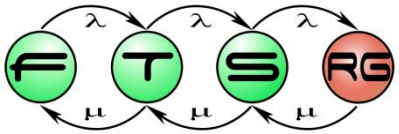
Outlierdetektálás

Salánki Ágnes

salanki@mit.bme.hu

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Definíció

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

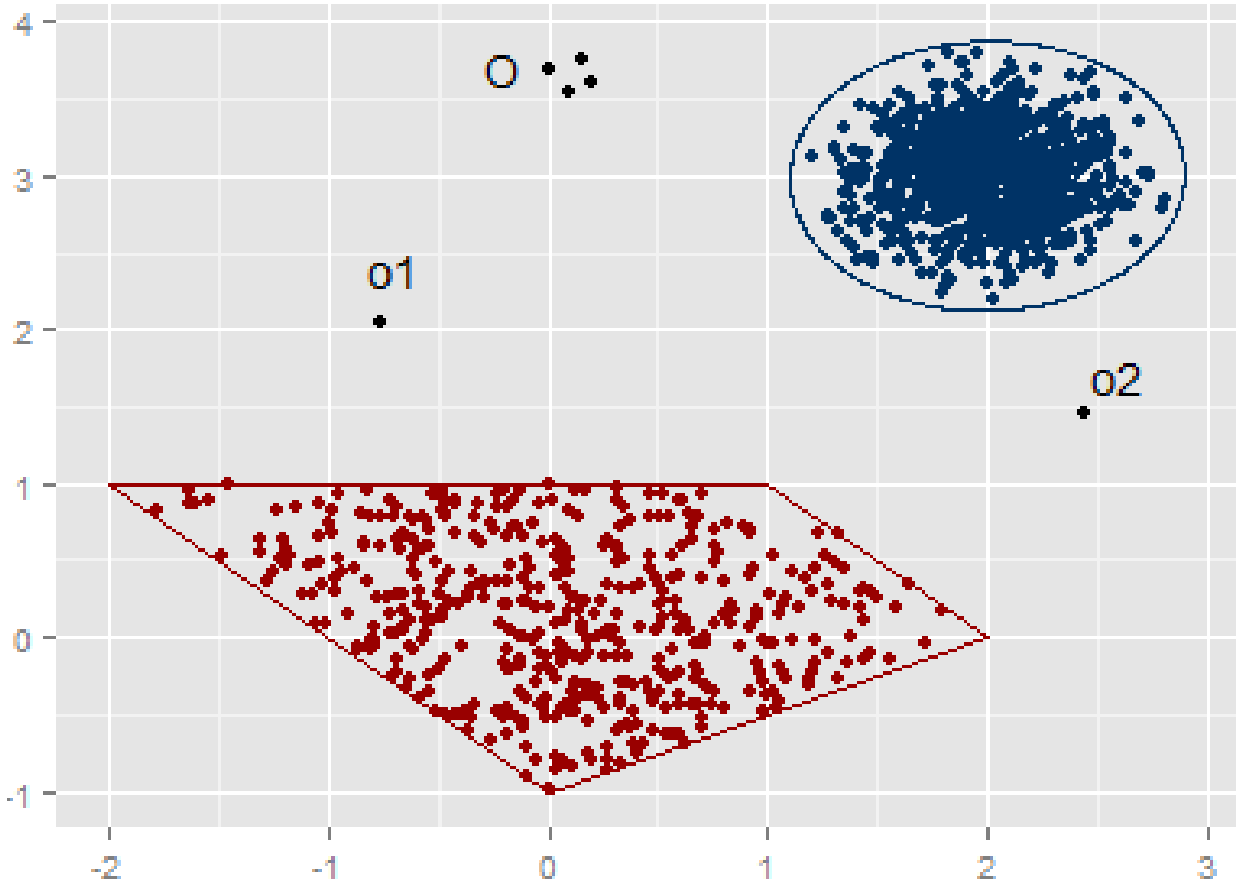
Definíció

- Kevés van belőlük
- „Gyanús”, hogy **más** a generáló folyamat/forrás
  - Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива **по-своему**.
  - Happy families are all alike;
    every unhappy family is unhappy **in its own way**.
  - A boldog családok mind hasonlók
    egymáshoz, minden boldogtalan család **a maga módján** az.

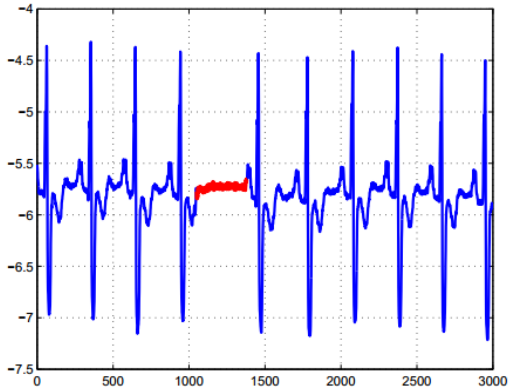
(Tolsztoj: Anna Karenina)

Pont- és kollektív anomála

- Pontanomália

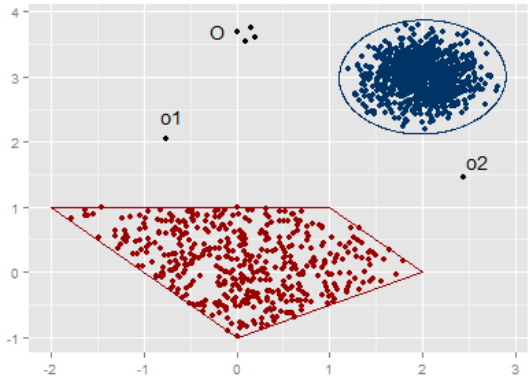


- Kollektív anomália

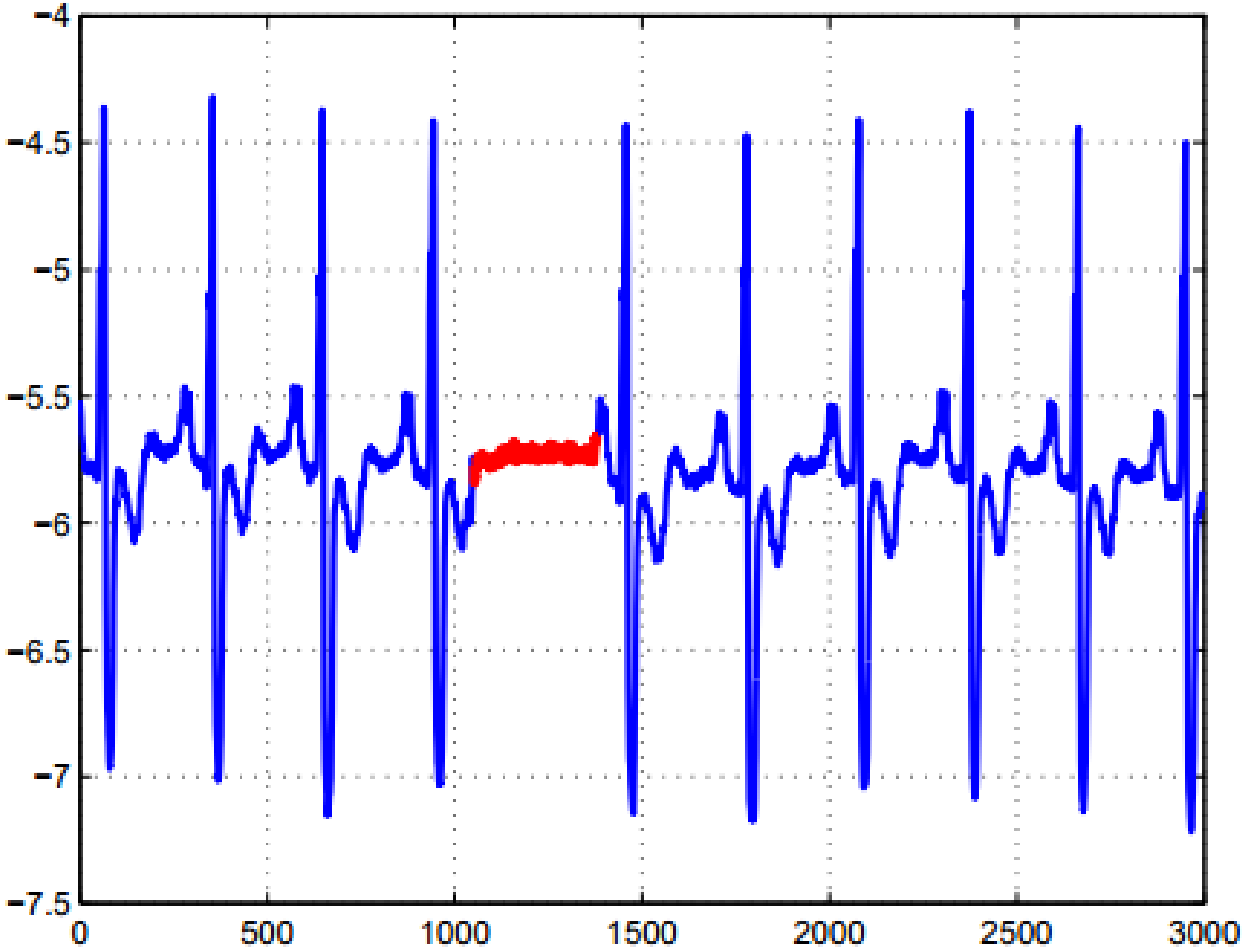


Pont- és kollektív anomála

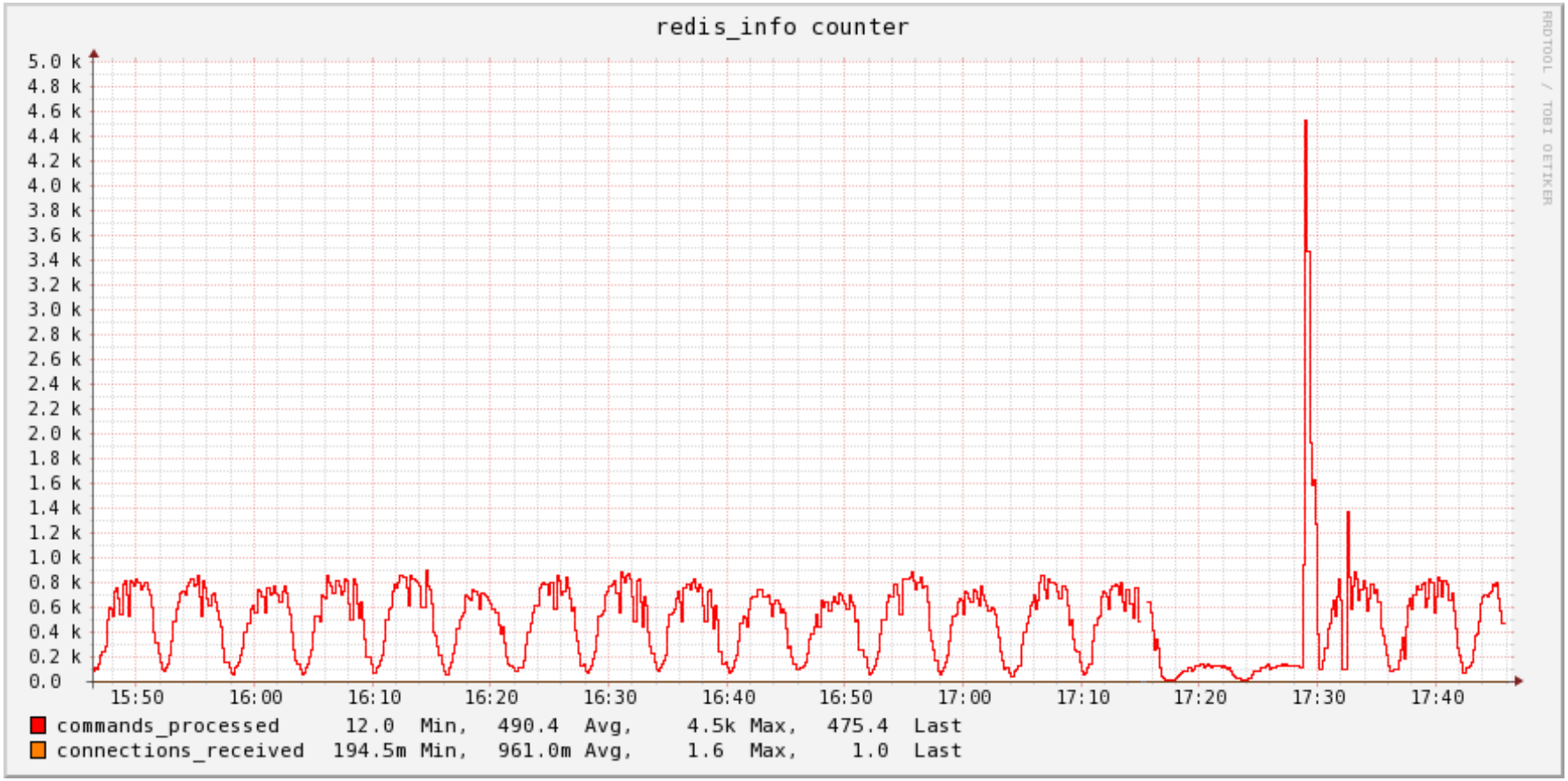
- Pontanomália



- Kollektív anomália

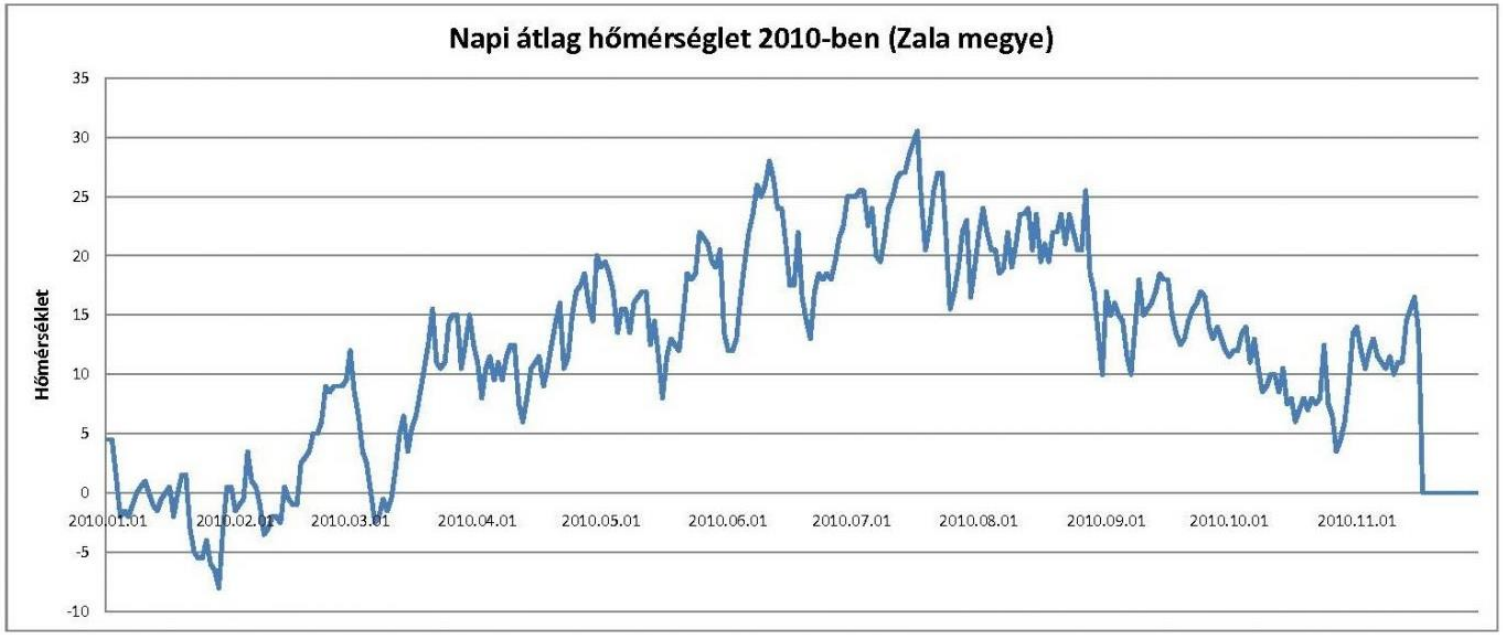


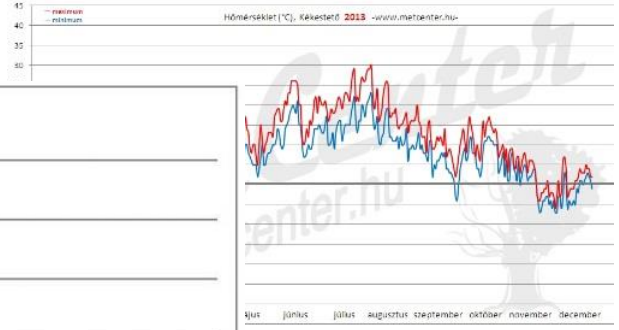
Pont- és kollektív anomália



Viselkedési és kontextusanomália

- Viselkedési
- Kontextus

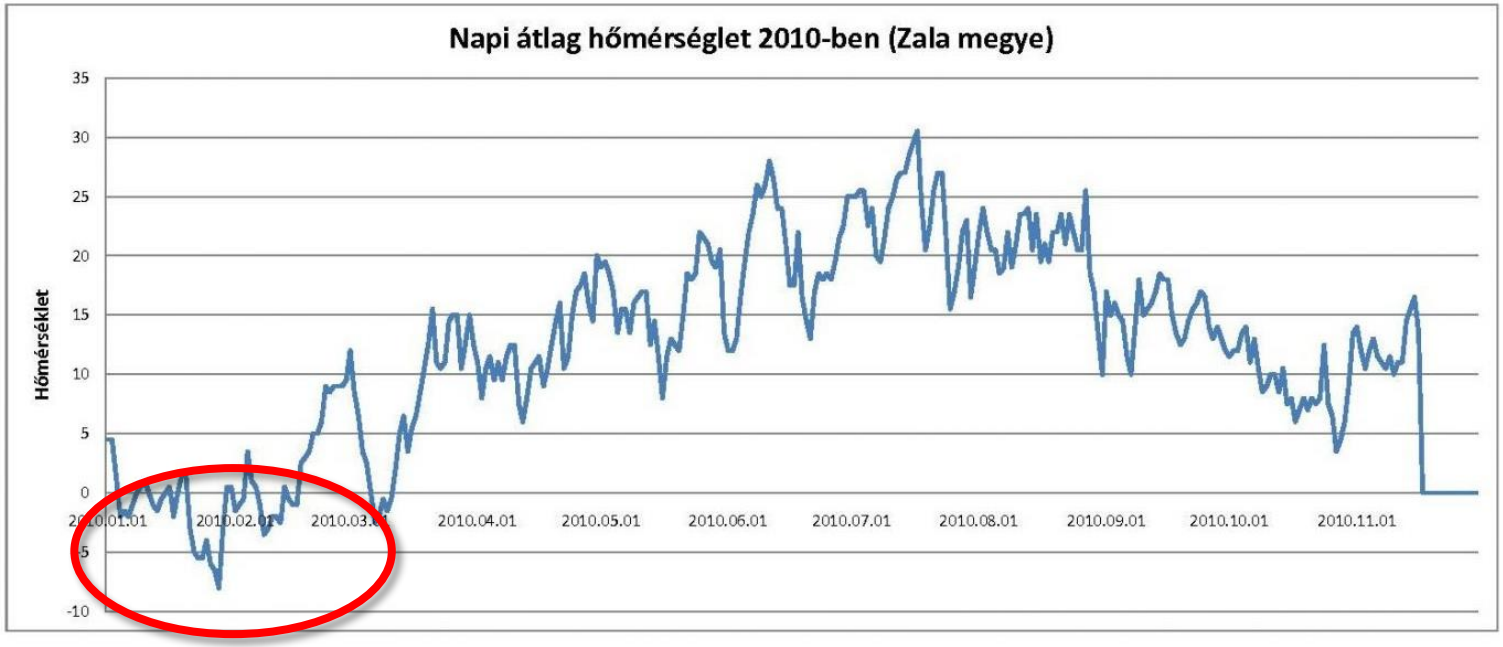


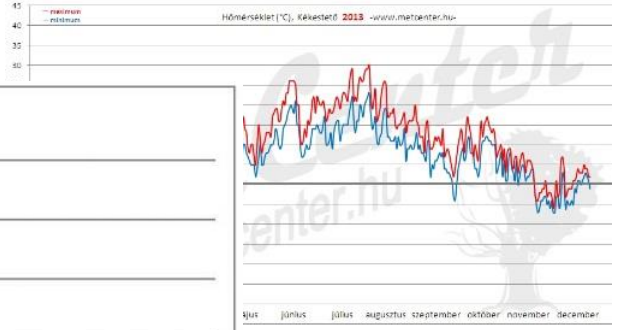


Ábrák forrása: http://www.baki-agrocentrum.hu/hu/szakmai-anyagok/agronomiai-informaciok/az-eghajlatvaltozas-zala-megye-szivebol-nezve
http://www.metcenter.hu

Viselkedési és kontextusanomália

- Viselkedési
- Kontextus

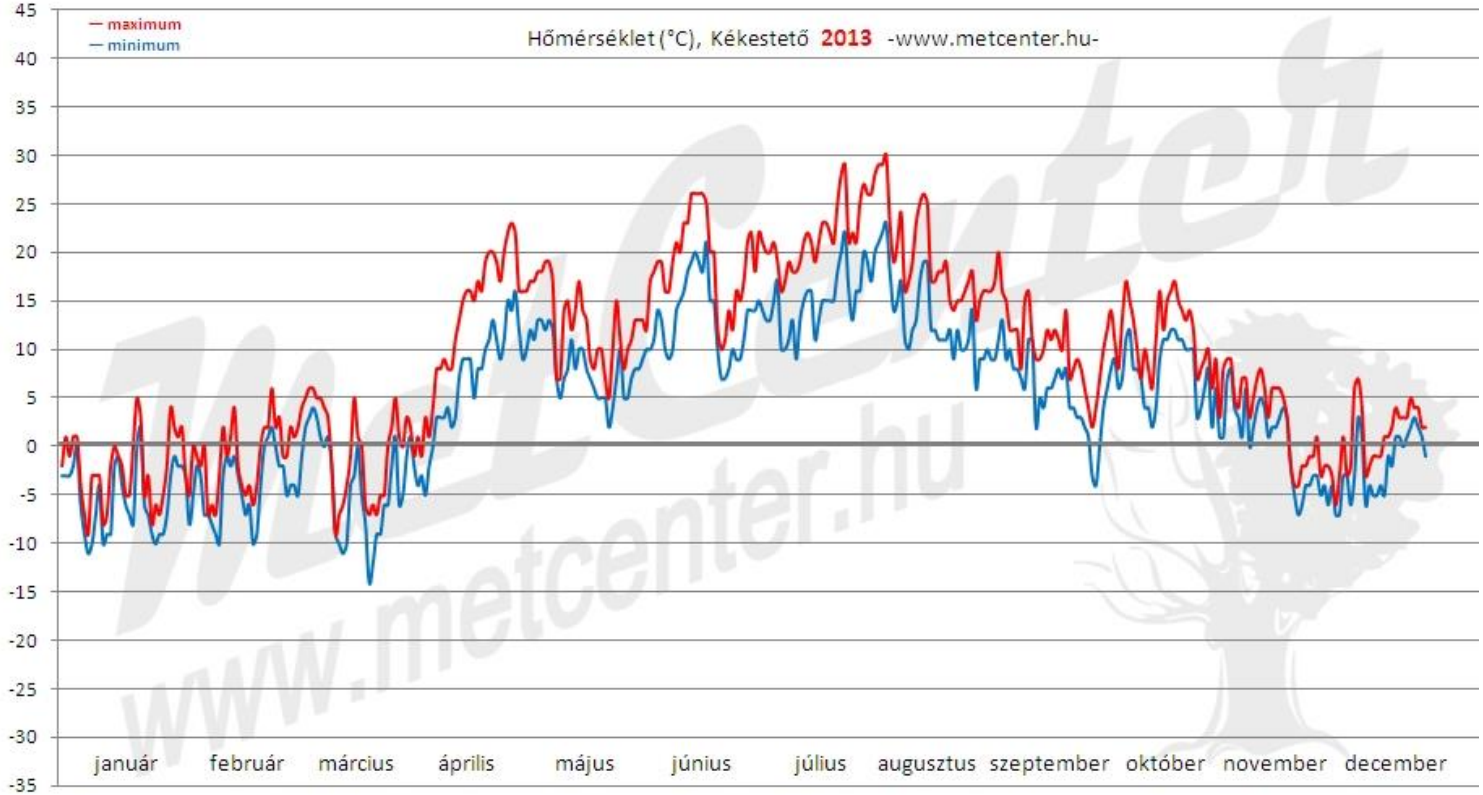




Ábrák forrása: http://www.baki-agrocentrum.hu/hu/szakmai-anyagok/agronomiai-informaciok/az-eghajlatvaltozas-zala-megye-szivebol-nezve
http://www.metcenter.hu

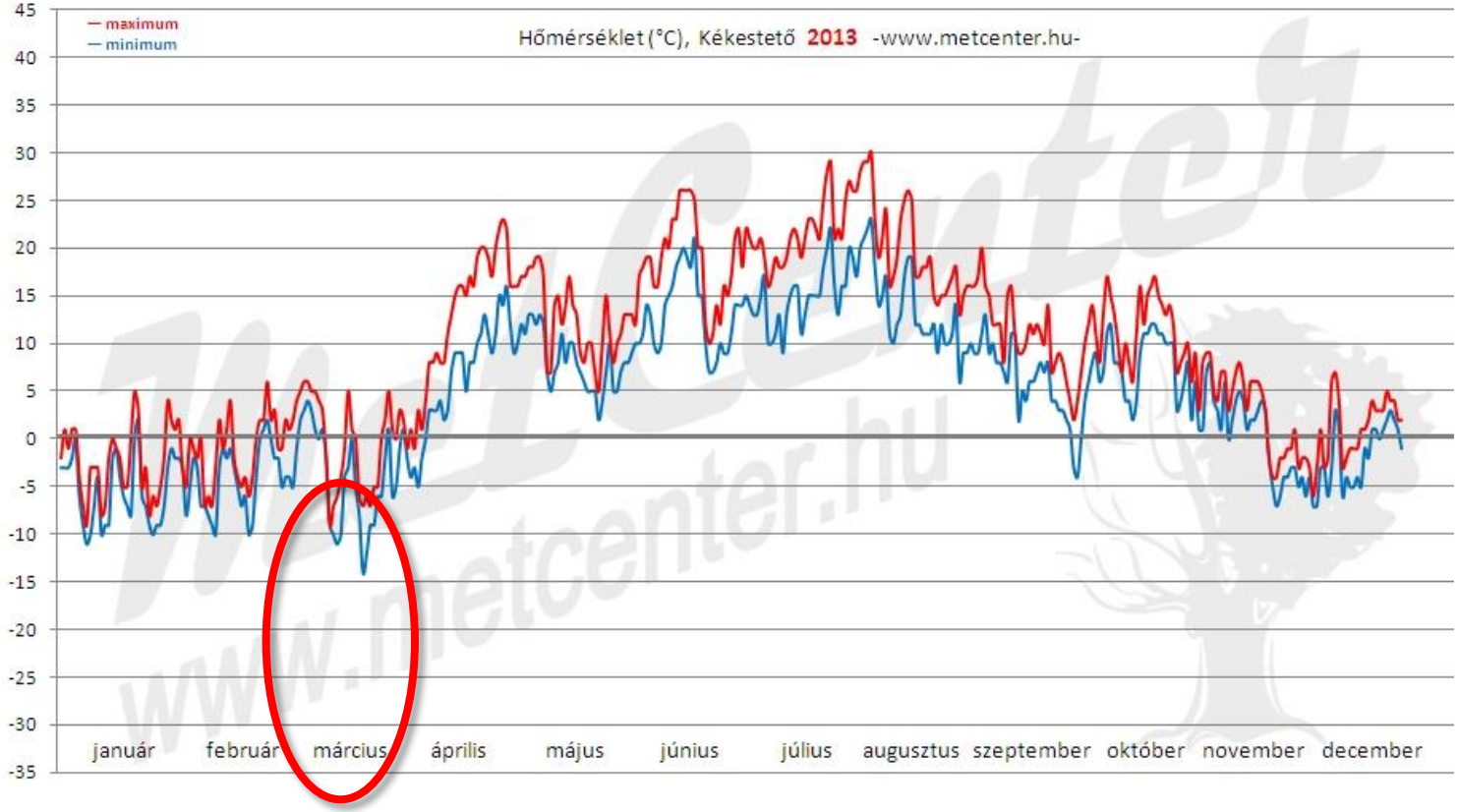
Viselkedési és kontextusanomália

- Viselkedési
- Kontextus



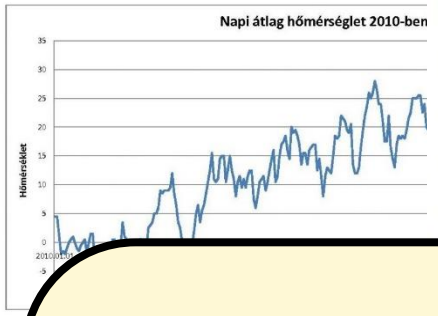
Viselkedési és kontextusanomália

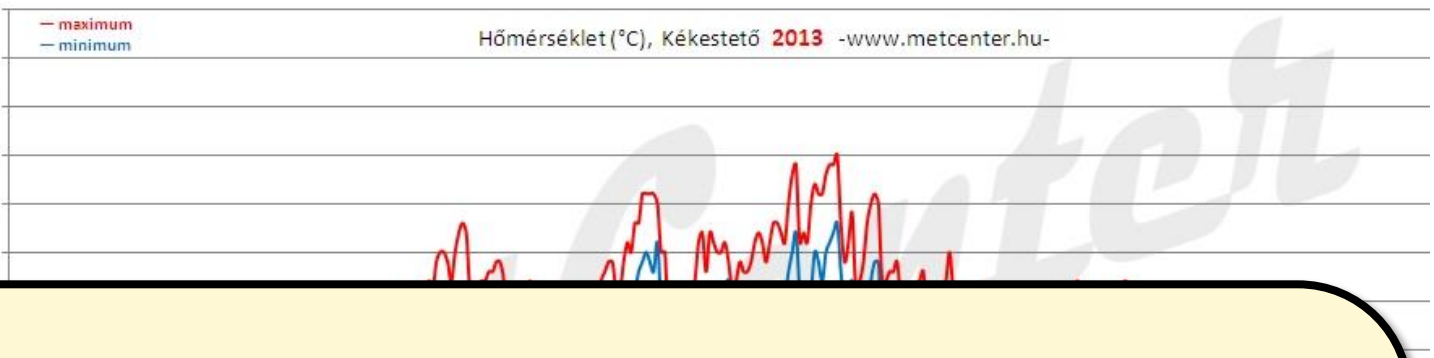
- Viselkedési
- Kontextus



Viselkedési és kontextusanomália

- Viselkedési
- Kontextus



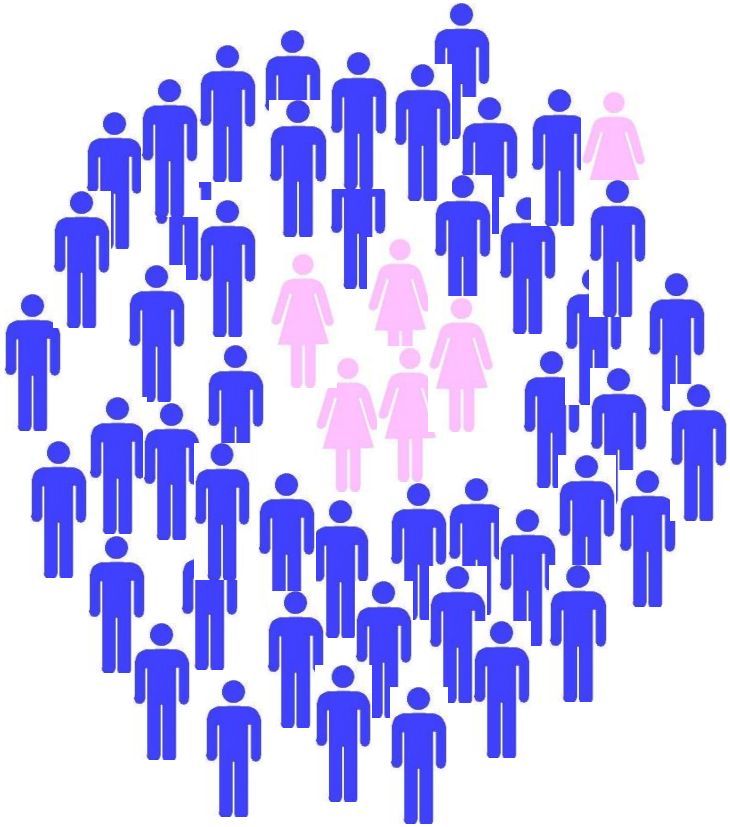


Itt: viselkedési és pontanomáliák

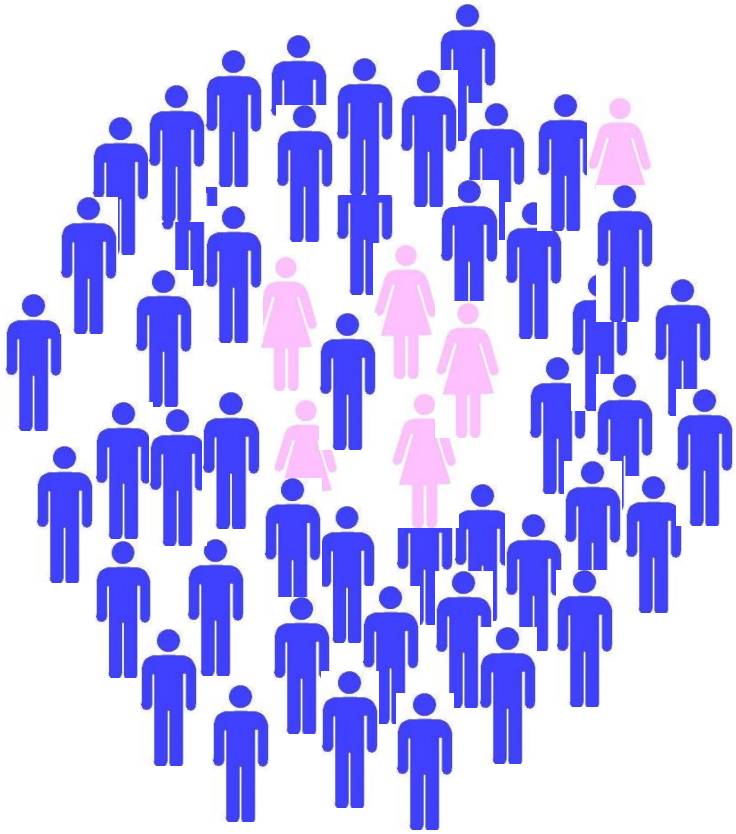
Ábrák forrása: http://www.baki-agrocentrum.hu/hu/szakmai-anyagok/agronomiai-informaciok/az-eghajlatvaltozas-zala-megye-szivebol-nezve
http://www.metcenter.hu

Megközelítések

- Globális outlierek

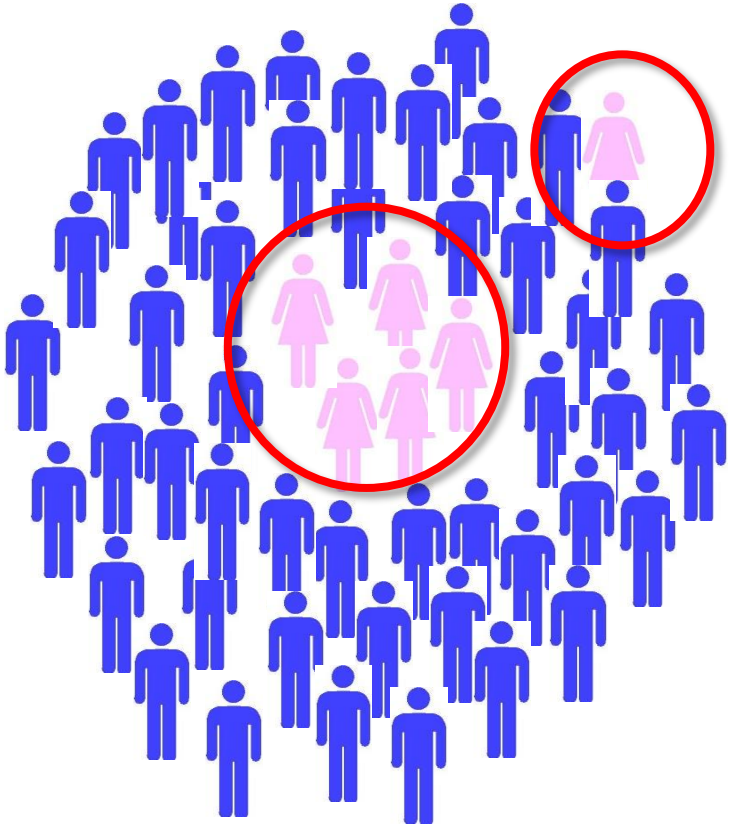


- Lokális outlierek

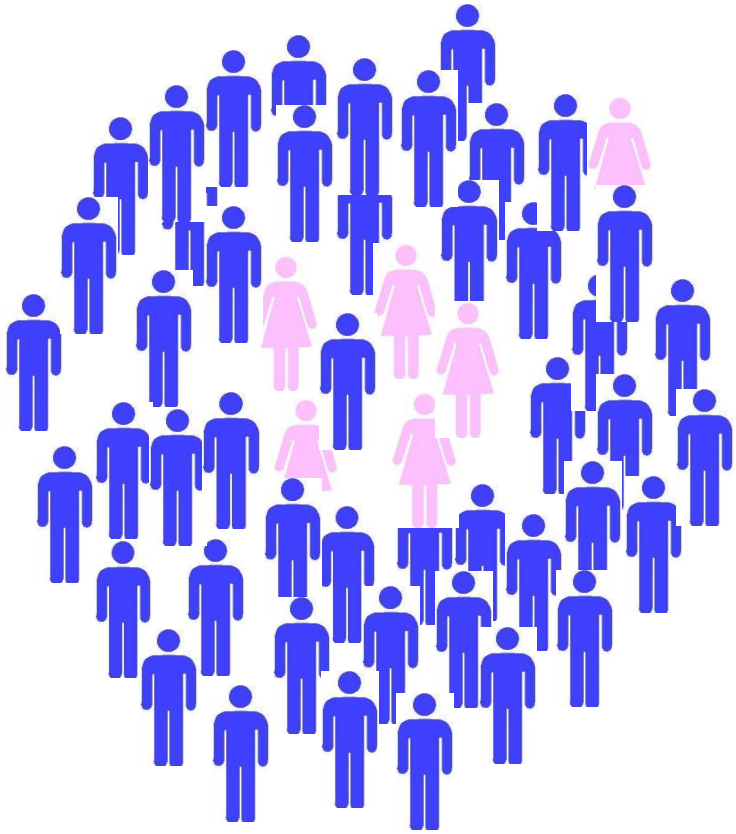


Megközelítések

- Globális outlierek

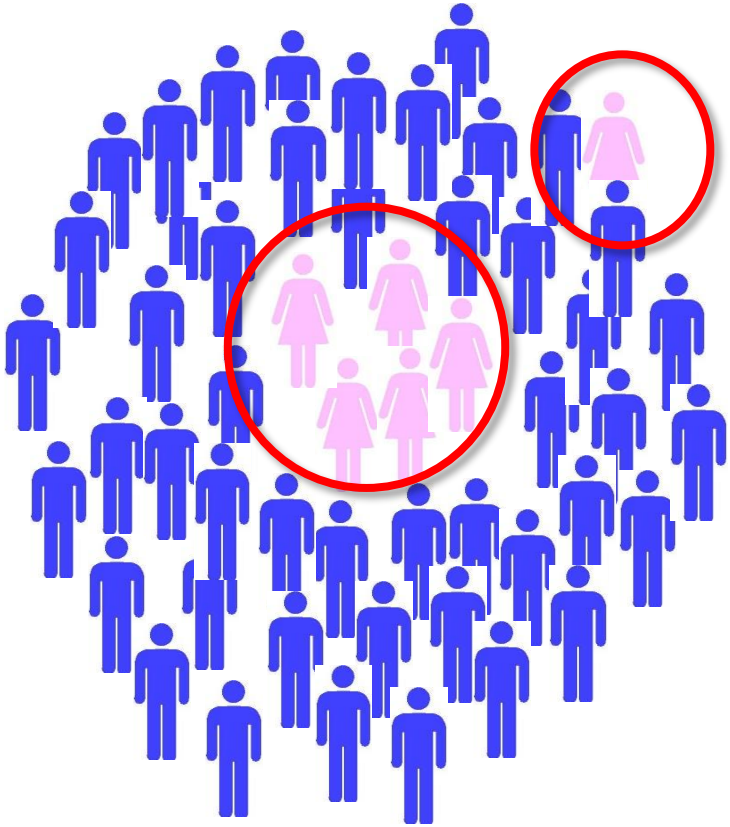


- Lokális outlierek

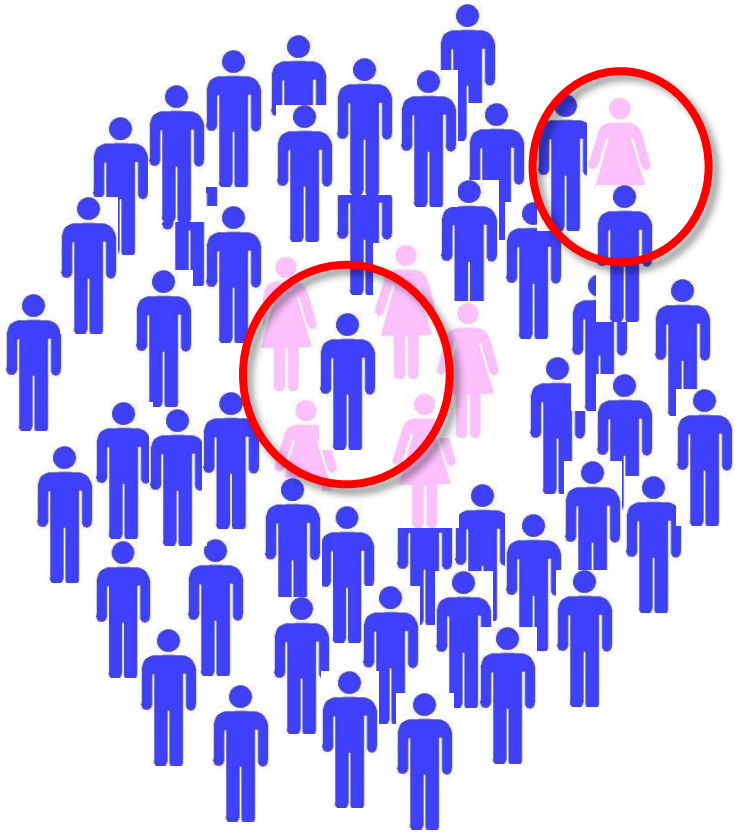


Megközelítések

- Globális outlierek

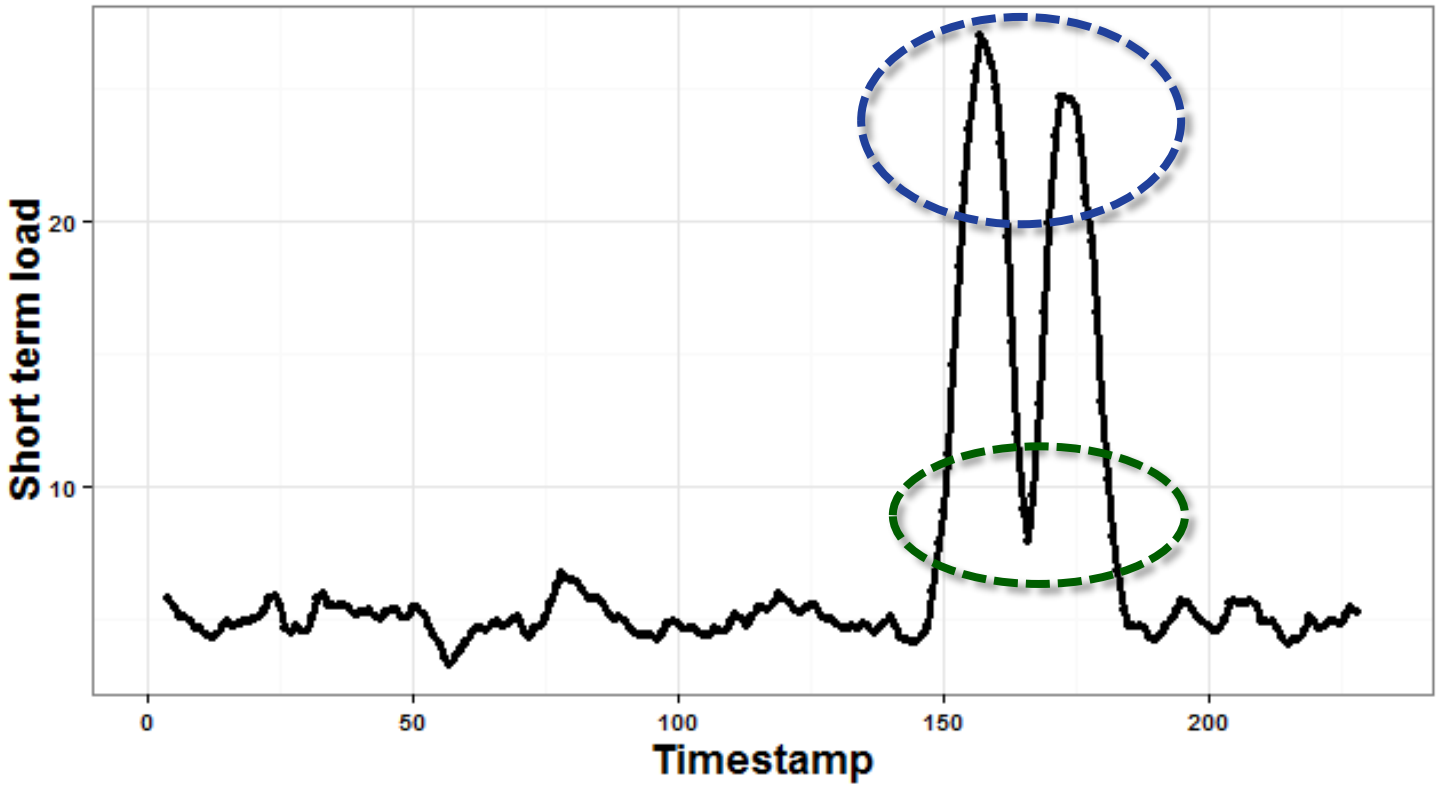


- Lokális outlierek



Megközelítések

- Globális outlierek
- Lokális outlierek

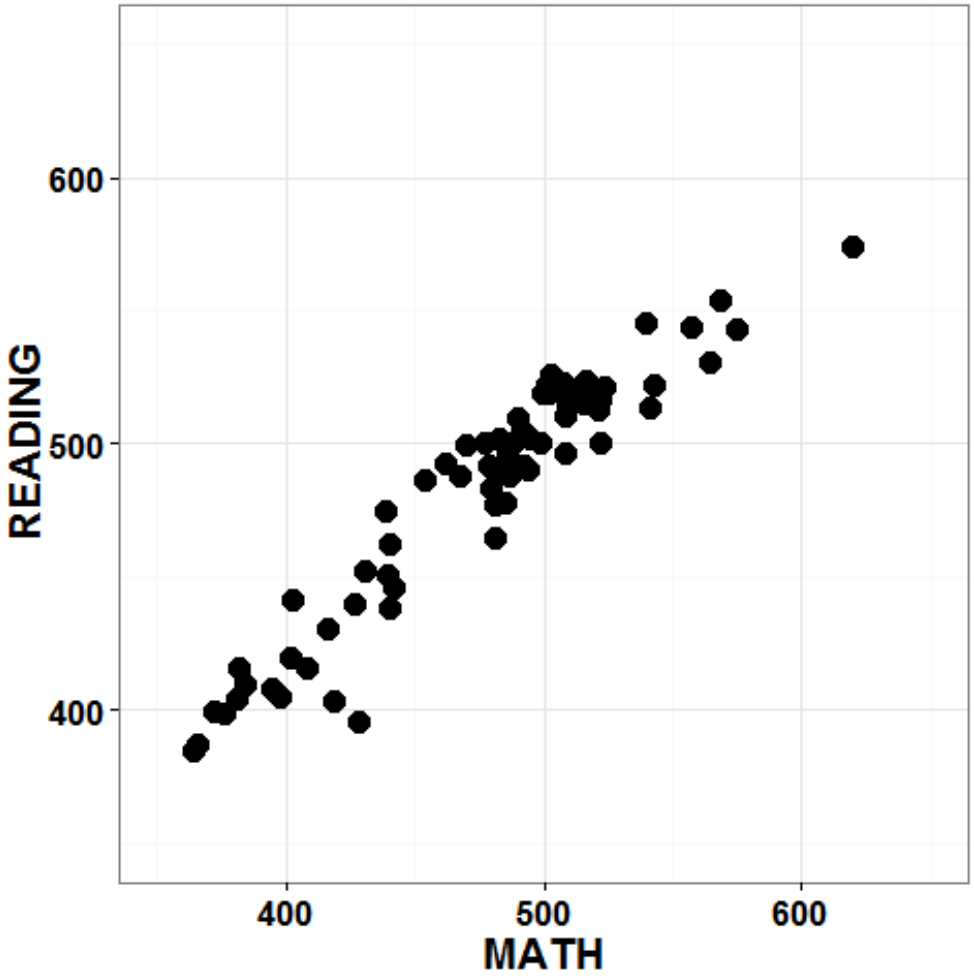


Esettanulmány: PISA 2012

- PISA 2012 results
  - Most: matematika és értő olvasás eredmények

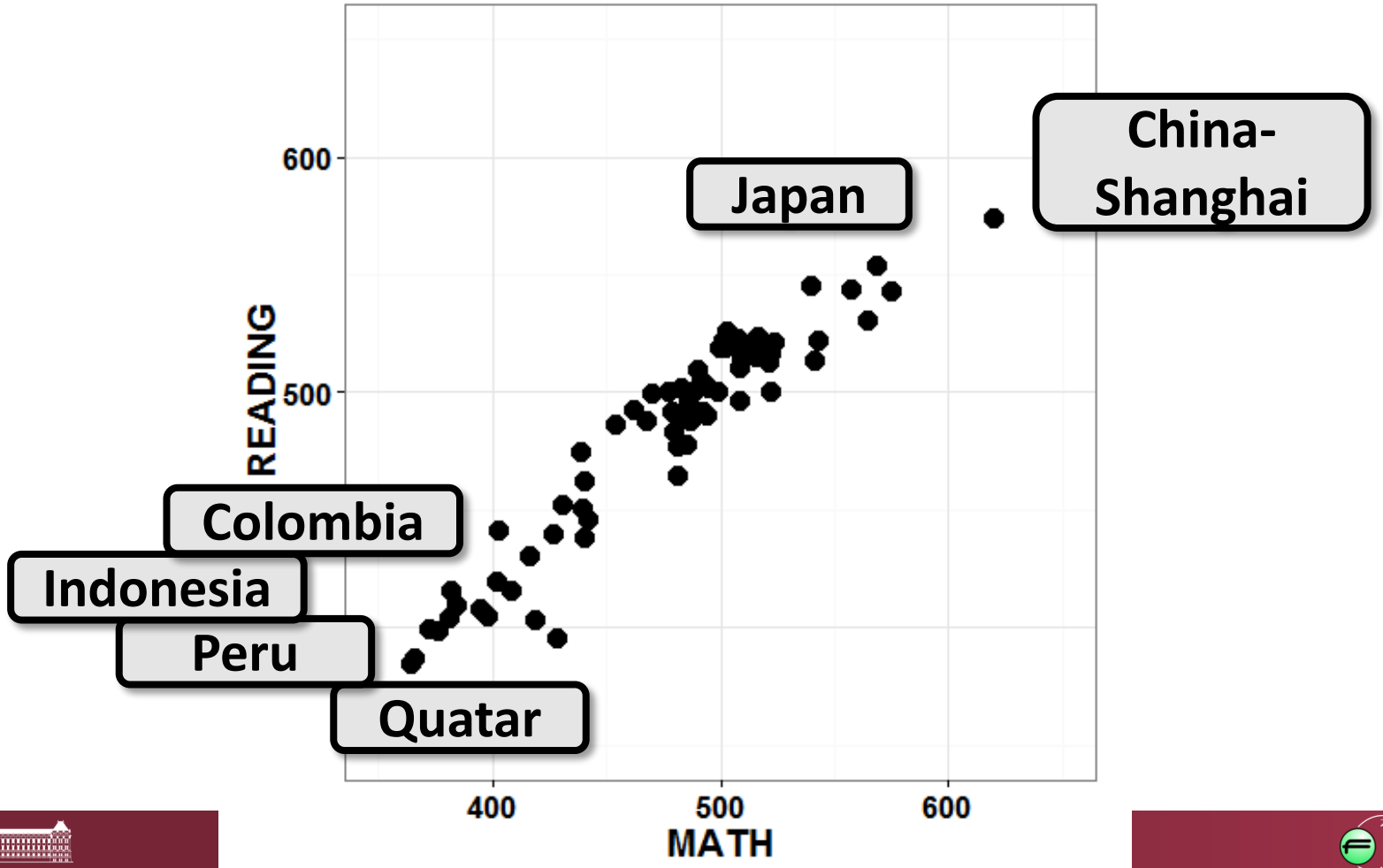
Esettanulmány: PISA 2012

- PISA 2012 results
  - Most: matematika és értő olvasás eredmények



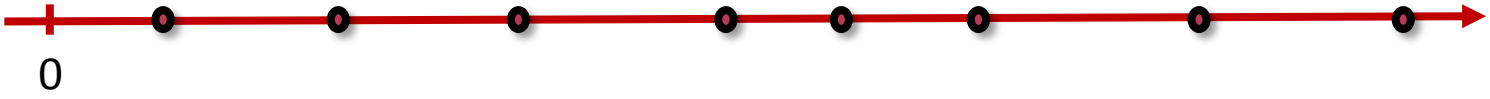
Esettanulmány: PISA 2012

- PISA 2012 results
  - Most: matematika és értő olvasás eredmények



Befoglaló burok

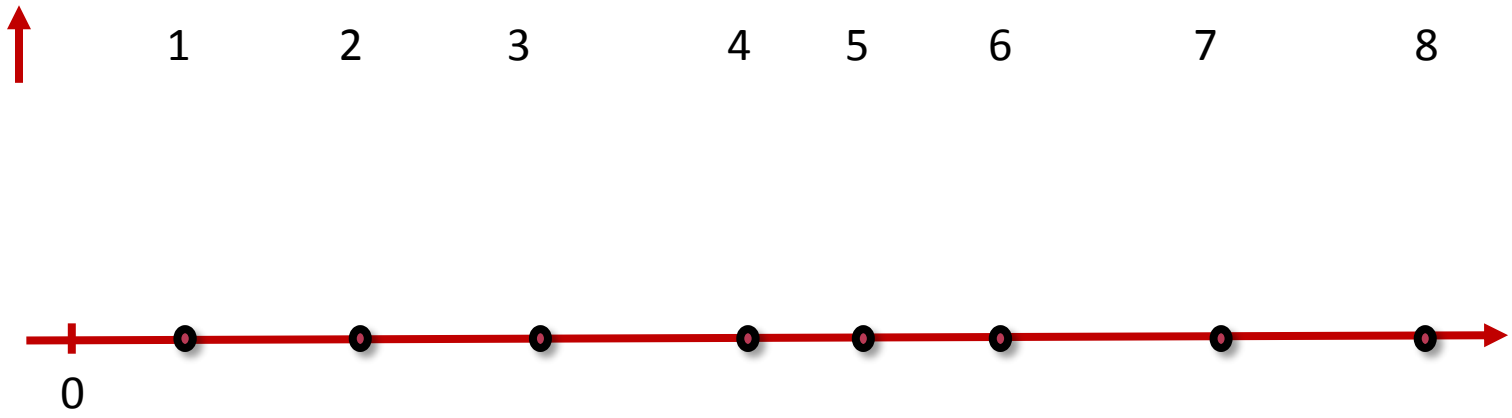
- Féltér-mélység: Tukey, 1974



$$hds(z):\ \min\{|x_i : x_i \leq z|, |x_j : x_j \geq z|\}$$

Befoglaló burok

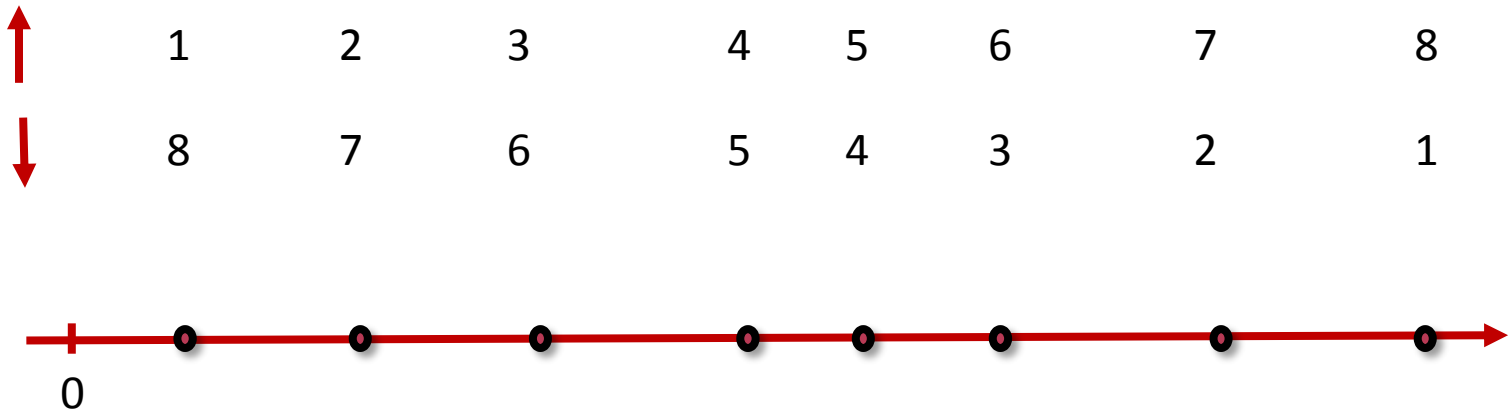
- Féltér-mélység: Tukey, 1974



$$hds(z):\ \min\{|x_i : x_i \leq z|, |x_j : x_j \geq z|\}$$

Befoglaló burok

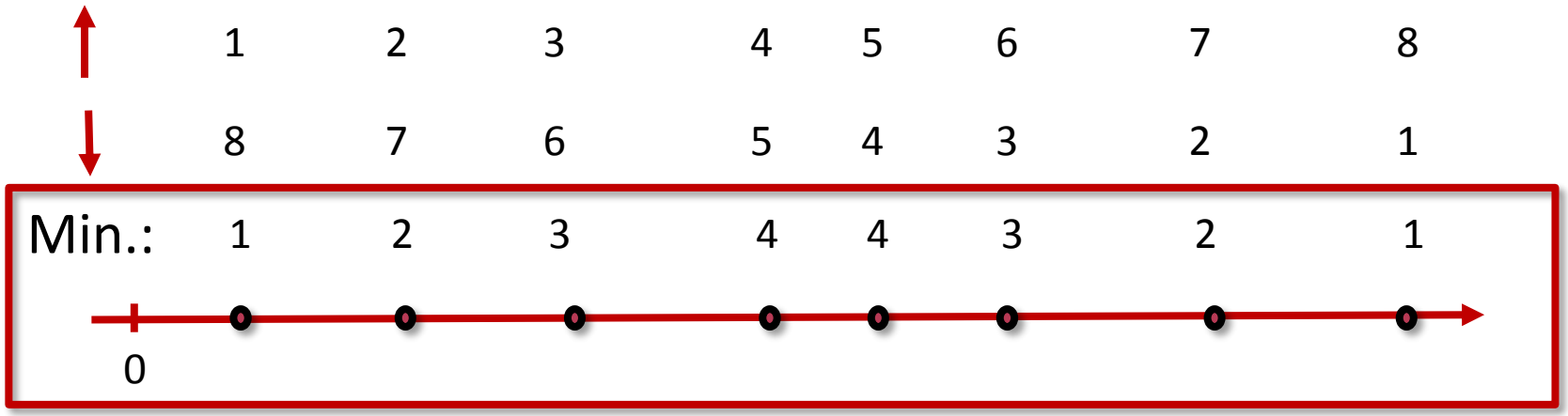
- Féltér-mélység: Tukey, 1974



$$hds(z):\ \min\{|x_i : x_i \leq z|, |x_j : x_j \geq z|\}$$

Befoglaló burok

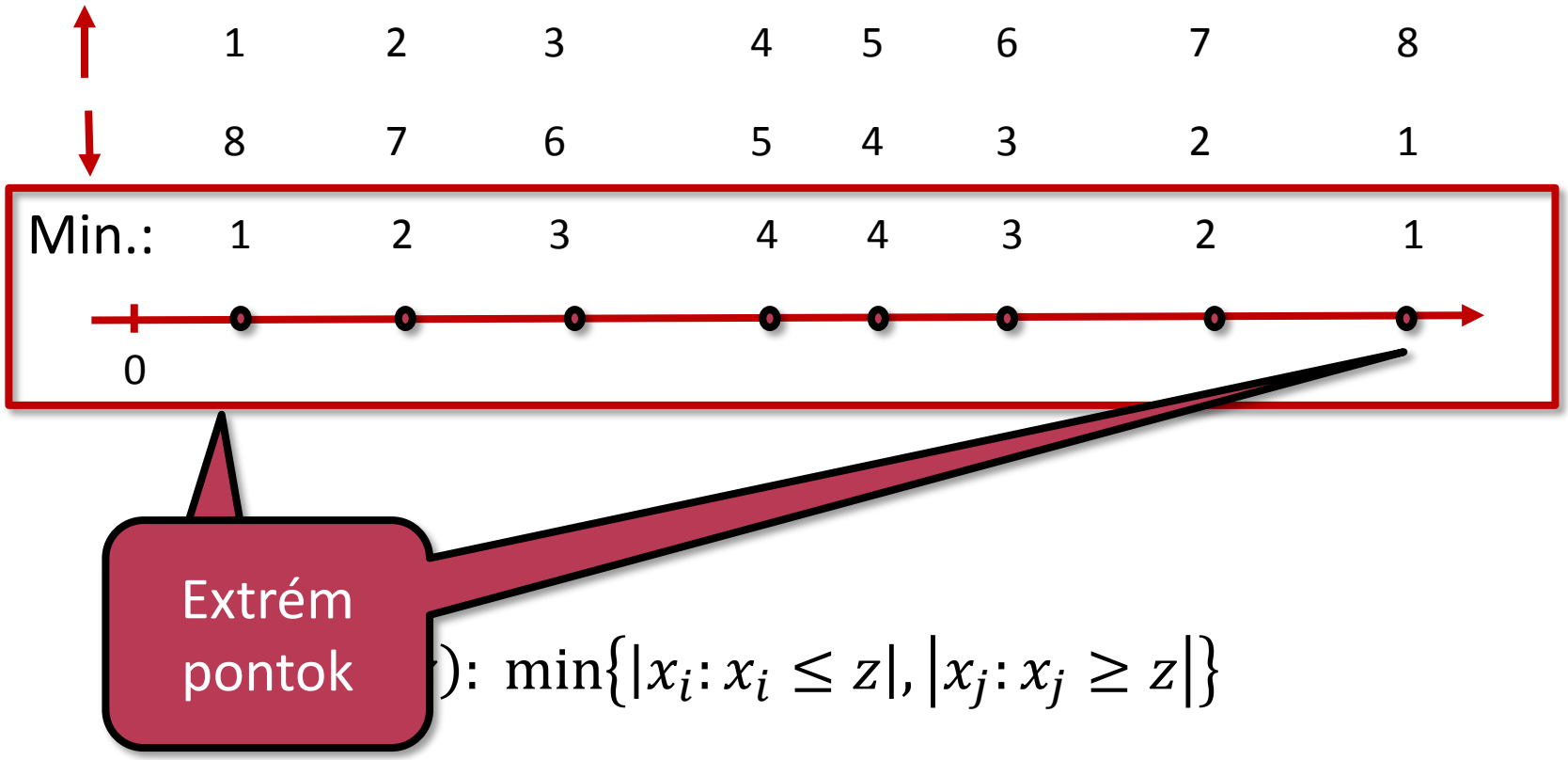
- Féltér-mélység: Tukey, 1974



$$hds(z)\colon \min\{|x_i : x_i \le z|, |x_j : x_j \ge z|\}$$

Befoglaló burok

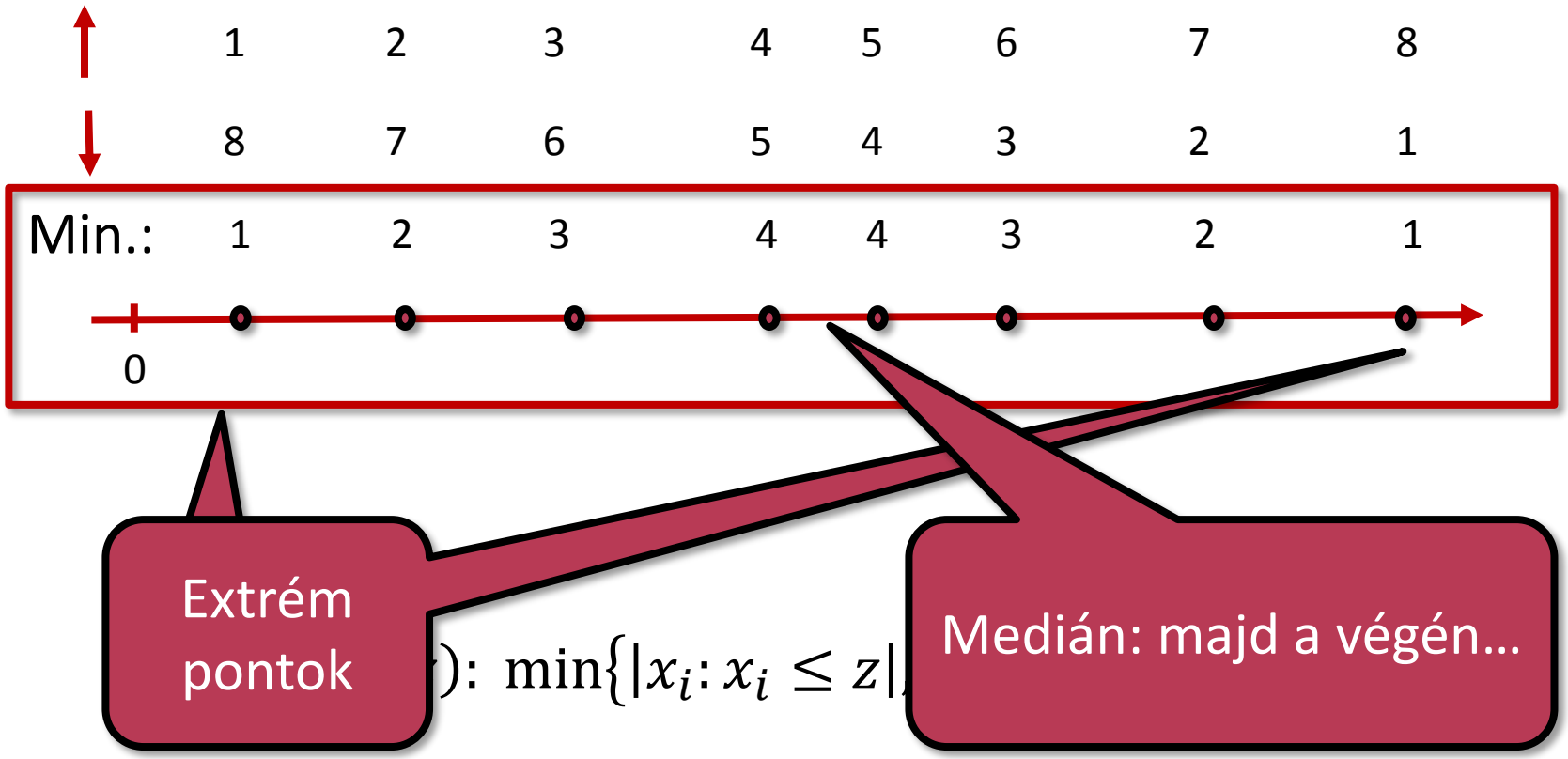
- Féltér-mélység: Tukey, 1974



$$\min\{|x_i : x_i \le z|, |x_j : x_j \ge z|\}$$

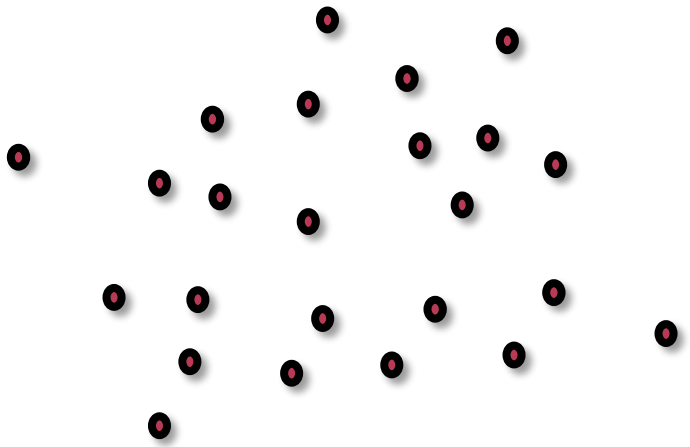
Befoglaló burok

- Féltér-mélység: Tukey, 1974



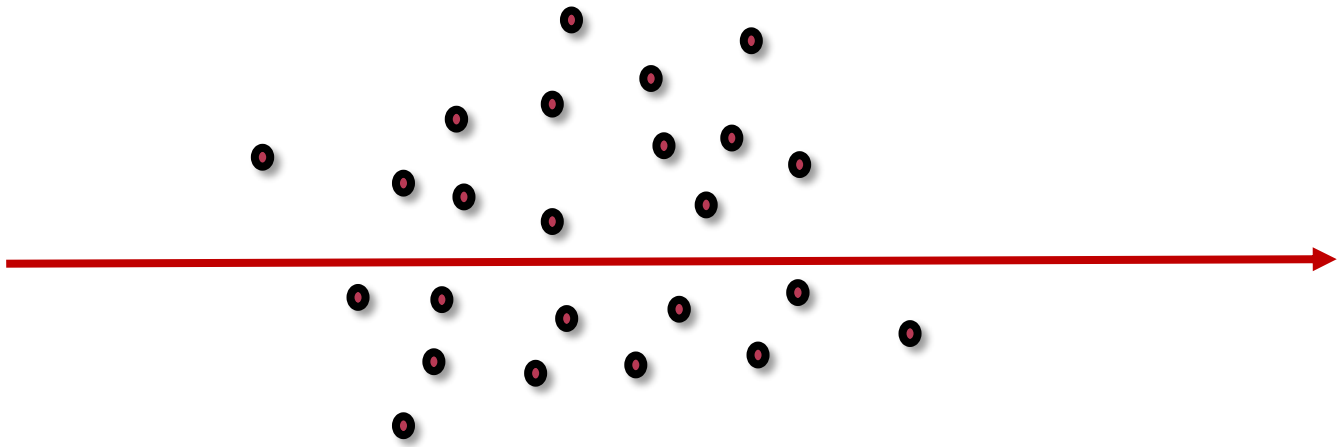
Befoglaló burok

- Féltér-mélység: Tukey, 1974



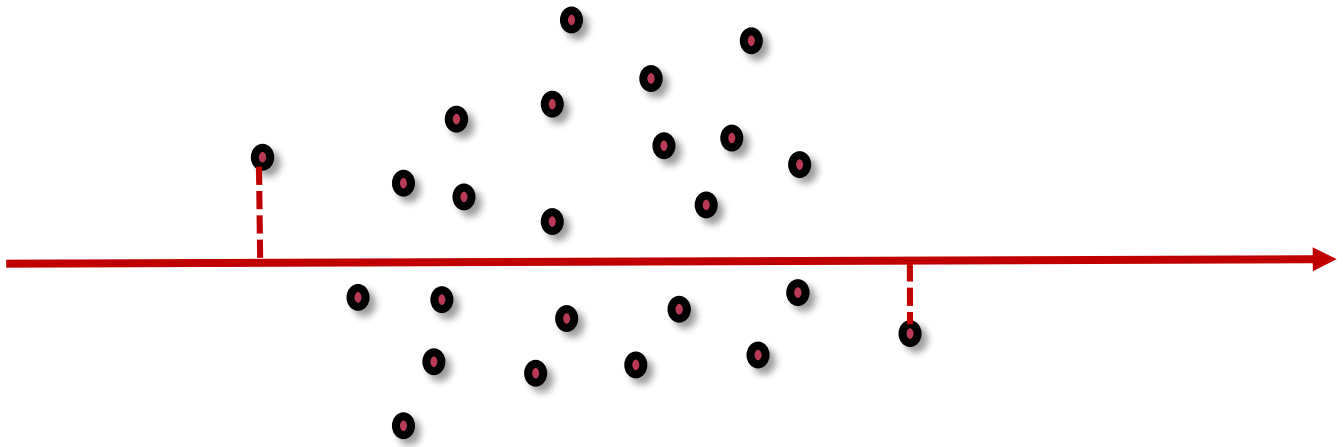
Befoglaló burok

- Féltér-mélység: Tukey, 1974



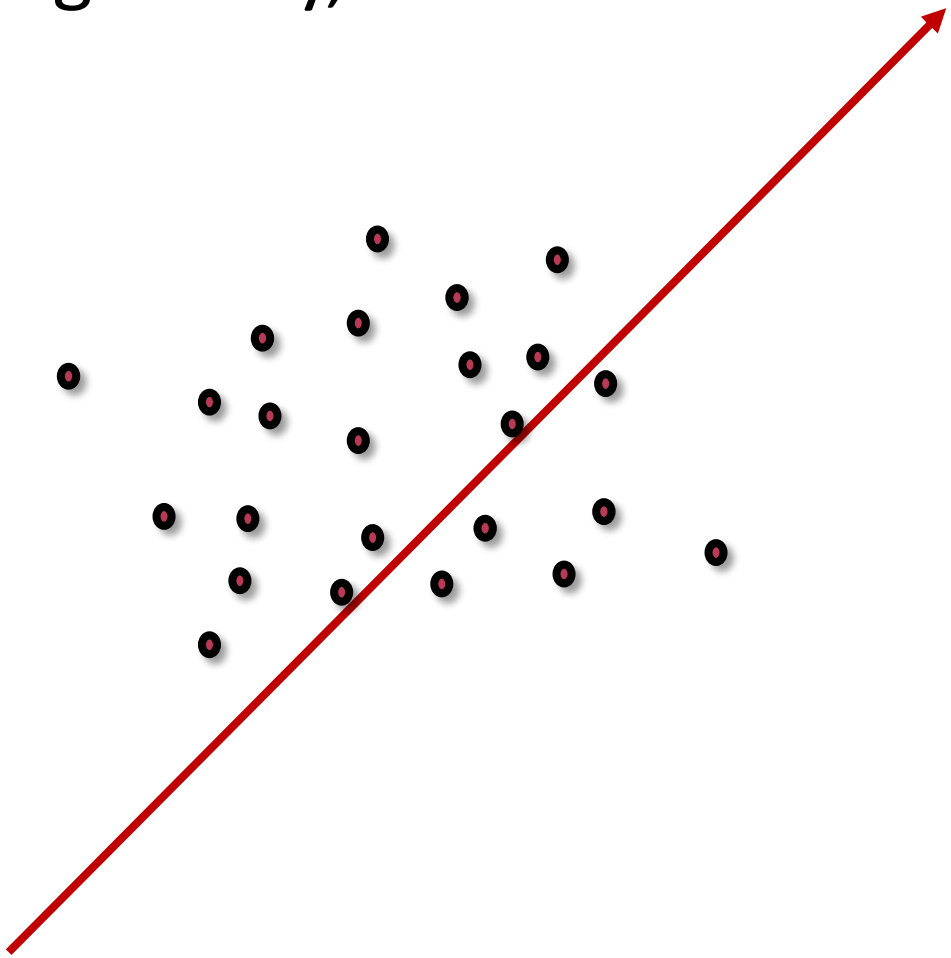
Befoglaló burok

- Féltér-mélység: Tukey, 1974



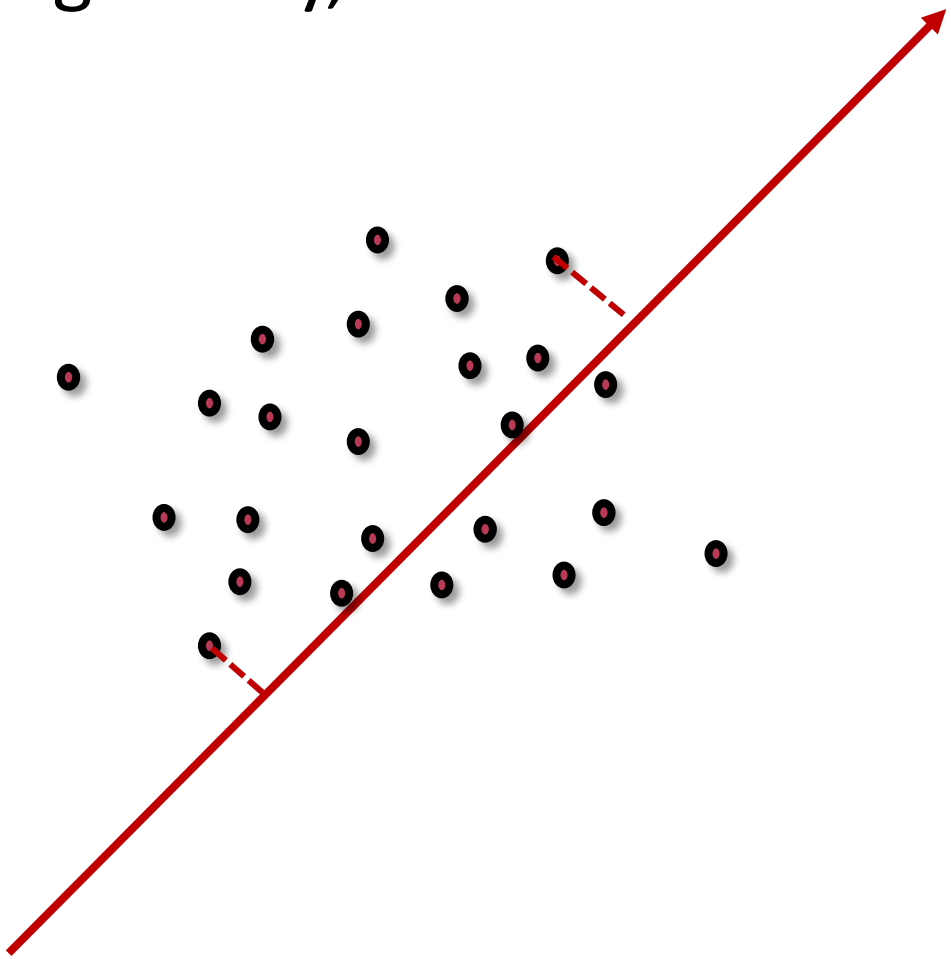
Befoglaló burok

- Féltér-mélység: Tukey, 1974



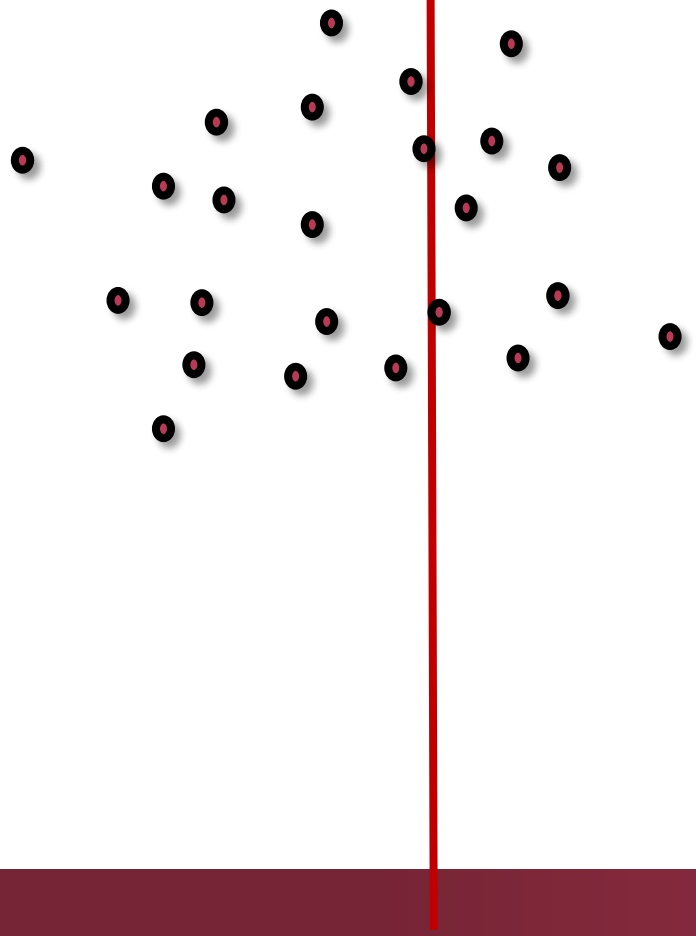
Befoglaló burok

- Féltér-mélység: Tukey, 1974



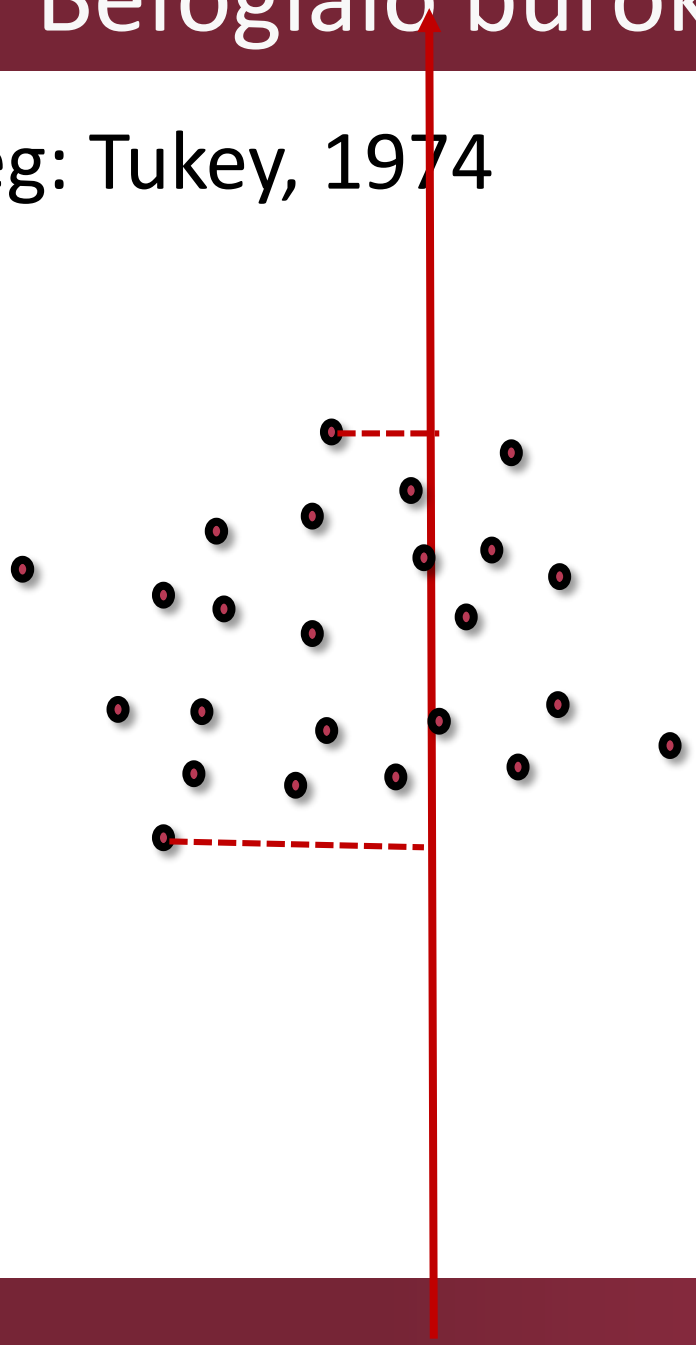
Befoglaló burok

- Féltér-mélység: Tukey, 1974



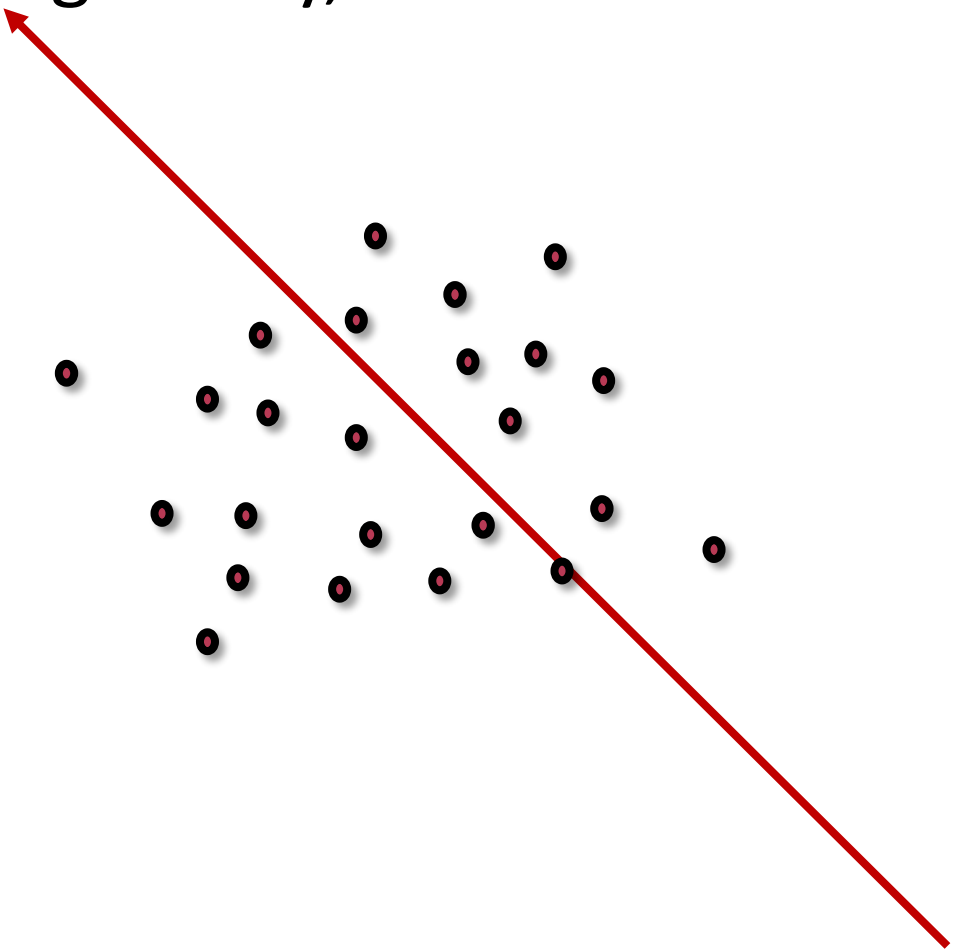
Befoglaló burok

- Féltér-mélység: Tukey, 1974



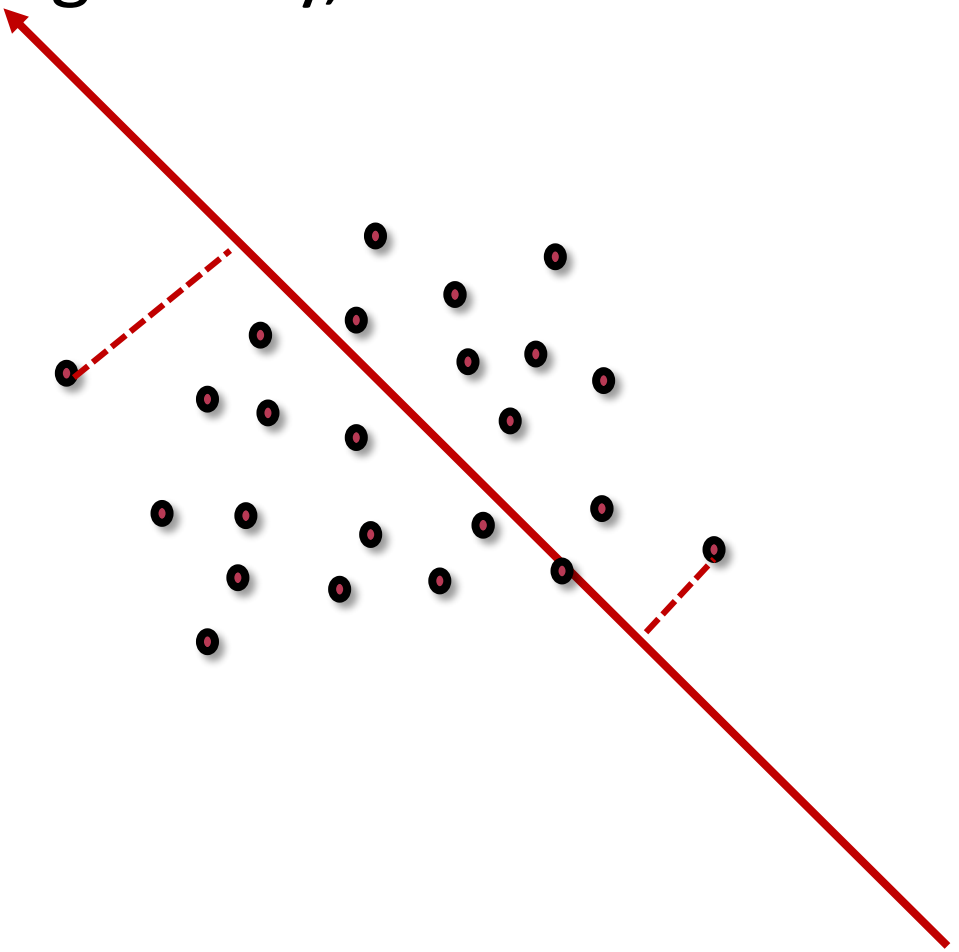
Befoglaló burok

- Féltér-mélység: Tukey, 1974



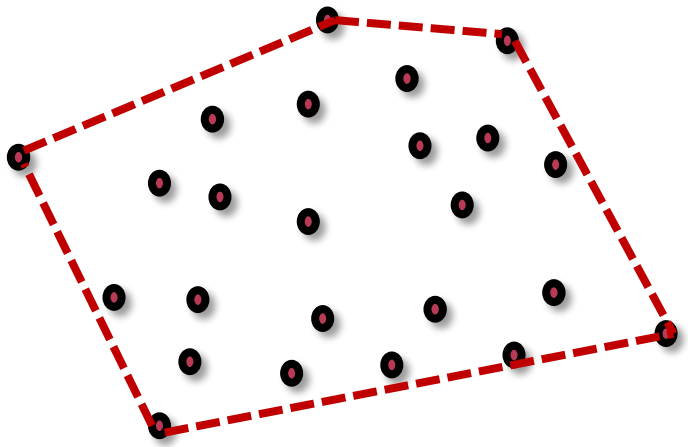
Befoglaló burok

- Féltér-mélység: Tukey, 1974

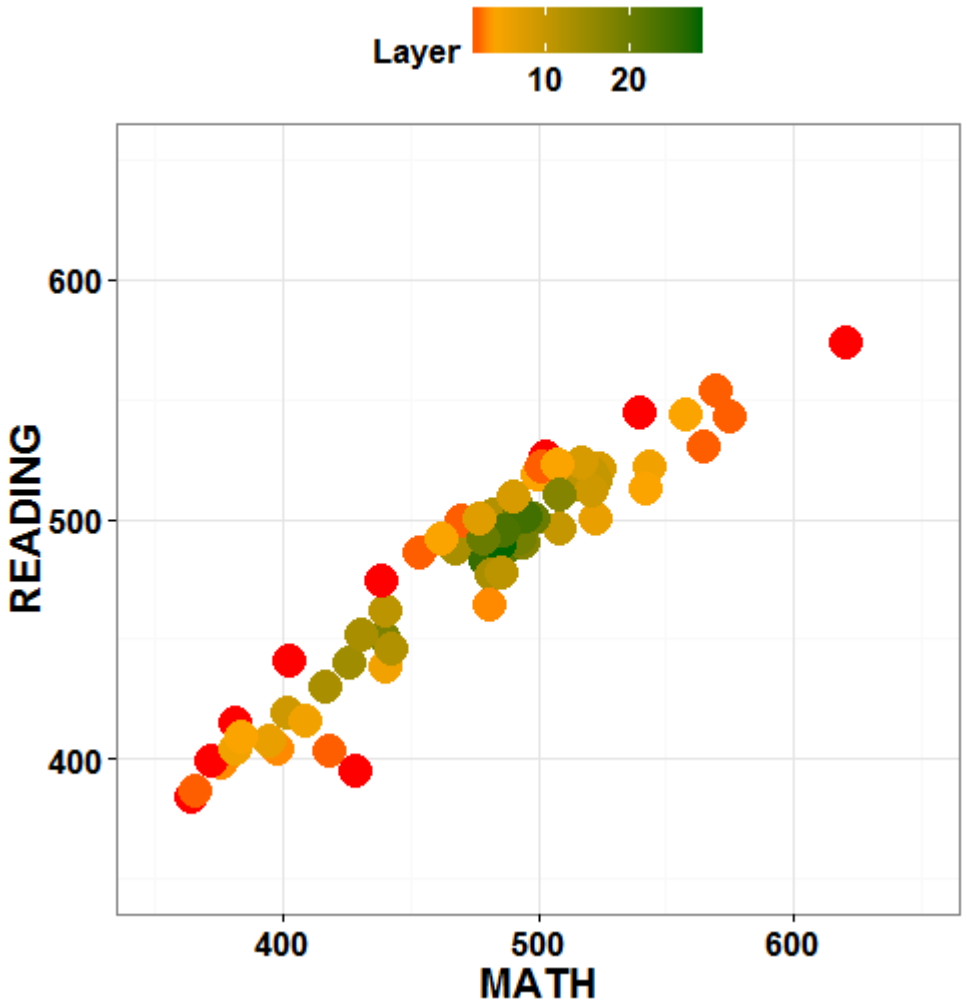


Befoglaló burok

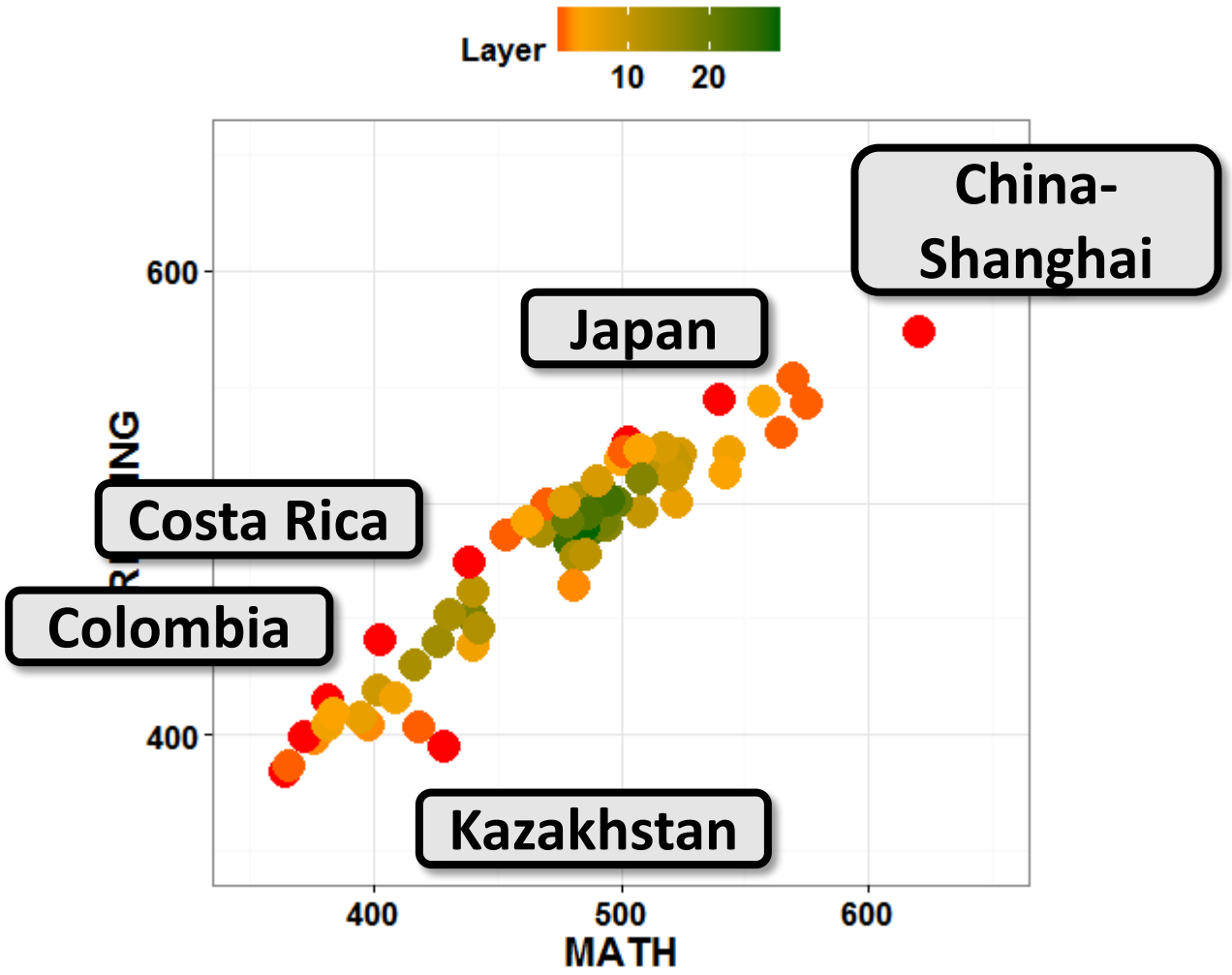
- Féltér-mélység: Tukey, 1974



Isodepth (Depth-Based)

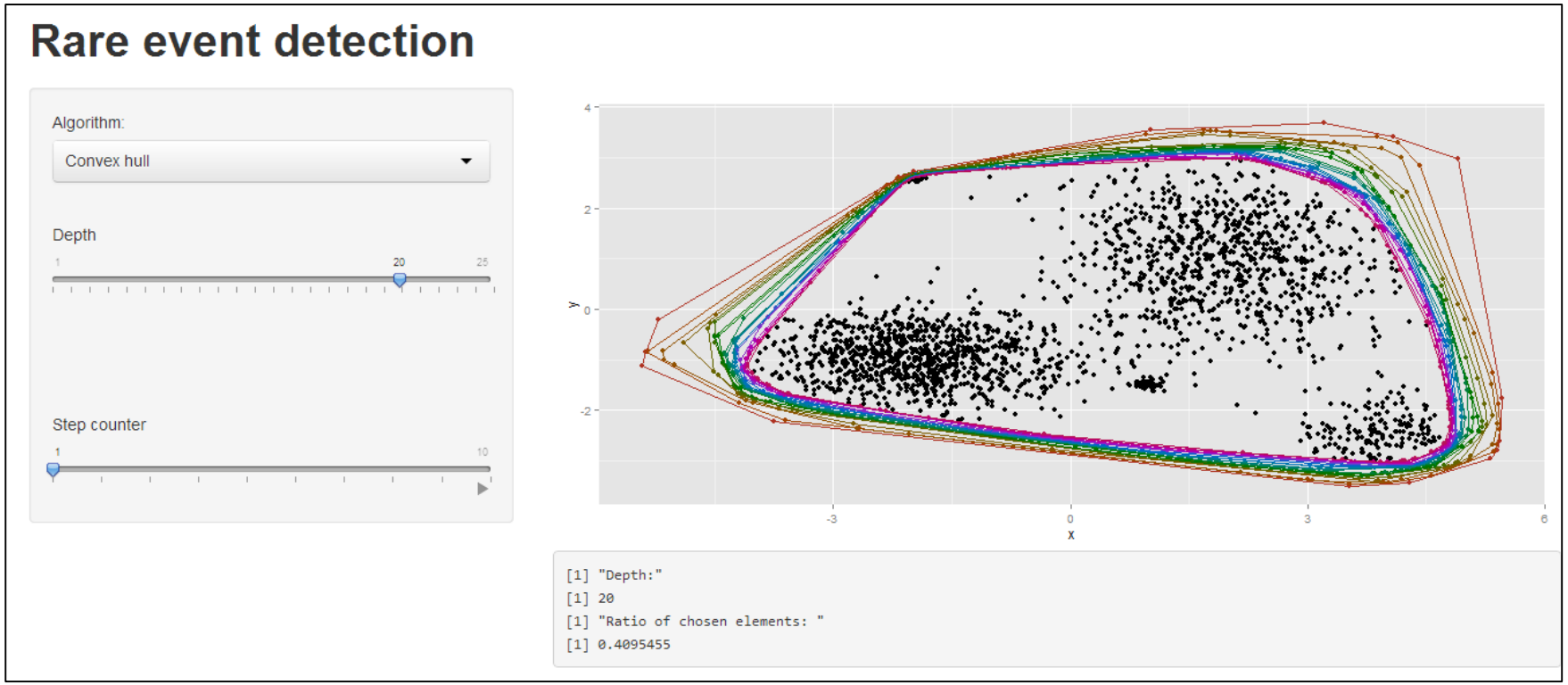


Isodepth (Depth-Based)



DEMO Befoglaló burok

- Csomag: *depth*
- Hasznos függvények: *depth, isodepth*
- Paraméterek: u pont, $dpth$ mélység

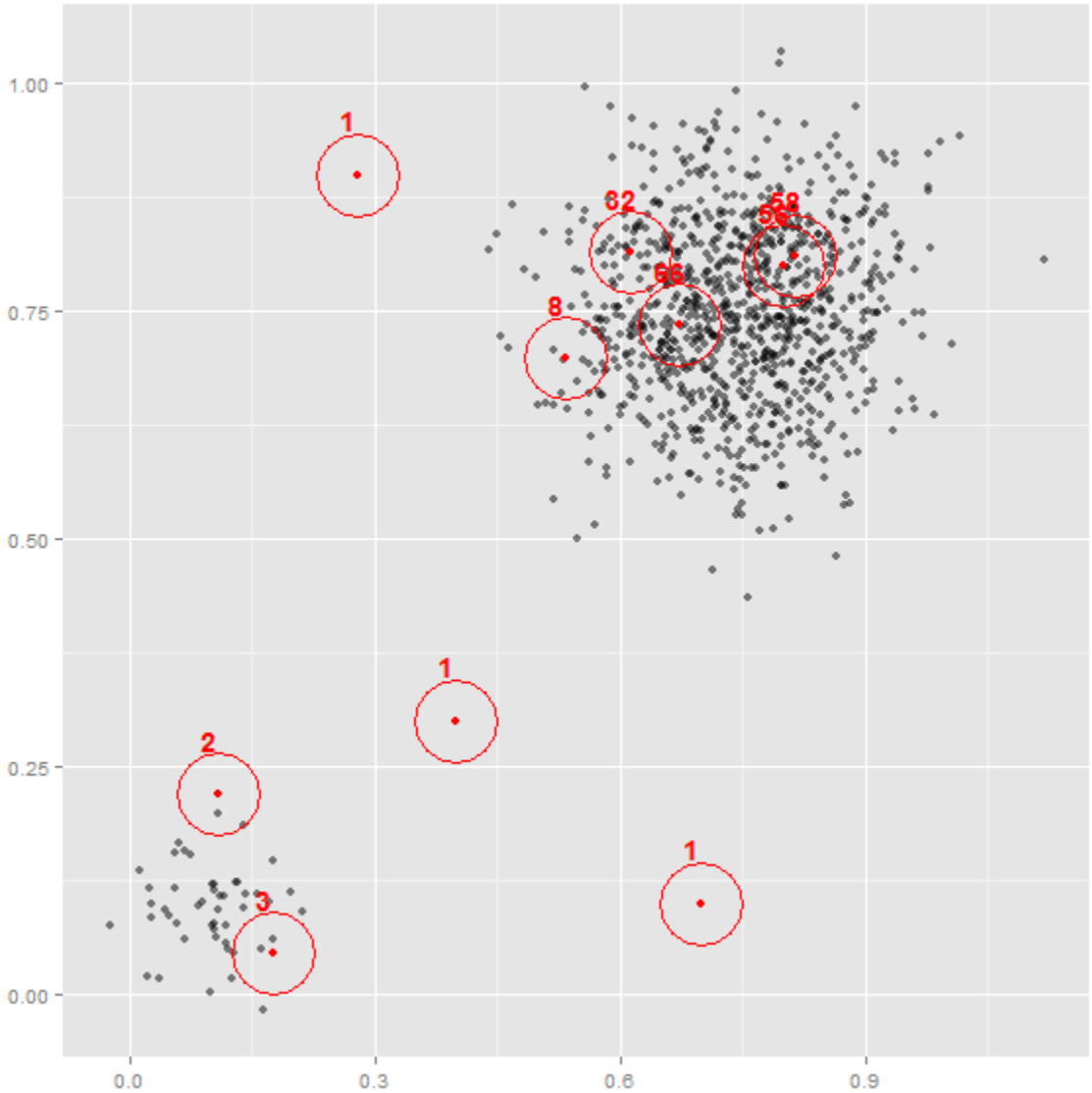


DB

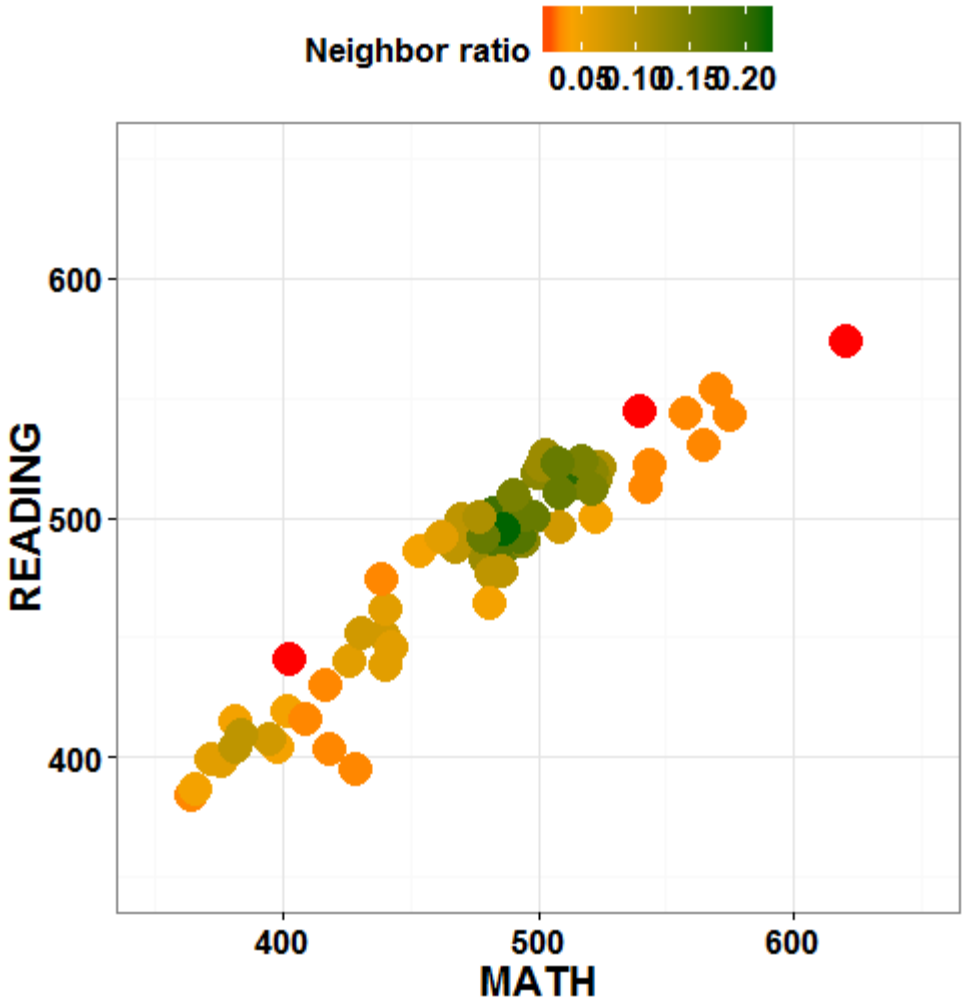
- Distance Based
- Outlier: szomszédok száma alacsony
- Paraméterek
  - r sugarú hipergömb
  - Szomszédok elvárt π aránya

DB

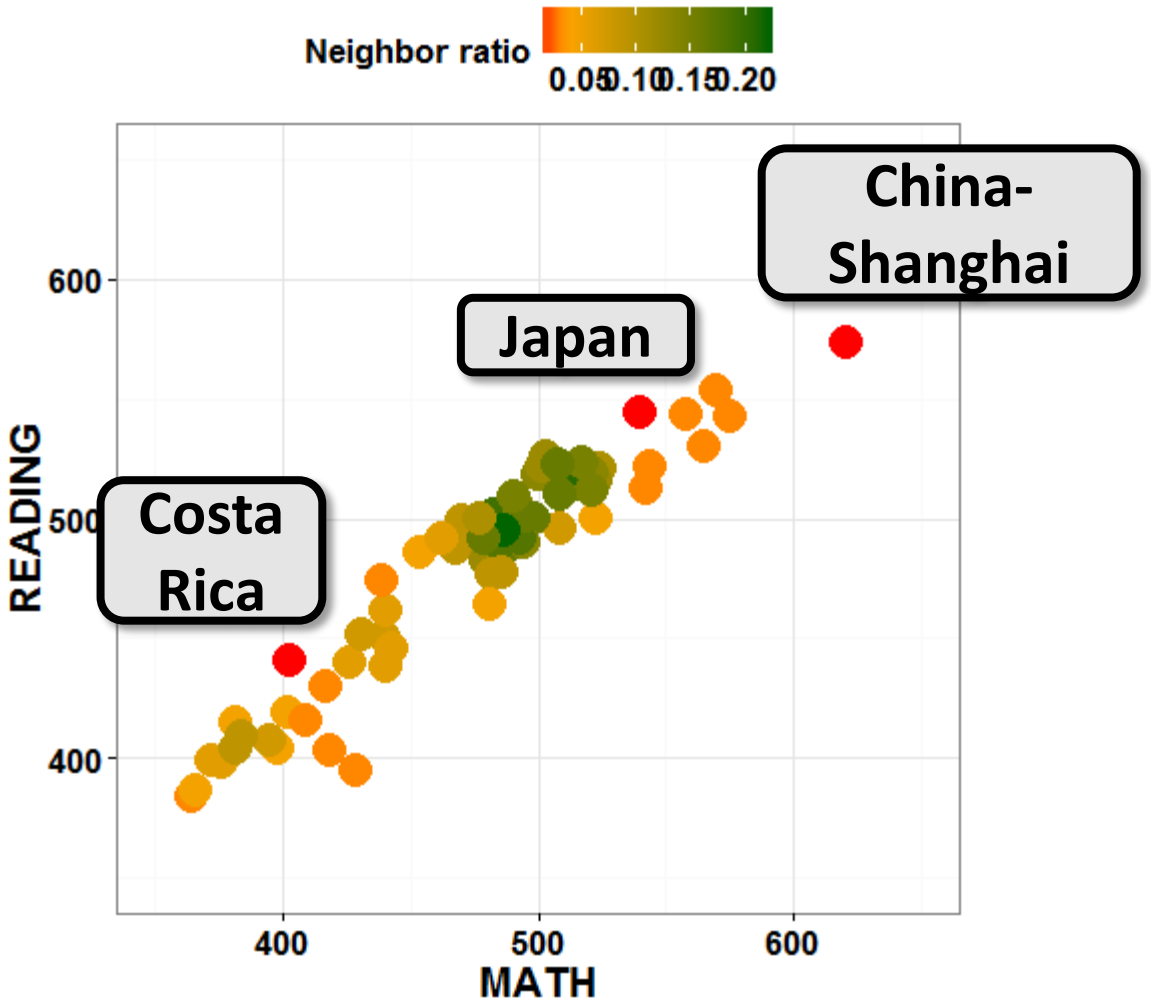
- Distanc
- Outlier
- Parame
  - r sug
  - Szom



db – Distance-Based

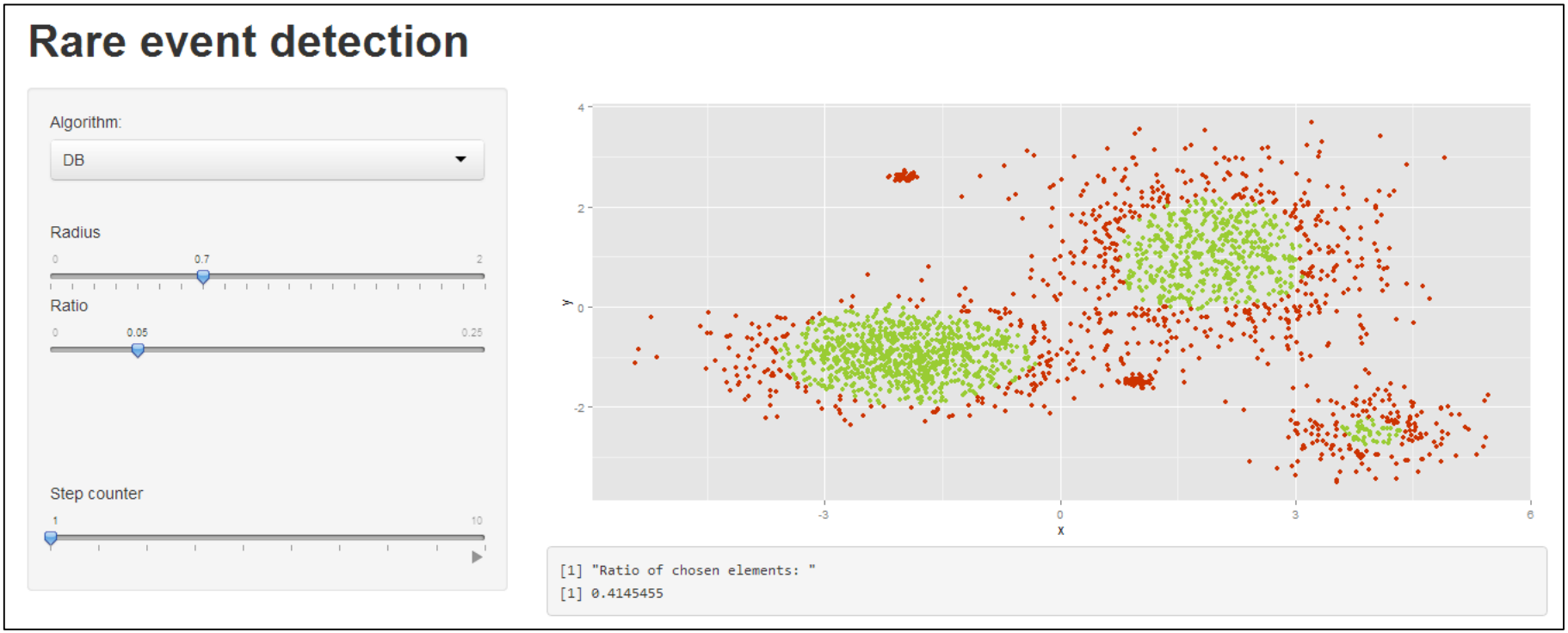


db – Distance-Based



DEMO DB

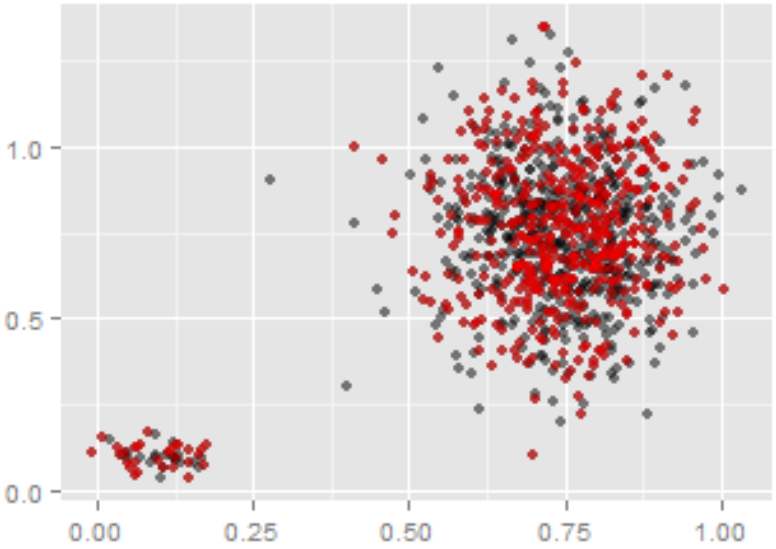
- Csomag: *fields*
- Függvény: *fields.rdist.near*
- Paraméterek: $delta$ sugár



MCD

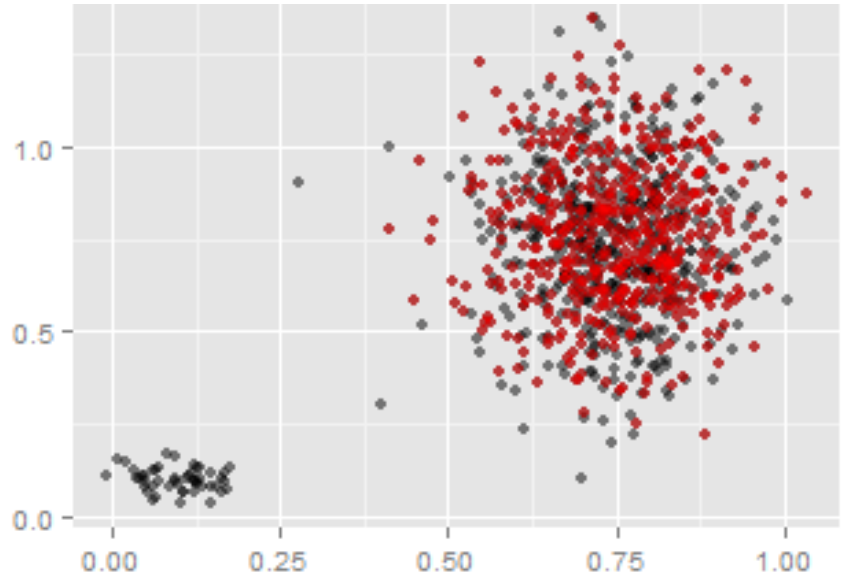
- Minimum Covariance Determinant
- Alapötlet
  - Keressük meg a legkompaktabb részhalmazt!

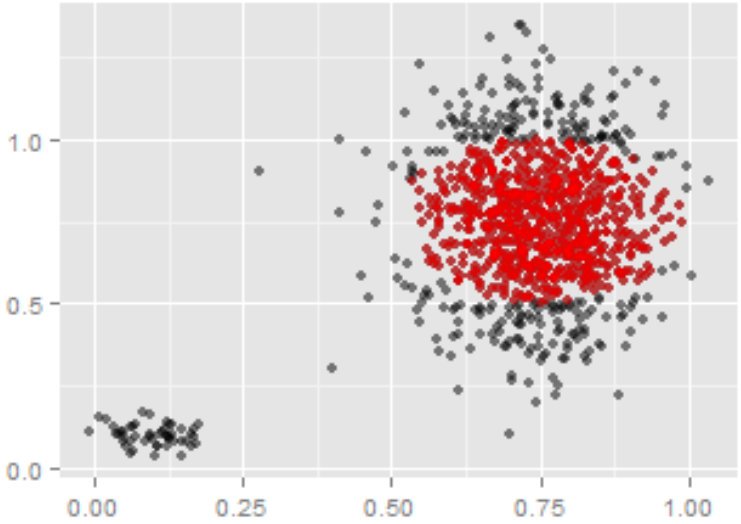
MCD


1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00

De

ıpa

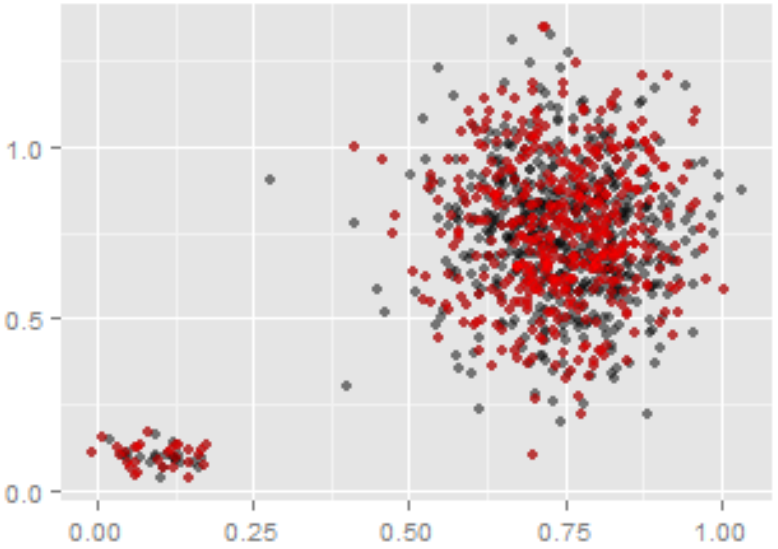

1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00


1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00

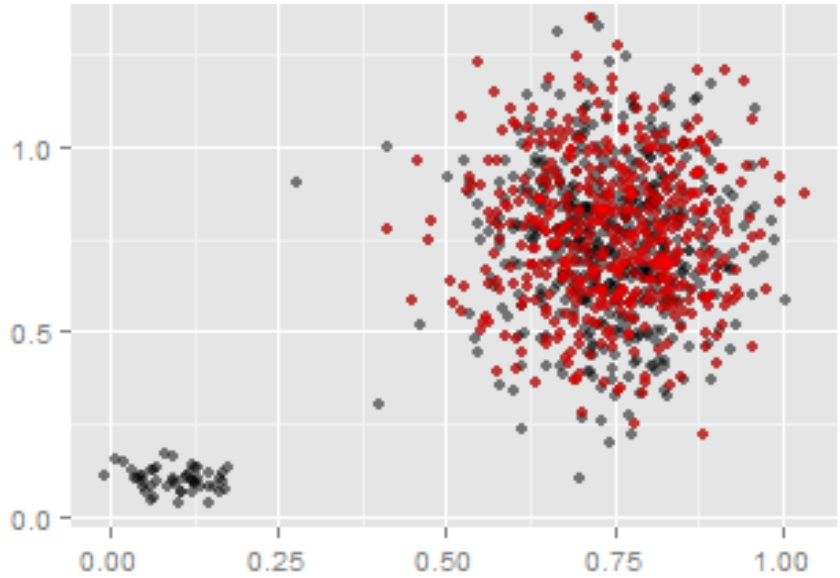
MCD

De

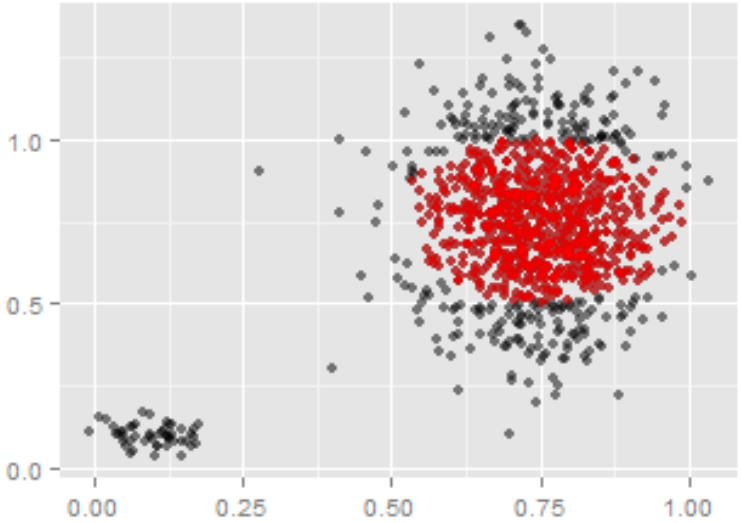
ıpa



0.0014



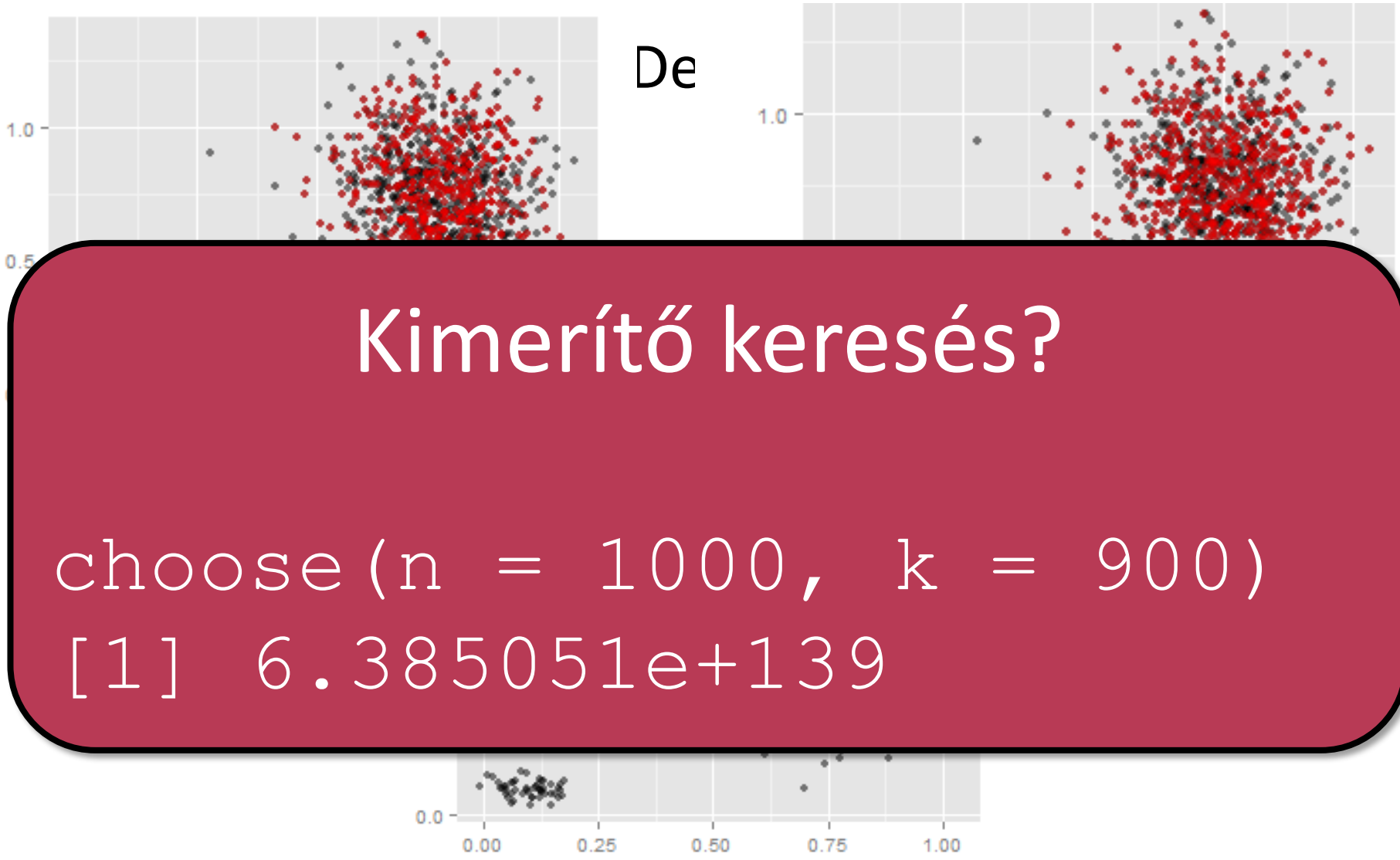
0.00041



0.00011

MCD

De



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 alapján

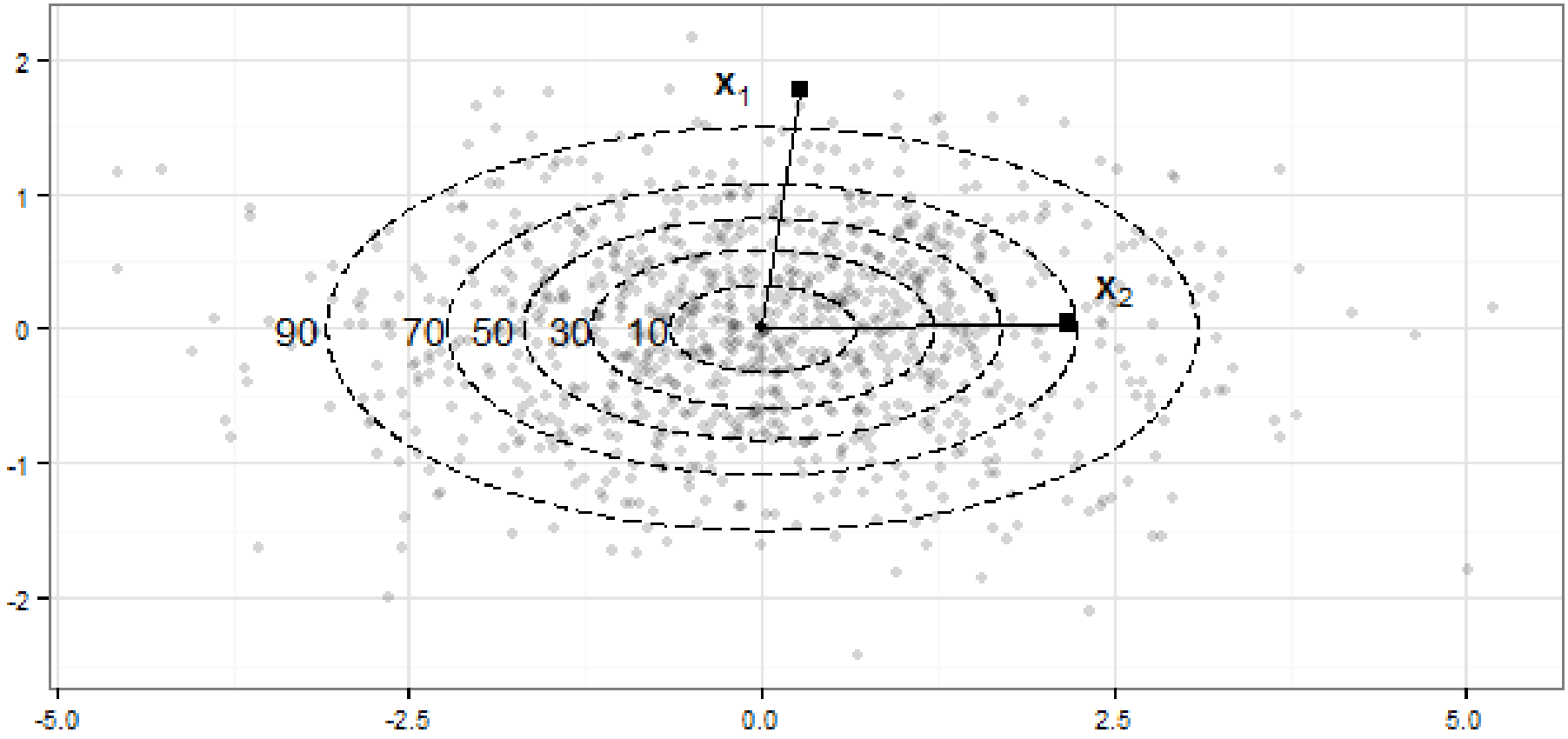
Mahalanobis távolság

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

Ábra forrása: http://stats.stackexchange.com/questions/62092/bottom-to-top-explanation-of-the-mahalanobis-distance

Mahalanobis távolság

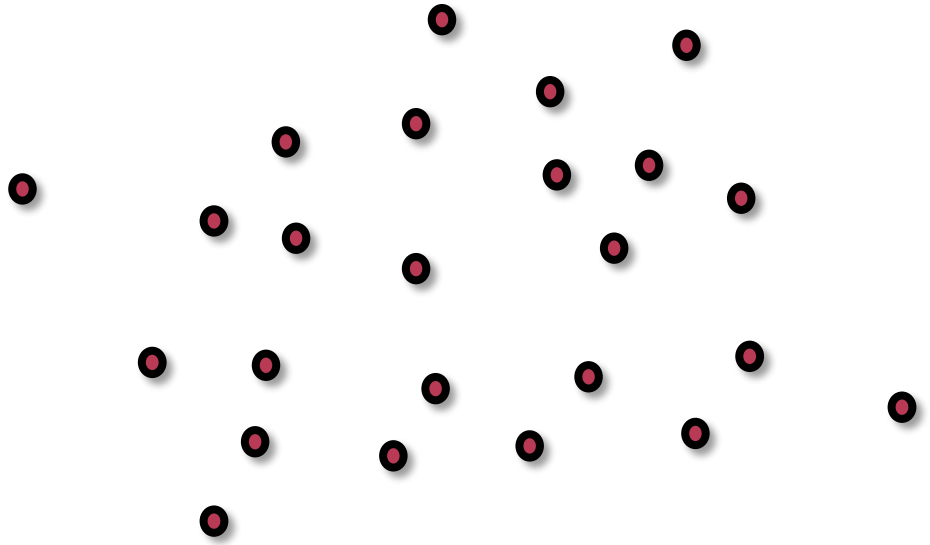
- $D(x, M) = \sqrt{(x - \vartheta)^T S^{-1} (x - \vartheta)}$
  - S – kovarianciamátrix



Ábra forrása: http://stats.stackexchange.com/questions/62092/bottom-to-top-explanation-of-the-mahalanobis-distance

FAST-MCD

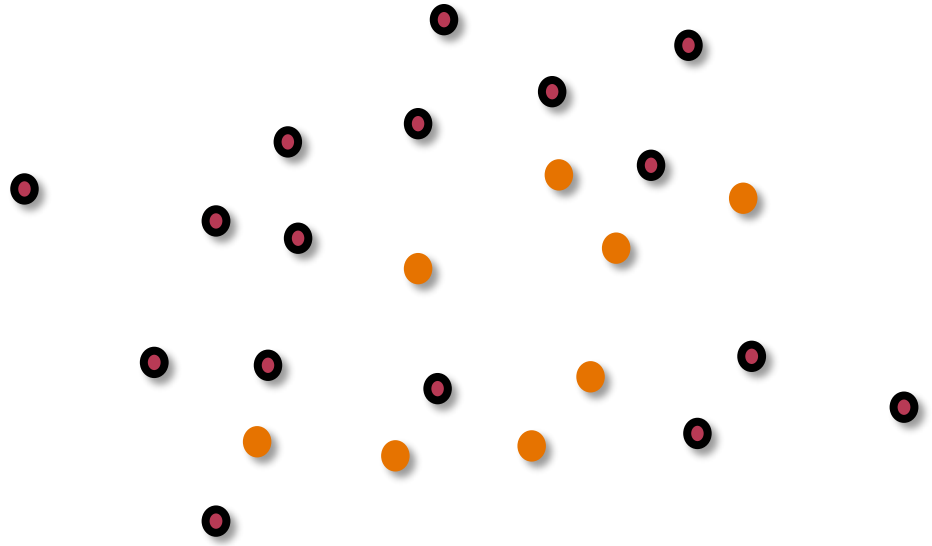
- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív



- Legközelebbi pontok kiválasztása
  - Mahalanobis távolság alapján
  - Legközelebbi $x\%$

FAST-MCD

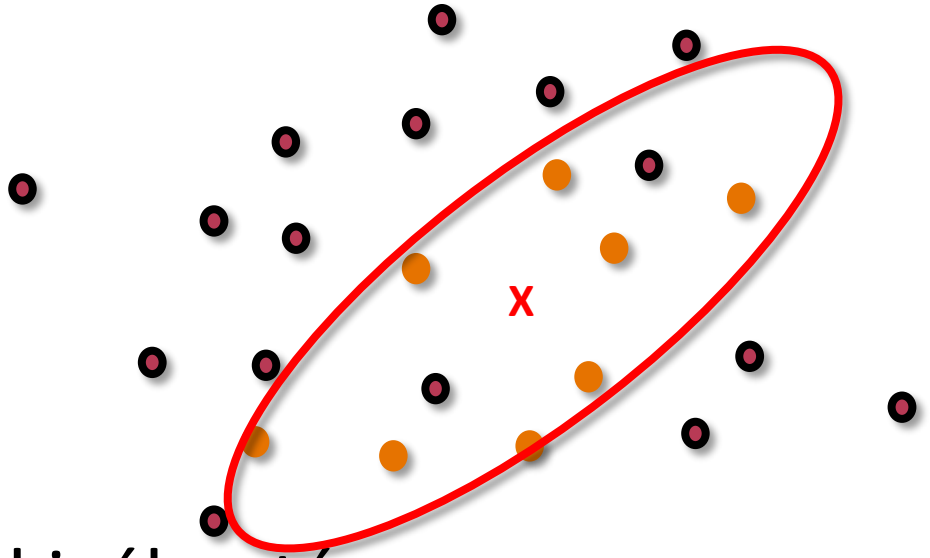
- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív



- Legközelebbi pontok kiválasztása
  - Mahalanobis távolság alapján
  - Legközelebbi $x\%$

FAST-MCD

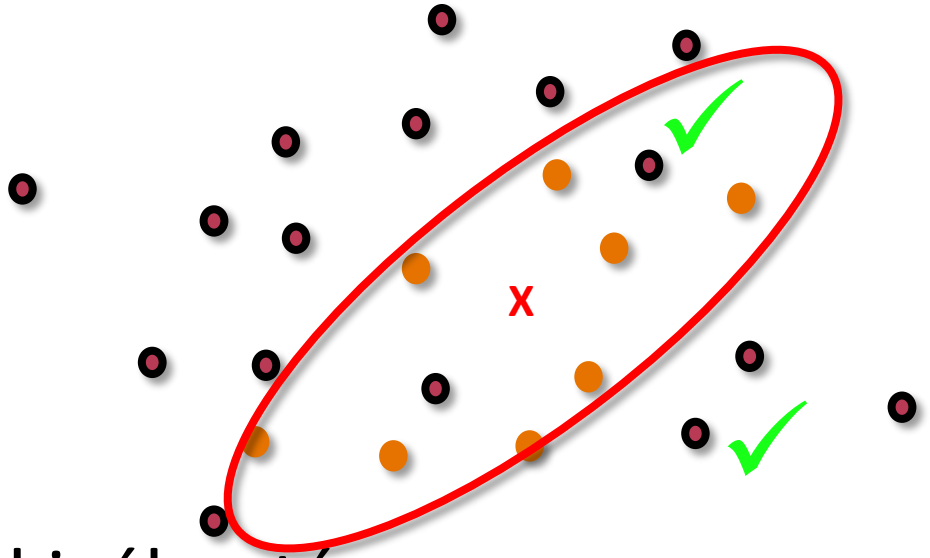
- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív



- Legközelebbi pontok kiválasztása
  - Mahalanobis távolság alapján
  - Legközelebbi $x\%$

FAST-MCD

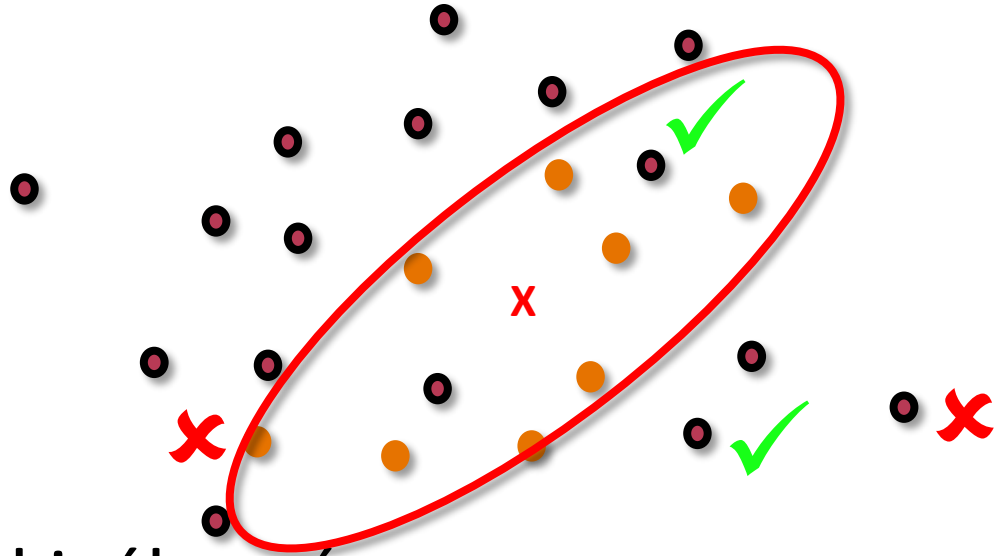
- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív



- Legközelebbi pontok kiválasztása
  - Mahalanobis távolság alapján
  - Legközelebbi $x\%$

FAST-MCD

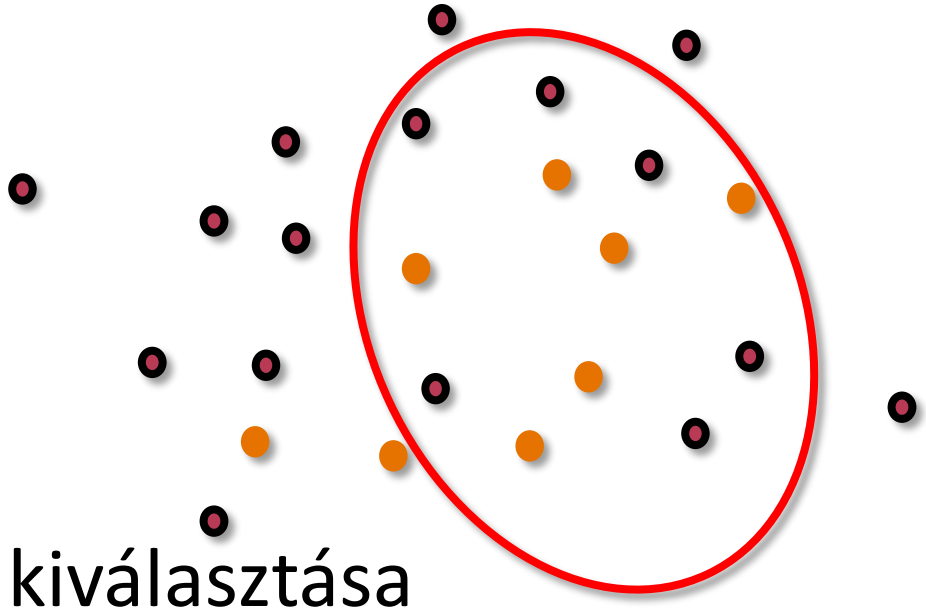
- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív



- Legközelebbi pontok kiválasztása
  - Mahalanobis távolság alapján
  - Legközelebbi $x\%$

FAST-MCD

- Közelítő algoritmus
- Véletlenszerűen választott kezdőhalmaz
- Iteratív



- Legközelebbi pontok kiválasztása
  - Mahalanobis távolság alapján
  - Legközelebbi $x\%$

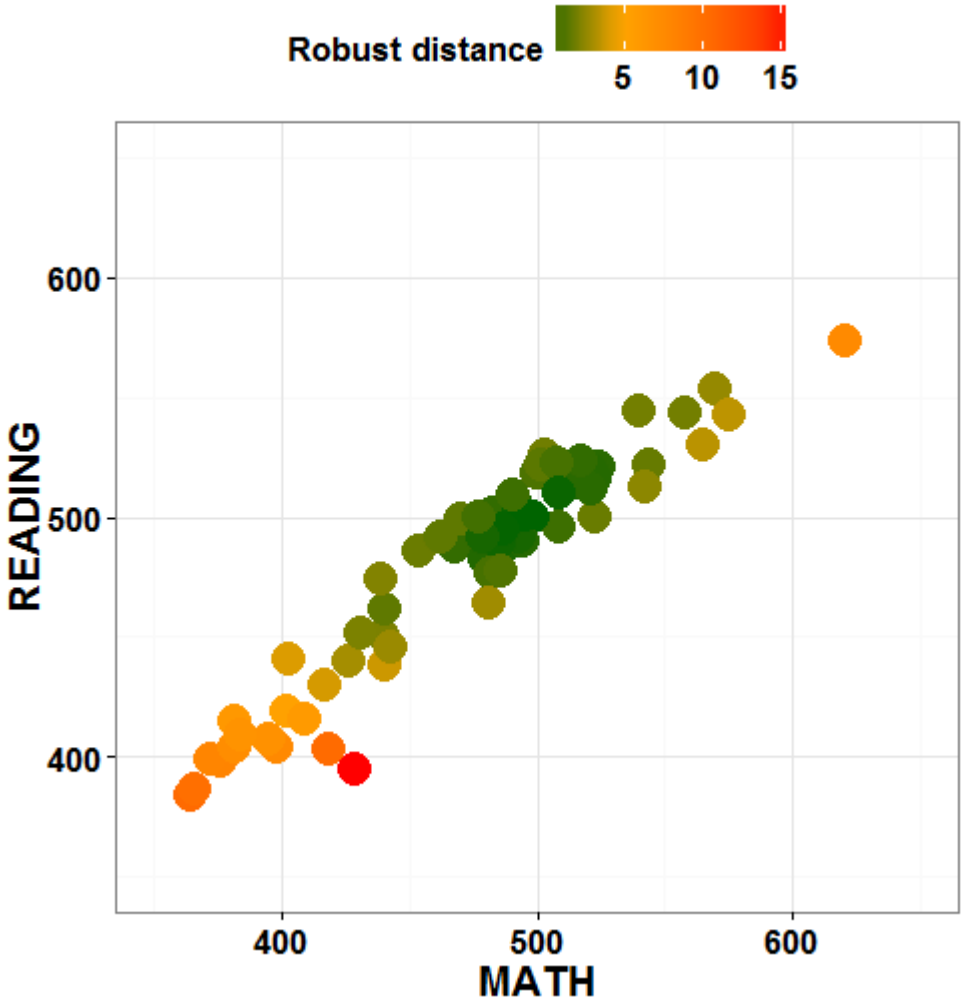
BACON

- Blocked Adaptive Computationally Efficient Outlier Nominators

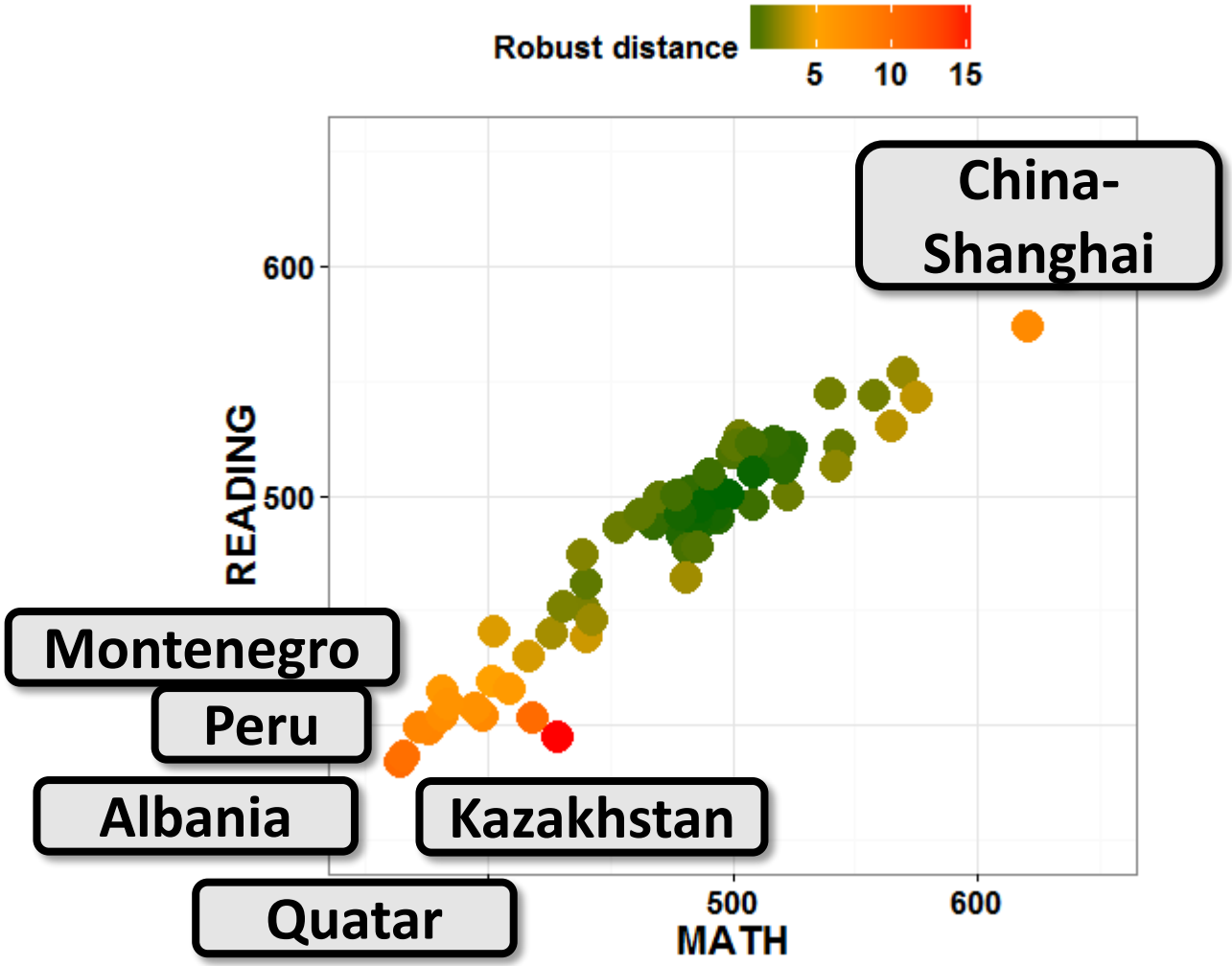
- Kiinduló halmaz félig felügyelt módban is!

- Új halmaz: küszöbérték alapján

BACON



BACON

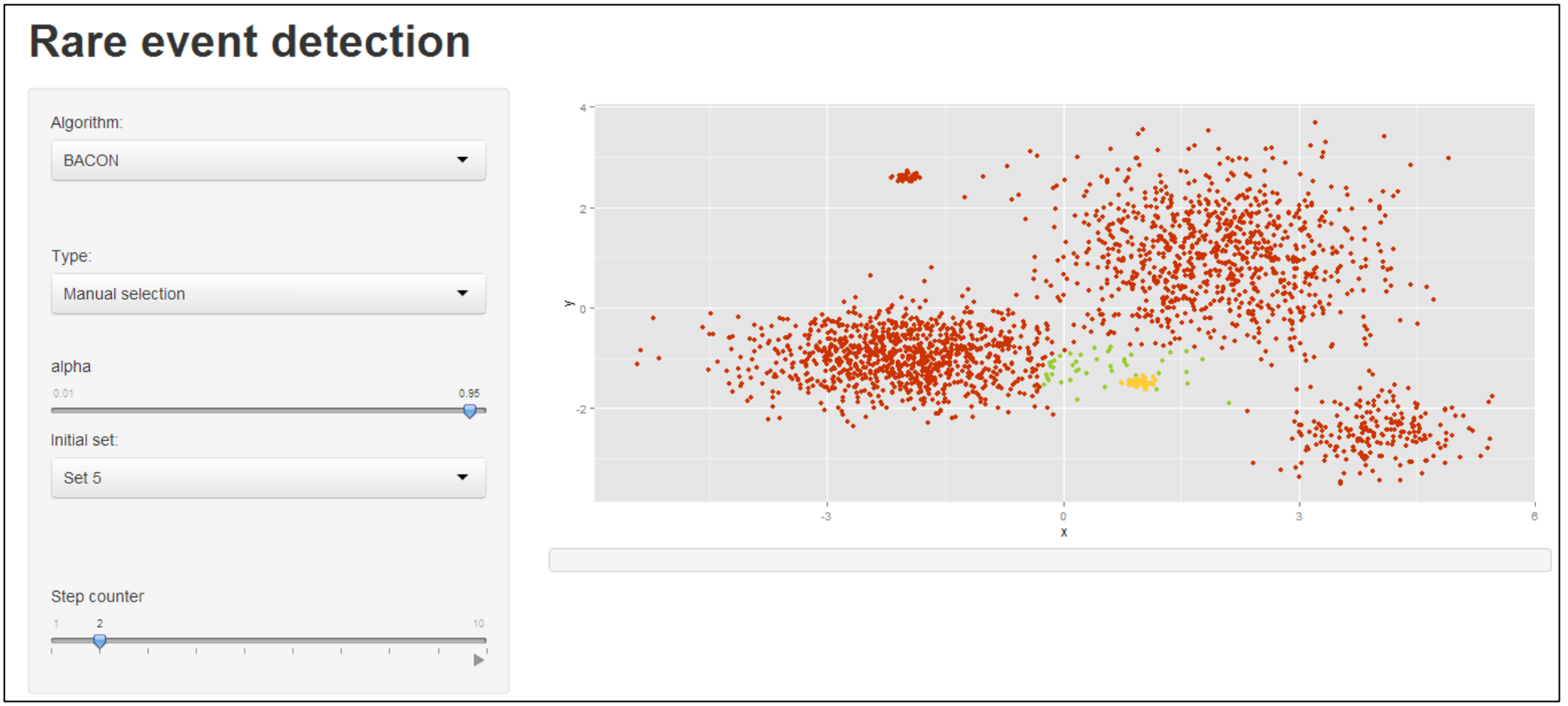

Robust distance
5
10
15
China-
Shanghai
600
500
READING
Montenegro
Peru
Albania
Kazakhstan
Quatar
500
600
MATH

DEMO BACON

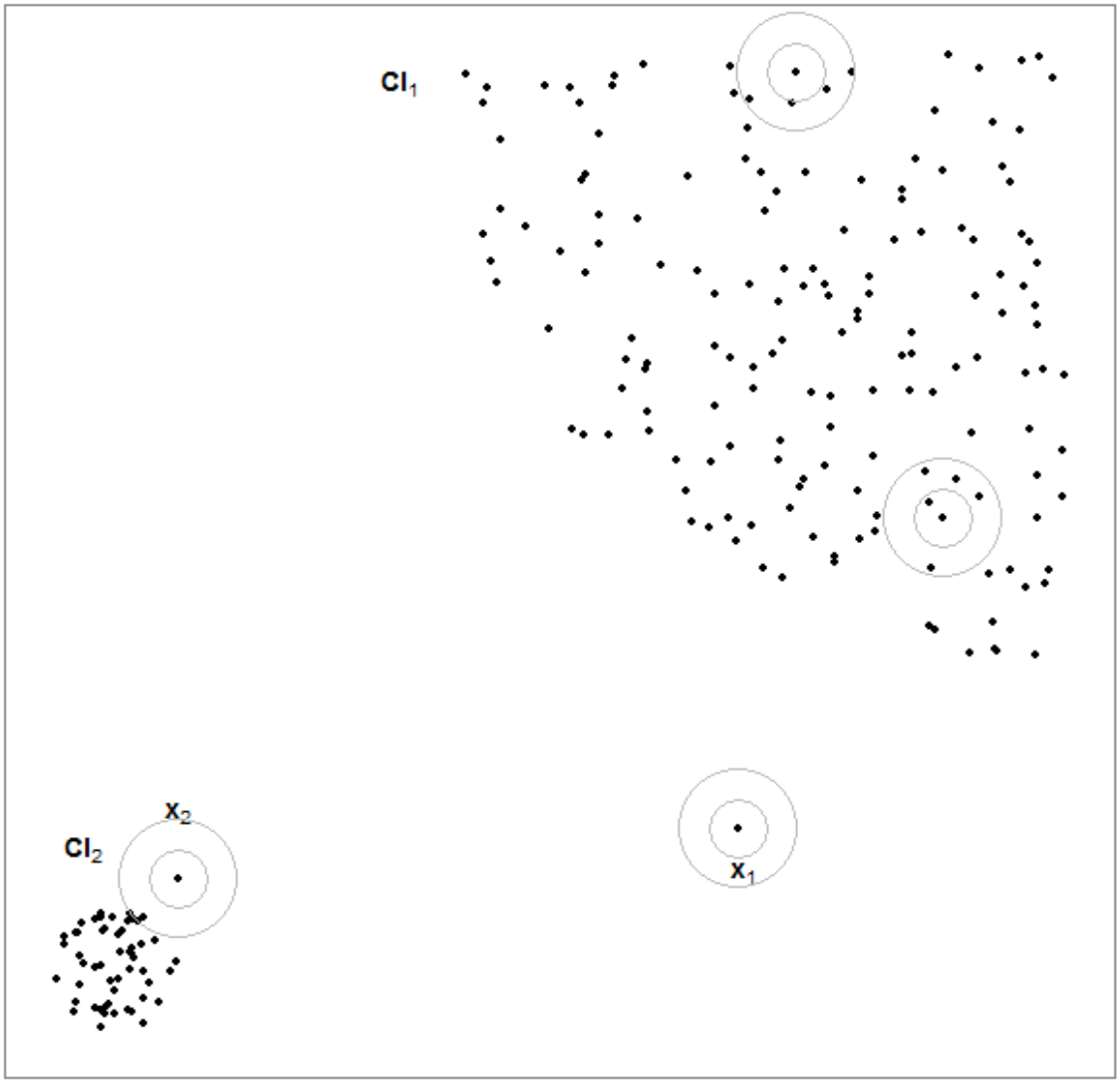
- Csomag: *robustX*
- Függvény: *mvBACON*
- Paraméterek
  - $init.sel$ kezdőhalmaz
    - „manual” – $man.sel$ kezdőhalmaz
    - „Mahalanobis”, „dUniMedian” – m kezdőhalmaz mérete

DEMO BACON

- Csomag: *robustX*



LOF motiváció



LOF

- Local Outlier Factor

- Alapötlet
  - csak a szomszéd számít
  - a távolság is módosul
  - lokális sűrűség

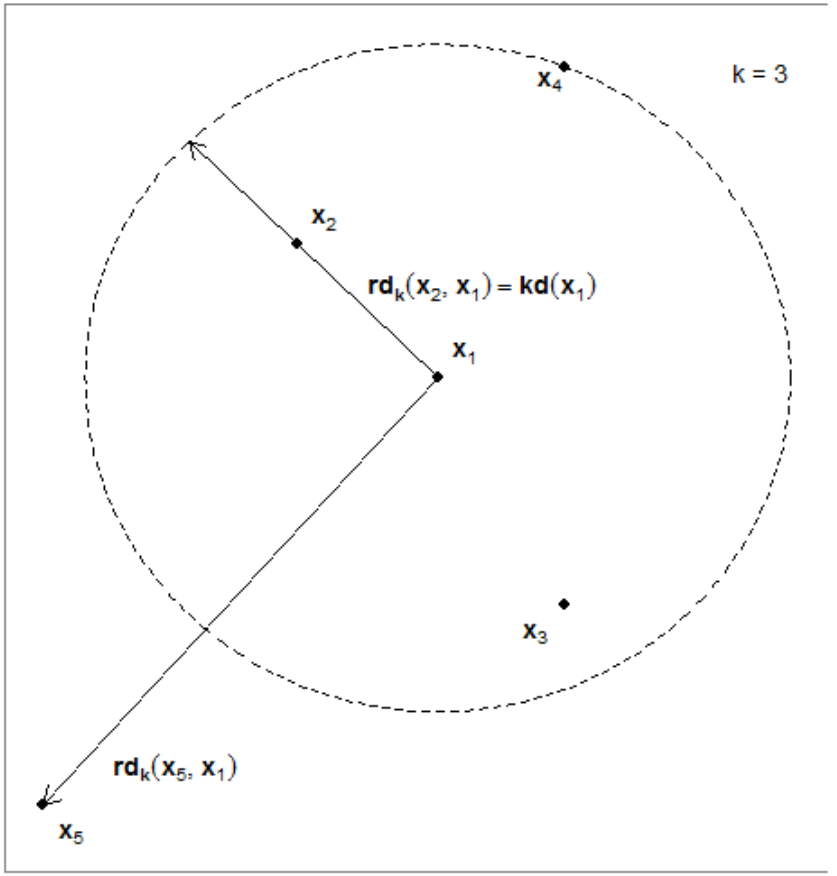
- Outlier kritérium
  - a lokális sűrűség jóval kisebb, mint a szomszédaimnak átlagosan

LOF

- Local Outlier Factor

- Alapötlet
  - csak a szomszéd számít
  - a távolság is módosul
  - lokális sűrűség

- Outlier kritérium
  - a lokális sűrűség jóval kisek átlagosan



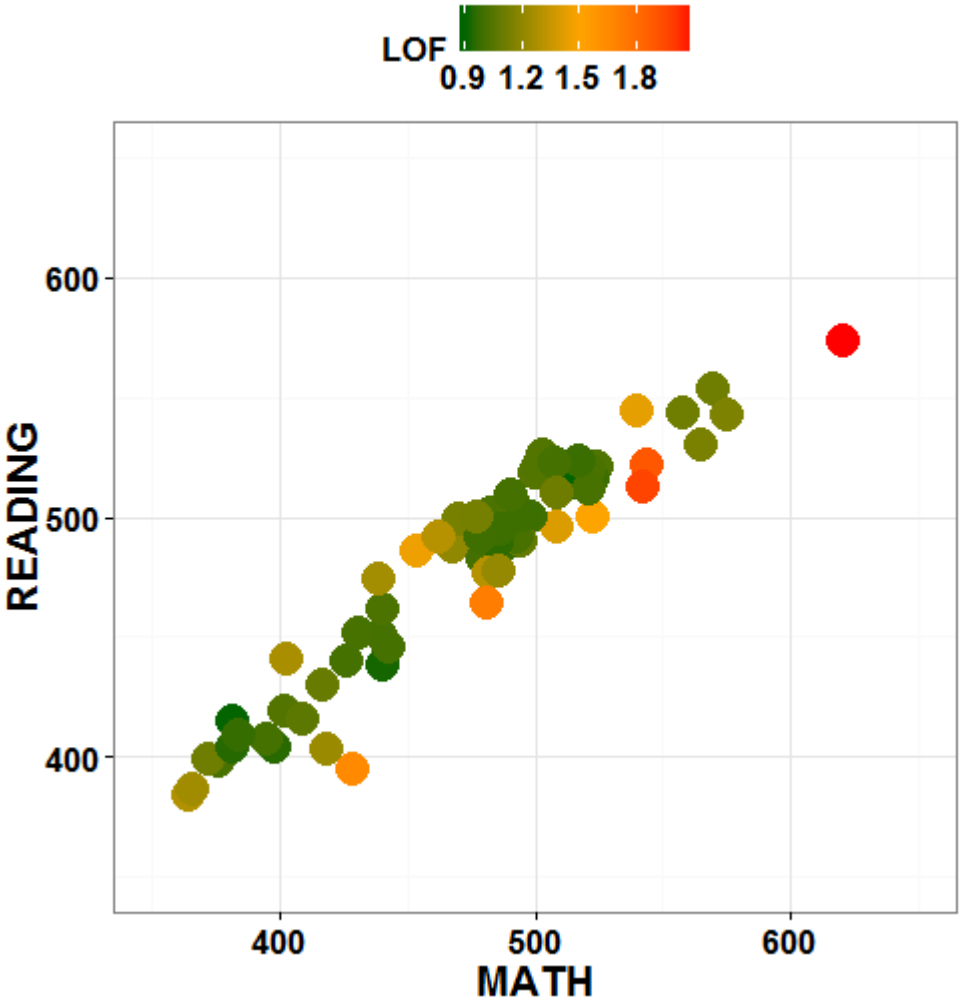
LOF

- Local Outlier Factor

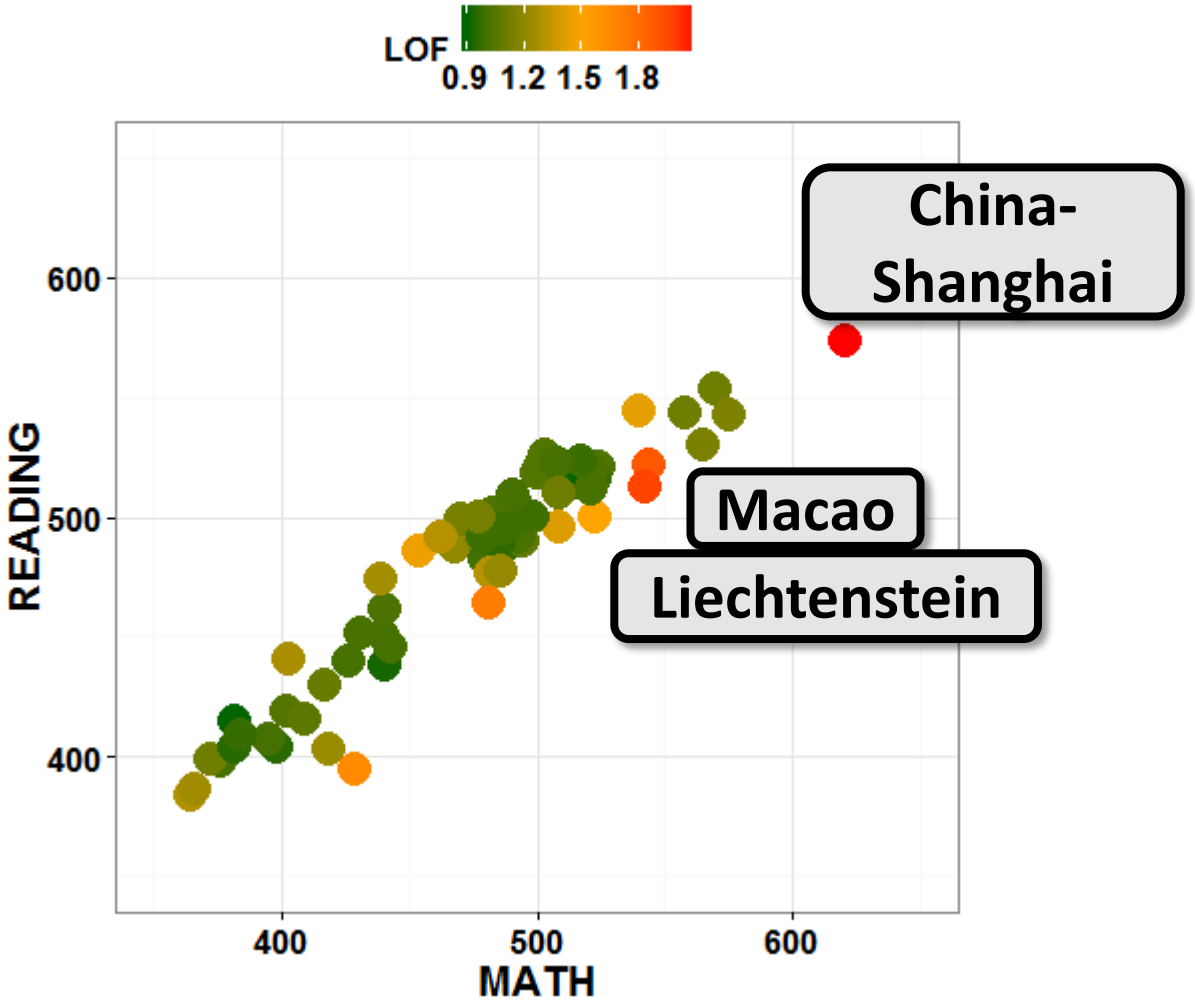
- Alapötlet
  - csak a szomszéd számít
  - a távolság is módosul
  - lokális sűrűség

- Outlier kritérium
  - a lokális sűrűség jóval kisebb, mint a szomszédaimnak átlagosan

LOF

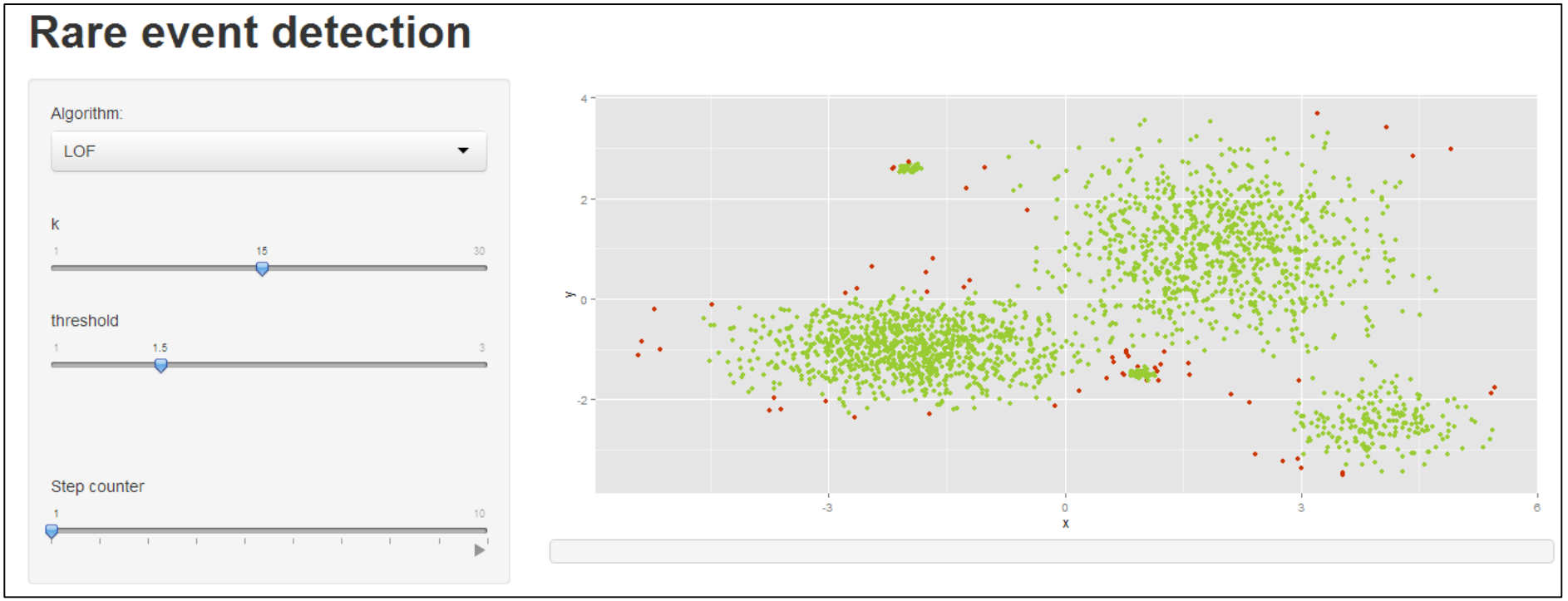


LOF


LOF
0.9 1.2 1.5 1.8
China-
Shanghai
Macao
Liechtenstein
READING
600
500
400
400
500
600
MATH

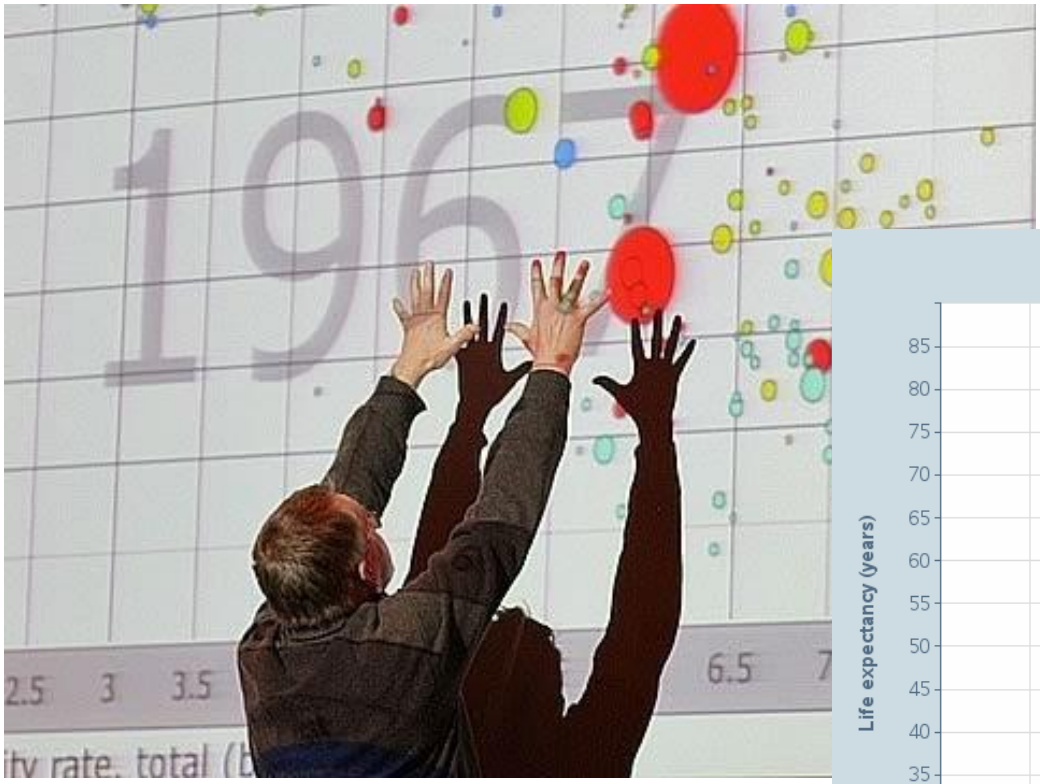
DEMO LOF

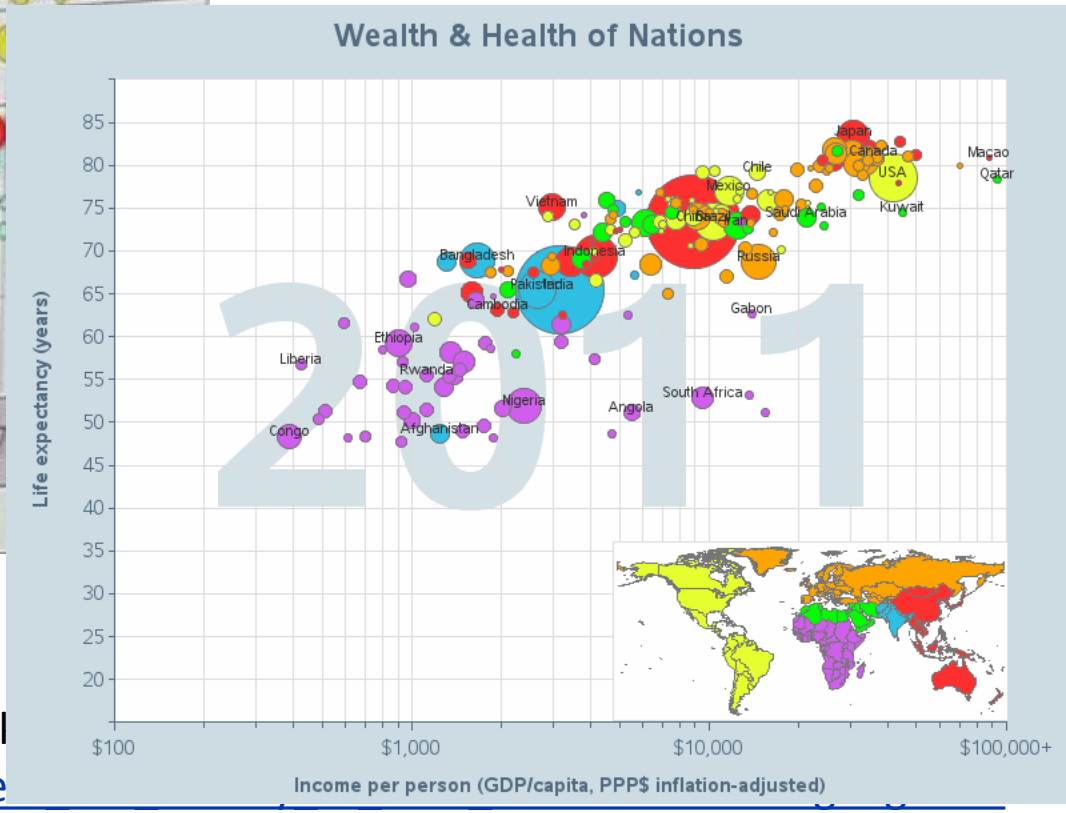
- Csomag: *DMwR* (Data Mining with R)
- Függvény: *lofactor*
- Paraméterek: k szomszédság mérete



Esettanulmány: wealth & health of nations

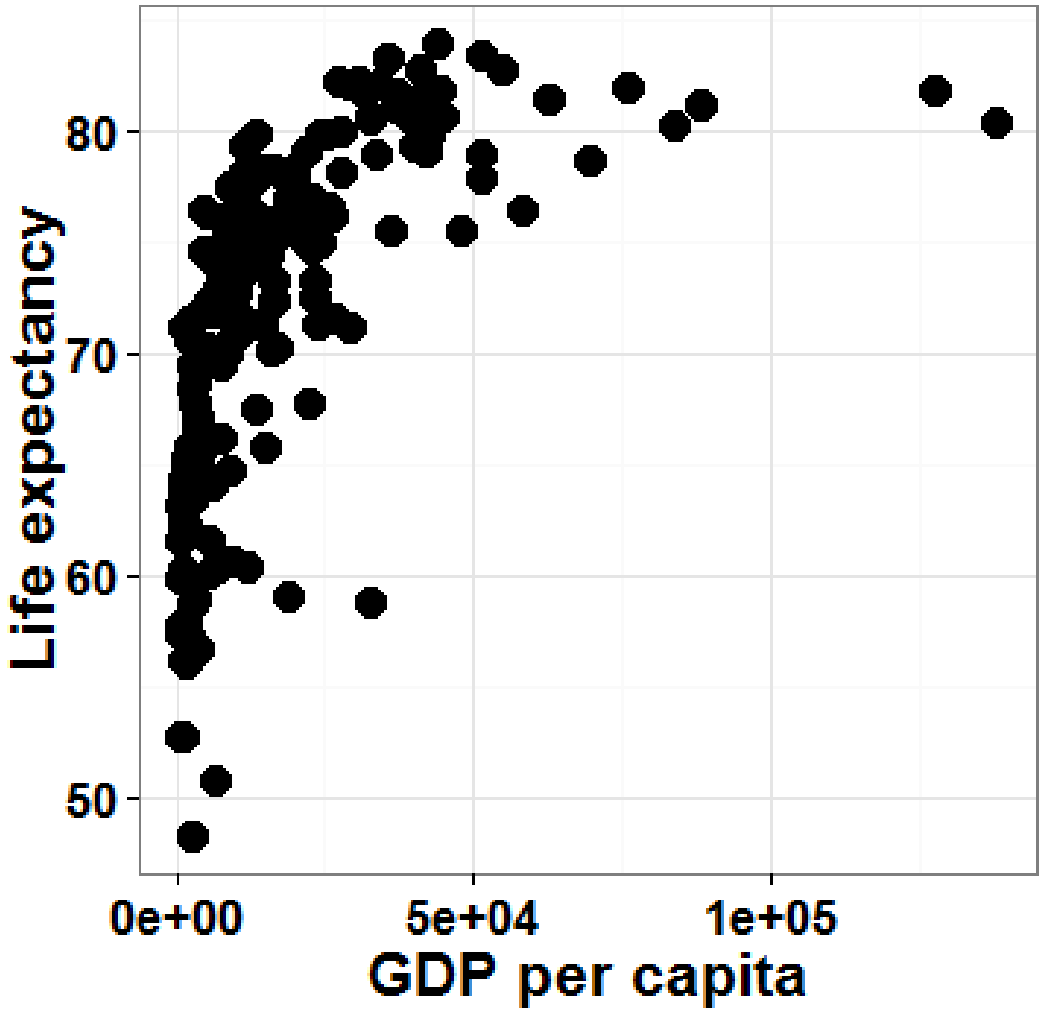
- Hans Rosling 2006-os TED talkja



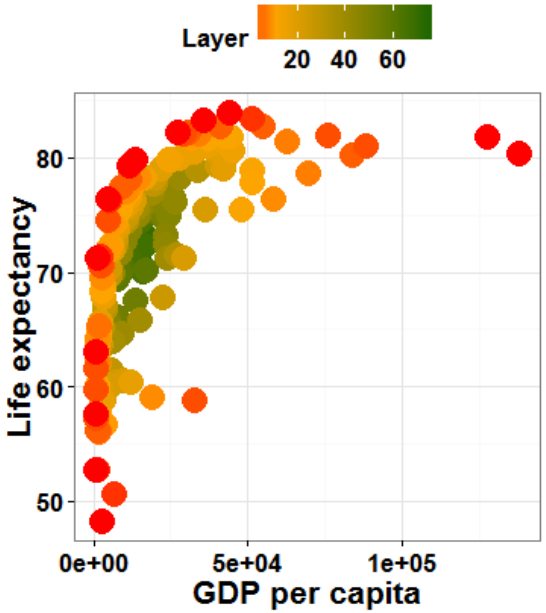


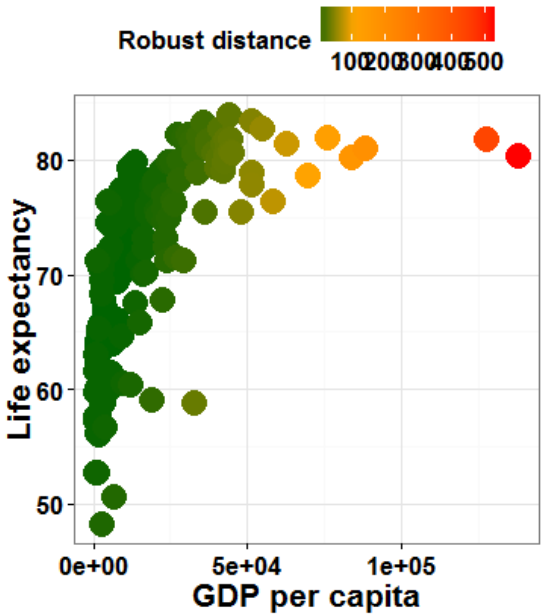
A másik teljesen irreleváns kedvenc TED tall
http://www.ted.com/talks/david_mccandle

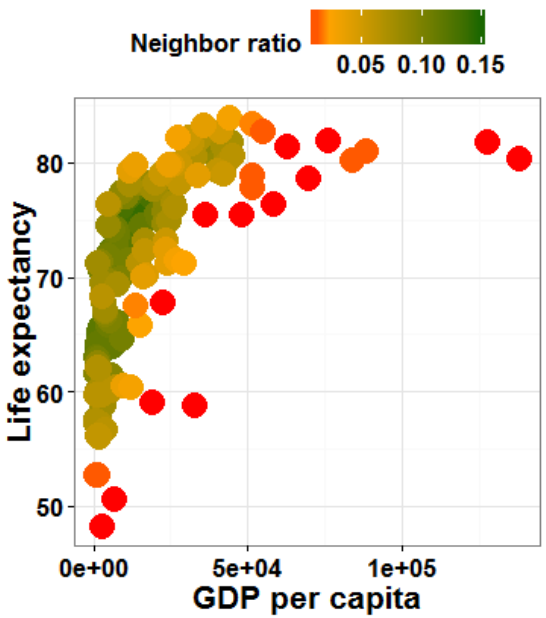
Wealth and health of nations

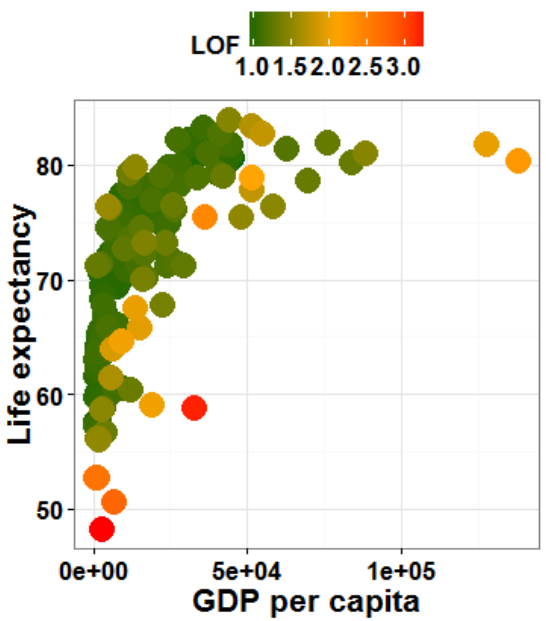


Wealth and health of nations



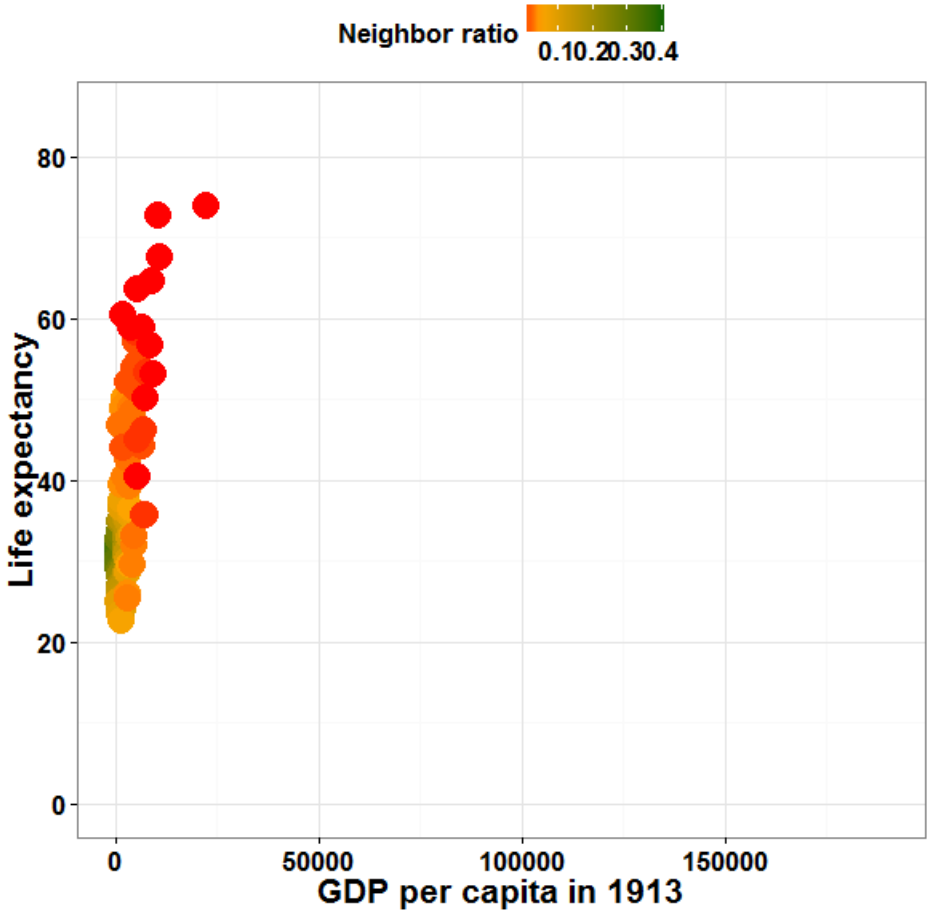


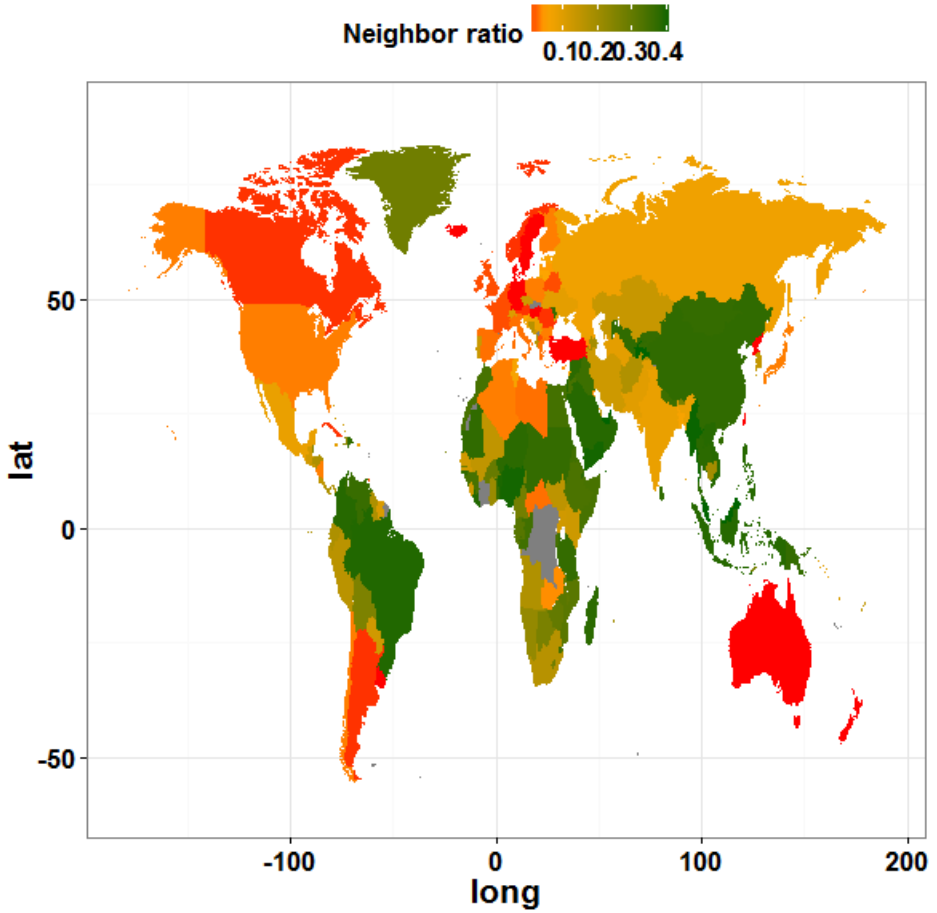




Happy families are all alike;
every unhappy family is unhappy **in its own way**.

Wealth and health of nations időbeli változás





Hivatkozásjegyzék

- Outlier Detection alapmű
  - Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3):15, 2009
- Outlier Detection demo
  - https://github.com/FTSRG/BigDANTE/tree/master/RareEventDetectionDemo