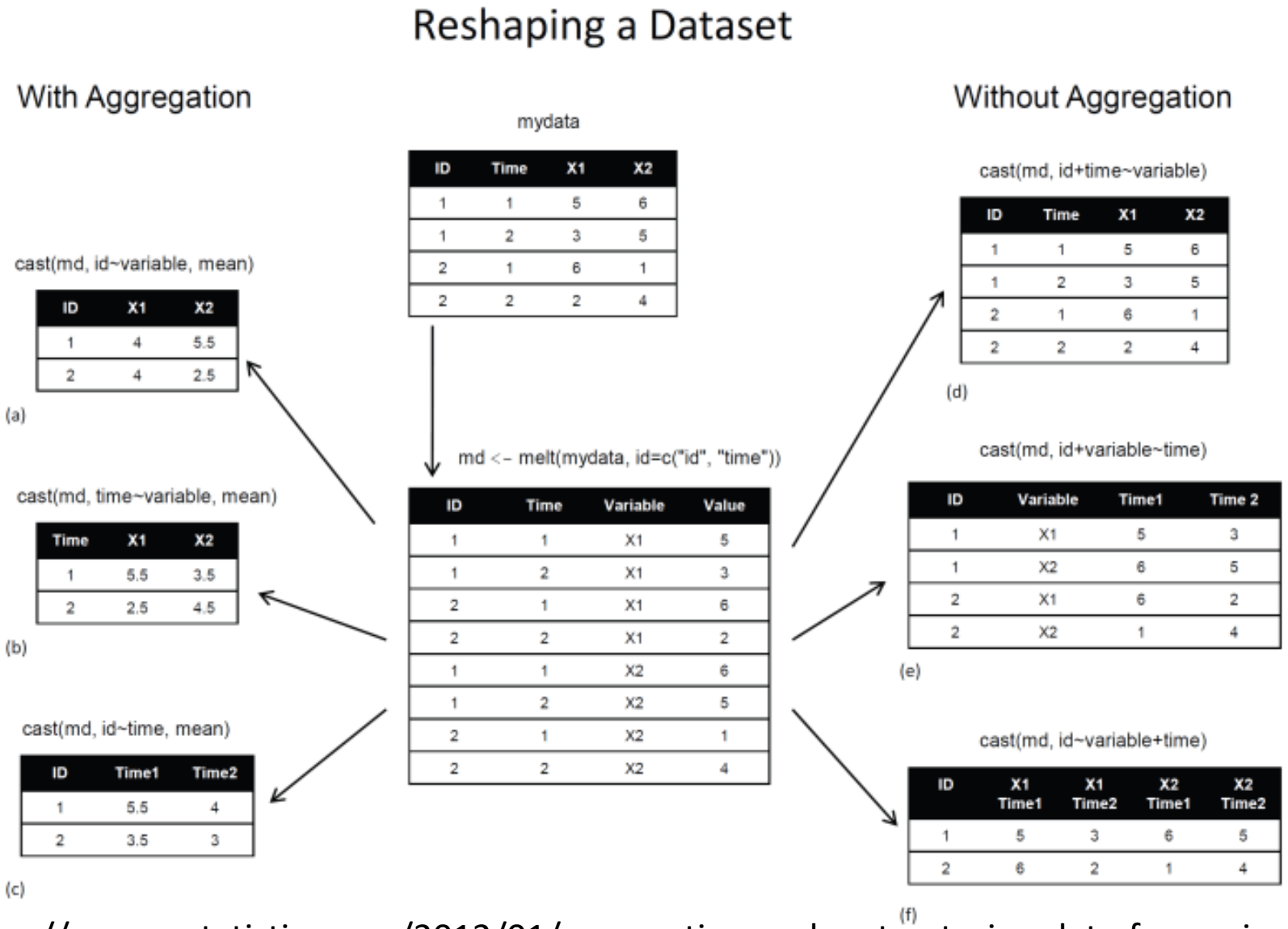
variables



observations

http://garrettgman.github.io/tidying/

Data restructuring examples (in R)



https://www.r-statistics.com/2012/01/aggregation-and-restructuring-data-from-r-in-action/

DATA STORAGE

Common data storage techniques

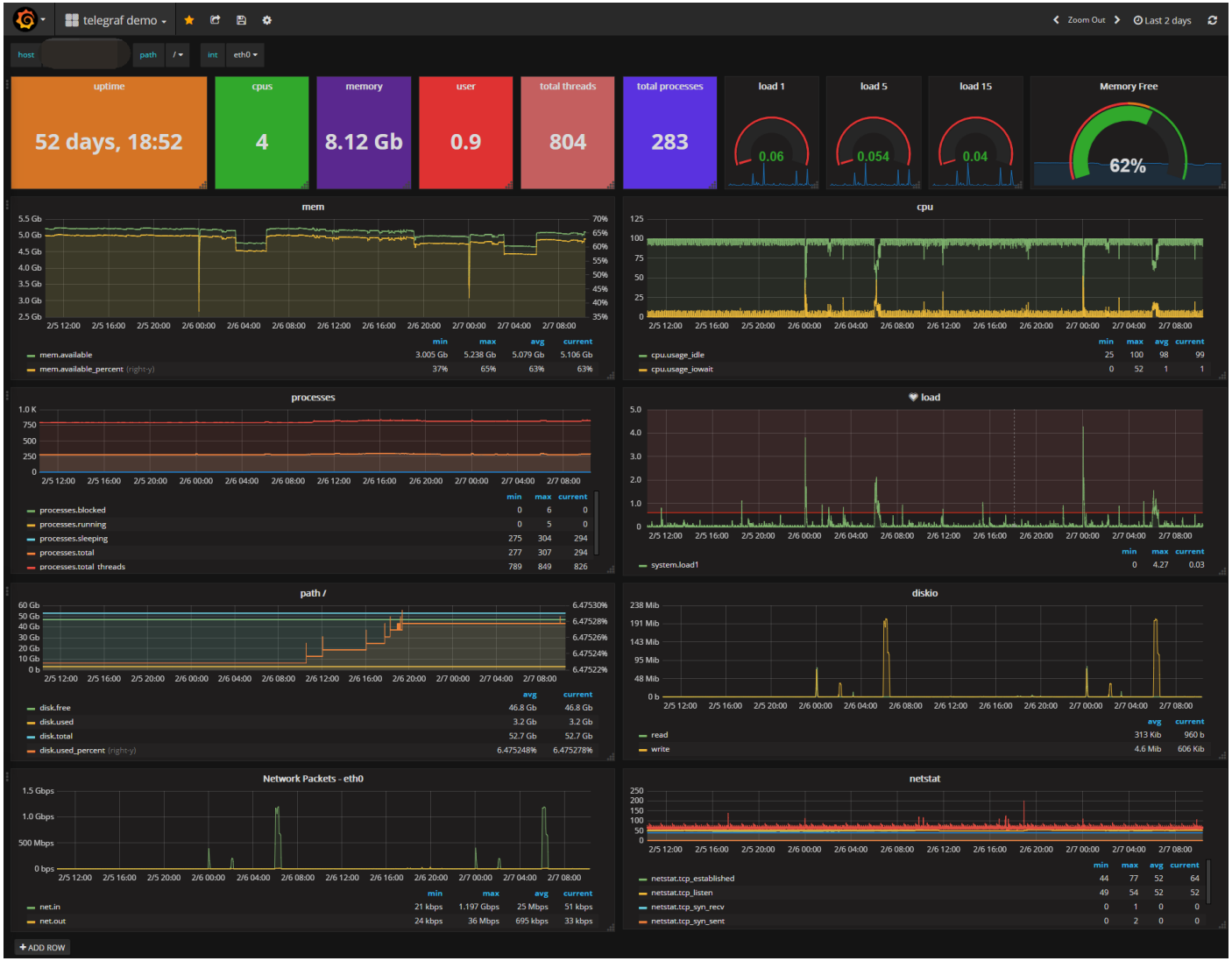
- .CSV
  - Majority of inputs
  - Length? Header? Encoding?
- DB with a schema (in memory?)
- Graph databases, ontologies, RDF…
- Key-value stores (redis)
- Time series databases (openTSDB, influxDB)
  - Time series + metadata
- „Data in motion”
  - Streams as input for processing/analysis

Time series example: influxDB

- Data: measurement
  - Fields, tags, timestamp

AGGREGATIONS	SELECTORS	TRANSFORMATIONS
COUNT()	BOTTOM()	CEILING()
DISTINCT()	FIRST()	DERIVATIVE()
INTEGRAL()	LAST()	DIFFERENCE()
MEAN()	MAX()	FLOOR()
MEDIAN()	MIN()	HISTOGRAM()
SPREAD()	PERCENTILE()	NON_NEGATIVE_DERIVATIVE()
SUM()	TOP()	STDDEV()

Dashboards… (e..g Grafana)

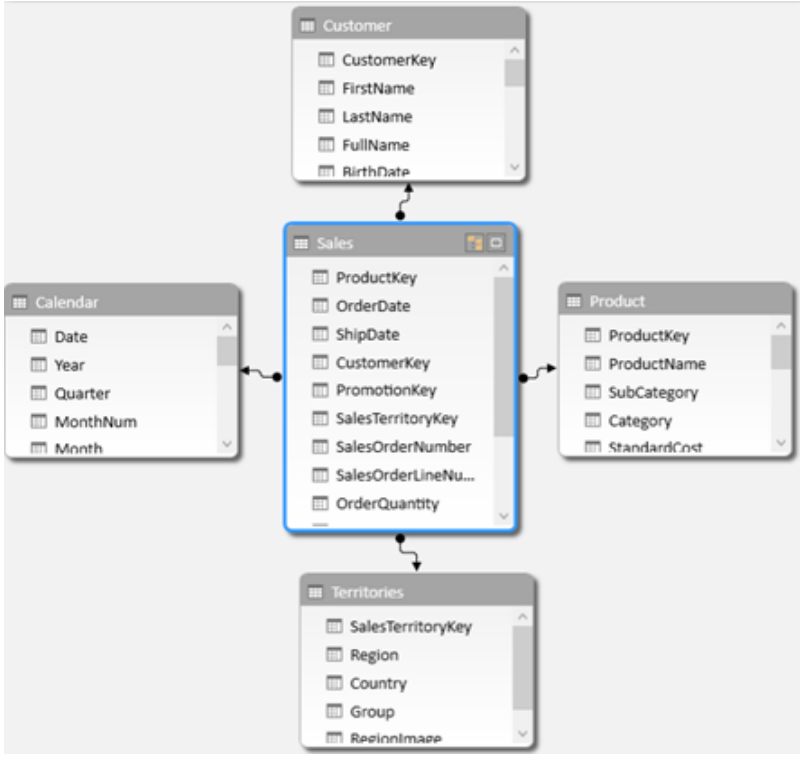


https://grafana.com/dashboards/1443

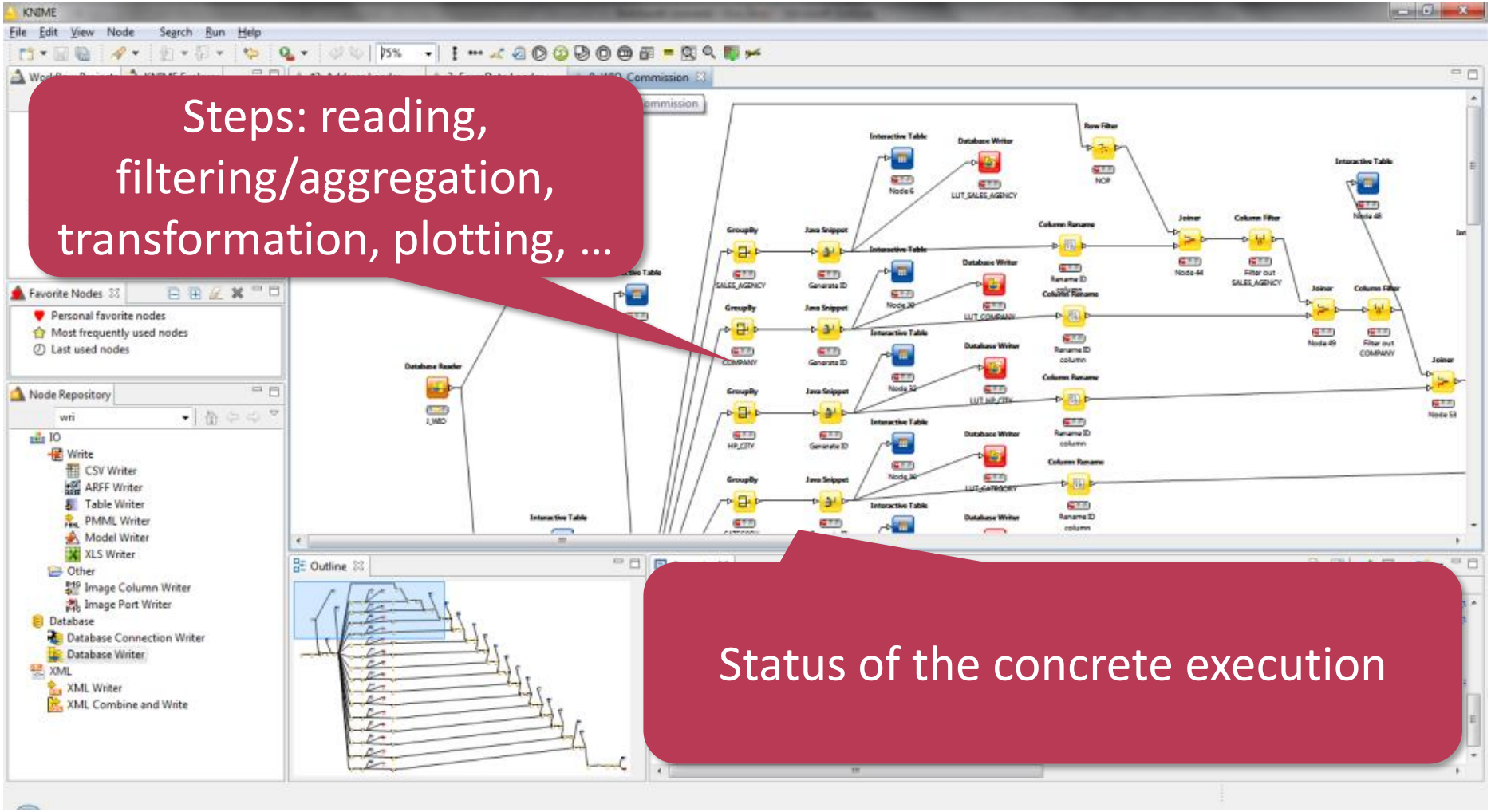
DATA PROCESING WORKFLOW & TOOLS

ETL

- „Extract-Transform-Load”
- Originally: to fill a snowflake/star schema
- In data science: create dataframes
- Cleaning tasks
  - Standardization
  - Normalization
  - Deduplication
  - Enrichment
  - Clear/fill NAs



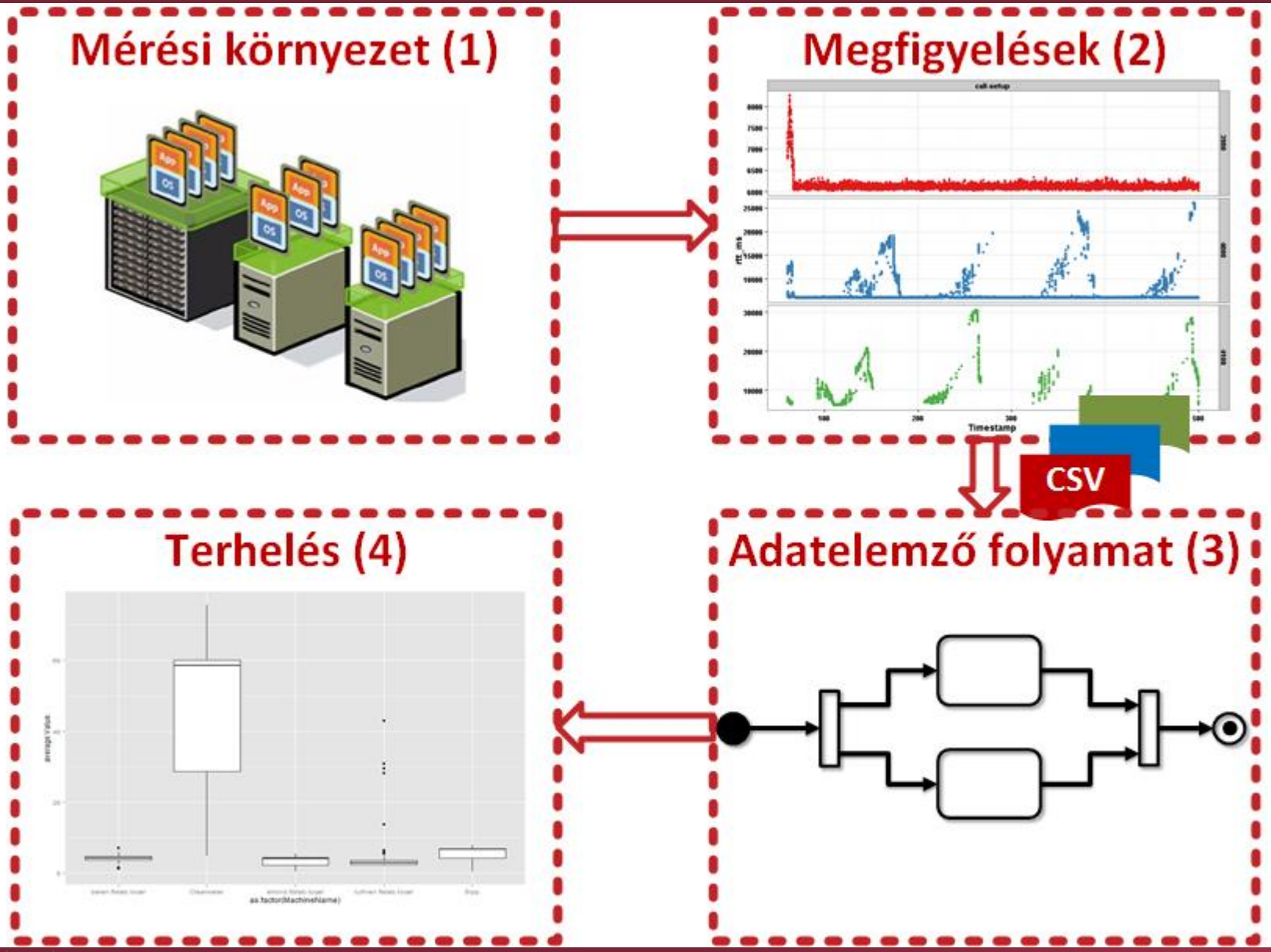
Example data processing workflow (KNIME)



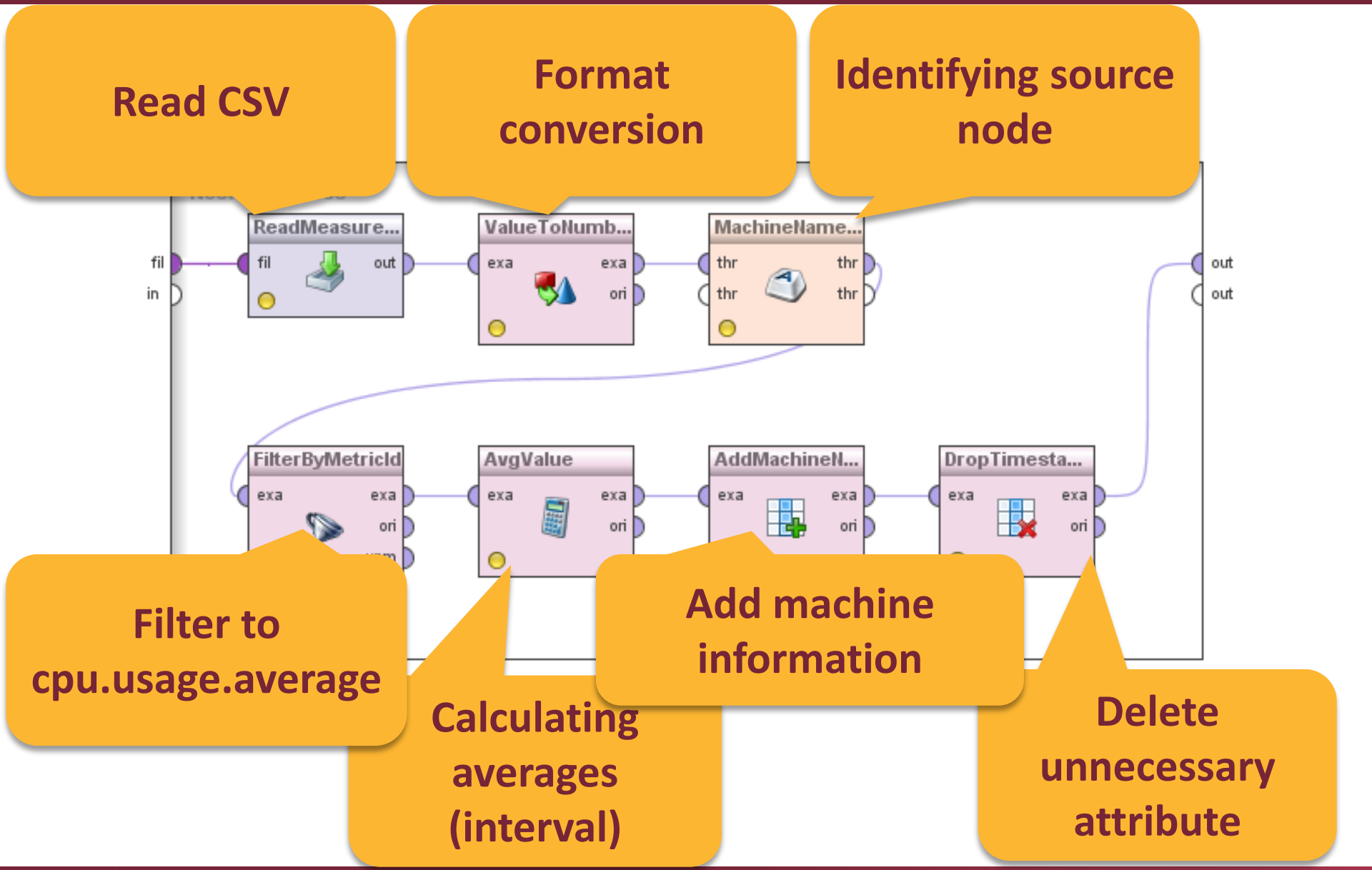
CASE STUDY

Processing of measurement data

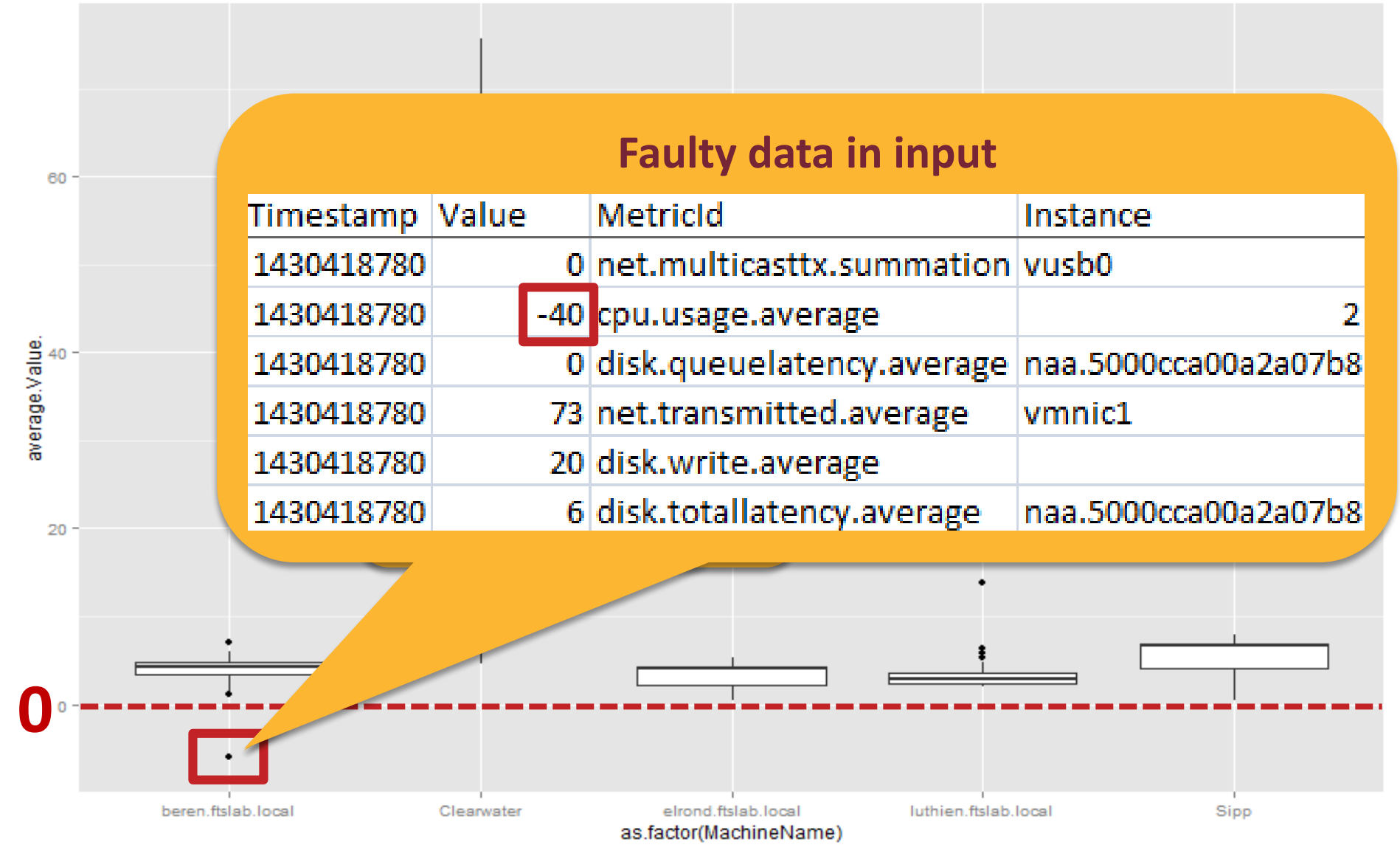
Motivating example



Measurement processing: RapidMiner



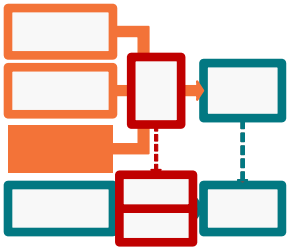
Constraint on data?

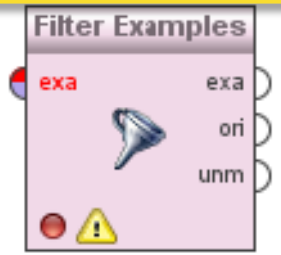
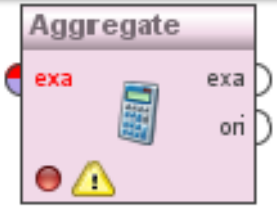
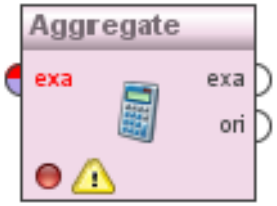


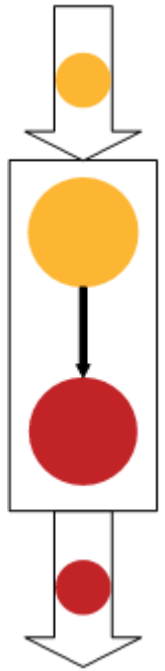
Faulty data in input

Timestamp	Value	MetricId	Instance
1430418780	0	net.multicasttx.summation	vusb0
1430418780	-40	cpu.usage.average	2
1430418780	0	disk.queuelatency.average	naa.5000cca00a2a07b8
1430418780	73	net.transmitted.average	vmnic1
1430418780	20	disk.write.average	
1430418780	6	disk.totallatency.average	naa.5000cca00a2a07b8

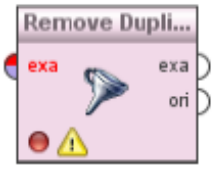
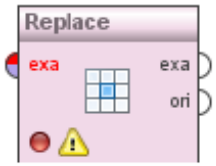
Data transformation rules

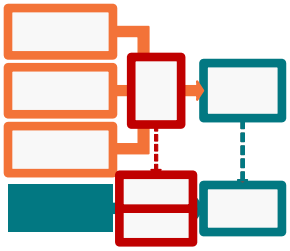


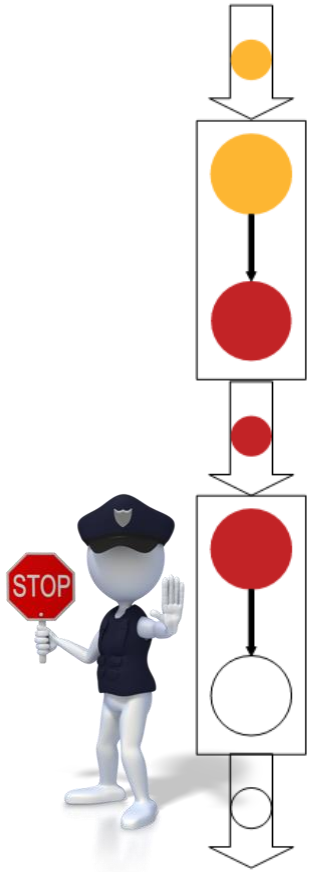
Hibatípus a bemeneten	Operátor	Hibatípus a kimeneten
Elgépelés következtében keletkezett hibák (*Typos*)		Hiányzó sor (*Missing* observation)
Hiányzó adatok (*Missing Value*)		Felesleges sor (*Superfluous* observation)
Értéktartományra vonatkozó kényszer megsértésére (*Min-Max*)		Hibás cellaérték (*Cell value subtle error*)



Prevent data faults

Adathiba	RapidMiner Operátor	
• Duplikátumok (Duplicate)		Remove Duplicates
• Hiányzó adatok (Missing Value)		Replace Missing Values
• Elgépelés következtében keletkezett hibák (Typos) • Helyesírási hibák (Misspellings)		Replace
• Felesleges, ismétlődő oszlopok (Structural feature/Superfluous)		Remove Correlated Attributes
• Értéktartományra vonatkozó kényszer megsértésére (Min-Max) • Attribútumok közötti függőségi kényszer megsértése (ViolatedAttributeDependency)		Filter Examples





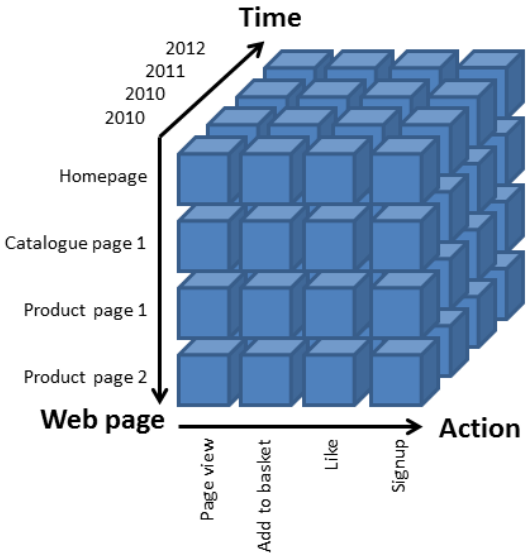
SOME BACKGROUND… OLAP

On-Line Analytical Processing (img: snowplowanalytics.com)

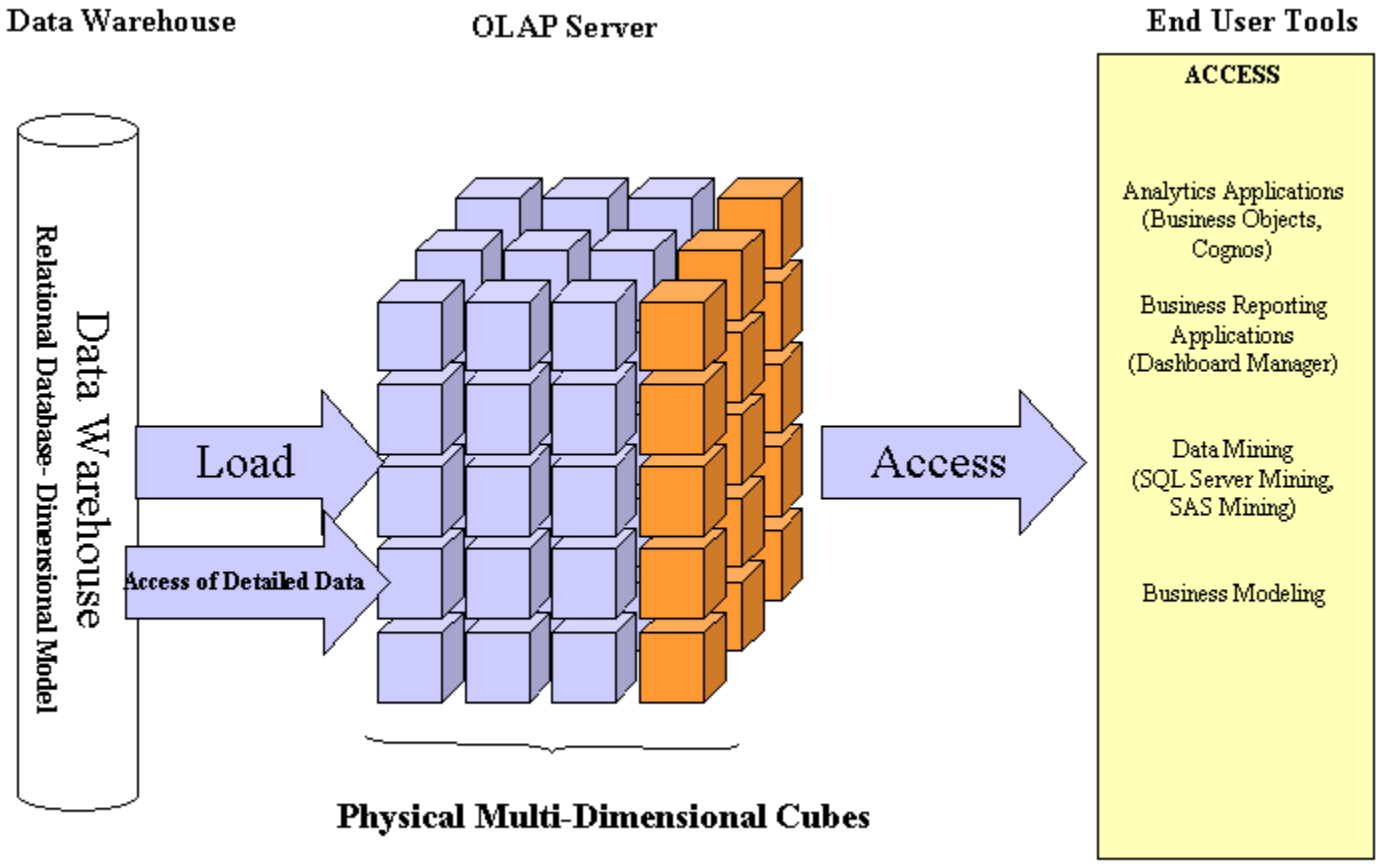
- Business intelligence approach
- Extensively used since early 2000s
  - Still! (although not that popular as it was – at least in academic research)
- Features
  - Multi-dimensional analysis
  - Fast query execution
  - Exploratory analysis of data
    - Support ad-hoc queries
  - Report generation
  - (Visualization)

On-Line Analytical Processing (img: snowplowanalytics.com)

- Central concept: OLAP cube
  - Multi-dimensional array:
    - set of separate data
      - Dimensionality >3
      - → technically a hypercube
      - ~ a multi-dimensional spreadsheet
  - Slicer: dimension held constant
    - For a given query
      (e.g. sales in a particular year)



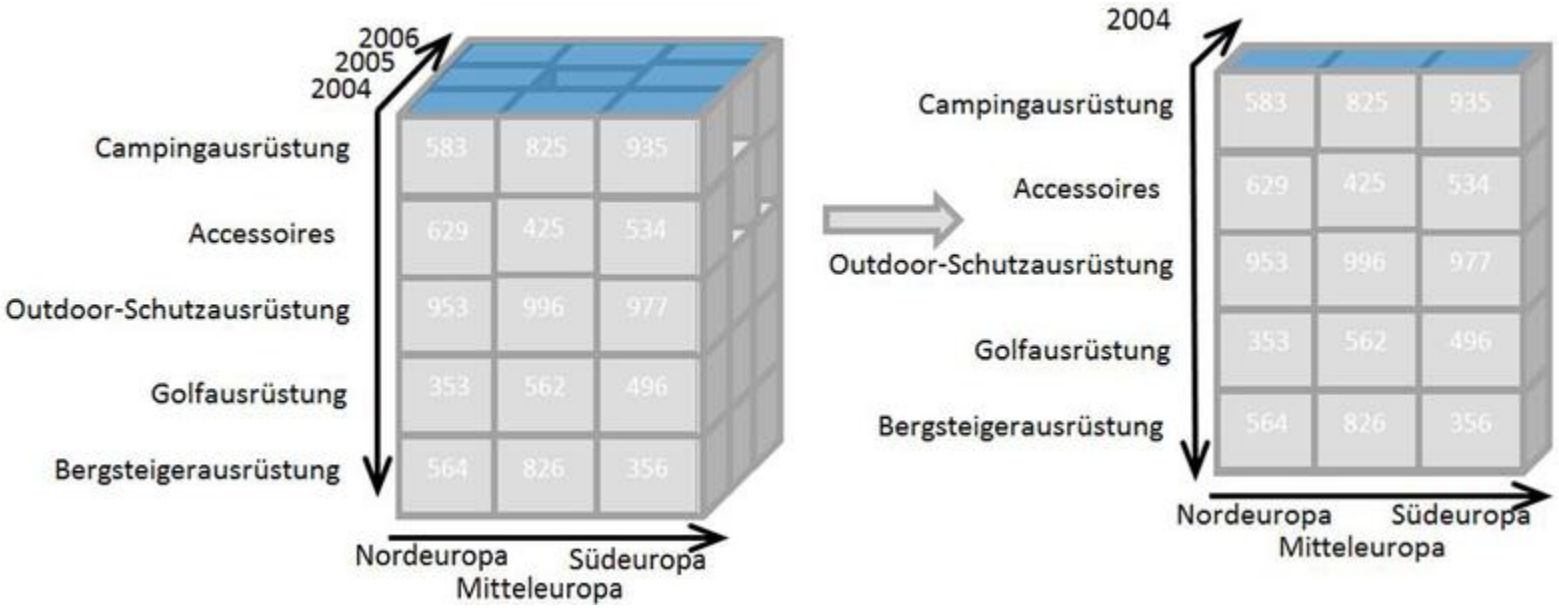
OLAP process (img: Pranav Joshi)



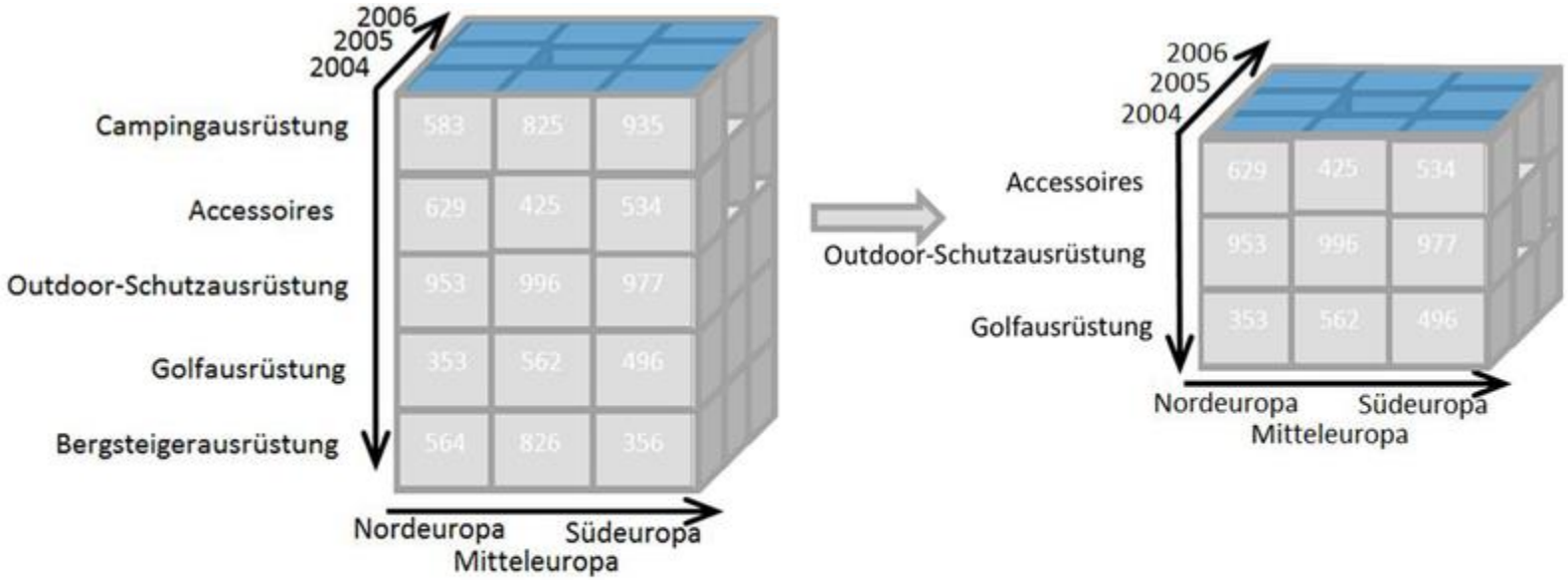
OLAP operations

- Operations
  - Slicing & dicing
  - Drill up & down
  - Pivoting
- Easy to visualize by the cube itself

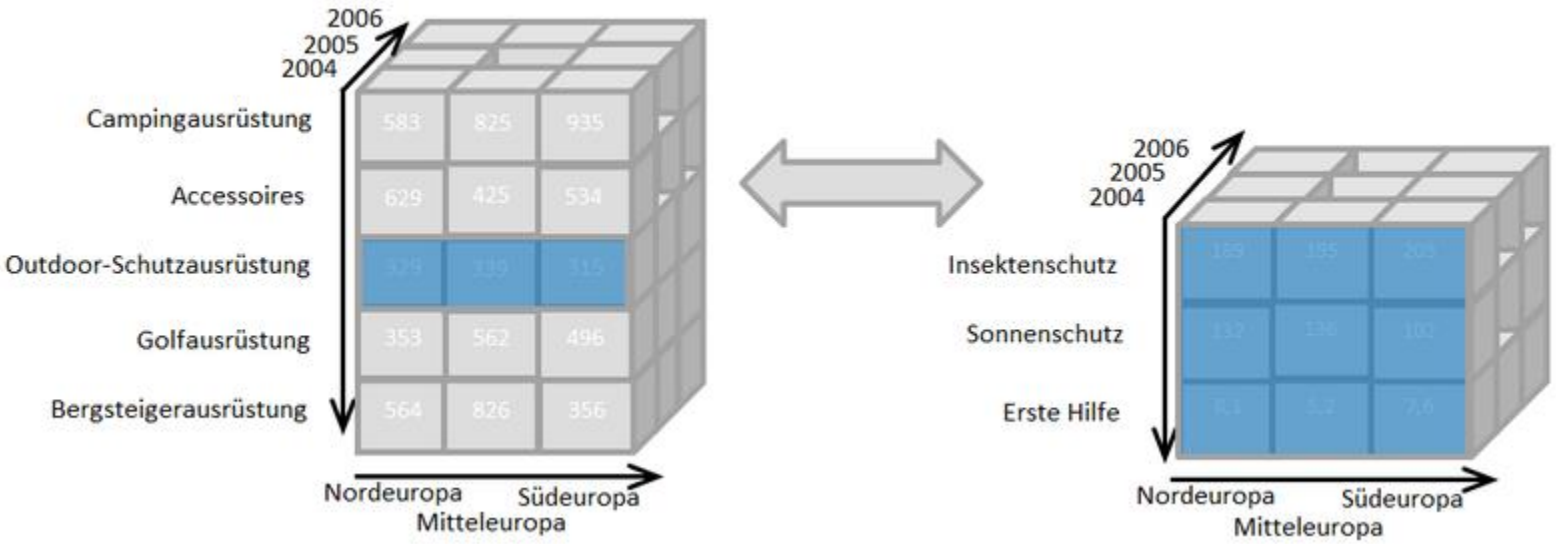
Slicing (img: Wikipedia)



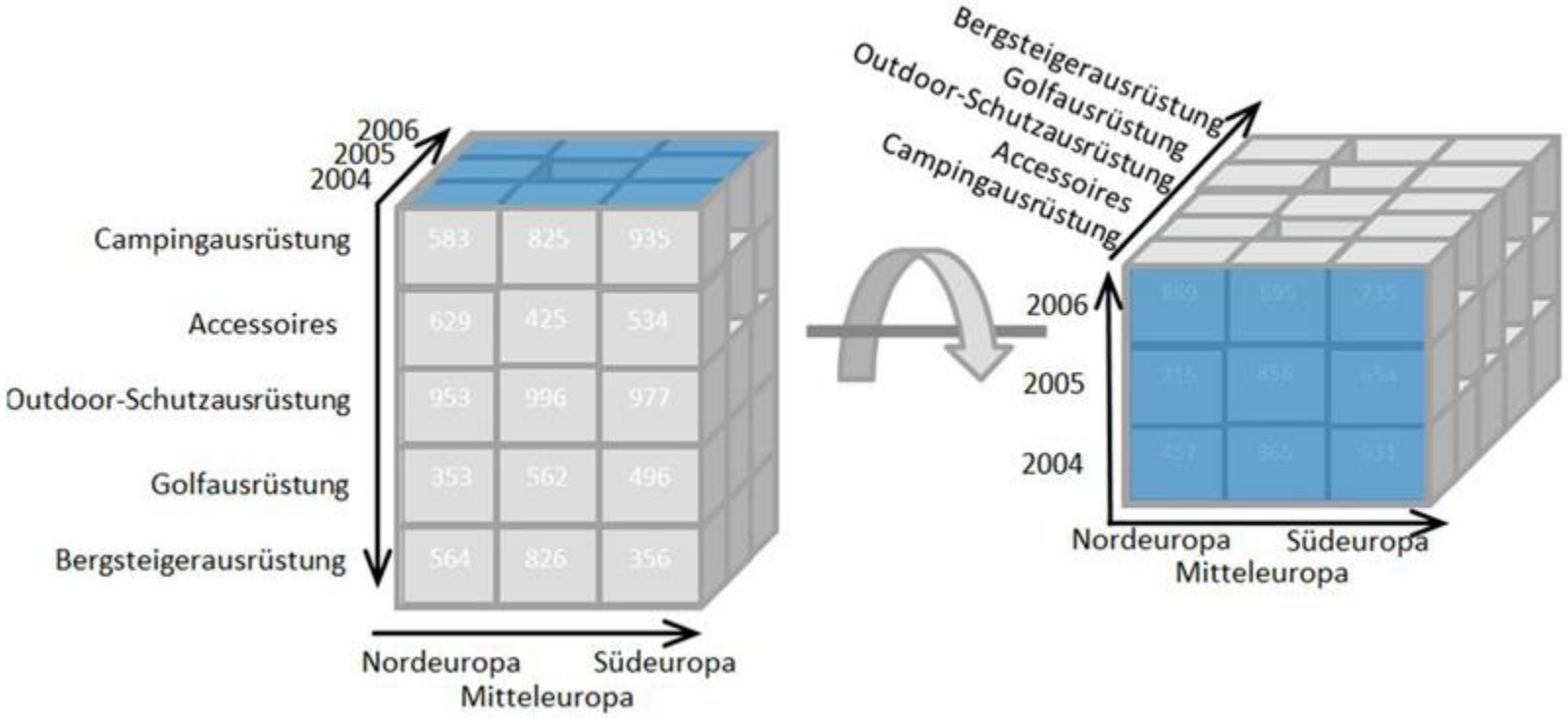
Dicing (img: Wikipedia)



Drill up & down (img: Wikipedia)



Pivoting (img: Wikipedia)



OLAP vs. “regular/modern” data analysis

- OLAP cube: like a set of spreadsheets
  - multi-dimensional
  - interlinked
- Modern data analysis: “flat” data frames
  - Modern machine learning algorithms:
    - require (?) single dataframes
- Operations: basically the same (slicing, dicing, drill up & down, pivoting)

VISUALIZATION: PLOT TYPES

Reminder: Tabular Representation

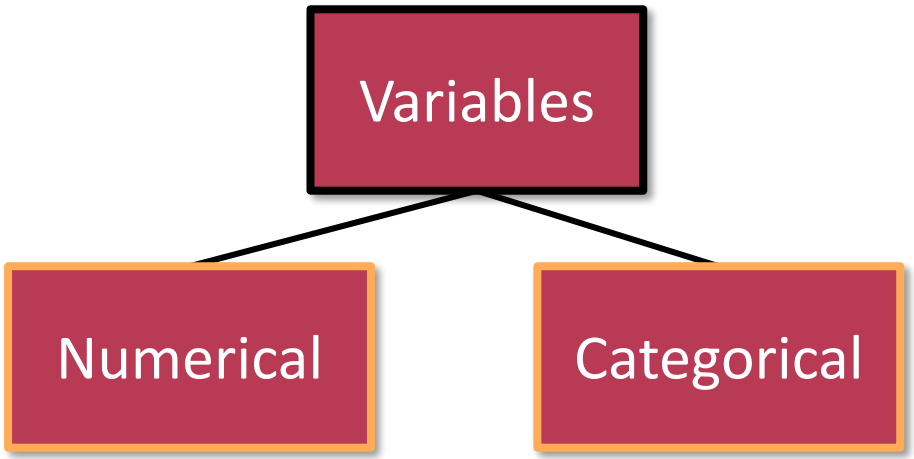
- **Rows of the table** = Model elements
- **Columns of the table** = Properties

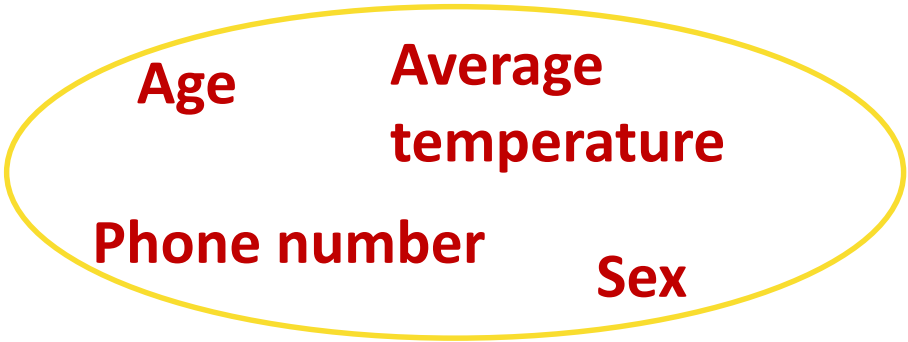
Name	Type	Size (kB)	Last modified
Documents	directory		2016.02.02
Contracts.pdf	file	569	2015.11.09
Pictures	directory		2016.02.02
Logo.png	file	92	2015.03.06
Groundplot.jpg	file	1226	2016.02.02

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

Numerical and Categorical Variables

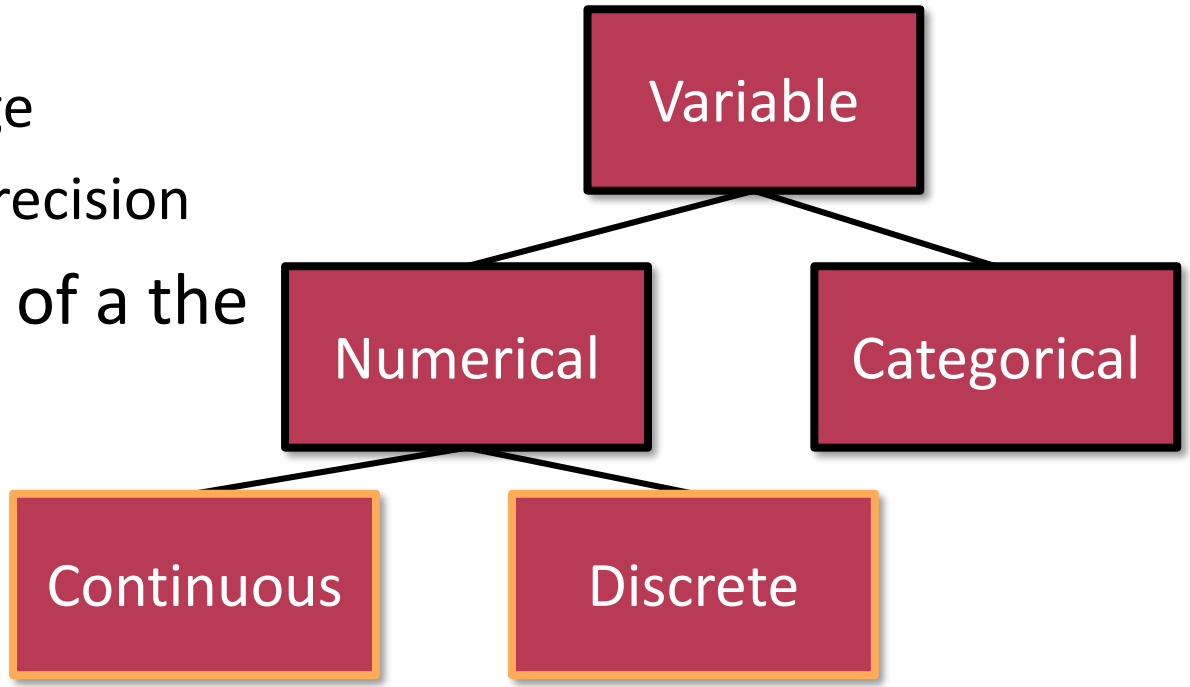
- Numerical
  - Arithmetic operations are interpreted meaningful (average, sum, inc, dec, …)
- Categorical
  - No operation between the values





Numerical Variables

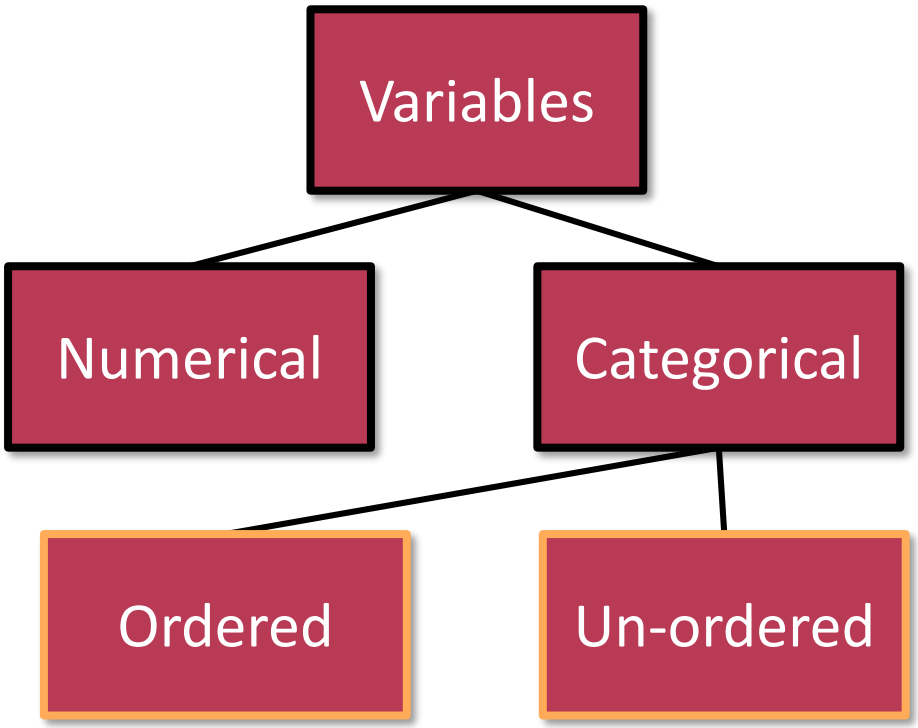
- Continuous
  - Measured
    - in a specific range
    - with a specific precision
  - e.g. temperature of a the server room



- 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

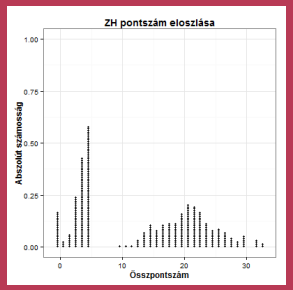
Single variable

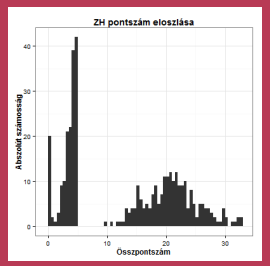
Variables

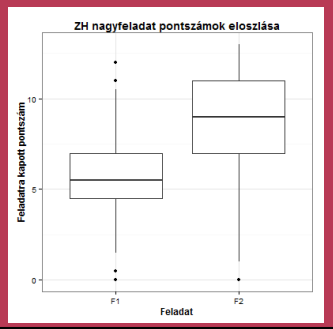
Numerical

Categorical

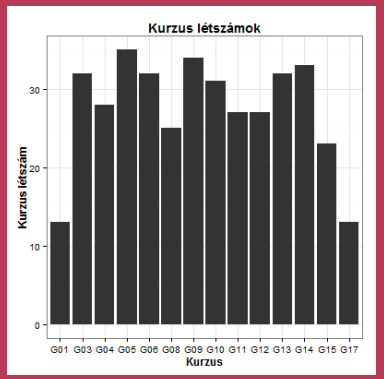
Test results: [28, 28, 30, …]







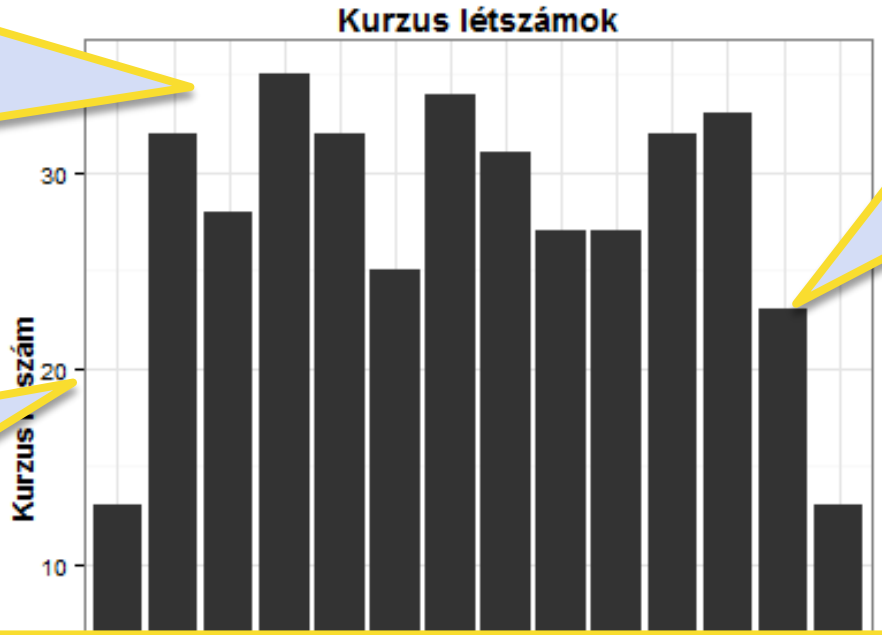
Training groups: [G01, G02, G03, …]



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?

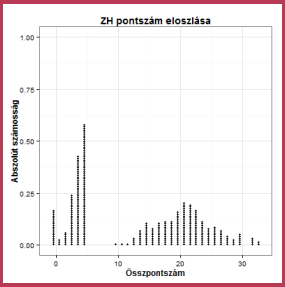
Single variable

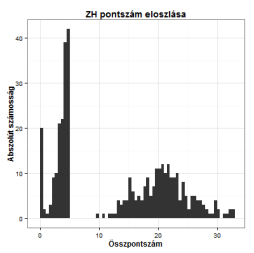
Variables

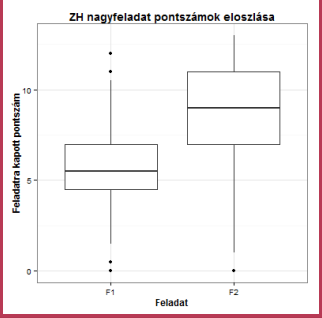
Numerical

Categorical

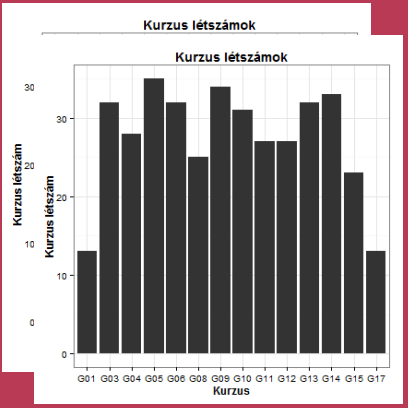
Test results: [28, 28, 30, …]





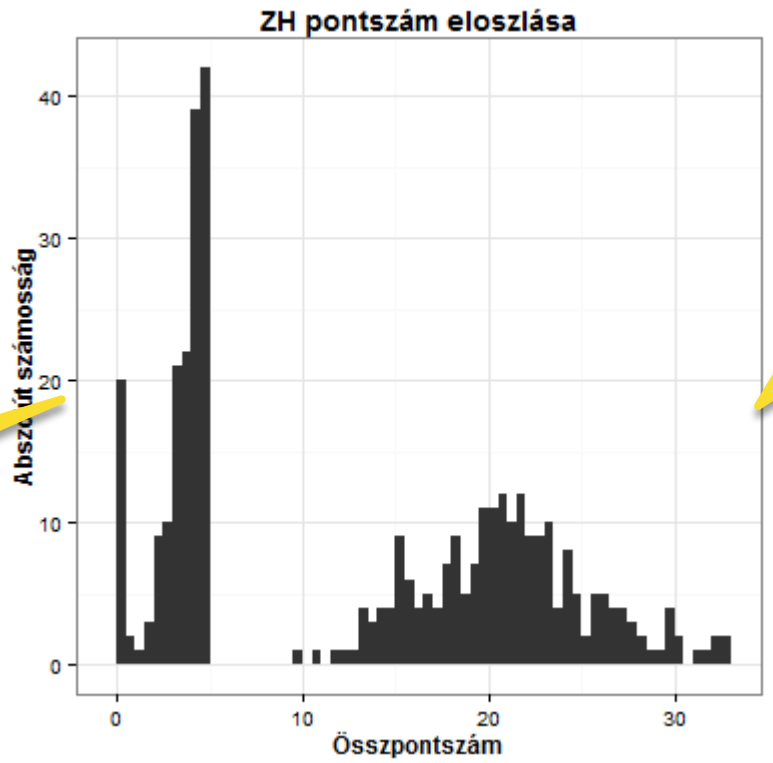


Training groups: [G01, G02, G03, …]



Histogram

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



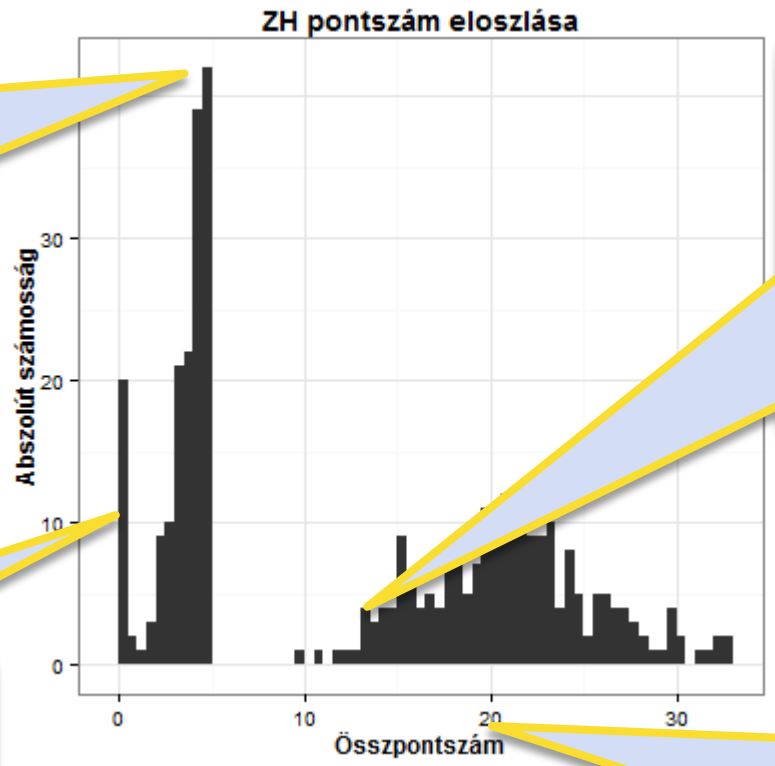
Height of bar : Frequency of the given **interval** of values

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.

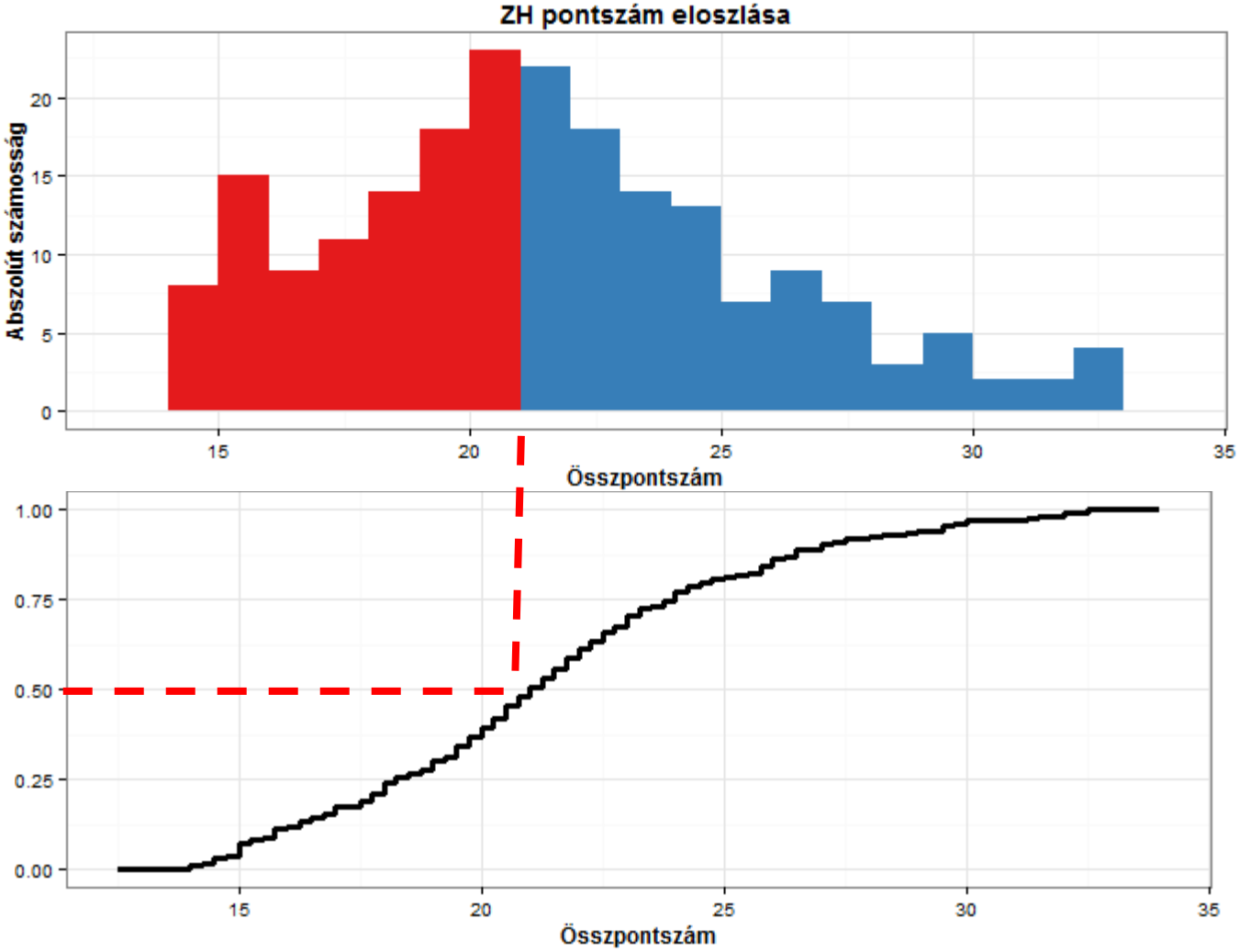
The most of those who made the entry test, made the whole test also.

Those not appeared at all.

The average/ median was at 20 points.

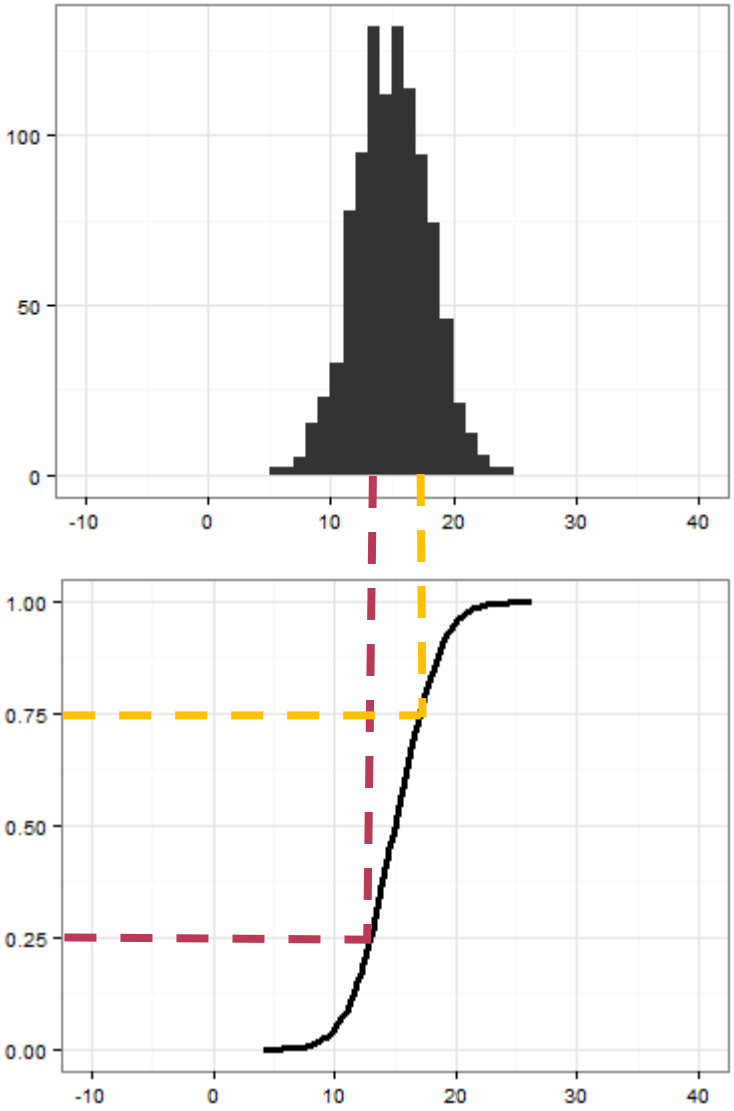
Simple Statistical Description

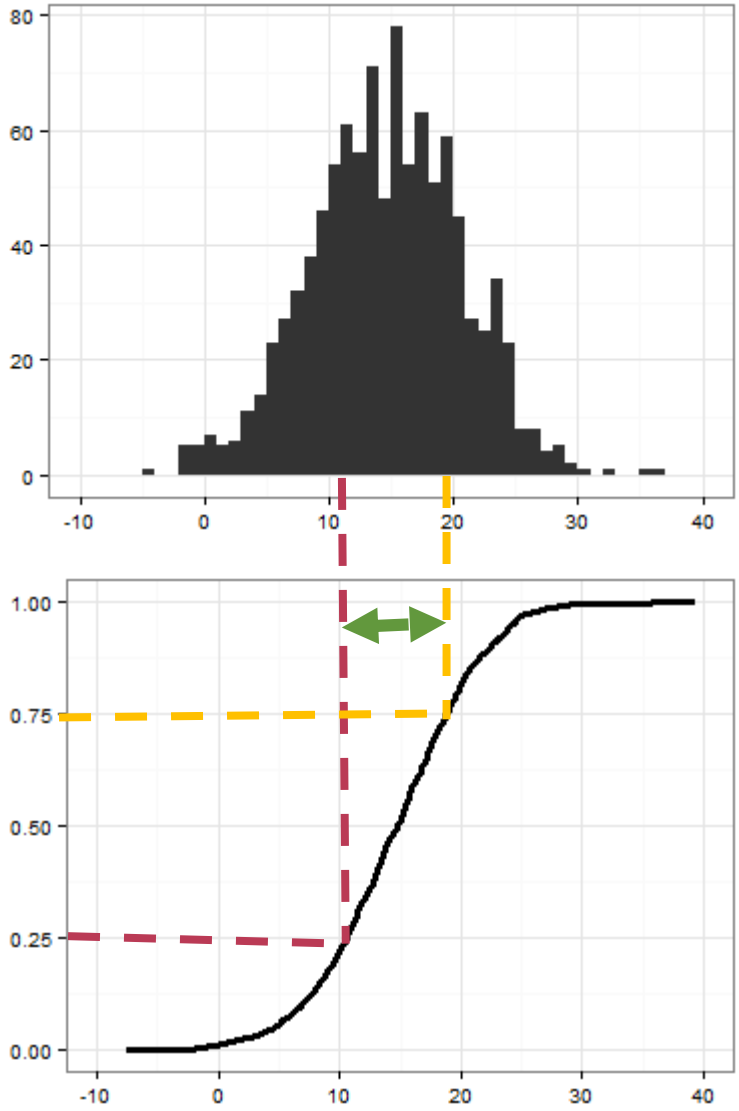
- Where is „the middle” of the values?



Simple Statistical Description

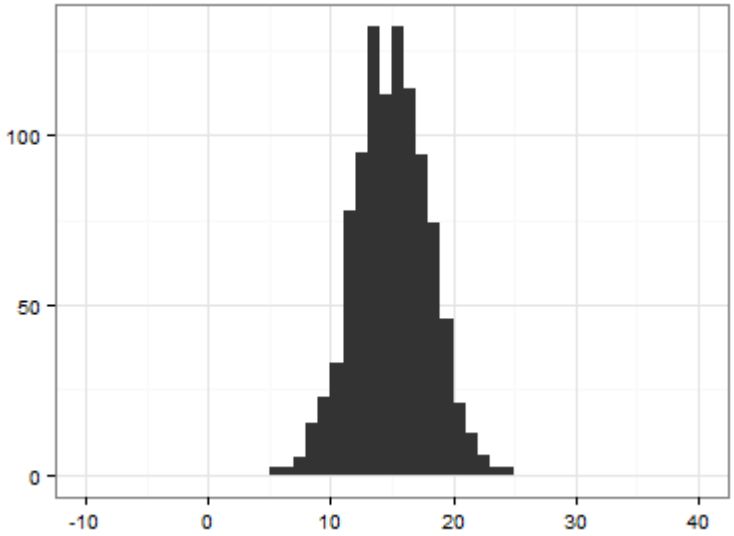
- How far are the values „scattered”?

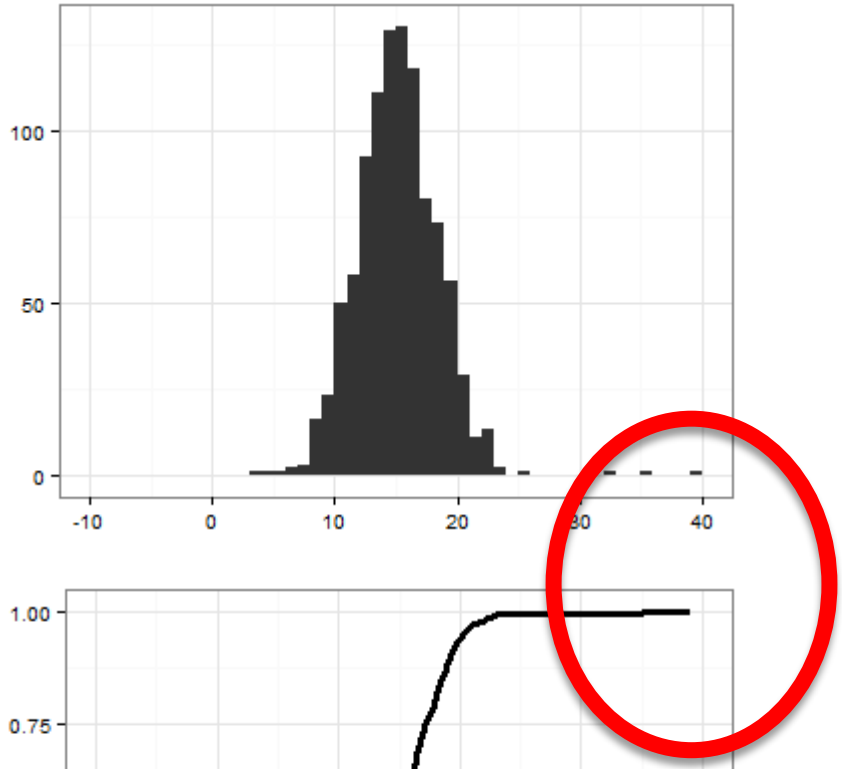


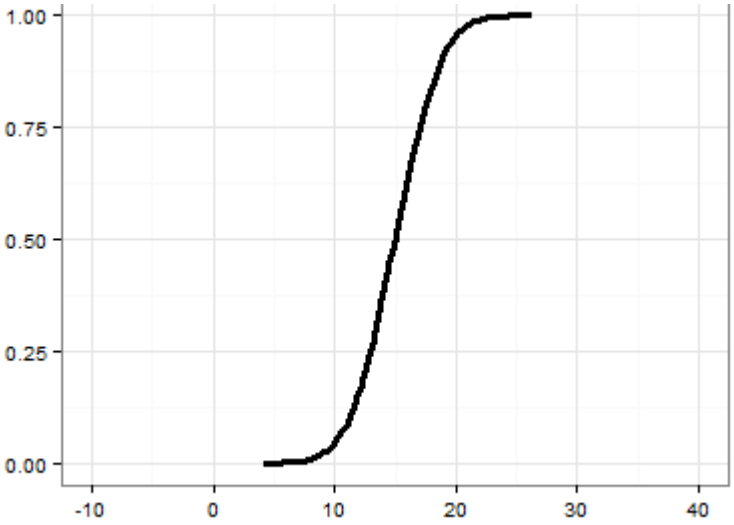


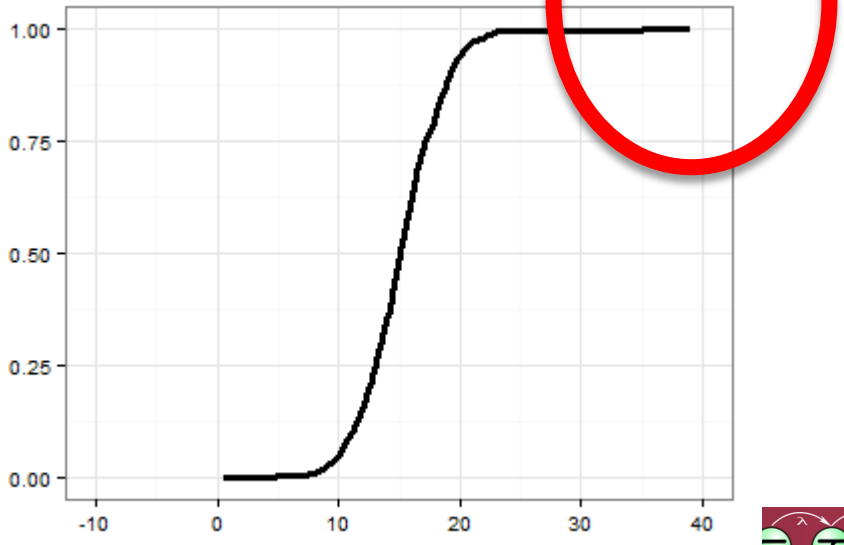
Simple Statistical Description

- Are there outliers?



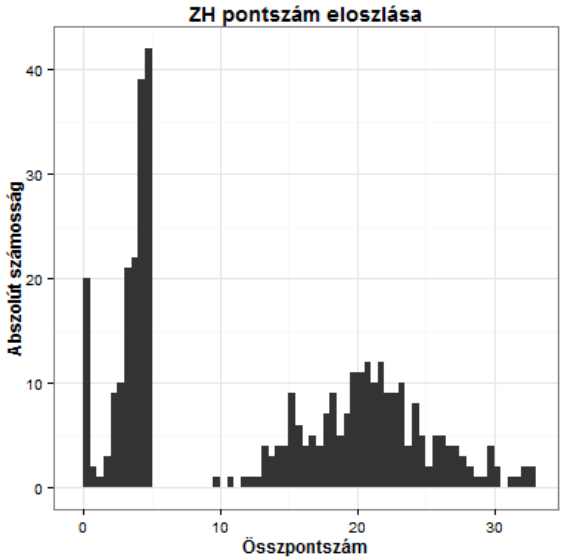






Box plots

- 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)

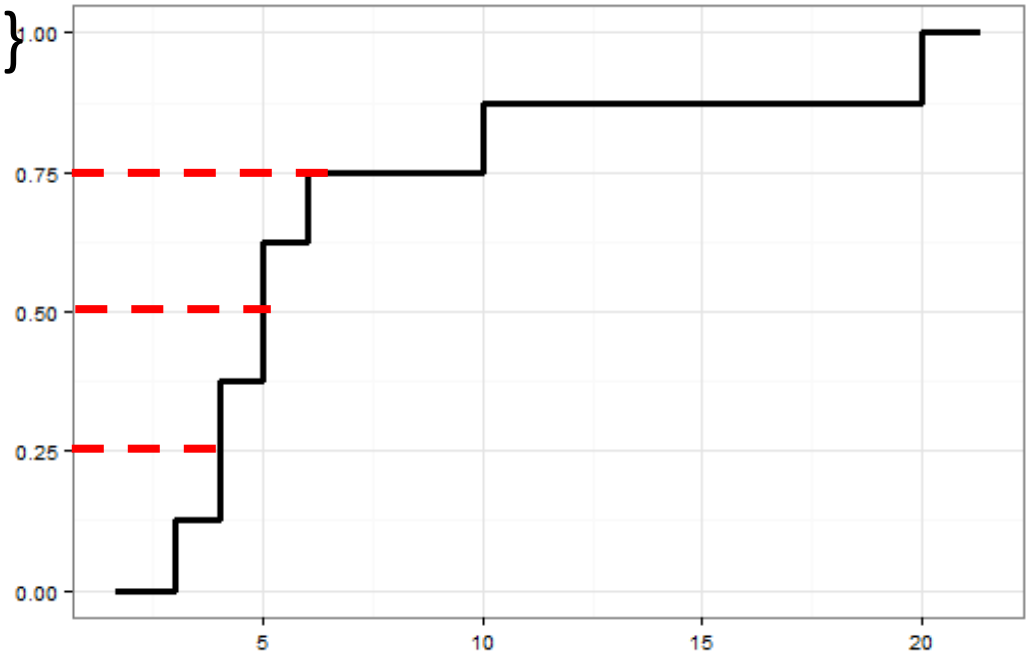
Describing (cont.) Observations

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.

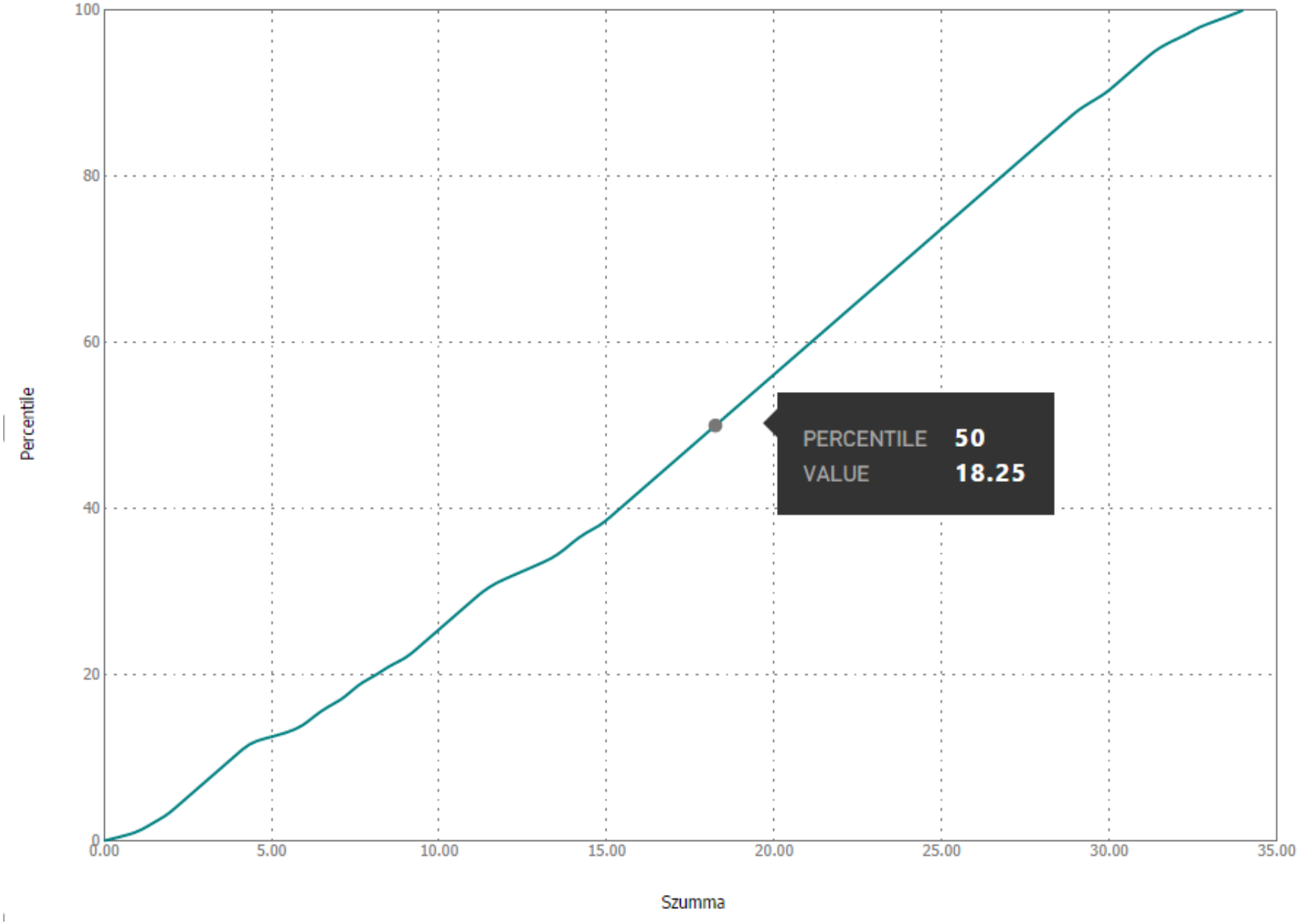
The mode is the most frequent element (the most frequent elements) of the data set. If there is no unique *most frequent element*, the data set has multiple Modes.

Percentiles

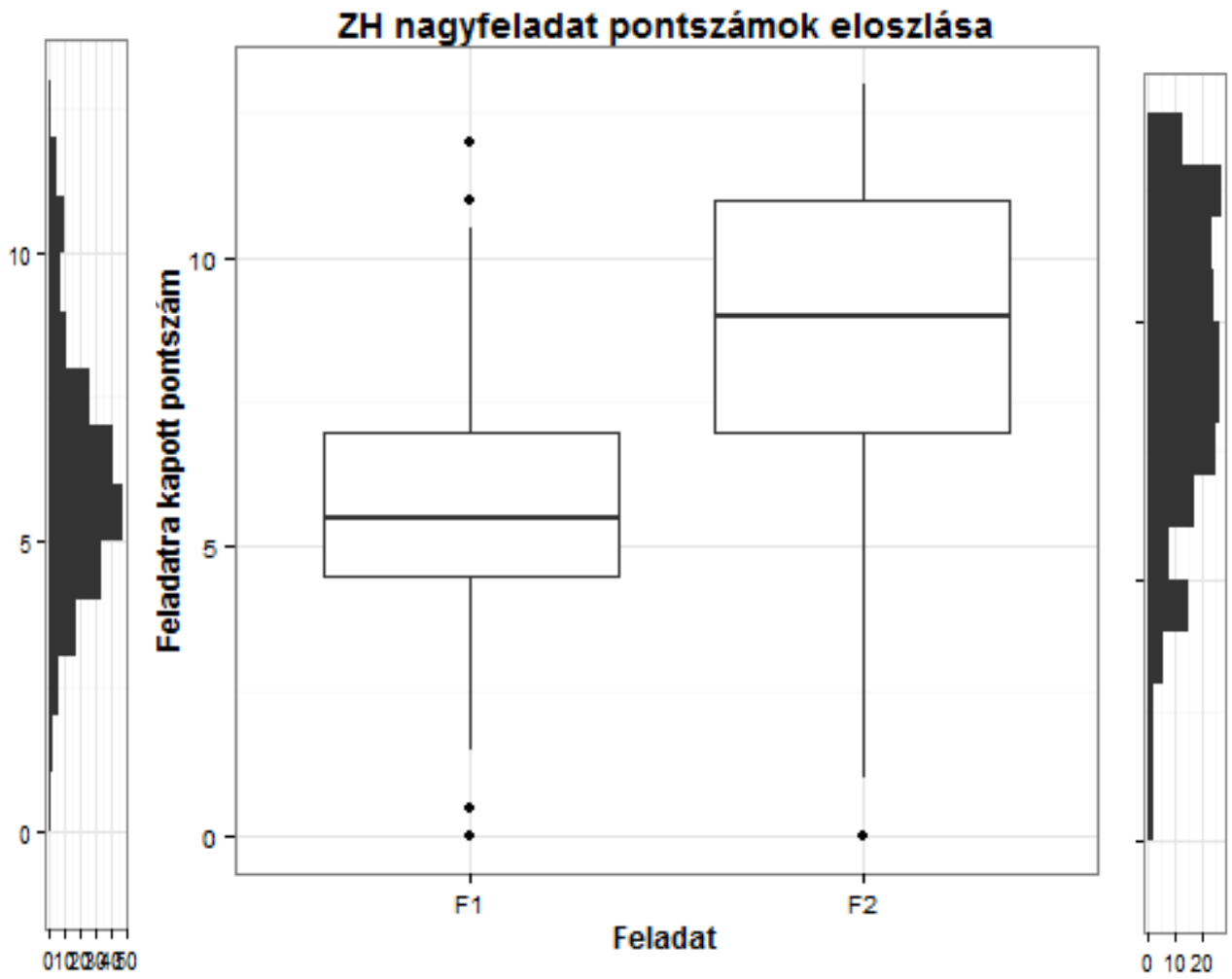
- Percentile
  - $n\%$ of the values are weaker than the n^{th} **percentile**
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - 50. percentile: 5
    - 25. percentile: 4
    - 75. percentile: 10
- Quartiles
  - Q1: 25. percentile
  - Q3: 75. percentile
  - **Q2: Median**



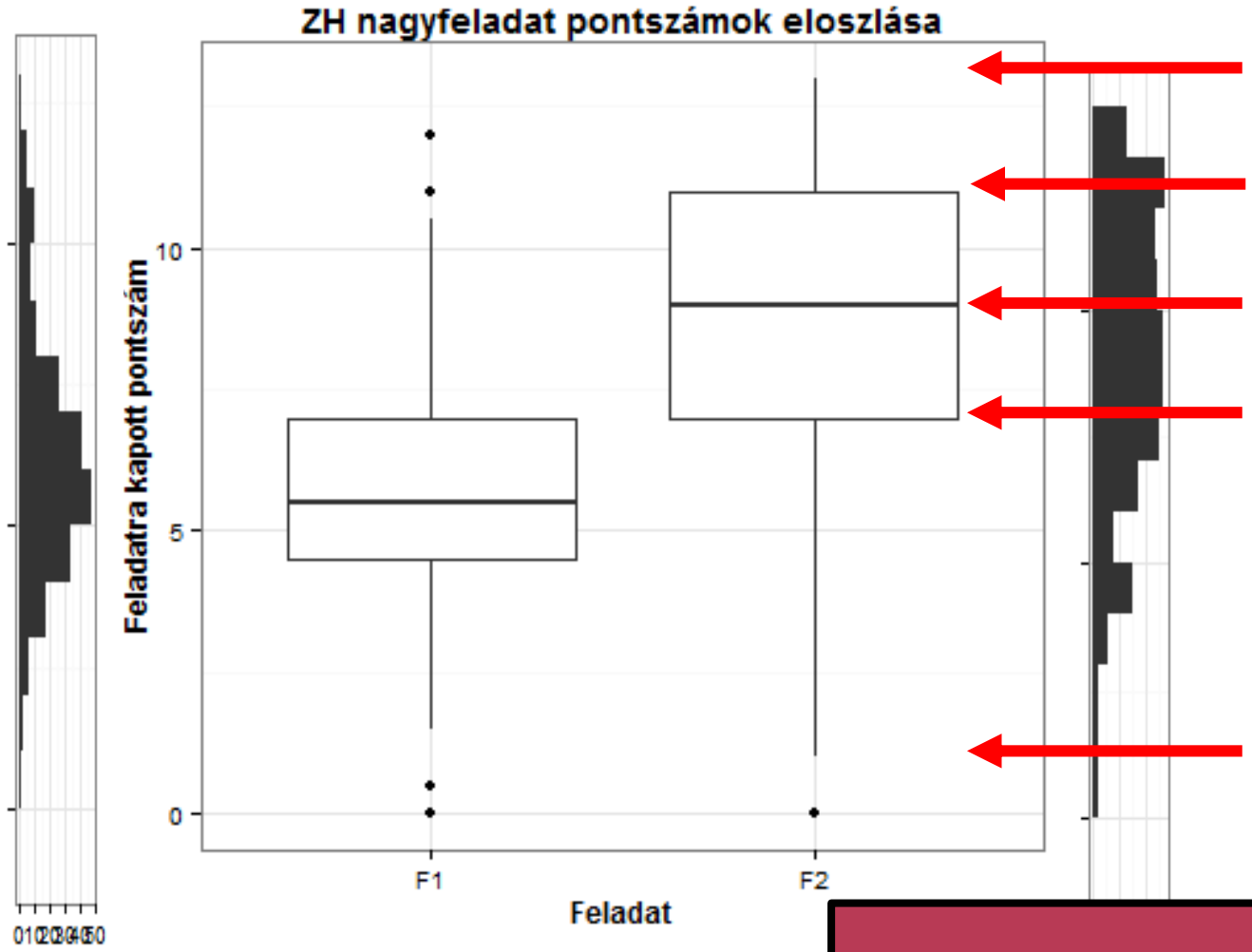
Example: Representation of Percentiles



Box and whisker plot



Box and whisker plot



Limit of the upper outliers

Q3

Median

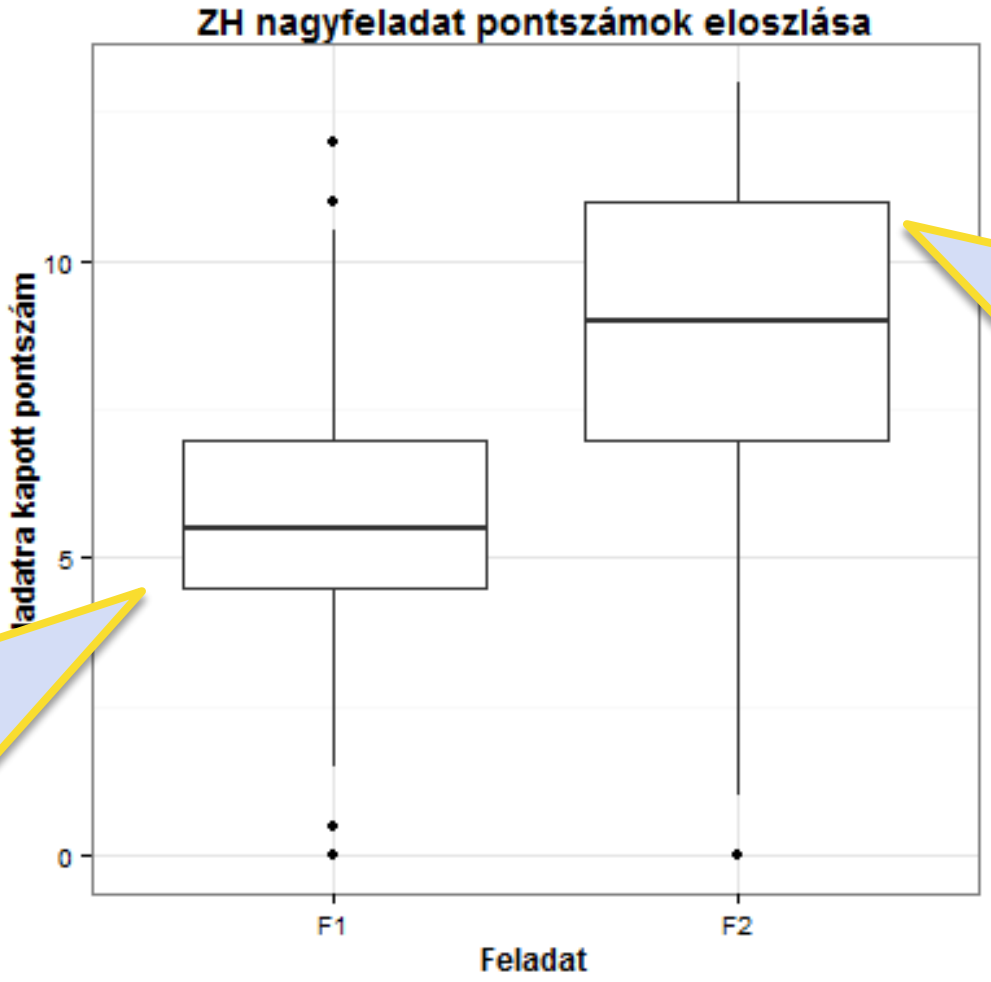
Q1

Limit of the lower outliers

How to do it in Excel?

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

Box and whisker plot

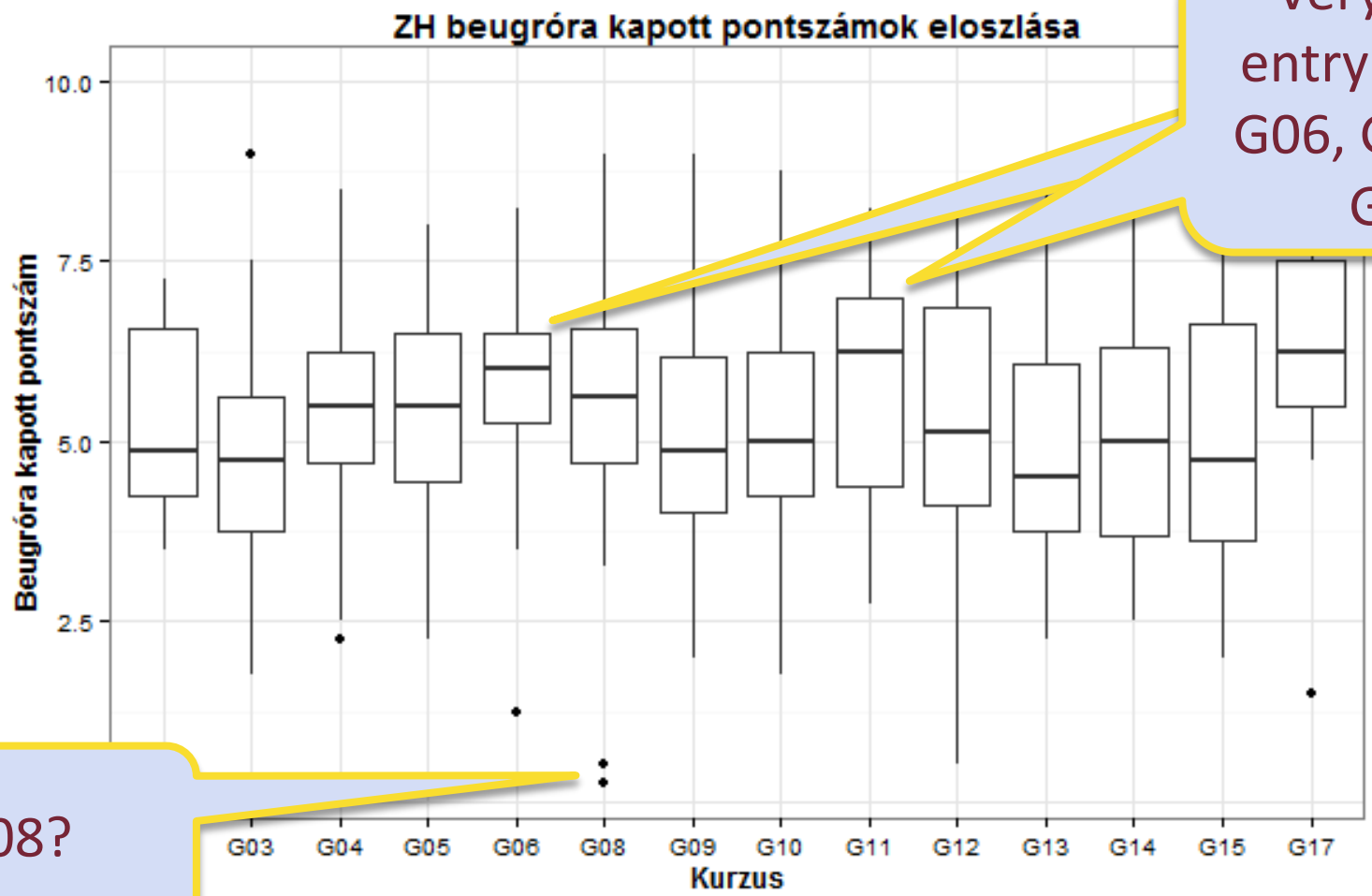


Task2 has produced (in average) better results than Task1.

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

Box and whisker plot

- How were the results per training groups?

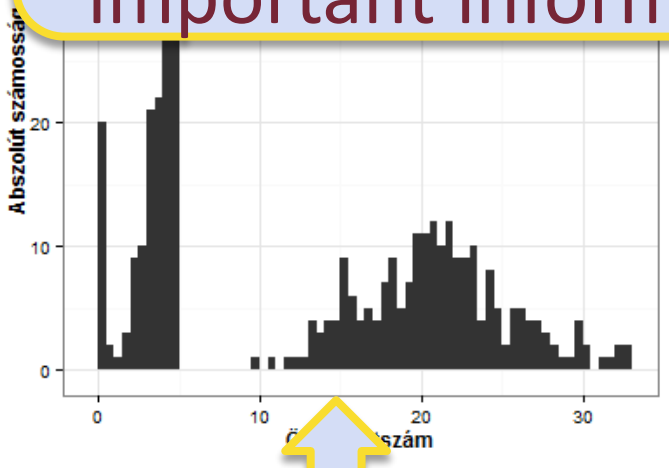


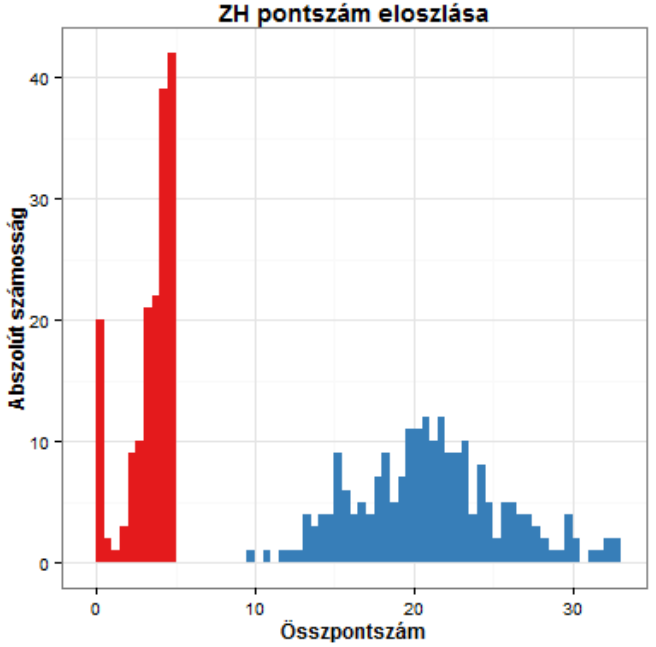
Very good entry tests in G06, G11 and G17

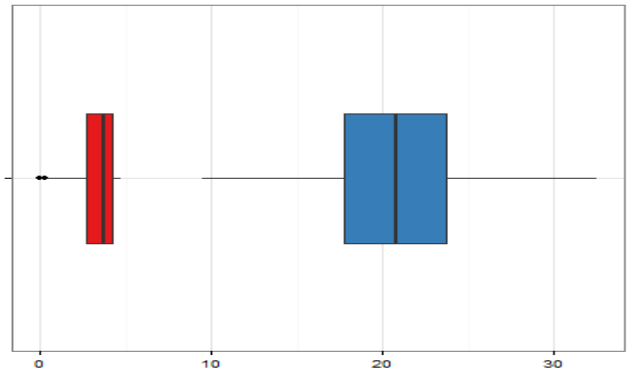
G08?

Box and whisker plot

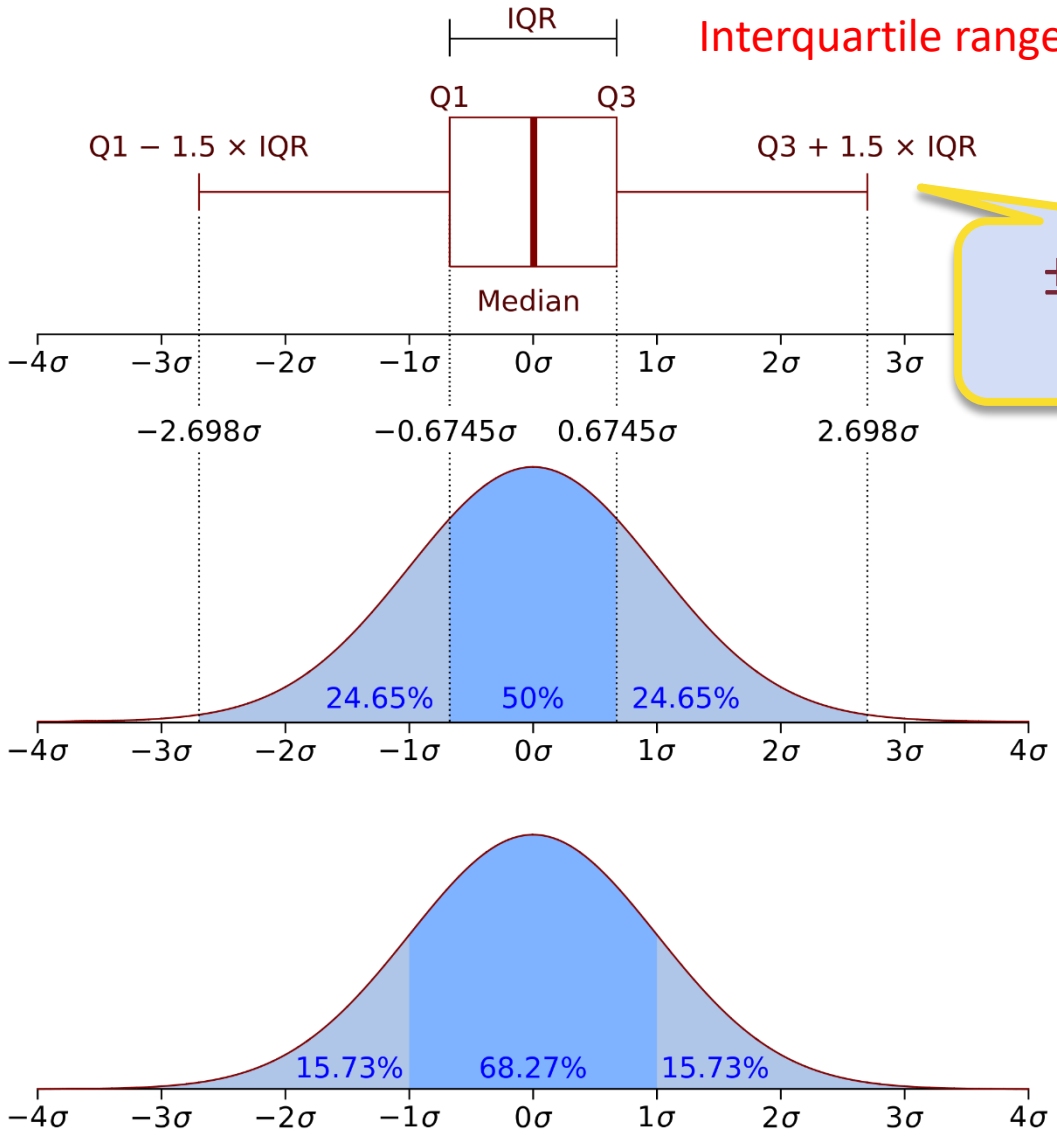
Abstraction: With box plot we can miss important information.



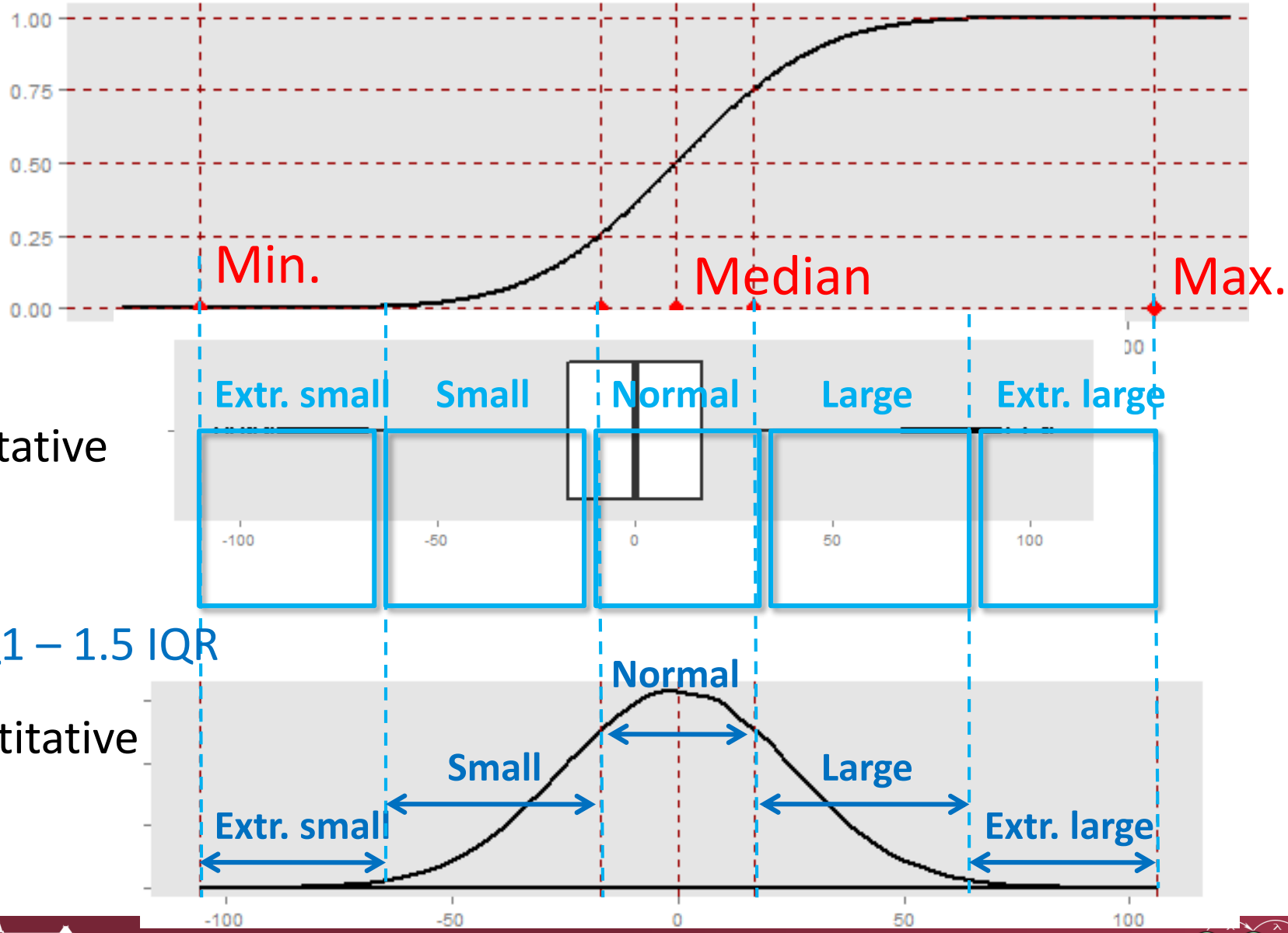




Boxplot (Box and whisker plot)

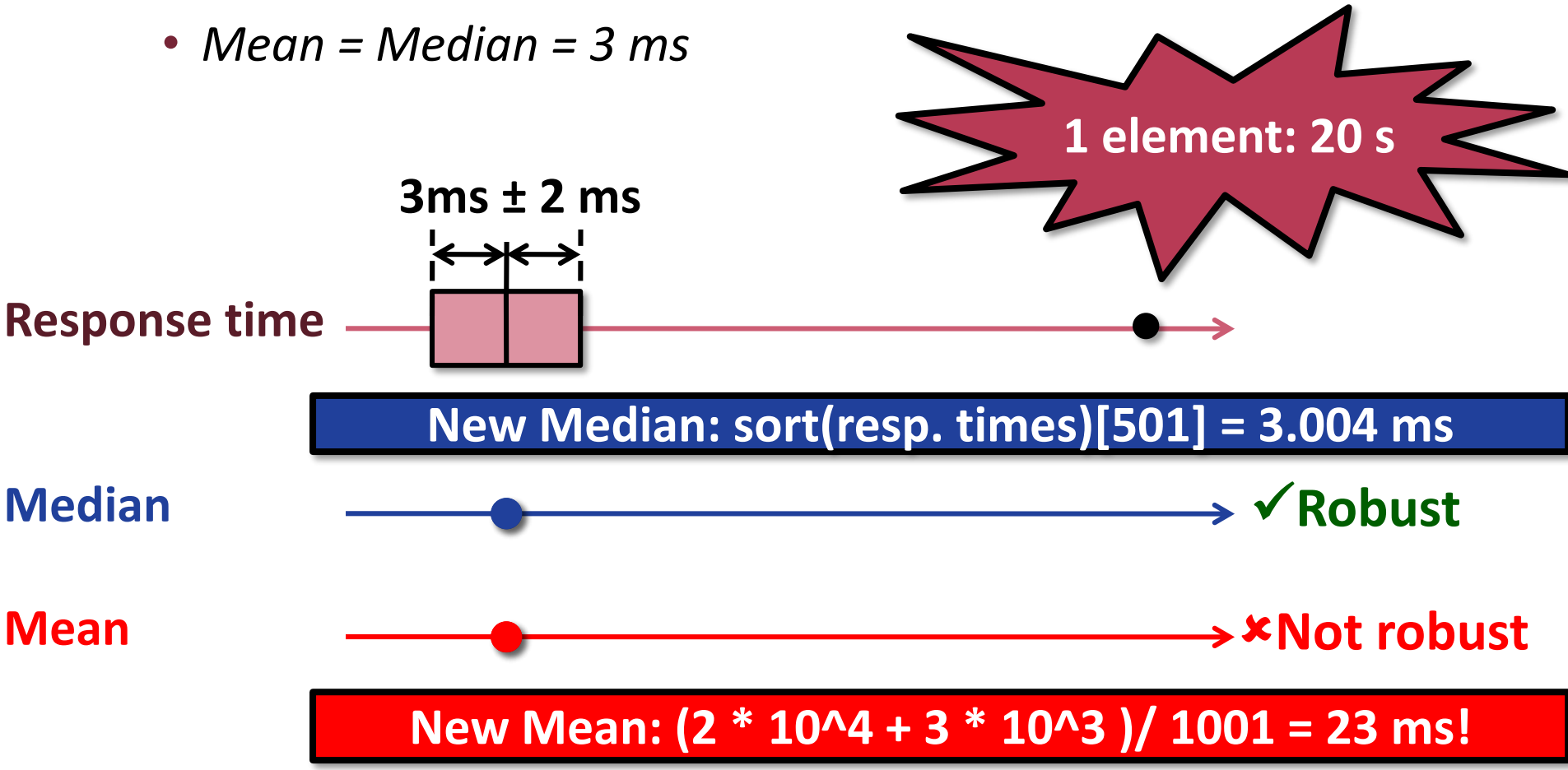


Boxplots vs qualitative domains

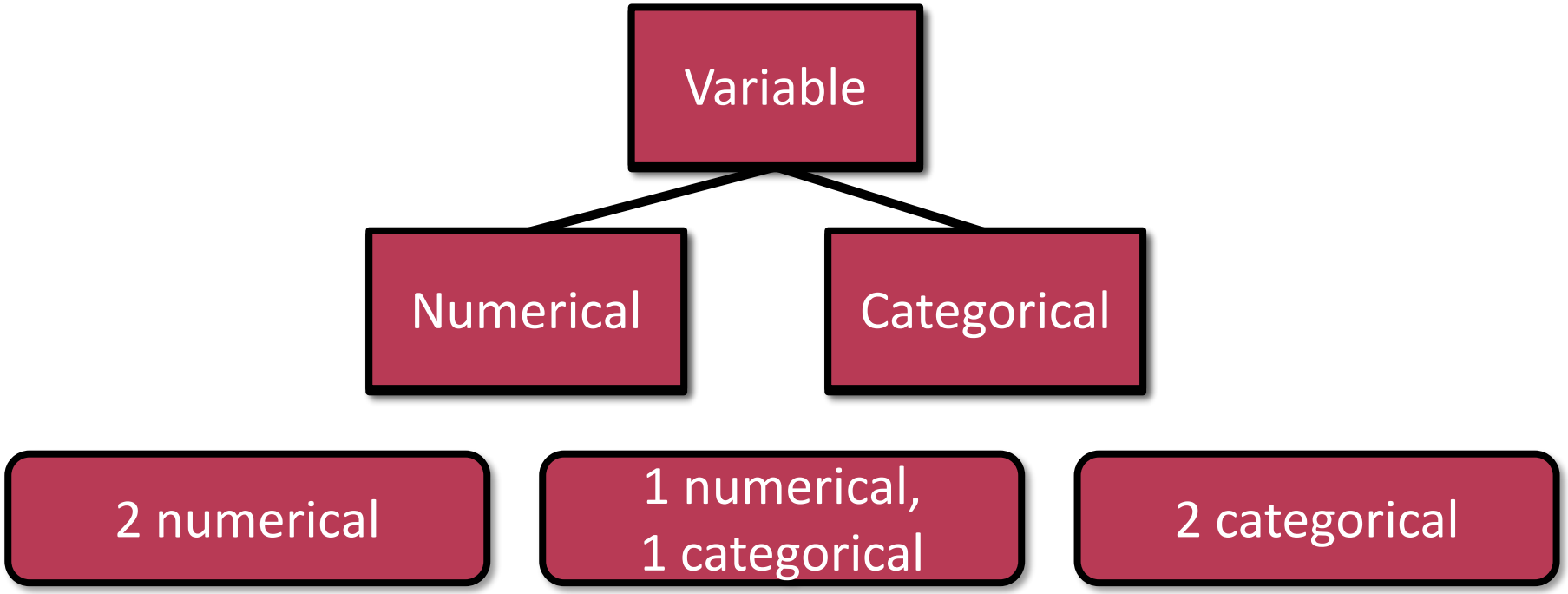


Median instead of Mean – Why?

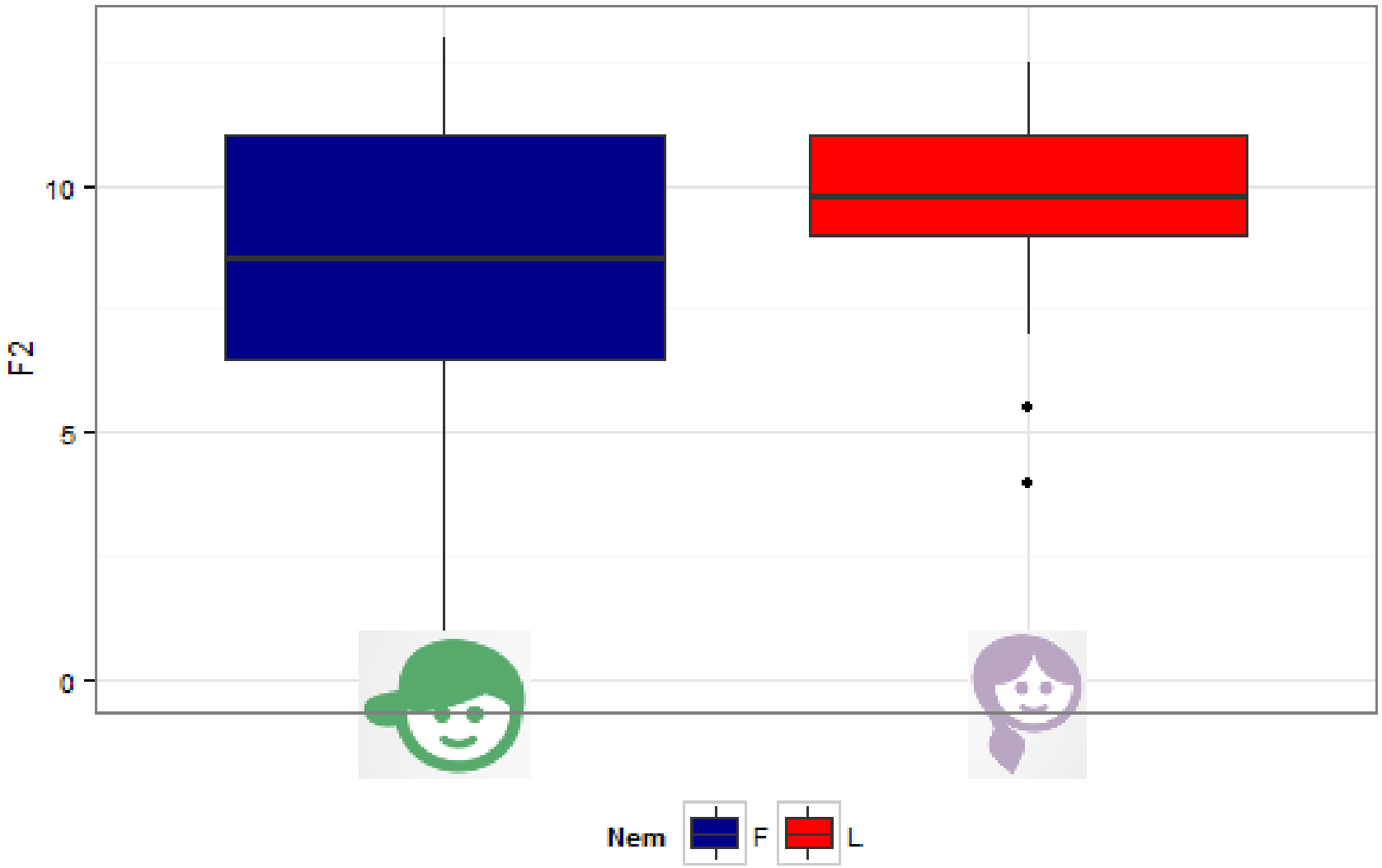
- Data set: 1000 Points in (1, 5) with uniform distribution
  - *Mean = Median = 3 ms*



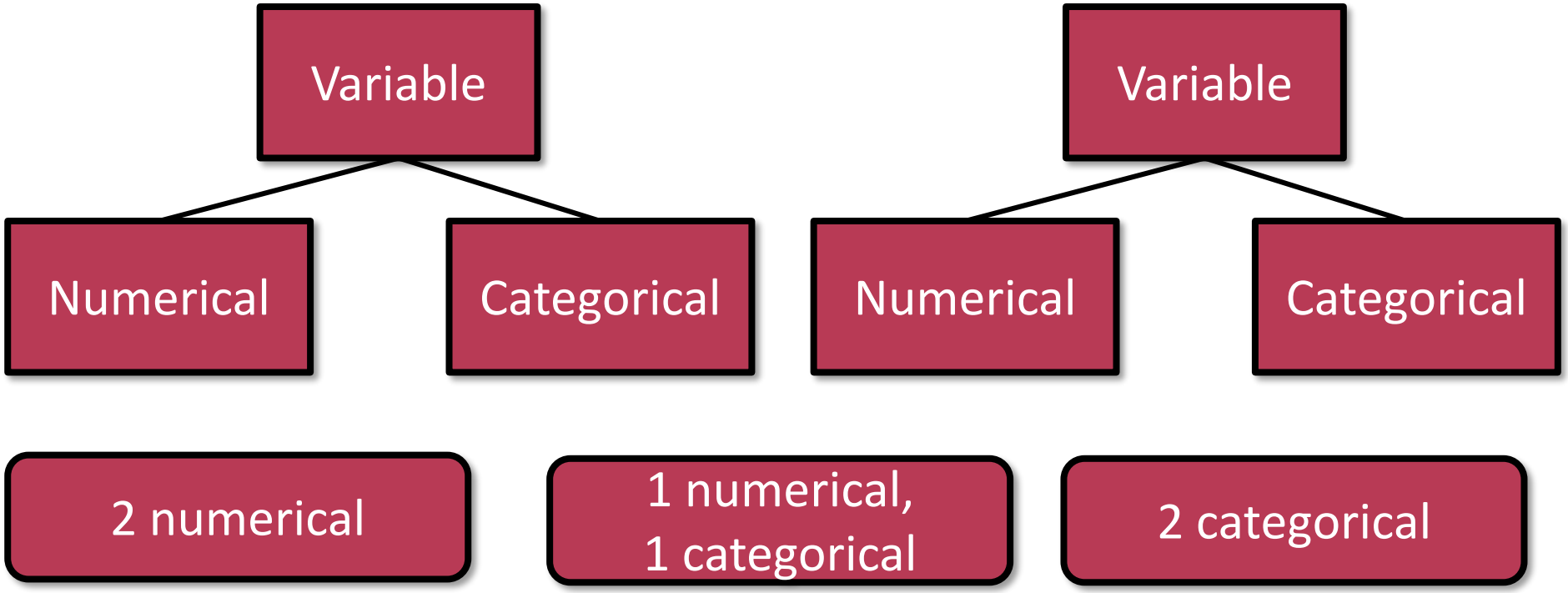
Relation between two Variables



Numerical, per Category



Relation between two Variables

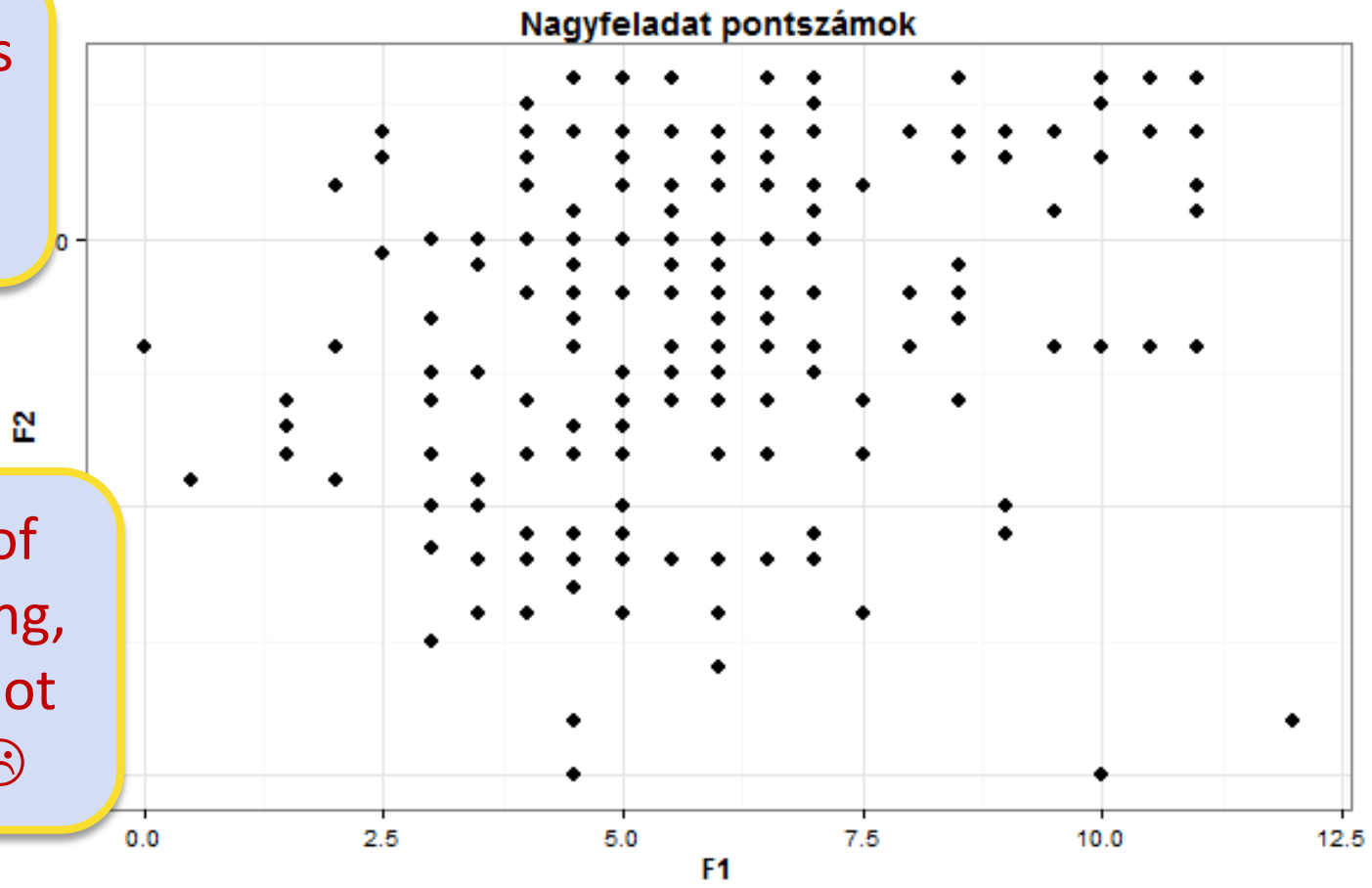


Scatter Plot

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

Pairs of results are displayed (visualised).

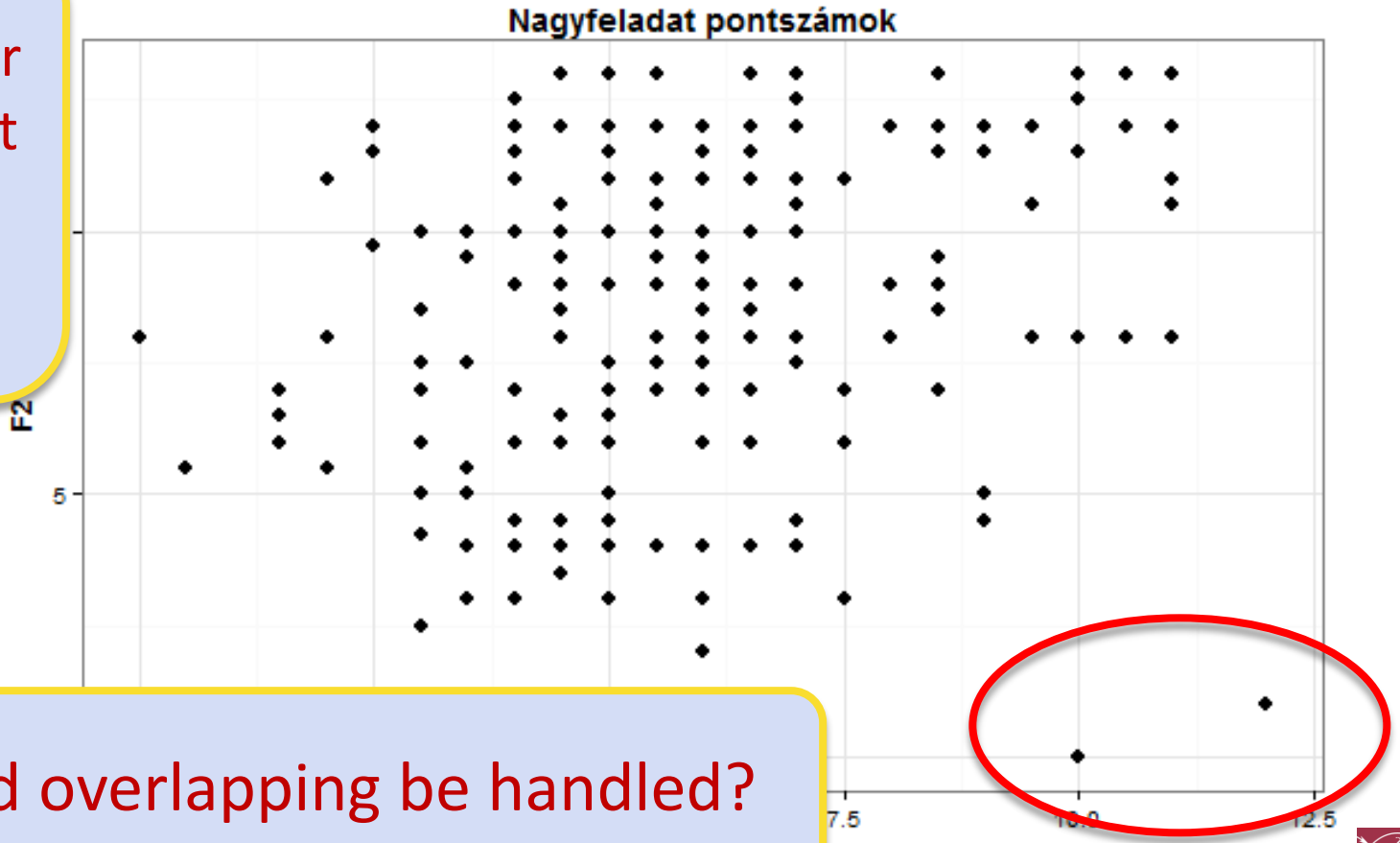
Where one of them is missing, the pair cannot be shown. ☹



Scatter Plot

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

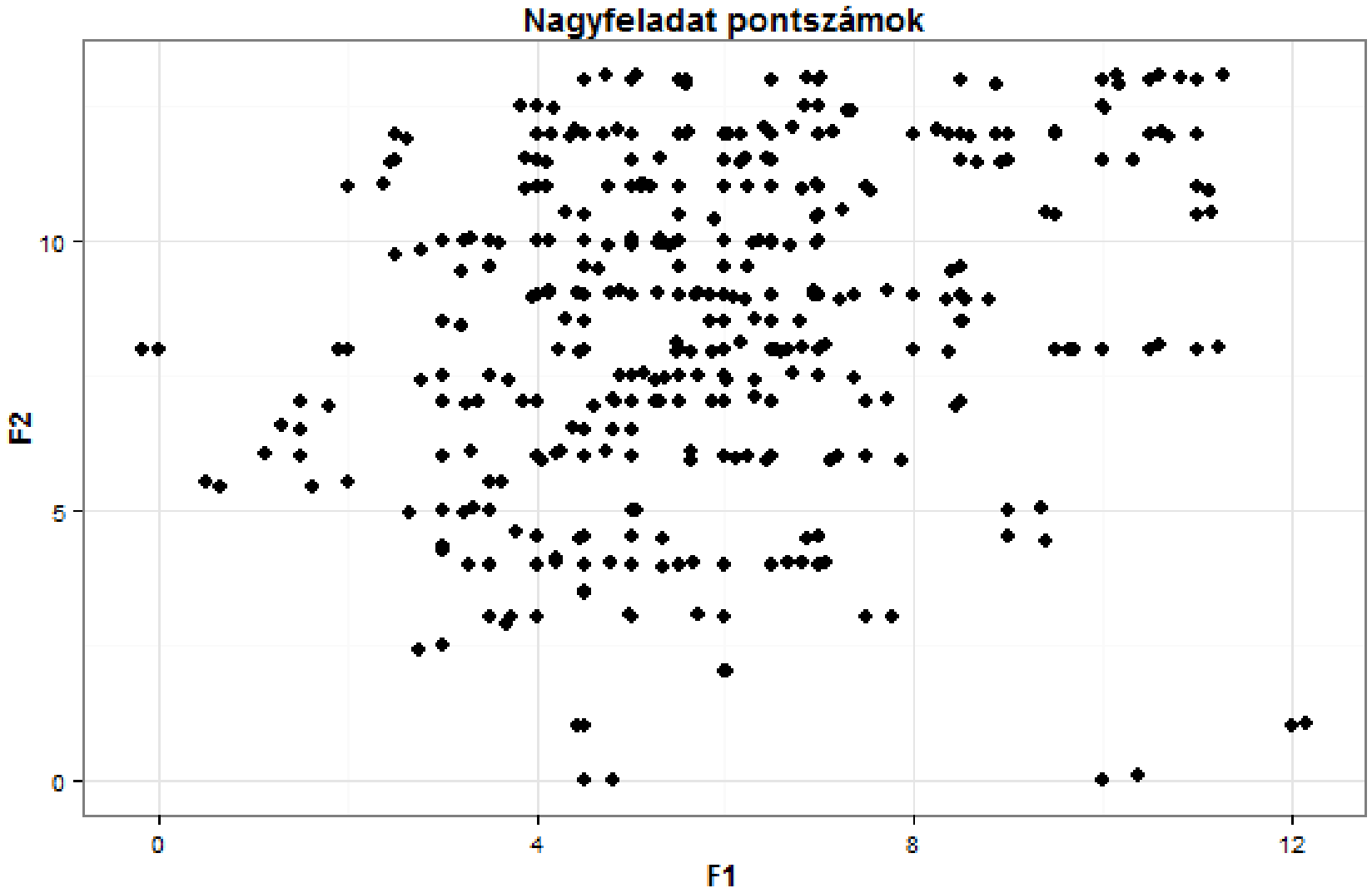
Who had a good result for Task1, had not necessarily a good one for Task2.



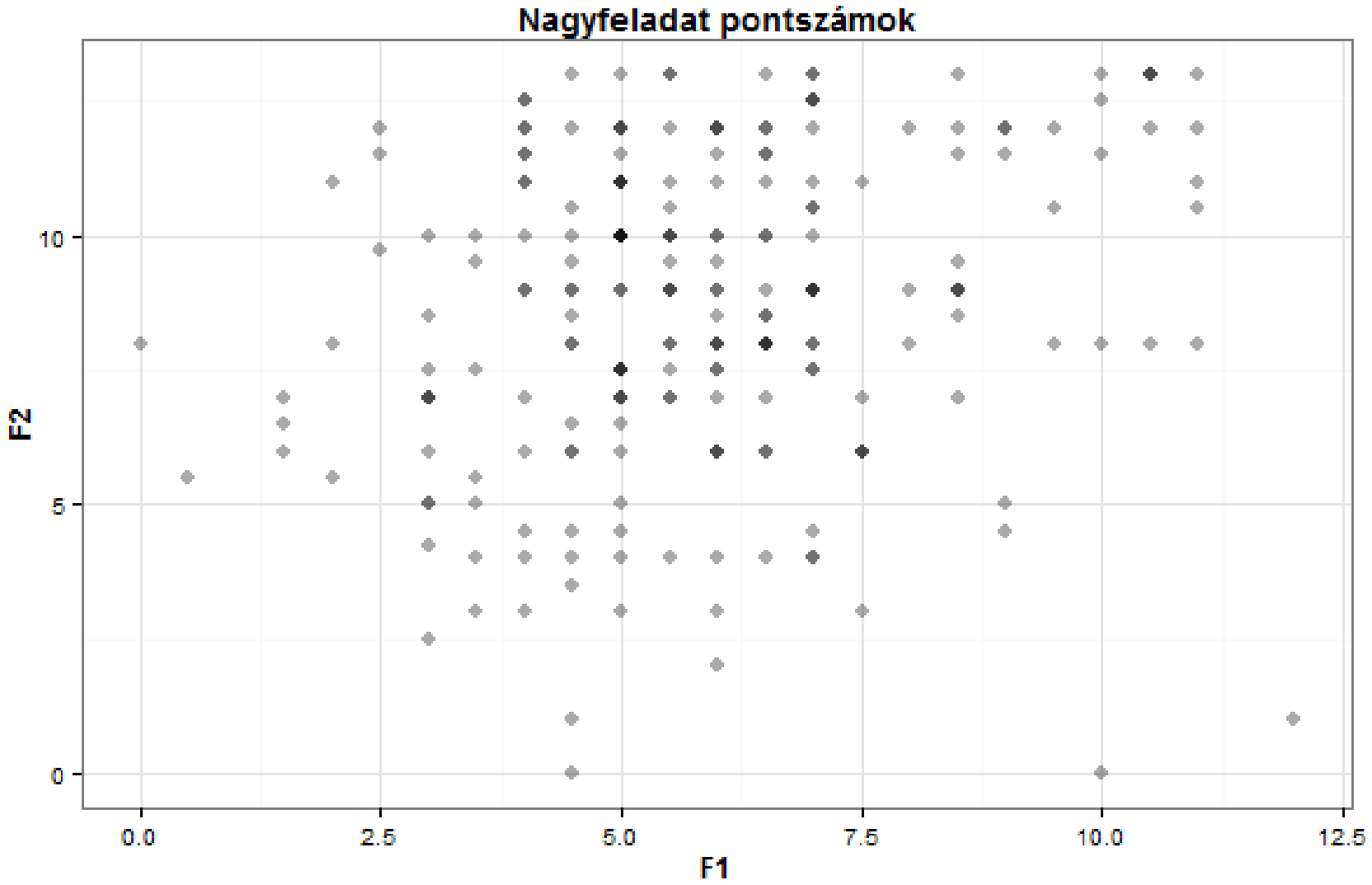
How should overlapping be handled?

Overplotting

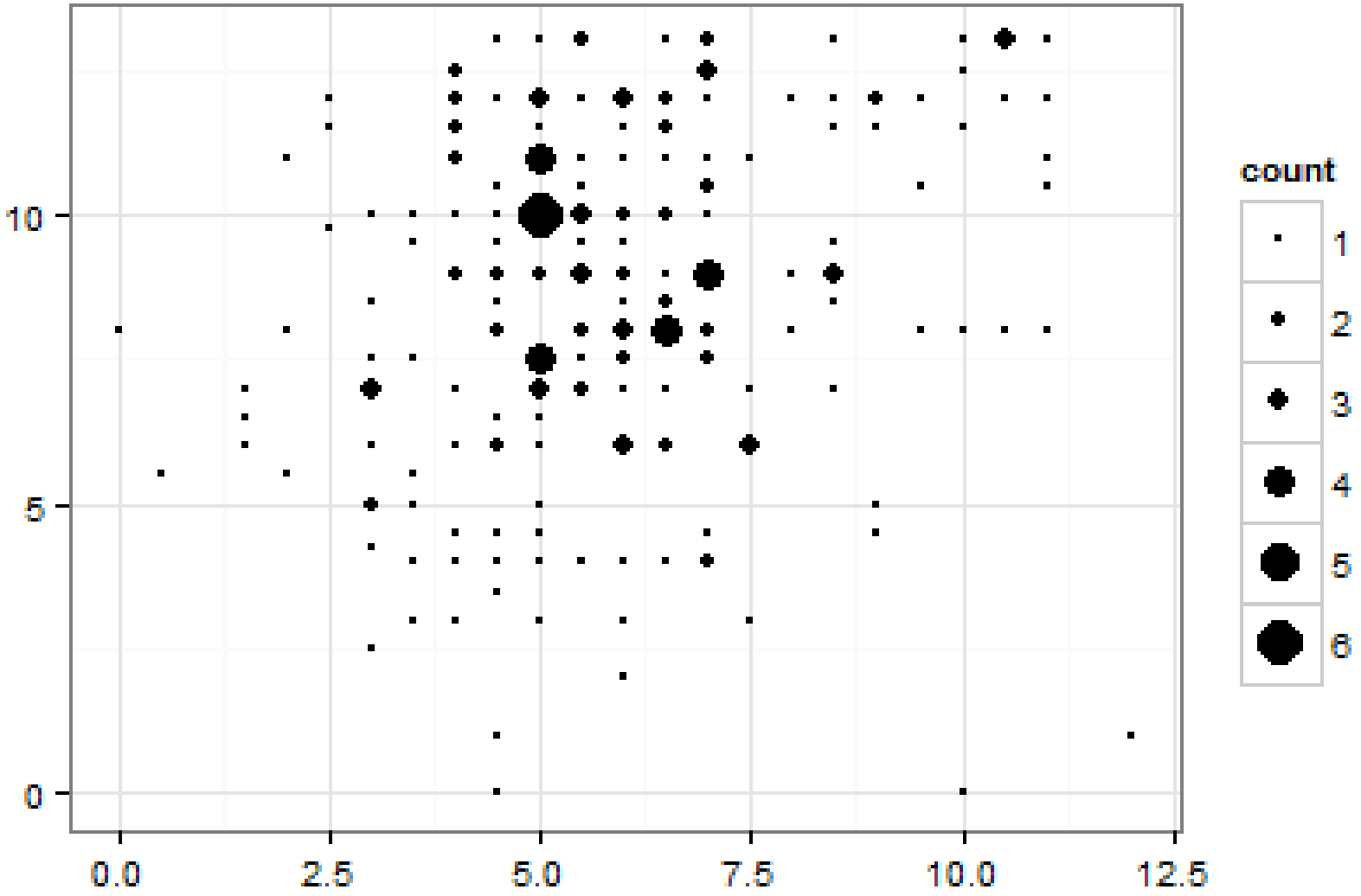
Overplotting – Solution 1: Jitter



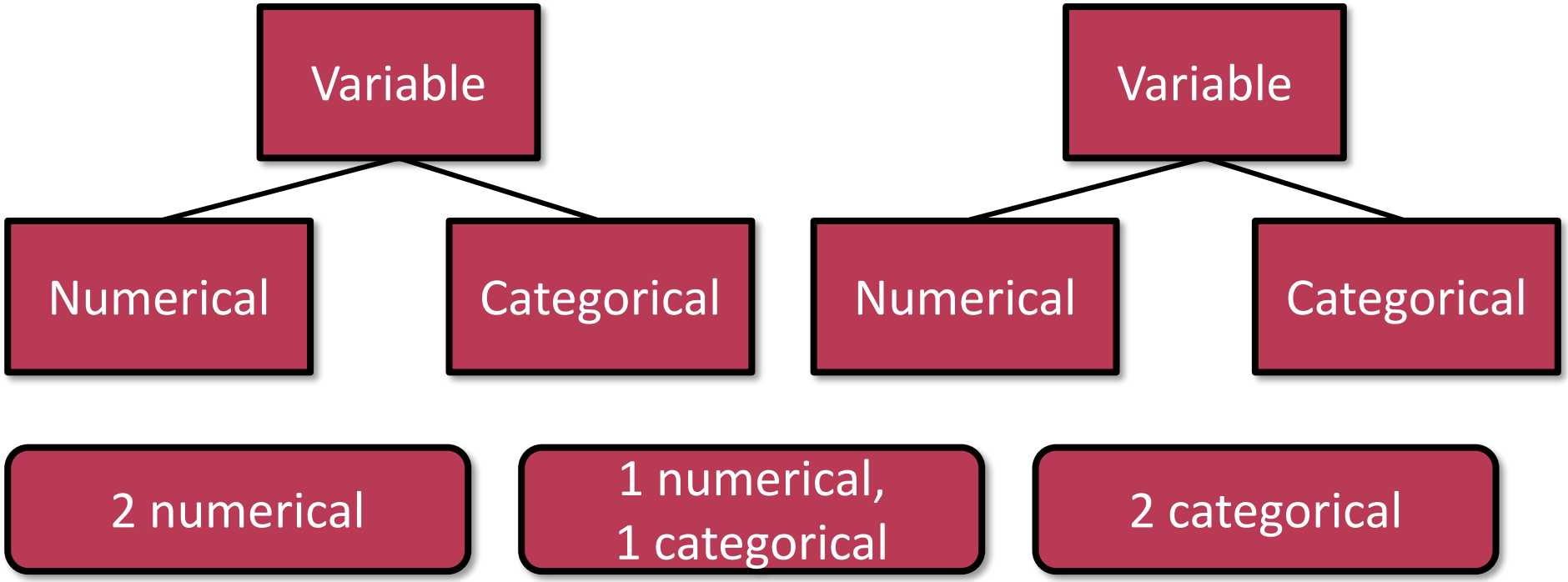
Overplotting – Solution 2: Transparency



Overplotting – Solution 3: Size

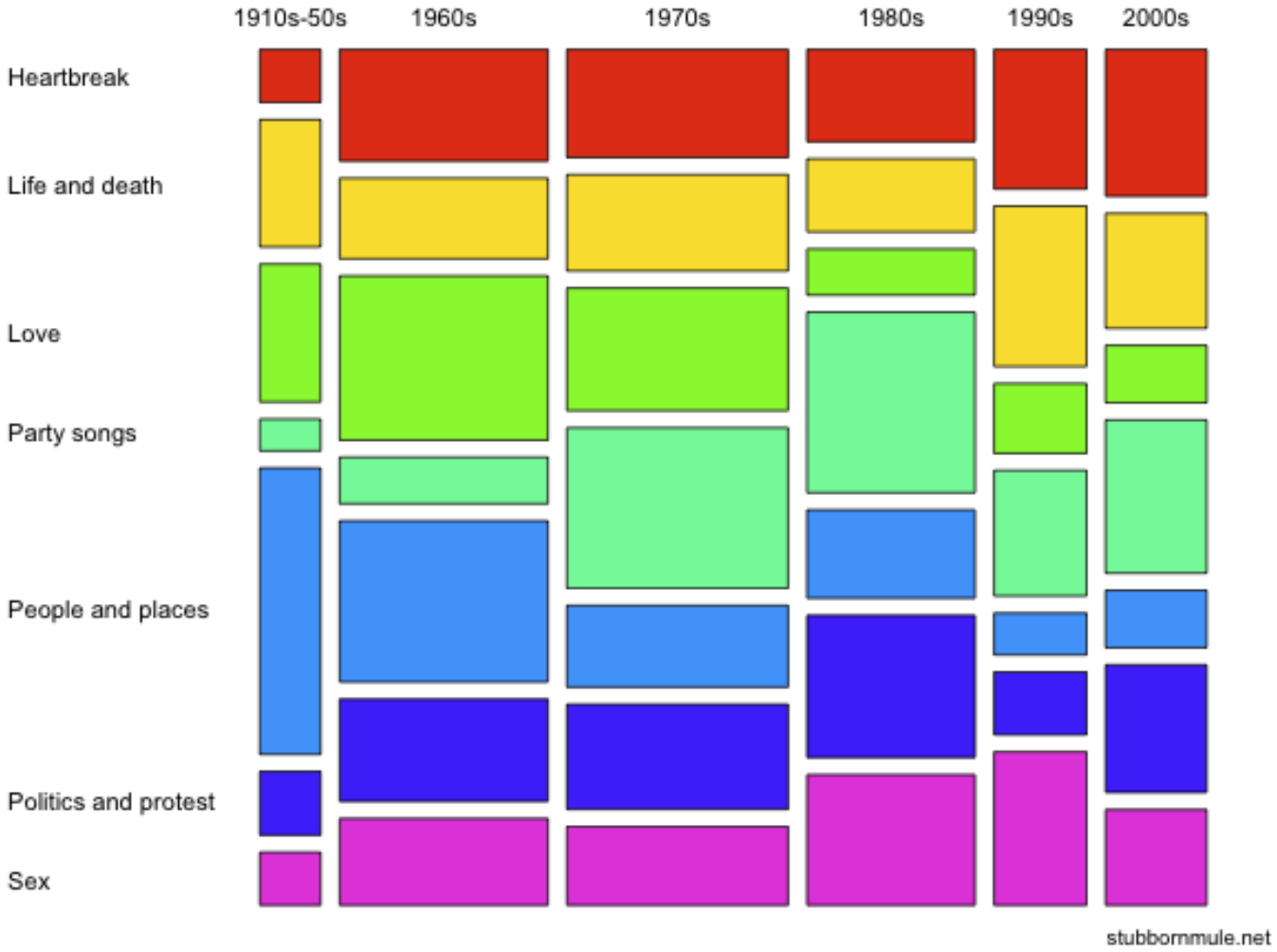


Relation between two Variables



Mosaic Plot

- Relation between 2 or more categorical variables



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

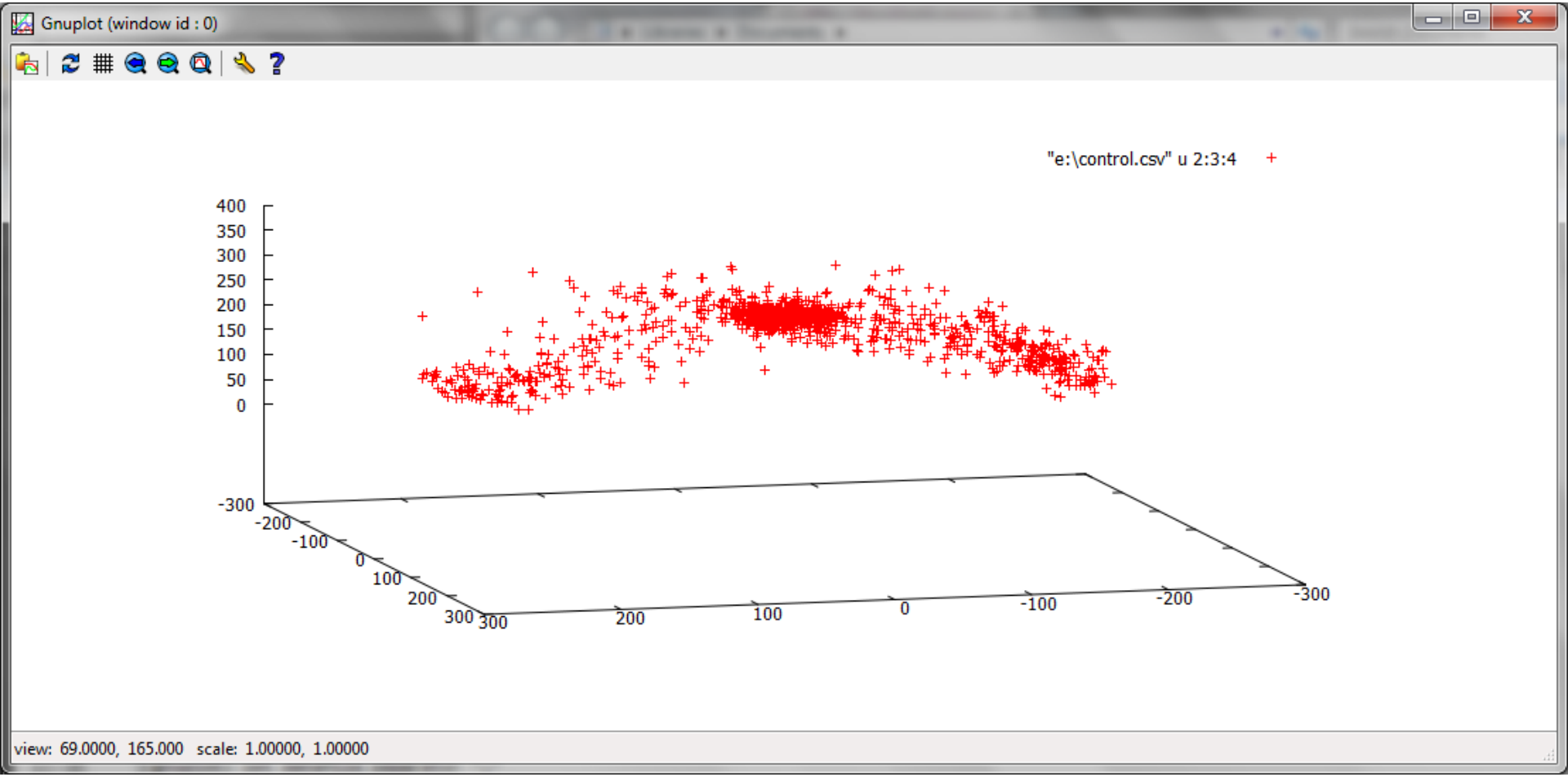
MULTIPLE VARIABLES

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)

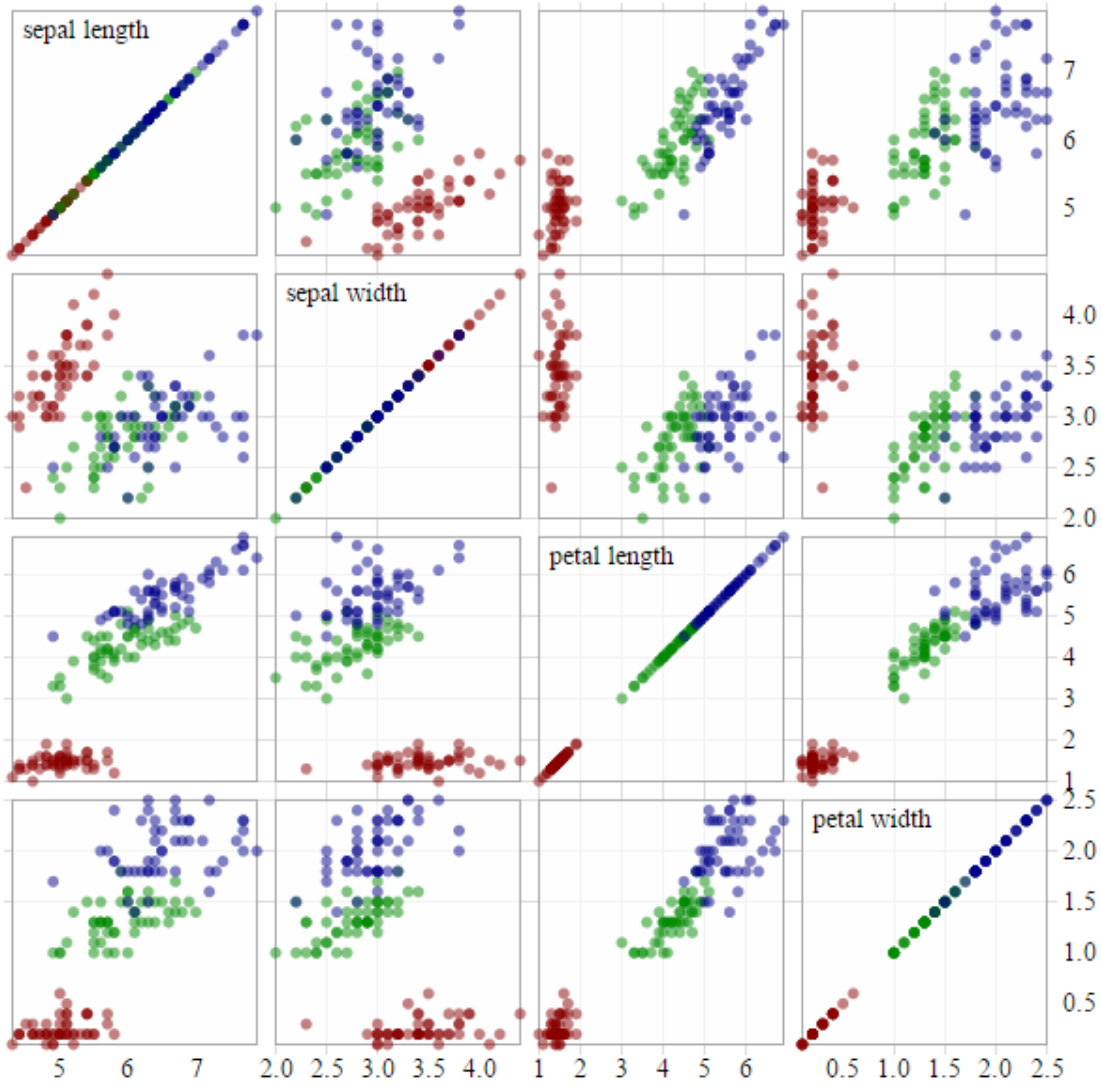
- E.g. bubble chart, heatmap, treemap

3D Plot

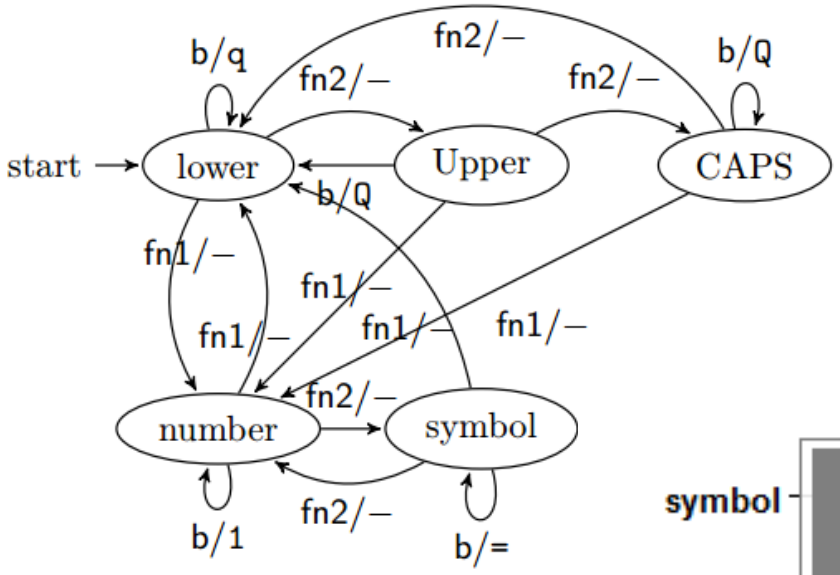

Gnuplot (window id : 0)
"e:\control.csv" u 2:3:4
400
350
300
250
200
150
100
50
0
-300
-200
-100
0
100
200
300
300
200
100
0
-100
-200
-300
view: 69.0000, 165.000 scale: 1.00000, 1.00000

Bubble Chart: Average Age by Regions

Scatterplot matrix

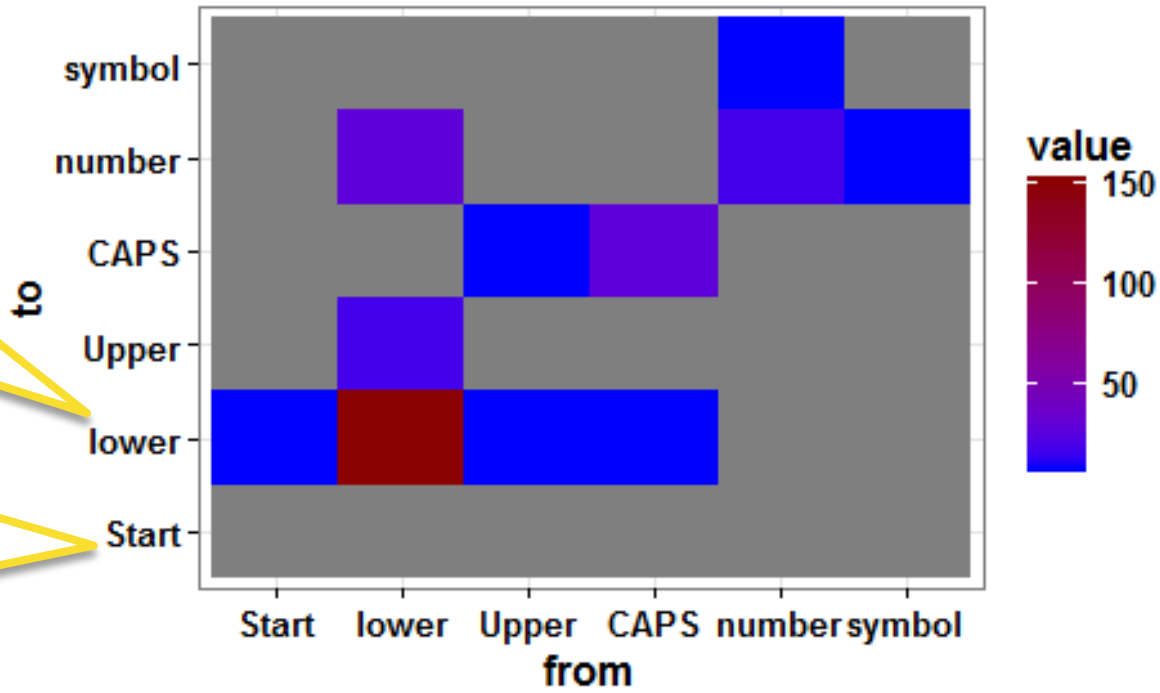


Heatmap: Operations Statistics

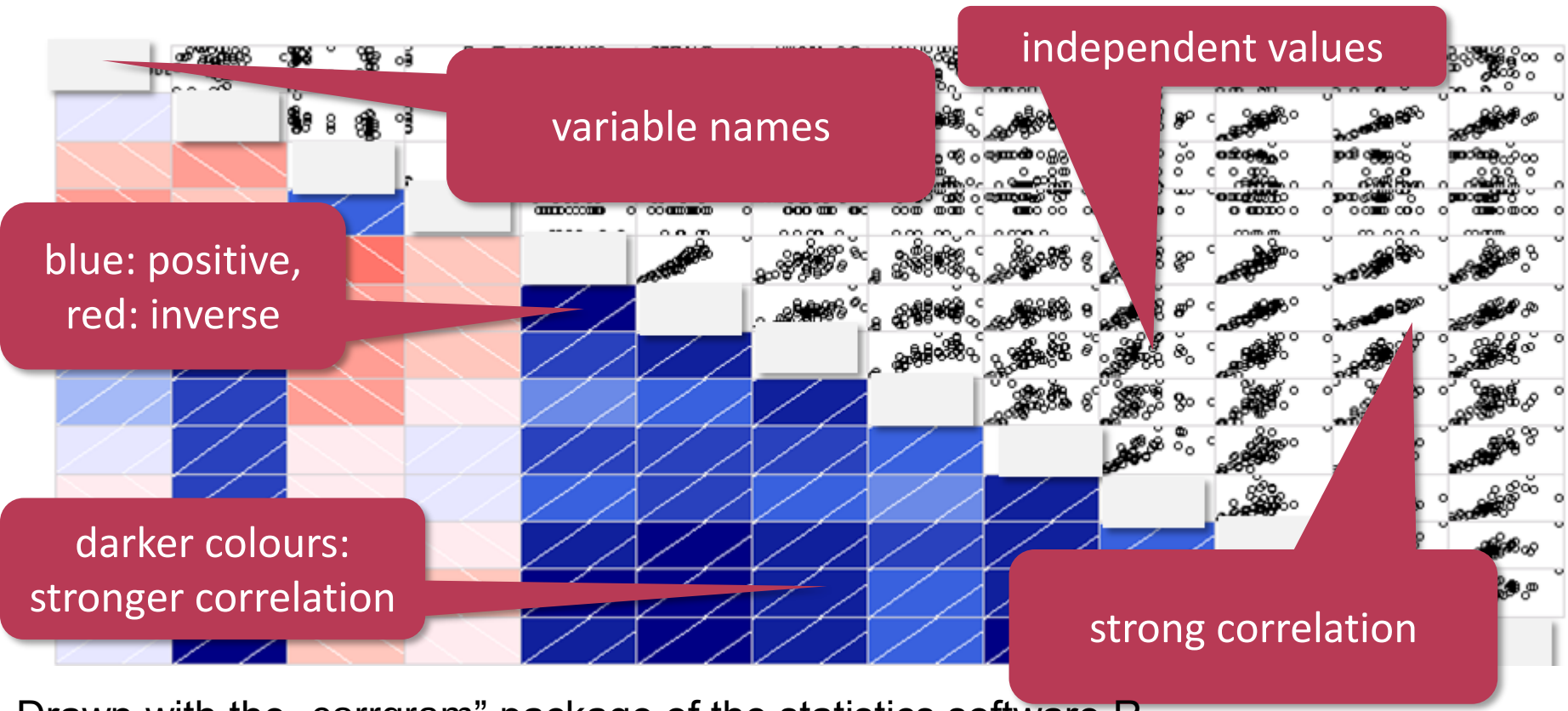


We mostly write simple texts.

In Start we only start, but we never return.



Pairwise Correlation of Multiple Values



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

Correlation (see Probability Theory):

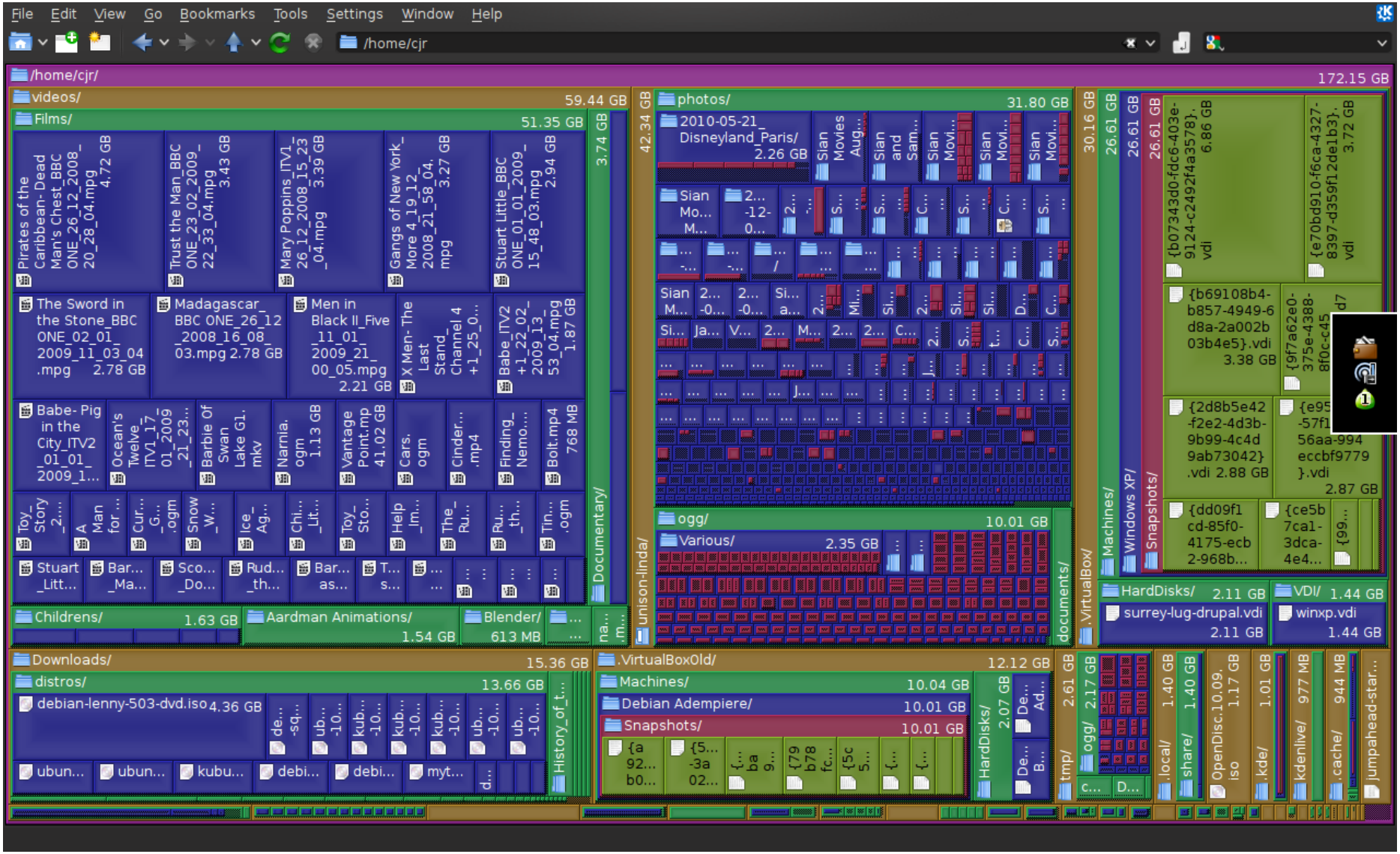
Strength and direction of the linear dependency between two variables

Over the diagonal: **scatterplot matrix**

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

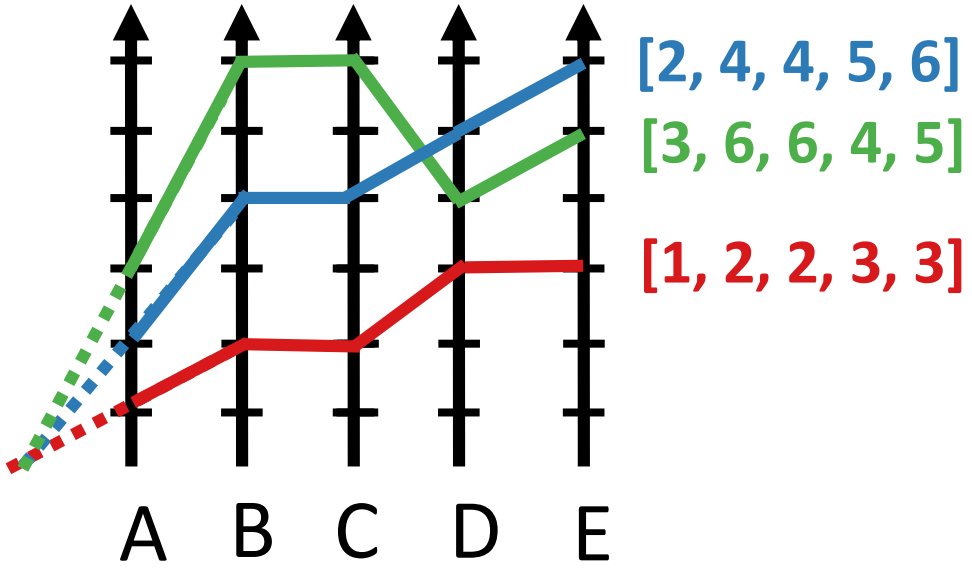
→ Which variables are important for the prediction of the load?

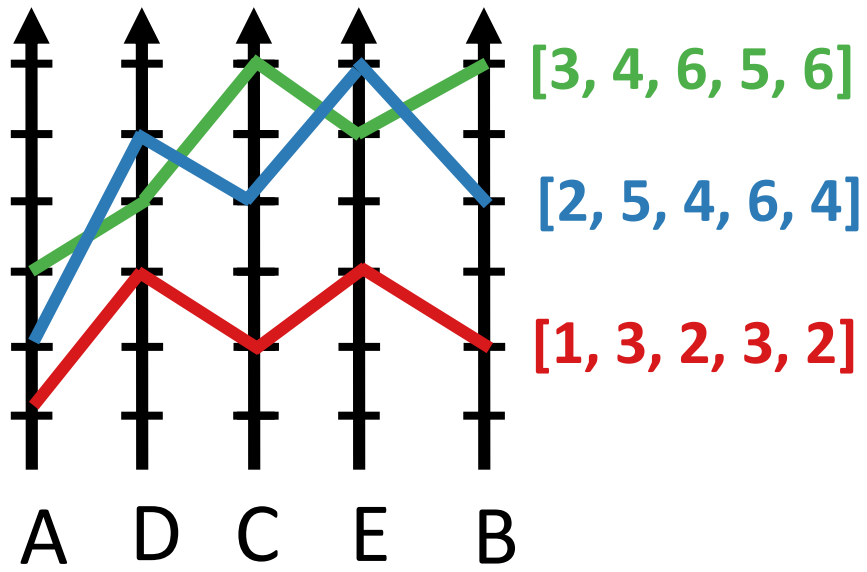
Tree Map: e.g. File System



Parallel Coordinates

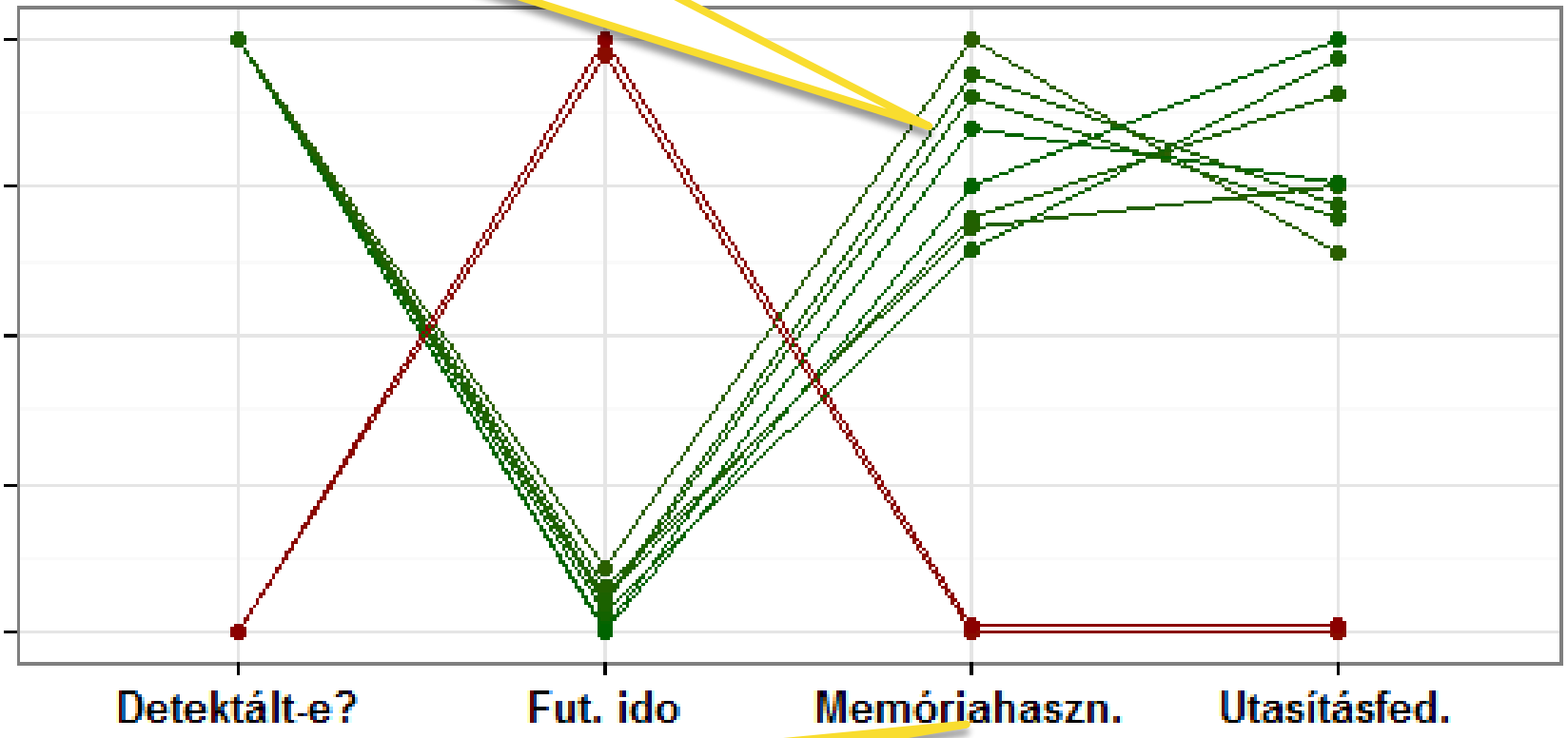
- Multi-dimensional visualization
- Compact, scalable
- Axis order?





Parallel Coordinates: Analysis of the Test Cases

1 test case: 1 broken line

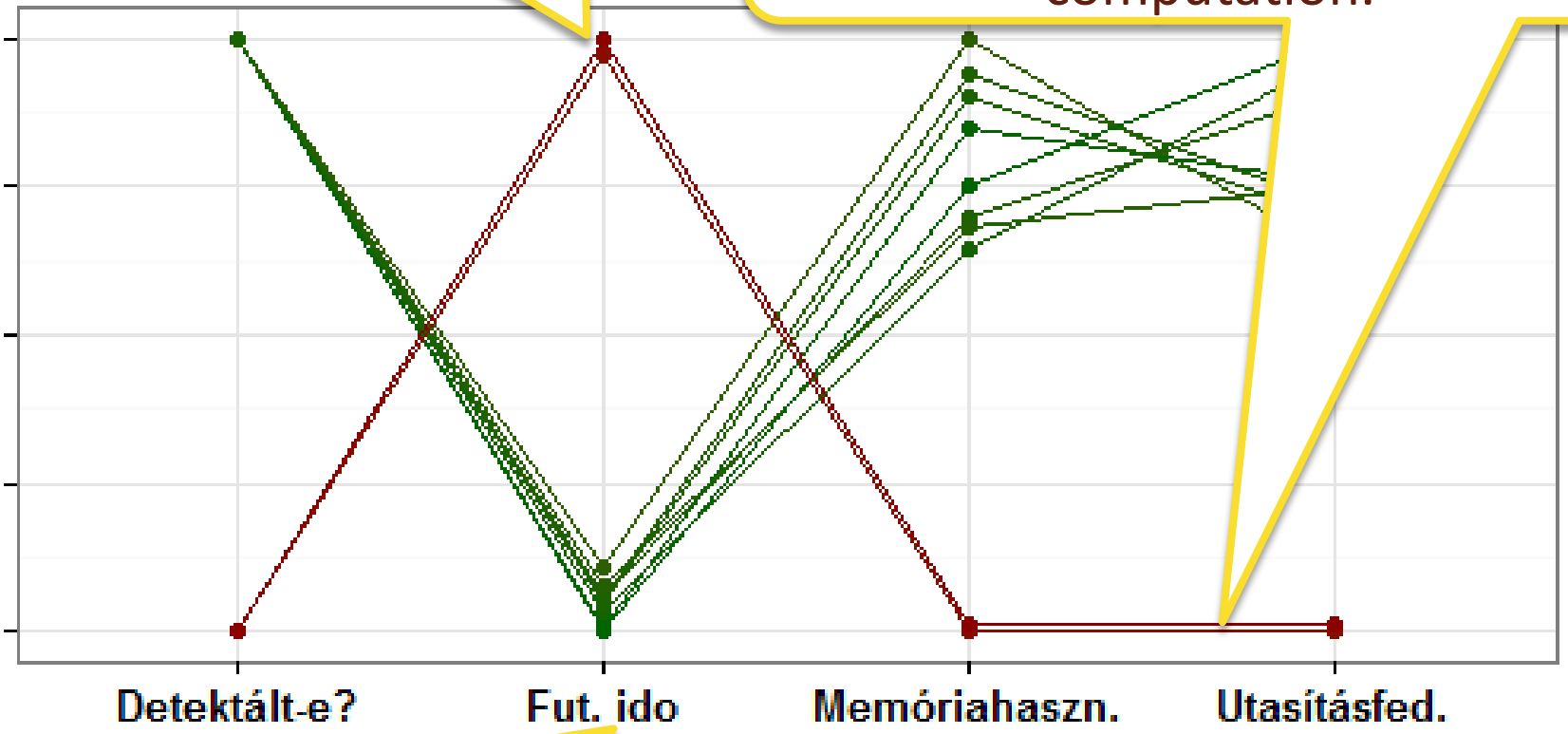


The variables appear on the x-axis

Parallel Coordinates: Analysis of the Test Cases

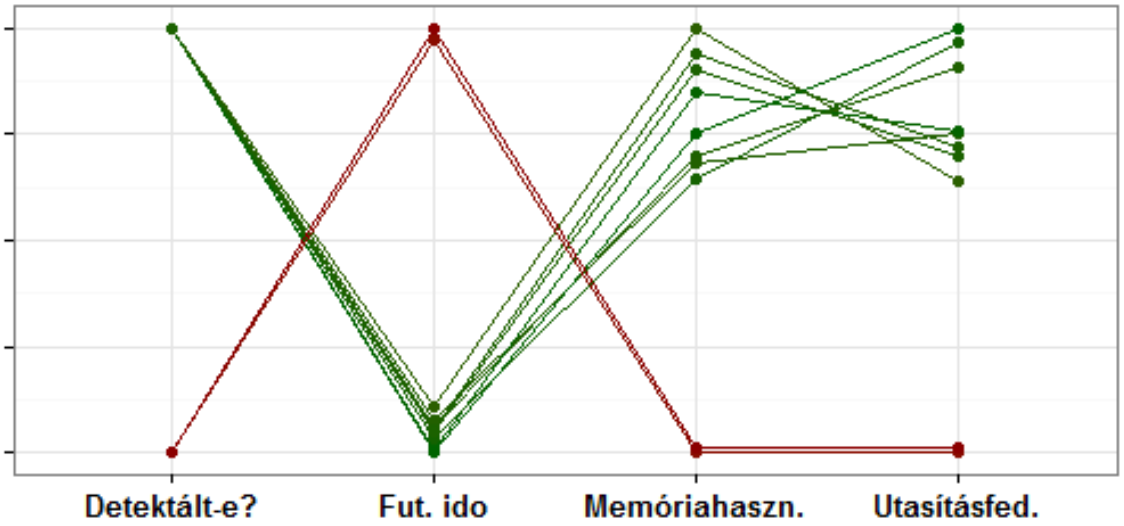
Timeout?

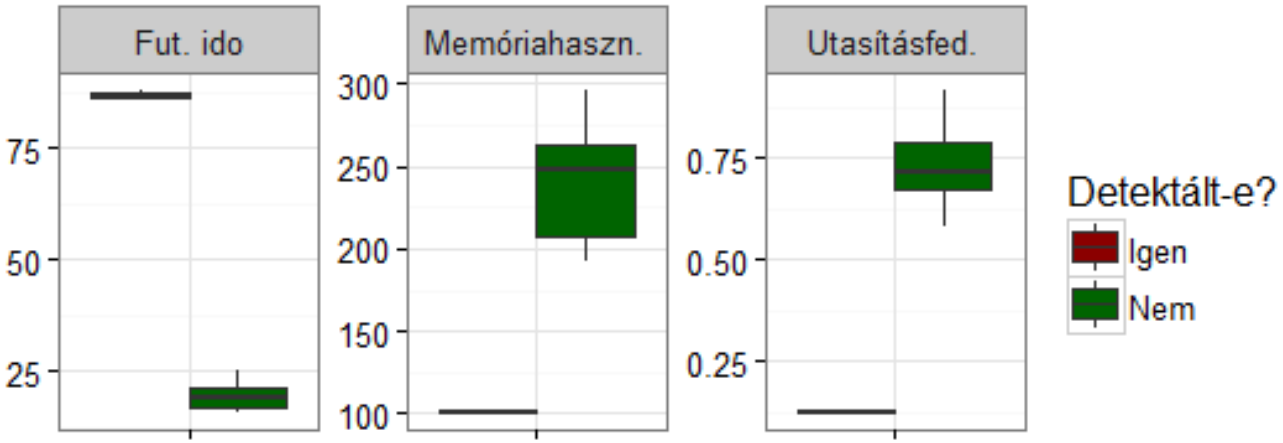
The ones detecting an error did not even came to the actual computation.



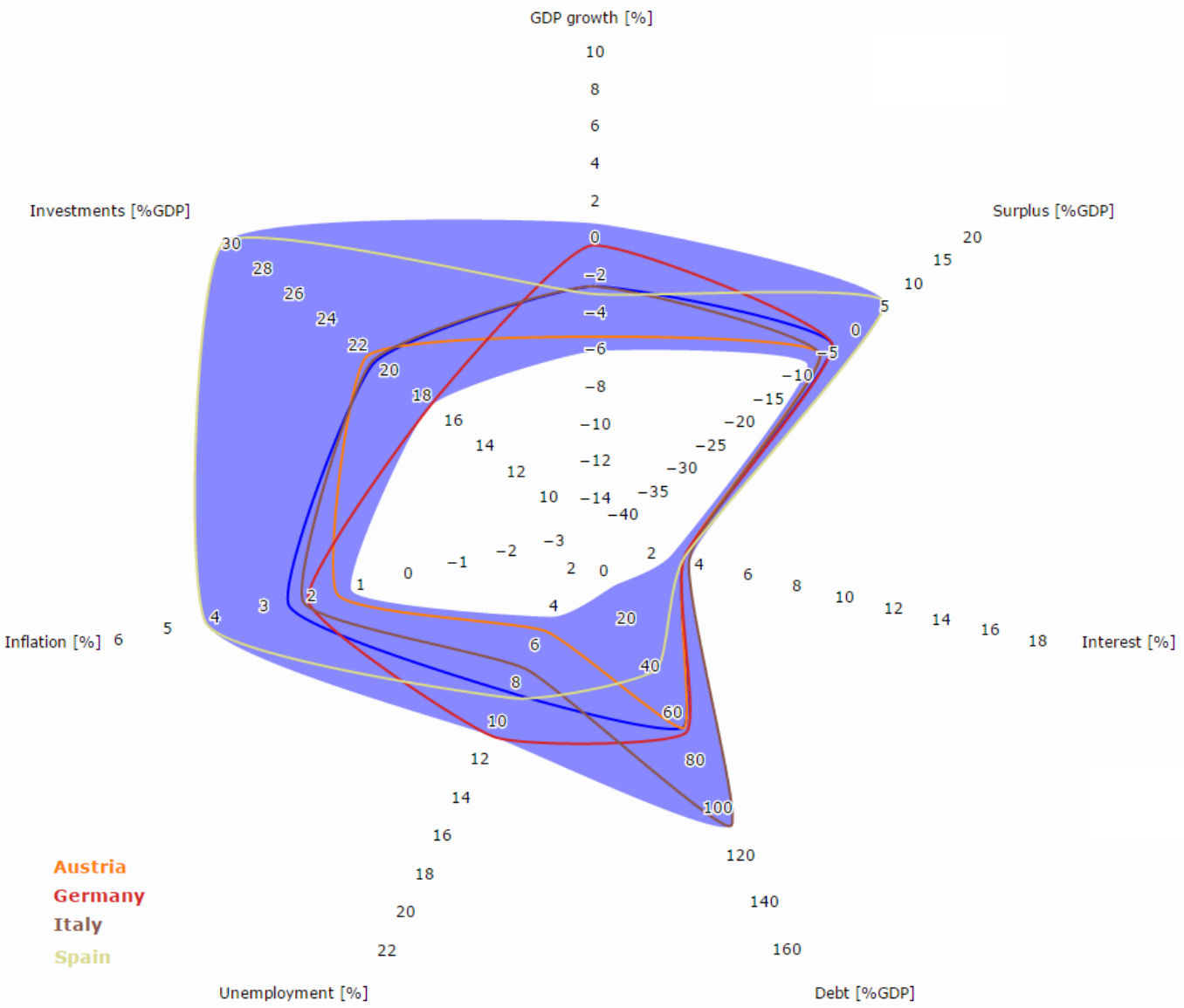
Run time and memory usage seem to be in a positive relation (if the test is successful)

Parallel Coordinates: the Alternatives





Radar Chart: An Extension of Parallel Coord.



OUTLOOK

Data mining „brickstones”

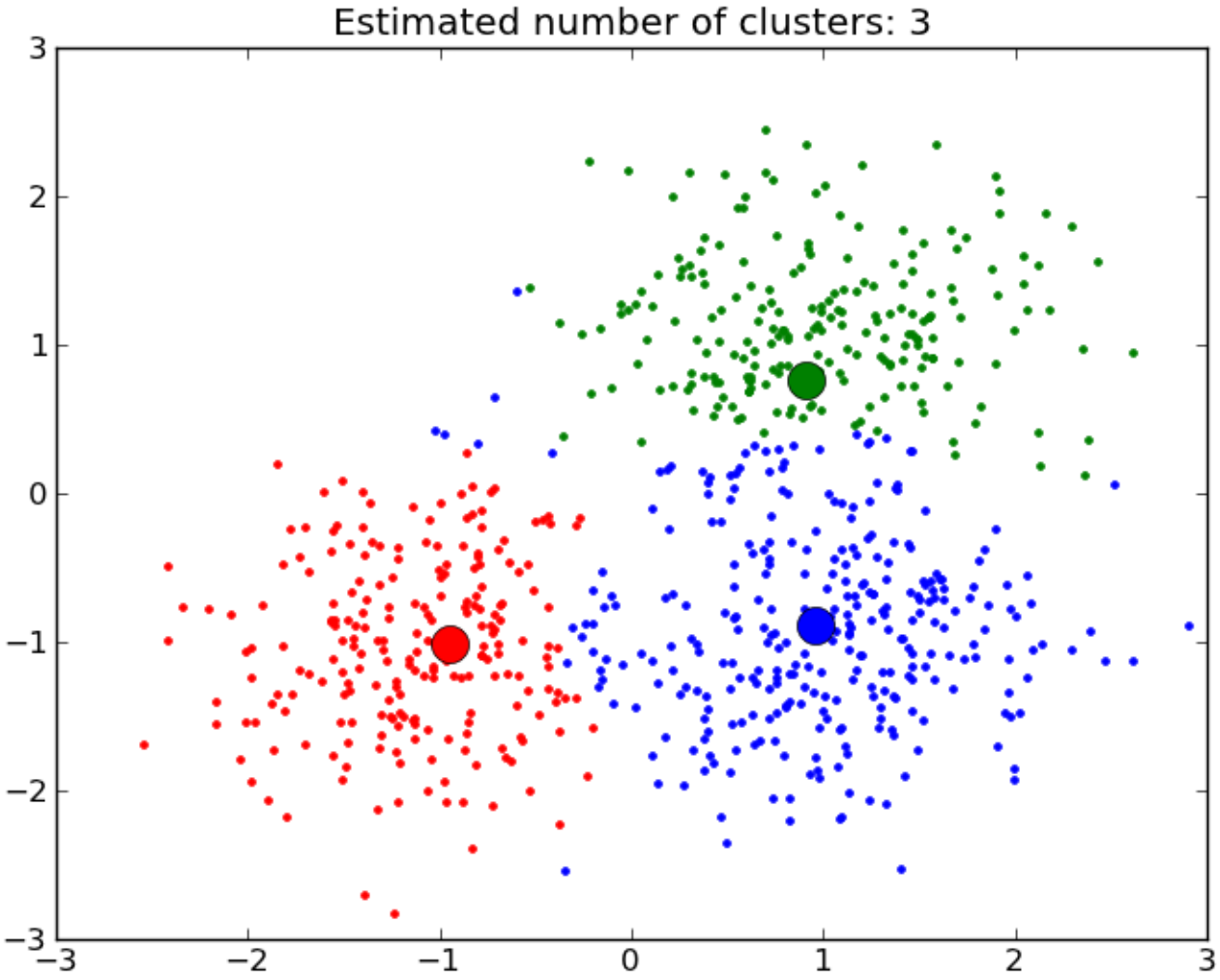
Clustering

Classification

Association rules

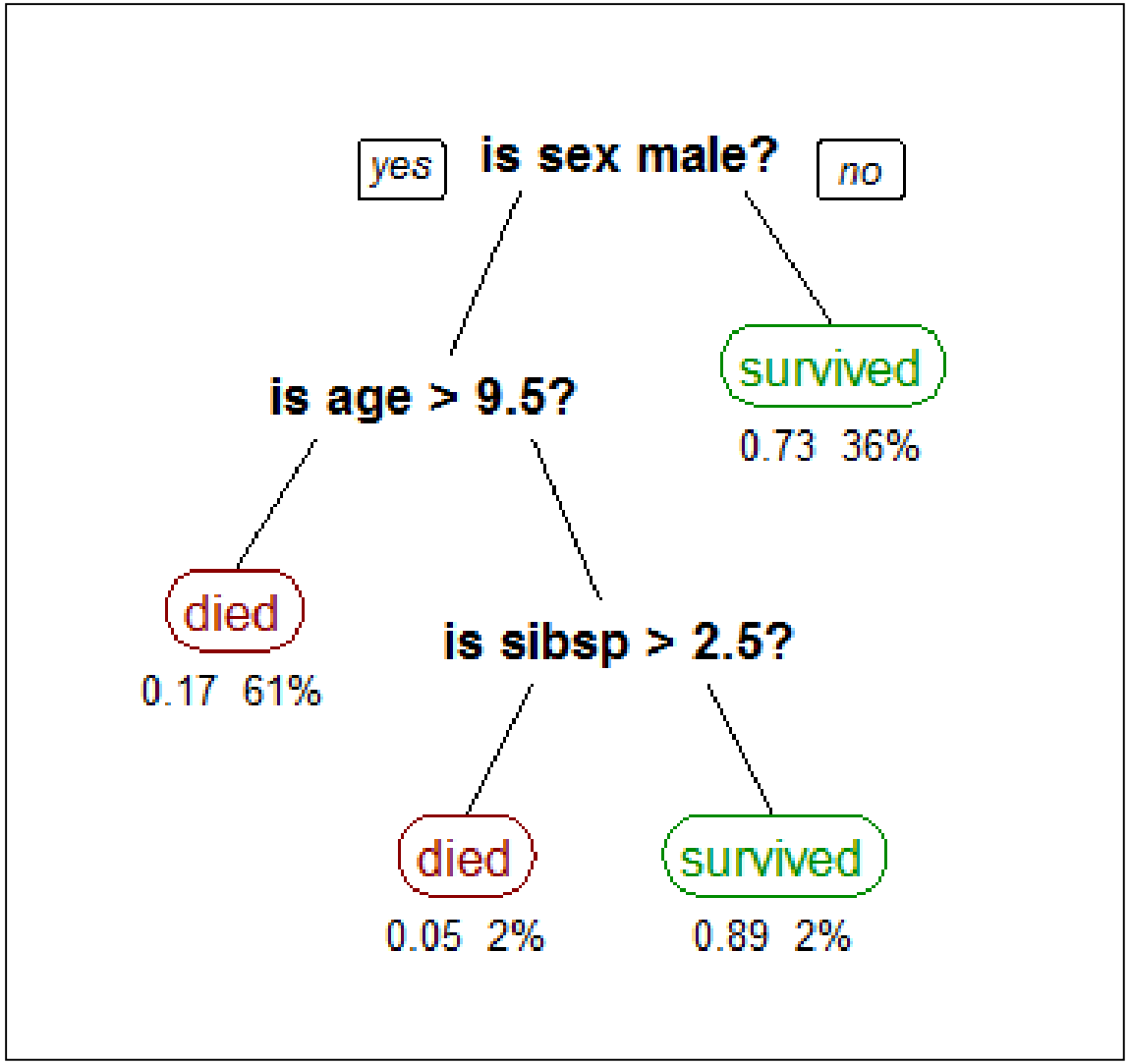
Regression

Clustering

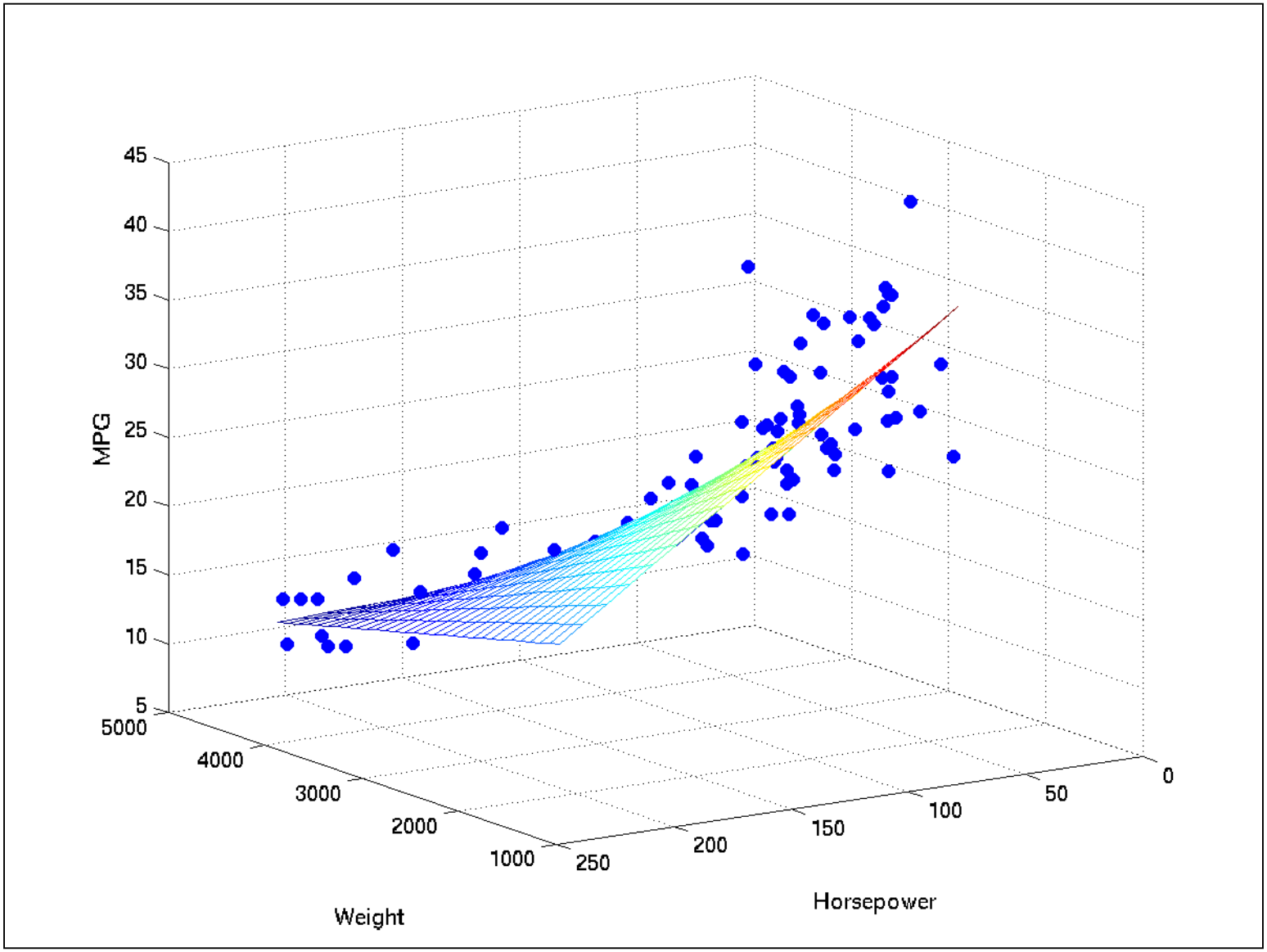


Association rules

Classification



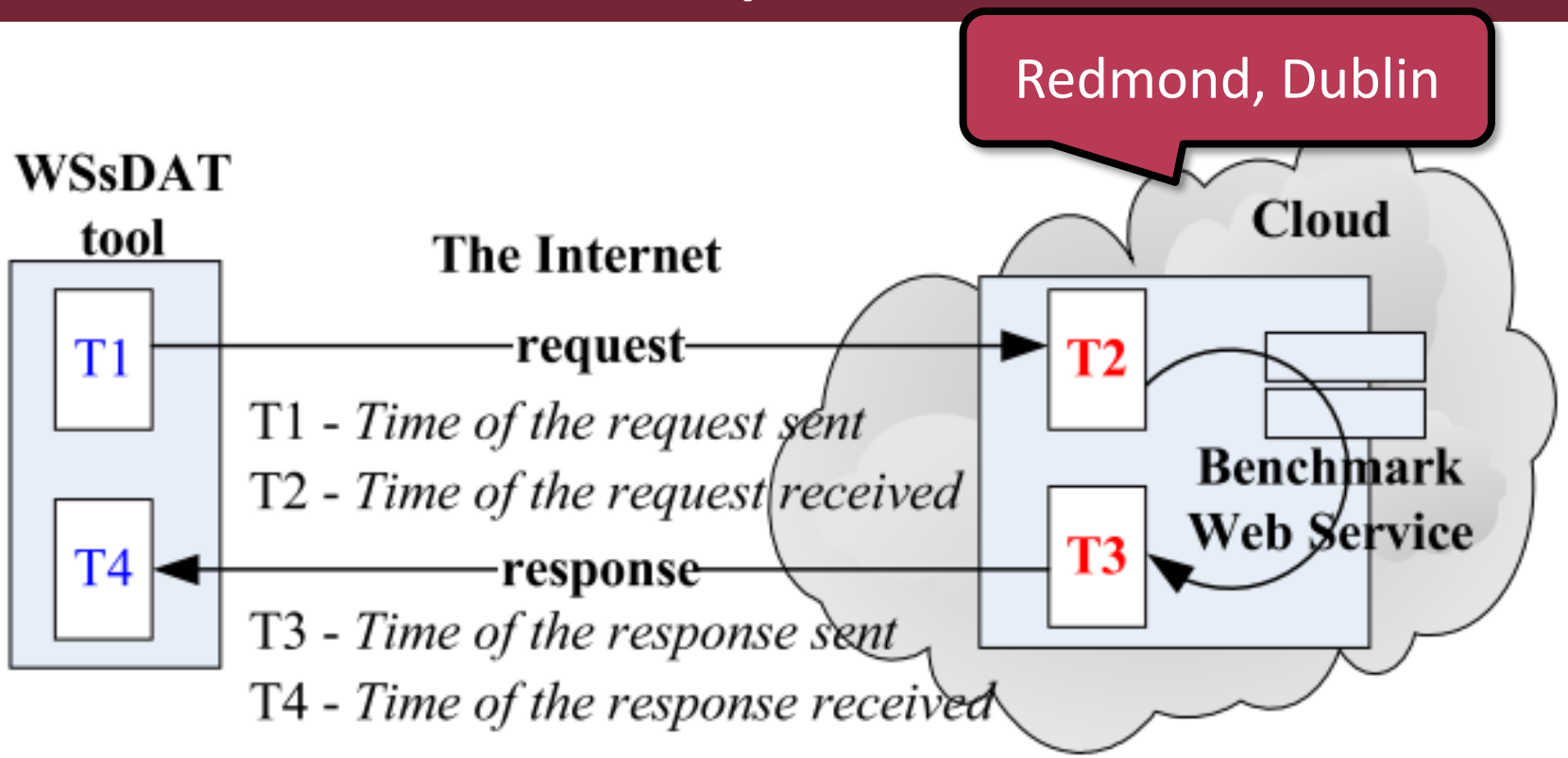
Regression


MPG
45
40
35
30
25
20
15
10
5
Weight
5000
4000
3000
2000
1000
Horsepower
250
200
150
100
50
0

CASE STUDY

How to make the analysis systematic?

The experiment



$$RT = T4-T1 \qquad RPT = T3-T2$$

$$RTT = (T2-T1) + (T4-T3) = (T4-T1) - (T3-T2) = RT-RPT$$

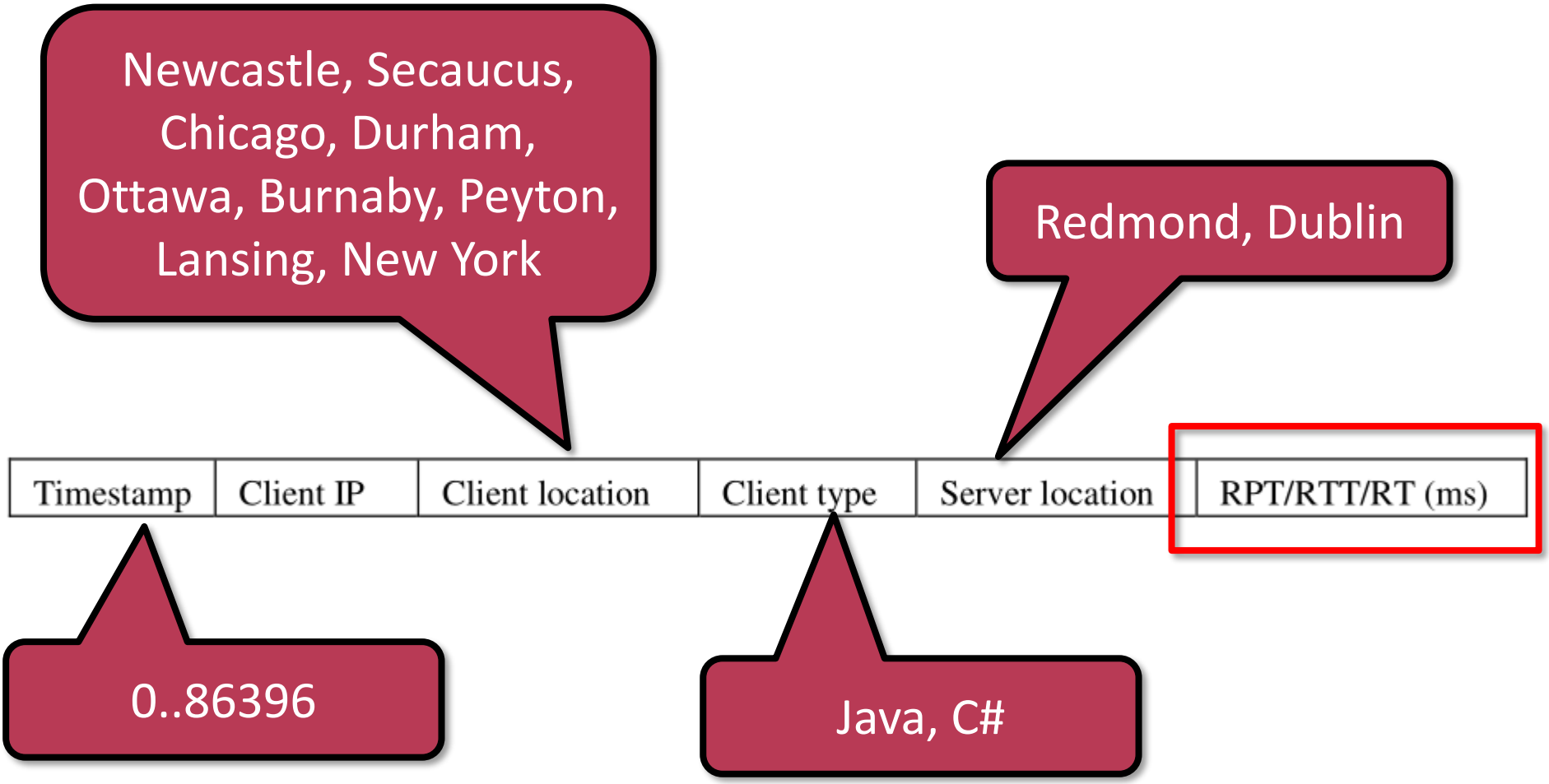
[8], p 186

Gorbenko, A., Kharchenko, V., Mamutov, S., Tarasyuk, O., Romanovsky, A.: Exploring Uncertainty of Delays as a Factor in End-to-End Cloud Response Time. In EDCC: 2012

Delay analysis

- Benchmarking a cloud-based web-service
- Analysis of daly factors in end-to-end RT
- Comparison
  - Client locations
  - Client technology
  - Cloud technology

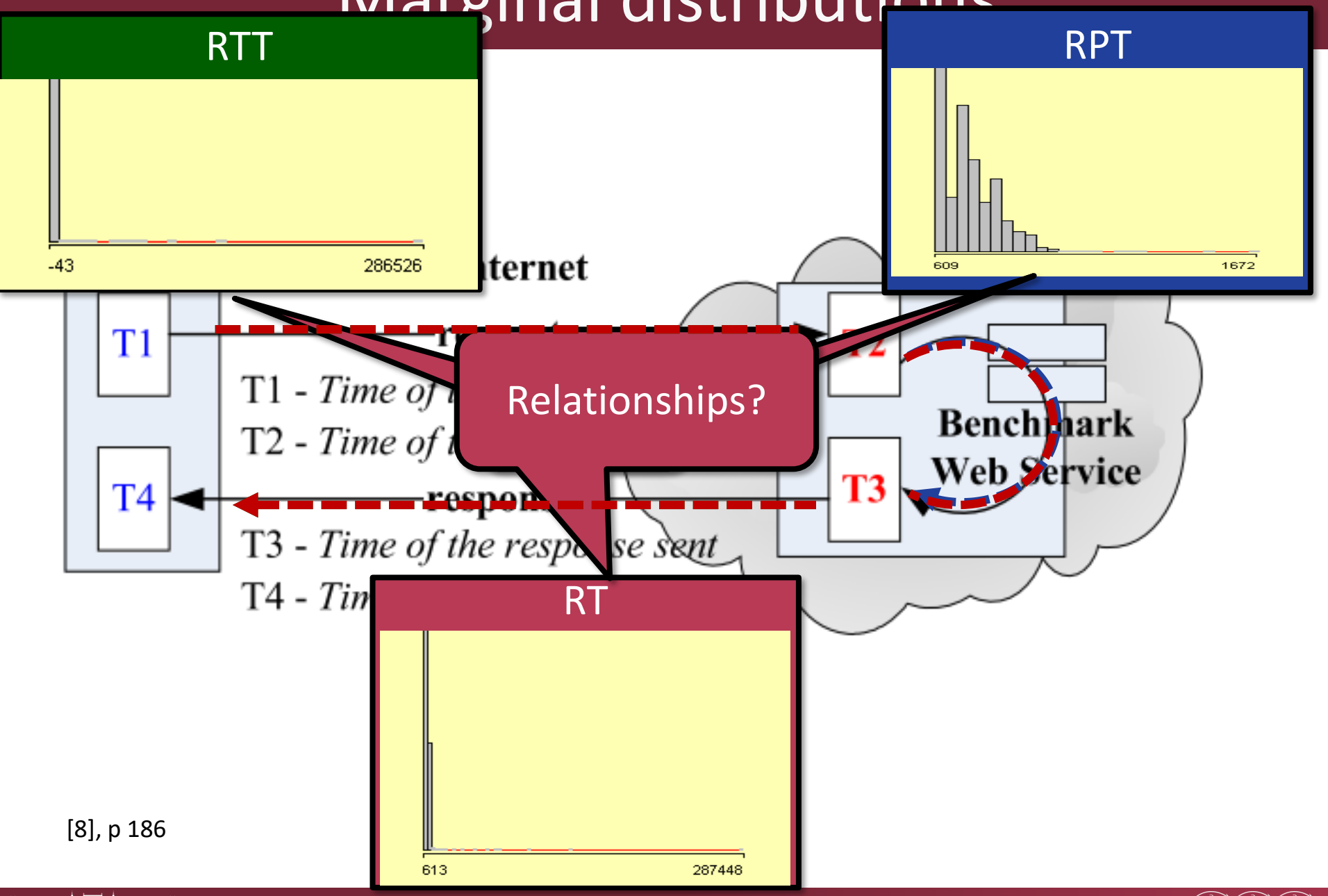
Dataset


Newcastle, Secaucus, Chicago, Durham, Ottawa, Burnaby, Peyton, Lansing, New York
Redmond, Dublin
Timestamp
Client IP
Client location
Client type
Server location
RPT/RTT/RT (ms)
0..86396
Java, C#

Delay analysis: conclusions

- RTT dominates RT
- Significantly differences in RTT values
  - Among different clients
  - Among different queries of the same client
- Even the UK client - Dublin DC requests go via Redmond
- Different RPT patterns by the in-house hosted WS (compared to the cloud services)
- Different RT values by Java and C# clients
- Hour-of-a-day dependency by clients in Newcastle
- The average exception propagation time for C# clients was significantly lower than for Java clients

Marginal distributions



[8], p 186