

Vizuális adatelemzés

Rendszermodellezés 2020.

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék



Tartalom

Miért vizualizálunk?



Mit vizualizálunk?



Hogyan vizualizálunk?



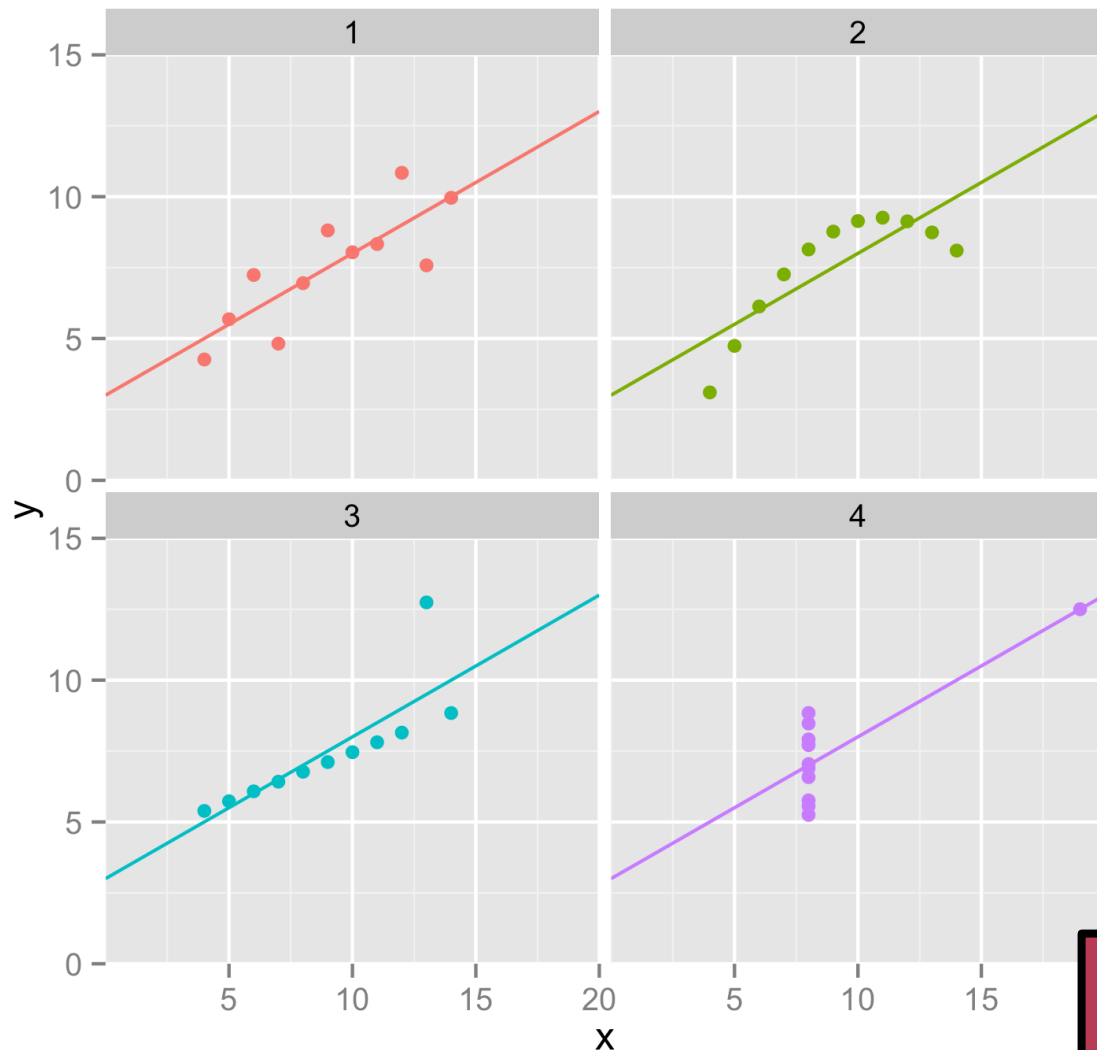
Mire következtetünk?

Vizuális adatelemzés célja

- Rendszermodell megalkotása
 - Hogyan működik a rendszer?
- Rendszermodell validálása
 - Igazolják-e az adatok az elképzelést?
- Rendszermodell paraméterezése
 - Milyen számszerű értékek a leghitelesebbek?
- Rendszer elemzésének/futásának kiértékelése
 - Jól működik-e a rendszer, teljesíti-e a követelményeket?

Számítások ellenőrzése

Anscombe's Quartet



■ Minden esetre:

Átlag:

$$M[x] = 9$$

$$M[y] \sim 7.5$$

Szórás:

$$\sigma[x] = 11$$

$$\sigma[y] \sim 4.12$$

Korreláció:

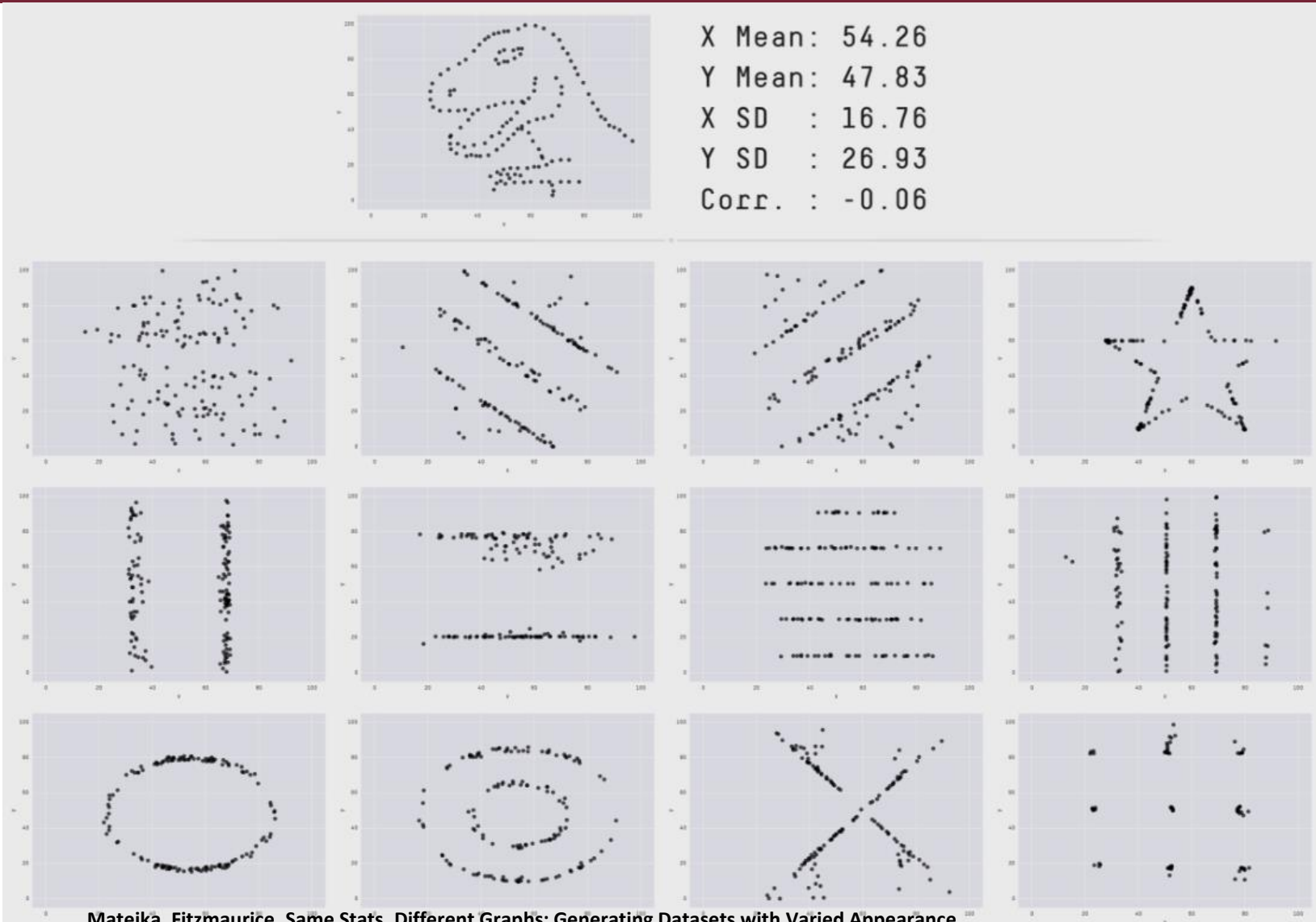
$$C(x, y) \sim 0.816$$

Becslés:

$$y \sim 4 + 0.5x$$

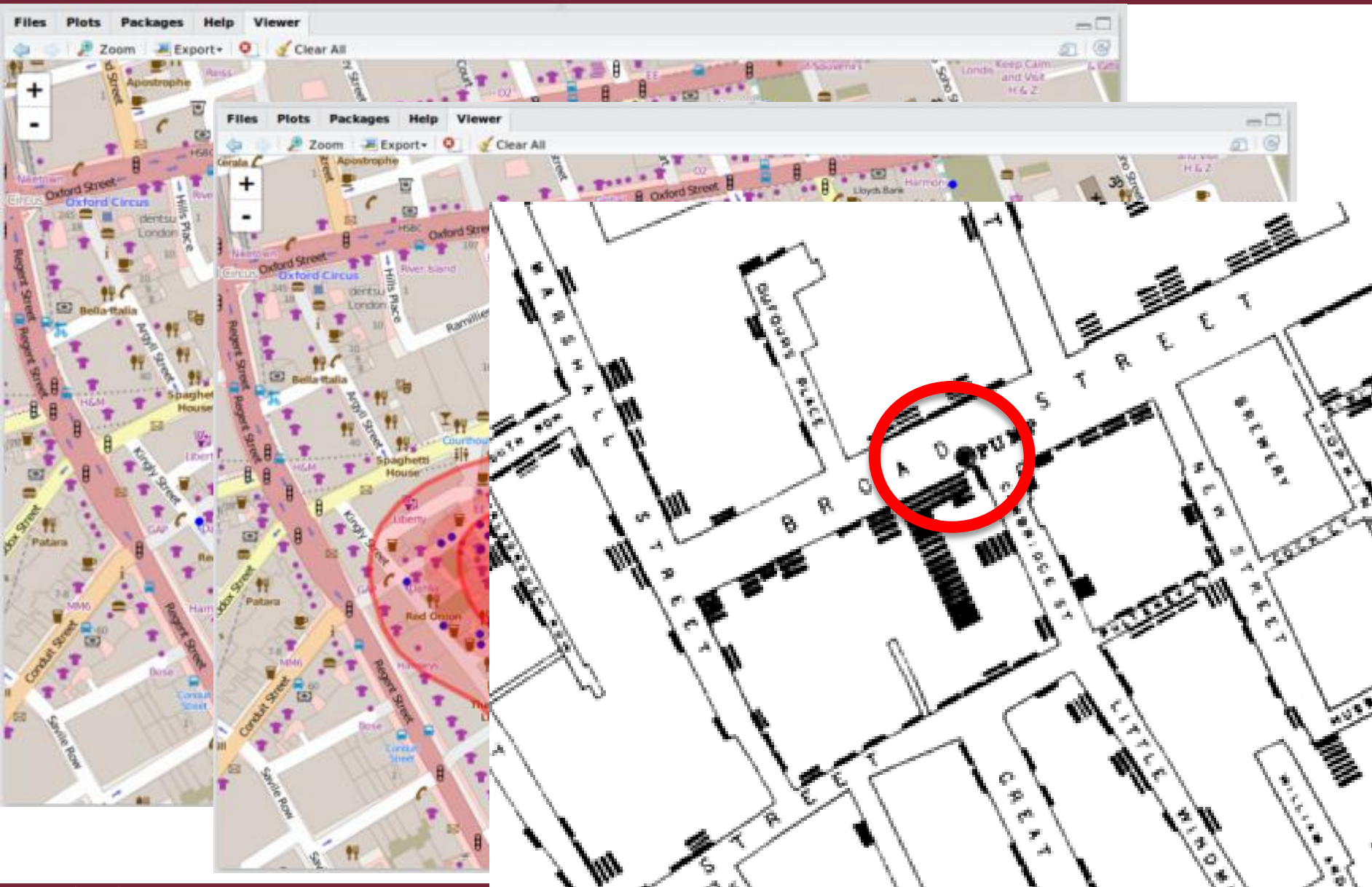
Hibás feltételezések
elkerülése... és intuíció

... egyéb példák



Matejka, Fitzmaurice. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. ACM SIGCHI Conference on Human Factors in Computing Systems
<https://www.autodeskresearch.com/publications/samestats>

Összefüggések feltárása



Mindent a szemnek!

„Masszív” erőforrások

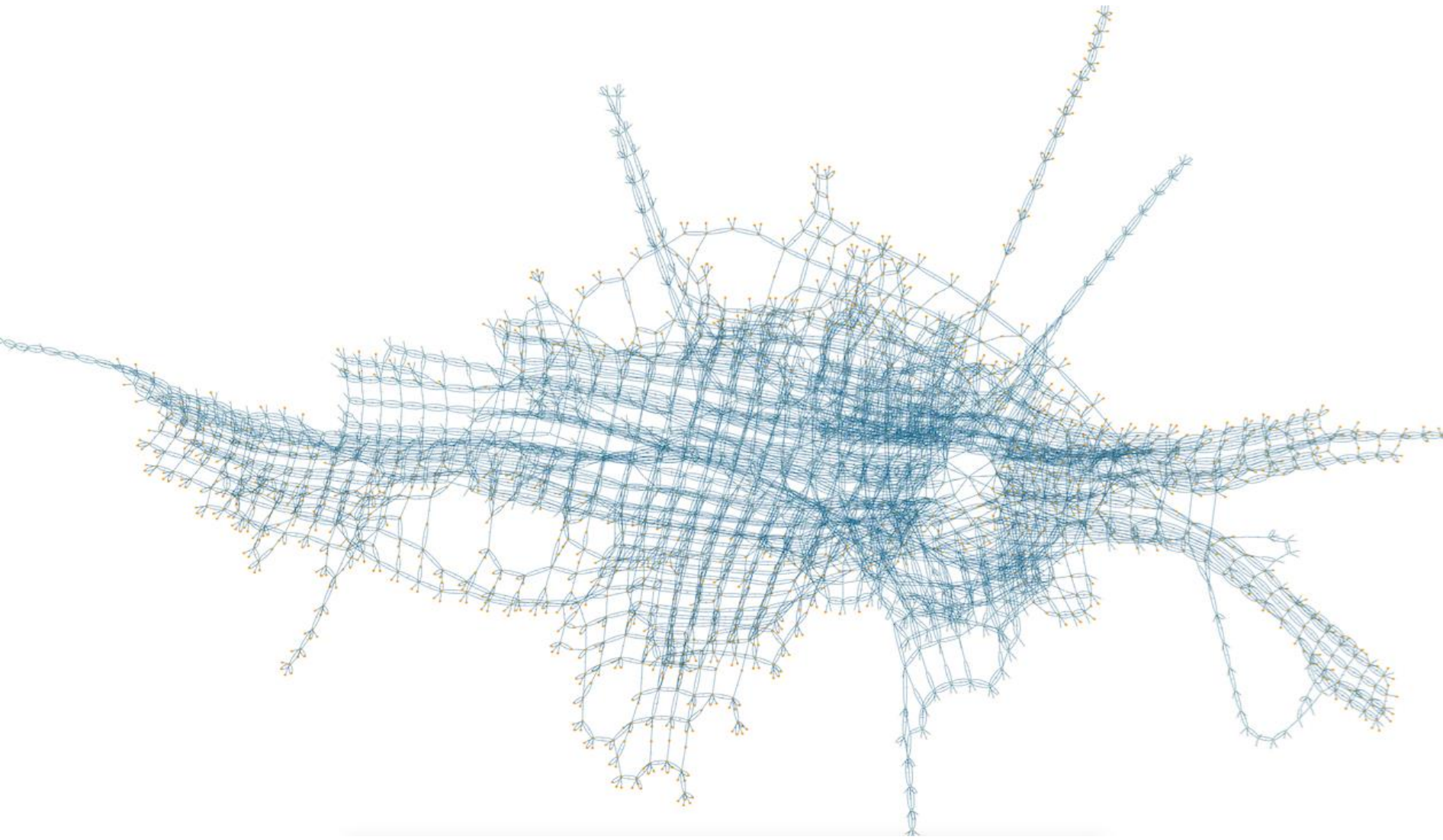
- 120.000.000 szenzor



3. Vizuális kiválasztás és manipuláció

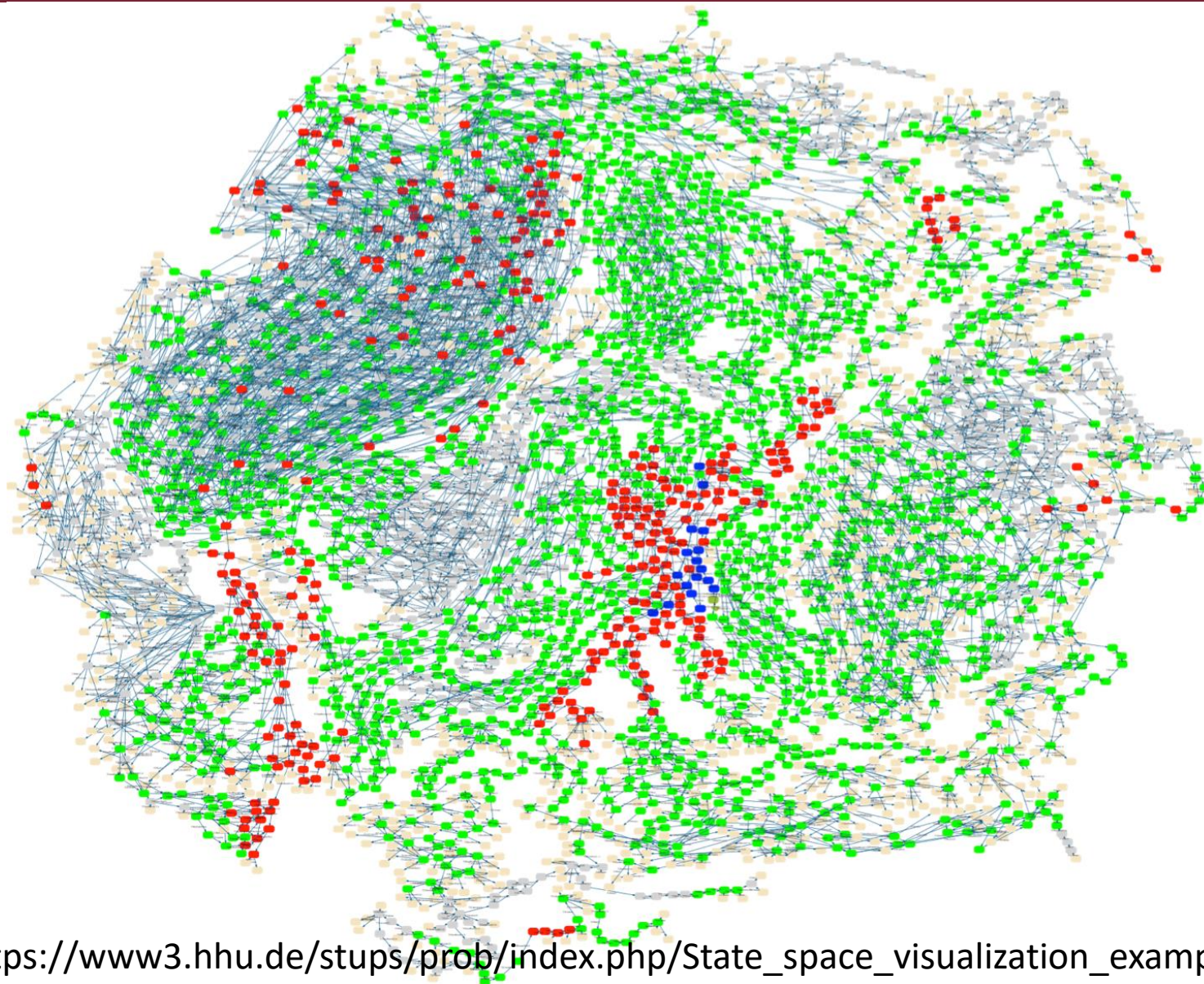
4. Interpretáció, korreláció más modellekkel, kiértékelés

Példa: állapottér vizualizáció (hálózat)



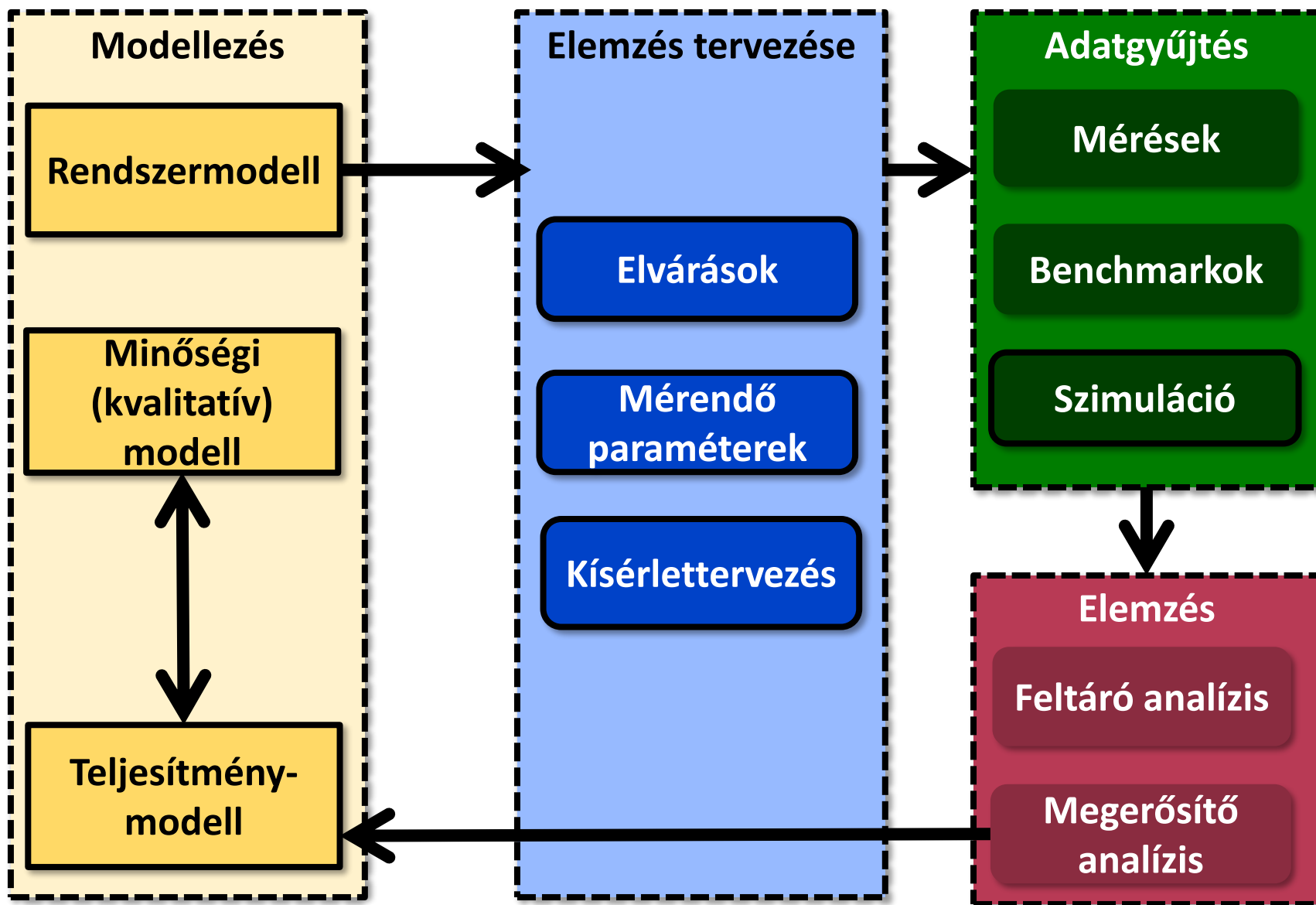
https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples

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

Reproducible
research

Reporting

Self-Service BI

Dashboard

EDA

Storytelling

Data tour

Data
journalism

Infographics

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

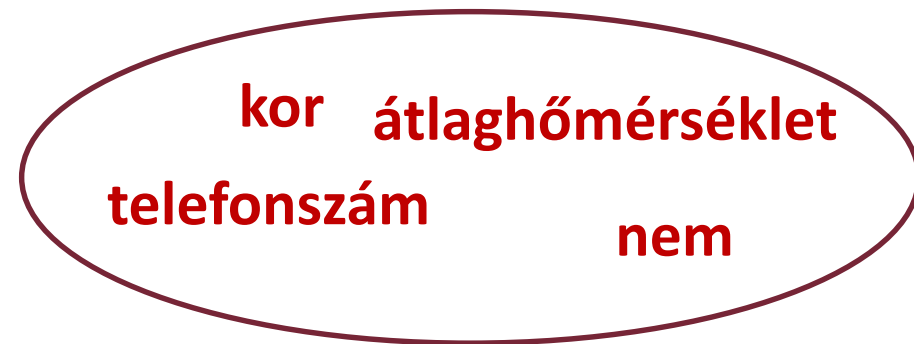
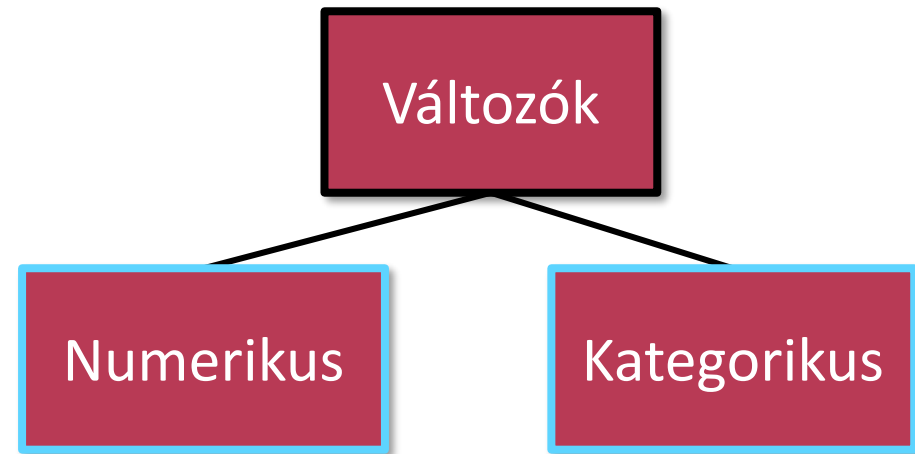
Numerikus és kategorikus változók

- Numerikus (numerical)

- az alapvető aritmetikai műveletek értelmesek

- Kategorikus (categorical)

- Matematikai műveletek nem értelmezhetőek rajtuk, legfeljebb sorba rendezés



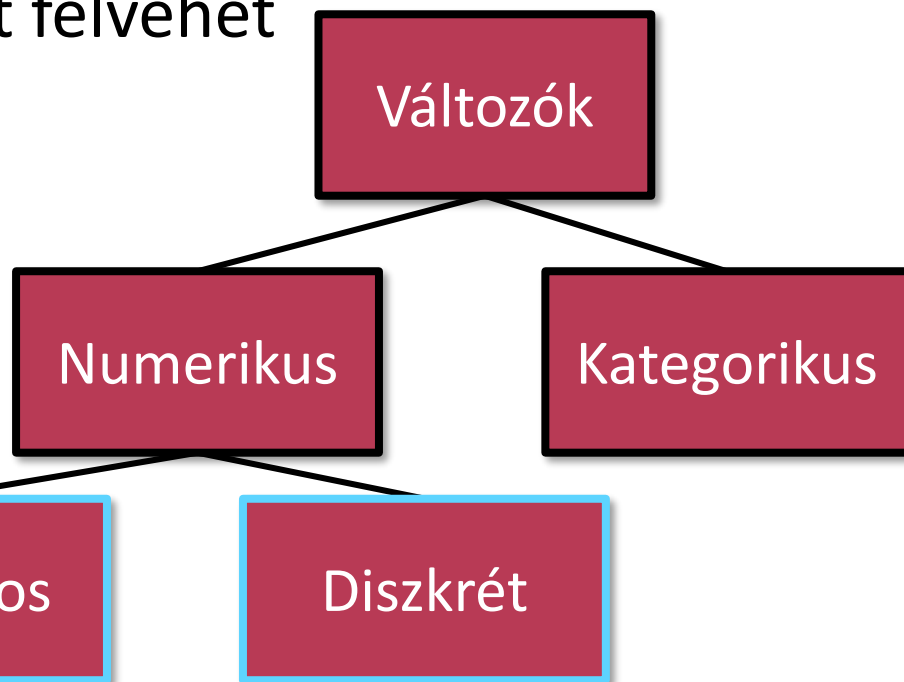
Numerikus változók

■ Folytonos

- Mért – tetszőleges értéket felvehet

- adott tartományon belül
- adott pontosság mellett

- Pl. a teremben ülők
ZH pontszámának átlaga



■ Diszkrét

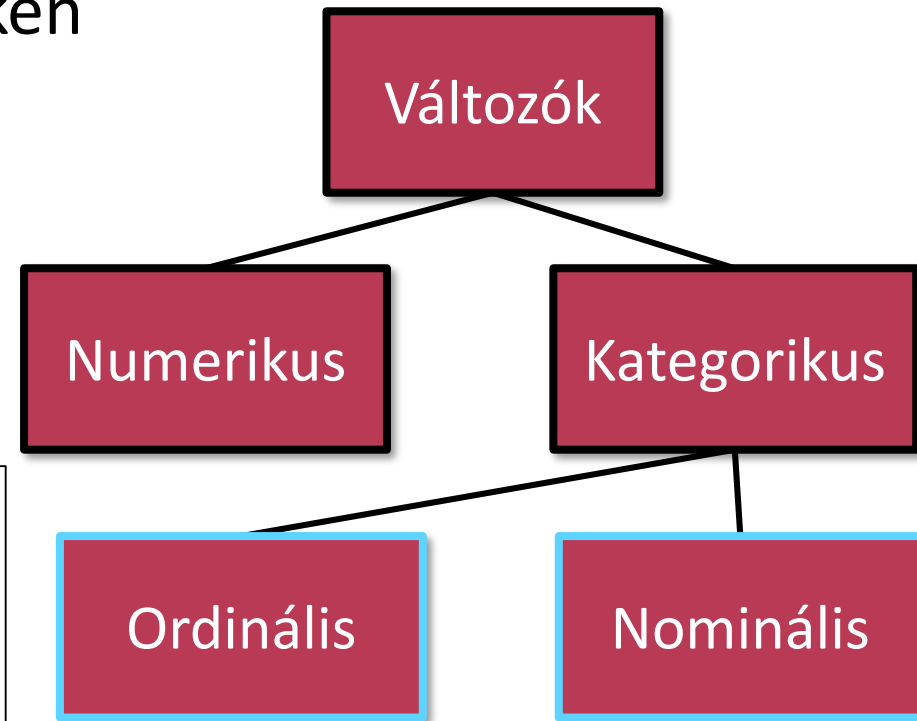
- Számolt – véges sok értéket vehet fel adott tartományban
- Pl. az előadáson ülők száma

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?

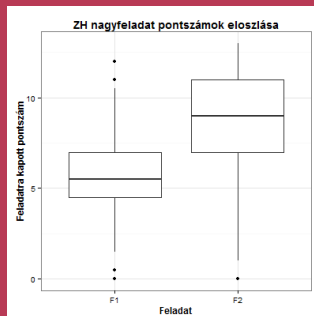
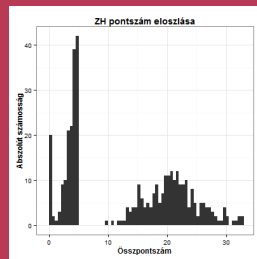
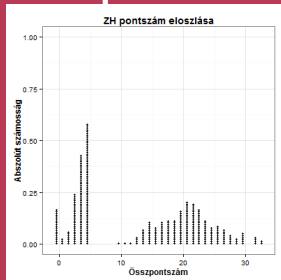
1 változó – eloszlásokra

Változók

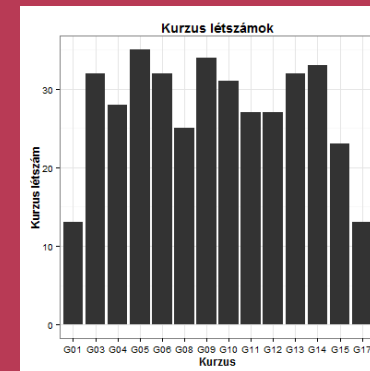
Numerikus

Kategorikus

ZH pontszám: [13, 15, 2, ...]

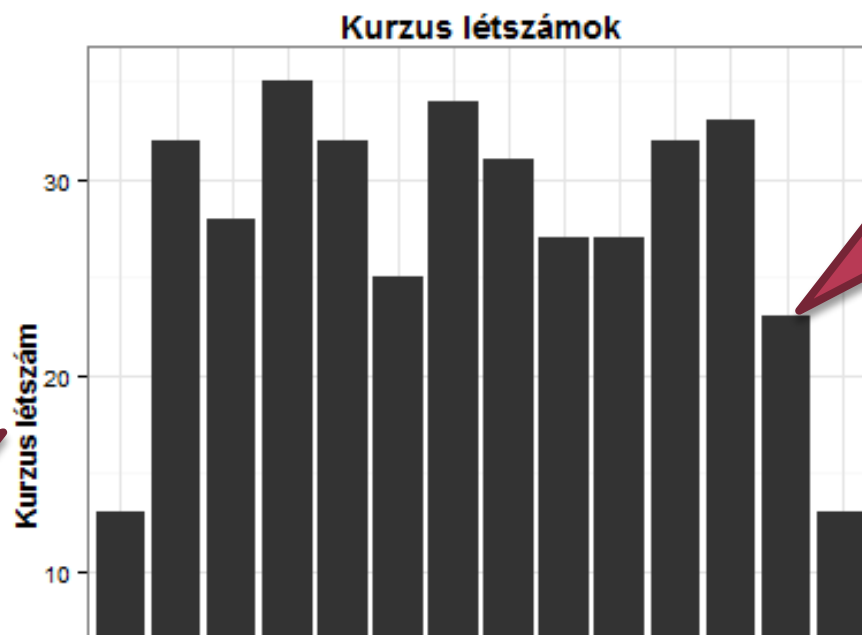


Kurzus: [G01, G03, G15, G17, ...]



Oszlopdiagram

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



abszolút
gyakoriság!

Oszlop-
magasság:
adott érték
gyakorisága

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

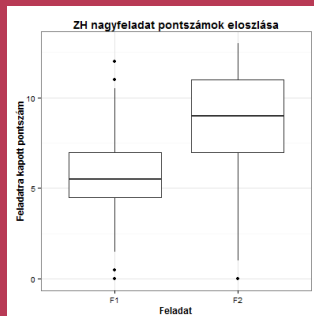
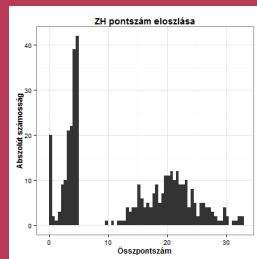
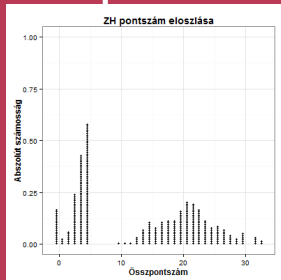
1 változó – eloszlásokra

Változók

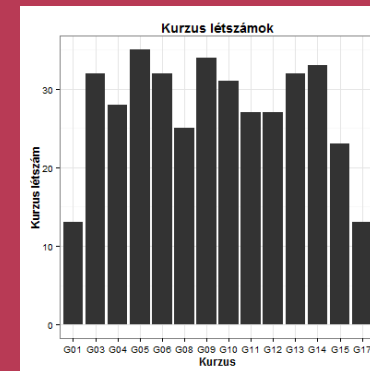
Numerikus

Kategorikus

ZH pontszám: [13, 15, 2, ...]



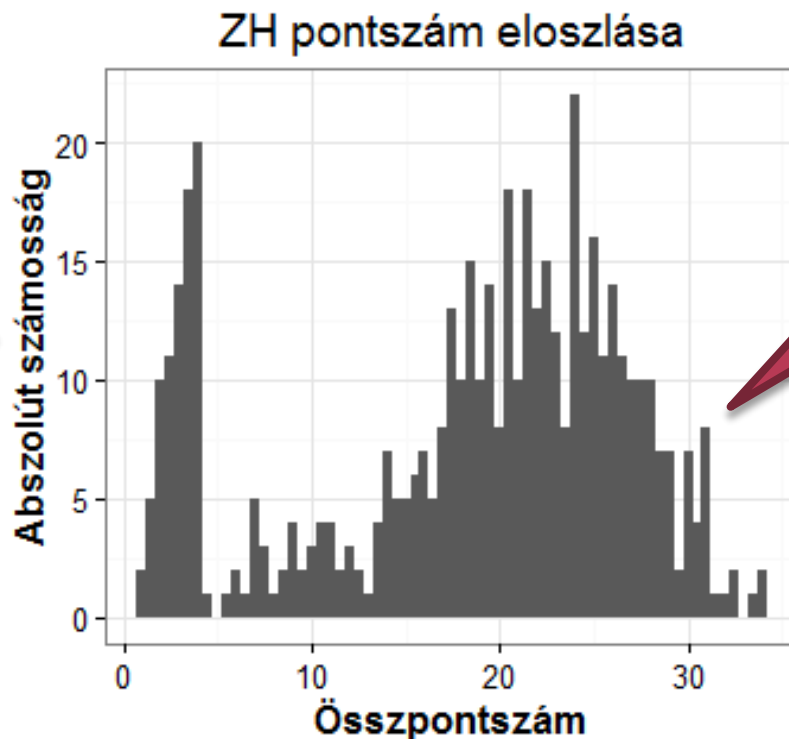
Kurzus: [G01, G03, G15, G17, ...]



Hisztogram

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

abszolút
gyakoriság!

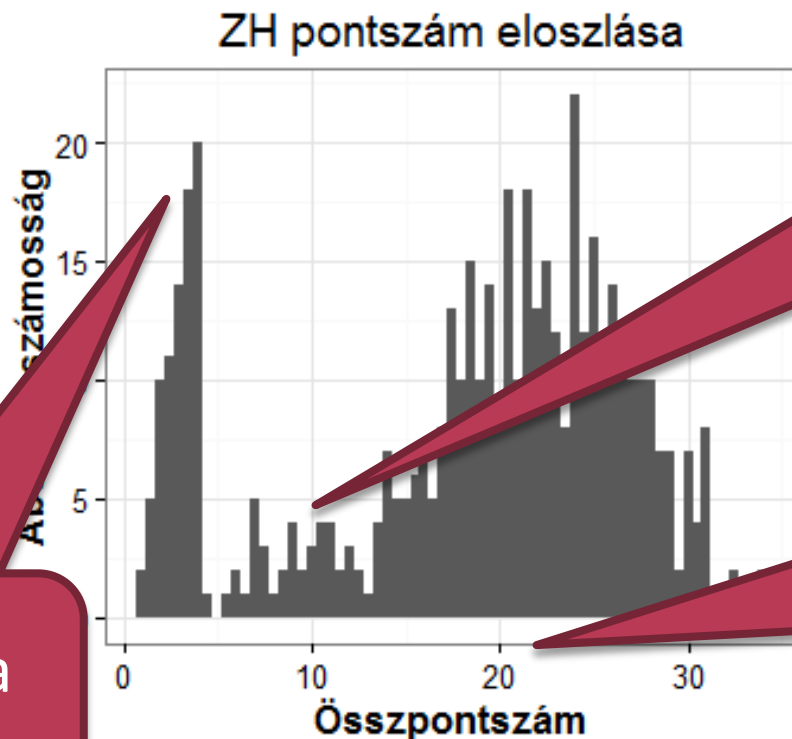


Oszlop-
magasság:
adott
intervallum
számossága

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?



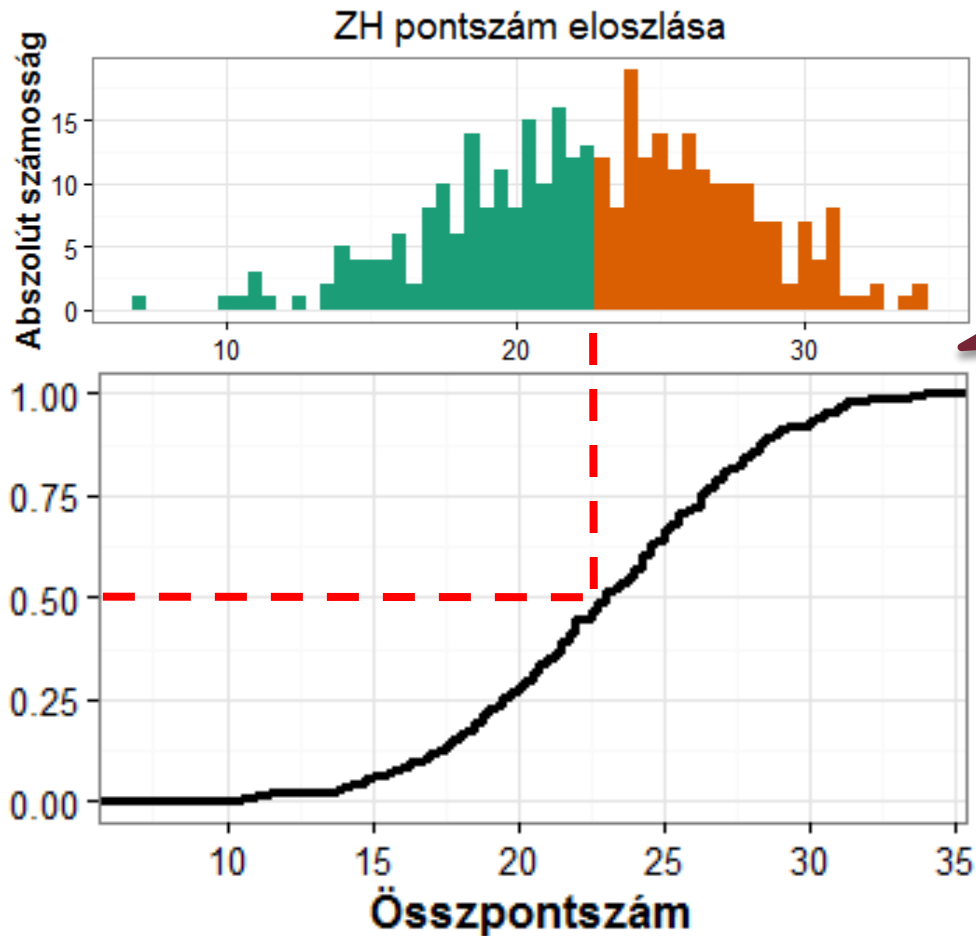
Sokan voltak a határon

Akik átmentek a beugrón, valószínűleg át is mentek

18 pont körül volt az átlag, 20 körül a medián

Egyszerű statisztikai jellemzés

- Hol van az adatok „közepe”?

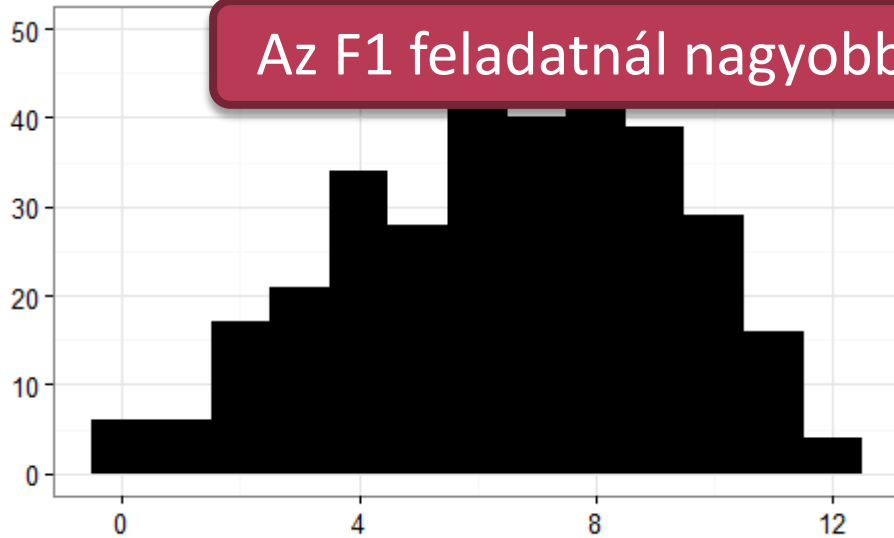


Az átmentek
összpontszám
mediánja 23

Egyszerű statisztikai jellemzés

■ Mennyire „szórtak” az adatok?

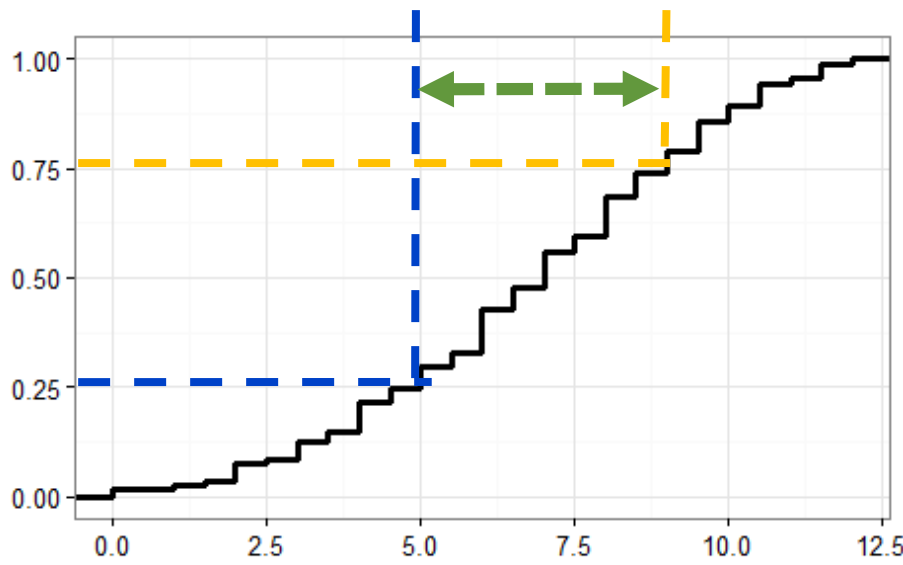
Az F1 feladatnál nagyobb volt a szórás, mint az F2-nél



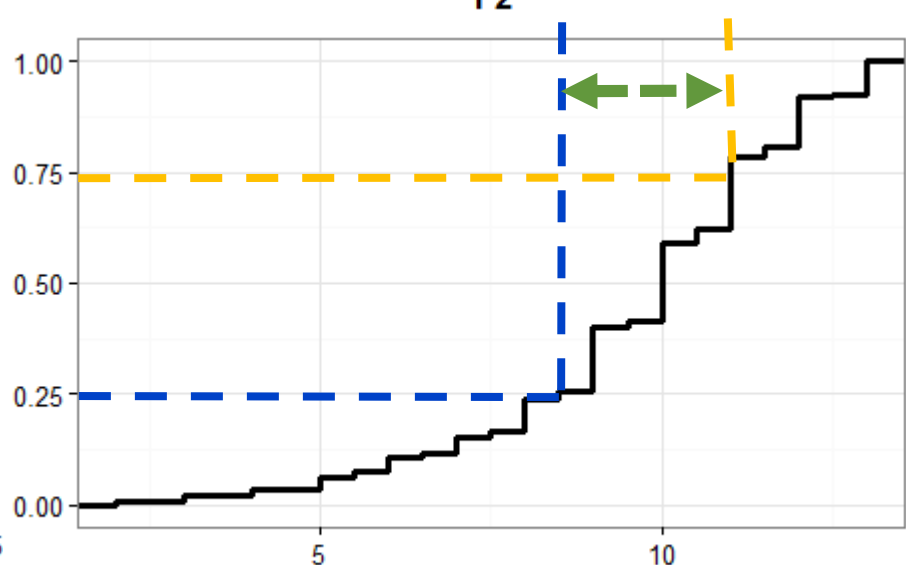
F1



F2



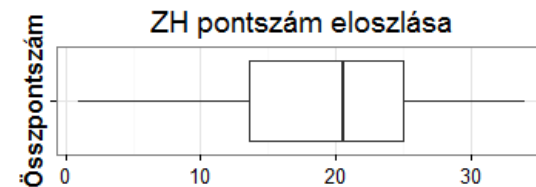
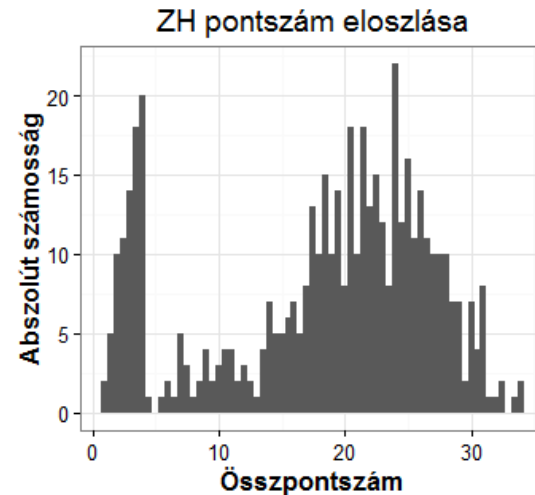
F1



Boxplot

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

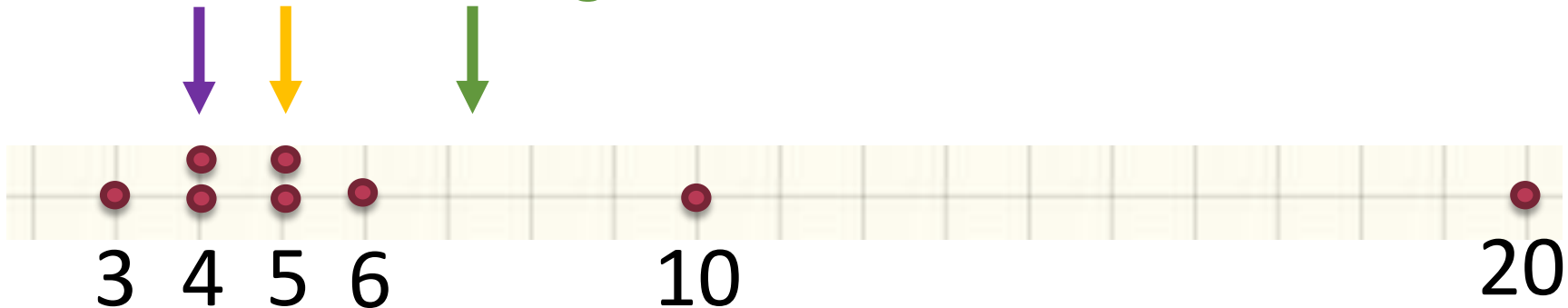
Egyfajta absztrakció itt is:
legyenek intervallumok,
felesleges minden pontot
kirajzolni



(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

módusz medián átlag



(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 módusról beszélünk.

Terjedelem jellemzése: percentilisek

Az n -edik **percentil**snél az értékek $n\%$ -a kisebb.

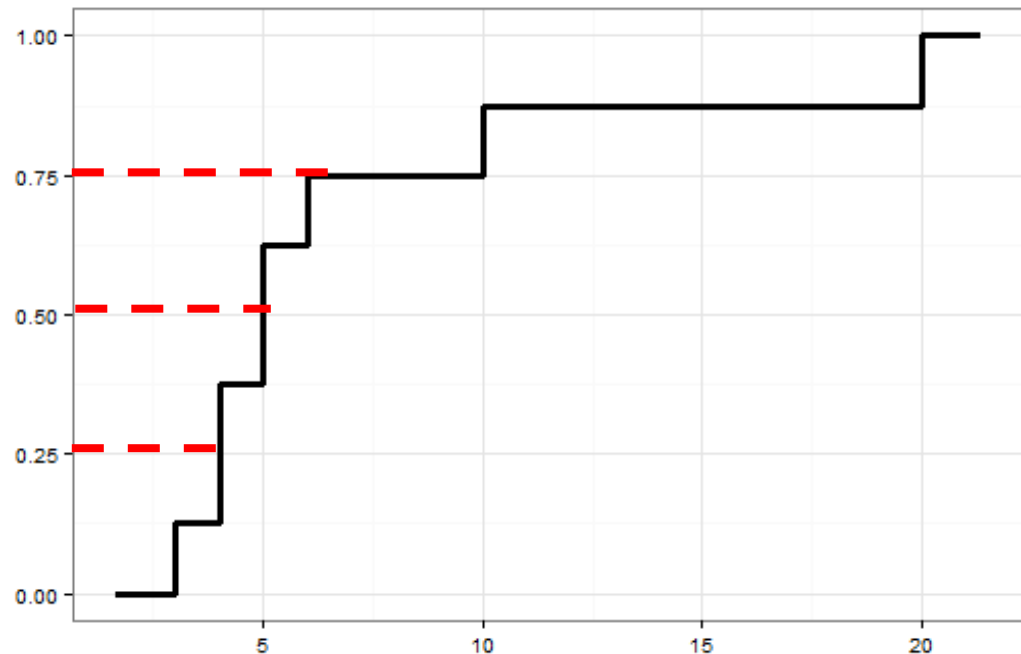
■ Percentilis

○ {3, 4, 4, 5, 5, 6, 10, 20}

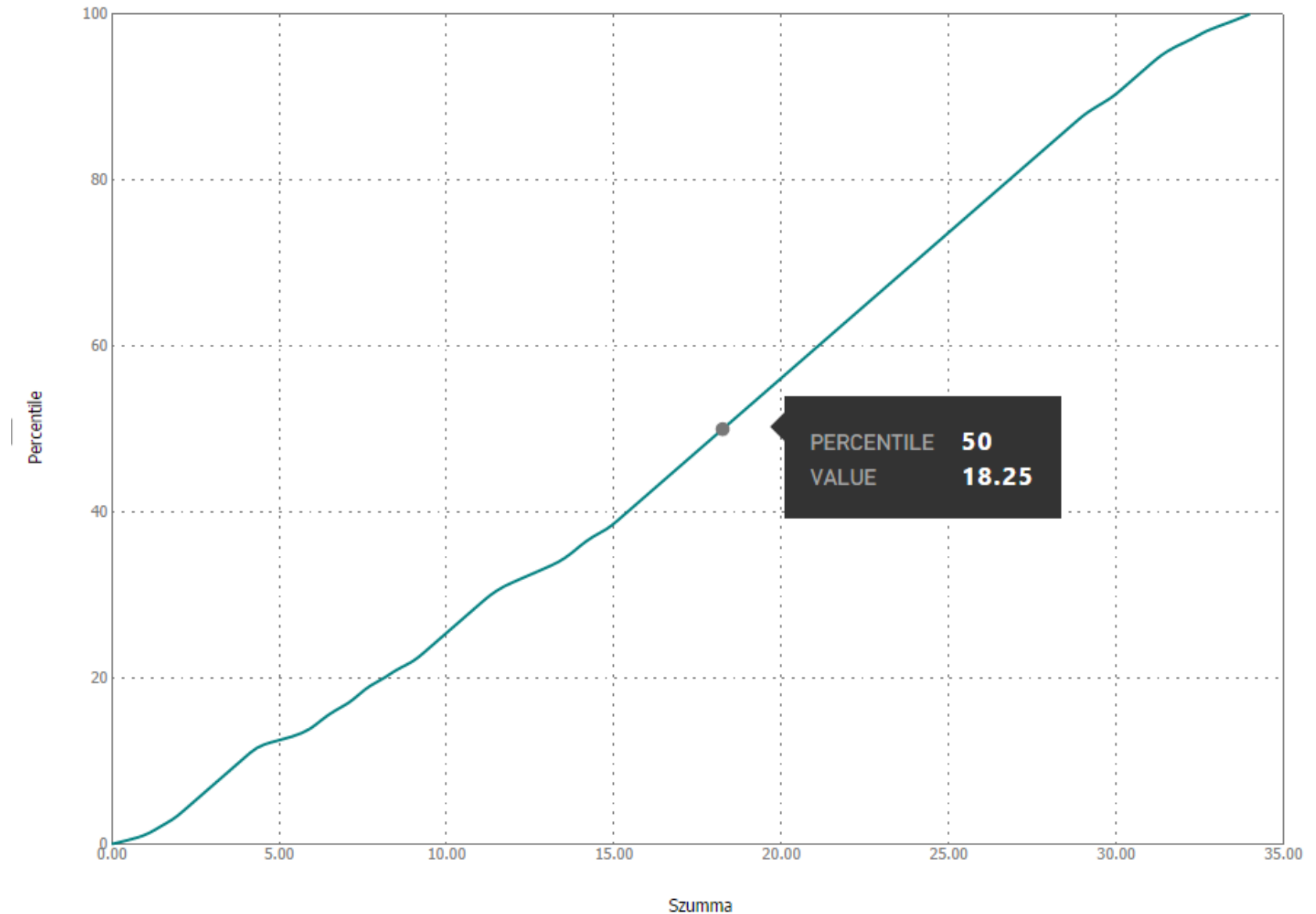
- 50. percentilis: 5
- 25. percentilis: 4
- 75. percentilis: 6

■ Kvartilis

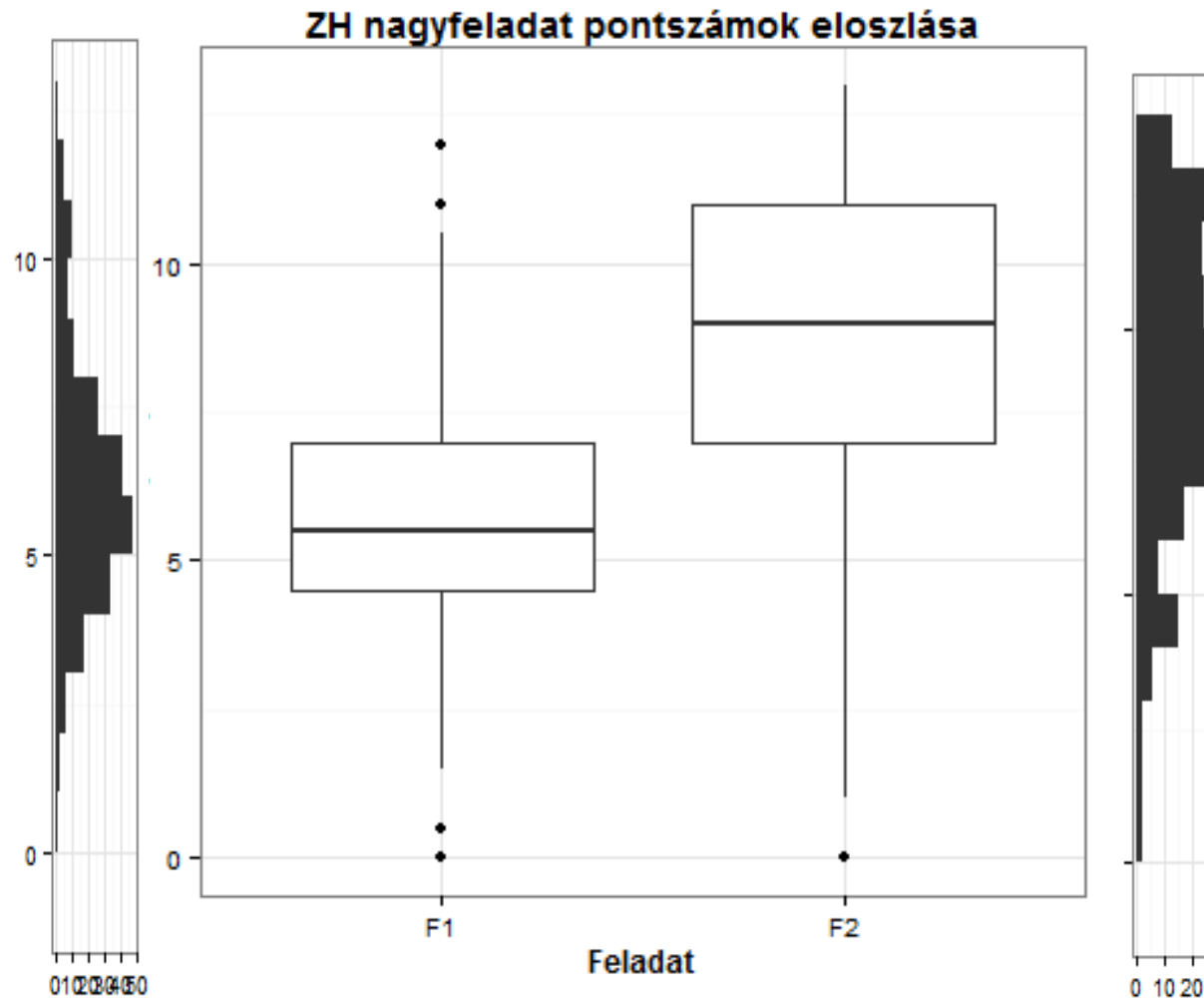
- Q1: 25. percentilis
- Q3: 75. percentilis
- **Q2: medián**



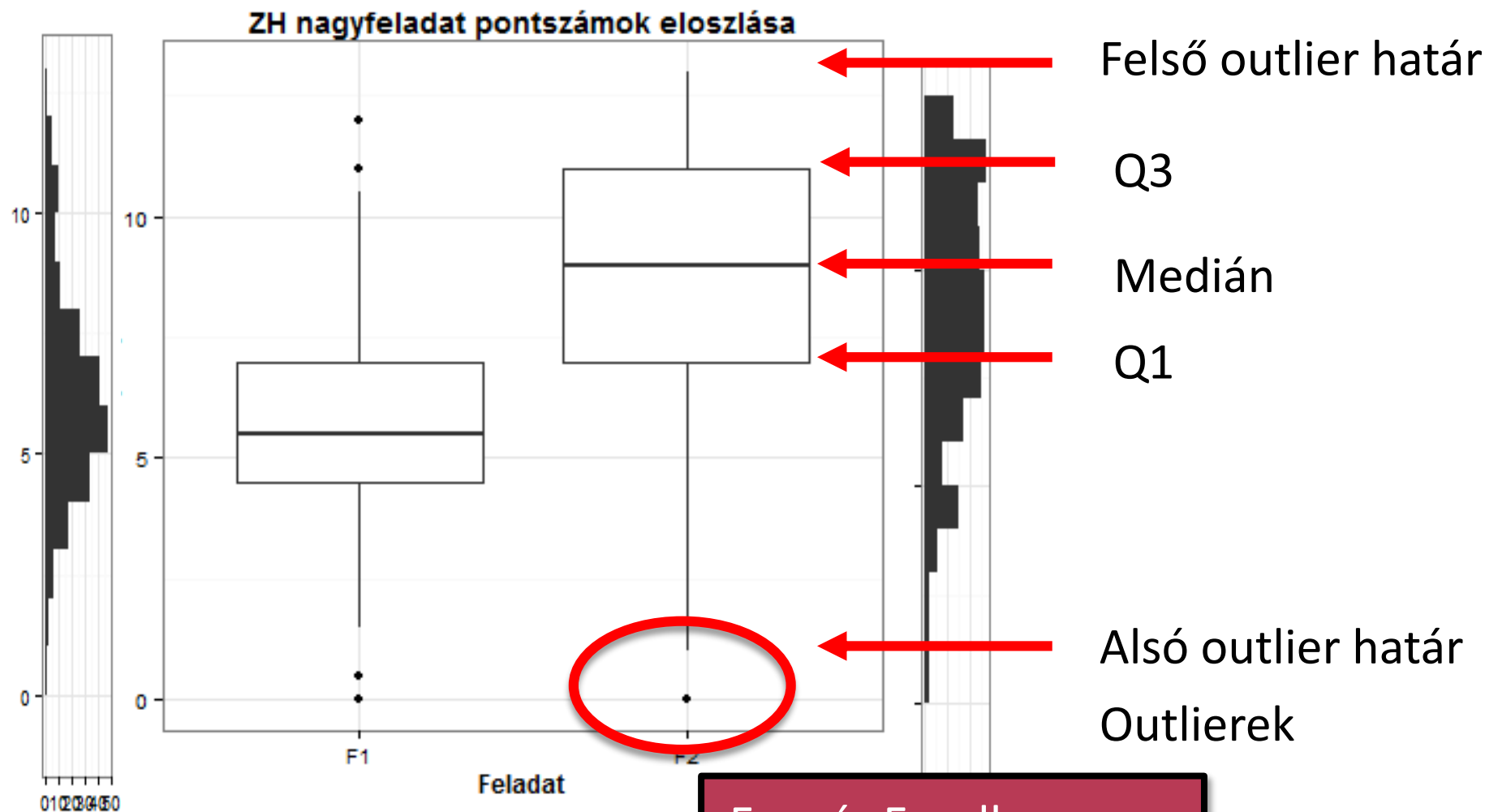
Példa: percentilis ábrázolás



Boxplot (Box and whisker plot)

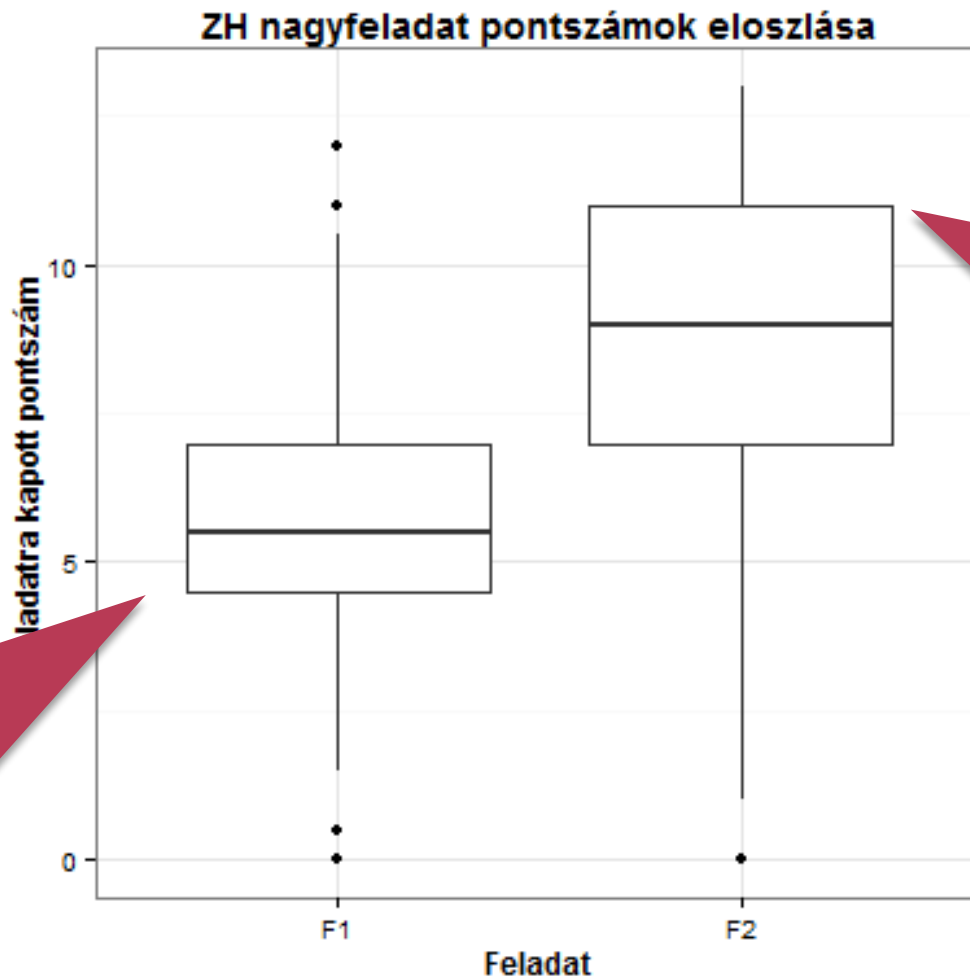


Boxplot (Box and whisker plot)



Ez már Excelben nem
könnyű...

Boxplot (Box and whisker plot)

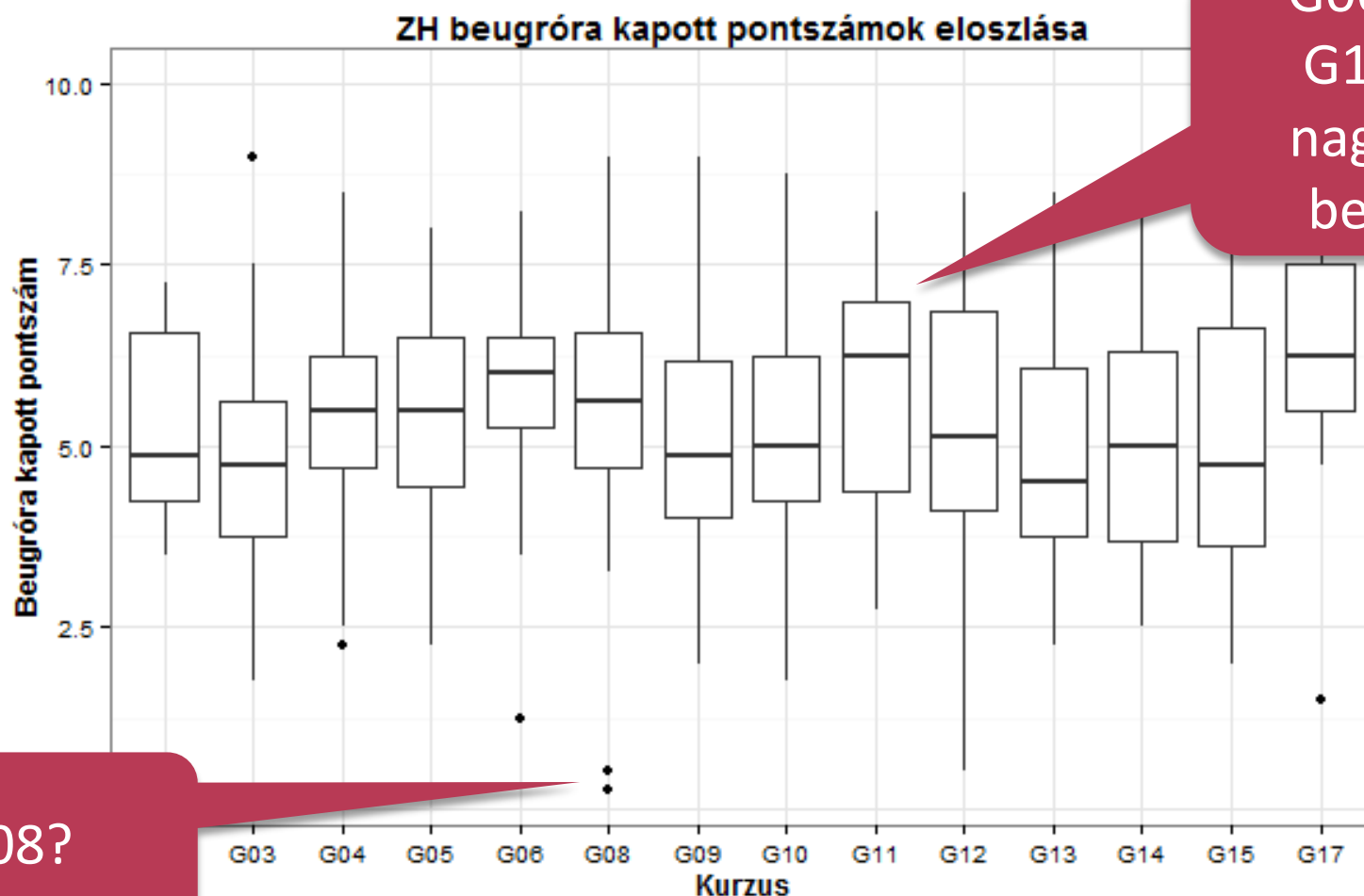


Az F1
pontszámok
50%-a
4.5 és 7.5
között volt

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

Boxplot (Box and whisker plot)

- Melyik csoportban hogyan sikerültek a beugrók?

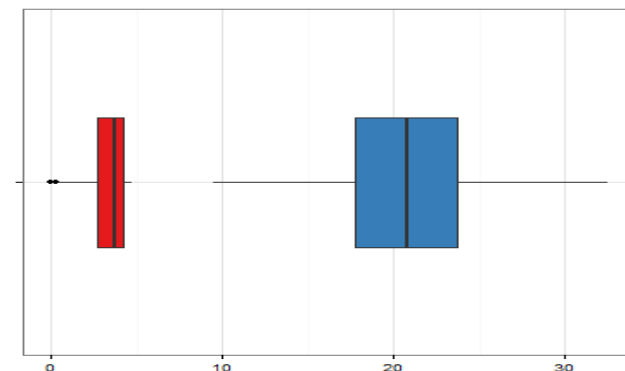
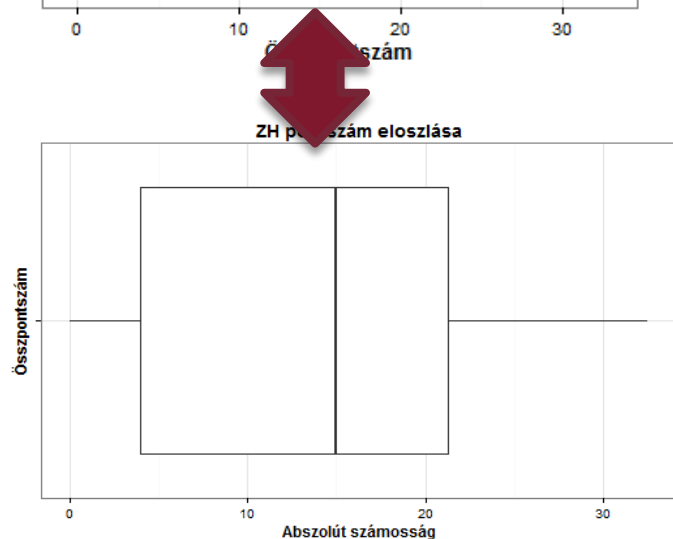
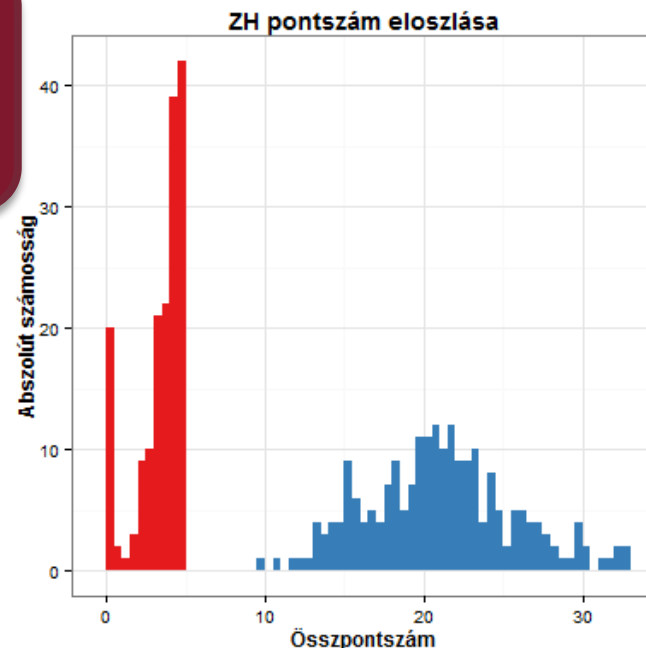
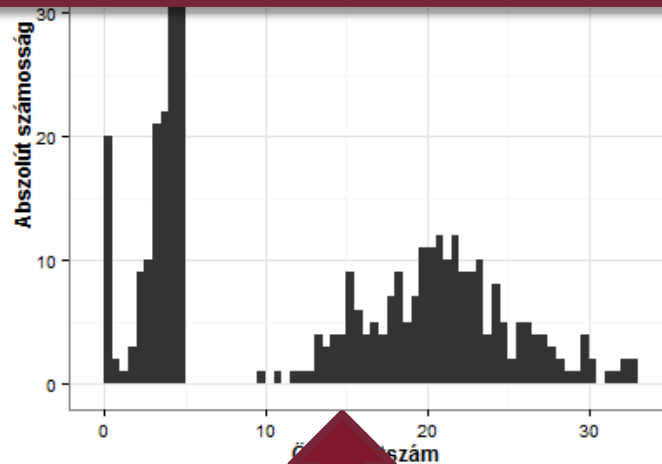


G06, G11,
G17-ben
nagyon jó
beugrók

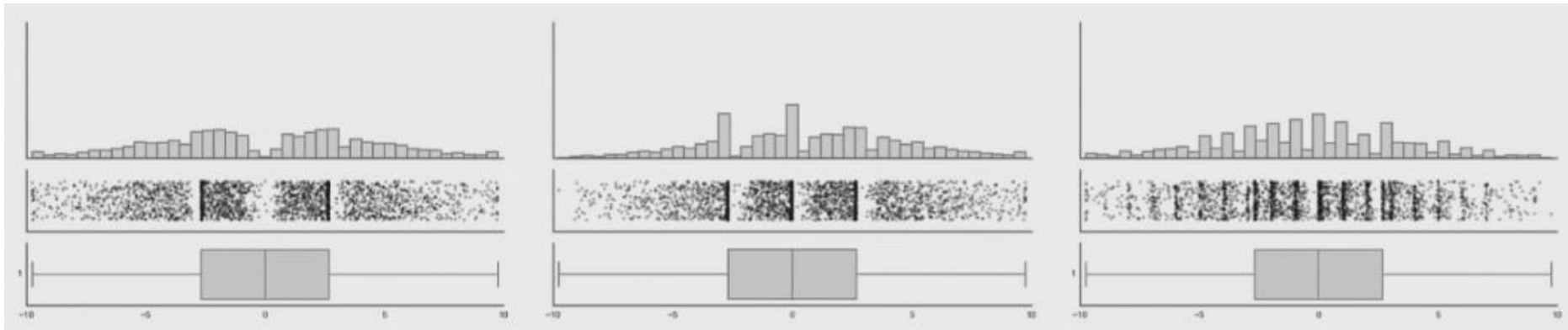
G08?

Boxplot (Box and whisker plot)

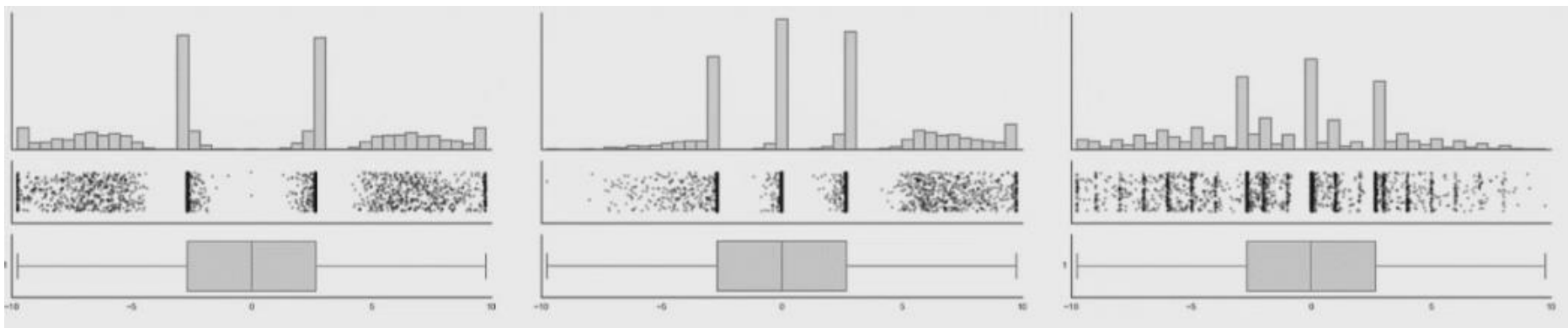
Absztrakció: a boxplottal fontos információt is veszíthetünk!



Eloszlás vs. összegzés

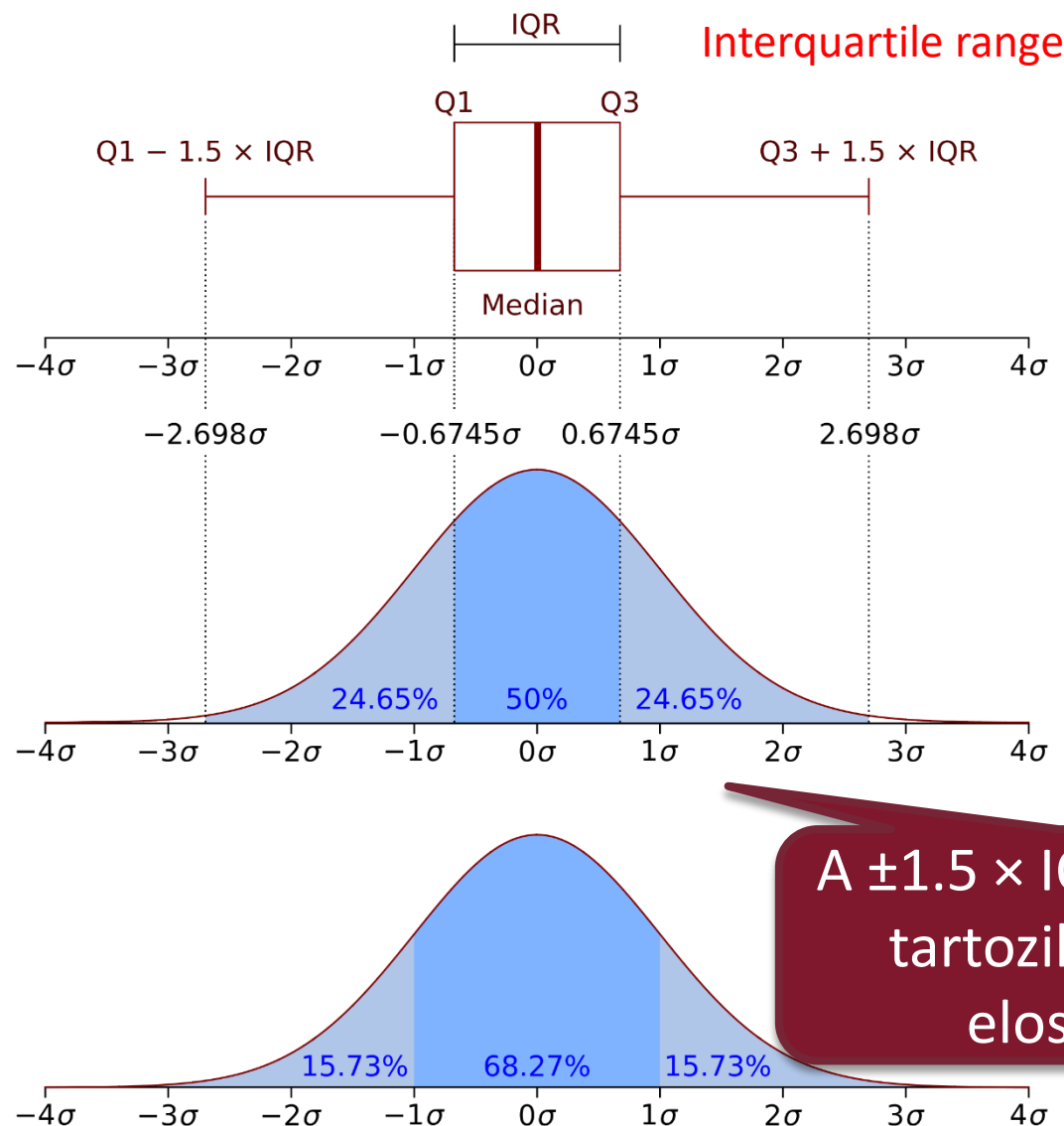


Ugyanaz az „összegzés” != ugyanaz az eloszlás

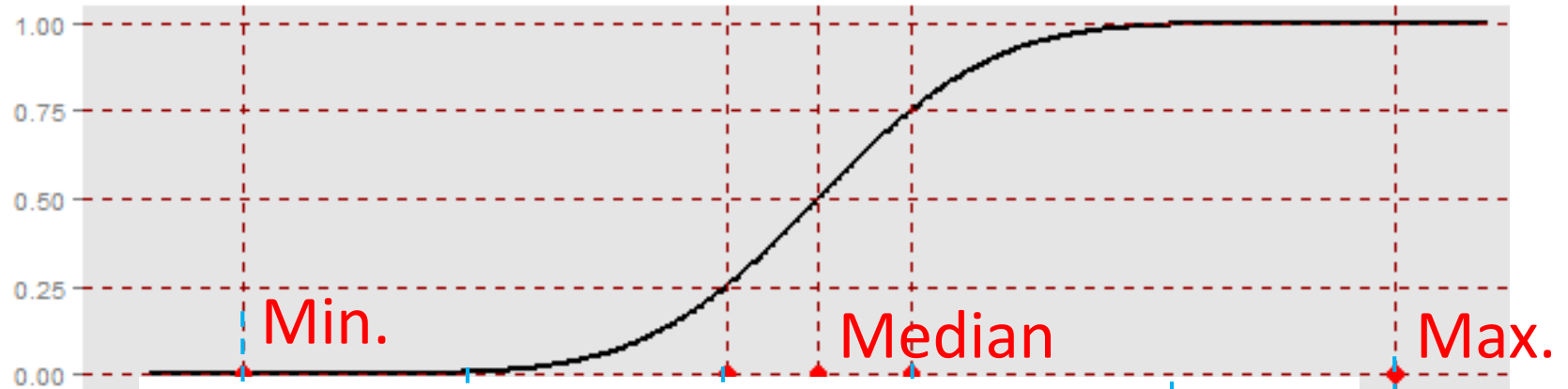


Matejka, Fitzmaurice. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. ACM SIGCHI Conference on Human Factors in Computing Systems
<https://www.autodeskresearch.com/publications/samestats>

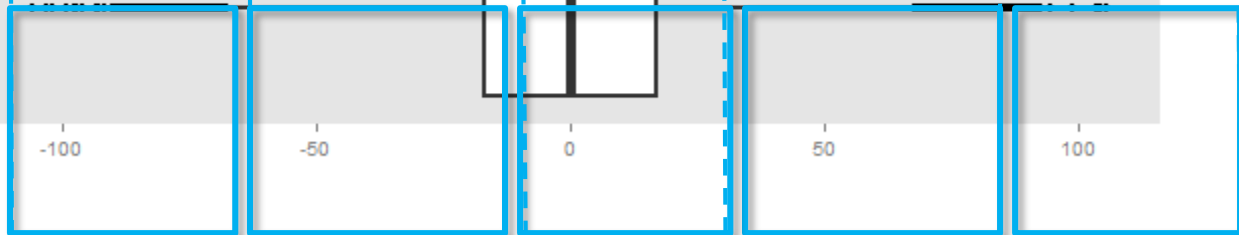
Boxplot (Box and whisker plot)



Boxplot: kvalitatív jellemzés



Extr. small Small Normal Large Extr. large



$Q1 - 1.5 \text{ IQR}$

Normal

Small

Large

Extr. small

Extr. large

-100 -50 0 50 100

Kvalitatív
tartomány

Kvantitatív
tartomány

Miért medián, miért nem átlag?

■ Alaphalmaz

○ 1000 adatpont $\sim U(1, 5)$ egyenletes eloszlás

- *átlag = medián = 3 ms*



3ms $\pm$ 2 ms



Válaszidő

Új medián: `sort(resp. times)[501] = 3.02 ms`

Vál. medián

→ ✓ Robusztus

Vál. átlag

→ ✗ Nem rob.

Új átlag: $(2 * 10^4 + 3 * 10^3) / 1001 = 25 \text{ ms!}$

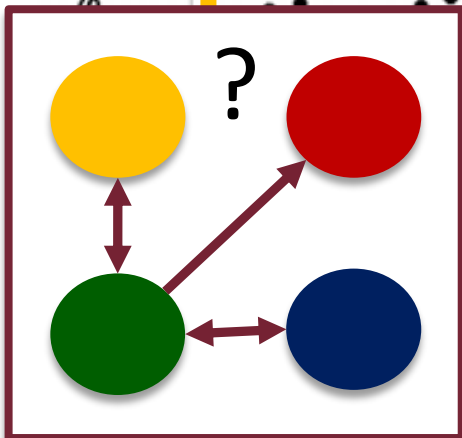
Példa: terhelés vs. kihasználtság

Kiugró értékek/
Háttérműveletek?

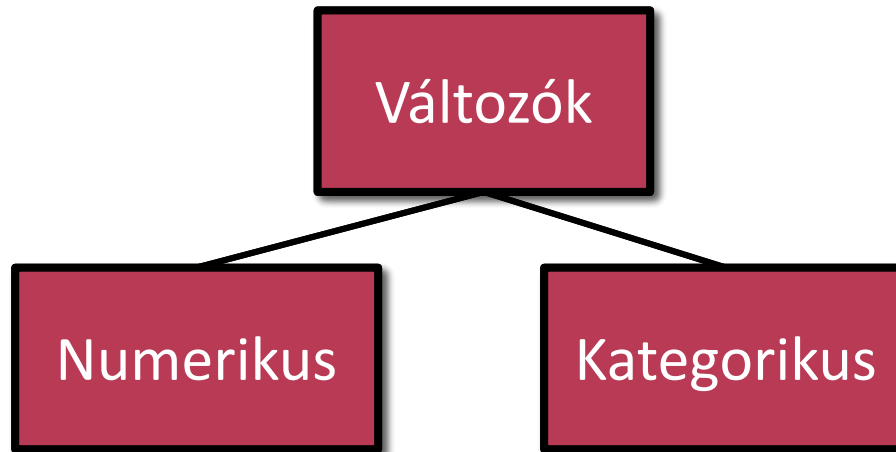
Eltérések?

Lineáris
kapcsolat

Nagy terhelésnél
nem jósolható
működés



2 változó kapcsolata

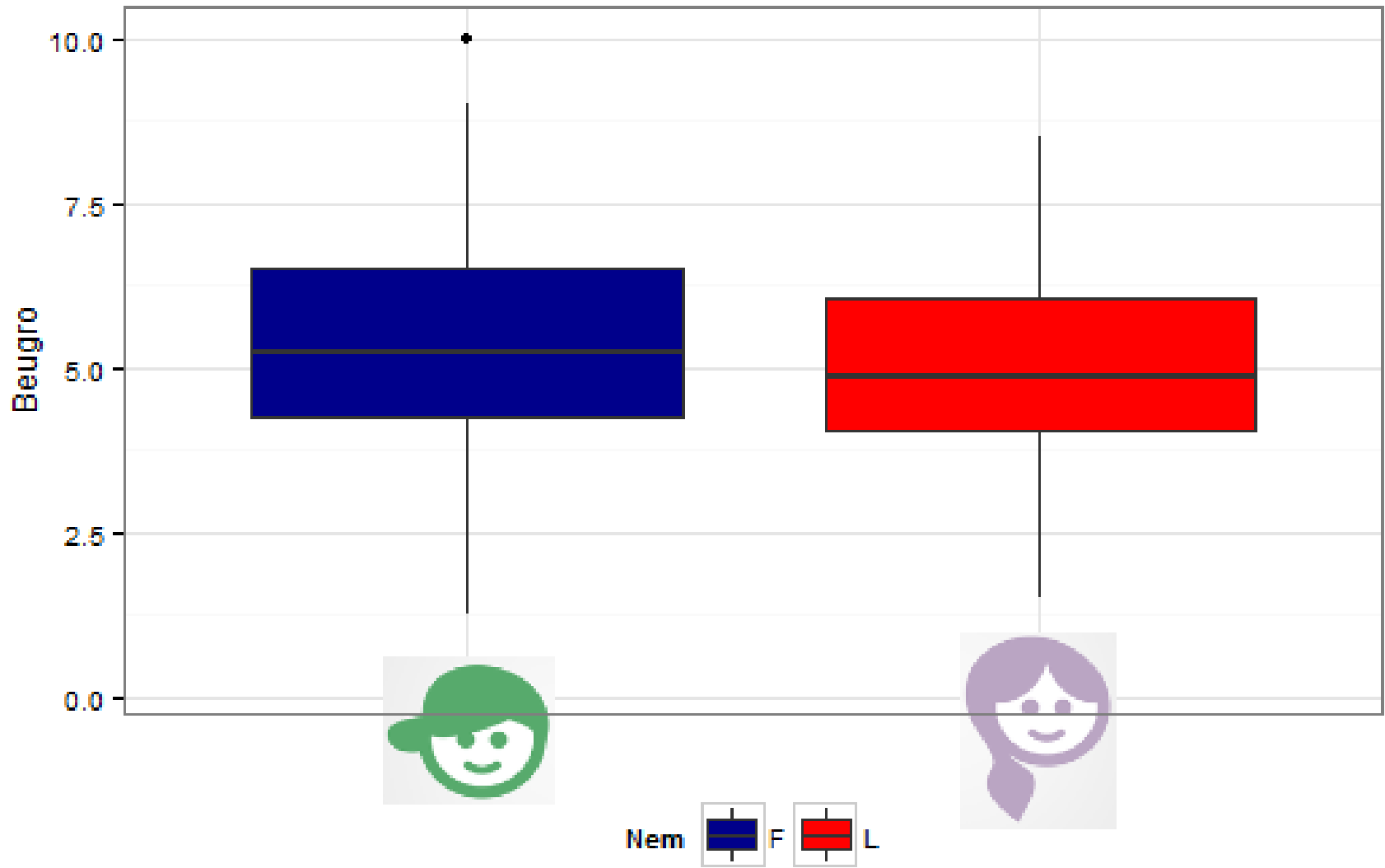


2 numerikus

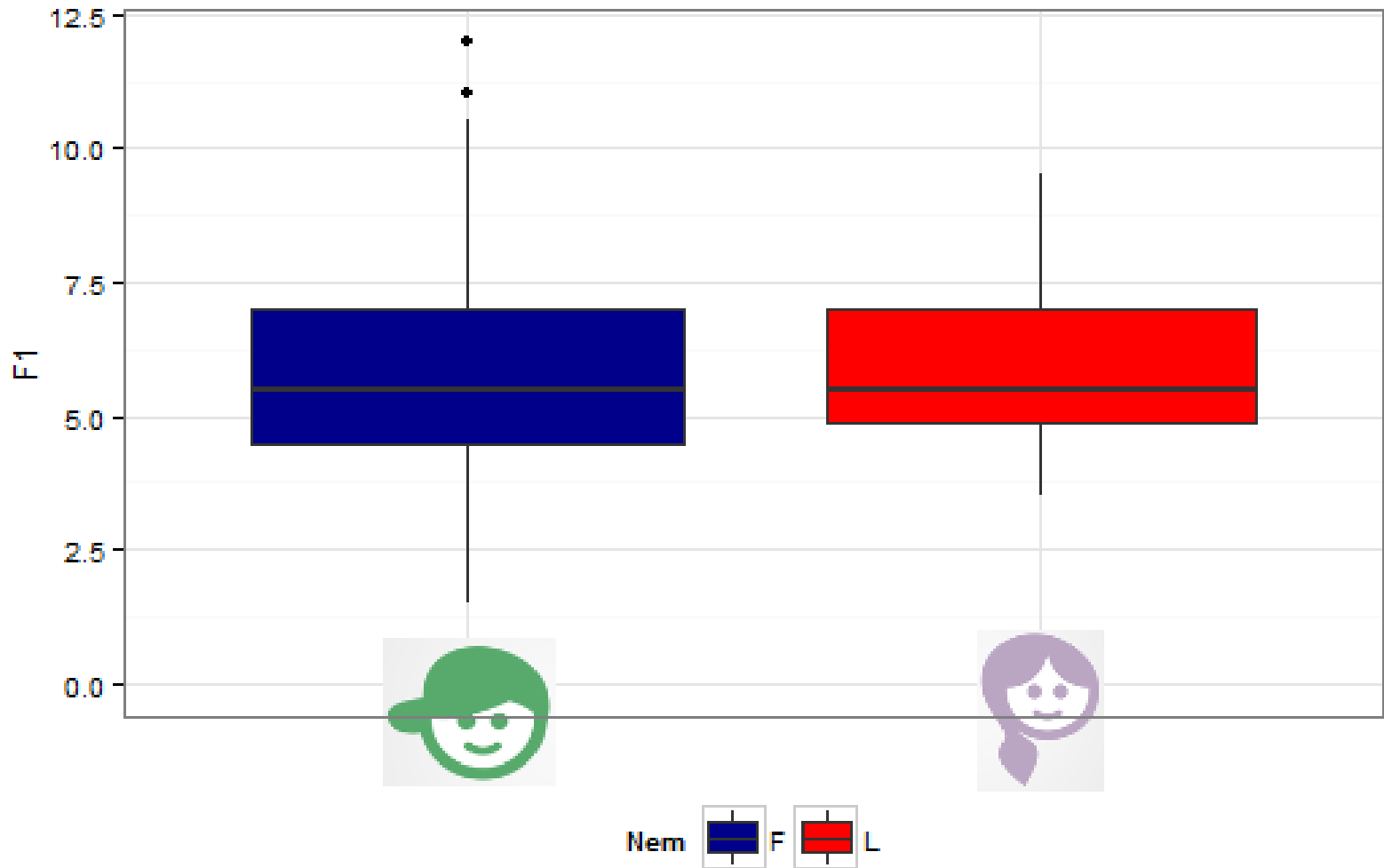
1 numerikus,
1 kategorikus

2 kategorikus

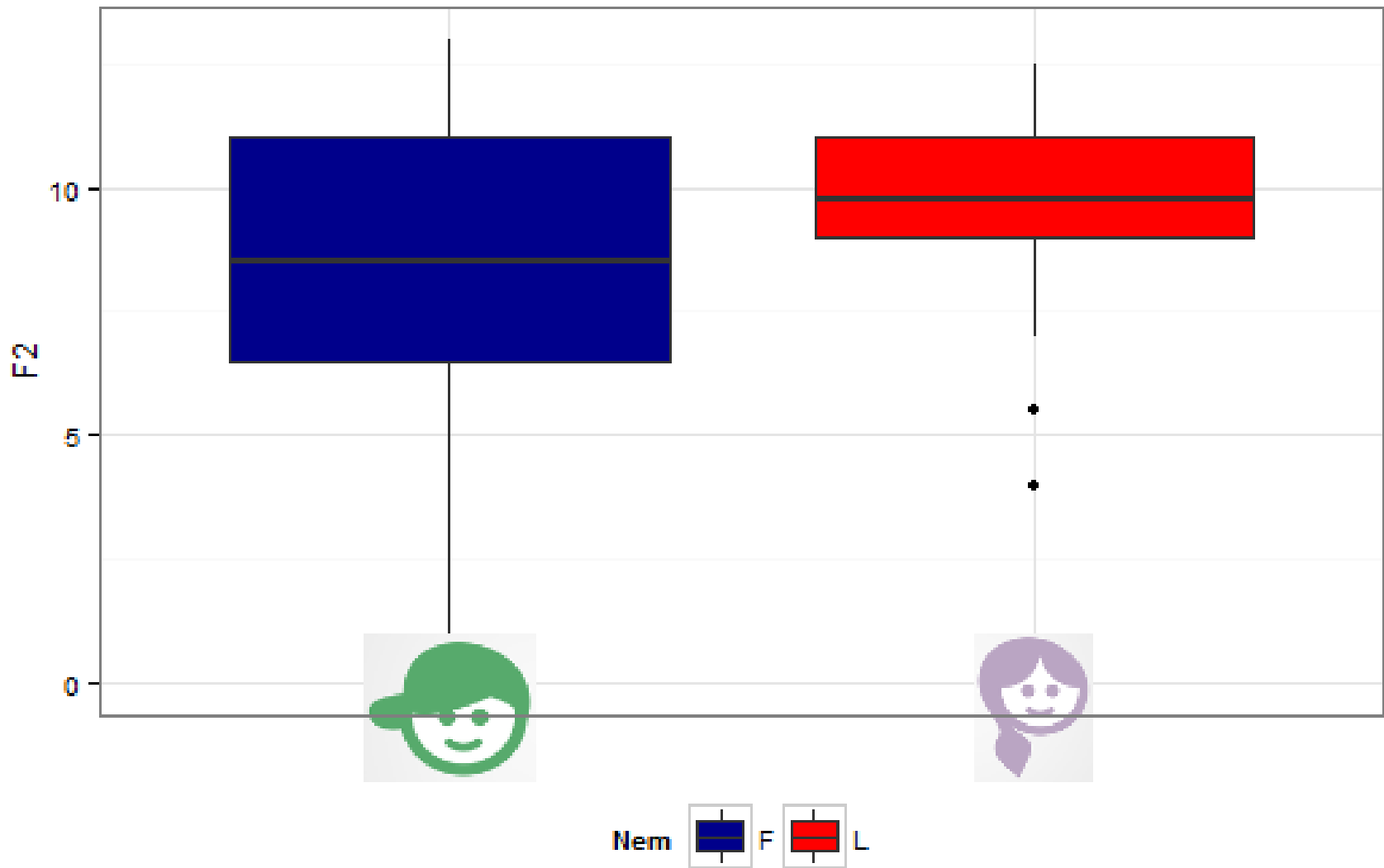
Numerikus kategóriánként



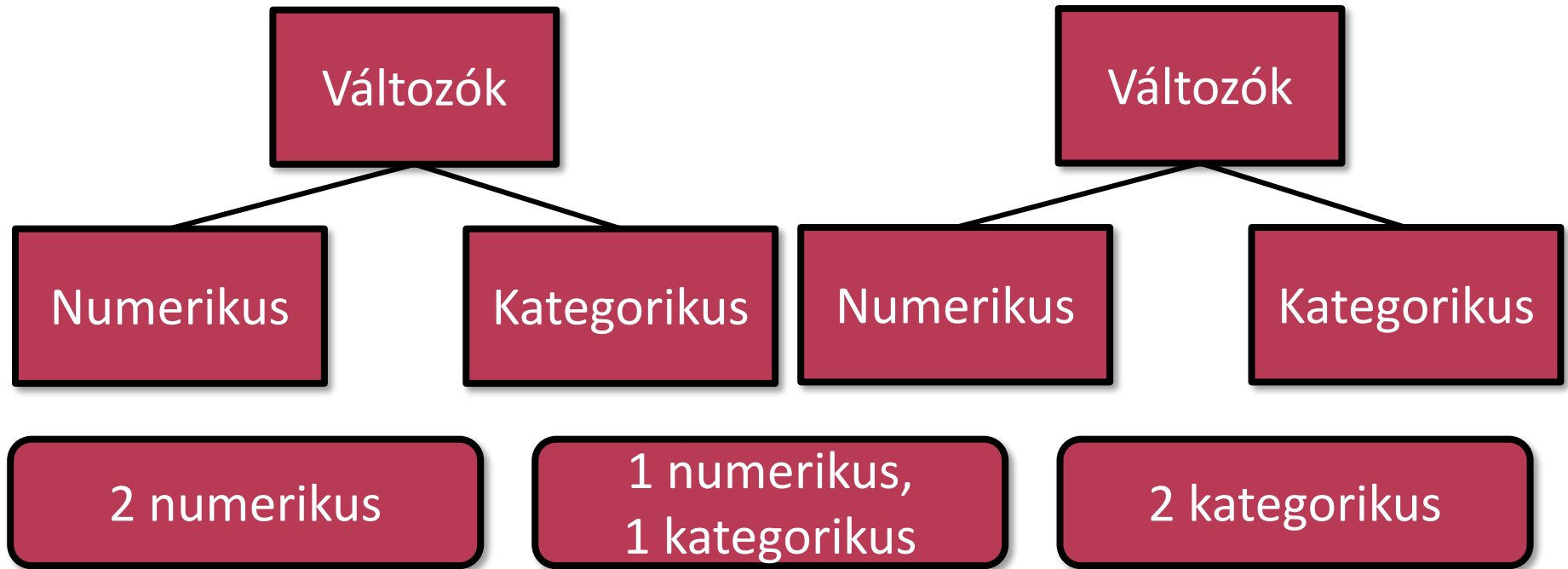
Numerikus kategóriánként



Numerikus kategóriánként



2 változó kapcsolata

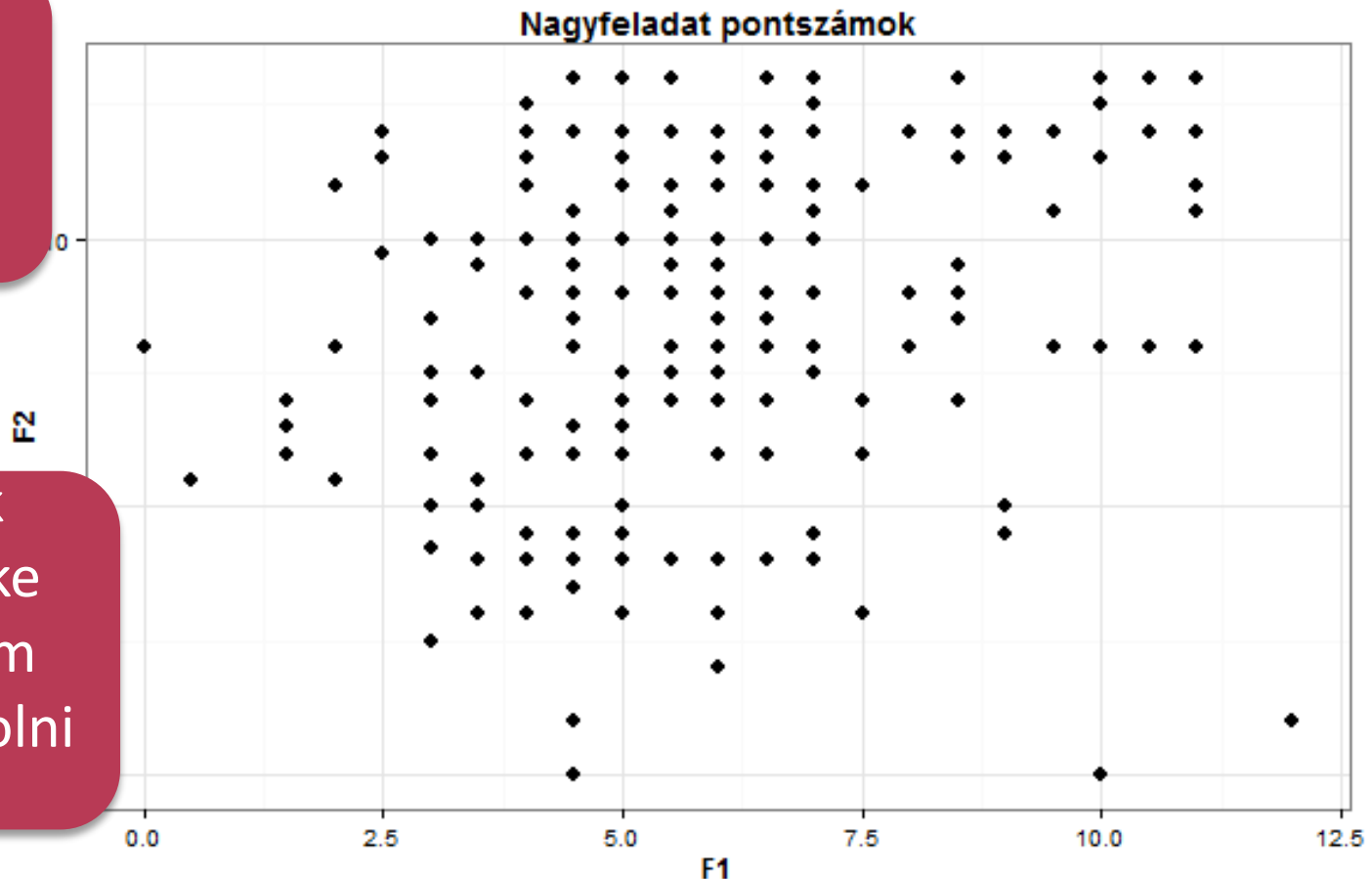


Pont – pont diagram (scatterplot)

- 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



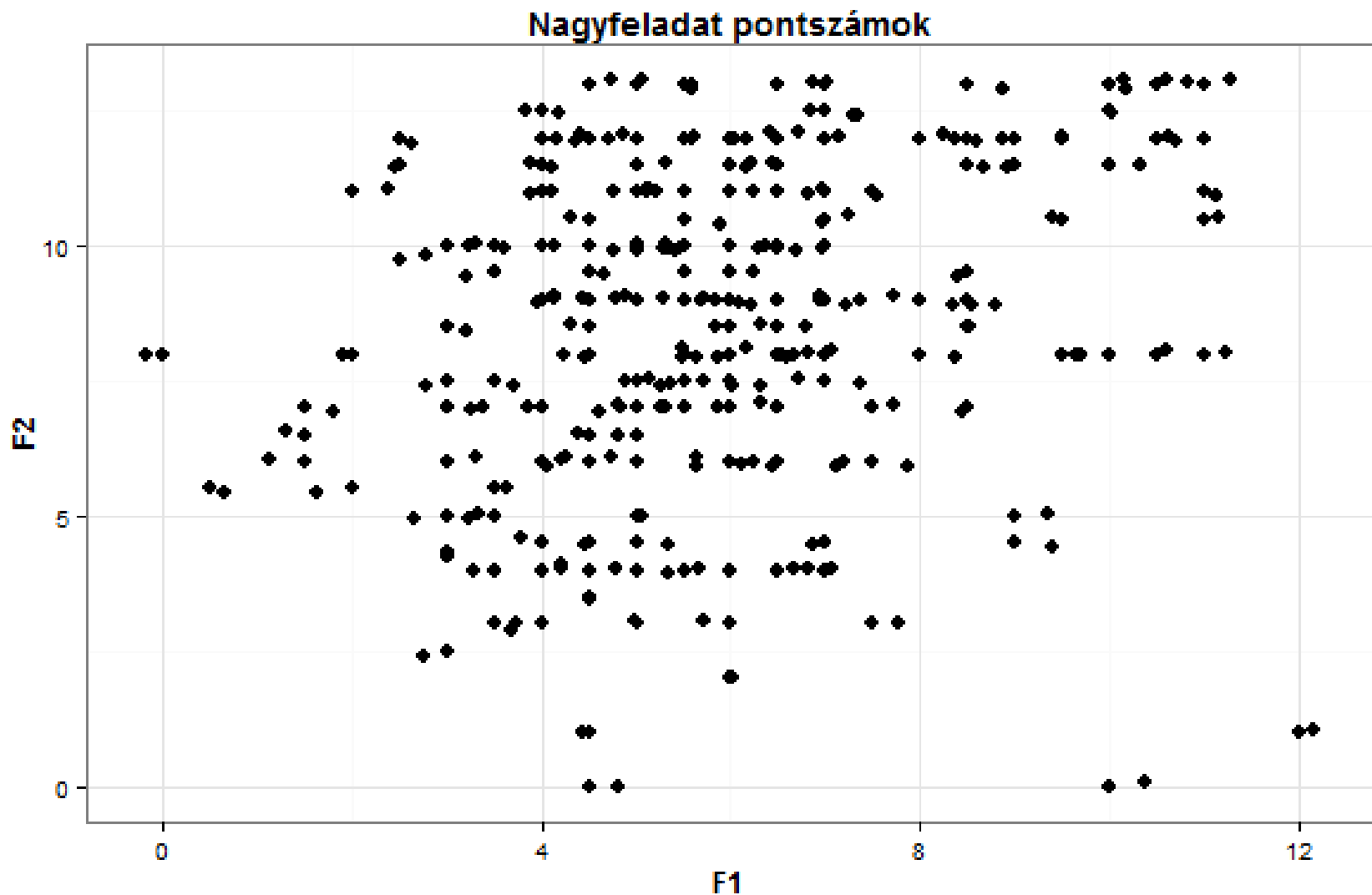
Hogyan kezeljük a takarásokat?

Overplotting

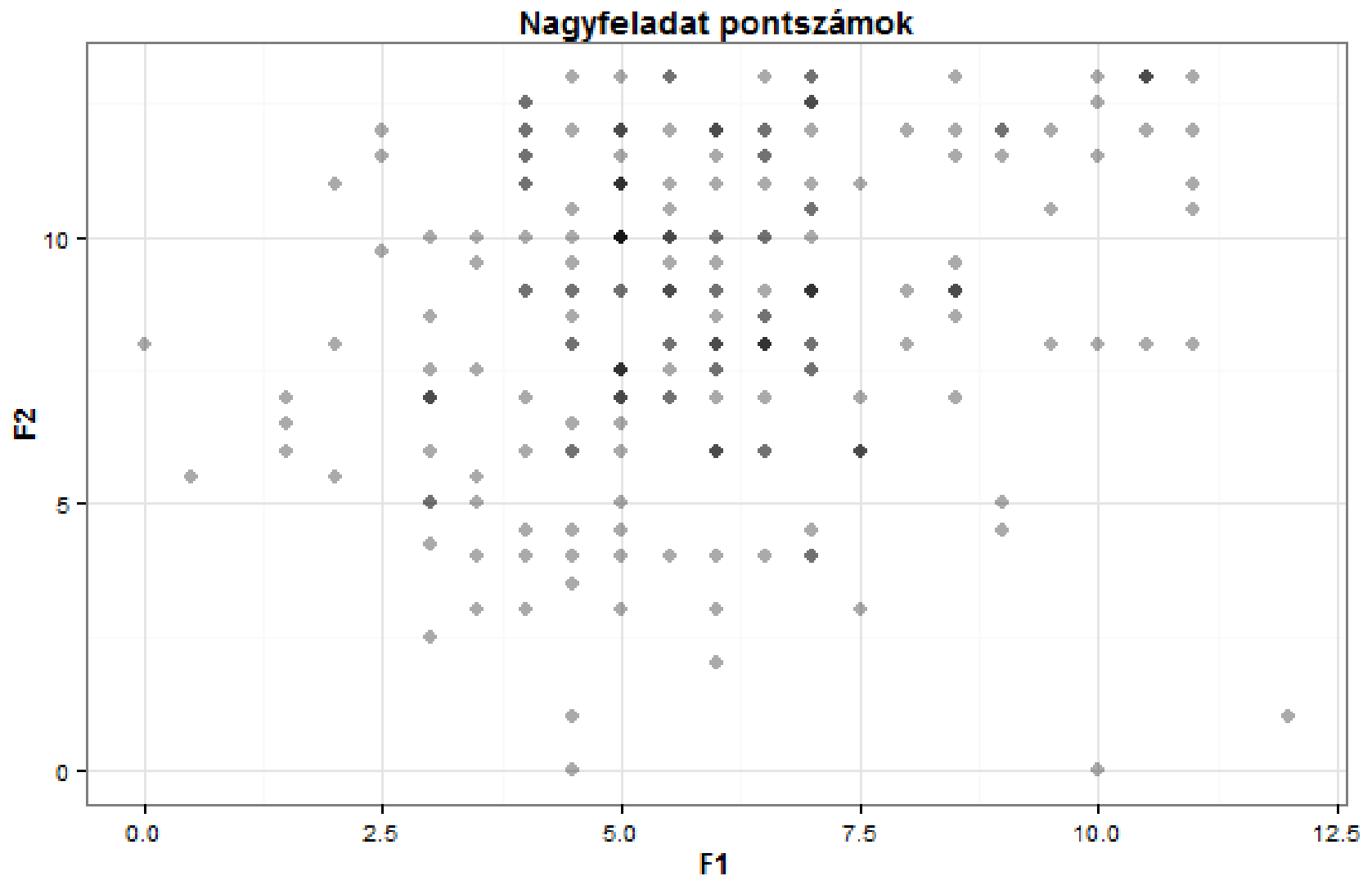


<http://www.holigyunkma.hu/>

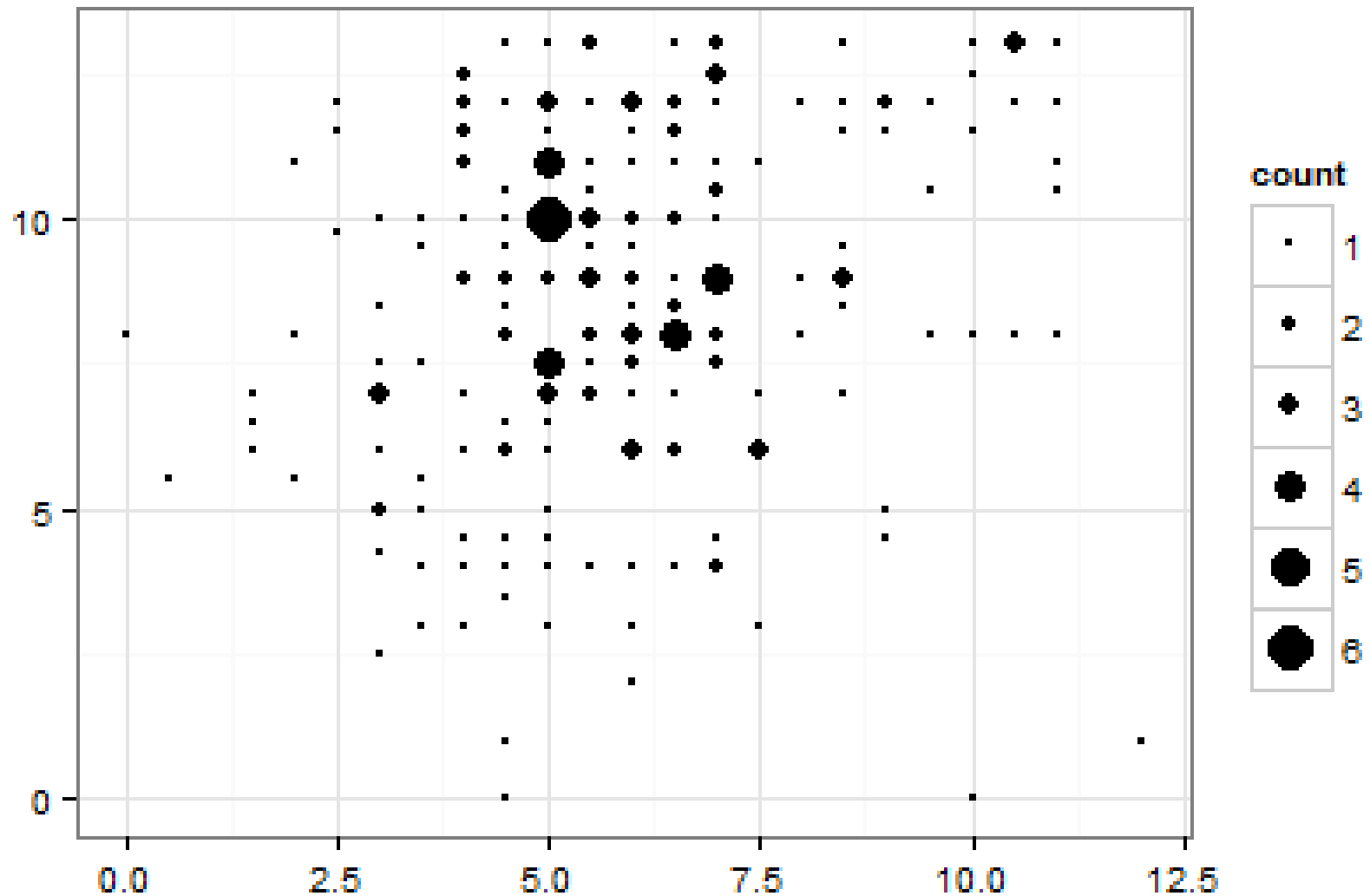
Overplotting megoldások 1: jitter



Overplotting megoldások 2: átlátszóság



Overplotting megoldások 3: méret

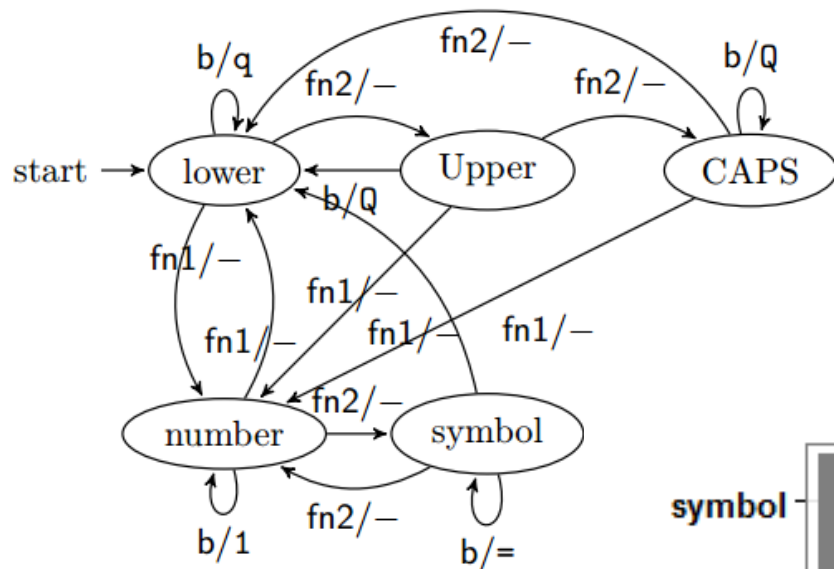


SOK VÁLTOZÓ

≥ 3 változó

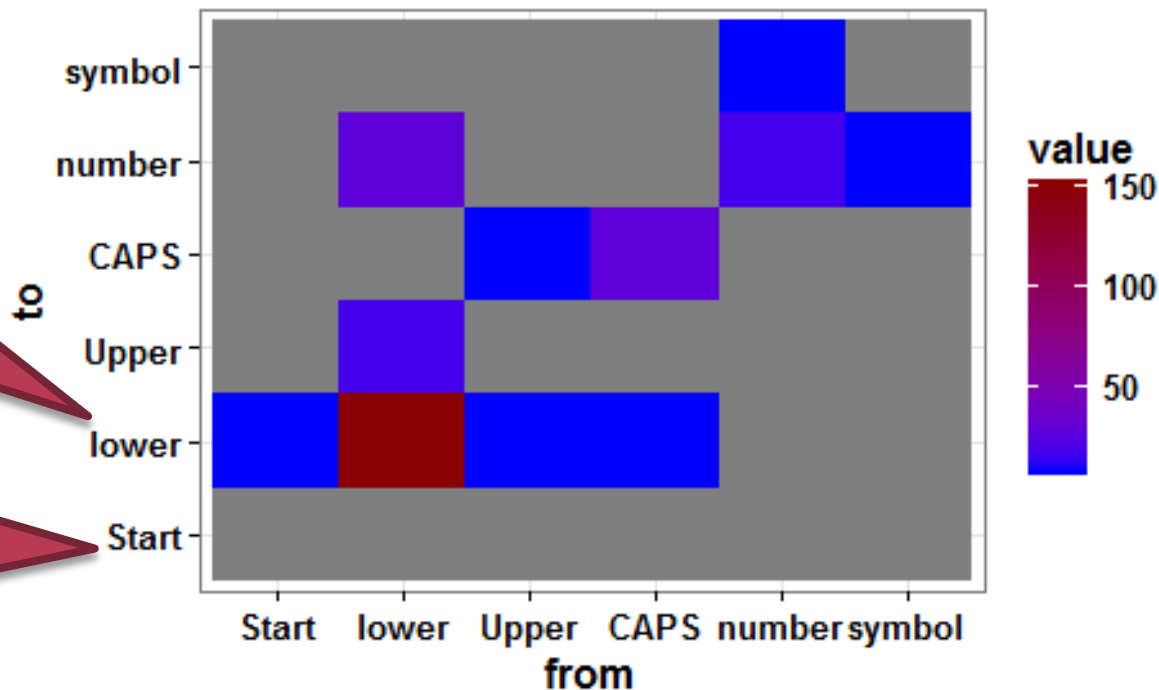
- A grafikai objektumok attribútumait változtatom
 - Szín
 - Méret
 - Textúra
 - Hely – ez triviálisnak tűnik, de a treemapnél van jelentősége
- Pl. heatmap, treemap

Heatmap: lefutási statisztikák

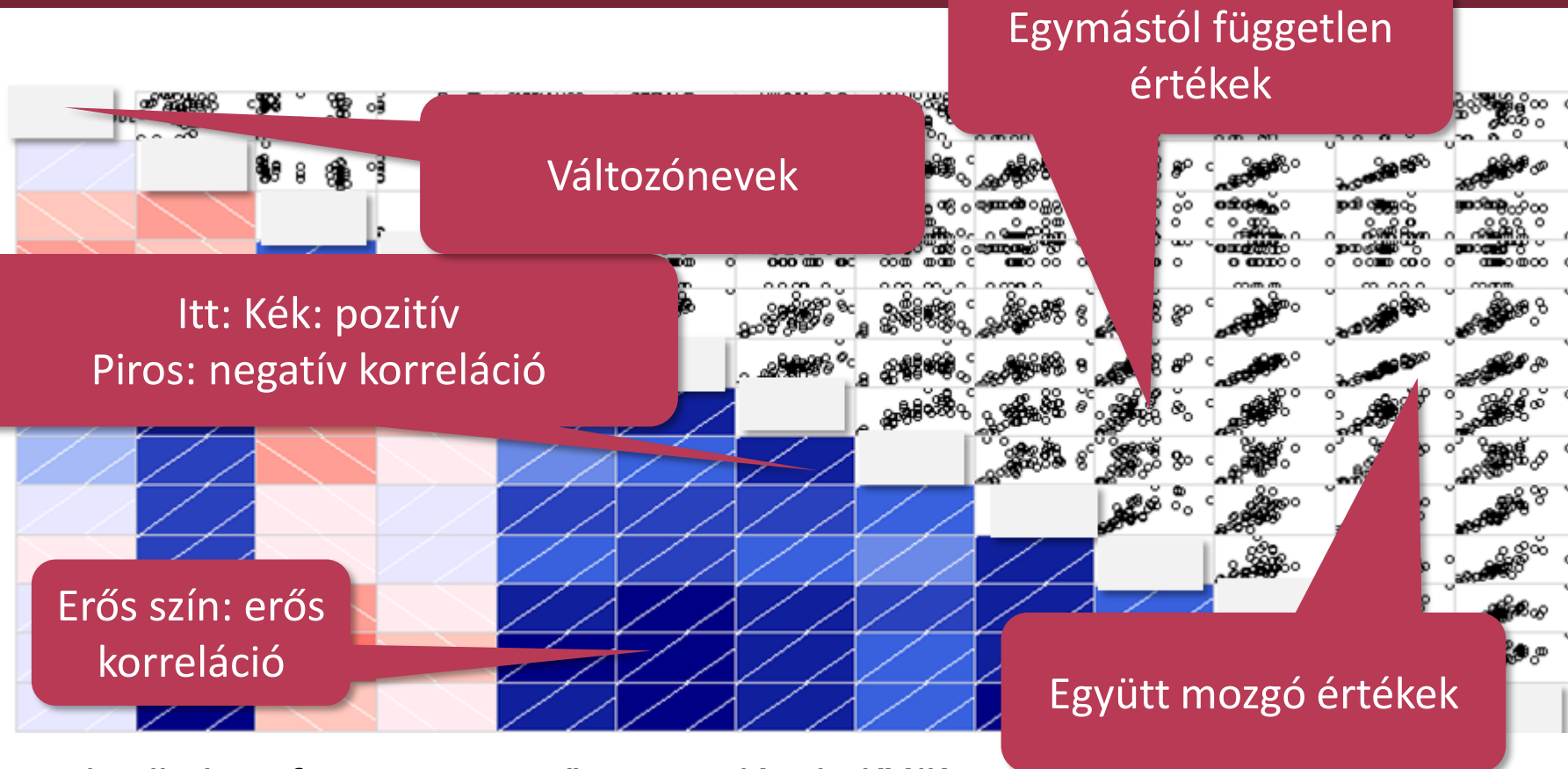


Inkább csak sima szöveget írunk

A Startban mindig csak kezdünk, oda nem jutunk vissza



Kitekintés: több érték páronkénti korrelációja



R statisztikai szoftver „corrgram” csomagjával előállítva.

Korreláció (ld. Valószínűségi számítás):

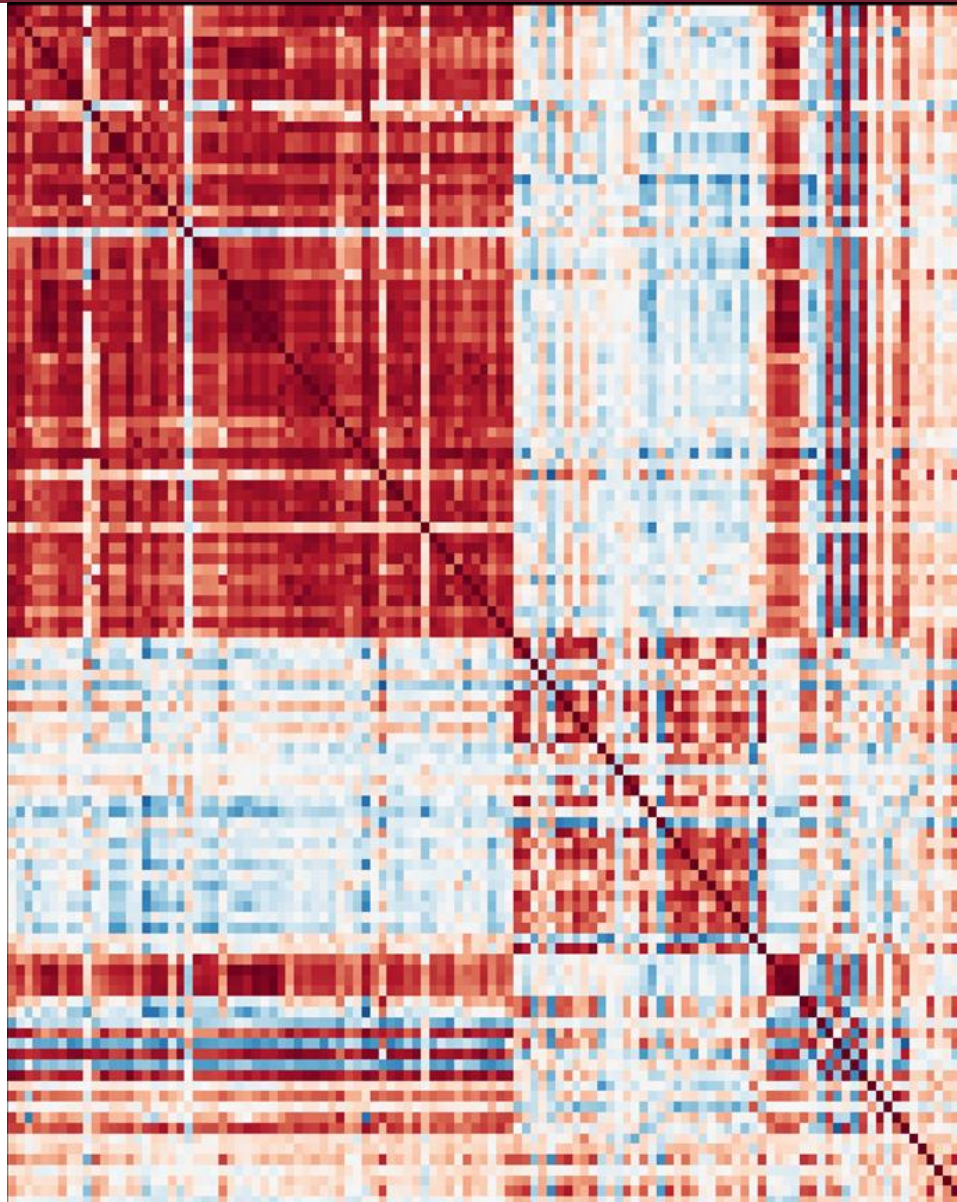
két érték közti lineáris kapcsolat erőssége és iránya

Átló felett: **scatterplot mátrix**

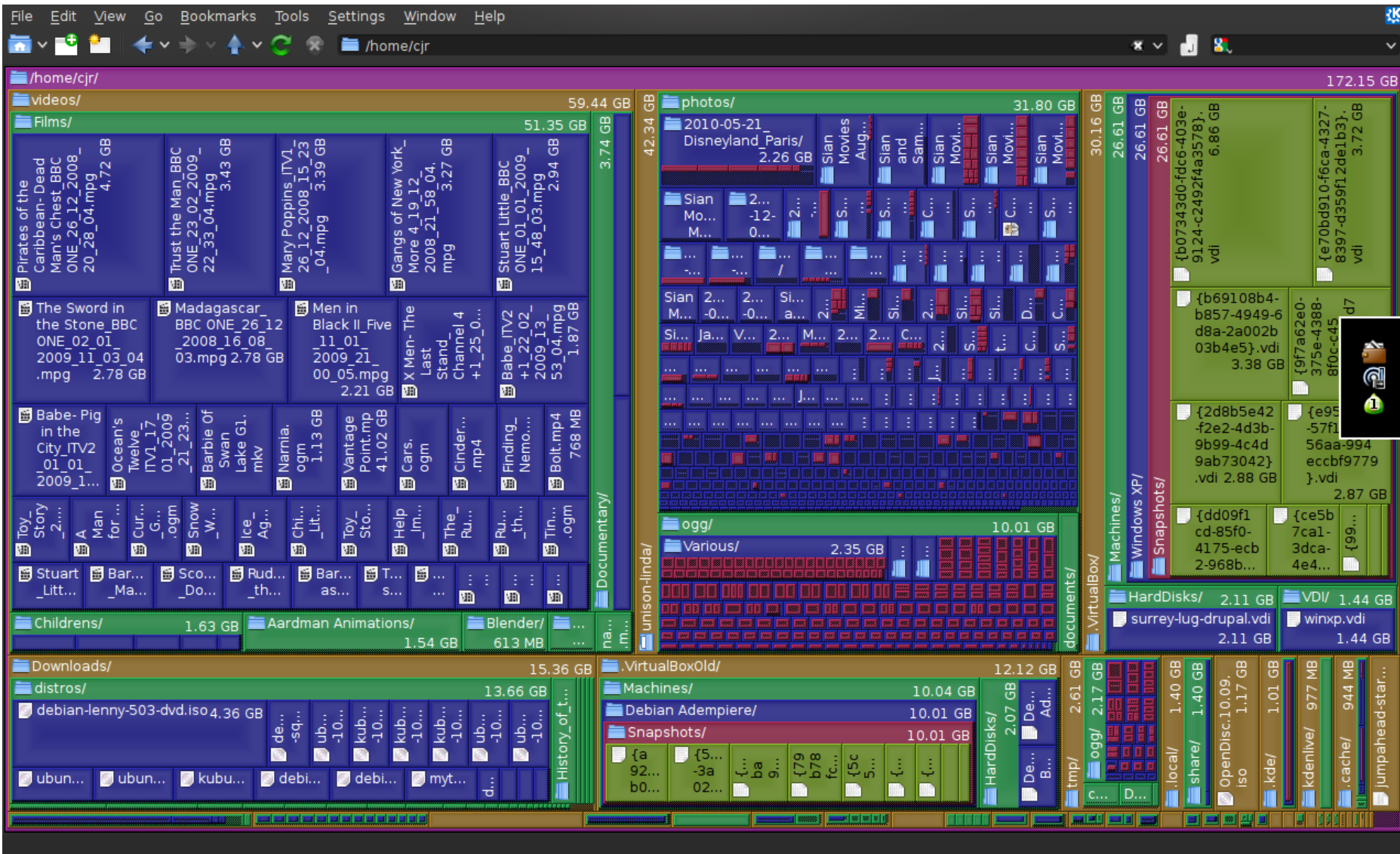
Cél: együtt mozgó értékek kiszűrése, **kiugró értékek (outlierek)** azonosítása.

→ Mik a terhelés/előrejelzés szempontjából lényeges változók?

Termelés – mi mivel függ össze?



Treemap: állományrendszer



Párhuzamos koordináták

- Tengelyek: dimenziók/koordináták

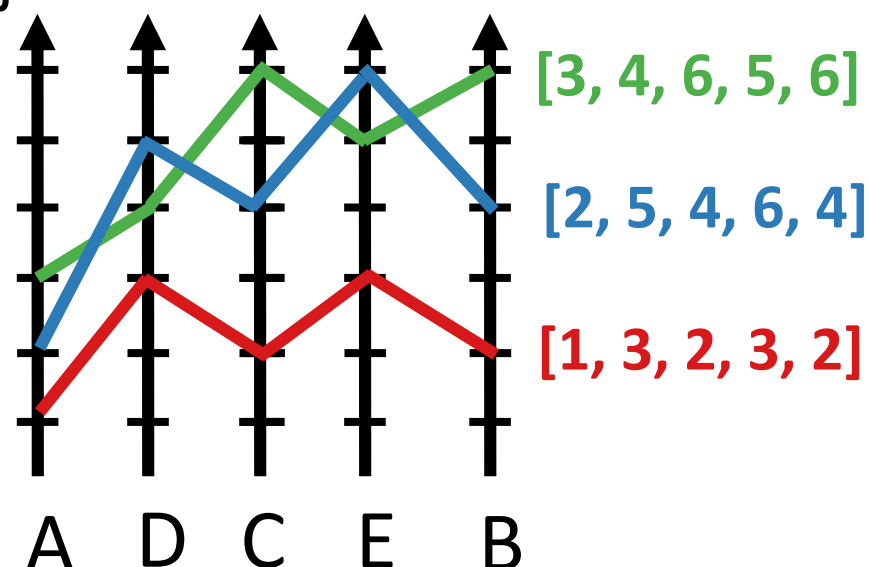
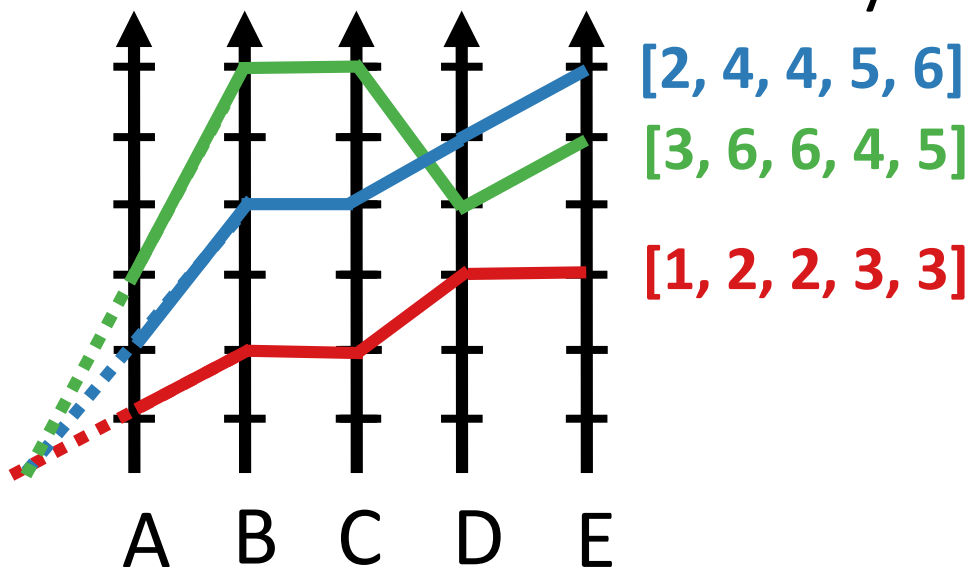
- tetszőleges számú

- tetszőleges skála

- Egy vonal egy mérés (darabszám?)

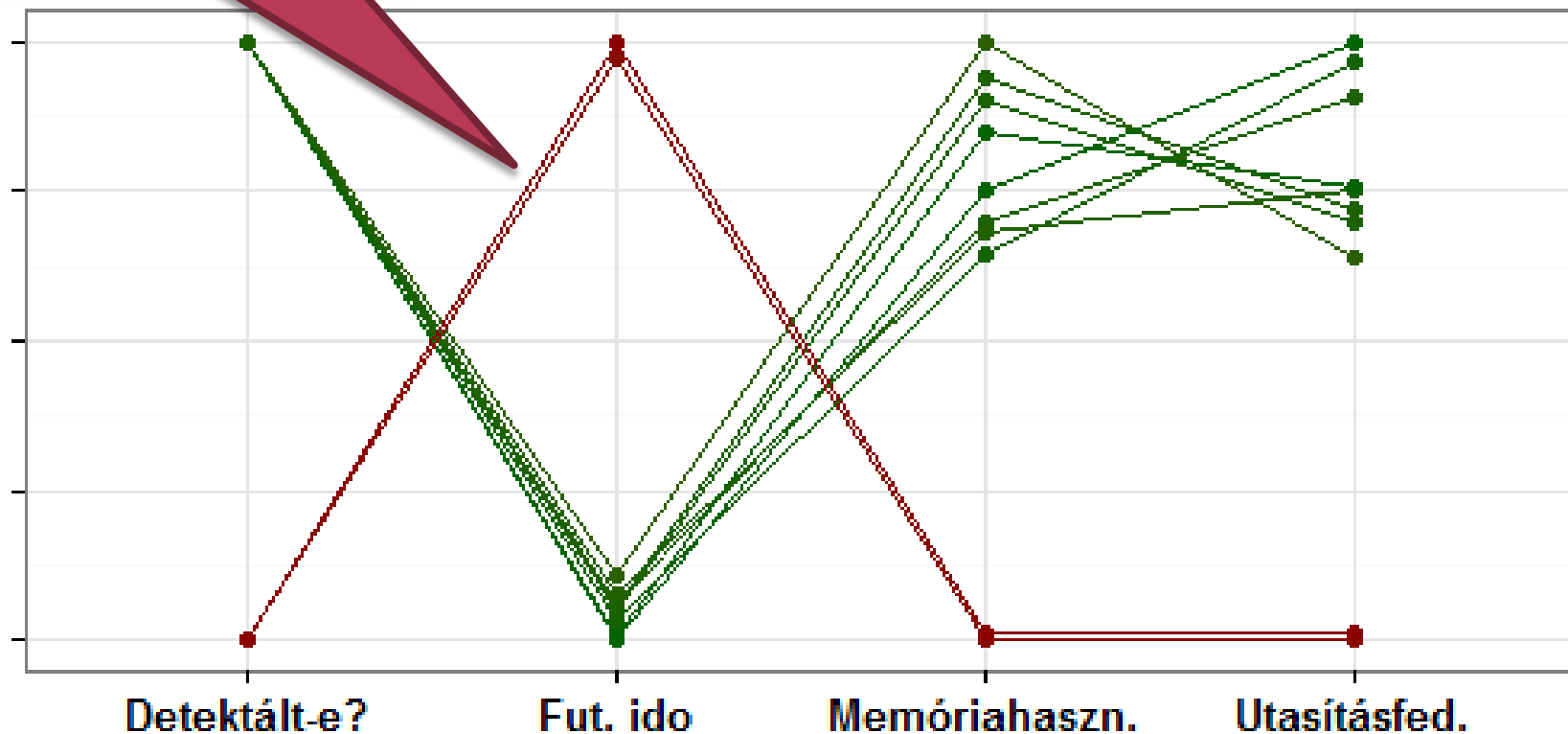
- Kompakt és skálázható

- Koordináta sorrend befolyásolja a kiértékelést



Párhuzamos koordináták: tesztesetek elemzése

1 teszteset 1 törött vonal

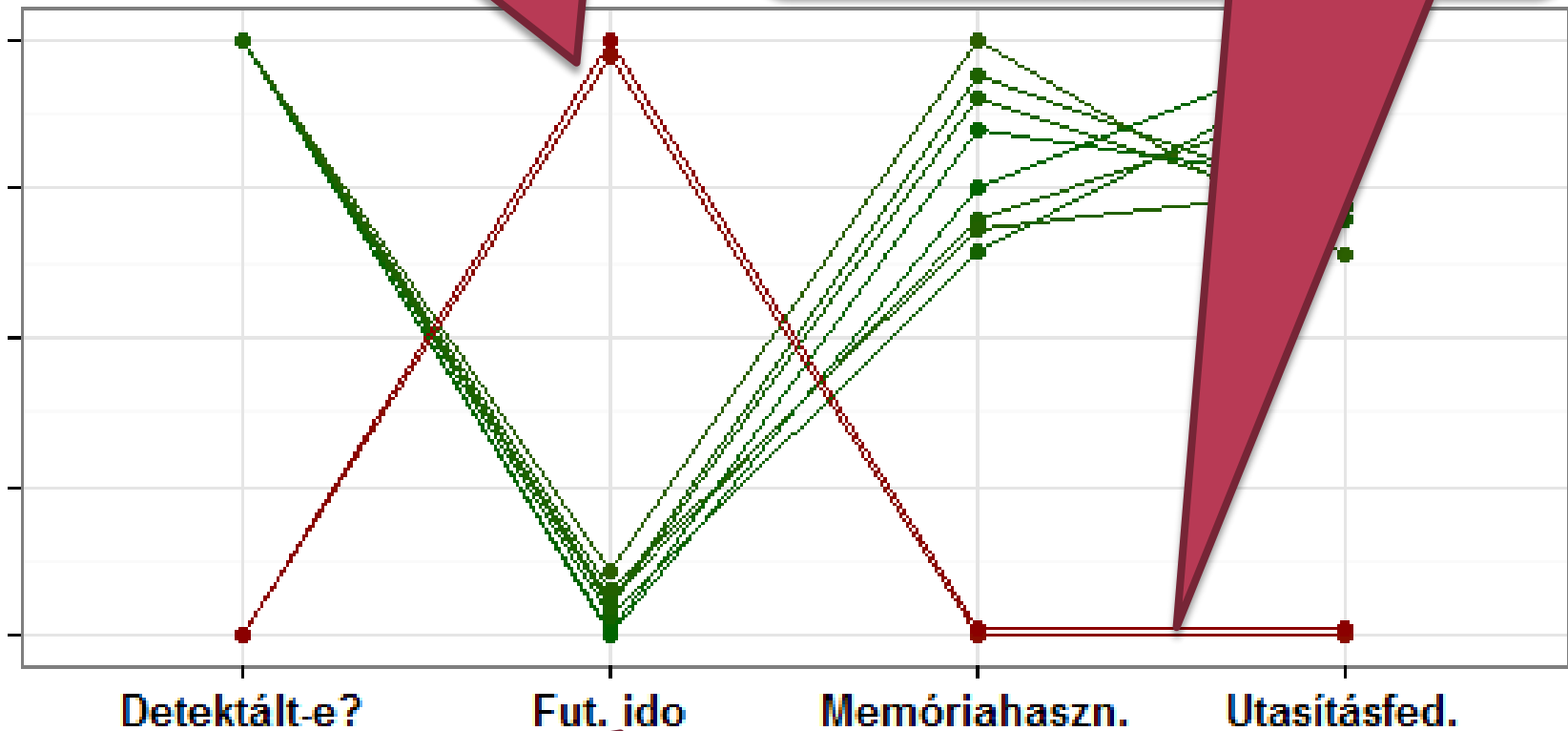


A változók az x tengelyen jelennek meg

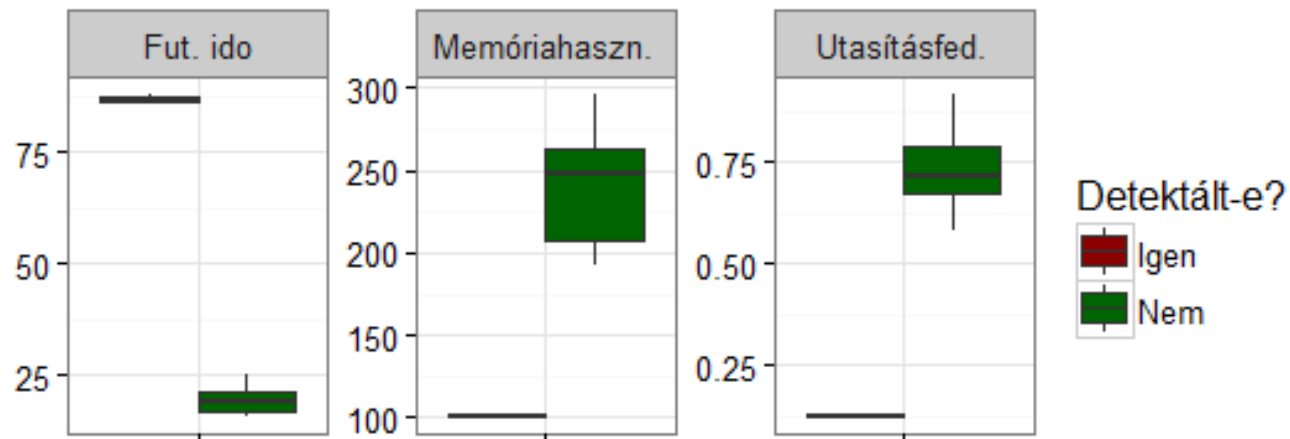
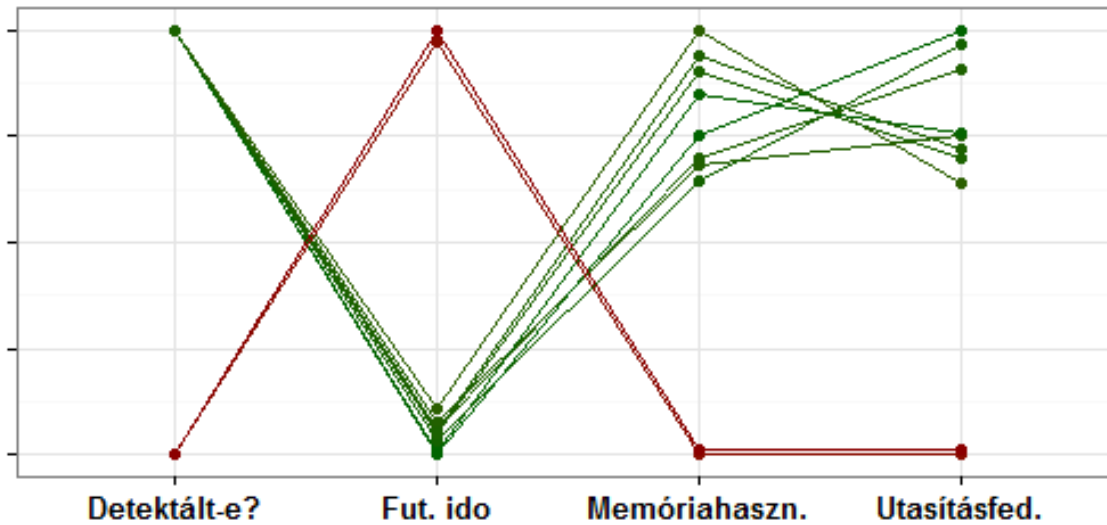
Párhuzamos koordináták: tesztesetek elemzése

Timeout?

A hibát detektálók az érdemi számításig valószínűleg el sem jutnak



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



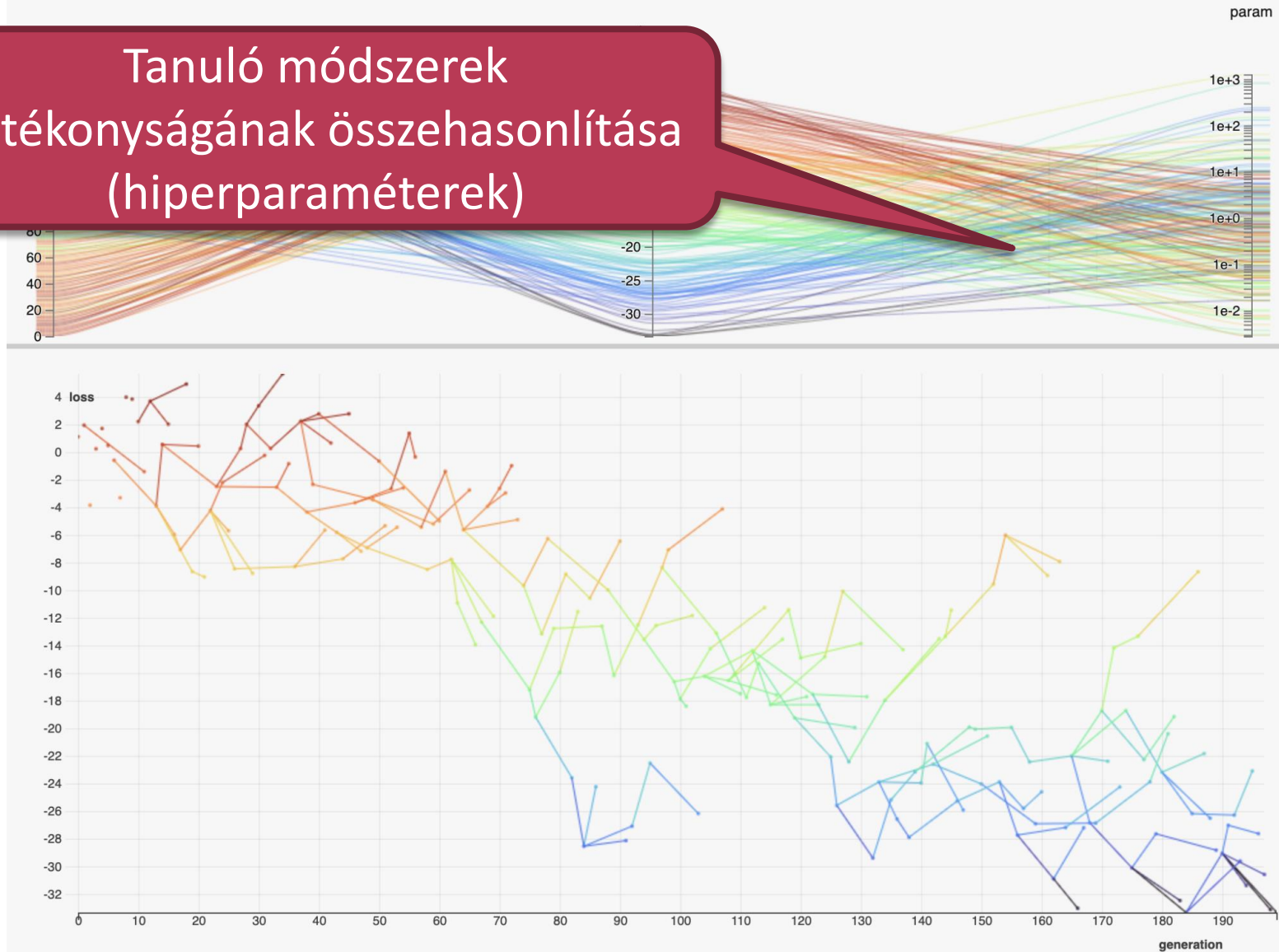
Mire használhatjuk?

Csoportosítás

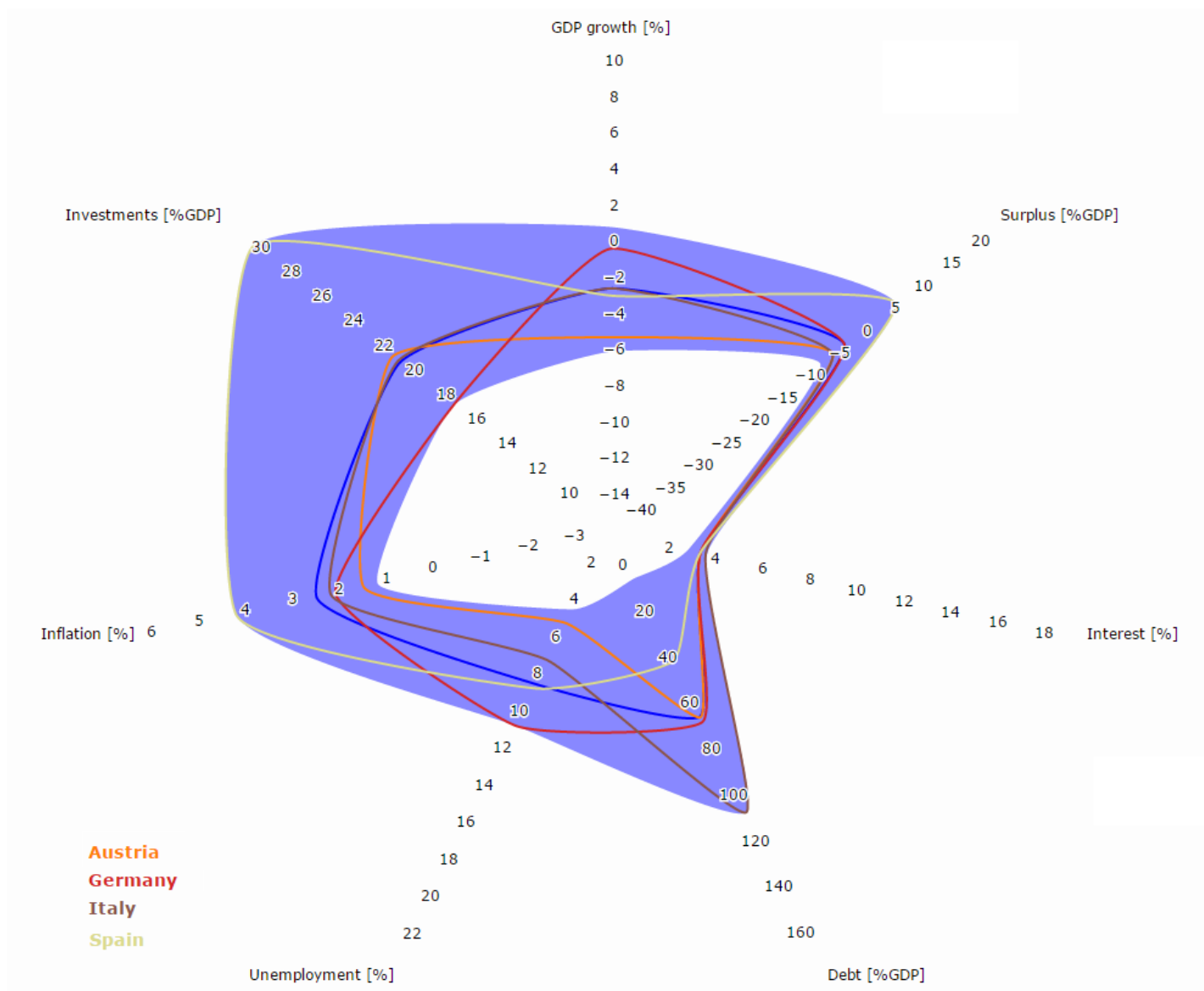
Korreláció

Hiplo (facebook)

Tanuló módszerek
hatékonyságának összehasonlítása
(hiperparaméterek)



Radar chart: egy párhuzamos koord. kiterjesztés

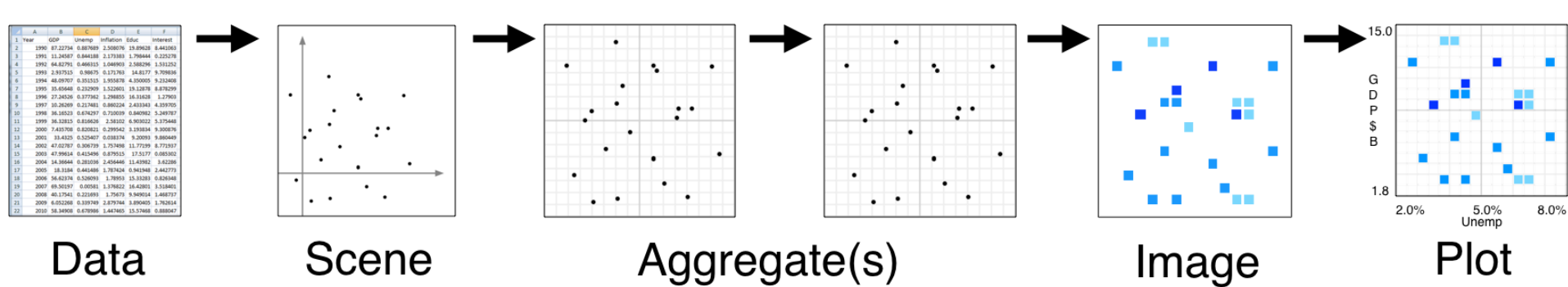


AGGREGÁLÁS + VIZUÁLIS MEGJELENÉS

Datashader

- Gyakran nagy az adat (1.000.000+)
- DE nem az egyedi adatvektor, hanem az értékek eloszlása érdekes

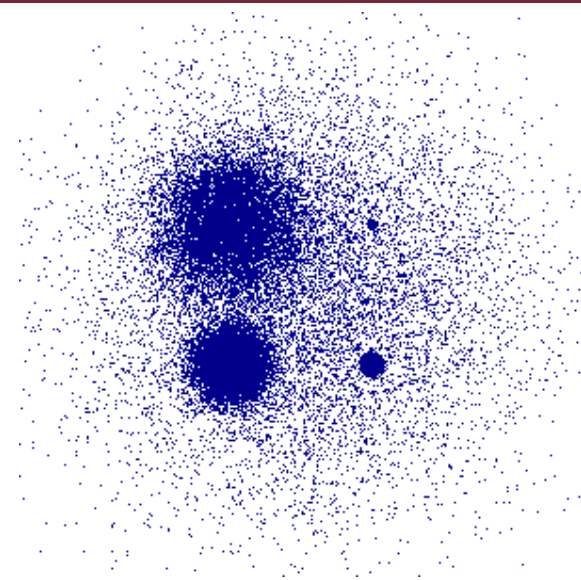
Projection Aggregation Transformation Colormapping Embedding



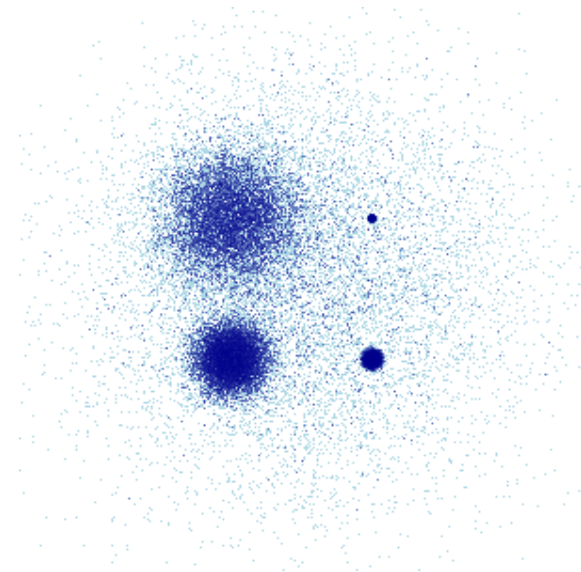
http://datashader.org/getting_started/2_Pipeline.html

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

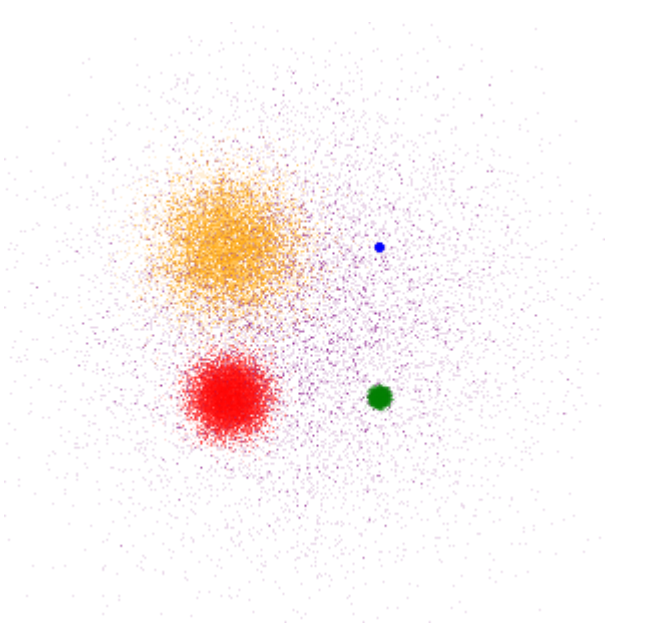


# Szűrés, színezés

- 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')
```



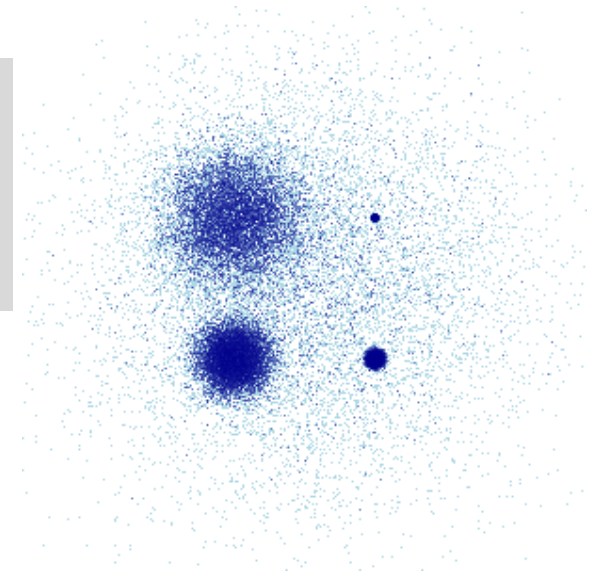
Szűrt előfordulás
Cellaintenzitás
Cellabeliek száma<--
Cellabeliek száma | d3)



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

