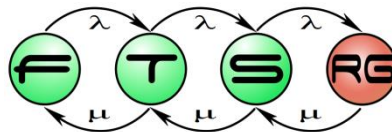


Mintavételezés, szűrés, kilógó esetek detektálása

Salánki Ágnes

salanki@mit.bme.hu

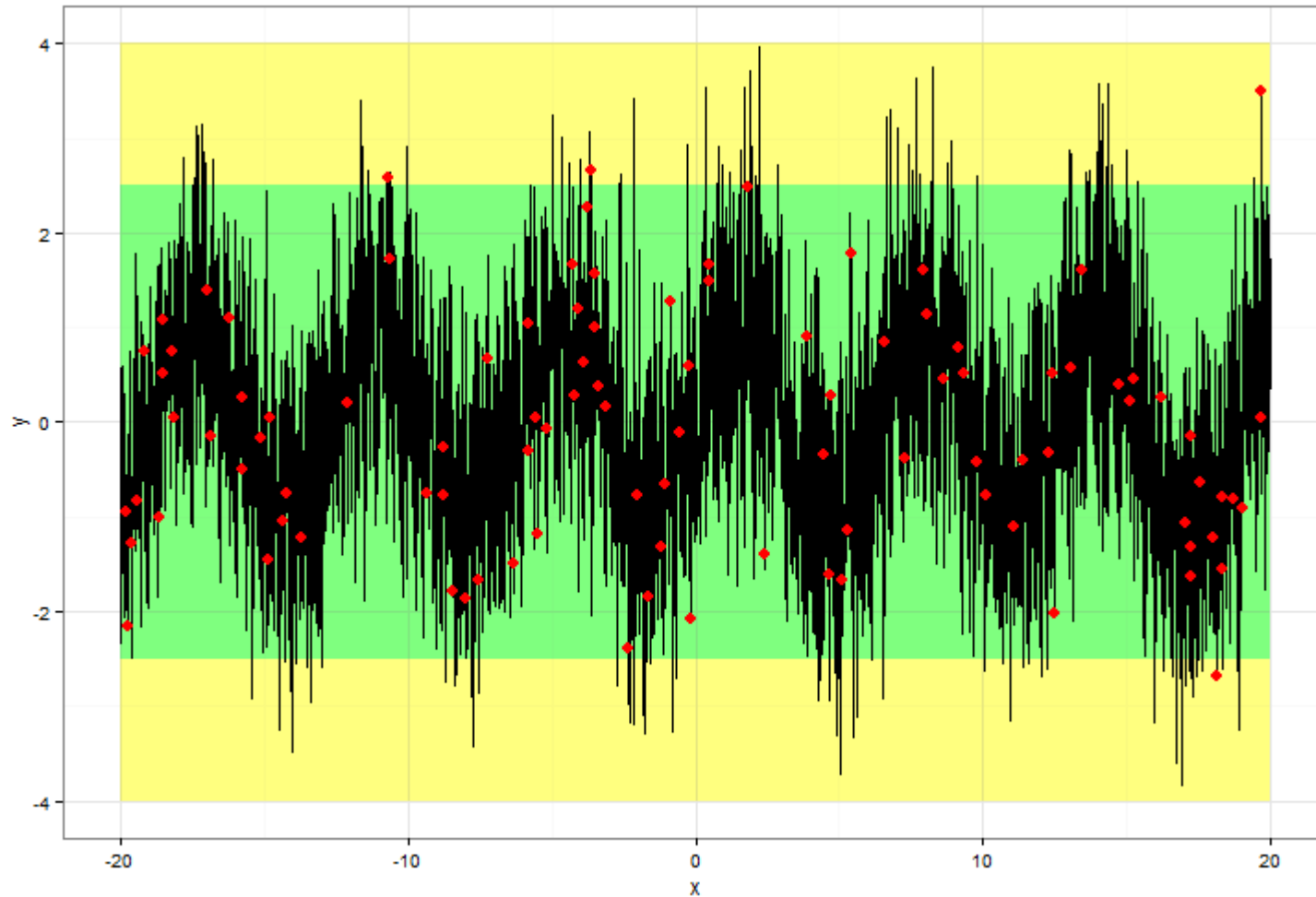
Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Alapfogalmak

■ Az alapfeladat ugyanaz

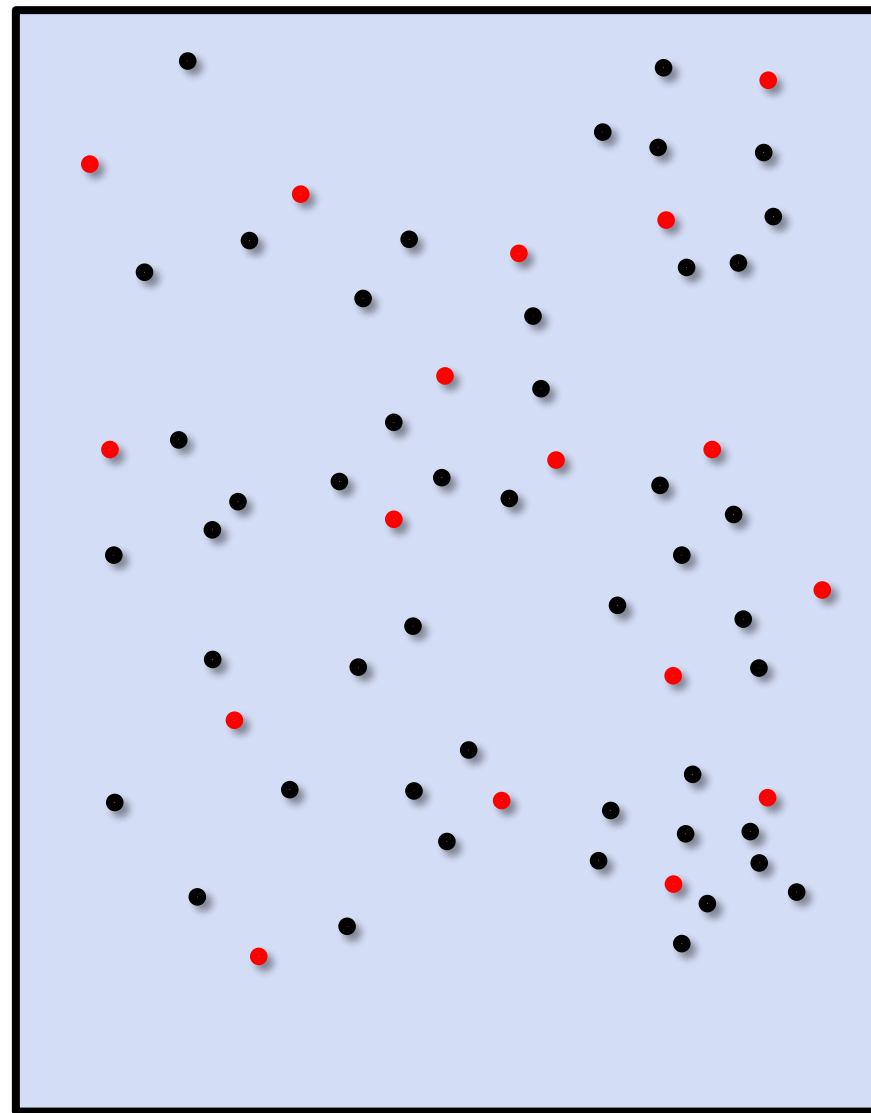
■ Az



MINTAVÉTELEZÉS

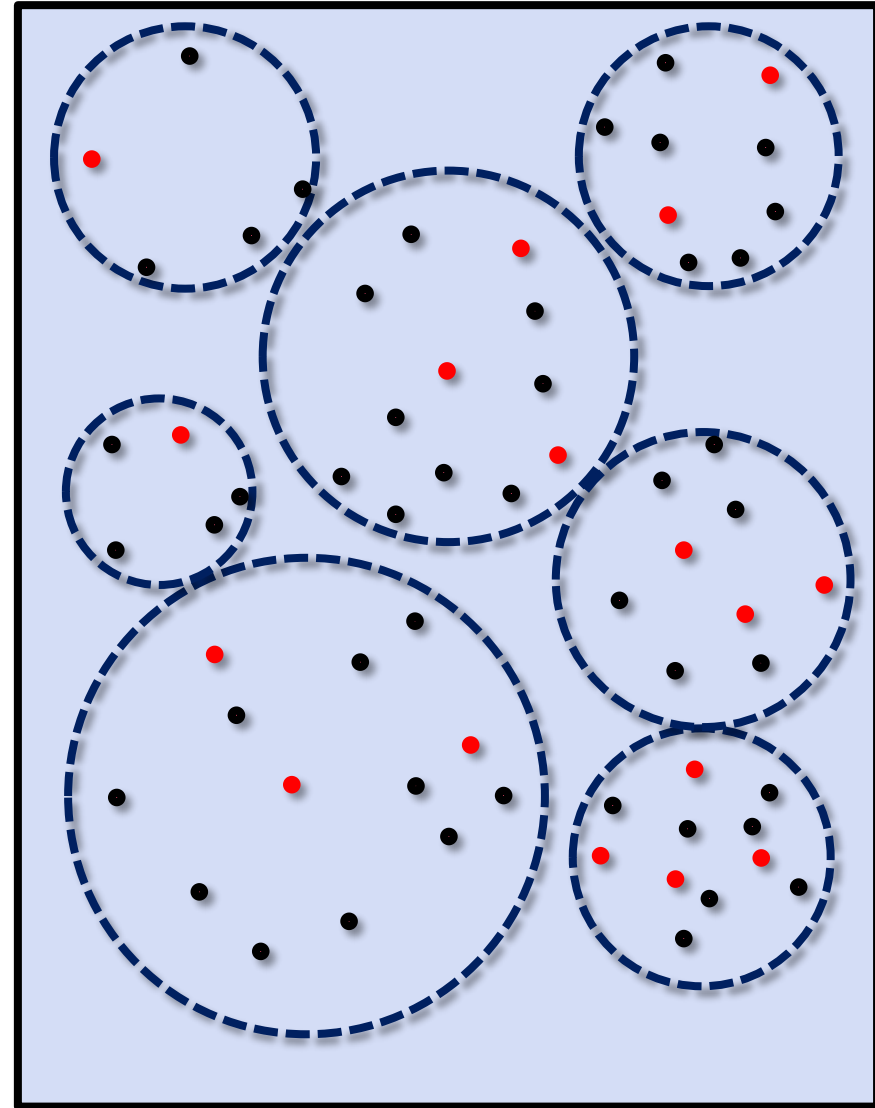
Mintavételezés

- SRS
 - Simple Random Sample
 - random mintavétel
- Stratified Sample
- Cluster sample



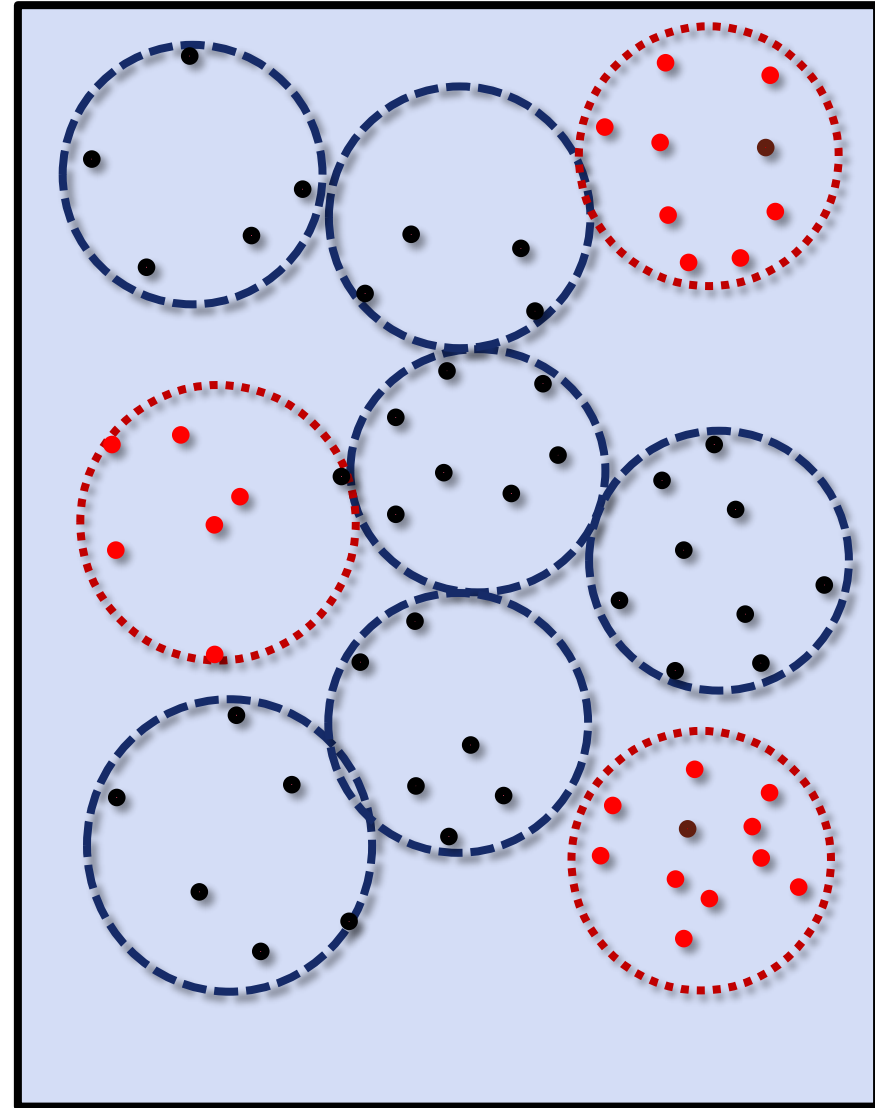
Mintavételezés

- SRS
 - Simple Random Sample
- Stratified Sample
 - Homogén „réteg”
 - Mindegyikből random m.
- Cluster sample

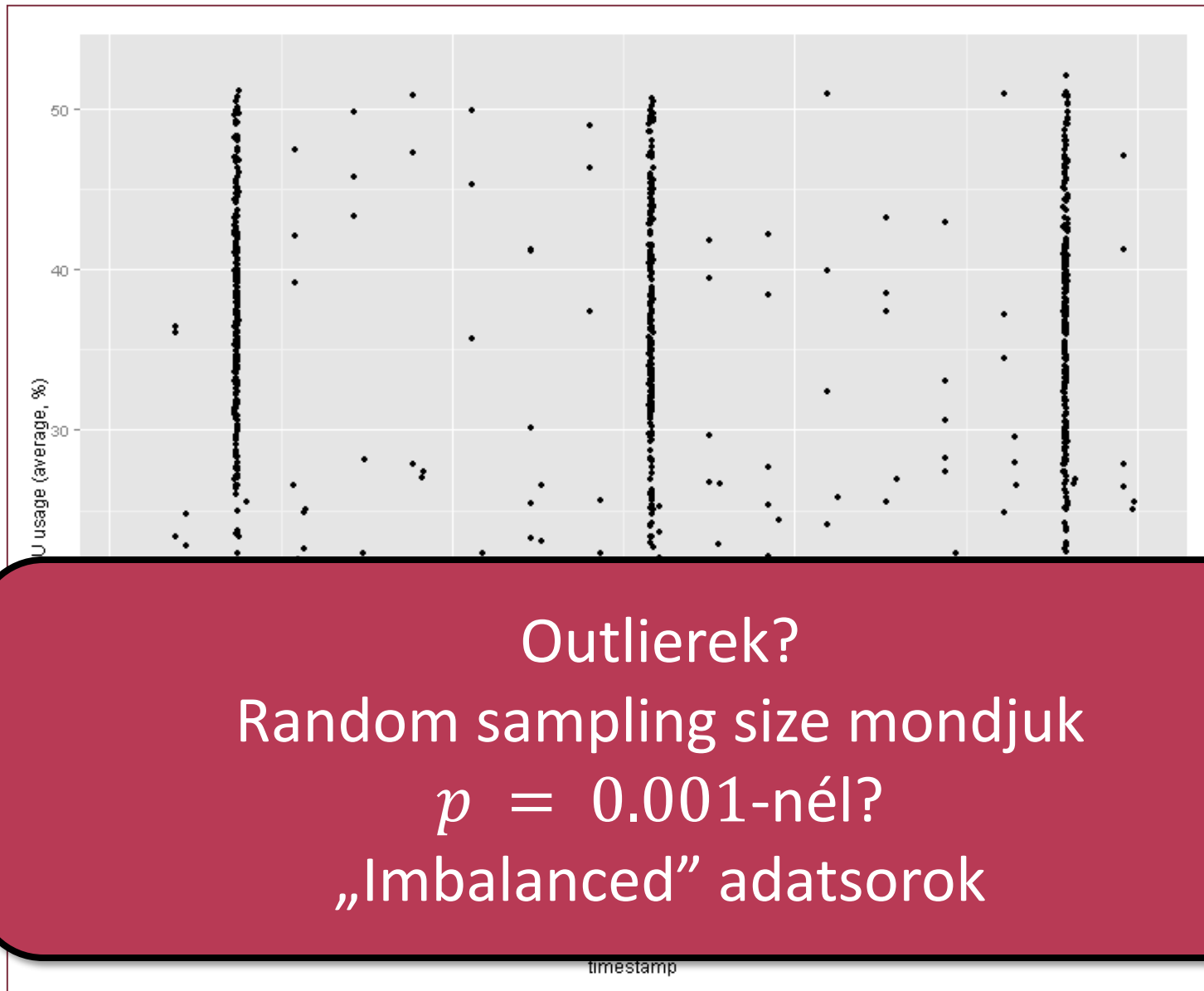


Mintavételezés

- SRS
 - Simple Random Sample
- Stratified Sample
- Cluster sample
 - ~azonos méretű klaszterek
 - Azokból random m.



Idősoroknál



Outlierek?

Random sampling size mondjuk

$p = 0.001$ -nél?

„Imbalanced” adatsorok

SZŰRÉS

Definíciók

- Jelfeldolgozás – frekvencia
- Fizika – hullámhosszok
- Matematika – részhalmaz
- Industrial rock – ??
- Számítástechnika
 - Értékek, jellemzően data streamben
 - Mail
 - Tartalom
 - Wordfilter



Adatfolyamok

Egyszer streamenként:
„Lokális maximum?”

Globális kérdések:
„Minden új maximumot
jelezzünk”

... 1, 5, 2, 7, 0, 9, 3

... a, r, v, t, y, h, b

... 0, 0, 1, 0, 1, 1, 0
time
→

Streams Entering

Ad-Hoc
Queries

Standing
Queries

Processor

Output

Buffer, megengedett
számítási memória
igény korlátos

Limited
Working
Storage

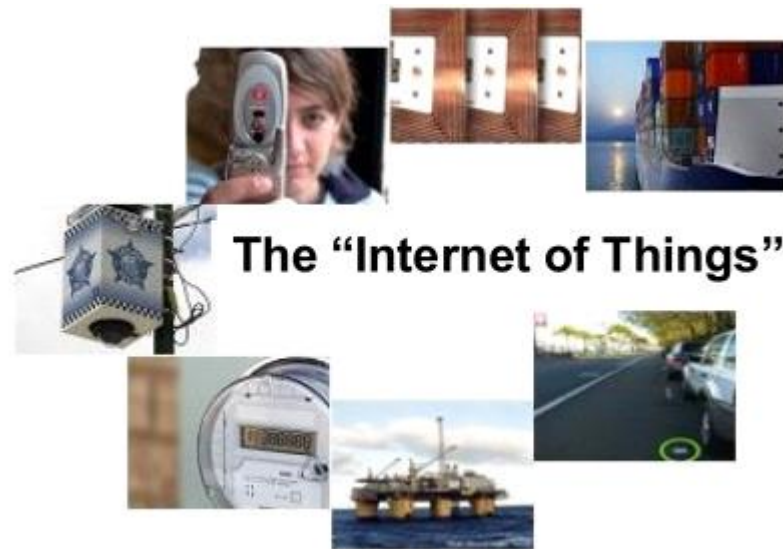
Archival
Storage

1. több forrásból,
2. ismeretlen sebességgel

Ábra és a számértékes példák forrása: [1]

Adatfolyam-források

- Szenzor-adatok
 - 1 millió szenzor x 10/s x 4B
- Képek
 - Szatelitek: n TB/nap
- Internetes szolgáltatók
 - „A legutóbbi órában melyik volt a legnépszerűbb weboldal”
- Hálózati forgalom
- Tőzsdei adatok



Traditional Computing



Fact finding with data at rest

Streams Computing



Insights from data in motion

Feldolgozás: időkorlát!

- Diszk nem használható
- Megengedett memóriaigény: korlátos
- Elemenkénti számítási igény: korlátos
- Szokásos megoldások:
 - Csúszóablakos tárolás és feldolgozás
 - Mintavételezés
 - Közelítő algoritmusok

Csúszóablakos tárolás

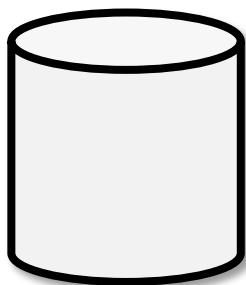
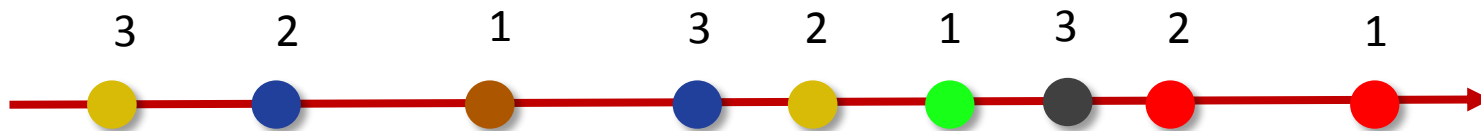
- Eltároljuk
 - A legutolsó t elemet
 - Az elmúlt T időben jövőket
 - A triggertől számító utolsókat
- Csúszóablakra tervezett algoritmusok
 - Pl. átlagszámítás

Mintavételezés streamekben

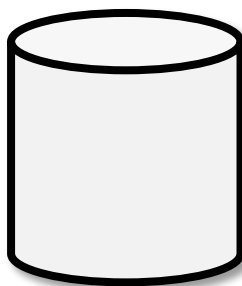
- Pl. „az elmúlt héten hány egyedi query jött?”
- Ezt kb. 10% minta alapján

Random mintavételezés

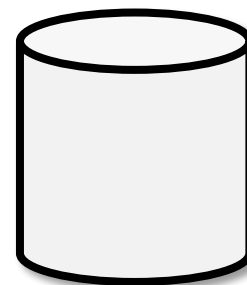
- 1/3-os mintavételezés



$$\hat{p} = 1.0$$



$$\hat{p} = 1.0$$



$$\hat{p} = 1.0$$

Mintavételezés streamekben

- Pl. „az elmúlt héten hány egyedi query jött?”
- Ezt kb. 10% minta alapján

- Random

- Ha té
- Ha ké
- Stb.

Nem tudunk a minta alapján
általánosítani a teljes streamre



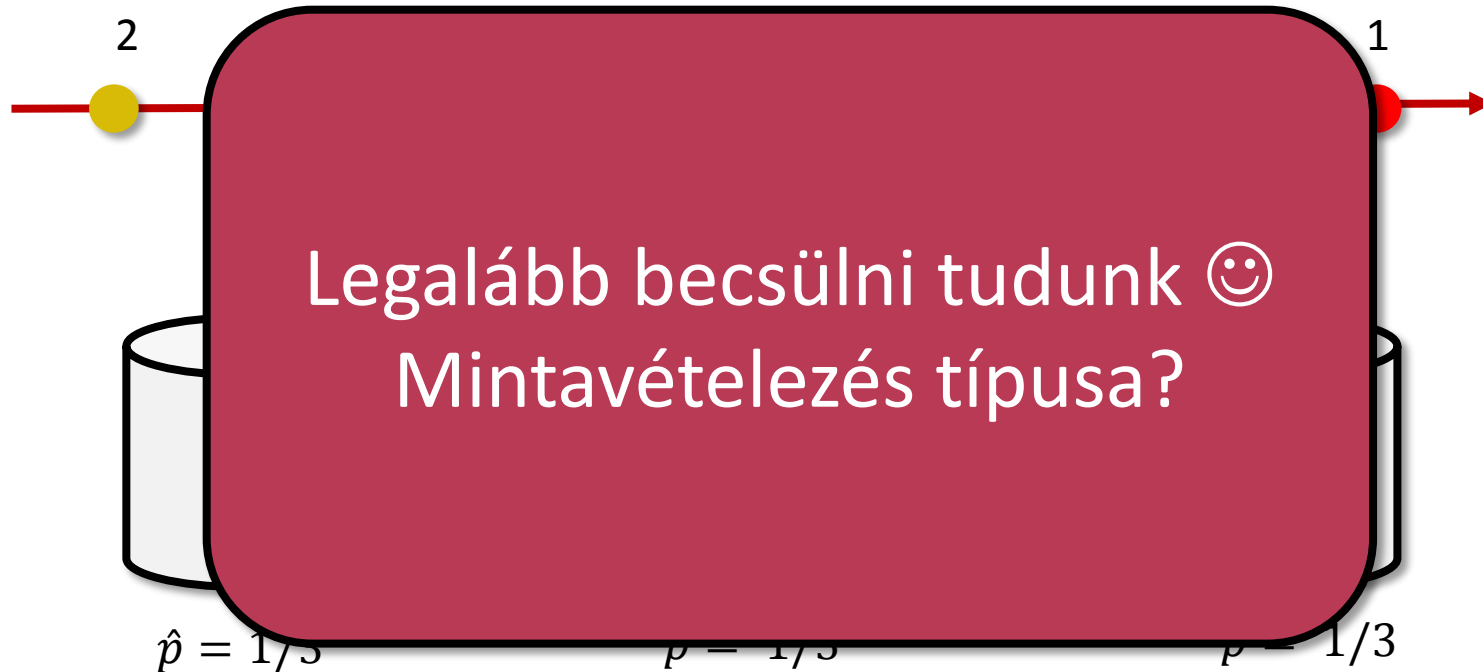
ntában

Mintavételezés streamekben: Hash

- Pl. „az elmúlt héten hány egyedi query jött?”
- Ezt kb. 10% minta alapján
- Érték alapján szűrünk
 - Pl. hash függvény 0-9 közé
 - Feltételezések
 - A hash egyenletes $\rightarrow$ az értékek 1/10-e kerül be a 0-ba

Mintavételezés streamekben: hash

- 1/3-os mintavételezés



Szűrés streamekben

- Pl. találjuk meg
 - a „különlegeseket”
 - az azonosokat
 - a különbözőeket

Már az első is nehéz
probléma

Bitszámlálás

- Adott: 0-1 stream
- N buffermérettel
- Ad-hoc query: „Hány 1-es van az utolsó k bitben?”
- Néha még erre sincs időnk...

Bloom filterek

- Pl. web crawling: „láttuk-e már ezt az URL-t?”
- Bloom filter
 - bitvektor
 - hash függvények (h_1, h_2 stb.)
- Folyamat: x inputra
 - BESZÚR: $h_1(x), h_2(x)$ stb. indexekre: $0 \rightarrow 1$ a vektorban
 - KERES: ha $h_1(x) \wedge h_2(x) \wedge \dots \neq 1$, akkor BESZÚR

Bloom filterek

- $N = 11$
- Input: egész számok
- $h_1(x)$: a páros bitekből képzett $y \bmod N$
- $h_2(x)$: a páratlan bitekből képzett $y \bmod N$

Bloom filterek

Folyam	$h_1(x)$	$h_2(x)$	Filter
			000000000000
BESZÚR(25 = 1 1 0 0 1)	5	2	00 1 00 1 00000
BESZÚR(159 = 1 0 0 1 1 1 1 1)	7	0	1 010010 1 000
BESZÚR(585 = 1 0 0 1 00 1 00 1 9)	9	7	101001010 1 0
KERES(118 = 1 1 1 0 1 1 0)	3	5	101 0 0 1 01010

Bloom filterek

- Persze van fals pozitív
- N mérettel
- h hash függvény
- eddig k beszűrt elemmel
 - $p(P) = \left[\frac{\#(1)}{N} \right]^h$
 - $\#(1) \leq k \times N$
- Pl. $N = 10^6, h = 5, k = 10^7$ értékekkel
 - $p(FP) = 0.00937$

OUTLIER DETEKTÁLÁS

Alapfeladat

Vannak-e
egyáltalán?

Vannak-e
egyáltalán?

Hogy néznek ki?

Szakterület
specifikus?

Hogyan
szeperálhatóak?

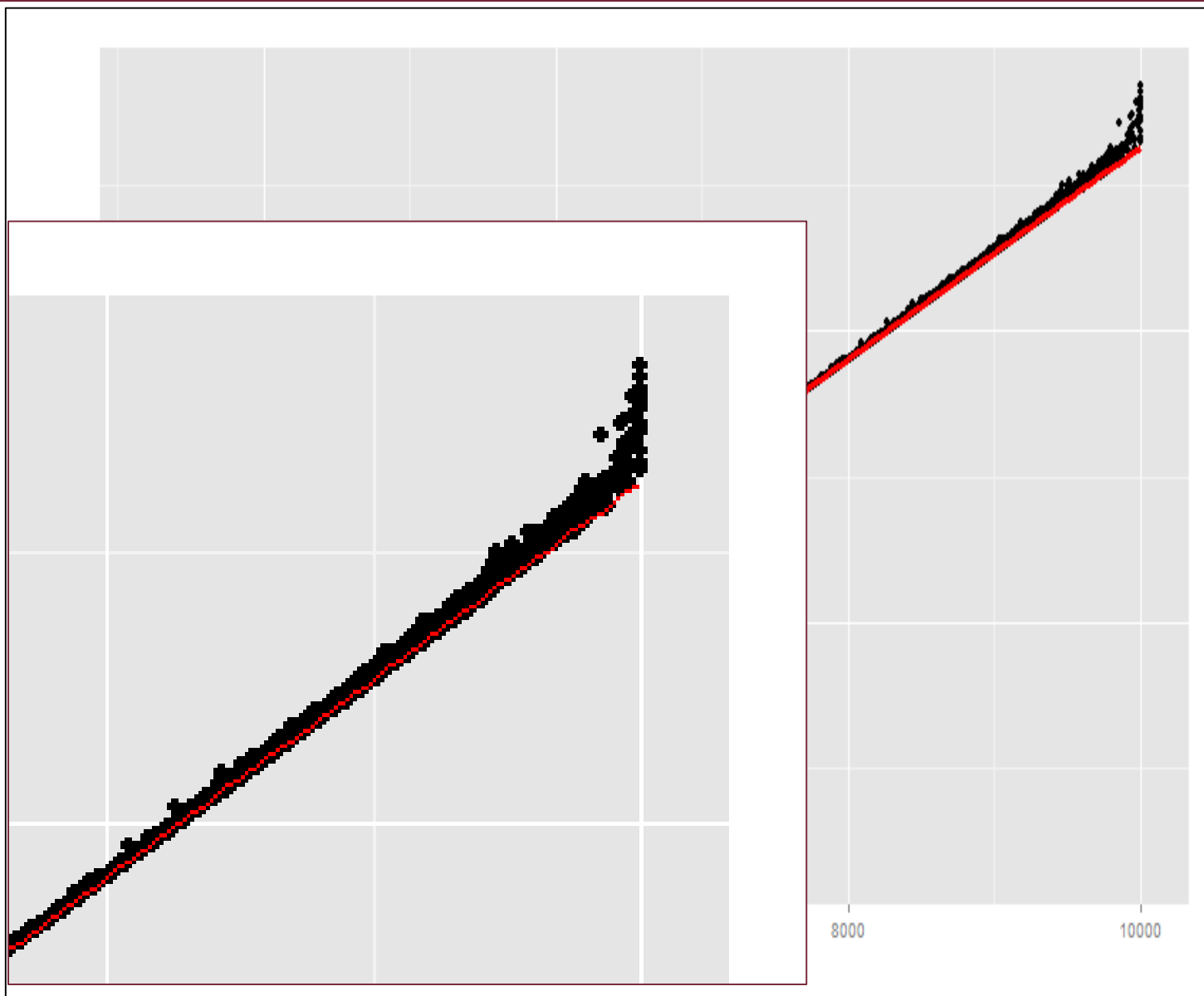
Nagy adat:
aggregálás?

Miért?

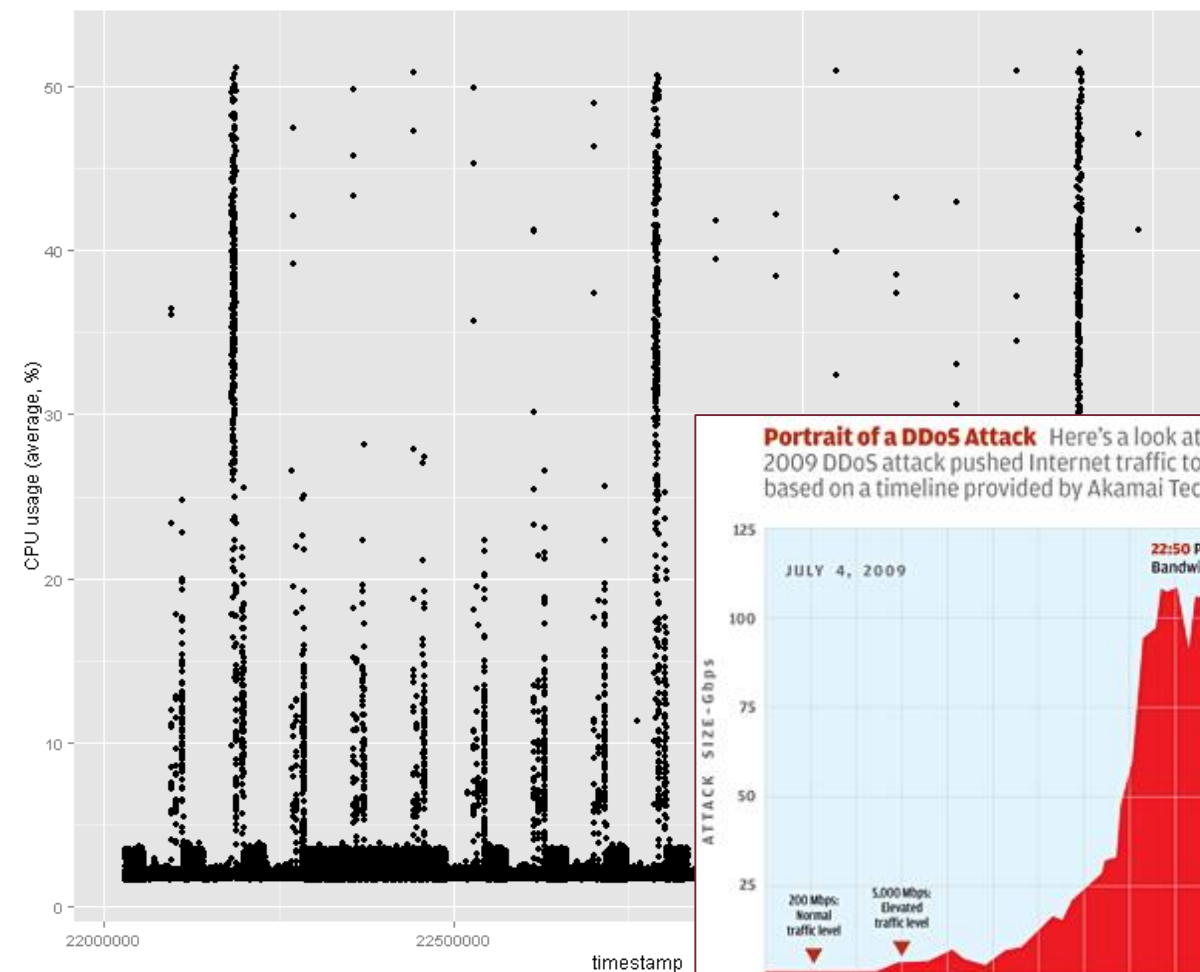
Hatások?



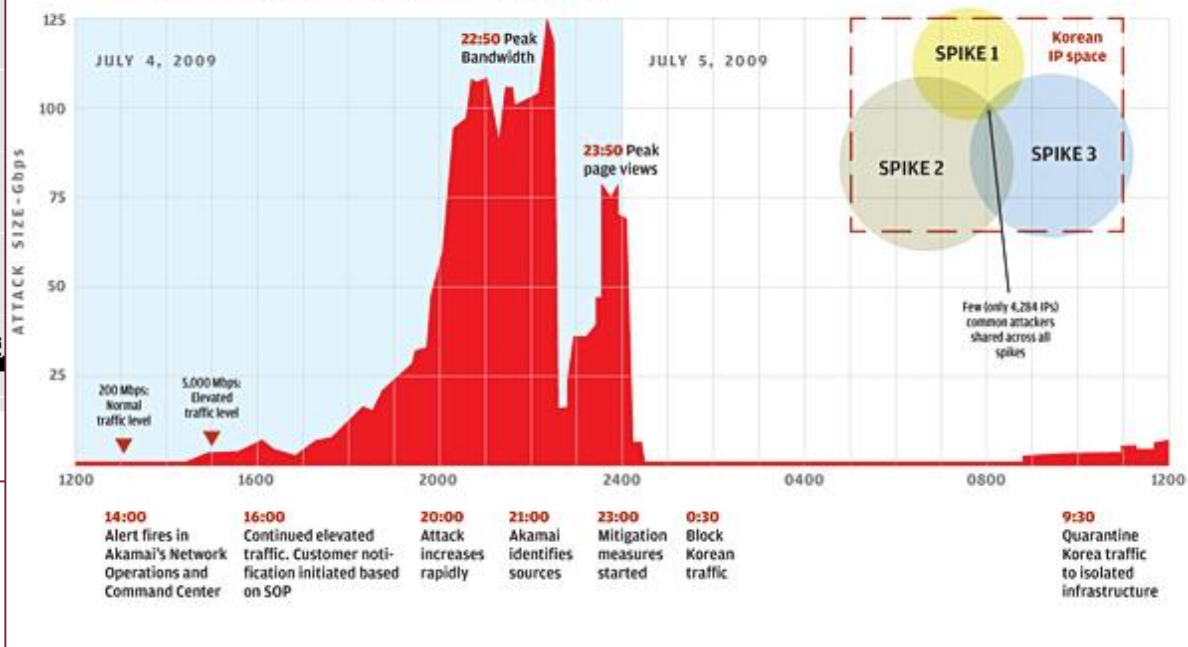
Alapfeladat



Használati esetek



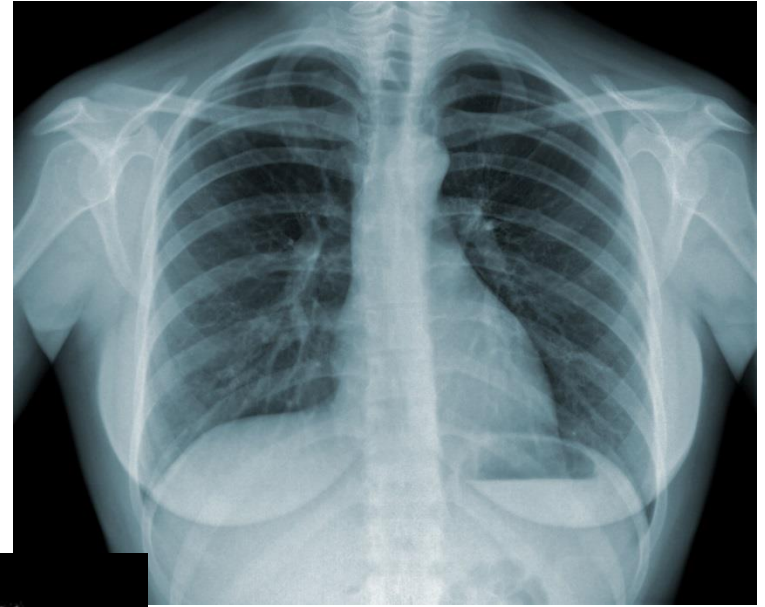
Portrait of a DDoS Attack Here's a look at how the July 4-5, 2009 DDoS attack pushed Internet traffic to the breaking point, based on a timeline provided by Akamai Technologies



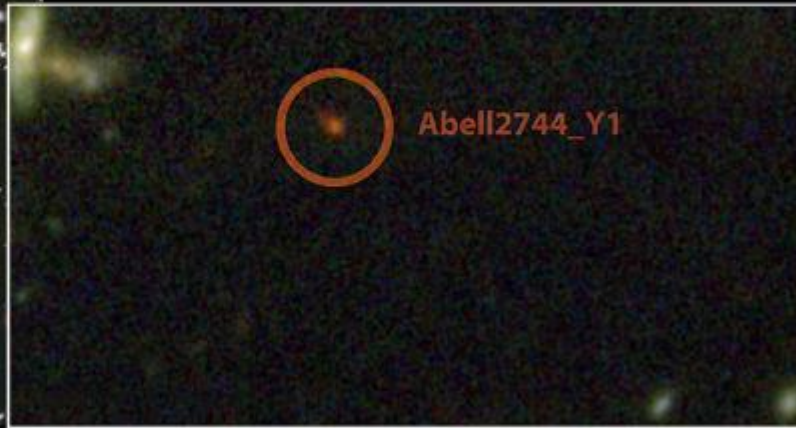
Használati esetek



Használati esetek

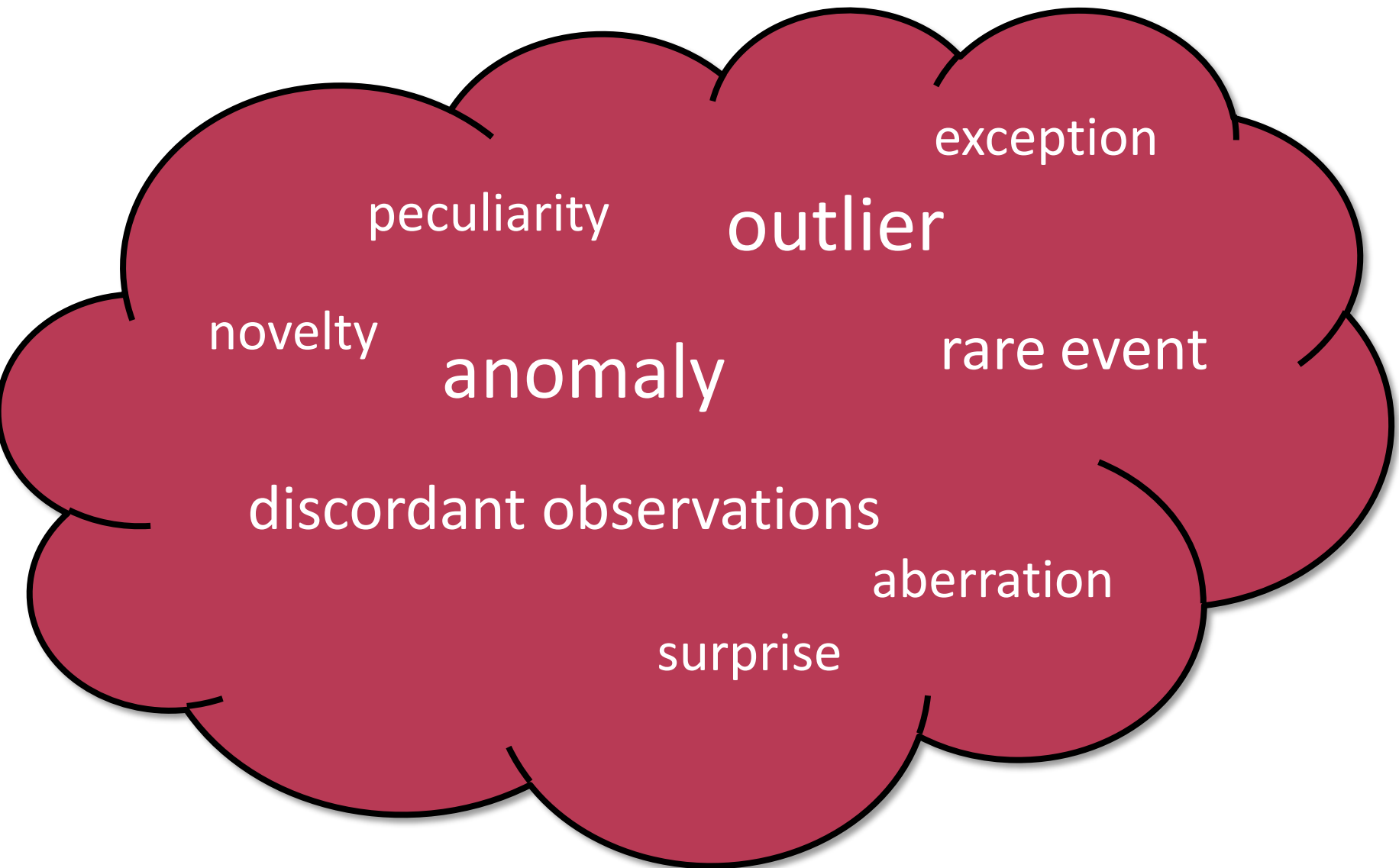


Distant Galaxy in Hubble Frontier Field Abell 2744



NASA and ESA • PRC14-17a

Alapfogalmak

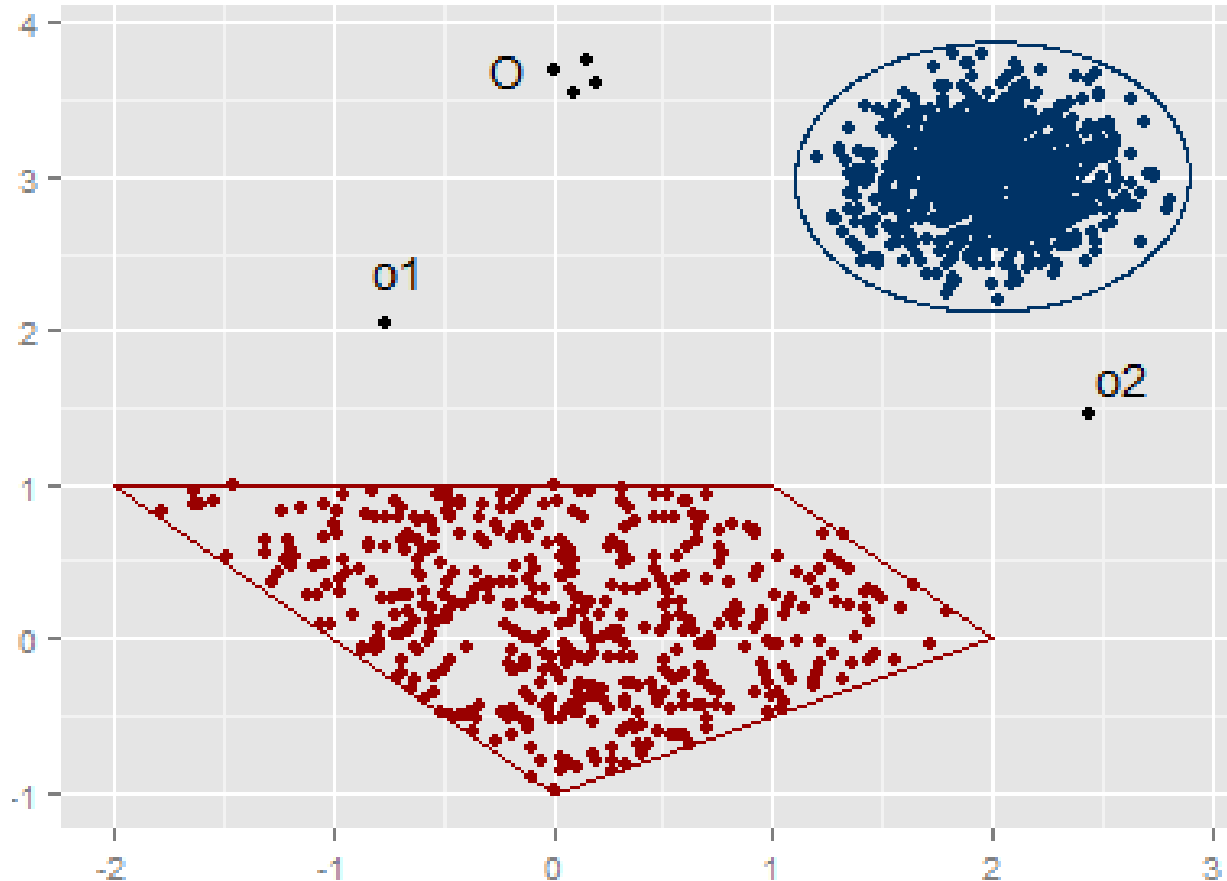


Definíció

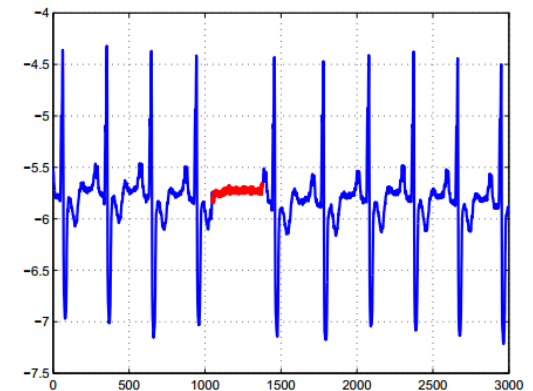
- Kevés van belőlük
- „Gyanús”, hogy más a generáló folyamat/forrás

Pont- és kollektív anomália

■ Pontanómália

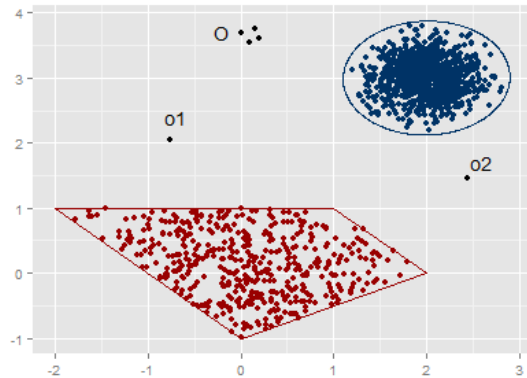


■ Kollektív anomália

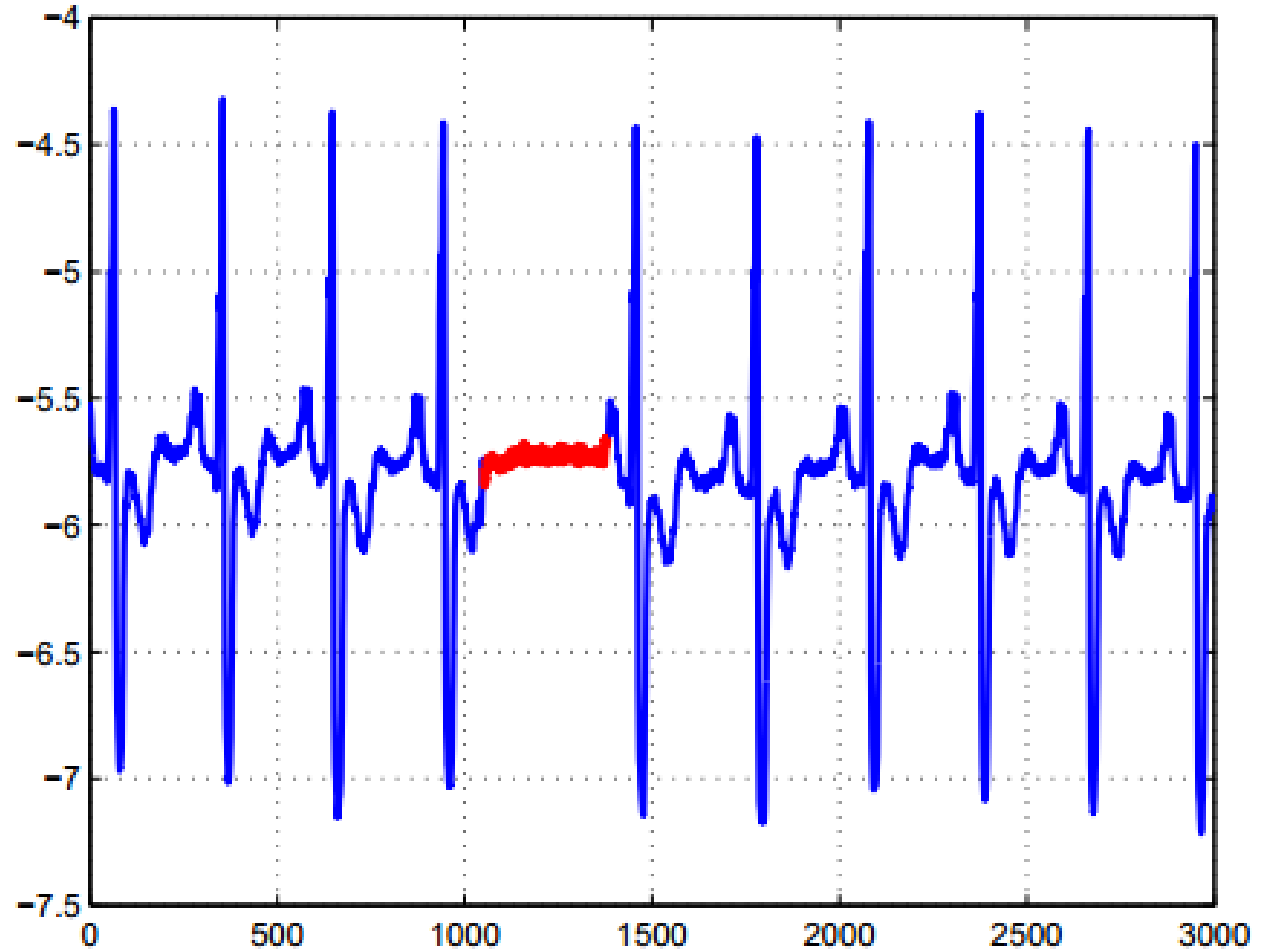


Pont- és kollektív anomália

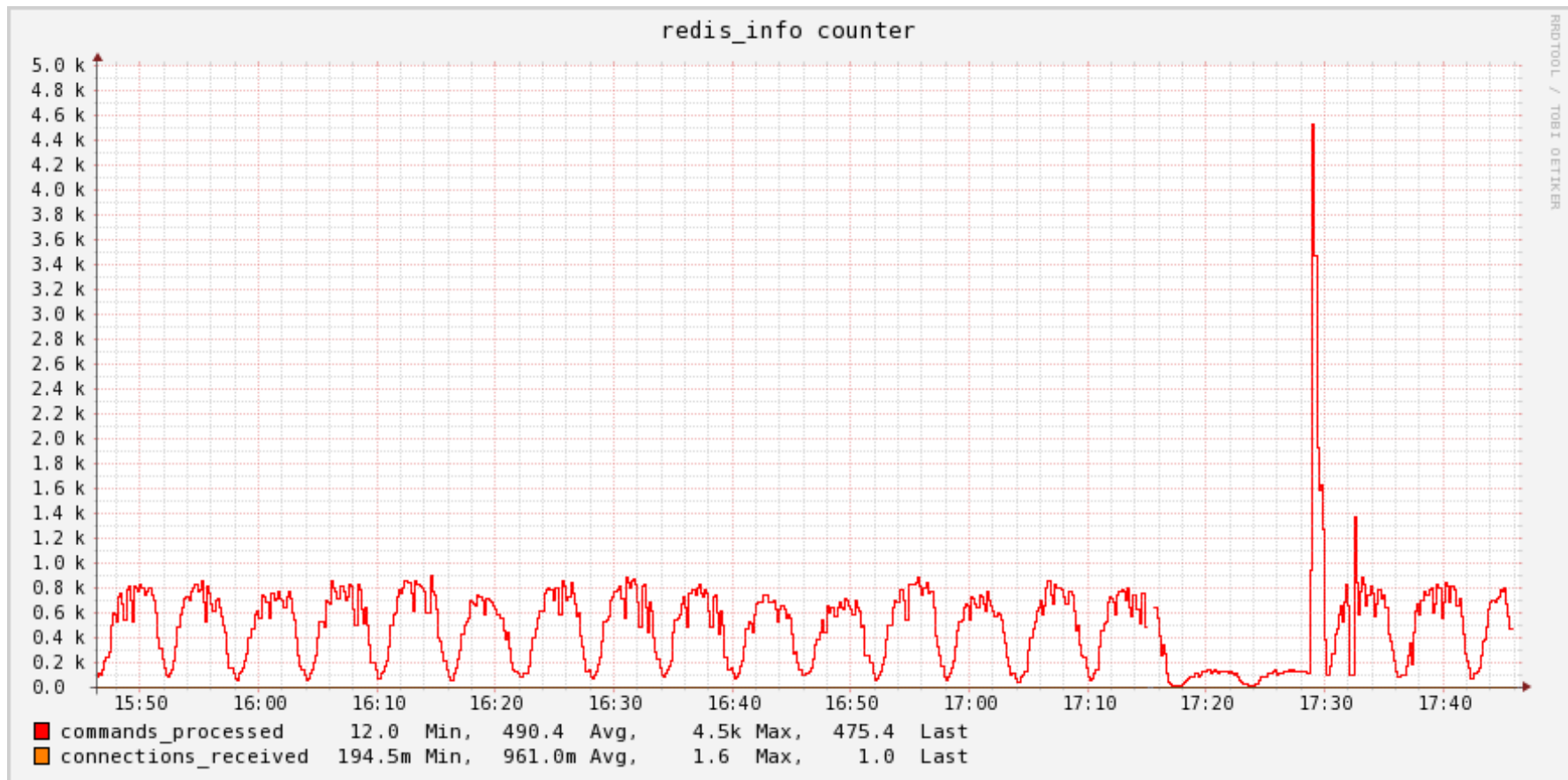
■ Pontanómália



■ Kollektív anomália



Pont- és kollektív anomália

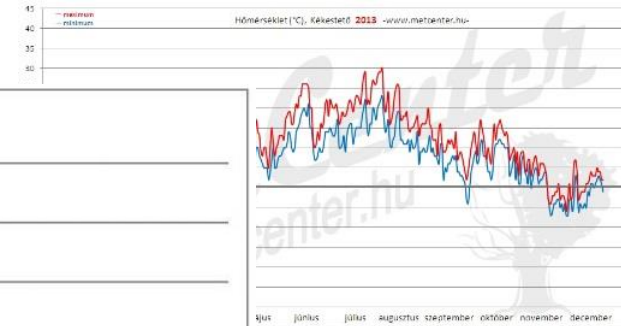
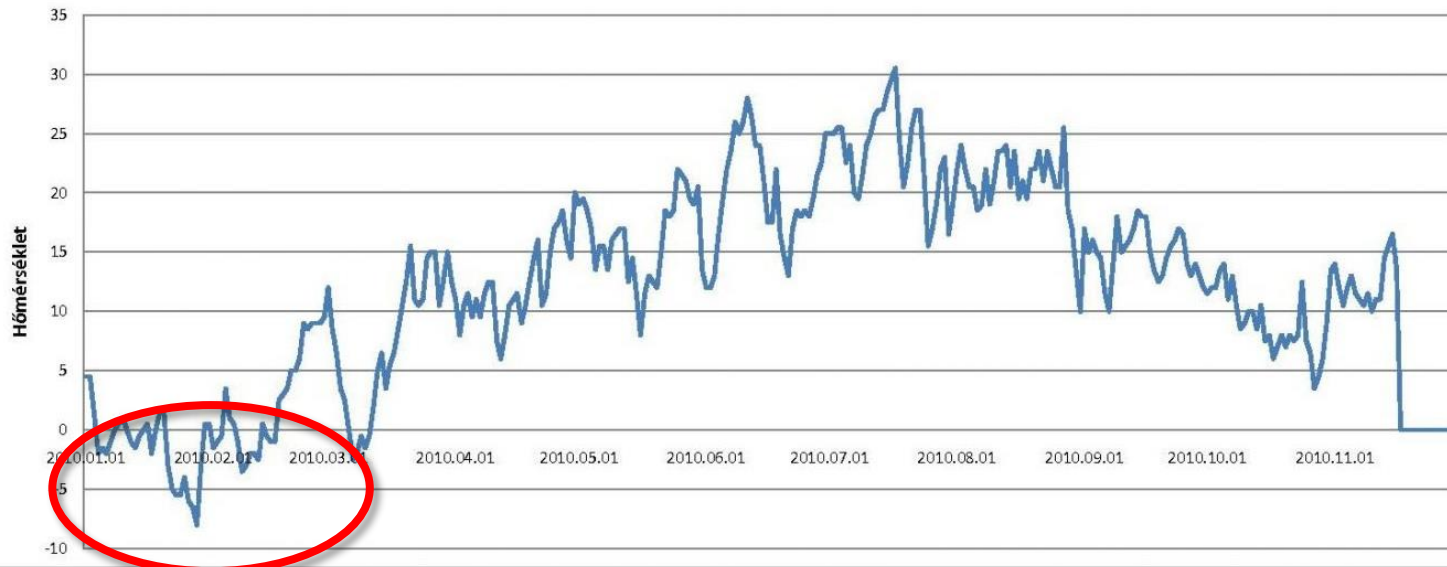


Viselkedési és kontextusanomália

■ Viselkedési

■ Kontextus

Napi átlag hőmérséglet 2010-ben (Zala megye)

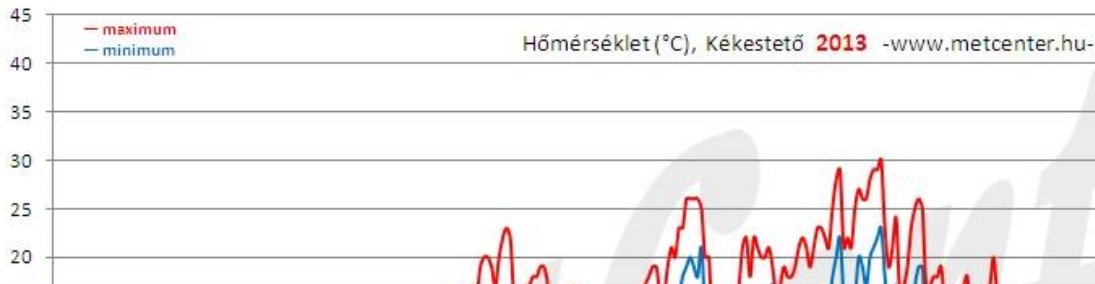


Viselkedési és kontextusanomália

■ Viselkedés:



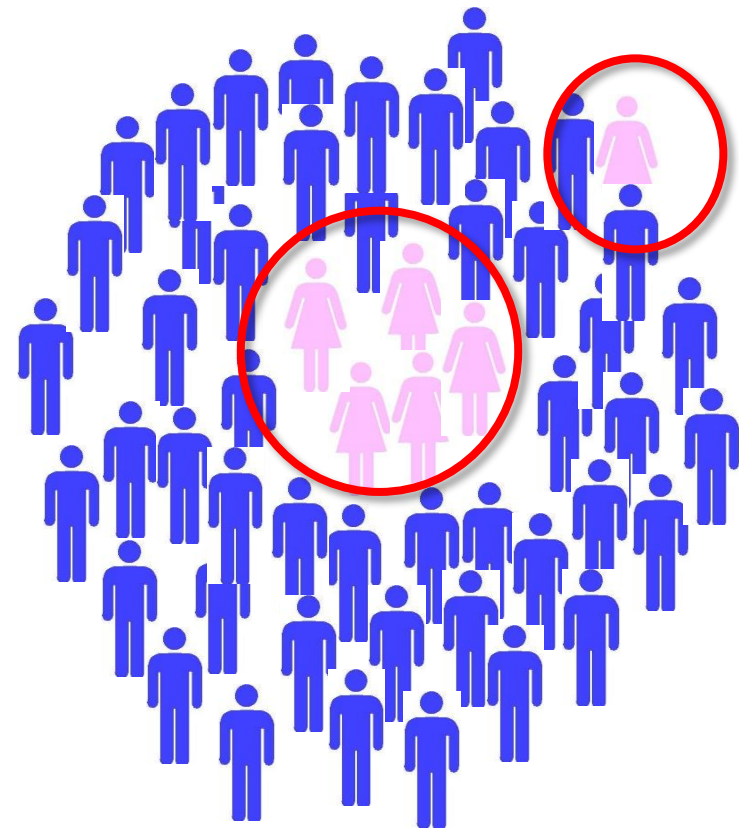
■ Kontextus



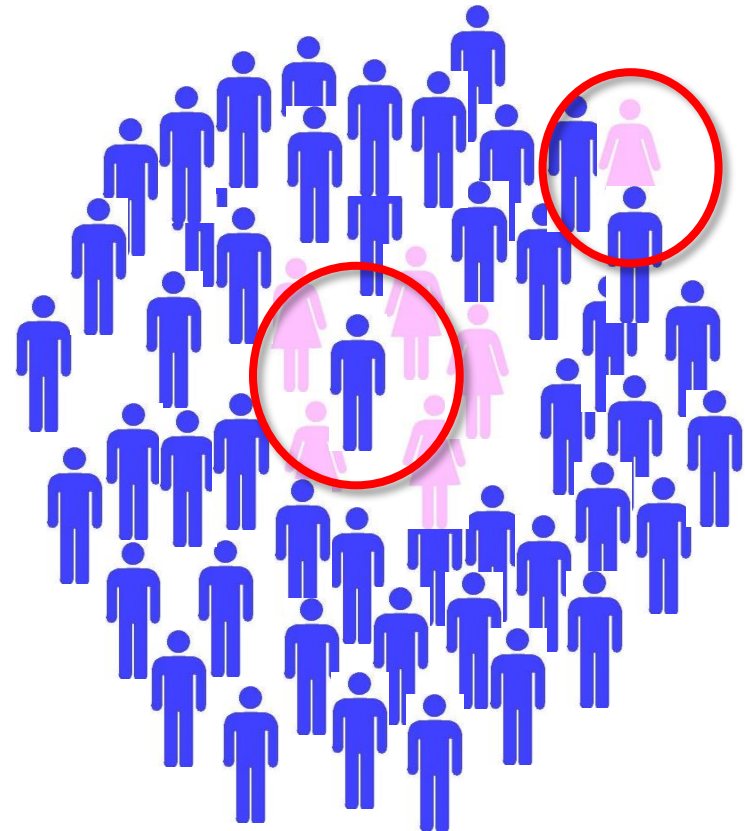
Itt: viselkedési és pontanomáliák

Megközelítések

- Távolság alapúak



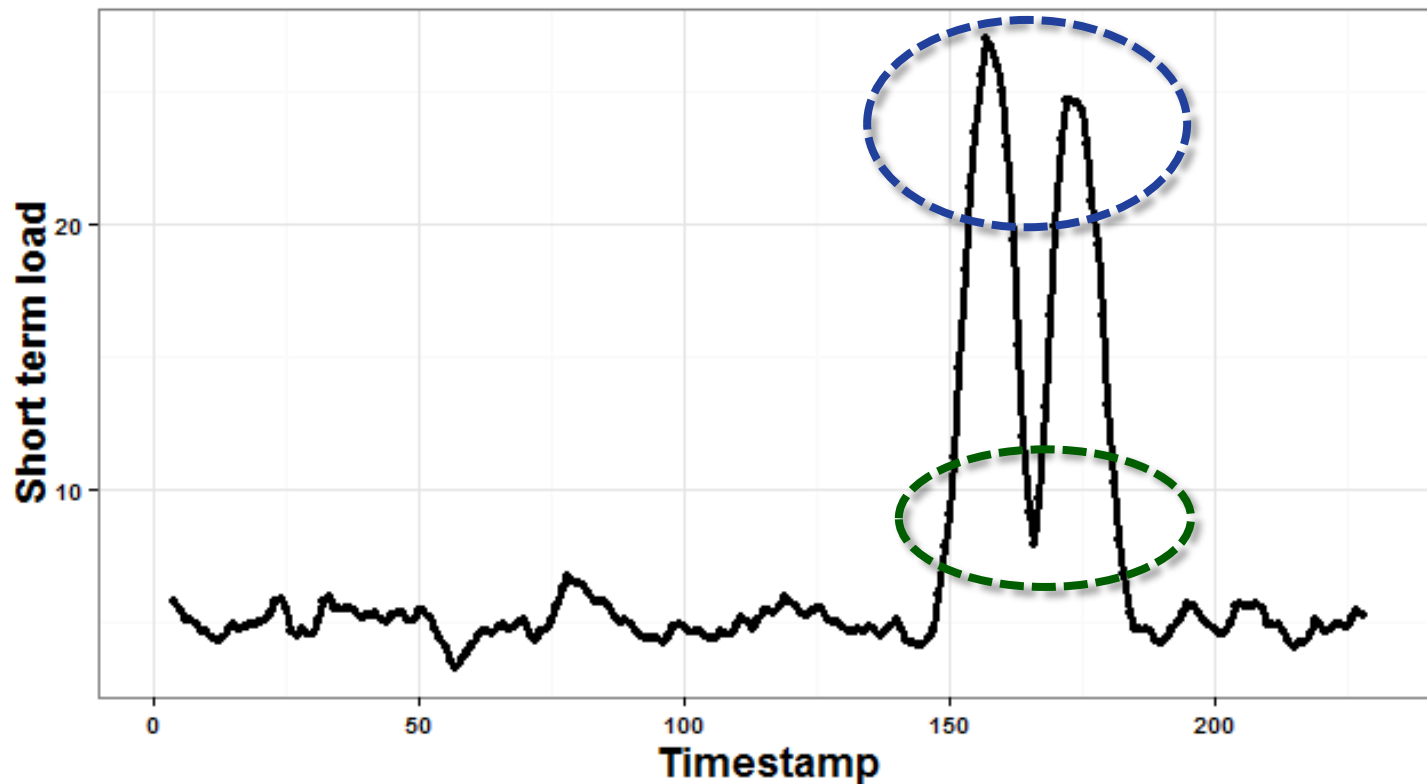
- Sűrűség alapúak



Megközelítések

- Távolság alapúak

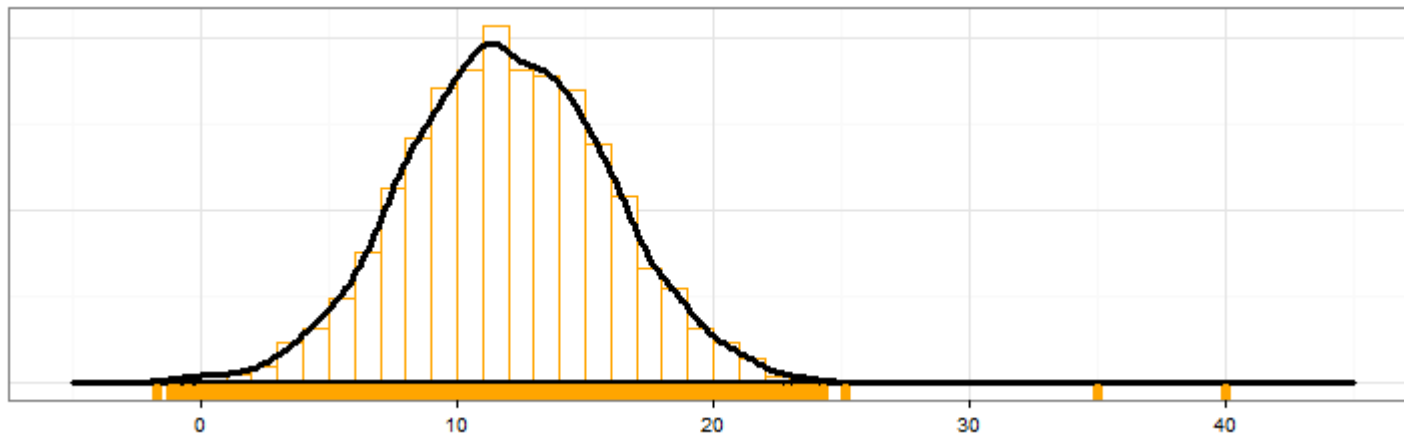
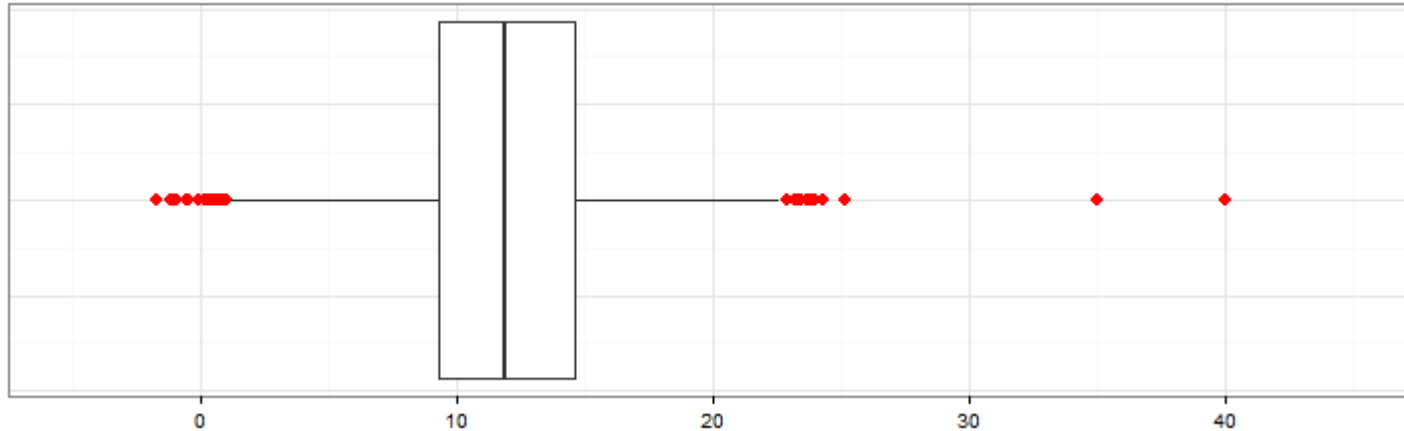
- Sűrűség alapúak



VIZUÁLIS MÓDSZEREK

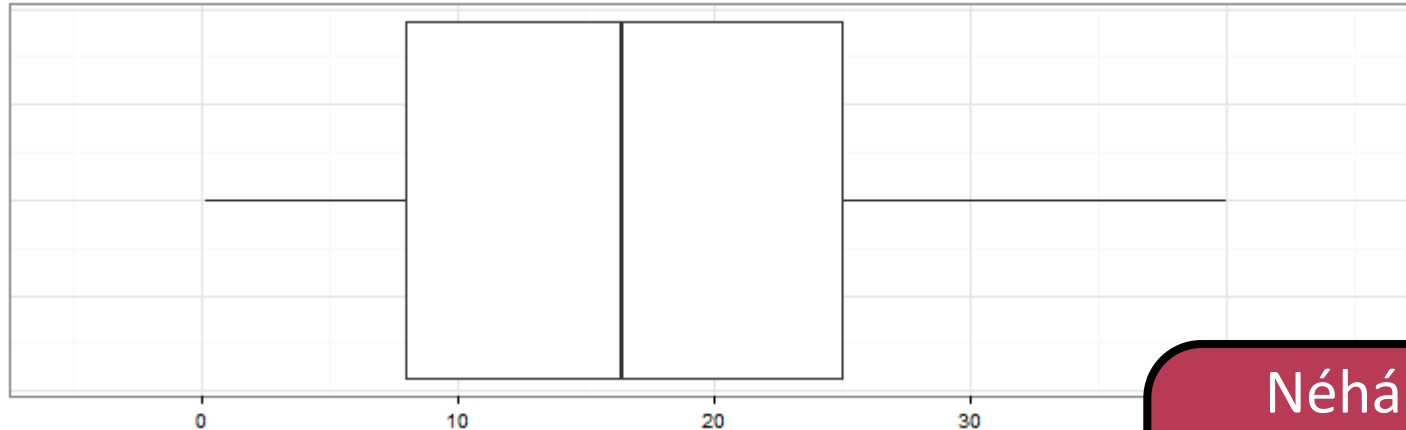
1D

- Az aggregálás általában nem jó

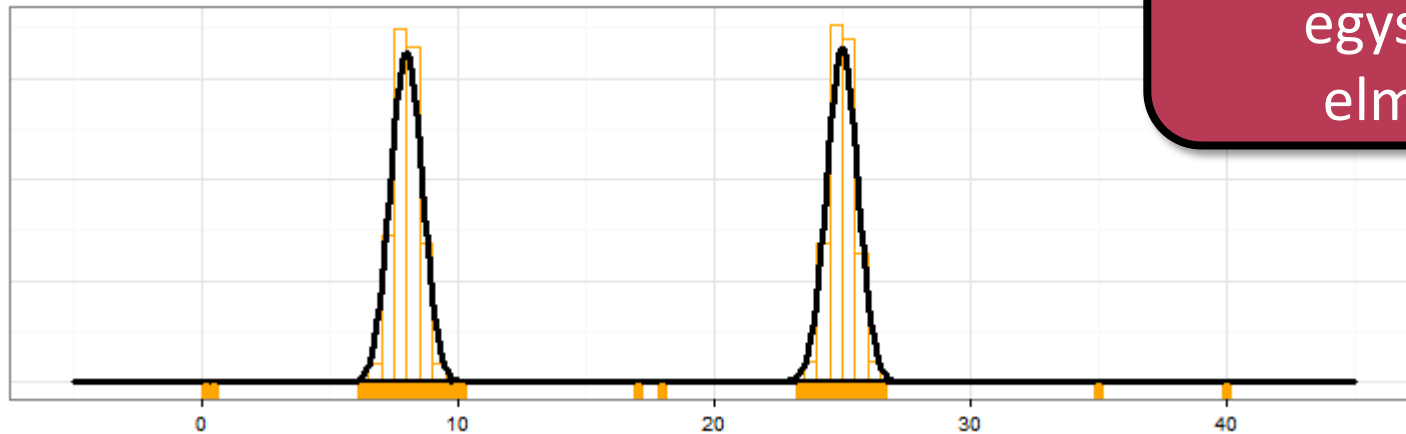


1D

- Az aggregálás általában nem jó

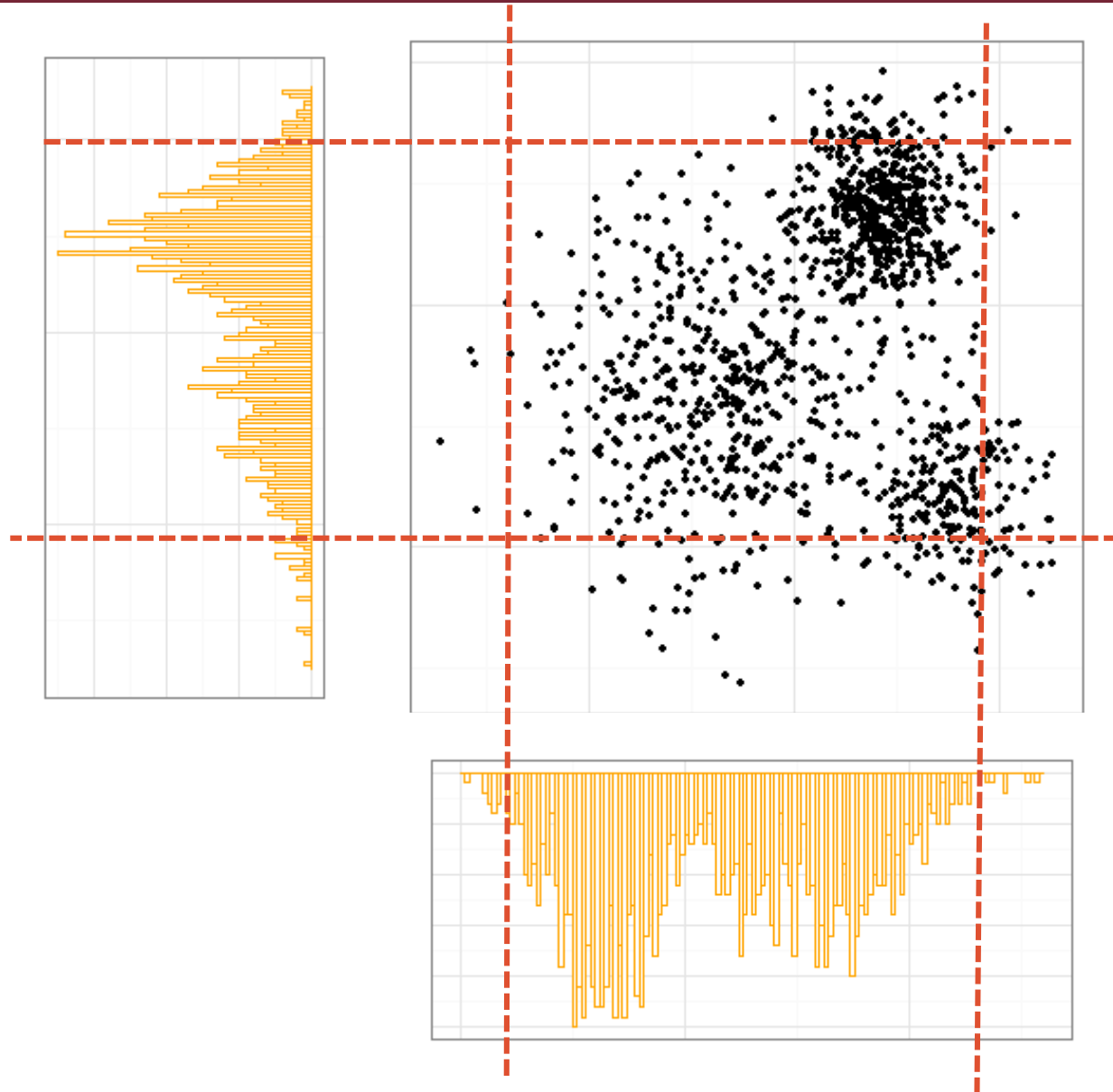


Néhány kilógó
értéket a boxplot
egyszerűen
elmaszkol

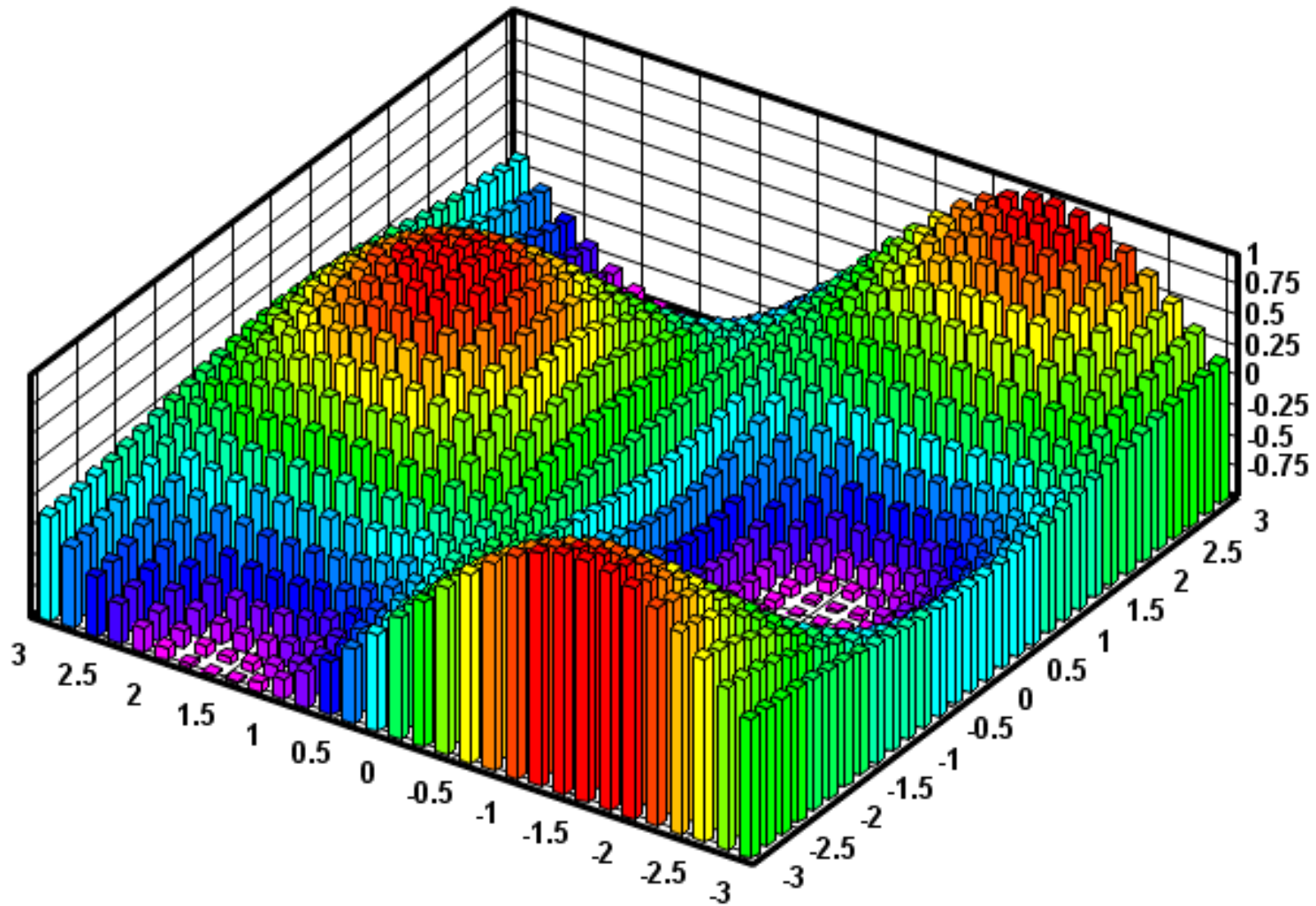


- Eddig: kilógó esetek az egyes dimenziókban
- Hogyan általánosítsunk?
 - Descartes szorzat?
 - Sűrűségfüggvény bevonása?

n D: Descartes szorzat



nD : Többdimenziós sűrűségfüggvény



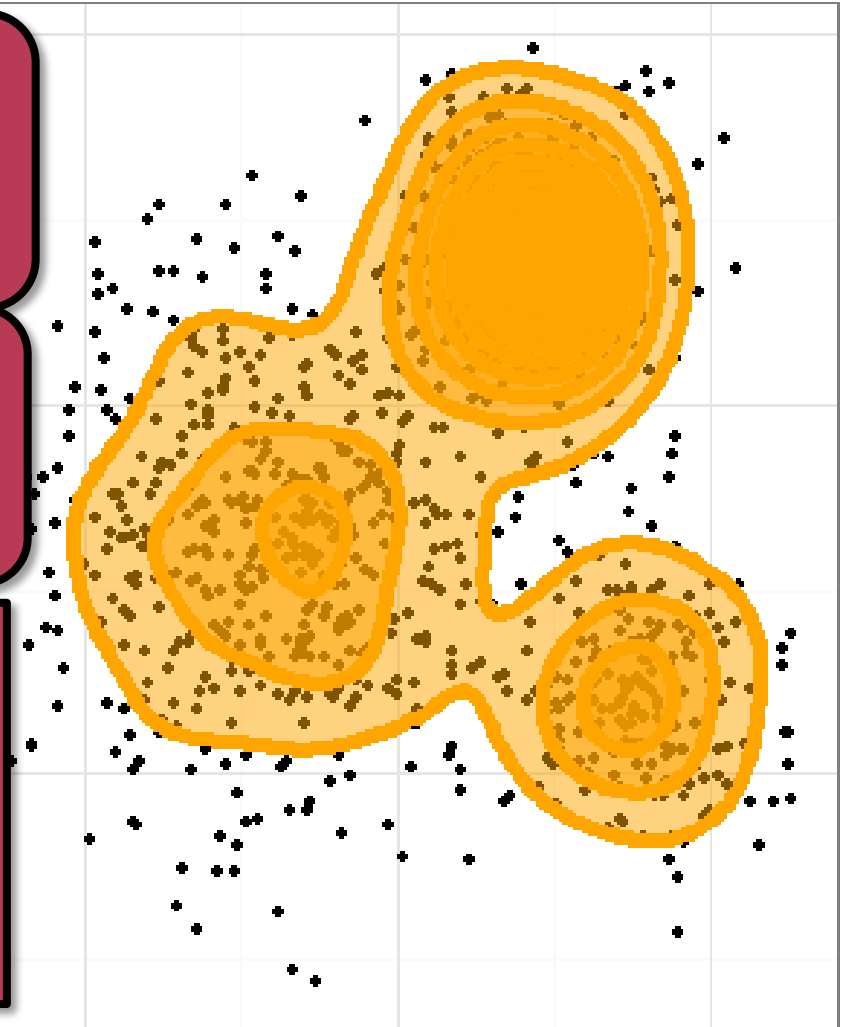
n D: Többdimenziós sűrűségfüggvény

- Egydimenziós általánosítása

Több „normál” kategória is létezik

Minden más outlier, esetleg átmenet a normálok között

Pontok nélkül a 0-1 nem látszik 😞



TÁVOLSÁG ALAPÚ TECHNIKÁK

Befoglaló burok

- Féltér-mélység: Tukey, 1974



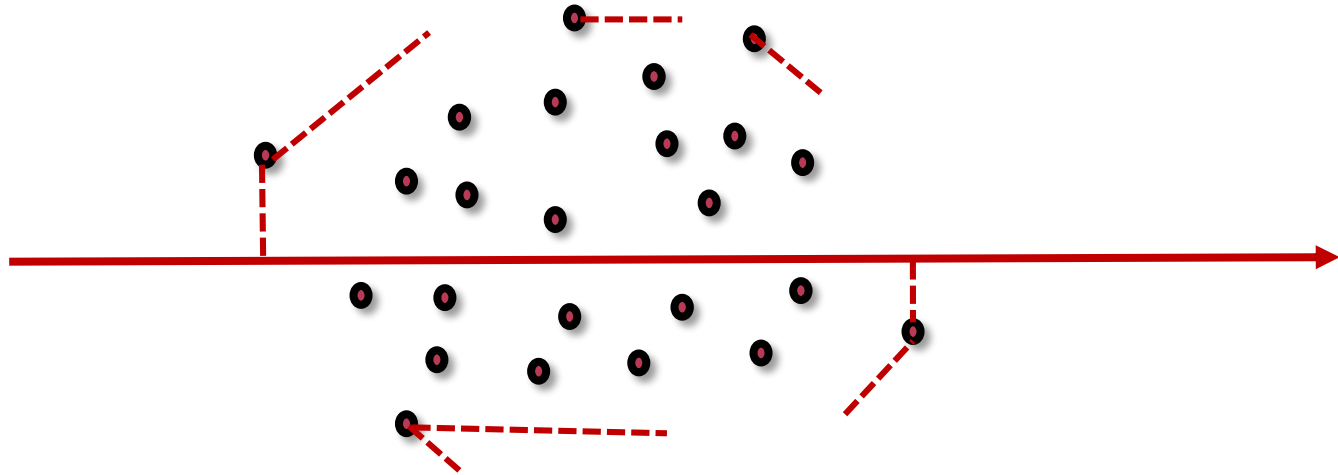
Extrém
pontok

$\min\{|x_i: x_i \leq z|\}$

Medián: majd a végén...

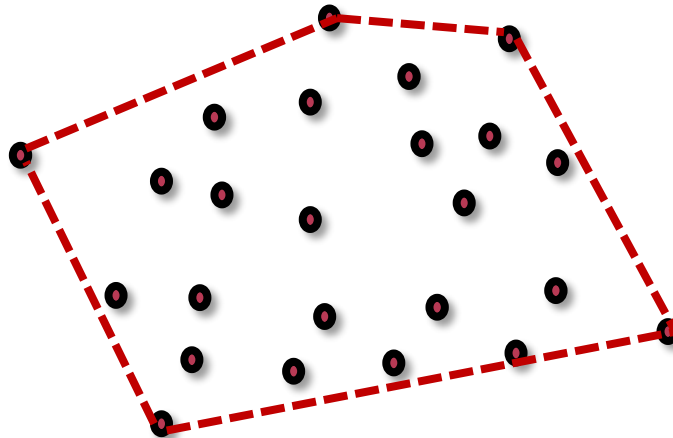
Befoglaló burok

- Féltér-mélység: Tukey, 1974



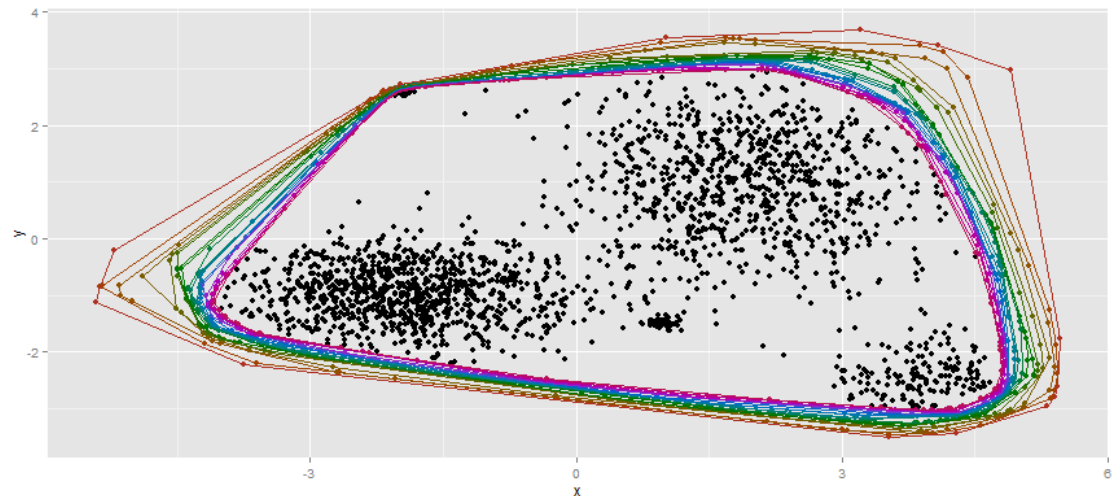
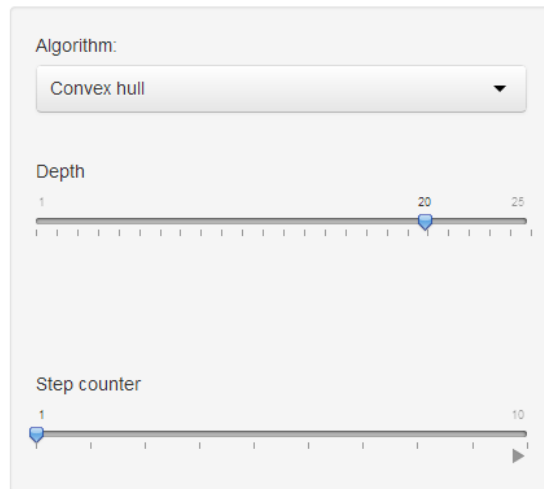
Befoglaló burok

- Féltér-mélység: Tukey, 1974



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

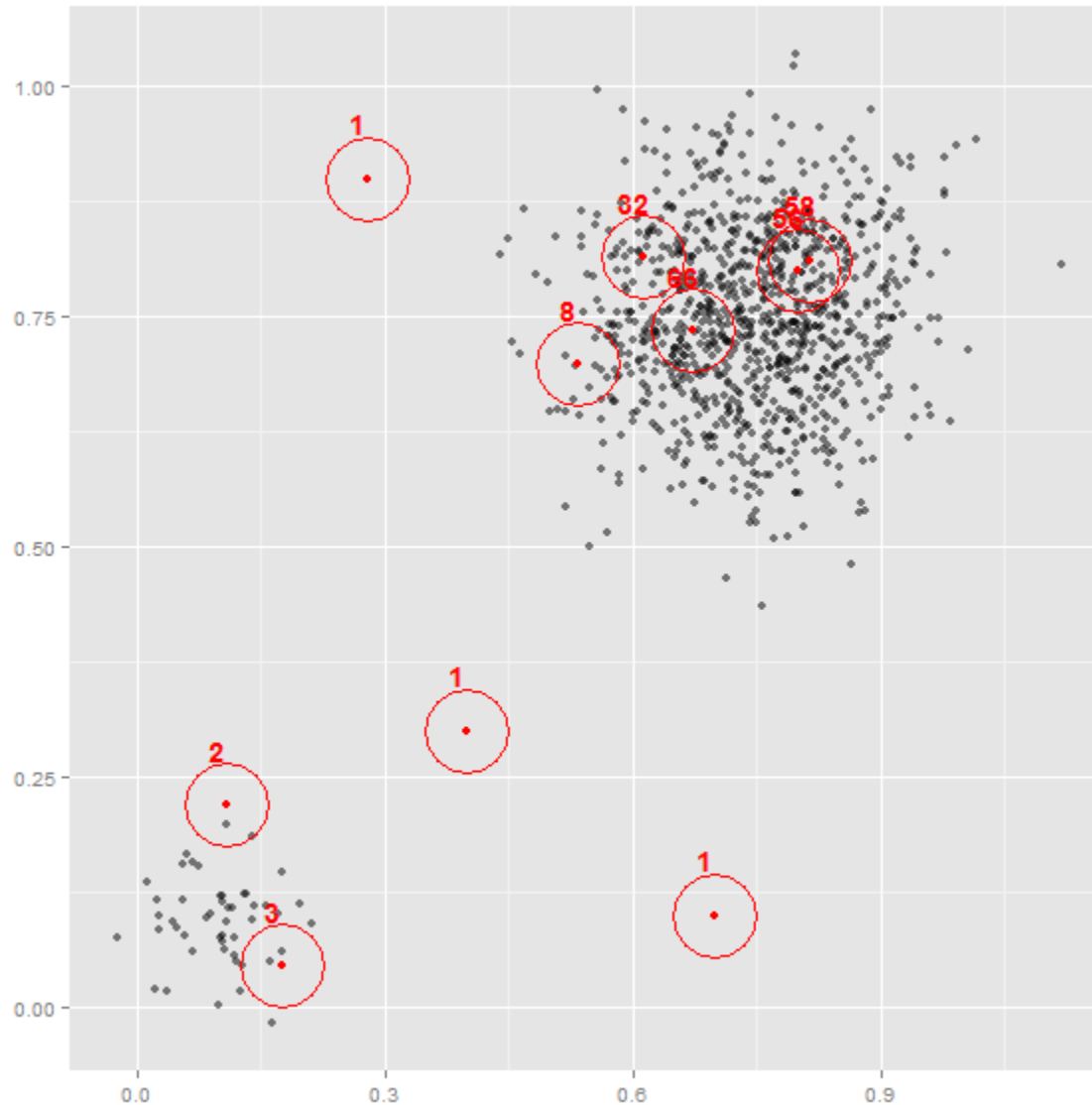
Rare event detection



```
[1] "Depth:"  
[1] 20  
[1] "Ratio of chosen elements: "  
[1] 0.4095455
```

DB

- Distance
- Outlier
- Parameter
 - r sug
 - Szomszéd



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

Rare event detection

Algorithm:

DB

Radius

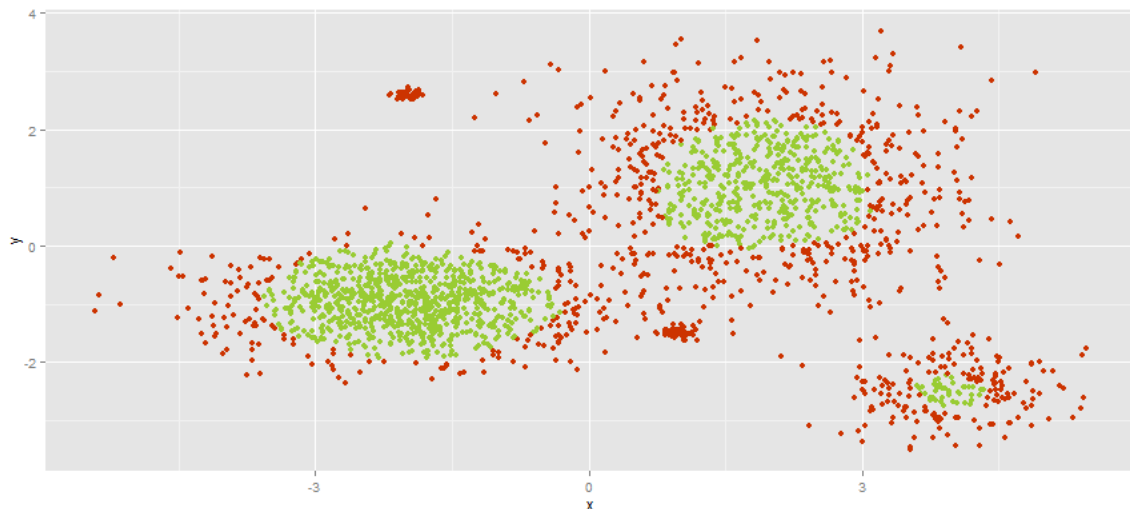
0 0.7 2

Ratio

0 0.05 0.25

Step counter

1 10

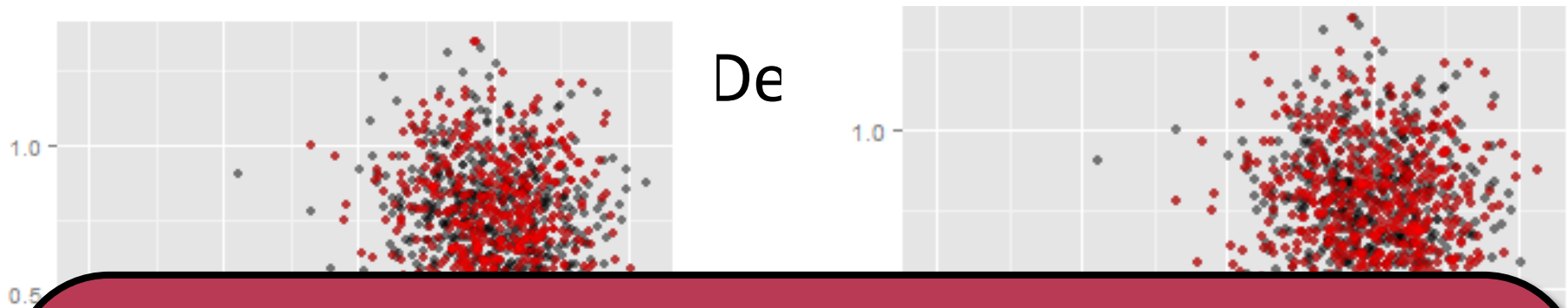


```
[1] "Ratio of chosen elements: "  
[1] 0.4145455
```

MCD

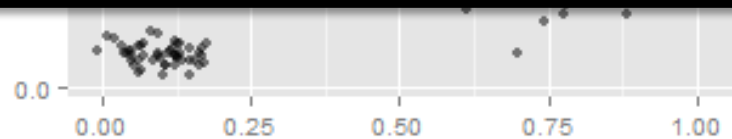
- Minimum Covariance Determinant
- Alapötlet
 - Keressük meg a legkompaktabb részhalmazt!

MCD



Kimerítő keresés?

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

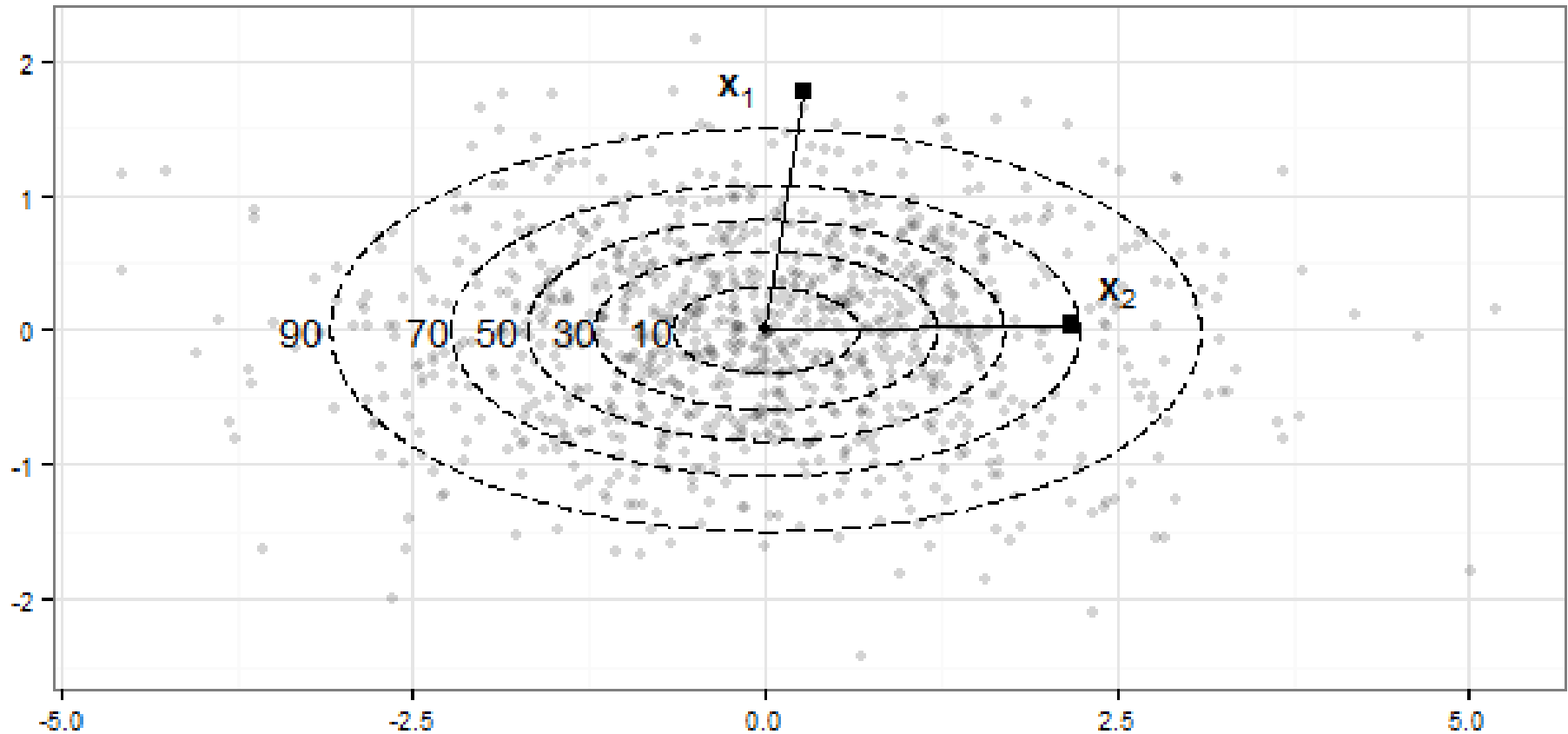


FAST-MCD

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

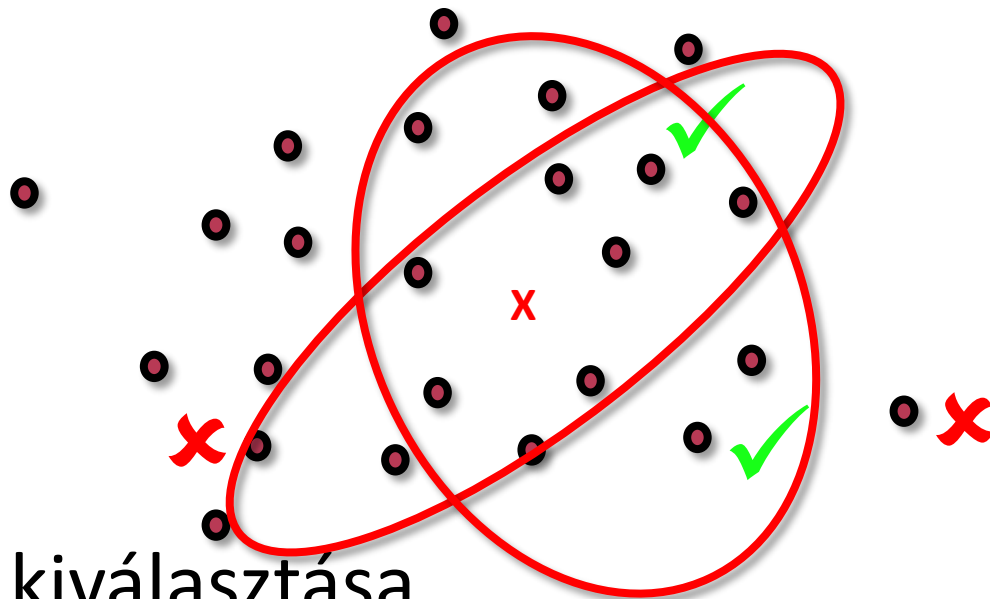
Mahalanobis távolság

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



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

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

■ Csomag: *robustX*

Rare event detection

Algorithm:

BACON

Type:

Manual selection

alpha

0.01

0.95

Initial set:

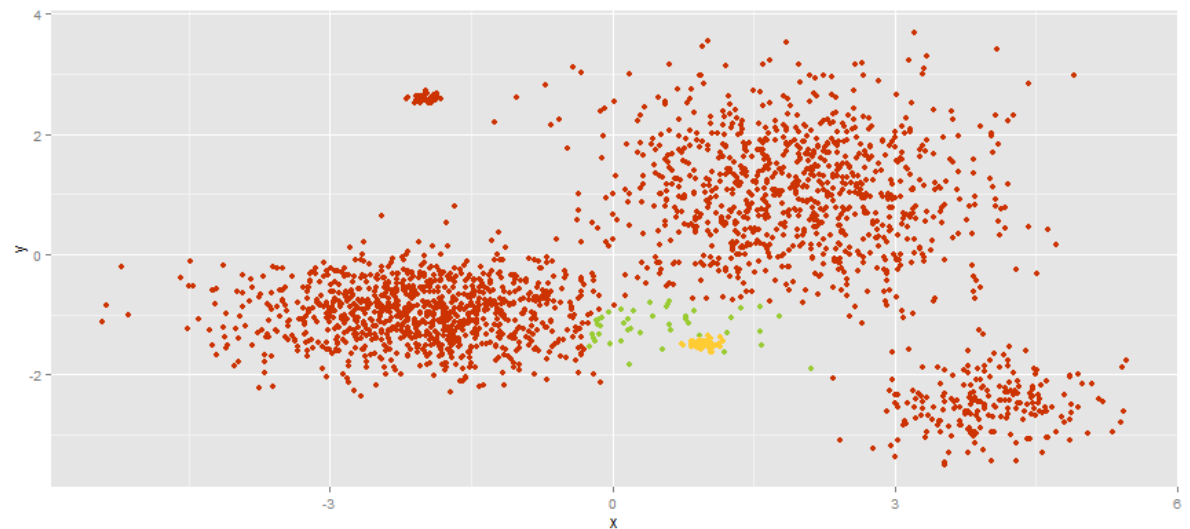
Set 5

Step counter

1

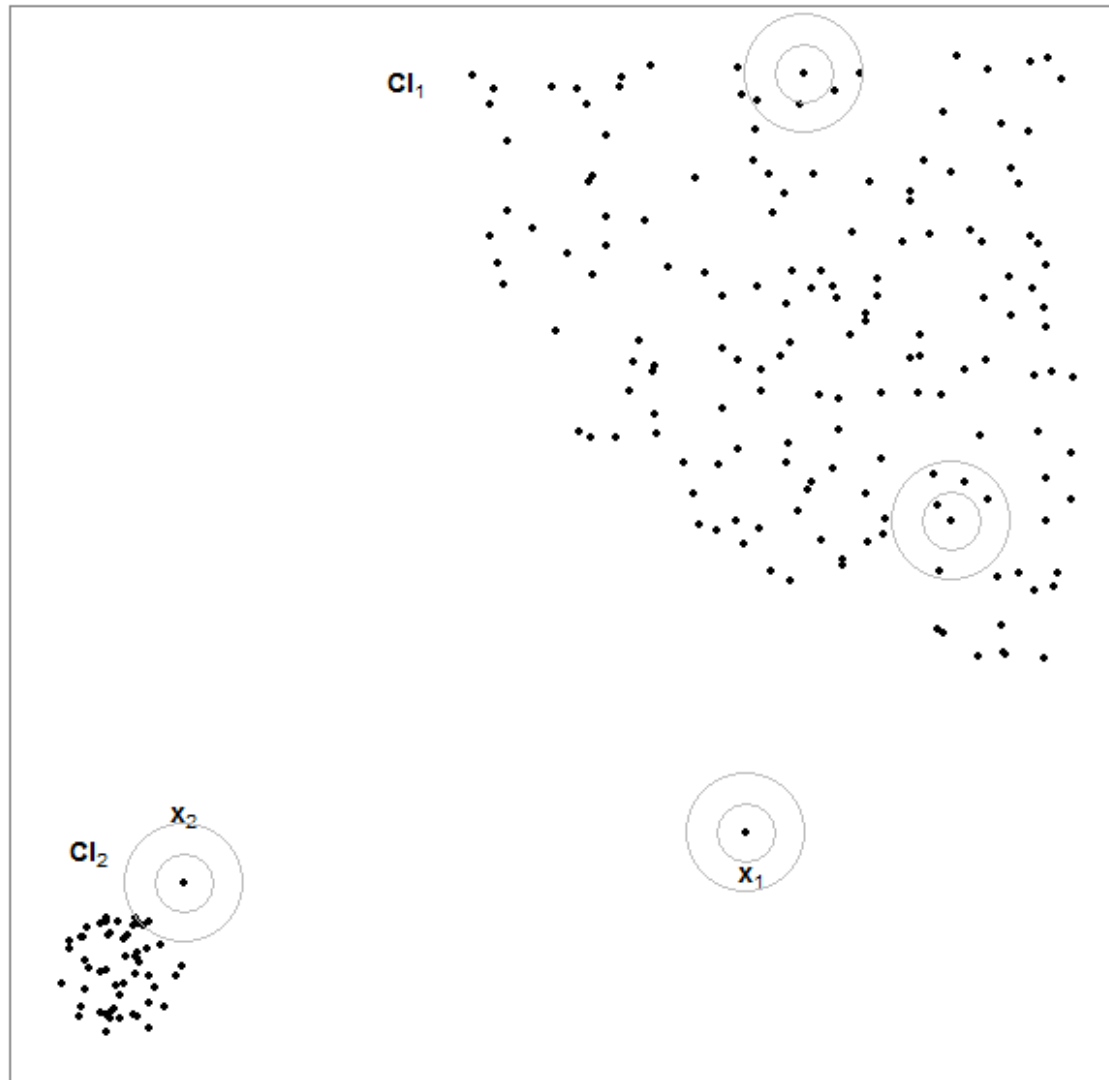
2

10

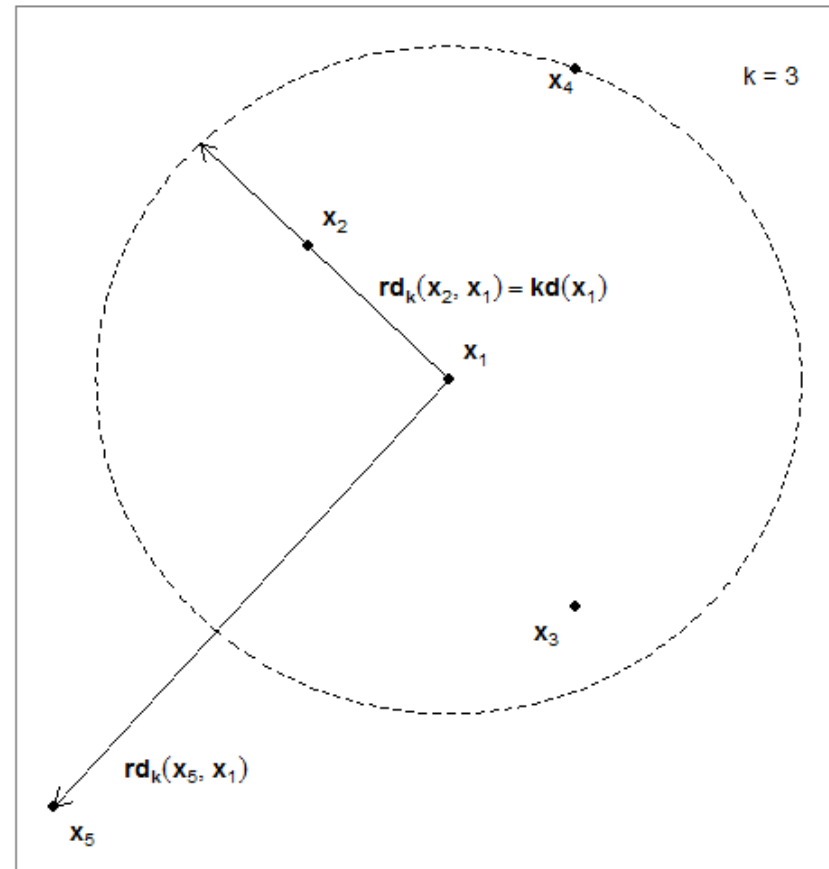


SŰRŰSÉG ALAPÚ TECHNIKÁK

LOF motiváció



- Local Outlier Factor
- Alapötlet: csak a szomszédok lokális sűrűség
- Outlier kritérium
 - a lokális sűrűség jóval kisebb, mint az átlagosan



DEMO LOF

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

Rare event detection

Algorithm:

LOF

k

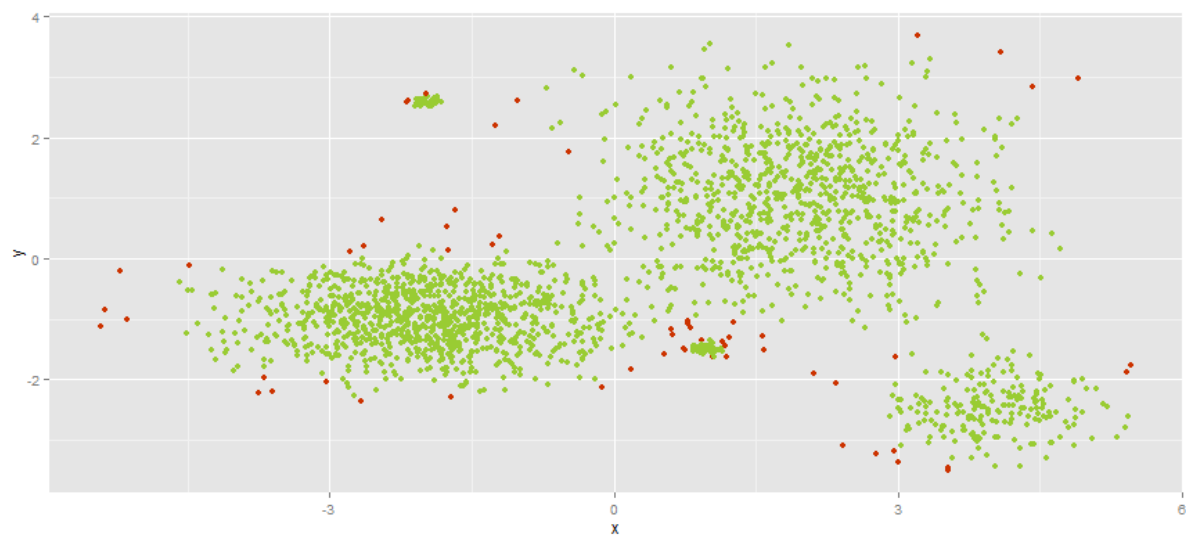
1 15 30

threshold

1 1.5 3

Step counter

1 10



Big Data környezet?

- Közelítő algoritmusok
- Stream processing

Hivatkozásjegyzék

- [1] Stream Processing, filtering: Mining of Massive Data Sets
 - Alapmű:
<http://infolab.stanford.edu/~ullman/mmds/book.pdf>
 - Coursera tárgy:
<https://www.coursera.org/course/mmds>
- [2] Outlier Detection
 - Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3):15, 2009