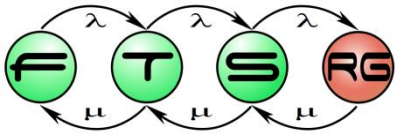
Mintavételezés, szűrés, outlierek detektálása

Salánki Ágnes

salanki@mit.bme.hu

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group

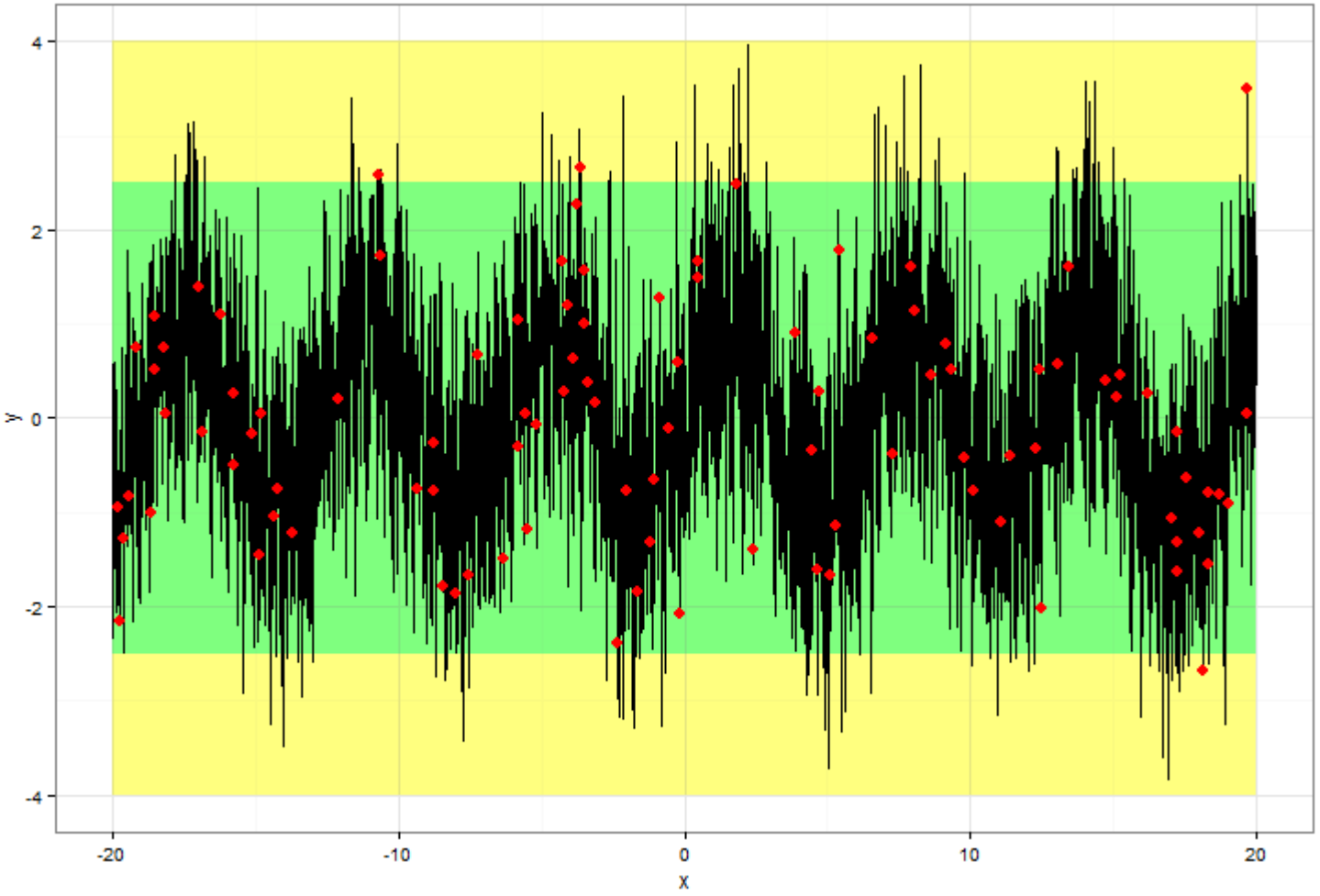


Alapfogalmak

- Az alapfeladat ugyanaz
- Az aspektus más

Alapfogalmak

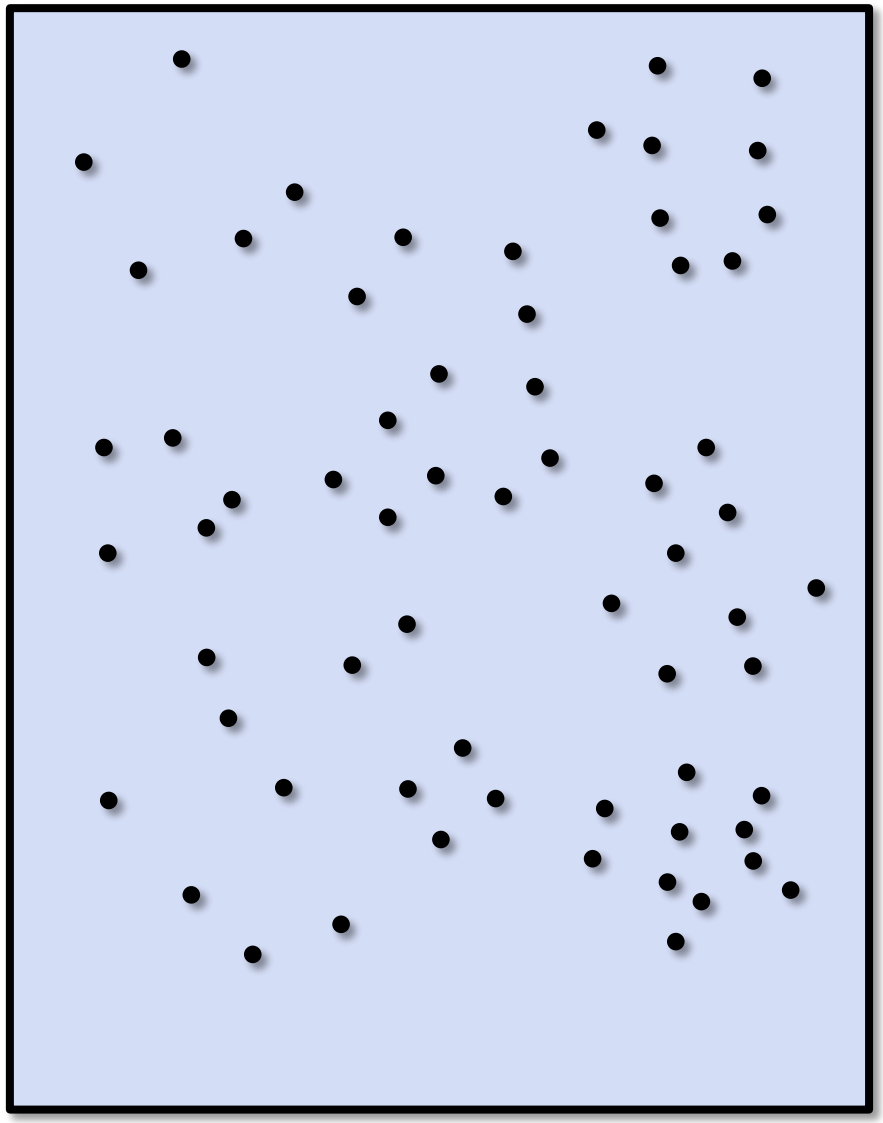
- Az alapfeladat ugyanaz
- Az



MINTAVÉTELEZÉS

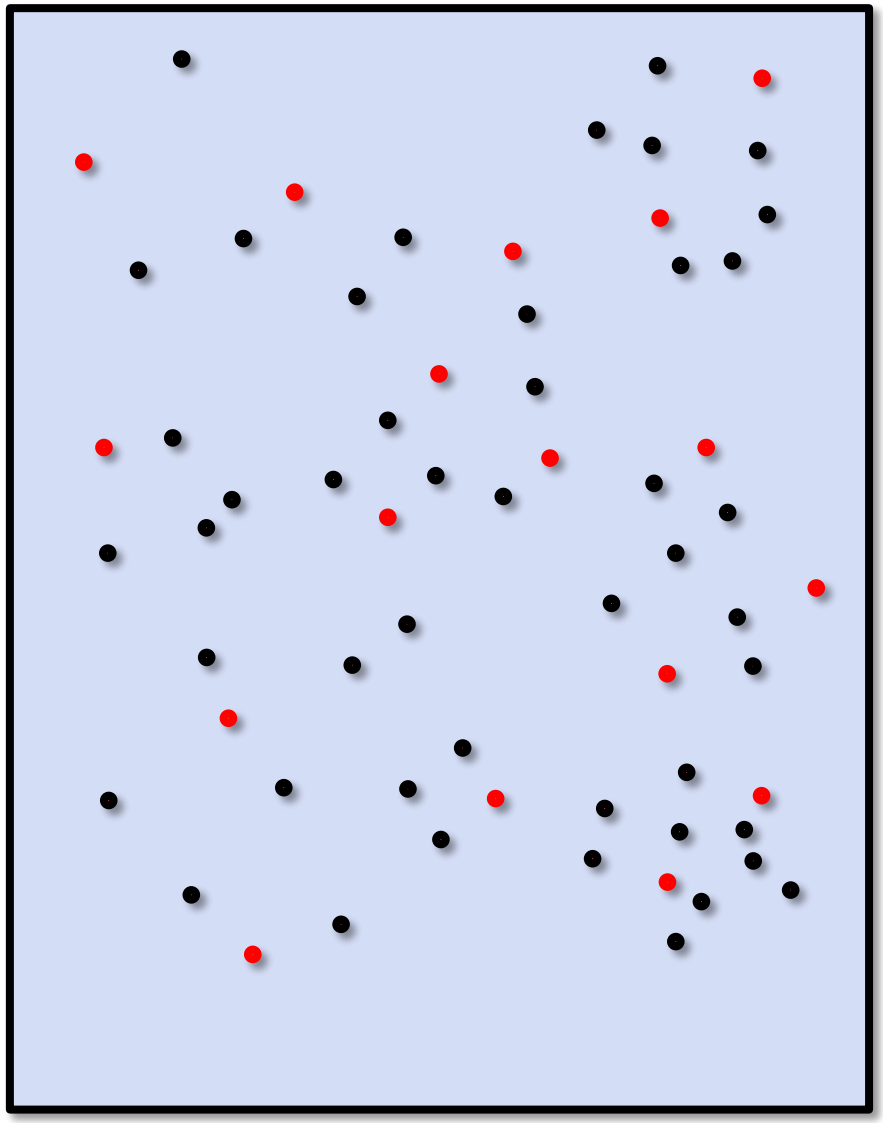
Mintavételezés

- SRS
- Stratified Sample
- Cluster sample



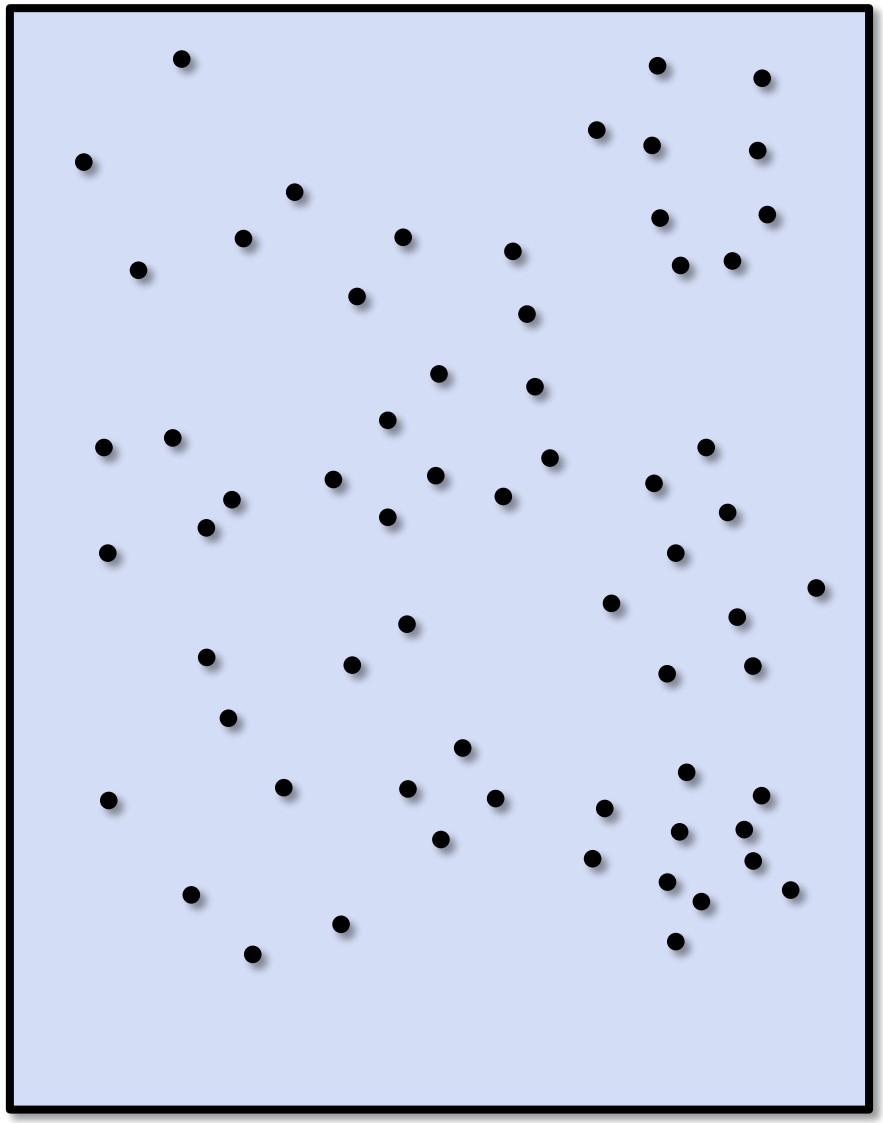
Mintavételezés

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



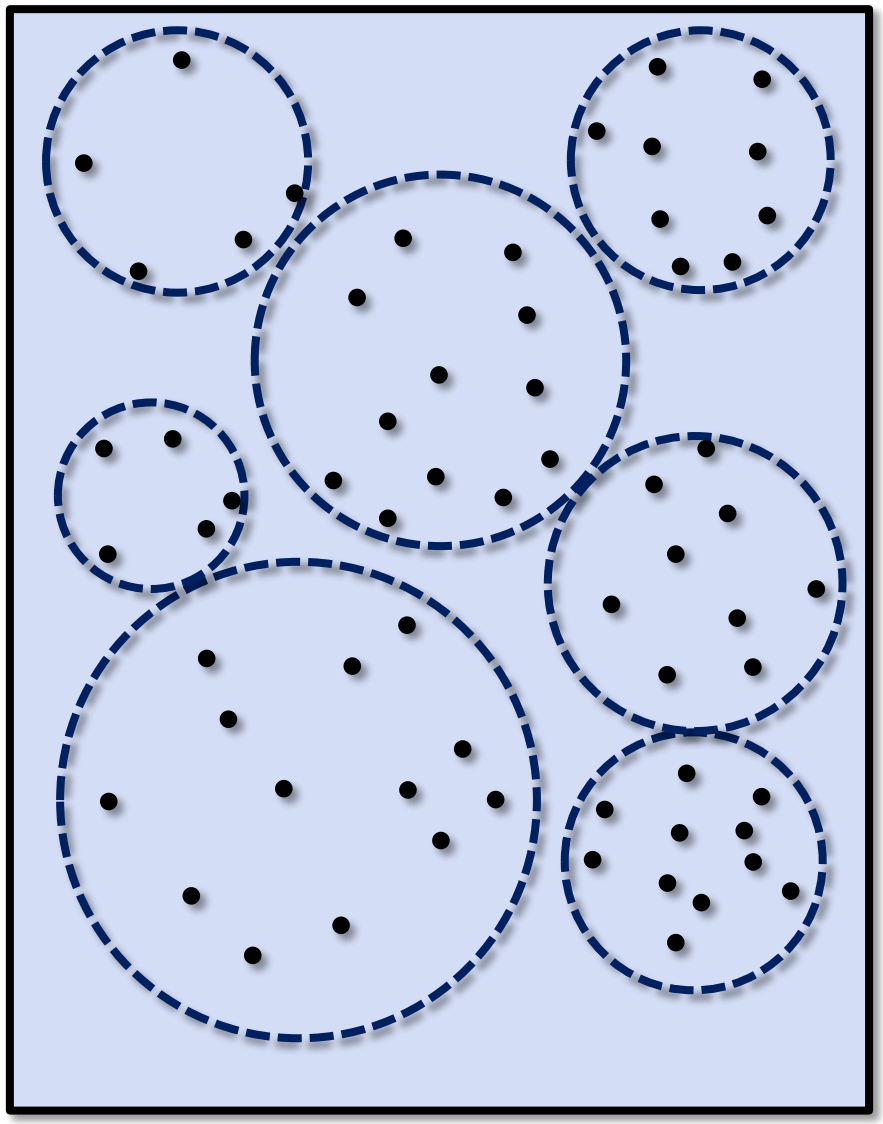
Mintavételezés

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



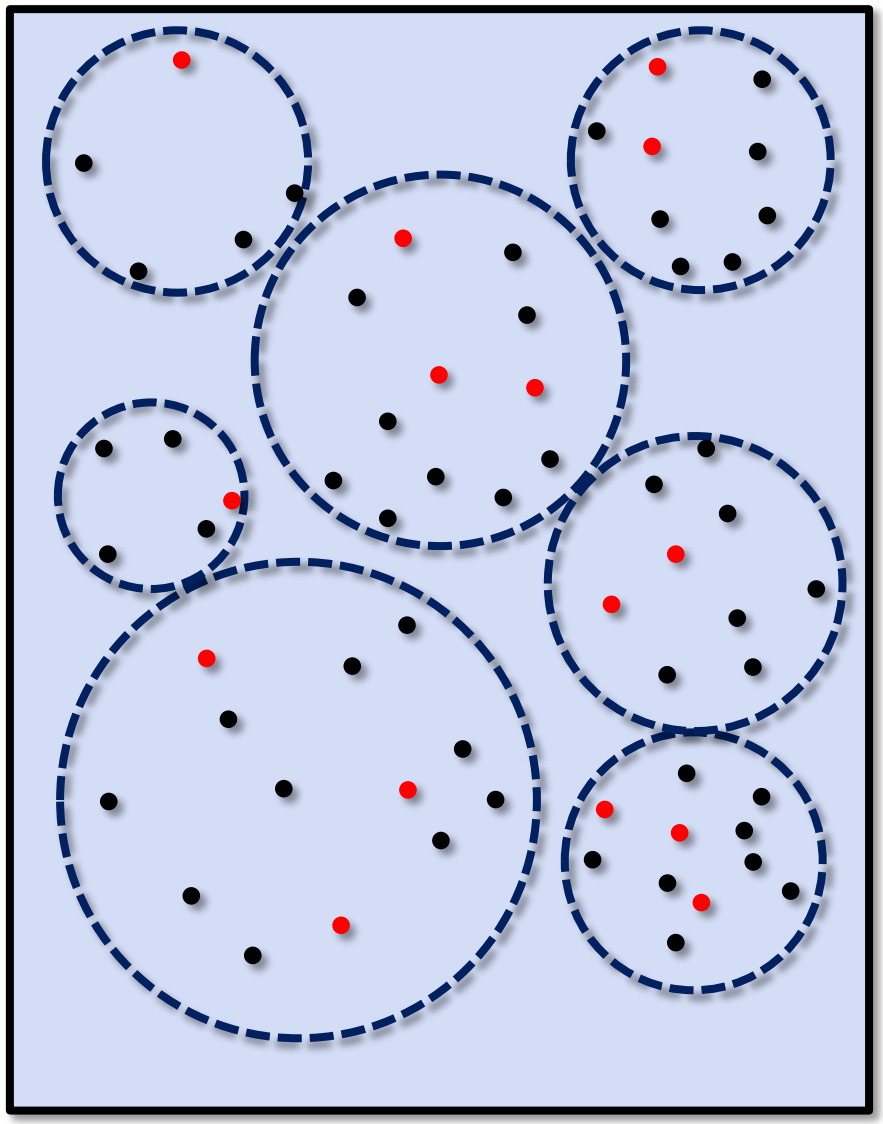
Mintavételezés

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



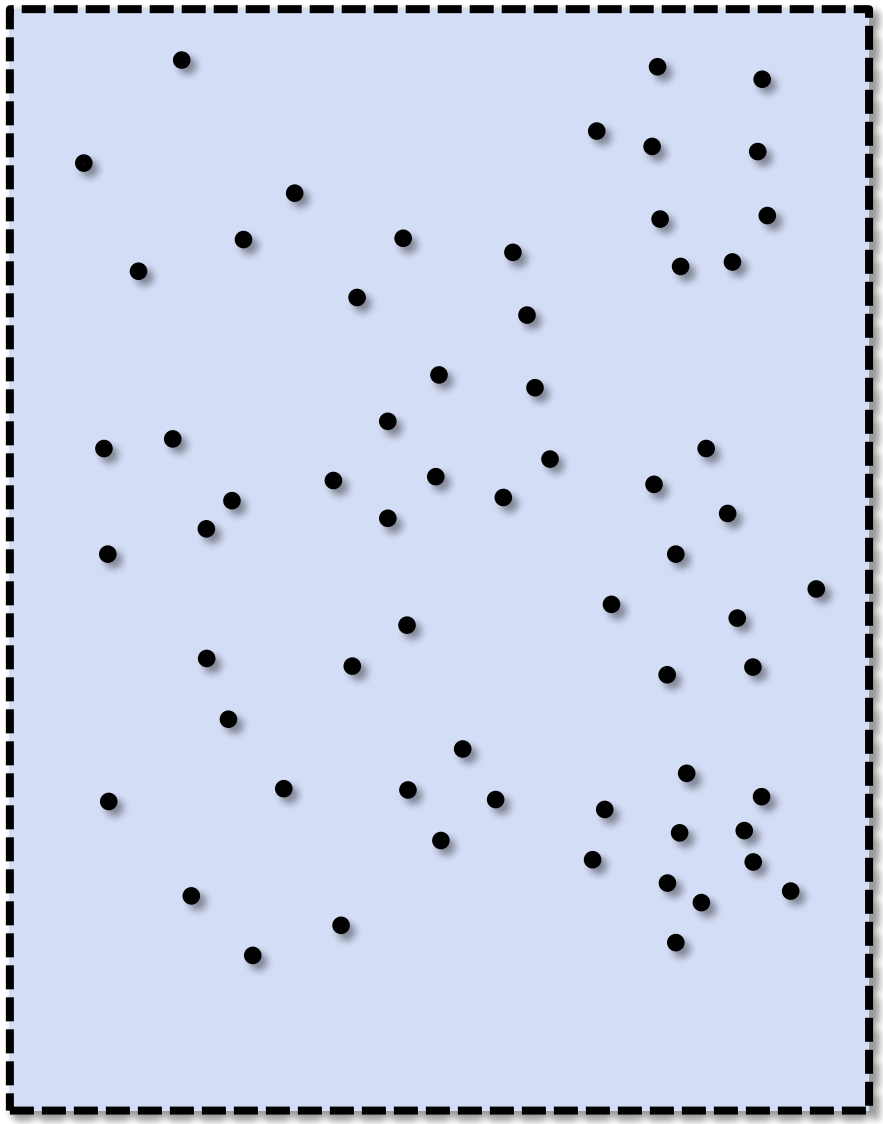
Mintavételezés

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



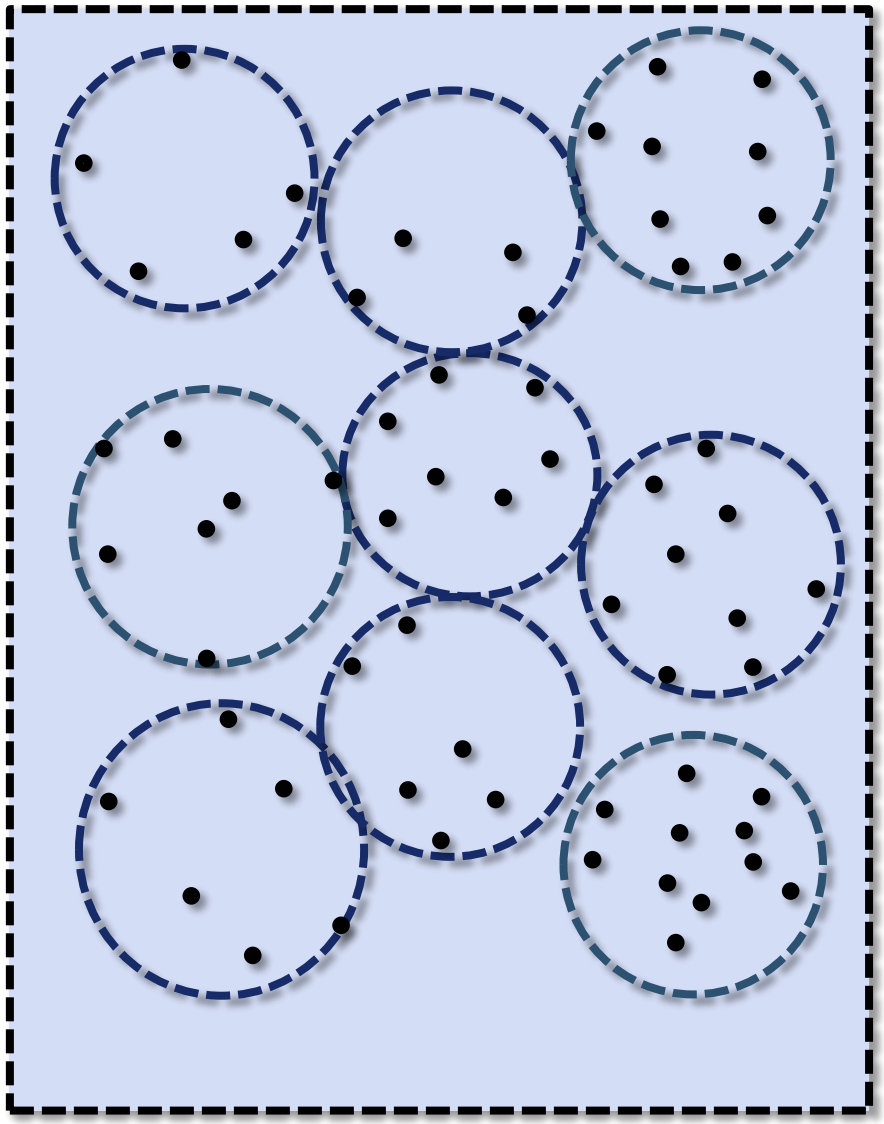
Mintavételezés

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



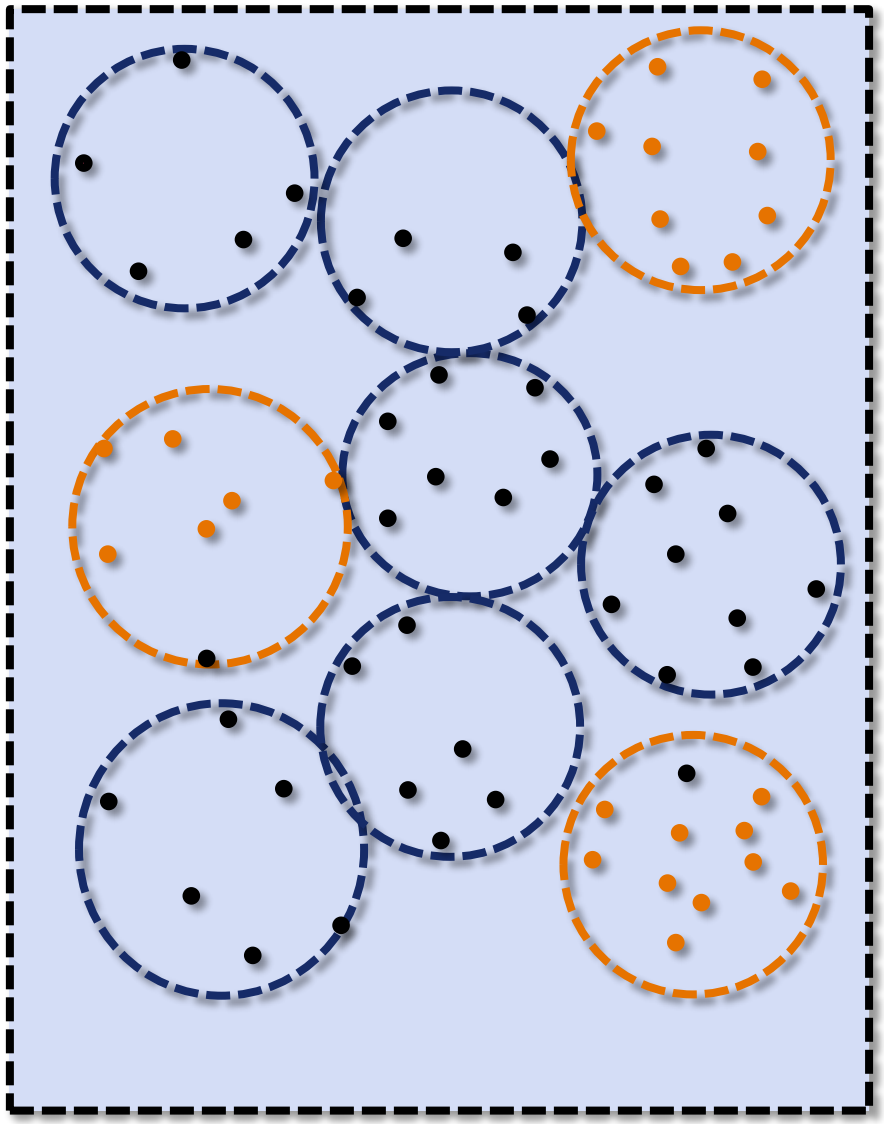
Mintavételezés

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

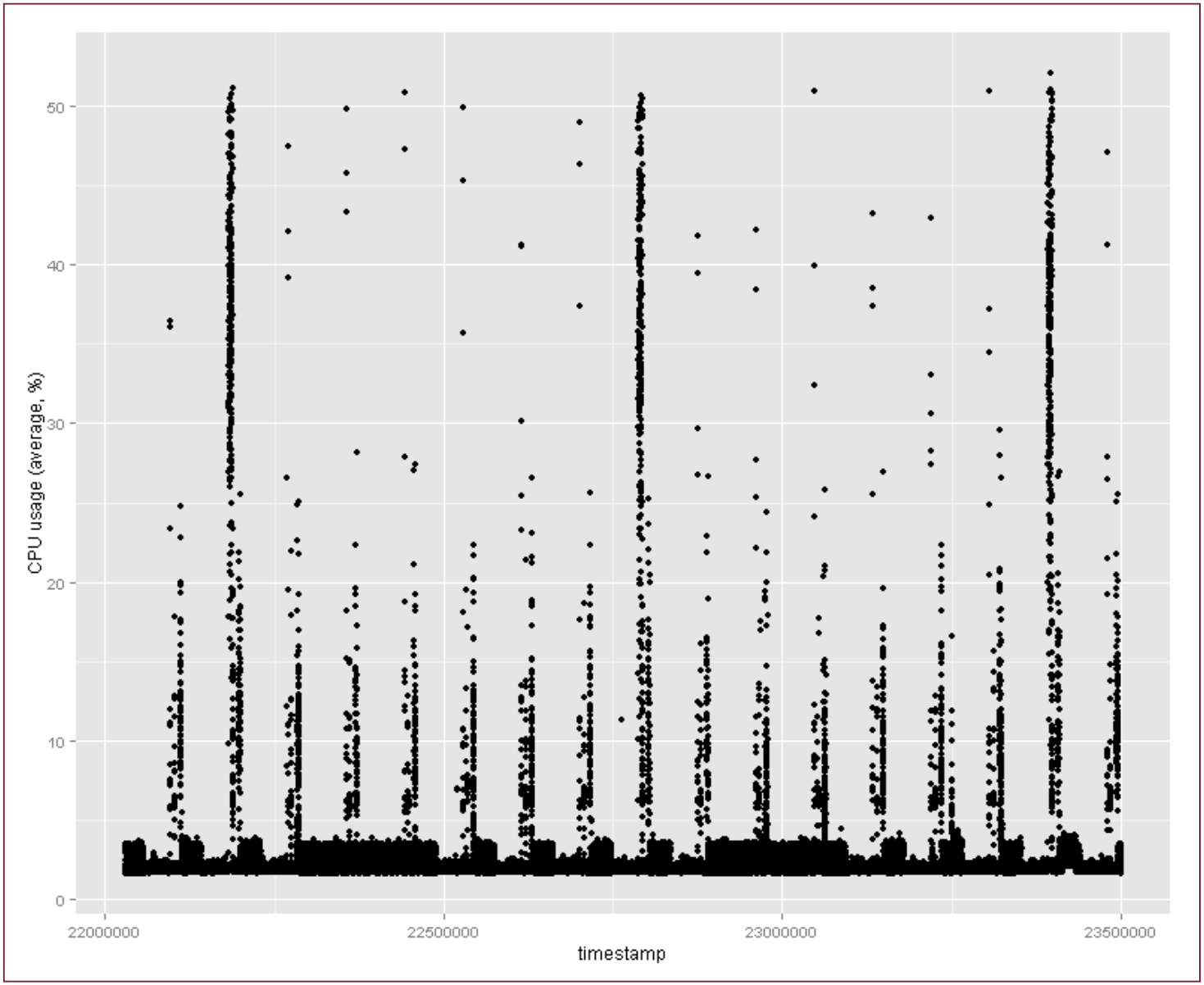


Mintavételezés

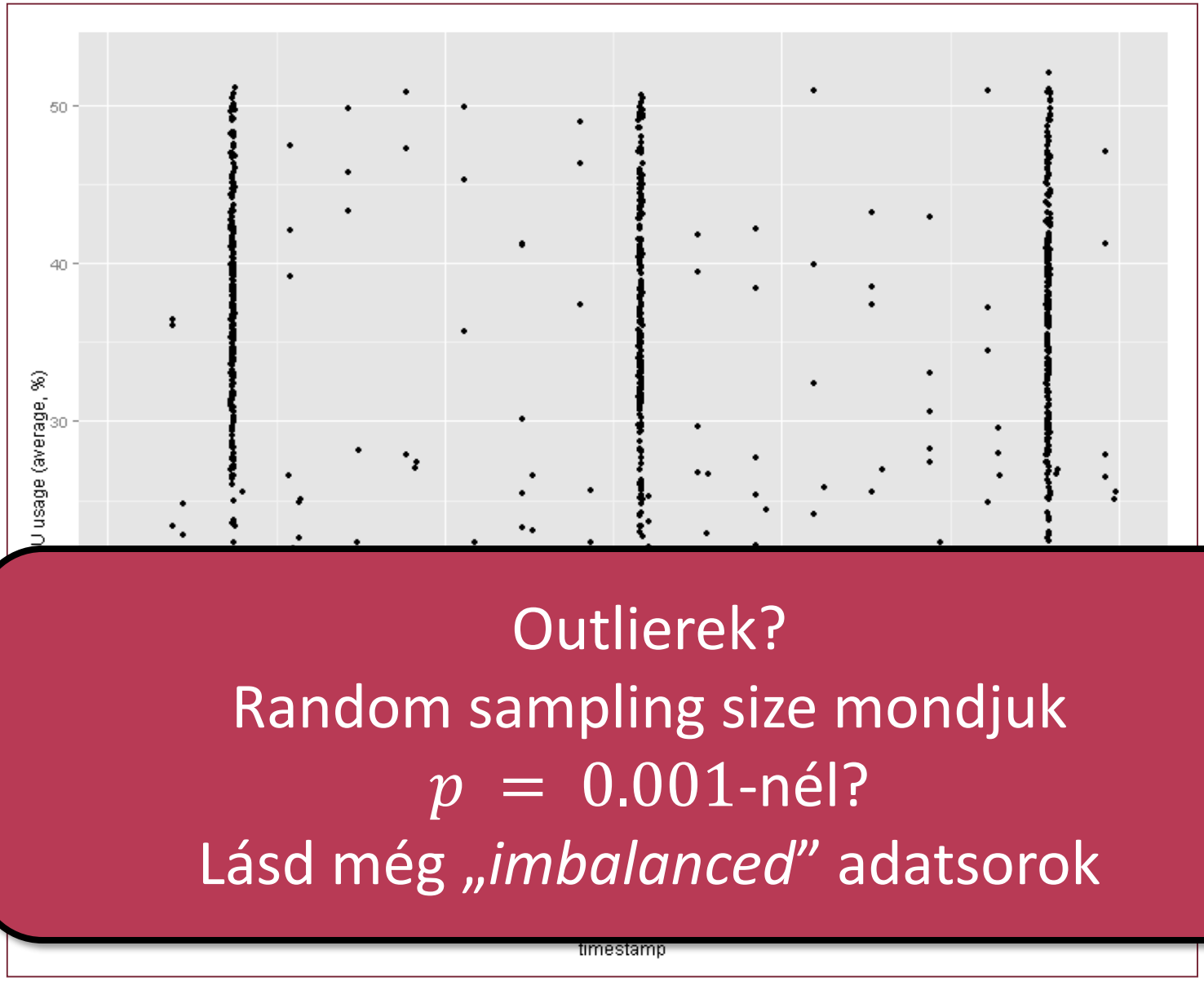
- SRS
  - Simple Random Sample
- Stratified Sample
- Cluster sample
  - ~azonos méretű klaszterek
  - Azokból random m.



Idősoroknál



Idősoroknál



Outlierek?

Random sampling size mondjuk

$p = 0.001$-nél?

Lásd még „*imbalanced*” adatsorok

Mintavételezés streamekben

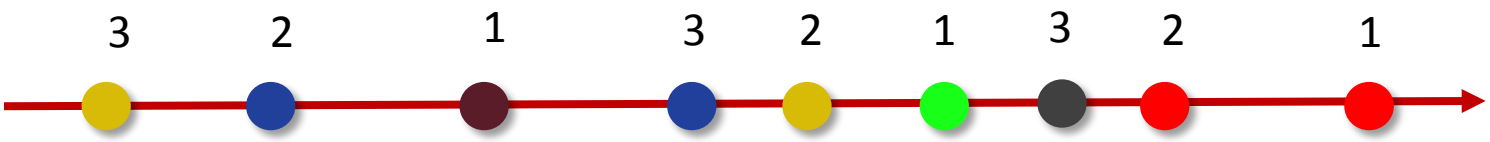
- Pl. „az elmúlt héten hány egyedi lekérdezés jött?” megválaszolása $n\%$ minta alapján

Random mintavételezés

- 1/3-os mintavételezés
  - egyedi lekérdezések aránya: 3/9
  - egyedi lekérdezések aránya egy kiválasztott mintában?

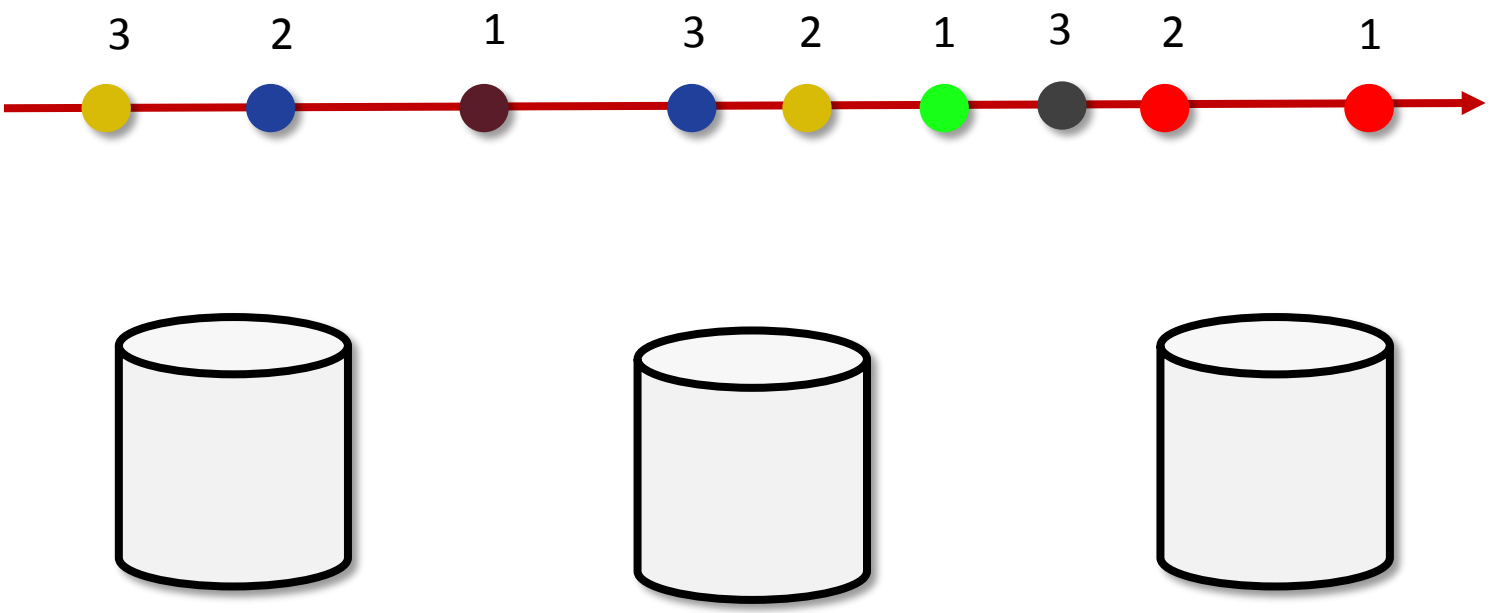
Random mintavételezés

- 1/3-os mintavételezés
  - egyedi lekérdezések aránya: 3/9
  - egyedi lekérdezések aránya egy kiválasztott mintában?



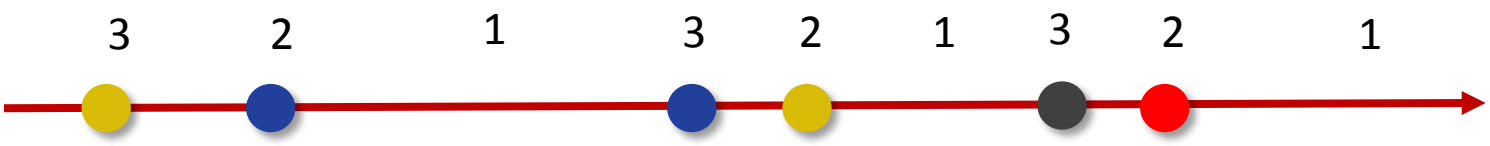
Random mintavételezés

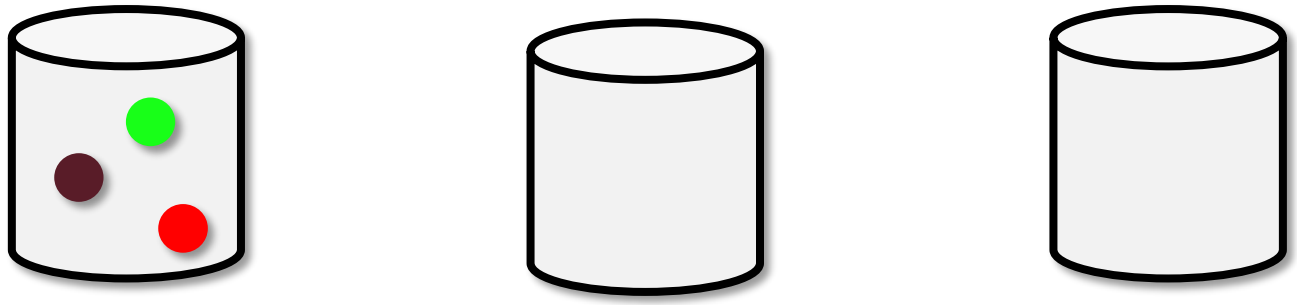
- 1/3-os mintavételezés
  - egyedi lekérdezések aránya: 3/9
  - egyedi lekérdezések aránya egy kiválasztott mintában?



Random mintavételezés

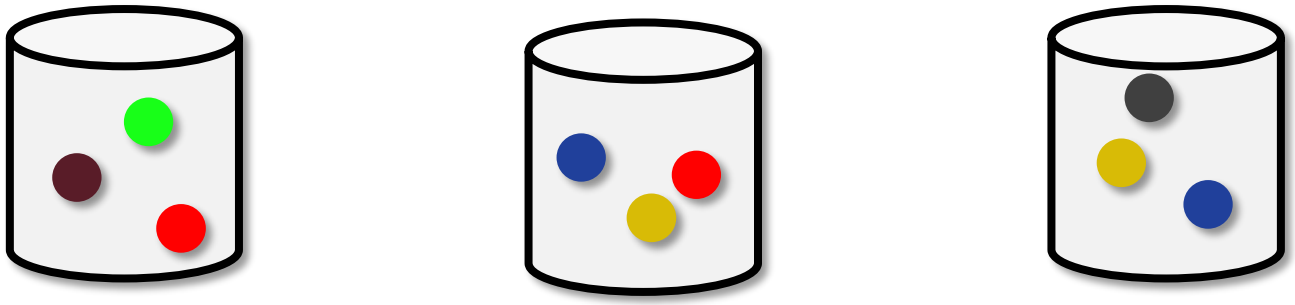
- 1/3-os mintavételezés
  - egyedi lekérdezések aránya: 3/9
  - egyedi lekérdezések aránya egy kiválasztott mintában?





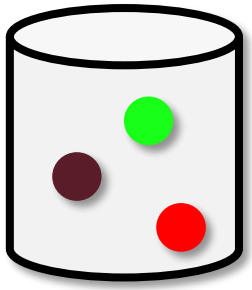
Random mintavételezés

- 1/3-os mintavételezés
  - egyedi lekérdezések aránya: 3/9
  - egyedi lekérdezések aránya egy kiválasztott mintában?

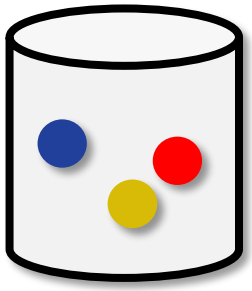


Random mintavételezés

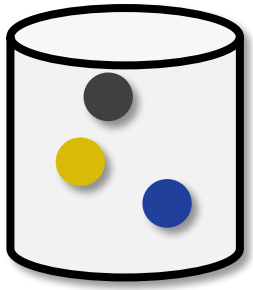
- 1/3-os mintavételezés
  - egyedi lekérdezések aránya: 3/9
  - egyedi lekérdezések aránya egy kiválasztott mintában?



$\hat{p} = 1.0$



$\hat{p} = 1.0$



$\hat{p} = 1.0$

Mintavételezés streamekben

- Random mintavételezés 10 vödörrel
  - Ha tényleg egyedi a streamben, p = 0.1 a mintában (egy adott megfigyelt vödörben)
  - Ha kétszer fordul elő, a mintába p = 0.18 valószínűséggel kerül csak egy stb.

Mintavételezés streamekben

- Random mintavételezés 10 vödörrel
  - Ha tényleg egyedi a streamben, p = 0.1 a mintában (egy adott megfigyelt vödörben)
  - Ha kétszer fordul elő, a mintába p = 0.18 valószínűséggel kerül csak egy stb.

Nem tudunk a minta alapján általánosítani a teljes streamre ☹

Mintavételezés streamekben: Hash

- Pl. „az elmúlt héten hány egyedi lekérdezés jött?” megválaszolása $n\%$ minta alapján

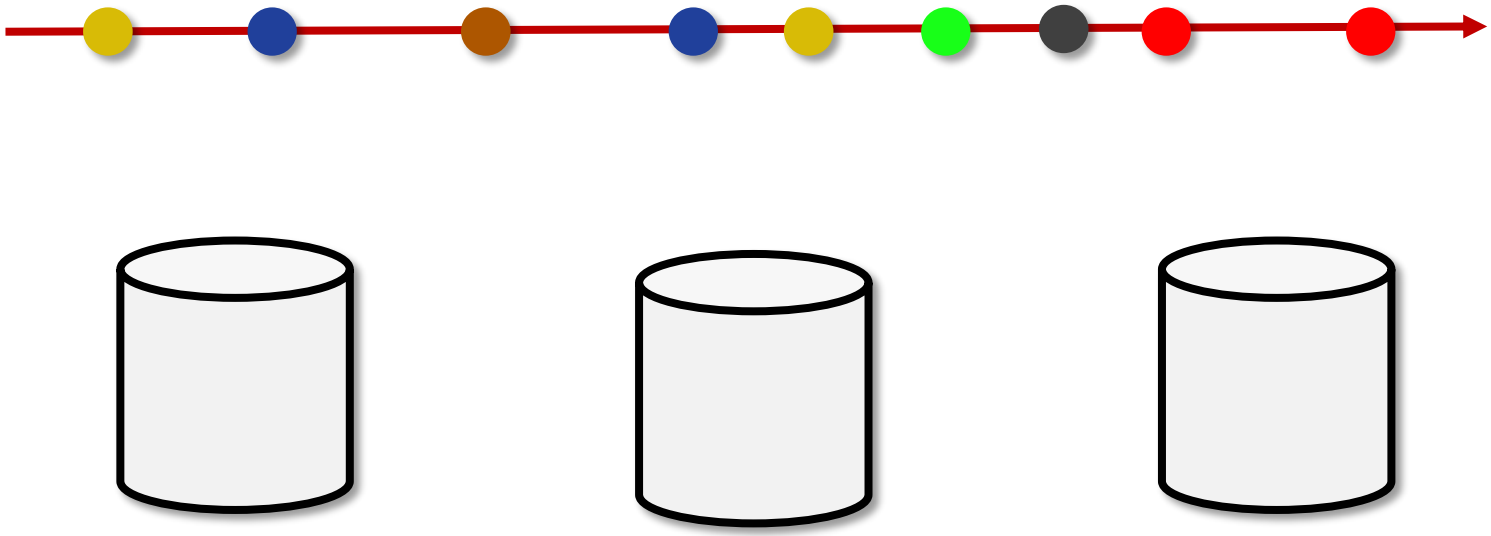
- Érték alapján szűrünk
  - Pl. hash függvény 0-9 közé
    - Az azonosak azonos vödörbe kerülnek
  - Feltételezések
    - A hash egyenletes → az értékek 1/10-e kerül be a 0-ba

Mintavételezés streamekben: hash

- 1/3-os mintavételezés

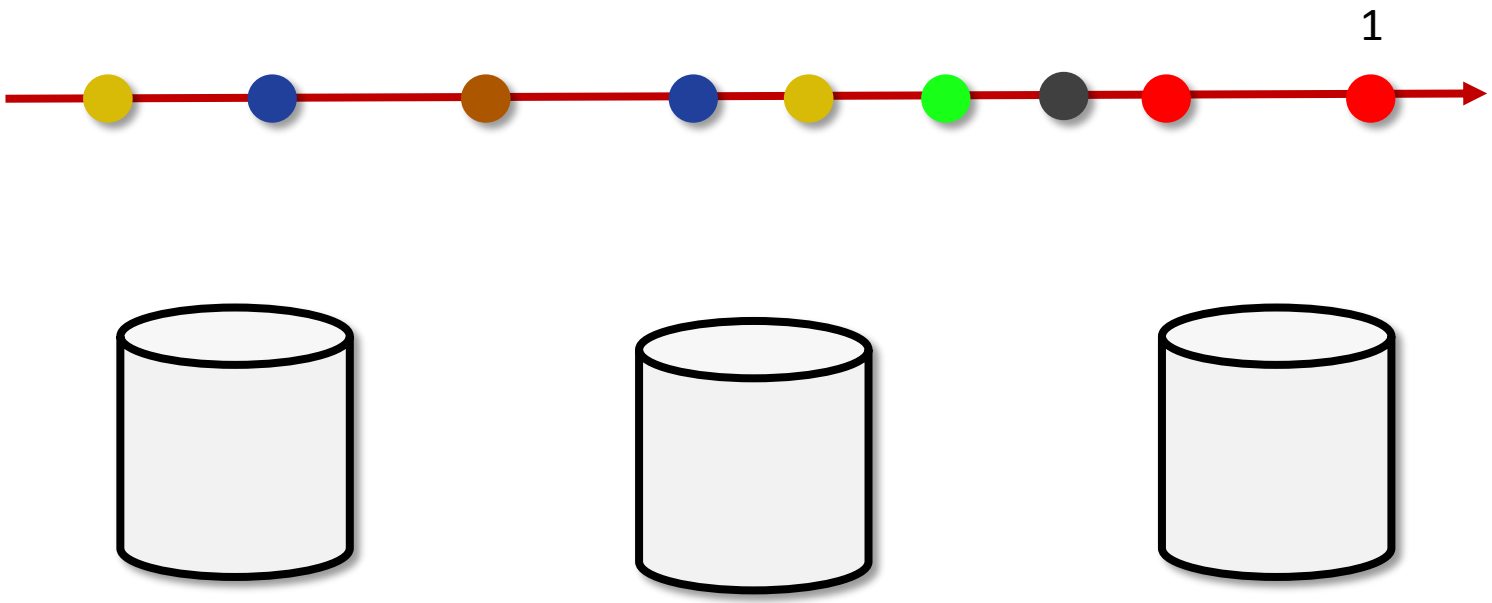
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



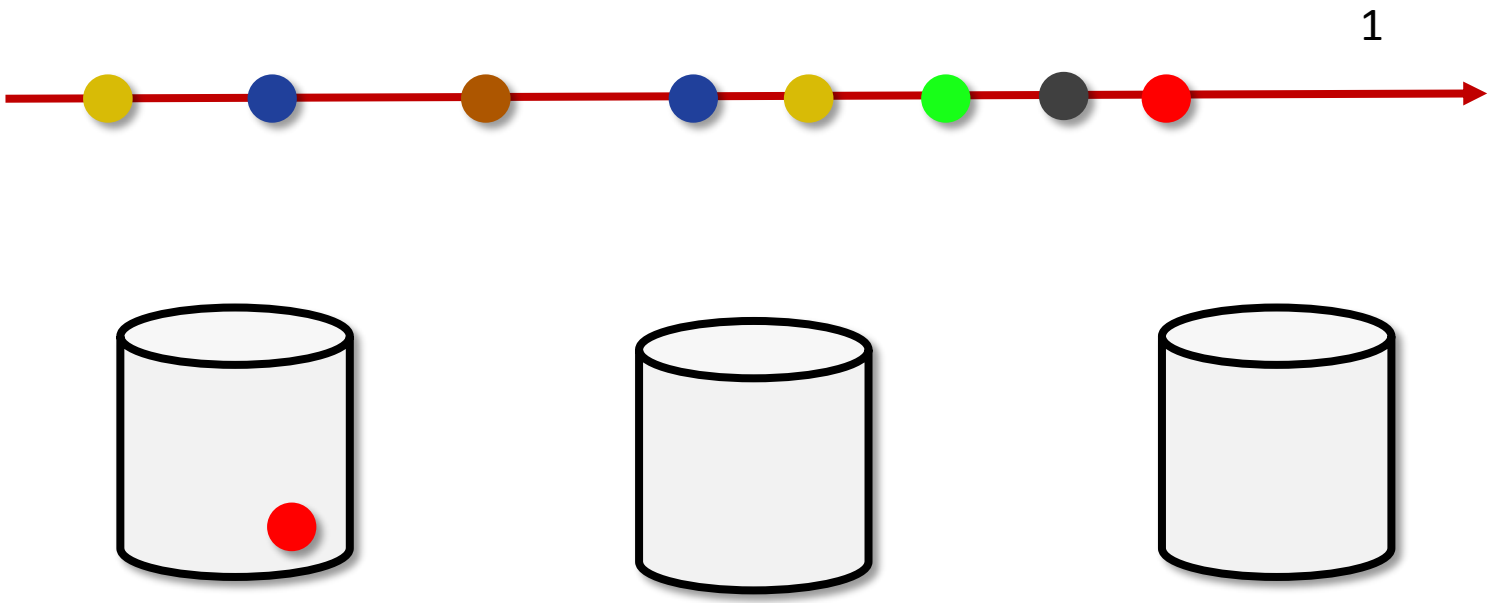
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



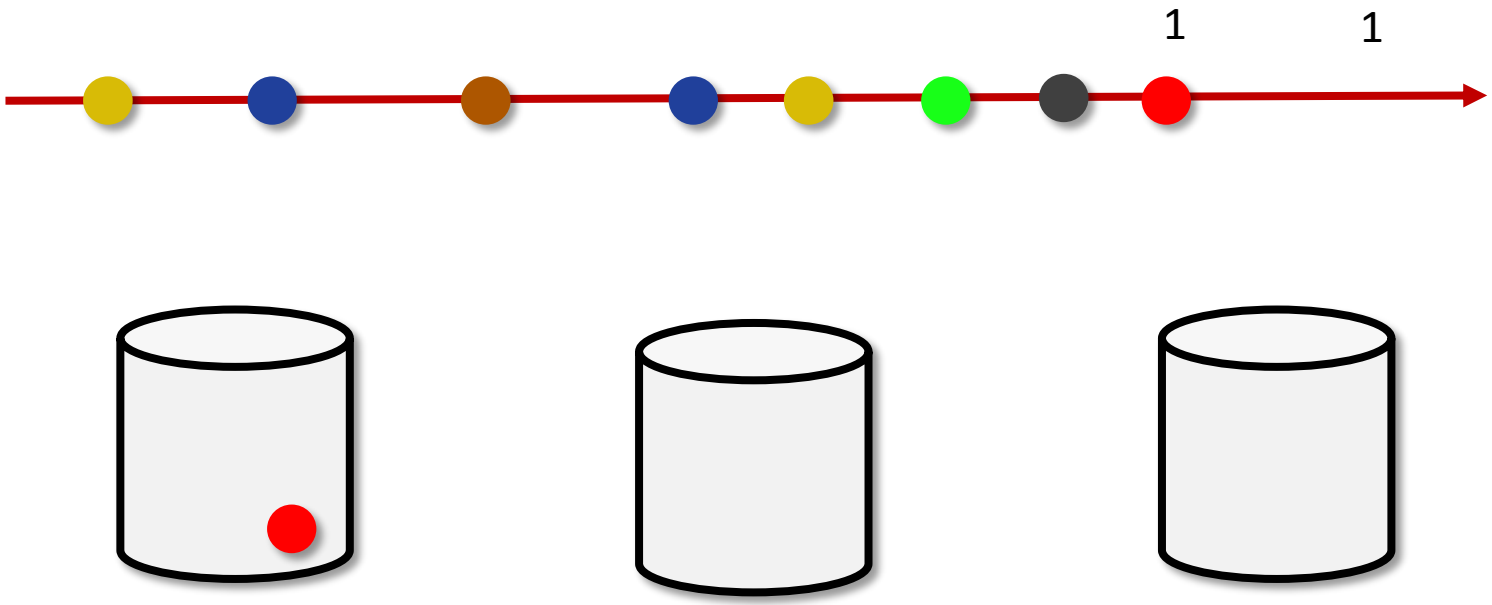
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



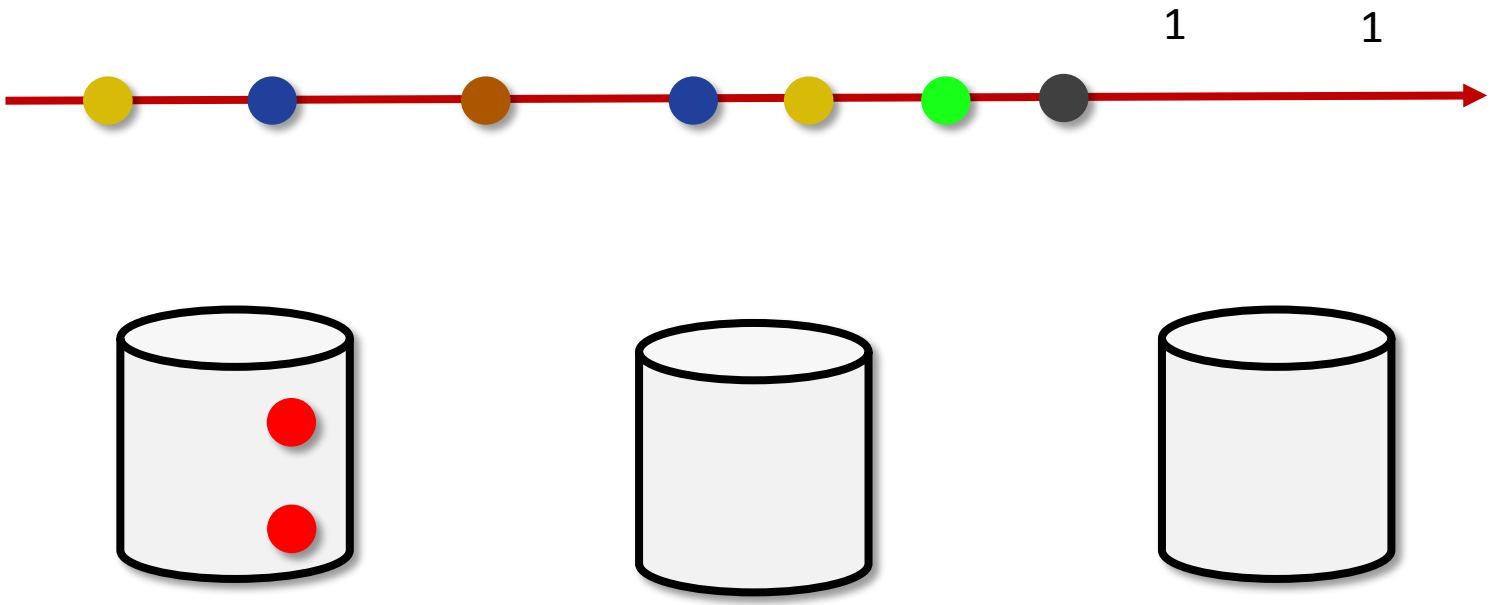
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



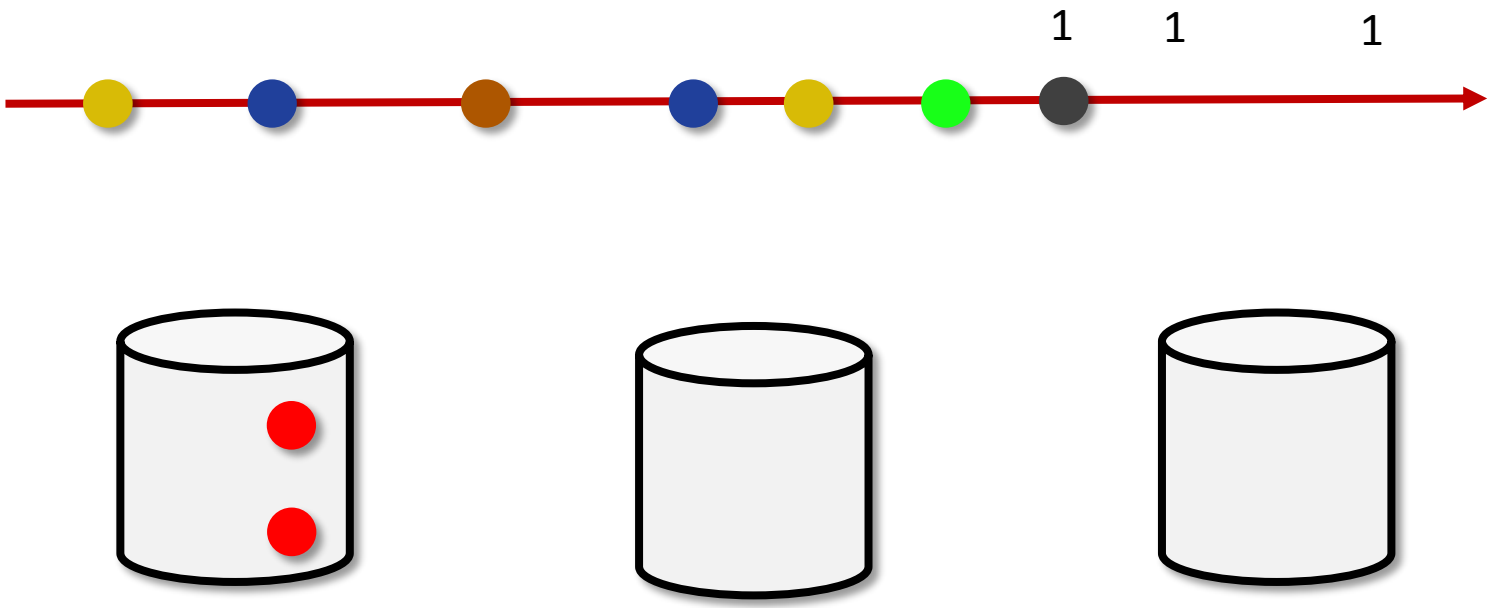
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



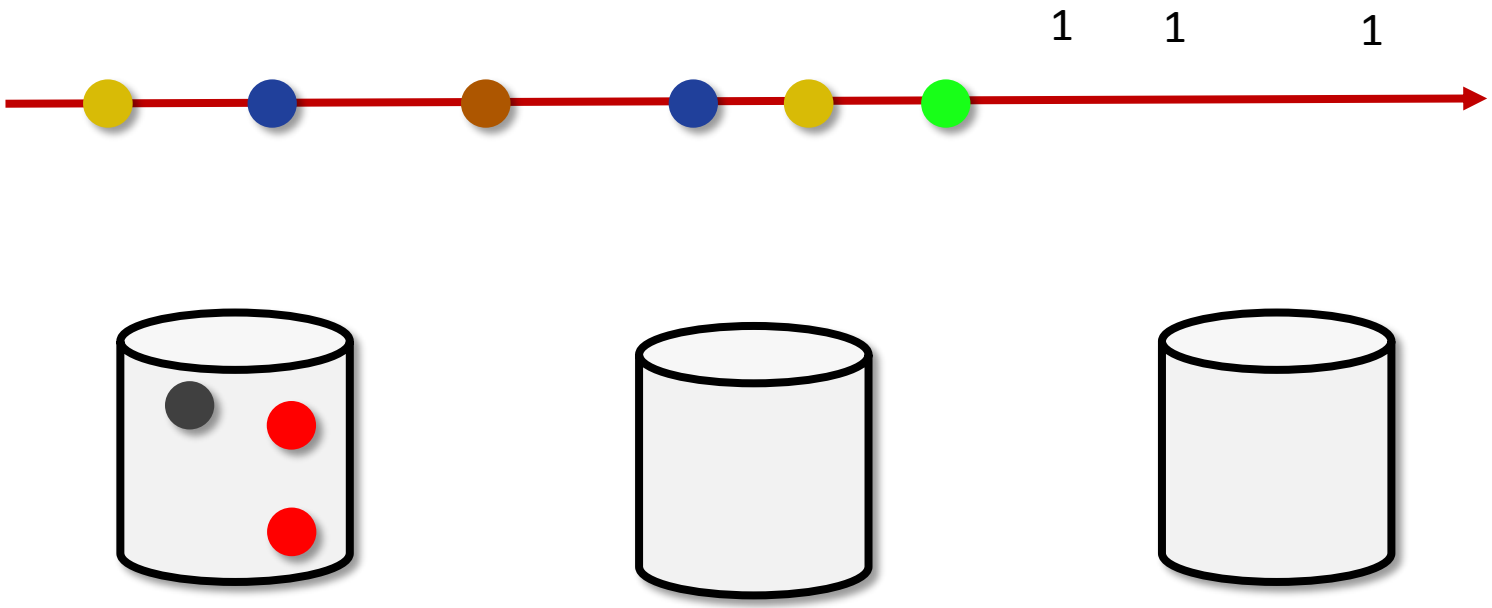
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



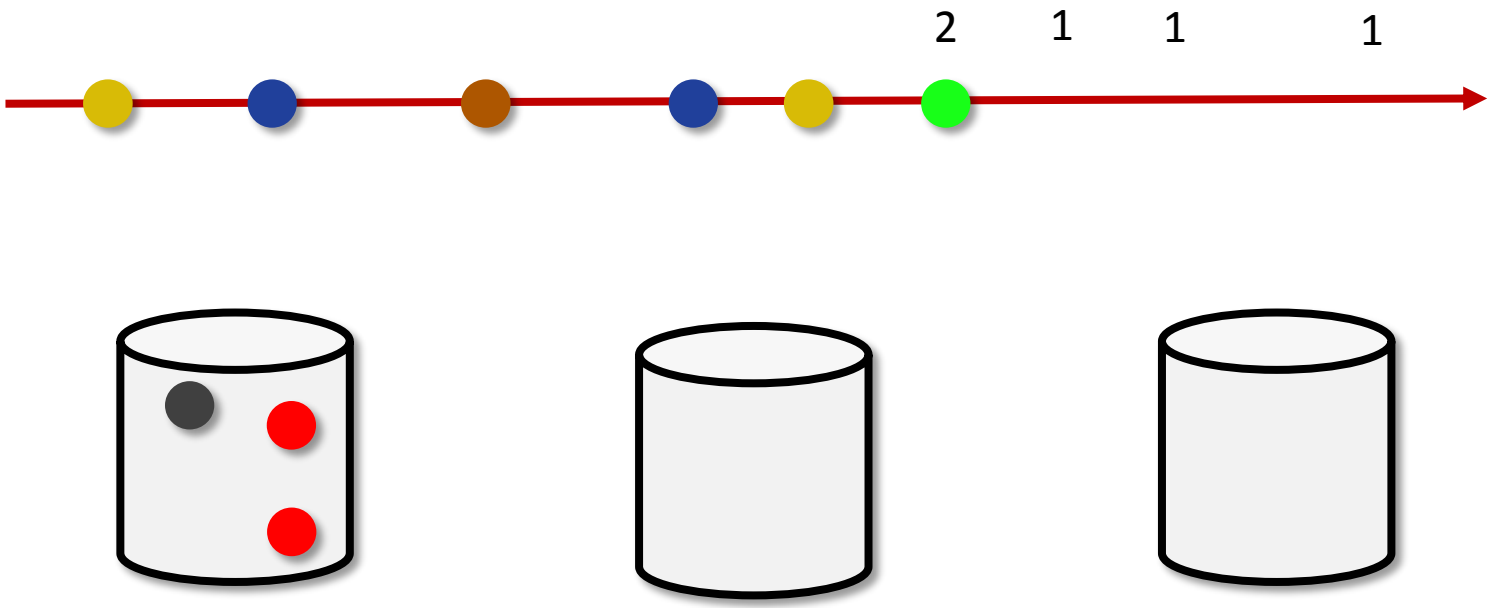
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



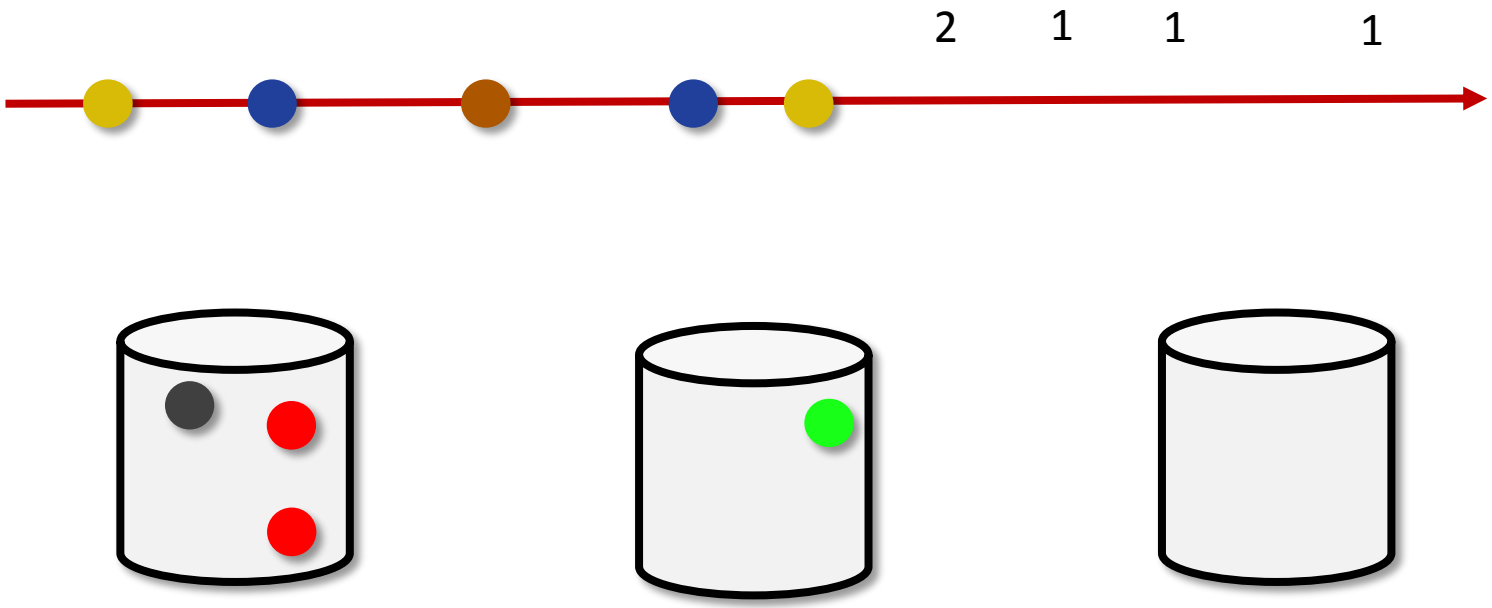
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



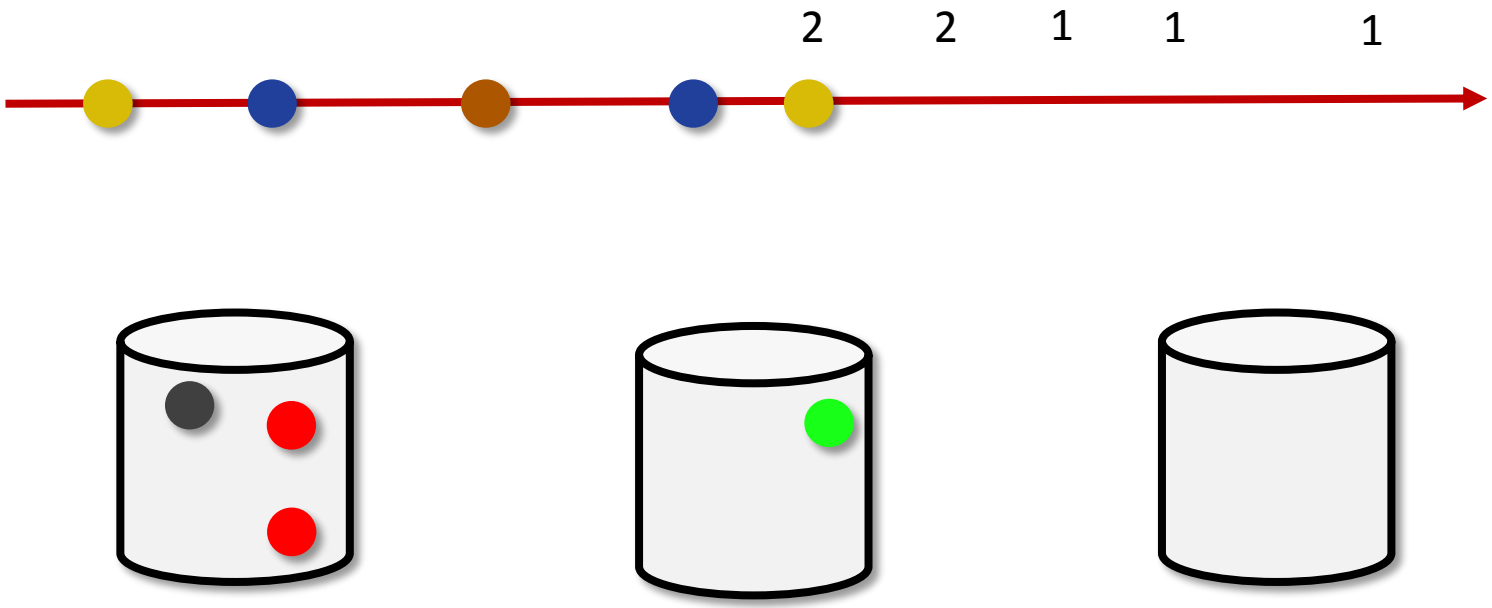
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



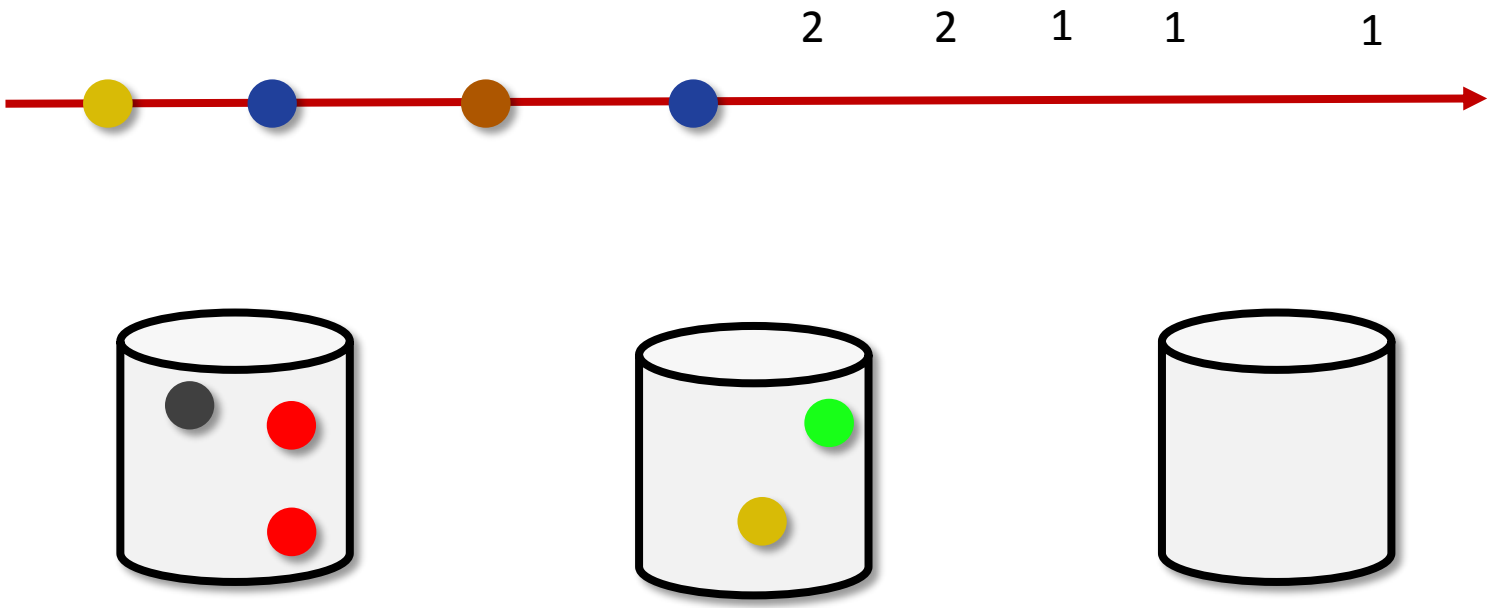
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



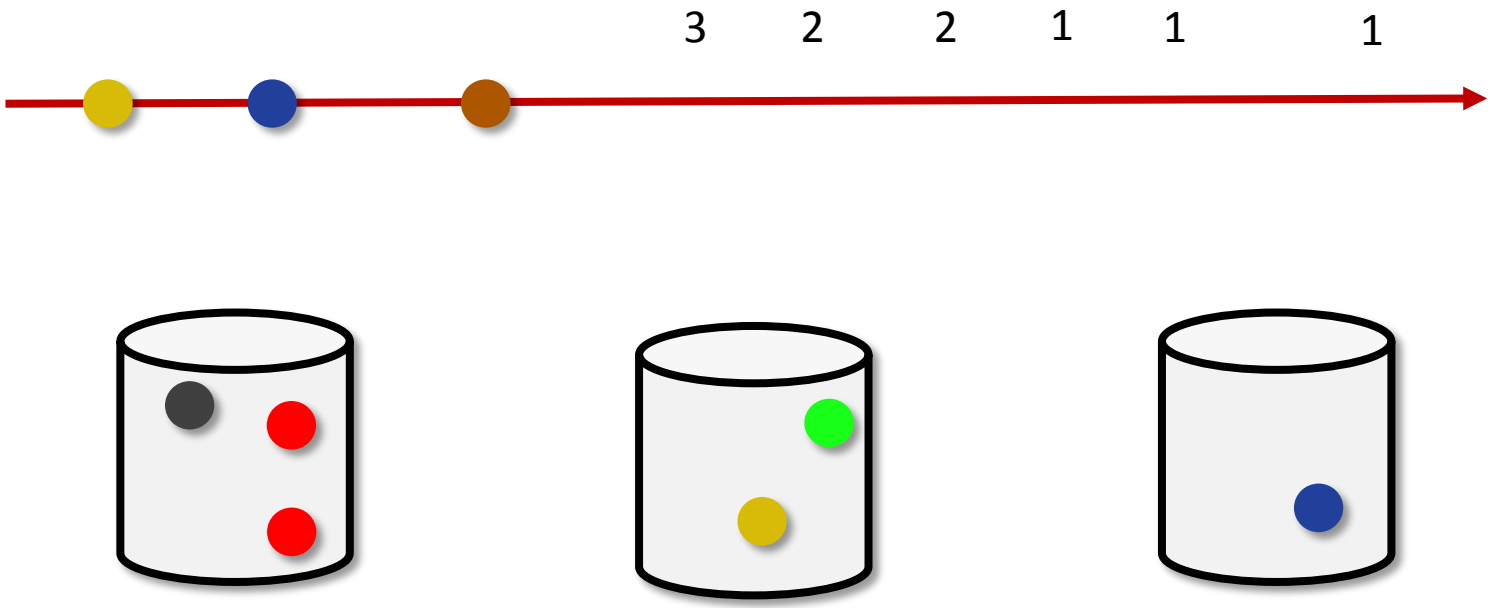
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



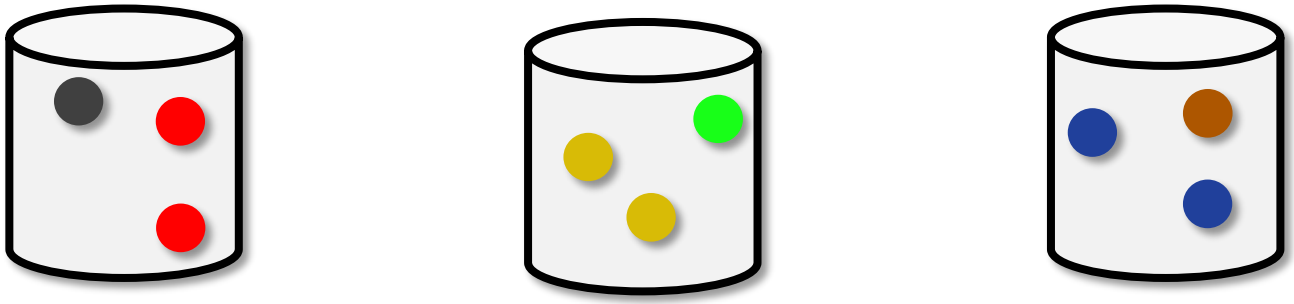
Mintavételezés streamekben: hash

- 1/3-os mintavételezés



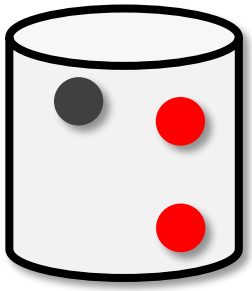
Mintavételezés streamekben: hash

- 1/3-os mintavételezés

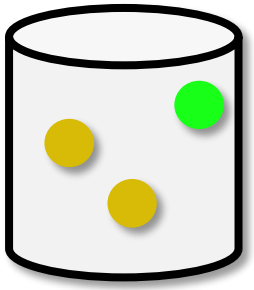


Mintavételezés streamekben: hash

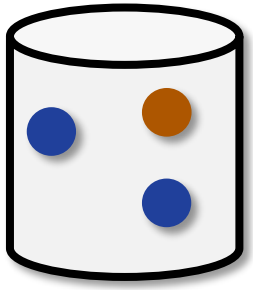
- 1/3-os mintavételezés



$\hat{p} = 1/3$



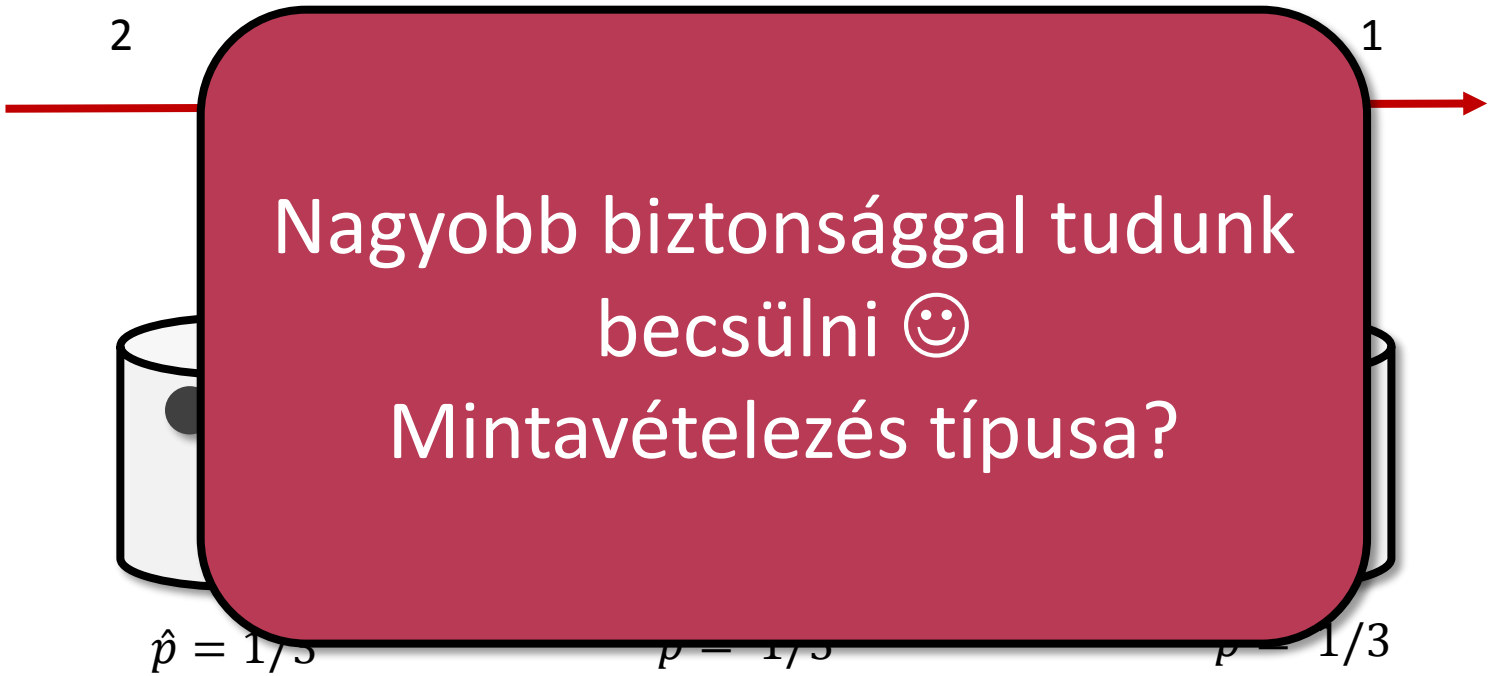
$\hat{p} = 1/3$



$\hat{p} = 1/3$

Mintavételezés streamekben: hash

- 1/3-os mintavételezés



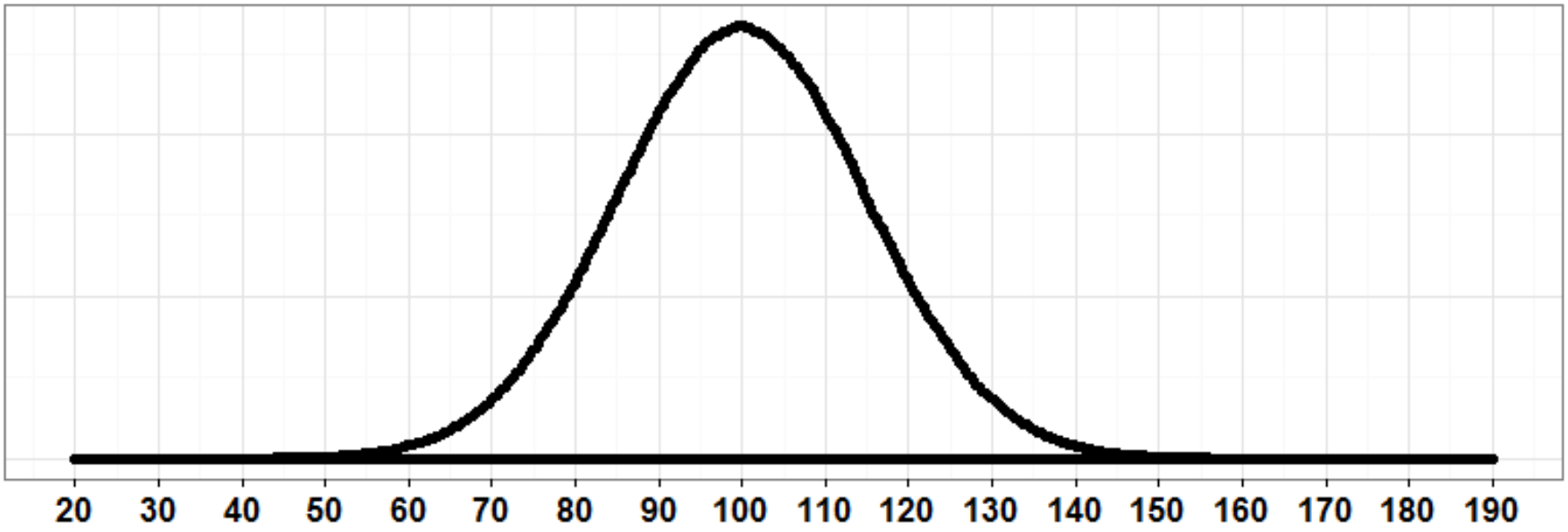
OUTLIER DETEKTÁLÁS

Outlier

„An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins 1980)

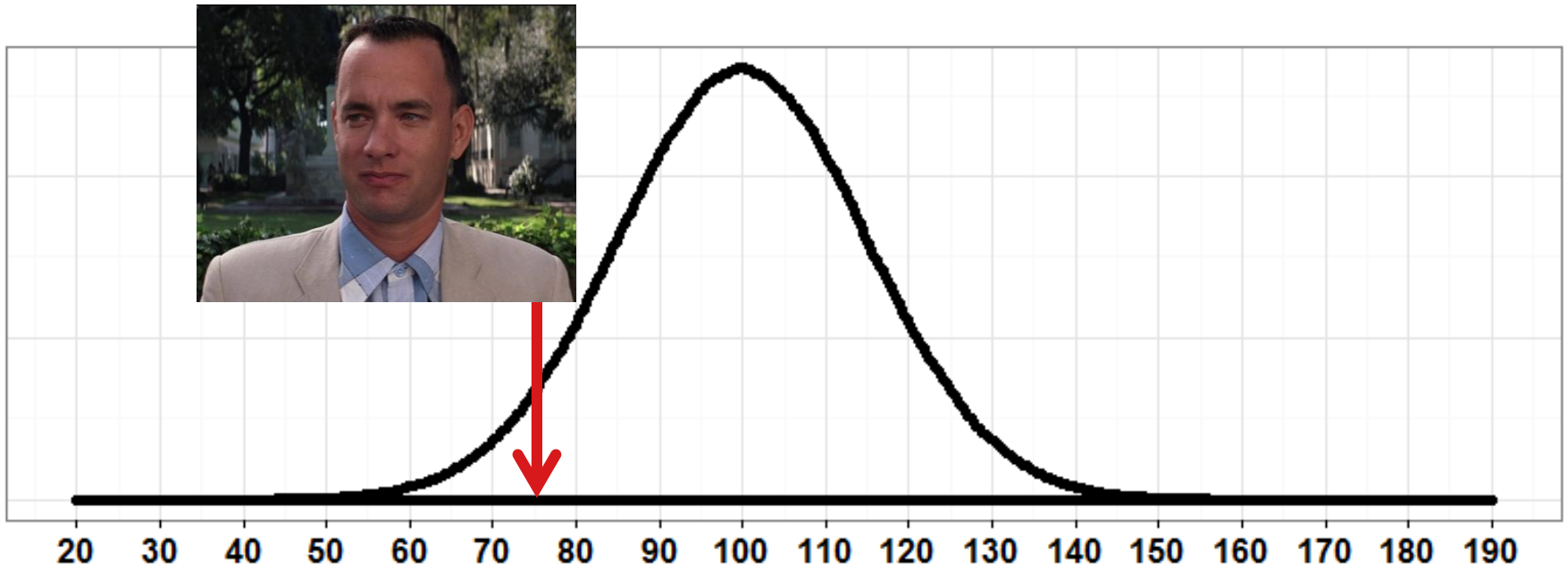
Outlier

„An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins 1980)



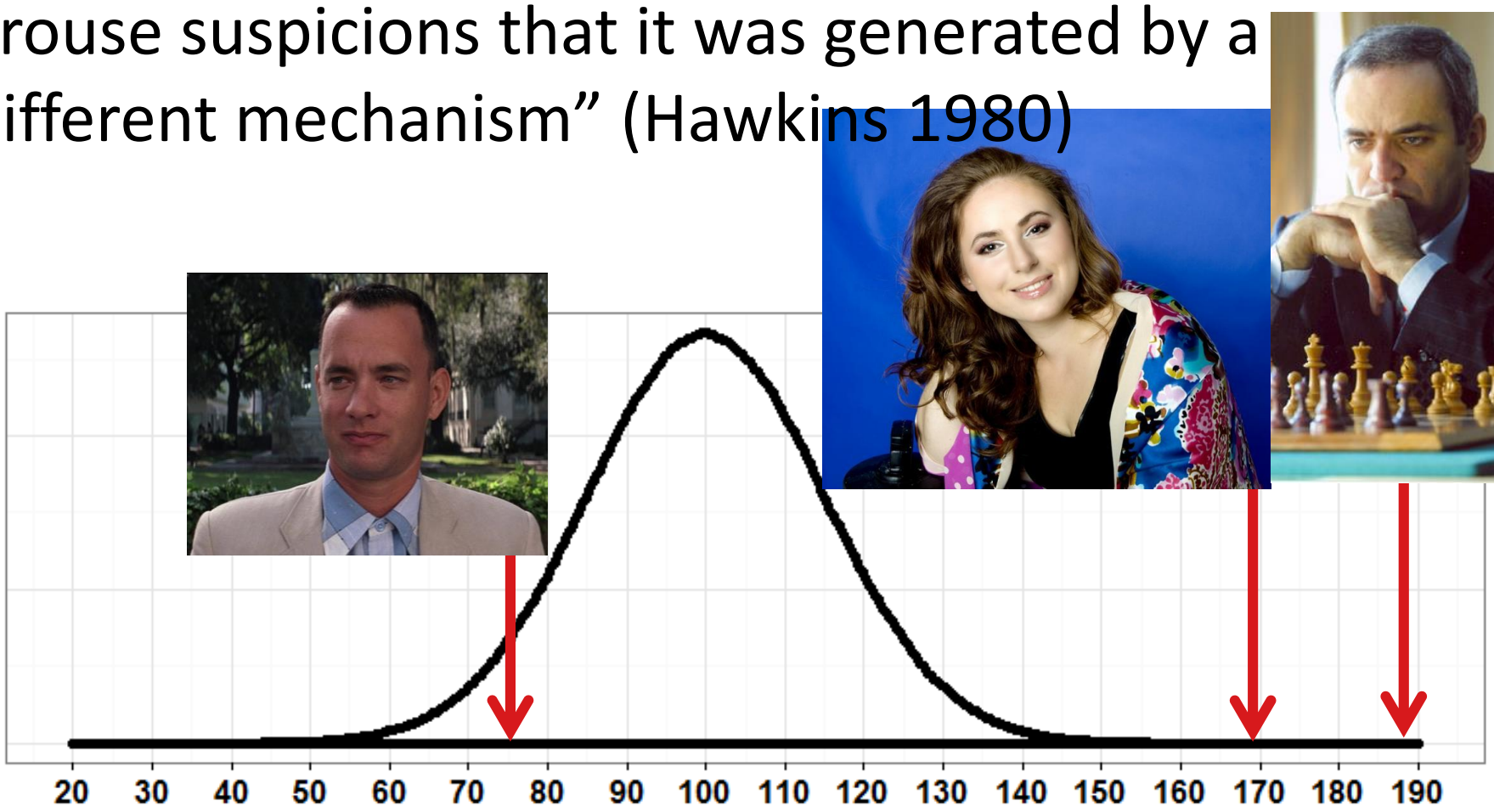
Outlier

„An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins 1980)

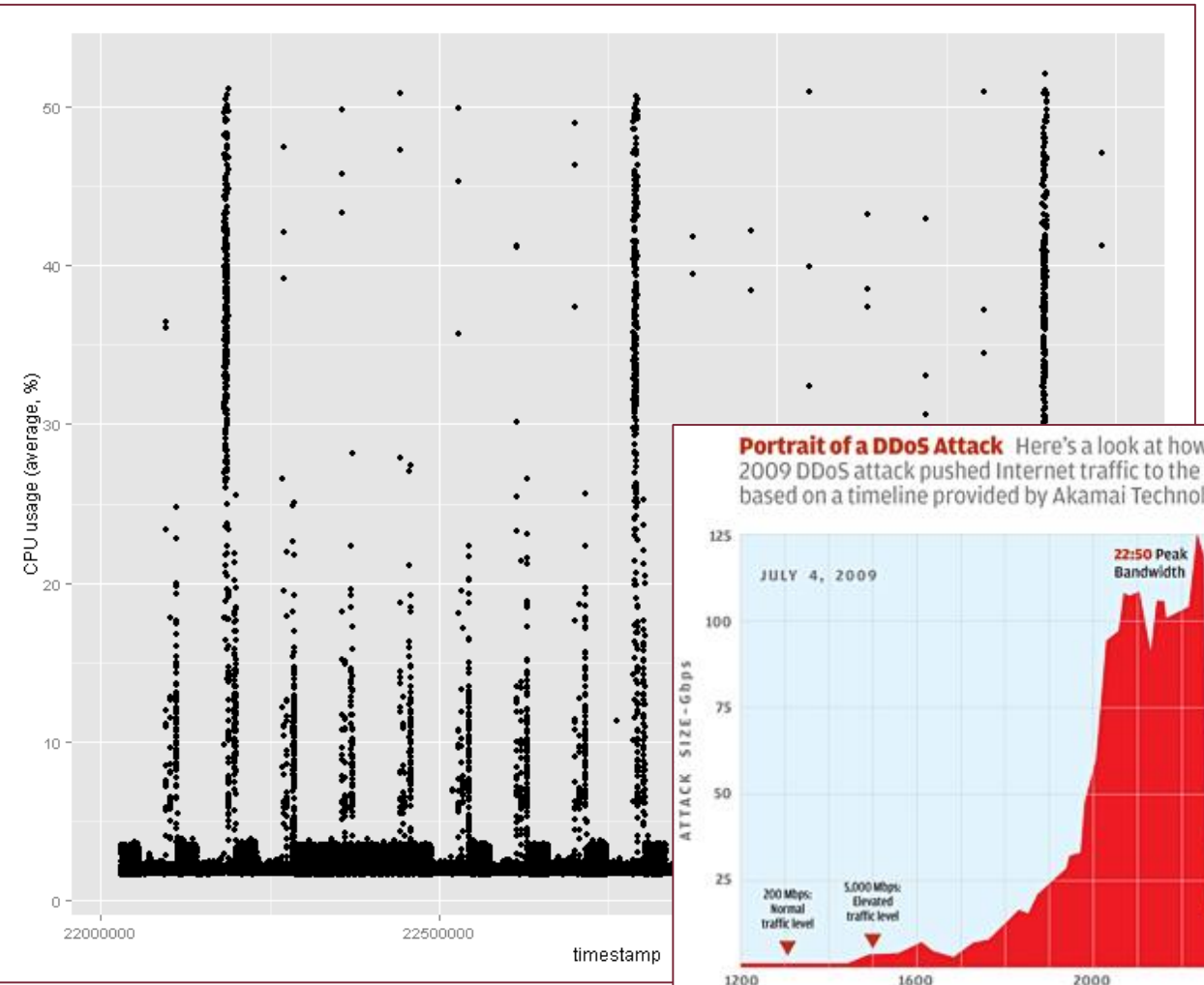


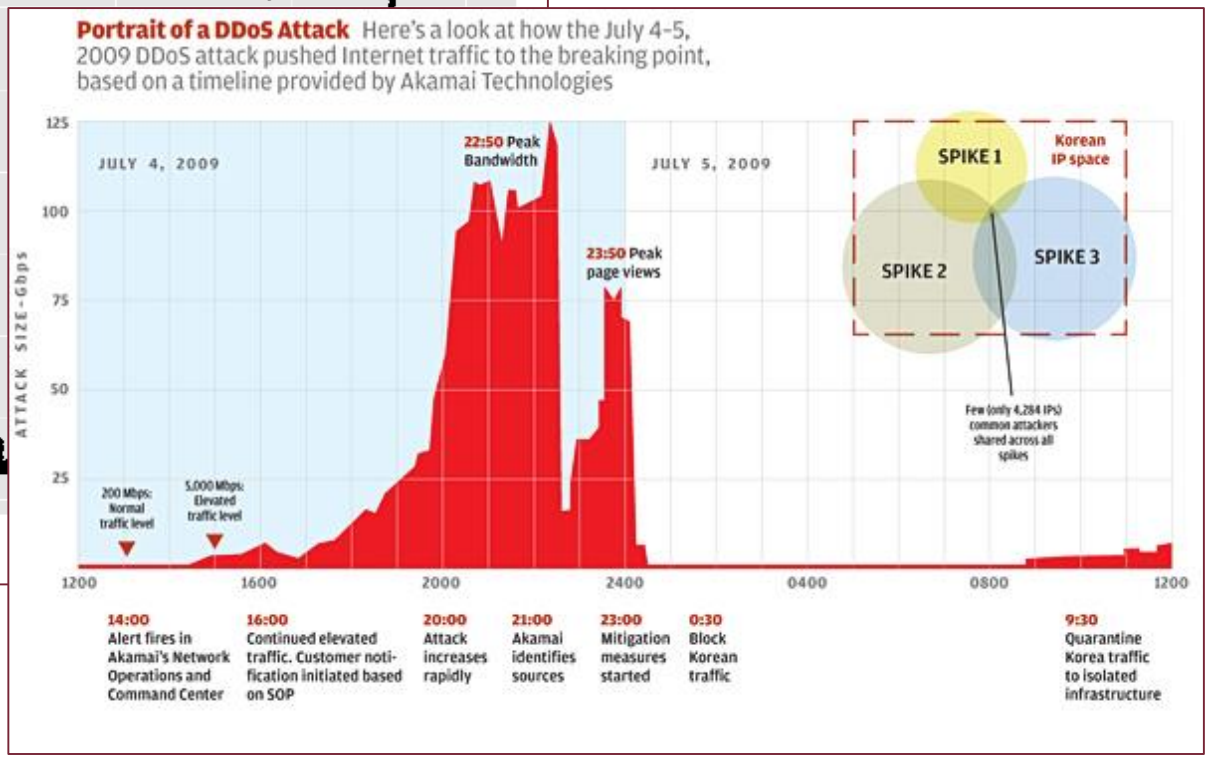
Outlier

„An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins 1980)



Használati esetek



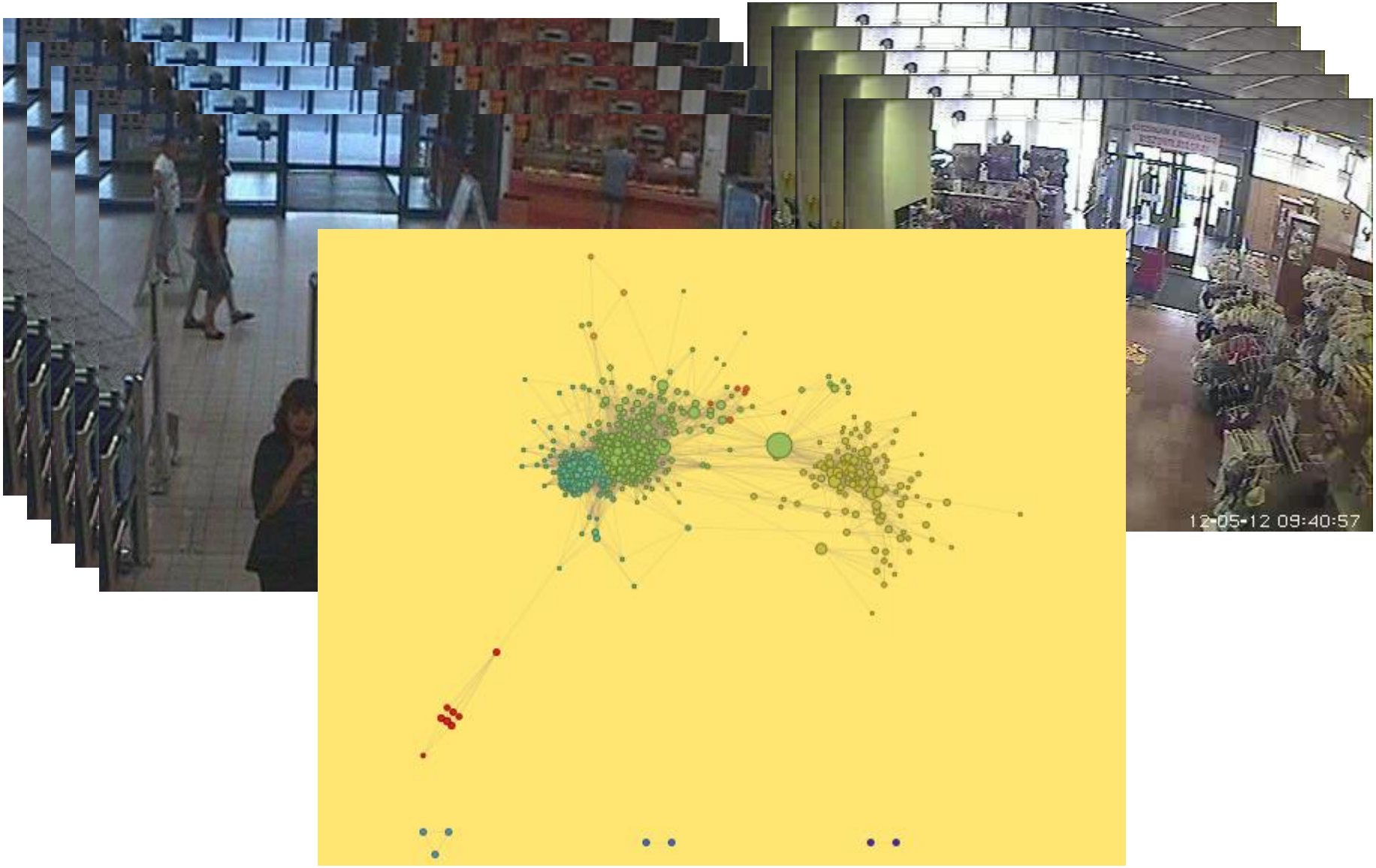


Kép forrása: http://www.csoonline.com/article/592776/the-ddos-attack-survival-guide-

Használati esetek

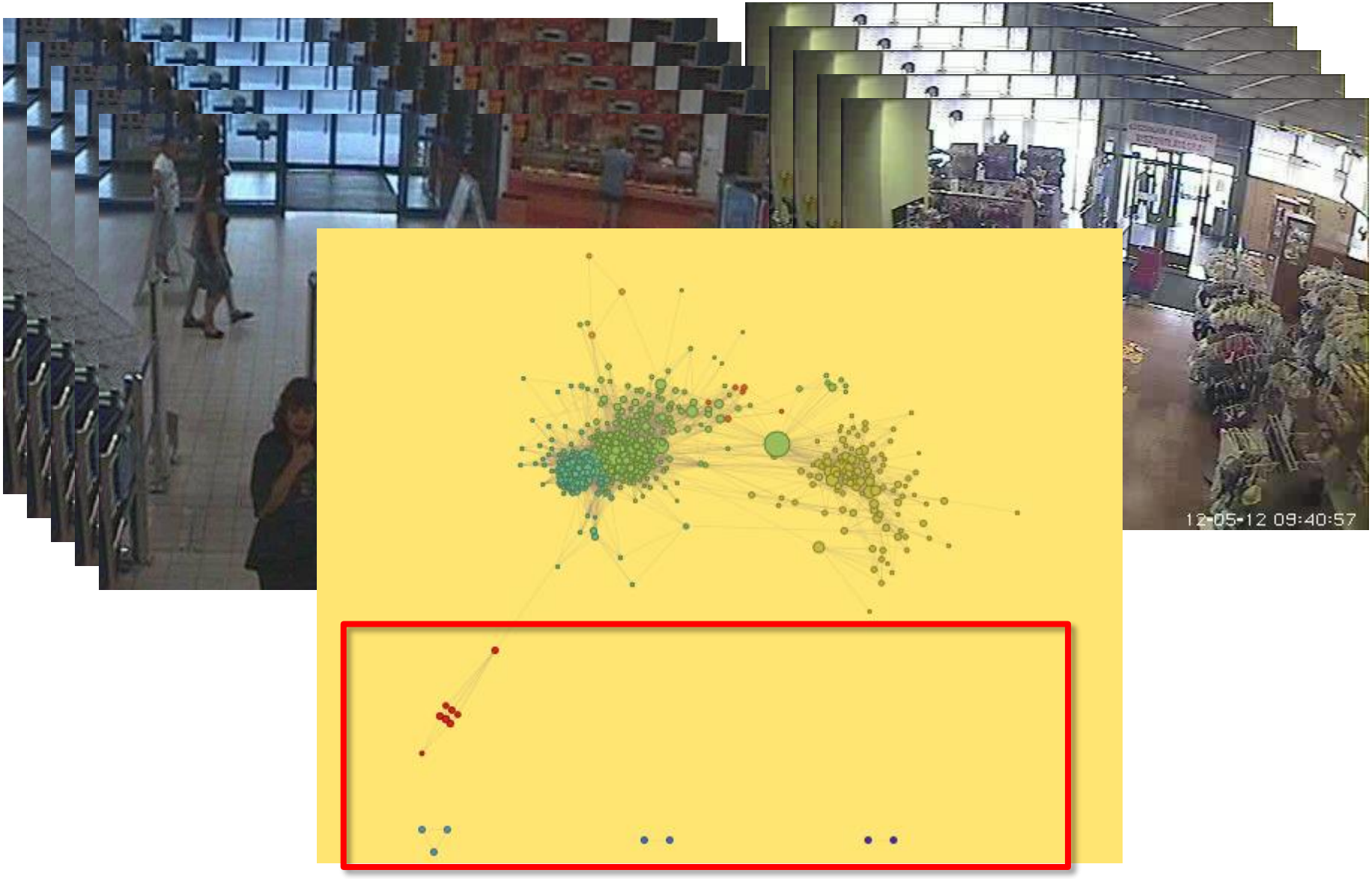
Képek forrása: http://www.szon.hu/

Használati esetek



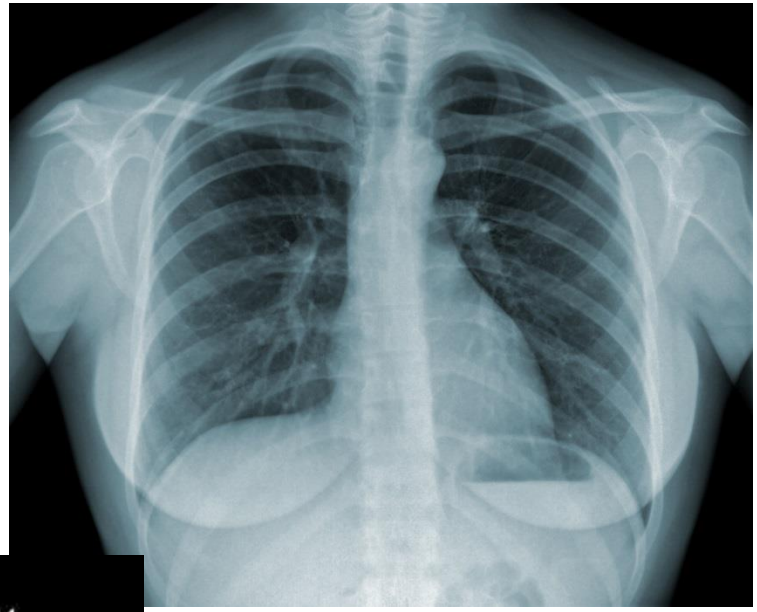
Képek forrása: http://www.szon.hu/

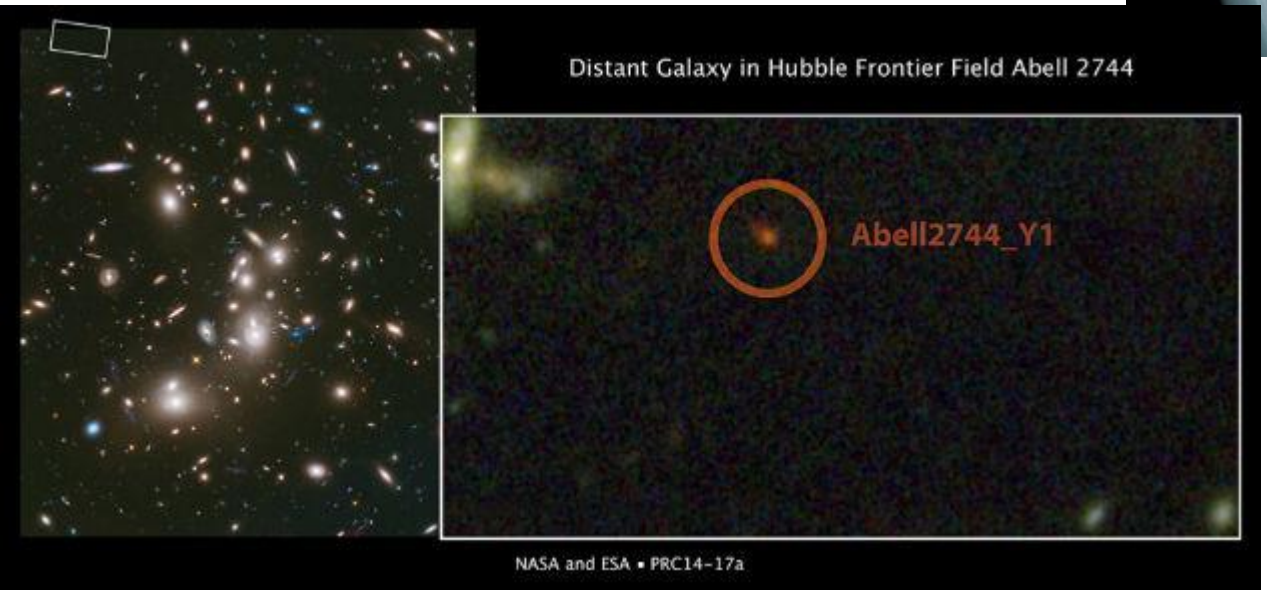
Használati esetek



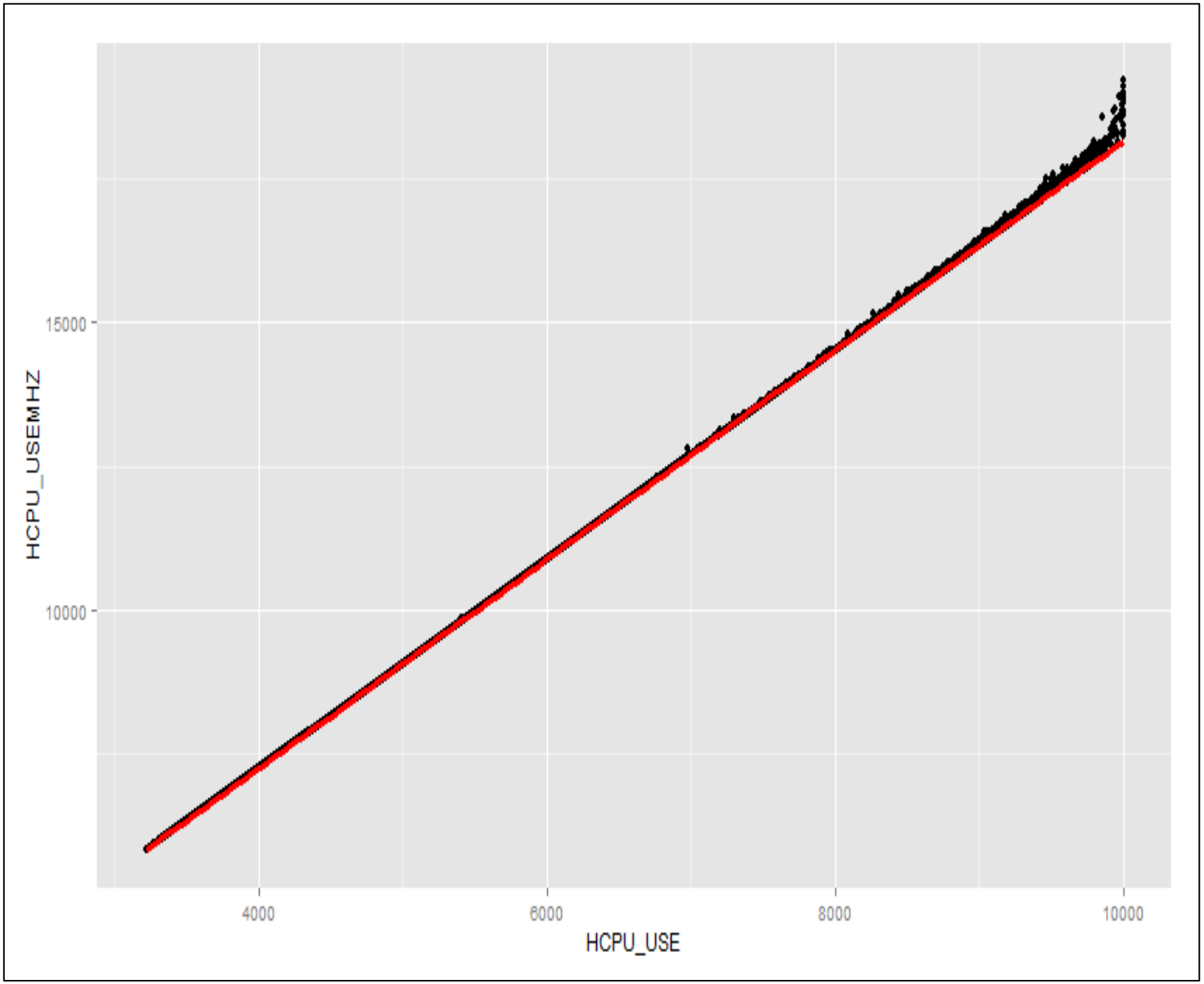
Képek forrása: http://www.szon.hu/

Használati esetek

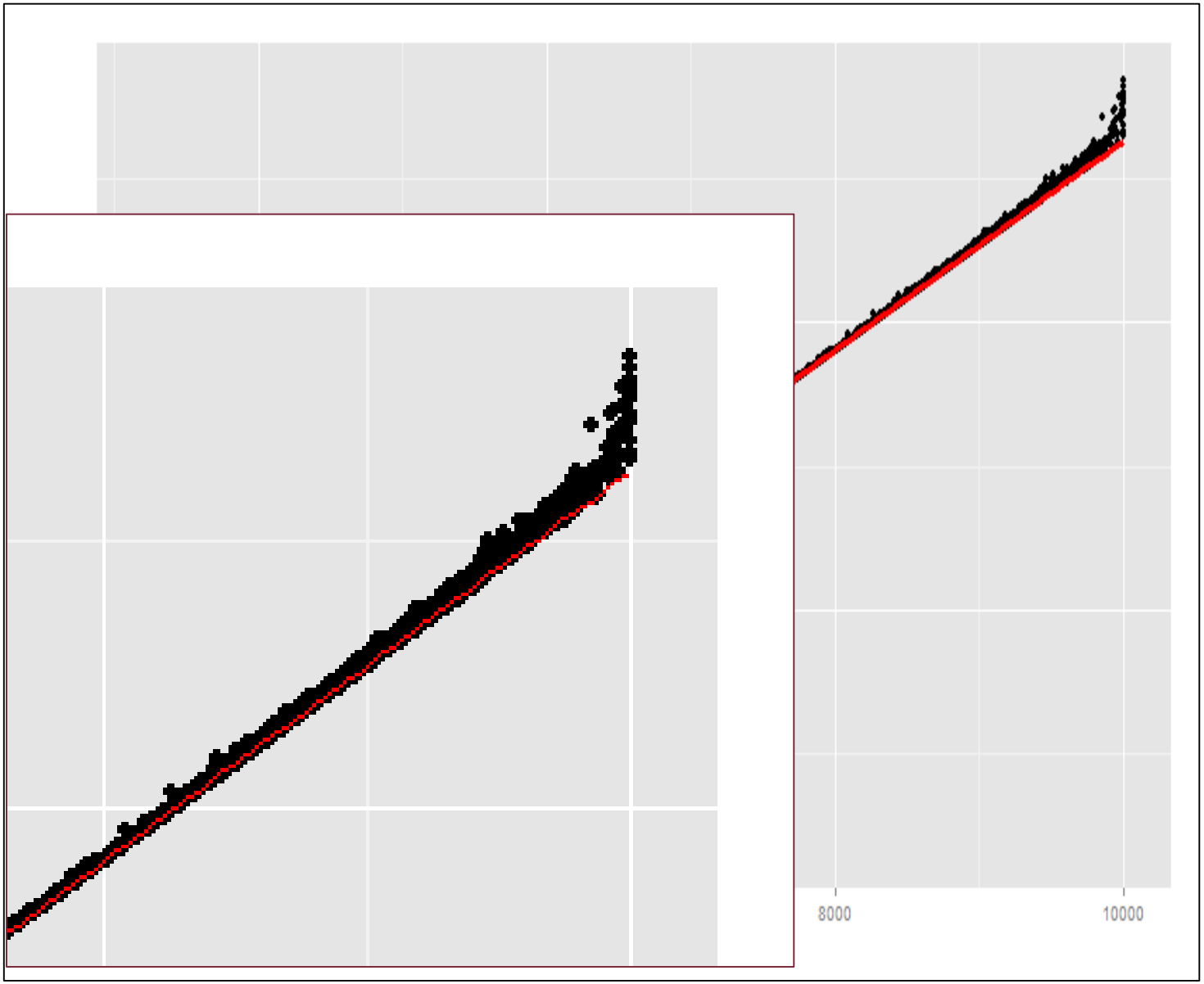




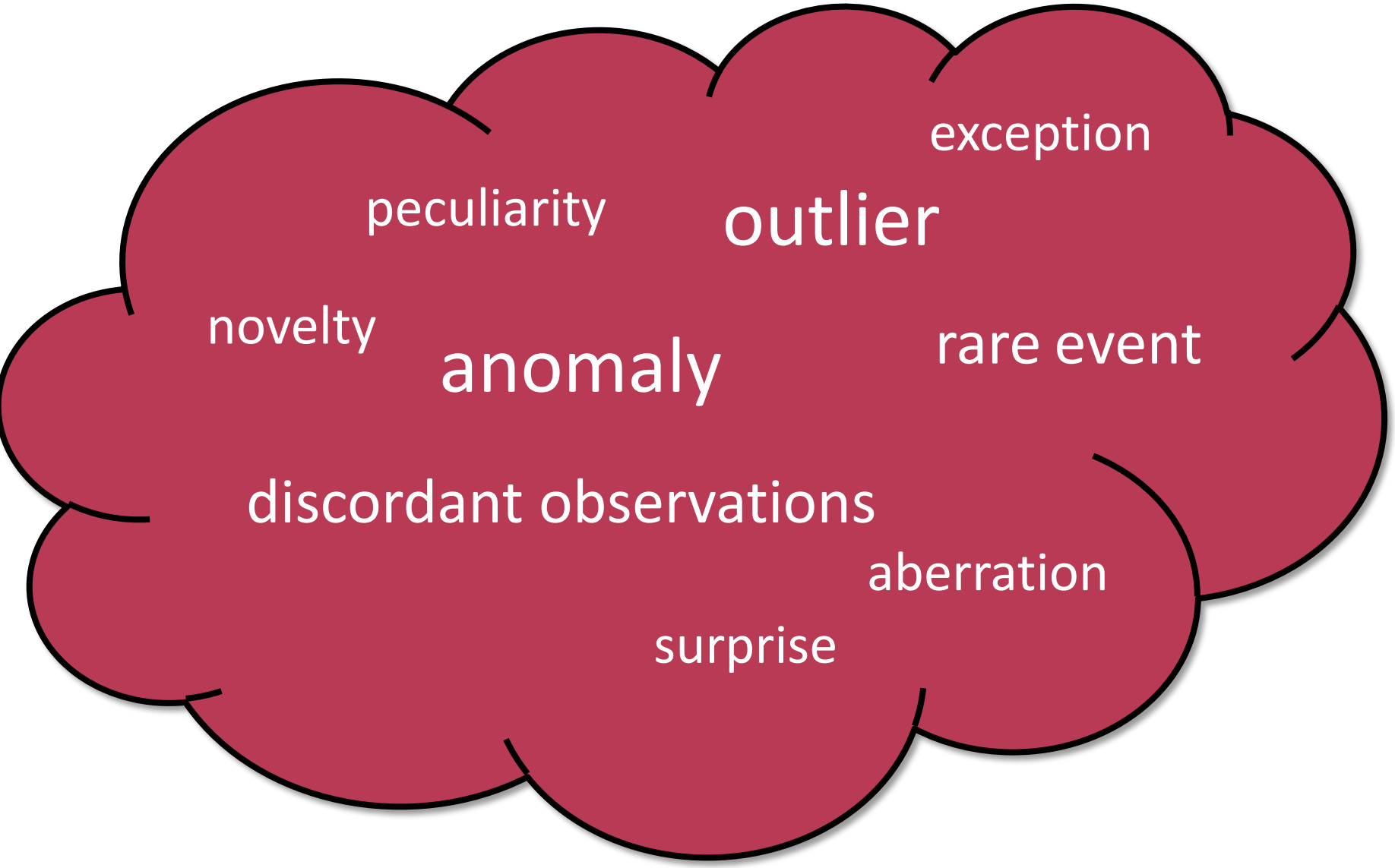
Használati esetek



Használati esetek



Alapfogalmak



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**.

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.

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)

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