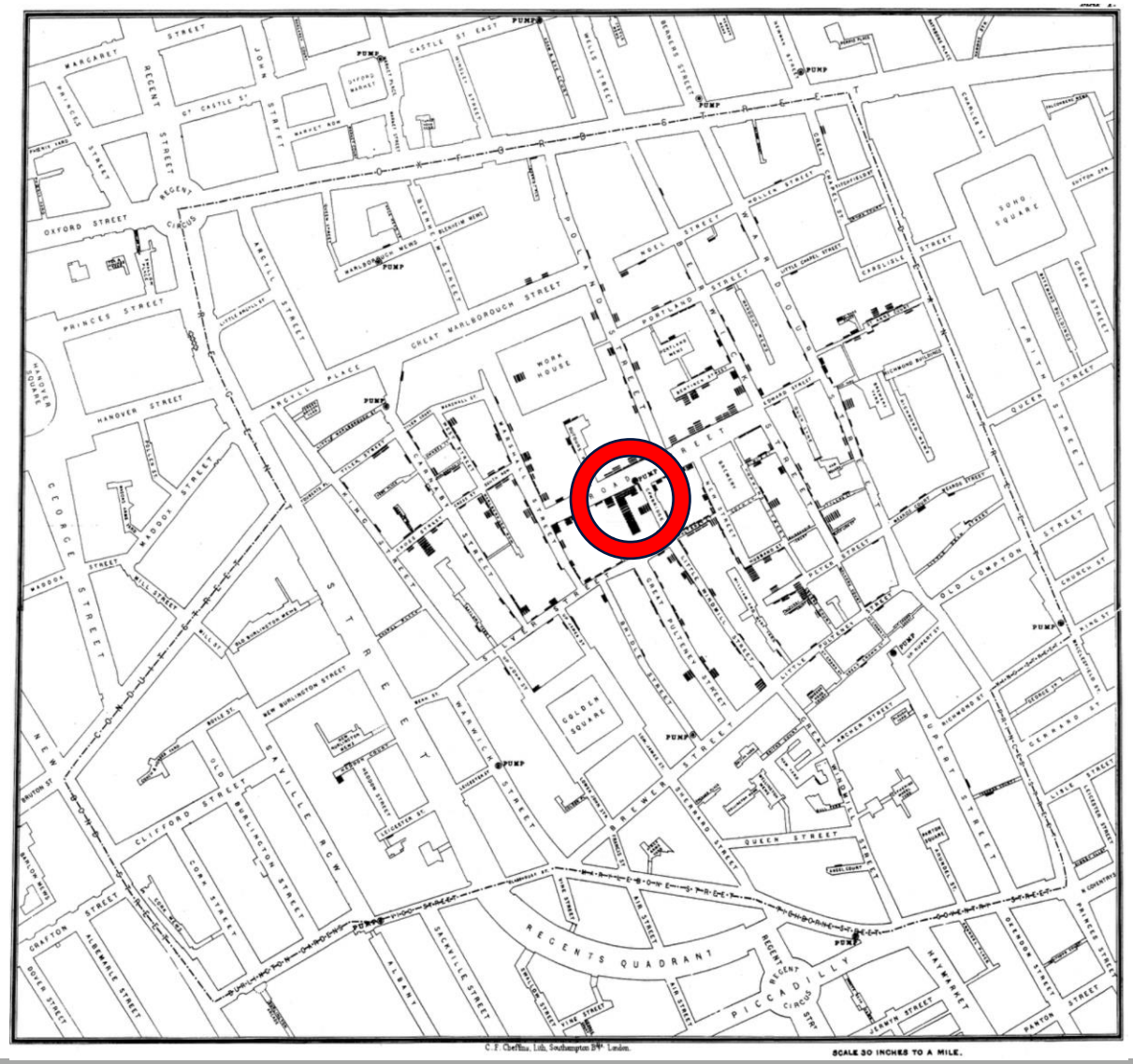
Data analysis

Andras Foldvari - andras.foldvari@edu.bme.hu

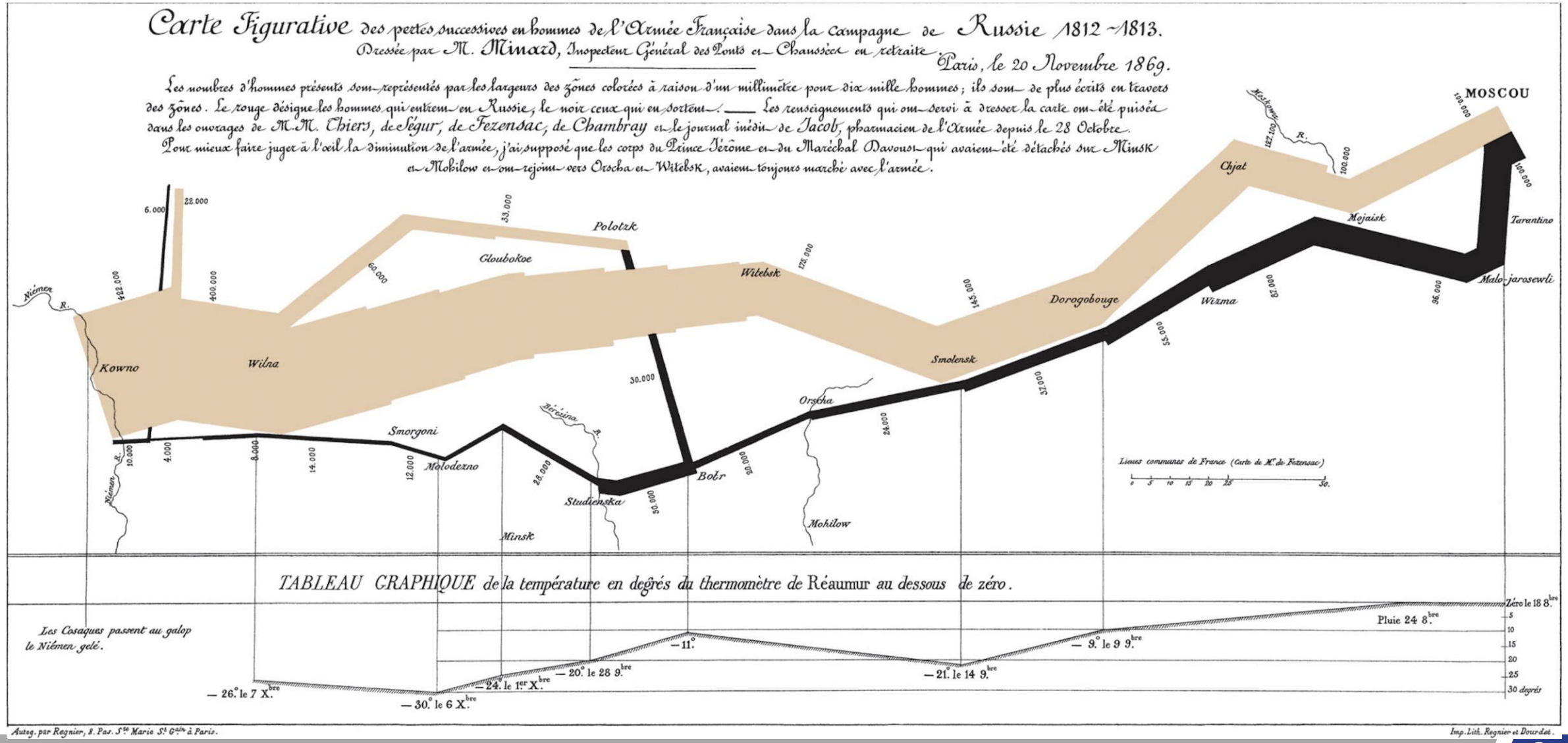
Sources: System Modeling, Empirical Systems Engineering

Budapest University of Technology and Economics
Department of Artificial Intelligence and Systems Engineering
ftsrg Research Group

Exploring Underlying Relations

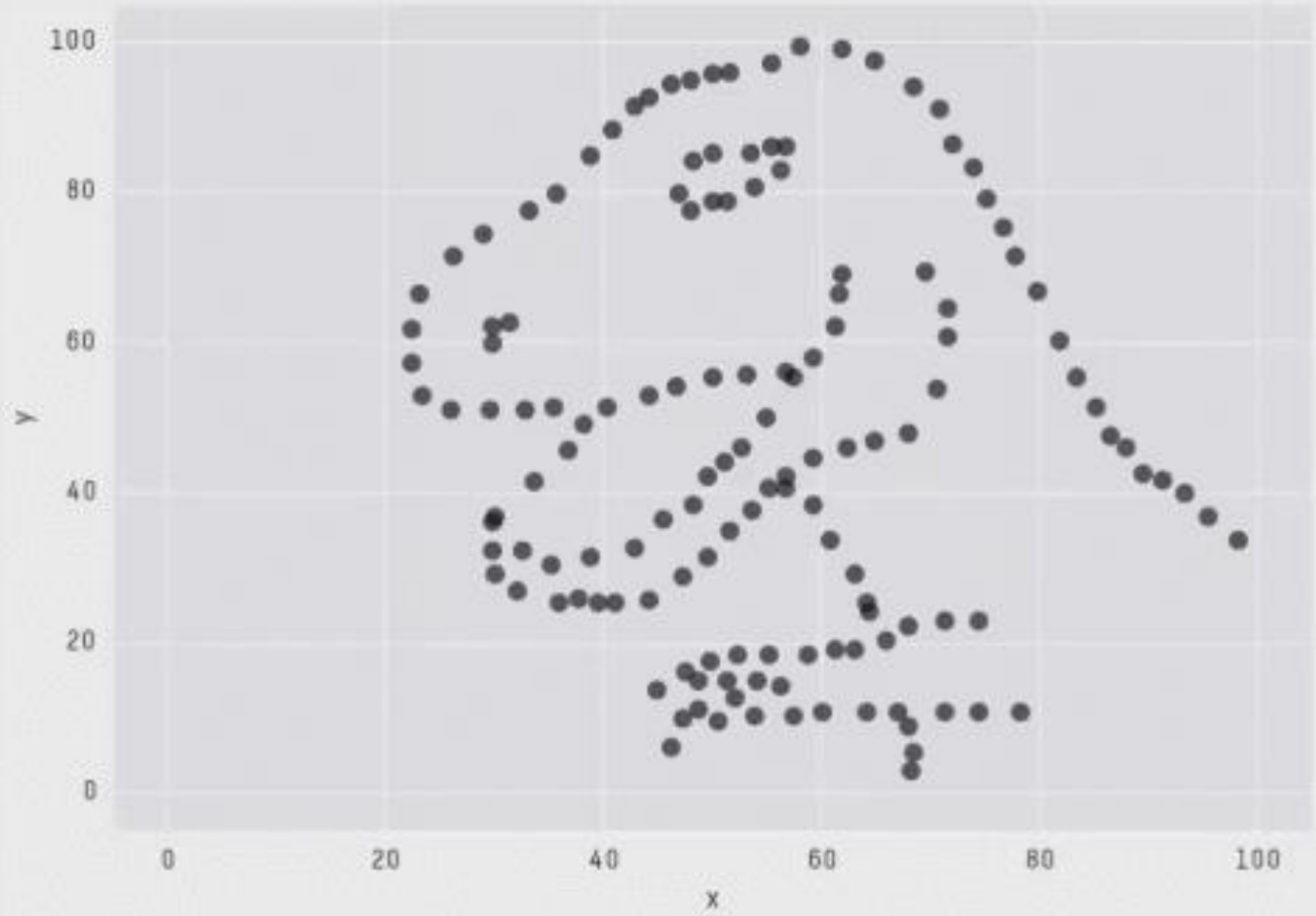


Infographics




Advance
Retreat

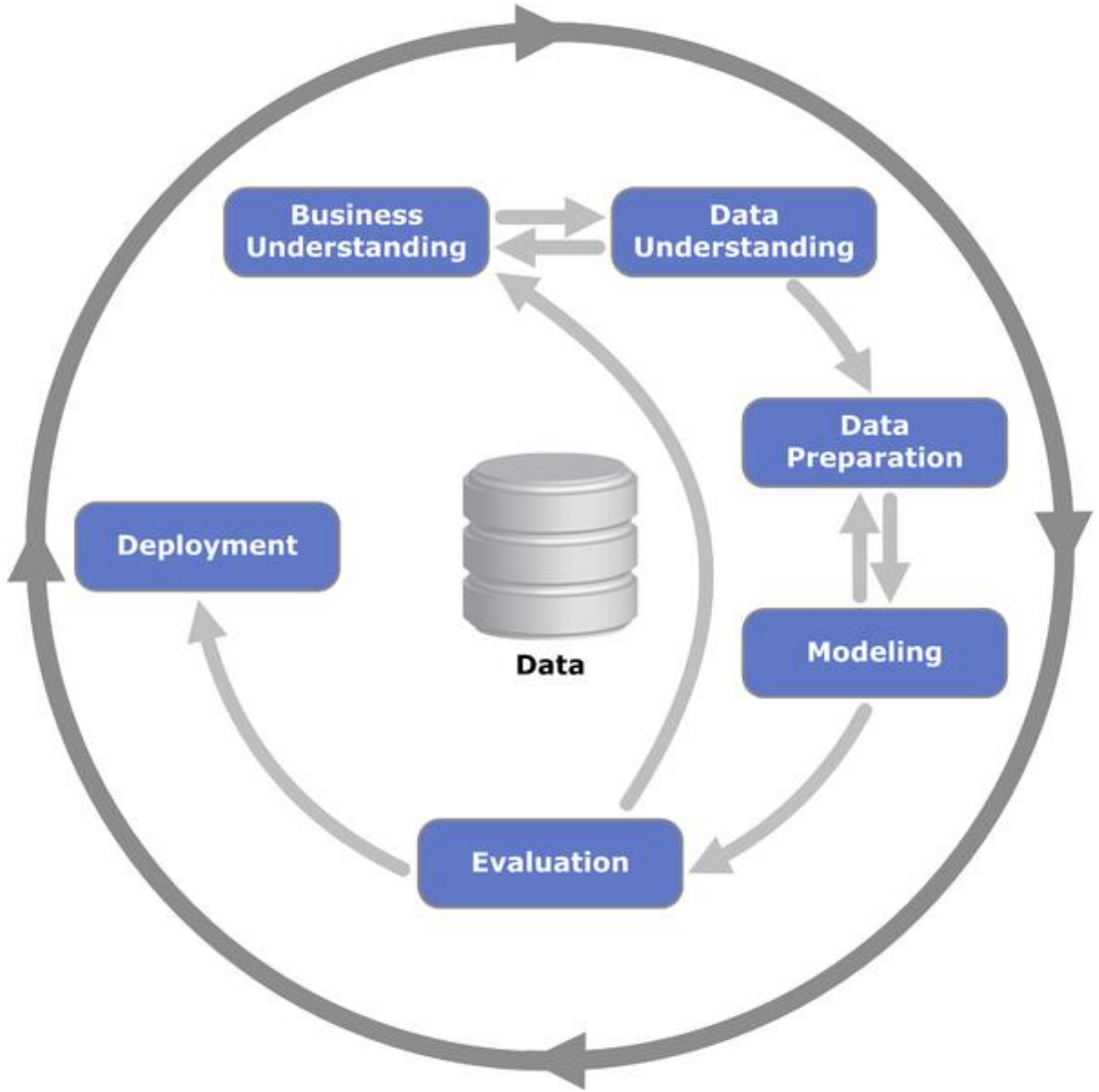
The datasaurus



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Alberto Cairo (2016): The datasaurus

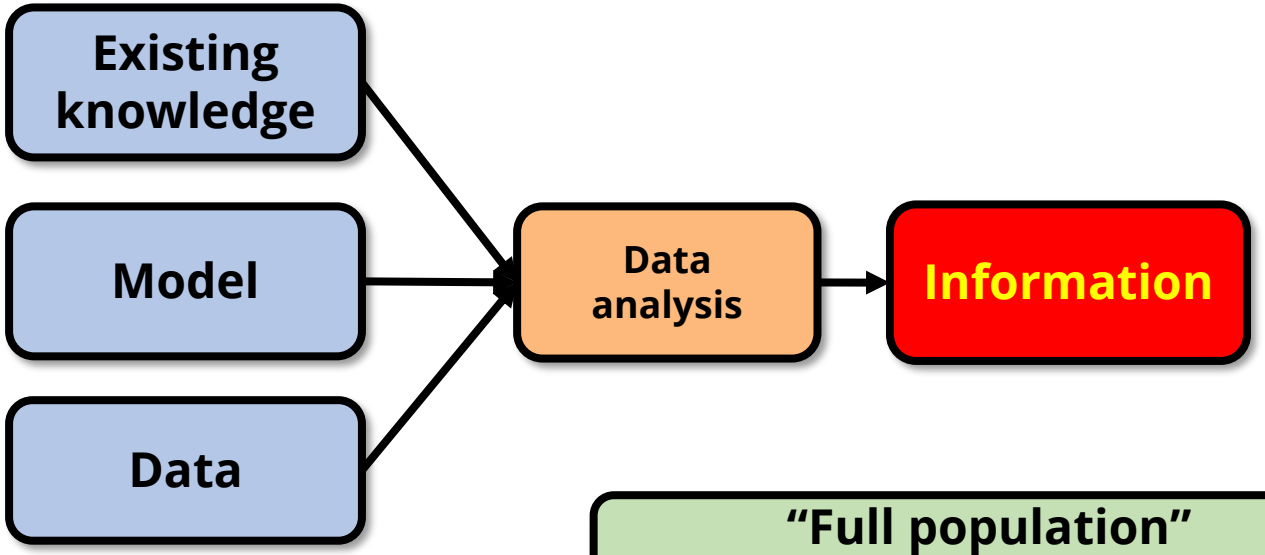
CRISP-DM folyamatmodell



- **Cross-industry Standard Process for Data Mining**
- **IBM 2015**

https://developer.ibm.com/articles/architectural-thinking-in-the-wild-west-of-data-science/?mhsrc=ibmsearch_a&mhq=ASUM-DM

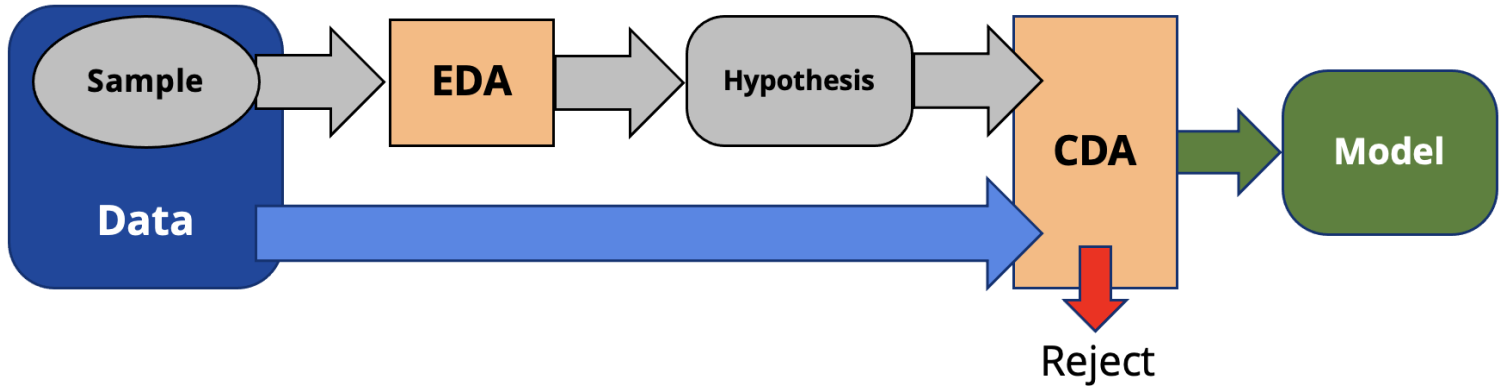
Informal model of data analysis



“Full population” characteristics	Classical stats
New insights	“Data mining”
“Good enough” classifiers/predictors/…	“Machine learning”
Visual patters, new insights	Exploratory Data Analysis

Exploratory Data Analysis (EDA)

- Approach to analyzing data sets
  - Summary of main characteristics
  - Easy-to-understand,
  - Often: visual graphs
  - Without using statistical models
  - Or having formulated a hypothesis



- Prepares Confirmatory Data Analysis
  - Hypothesis testing
  - Statistical model building and checking

- (Dynamic) visualization delivers understanding: outliers, trends and patterns

- *Not* statistics-free - but "inference" sensitivity to errors should be managed
  - Robust statistics - low sensitivity to outliers
  - Nonparametric statistics – no a priori assumptions

Data models

Tabular Data

- Data is stored in rows and columns, in a tabular format.
- Columns represent attributes, and rows represent individual records.
- E.g., Pandas, MySQL, PostgreSQL

Graph data model

- Data is represented as nodes and edges.
- Useful for visualizing relationships and connections.
- E.g., Neo4j, GraphDB

Ducument data model

- Data is stored in structured document formats.
- Offers high flexibility, especially for handling unstructured or semi-structured data.
- JSON, XML
- E.g., MongoDB, CouchDB

Column data model

- Only relevant columns are read during queries.
- Commonly used for Big Data.
- E.g., Apache Cassandra, HBase

Further data models

- **Key-Value**
  - **Redis**
- **Object-Oriented**
- **Semantic Models**
  - **RDF, OWL**
- **File-Based**
  - **AWS S3**
- **Time Series**
  - **InfluxDB**

Tabular Data Model

- **Table row**= model element
- **Table column**= attribute

Név	Típus	Méret (kB)	Utolsó módosítás
Dokumentumok	könyvtár		2016.02.02
szerződés.pdf	fájl	569	2015.11.09
Képek	könyvtár		2016.02.02
logó.png	fájl	92	2015.03.06
alaprajz.jpg	fájl	1226	2016.02.02

- Data analysis tools (e.g., R, Python): **dataframe**
  - Each row represents a measurement
  - Each column has a specific type

Data Cleaning

- The goal of data cleaning is to remove or correct errors, gaps, and inconsistencies that may distort analysis results.

- Handling missing values
  - They may indicate errors!
- Duplicates
- Erroneous data and outliers
  - Outliers can be valuable!
- Standardizing data formats
- Standardizing categories and labels
- Logical checks

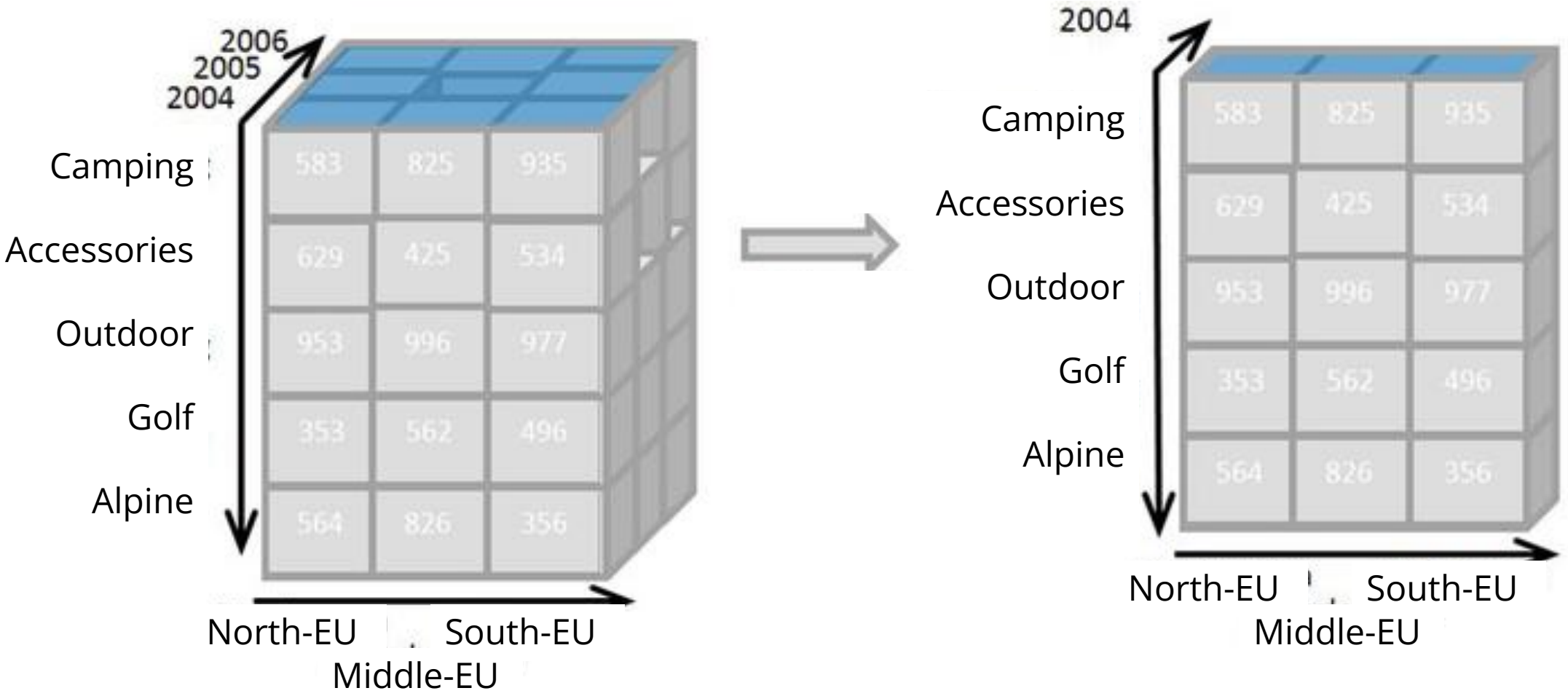
- Tools: Python (pandas, NumPy), OpenRefine, Excel, Apache NiFi

Example data cleaning

- Uniqueness
- Mixed columns (weight): kg and Unit are incompatible
- Multiple data in one cell (status)
  - Multiple statuses
  - State machine
- Uninterpreted textual description

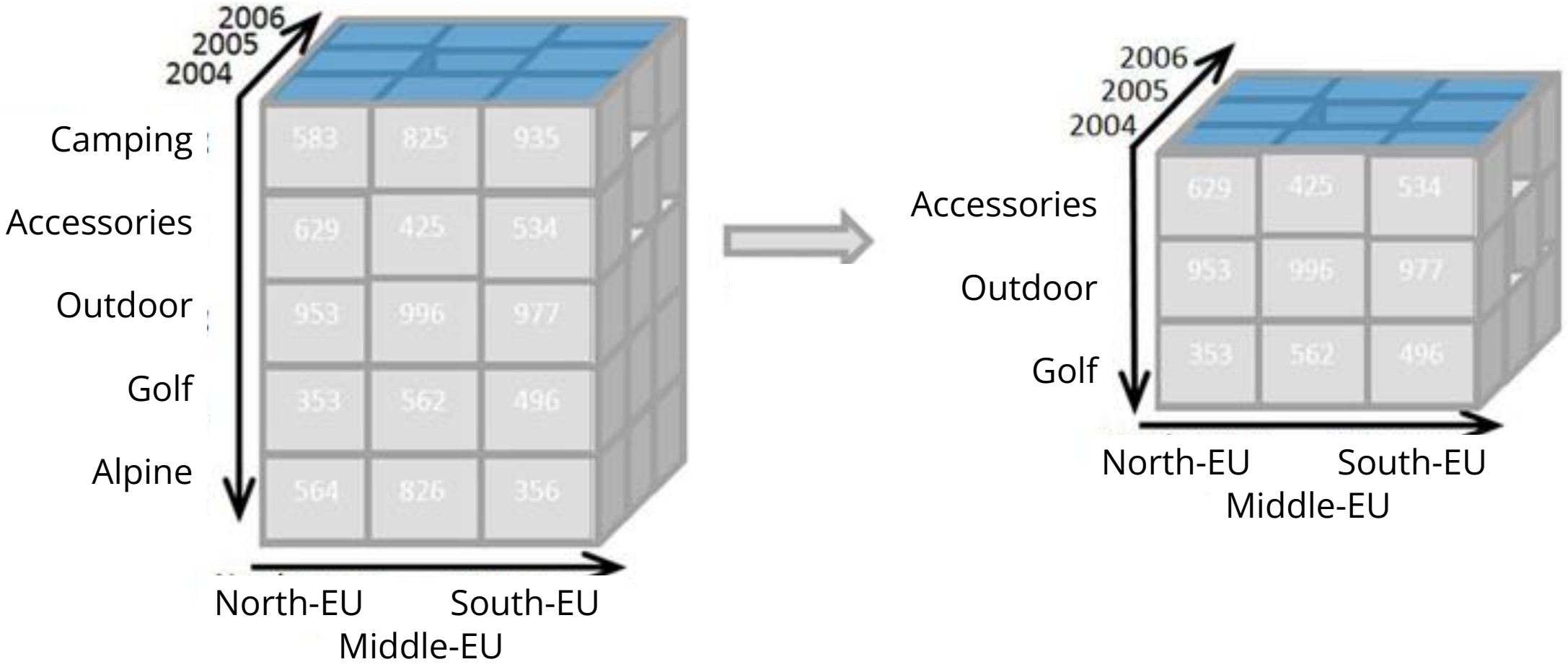
Mission name	Organisation	Nation	Type (U/mass)	Launch date	Status	Mission description	Photo
TUBSAT-N	TU Berlin	Germany	8 kg	1998-07-07	Reentered, Was operational	Store and forward communication.	
TUBSAT-N1	Technische Unviversiat Berlin	Germany	2U	1998-07-07	Reentered, Was operational	Store and forward communication.	
Artemis JAK (MASat, Picosat 5)	Santa Clara University	US	0.2 kg	2000-01-27	No signal	Simple beacon transmitter.	

Slicing



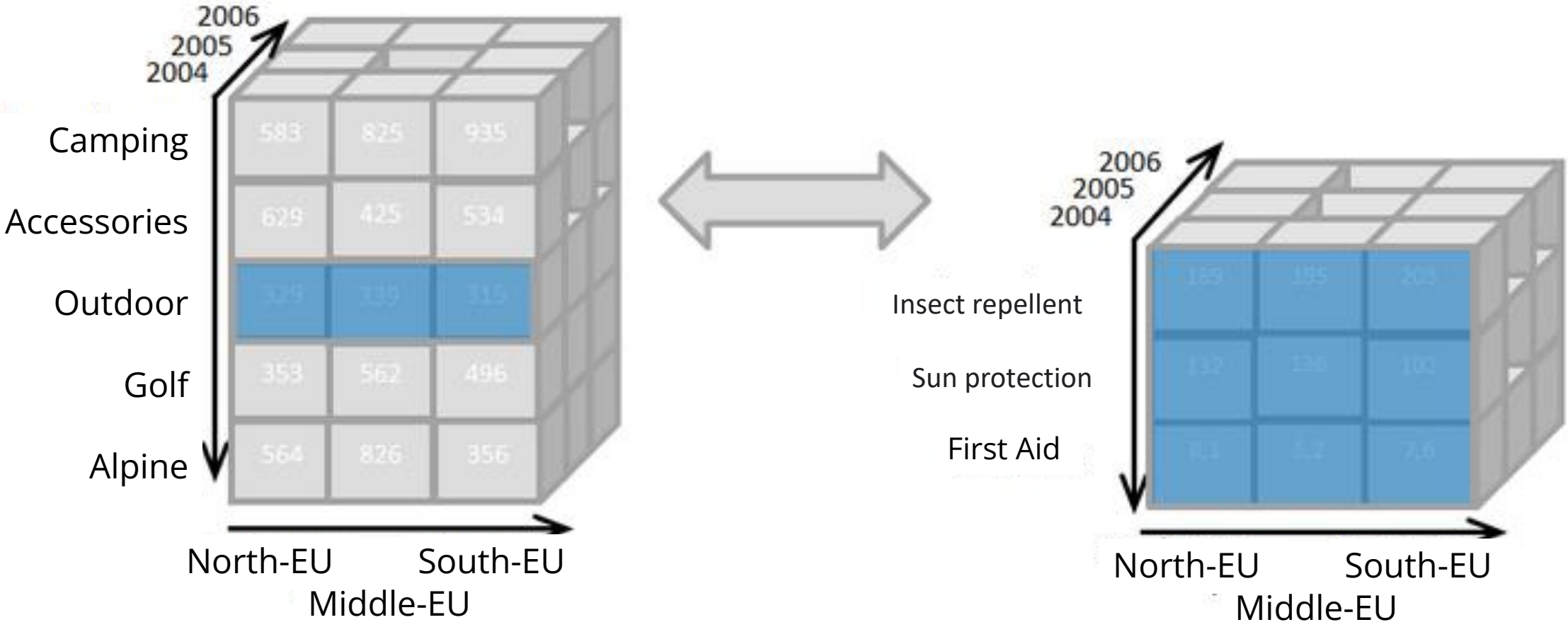
(img: Wikipedia)

Dicing



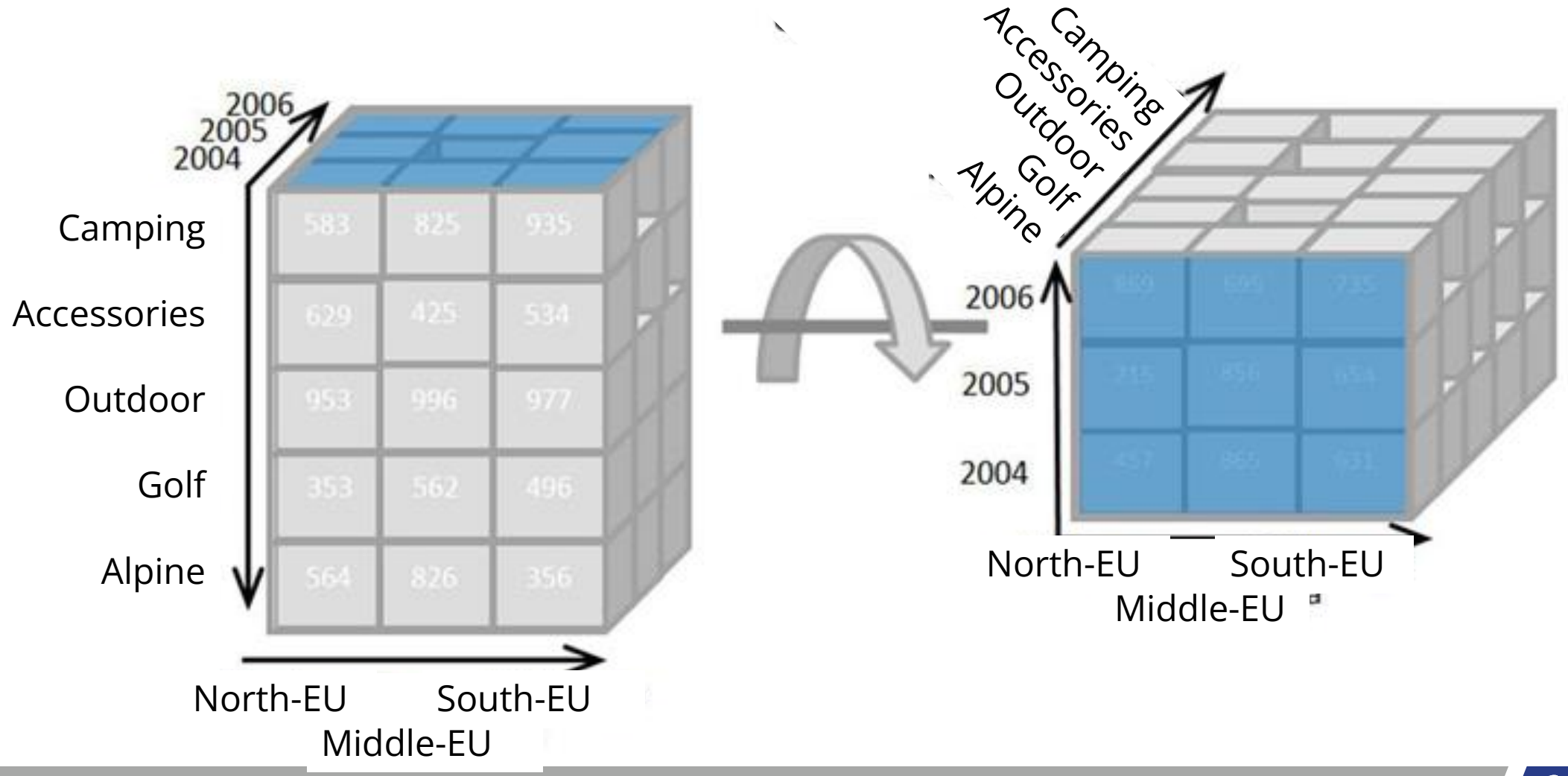
(img: Wikipedia)

Drill up & down



(img: Wikipedia)

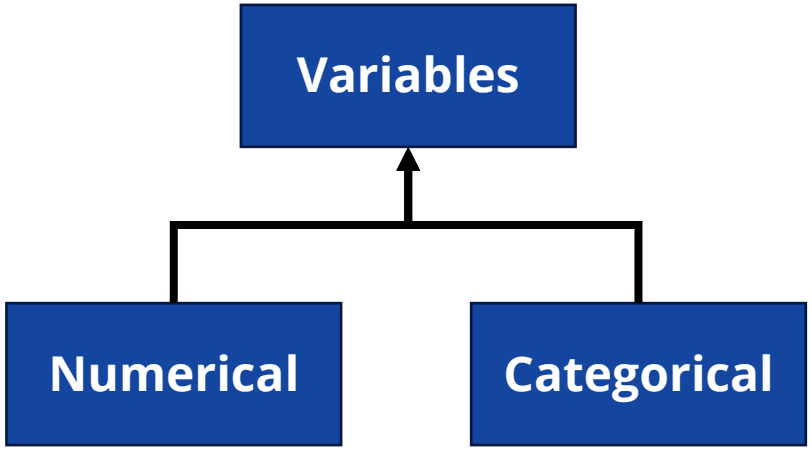
Pivoting



(img: Wikipedia)

Numerical and Categorical Variables

- **Numerical**
  - Basic arithmetic operations are interpretable.



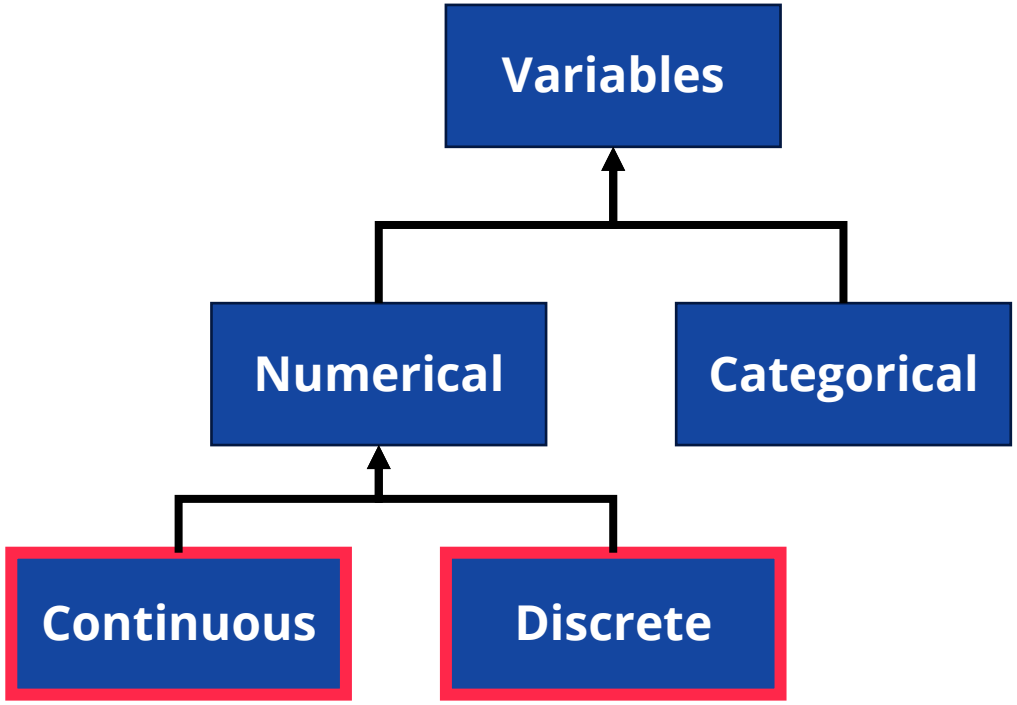
- **Categorical**
  - Mathematical operations are not interpretable; at most, they can be ordered.

age, average temperature, phone number

Numerical Variables

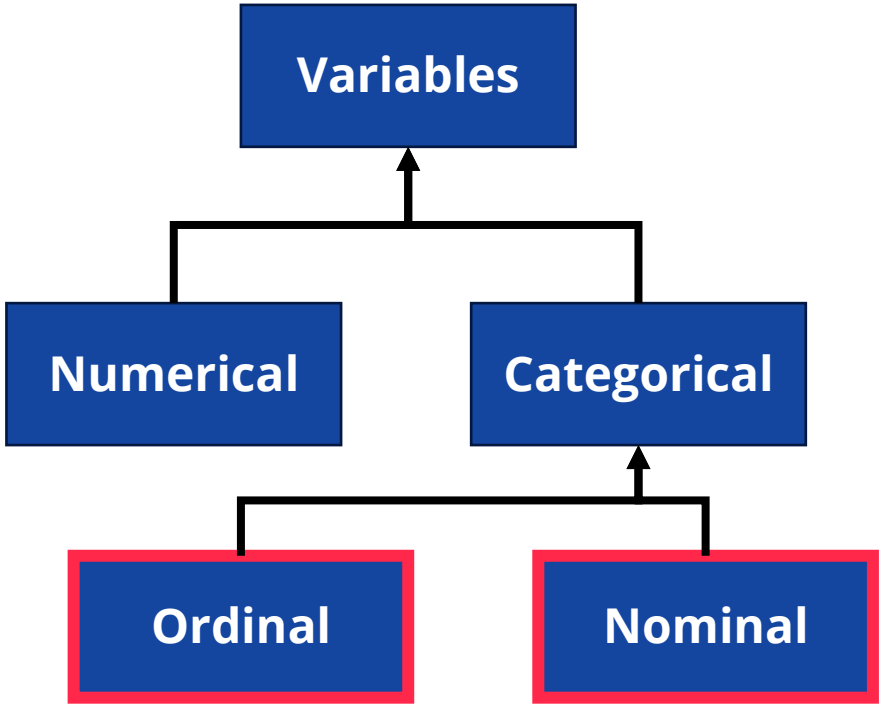
- Continuous
  - Measured – can take any value
    - within given range
    - with given precision
  - Pl. the average test score of the people in the room

- Discrete
  - Counted– can take finite number of values within a given range
  - Pl. the number of people attending the lecture



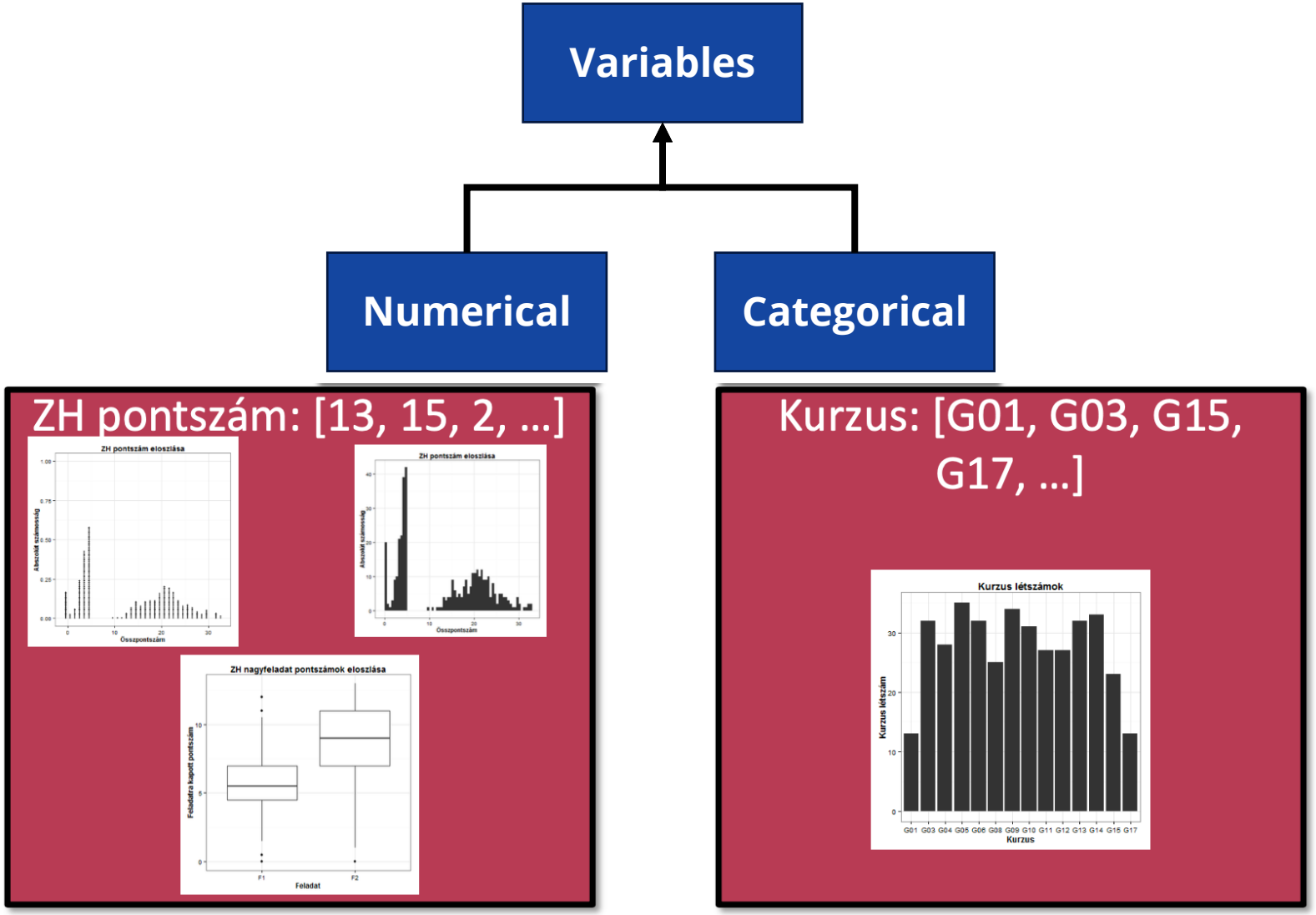
Categorical Variables

- Ordinal
  - Total ordering of values
  - E.g., stars of a hotel
- Nominal
  - E.g., colors, car brands



Data visulaization

1 variable – distributions

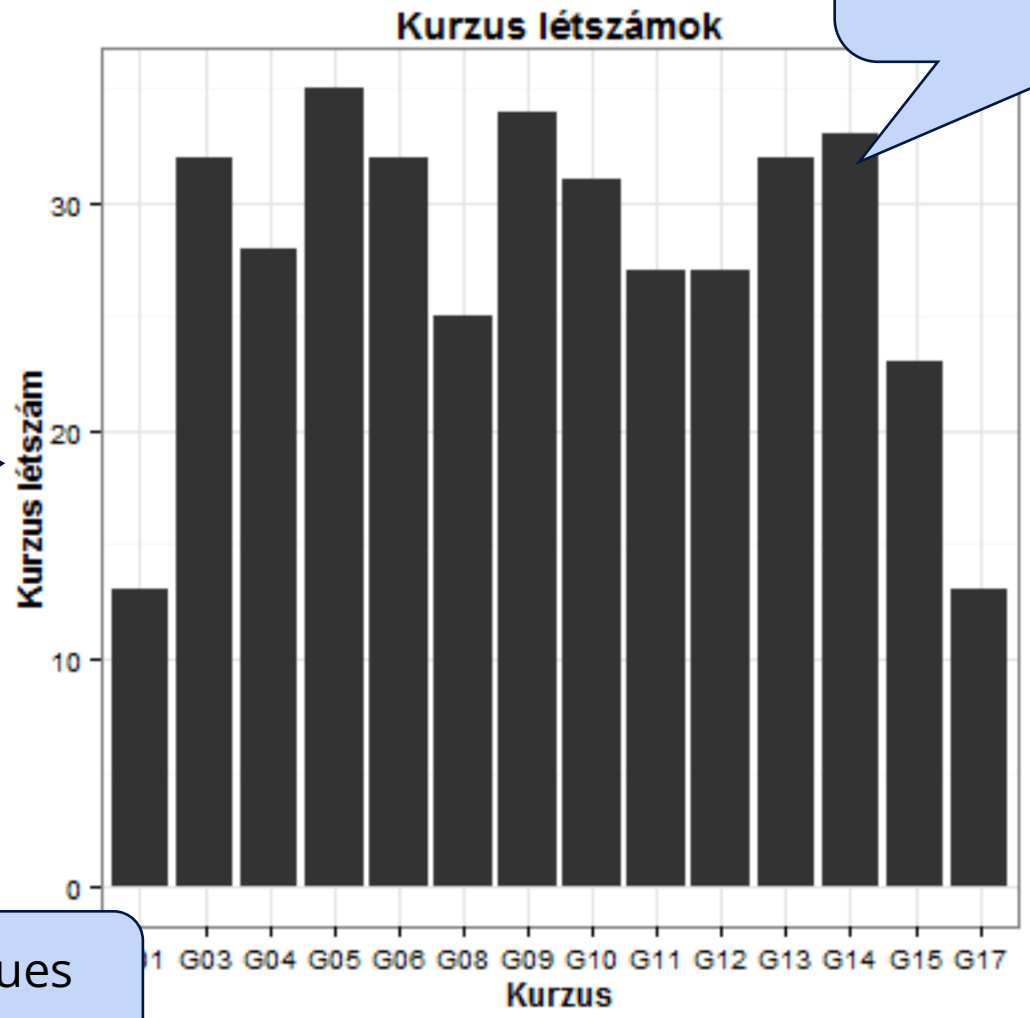


Bar chart

Input variable: course code
Question: how many students are enrolled in each course?

Bar height: frequency of a given value

Absolute frequency

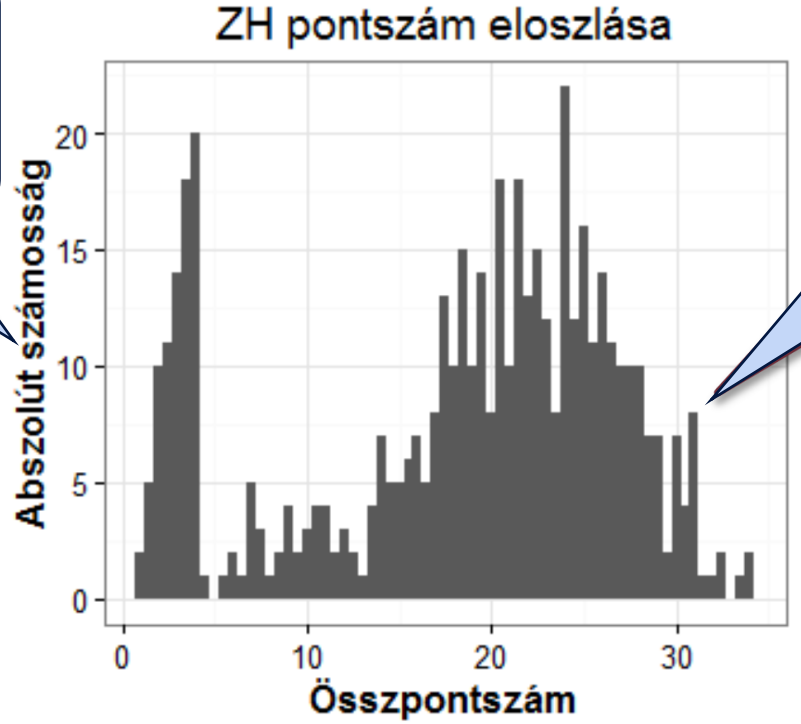


Design decision: breaking down the set of values
Example: Tuesday–Thursday–Friday

Histogram

Input variable: Total score of midterm (ZH)
Question: How did the midterm scores turn out?

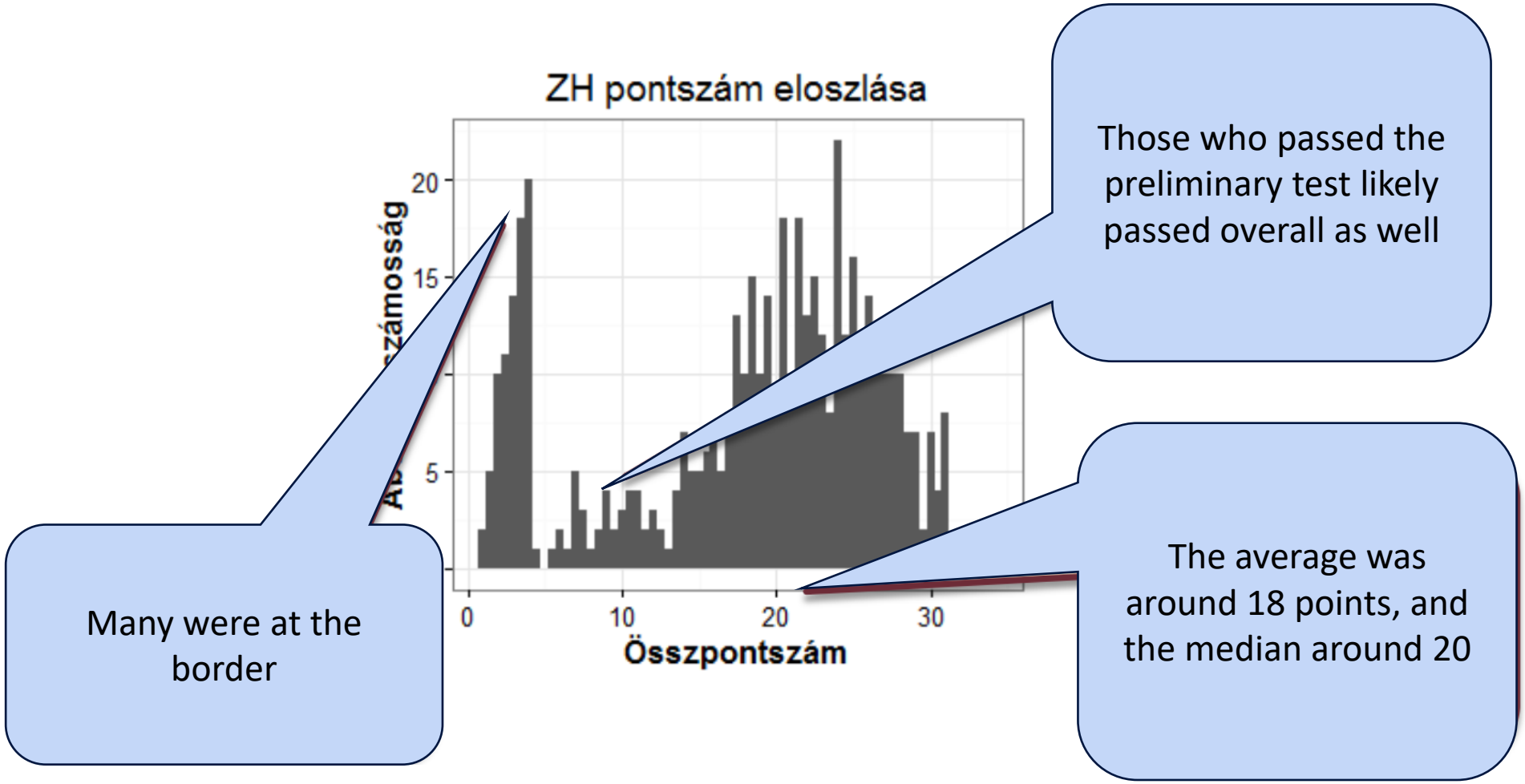
Absolute frequency



Bar height: count of a given interval

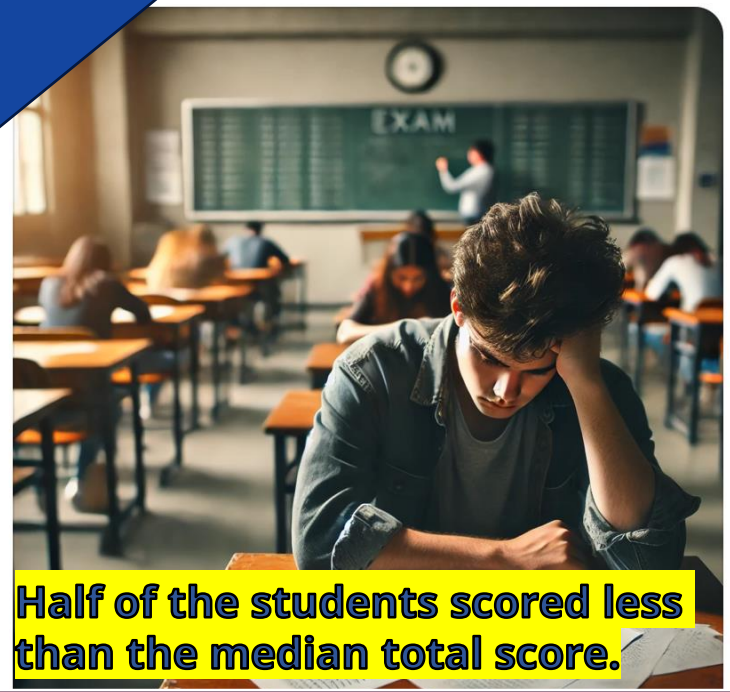
Design decision: how long should the interval be?
Example: is a resolution of 1 point enough, or should we go down to half-point intervals?

Histogram


ZH pontszám eloszlása
20
15
5
0
10
20
30
Összpontszám
Those who passed the preliminary test likely passed overall as well
Many were at the border
The average was around 18 points, and the median around 20

Interesting titles

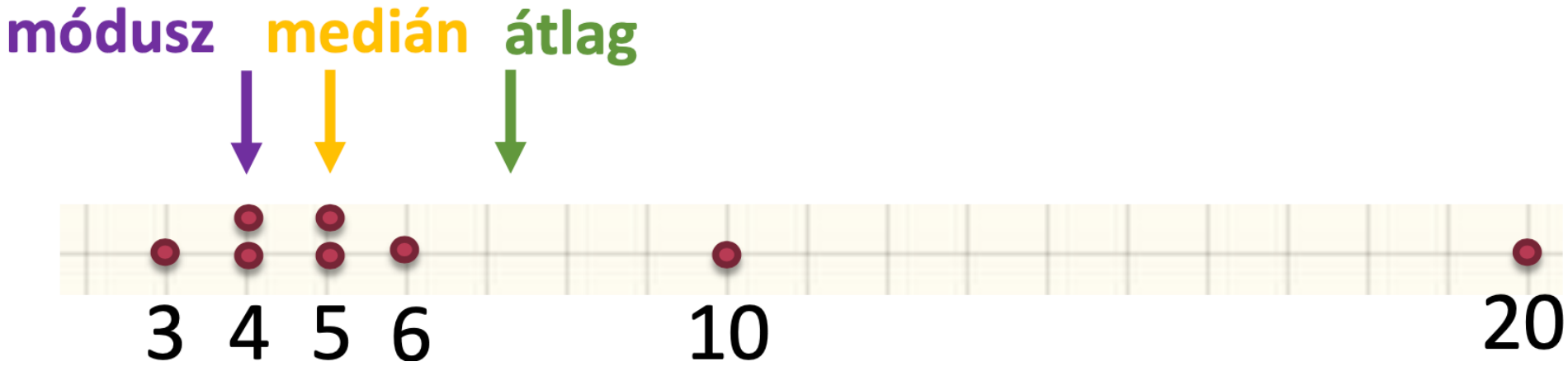
"Half of Hungarians earn less than the median wage."



Bonus question: *What is wrong with the picture?*

Mean, median, mode – the “middle”

- {3, 4, 4, 5, 5, 6, 10, 20}



Mode and Median

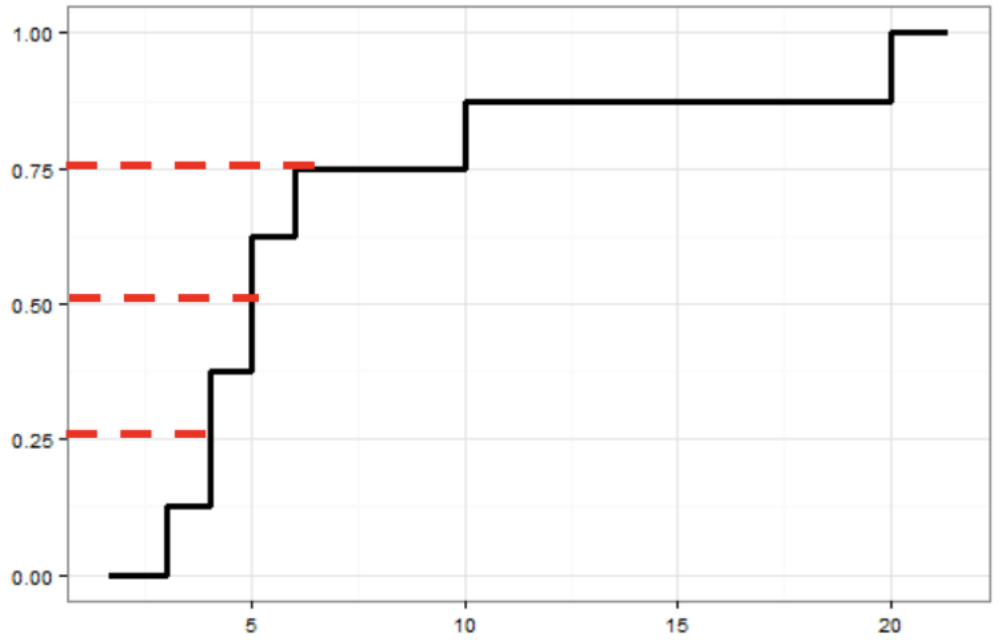
- If we arrange the values in ascending order, the middle value is the **median** of the dataset. If there is no middle value (in the case of an even number of values), the median is the average (arithmetic mean) of the two middle values.

- The **mode** is the value that occurs most frequently in the dataset. This is not necessarily unique; in such cases, we talk about multiple modes.

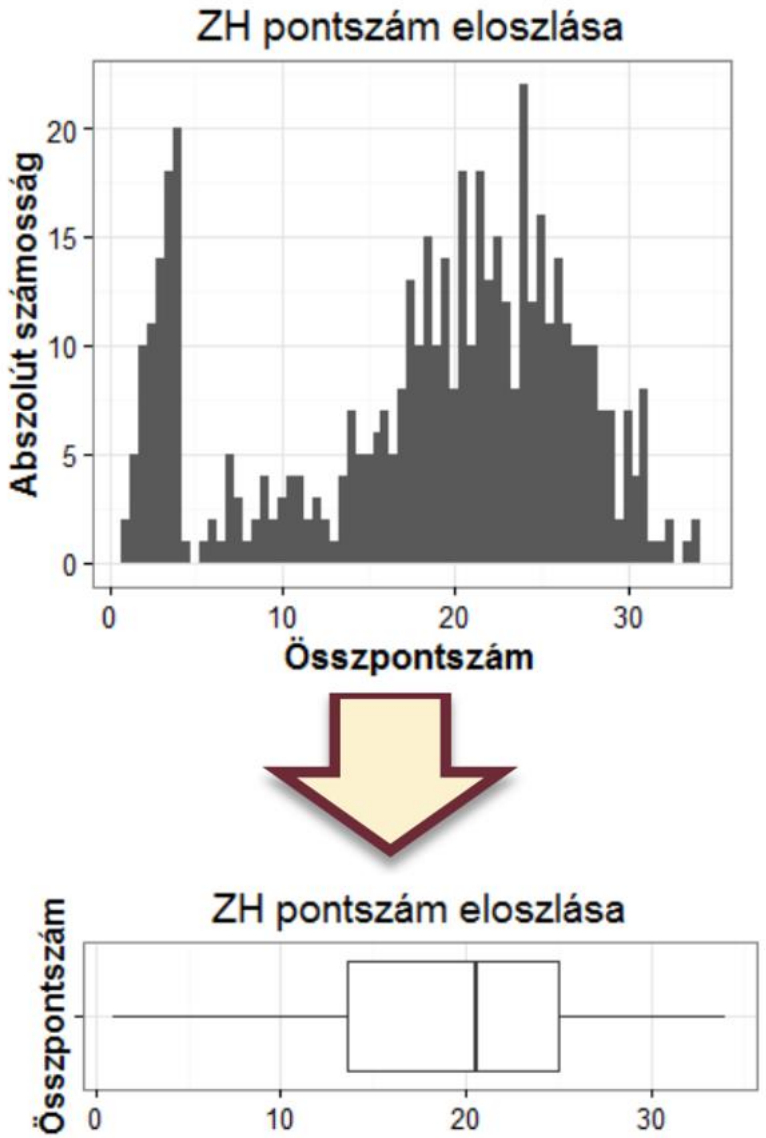
Percentile

- **At the n-th percentile, n% of the values are smaller.**

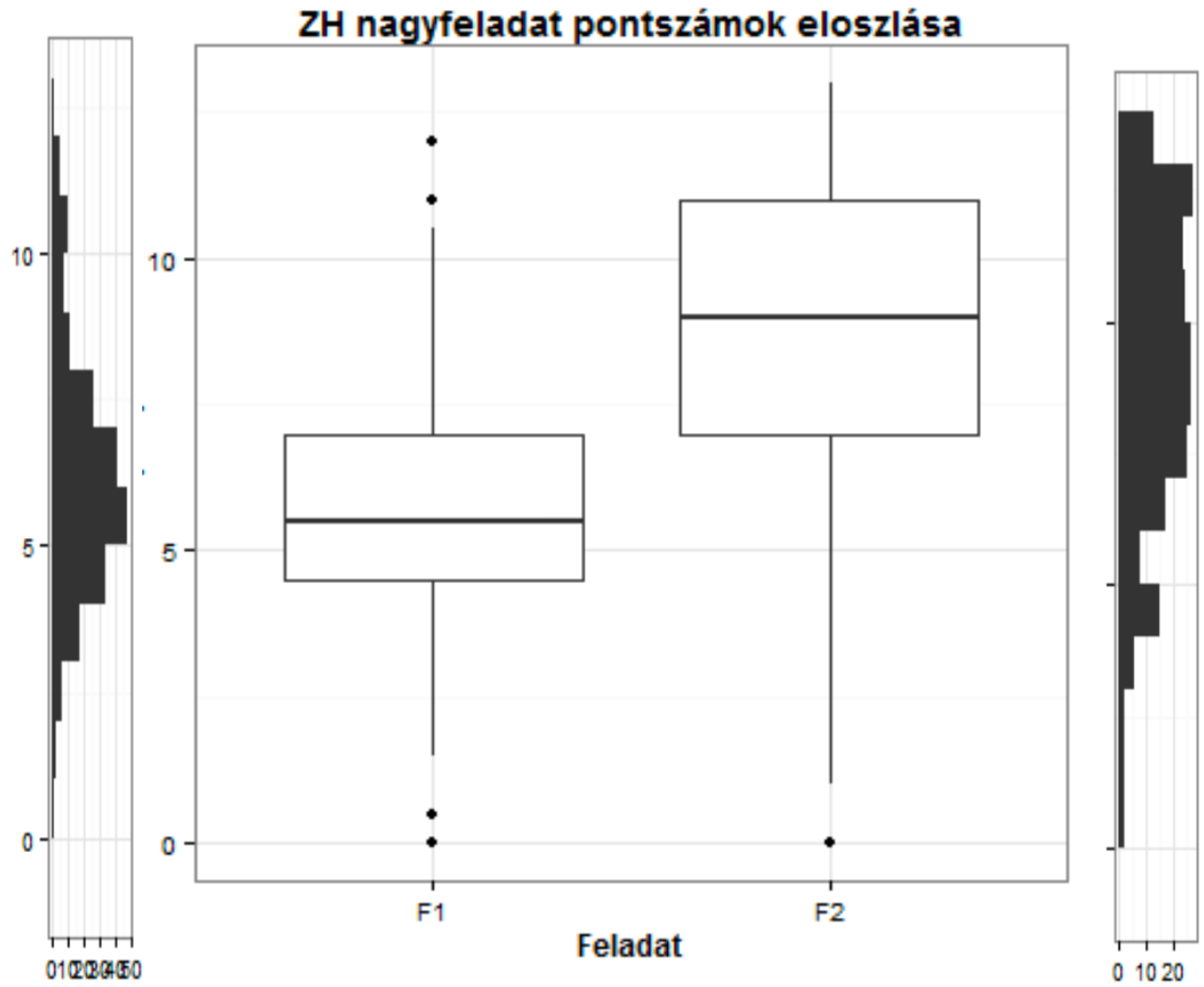
- **Percentile**
  - {3, 4, 4, 5, 5, 5, 6, 10, 20}
    - 50th percentile: 5
    - 25th percentile: 4
    - 75th percentile: 6
- **Quartile**
  - Q1: 25th percentile
  - Q3: 75th percentile
  - Q2: median



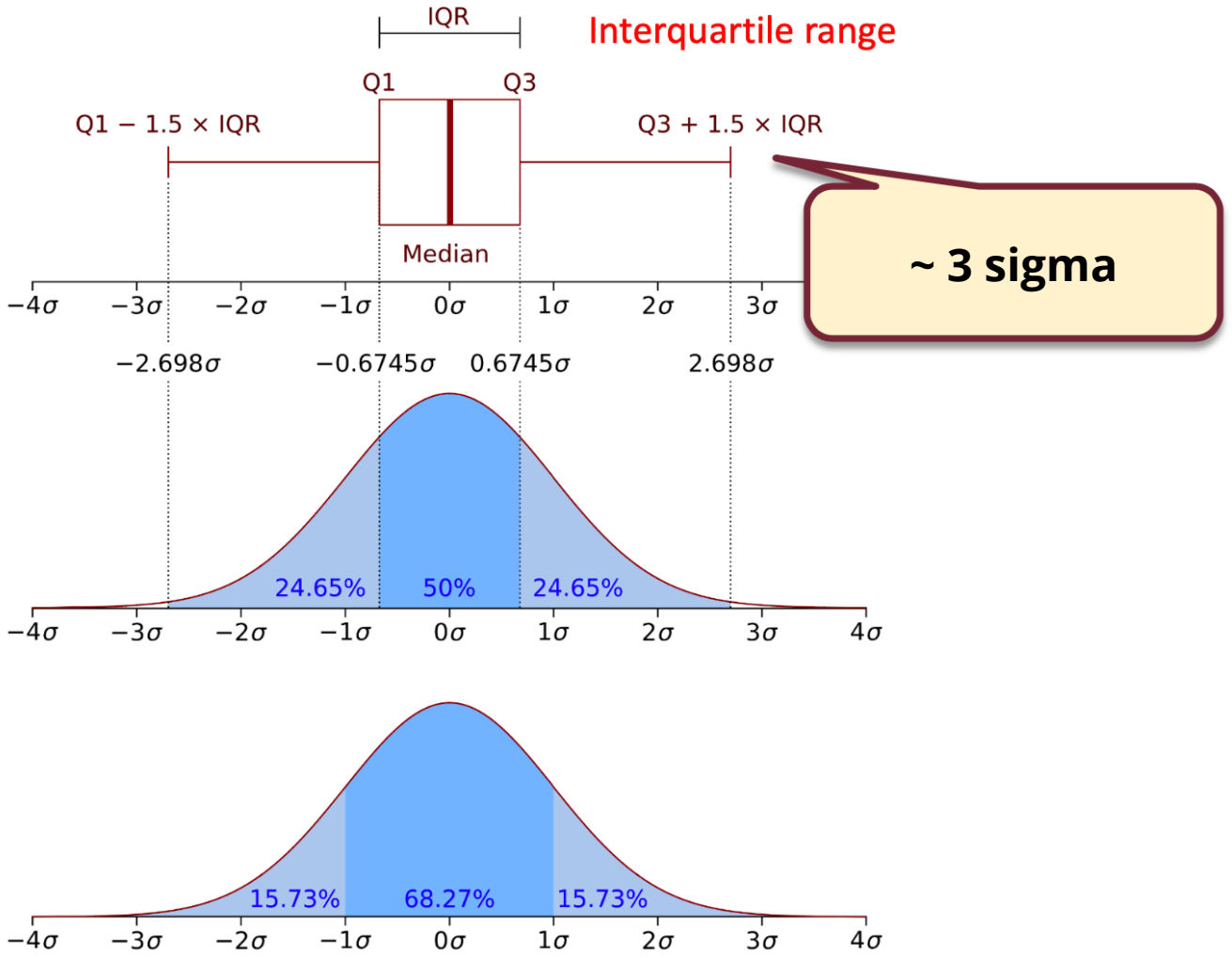
Boxplot


ZH pontszám eloszlása
Abszolút számosság
0
5
10
15
20
0
10
20
30
Összpontszám
ZH pontszám eloszlása
Összpontszám
0
10
20
30

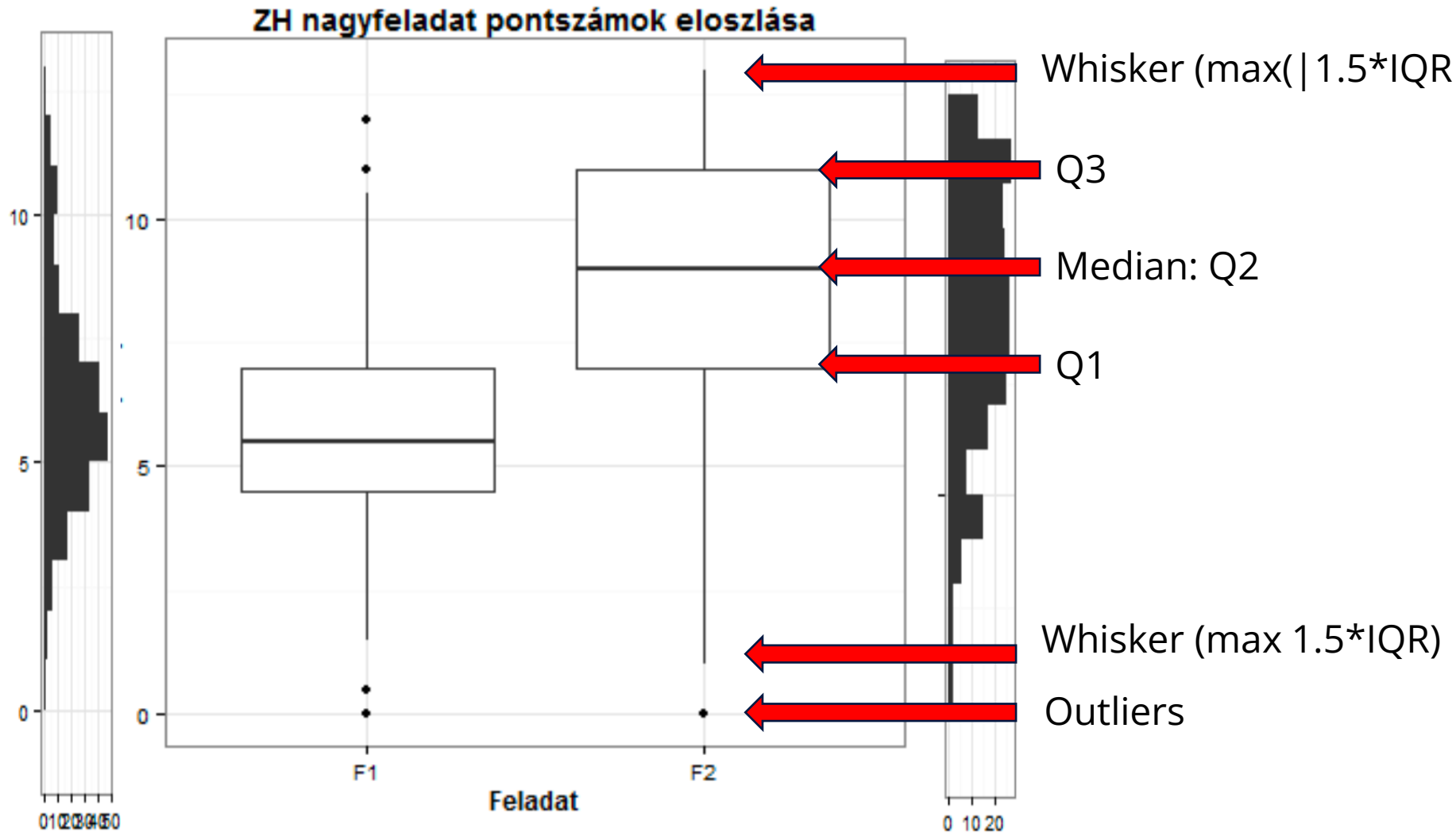
Boxplot


ZH nagyfeladat pontszámok eloszlása
10
5
0
F1
F2
Feladat
0 10 20

IQR

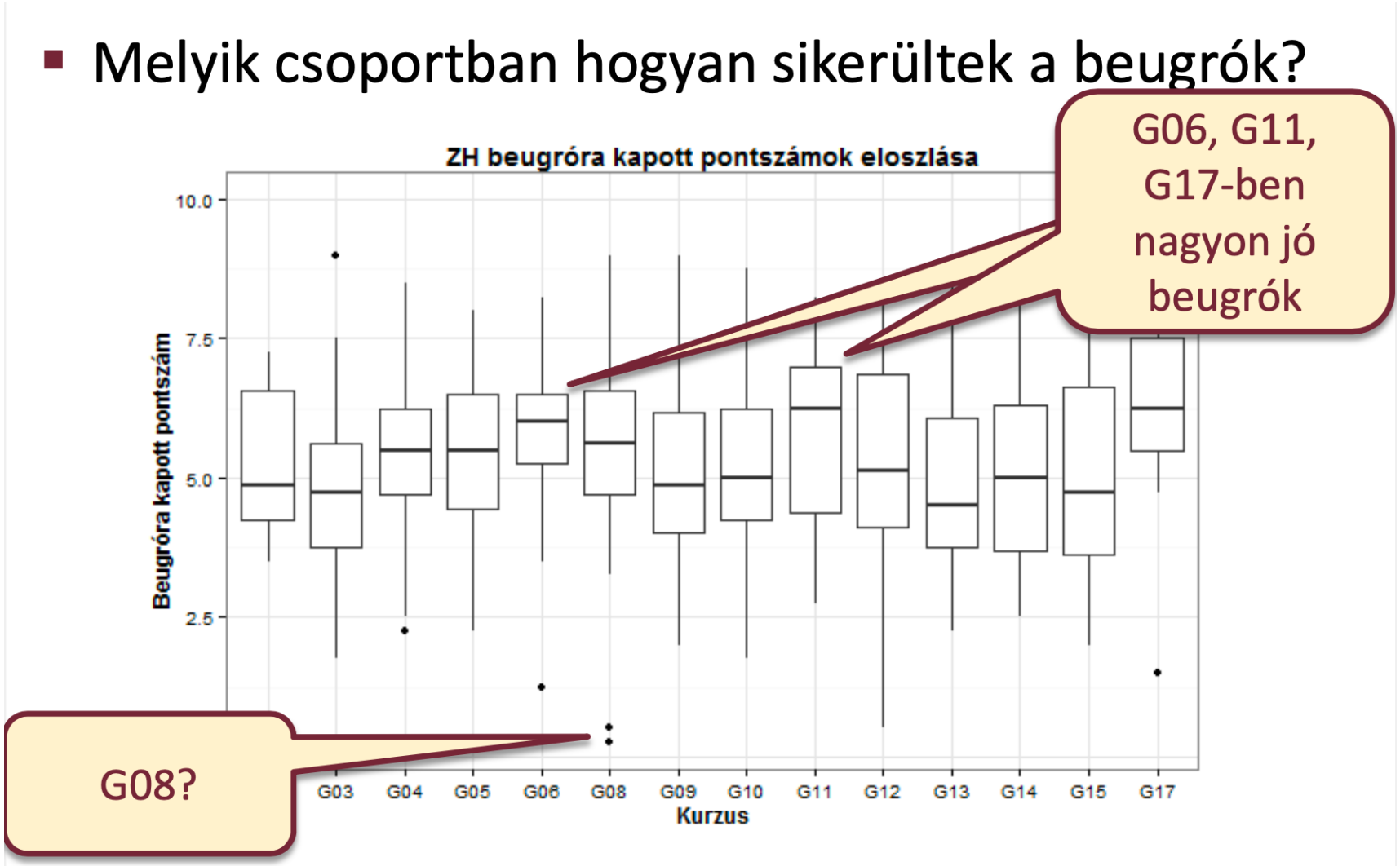

IQR
Interquartile range
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
~ 3 sigma
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73%
68.27%
15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

Boxplot


ZH nagyfeladat pontszámok eloszlása
10
5
0
F1
F2
Feladat
0 10 20
Whisker (max(|1.5*IQR|))
Q3
Median: Q2
Q1
Whisker (max 1.5*IQR)
Outliers

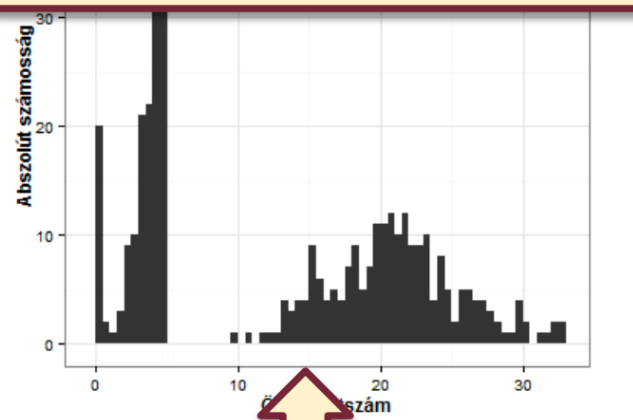
Boxplot example

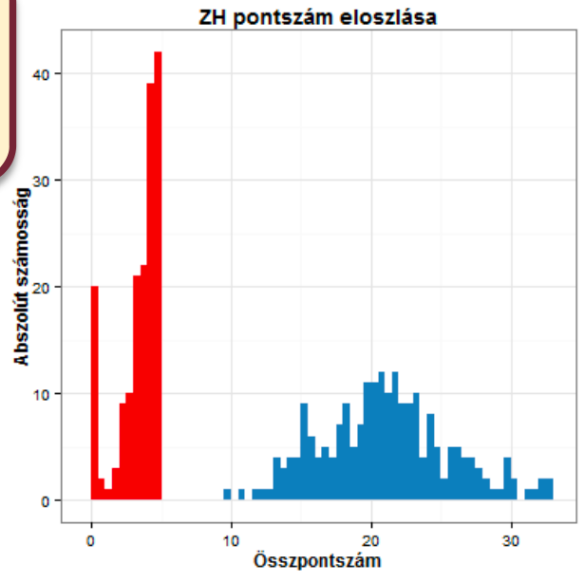
- Melyik csoportban hogyan sikerültek a beugrók?

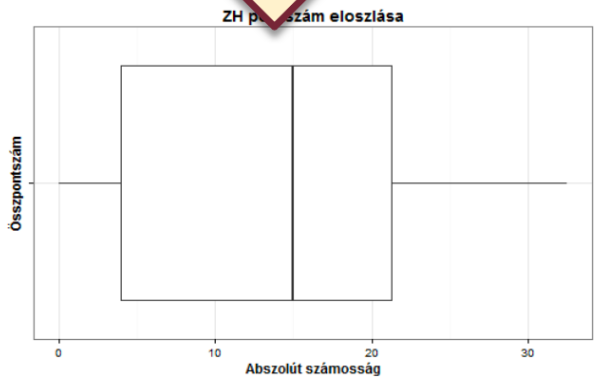


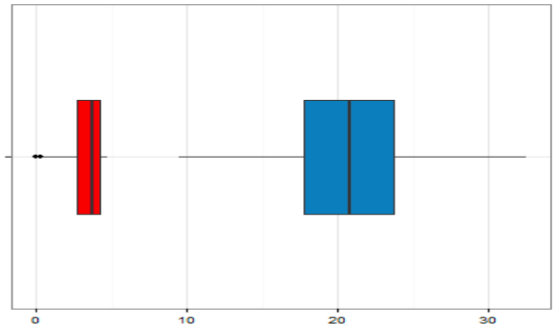
Boxplot abstraction

Abstraction: with a boxplot, we can lose important information!



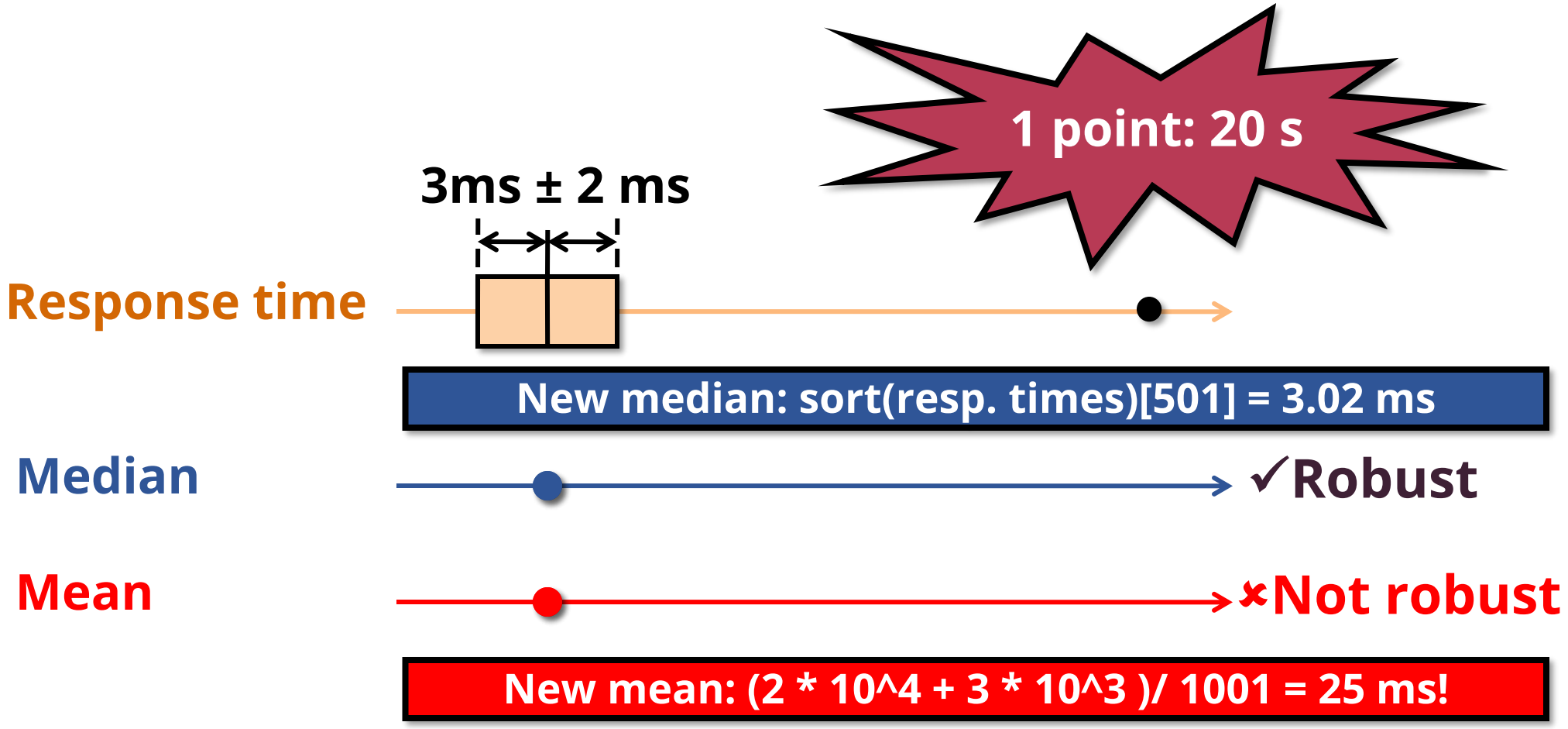




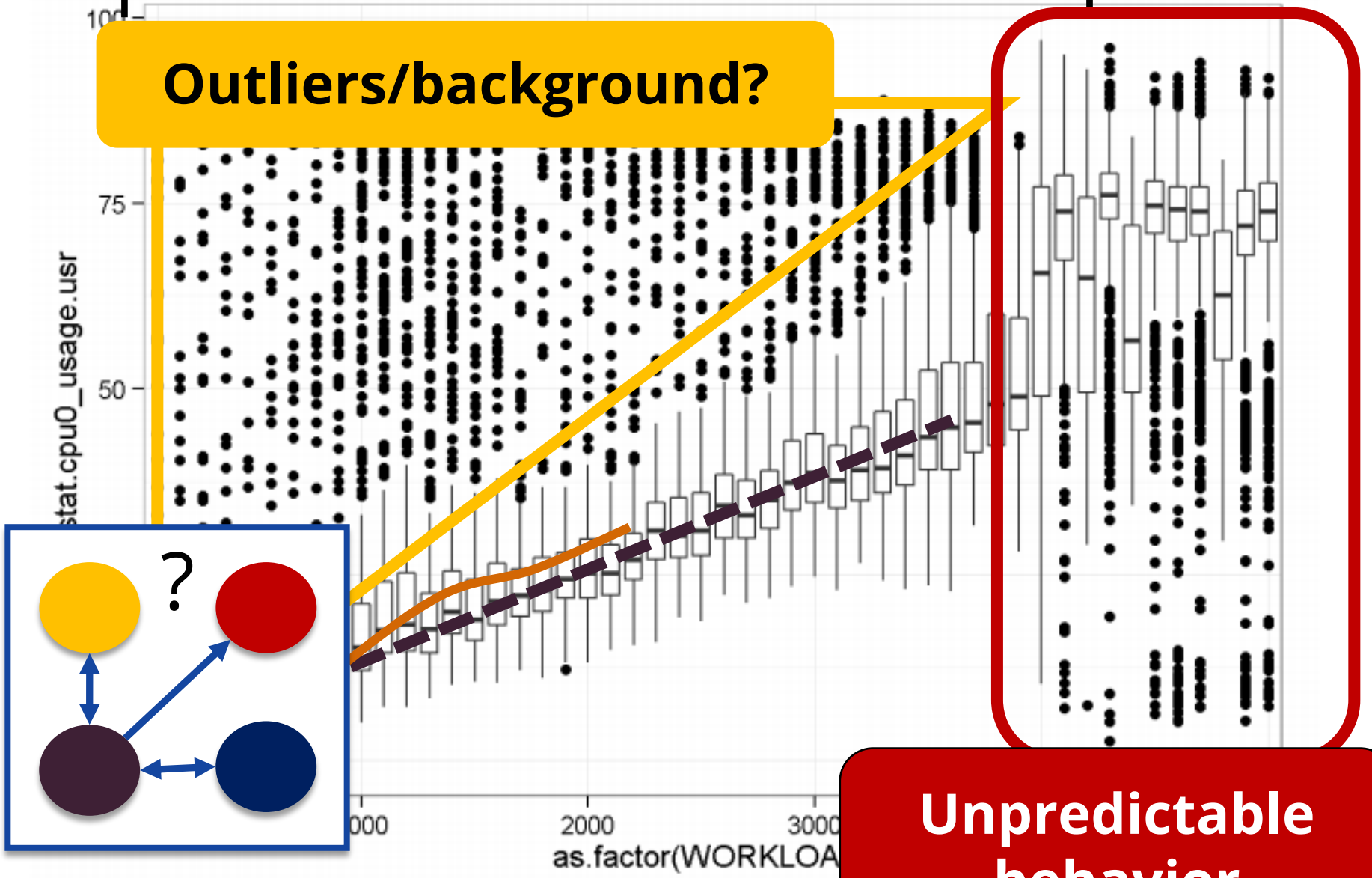


Robust statistics

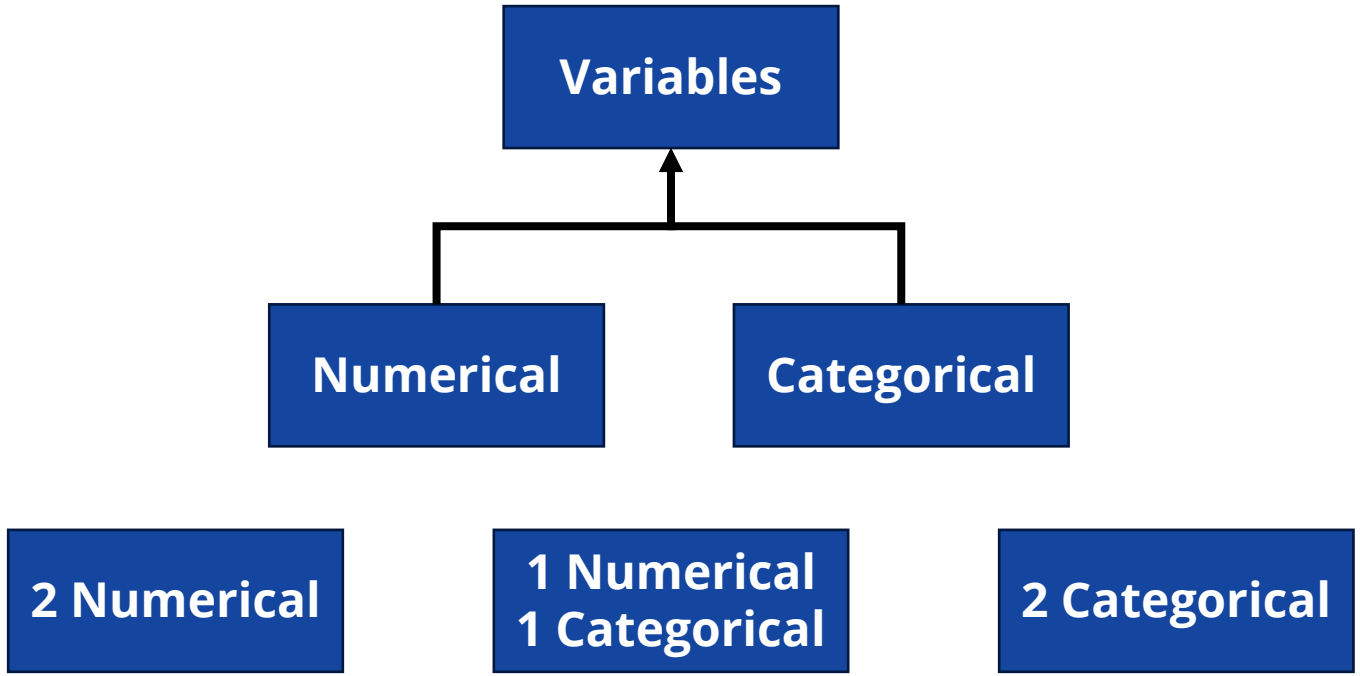
- Base population: 1000 samples from U(1, 5)



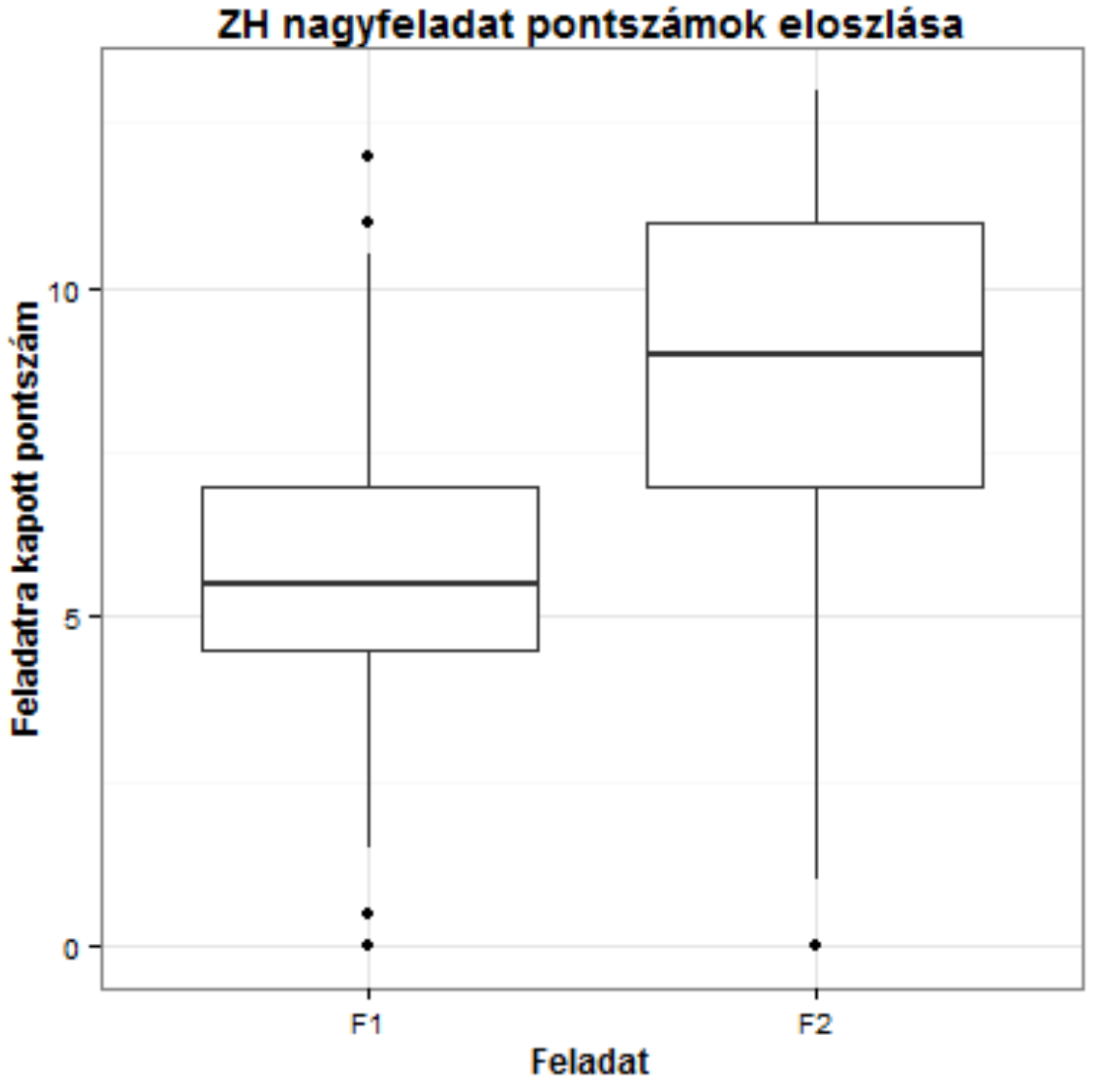
Some qualitative fun with boxplots



Visualization of 2 variables



Numerical by category

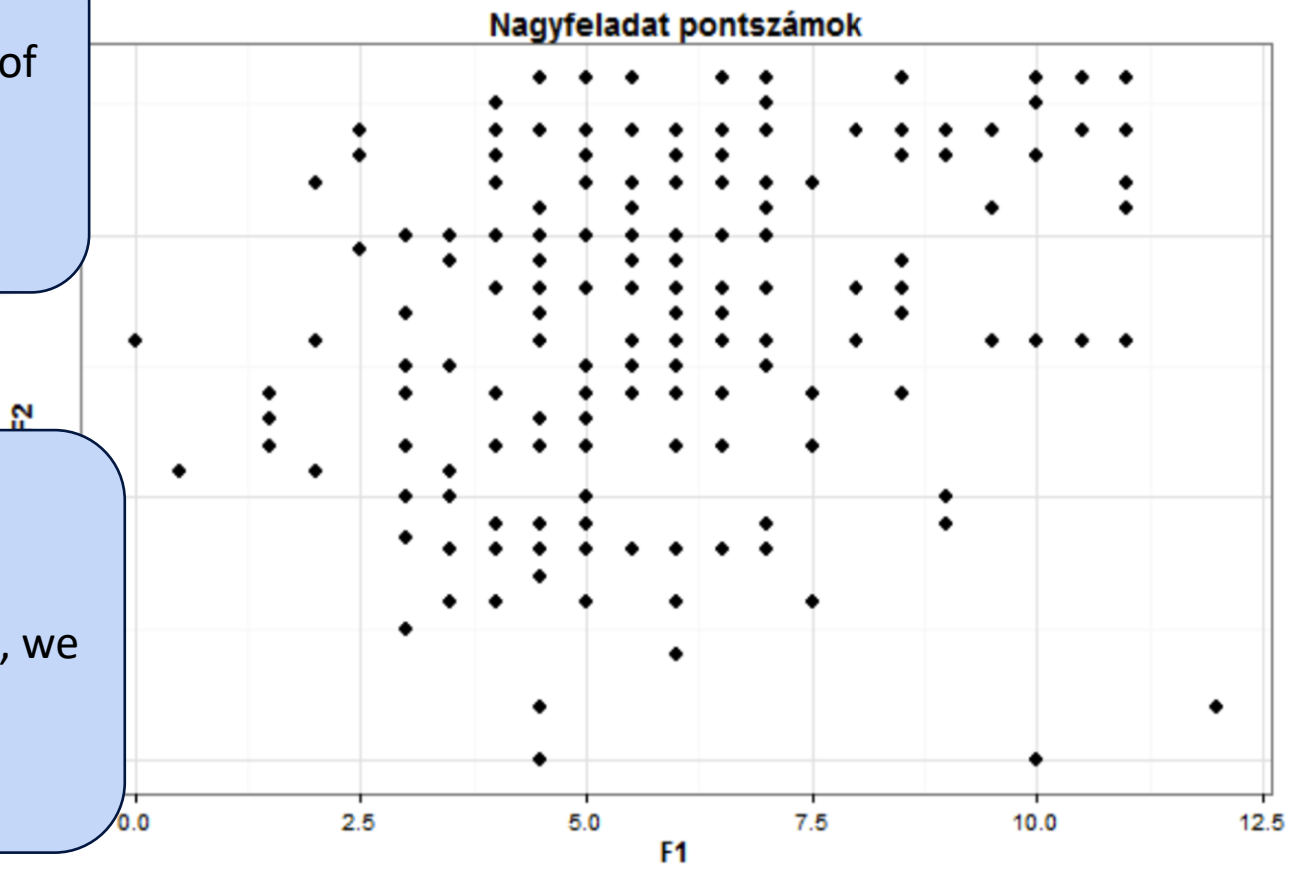


Scatterplot

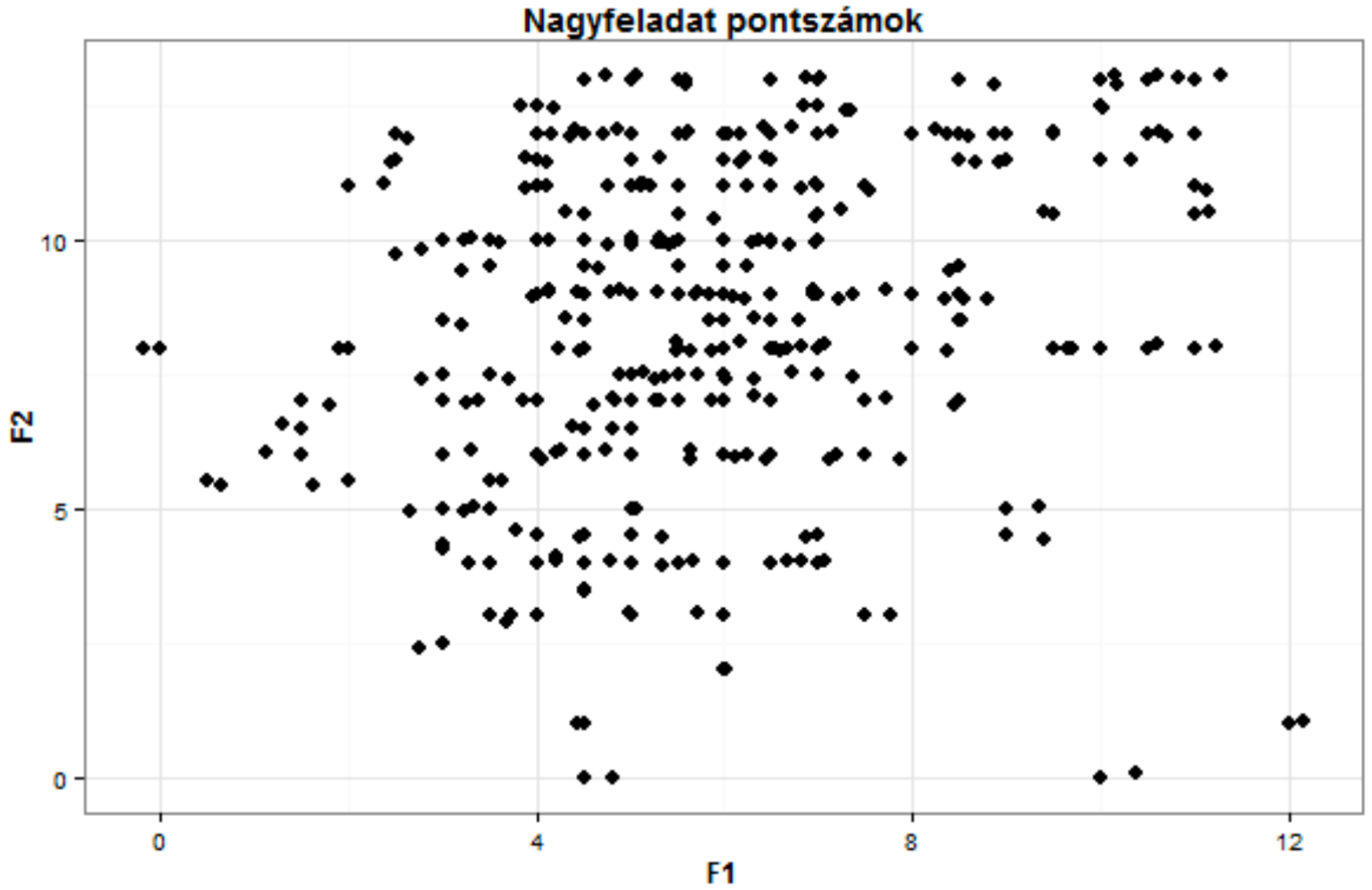
- **Input variable**: points received on tasks
- **Question**: how do they relate to each other?

We visualize pairs of points that occur together

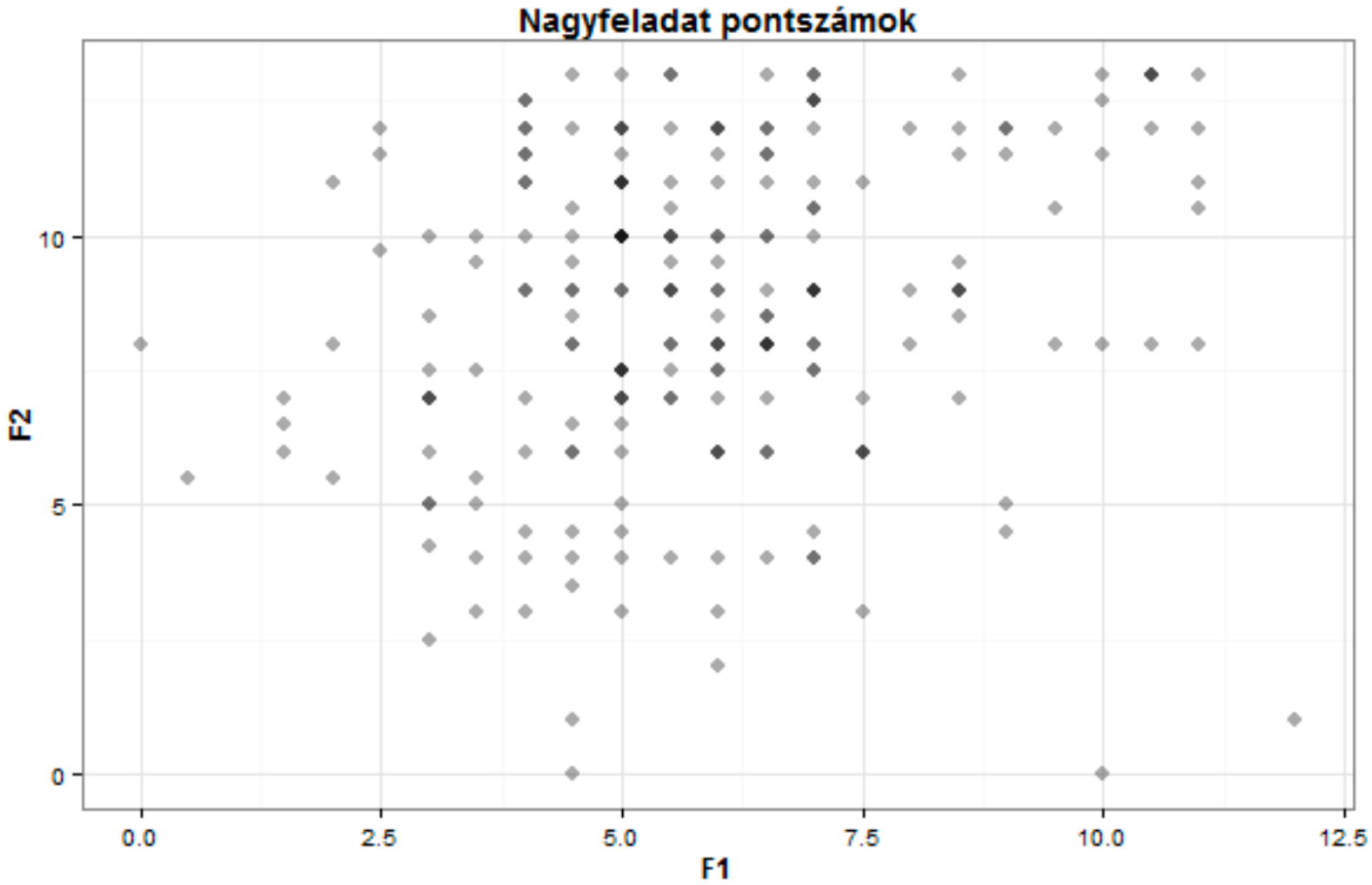
If the value of one variable is missing, we cannot plot it



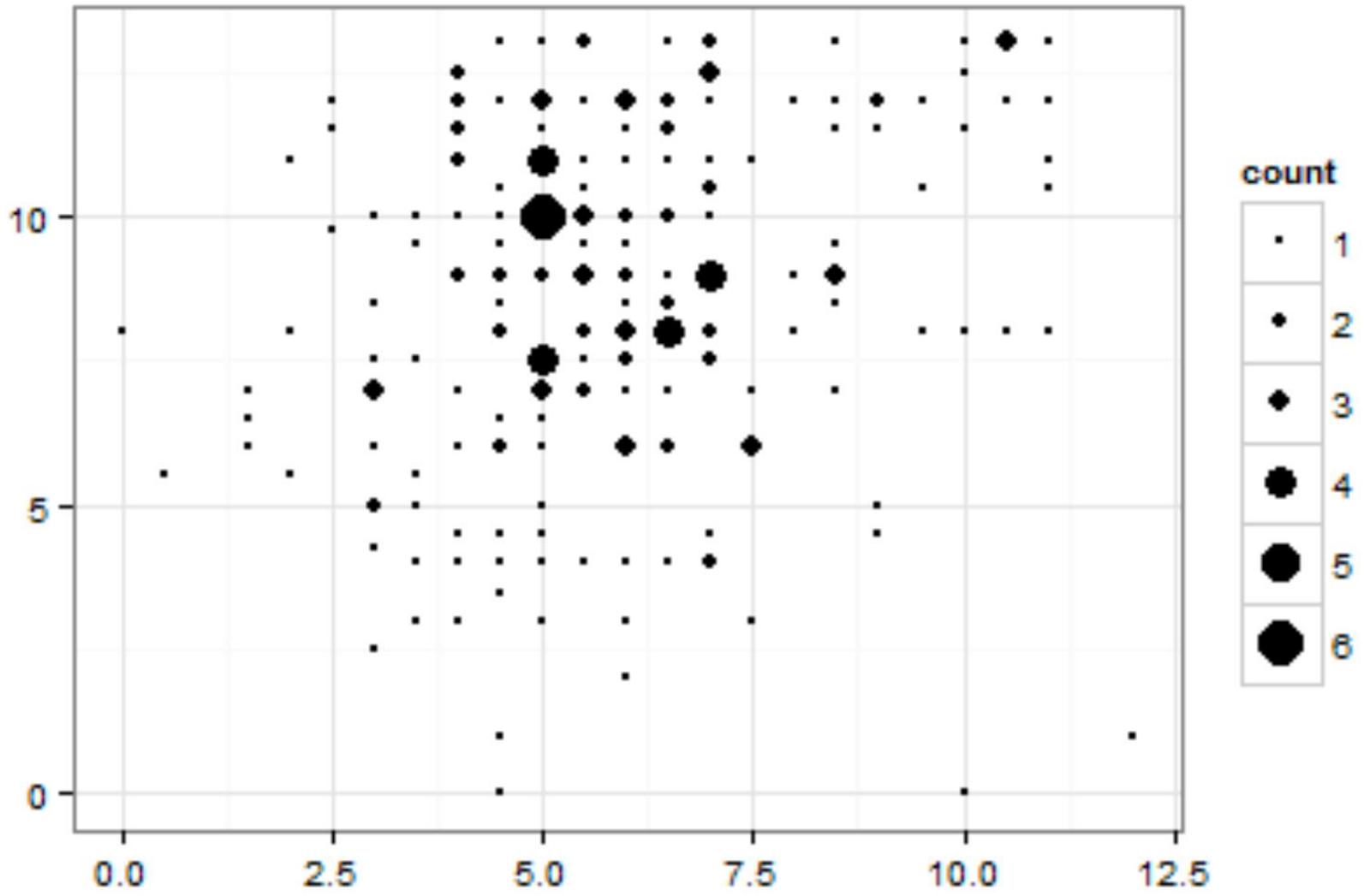
Jitter



Transparency

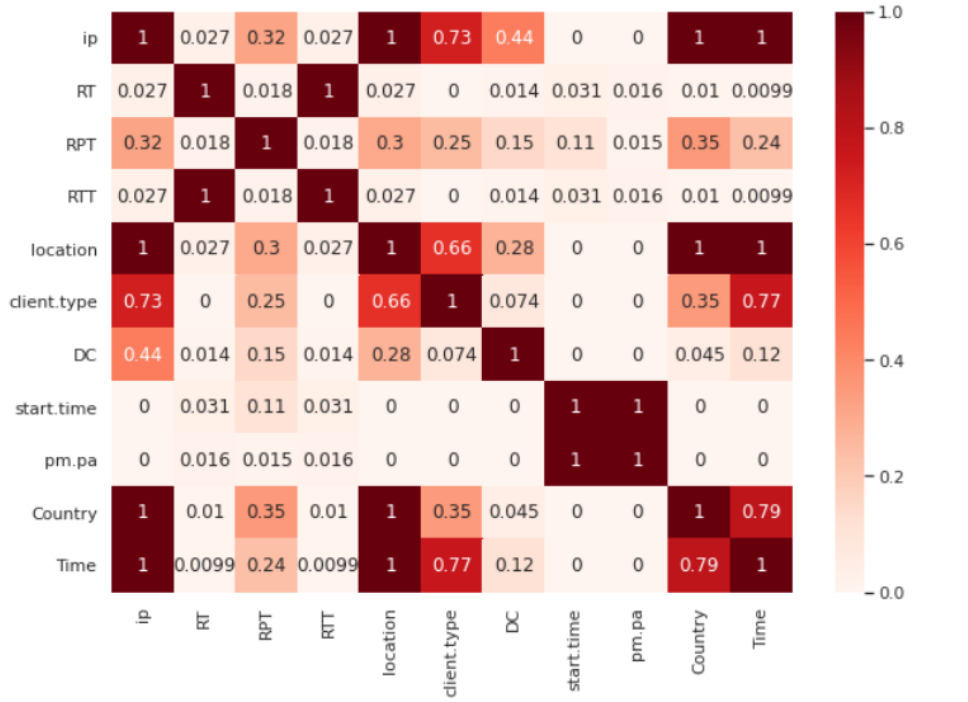


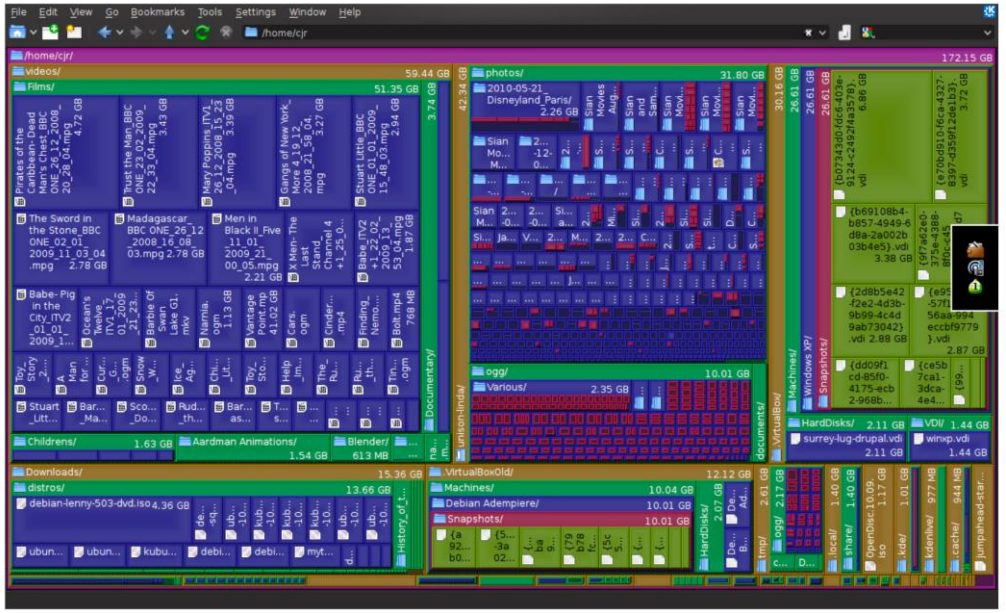
Size



>= 3 variable

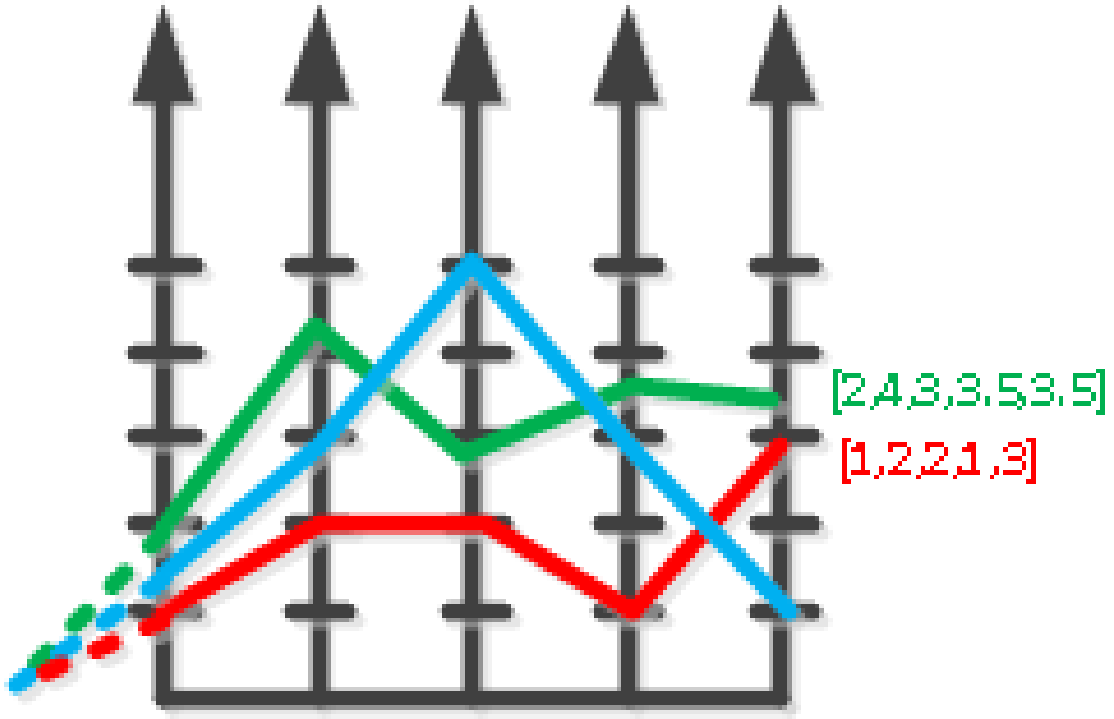
- Modifying the attributes of graphical objects
  - Color
  - Size
  - Texture
  - Position – this may seem trivial, but it is important in a treemap
  - PI. heatmap, treemap





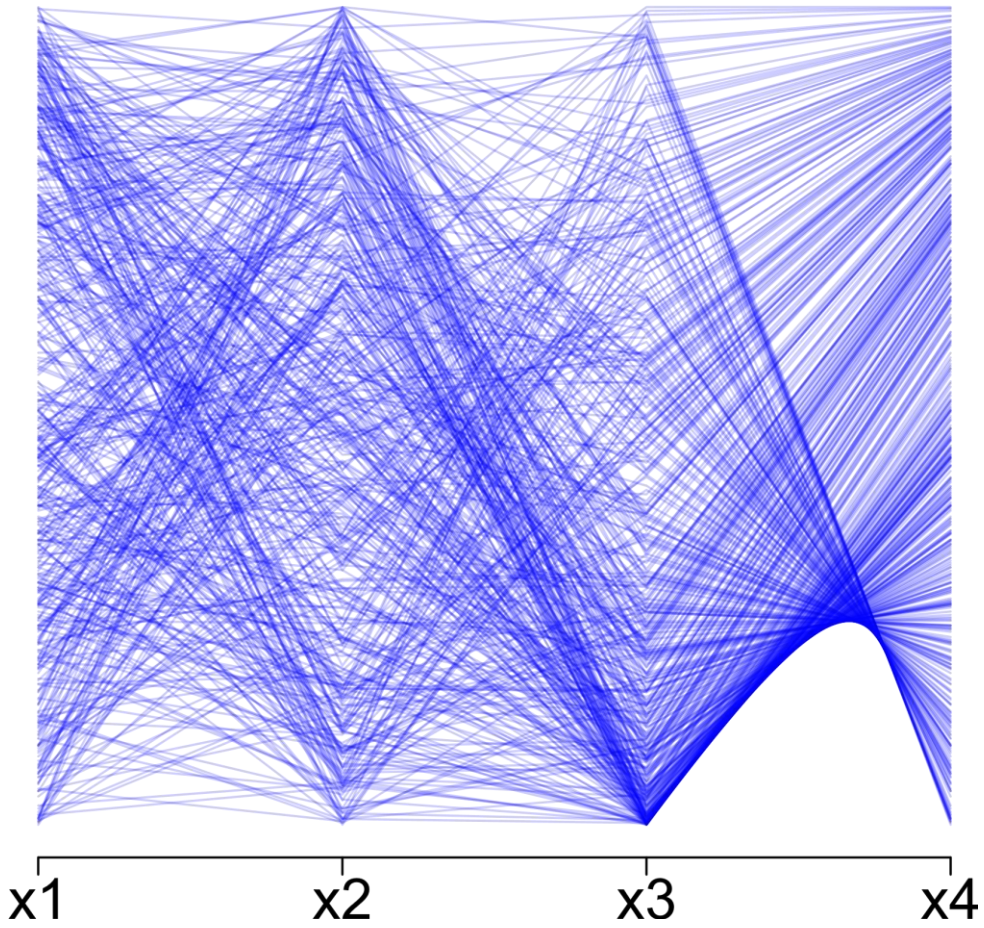
Parallel coordinates

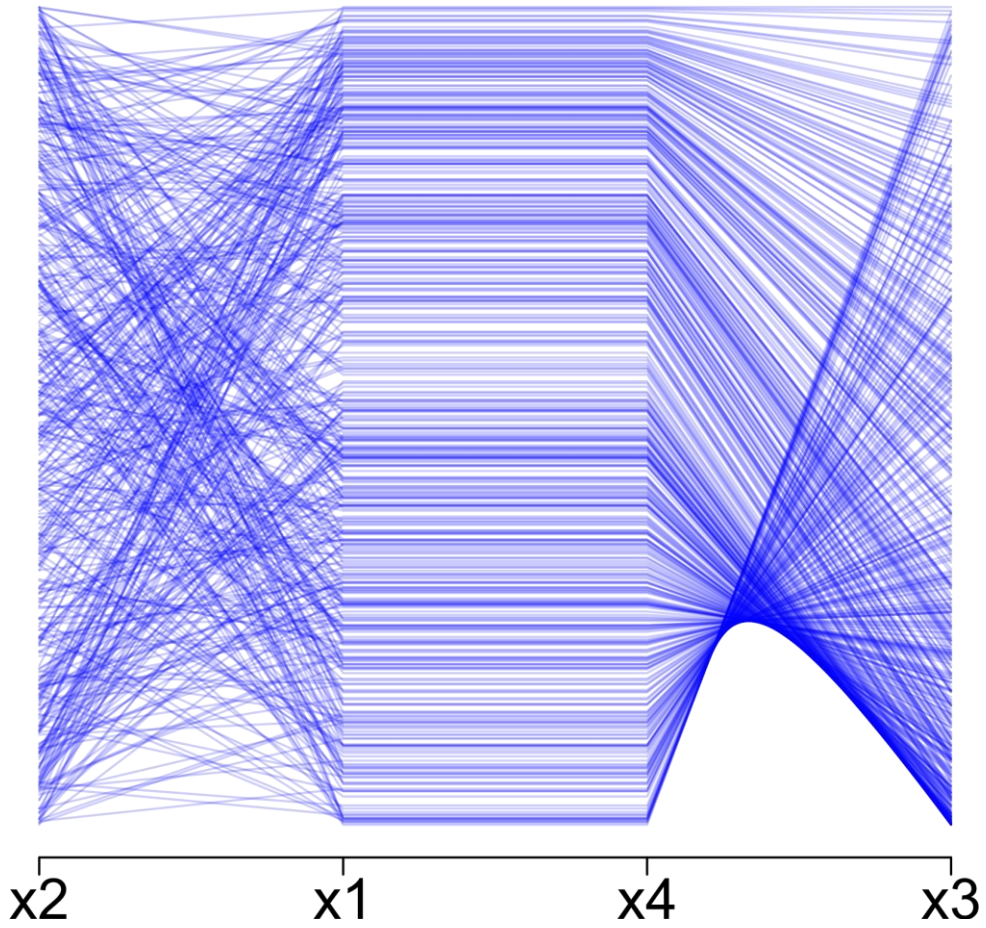
- The trained eye can…
  - Remove redundancy
  - Recognize linear correlation
  - Coplanarity and other structures
  - Clustering
  - …
- Variable order is crucial



Inselberg, A: Parallel Coordinates, Visual Multidimensional Geometry and Its Applications, Springer 2009

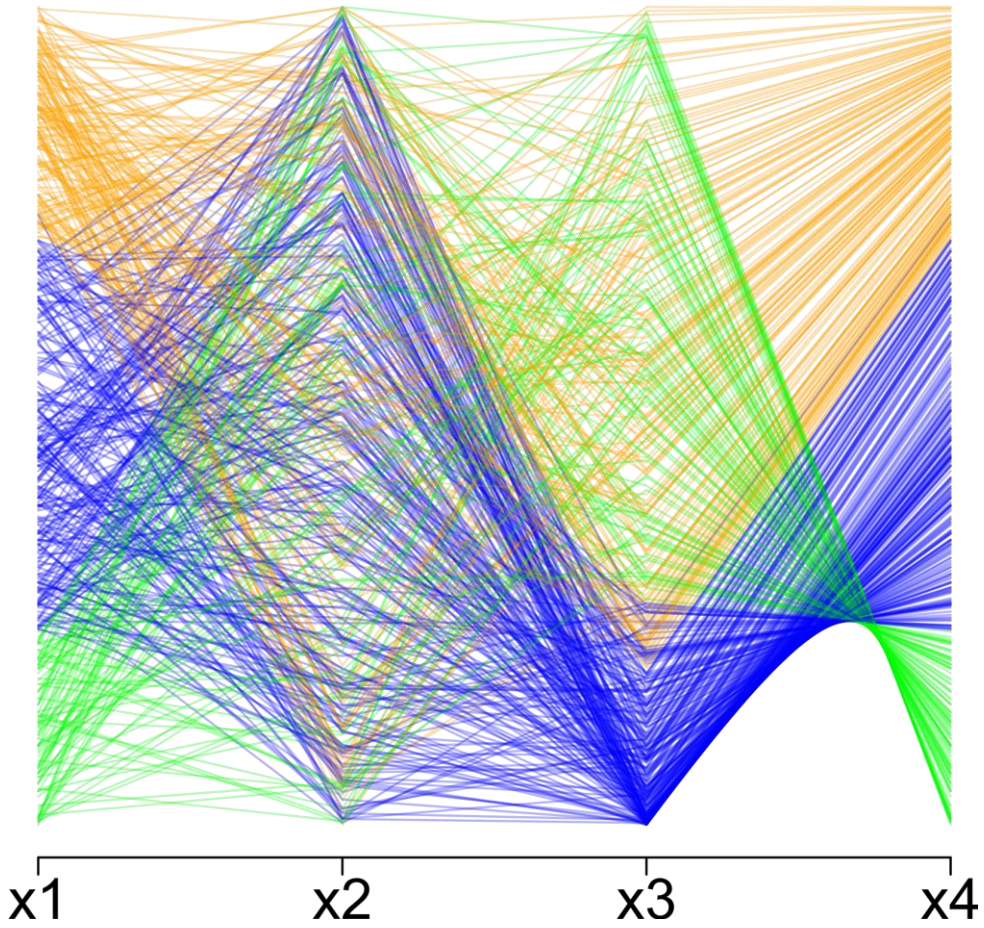
The impact of axis ordering

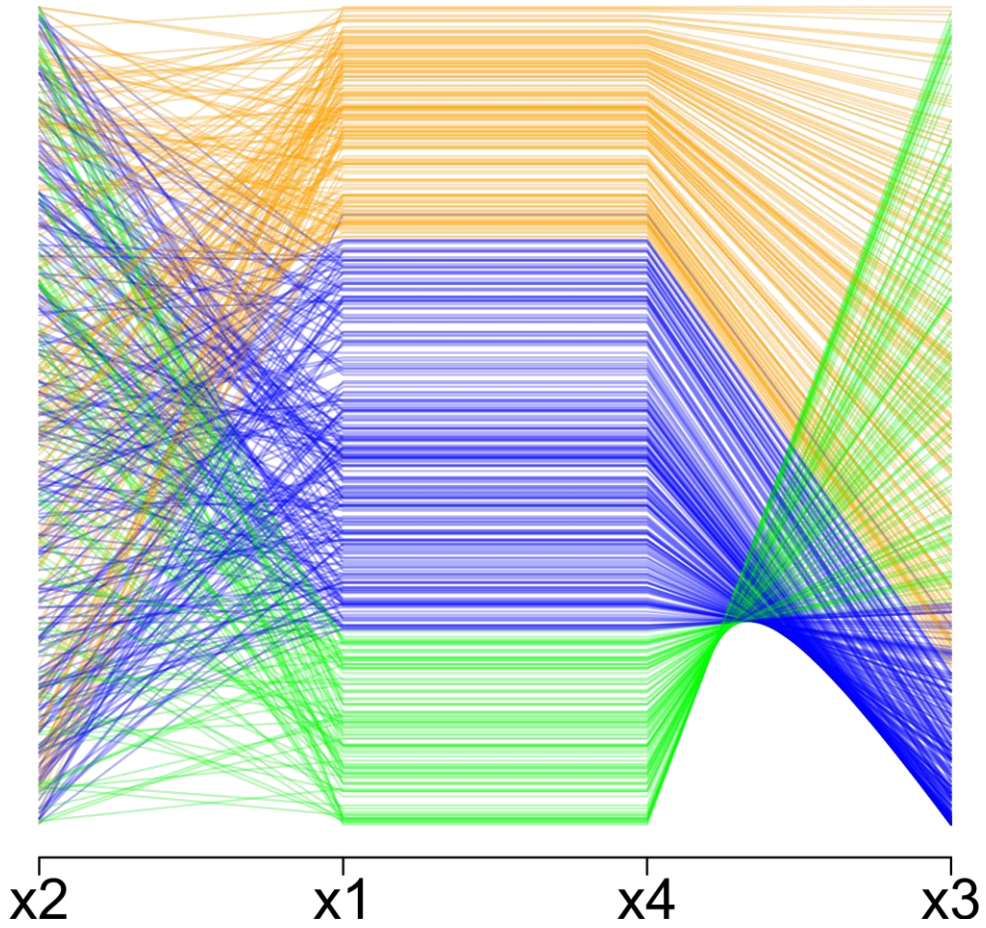




S. Tilouche and V. Nia and S. Bassetto : Parallel Coordinate Order for High-Dimensional Data, ArXiv, 2019, 1905.10035

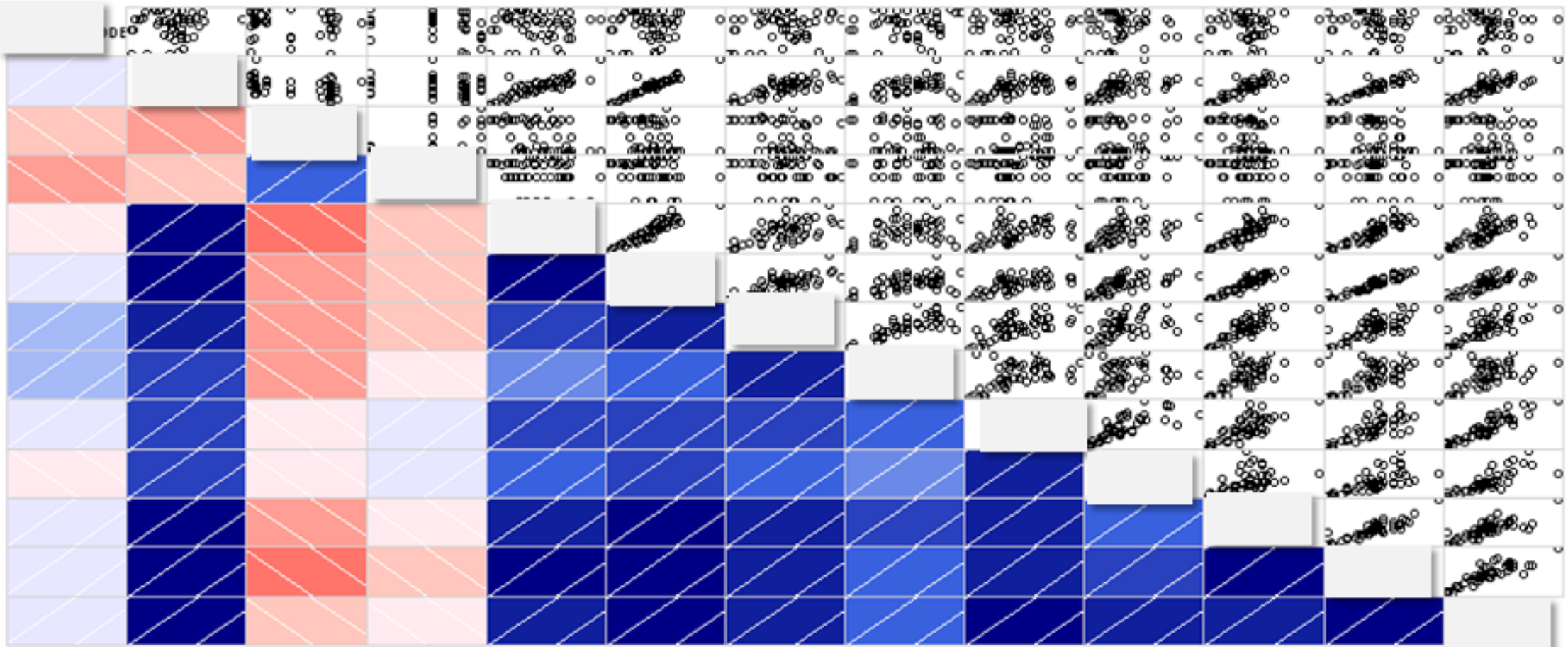
Model extraction



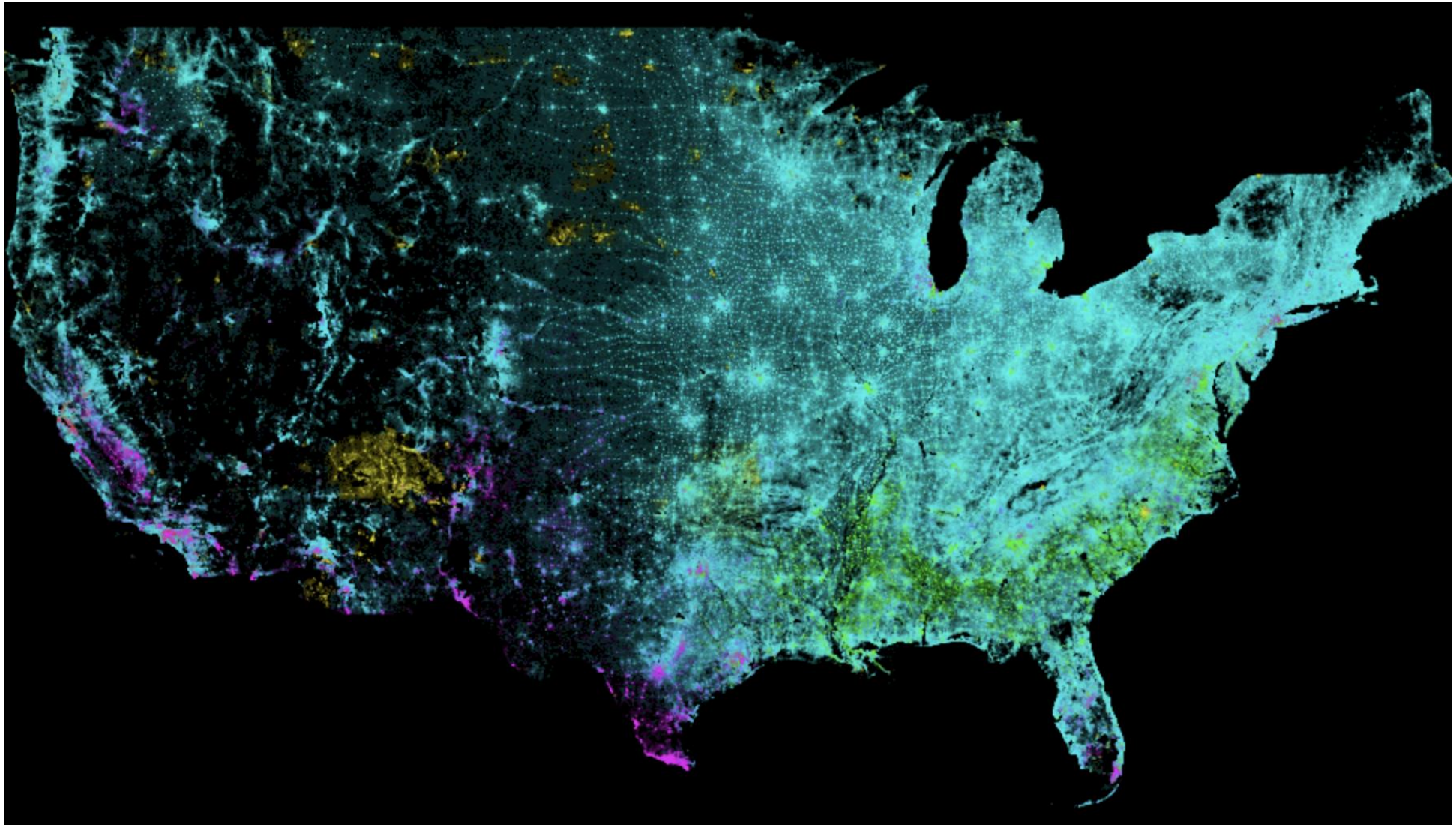


S. Tilouche and V. Nia and S. Bassetto : Parallel Coordinate Order for High-Dimensional Data, ArXiv, 2019, 1905.10035

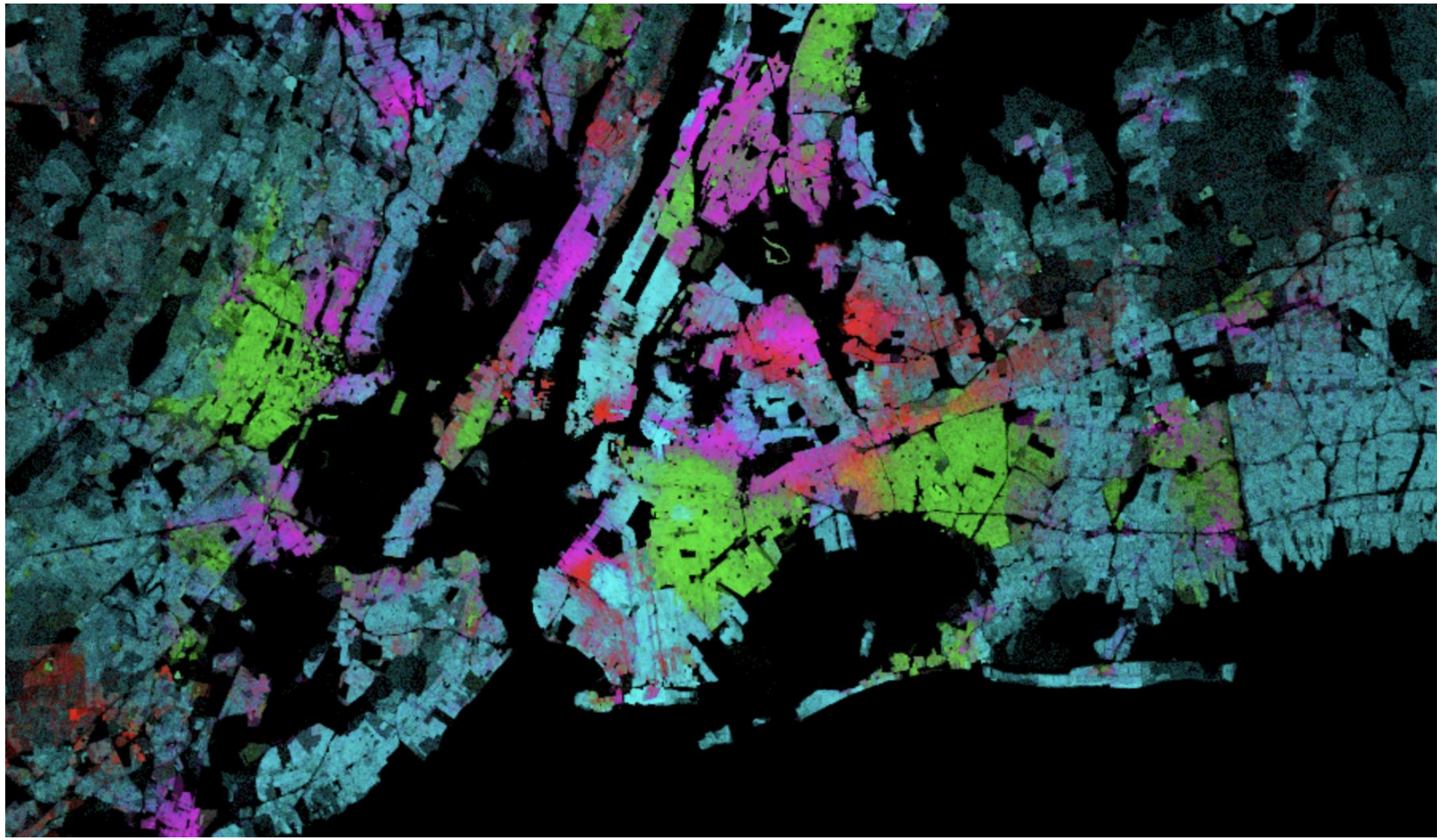
Correlation matrices/correlograms



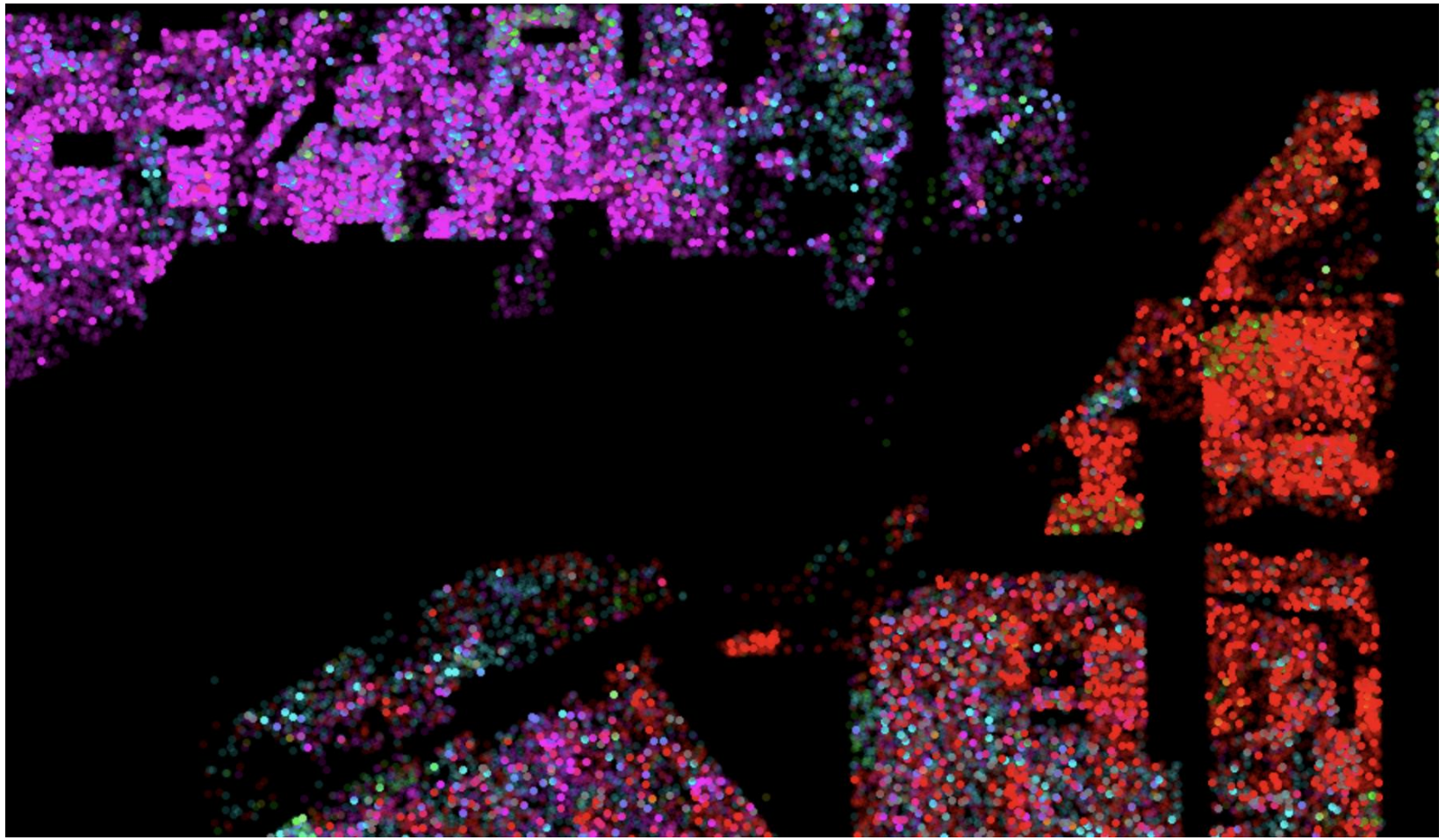
Big data



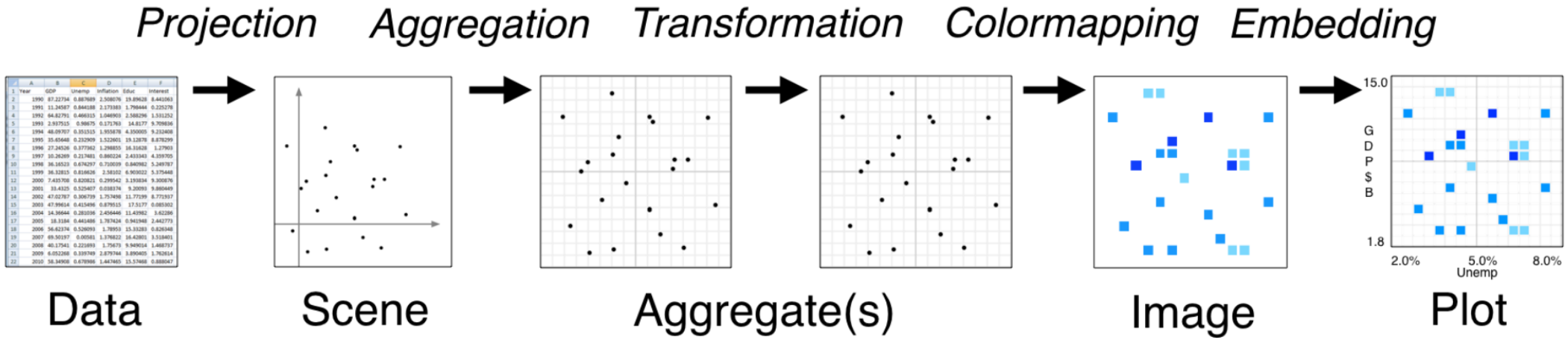
Big data



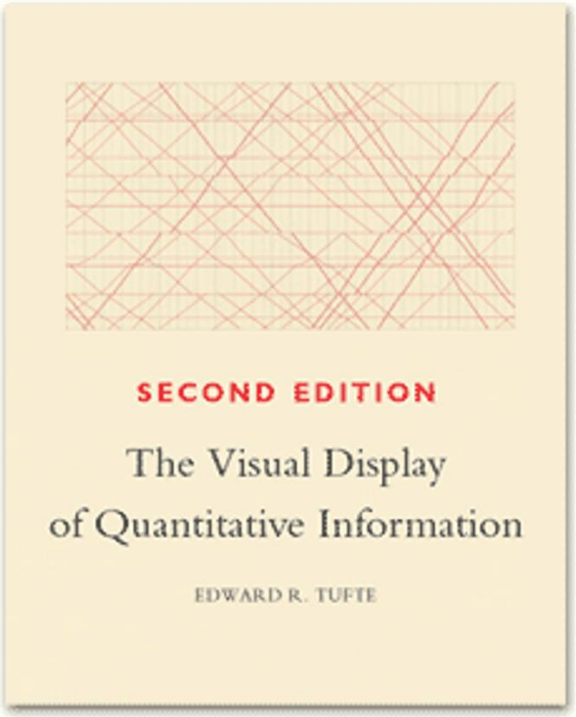
Big data

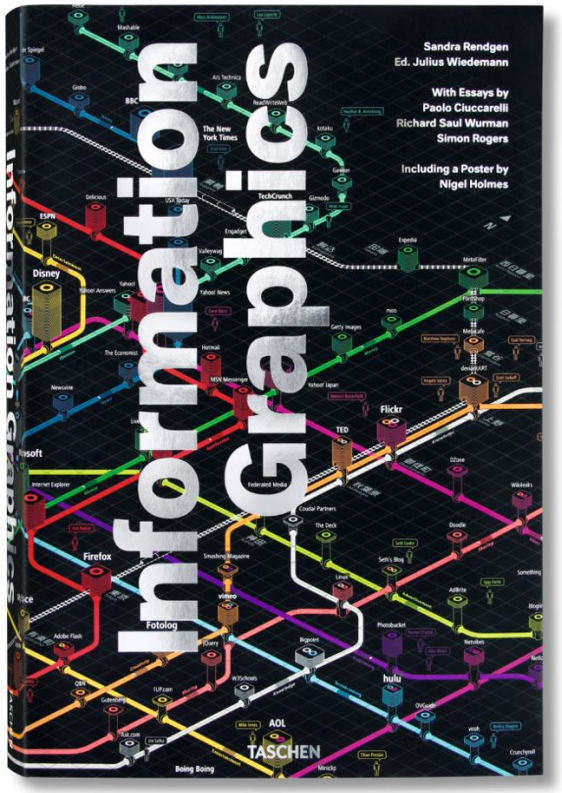


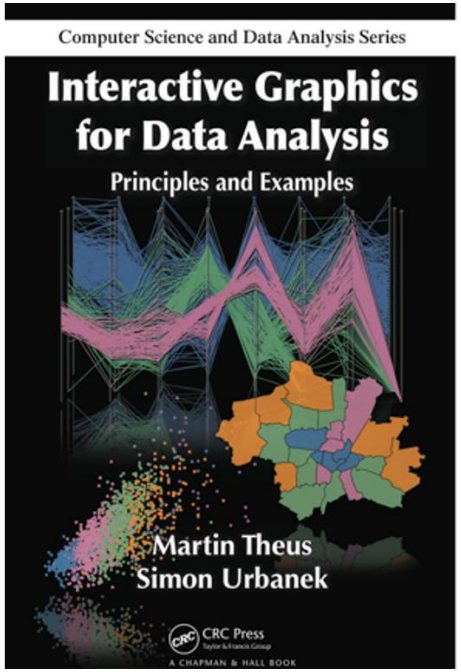
Big data visualization



Highly recommended learning sources







https://guides.library.duke.edu/datavis/vis_types

About Data Visualization | **Visualization Types** | Designing a Visualization | Helpful Tools and Tutorials | Learning About Visualization

ftsrg

Data analysis topics at ftsrg

- Model and knowledge-driven visual exploratory data analysis
- Extracting qualitative models from observations
- Data-driven model extraction (Process mining)
- Secure data sharing in data spaces
- Logic reasoning-based decision support
- Tool support for automated and visual data analysis
- LLM-based interactive data exploration with knowledge graph support