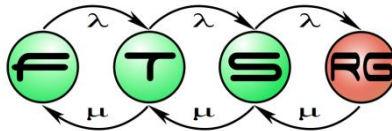


Anomáliadetektálás IT adatokon

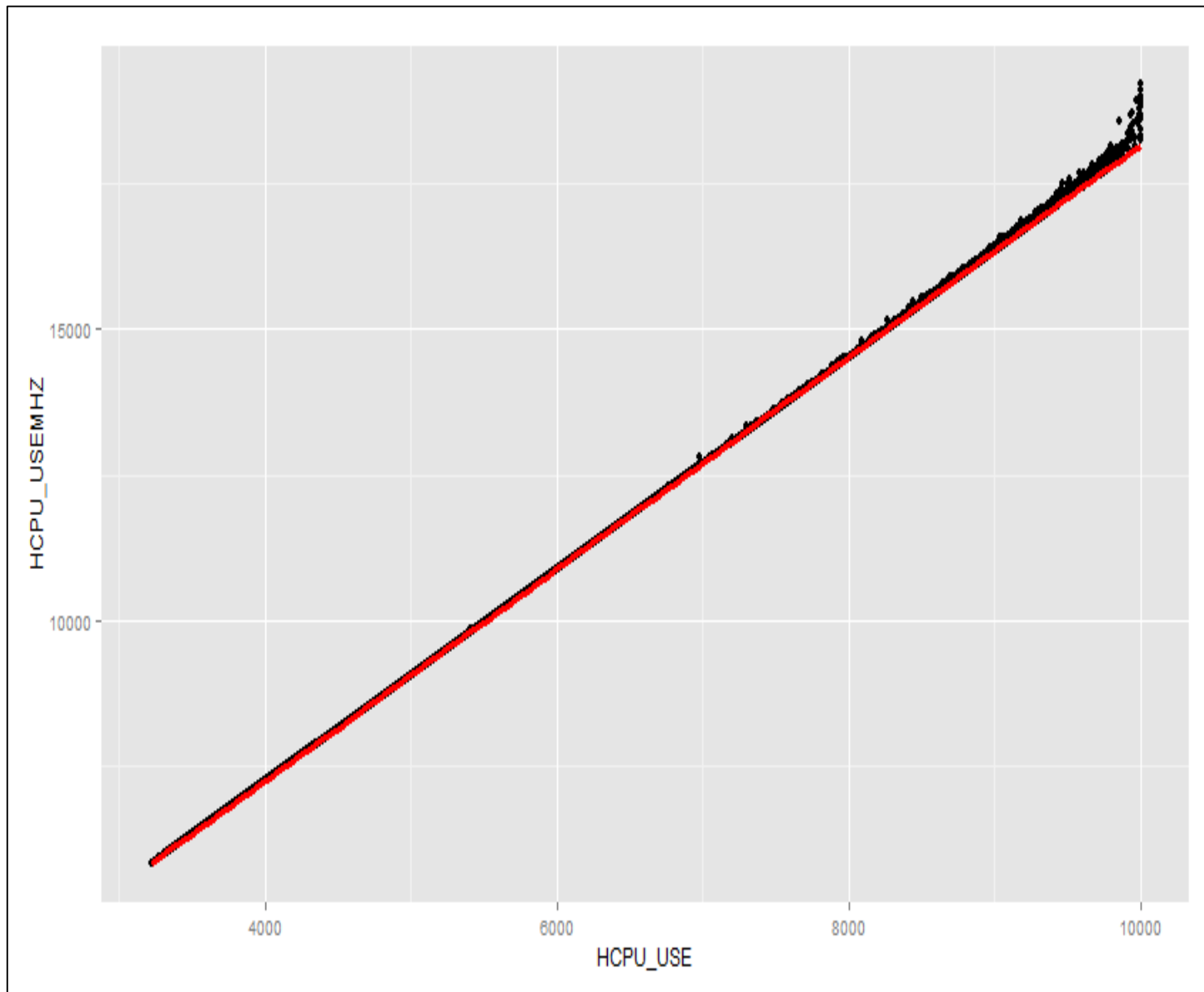
Salánki Ágnes

salanki@mit.bme.hu

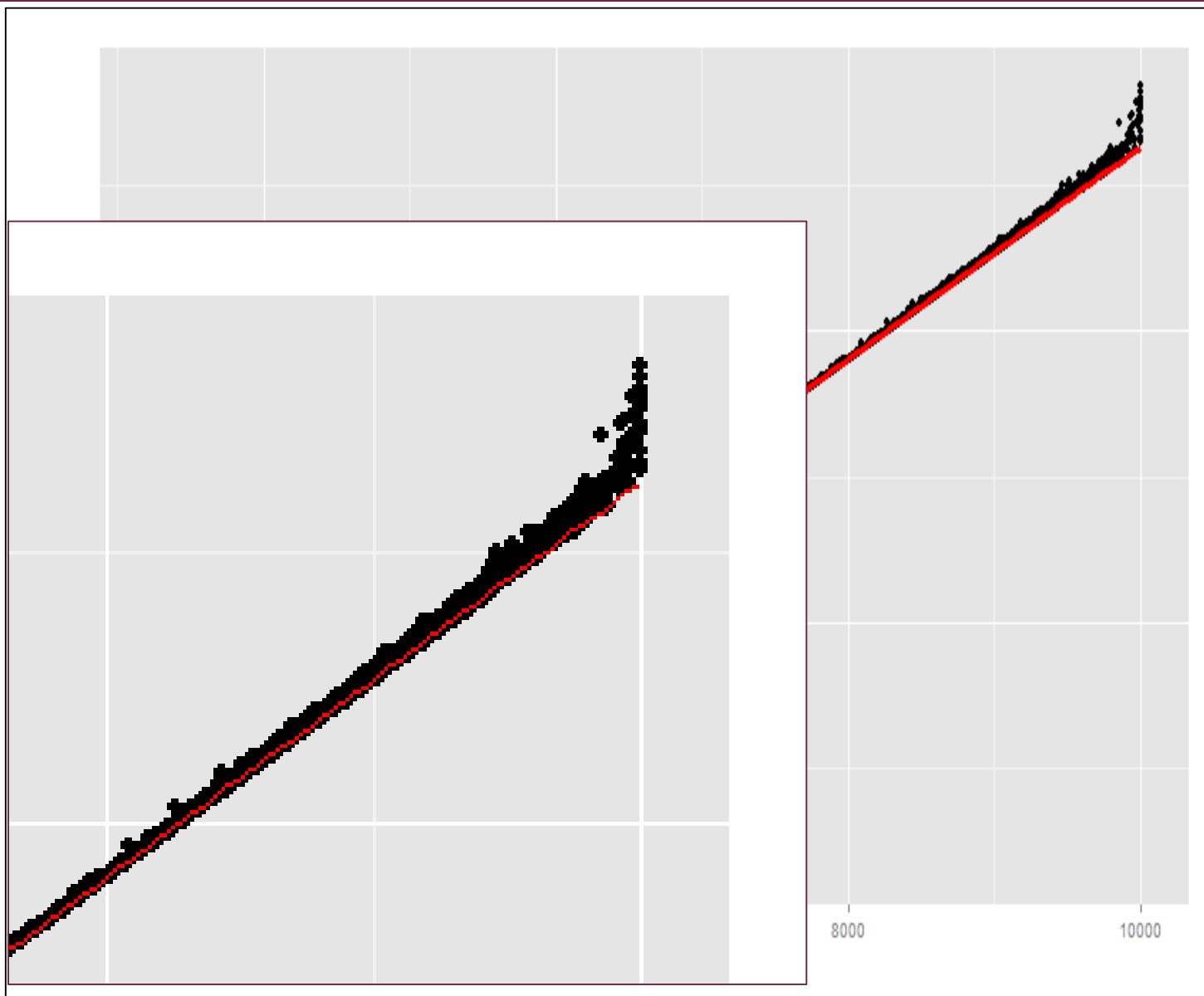
2014.05.10.



Alapfeladat



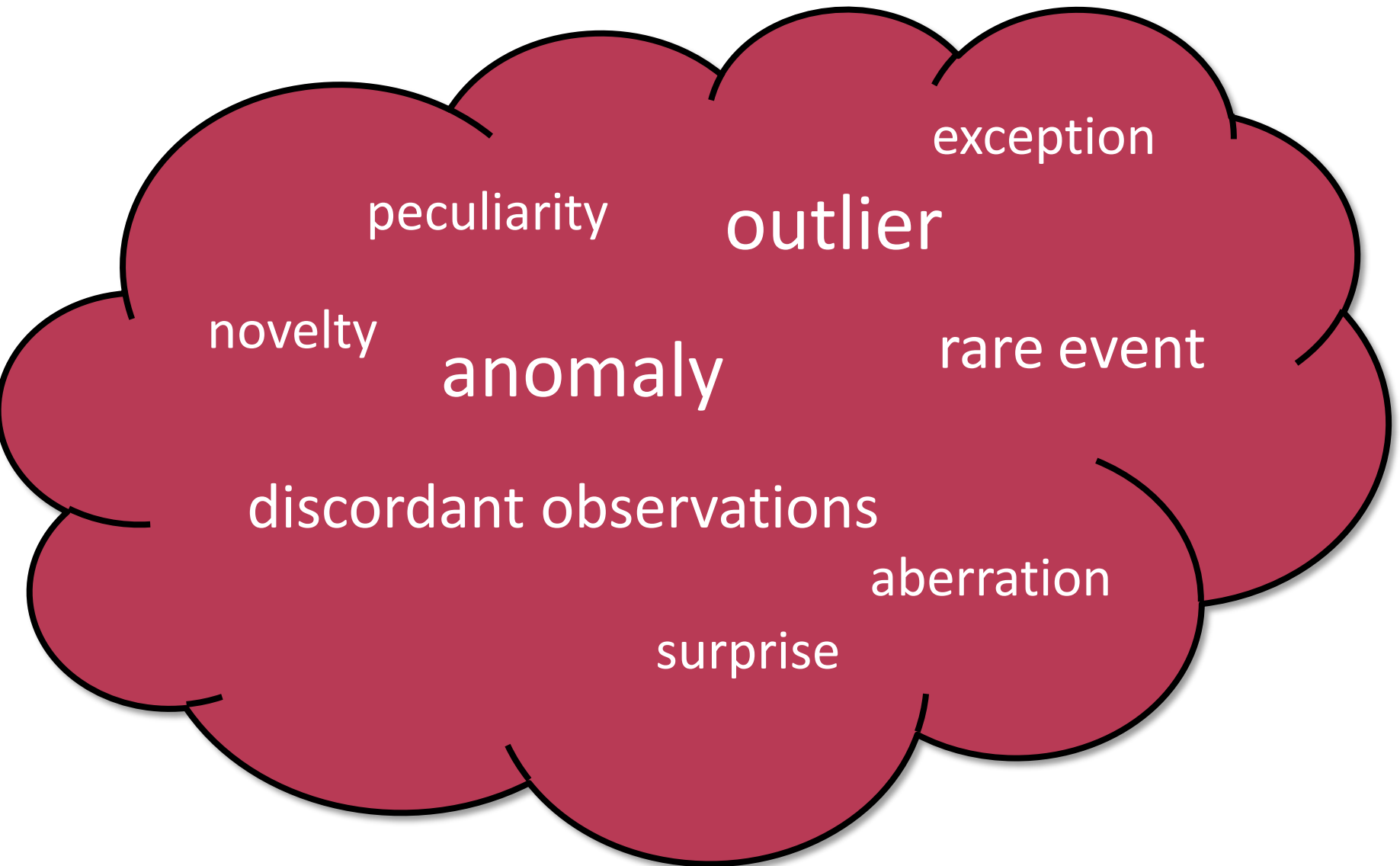
Alapfeladat



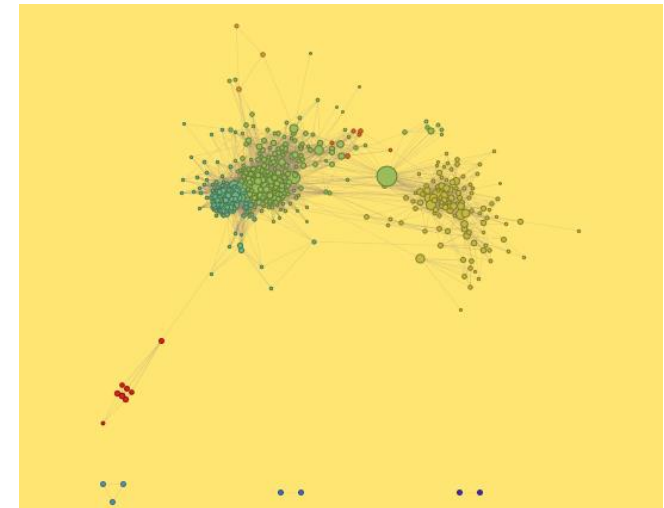
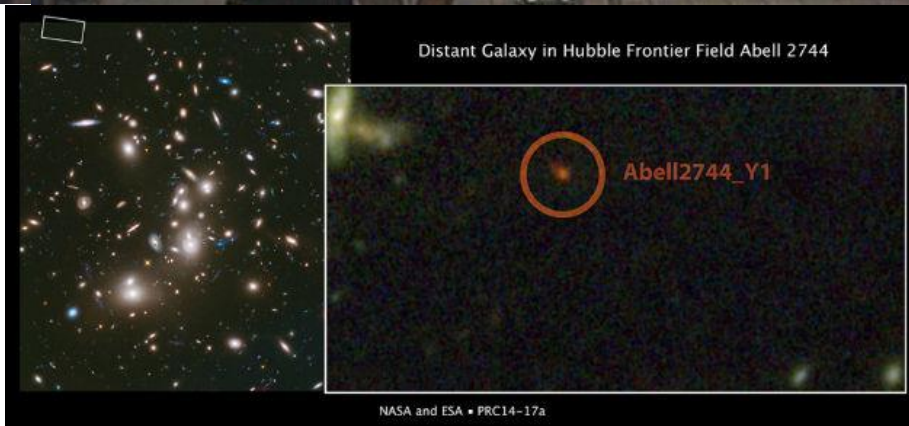
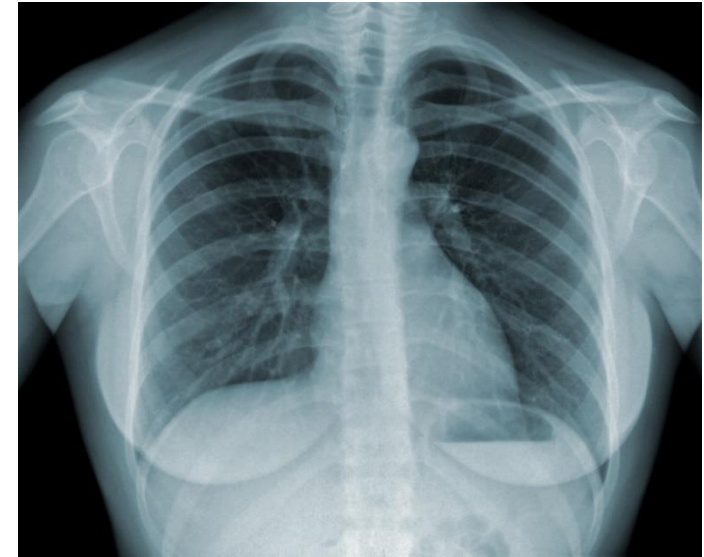
Alapfeladat

- Nagy rendelkezésre állású rendszerek
 - A hibák ritkán fordulnak elő
 - Általában csomósodnak
 - Időben
 - Kiváltó tényezők szerint
- Skálázható
- Félig felügyelt módszerek?

Alapfogalmak

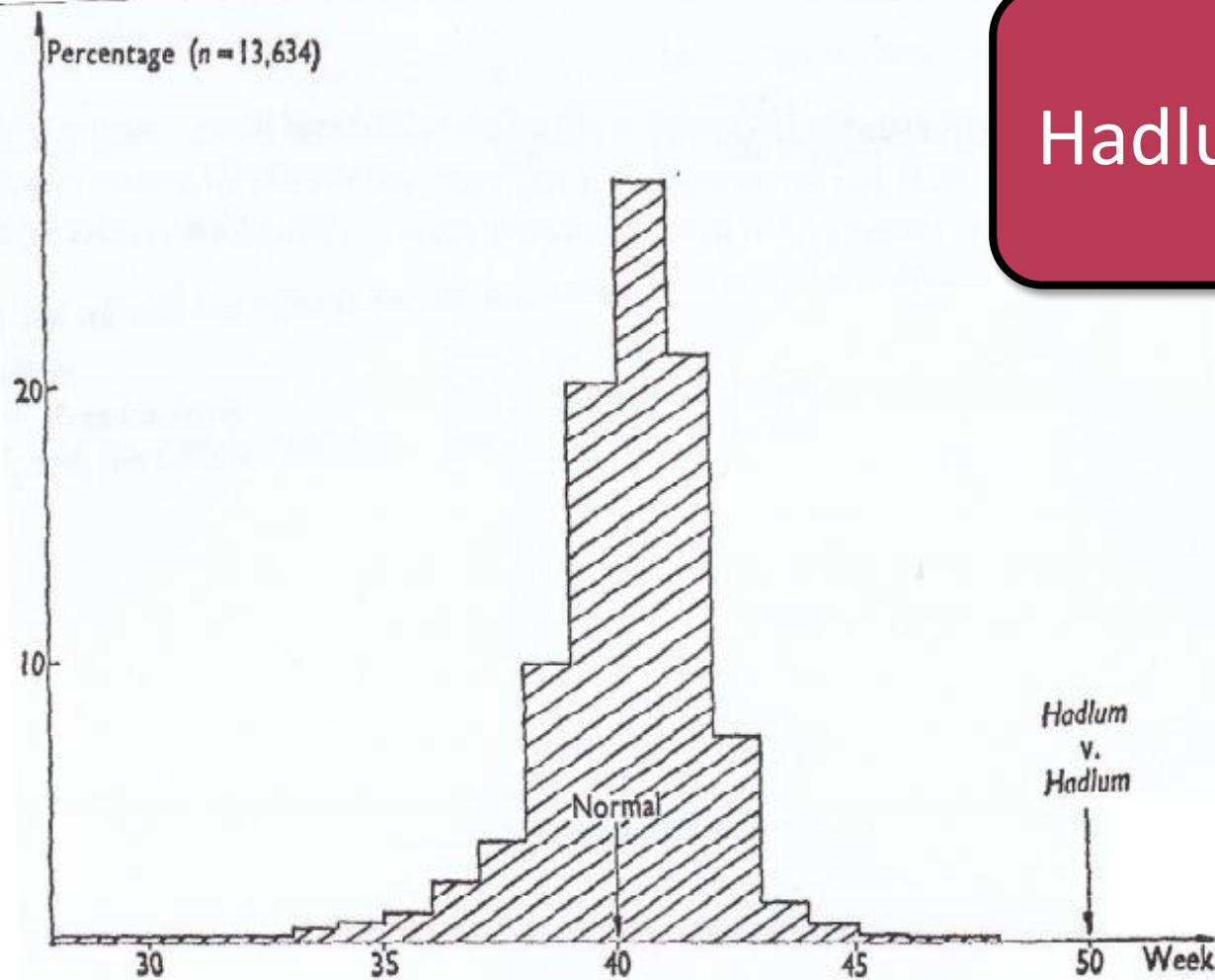


Használati esetek



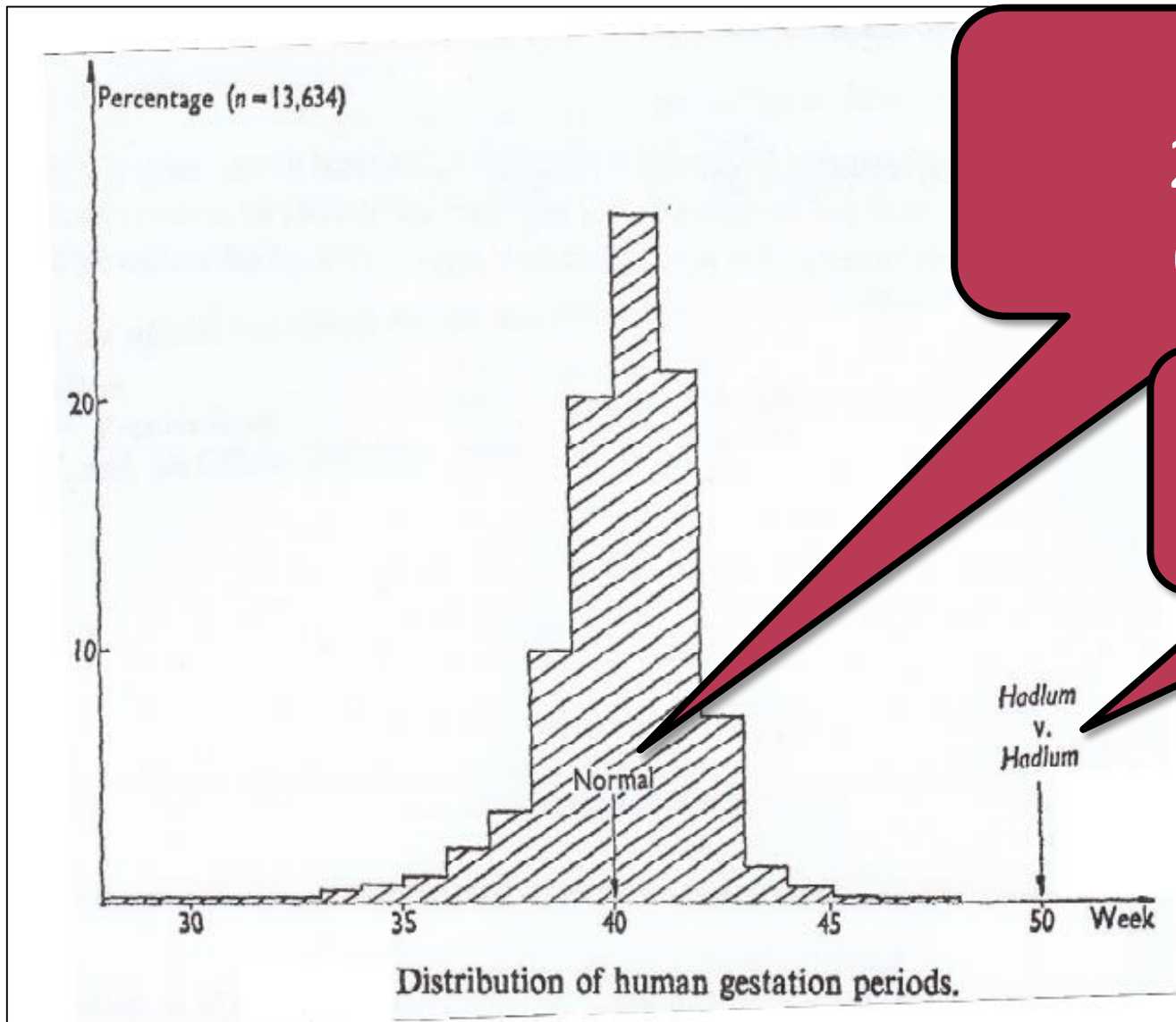
Történelem

Hadlum vs. Hadlum



Distribution of human gestation periods.

Történelem



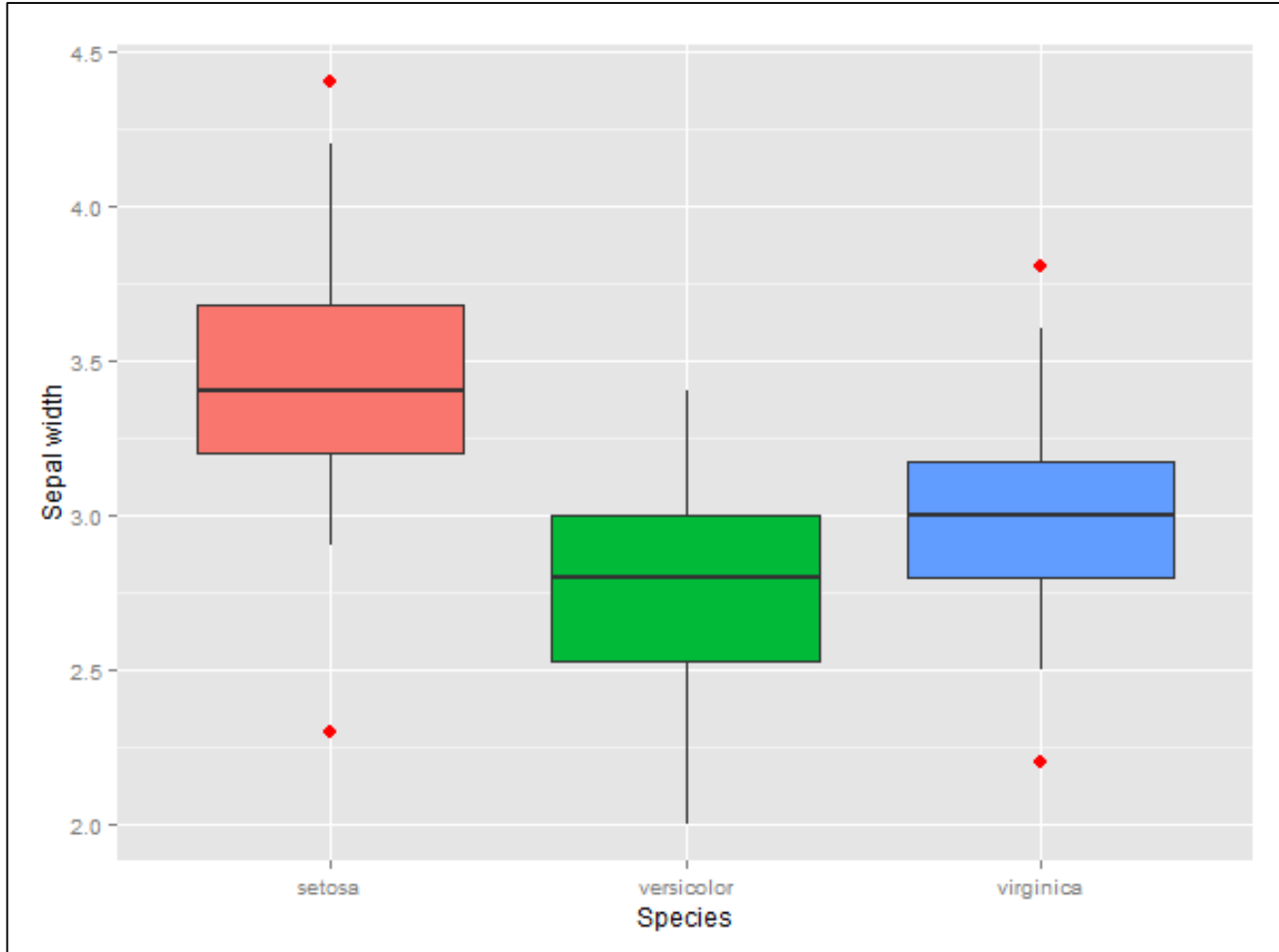
Átlag:
280 nap
(40 hét)

Mrs. Hadlum:
349

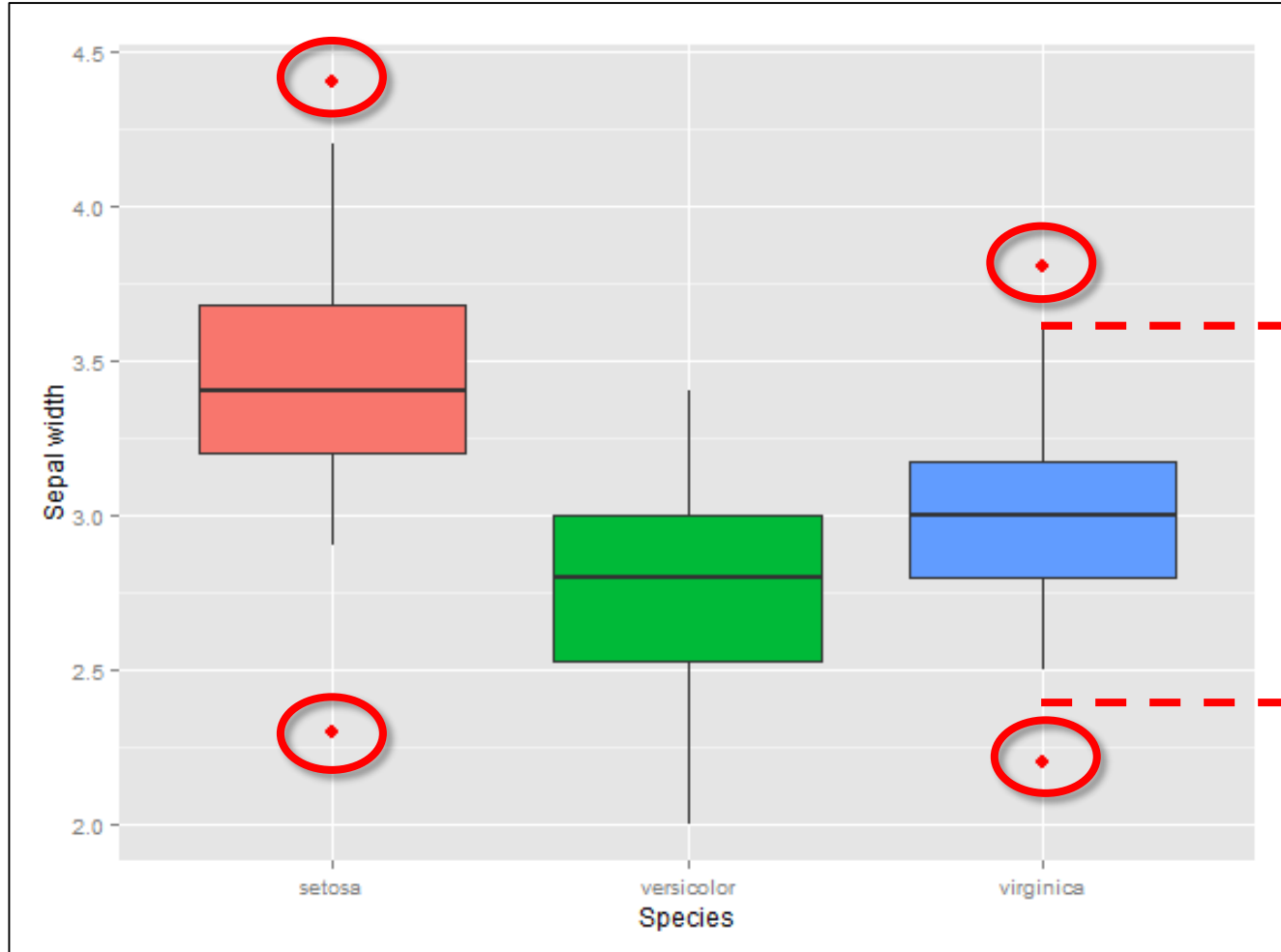
Definíció

- Kevés van belőlük
- „Gyanús”, hogy más a generáló folyamat/forrás

0. lépés: statisztikai outlierok



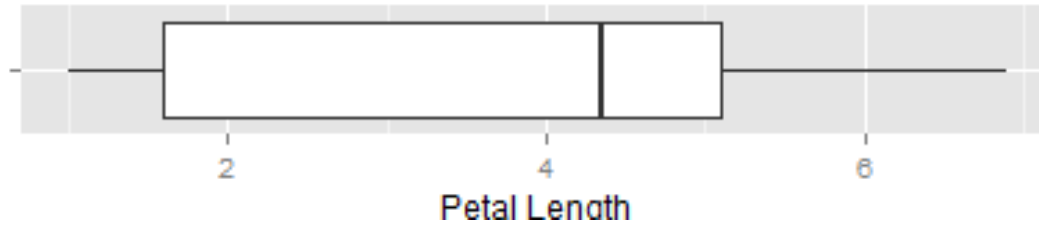
0. lépés: statisztikai outlierok



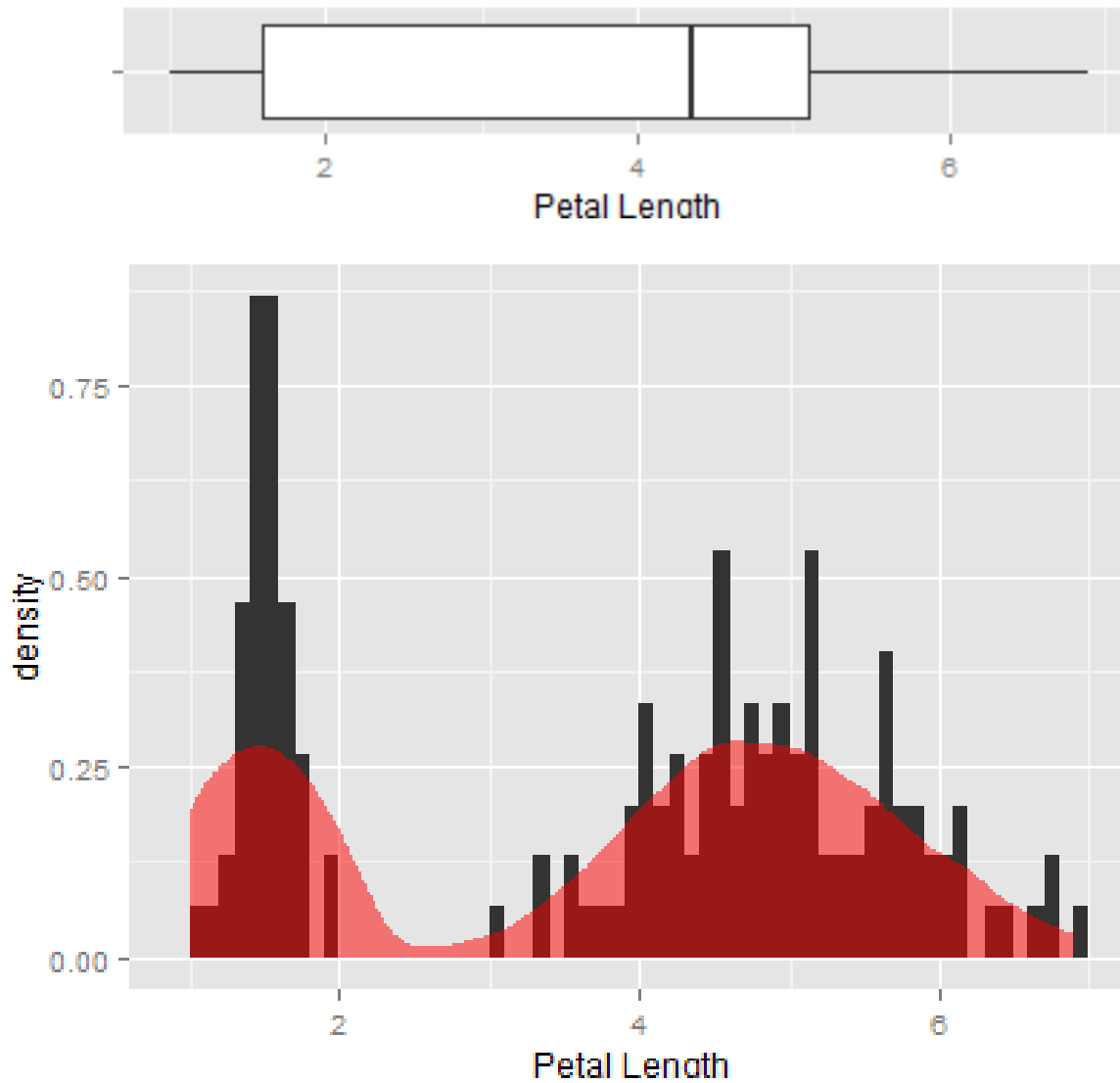
--- $Q3 + 1.5IQR$

--- $Q1 - 1.5IQR$

0. lépés: statisztikai outlierok

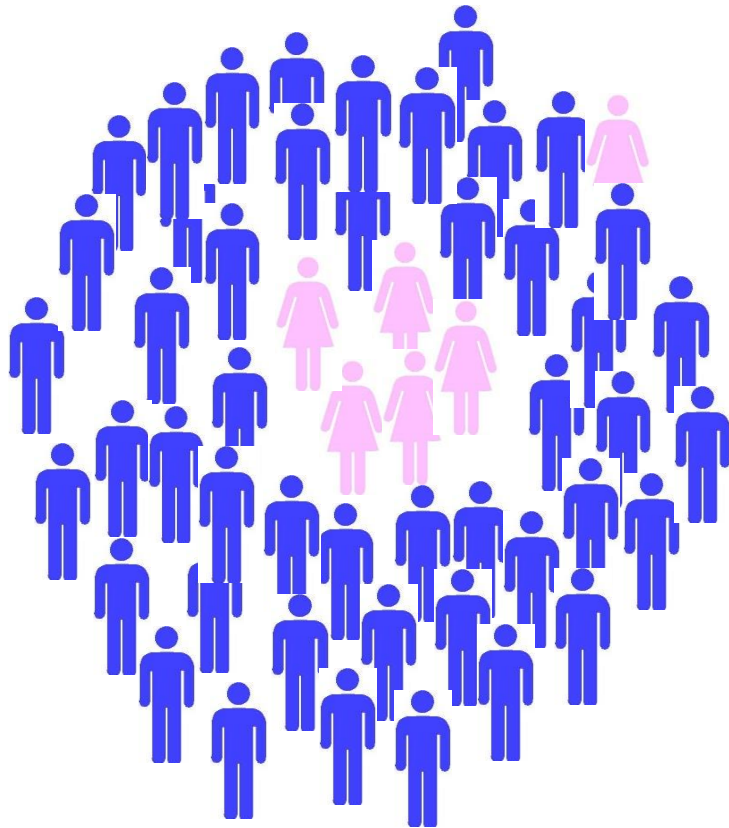


0. lépés: statisztikai outlierok

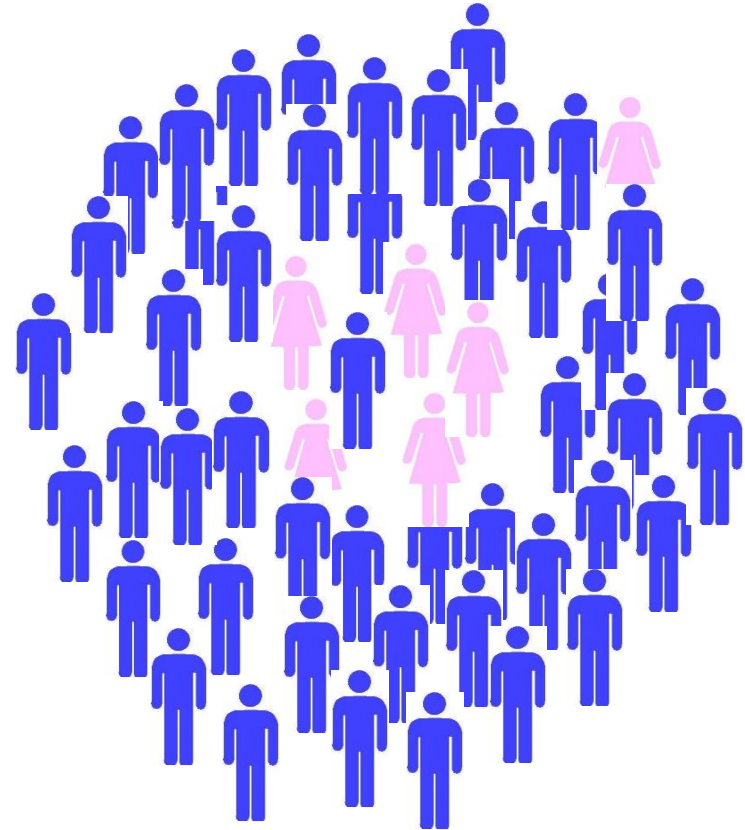


Megközelítések

- Távolság alapúak

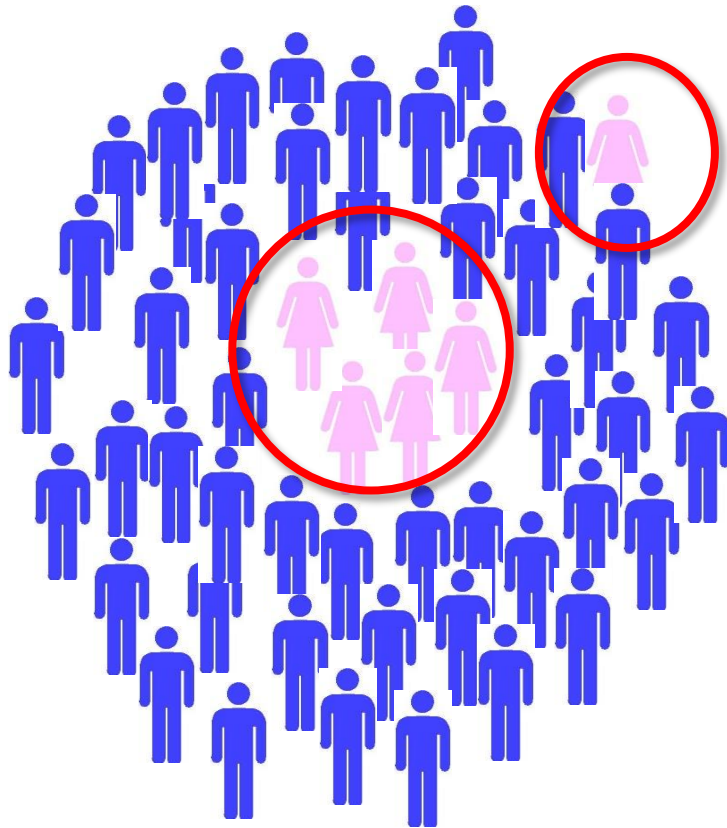


- Sűrűség alapúak

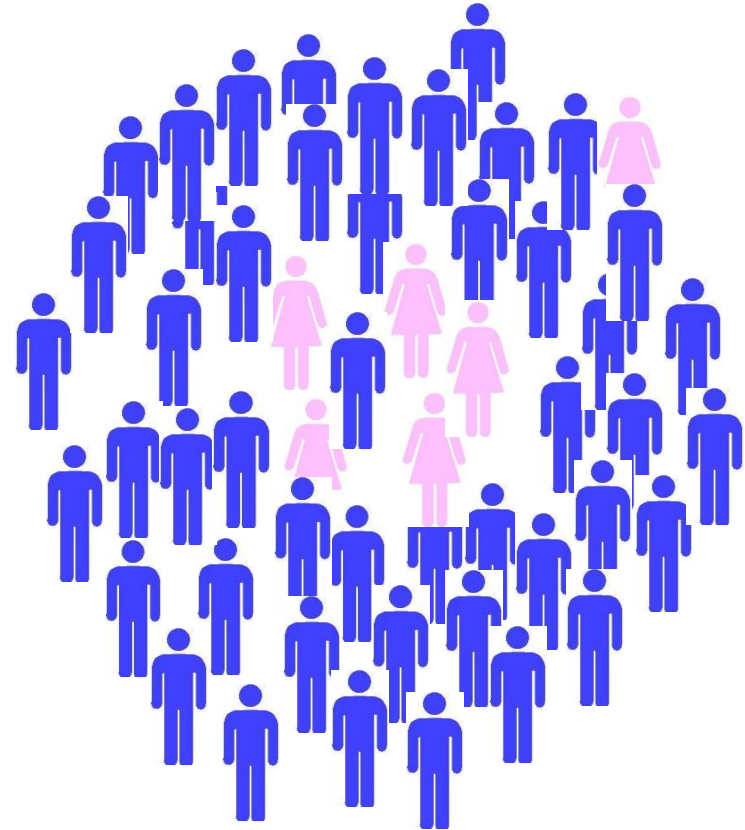


Megközelítések

- Távolság alapúak

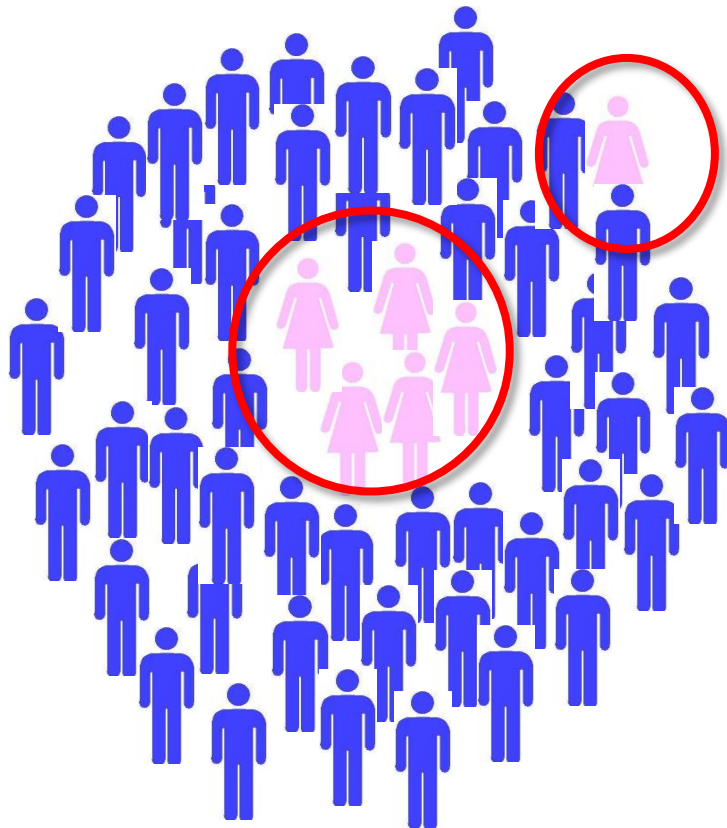


- Sűrűség alapúak

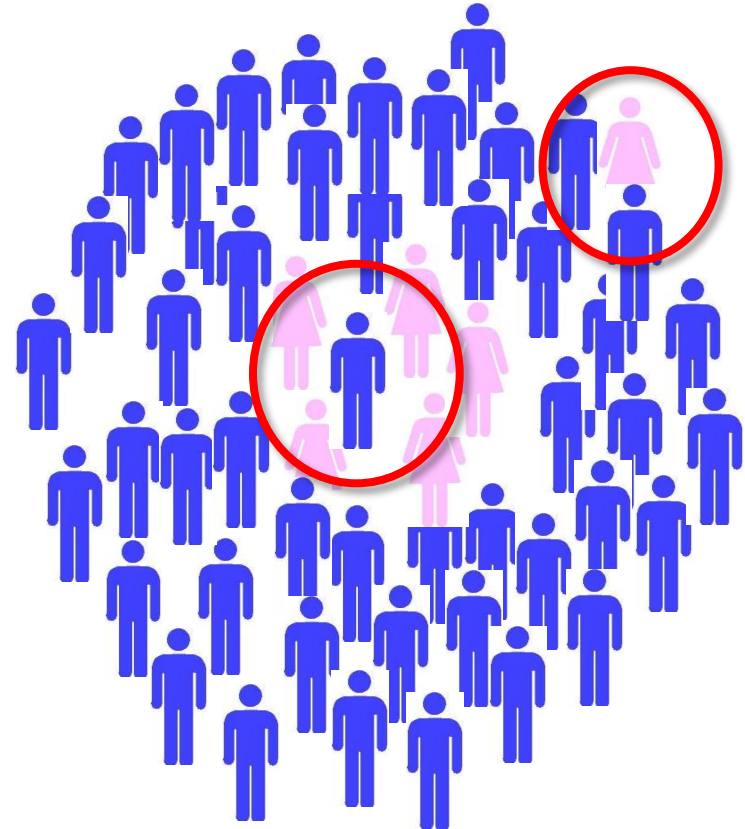


Megközelítések

- Távolság alapúak

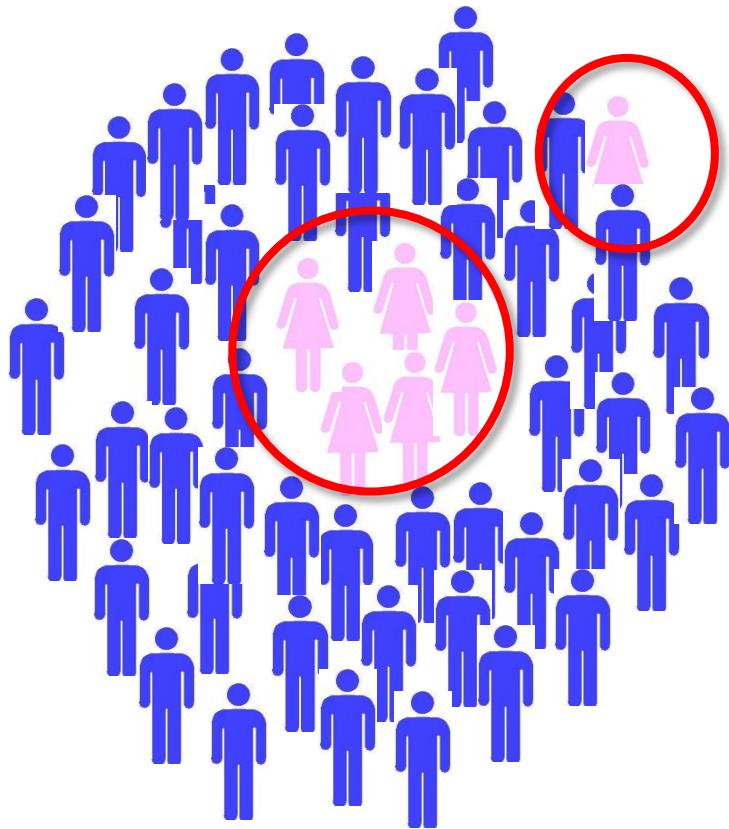


- Sűrűség alapúak

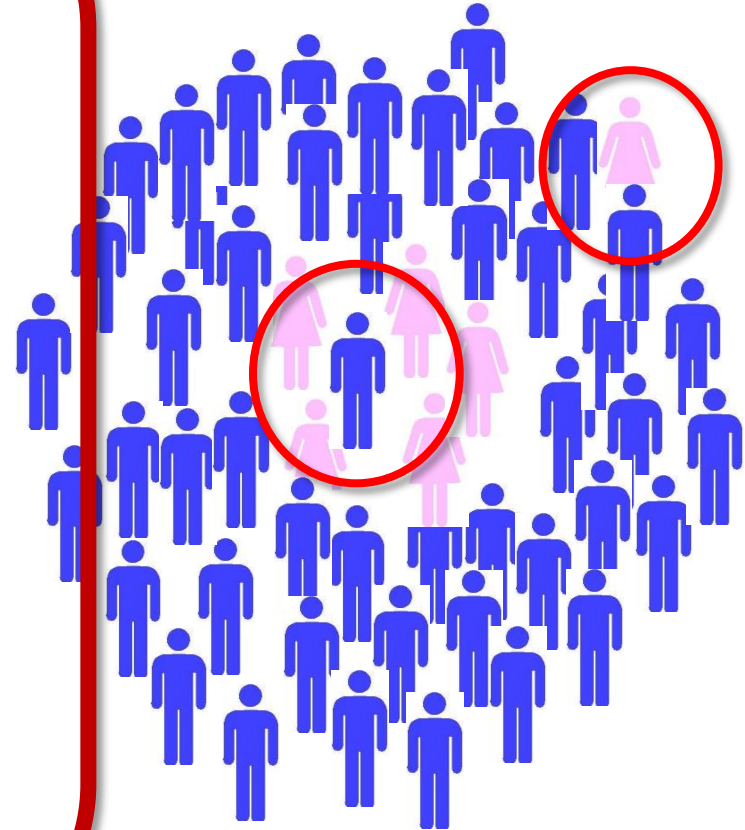


Megközelítések

Távolság alapúak

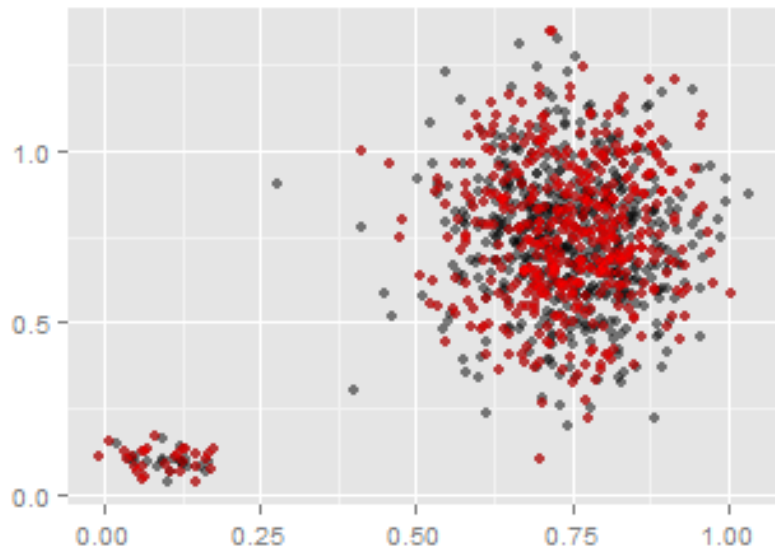


Sűrűség alapúak



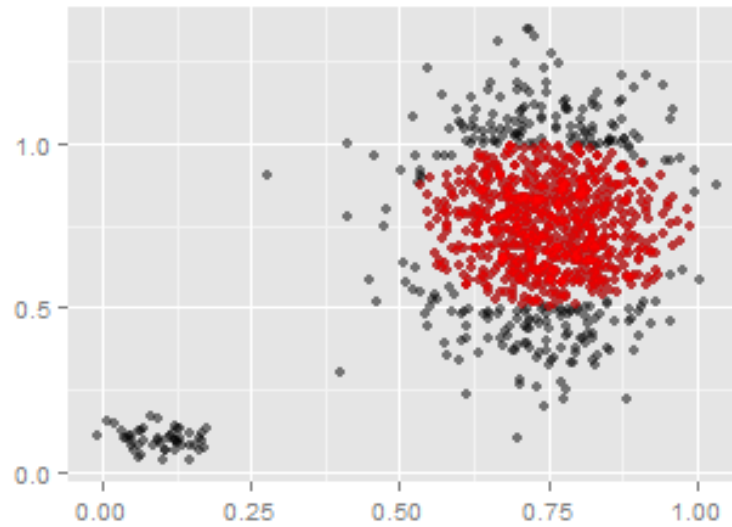
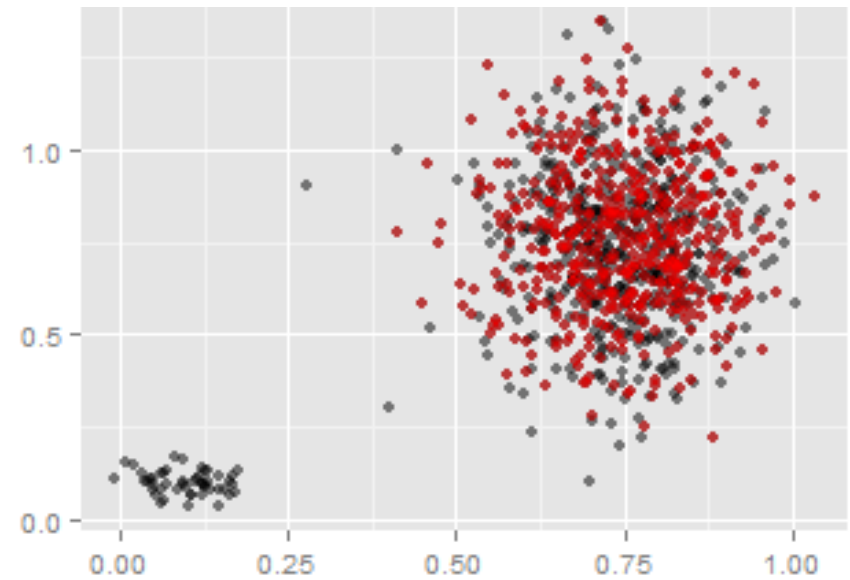
- Minimum Covariance Determinant
- Alapötlet
 - Keressük meg a legkompaktabb részhalmazt!

MCD

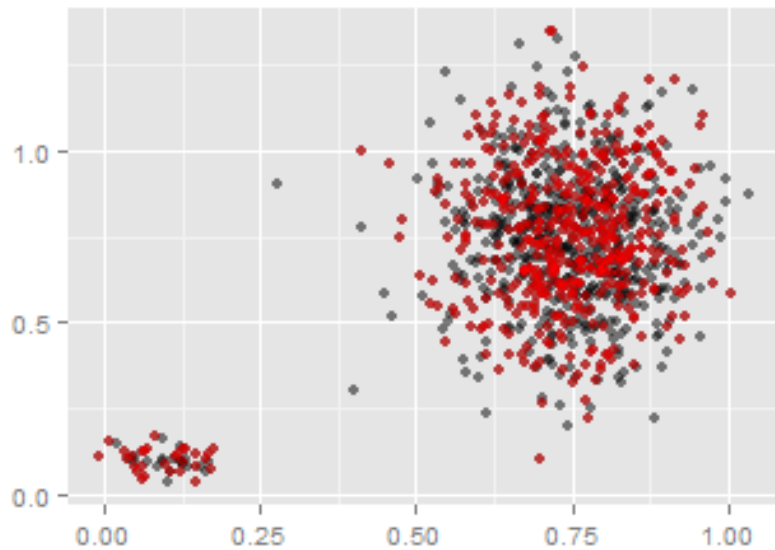


De

ipá

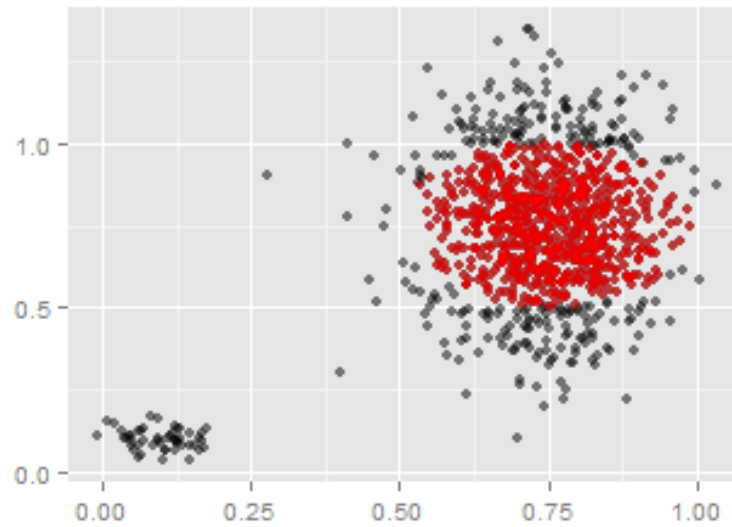
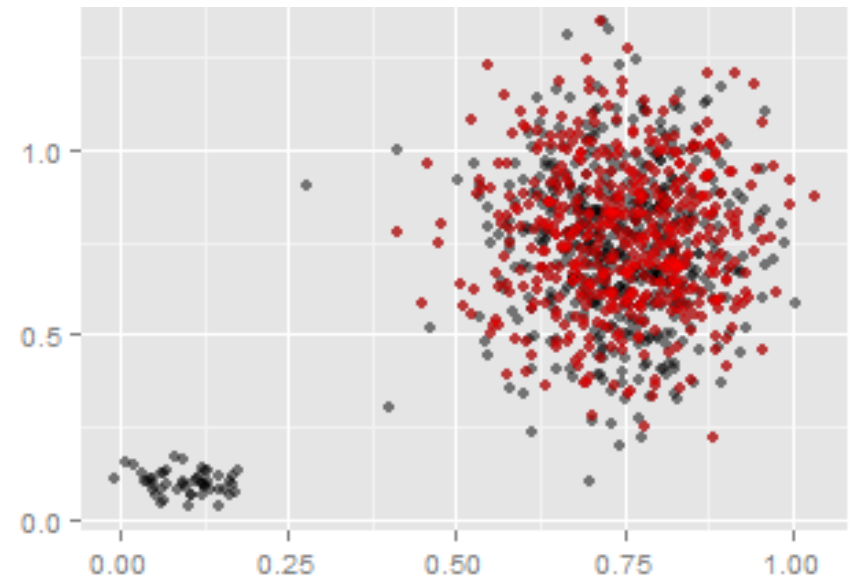


MCD

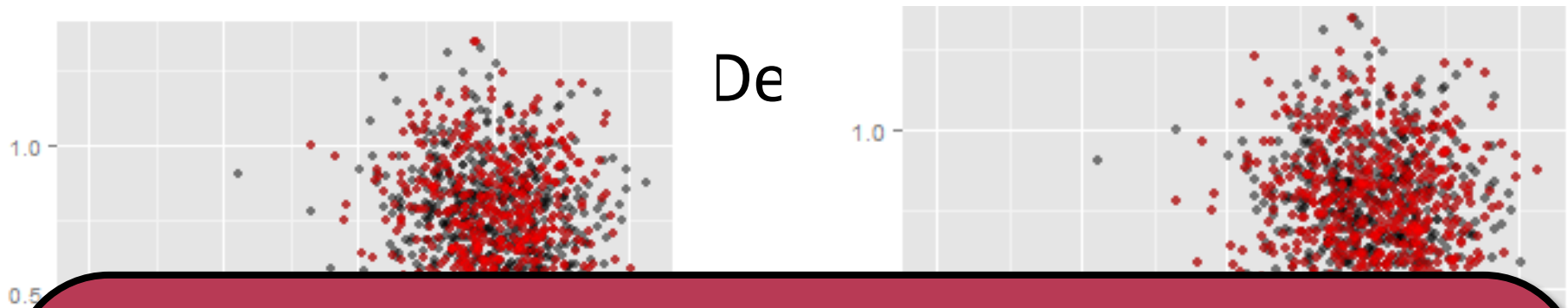


De

ipá

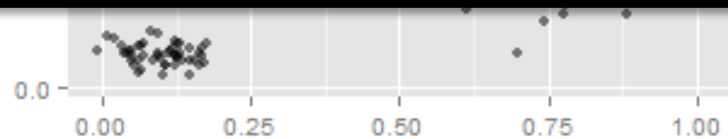


MCD



Kimerítő keresés?

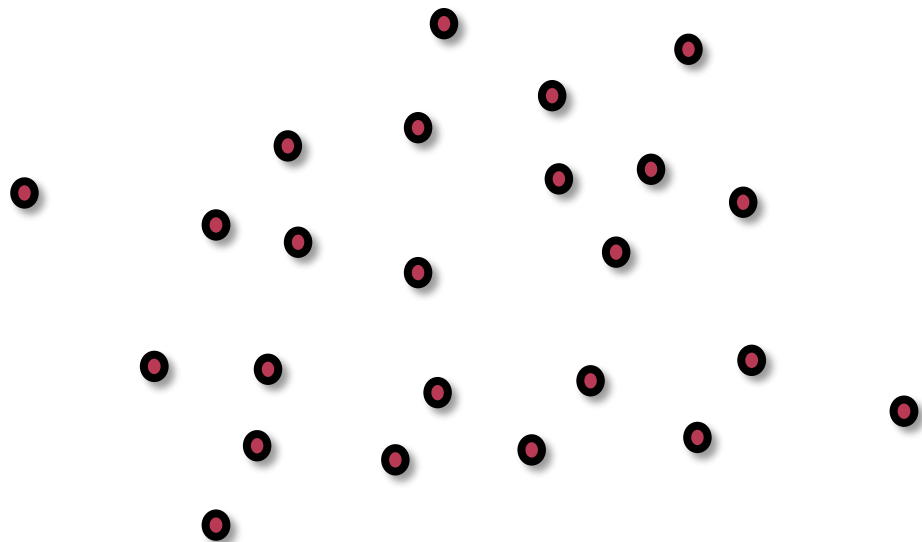
```
choose(n = 1000, k = 900)  
[1] 6.385051e+139
```



0.00011

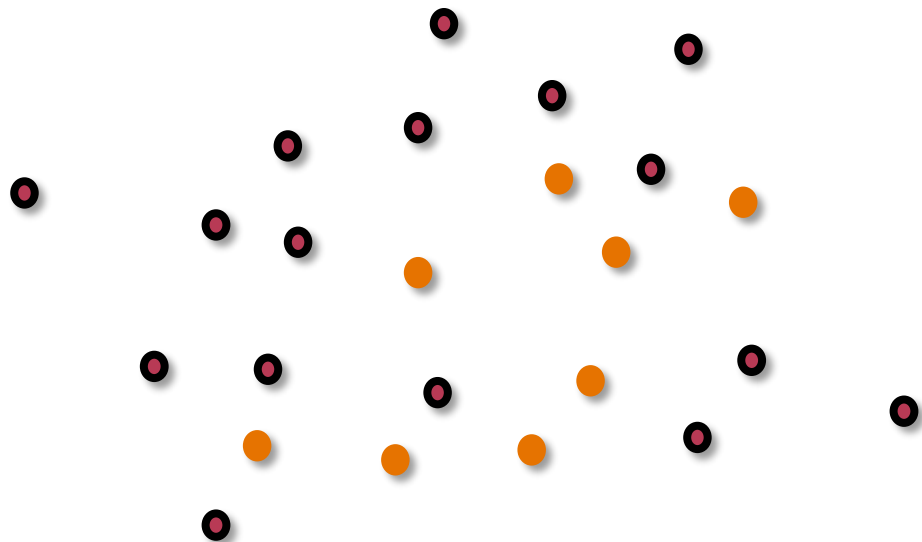
FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív
- Legközelebbi pontok kiválasztása
 - Mahalanobis távolság



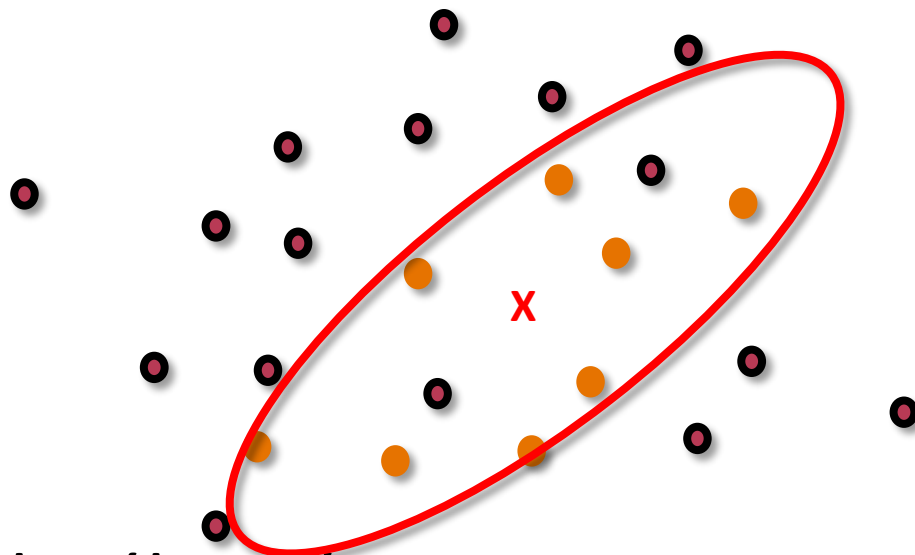
FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív
- Legközelebbi pontok kiválasztása
 - Mahalanobis távolság



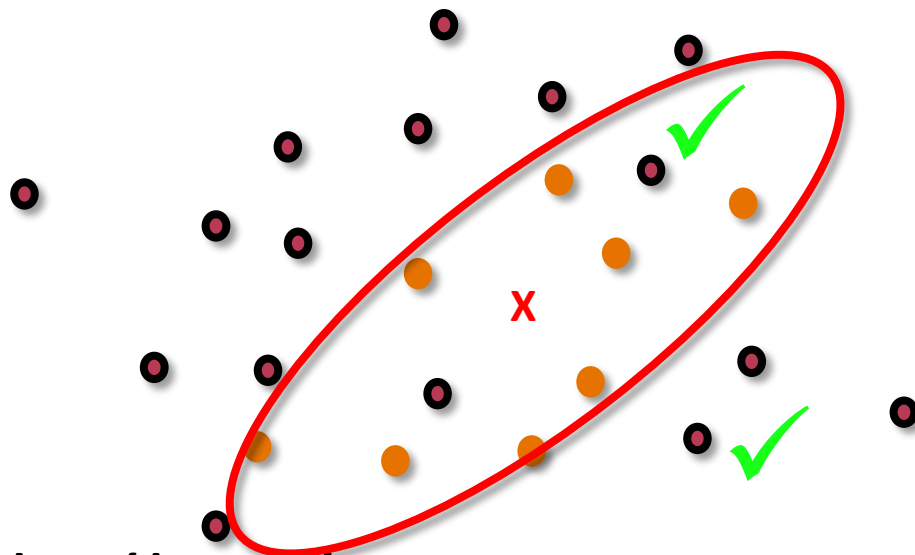
FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív
- Legközelebbi pontok kiválasztása
 - Mahalanobis távolság



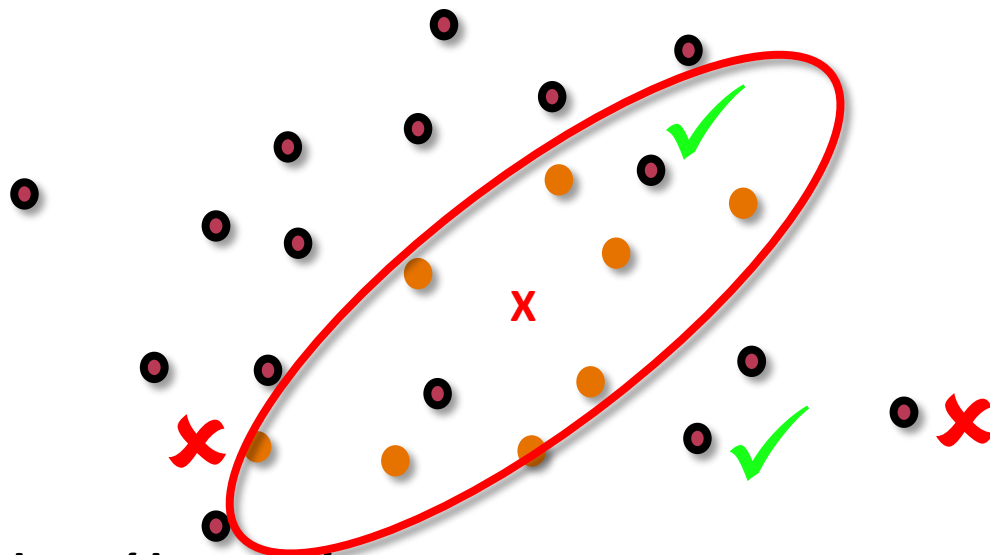
FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív
- Legközelebbi pontok kiválasztása
 - Mahalanobis távolság



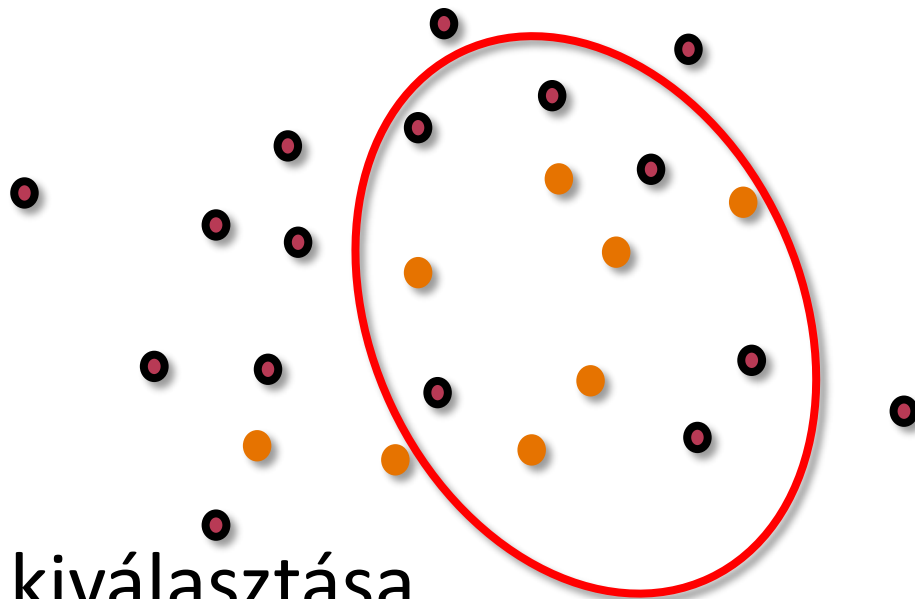
FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív
- Legközelebbi pontok kiválasztása
 - Mahalanobis távolság



FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív
- Legközelebbi pontok kiválasztása
 - Mahalanobis távolság

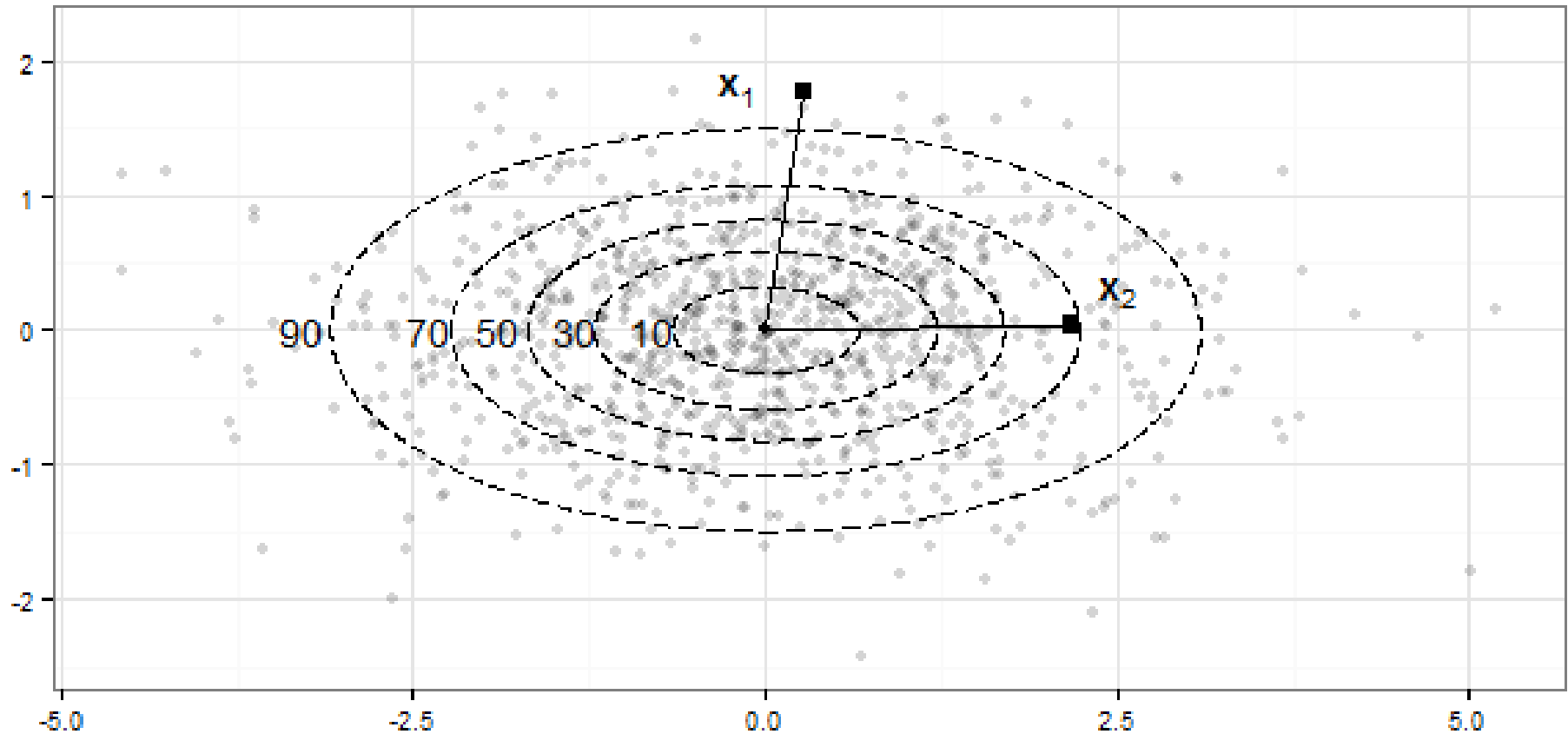


Mahalanobis távolság

- $D(x, M) = \sqrt{(x - \vartheta)^T S^{-1} (x - \vartheta)}$
 - S – kovarianciamátrix
 - ϑ – súlypont

Mahalanobis távolság

- $$D(x, M) = \sqrt{(x - \vartheta)^T S^{-1} (x - \vartheta)}$$



- Blocked Adaptive Computationally Efficient Outlier Nominators
- Kiinduló halmaz félig felügyelt módon is!
- Új halmaz: küszöbérték alapján
- Paraméterezés
 - *init.sel* kezdőhalmaz
 - „manual” – *man.sel* kezdőhalmaz
 - „Mahalanobis”, „dUniMedian” – *m* kezdőhalmaz mérete
 - α paraméter a küszöbérték befolyásolására

TELJESÍTMÉNYMENEDZSMENT

Mérési környezet

- Apache VCL

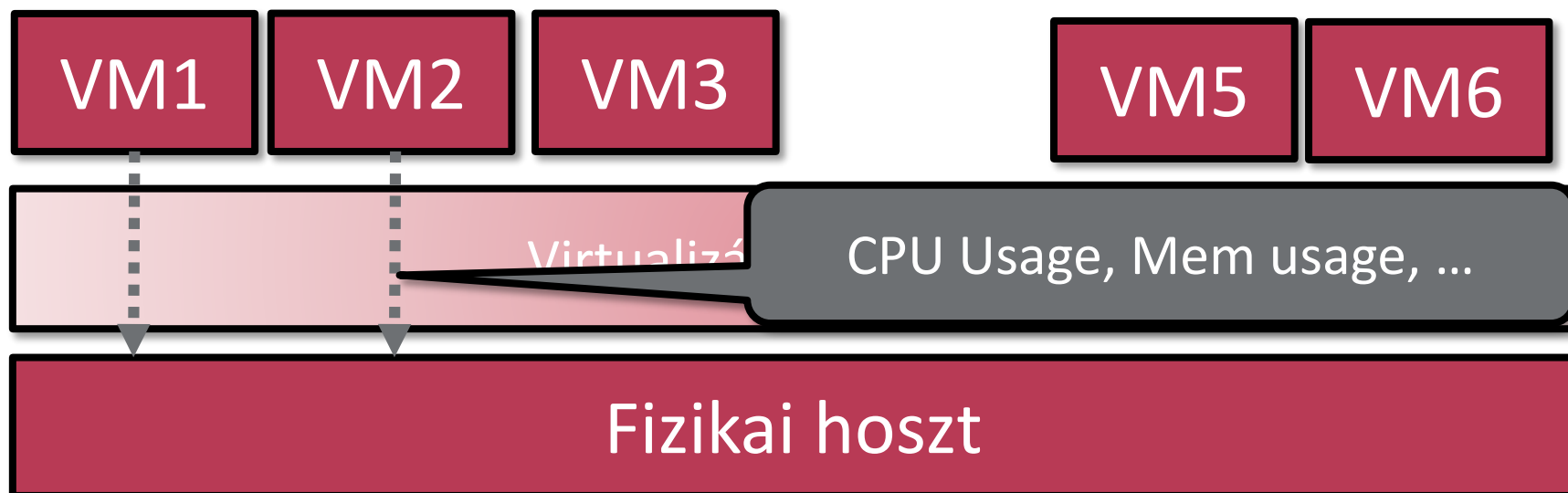
- Oktatási felhő
- 10 fizikai hoszt
- 38 tárolt platform szintű metrika
- ≈ 7 millió adatpont (inkább ≈ 1 millió...)



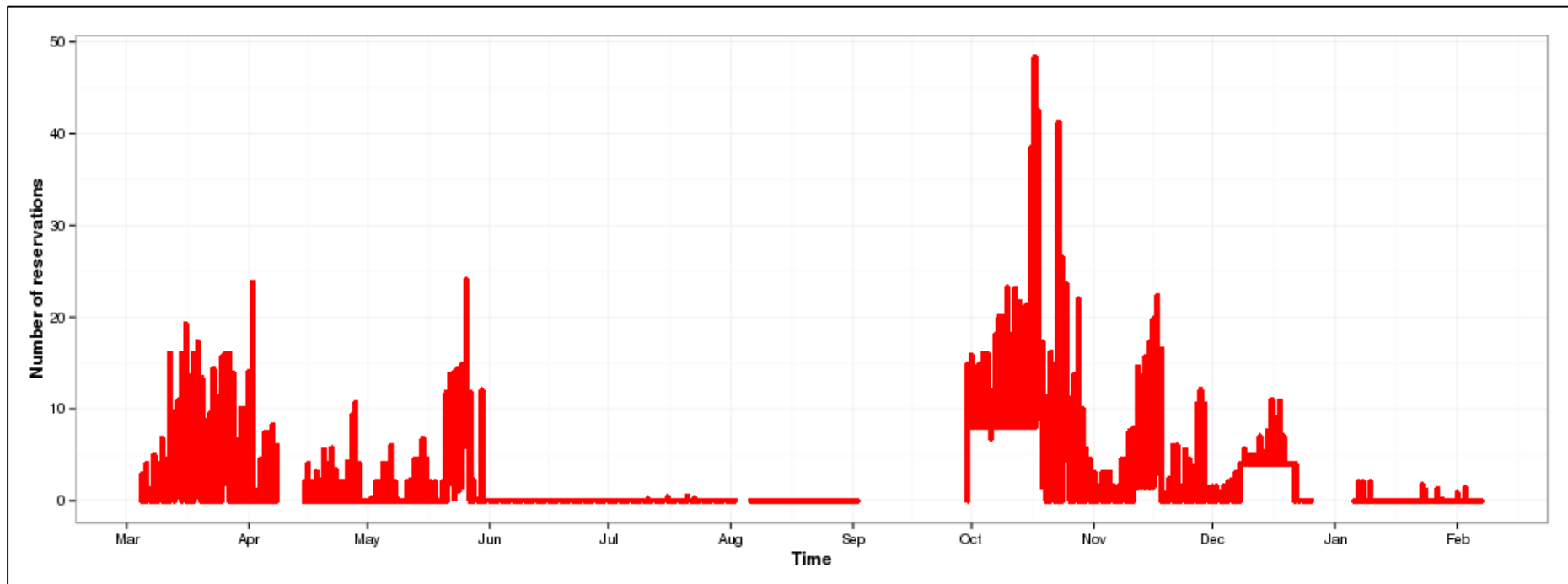
Mérési környezet

- Apache VCL

- Oktatási felhő
- 10 fizikai hoszt
- 38 tárolt platform szintű metrika
- ≈ 7 millió adatpont (inkább ≈ 1 millió...)



Előfeldolgozás: terhelési tartományok



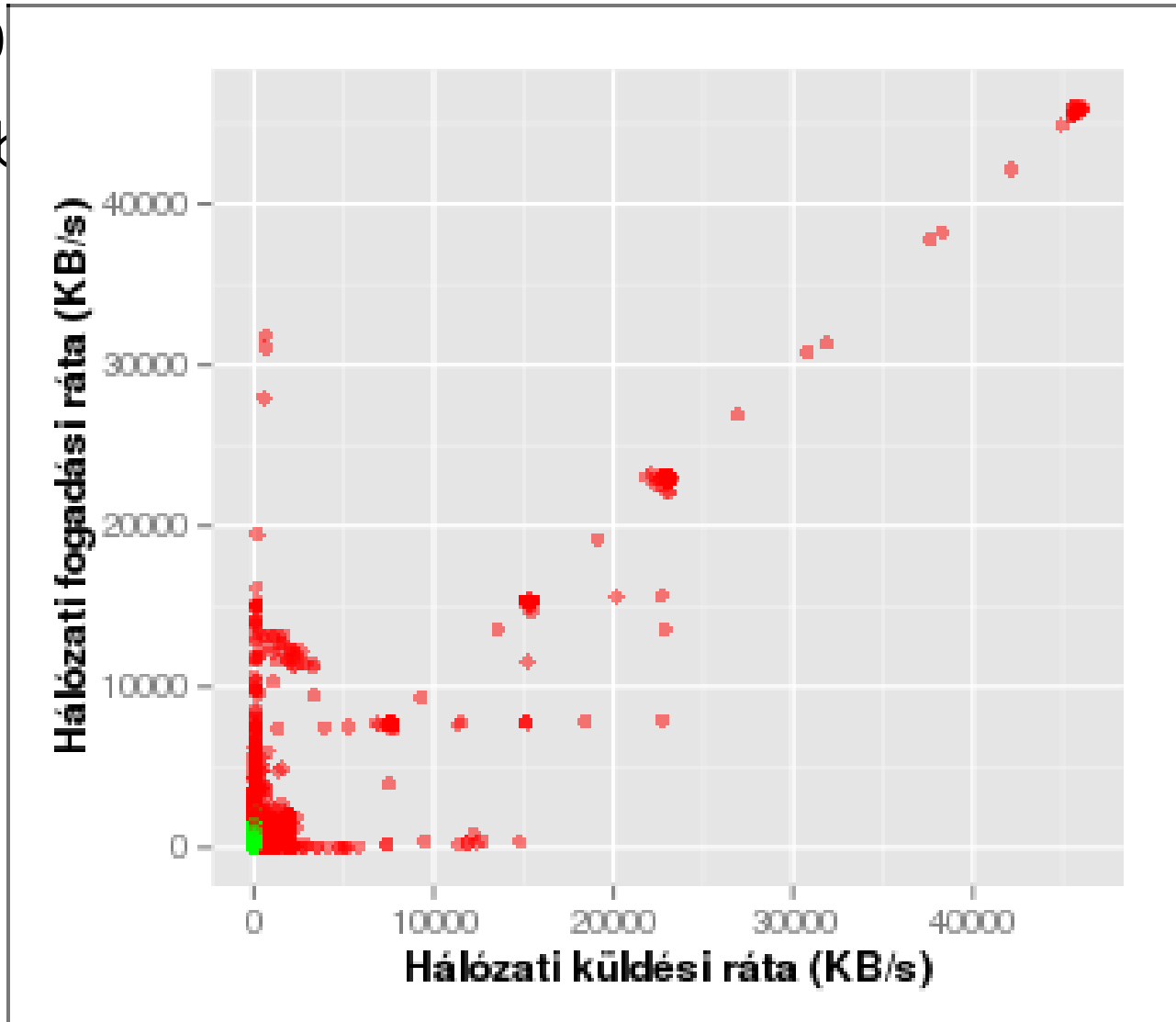
Mérési környezet

- BACON
 - Előbb nem felügyelt módon...

Mérési környezet

- BACO

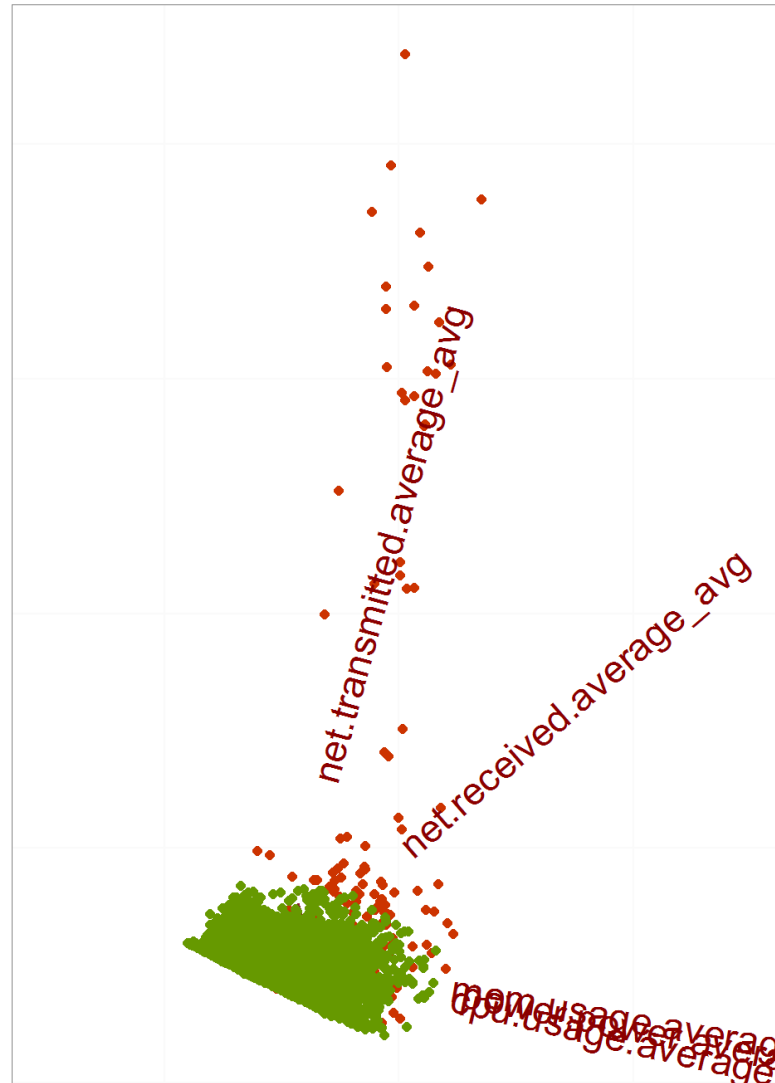
○ Elők



Mérési környezet

■ BACC

○ Elő

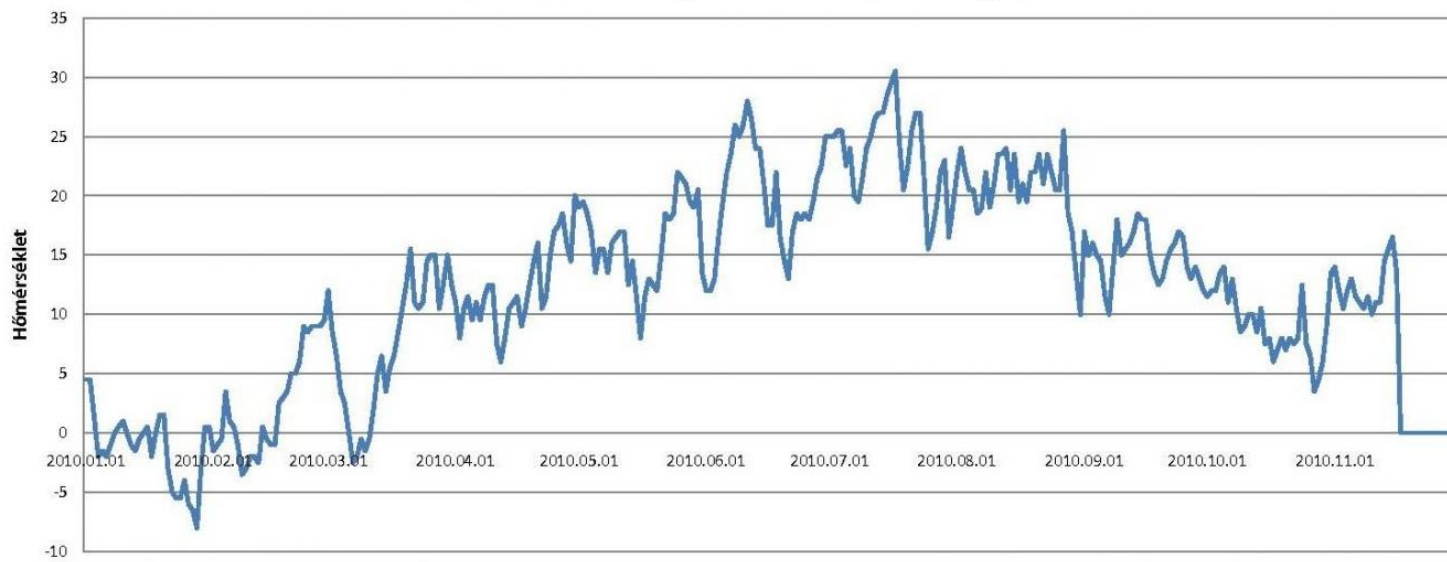


Mire nem jó? Kontextusanomáliák

■ Viselkedési

■ Kontextus

Napi átlag hőmérséglet 2010-ben (Zala megye)

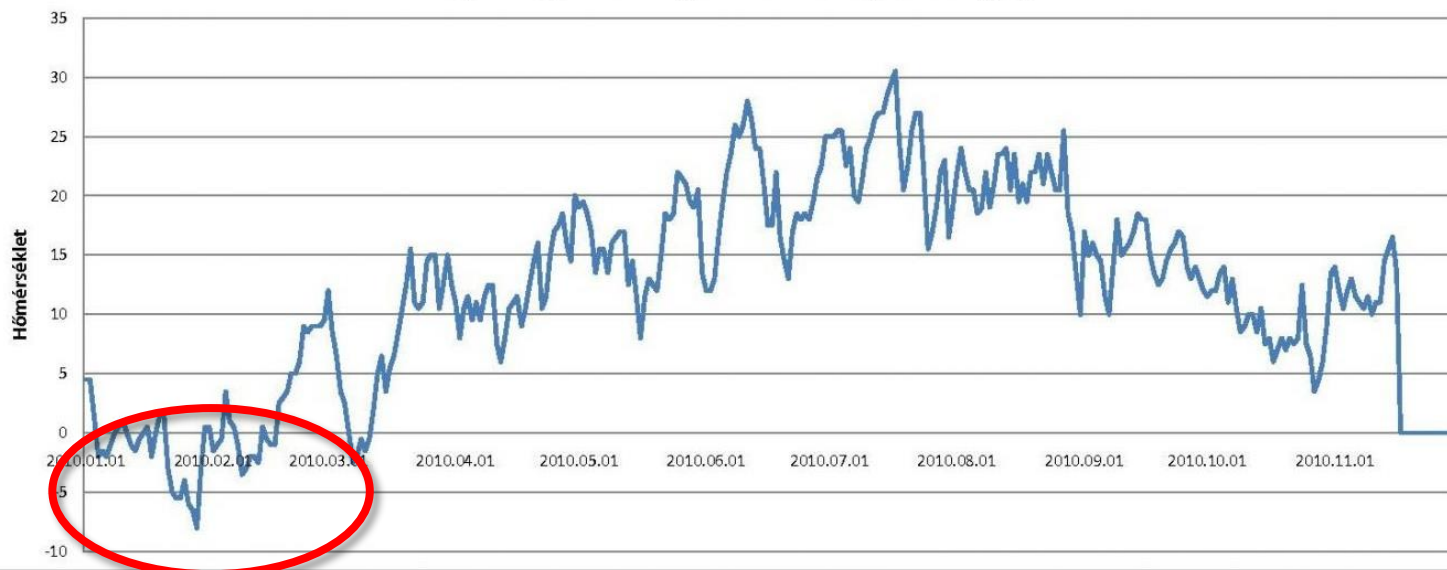


Mire nem jó? Kontextusanomáliák

■ Viselkedési

■ Kontextus

Napi átlag hőmérséglet 2010-ben (Zala megye)

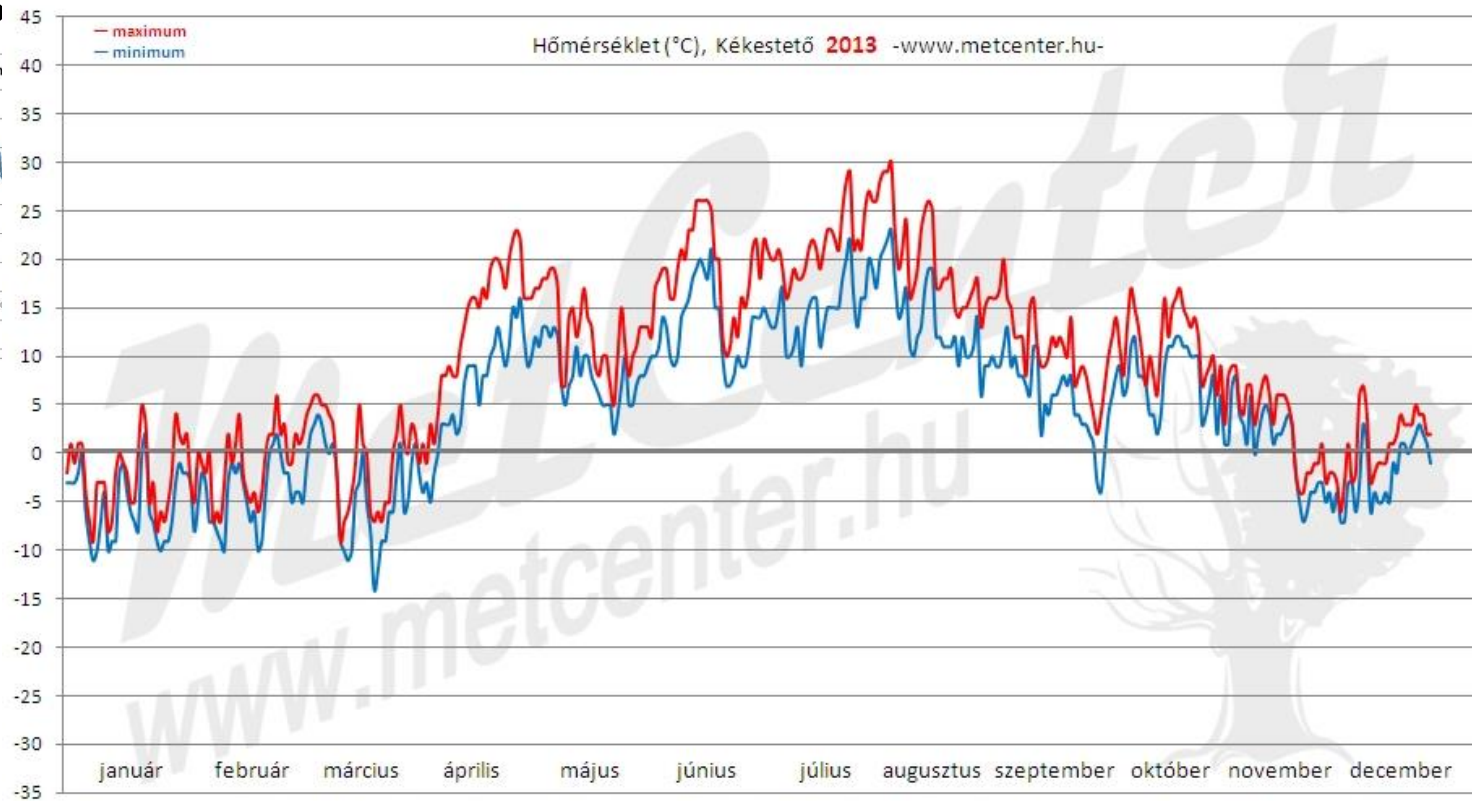


Mire nem jó? Kontextusanomáliák

■ Viselkedés:



■ Kontextus

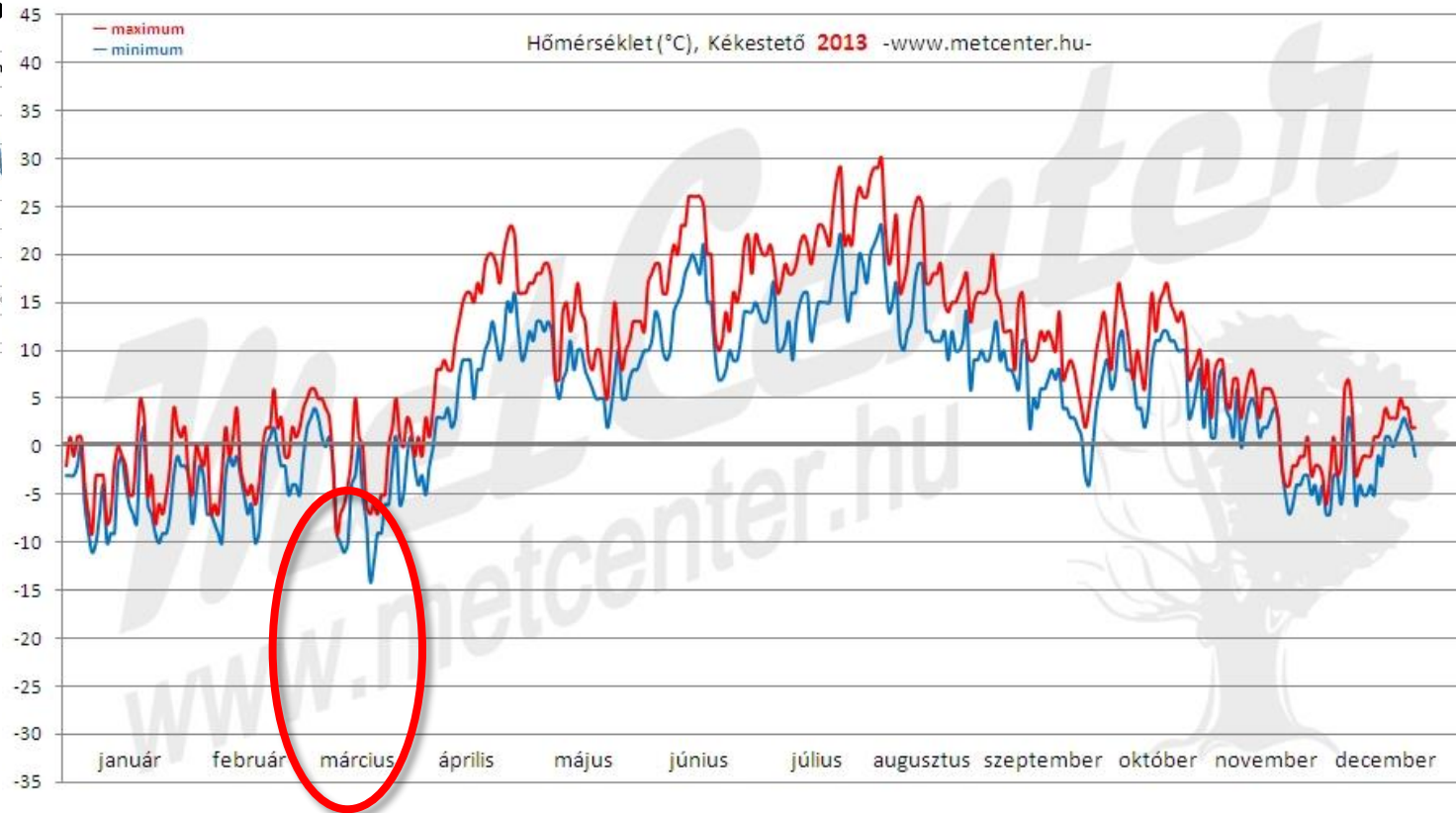


Mire nem jó? Kontextusanomáliák

■ Viselkedés:



■ Kontextus



Ahol kezdeni érdeemes

- Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR)* 41.3 (2009): 15.
 - A bővített változatban idősor-elemzés is!
- Billor et al. "BACON: blocked adaptive computationally efficient outlier nominators." *Computational Statistics & Data Analysis* 34.3 (2000): 279-298.
- Intelligens adatelemzés tankönyv:
<http://www.tyopotex.hu/index.php?page=BME-SOTE>