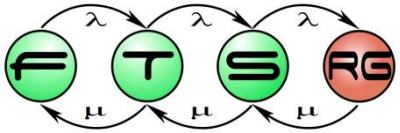
Alapfogalmak az adatelemzésben

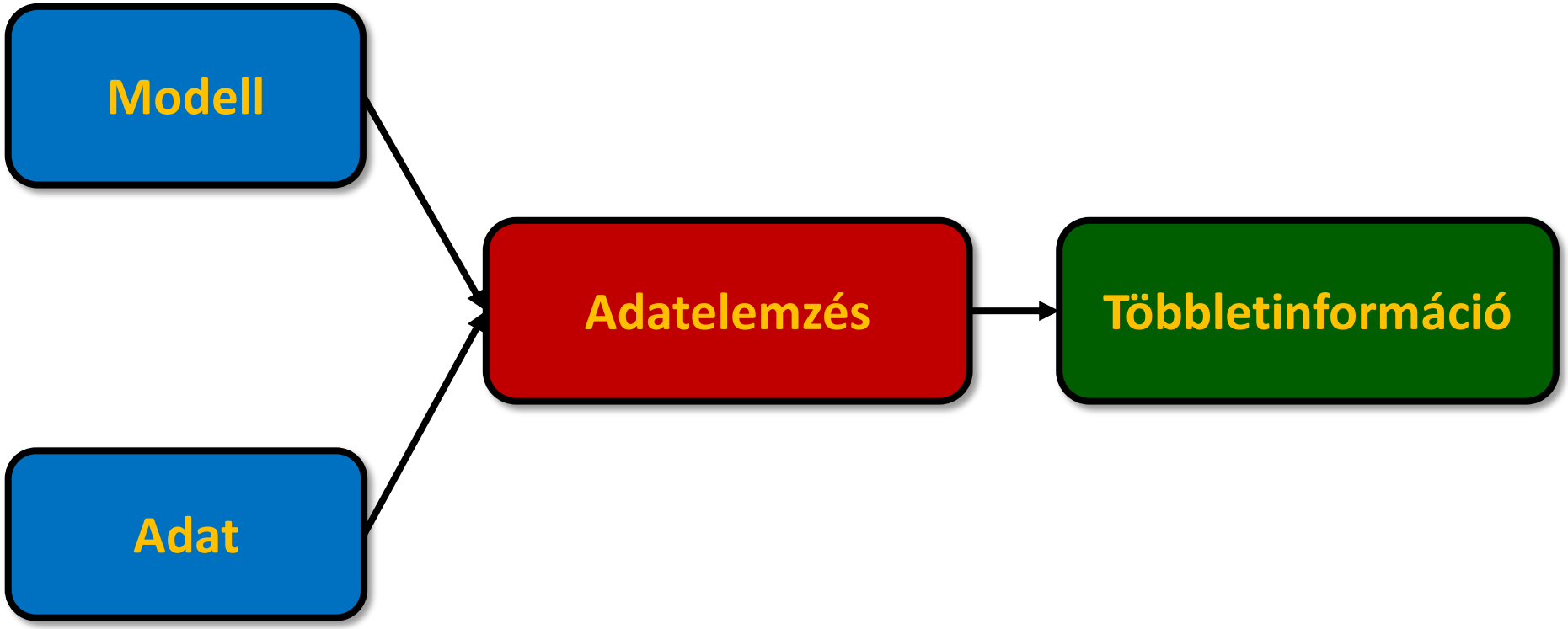
„Big Data” elemzési módszerek

Kocsis Imre, Salánki Ágnes
ikocsis@, salanki@

2015



Adatelemzés



Modell

- Szakértői tudás
  - Elvárt összefüggések
  - Háttértudás a kísérletről
  - …

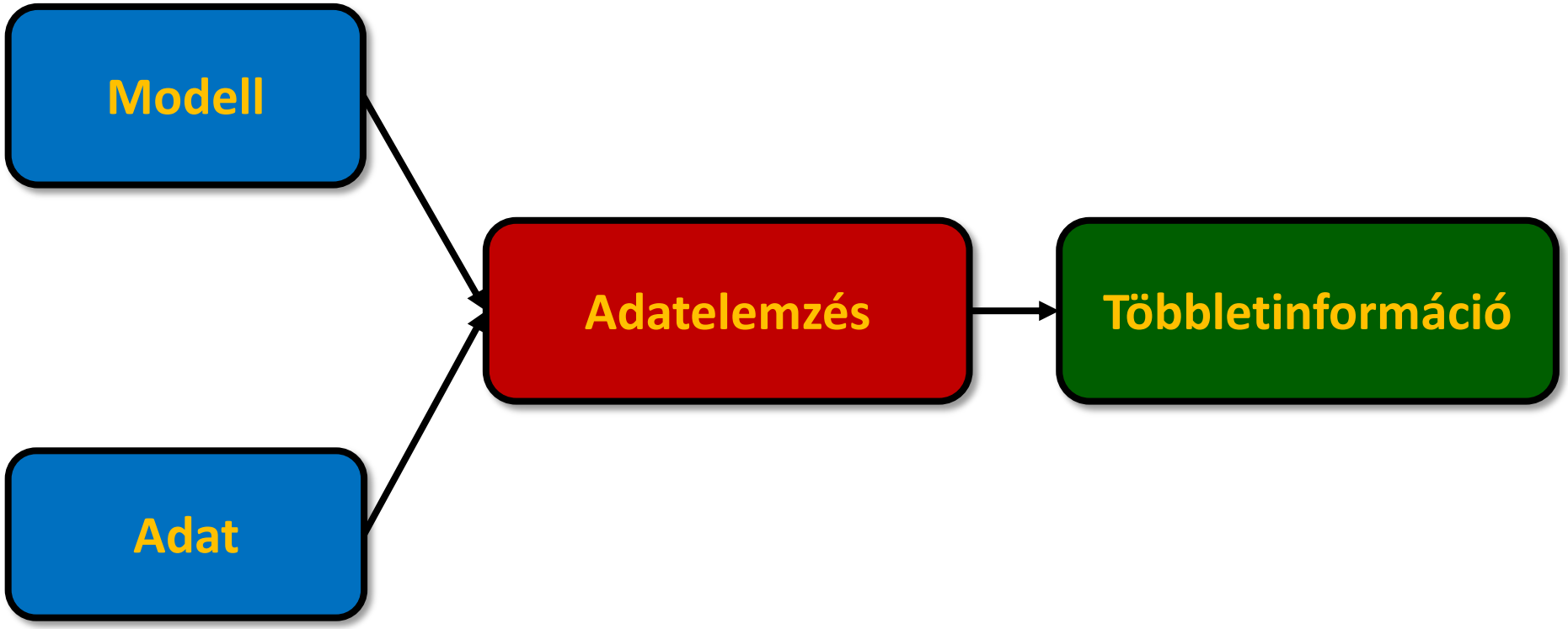
	timestamp	mem.free	mem.used	mem.system
1	1411466821	0.10080218	0.40543602	0.49376180
2	1411466841	0.15876118	0.30825174	0.53298707
3	1411466861	0.36164311	0.38978400	0.24857288
4	1411466881	0.27752493	0.16036882	0.56210626
5	1411466901	0.38525492	0.24269648	0.37204860

Modell

- Szakértői tudás
  - Elvárt összefüggések
  - Háttértudás a kísérletről
  - …

	date	temp_avg	tmp_max	tmp_min	prec_sum	prec_type
1	2000-12-27	4.1	6.1	2.9	3.5	1
2	2000-12-28	6.2	8.4	4.3	NA	NA
3	2000-12-29	4.8	6.2	3.7	14.6	1
4	2000-12-30	2.7	4.4	1.4	5.5	4
5	2000-12-31	1.5	3.7	0	NA	NA

Adatelemzés



Adat

- Strukturált
  - Rögzített formátum
  - Általában $n \times p$ [illegible]
- Nemstrukturált
  - Nincs előre rögzített tárolási/értelmezési modell

Változó/ attribútum

	timestamp	mem.free	mem.used	mem.system
1	1411067907	0.185196470	0.352313839	0.4624897
2	1411067927	0.324186013	0.307872234	0.3679418
3	1411067947	0.225055351	0.141652608	0.6332920
4	1411067967	0.337036606	0.158649255	0.5043141
5	1411067987	0.228445467	0.325578194	0.4459763
6	1411068007	0.275945384	0.377976430	0.3460782
7	1411068027	0.180429998	0.235329471	0.5842405
8	1411068047	0.178786330	0.161953066	0.6592606
9	1411068067	0.263418770	0.238677062	0.4979042

Rekord/ megfigyelés

Adat

- Strukturált
  - Rögzített formátum
  - Általában $n \times p$-s táblázat
  - „Tidy”
    - Sor: pontosan egy megfigyelés
    - Oszlop: pontosan egy változó
- Nemstrukturált
  - Nincs előre rögzített tárolási/értelmezési modell

	id	var1	var2
1	1	A	A
2	2	B	A
3	3	A	B
4	4	B	B
5	5		

Széles

	row.names	id	variable	value
1	1	1	var1	A
2	6	1	var2	A
3	2	2	var1	B
			var2	A
		3	var1	A

Hosszú

Adat

- Strukturált
  - Rögzített formátum
  - Általában $n \times p$-s táblázat
  - „Tidy”
- Nemstrukturált
  - Nincs előre rögzített tárolási/értelmezési modell
  - Csak metaadat

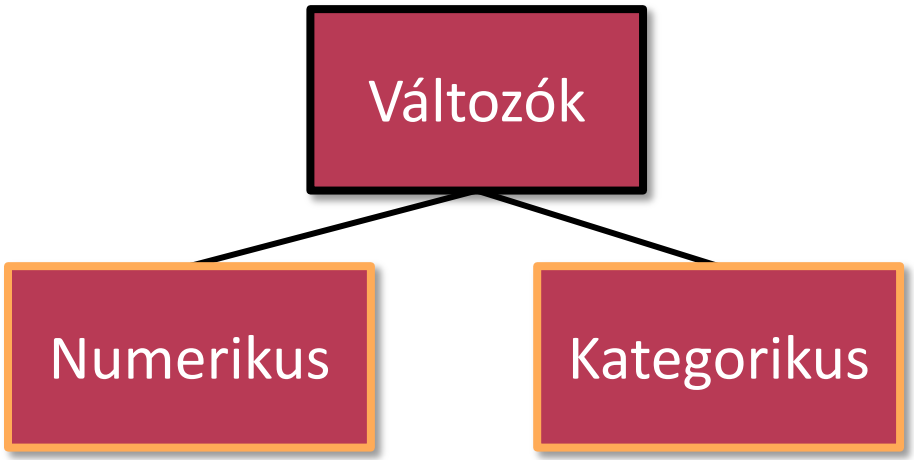
k

**** Szemistrukturált adat?**

- **⊂ Strukturált**
- **Nem ábrázolható hatékonyan adattáblában**
  - **Azonos típusú objektumok más attribútumokkal is**
  - **Az attribútumok sorrendje nem számít**
- **Pl. XML, JSON**

Numerikus és kategorikus változók

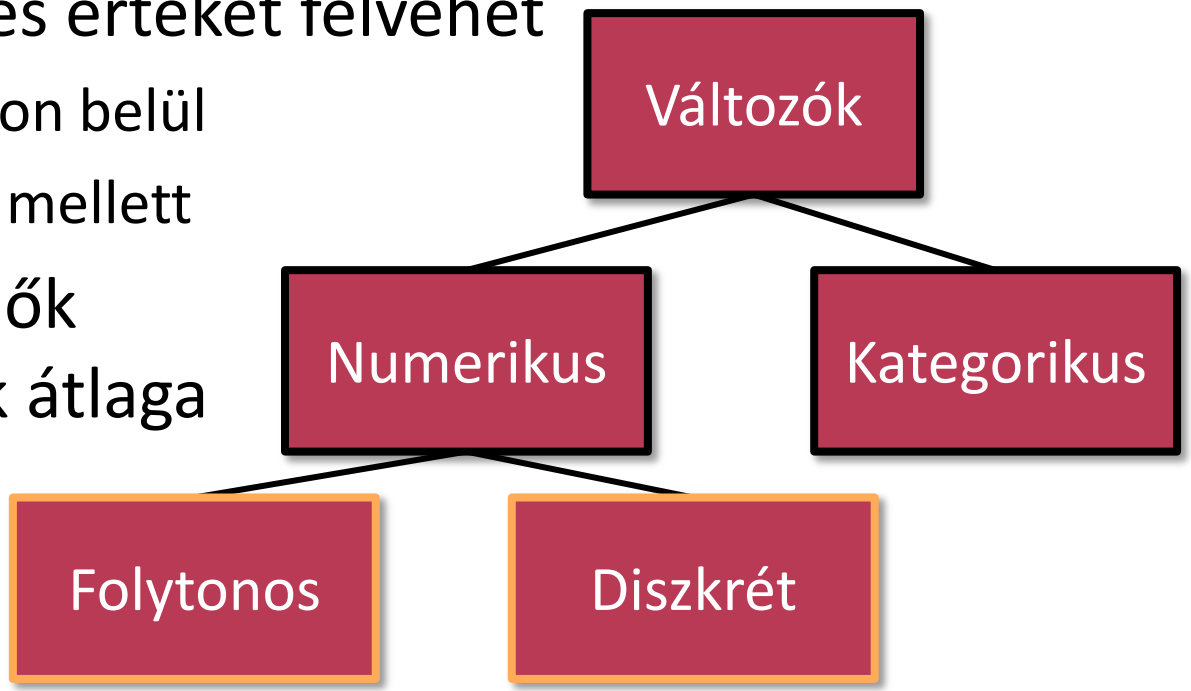
- Numerikus (numerical)
  - az alapvető aritmetikai műveletek értelmesek
  - Pl. átlaghőmérséklet, kor



- Kategorikus (categorical)
  - Csak megkülönböztetés miatt
  - Pl. telefonszám, nem

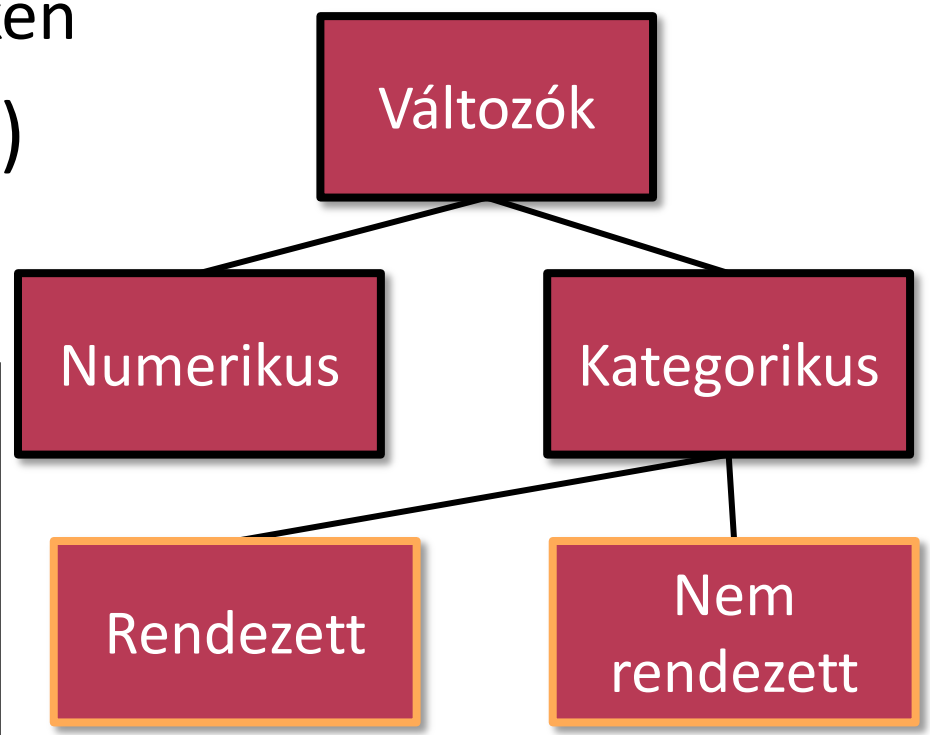
Numerikus változók

- Folytonos
  - Mért – tetszőleges értéket felvehet
    - adott tartományon belül
    - adott pontosság mellett
  - Pl. a teremben ülők BigData jegyének átlaga
- Diszkrét
  - Számolt – véges sok értéket vehet fel adott tartományban
  - Pl. BigData előadáson ülők száma



Kategorikus változók

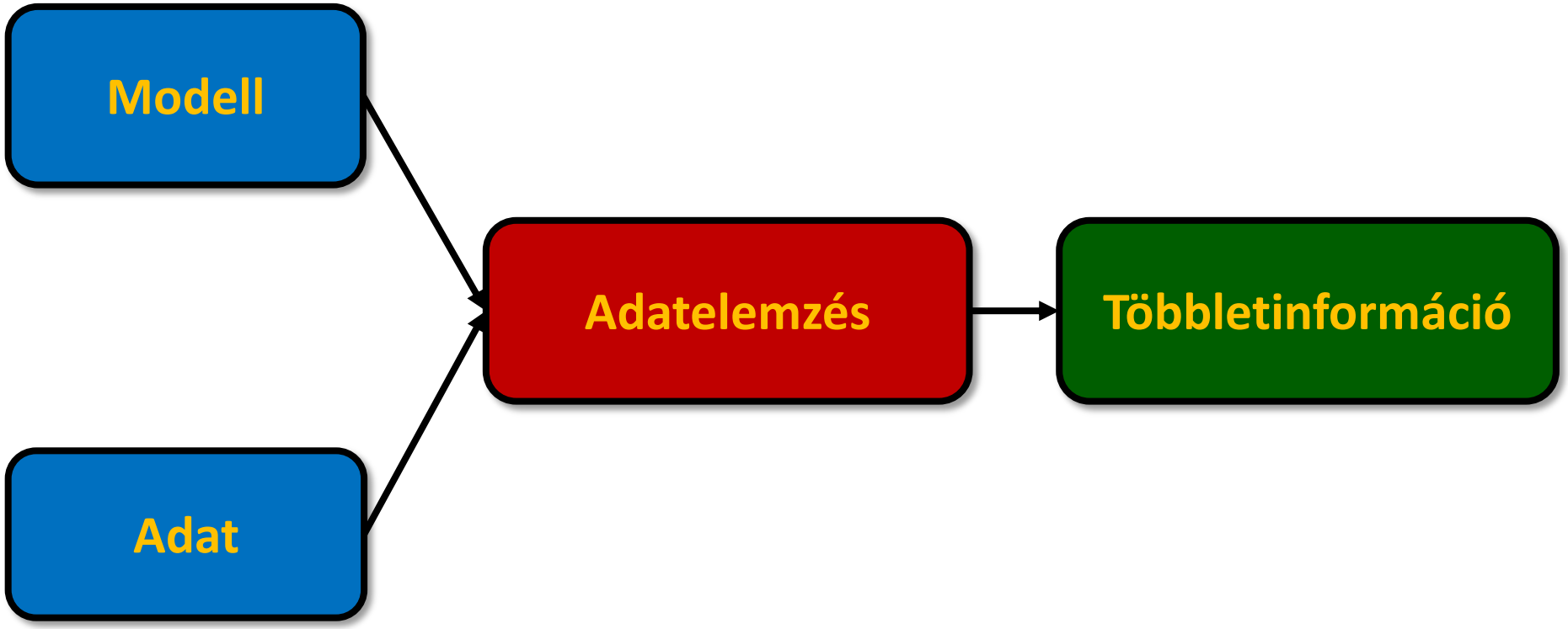
- Rendezett
  - Teljes rendezés az értékeken
- Nem rendezett (reguláris)



9. Ajánlanád-e a tárgyat másoknak?

- ○ Mindenkit rábeszélnék
- ○ Nyugodtan ajánlanám
- ○ Esetleg ajánlanám
- ○ Inkább lebeszélném róla
- ○ Feltétlenül lebeszélném
- ○ Nem kívánok válaszolni

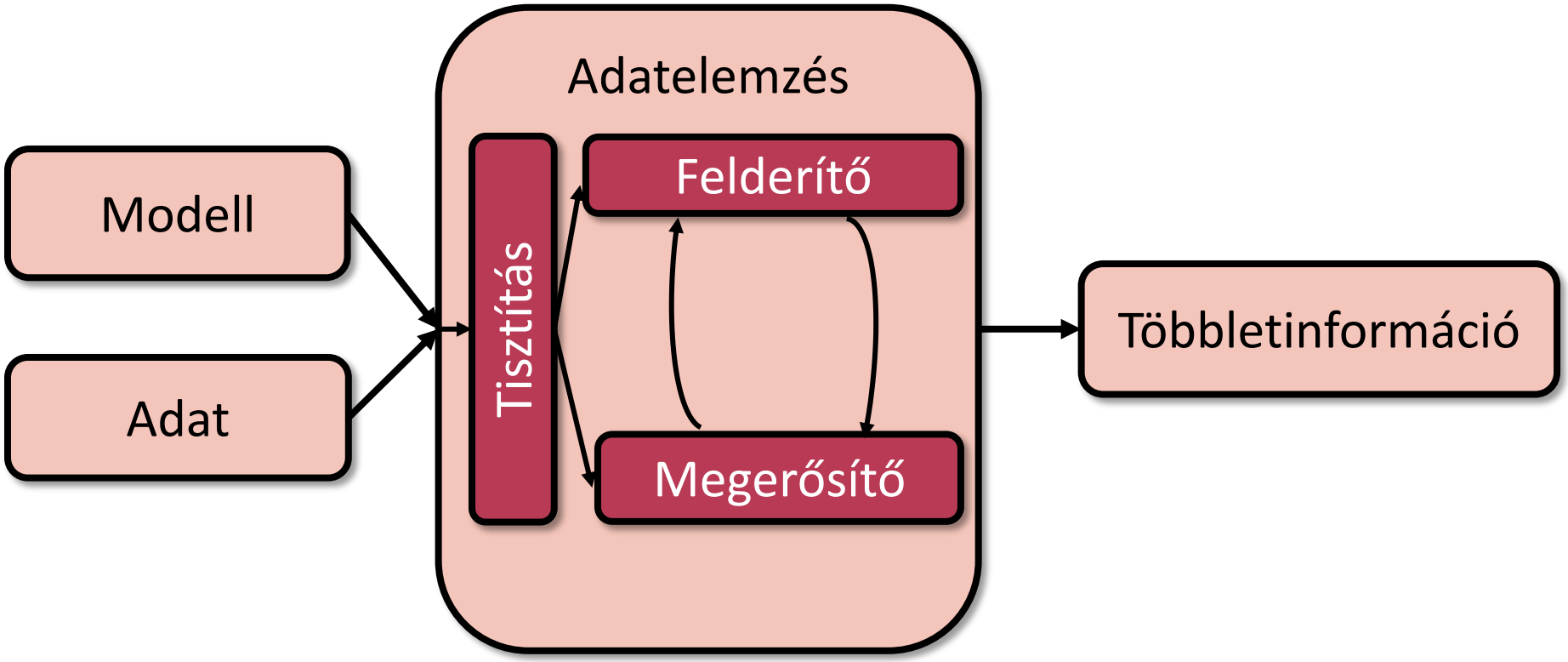
Adatelemzés



Többletinformáció

- „A kávét vásárlók gyakran vásárolnak tejet”
- „A férfiak és nők előléptetési aránya különbözik”
- „Az alkalmazás memóriaigénye a kiszolgálandó kérések számával exponenciálisan növekszik”
- „A teljes populáció IQ-ja $N(100, 15)$ eloszlást követ”
- „Az Apple részvények vételi árának prediktált ingadozása a következő hónapban 2”
- „A BME-s hallgatók tanulmányi átlaguk alapján 3 jól elkülöníthető csoportba tartoznak”

Adatelemzés



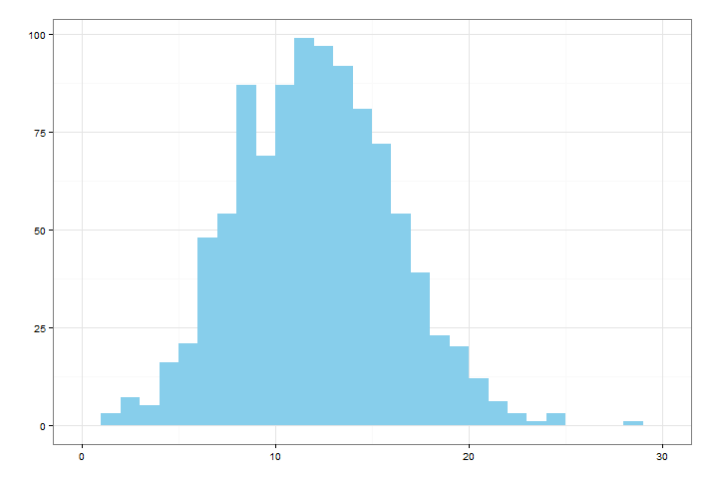
Adatelemzés

Adatelemzés

- Pl. eloszláselemzés

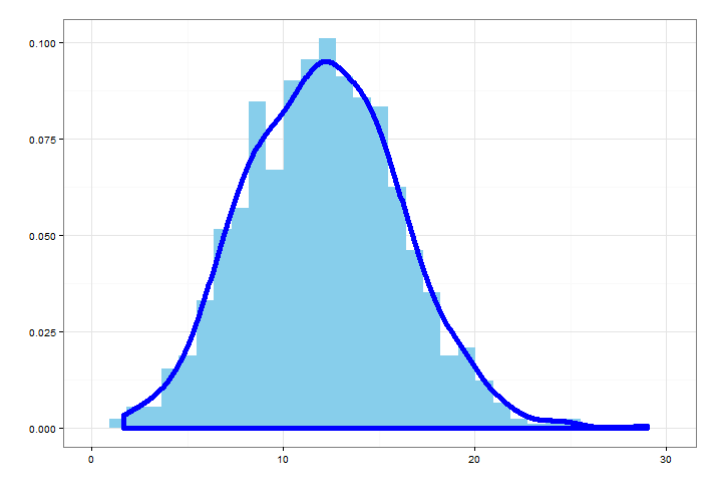
Felderítő analízis

Sejtés: az x változó normális eloszlású



Megerősítő analízis

Az x változó hihetően $N(12, 4)$ eloszlást követ

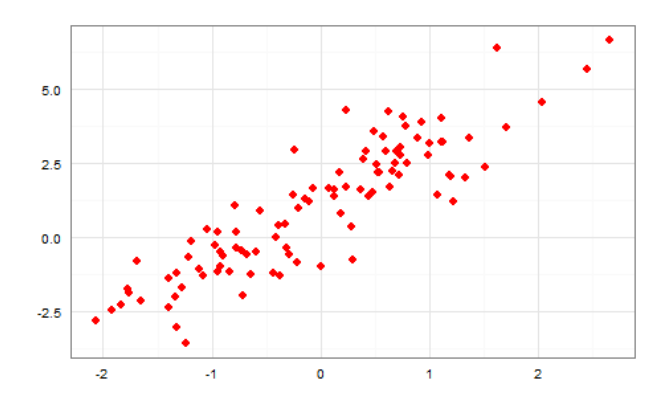


Adatelemzés

- Pl. lineáris regresszió

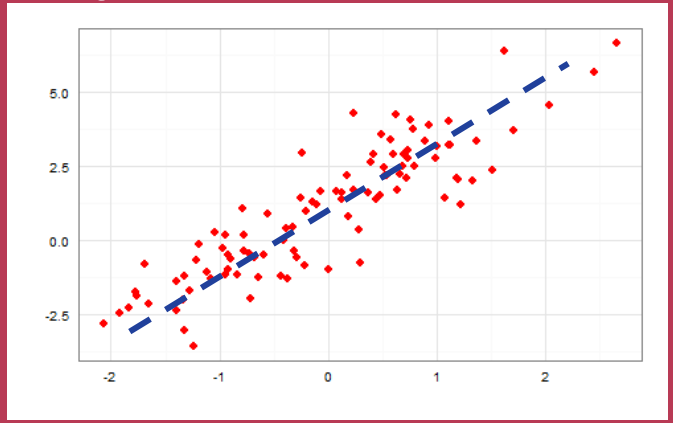
Felderítő analízis

Sejtés: az x és y változó között valamilyen lineáris kapcsolat van

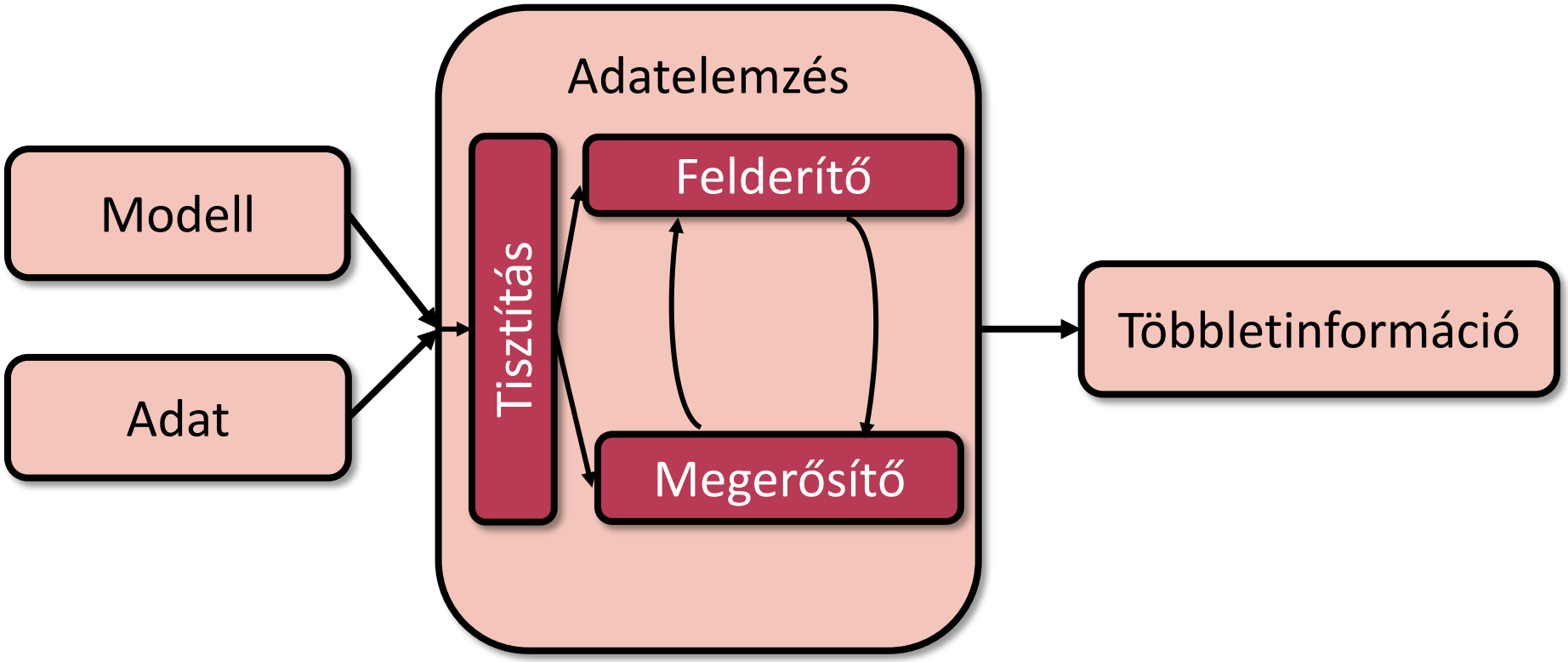


Megerősítő analízis

Az x és y változó között az $Y = 3.2 \times x - 1.2$ írható fel, $R^2 = 0.1234$



Adatelemzés



Adattisztítás

- Adattisztítás!
  - Meglepően hosszú tud lenni
  - Legtöbbször nem tökéletes
  - Big Data?
- Inkonzisztenciák
  - Beviteli/mérési hibák, „hibás join”, részleges megfigyelés, hamisítás, …
- Kieső értékek (*outliers*)
  - =/= „durva hiba” (*gross error*)
  - Nem feltétlenül előnytelen, de klasszikusan az
  - Magas dimenziószámnál nehéz lehet detektálni
  - Alacsony dimenziós vizualizáció segíthet

Adattisztítás

- Hiányzó adatok (*missing data, „NA”, „null”*)
  - Hol lehet probléma?
  - Mesterséges feltöltés („*imputation*”)

- Több változó, mint megfigyelés/minta
  - Génkifejeződési vizsgálatok
  - Műholdképek spektrális vizsgálata
  - …

Leíró statisztika

Leíró statisztika

- Vizsgált adatok alapvető jellemzői
  - Kvantitatív
  - Erősen absztrahál, „összefoglal”

- Egyfajta ellentéte: következtető (*inferential*) stat.
  - Megfigyelt mintán túlmutató következtetések
  - Pl. populáció tulajdonságaira következtetés mintából

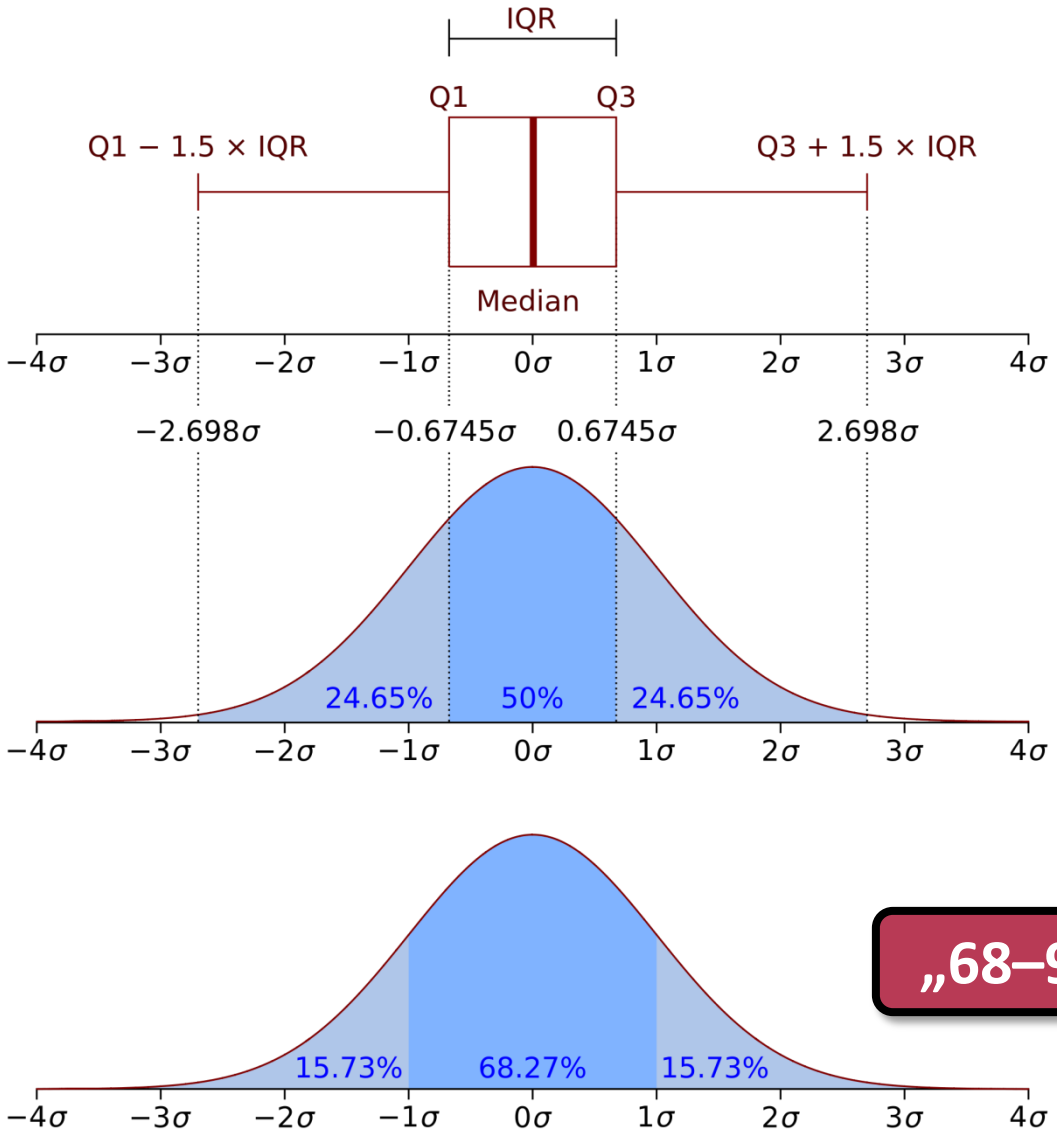
(Folytonos) megfigyelések jellemzése

- Átlag, medián, módusz

- Percentilis
  - Az n-edik percentilisnél az adatok $n\%$-a kisebb
- Kvartilis
  - Q1, Q3: 25. és 75. percentilis
  - Q2: medián

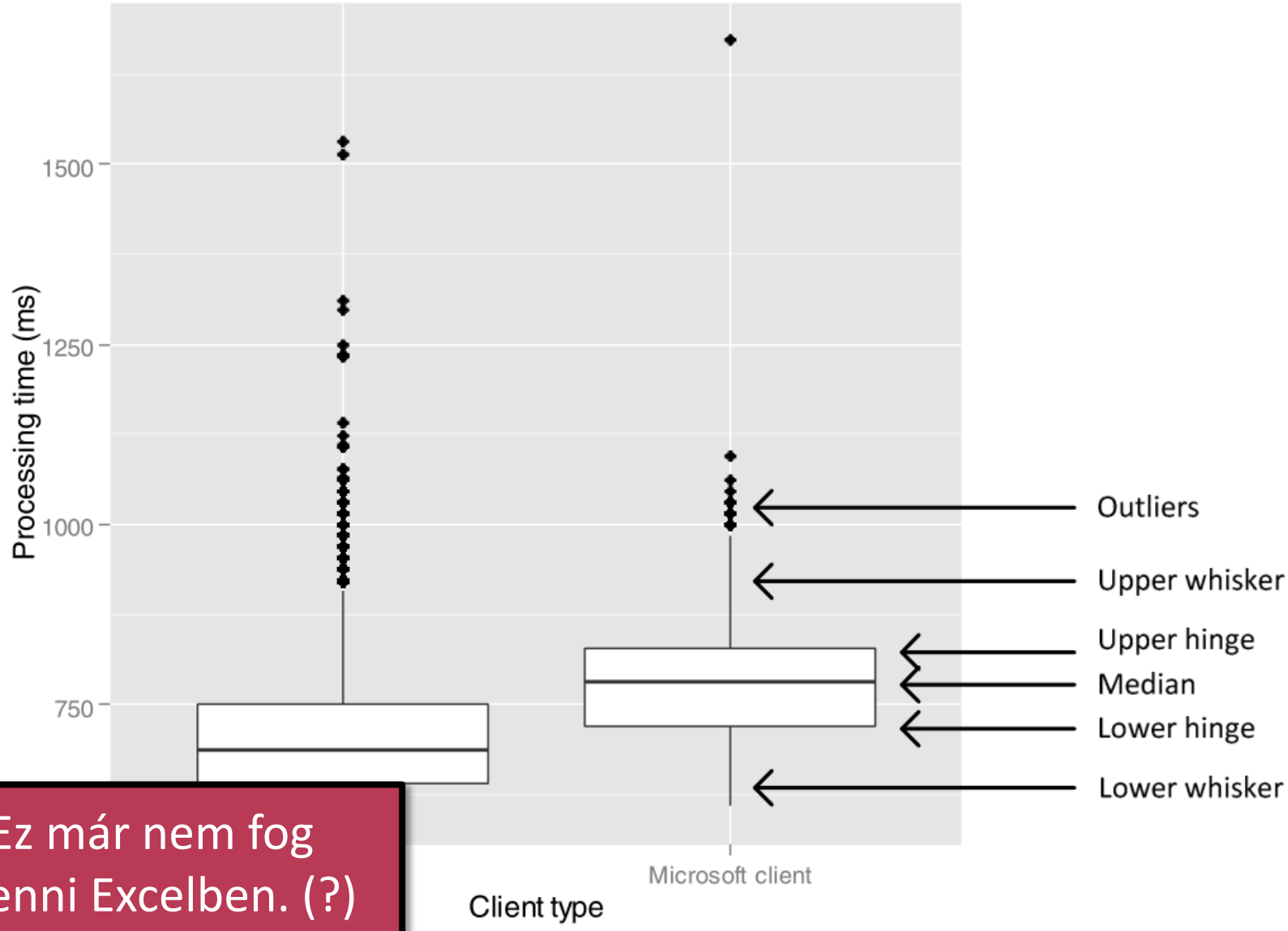
- Inter-quartile range (IQR)
  - Q3 – Q1

Kvartilisek szerepe



„68–95–99.7 rule”

Boxplot (Box and whisker plot)



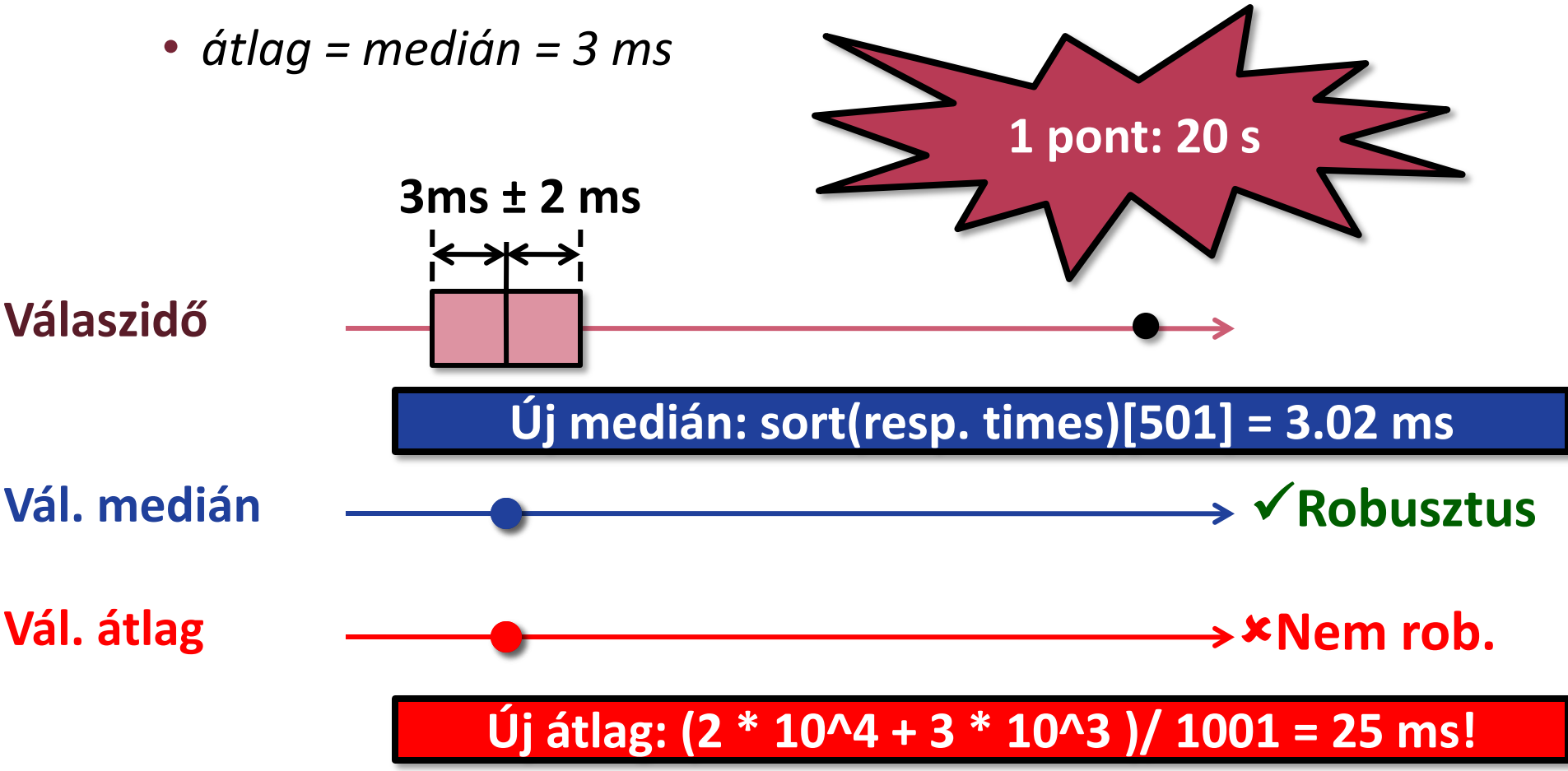
Ez már nem fog menni Excelben. (?)

Centrális tendencia és diszperzió

- Centrális jelleg jellemzői:
  - Átlag, medián, multimodalitás (illetve módus)
- „Diszperzió” jellemzői
  - Percentilisek, szórás(ok), variancia
- Melyik mennyire érzékeny a kiugró értékekre?
- Megj.: a mintaátlag vs. populáció-átlag jellegű kérdésekkel itt nem foglalkozunk
  - (Mi minek hogyan milyen becslője…)

Robusztus mérőszámok

- Alaphalmaz
  - 1000 pont ~ U(1, 5) egyenletes eloszlás
    - *átlag = medián = 3 ms*



Minta-variancia; minta kovariancia-mátrix

$$s^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

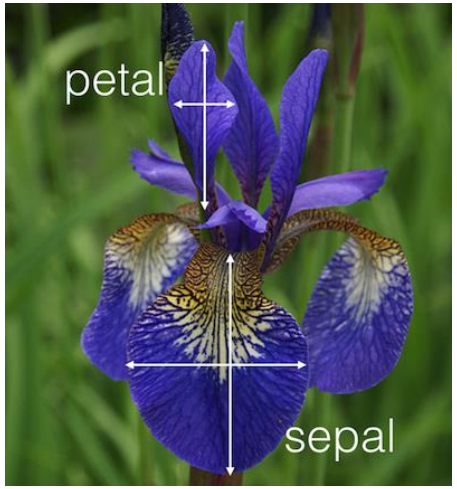
$$cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{x})(Y_i - \bar{y})$$

Breakdown point (becslőé): „rossz” megfigyelések max. aránya, ami után már tetszőlegesen rossz eredményt ad

Mennyire robosztusak?

Példa - felvezetés

- Fisher “Iris” adatkészlete
  - “The use of multiple measurements in taxonomic problems” (Fisher, 1936)
  - Cél: osztályozás folytonos jellemzők alapján
  - 50 minta, morfológiai jellemzők
  - 3 faj: setosa, versicolor, virginica



Példa - felvezetés

- A **csésze** (*kalyx*; virágképletbeli jele: **K**) a kétnemű virágtakarójú virágok külső takaróköre, a**csészelevelek** *(sepala)* összessége. A csésze a pártát övezi.

- A **párta** (*corolla*; virágképletbeli jele: **C**) a kétnemű virágtakarójú virágok belső takaróköre, a**sziromlevelek** *(petala)* összessége. A pártát a csésze övezi.

Variancia, kovariancia: példa

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
```

Variancia, kovariancia: példa

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> 
```

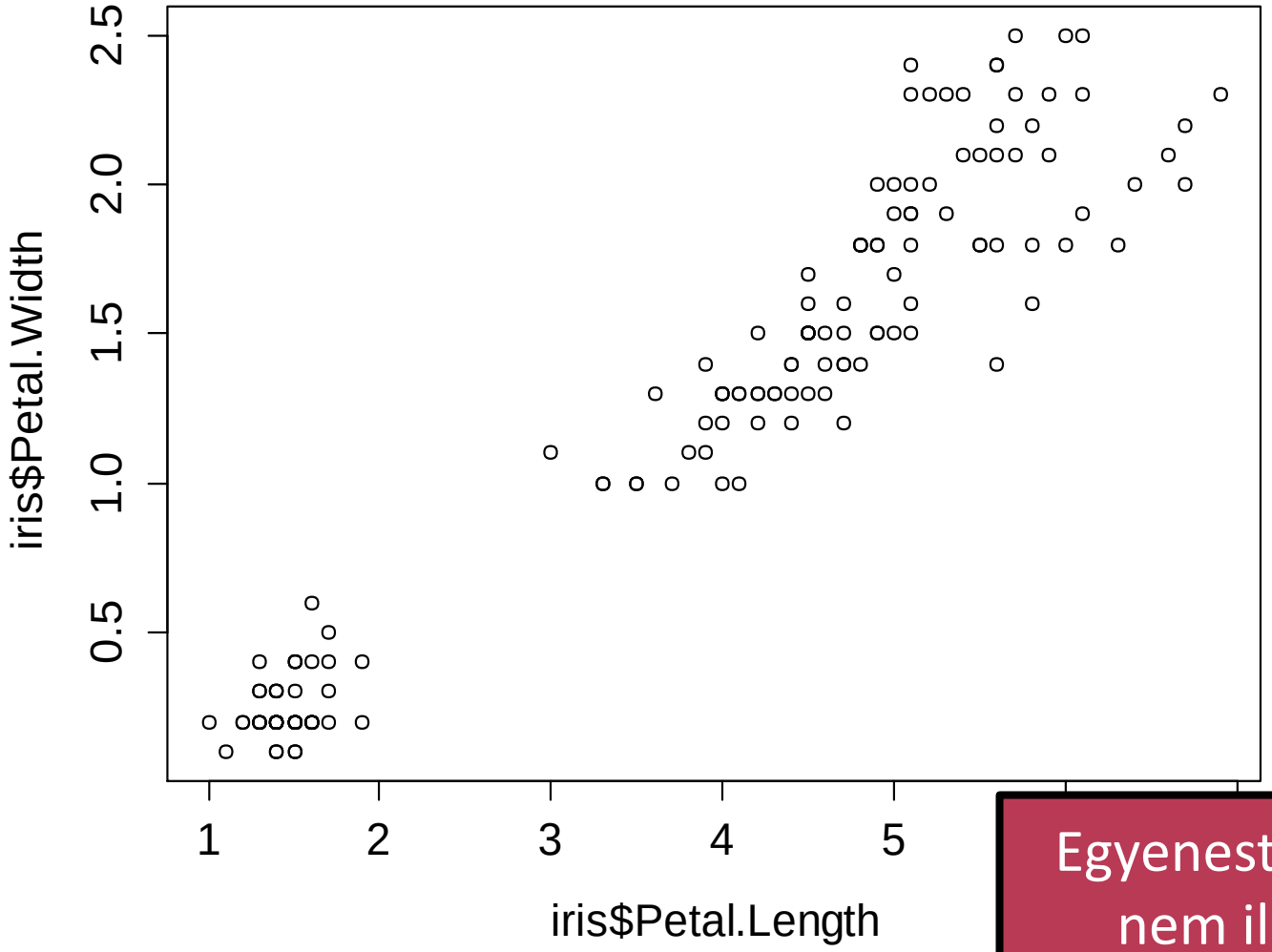
(Minta) variancia, kovariancia: példa

```
> cov(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

Normalizálás (szórások szorzatával): Pearson-féle lineáris korrelációs koefficiens

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> 
```

Lineáris korrelációs koefficiens



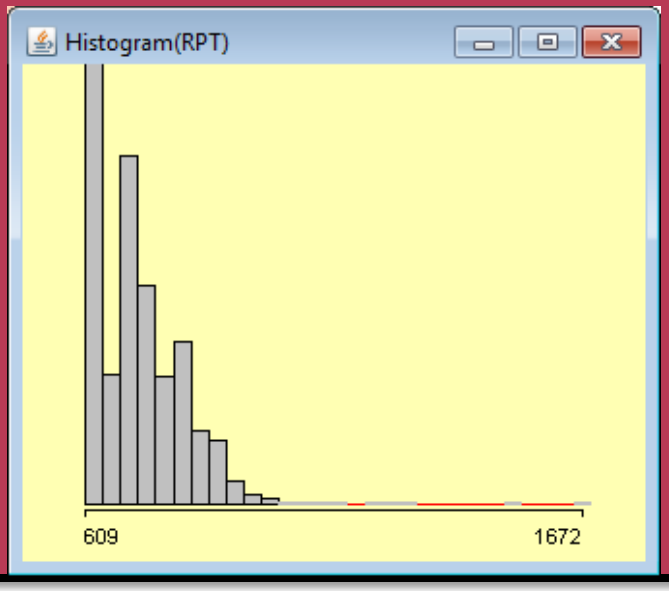
Egyenest most még nem illesztünk

Eloszlás jellemzése?

Változók

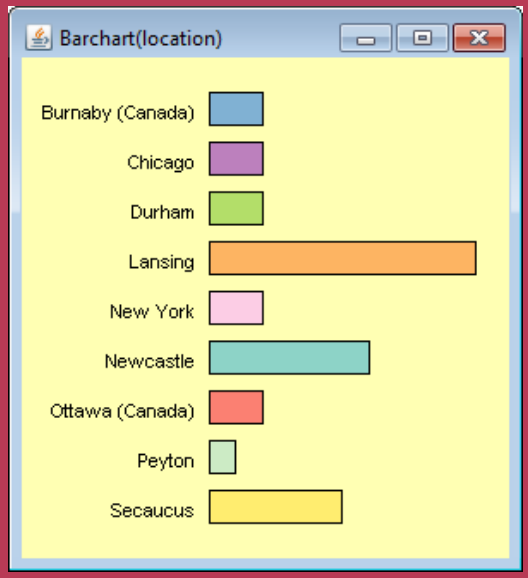
Numerikus

{RPT: 609, 613, 913, …}



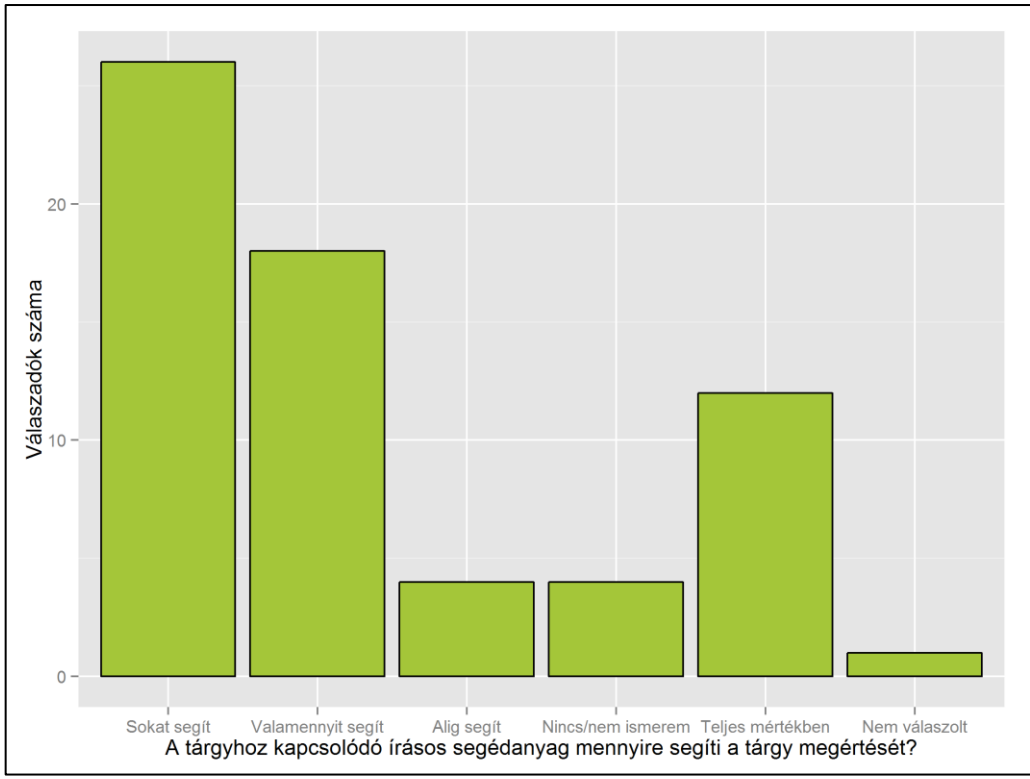
Kategorikus

{location: Peyton, Durham, …}



Oszlopdiagram (bar chart)

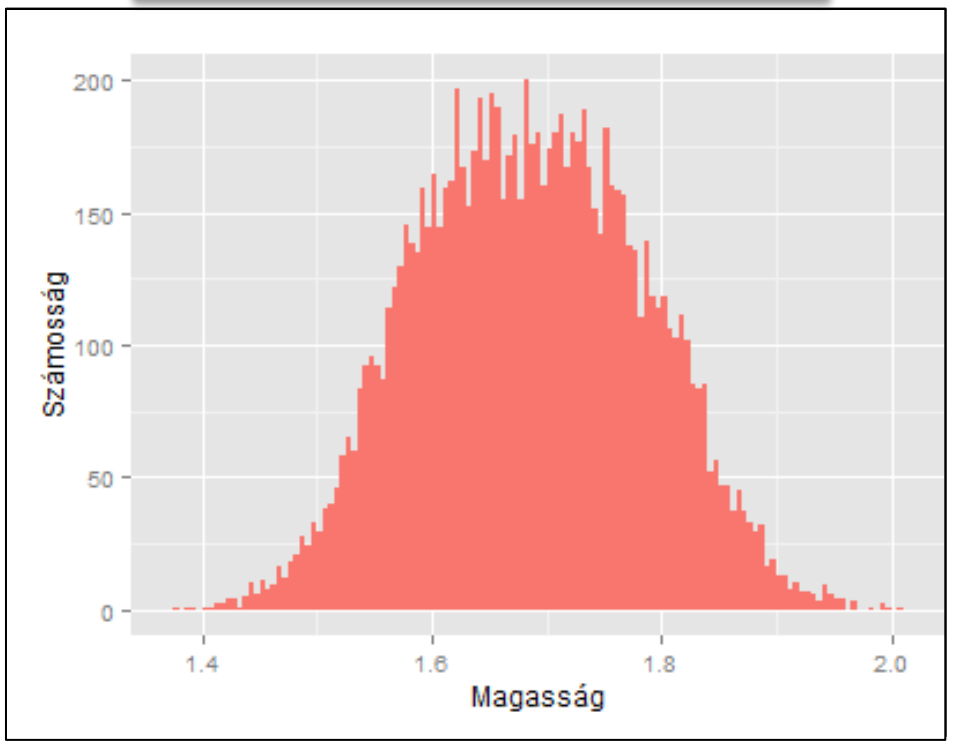
- Ábrázolt összefüggés:
  - Kategorikus változó egyes értékeinek abszolút gyakorisága
- Adategység:
  - Oszlop – magassága: adott érték gyakorisága
- Tervezői döntés:
  - Értékkészlet darabolása?



Hisztogram

- Ábrázolt összefüggés:
  - Folytonos változó egyes értékeinek abszolút gyakorisága
- Adategység:
  - Oszlop – magassága: adott érték gyakorisága
- Tervezői döntés:
  - Oszlopszélesség/kezdőpont?

Fontos percentilisek?

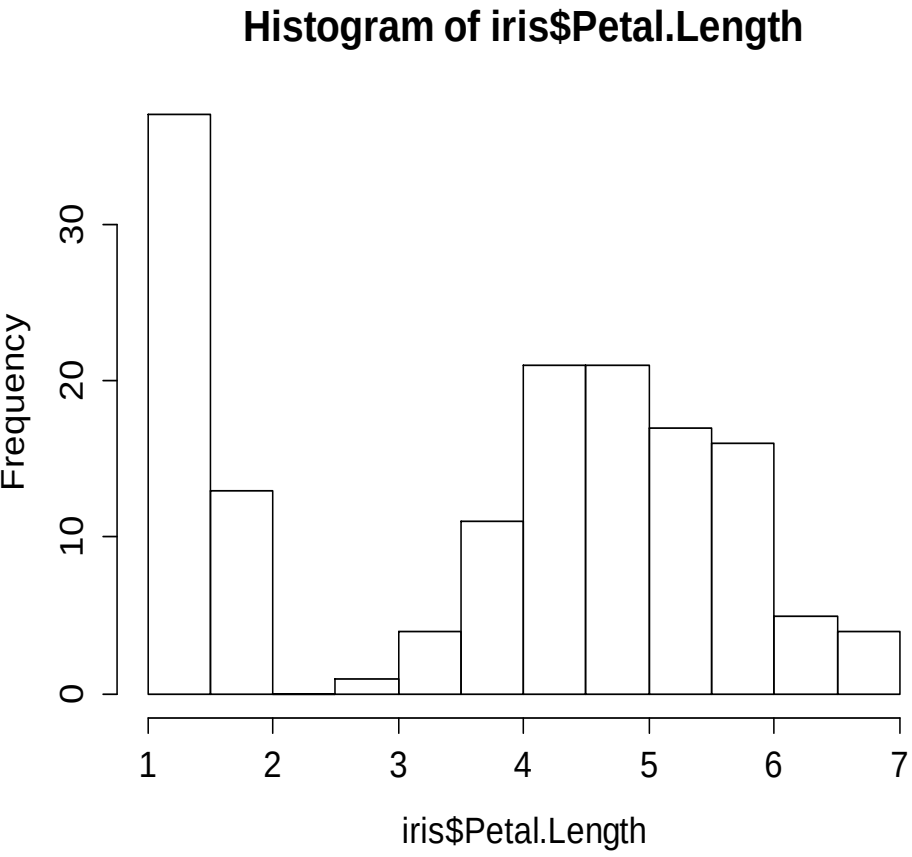


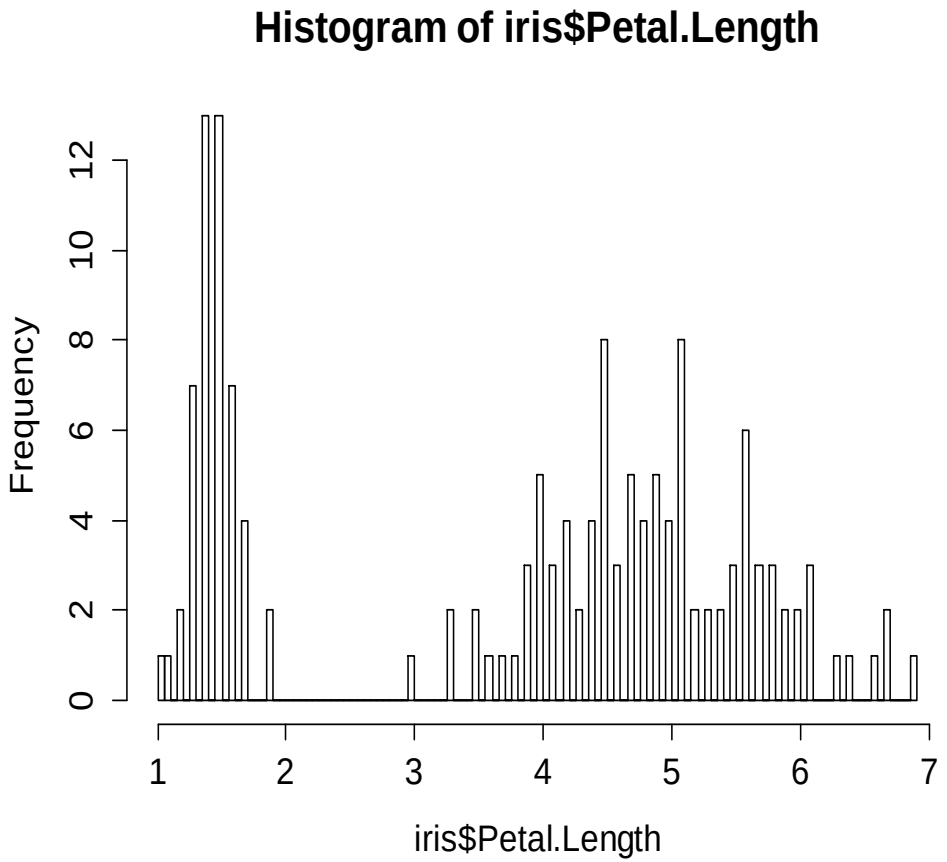
Problémák a hisztogrammal?

- ~~Általánosságban nem elfogulatlan~~
- ~~Akkor konzisztens, ha nem csökkentjük túl gyorsan a bin-méretet~~
- ~~(Ronda „zárt” alak)~~

- Érzékeny például az „origó” választására
- A „query value” a határon „ugrik”
- Az ismert algoritmusok ellenére a gyakorlatban jórészt manuálisan paraméterezzük
- Vagy „darabos”, vagy „nem folytonos”

Bin-szélesség hatása

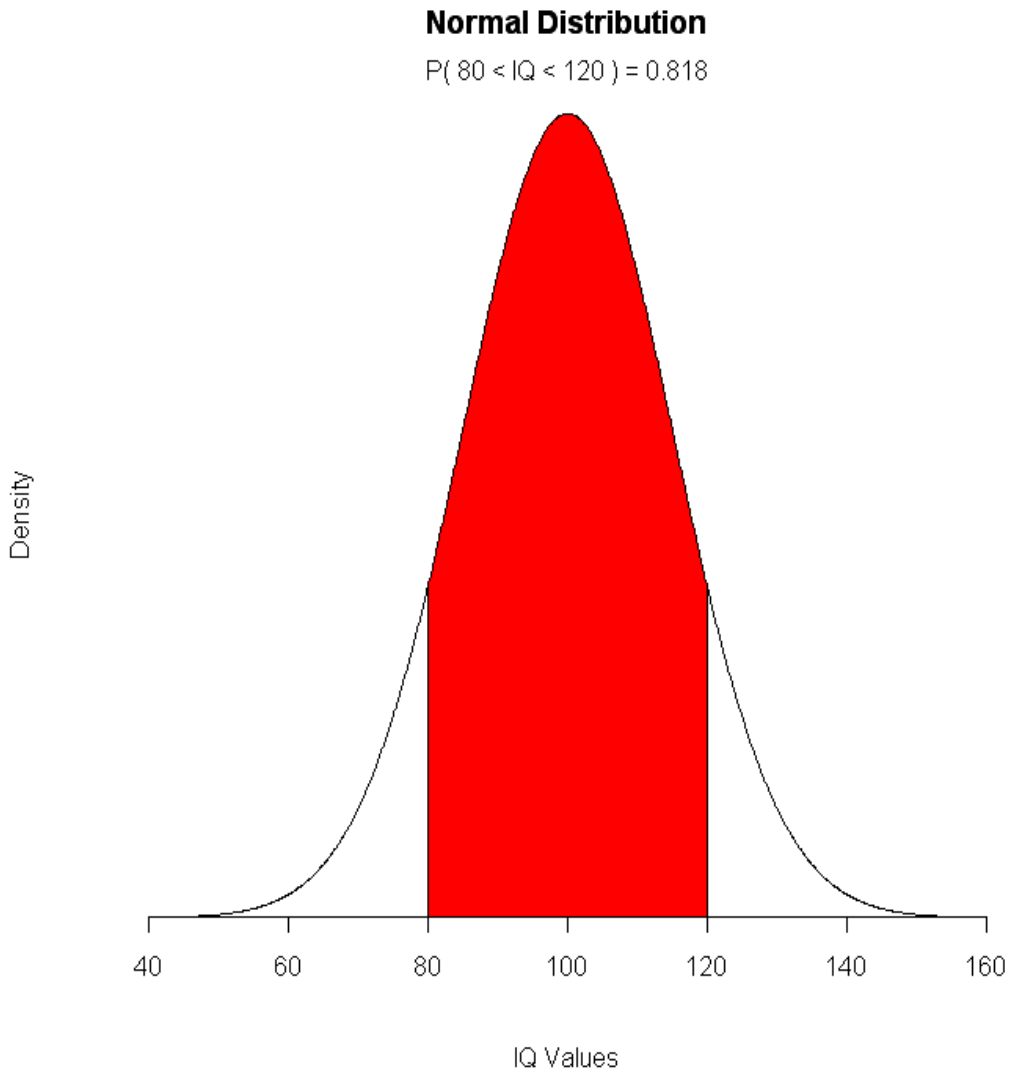




A többváltozós hisztogramokkal itt nem foglalkoztunk

Sűrűségfüggvény

- Ha mégis kevésbé akarunk absztrahálni
- Problémák
- 1. Biztos, hogy normál eloszlású a populáció?
- 2. Paramétereket kell becsülnünk a mintából



Nemparametrikus sűrűségbecslés

- Legyen X egy r komponensű val. vektorvált.
- Bármely p, amire

$$p(\boldsymbol{x}) \geq 0, \int_{\Re^r} p(\boldsymbol{x}) d\boldsymbol{x} = 1,$$

- Ún. „bona fide sűrűségbecslő” (*bona fide density est.*)
- NPDE: p parametrikus struktúra nélkül
  - Pl. elég nagy családba tartozik ahhoz, hogy esélytelen legyen véges paraméterkészlettel reprezentálni.
  - (Vagy csak nem akarunk ezzel foglalkozni…)

Nemparametrikus sűrűségbecslés

- Egy $\hat{p}$ becslő elfogulatlan (*unbiased*) p-re, ha minden $\boldsymbol{x} \in \Re^r$

$$E\{\hat{p}(\boldsymbol{x})\} = p(\boldsymbol{x})$$

- Véges adatkészleten nincs bona fide becslő, ami ezt minden folytonos sűrűségre teljesítené.
  - ***Aszimptotikus becslők vannak: mintaszámmal „egyre jobb" a megfelelés***

- Konzisztencia-kritériumok
  - MSE(**x**) → 0 minden **x**-re a mintaméret növelésével: „kvadratikus átlagban pontonként konzisztens becslő"
  - MSE: becslési hiba várhatóértékének négyzete

Kernel-módszerek

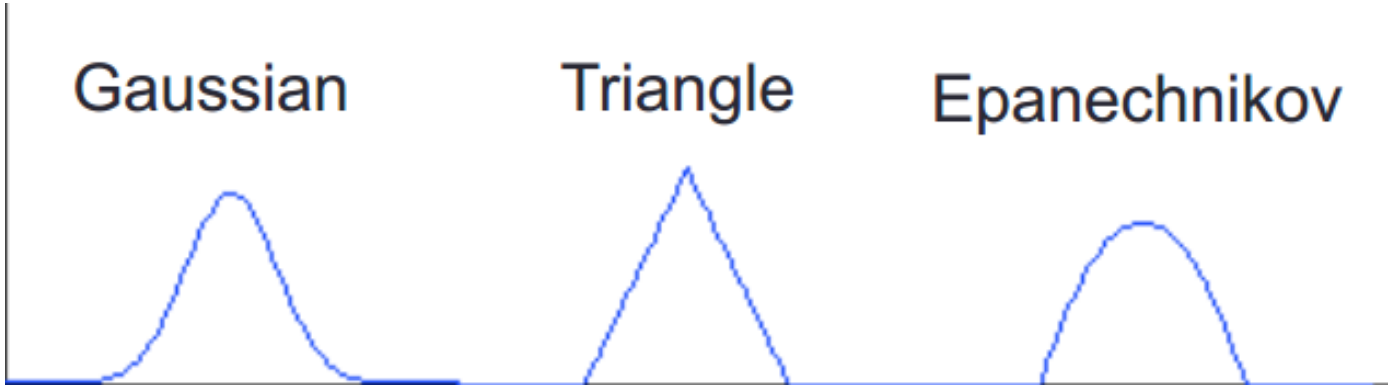
- Megpróbálunk „folytonos vonalat húzni”
- Legyen X_i egy ismeretlen f eloszlásból vett n elemű minta.
- Egy „*kernel density estimator*” függvény ezt közelíti:

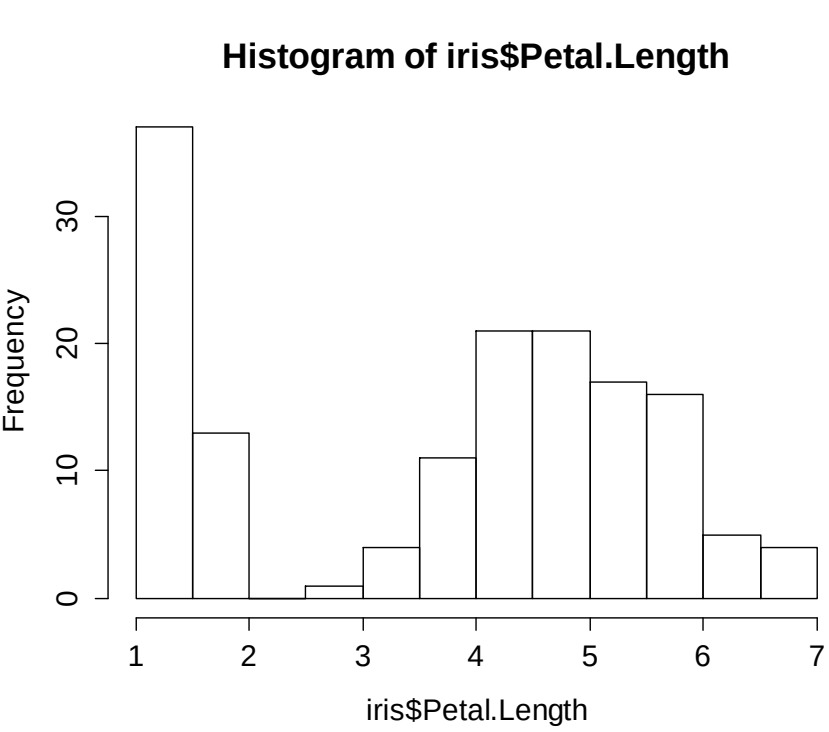
$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

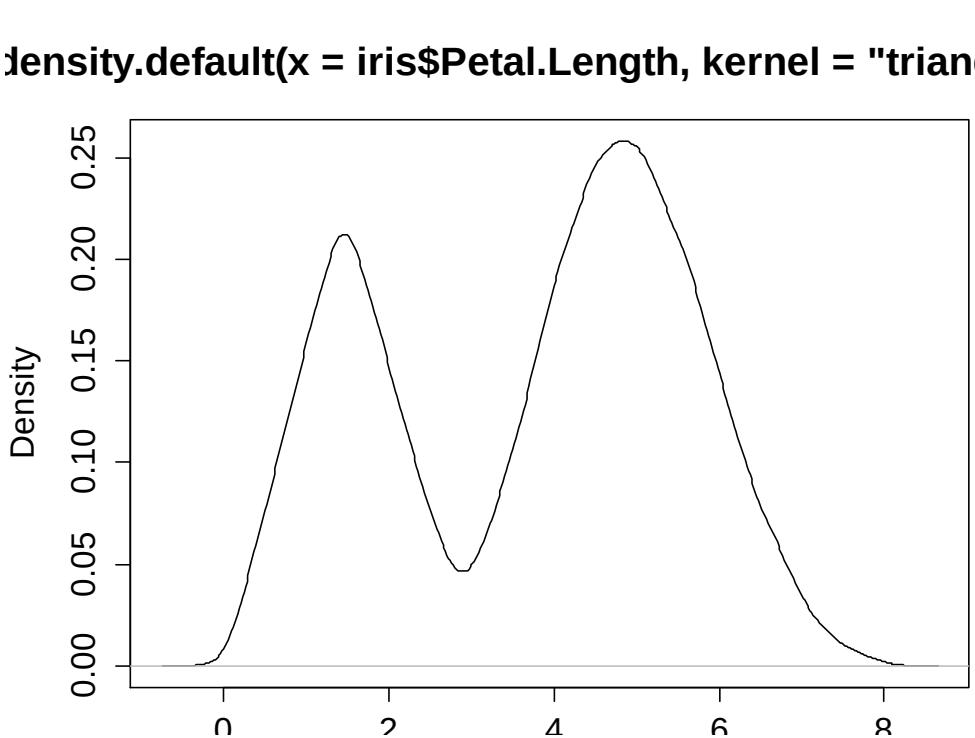
- h egy „ablakszélesség-paraméter”; K egy „magfüggvény”.

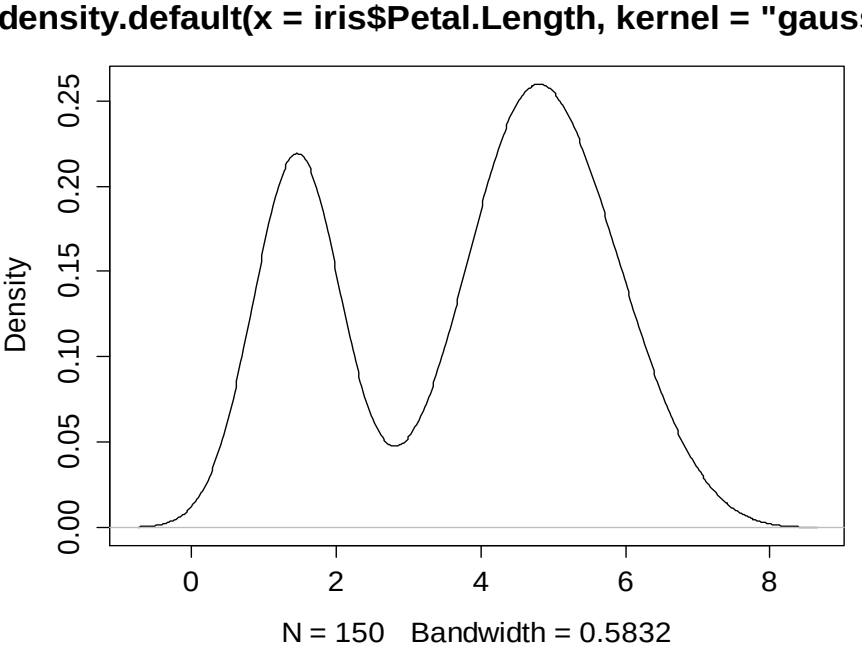
Magfüggvény-példák [4]

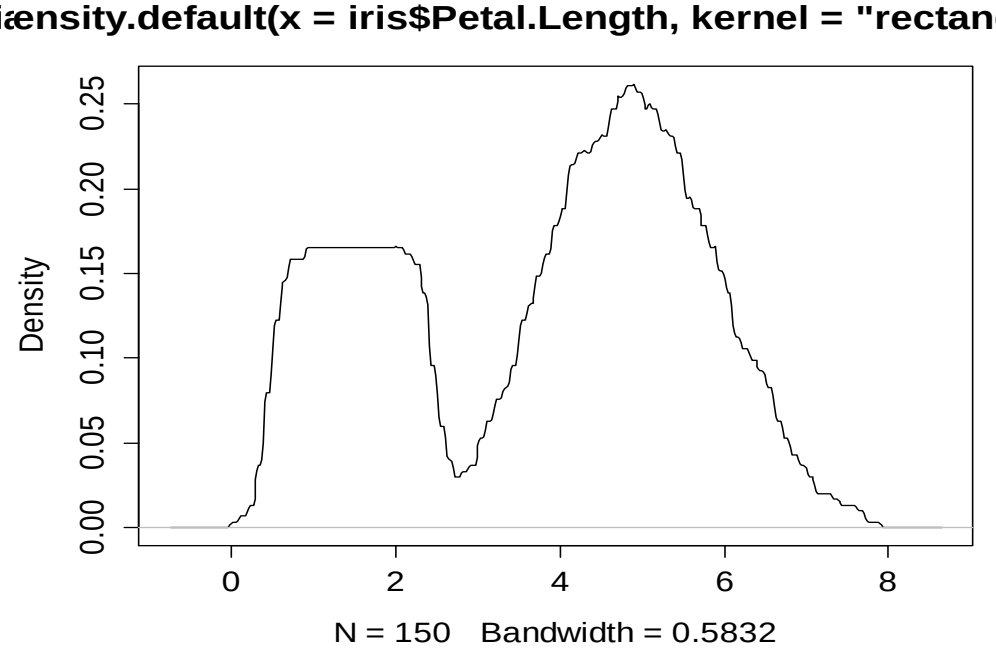
- Négyszög (*rectangular*): $\frac{1}{2} I_{[|x|\leq 1]}$
- Háromszög (*triangular*): $(1 - |x|) I_{[|x|\leq 1]}$
- Bartlett-Epanechnikov: $\frac{3}{4}(1 - x^2) I_{[|x|\leq 1]}$
- Nem korlátos bázisú Gauss (*Gaussian*): $(2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}, x \in \Re$




Histogram of iris$Petal.Length
Frequency
0
10
20
30
1
2
3
4
5
6
7
iris$Petal.Length

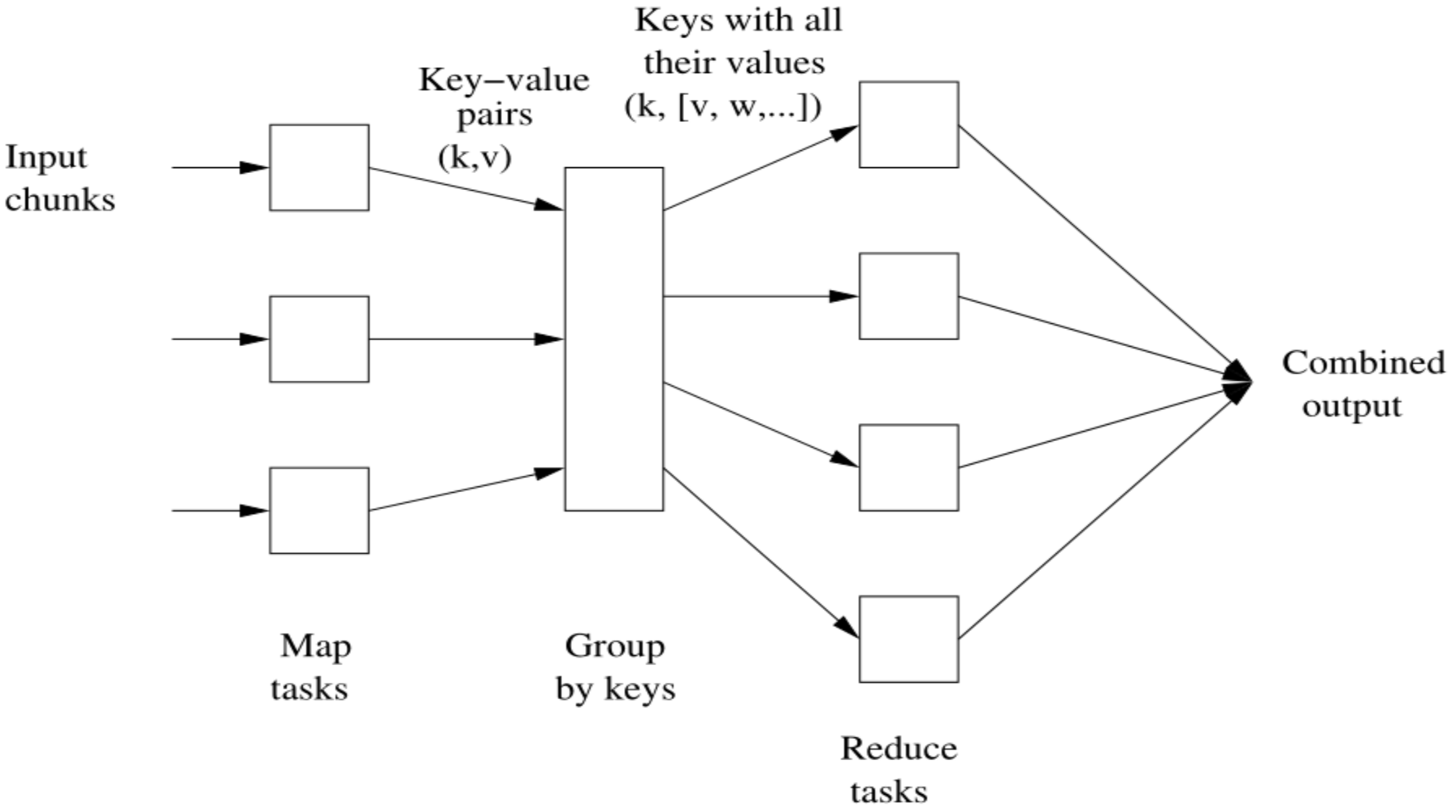

density.default(x = iris$Petal.Length, kernel = "triangul
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8


density.default(x = iris$Petal.Length, kernel = "gaussi
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
N = 150 Bandwidth = 0.5832


ensity.default(x = iris$Petal.Length, kernel = "rectangu
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
N = 150 Bandwidth = 0.5832

Big Data és leíró statisztika?

- A MapReduce programozási modellt láttuk. [5]



MapReduce és leíró statisztika?

- MIN/MAX/AVG…
  - Folytonos esetben?
  - Diszkrét esetben?
- Oszlopdiagram?
- **Hisztogram?**
- **Kernel sűrűség-közelítés nagy adatra?**
  - Tényleg nagy adatra drága „lekérdezni”: O(n) tag!
  - SIGMOD 2013: approximáció a minták csak egy mintáján számolással

Leíró statisztikák MapReduce becslése

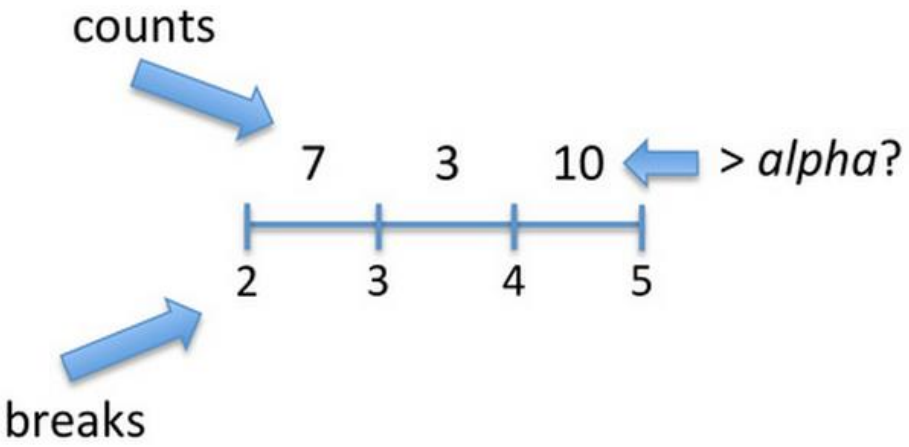
- Közelítő hisztogram újrahasznosítása
  - Kvantilisek becslése
  - Medián becslése
  - … De hogyan?
- Faktor/nominális változók: wordcount! ☺
- Variancia/szórás: pl. két menetben
  - Empirikus átlag kell hozzá
- Kovariancia, korreláció: két menetben, egy változó-párra egyszerű

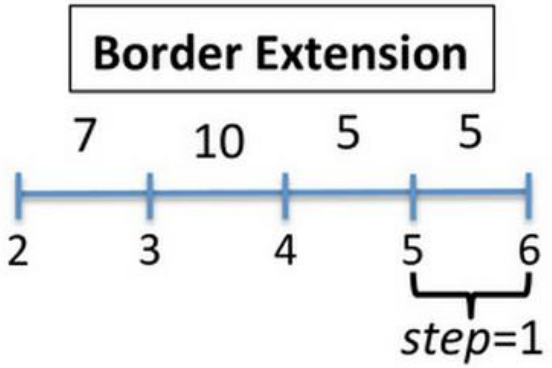
Hisztogram MapReduceban – nem ZH anyag

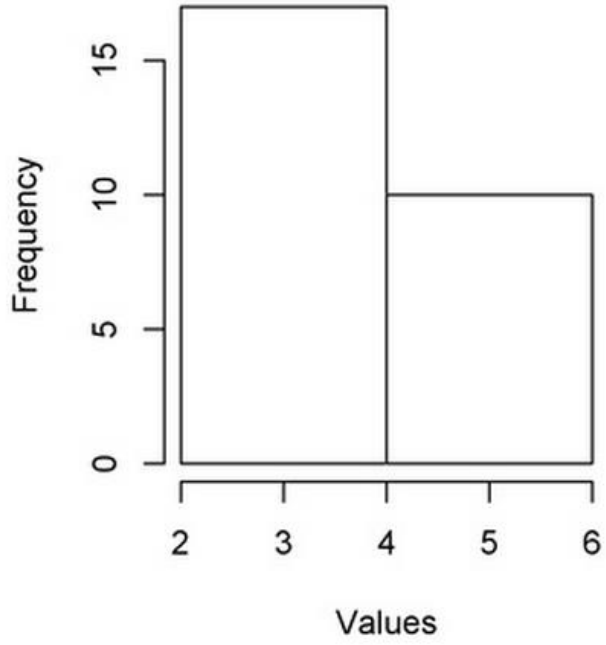
MapReduce és hisztogram (közelítés)

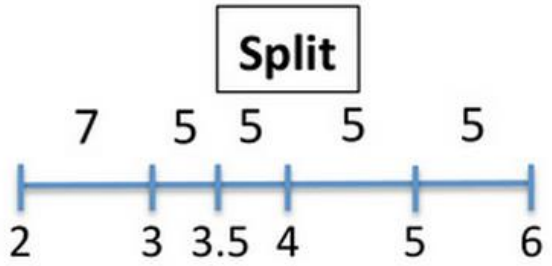
- Tfh. Nem feltételezhetjük az ún. *range partitioning*-et a vizsgált változóra
  - Pl. óriási CSV-t dolgozunk fel – Hadoop + HDFS
  - Különben nem lenne problémánk
- Partition Incremental Discretization (PiD) [4]
  - Módosítva [5]
- Layer1
  - Párhuzamosan több hisztogram építése
  - Azonos (igen kicsi) szélességű bin-ekkel kezdünk feltételezett intervallumon
  - Egy bin átlép egy *thresholdot*: *split*
  - N.B. adatfolyamra is működik
- Layer2: Layer1 hisztogramok összefűzése

Layer1 karbantartás [5]









Layer1 karbantartás [4]

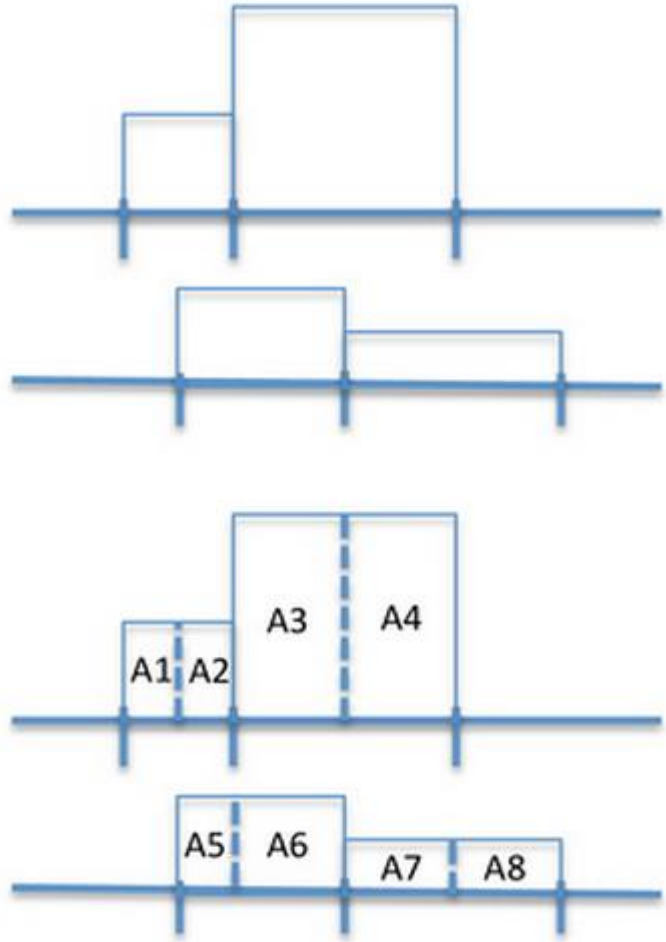
```
Update-Layer1(x, breaks, counts, NrB, alfa, Nr)
x - Observed value of the random variable
breaks - Vector of actual set of break points
counts - Vector of actual set of frequency counts
NrB    - Actual number of breaks
alfa   - Threshold for Split an interval
Nr     - Number of observed values

 If (x < breaks[1]) k = 1; Min.x = x
 Else If (x > breaks[NrB] k = NrB; Max.x = x
 Else k = 2 + integer((x - breaks[1]) / step)

 while(x < breaks[k-1]) k <- k - 1
 while(x > breaks[k]) k <- k + 1
```

Layer1 karbantartás [4]

```
counts[k] = 1 + counts[k]
Nr = 1 + Nr
If ((1+counts[k])/(Nr+2) > alfa)  {
      val = counts[k] / 2
      counts[k] = val
      if (k == 1) {
        breaks = append(breaks[1]-step, breaks)
        counts <- append(val,counts)
      }
      else {
        if(k == NrB) {
          breaks <- append(breaks, breaks[NrB]+step)
          counts <- append(counts,val)
        }
        else {
          breaks <- Insert((breaks[k]+ breaks[k+1])/2, breaks, k)
          counts <- Insert(val, counts, k)
        }
      }
  NrB = NrB + 1
}
```

Összefűzés - hibaforrások

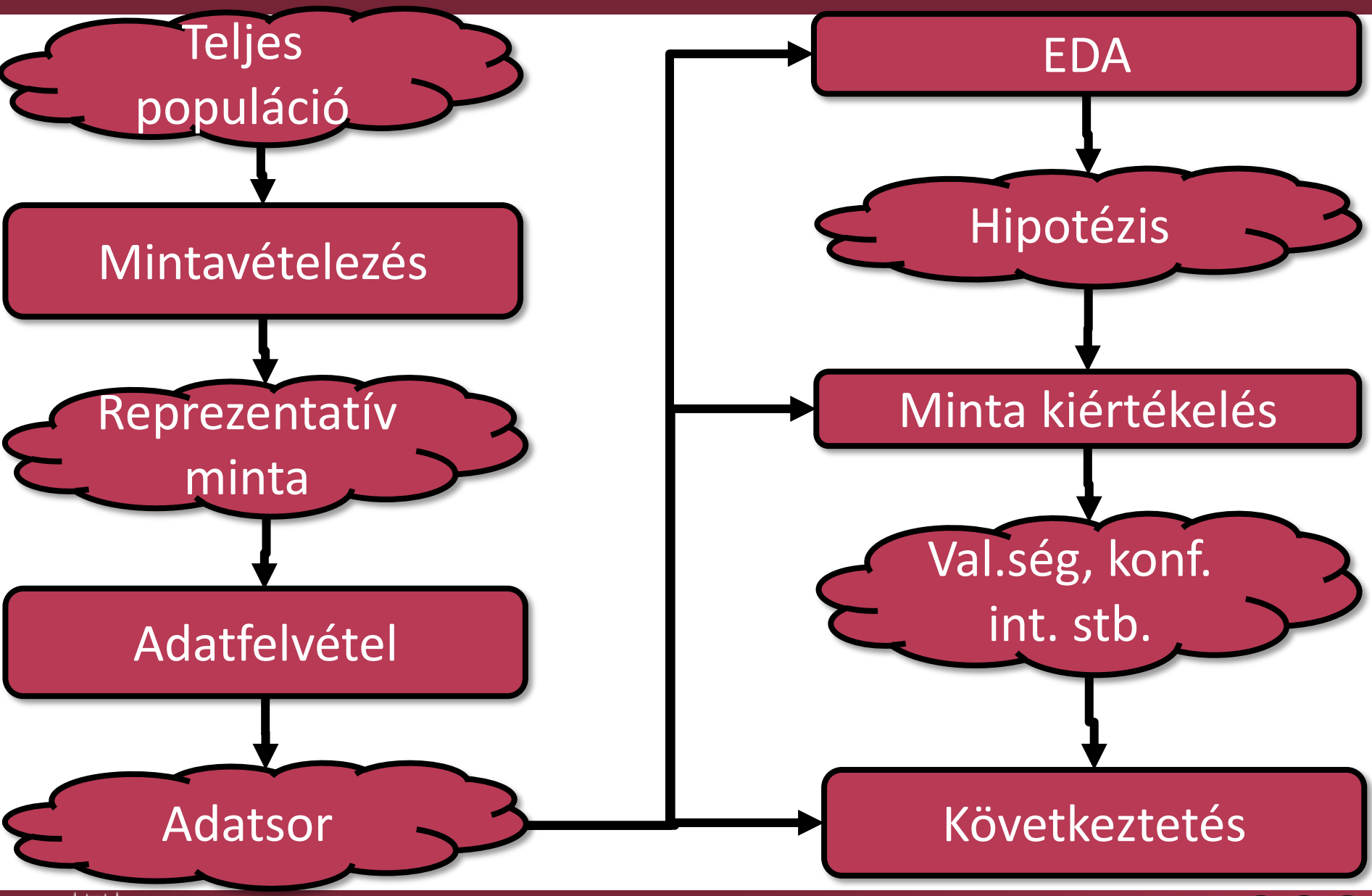
- Csak a Layer1 töréspontjai
- Split → pontatlan számlálók
- + „split” az összefűzés során



Következtető statisztika

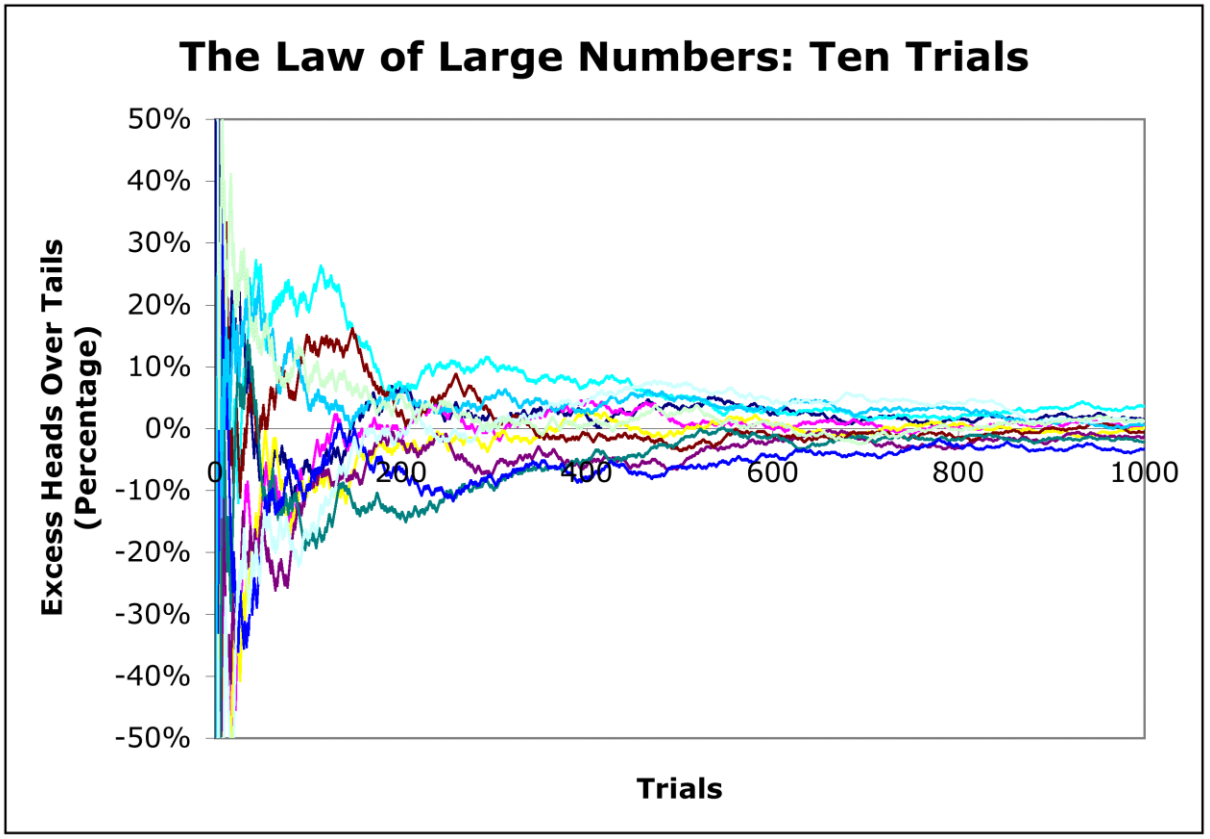
Következtető statisztika

Következtető statisztika



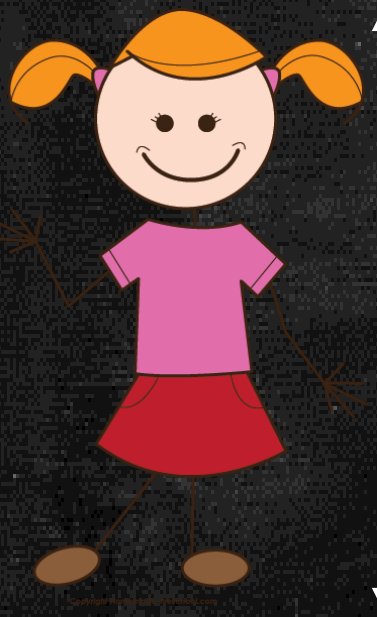
Ökölszabályok

- LLN (Law of Large Numbers)
  - Ha a kísérletek száma tart a végtelenhez, az előfordulási gyakoriság az elméleti valószínűséghez konvergál



Ökölszabályok

- CLT (Central Limit Theorem)
  - A minták statisztikáinak átlaga normális eloszlást követ (bizonyos feltételek mellett).
  - $\bar{x} \sim N\left(mean = \mu, SE = \frac{s}{\sqrt{n}}\right)$
    - $\bar{x}$ a mintaátlag
    - μ a populáció várható értéke
    - s a populáció (empirikus) szórása
    - n a mintaméret



?

Magyarországi kamaszlányok

Békés

Heves

Vas

$\overline{x}_{Békés}$

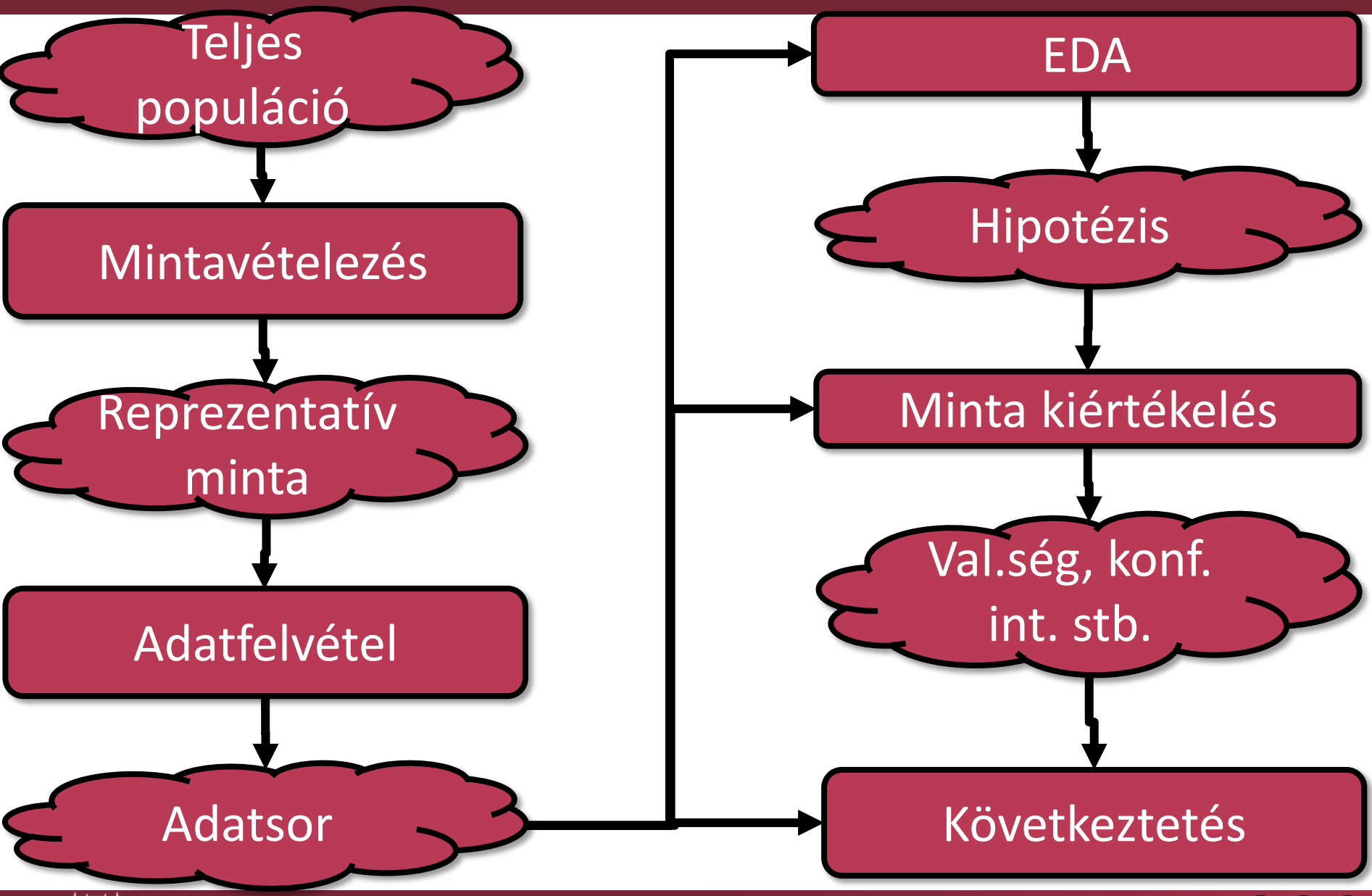
$\overline{x}_{Heves}$

$\overline{x}_{Vas}$

$$\mu = \frac{X_1 + X_2 + \ \ldots + X_N}{N}$$

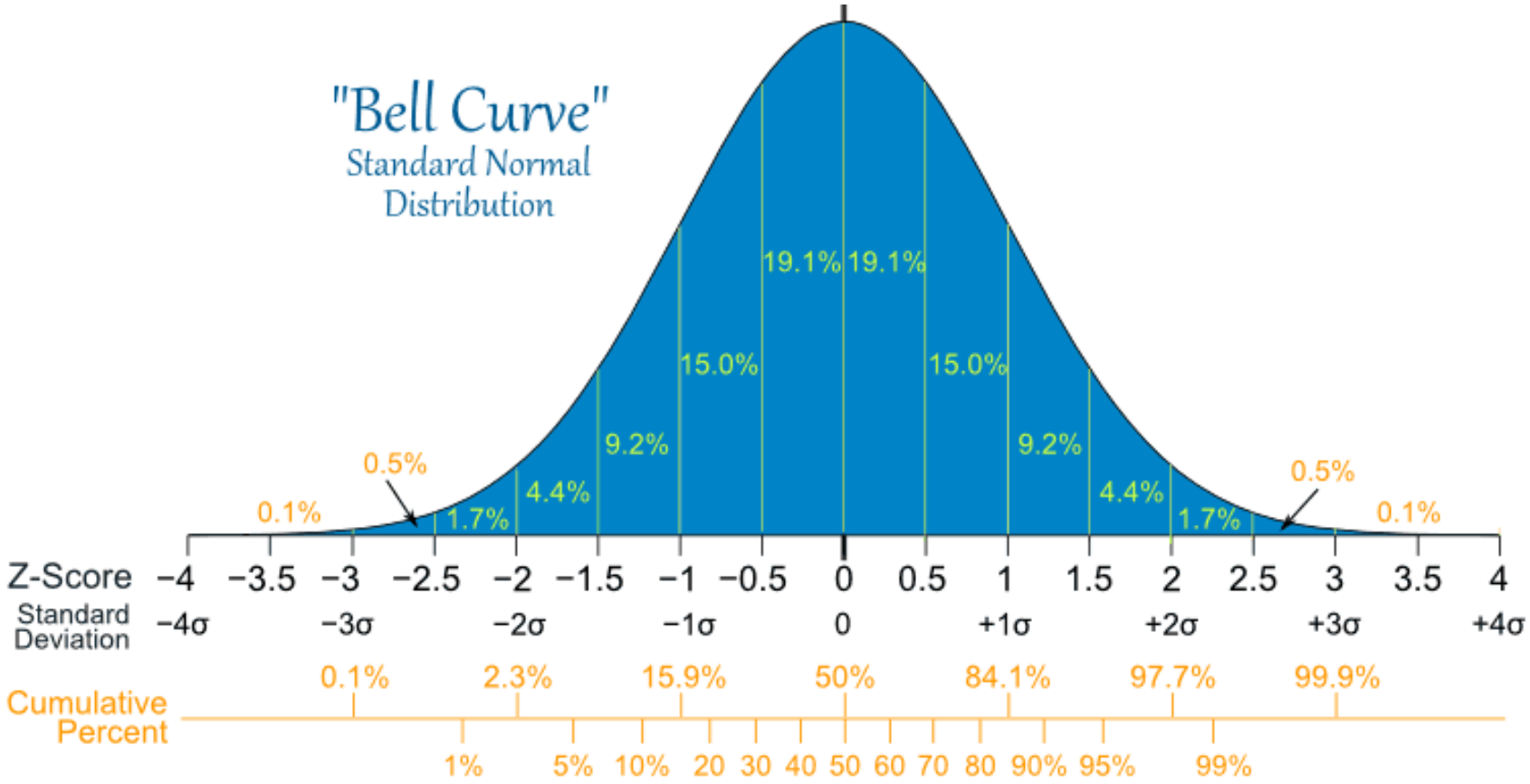
$$\mu \approx mean(\overline{x})$$

Következtető statisztika



Minta kiértékelés

- EDA ~ nyomozás
- Kiértékelés ~ a per maga

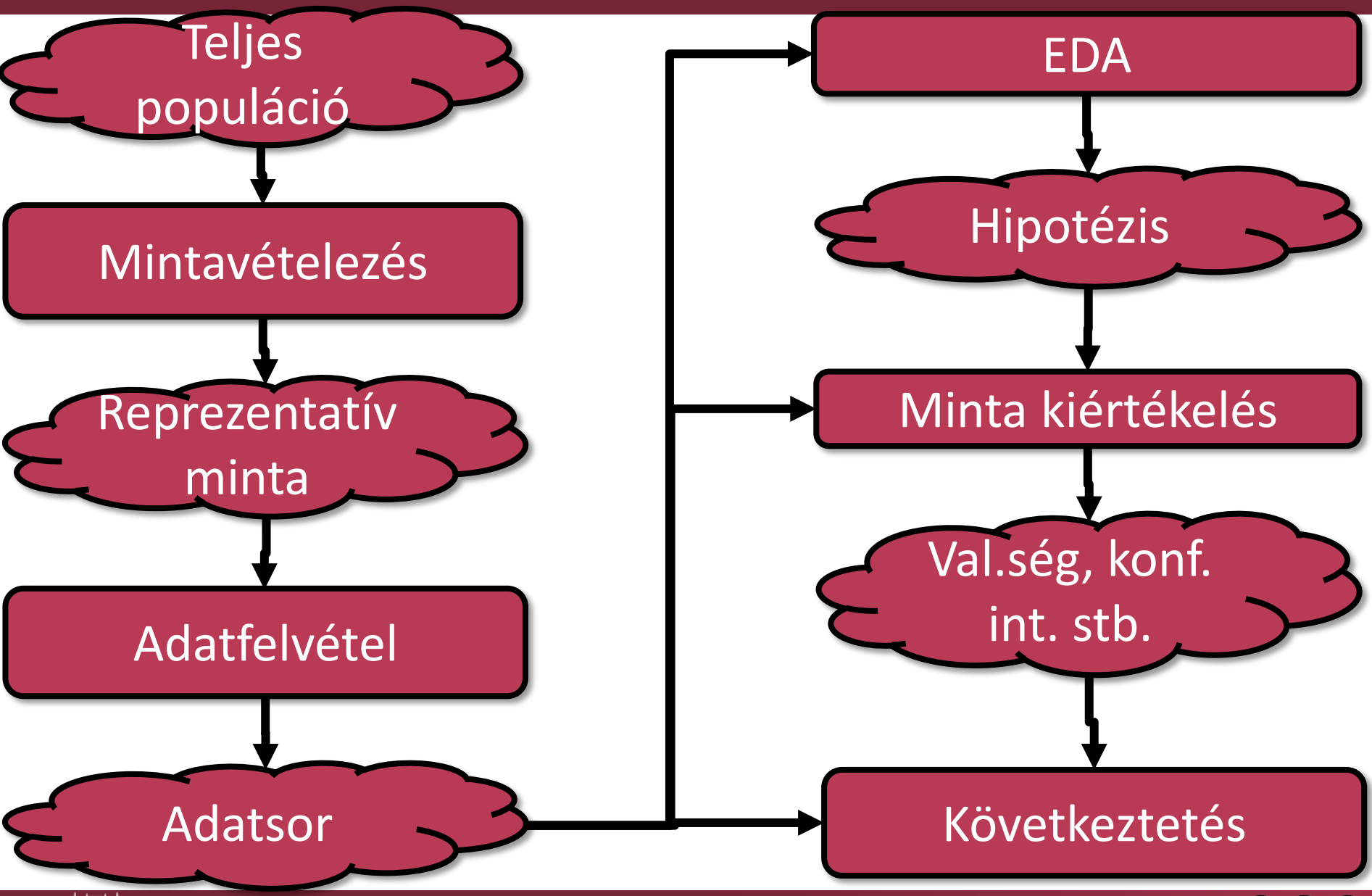


Mit tesztelünk tipikusan?

- Parametrikus tesztek
  - Egy minta eloszlás egy paraméterét próbáljuk kitalálni
  - Két minta eloszlásának a paramétere megegyezik-e?

- Nemparametrikus tesztek
  - Illeszkedésvizsgálat → adott eloszlású-e egy minta?
  - Függetlenségi vizsgálat → független-e két minta?
  - Homogenitásvizsgálat → két minta eloszlása megegyezik-e?

Következtető statisztika



Következtetés

- Döntési bemenet
  - Valami küszöbérték
- Adatsor típusa
  - Megfigyelési tanulmány (observational study)
  - Kísérlet (experiment)

Különbség: a *köztes változók* eliminálása

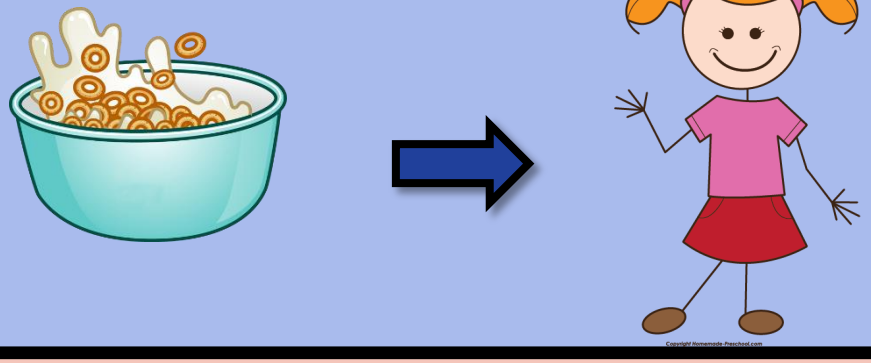
Esettanulmány

„Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast.„

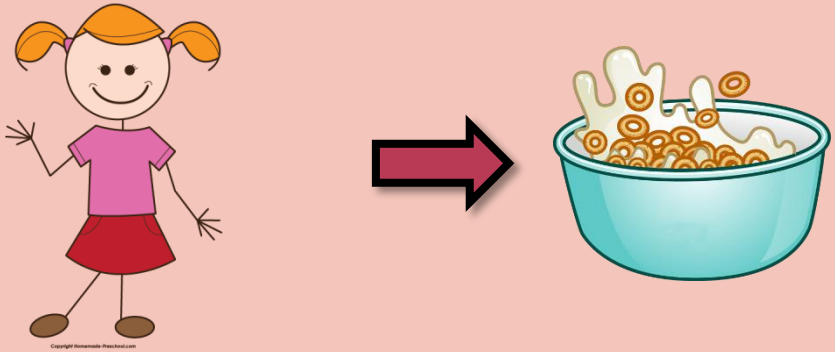
Forrás: http://usatoday30.usatoday.com/news/health/2005-09-08-cereal-slimming_x.htm

Esettanulmány

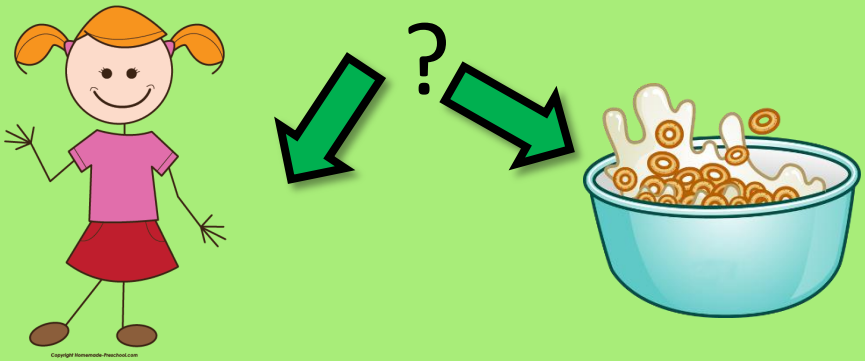
1. „Breakfast, cereal keep girls slim”



2. „Being slim causes girls to eat breakfast„



3. „A confounding variable is responsible for both”



Forrás: http://usatoday30.usatoday.com/news/health/2005-09-08-cereal-slimming_x.htm

Következtetés

- Döntési bemenet
  - Valami küszöbérték
- Adatsor típusa
  - Megfigyelési tanulmány (observational study)
    - A köztes változók kiléte bizonytalan
    - Csak korreláció, kauzális következtetések nem
  - Kísérlet (experiment)
    - A köztes változókat kiszűrtük (mintavételezés!)
    - Kauzális következtetések is

Adatelemzési módszerek

Adatbányászati építőkövek

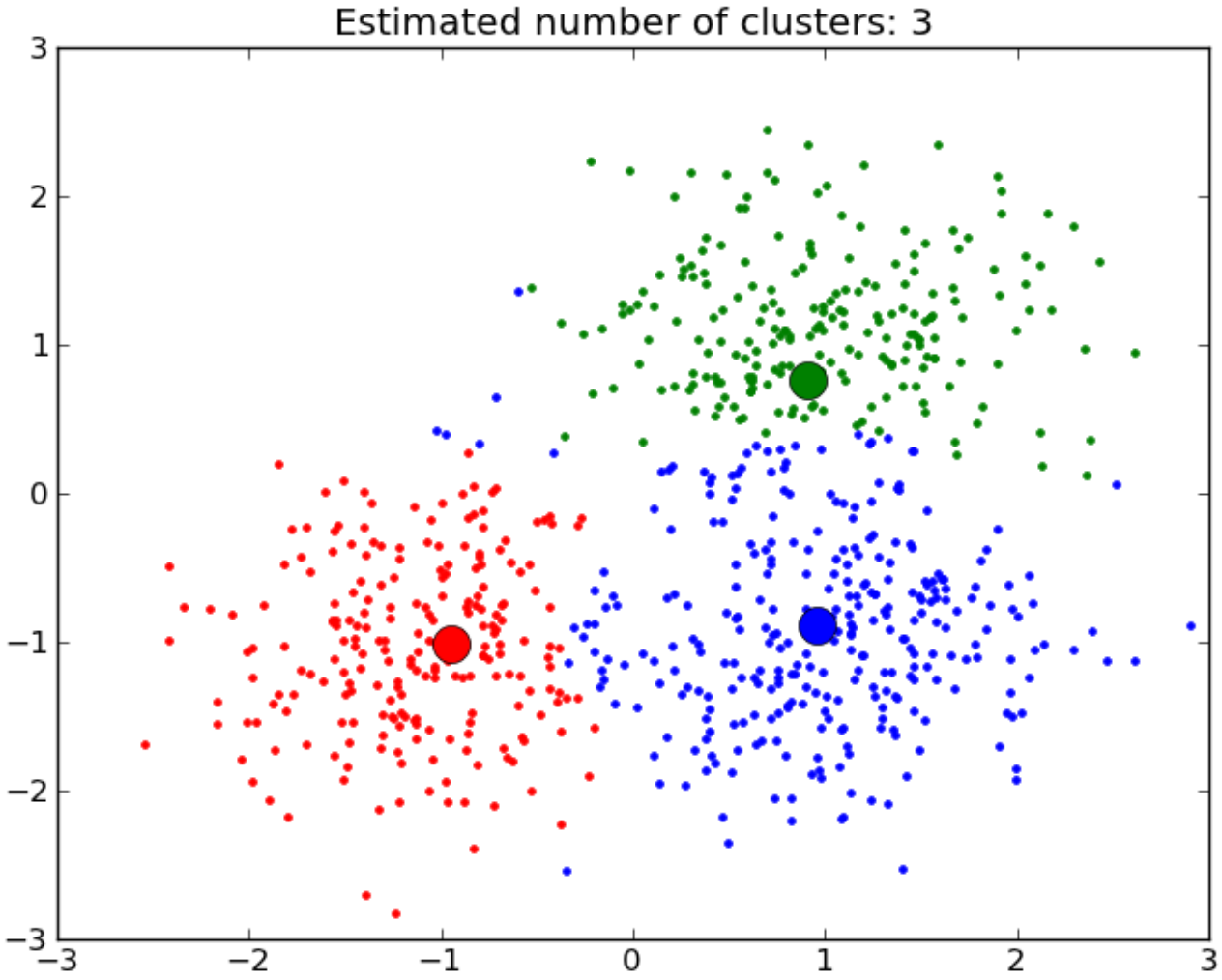
Klaszterezés

Osztályozás

Asszociációs szabályok

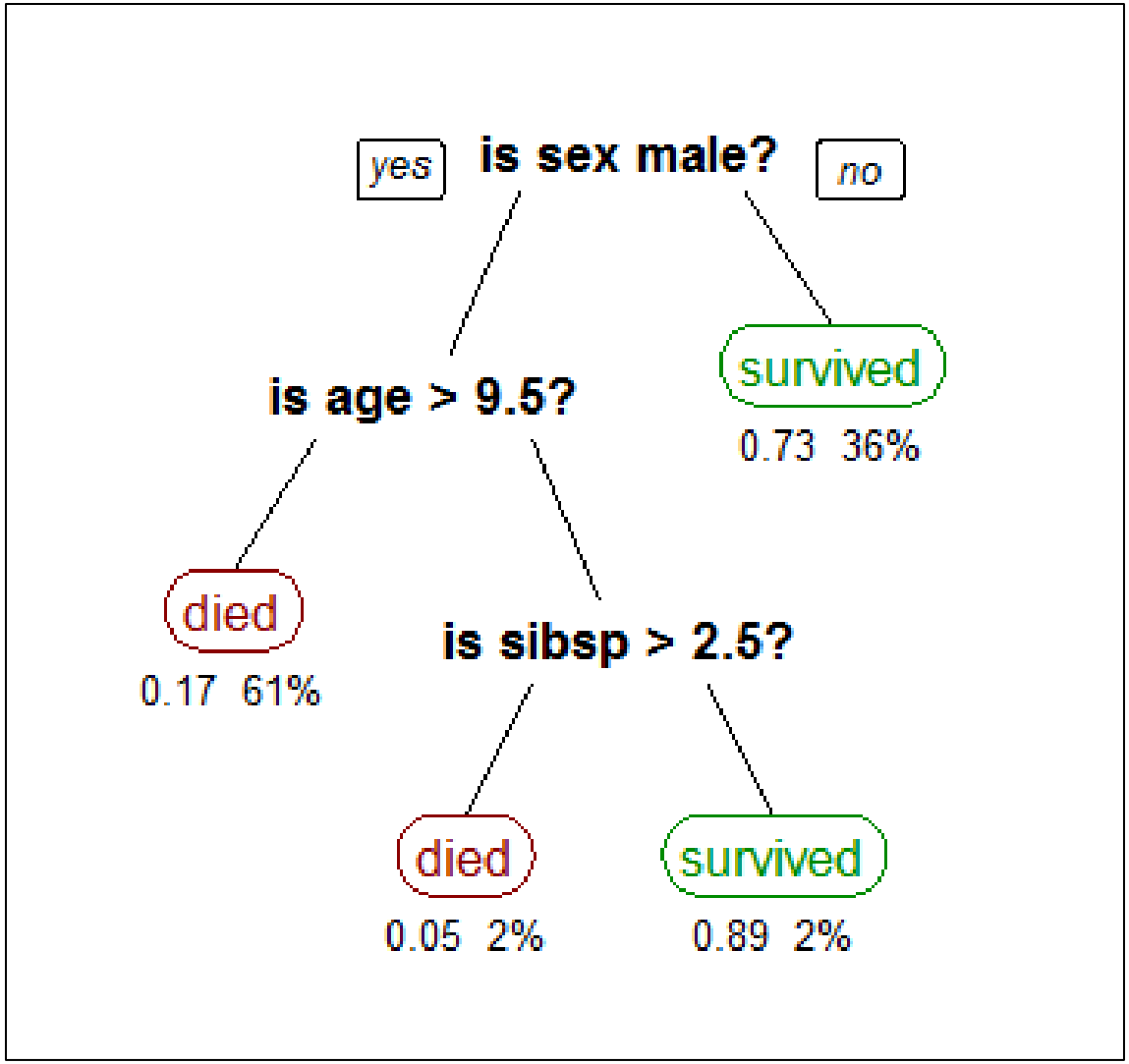
Regresszió

Klaszterezés

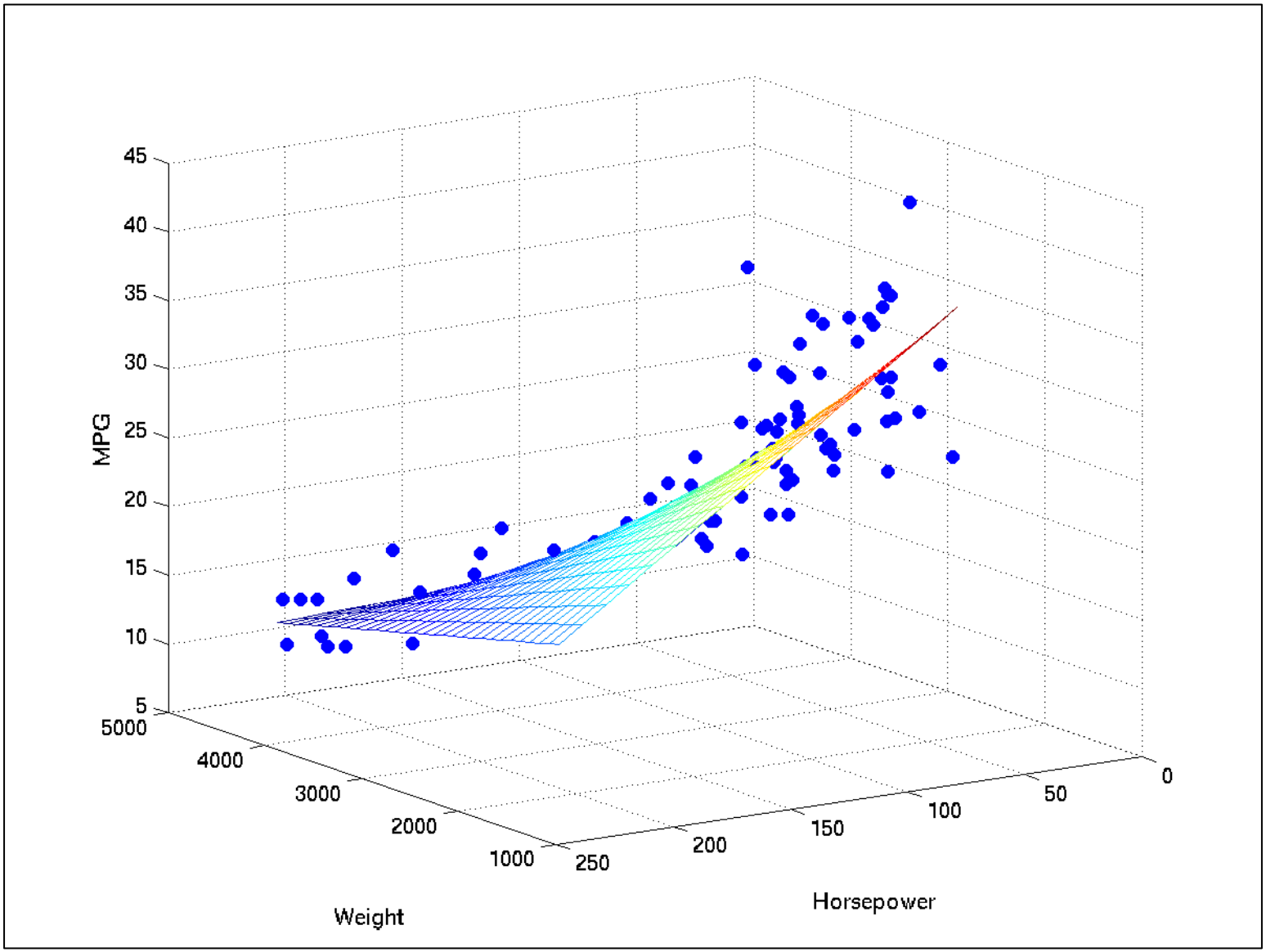


Asszociációs szabályok

Osztályozás



Regresszió


45
40
35
30
25
20
15
10
5
MPG
5000
4000
3000
2000
1000
Weight
250
200
150
100
50
0
Horsepower

Források

- [1] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. New York, NY: Springer New York. doi:10.1007/978-0-387-78189-1
- [2] Zheng, Y., Jestes, J., Phillips, J. M., & Li, F. (2013). Quality and efficiency for kernel density estimates in large data. In *Proceedings of the 2013 international conference on Management of data - SIGMOD '13* (p. 433). New York, New York, USA: ACM Press. doi:10.1145/2463676.2465319
- [3] Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139058452
- [4] Gama, J., & Pinto, C. (2006). Discretization from data streams. In *Proceedings of the 2006 ACM symposium on Applied computing - SAC '06* (p. 662). New York, New York, USA: ACM Press. doi:10.1145/1141277.1141429
- [5] http://www.slideshare.net/Hadoop_Summit/creating-histograms-from-data-stream-via-map-reduce