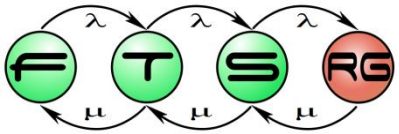
Vizuális adatelemzés

Rendszermodellezés 2019.

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Tartalom

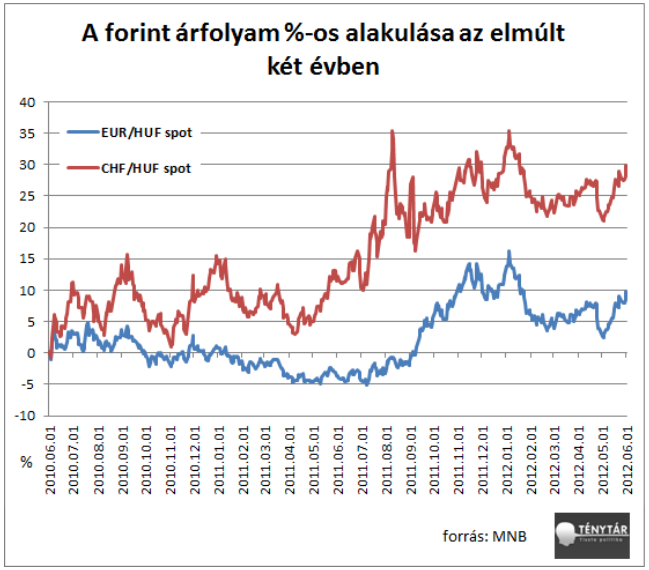
- Miért vizualizálunk?
- Mit vizualizálunk?
- Hogyan vizualizálunk?
- Mire következtetünk?

Tartalom

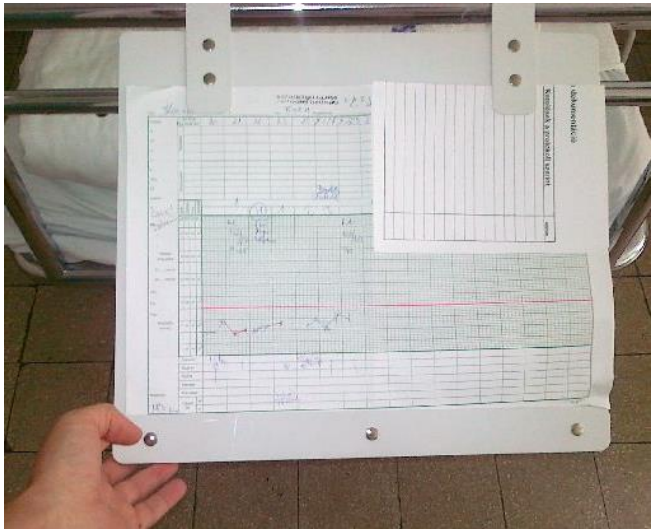
- Miért vizualizálunk?
- Mit vizualizálunk?
- Hogyan vizualizálunk?
- Mire következtetünk?

A vizualizáció alkalmazásai

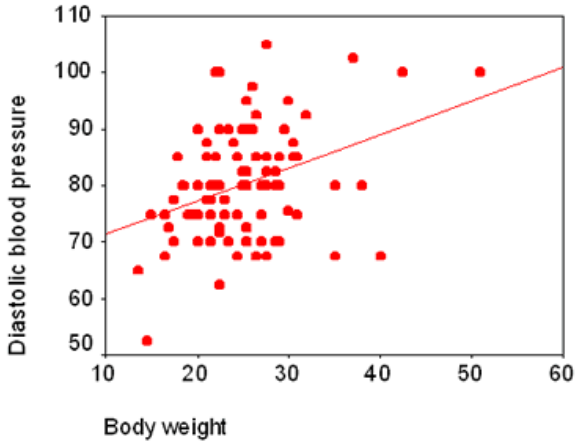
Trend analízis és előrejelzés



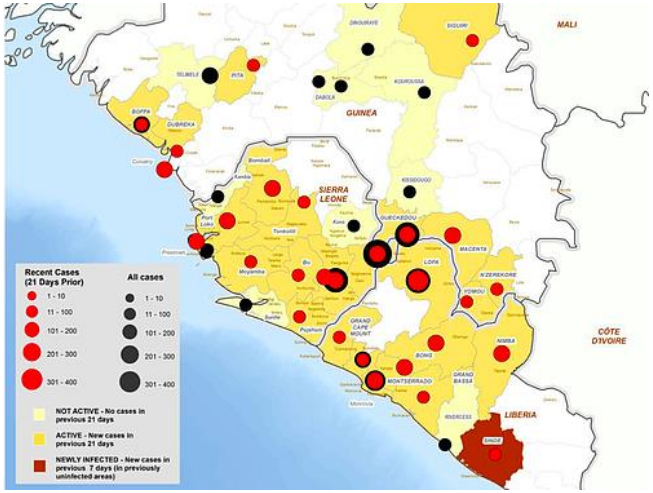
Idősor analízis



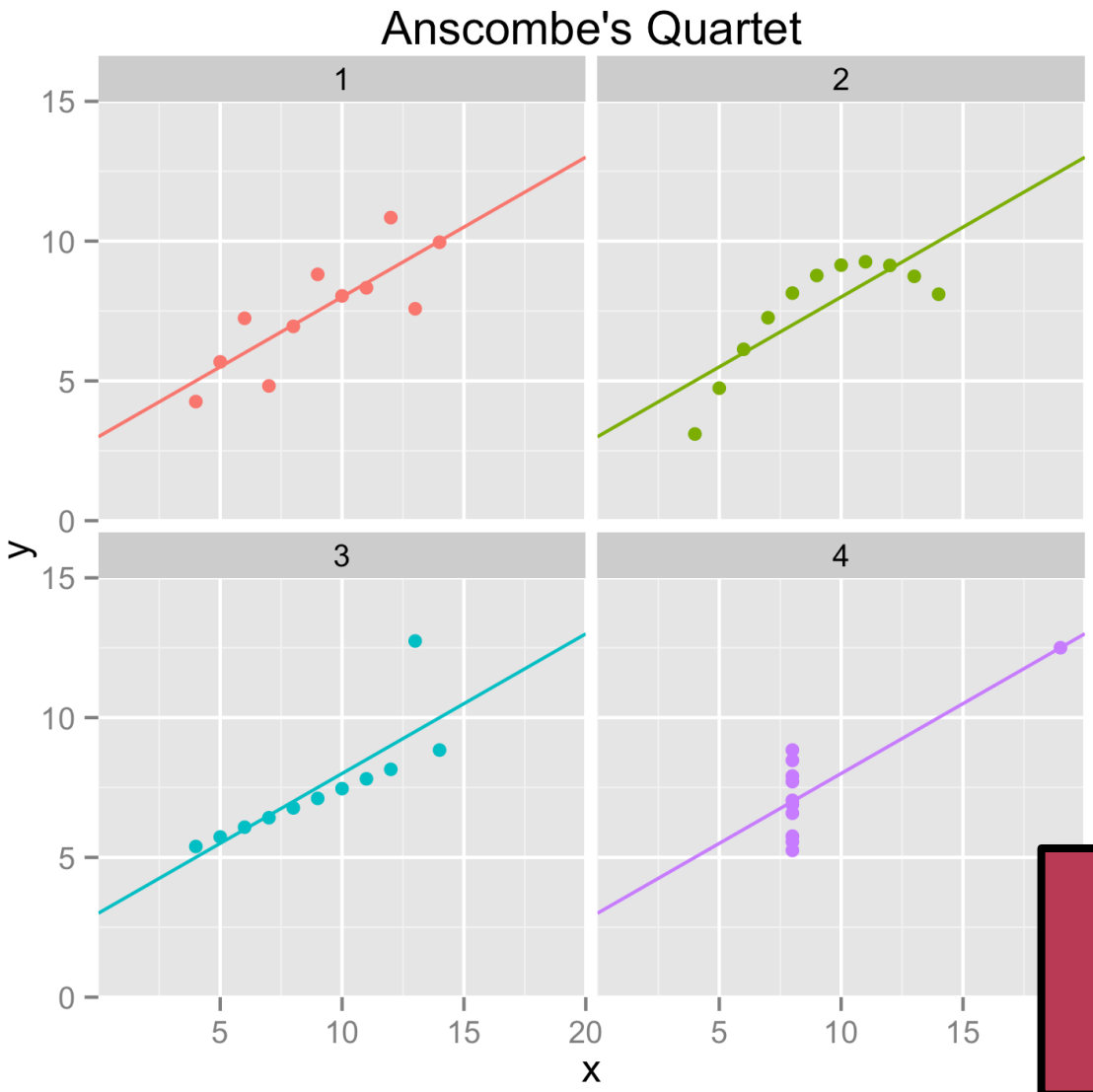
Korrelációanalízis



Térbeli analízis

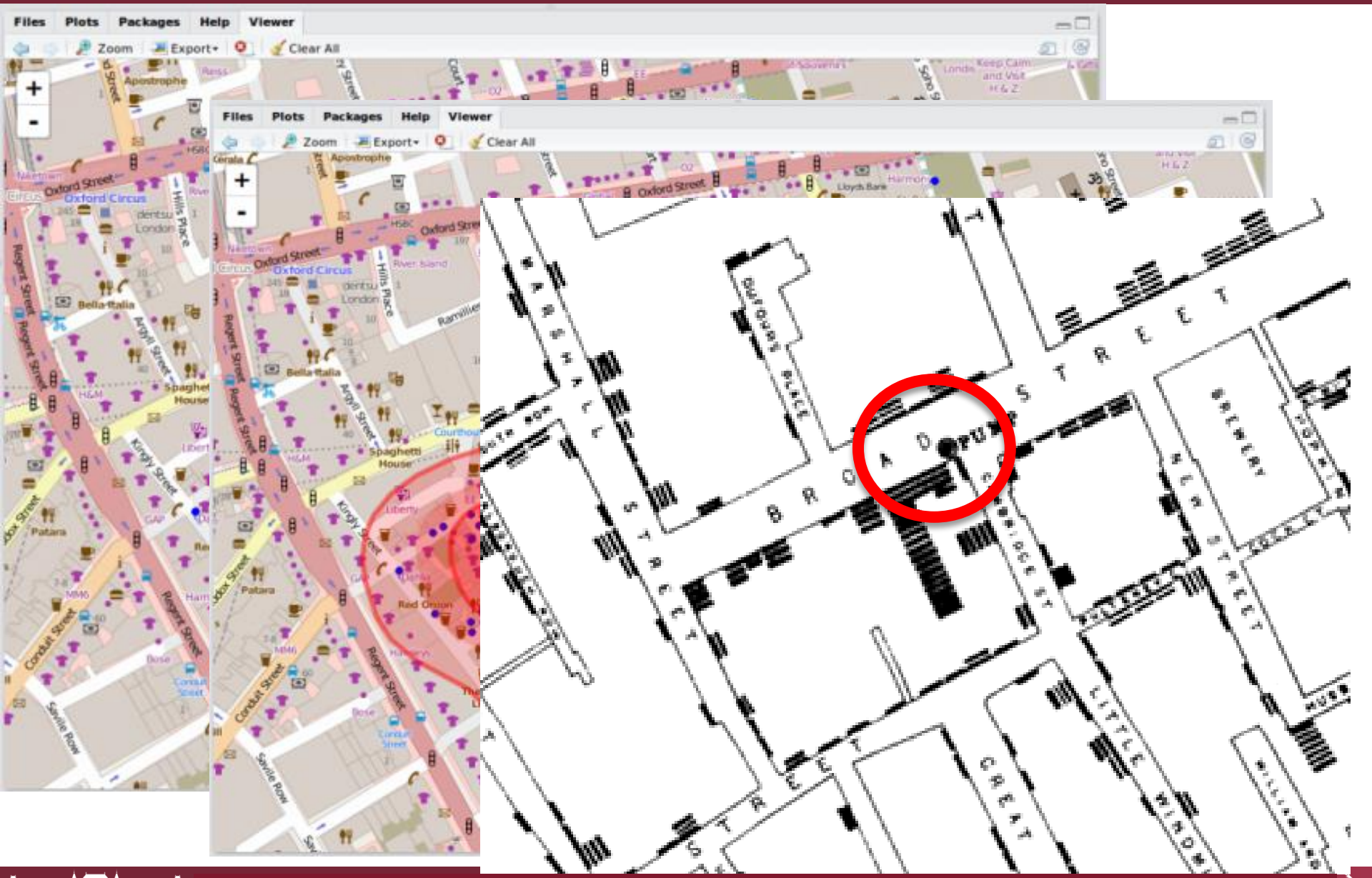


Számítások ellenőrzése



Hibás feltételezések elkerülése… és intuíció

Összefüggések feltárása



Forrás: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

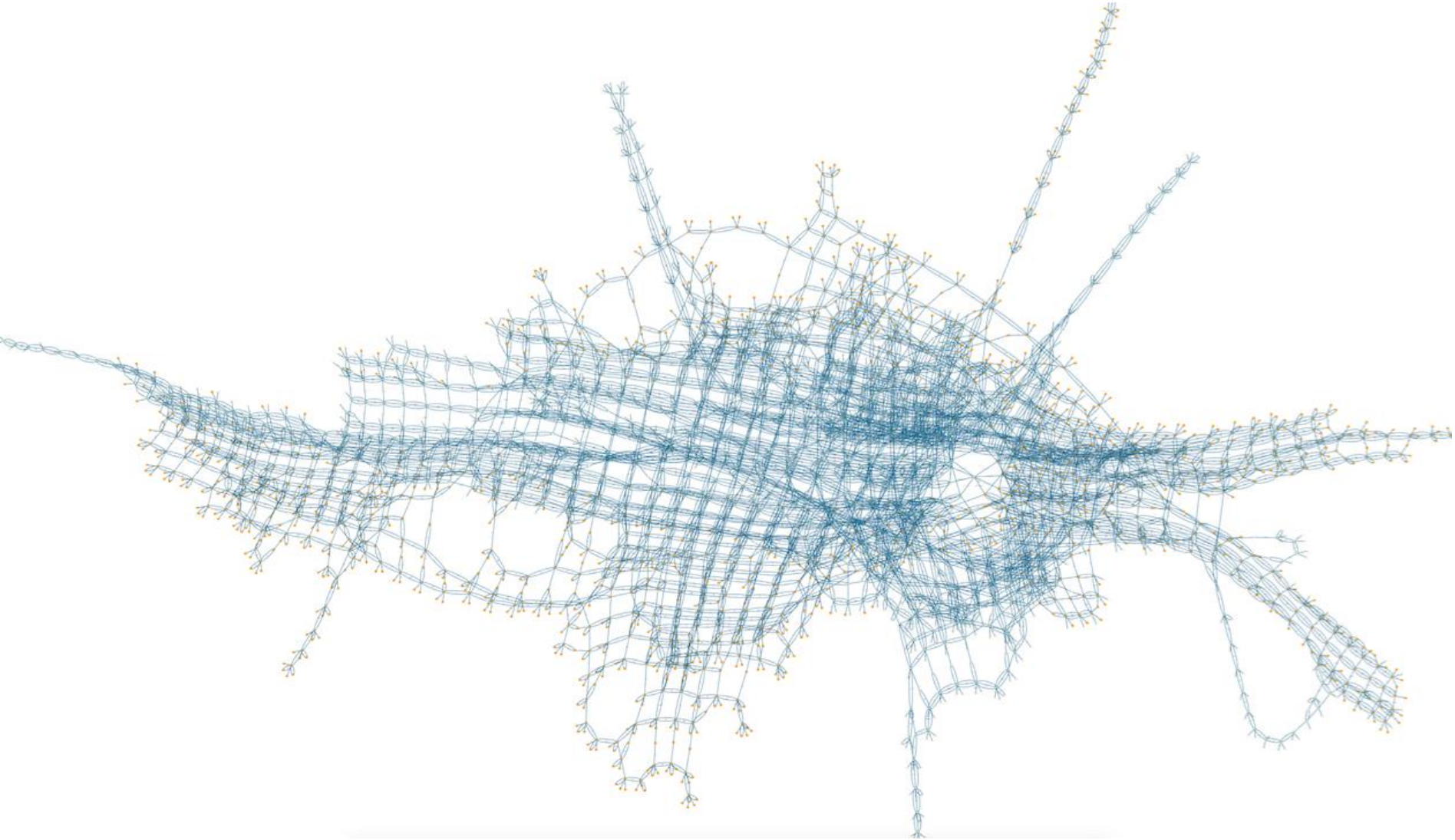
Mindent a szemnek!

„Masszív” erőforrások

- 120.000.000 szenzor

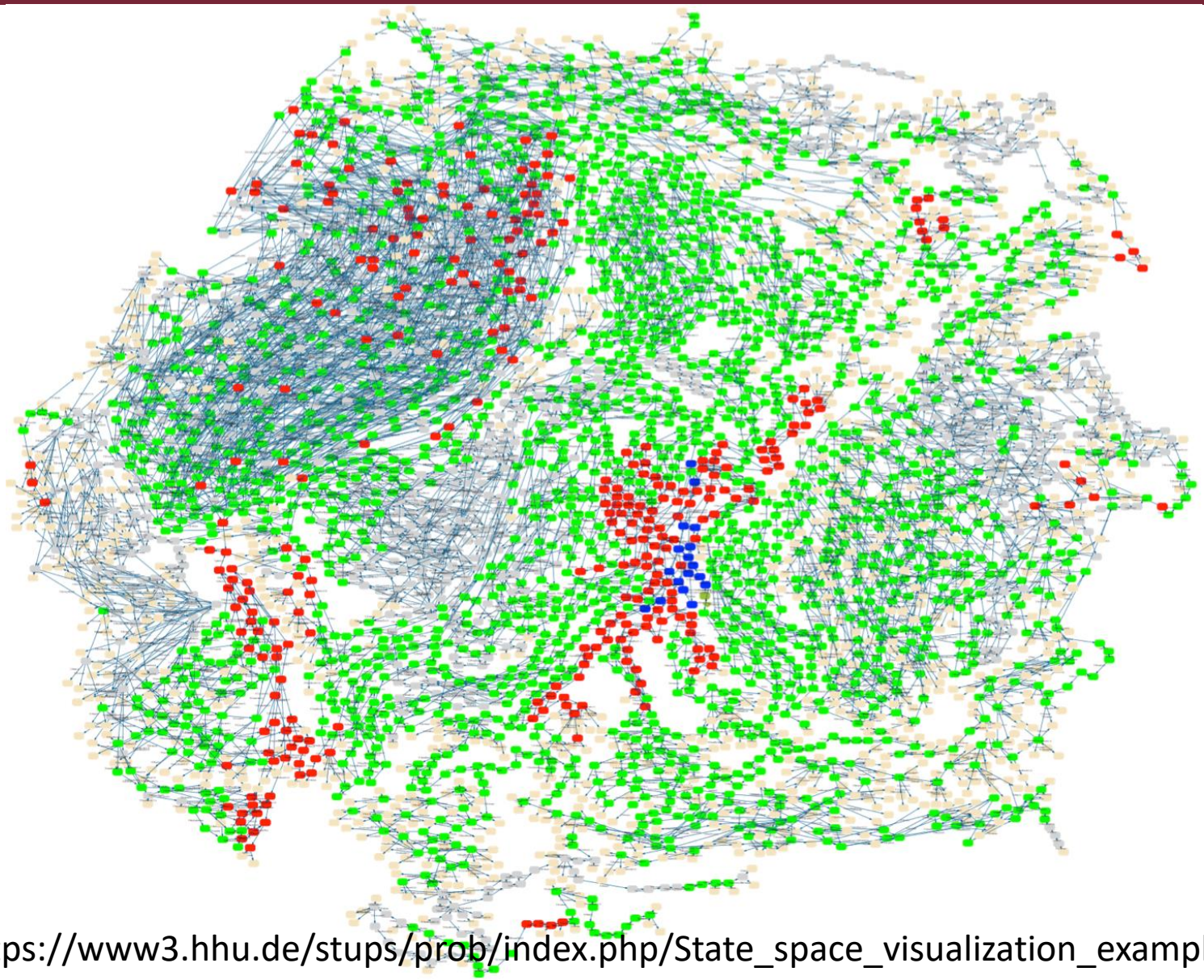
3. Vizuális kiválasztás és manipuláció
4. Interpretáció, korreláció más modellekkel, kiértékelés

Példa: állapottér vizualizáció (hálózat)



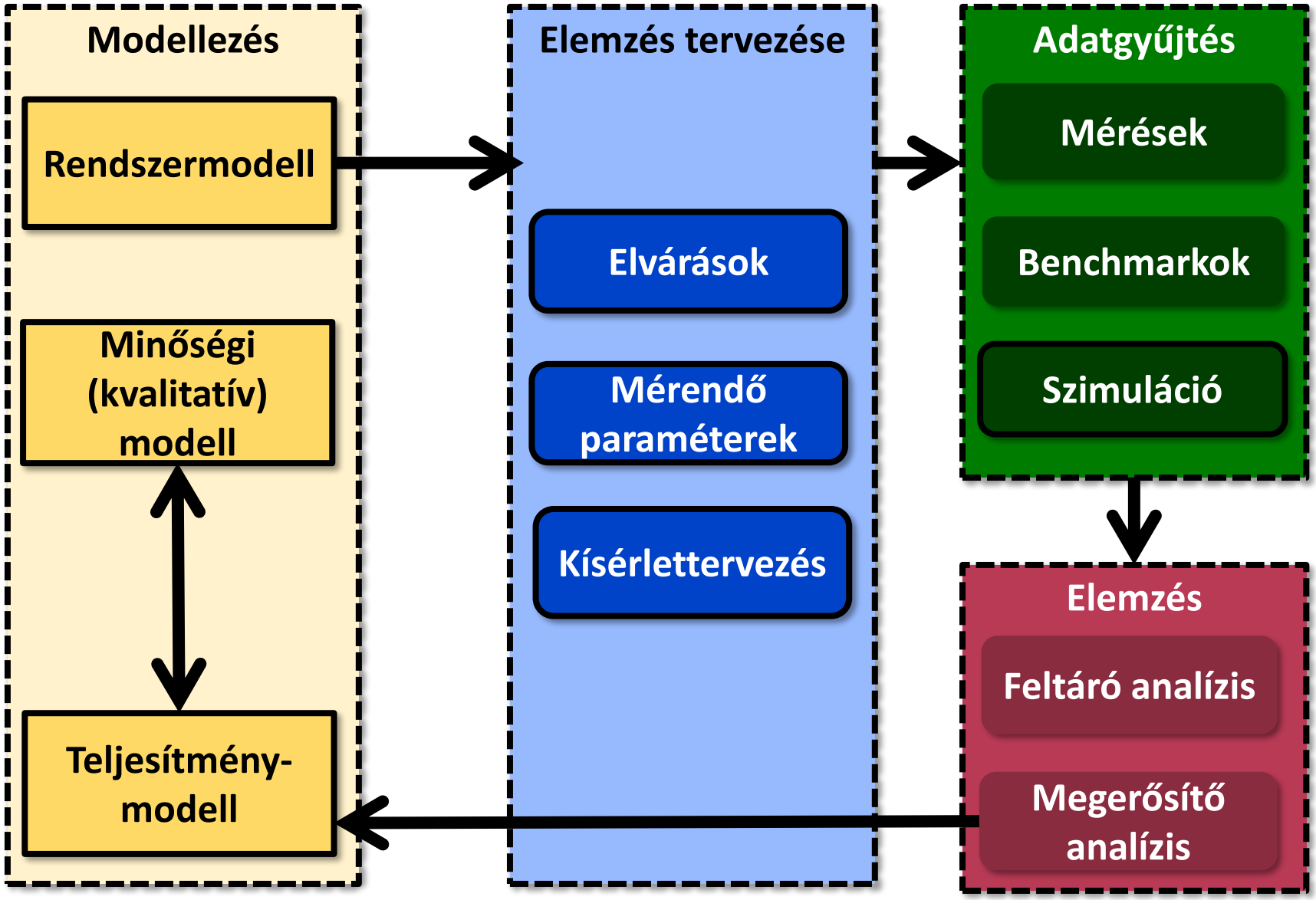
https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples

Példa: CAN bus állapottér



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

Példa: Rendszermodell → teljesítménymodell



Mi is lesz?

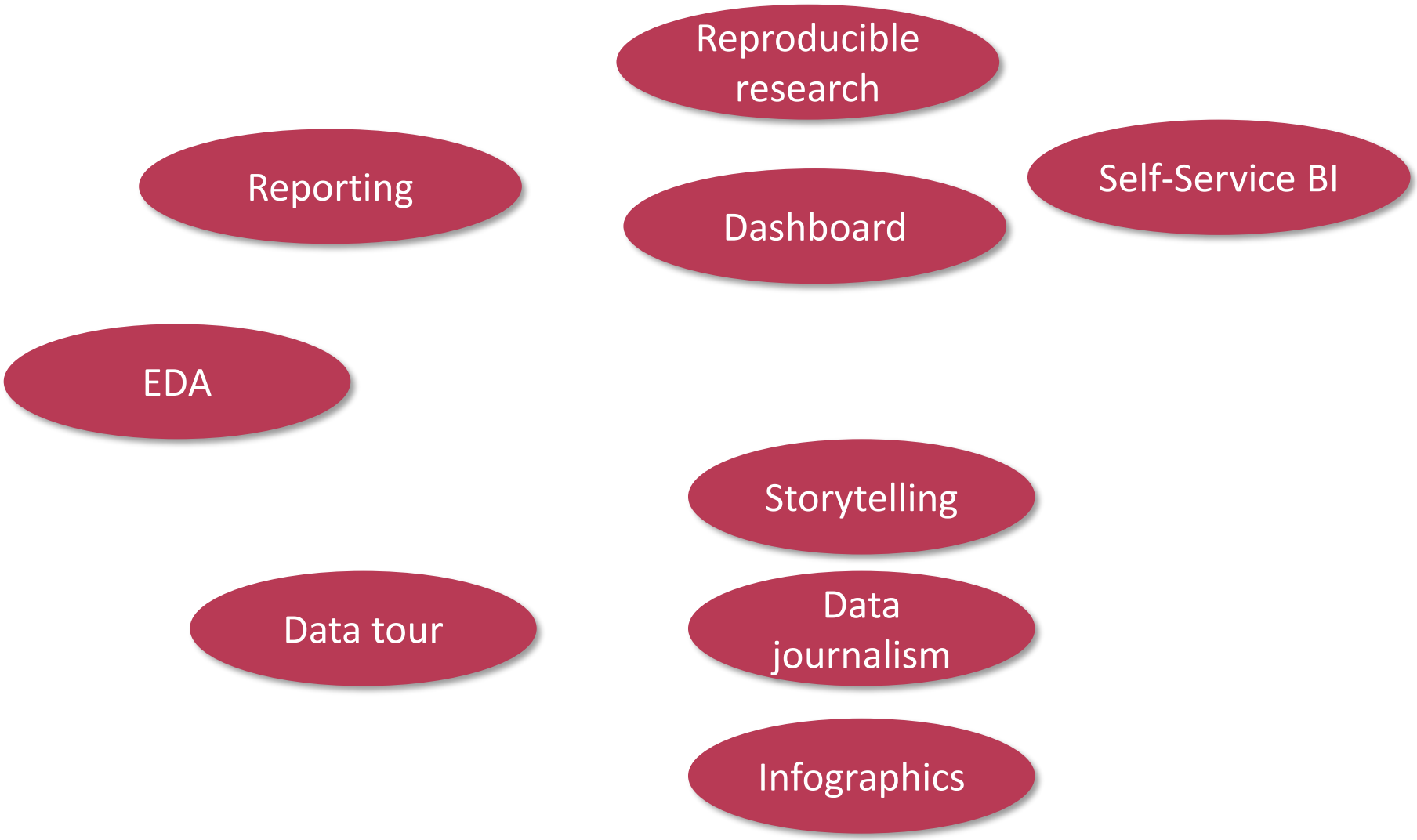
Miért vizualizálunk?

Mit vizualizálunk?

Hogyan vizualizálunk?

Mire következtetünk?

Főbb vizualizációs célok



Emlékeztető: táblázatos ábrázolás

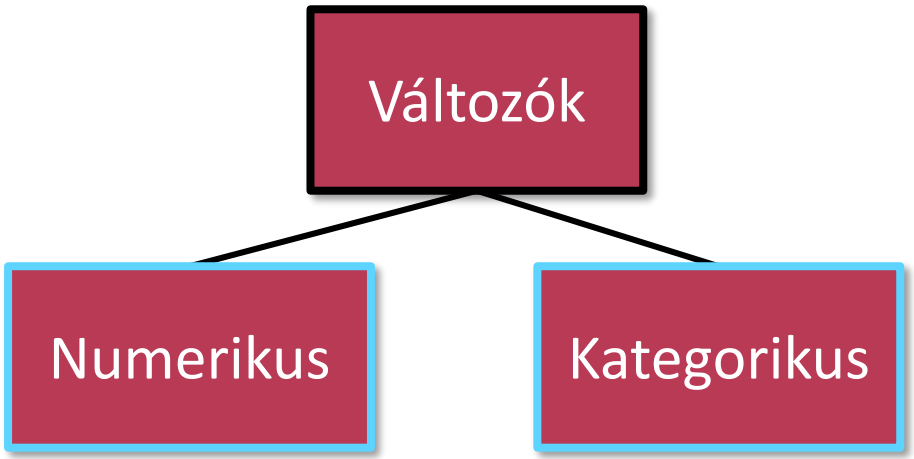
- **Táblázat sora** = modellelem
- **Táblázat oszlopa** = tulajdonság

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

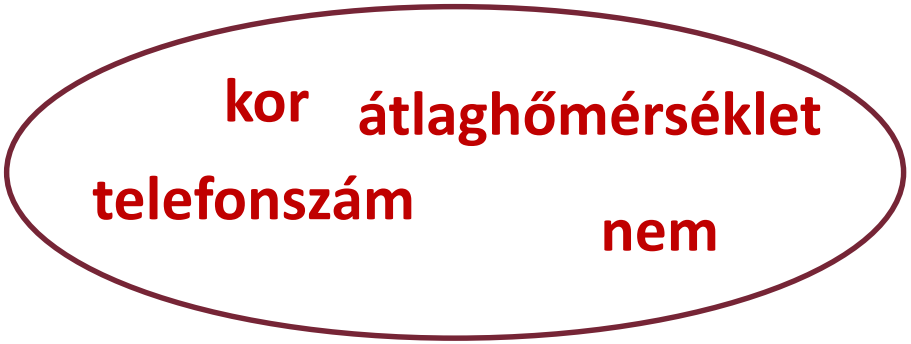
- Adatelemzési eszközök (pl. R, Python): **dataframe**
  - Egy sor egy mérés
  - Egyes oszlopoknak **típusai** vannak

Numerikus és kategorikus változók

- Numerikus (numerical)
  - az alapvető aritmetikai műveletek értelmesek

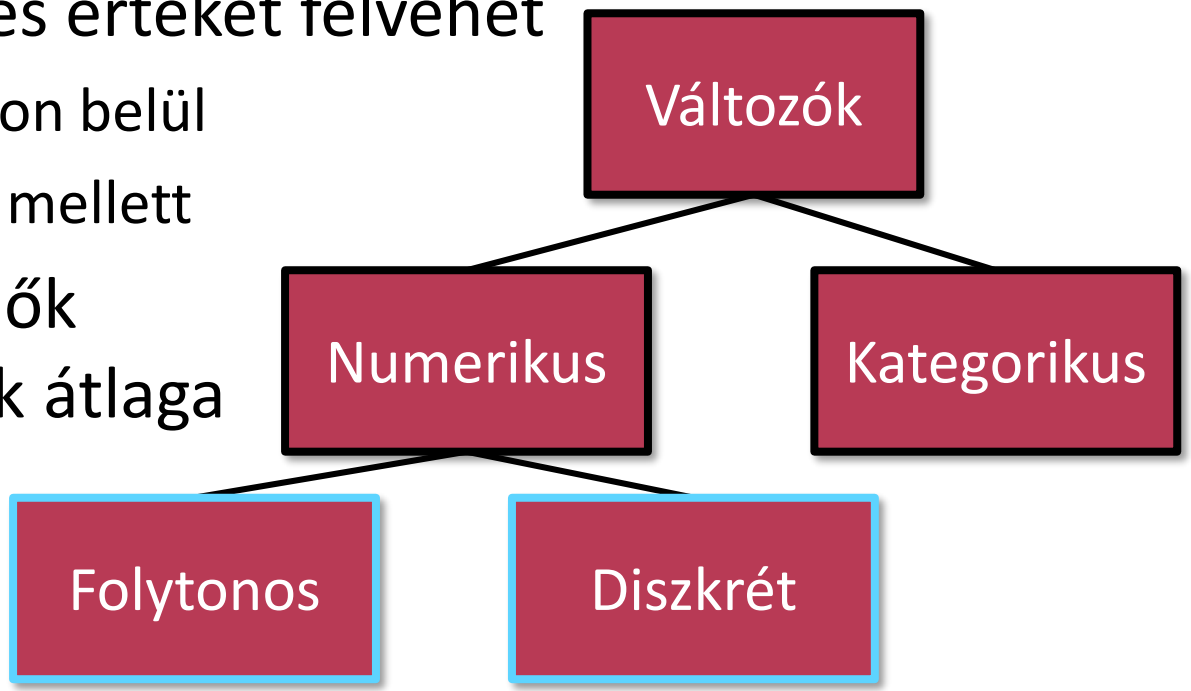


- Kategorikus (categorical)
  - Matematikai műveletek nem értelmezhetőek rajtuk, legfeljebb sorba rendezés



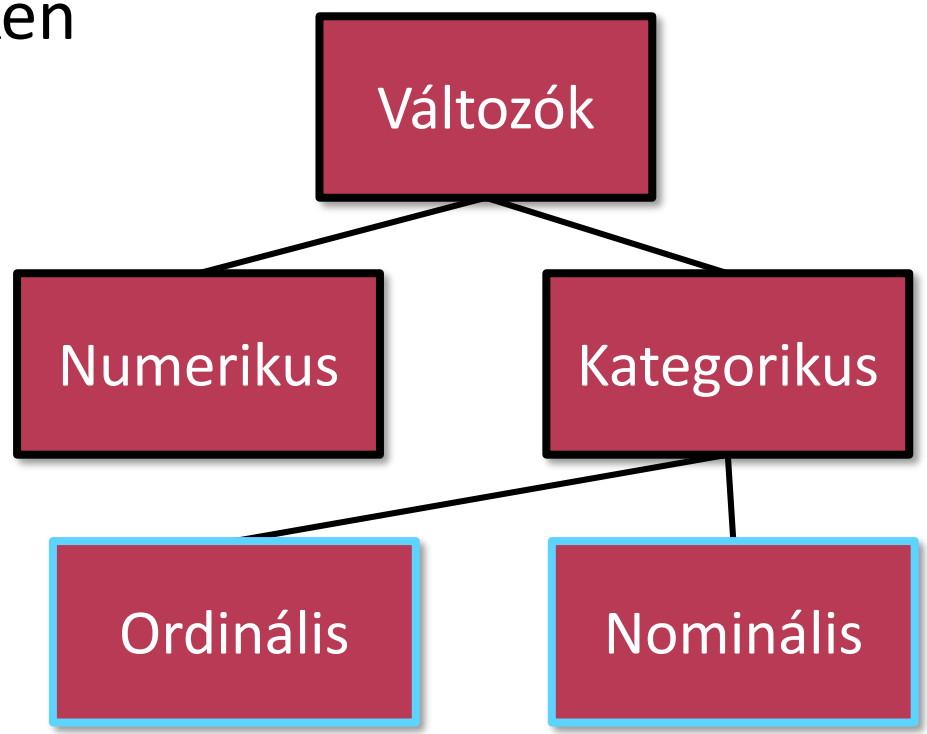
Numerikus változók

- Folytonos
  - Mért – tetszőleges értéket felvehet
    - adott tartományon belül
    - adott pontosság mellett
  - Pl. a teremben ülők ZH pontszámának átlaga
- Diszkrét
  - Számolt – véges sok értéket vehet fel adott tartományban
  - Pl. az előadáson ülők száma



Kategorikus változók

- Ordinális
  - Teljes rendezés az értékeken
  - Pl. szállodai csillagok
- Nominális



9. Ajánlanád-e a tárgyat másoknak?

- ○ Mindenkit rábeszélnék
- ○ Nyugodtan ajánlanám
- ○ Esetleg ajánlanám
- ○ Inkább lebeszélném róla
- ○ Feltétlenül lebeszélném
- ○ Nem kívánok válaszolni

Mi is lesz?

Miért vizualizálunk?

Mit vizualizálunk?

Hogyan vizualizálunk?

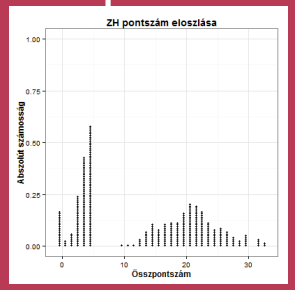
Mire következtetünk?

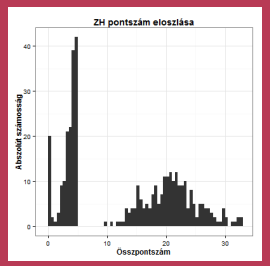
1 változó – eloszlásokra

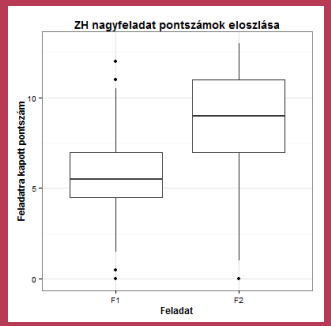
Változók

Numerikus

ZH pontszám: [13, 15, 2, …]

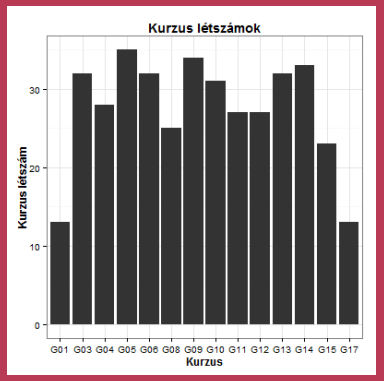






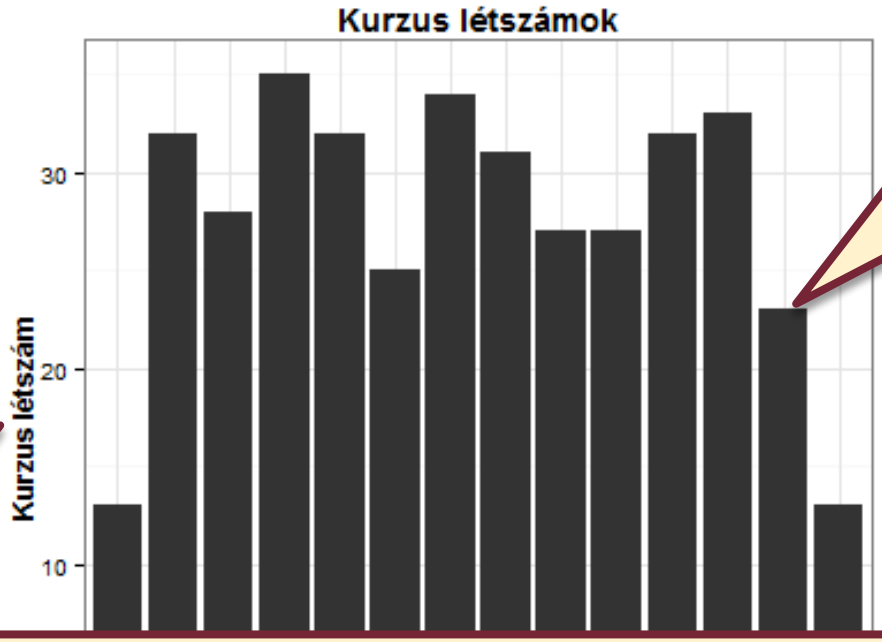
Kategorikus

Kurzus: [G01, G03, G15, G17, …]



Oszlopdiagram

- Bemenő változó: kurzus kód
- Kérdés: az egyes kurzusokra hányan járnak?



Oszlop-magasság: adott érték gyakorisága

abszolút gyakoriság!

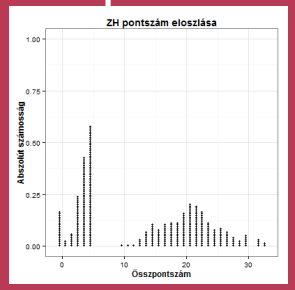
Tervezői döntés: értékkészlet darabolása
Pl.: kedd-csütörtök-péntek

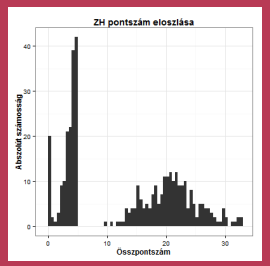
1 változó – eloszlásokra

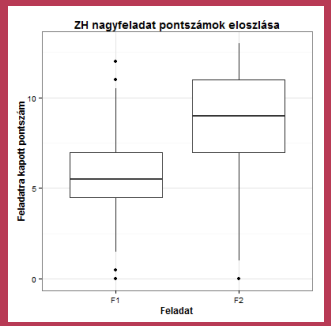
Változók

Numerikus

ZH pontszám: [13, 15, 2, …]

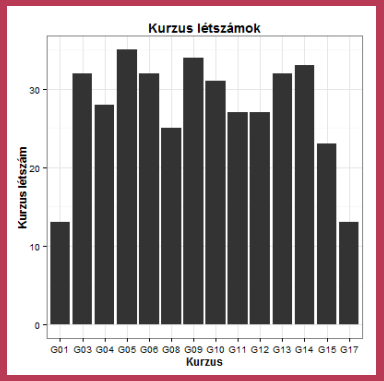






Kategorikus

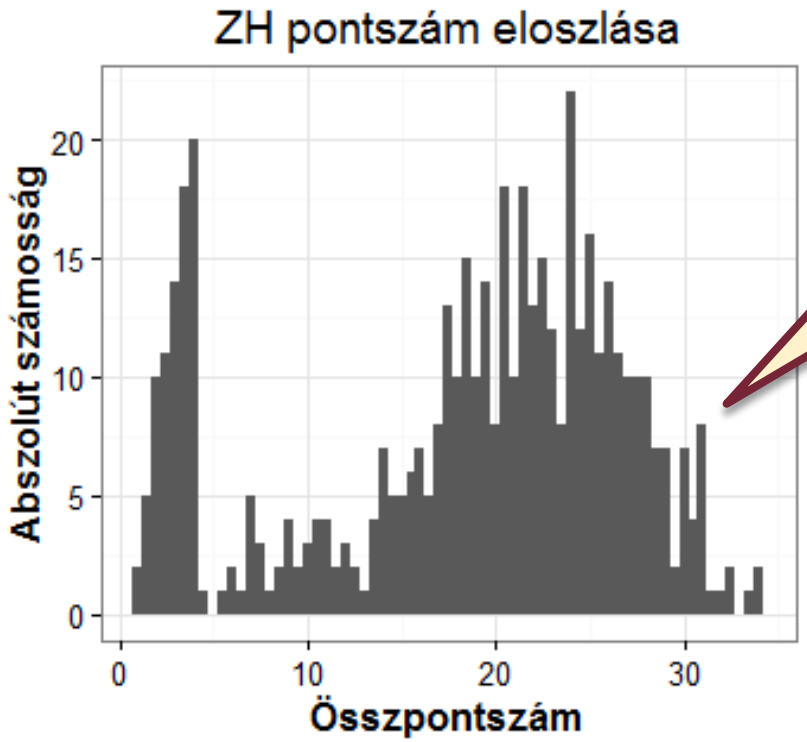
Kurzus: [G01, G03, G15, G17, …]



Hisztogram

- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok?

abszolút gyakoriság!

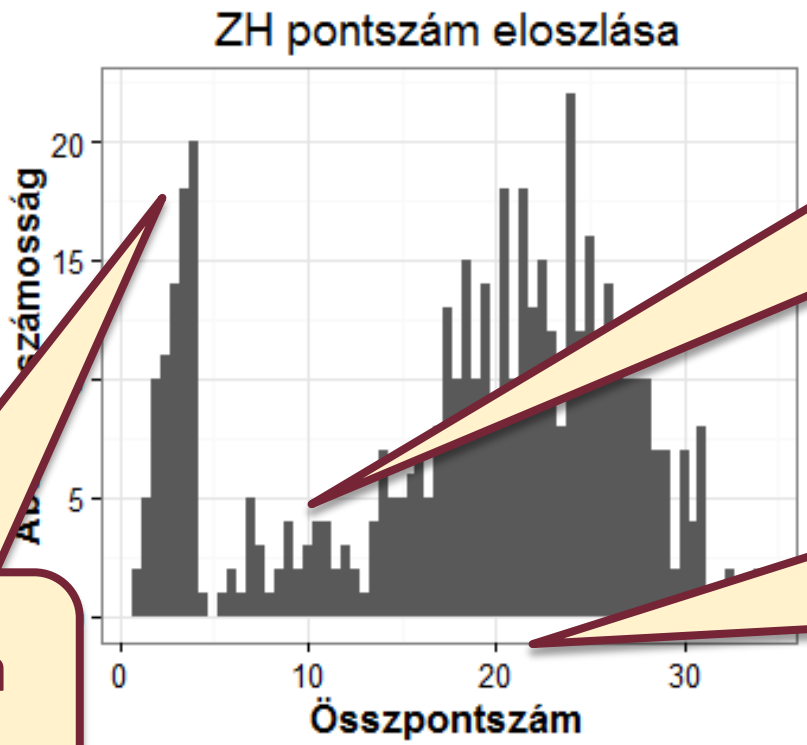


Oszlop-magasság: adott **intervallum** számossága

Tervezői döntés: mekkora legyen az intervallum hossza (bin size)?
Pl.: elég 1 pontos felbontással, vagy menjünk fél pontokig?

Hisztogram

- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok?



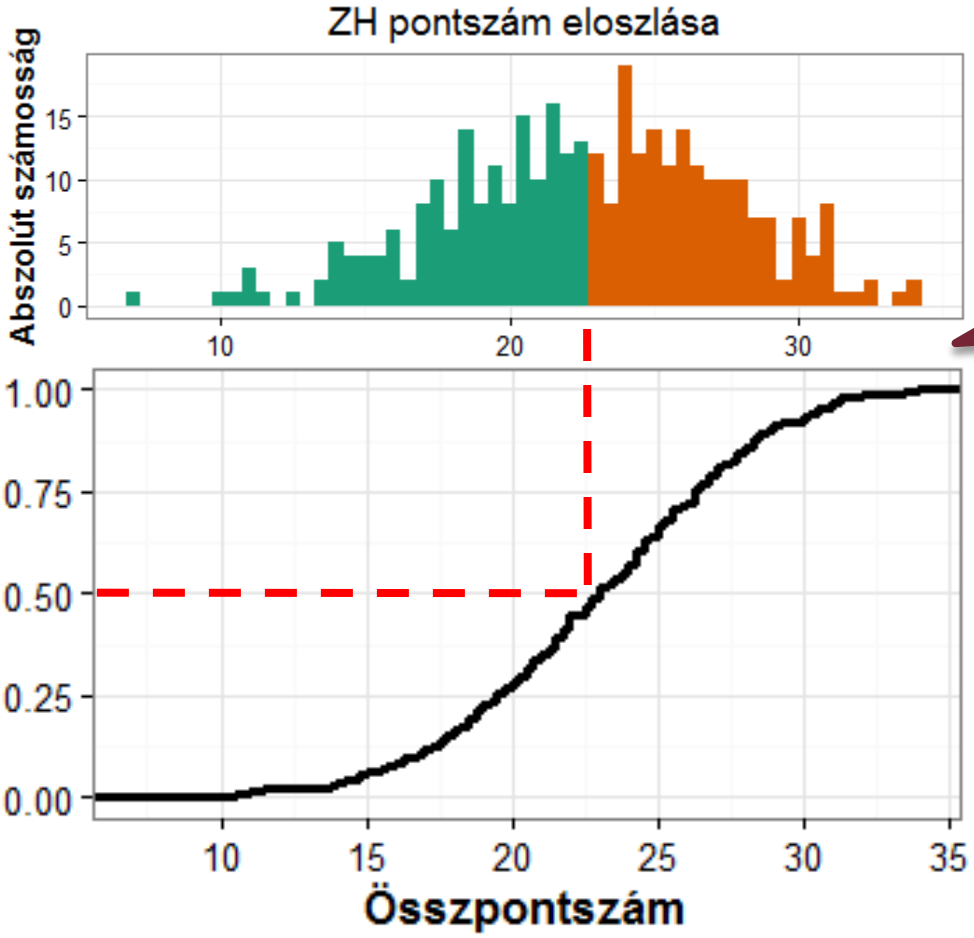
Akik átmentek a beugrón, valószínűleg át is mentek

Sokan voltak a határon

18 pont körül volt az átlag, 20 körül a medián

Egyszerű statisztikai jellemzés

- Hol van az adatok „közepe”?

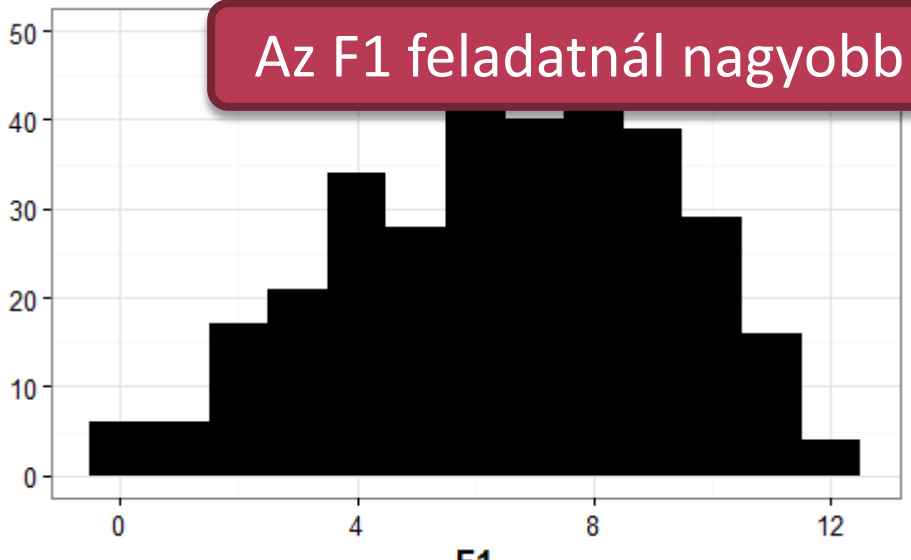


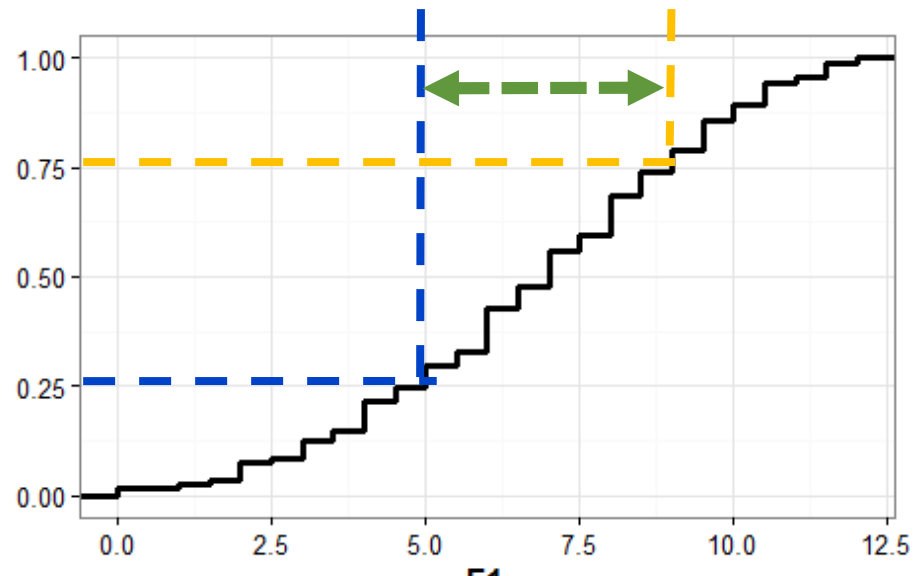
Az átmentek összpontszám mediánja 23

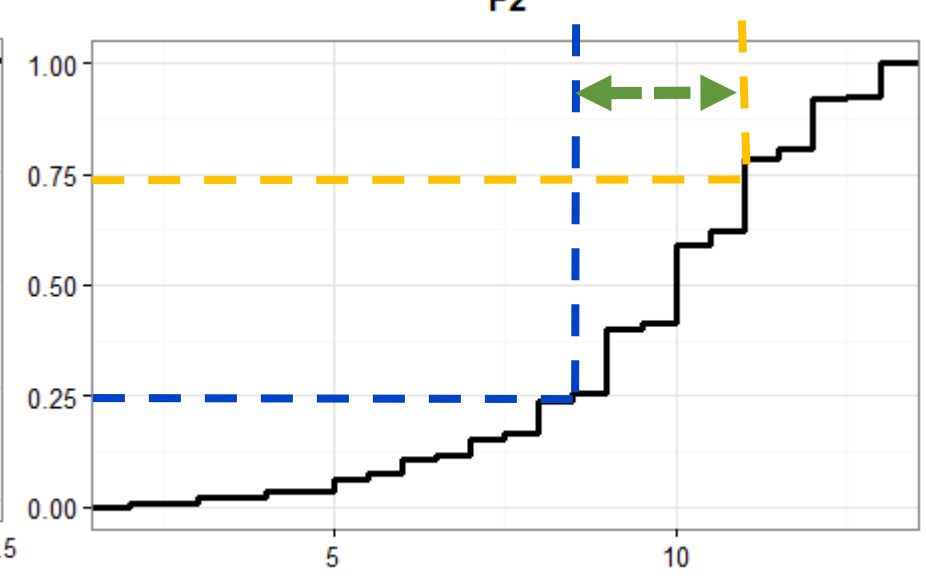
Egyszerű statisztikai jellemzés

- Mennyire „szórtak” az adatok?

Az F1 feladatnál nagyobb volt a szórás, mint az F2-nél



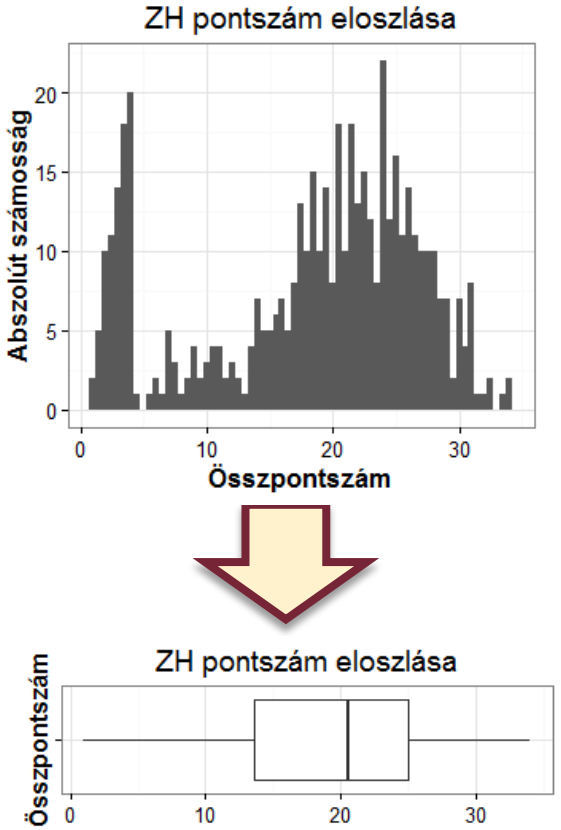




Boxplot

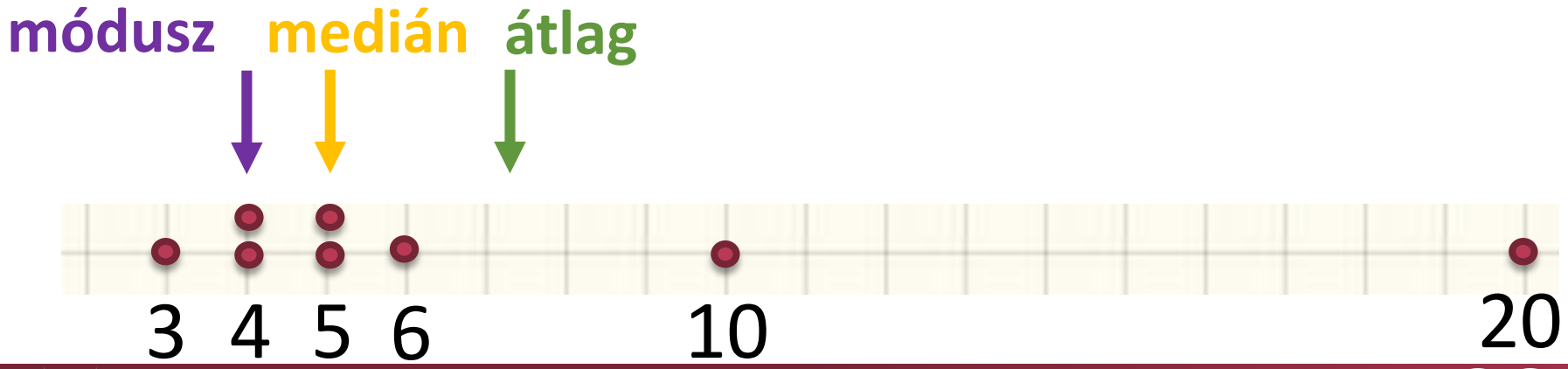
- Bemenő változó: ZH összpontszám
- Kérdés: hogyan alakultak a ZH pontszámok **úgy nagyjából**?

Egyfajta absztrakció itt is: legyenek intervallumok, felesleges minden pontot kirajzolni



(Folytonos) megfigyelések jellemzése

- A „központ” jellemzése
  - Átlag, **medián**, módusz
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - Átlag: ~ 7.125
    - Medián: 5
    - Módusz: 4 és 5



(Folytonos) megfigyelések jellemzése

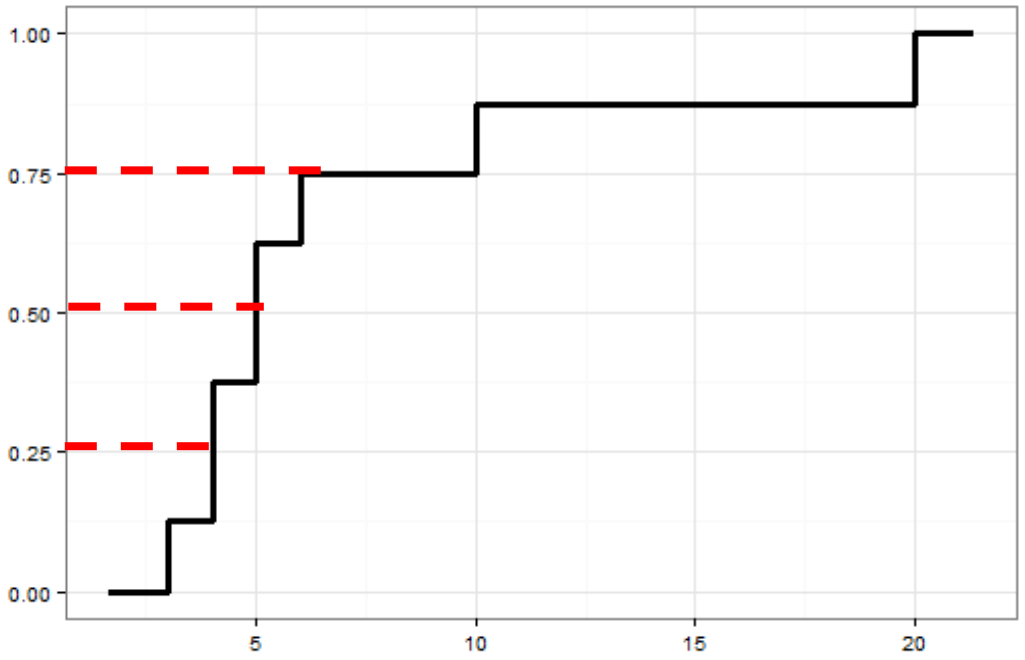
Ha az értékeket növekvően sorba rendezzük, akkor a középső adat az adathalmaz **mediánja**. Ha nincs középső adat (páros számú érték esetén), akkor a **medián** a két középső érték átlaga (számtani közepe).

A **módusz** az adathalmazban legtöbbször előforduló érték. Ez nem feltétlenül egyértelmű, ilyenkor több móduszról beszélünk.

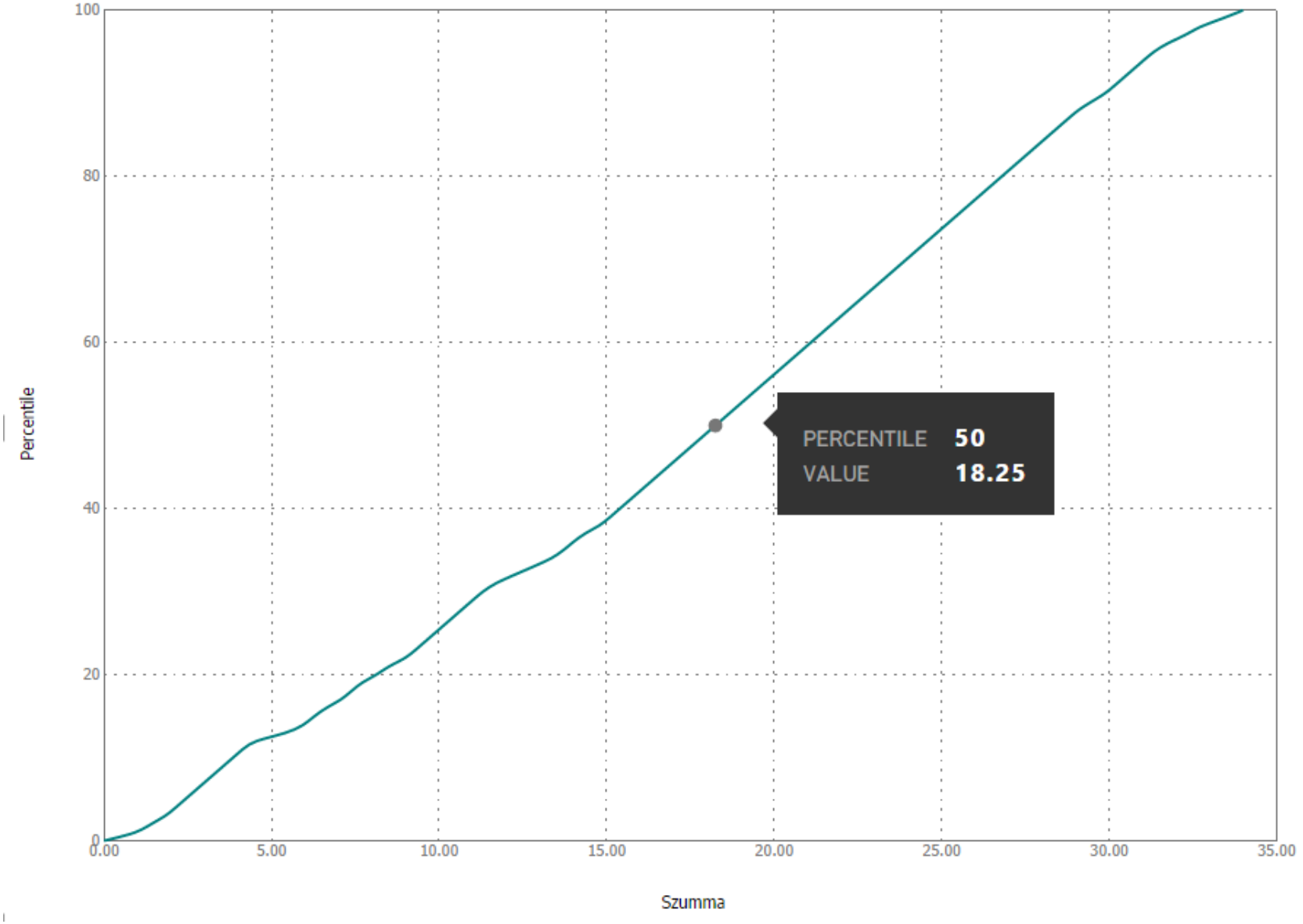
Terjedelem jellemzése: percentilisek

Az n-edik **percentilisnél** az értékek $n\%$-a nem nagyobb.

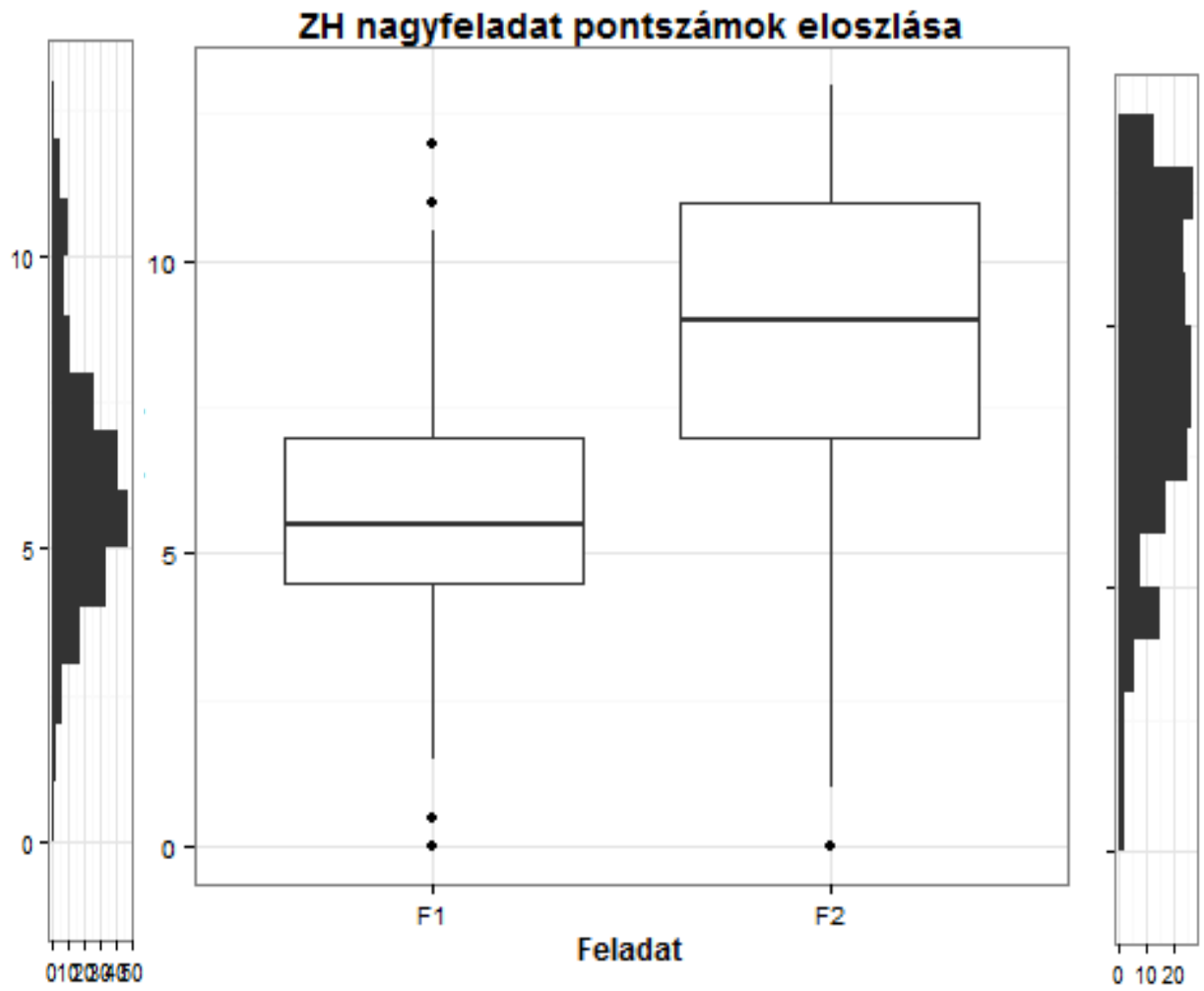
- Percentilis
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - 50. percentilis: 5
    - 25. percentilis: 4
    - 75. percentilis: 6
- Kvartilis
  - Q1: 25. percentilis
  - Q3: 75. percentilis
  - **Q2: medián**



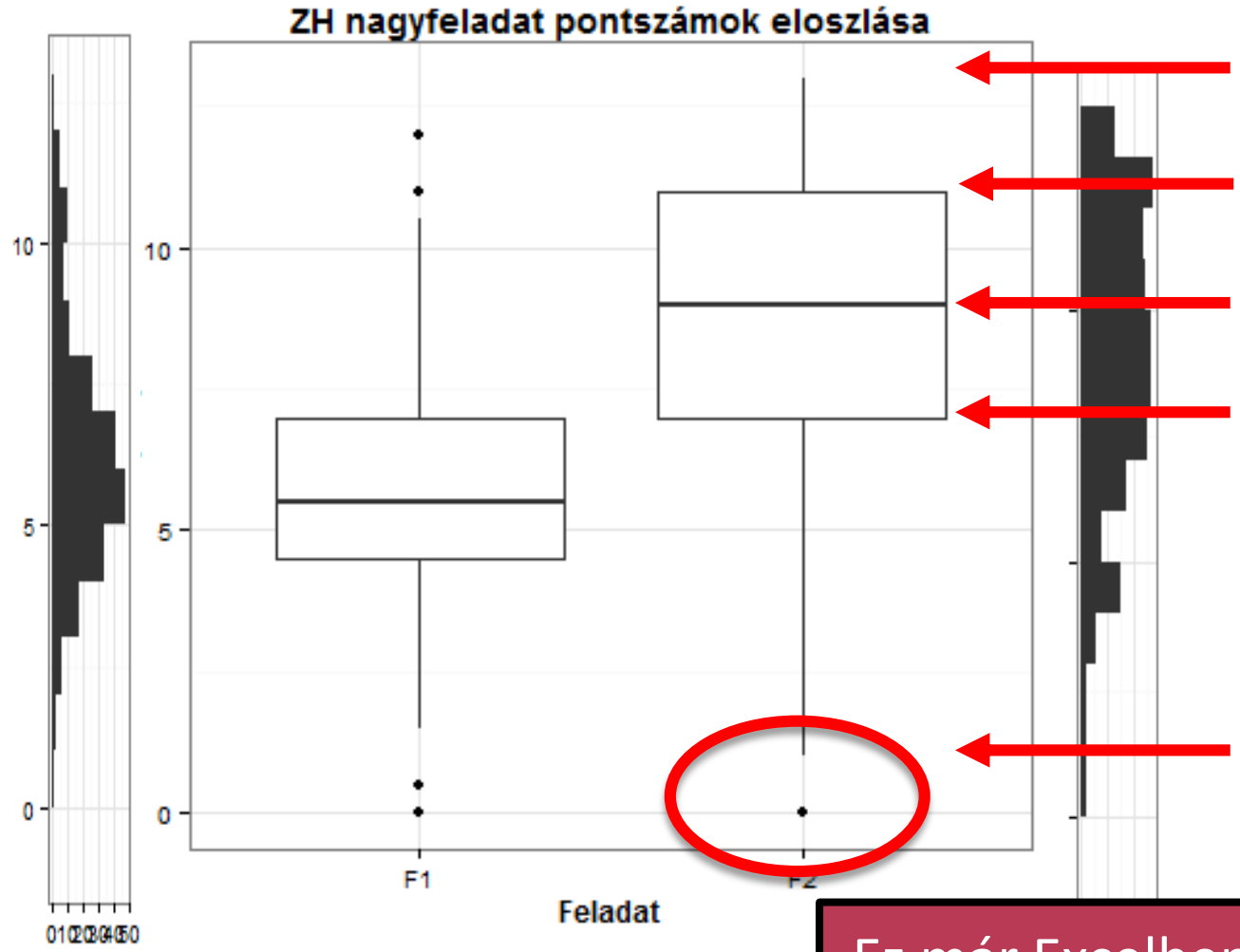
Példa: percentilis ábrázolás



Boxplot (Box and whisker plot)



Boxplot (Box and whisker plot)



Felső outlier határ

Q3

Medián

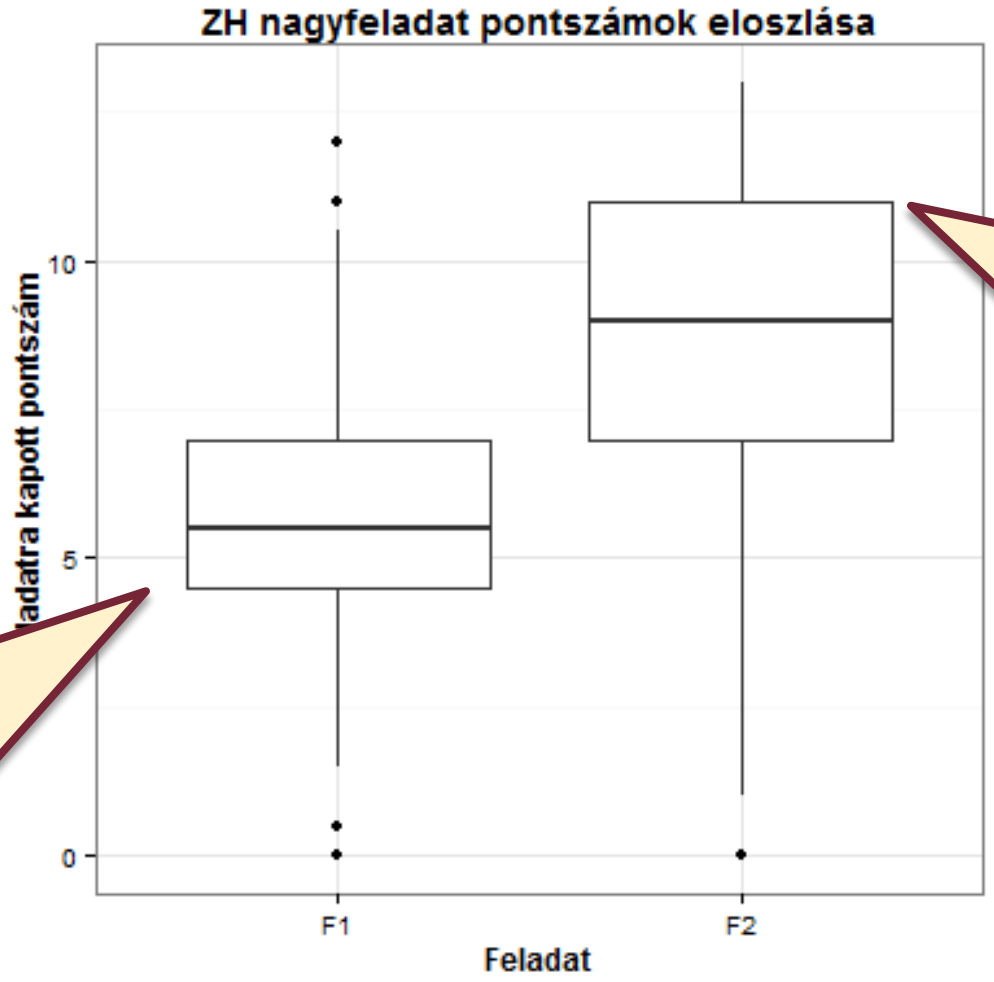
Q1

Alsó outlier határ

Outlierek

Ez már Excelben nem könnyű…

Boxplot (Box and whisker plot)

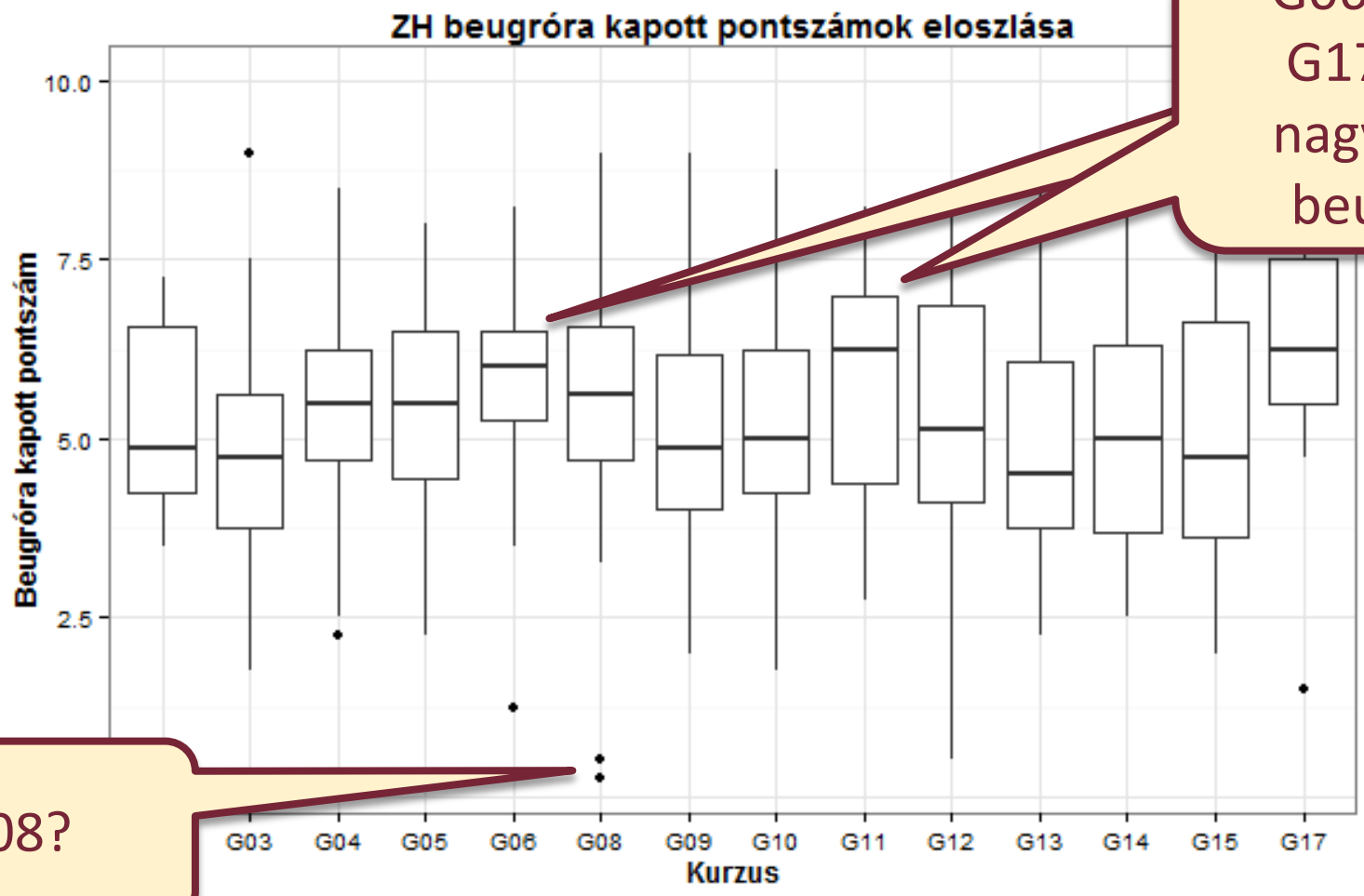


F2-re általában több pontot kaptak, mint F1-re

Az F1 pontszámok 50%-a 4.5 és 7.5 között volt

Boxplot (Box and whisker plot)

- Melyik csoportban hogyan sikerültek a beugrók?

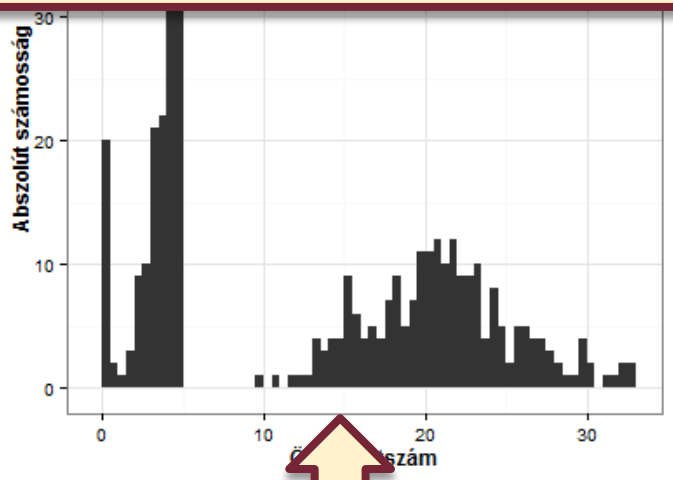


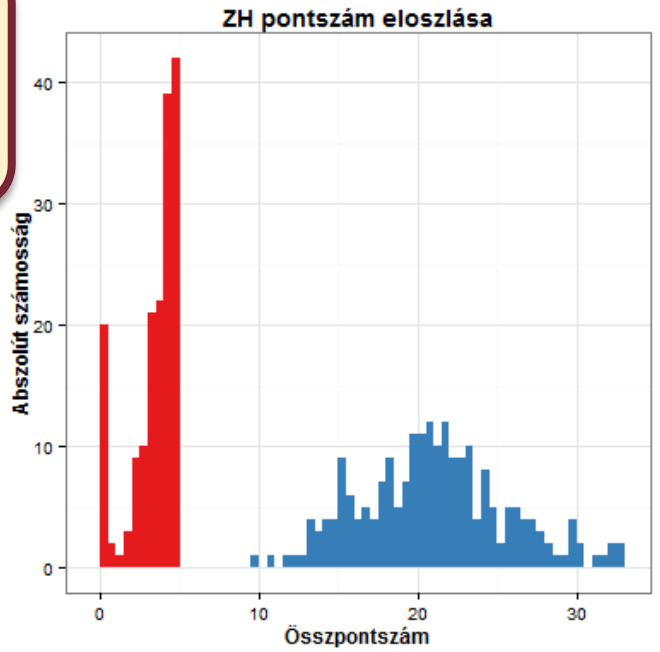
G06, G11, G17-ben nagyon jó beugrók

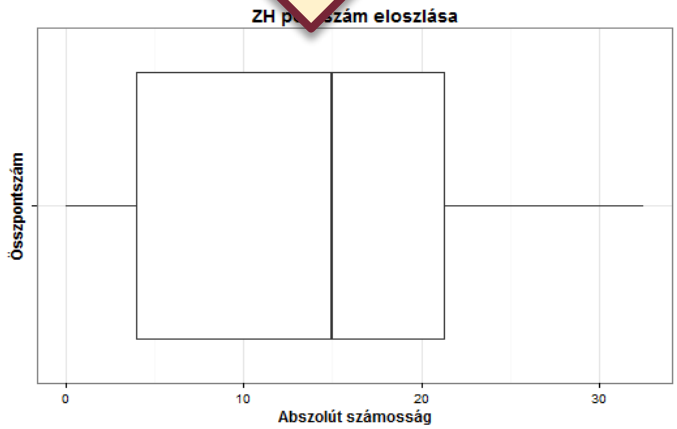
G08?

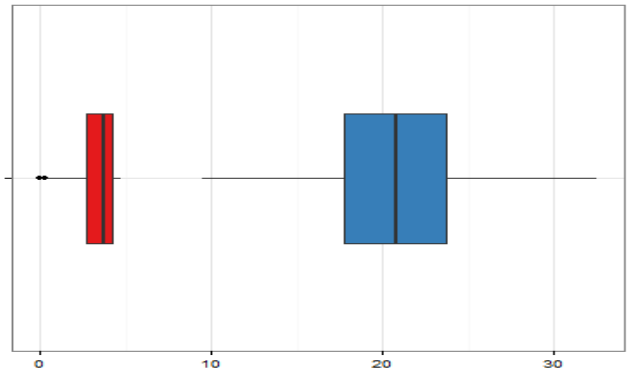
Boxplot (Box and whisker plot)

Absztrakció: a boxplottal fontos információt is veszíthetünk!

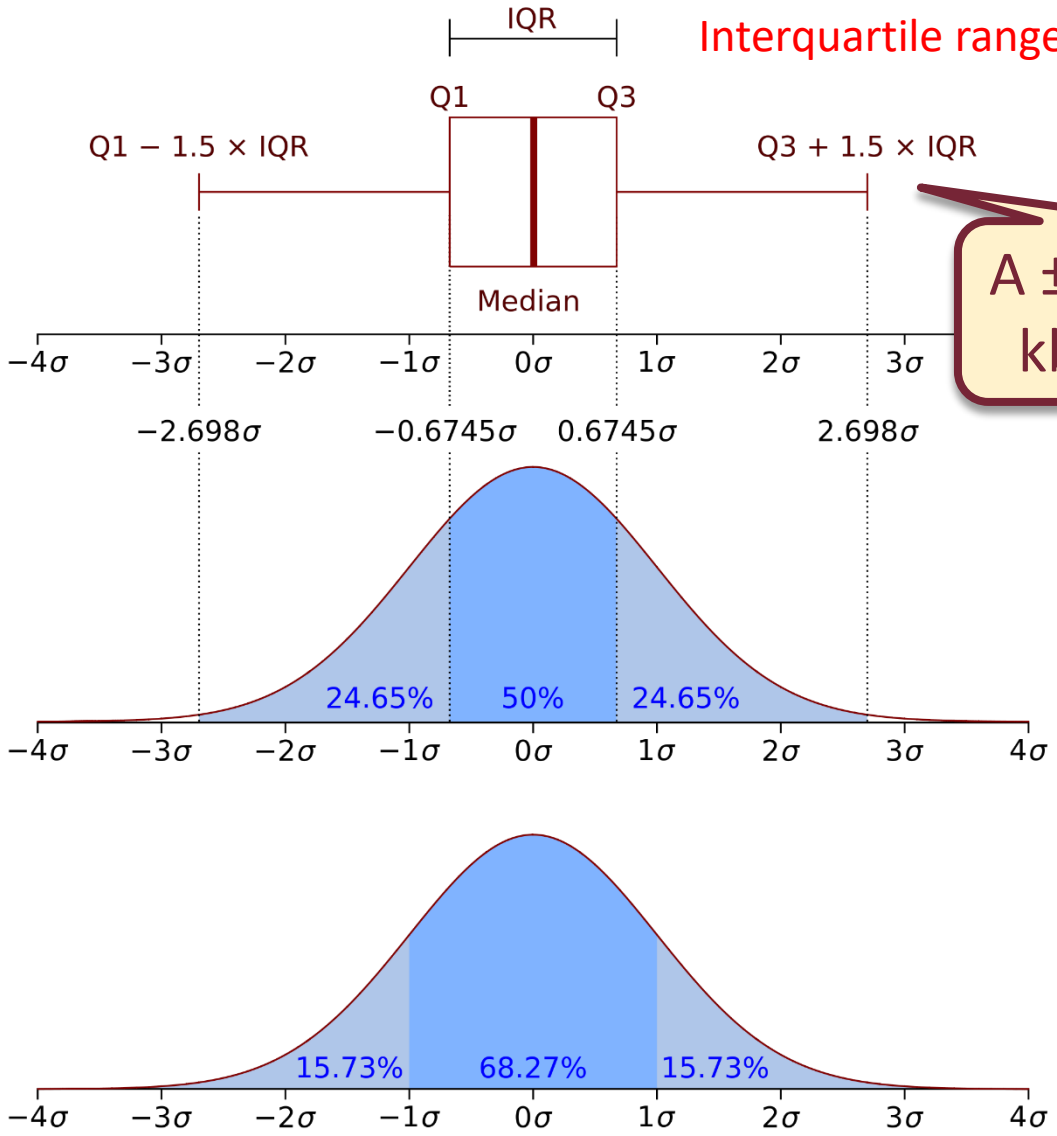






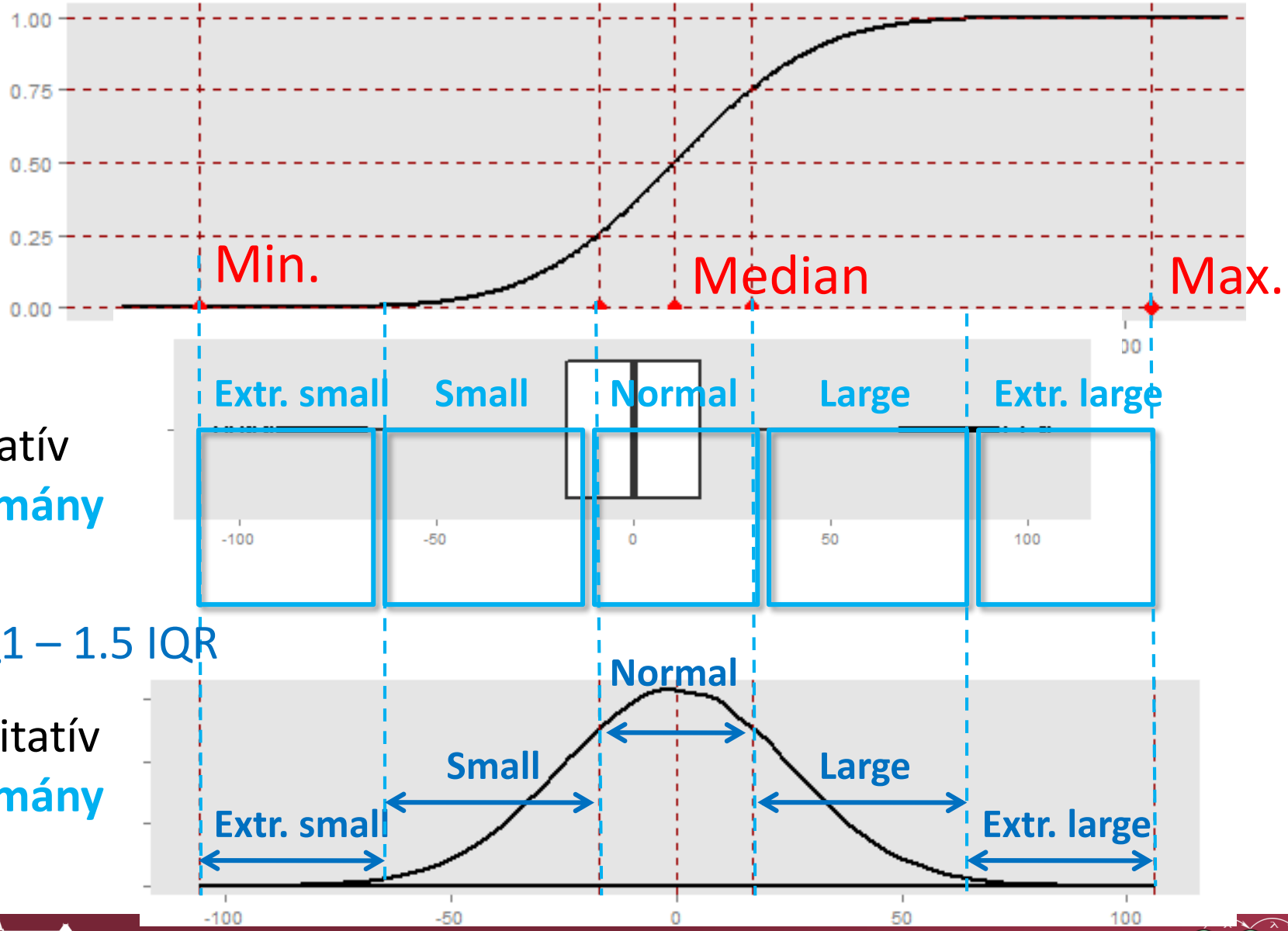


Boxplot (Box and whisker plot)



A ±1.5 × IQR-hez kb. 3σ tartozik

Boxplot: kvalitatív jellemzés



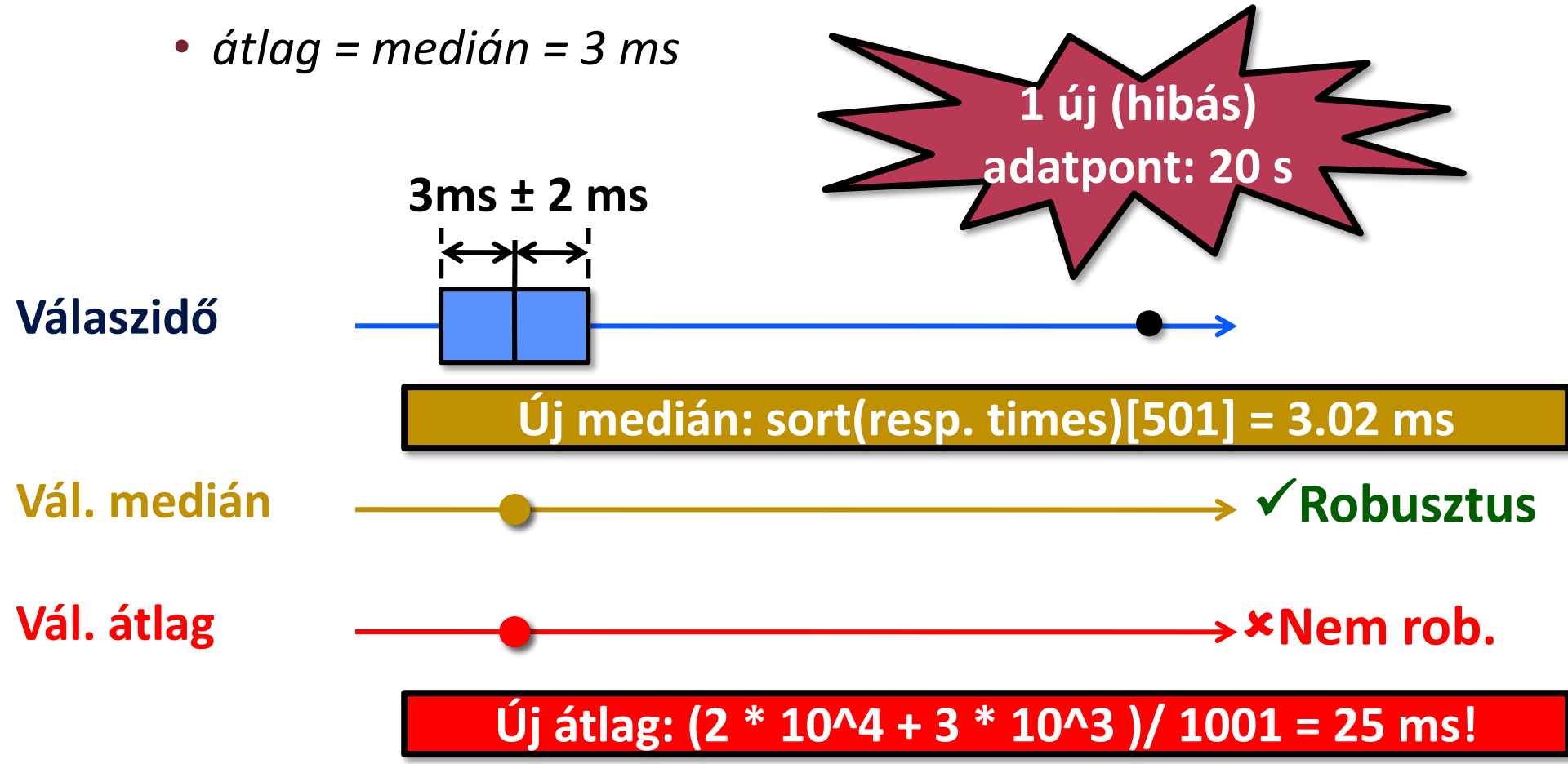
Kvalitatív **tartomány**

Q1 – 1.5 IQR

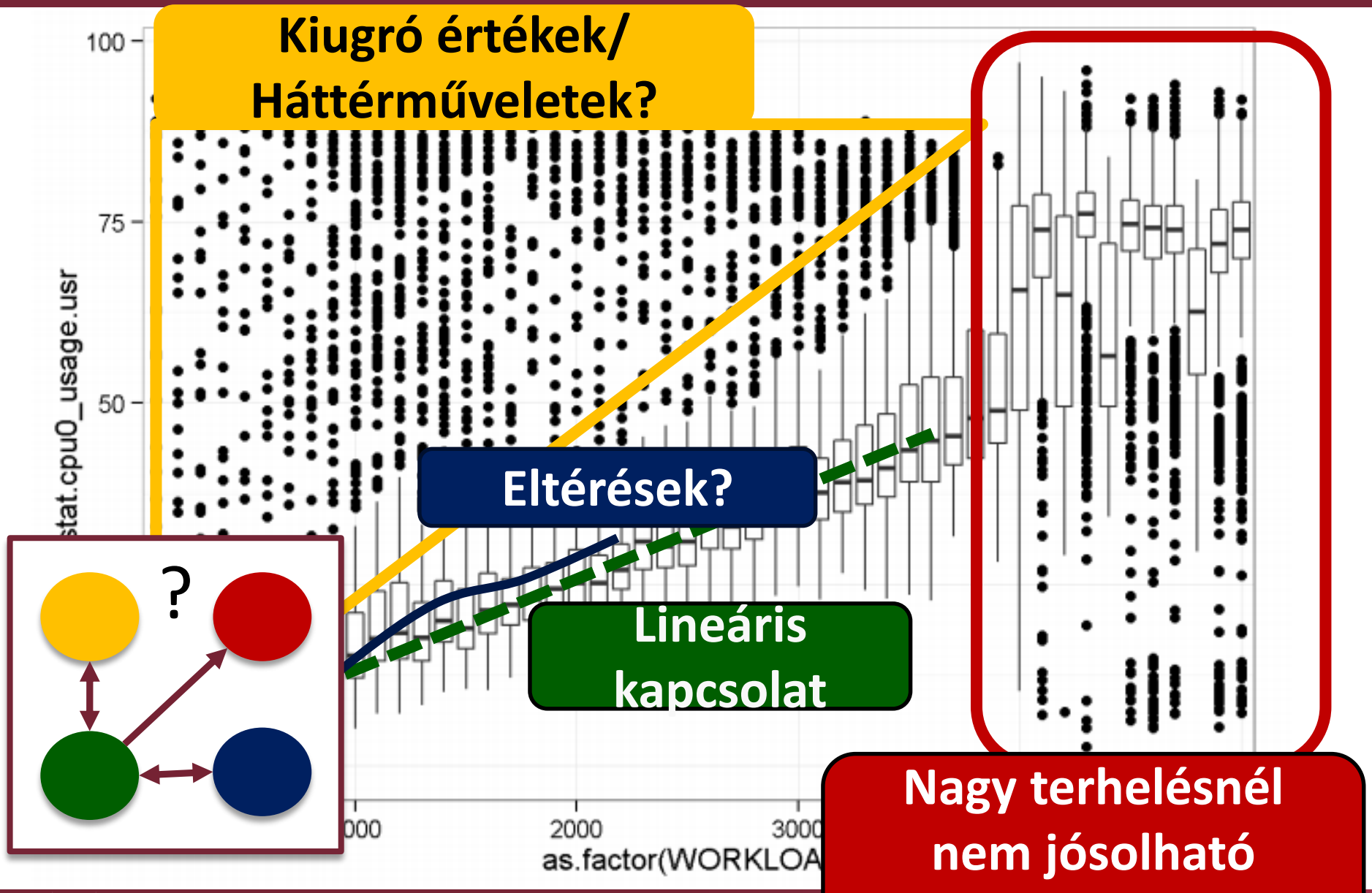
Kvantitatív **tartomány**

Miért medián, miért nem átlag?

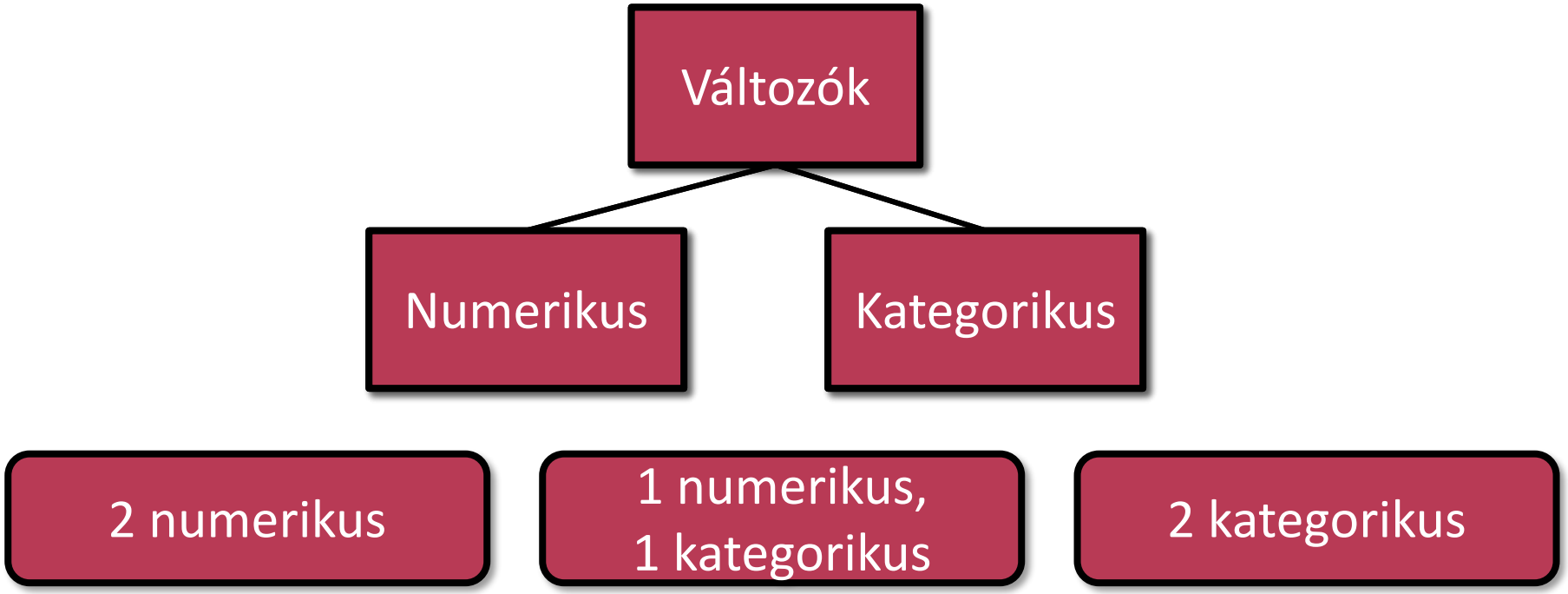
- Alaphalmaz
  - 1000 adatpont ~ U(1, 5) egyenletes eloszlás
    - *átlag = medián = 3 ms*



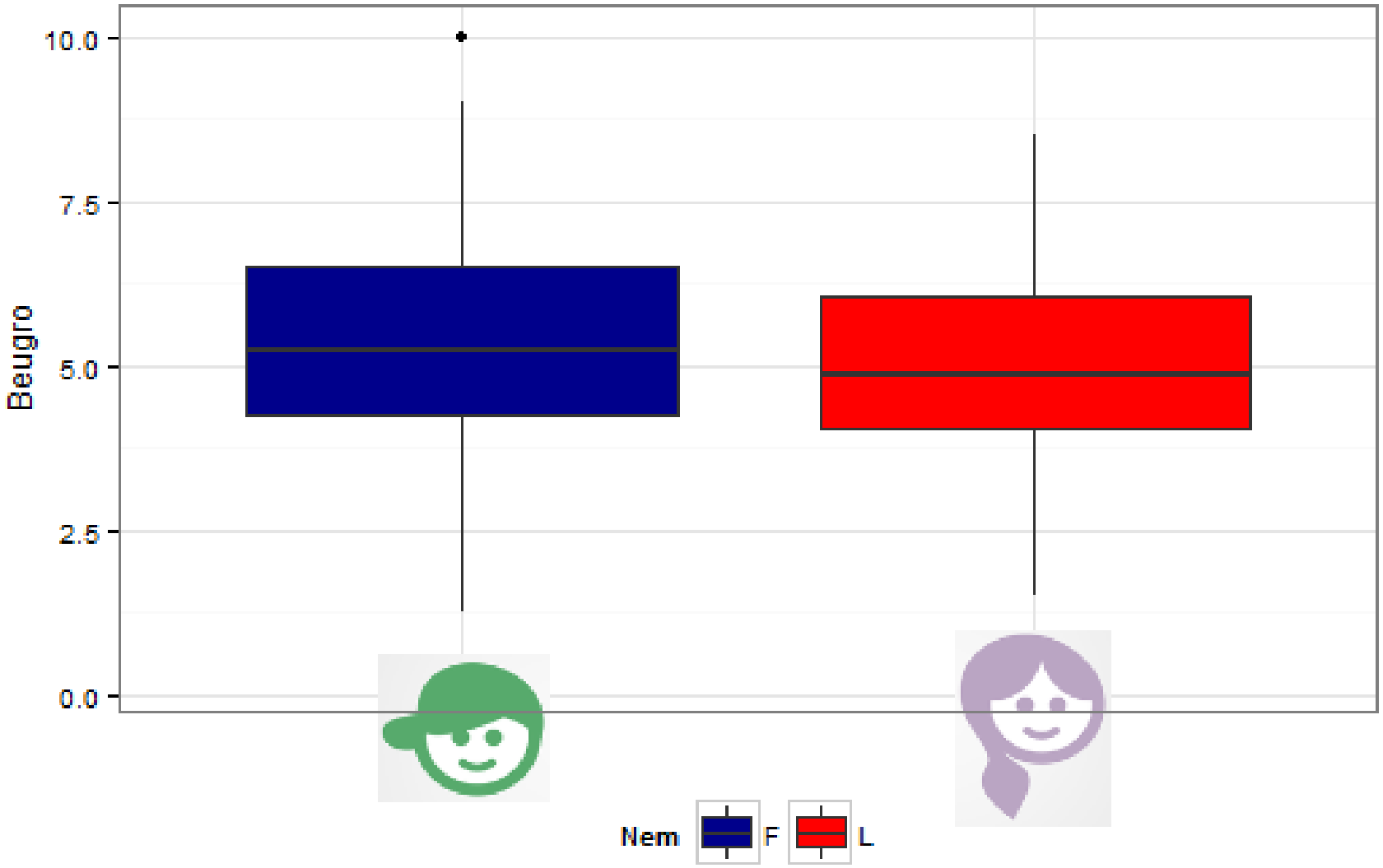
Példa: terhelés vs. kihasználtság



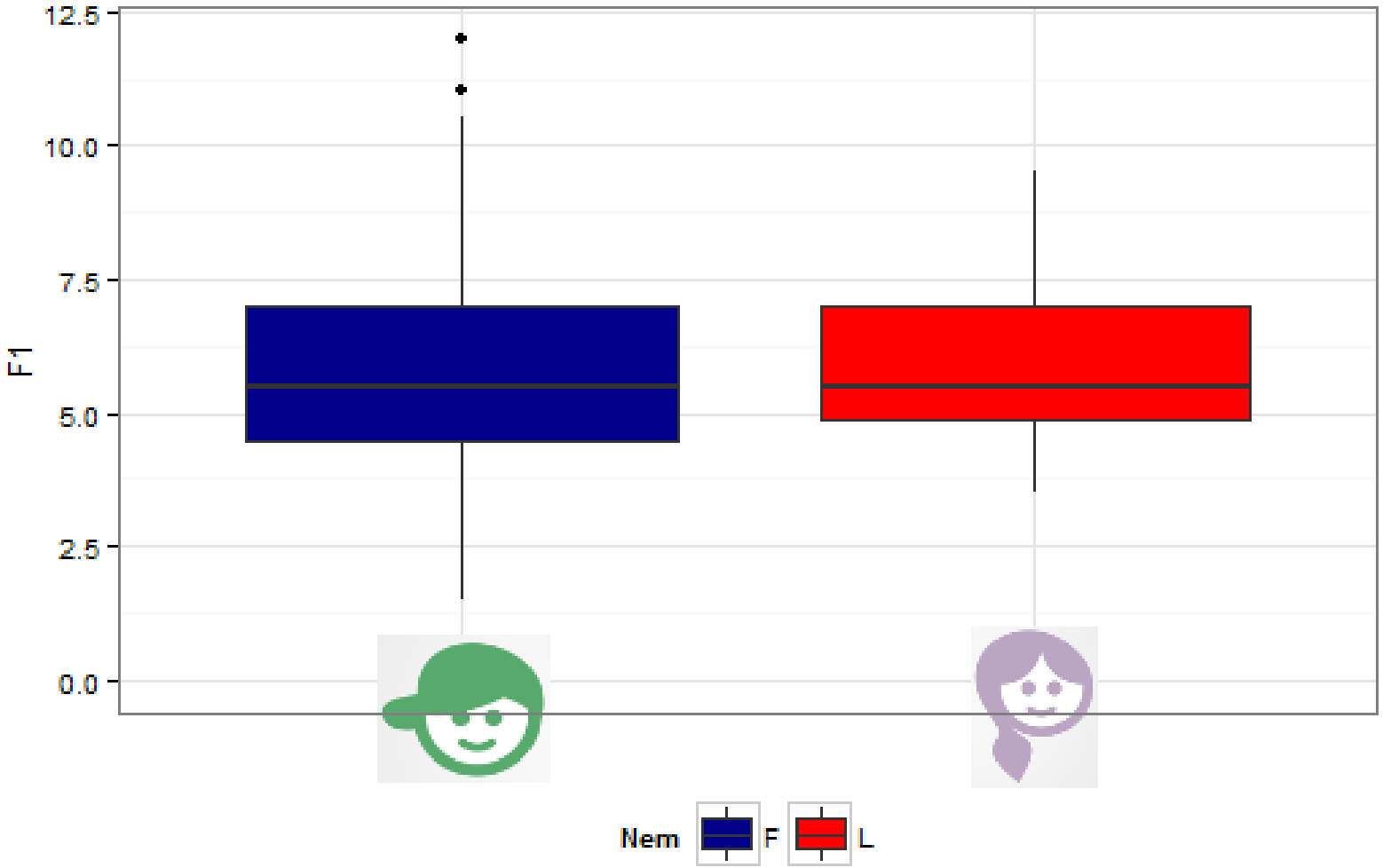
2 változó kapcsolata



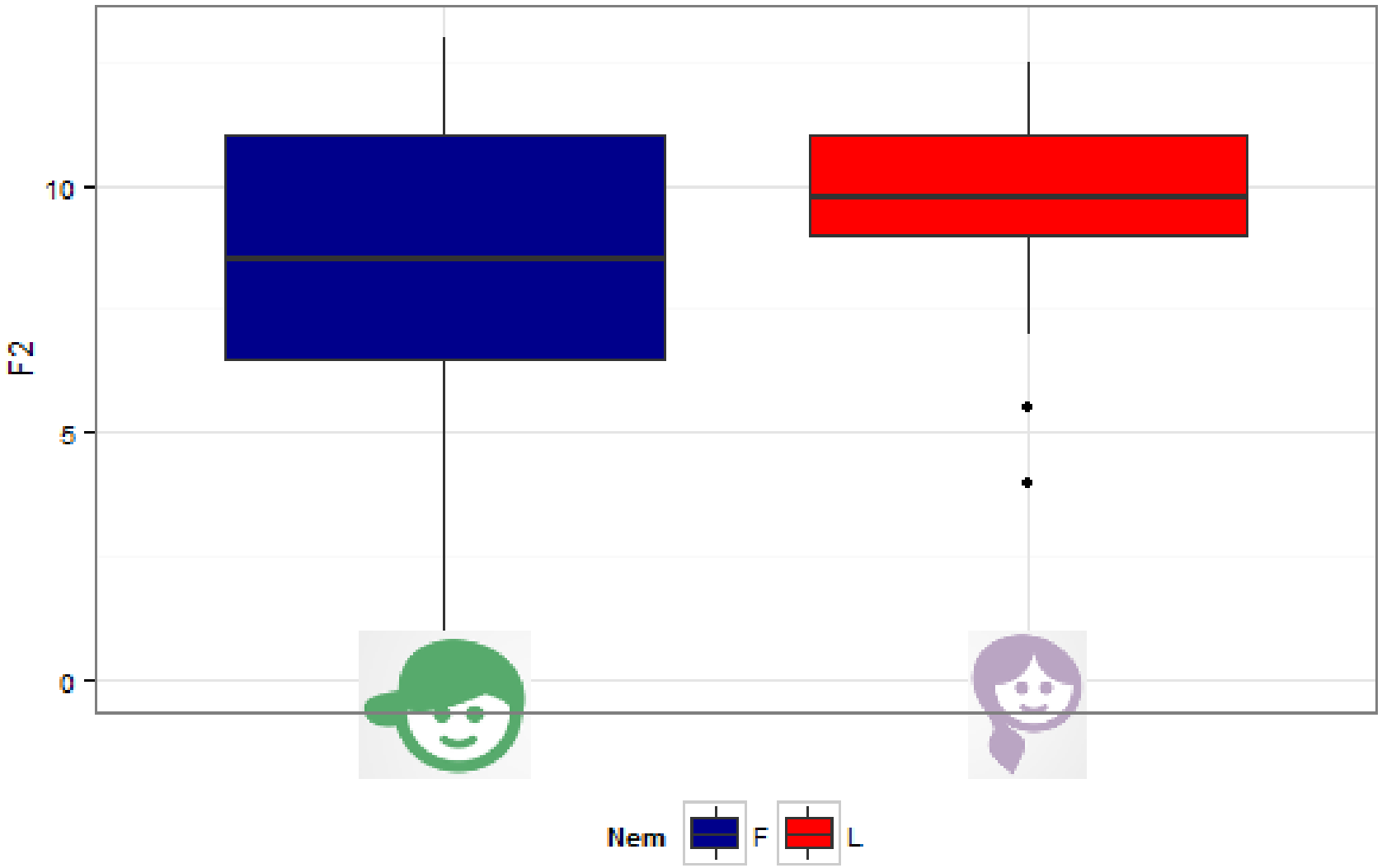
Numerikus kategóriánként



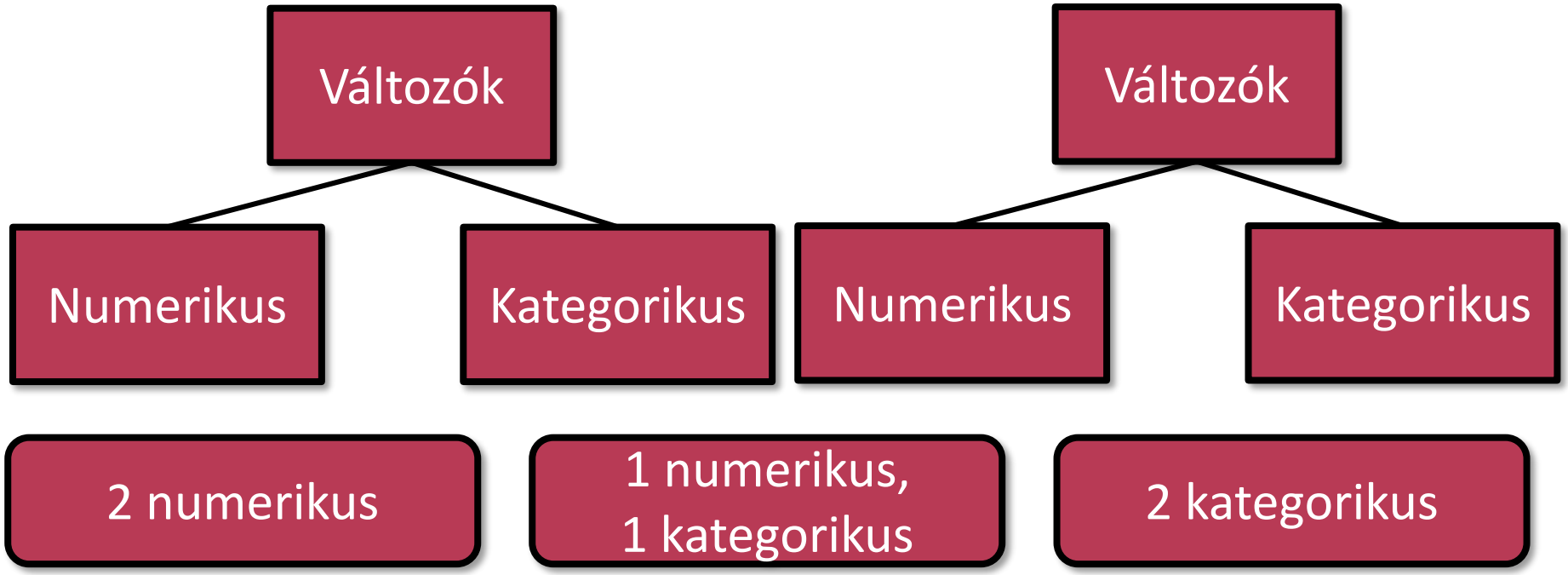
Numerikus kategóriánként



Numerikus kategóriánként



2 változó kapcsolata

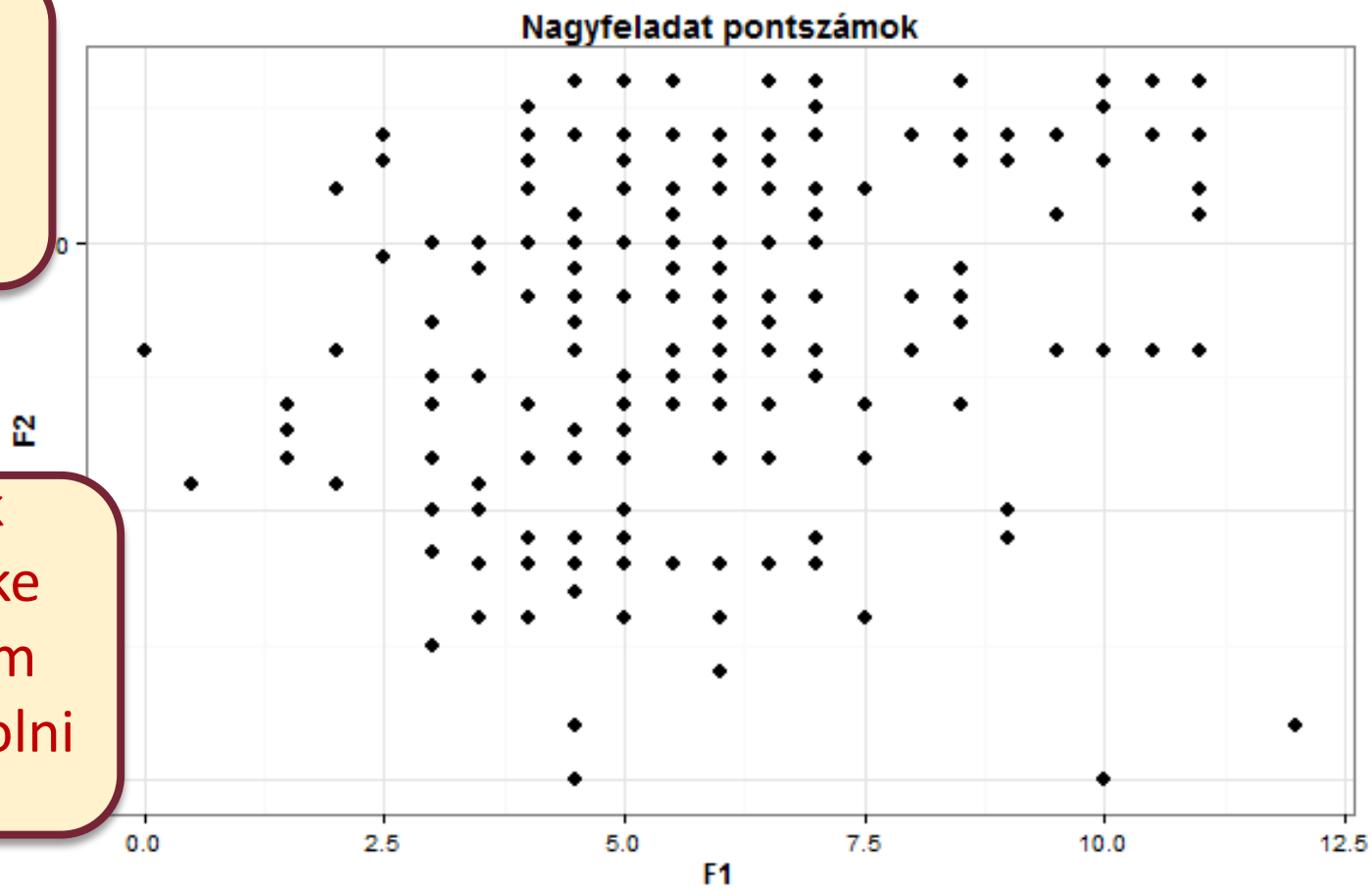


Pont – pont diagram (scatterplot)

- Bemenő változó: nagyfeladatokra kapott pontok
- Kérdés: hogyan viszonyulnak egymáshoz?

Együttesen előforduló pontpárokat vizualizálunk

Ha az egyik változó értéke hiányzik, nem tudjuk felrajzolni ☹



Pont – pont diagram (scatterplot)

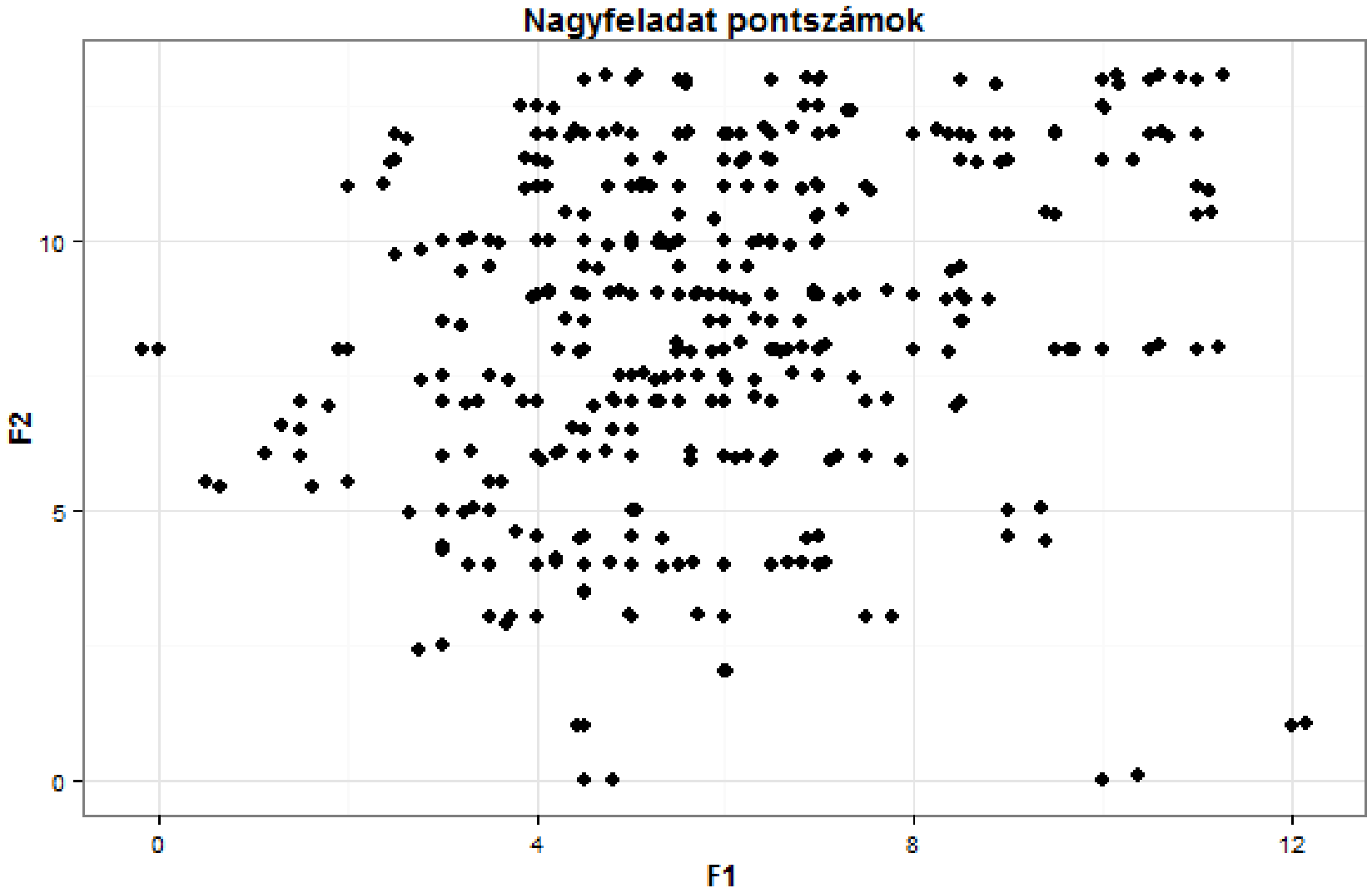
- Bemenő változó: nagyfeladatokra kapott pontok
- Kérdés: hogyan viszonyulnak egymáshoz?

Nem biztos, hogy akinek megy az F1, megy az F2 is

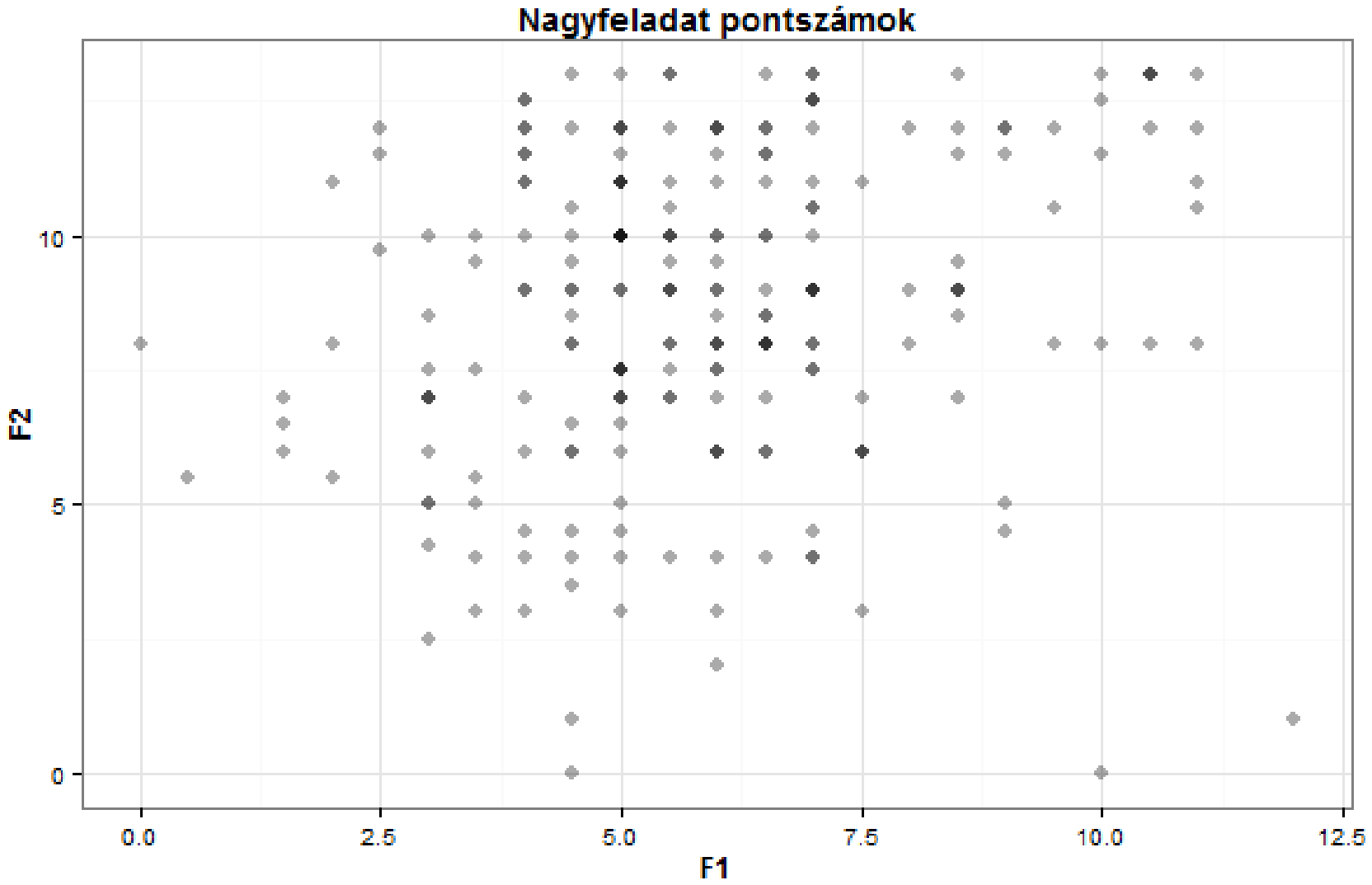
Hogyan kezeljük a takarásokat?

Overplotting

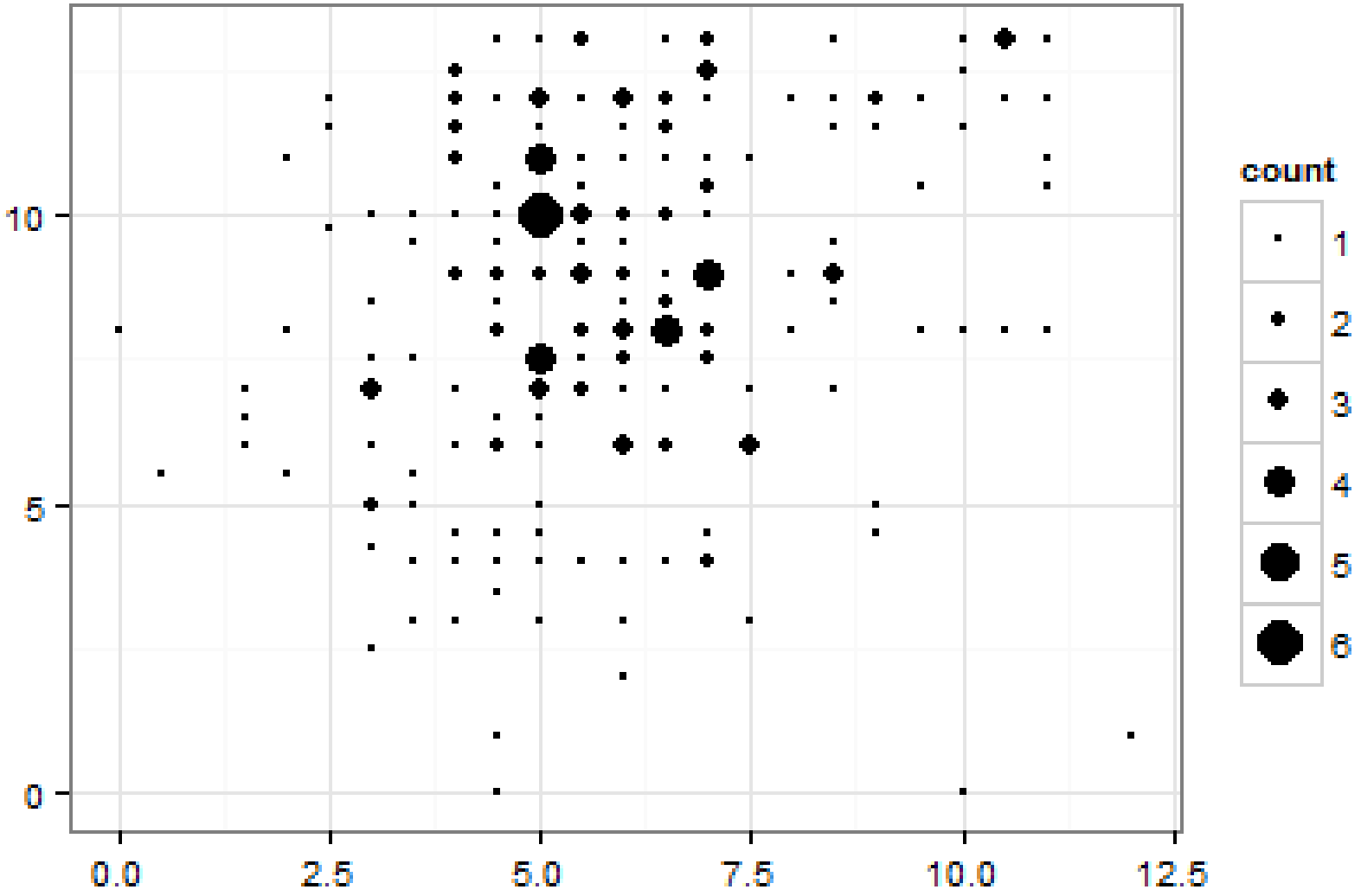
Overplotting megoldások 1: jitter



Overplotting megoldások 2: átlátszóság



Overplotting megoldások 3: méret

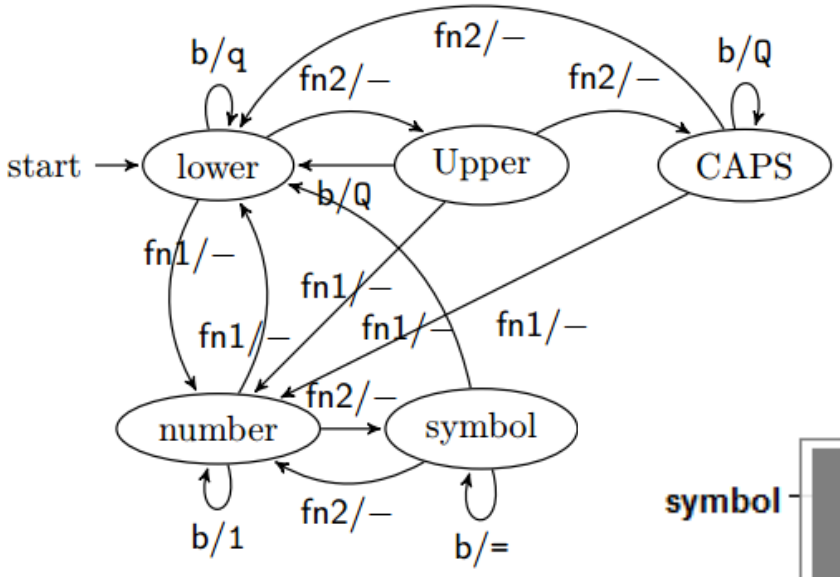


SOK VÁLTOZÓ

>=3 változó

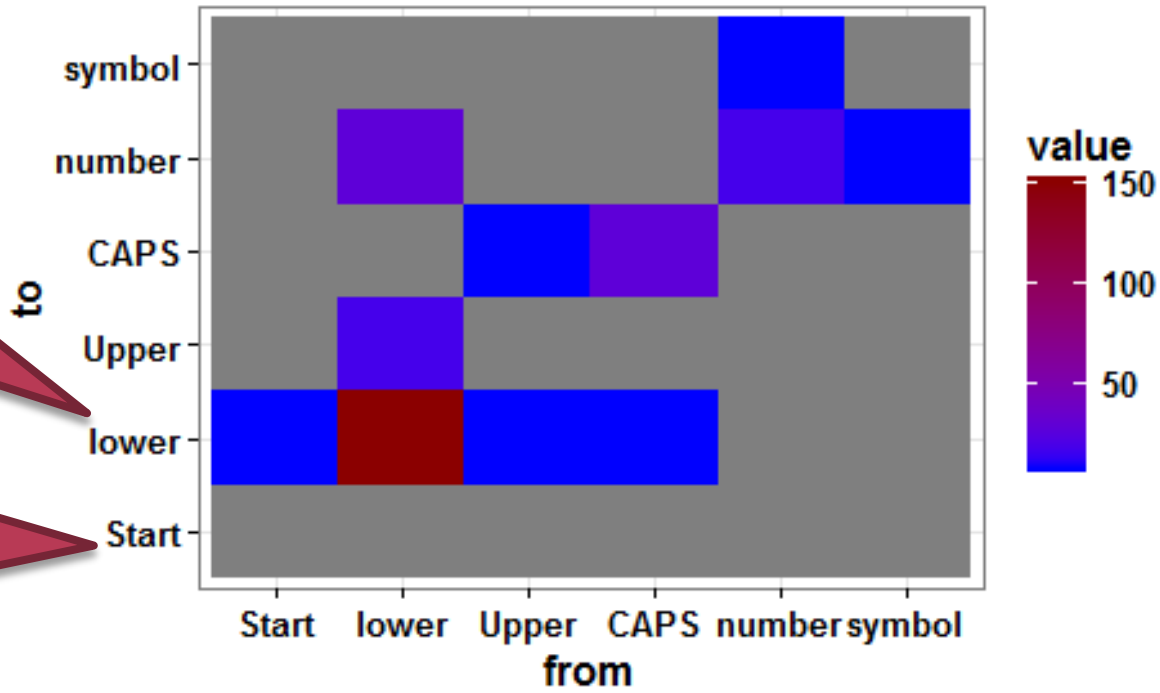
- A grafikai objektumok attribútumait változtatom
  - Szín
  - Méret
  - Textúra
  - Hely – ez triviálisnak tűnik, de a treemapnél van jelentősége
- Pl. heatmap, treemap

Heatmap: lefutási statisztikák

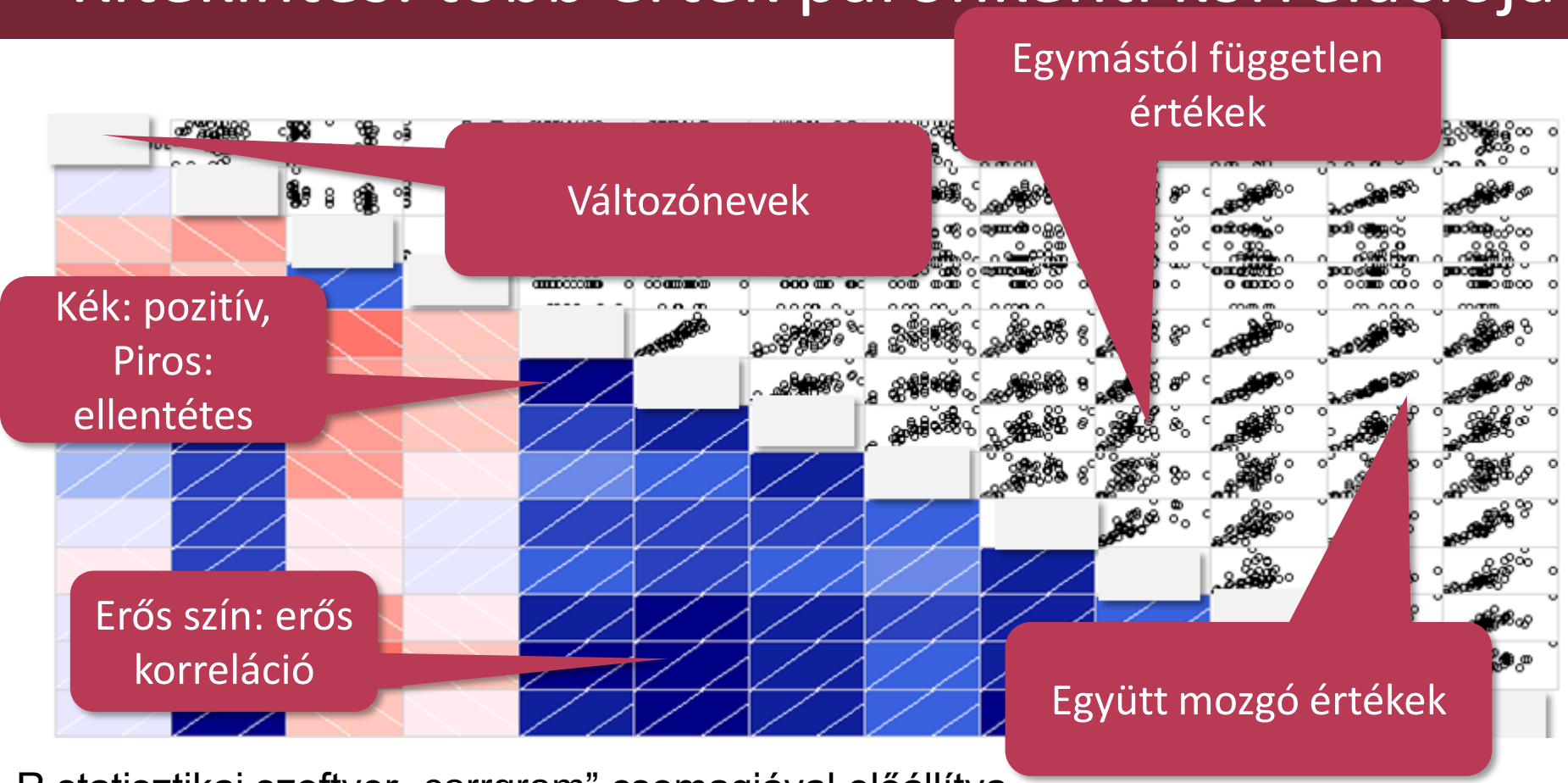


Inkább csak sima szöveget írunk

A Startban mindig csak kezdünk, oda nem jutunk vissza

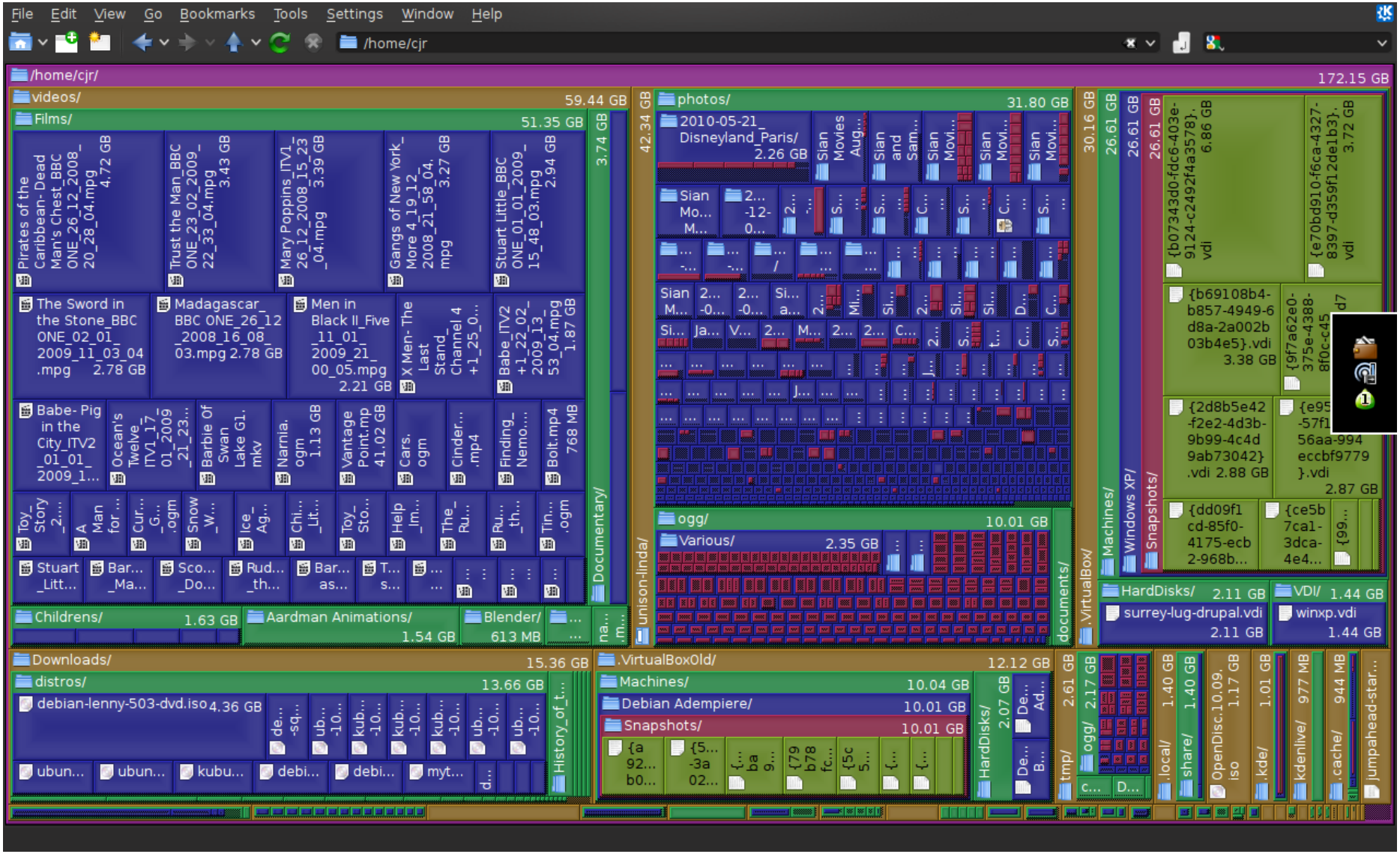


Kitekintés: több érték páronkénti korrelációja



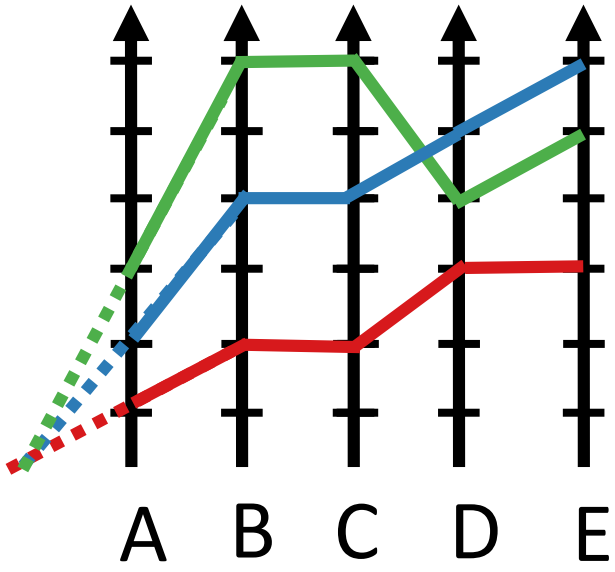
R statisztikai szoftver „corrgram” csomagjával előállítva.
Korreláció (ld. Valószínűségszámítás):
két érték közti lineáris kapcsolat erőssége és iránya
Átló felett: **scatterplot mátrix**
Cél: együtt mozgó értékek kiszűrése, **kiugró értékek (outlierek)** azonosítása.
→Mik a terhelés/előrejelzés szempontjából lényeges változók?

Treemap: állományrendszer



Párhuzamos koordináták

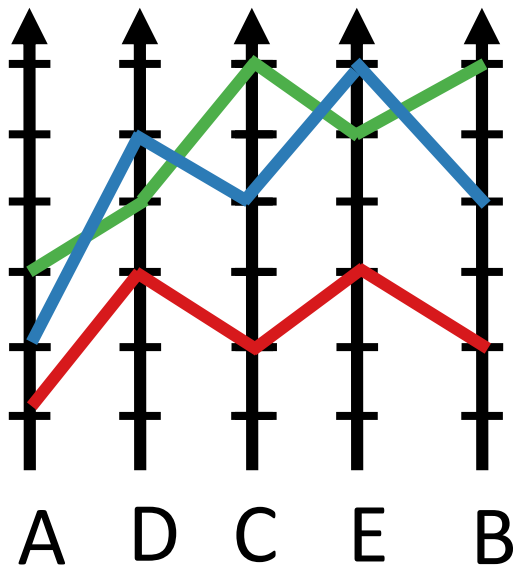
- Tengelyek: dimenziók/koordináták
  - tetszőleges számú
  - tetszőleges skála
- Egy vonal egy mérés (darabszám?)
- Kompakt és skálázható
- Koordináta sorrend befolyásolja a kiértékelést



[2, 4, 4, 5, 6]

[3, 6, 6, 4, 5]

[1, 2, 2, 3, 3]



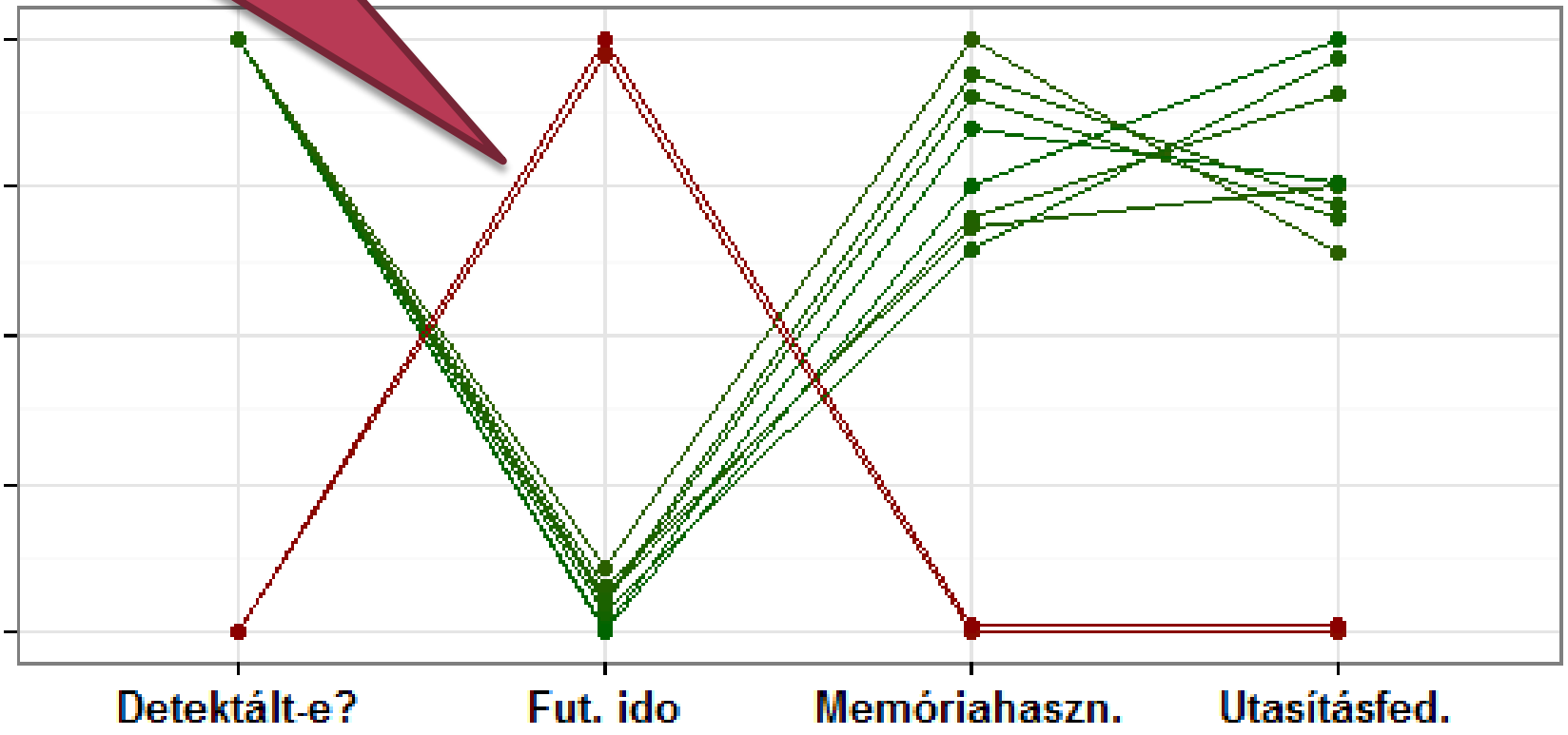
[3, 4, 6, 5, 6]

[2, 5, 4, 6, 4]

[1, 3, 2, 3, 2]

Párhuzamos koordináták: tesztesetek elemzése

1 teszteset 1 törött vonal

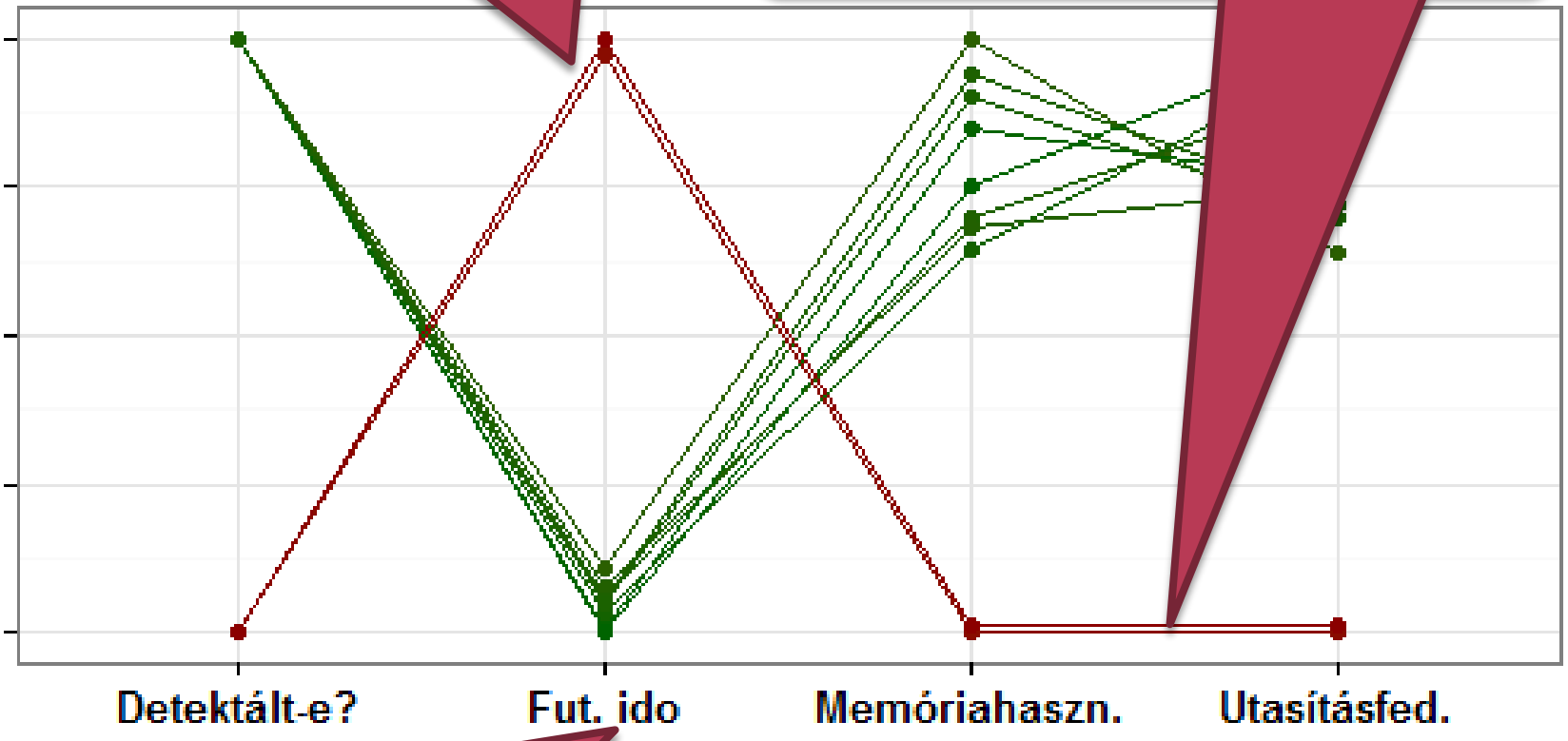


A változók az x tengelyen jelennek meg

Párhuzamos koordináták: tesztesetek elemzése

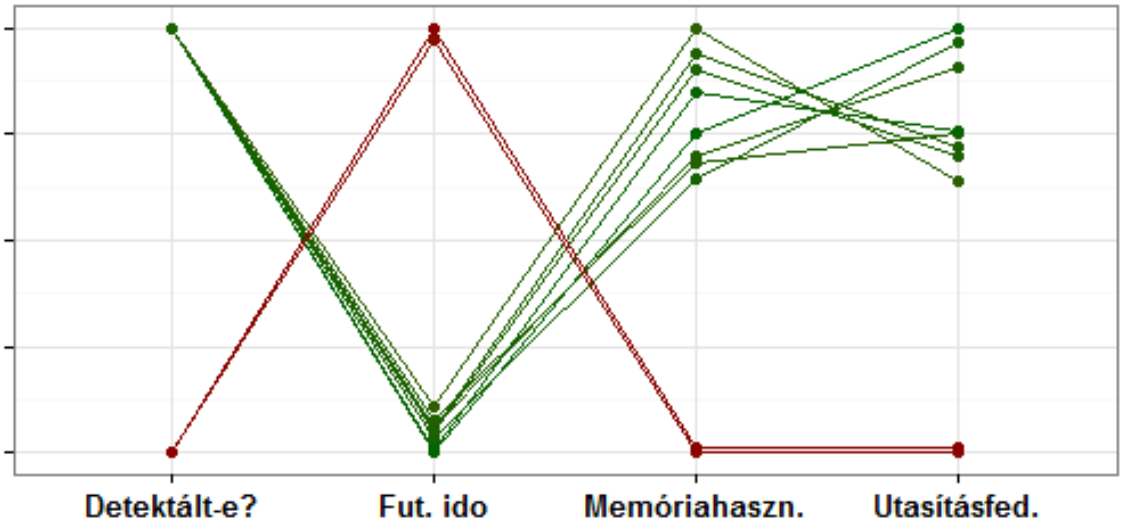
Timeout?

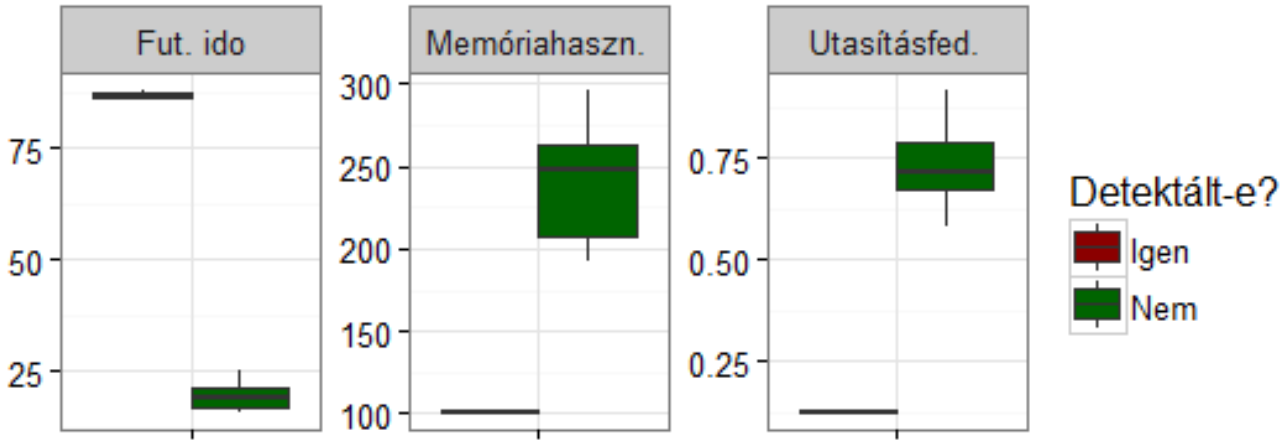
A hibát detektálók az érdemi számításig valószínűleg el sem jutnak



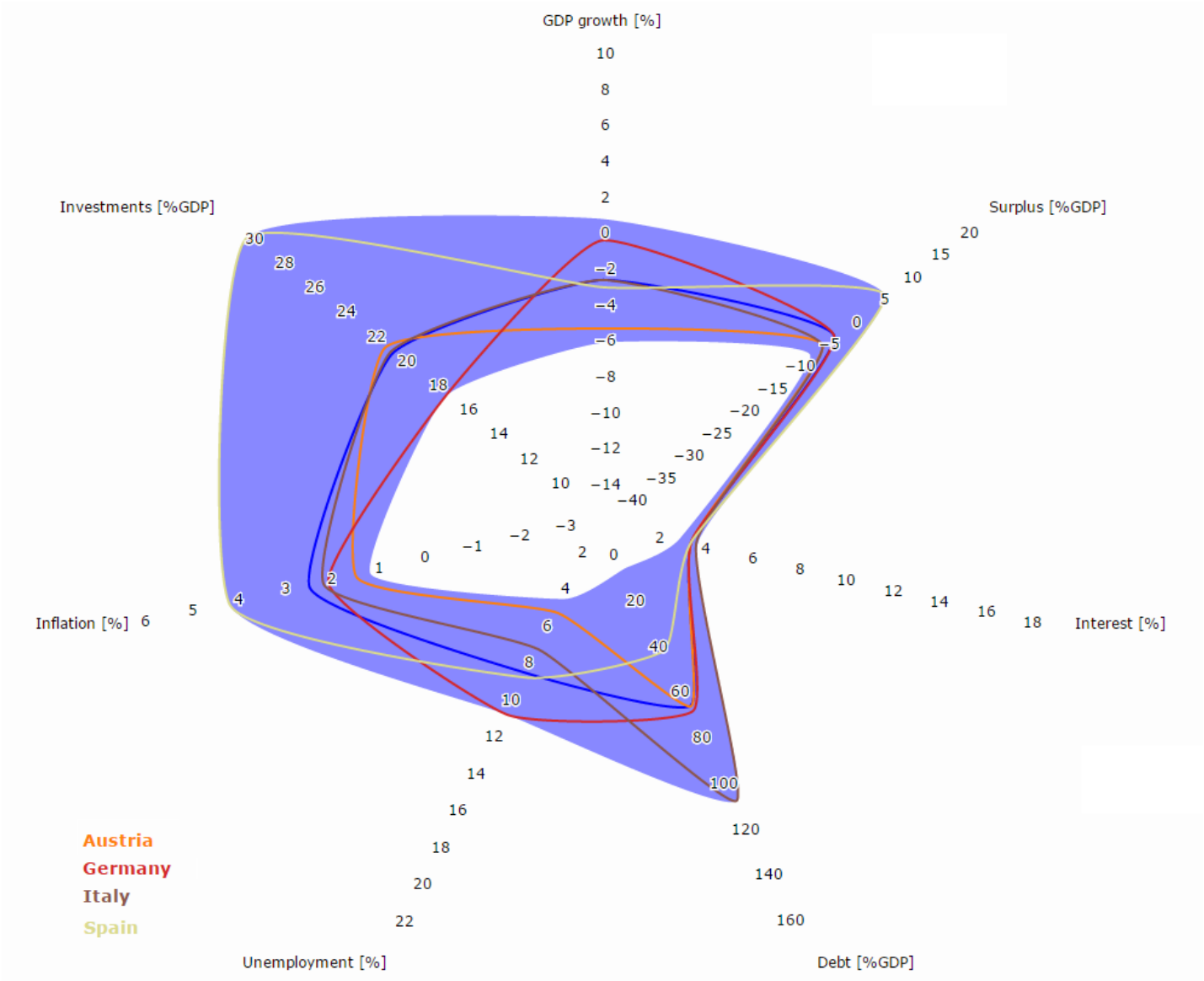
A futási idő és a memóriahasználat valószínűleg pozitív kapcsolatban állnak (sikeres teszteknél)

Párhuzamos koordináták: viz. alternatívák





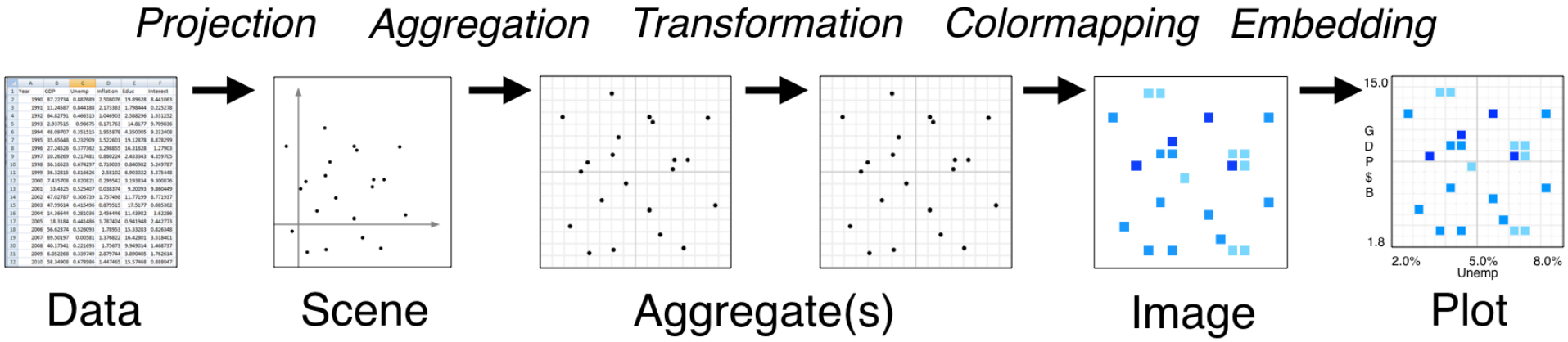
Radar chart: egy párhuzamos koord. kiterjesztés



AGGREGÁLÁS + VIZUÁLIS MEGJELENÉS

Datashader

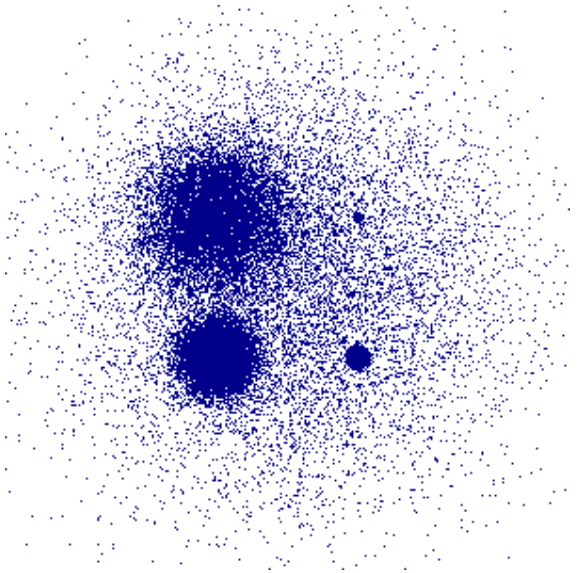
- Gyakran nagy az adat (1000000+)
- DE nem az egyedi adatvektor, hanem az értékek eloszlása érdekes



http://datashader.org/getting_started/2_Pipeline.html

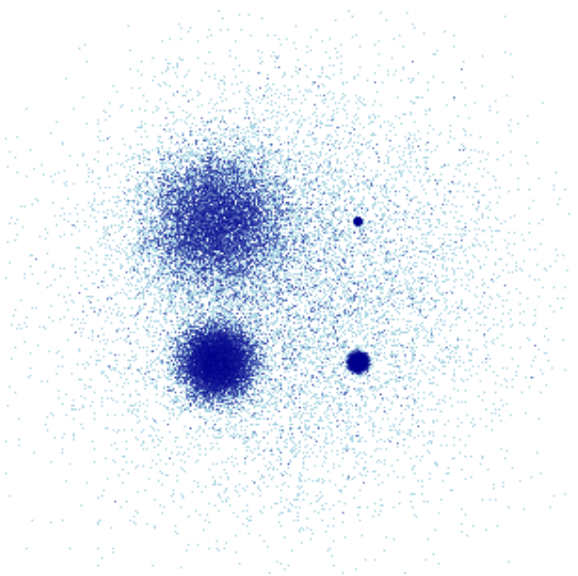
- Adat
- for x, y, val, cat in
[(2, 2, 10, "d1"),
(2, -2, 20, "d2"),
(-2, -2, 30, "d3"),
(-2, 2, 40, "d4"),
(0, 0, 50, "d5")]

Előfordulás
Cellaintenzitás :
Van-e adatpont a cellában?



Gyakoriság
Cellaintenzitás :

$$\frac{\text{Cellabeliek száma}}{\text{Összes minta}}$$

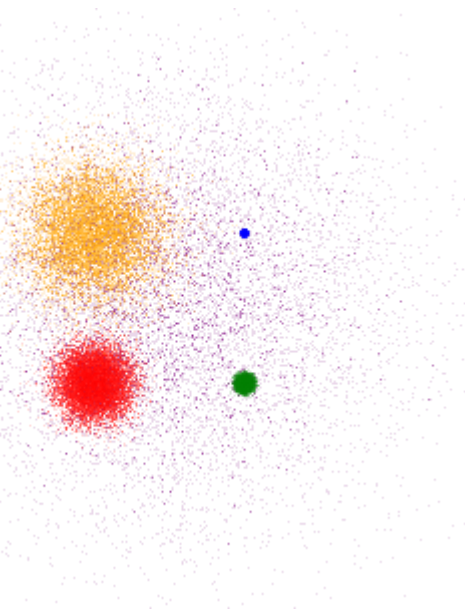


Szűrés, színezés

- Adat
- for x, y, val, cat in
  [(2, 2, 10, "d1"),
  (2, -2, 20, "d2"),
  (-2, -2, 30, "d3"),
  (-2, 2, 40, "d4"),
  (0, 0, 50, "d5")]

Színezés
kategóriánként

```
color_key = dict(
        d1='blue',
        d2='green',
        d3='red',
        d4='orange',
        d5='purple')
```

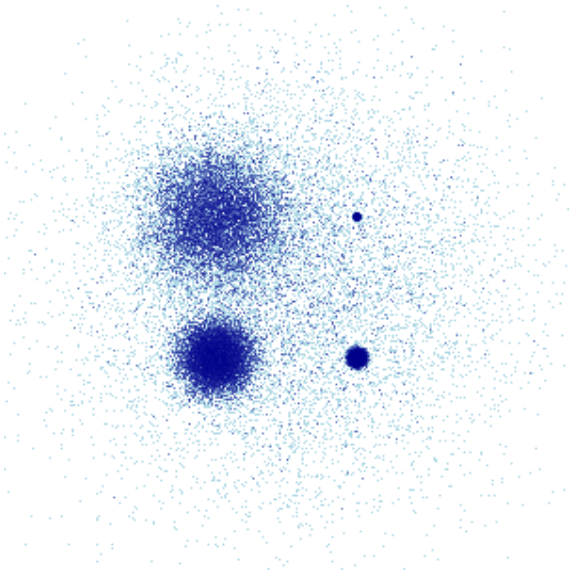


Szűrt előfordulás
Cellaintenzitás
 Cellabeliek száma<––
 Cellabeliek száma|d3)

Súlyozás

- Adat
- for x, y, val, cat in
  [(2, 2, 10, "d1"),
  (2, -2, 20, "d2"),
  (-2, -2, 30, "d3"),
  (-2, 2, 40, "d4"),
  (0, 0, 50, "d5")]

Gyakoriság
Cellaintenzitás :
Gyakoriság
négyzete



Cellaintenzitás
Cellabeliek
kivéve kiesők