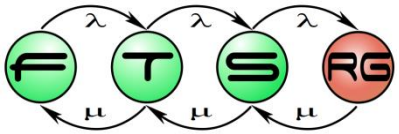
Az adatelemzés alapfeladatai

2017 ősz 6./7. alkalom
Kocsis Imre
ikocsis@mit.bme.hu

Budapesti Műszaki és Gazdaságtudományi Egyetem
Hibatűrő Rendszerek Kutatócsoport



MACHINE LEARNING / DATA MINING: ALAPFELADATOK

Adatelemzés: legfontosabb problémák

Csoportosítás (***clustering***)

Osztályozás (***classification***)

Asszoc. szabályok (***assoc. rules***)

Regresszió (***regression***)

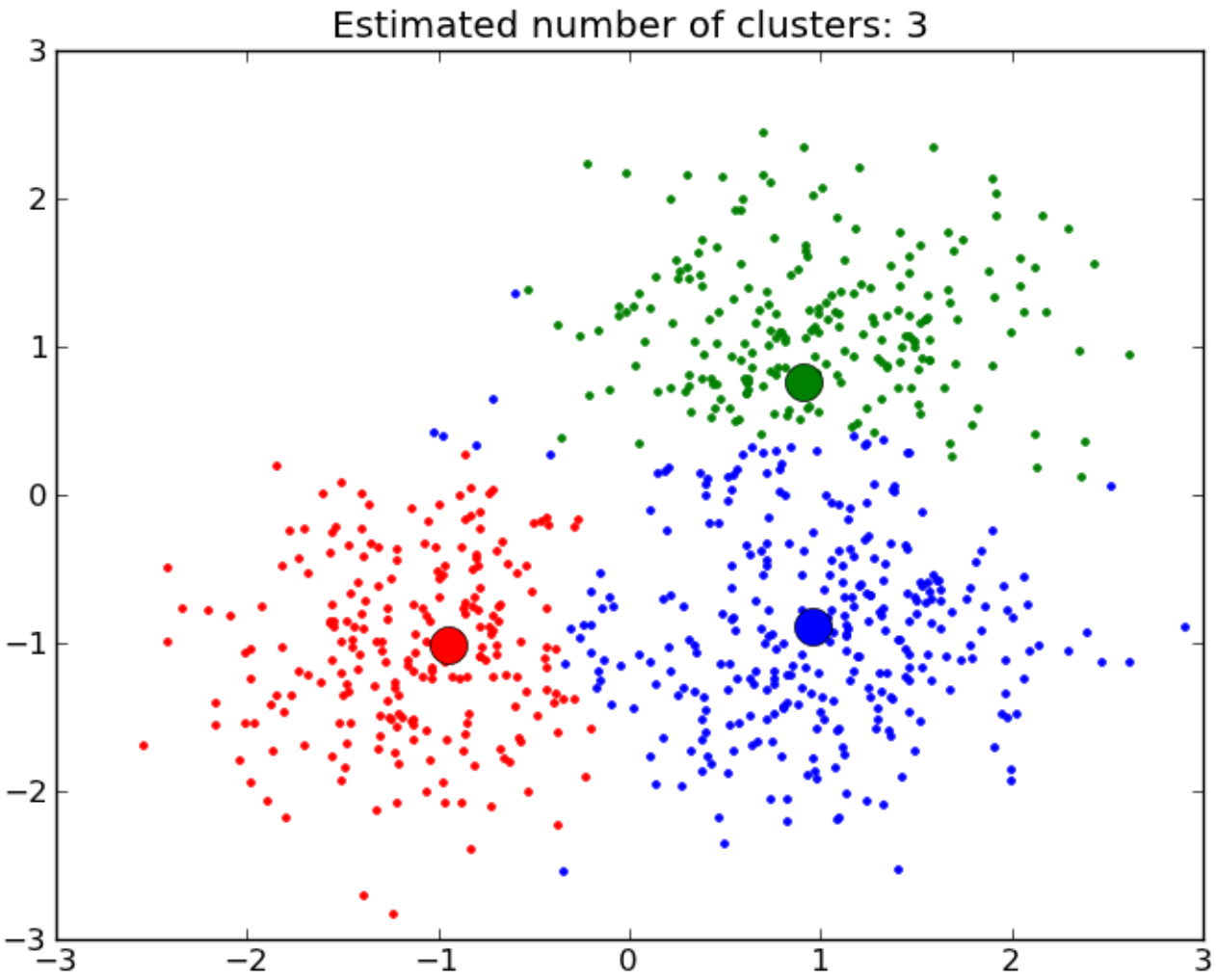
Megközelítés

- Probléma-osztály;
- Szemléltetés;
- (Egy algoritmus kivonata)

- Fő cél: orientáció
- ~~Deep learning, TensorFlow, ...~~

- N.B. egyre kevésbé „kézműves” tevékenység
  - „10 Algorithms every Data Scientist has to know”
  - SaaS/PaaS
  - Lásd később

Csoportosítás



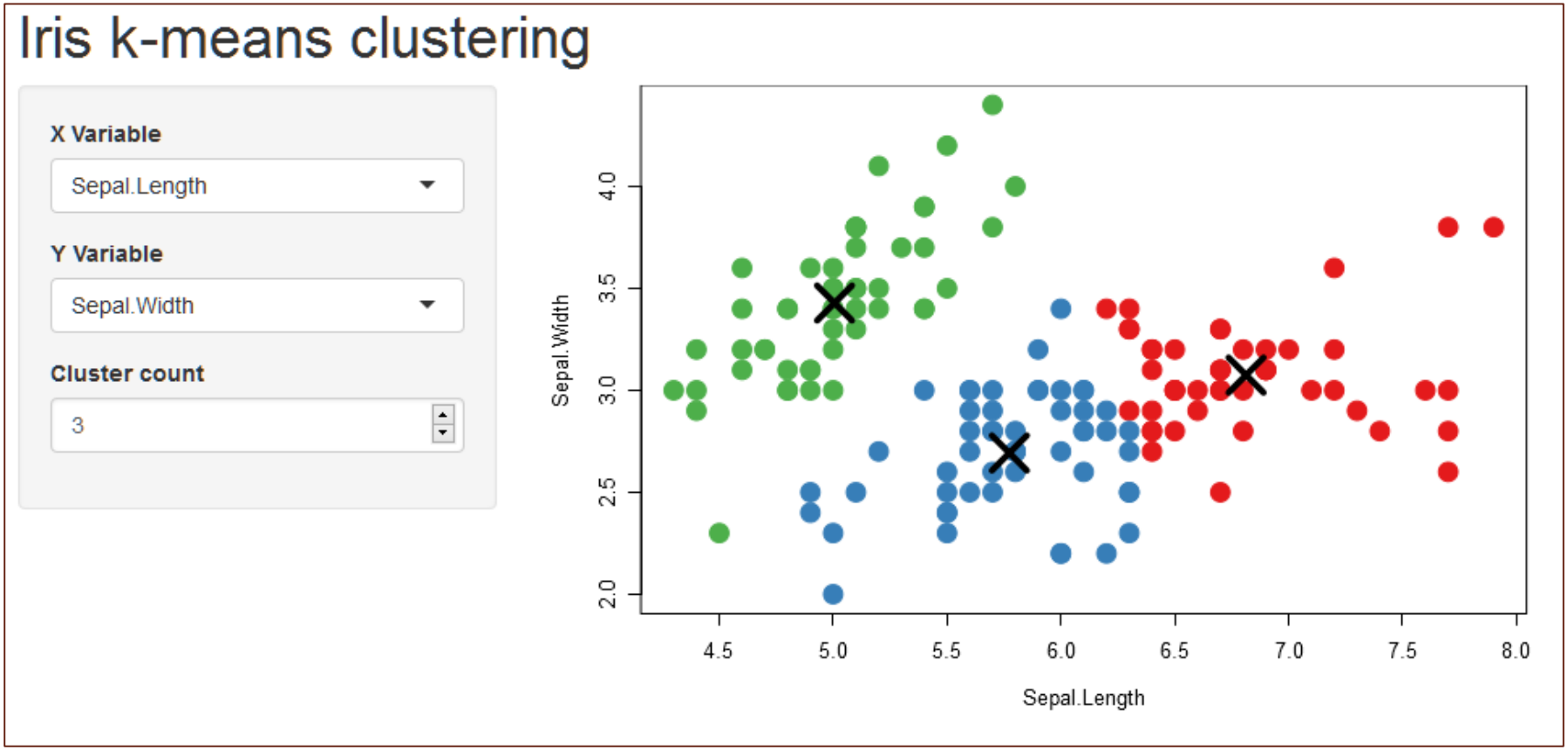
K-means

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

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

Demo

- https://github.com/rstudio/shiny-examples/tree/master/050-kmeans-example



Egy megoldás

$\{r(C_1), r(C_2), \ldots, r(C_k)\} \leftarrow$ repr. kezdeti halmaza
while $r(C_i)$ változik
do for $\forall u \in D$ adott sorrendben
do
$h \leftarrow u$ klaszter-indexe
$j \leftarrow argmin_i d(u, r(C_i))$
if $h \neq j$ **then** {
$C_j \leftarrow C_j \cup \{u\}$
$C_i \leftarrow C_i \setminus \{u\}$
$r(C_j) \leftarrow \frac{1}{|C_j|} \sum_{v \in C_j} v$
$r(C_h) \leftarrow \frac{1}{|C_h|} \sum_{v \in C_h} v\}$
return C

Régi klaszter

Új klaszter

Itt rögtön újra is számoljuk

Alapkérdések

- Klasztereken belül maximális homogenitás
  - „nagy hasonlóság”
  - „kis távolság”
- Klaszterek között „nagy távolság”
  - „kis hasonlóság” különböző klaszterek elemei között
- Hasonlósági mérce: „similarity metric”
  - Kategorikus változókra nehéz
  - Skálatranszformáció kellhet
  - Lehet választani...
  - Hierarchikus klaszterezés: dissimilarity metric
- Hasonlósági küszöb választása?
- Optimális klaszterszám?

Néhány távolság-mérték

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$

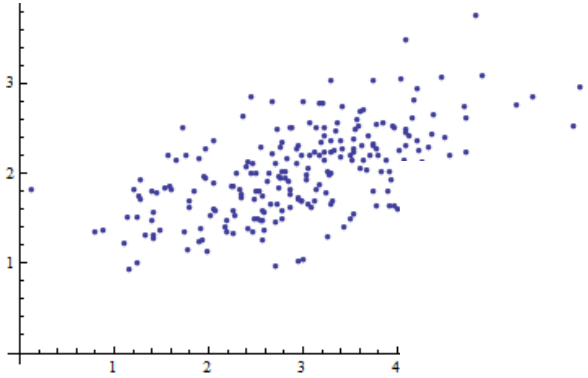
- (0,0) és (1,1) távolsága?

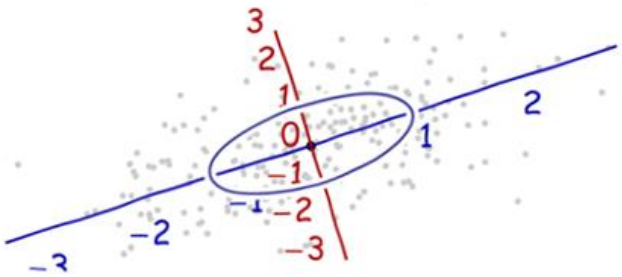
Mahalanobis távolság…?

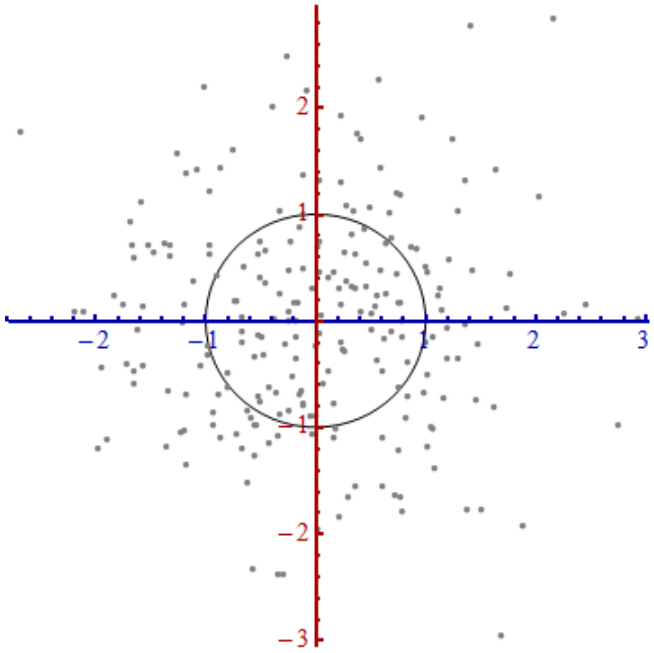
- Pont és eloszlás távolsága (S: kov-mátrix)

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

- “Szemléletesen”:



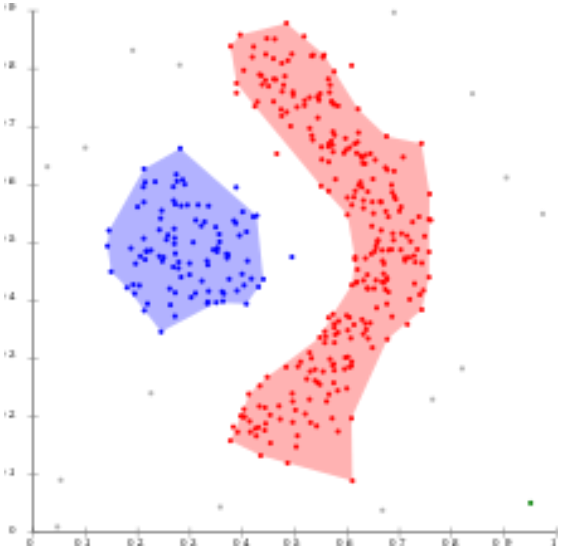




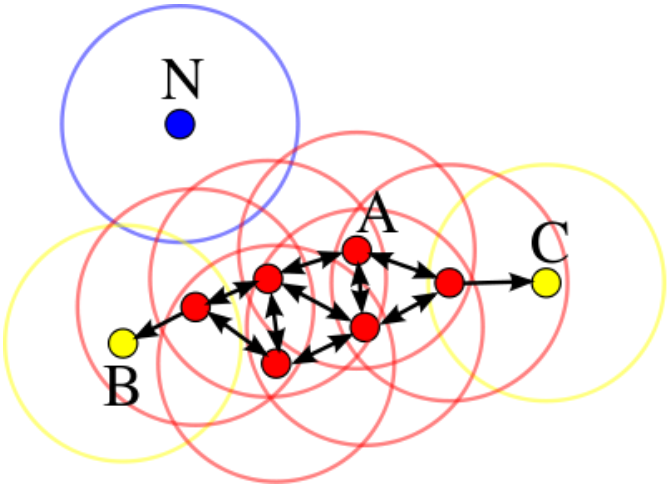
https://stats.stackexchange.com/questions/62092/bottom-to-top-explanation-of-the-mahalanobis-distance

További változatok

- Lineárisan nem szeparálható klaszterek: sűrűség alapú klaszterezés

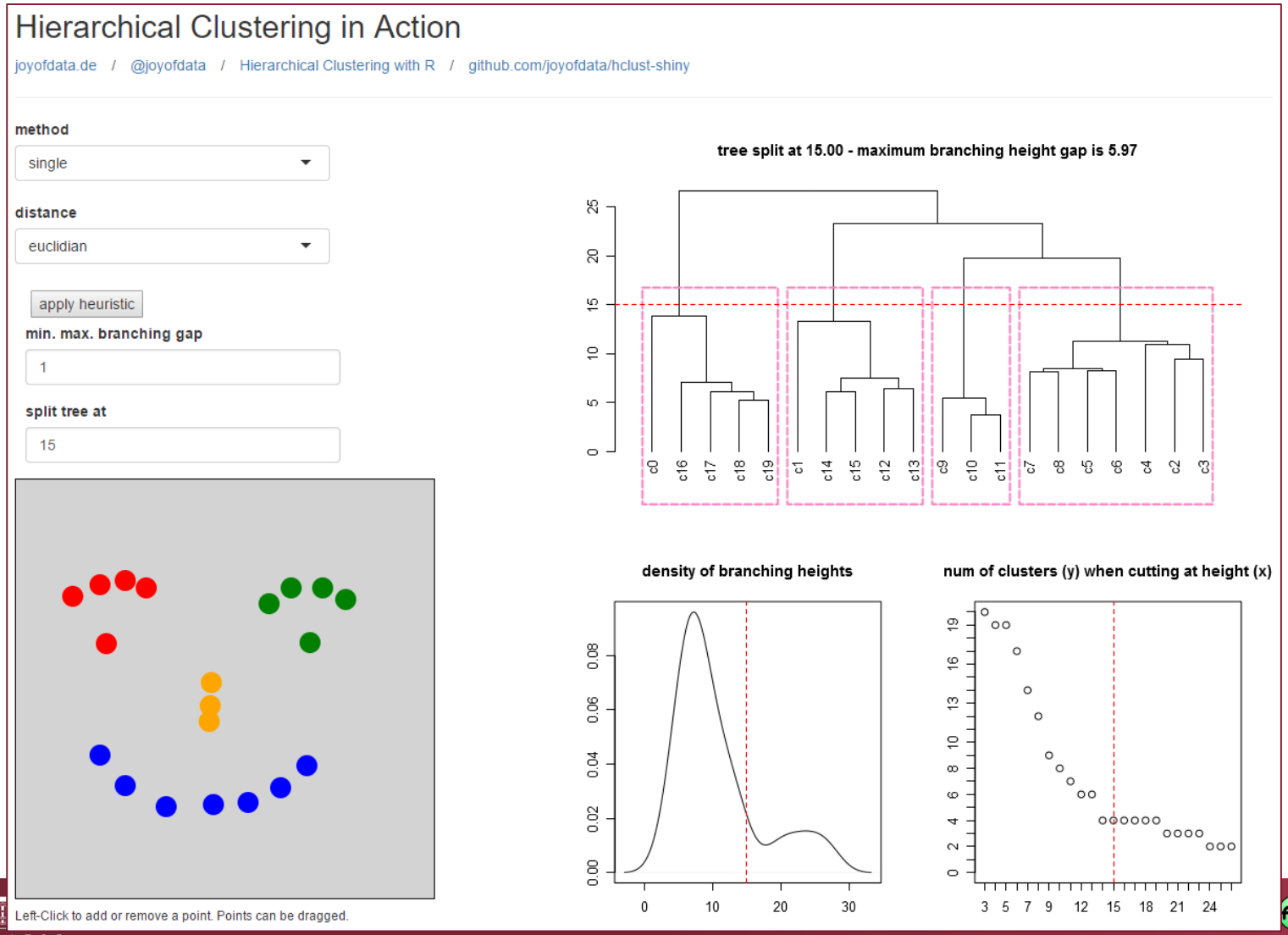


- “packing together closely grouped points”
- Pl. DBSCAN (“Density-based spatial clustering of applications with noise”)
  - Magpontok: legalább minPts pont e távolságon belül
  - Sűrűség-elérhető pontok
  - Outlierek



(Agglomeratív) hierarchikus klaszterezés

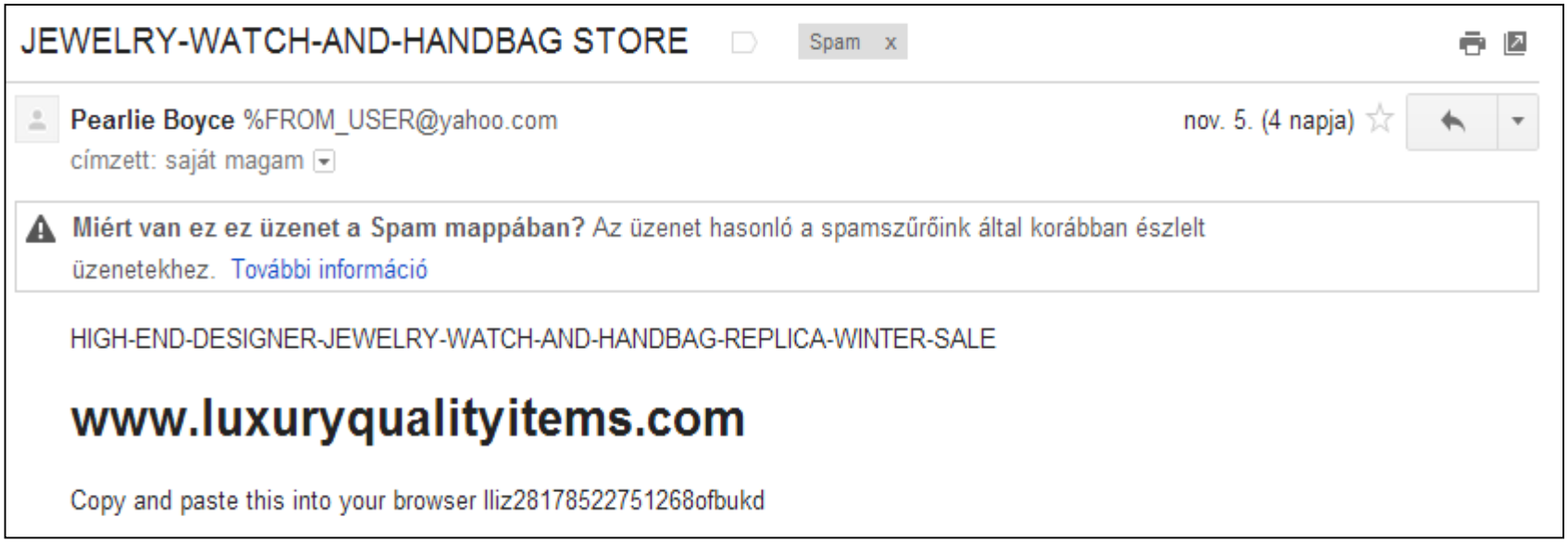
- CSAK KITEKINTÉS
- https://github.com/joyofdata/hclust-shiny



Osztályozás

Képosztályozás: a képen látható objektum madár vagy repülő?

Osztályozás


JEWELRY-WATCH-AND-HANDBAG STORE

Spam x

Pearlie Boyce %FROM_USER@yahoo.com

címzett: saját magam

nov. 5. (4 napja)

Miért van ez ez üzenet a Spam mappában? Az üzenet hasonló a spamszűrőink által korábban észlelt üzenetekhez. További információ

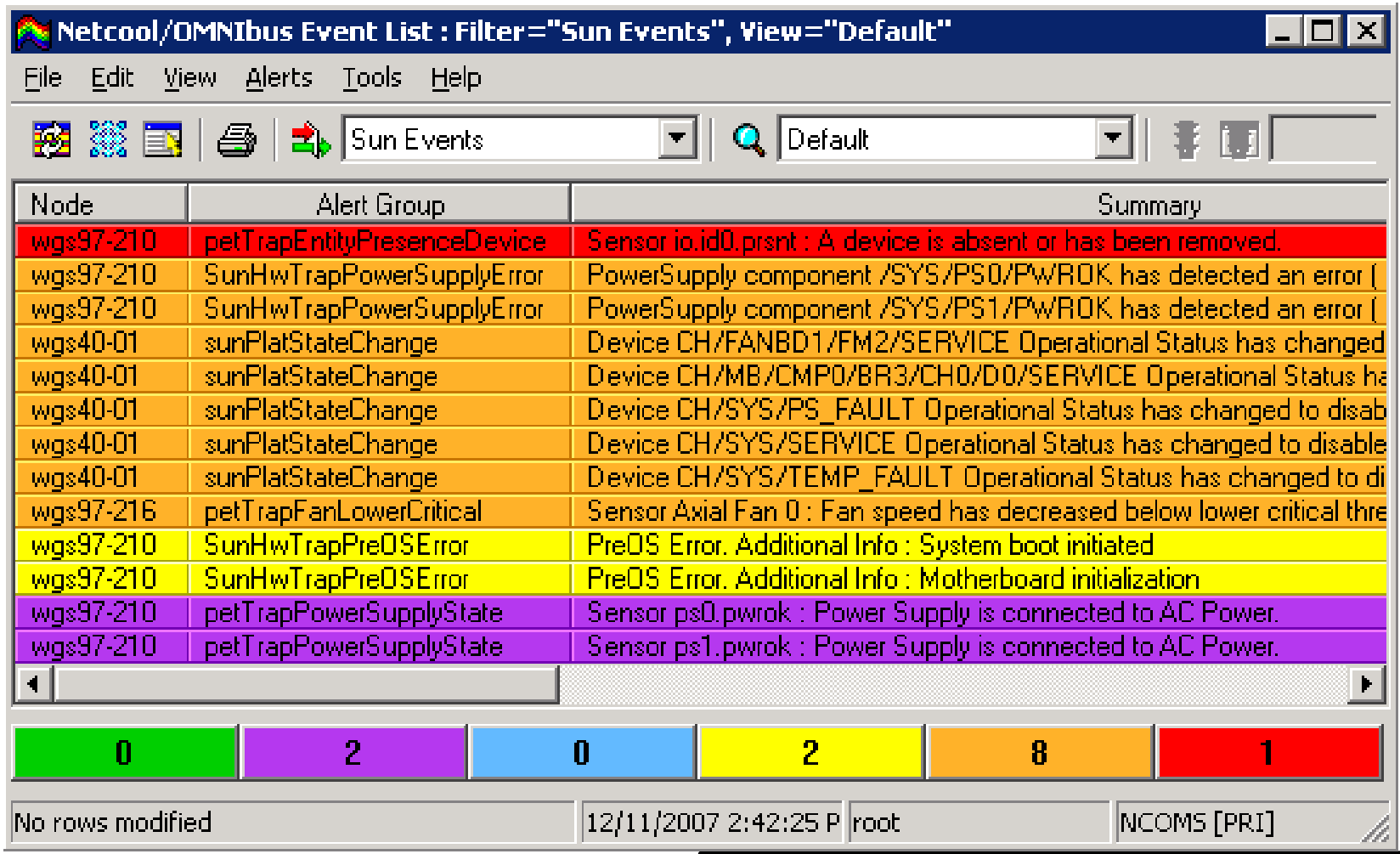
HIGH-END-DESIGNER-JEWELRY-WATCH-AND-HANDBAG-REPLICA-WINTER-SALE

www.luxuryqualityitems.com

Copy and paste this into your browser lliz28178522751268ofbukd

Levelek osztályozása: SPAM vagy nem SPAM?

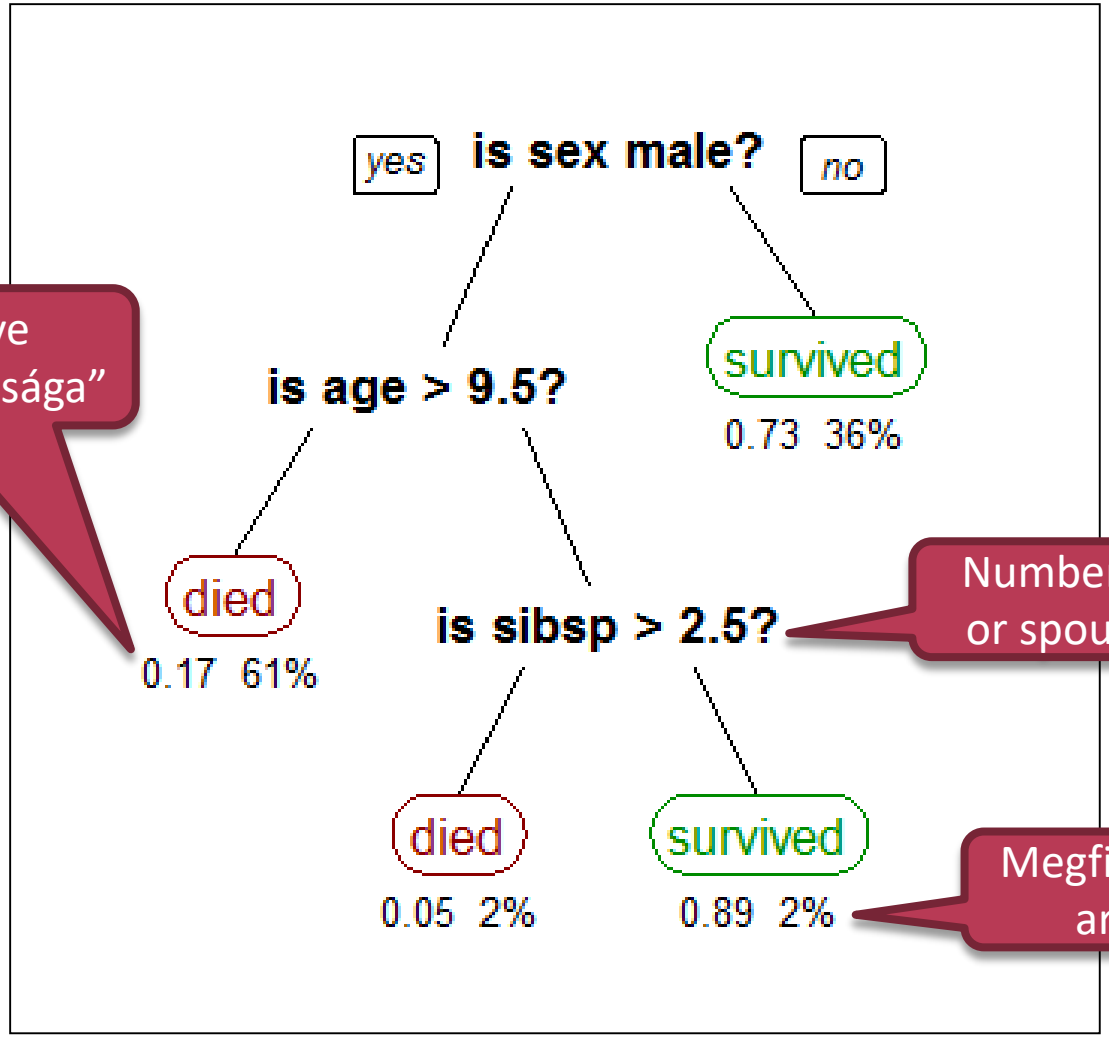
Osztályozás



Szabályok alapján Severity osztályozása

Kép forrása: http://192.9.172.90/bigadmin/features/articles/3pmi_mgmt.full.jsp

Döntési fák (Titanic, túlélési esélyek)



Túlélés esélye
-> osztály “tisztasága”

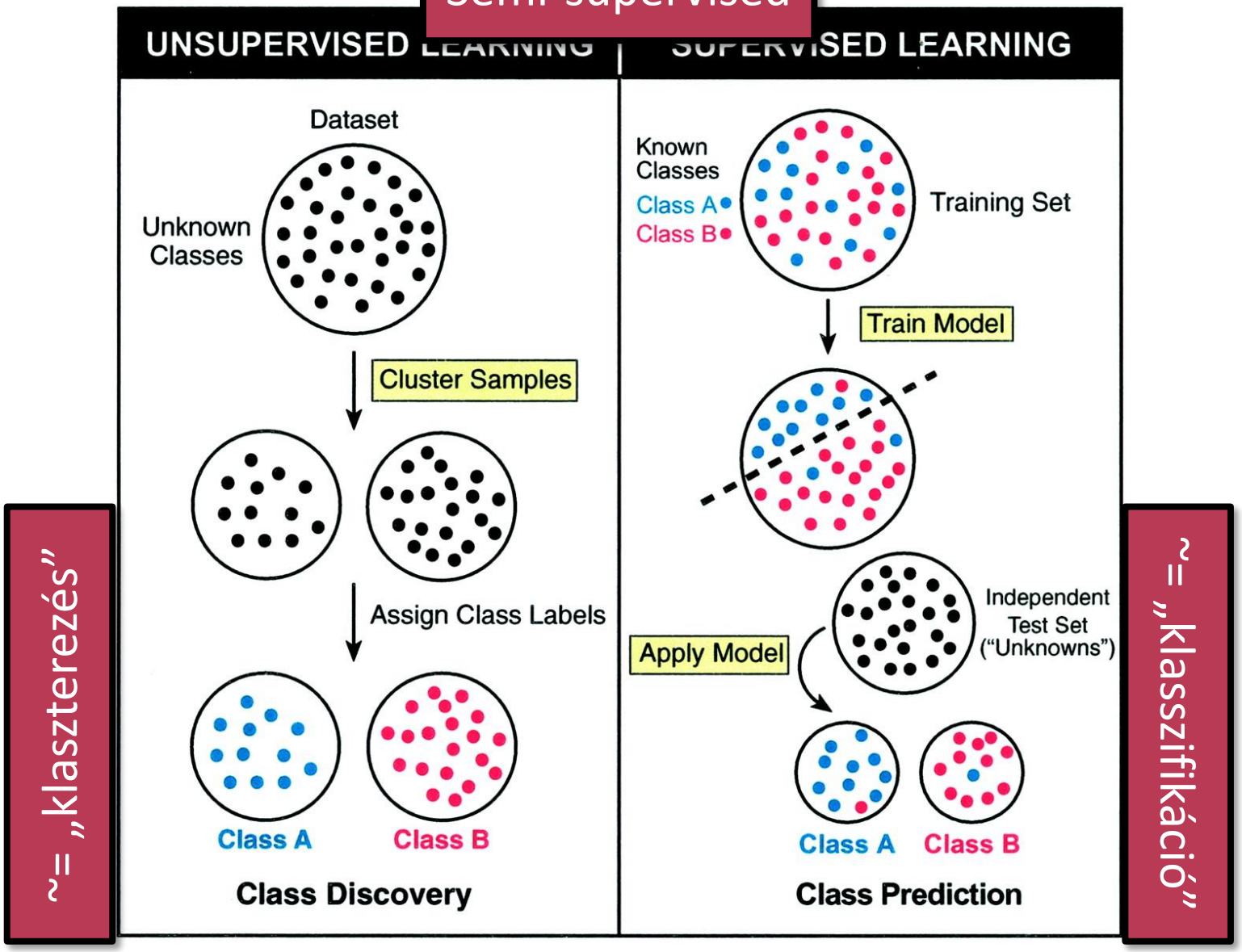
Number of siblings or spouses aboard

Megfigyelések aránya

https://en.wikipedia.org/wiki/Decision_tree_learning

Klaszterezés és klasszifikáció alapfeladat

Semi-supervised



~= „klaszterezés”

~= „klasszifikáció”

Kép forrása: Ramaswamy S , Golub T R JCO 2002;20:1932-1941

Felügyelt és nem felügyelt tanulás

- Felügyelt tanulás
  - Adott néhány pontra az elvárt kimenet is
  - ≈ a tanuló példákból való általánosítás
  - Output: függvény
    - a meglévő mintapontokra jól képez le
    - megfelelően általánosítható

Tanulóhalmaz: amin építjük a modellt
Teszthalmaz: amin ellenőrizzük

- Nem felügyelt tanulás
  - Nincs meg az elvárt kimenet
  - Visszajelzés nélkül építi a modellt
  - ≈ szabályok, összefüggések keresése (ismeretfeltárás)

Demo

- R „party” csomag
- Conditional Inference Tree, iris
  - Hogy ez mi, azt nem kezdjük itt nagyon részletezni
  - Party ctree dokumentáció: “*Roughly, the algorithm works as follows:*
    - *1) Test the global null hypothesis of independence between any of the input variables and the response (which may be multivariate as well).*
    - *Stop if this hypothesis cannot be rejected.*
    - *Otherwise select the input variable with strongest association to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response.*
    - *2) Implement a binary split in the selected input variable.*
    - *3) Recursively repeat steps 1) and 2).”*

Demo

- R „party” csomag
- Conditional Inference Tree, iris
  - Hogy ez mi, azt nem kezdjük itt nagyon részletezni
  - Party ctree dokumentáció: “*Roughly, the algorithm works as follows:*

Kis érték (tip. 0.05) indikatív a (függetlenségi) hipotézis ellen

    - *as well).*
    - *Stop if this hypothesis cannot be rejected.*
    - *Otherwise select the input variable with strongest association to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response.*
    - *2) Implement a binary split in the selected input variable.*
    - *3) Recursively repeat steps 1) and 2).”*

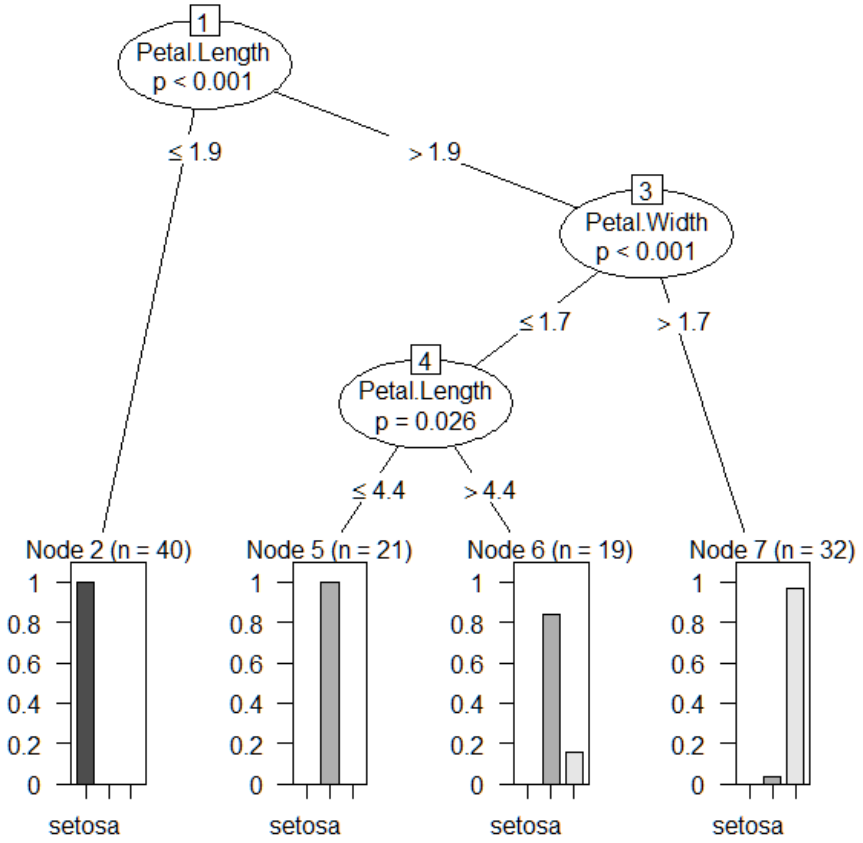
Pl. entrópia-csökkenés maximalizálásával: a split csökkenti az össz-entrópiát (alágak entrópiájának súlyozott összege)

Demo

```
> myFormula <- Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
> iris_ctree <- ctree(myFormula, data = train.data)
> table(predict(iris_ctree), train.data$Species)

             setosa versicolor virginica
  setosa         40          0         0
  versicolor      0         37         3
  virginica       0          1        31
> plot(iris_ctree)
> testPred <- predict(iris_ctree, newdata = test.data)
> table(testPred, test.data$Species)

testPred     setosa versicolor virginica
  setosa         10          0         0
  versicolor      0         12         2
  virginica       0          0        14
```



Bináris döntések jóságának mérése

Confusion matrix

Valós állapot \ Jósolt állapot (predikció)	P	N
P	TP	FN (II. típusú hiba)
N	FP (I. típusú hiba)	TN

Érzékenység vagy specifikusság a fontos?

- **Érzékenység (sensitivity): TP/(TP+FN)**
  - érzékenység magas -> FN alacsony
  - Pl:Legyen kicsi az esélye, hogy egy fertőző betegség tesztje tévesen negatív -> aki negatív eredményű az valójában is az.
- **Specifikusság (specificity): TN/(TN+FP)**
  - specifikusság magas -> FP alacsony
  - Pl:Legyen kicsi az esélye, hogy egy genetikai teszt (egy kezelés megkezdéséhez) tévesen pozitív -> aki pozitív eredményű az valójában is az.

További jellemzők:
https://en.wikipedia.org/wiki/Confusion_matrix
http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf

Asszociációs szabályok

Alapfogalmak

- Asszoc. szabályok: elemhalmazok közötti asszociáció vagy korreláció
  - `if LEFT then RIGHT`
- Pl. Tx DB-k: „sör + pelenka + fejfájáscsillapító”
- Adatkeretre is működik

„Sokszor igaz”

$$support(LEFT \rightarrow RIGHT) = \frac{N_{BOTH}}{N_{TOTAL}}$$

„Ritkán téved”

$$confidence(LEFT \rightarrow RIGHT) = \frac{N_{BOTH}}{N_{LEFT}}$$

„Nem véletlen”

$$lift(LEFT \rightarrow RIGHT) = \frac{confidence(LEFT \rightarrow RIGHT)}{N_{RIGHT}}$$

Alapfogalmak

- $lift(B \rightarrow 1) = \frac{P(\{B;1\})}{P(\{B;*\})P(\{*;1\})} \approx 1.56$

- >1: “többször fordulnak elő együtt, mint várható”
  - “ha pelenka, akkor többször sör”
- <1: “kevesebbszer fordulnak elő együtt, mint várható”

Antecedent	Consequent
A	0
A	0
A	1
A	0
B	1
B	0
B	1

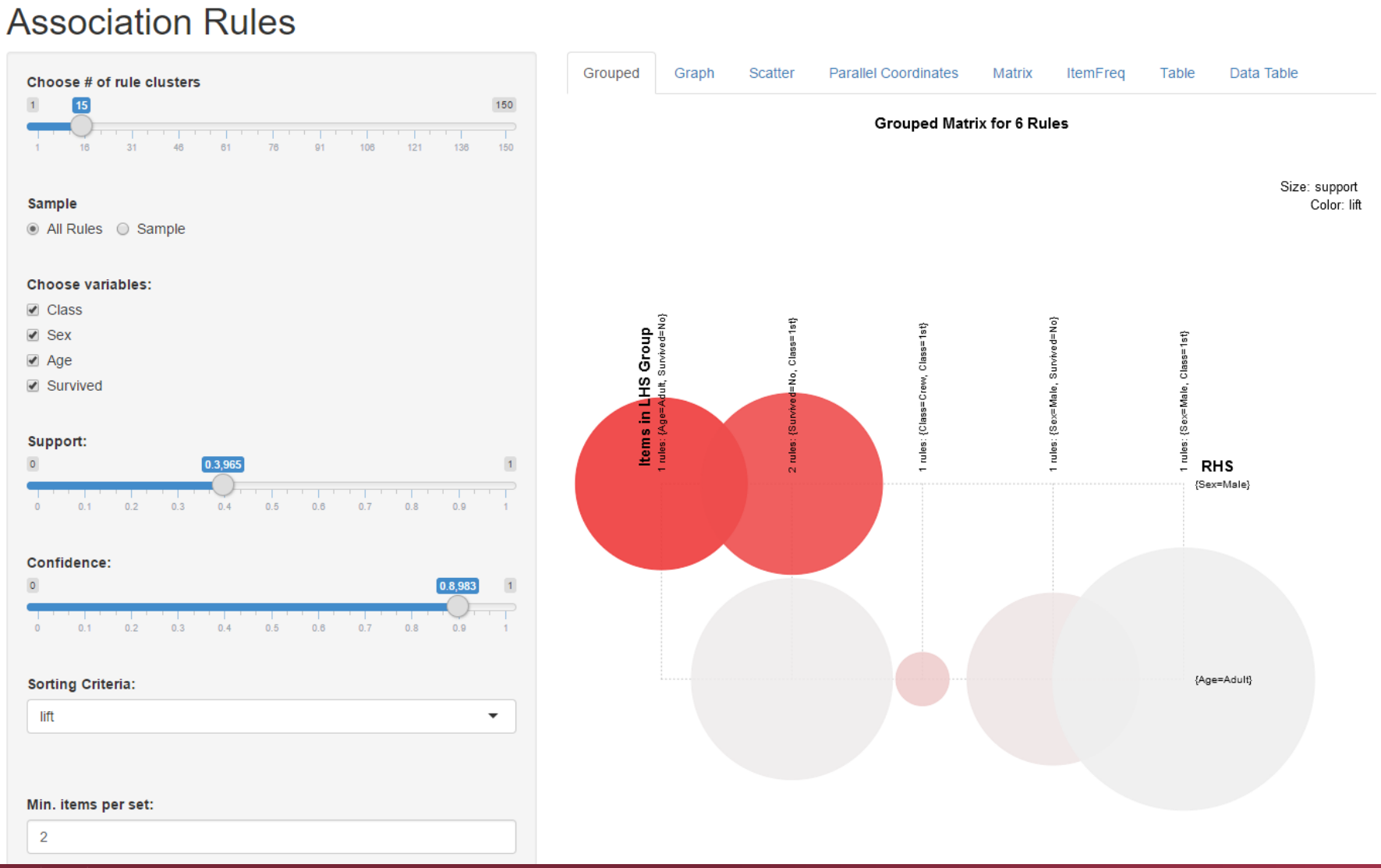
Lásd még: https://en.wikipedia.org/wiki/Lift_(data_mining)

Néhány megfontolás

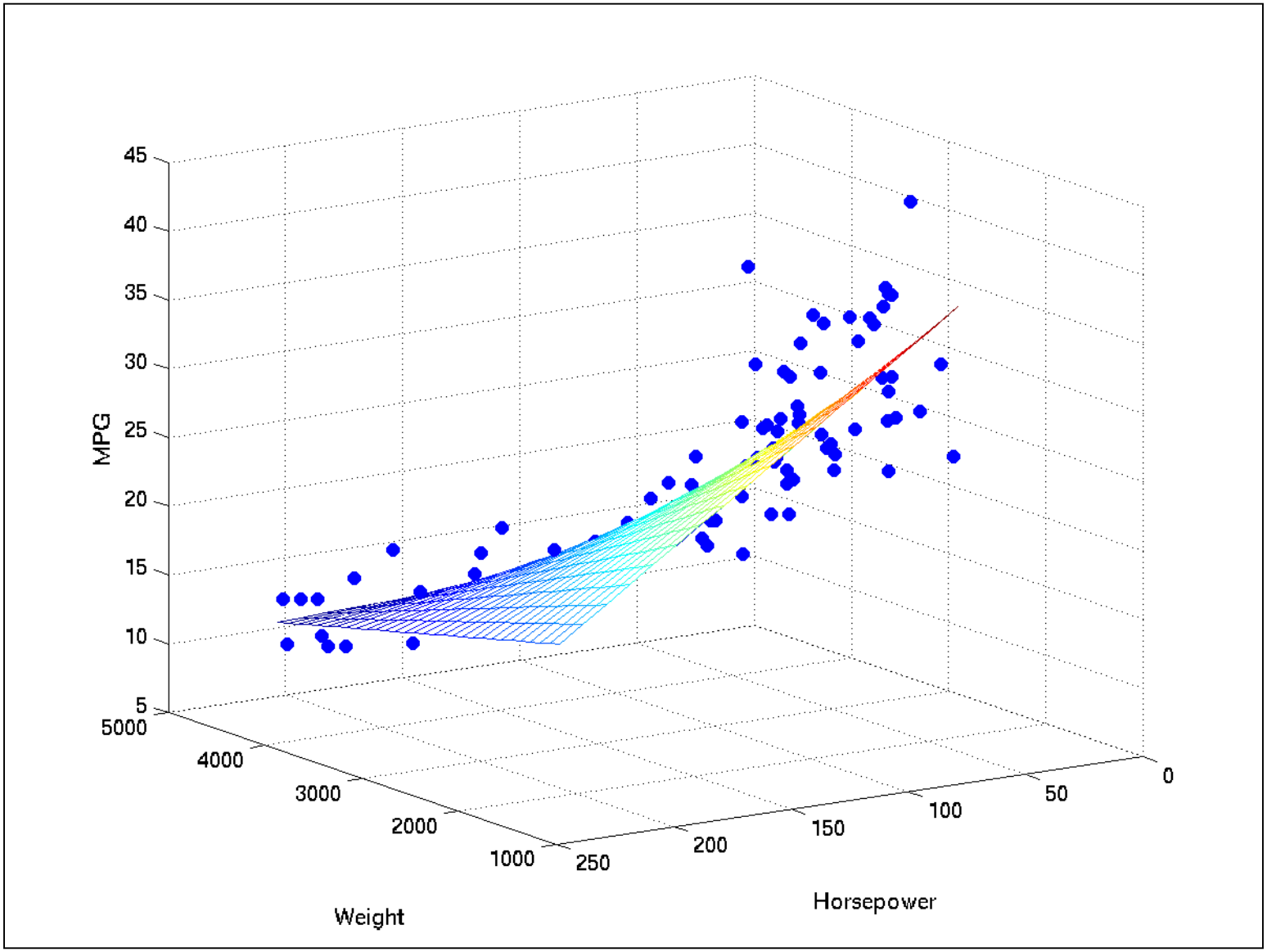
- Potenciálisan rengeteg szabály
- Amit szeretnénk: „elég magas” confidence ***és*** support
- Redundanciák is lehetnek
- A „legérdekesebbeket” bányásszuk ki
  - Valamilyen „érdekességi” metrika alapján

Demo (Titanic, ismét)

- http://brooksandrew.github.io/simpleblog/articles/association-rules-explore-app/



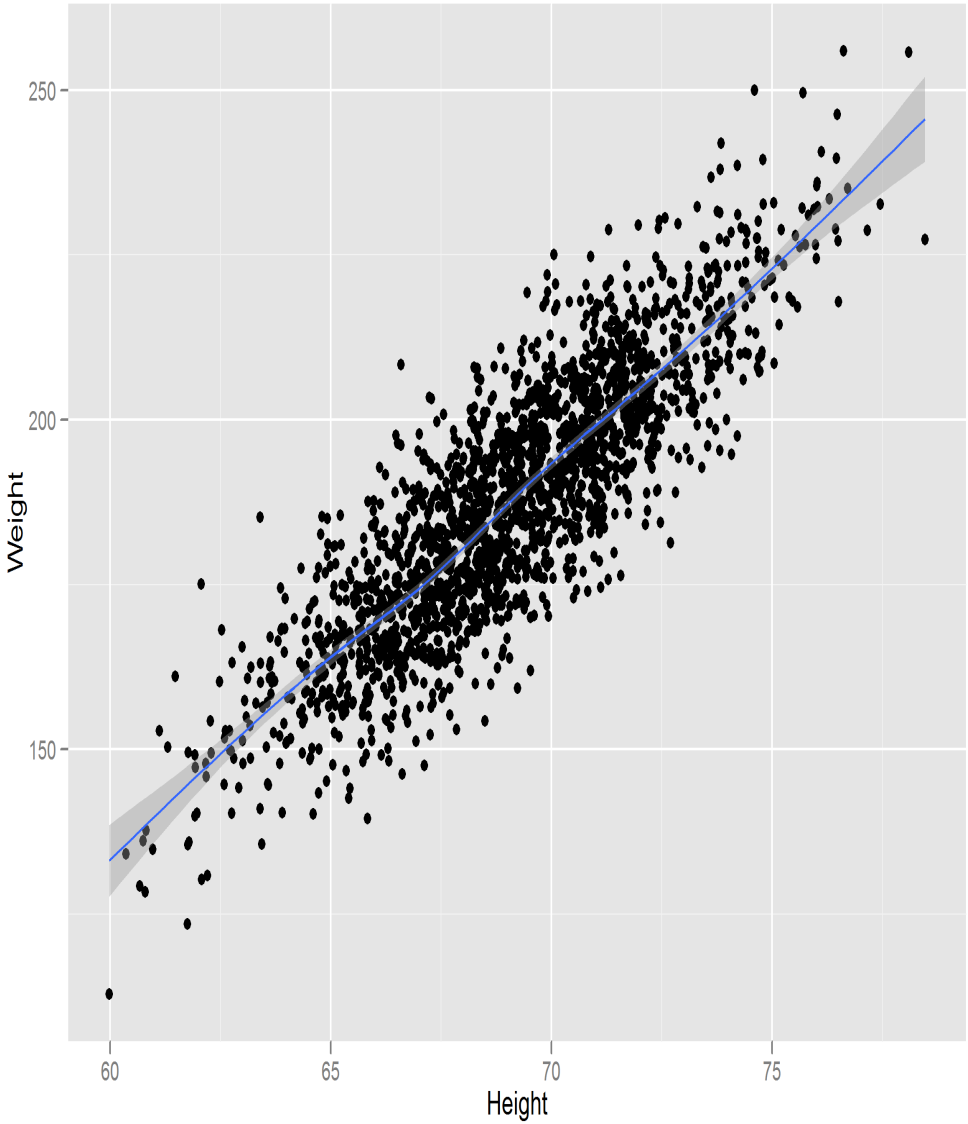
Regresszió


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

Regresszió

f függvény,

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



Regressziós módszerek

- Alapelv:

$$Y_t = f(\bullet) + \varepsilon_t$$

Véletlen változó

Közelítés

Hiba

Jósolt esemény

$$Y = f(X_1, X_2, \ldots, X_n)$$

Megfigyelhető változók

- Átlagos hiba (mean error)

Becsült érték

$$ME = \frac{\sum_{t=1}^{n} (Y_t - F_t)}{n}$$

Mért érték

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

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

- cél:

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

Zárt alak: levezetés (parc. deriválás)

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

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

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

Xi, Yi a mért értékpárok (pl. idő, terhelés)

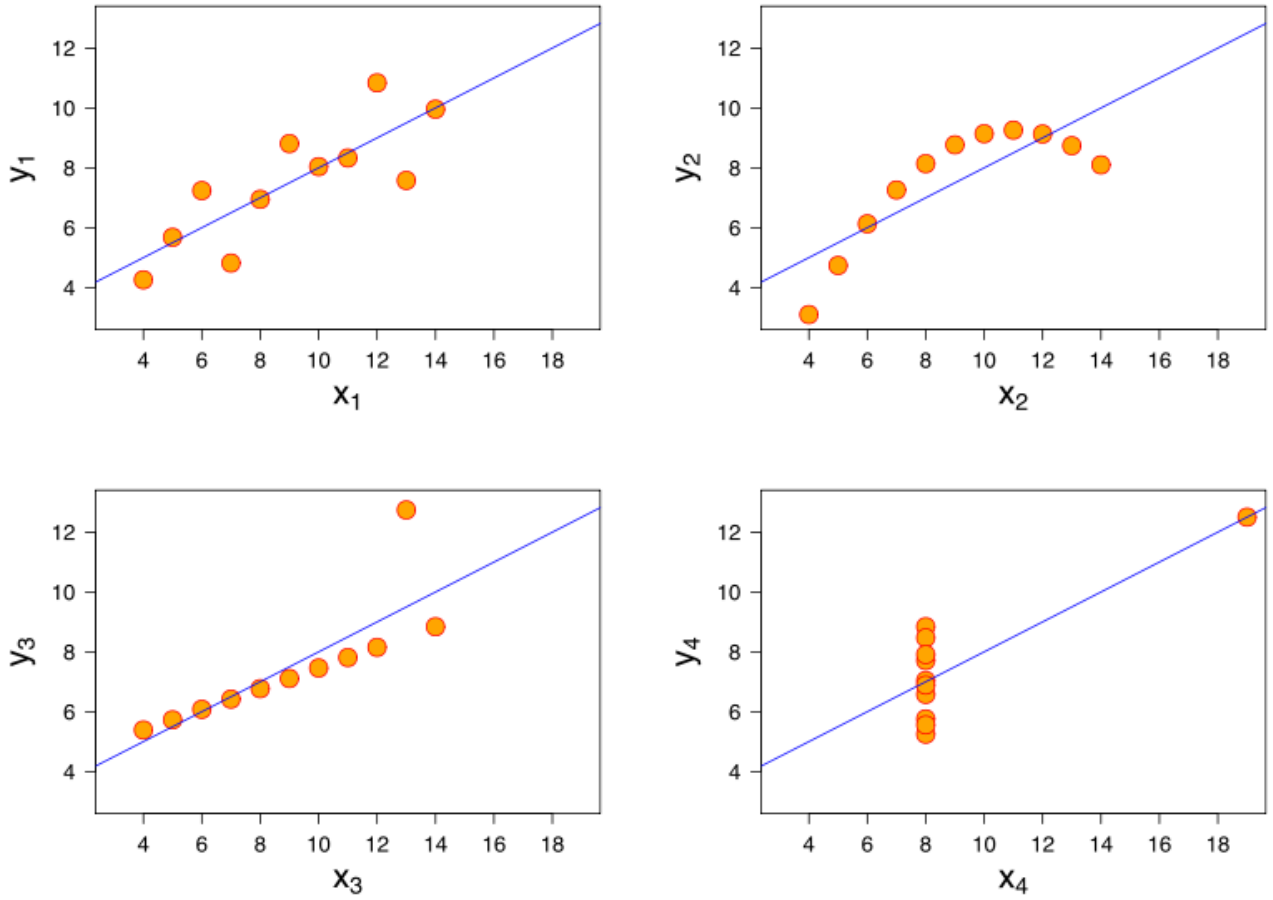
$$\frac{d\sum_{t=1}^{n}\left[Y_t-\left(a+bX_t\right)\right]^2}{db}=\sum_{t=1}^{n}X_t\left[Y_t-\left(a+bX_t\right)\right]=0$$

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

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

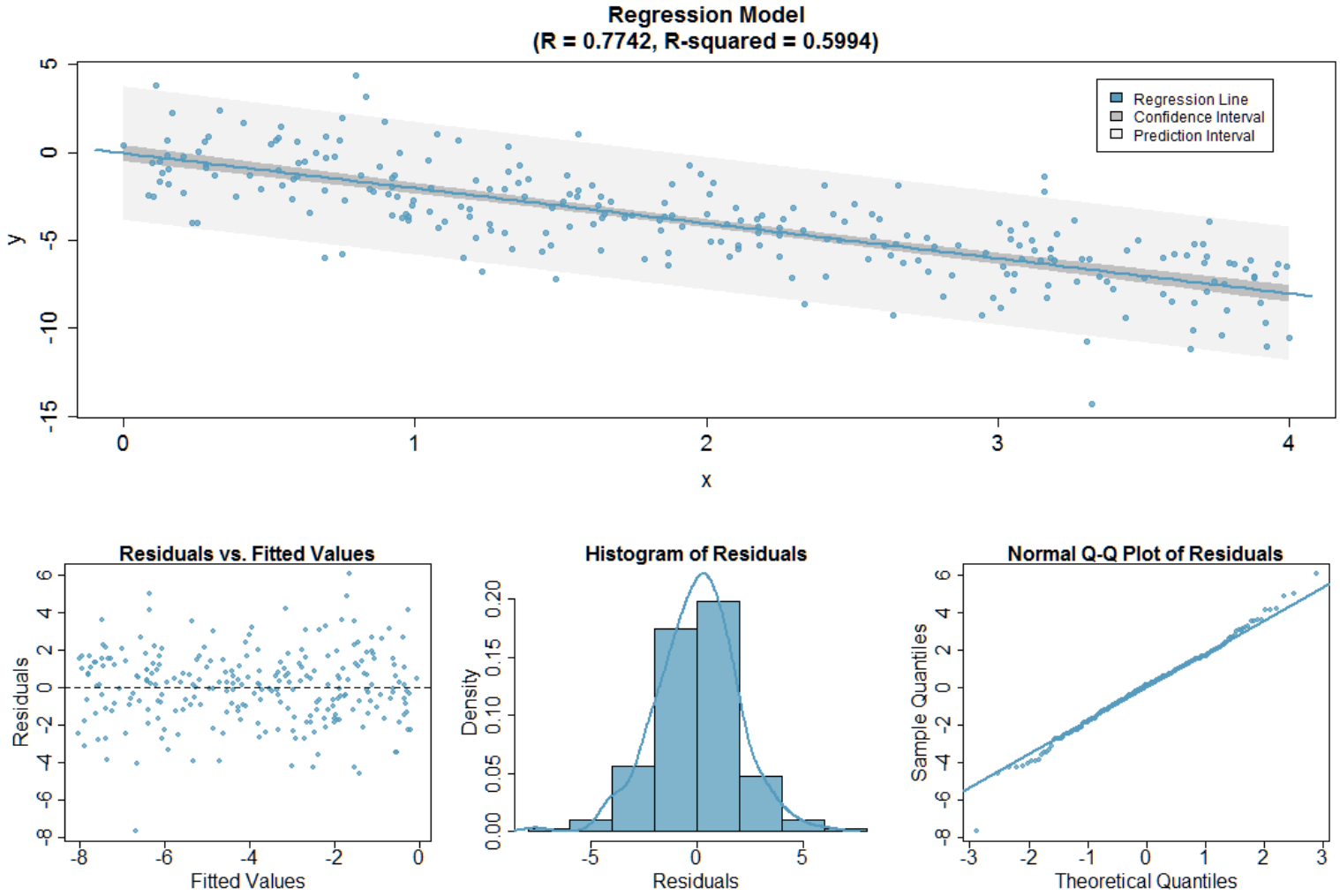
Ismétlés: Anscombe négyese

- Legjobban illeszkedő egyenes mindenre van
- Minőségileg különböző adatpontokra is



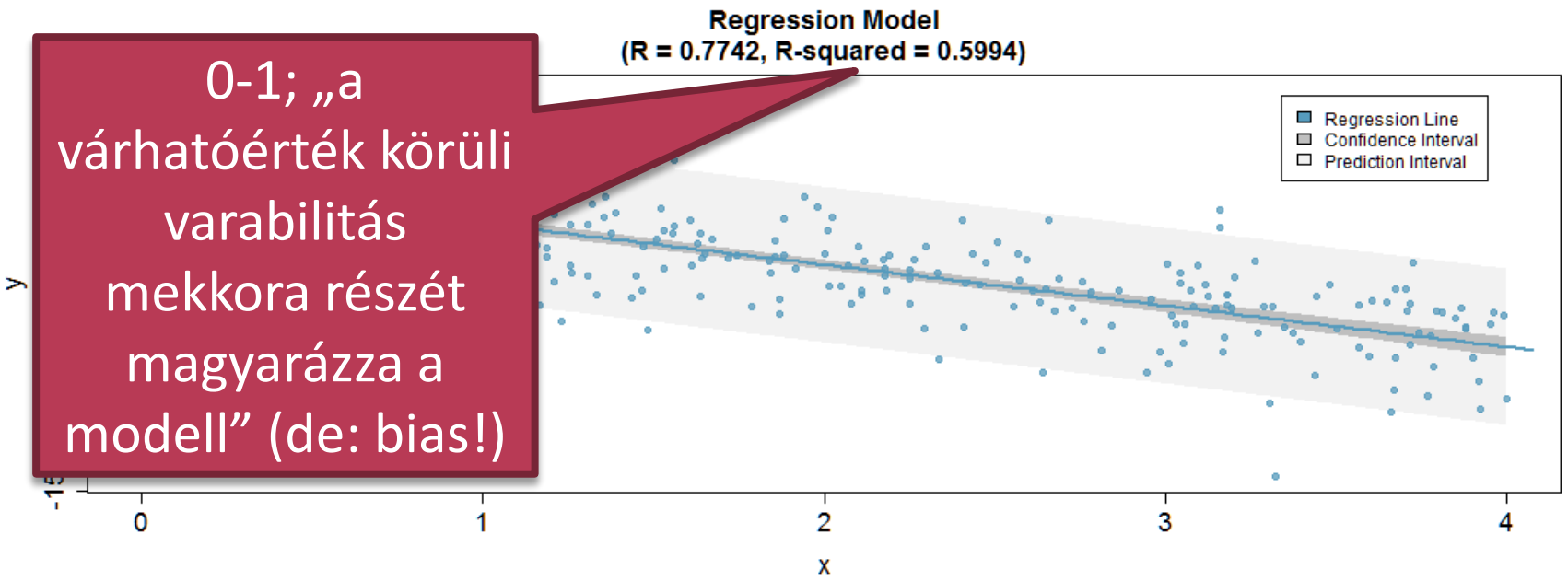
Demo

- https://github.com/ShinyEd/ShinyEd/tree/master/slr_diag

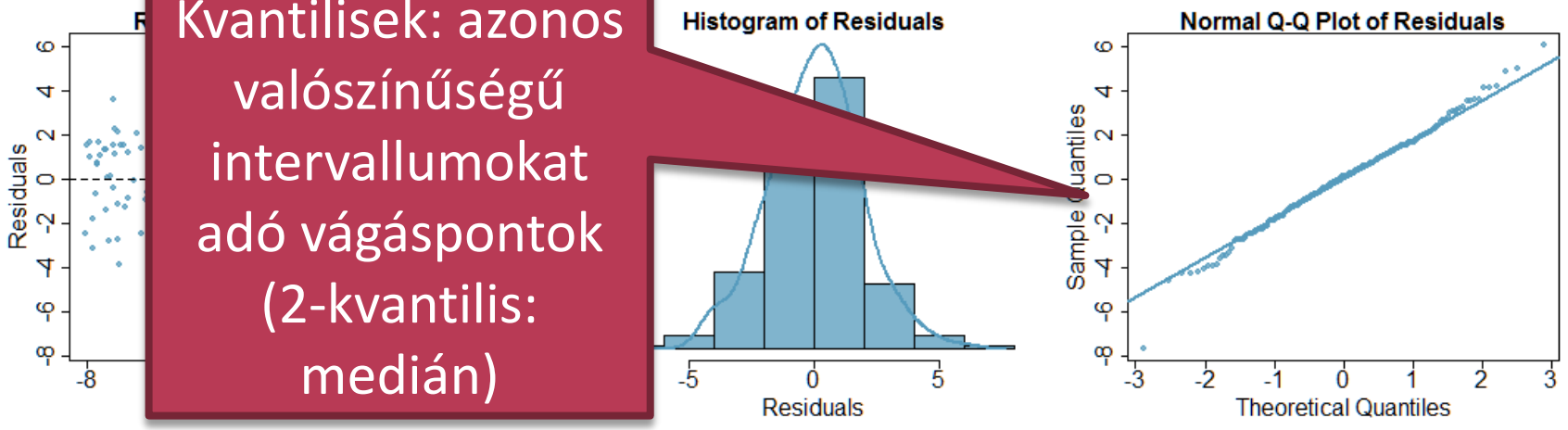


Néhány tulajdonság

0-1; „a várhatóérték körüli varabilitás mekkora részét magyarázza a modell” (de: bias!)

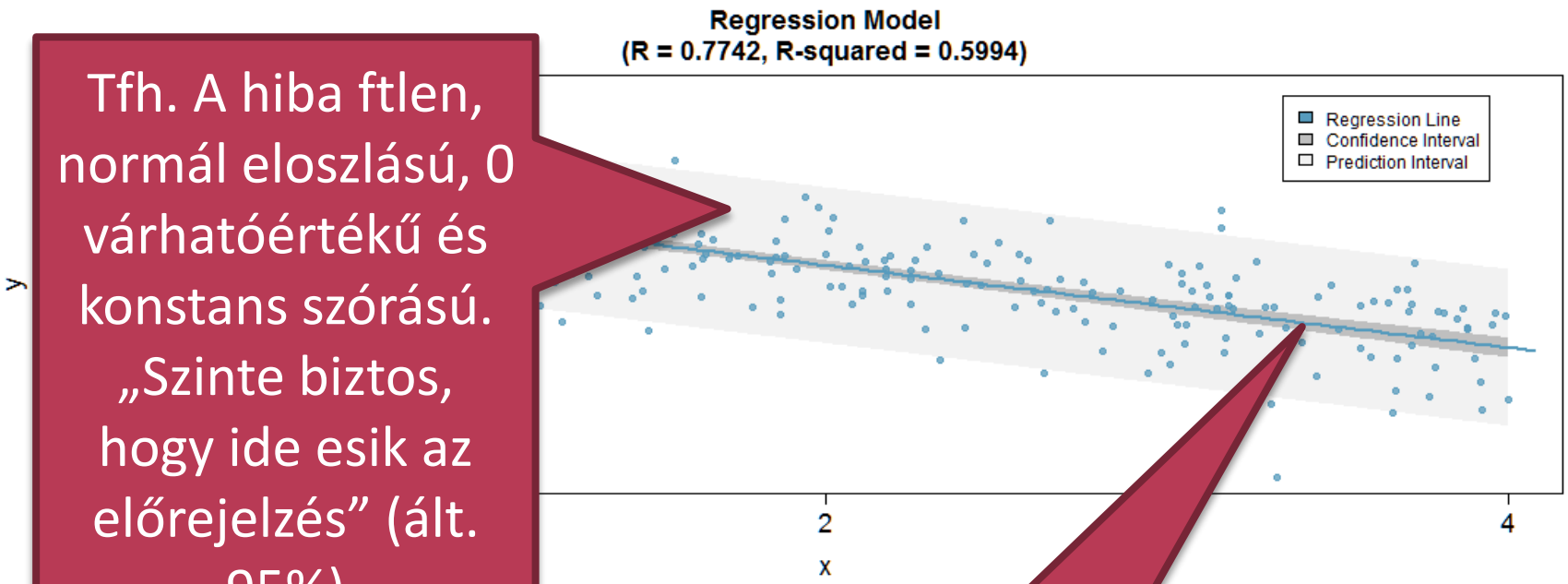


Kvantilisek: azonos valószínűségű intervallumokat adó vágáspontok (2-kvantilis: medián)

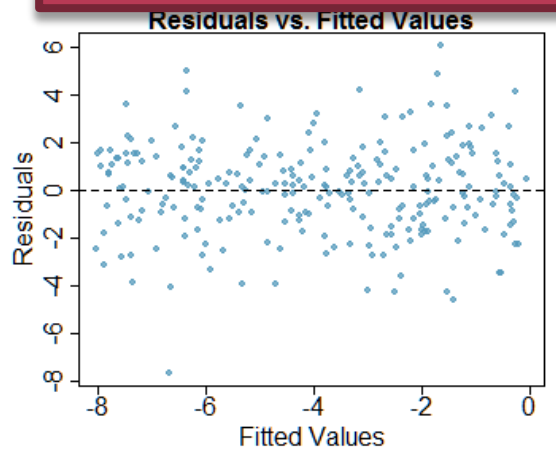


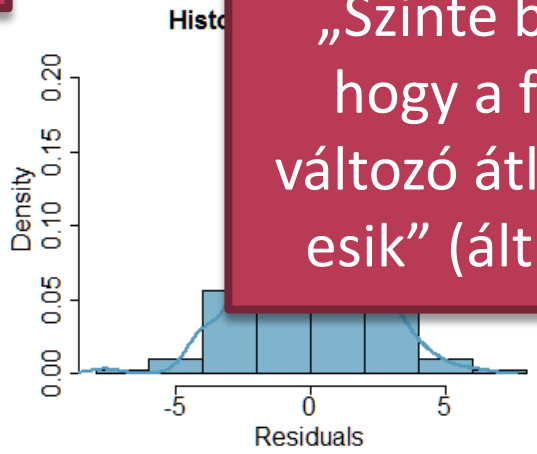
Néhány tulajdonság

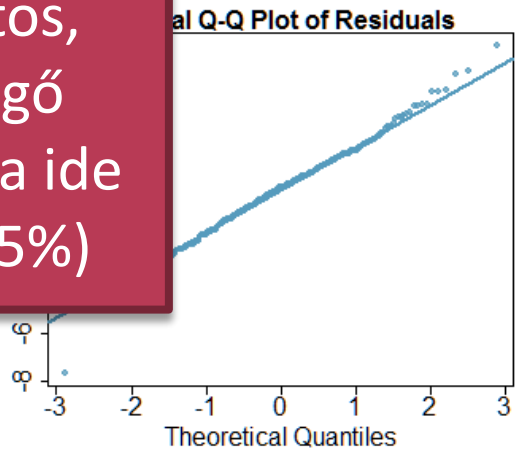
Tfh. A hiba ftlen, normál eloszlású, 0 várhatóértékű és konstans szórású. „Szinte biztos, hogy ide esik az előrejelzés” (ált. 95%)



„Szinte biztos, hogy a függő változó átlaga ide esik” (ált. 95%)







Adatelemzés: legfontosabb problémák

Idősorelemzés- és előrejelzés

Anomália-detektálás (***anomaly detection***)

„Structured prediction”

Osztályozás (***classification***)

Asszoc. szabályok (***assoc. rules***)

Csoportosítás (***clustering***)

Mintakeresés (***freq. pattern mining***)

Regresszió (***regression***)

Graph analysis

Feature selection

Dimensionality reduction

Feature extraction

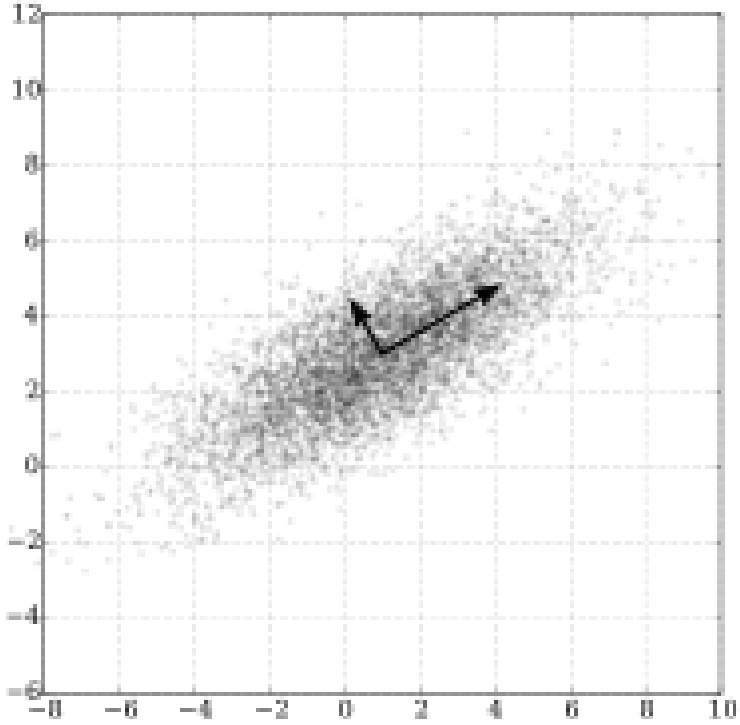
Principal Component Analysis

- Ortogonális transzformáció olyan ortogonális bázisra (lineárisan független változók; “főkomponensek”), ahol
  - Az első komponens varianciája a lehető legnagyobb
  - Az i-edik varianciája a lehető legnagyobb úgy, hogy merőleges legyen az eddigiekre.
- Értelme: lehet, hogy összvarianciát leíró komponens jóval kevesebb lesz, mint változó

- Nem faktoranalízis; az látens változók által okozott közös varianciát keres
  - És ezzel felteszi egy mögöttes modell létezését
  - Az “Exploratory Factor Analysis” tartalmazhat PCA-szerű lépést…

Principal Component Analysis

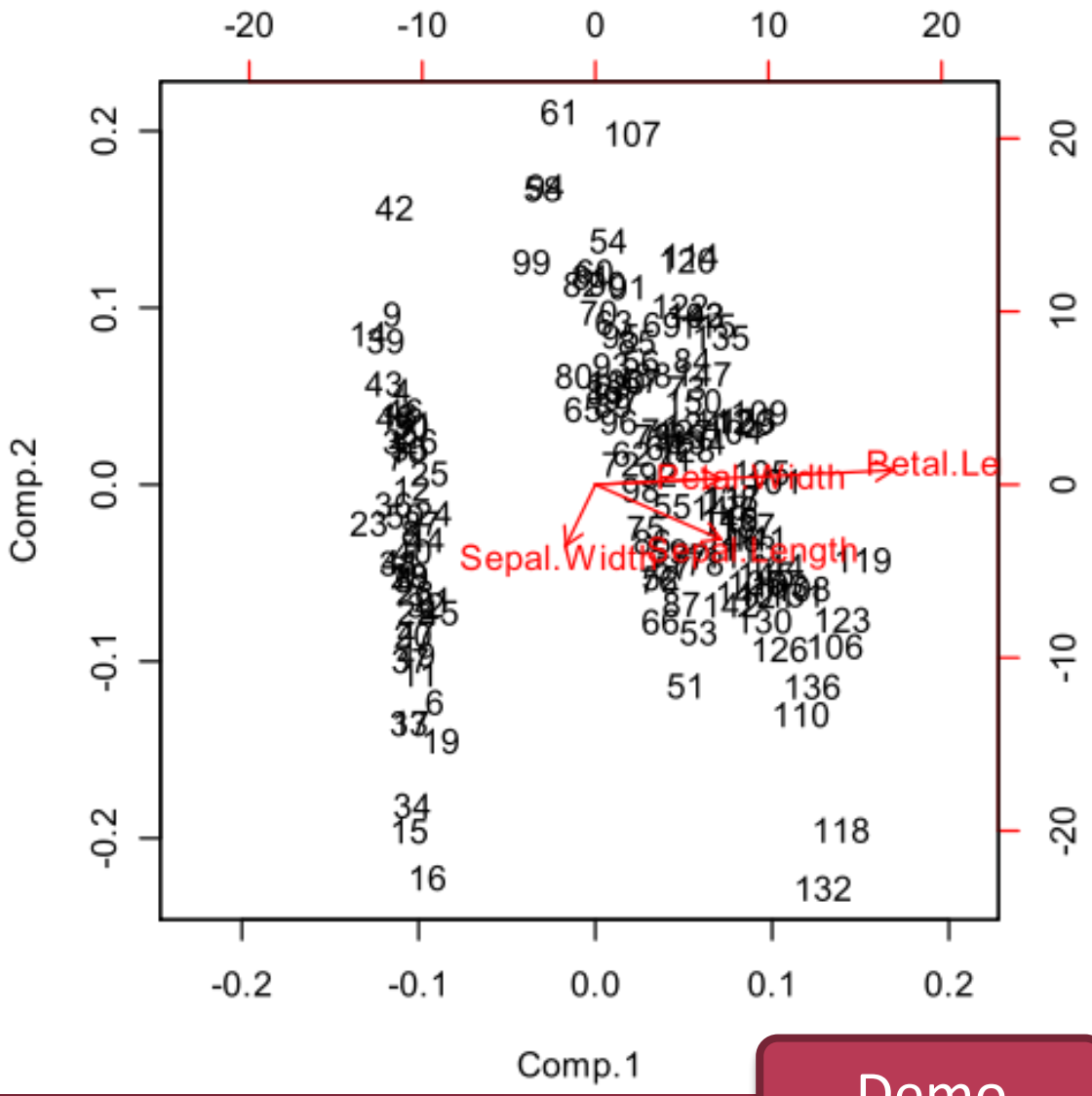
- Eltolás a várhatóértékbe + utána forgatás
- Skálaérzékeny, de lehet normalizálva is csinálni (pl. Z-score-ra)
- Ennek persze megint lehetnek nem kívánt hatásai



- Matematikáját nem tárgyaljuk

Biplot

- (Implicit) feltételezés: 2 komponens “elégségesen leírja”
- Komponensenkénti “koordináták” (score-ok)
- + Az eredeti változók súlya (“loading”) az első két faktorban



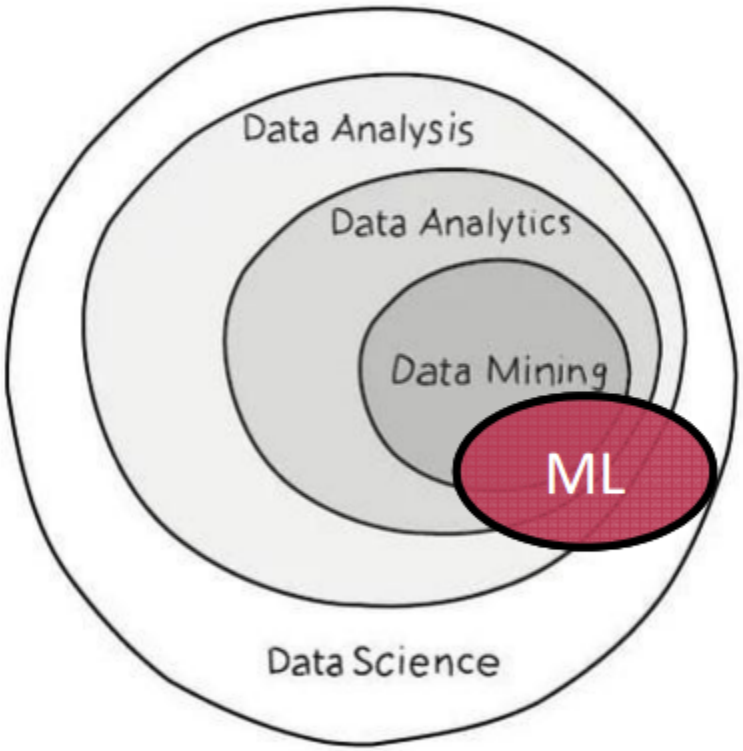
Demo

Gépi tanulás

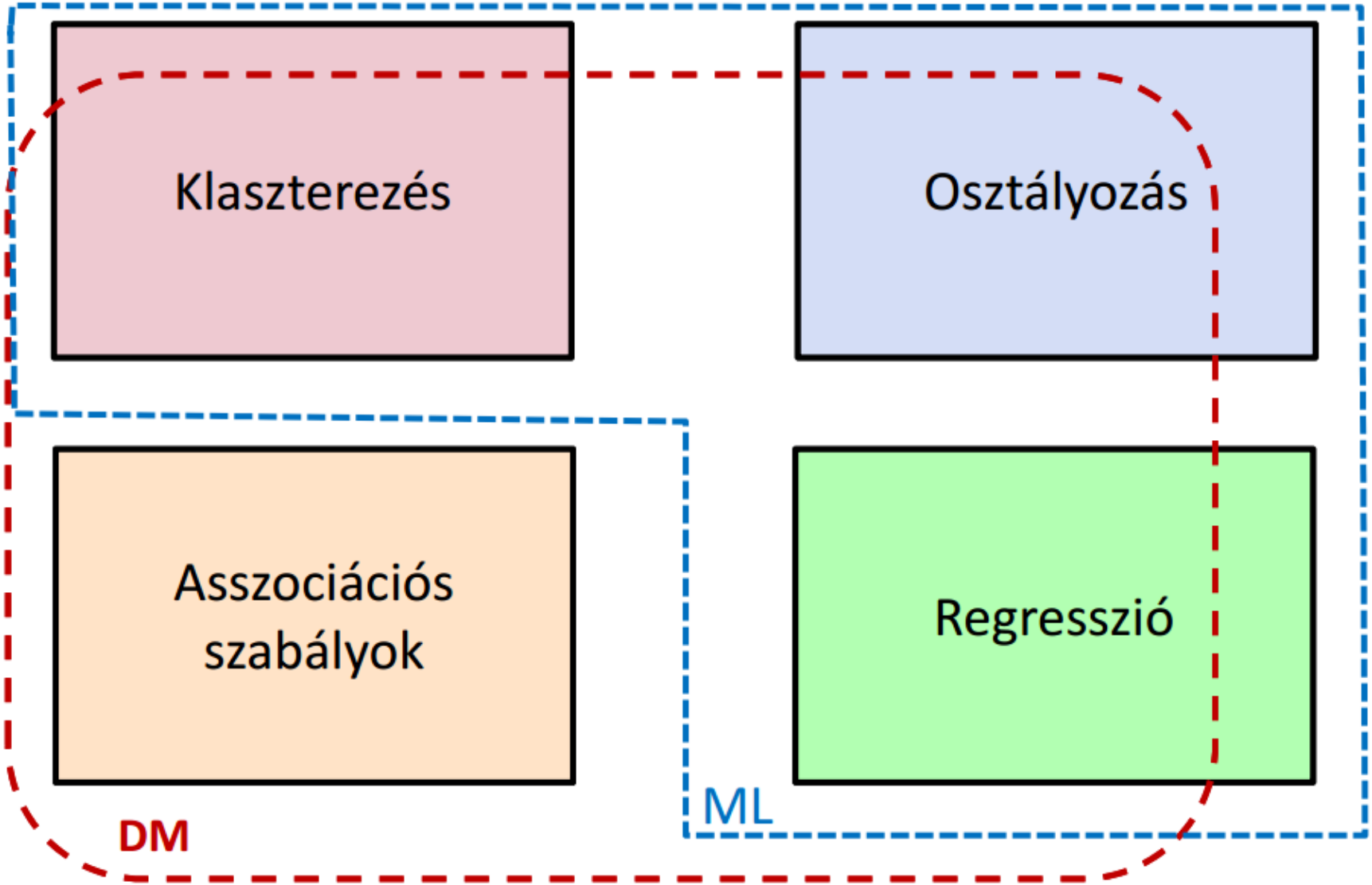
- **Gépi tanulás** (Machine Learning - ML)
  ~ Olyan módszerek és algoritmusok, melyek a rendelkezésre álló adathalmazból egy modellt tanulnak, és ezt felhasználva következtetést (predikciót) tesznek lehetővé (a meglévő vagy más/újabb adatokról).
- Adatorientált módszereket használ, tehát lényeges a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

Adatbányászat

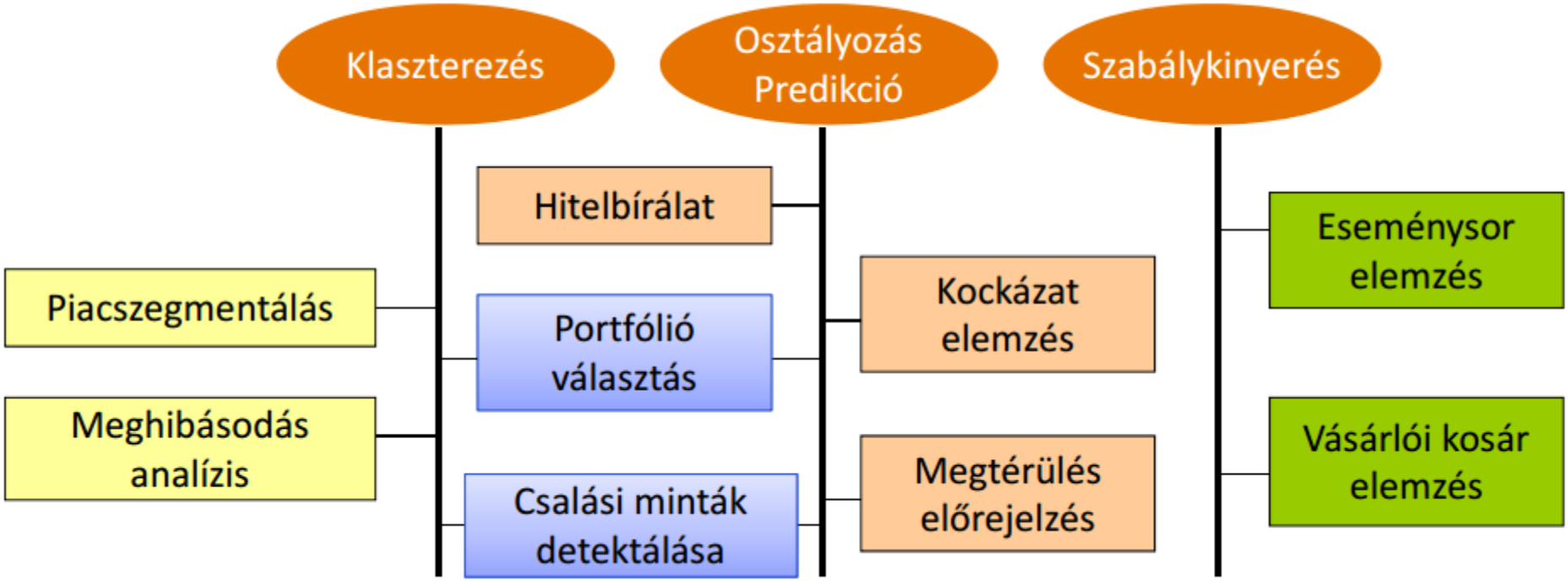
- Gépi tanulás
  - Előrejelző modellt épít
- Adatbányaszat
  - ML módszereket felhasználva nyer ki új információt
  - Inkább feltáró jellegű
  - Inkább nem felügyelt tanulás



ML vs DM


Klaszterezés
Osztályozás
Asszociációs szabályok
Regresszió
DM
ML

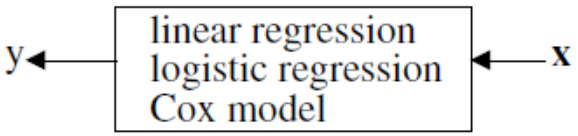
ML alapú üzleti intelligencia alkalmazások



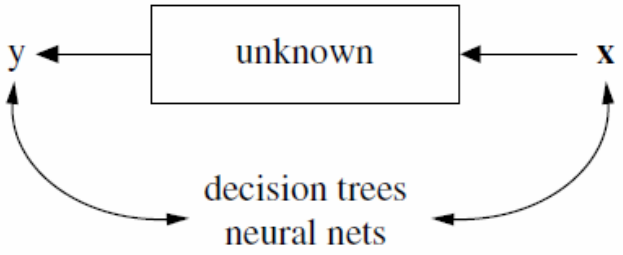
Machine Learning, Data Mining, statisztika

- Az eszközök és a „DNS”ugyanaz; különböző kultúrák

- Statisztika: (stochasztikus) adatmodellezés
- ML: informatikusok tanuló programot akartak
- DM: tudás kinyerése az adatokból (EDA...stat. Modellezés)



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Lásd [2][3]

“Közel ugyanaz”

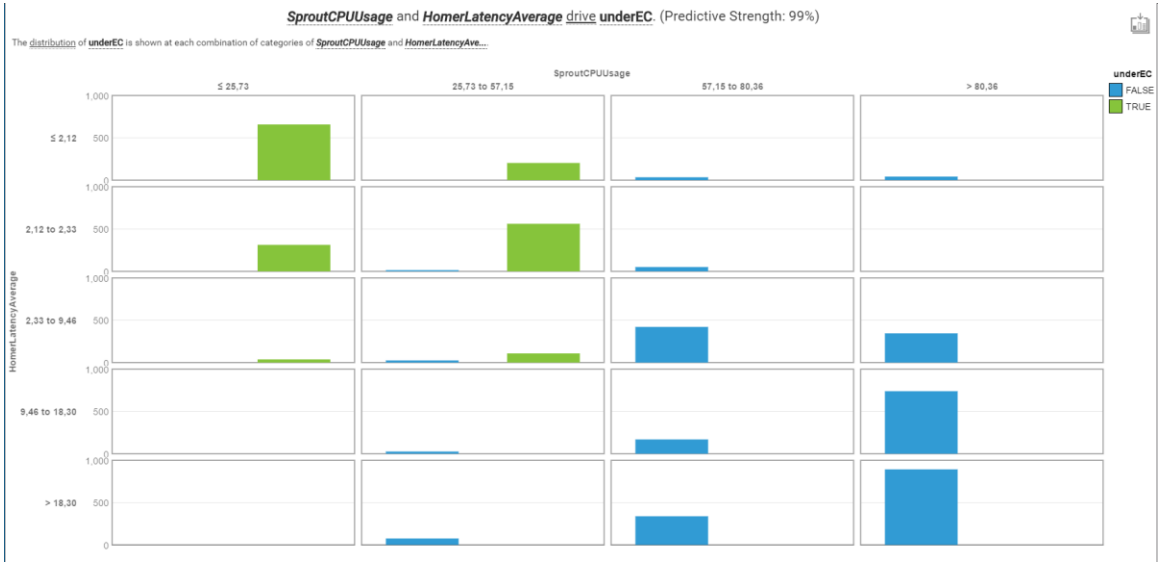
Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

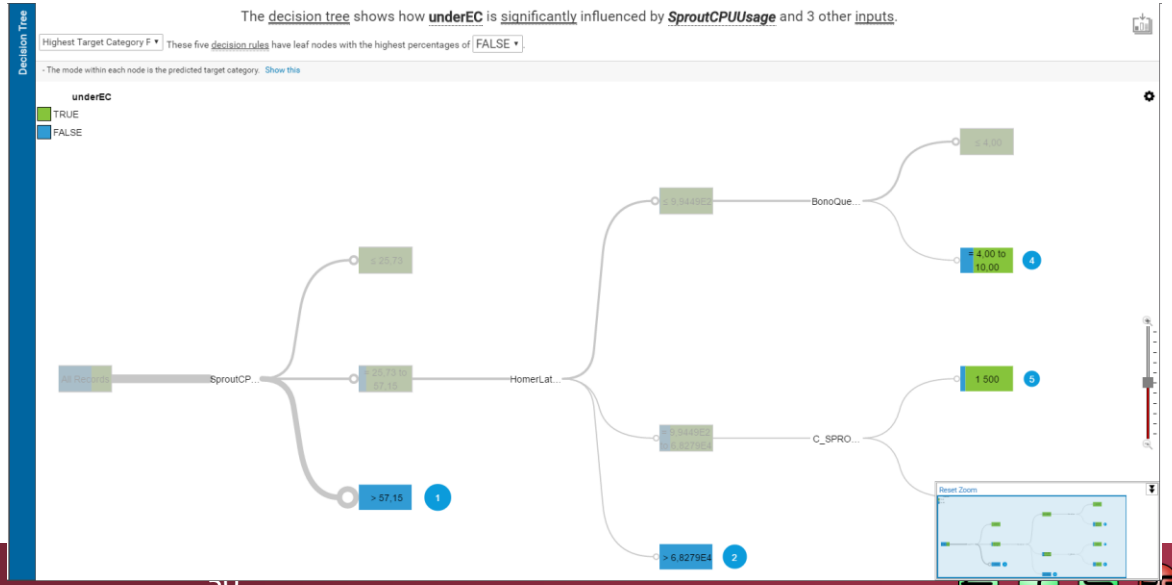
Lásd [4]

AUTOMATIZÁLT ADATELEMZÉS: IBM WATSON ANALYTICS

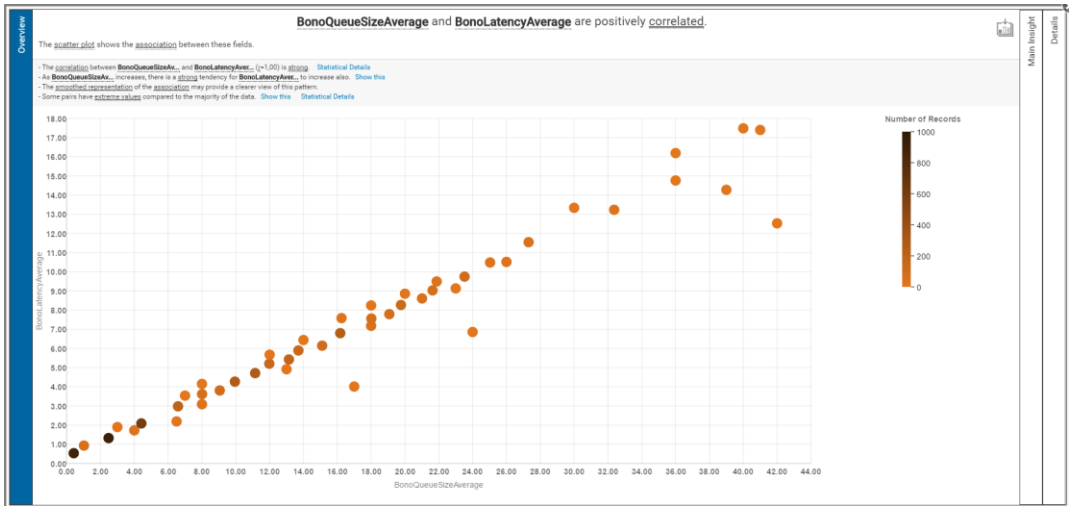
Automatizált adatelemzés – IBM Watson Analytics



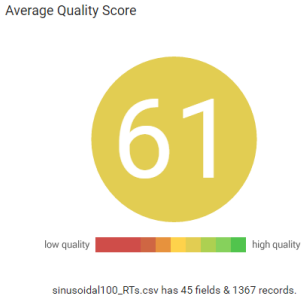
Automatikus “predikció”

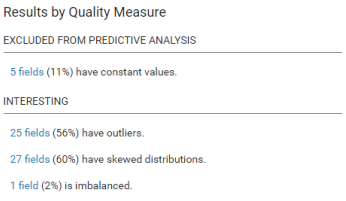


Automatizált adatelemzés – IBM Watson Analytics



"Érdekes asszociációk" automatikus feltárása





Automatizált adatminőség-értékelés

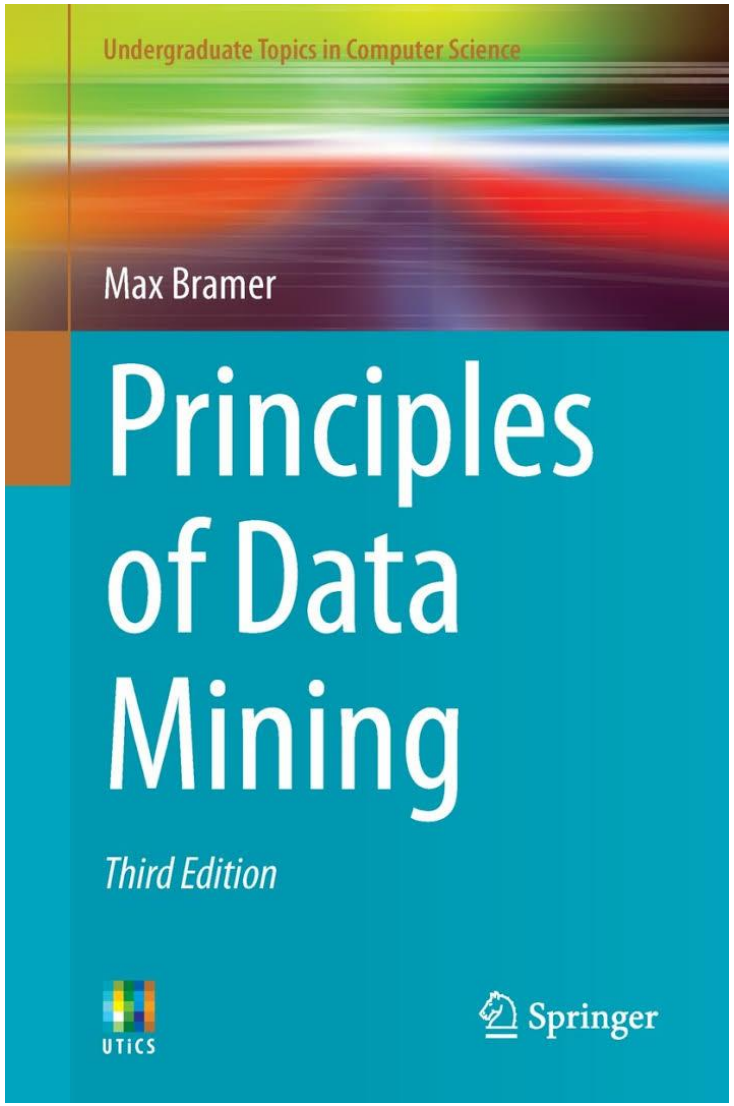
IRODALOM

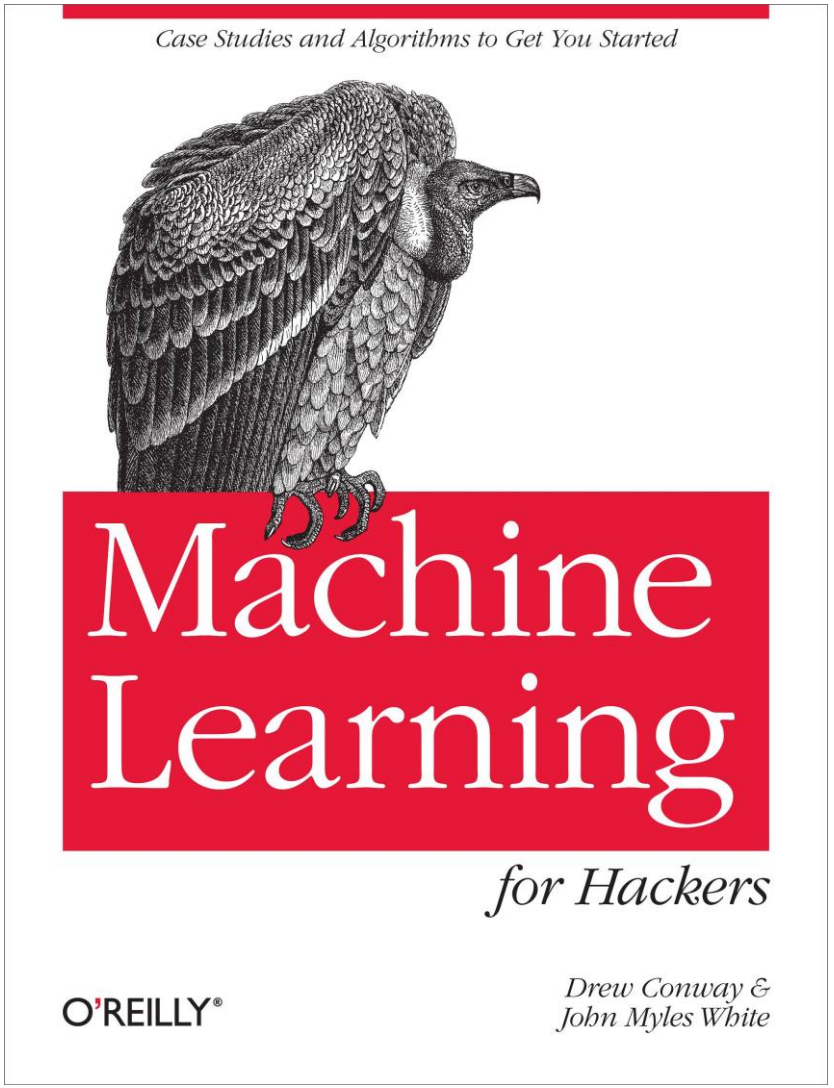
Javasolt magyar nyelvű anyagok

- Helyenként nagyon mély, de kiváló tanulmány/jegyzet a Számításelméleti és Inf. Tud. Tanszékről:
  - http://www.cs.bme.hu/nagyadat/bodon.pdf

- Egyéb:
  - Dr. Abonyi János: Adatbányászat a hatékonyság eszköze
  - Iványi Antal: Informatikai Algoritmusok 2. kötet (28-29. fejezetek), ELTE Eötvös Kiadó

További javasolt kezdőirodalom





Hivatkozások

- [1] Theus, M., Urbanek, S.: Interactive graphics for data analysis: principles and examples. CRC Press (2011)
- [2] https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726
- [3] https://www.r-bloggers.com/whats-the-difference-between-machine-learning-statistics-and-data-mining/
- [4] http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf