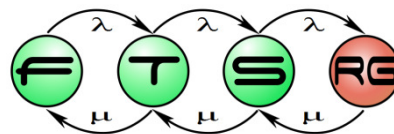


# Gépi tanulási (Machine Learning) módszerek alkalmazása

Hullám Gábor (hullam.gabor@ )  
Salánki Ágnes, Kocsis Imre  
salanki@, ikocsis@

2016.11.03.

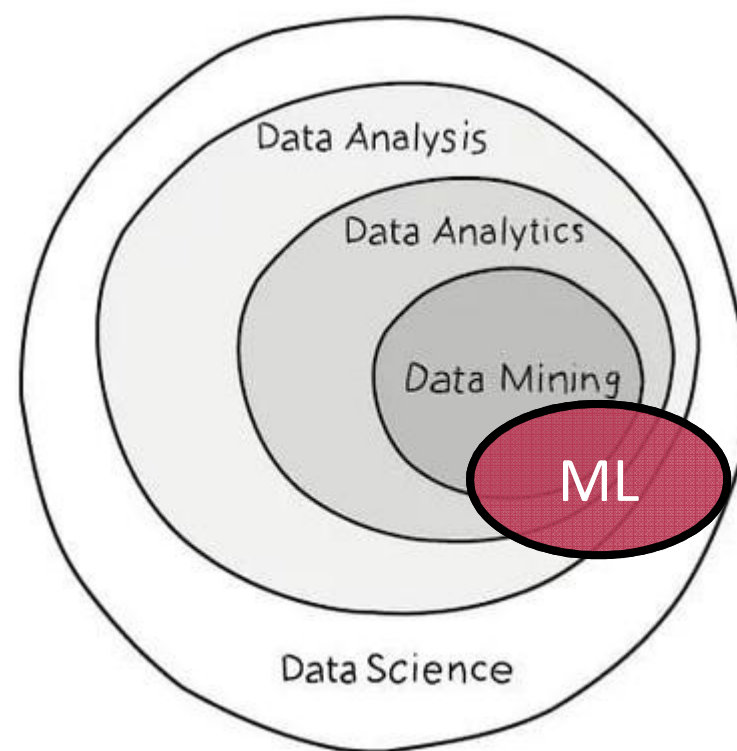


# Gépi tanulás

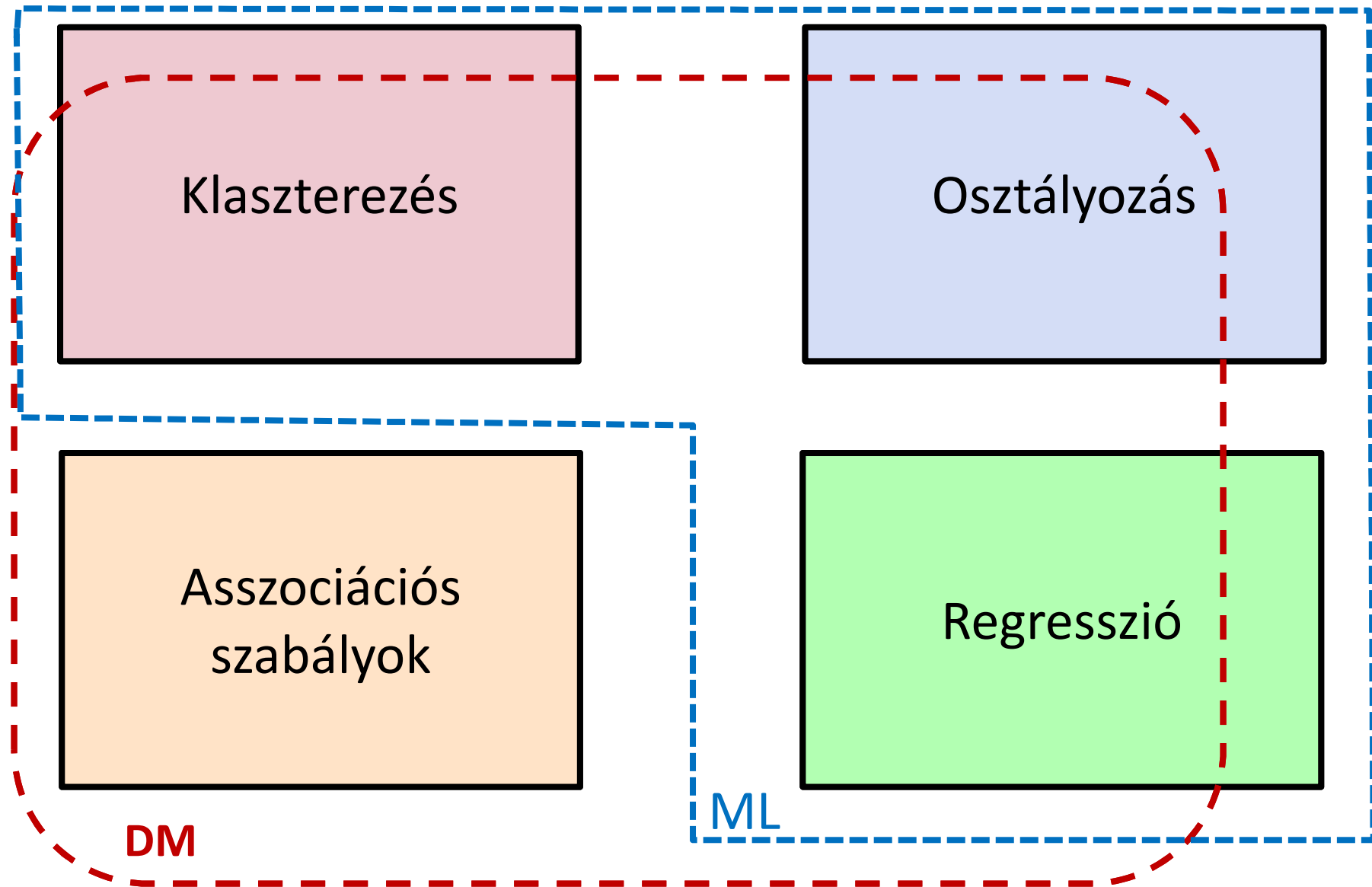
- **Gépi tanulás** (Machine Learning - ML)
  - ~ Olyan módszerek és algoritmusok, melyek a rendelkezésre álló adathalmazból egy **modellt tanulnak**, és ezt felhasználva **következtetést (predikciót) tesznek lehetővé** (a meglévő vagy más/újabb adatokról).
- Adatorientált módszereket használ, tehát lényeges a
  - Változósám
  - Mintaszám
- Ezek aránya jelentősen befolyásolja a tanult modell minőségét

# Gépi tanulás vs adatbányászat

- Gépi tanulás
  - Előrejelző modellt épít
- Adatbányászat
  - ML módszereket felhasználva nyer ki új információt
  - Inkább feltáró jellegű
  - Inkább nem felügyelt tanulás



# Alapfeladatok



# Felügyelt és nem felügyelt tanulás

## ■ Felügyelt tanulás

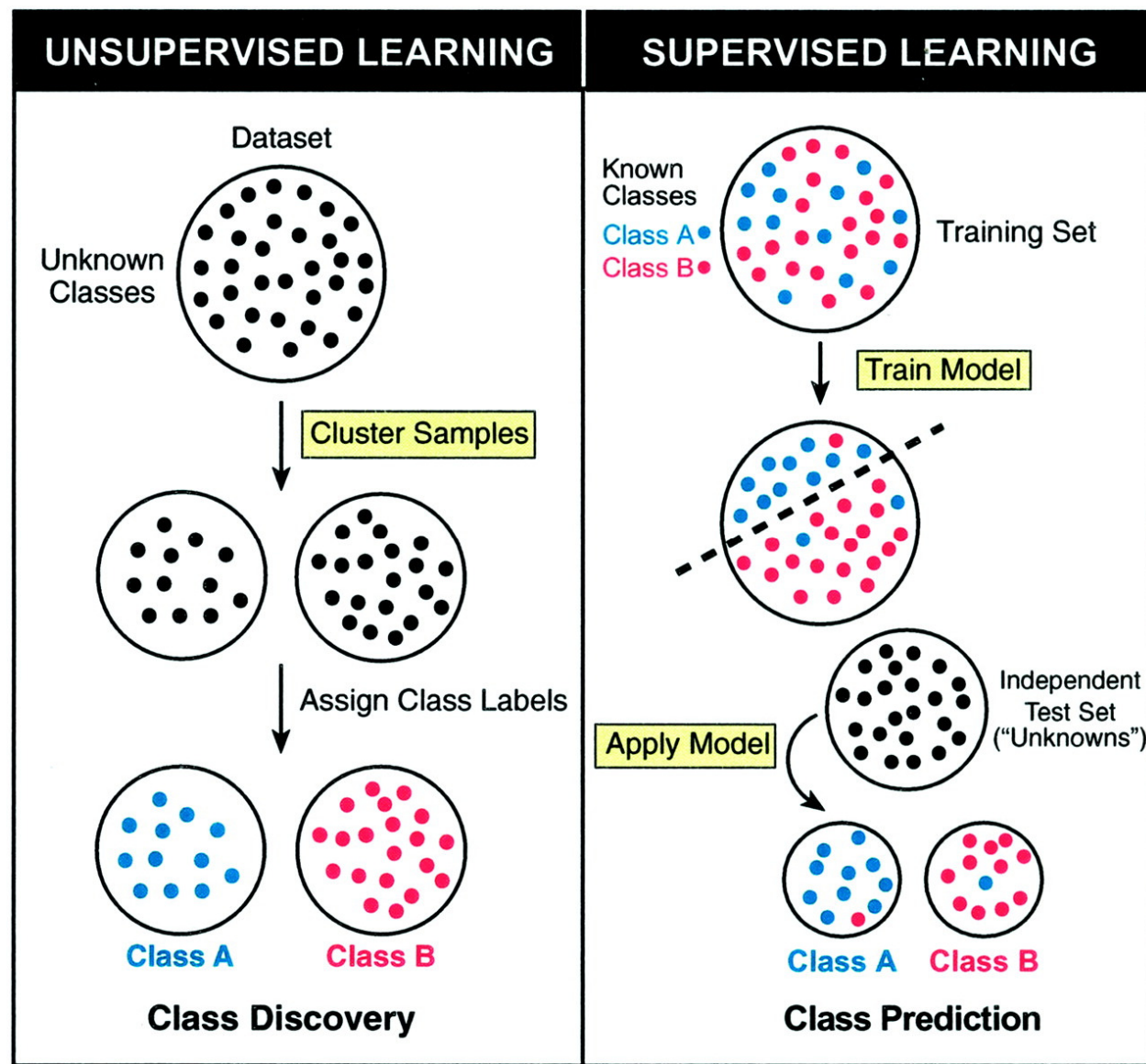
- Adott néhány pontra az elvárt kimenet is
- $\approx$  a tanuló példákból való általánosítás
- Output: függvény
  - a meglévő mintákra
  - megfelelően általánosítva

**Tanulóhalmaz** – amin építjük a modellt  
**Teszthalmaz** – amin ellenőrizzük

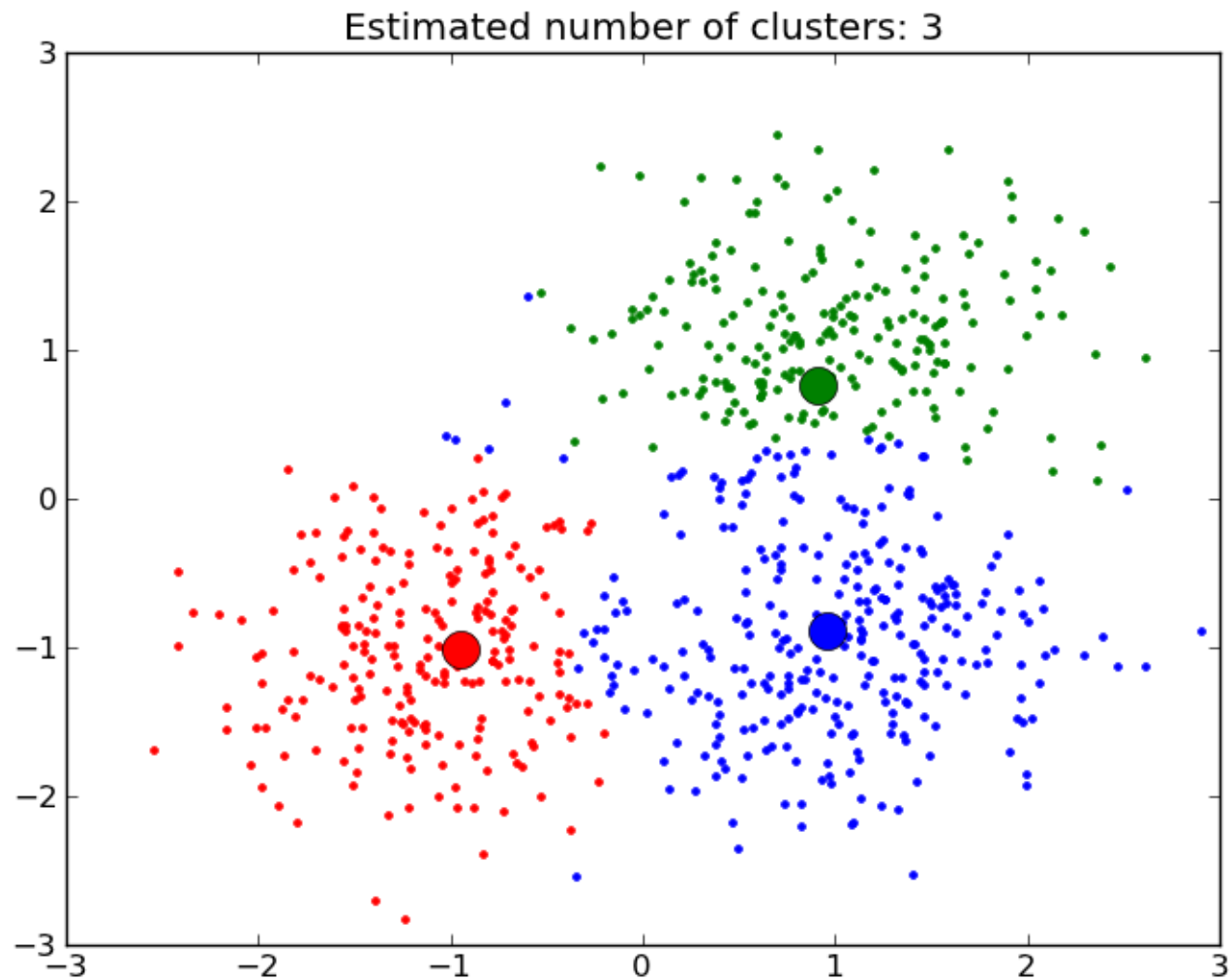
## ■ Nem felügyelt tanulás

- Nincs meg az elvárt kimenet
- Visszajelzés nélkül építi a modellt
- $\approx$  szabályok, összefüggések keresése (ismeretfeltárás)

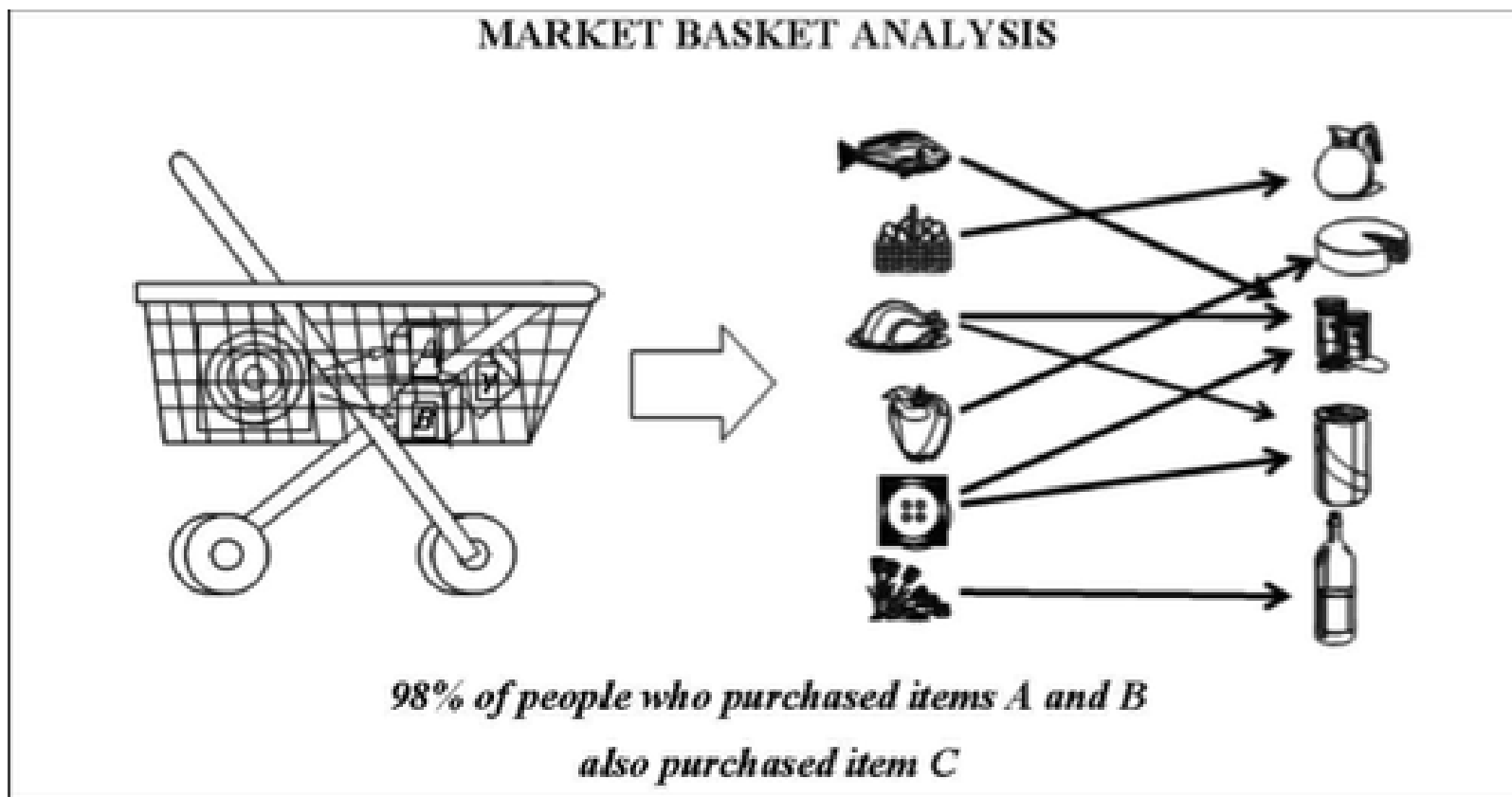
# Osztályozás és csoportosítás alapfeladat



# Klaszterezés

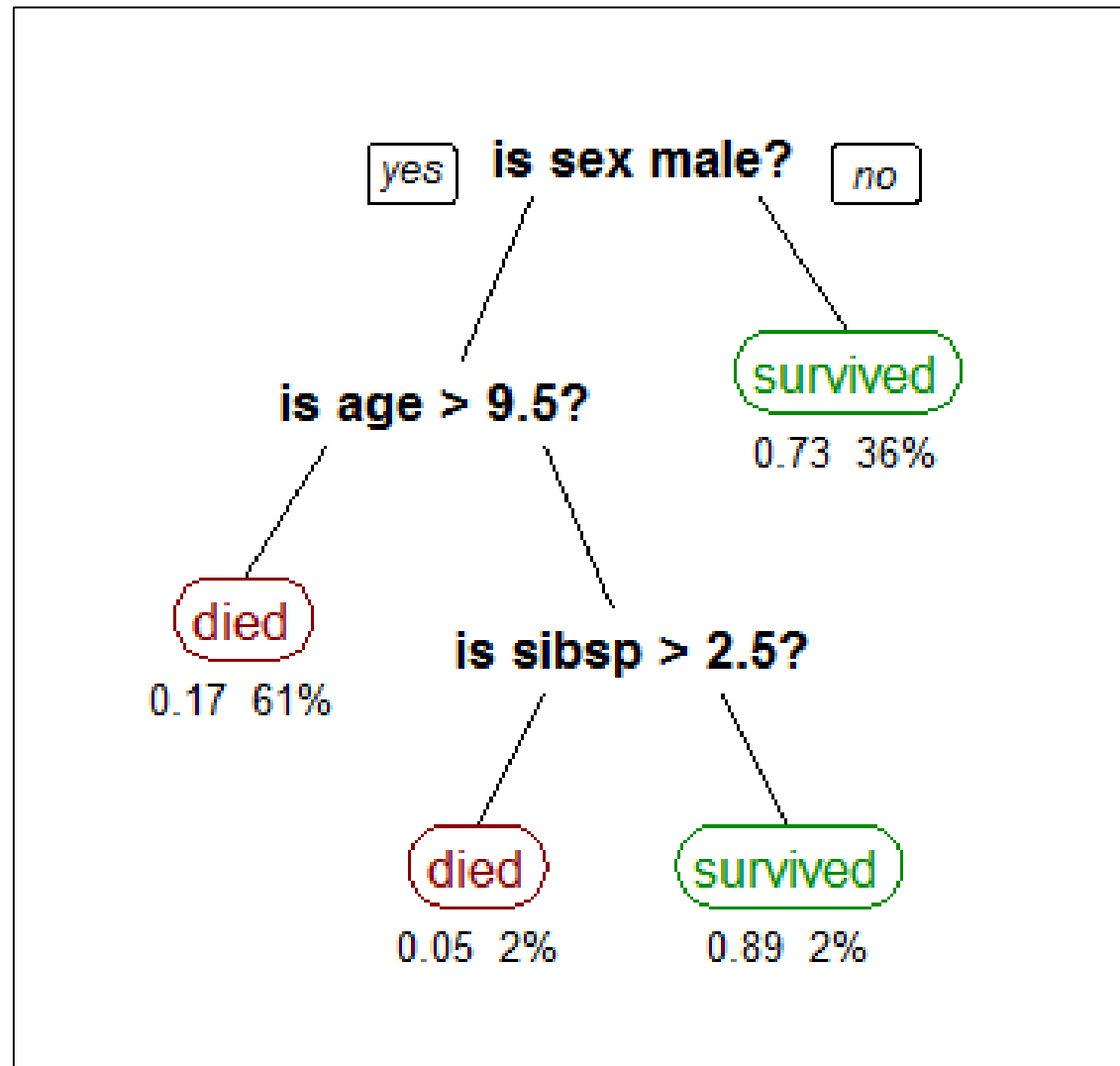


# Asszociációs szabályok

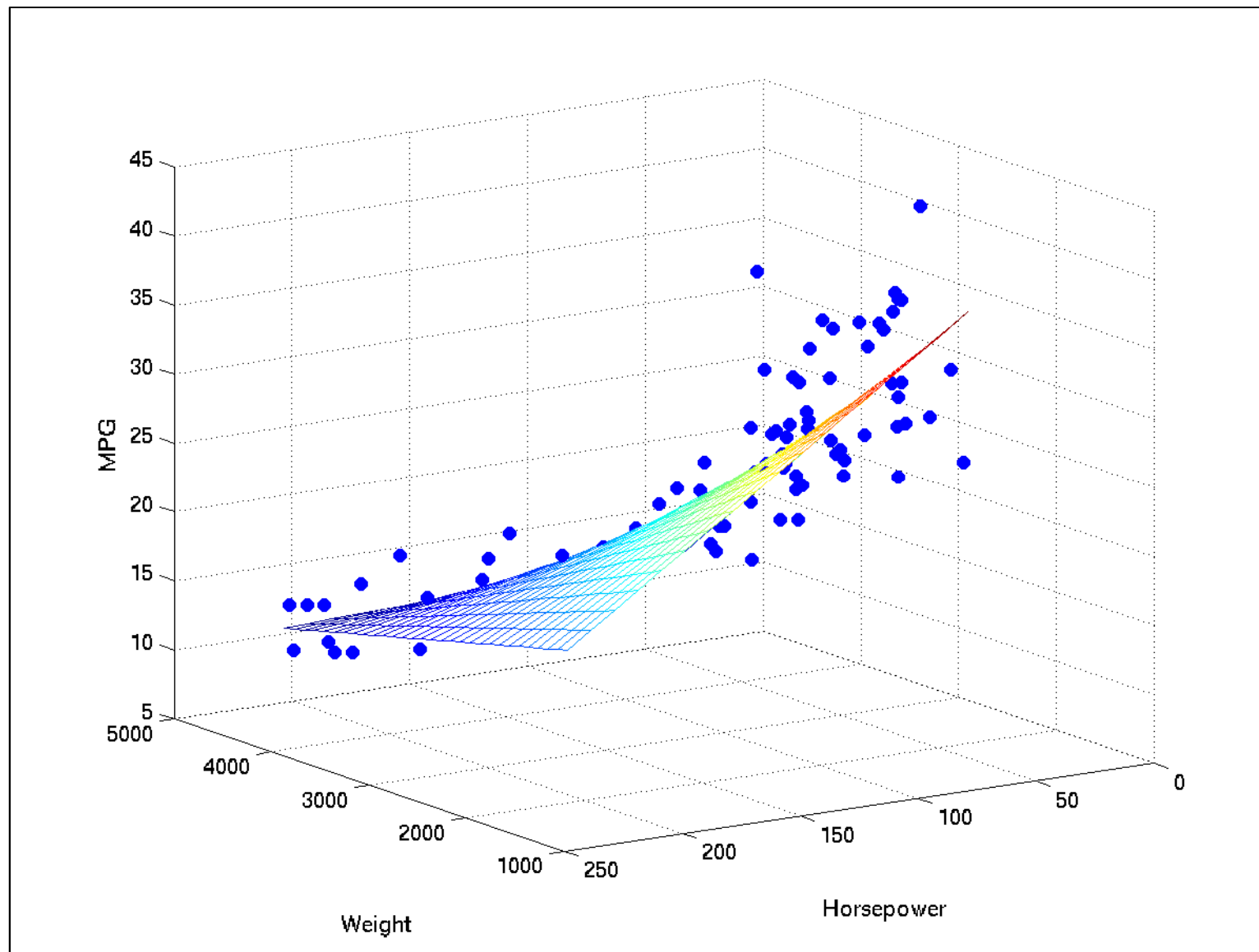




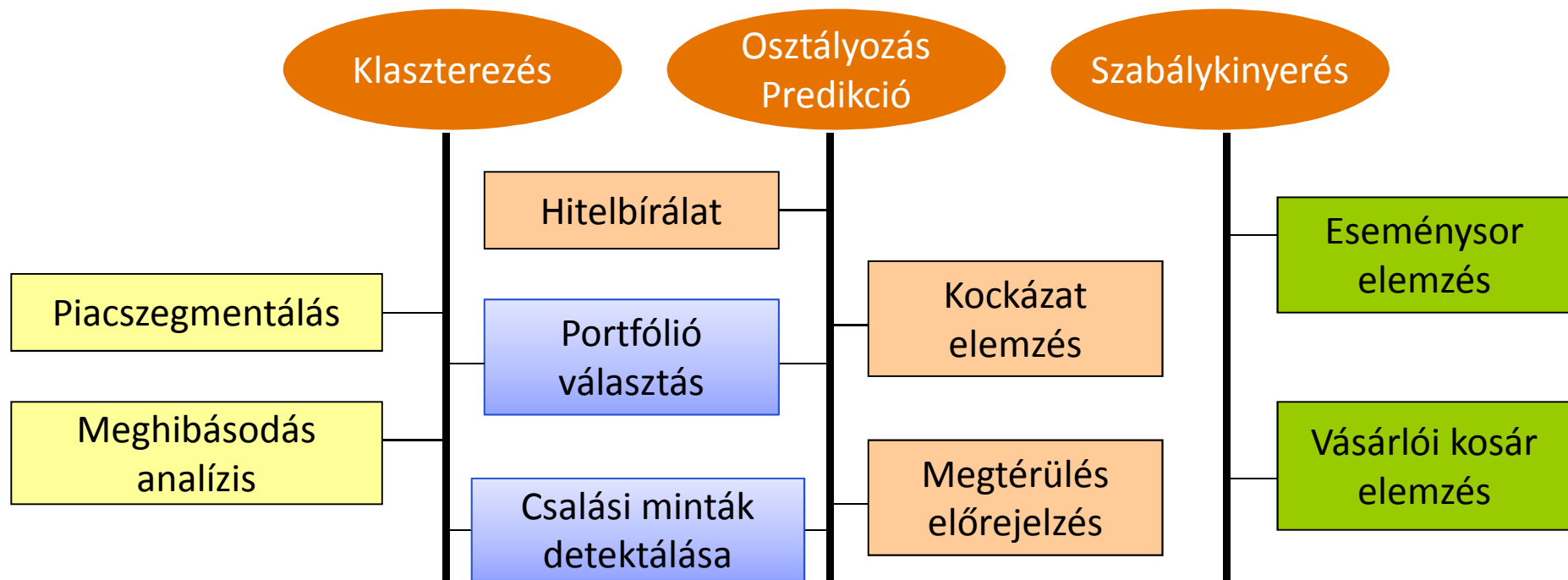
# Osztályozás



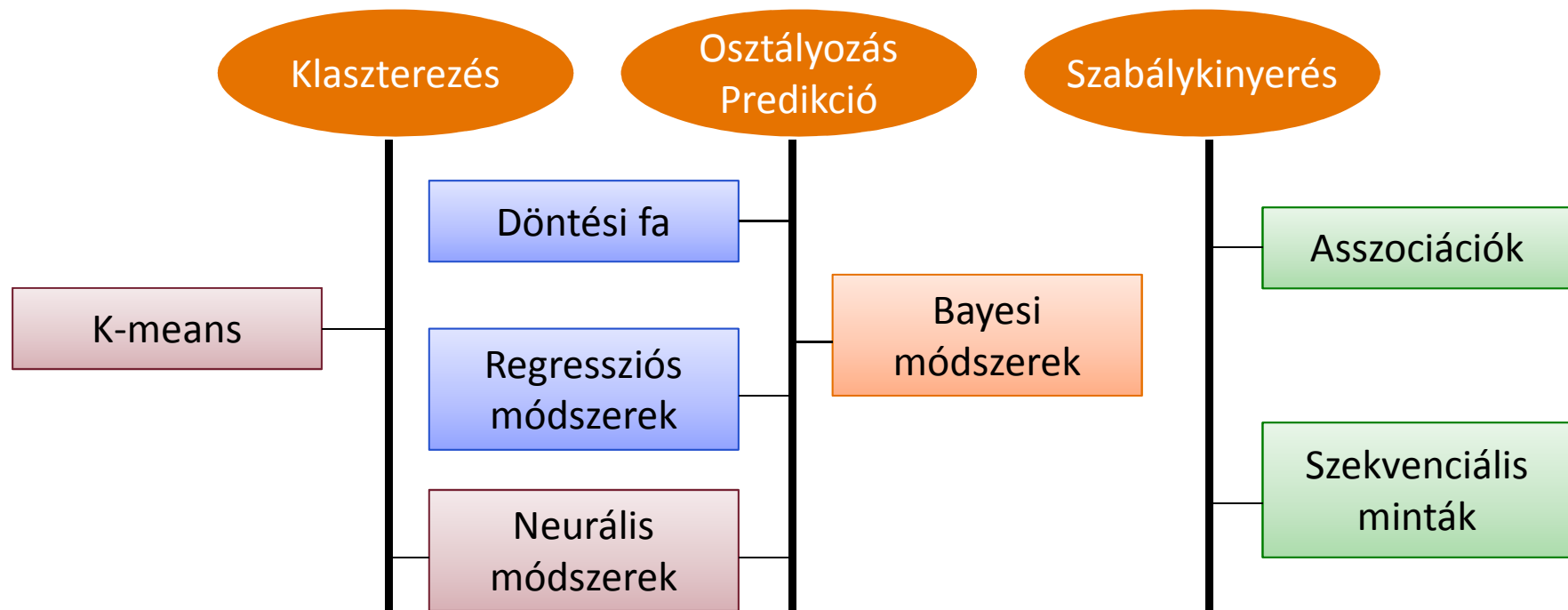
# Regresszió



# ML alapú üzleti intelligencia alkalmazások

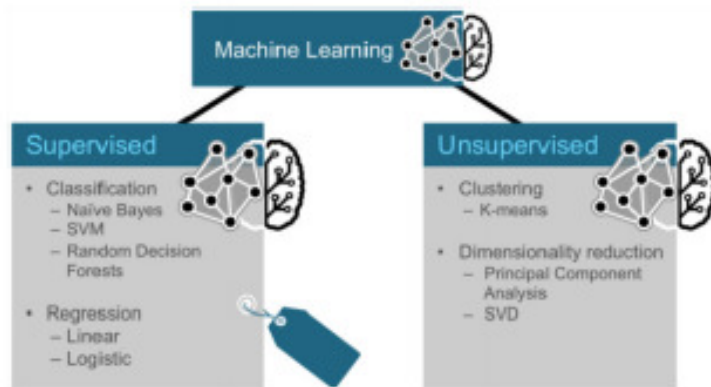


# Módszerek



# ML módszerek - BigData környezetben

Apache Spark MLlib (Machine language library) 



Collaborative Filtering  
(Recommendation)



<https://www.mapr.com/blog/apache-spark-machine-learning-tutorial>

# ML módszerek - BigData környezetben

## H2O

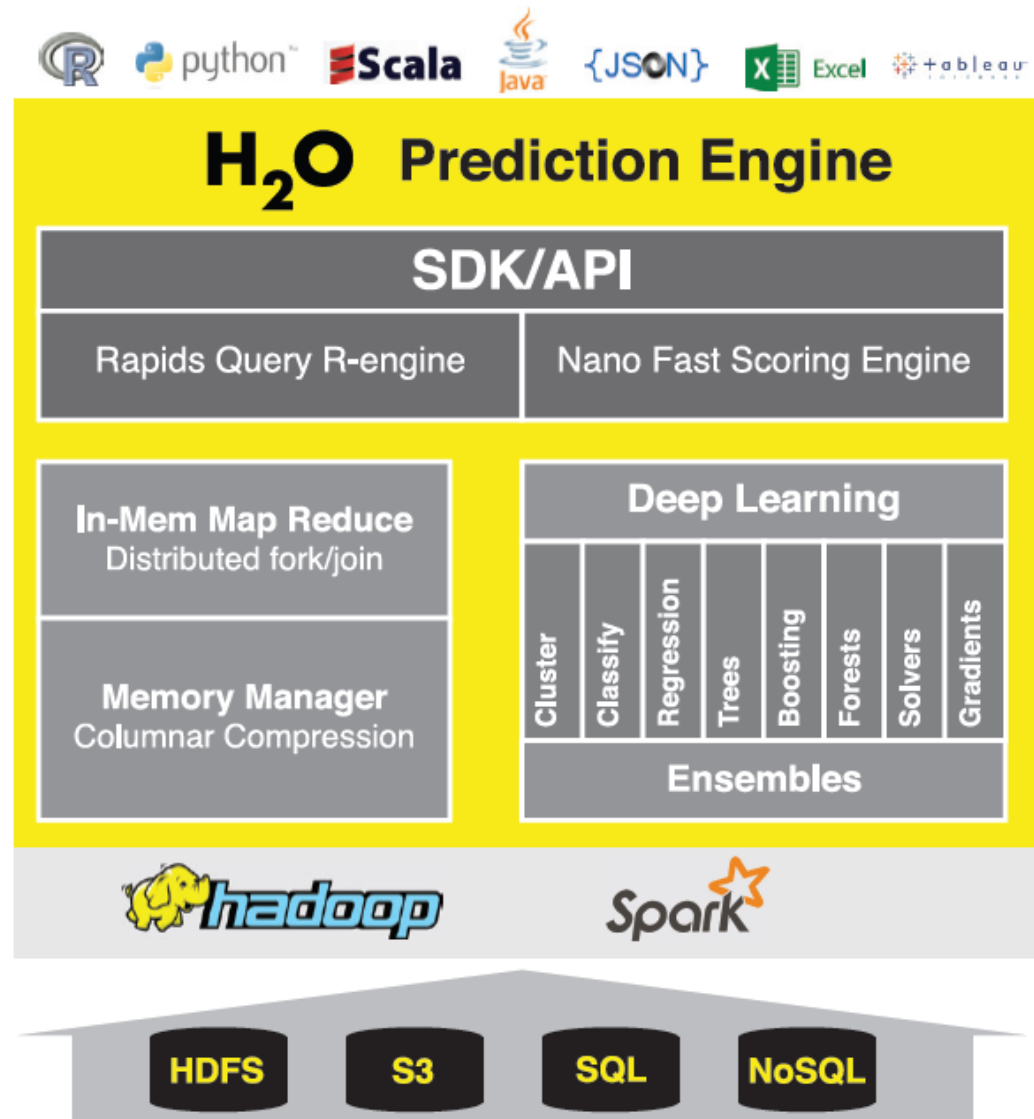
### Supervised Learning

Generalized Linear Modeling (GLM)  
Gradient Boosting Machine (GBM)  
Deep Learning  
Distributed Random Forest  
Naive Bayes Tutorial Booklet  
Ensembles (Stacking)

### Unsupervised Learning

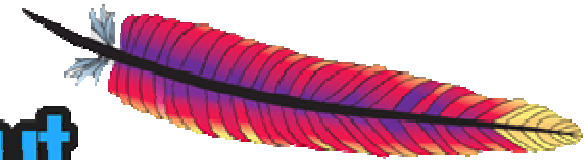
Generalized Low Rank Models (GLRM)  
K-Means Clustering  
Principal Components Analysis (PCA)

<http://docs.h2o.ai/h2o/latest-stable/index.html#algorithms>

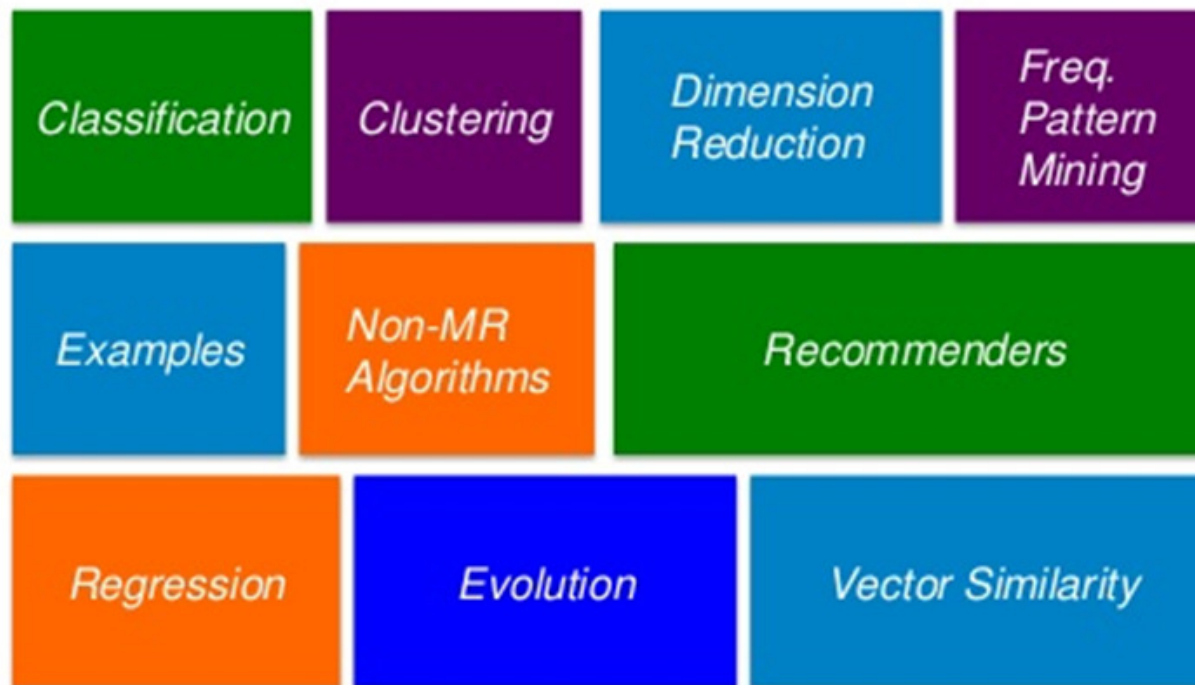


# ML módszerek - BigData környezetben

Mahout – Samsara

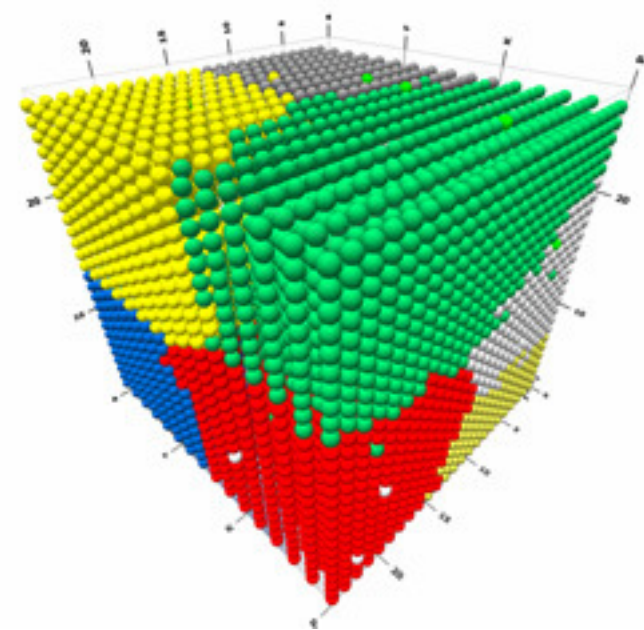
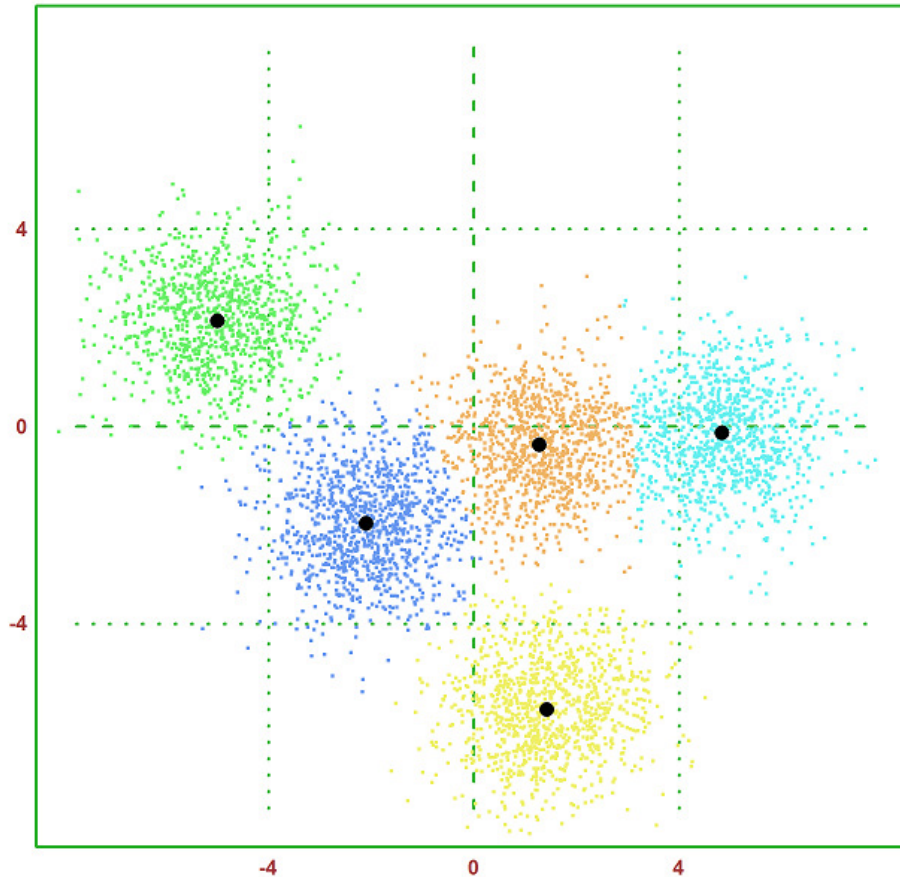


## Algorithms in Mahout



<https://mahout.apache.org/>

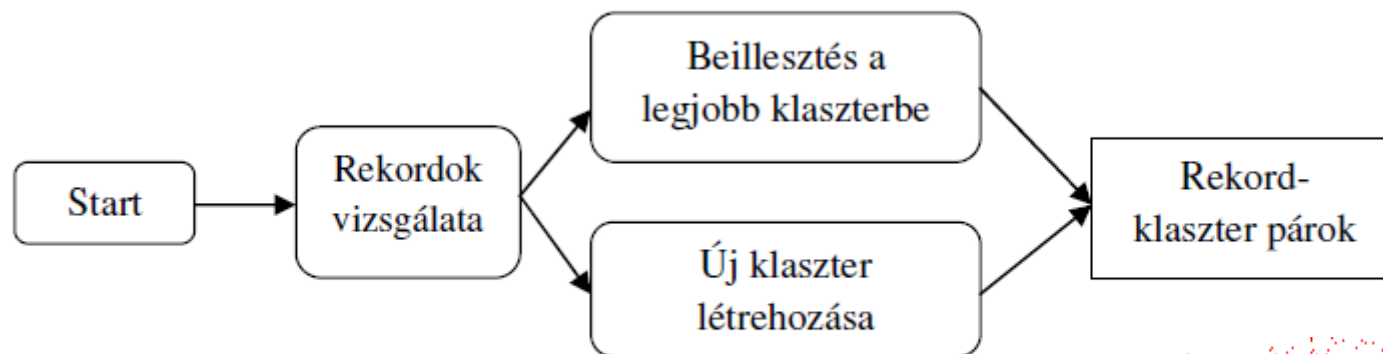
# Klaszterezés



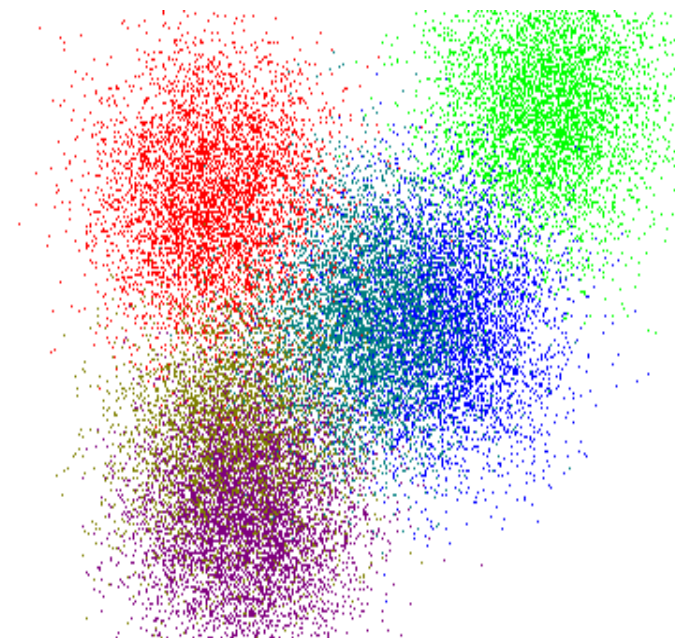
<http://gplsi.dlsi.ua.es/gplsi11/content/clustering-based-approach-unsupervised-word-sense-disambiguation>



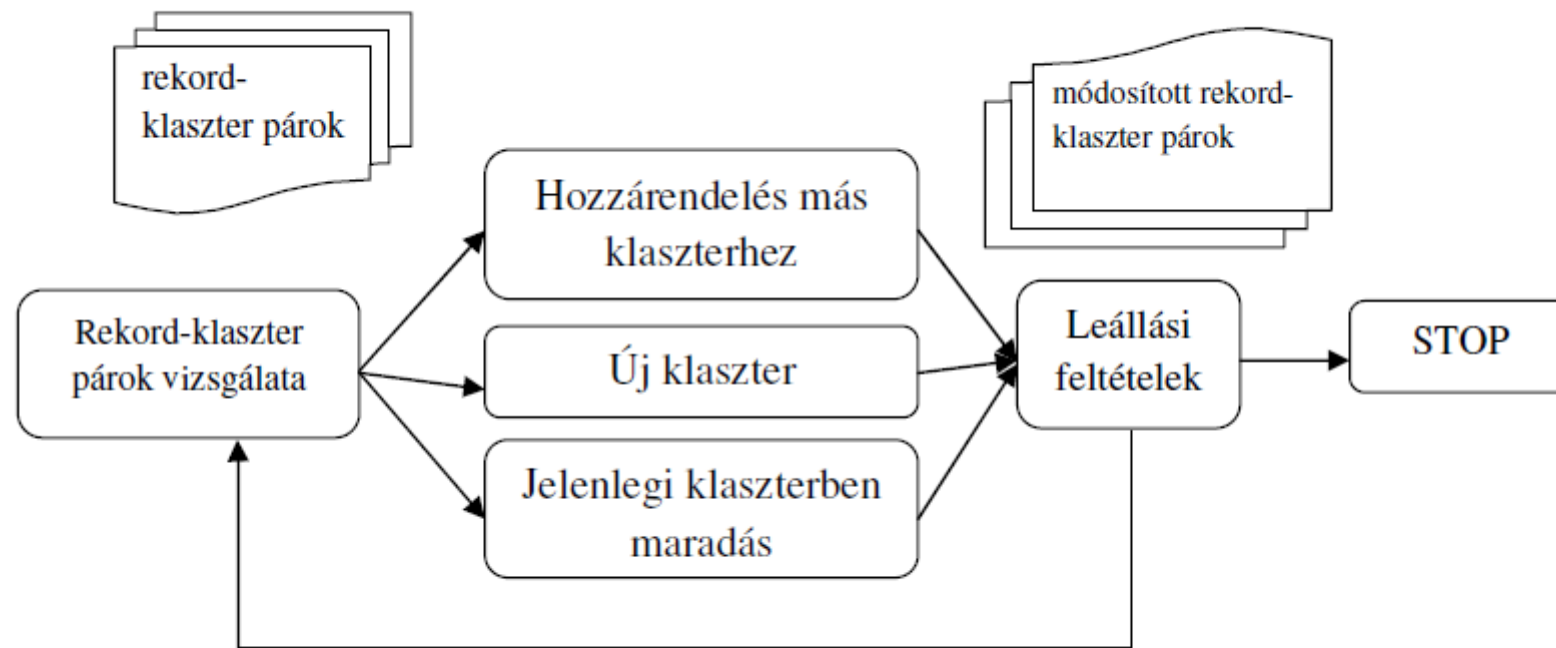
# Klaszterezés lépései I.



Az algoritmus lépései az adatbázis első átvizsgálásakor



# Klaszterezés lépései II.



Az algoritmus működése a további ciklusokban

# Klaszterezés alapkérdései – I.

2 fő célkitűzés:

- Az egyes **klasztereken belül** legyen a lehető legnagyobb a homogenitás
  - = nagy 'hasonlóság' az elemek között
  - = kis 'távolság' az elemek között
- Az különböző **klaszterek között** legyen nagy a 'távolság'
  - = kis 'hasonlóság' a különböző klaszterekbe tartozó elemek között
- A két kritérium között kell egyensúlyt találni

# Klaszterezés alapkérdései – II.

- Mi legyen a **hasonlósági mérce** (similarity metric)?
  - Folytonos és diszkrét numerikus változóknál egyszerűen értelmezhető
  - Kategorikus változóknál külön megfontolást igényel
- Mekkora legyen a **hasonlósági küszöb**?
  - Túl alacsony: jelentősen különböző elemek kerülnek egy klaszterbe
  - Túl magas: elég hasonló elemeket sem enged egy klaszterbe kerülni

# Klaszterezés alapkérdései – III.

- Mekkora az **optimális klaszterszám**?
  - Extrém esetek: → **használatatlan**
    - Annyi klaszter ahány minta
    - Egy óriási klaszter (benne minden minta)
  - Túl sok klaszter → **nehezen értelmezhető**
  - Túl kevés klaszter → **nem különülnek el a csoportok**
- Nincs általános megoldás
- Adott problémára megadható egy ideális klaszterszám

# K-közép (k-means) klaszterezés

- Adatpontok: vektortér
- Klaszter reprezentációja: súlyponttal / "középponttal" (vektor-átlag)
- $r(C_i)$ : i-edik klaszter reprezentánsa
- Minimalizálandó pontok négyzetes távolságösszege, mint hiba:

$$E(C) = \sum_{i=1}^k \sum_{u \in C_i} d(u, r(C_i))^2$$

# Egy megoldás

$\{r(C_1), r(C_2), \dots, r(C_k)\} \leftarrow \text{repr. kezdeti halmaza}$

**while**  $r(C_i)$  változik

**do for**  $\forall u \in D$  adott sorrendben

**do**

$h \leftarrow u$  klaszter-indexe

$j \leftarrow \text{argmin}_i d(u, r(C_i))$

**if**  $h \neq j$  **then** {

$C_j \leftarrow C_j \cup \{u\}$

$C_h \leftarrow C_h \setminus \{u\}$

$r(C_j) \leftarrow \frac{1}{|C_j|} \sum_{v \in C_j} v$

$r(C_h) \leftarrow \frac{1}{|C_h|} \sum_{v \in C_h} v$

Régi klaszter

Új klaszter

Itt rögtön újra is számoljuk

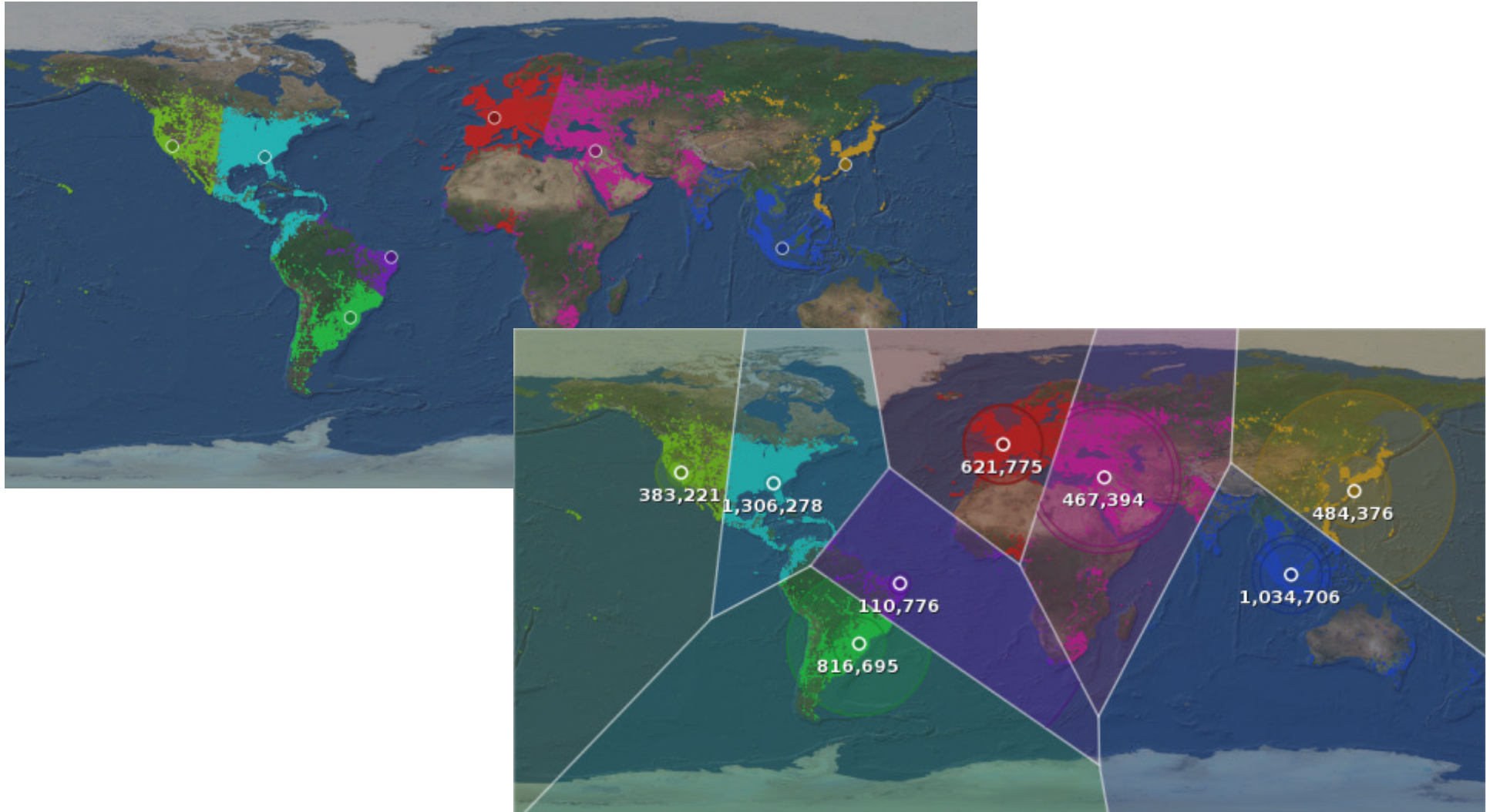
**return**  $\mathcal{C}$

# K-means klaszterezés példa

- Egy Twitter stream felhasználóinak klaszterezése földrajzi elhelyezkedés alapján
- <https://chimpler.wordpress.com/2014/07/11/segmenting-audience-with-kmeans-and-voronoi-diagram-using-spark-and-mllib/>
- Lehetséges változók:
  - Földrajzi elhelyezkedés
  - Tweet hossz
  - Első tweethez képest eltelt idő
- Vizualizáció: Voronoi cellákkal



# K-means klaszterezés példa



- <https://chimpler.wordpress.com/2014/07/11/segmenting-audience-with-kmeans-and-voronoi-diagram-using-spark-and-mllib/>

# Condorcet kritérium alapú klaszterezés

- **n** jelölt :  $x^{(i)}$
- **m** darab szavazó :  $v$  (voter)
- Mindegyik szavazó  $v_k$  egy rangsort állít fel  $\pi_k$  (permutáció)
  - Pl:  $x^{(\pi_k(i))} < x^{(\pi_k(j))}$  az  $i$ . helyen állót favorizálja  $j$ . helyen lévőhöz képest
- Az a jelölt győz aki a szavazók rangsorát páronként összemérve többször favorizált (paired majority rule)
- Ez alapján egy globális rangsort hoz létre  **$\lambda$** :
  - $x^{(\lambda(1))} < x^{(\lambda(2))} < \dots$

$$\sum_{1 \leq i < j \leq n} \#\{k | \pi_k^{-1}(\lambda(i)) < \pi_k^{-1}(\lambda(j))\}$$

# A szavazás

| 1. változó        | „Ellenfél” klaszter |          |          |          |          |
|-------------------|---------------------|----------|----------|----------|----------|
| „Jelölt” klaszter |                     | <b>1</b> | <b>2</b> | <b>3</b> | <b>4</b> |
|                   | <b>1</b>            | —        | 0        | 0        | 1        |
|                   | <b>2</b>            | 1        | —        | 1        | 1        |
|                   | <b>3</b>            | 1        | 0        | —        | 1        |
|                   | <b>4</b>            | 0        | 0        | 0        | —        |

- Preferencia sorrend (2,3,1,4)
- 0 jelentése: a jelölt rosszabb, mint az ellenfél
- 1 jelentése: a jelölt jobb, mint az ellenfél

# A szavazás

| 1I. változó       | „Ellenfél” klaszter |          |          |          |          | (4,1,3,2) |
|-------------------|---------------------|----------|----------|----------|----------|-----------|
| „Jelölt” klaszter |                     | <b>1</b> | <b>2</b> | <b>3</b> | <b>4</b> |           |
|                   | <b>1</b>            | —        | 1        | 1        | 0        |           |
|                   | <b>2</b>            | 0        | —        | 0        | 0        |           |
|                   | <b>3</b>            | 0        | 1        | —        | 0        |           |
|                   | <b>4</b>            | 1        | 1        | 1        | —        |           |

| 1II. változó      | „Ellenfél” klaszter |          |          |          |          | (1,3,2,4) |
|-------------------|---------------------|----------|----------|----------|----------|-----------|
| „Jelölt” klaszter |                     | <b>1</b> | <b>2</b> | <b>3</b> | <b>4</b> |           |
|                   | <b>1</b>            | —        | 1        | 1        | 1        |           |
|                   | <b>2</b>            | 0        | —        | 0        | 1        |           |
|                   | <b>3</b>            | 0        | 1        | —        | 1        |           |
|                   | <b>4</b>            | 0        | 0        | 0        | —        |           |

# Végeredmény

| $\Sigma$             | „Ellenfél” klaszter |          |          |          |          |
|----------------------|---------------------|----------|----------|----------|----------|
| „Jelölt”<br>klaszter |                     | <b>1</b> | <b>2</b> | <b>3</b> | <b>4</b> |
|                      | <b>1</b>            | —        | 2        | 2        | 2        |
|                      | <b>2</b>            | 1        | —        | 1        | 2        |
|                      | <b>3</b>            | 1        | 2        | —        | 2        |
|                      | <b>4</b>            | 1        | 1        | 1        | —        |

# Condorcet kritérium alkalmazása klaszter jelöltek minősítésére

- $n$  jelölt  $\rightarrow n$  darab klaszter
- $m$  szavazó  $\rightarrow m$  darab változó

## Jelölés:

$x, y$  : két összehasonlítandó objektum

$v$  : egy adott változó,  $v(x)$  :  $x$  objektum  $v$  változó szerinti értéke

$\mathcal{V}$  : az összes változó halmaza

$\mathcal{C}$  : egy adott klaszter

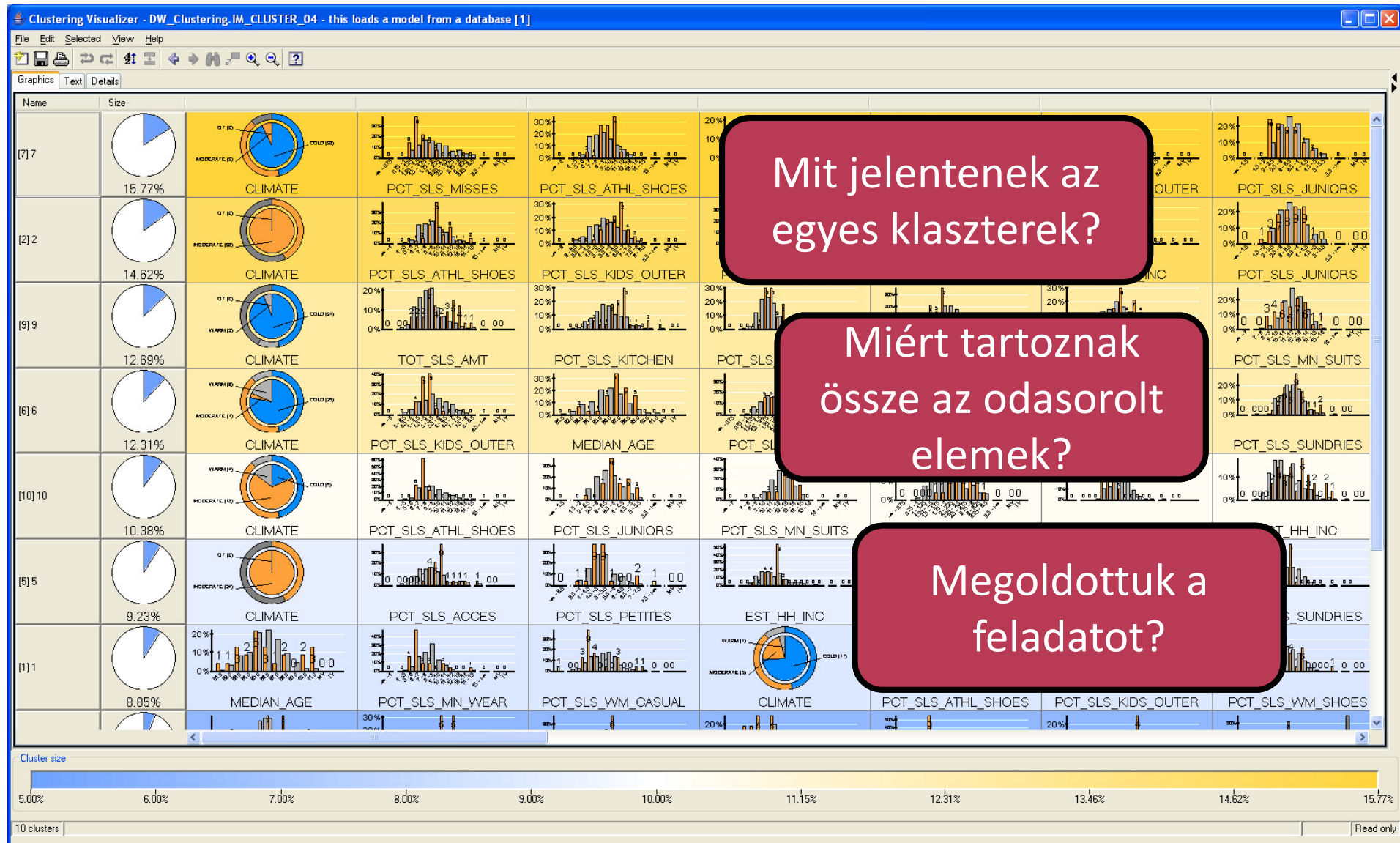
$\mathcal{C}$  : teljes klaszterezés

$$c(\mathcal{C}) := \sum_{C \in \mathcal{C}} \sum_{\substack{x, y \in C \\ x \neq y}} \sum_{v \in \mathcal{V}} [v(x) \sim v(y)] + \sum_{C \in \mathcal{C}} \sum_{\substack{x \in C \\ y \in \mathcal{O} \setminus C}} \sum_{v \in \mathcal{V}} [v(x) \not\sim v(y)]$$

# Demografikus klaszterezés: 1. példa - 1

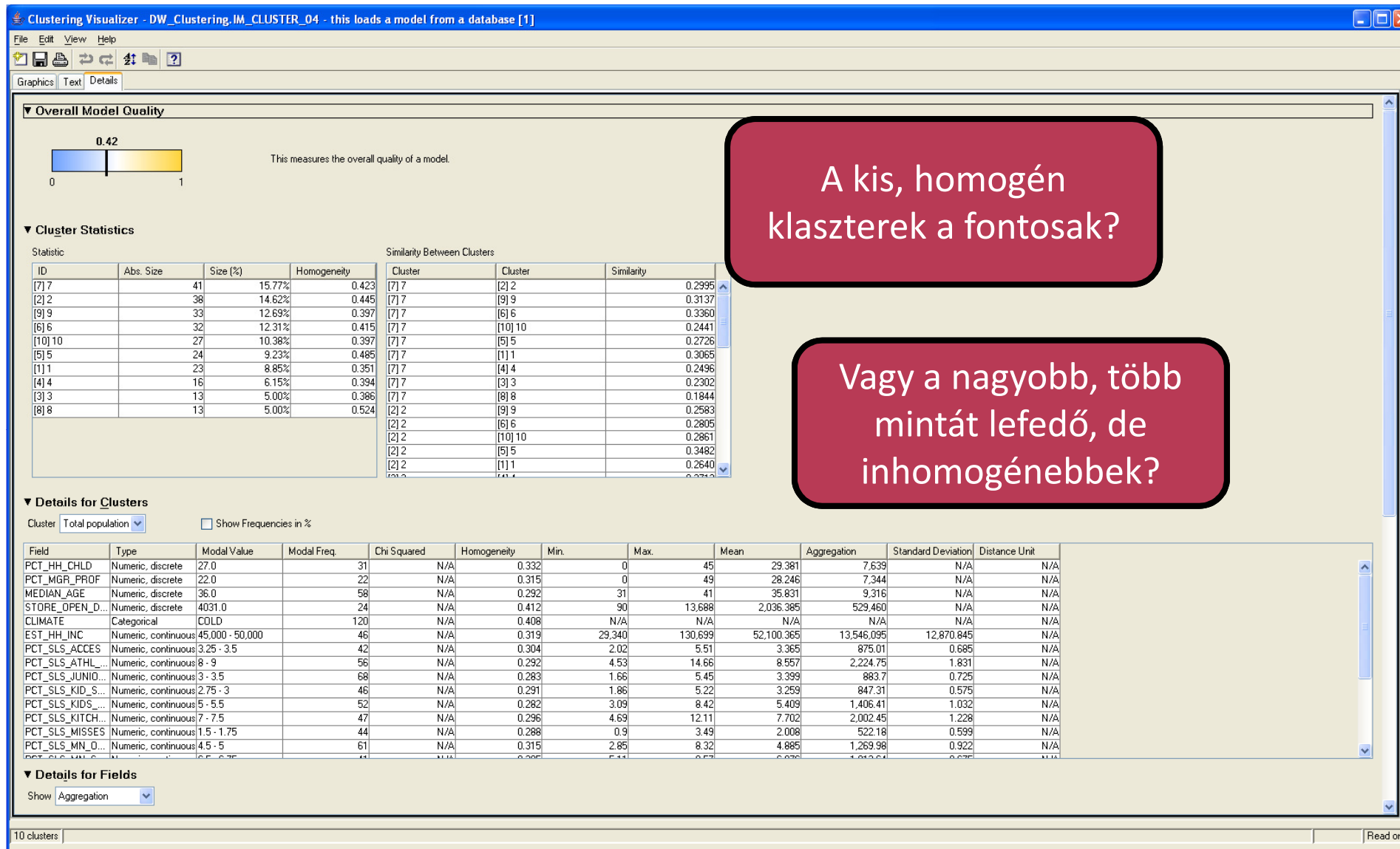
- Feladat: Áruházak klaszterezése
- Változók:
  - Áruházban kialakított osztályok bevétele
  - Elhelyezkedés
  - Éghajlat/időjárás
  - Vásárlók életkora
  - Vásárlók becsült jövedelme
- Kérdés: Ha új áruházat szeretnénk létrehozni, hogyan alakítsuk ki?

# Demografikus klaszterezés: 1. példa - 2





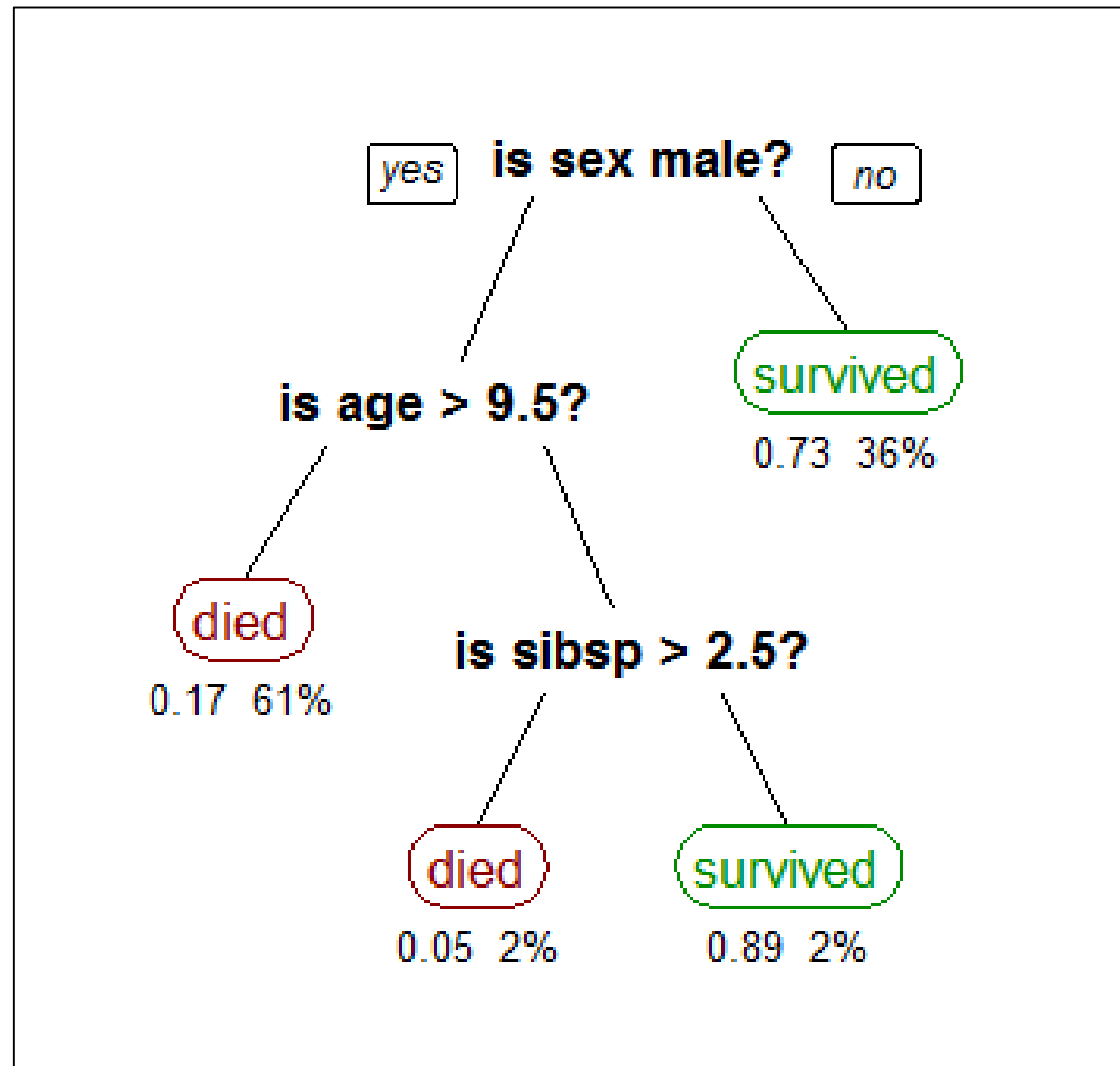
# Demografikus klaszterezés: 1. példa - 3



A kis, homogén  
klaszterek a fontosak?

Vagy a nagyobb, több  
mintát lefedő, de  
inhomogénebbek?

# Osztályozás

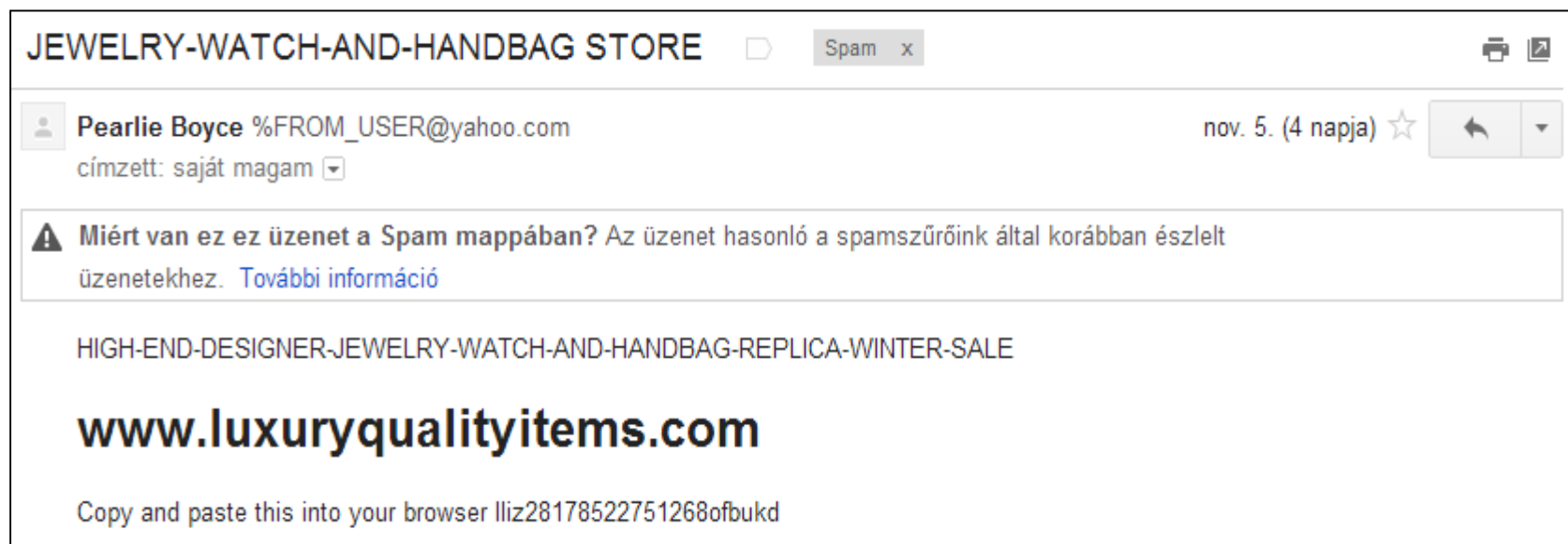


# Osztályozás alapfeladat



Képosztályozás: a képen látható objektum madár vagy repülő?

# Osztályozás alapfeladat



Levelek osztályozása: SPAM vagy nem SPAM?

# Osztályozás alapfeladat

Netcool/OMNibus Event List : Filter="Sun Events", View="Default"

File Edit View Alerts Tools Help

Sun Events Default

| Node      | Alert Group                 | Summary                                                                |
|-----------|-----------------------------|------------------------------------------------------------------------|
| wgs97-210 | petTrapEntityPresenceDevice | Sensor io.id0.prsnt : A device is absent or has been removed.          |
| wgs97-210 | SunHwTrapPowerSupplyError   | PowerSupply component /SYS/PS0/PWRDK has detected an error (           |
| wgs97-210 | SunHwTrapPowerSupplyError   | PowerSupply component /SYS/PS1/PWRDK has detected an error (           |
| wgs40-01  | sunPlatStateChange          | Device CH/FANBD1/FM2/SERVICE Operational Status has changed            |
| wgs40-01  | sunPlatStateChange          | Device CH/MB/CMP0/BR3/CH0/D0/SERVICE Operational Status ha             |
| wgs40-01  | sunPlatStateChange          | Device CH/SYS/PS_FAULT Operational Status has changed to disab         |
| wgs40-01  | sunPlatStateChange          | Device CH/SYS/SERVICE Operational Status has changed to disable        |
| wgs40-01  | sunPlatStateChange          | Device CH/SYS/TEMP_FAULT Operational Status has changed to di          |
| wgs97-216 | petTrapFanLowerCritical     | Sensor Axial Fan 0 : Fan speed has decreased below lower critical thre |
| wgs97-210 | SunHwTrapPreQSError         | PreOS Error. Additional Info : System boot initiated                   |
| wgs97-210 | SunHwTrapPreQSError         | PreOS Error. Additional Info : Motherboard initialization              |
| wgs97-210 | petTrapPowerSupplyState     | Sensor ps0.pwrok : Power Supply is connected to AC Power.              |
| wgs97-210 | petTrapPowerSupplyState     | Sensor ps1.pwrok : Power Supply is connected to AC Power.              |

0 2 0 2 8 1

No rows modified 12/11/2007 2:42:25 P root NCOMS [PRI]

Szabályok alapján Severity  
osztályozása

# Bináris döntések jóságának mérése

Confusion  
matrix

Jósolt állapot (predikció)

Valós állapot

|               |   | Jósolt állapot (predikció) |                         |
|---------------|---|----------------------------|-------------------------|
|               |   | P                          | N                       |
| Valós állapot | P | TP                         | FN<br>(II. típusú hiba) |
|               | N | FP<br>(I. típusú hiba)     | TN                      |

# Bináris döntés (osztályozás) jósági mutatói

- **Érzékenység (sensitivity):  $TP/(TP+FN)$**  az összes eredetileg pozitív esetből az osztályozó mennyit osztályozott jól
- **Specifikusság (specificity):  $TN/(TN+FP)$**
- **Negatív prediktív érték:  $TN/(TN+FN)$**  a valójában negatív esetek aránya az osztályozó által negatívnak osztályozott esetekhez képest
- **Pozitív prediktív érték:  $TP/(TP+FP)$**
- **Pontosság:  $TP+TN/(TP+FP+TN+FN)$**
- **AUC (area over the ROC curve)**  
: sensitivity (->y) versus 1- specificity(-> x))

# Érzékenység vagy specifikusság a fontos?

- **Érzékenység (sensitivity):  $TP/(TP+FN)$**

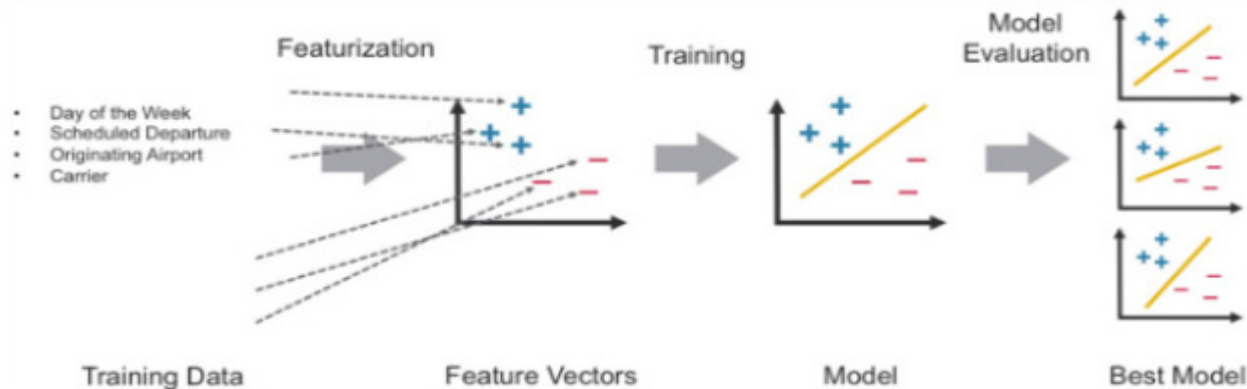
- érzékenység magas  $\rightarrow$  FN alacsony
- Pl: Legyen kicsi az esélye, hogy egy fertőző betegség tesztje tévesen negatív  $\rightarrow$  aki negatív eredményű az valójában is az.

- **Specifikusság (specificity):  $TN/(TN+FP)$**

- specifikusság magas  $\rightarrow$  FP alacsony
- Pl: Legyen kicsi az esélye, hogy egy genetikai teszt (egy kezelés megkezdéséhez) tévesen pozitív  $\rightarrow$  aki pozitív eredményű az valójában is az.



# Az osztályozás további tényezői - 1



<https://www.mapr.com/blog/apache-spark-machine-learning-tutorial>

- Jegykiválasztás (= változó halmaz kiválasztás)
  - Legyen **releváns** a változóhalmaz
  - **Ne legyenek** benne **redundáns** változók
- Jegykiválasztás és modellépítés mennyire integrált
- Számos kiértékelési módszer  
pl: keresztvalidáció típusok (k-fold, leave-1-out)

# Az osztályozás további tényezői - 2

- **Modell komplexitás vs. osztályozó jósága**
  - Sok esetben adott egy trade-off a két tulajdonság közt
  - Triviális: Két közel azonos jóságú modell közül az egyszerűbbet kell választani
  - De: mekkora komplexitás növekedést érdemes még elfogadni kis mértékű javulás fejében?
- **vagy **black box** módszer a megfelelő?**
  - : a felépített modell változói, paraméterei értelmezhetőek. **Pl: döntési fák, Bayes-hálók**
  - **Black box:** a modell elemei önmagukban nem értelmezhetőek, de jól megoldja az osztályozó a feladatot. **Pl: neurális hálózatokon alapuló osztályozók**

# Osztályozási feladat

## Adat:

- Adott egy filmes adatbázis

## Modelltanulás:

- Szeretnénk meghatározni, hogy mitől lesz sikeres egy film (=többet hoz vissza az eladott jegyek árából, mint amennyibe került)

## Predikció (jóslás):

- Szeretnénk tervezett jövőbeli filmekről vagy készülő filmekről megállapítani, hogy sikeresek lesznek-e

# Osztályozási feladat

**Célváltozó:** „Sikeres film” Y0: Nem Y1: Igen

Y1:Kiváló

Lehetséges magyarázó változók:

○ X1: Film típusa pl: akció vagy vígjáték

X2:2

○ X2: A filmben szereplő világsztárok száma >1

○ X3: Speciális effektusok minősége

X3:Kiváló

○ X4: Filmzene minősége

○ X5: Eladható termékek megjelenítése

○ X6: Film megjelenésének időzítése

X6:Kiváló

○ X7: Időjárás

X7:Jó

○ X8: Eredeti/adaptáció

○ X9: Marketing kampány sikeressége

○ X10: Sorozat része vagy egyedi

Releváns  
változók

# Osztályozási feladat-> regressziós feladat

**Célváltozó:** „Sikeres film” Y0: Nem Y1: Igen

Y1:Kiváló

Mi lenne, ha számszerűsíteni lehetne a sikert?  
Pl. dollárban kifejezve?

X2:2

Az egyes változókhoz X1, X2, X3 ...  
hozzárendelhetünk egy-egy tagot, ami leírja, hogy  
mennyivel járul hozzá a sikerhez:  $\beta_1, \beta_2, \beta_3...$

X3:Kiváló

Ekkor: Célváltozó nem pusztán a siker, hanem a  
várható árbevétel.

X6:Kiváló

X7:Jó

X8: Eredeti/adaptáció

Tegyük fel, hogy a várható árbevétel egy lineáris  
egyenlettel leírható

Releváns  
változók

# Regressziós feladat

**Célváltozó: Mekkora lesz a film bevétele?**

$Y=50M\$$

Lehetséges magyarázó változók:

○ X1: Film típusa pl: akció vagy vígjáték

$\beta_1 = 0$

○ X2: A filmben szereplő világsztárok

$\beta_2 * X_2 = 8M\$ * 2 = 16M\$$

○ X3: Speciális effektusok minősége

$\beta_3 * X_3: 6M\$ * 4 = 24M\$$

○ X4: Filmzene minősége

$\beta_4 = 0$

$\beta_5 = 0$

○ X5: Eladható termékek megjelenítése

○ X6: Film megjelenésének időzítése

$\beta_6 * X_6: 2M\$ * 4 = 8M\$$

○ X7: Időjárás

$\beta_7 * X_7: 1M\$ * 2 = 2M\$$

○ X8: Eredeti/adaptáció

$\beta_8 = 0$

$\beta_9 = 0$

$Y = \alpha + \beta_i * X_i$

$\alpha$ : offset (eltolás)

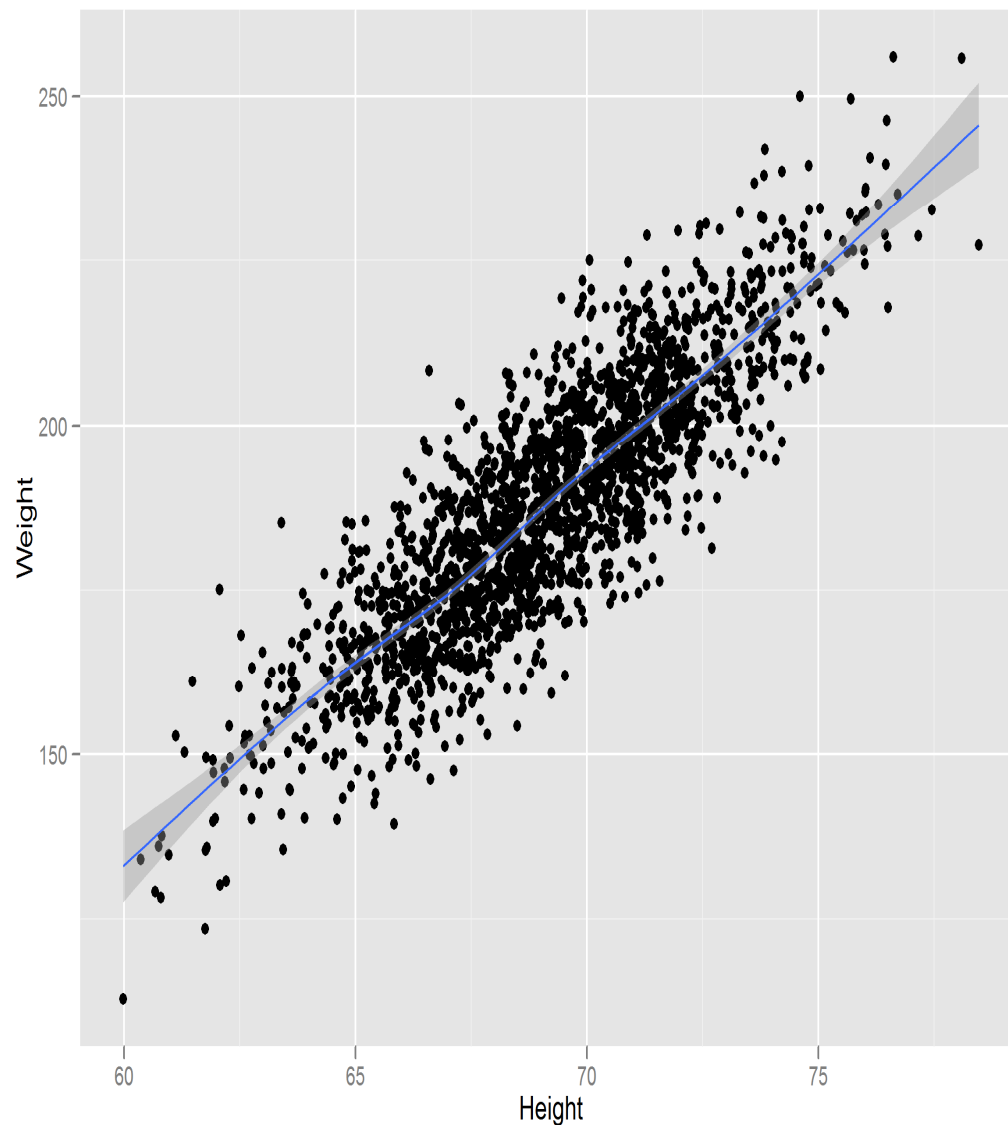
$\beta_i$ : regressziós együtthatók

$\beta_{10} = 0$

# Regresszió

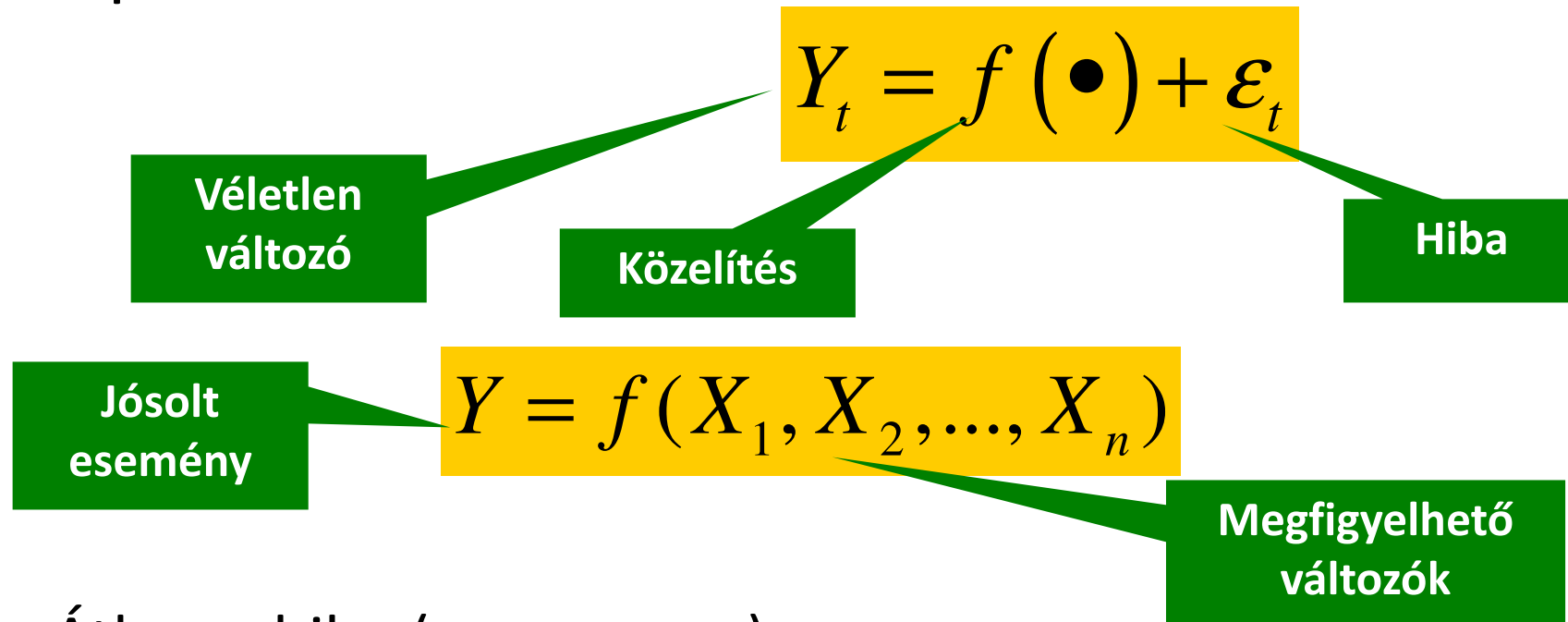
$f$  függvény,

- bemenet:  
az attribútumok értéke
- kimenet: megfigyelések  
legjobb közelítése
- „ökölszabály”
- Példa:  
testtömeg/magasság  
együttes eloszlás  
valójában egyenesre  
illeszthető,

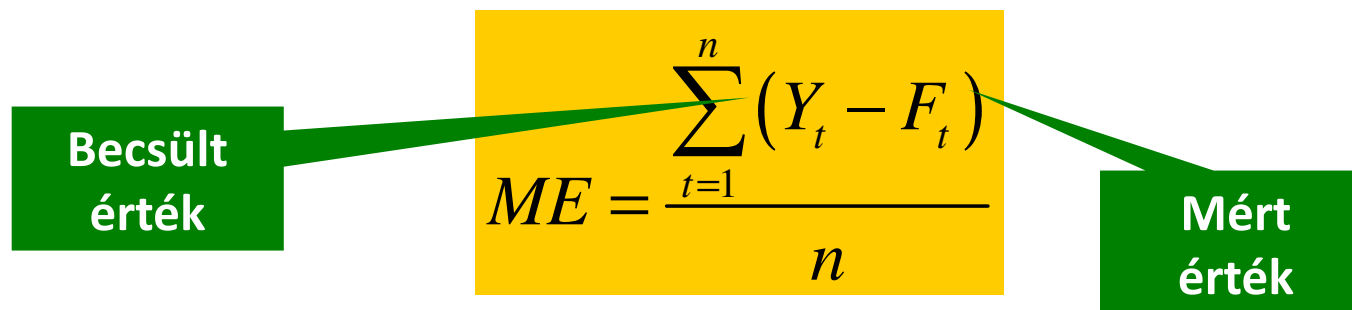


# Regressziós módszerek

- Alapelv:



- Átlagos hiba (mean error)





# Lineáris regresszió

- Egyszerű lin. függvény illesztése az adatokra
  - nem vár alapvető változást a rendszer viselkedésében

$$Y = a + bX$$

- Legkisebb négyzetek módszere
  - keressük azokat az  $a, b$  paramétereket, amelyekre

$$SSE = \sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n (Y_t - F_t)^2 \quad \text{minimális (Sum of Squared Errors)}$$

- cél:

$$\sum_{t=1}^n (Y_t - F_t)^2 = \sum_{t=1}^n [Y_t - (a + bX_t)]^2$$

# Levezetés (parc. deriválás)

$$\frac{d \sum_{t=1}^n [Y_t - (a + bX_t)]^2}{da} = \sum_{t=1}^n (-2) [Y_t - (a + bX_t)] = 0$$

$$na = \sum_{t=1}^n (Y_t - bX_t)$$

$$a = \bar{Y} - b\bar{X}$$

**$X_i, Y_i$  a mért értékpárok  
(pl. idő, terhelés)**

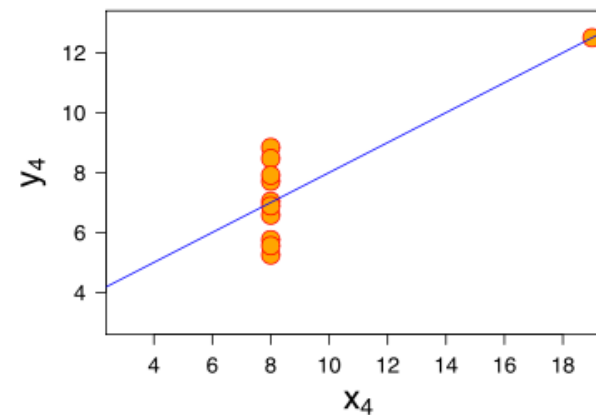
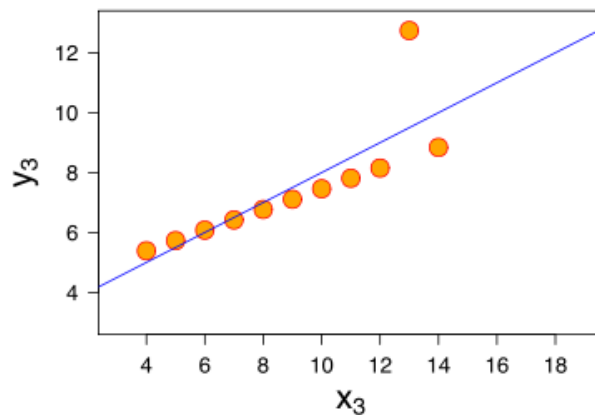
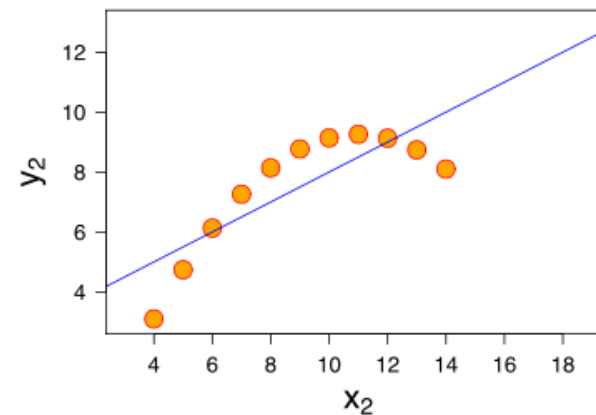
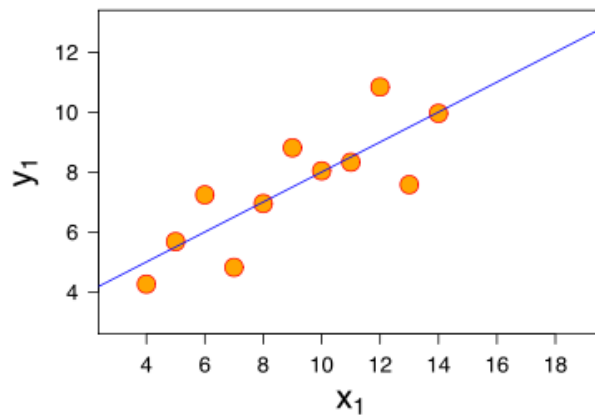
$$\frac{d \sum_{t=1}^n [Y_t - (a + bX_t)]^2}{db} = \sum_{t=1}^n X_t [Y_t - (a + bX_t)] = 0$$

$$\sum_{t=1}^n X_t \left[ Y_t - \frac{1}{n} \sum_{t=1}^n (Y_t - bX_t) - bX_t \right] = \sum_{t=1}^n X_t Y_t - \frac{1}{n} \left( \sum_{t=1}^n X_t \right) \left( \sum_{t=1}^n Y_t \right) + \frac{1}{n} b \left( \sum_{t=1}^n X_t \right) \left( \sum_{t=1}^n X_t \right) - b \sum_{t=1}^n X_t^2 = 0$$

$$b = \frac{n \sum_{t=1}^n X_t Y_t - \left( \sum_{t=1}^n X_t \right) \left( \sum_{t=1}^n Y_t \right)}{n \sum_{t=1}^n X_t^2 - \left( \sum_{t=1}^n X_t \right)^2}$$

# Anscombe négyese

- Legjobban illeszkedő egyenes mindenre van
- Minőségileg különböző adatpontokra is



# Lineáris regresszió: általános alak

- $y = \Theta^T x$
- Solve:  $\Theta^* = \min_{\Theta} \sum_{i=1}^m (\Theta^T x_i - y_i)$
- $X \in \mathbb{R}^{m \times n}$ : „design matrix” - tanító-minták a sorok
- $\vec{y} = [y_1, \dots, y_m]^m$ : „target labels”
- $\Theta^* = (X^T X)^{-1} X^T \vec{y}$

# „Summation form”

- $\Theta^* = (X^T X)^{-1} X^T \vec{y}$
- $\Theta^* = A^{-1} b$
- $A = \sum_{i=1}^m x_i y_i^T$  és  $b = \sum_{i=1}^m x_i y_i$
- ... és ezek a szummák már párhuzamosíthatóak m-ben.

# Asszociációs szabályok tanulása

$$R1: \quad [X_1, X_2, \dots, X_n] \Rightarrow Y$$

Szabály: HA (feltétel)  $\Rightarrow$  AKKOR (következmény)  
(törzs – body) (fej – head)

- asszociációs kapcsolatok (szabályok) megismerésére irányul
- Lényege: ha egy **S** elemhalmazban szerepel egy (vagy több) adott elem  $X_1, X_2, \dots, X_n$  akkor egy másik  $Y$  elem is nagy valószínűséggel szerepel benne.

# Szekvenciális szabályok tanulása

$$\underbrace{[X_1, X_2, \dots, X_n]}_{\substack{t_1:S_1 \\ t_1:S_2}} + \underbrace{[Z_1, Z_2, \dots, Z_n]}_{\substack{t_2:S_1 \\ t_2:S_2}} + \underbrace{[\dots]}_{\substack{t_k:S_1 \\ t_k:S_2}} \rightarrow \underbrace{Y}_{\substack{t_{k+1}:S_1 \\ t_{k+1}:S_2}}$$

Elemhalmaz szekvencia  
(törzs – body)

→ következmény  
(fej – head)

- asszociációs kapcsolatok megismerésére irányul több eltérő (időí) csoportba tartozó elemhalmaz között
- Lényege: ha a  $t_1$  (időí) csoportba tartozó **S** elemhalmazban és  $t_2 \dots t_k$  (időí) csoportba tartozó **S** elemhalmazban szerepel egy (vagy több) adott elem  $X_1, X_2, \dots$ , akkor egy  $Y$  elem nagy valószínűséggel szerepel egy  $t_{k+1}$  (időí) csoportba tartozó **S** halmazban.

# Asszociációs vs Szekvenciális szabálytanulás

## Asszociációs szabálytanulás

- Célja az tranzakciók (elemhalmazok) elemzése
- Gyakori elemhalmazok (itemset), gyakran együtt előforduló elemek vizsgálata

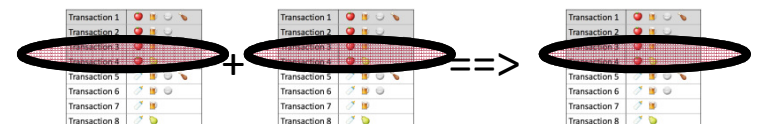
Pl.: vásárlói kosarak elemzése

|               |                                                                                                                                                                                                                                                                                                                                                 |
|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Transaction 1 |     |
| Transaction 2 |                                                                                        |
| Transaction 3 |                                                                                                                                                                           |
| Transaction 4 |                                                                                                                                                                           |
| Transaction 5 |     |
| Transaction 6 |                                                                                        |
| Transaction 7 |                                                                                                                                                                           |
| Transaction 8 |                                                                                                                                                                           |

## Szekvenciális szabálytanulás

- az idői dimenziót is figyelembe veszi (a halmazok csoportosításának alapját)
- célja az összetartozó tranzakciók láncolatának elemzése, időbeli kapcsolatok azonosítása

Pl.: meghibásodás elemzés (meghibásodások időbeli láncolatának elemzésével)





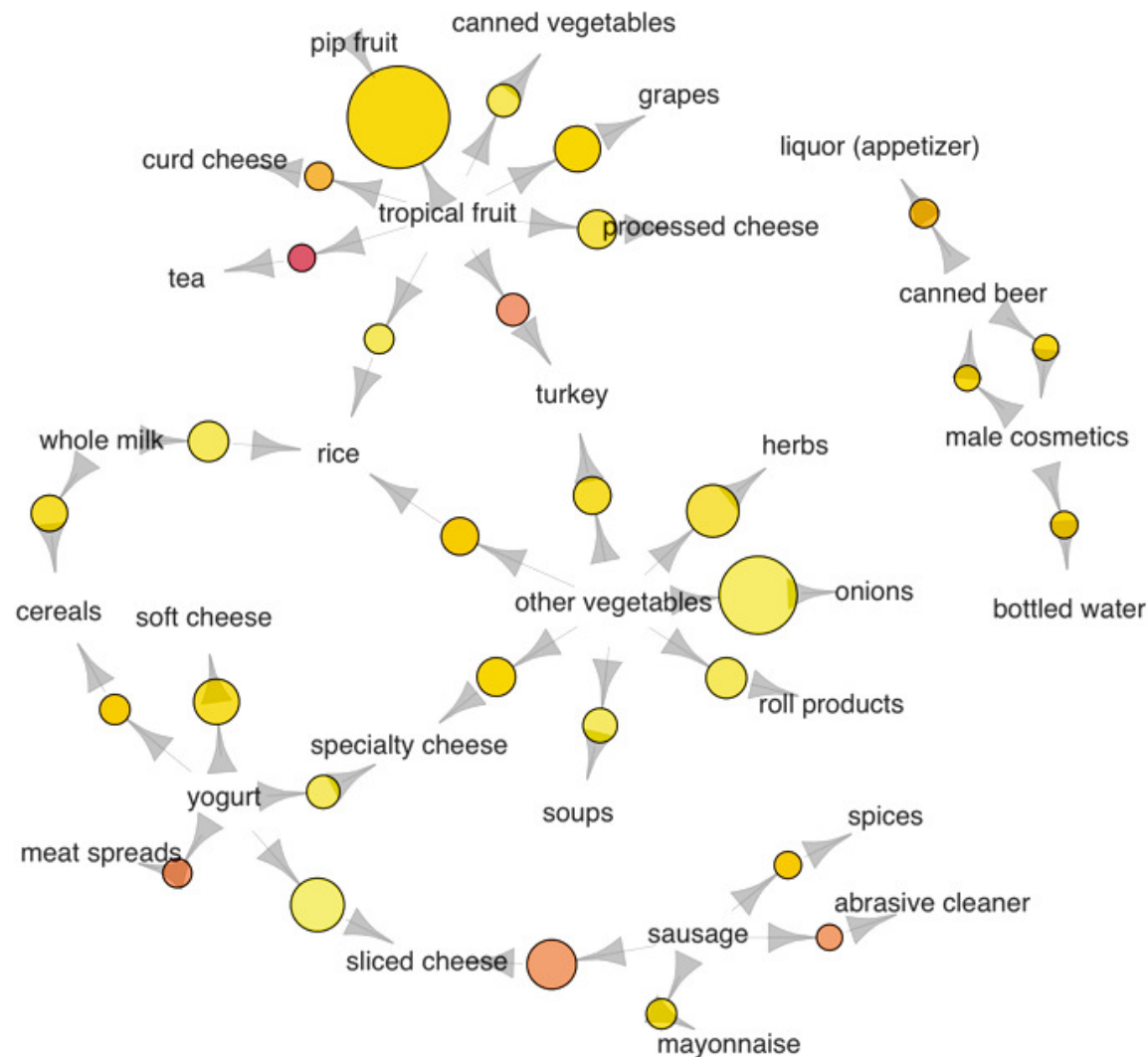
# Szabályok jellemzése - 1

- **Maximális szabályhossz** (rule length): a szabálytörzsben lévő elemek száma + a fej
- **Támogatottság (support)**: egy adott szabály a rekordok mekkora részében igaz (%-os megadás: relatív szupport, darabszám: abszolút szupport)
- **Bizonyosság(confidence)**: a rekordok mekkora részében igaz az, hogy ha az adott szabálytörzs igaz, akkor ott a szabály fejrésze is igaz:  $P(Y|X)$
- **Lift**: a szabály által leírt elemek megfigyelt együttes gyakoriságának az elvárt gyakorisághoz mért aránya:  $P(Y, X) / (P(Y)P(X))$

## Szabályok jellemzése - 2

- **Lift**: a szabály által leírt elemek megfigyelt együttes gyakoriságának az elvárt gyakorisághoz mért aránya:  
 $P(Y, X) / (P(Y)P(X))$
- **Lift>1** azt jelenti, hogy a két elem kiegészíti egymást, tehát az elemhalmaz nagyobb valószínűséggel tartalmaz 'Y'-t, ha van benne 'X' is.
- **Lift<1** azt jelenti, hogy a két elem helyettesíti egymást, tehát ha 'X' szerepel az elemhalmazban, akkor 'Y' kisebb valószínűséggel fog szerepelni.
- **Lift=1** jelentése az, hogy a 'X' nem befolyásolja 'Y'-t, vagyis attól, hogy 'X' benne van az elemhalmazban, nem változik annak a valószínűsége, hogy 'Y' szerepel-e benne. Tehát 'X' és 'Y' függetlenek.

# Szabályok jellemzése - 3



Kör nagysága:  
support

Kör színe: lift

- Illusztráció: <http://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html>

# Szabályok jellemzése - 4

- Asszociációs szabály  $\neq$  Oksági szabály !!

Ha X és Y korrelációban lévő események, akkor vagy

- X okozza Y-t
  - Y okozza X-et
  - Z okozza X-et és Y-t (közös ok)
- 
- Két változó közötti asszociáció szimmetrikus reláció, nincs iránya
  - Az oksági relációnak van iránya, ok  $\rightarrow$  okozat
  - Asszociációból nem következik oksági reláció

# Szabálytanulás - példa

## Walmart adatbázis vizsgálata

- 3600 üzleti egység
- 100 Millió vásárló
- 460 TB !! a teljes adatvagyon mérete
- Kérdés: Mit vásárolnak leginkább az emberek hurrikán veszély esetén (mivel töltsék fel a boltokat)?
- Ismert, hogy ásványvizet, elemet, elemlámpát, sört vesznek ilyenkor. Mit még?

<http://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html>

# Szabálytanulás - példa

## Walmart adatbázis vizsgálata

8-szor gyakrabban fordult elő hurrikán veszély esetén a vásárlói kosarakban:

**Strawberry pop-tart**



Kép: <https://commons.wikimedia.org/wiki/File:Rasp-Pop-Tart.jpg>

# Kollaboratív szűrés (Collaboratív filtering)



Books

Departments  Your Amazon.com Today's Deals Gift Cards & Registry Sell Help

1-12 of 155,767 results for Books : "order"

Show results for

< Any Category

Books

Thrillers & Suspense (1,441)

Suspense (543)

Mystery, Thriller & Suspense (1,878)

Political (137)

Military (95)

Espionage (127)

Literature & Fiction (10,076)

+ See more

Refine by

International Shipping

☐ Ship to Hungary

Amazon Prime

☐ Prime

Eligible for Free Shipping

☐ Free Shipping by Amazon

New Releases

Last 30 days (1,277)

Last 90 days (3,613)

Coming Soon (166)

Author

Vince Flynn

Kyle Mills

Michael A. Stackpole

Walter Jon Williams

Philippa Gregory

James Luceno

James Rollins

Rebecca Cantrell

Sean Williams

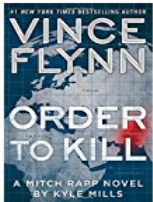
Nicola Barber

Stephanie Rowe



Your Orders – View, Modify, Cancel or Track an Order

[Recent Orders](#) | [Open Orders](#) | [Digital Orders](#)



Order to Kill: A Novel (A Mitch Rapp Novel)

Oct 11, 2016

by Vince Flynn and Kyle Mills

Hardcover

~~\$17.39~~ ~~\$28.99~~ Prime

Get it by **Friday, Nov 11**

More Buying Choices

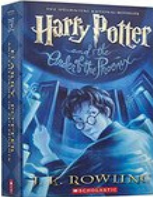
**\$5.97** used & new (94 offers)

Kindle Edition

**\$19.04**

Whispersync for Voice-ready

Other Formats: [Paperback](#), [Audible Audio Edition](#), [Mass Market Paperback](#), [Audio CD](#)



Harry Potter And The Order Of The Phoenix

Aug 10, 2004

by J.K. Rowling

Paperback

~~\$7.86~~ ~~\$12.99~~ Prime

Get it by **Friday, Nov 11**

More Buying Choices

**\$1.98** used & new (441 offers)

Kindle Edition

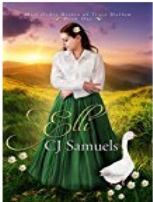
**\$0.00** kindleunlimited

Read this and over 1 million books with [Kindle Unlimited](#).

**\$10.12** to buy

Auto-delivered wirelessly

Other Formats: [Hardcover](#), [Audible Audio Edition](#), [Mass Market Paperback](#), [Audio CD](#)



Elli: Mail Order Brides of Trace Hollow Book 1

Nov 4, 2016

by CJ Samuels

Kindle Edition

**\$0.00** kindleunlimited

Read this and over 1 million books with [Kindle Unlimited](#).

**\$1.26** to buy

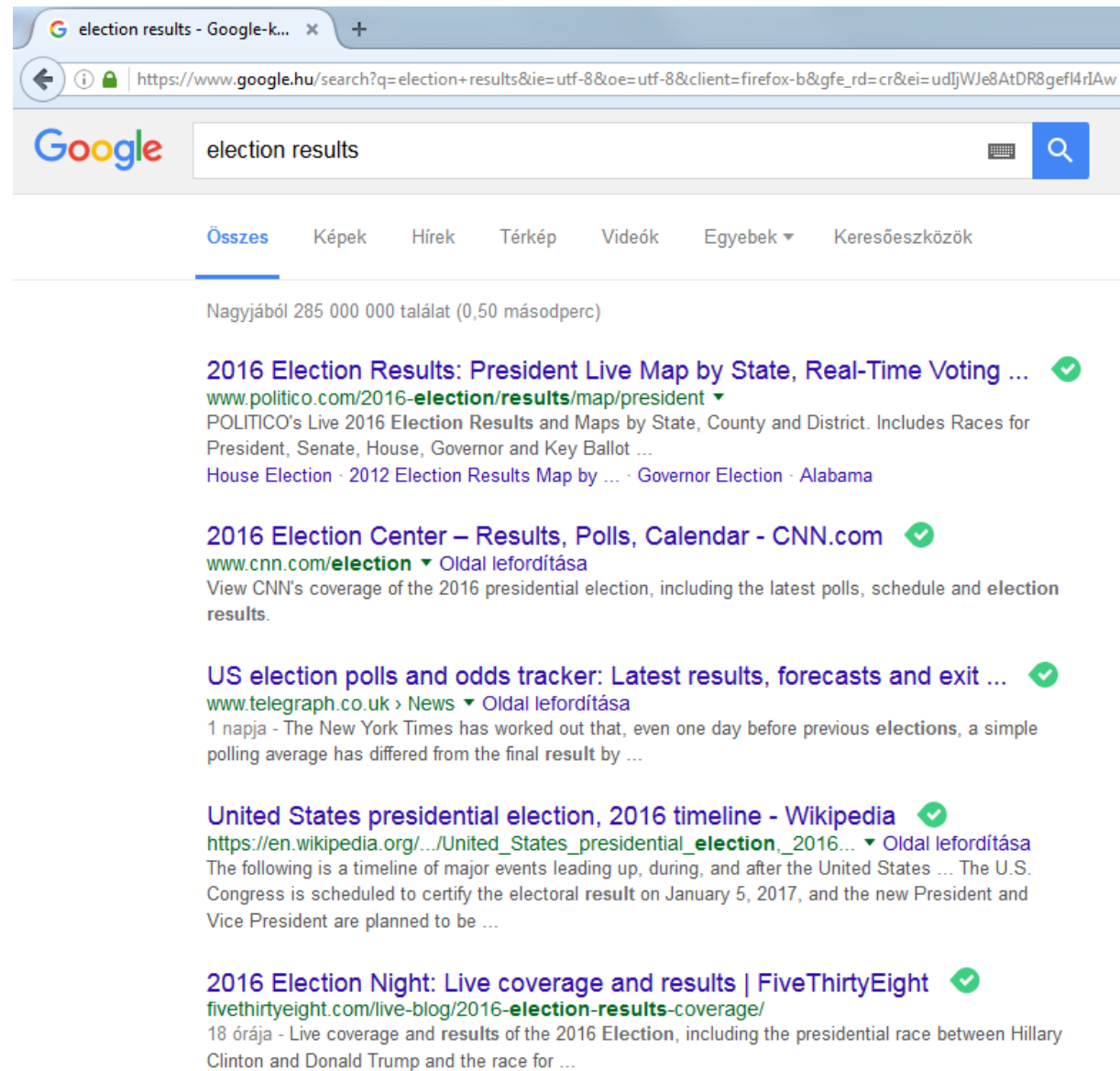
Auto-delivered wirelessly

MŰEGYETEM 1782

FTES RG



# Kollaboratív szűrés (Collaboratív filtering)



The screenshot shows a Google search interface with the query "election results". The search results are displayed in a list format. The first result is from POLITICO, titled "2016 Election Results: President Live Map by State, Real-Time Voting ...". The second result is from CNN, titled "2016 Election Center – Results, Polls, Calendar - CNN.com". The third result is from The Telegraph, titled "US election polls and odds tracker: Latest results, forecasts and exit ...". The fourth result is from Wikipedia, titled "United States presidential election, 2016 timeline - Wikipedia". The fifth result is from FiveThirtyEight, titled "2016 Election Night: Live coverage and results | FiveThirtyEight". Each result includes a green checkmark icon, indicating it has been verified or approved.

election results - Google-k... x +

https://www.google.hu/search?q=election+results&ie=utf-8&oe=utf-8&client=firefox-b&gfe\_rd=cr&ei=udljWJe8AtDR8gefI4rIAw

Google election results

Összes Képek Hírek Térkép Videók Egyebek ▼ Keresőeszközök

Nagyjából 285 000 000 találat (0,50 másodperc)

**2016 Election Results: President Live Map by State, Real-Time Voting ...** ✓  
[www.politico.com/2016-election/results/map/president](http://www.politico.com/2016-election/results/map/president) ▼  
POLITICO's Live 2016 Election Results and Maps by State, County and District. Includes Races for President, Senate, House, Governor and Key Ballot ...  
[House Election](#) · [2012 Election Results Map by ...](#) · [Governor Election](#) · [Alabama](#)

**2016 Election Center – Results, Polls, Calendar - CNN.com** ✓  
[www.cnn.com/election](http://www.cnn.com/election) ▼ Oldal lefordítása  
View CNN's coverage of the 2016 presidential election, including the latest polls, schedule and election results.

**US election polls and odds tracker: Latest results, forecasts and exit ...** ✓  
[www.telegraph.co.uk](http://www.telegraph.co.uk) > News ▼ Oldal lefordítása  
1 napja - The New York Times has worked out that, even one day before previous elections, a simple polling average has differed from the final result by ...

**United States presidential election, 2016 timeline - Wikipedia** ✓  
[https://en.wikipedia.org/.../United\\_States\\_presidential\\_election\\_2016...](https://en.wikipedia.org/.../United_States_presidential_election_2016...) ▼ Oldal lefordítása  
The following is a timeline of major events leading up, during, and after the United States ... The U.S. Congress is scheduled to certify the electoral result on January 5, 2017, and the new President and Vice President are planned to be ...

**2016 Election Night: Live coverage and results | FiveThirtyEight** ✓  
[fivethirtyeight.com/live-blog/2016-election-results-coverage/](http://fivethirtyeight.com/live-blog/2016-election-results-coverage/)  
18 órája - Live coverage and results of the 2016 Election, including the presidential race between Hillary Clinton and Donald Trump and the race for ...



# Ajánlórendszerek - 1

- Felhasználói profil hasonlósága alapján
  - lehetne klaszterezéssel
  - keressen a rendszer hasonló felhasználót
  - az ő választásai alapján adjon ajánlást
- Problémák:
  - klaszterszám
  - számítási igény
  - mi alapján lesz hasonló?
  - van-e elég információ ehhez?

# Ajánlórendszerek - 2

- Együtt előforduló elemek (vásárlói kosarak) alapján
  - Pl.: online áruház termékajánló rendszere
  - megoldható szabálytanulással
  - Problémák:
    - számítási igény
    - így nem lesz személyre szabott



<https://sarahtianhua.wordpress.com/portfolio/market-basket-analysis/>  
<http://thecentsiblefamily.com/category/stores/market-basket/>

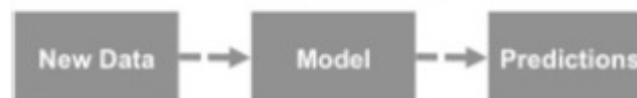
# Ajánlórendszerek – CF

- Kollaboratív szűrés
  - Preferencia információ sok felhasználótól
  - Feltételezés: akik hasonló dolgokat kedveltek a múltban, a jövőben is így tesznek majd
  - Vannak elemek, melyekről van preferencia megadva
  - Sok esetben nincs megadva

Ted and Carol like movies B and C



Bob likes movie B, what might he like?



Bob likes movie B, predict C

User Item Rating Matrix

|       | A | B | C |
|-------|---|---|---|
| Ted   | 4 | 5 | 5 |
| Carol |   | 5 | 5 |
| Bob   |   | 5 | ? |

<https://www.mapr.com/blog/apache-spark-machine-learning-tutorial>

# Ajánlórendszerek – CF-2

- Kérdés: Hogyan döntenének a felhasználók a még nem értékelt elemekről?
- Ez kellene ahhoz, hogy az értékelések alapján másoknak ajánlásokat adhassunk
- Mátrix alapú reprezentáció (ez sokszor hiányos)
- Cél: mátrix „kitöltése”
- Módszer: mátrix faktorizáció

|    | D1 | D2 | D3 | D4 |
|----|----|----|----|----|
| U1 | 5  | 3  | -  | 1  |
| U2 | 4  | -  | -  | 1  |
| U3 | 1  | 1  | -  | 5  |
| U4 | 1  | -  | -  | 4  |
| U5 | -  | 1  | 5  | 4  |

# Mátrix faktorizáció -1

## Mátrix faktorizáció alkalmazása

- Feltételezés: a háttérben léteznek nem ismert jegyek **K** (features)
- jegyek alapján értékelnek a felhasználók: **U**
- jegyek száma jóval kisebb a felhasználókénál : **K << U**
- elemek (pl.: filmek): **D**
- Preferencia információ (rating mátrix) : **R**

|    | D1 | D2 | D3 | D4 |
|----|----|----|----|----|
| U1 | 5  | 3  | -  | 1  |
| U2 | 4  | -  | -  | 1  |
| U3 | 1  | 1  | -  | 5  |
| U4 | 1  | -  | -  | 4  |
| U5 | -  | 1  | 5  | 4  |

# Mátrix faktorizáció -2

A cél megtalálni 2 mátrixot:

- $P = U \times K$  ~ felhasználók és jegyek közötti asszociáció
- $Q = D \times K$  ~ elemek és jegyek közötti asszociáció

melyekre igaz, hogy:

$$P \times Q^T = R^{\wedge} \text{ (} R \text{ közelítése)}$$

Az  $u_i$  felhasználó által a  $d_j$  elemre adott rating :  $r_{i,j}$

$$\hat{r}_{ij} = p_i^T q_j = \sum_{k=1}^K p_{ik} q_{kj}$$

$P$  és  $Q$  „megtalálásához” két (valamilyen módon inicializált) mátrix szorzatát kell összevetni az eredeti  $R$  mátrixszal és törekedni iteratívan az eltérés csökkenésére.

Eltérés:

$$e_{ij}^2 = (r_{ij} - \hat{r}_{ij})^2 = (r_{ij} - \sum_{k=1}^K p_{ik} q_{kj})^2$$

# Mátrix faktORIZÁCIÓ -3

- Gradiens számítás  $p_{ik}$  és  $q_{kj}$  módosításához

$$\begin{aligned}\frac{\partial}{\partial p_{ik}} e_{ij}^2 &= -2(r_{ij} - \hat{r}_{ij})(q_{kj}) = -2e_{ij}q_{kj} \\ \frac{\partial}{\partial q_{kj}} e_{ij}^2 &= -2(r_{ij} - \hat{r}_{ij})(p_{ik}) = -2e_{ij}p_{ik}\end{aligned}$$

- $p_{ik}$  és  $q_{kj}$  módosítása a gradiens alapján

$$\begin{aligned}p'_{ik} &= p_{ik} + \alpha \frac{\partial}{\partial p_{ik}} e_{ij}^2 = p_{ik} + 2\alpha e_{ij}q_{kj} \\ q'_{kj} &= q_{kj} + \alpha \frac{\partial}{\partial q_{kj}} e_{ij}^2 = q_{kj} + 2\alpha e_{ij}p_{ik}\end{aligned}$$

- Leállni akkor célszerű, amikor a teljes hiba ( $E$ ) minimális

$$E = \sum_{(u_i, d_j, r_{ij}) \in T} e_{ij} = \sum_{(u_i, d_j, r_{ij}) \in T} (r_{ij} - \sum_{k=1}^K p_{ik}q_{kj})^2$$

# Mátrix faktorizáció -4

- Találtunk egy olyan  $\mathbf{P}$  és  $\mathbf{Q}$  mátrixot, ami jól közelíti  $\mathbf{R}$  ismert részeit
- A nem ismert részeire pedig ad egy becslést
- Így már  $\mathbf{R}$  használható ajánláshoz