

Adatbányászat: Adatfeltárás

3. fejezet

Tan, Steinbach, Kumar
Bevezetés az adatbányászatba
előadás-fóliák
fordította
Ispány Márton

Mi az adatfeltárás?

Az adatok előzetes feltárása (vizsgálata) segít jellemzőinek jobb megértésében.

- Az adatfeltárás alapvető motivációi
 - Segíti a helyes módszer kiválasztását az előfeldolgozásnál és az elemzésnél.
 - Lehetővé teszi az emberi képességek felhasználását a mintázatok felismerésében.
 - ◆ Az ember az elemző szoftverek által nem felismert mintázatokat is megtalálhatja.
- Összefügg a feltáró adatelemzéssel (EDA)
 - A módszer John Tukey statisztikustól származik.
 - Alapvető irodalom: Tukey, Exploratory Data Analysis
 - Online bevezetés: Chapter 1, NIST Engineering Statistics Handbook
<http://www.itl.nist.gov/div898/handbook/index.htm>

Az adatfeltárás módszerei

- Az EDA-ban ahogy Tukey eredetileg definiálta:
 - A hangsúly a vizualizáción van.
 - A klaszterosítást és eltérés keresést a feltárási módszerekbe sorolja.
 - Az adatbányászatban a klaszterosítás és eltérés keresés az érdeklődés központjában van és nem csupán a feltárás egy része.
- Az adatfeltárásban az alábbiakra fókuszálunk:
 - Leíró statisztikák.
 - Megjelenítés, grafikus eszközök.
 - OLAP: közvetlen analitikus feldolgozás.

Az írisz adatállomány

- Sok adatfeltárási módszert szemléltethetünk az írisz növénnel kapcsolatos adatokkal.
 - Letölthető: UCI Machine Learning Repository
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
 - Douglas Fisher statisztikustól származik
 - Három virág alfaj (osztályok):
 - ◆ Setosa
 - ◆ Virginica
 - ◆ Versicolour
 - Négy (folytonos) attributum
 - ◆ Levél szélesség és hosszúság
 - ◆ Szirom szélesség és hosszúság



Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Délnyugati lápos növényzet. Terep kalauz növényfajtákhoz. Northeast National Technical Center, Chester, PA. A USDA NRCS Wetland Science Institute engedélyével.

Leíró statisztikák

- A leíró statisztikák olyan mutató számok, amelyek az adatok tulajdonságait összegzik, tömörítik.
 - Ezek a tulajdonságok lehetnek gyakoriságok, helyzet, szóródás és alakmutatók.
 - ◆ Példák: helyzet mutatók: átlag, medián, módusz
 szóródás mutatók: variancia, std. dev.
 - A legtöbb leíró statisztika az adatállomány egyszeri átfésülésével számolható.

Gyakoriság és módusz

- Egy attributum érték gyakorisága annak száma, hogy az érték hányszor fordul elő az adatállományban.
 - Például a „nem” attributum esetén egy reprezentatív mintánál a nők relatív gyakorisága 50% körül van.
- Egy attributum módusza a leggyakoribb attributum érték.
- A gyakoriság és a módusz fogalmát általában kategórikus (diszkrét) attributumoknál használják.

Percentilisek, kvantilisek

- Folytonos attributumra a percentilis (kvantilis) fogalma a hasznosabb.
- Egy sorrendi vagy különbségi skálán mért X attributum és egy p 0 és 100 közötti szám esetén a p -edik percentilis az az x_p érték, amelynél az X -re megfigyelt értékek $p\%$ -a kisebb.
- Például az 50%-os percentilis (medián) az az $x_{50\%}$ érték, amelynél az attributum értékek 50% kisebb.

Helyzet mutatók: átlag és medián

- Az átlag a legáltalánosabban használt mutató rekordok (pontok) egy halmazának helyzetére.
- Az átlag nagyon érzékeny a kiugró értékekre.
- Ennek kivédésére a mediánt vagy a nyírott átlagot használják.

$$\text{átlag } (x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{medián } (x) = \begin{cases} x_{(r+1)} & \text{ha } m = 2r + 1 \\ \frac{1}{2} (x_{(r)} + x_{(r+1)}) & \text{ha } m = 2r \end{cases}$$

Példa

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Átlag: 1090K

Nyírott átlag (min és max nélkül): 105K

Medián: $(90+100)/2 = 95K$

Szóródás mutatók: terjedelelem, variancia

- A terjedelelem a maximum és a minimum eltérése.
- A variancia (standard deviáció) egy pontthalmaz szóródásának legelterjedtebb mérőszáma.

$$\text{var}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$$

- Mivel a variancia szintén érzékeny a kiugró értékekre ezért más mérőszámokat is használnak.

$$AAD(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$$

Átlagos abszolút eltérés

$$MAD(x) = \text{medián}(|x_i - \bar{x}|, i = 1, \dots, m)$$

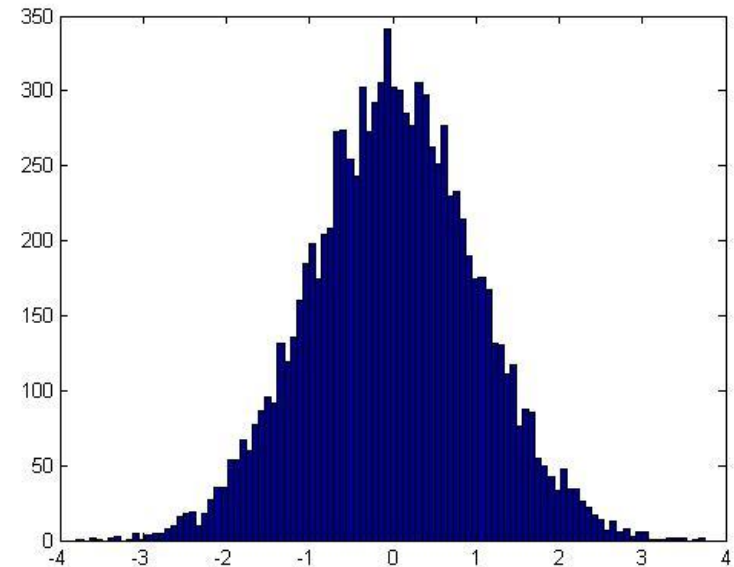
Medián abszolút eltérés

$$IQR(x) = x_{75\%} - x_{25\%}$$

Interkvartilis terjedelelem

Normális eloszlás

$$\bullet \phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

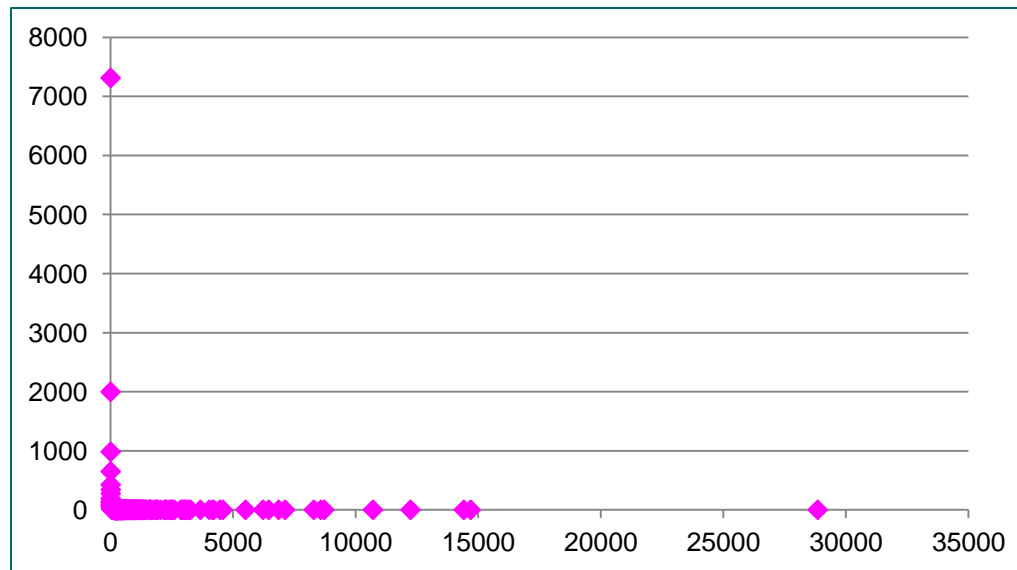


Egy histogram

- Az a nevezetes eloszlás, amely számos változót jellemez és központi szerepet játszik a valószínűségszámításban és a statisztikában.
 - A központi határeloszlás tételben is megjelenik
- Teljesen jellemzi a μ várható érték és a σ szórás.

Nem minden normális eloszlású

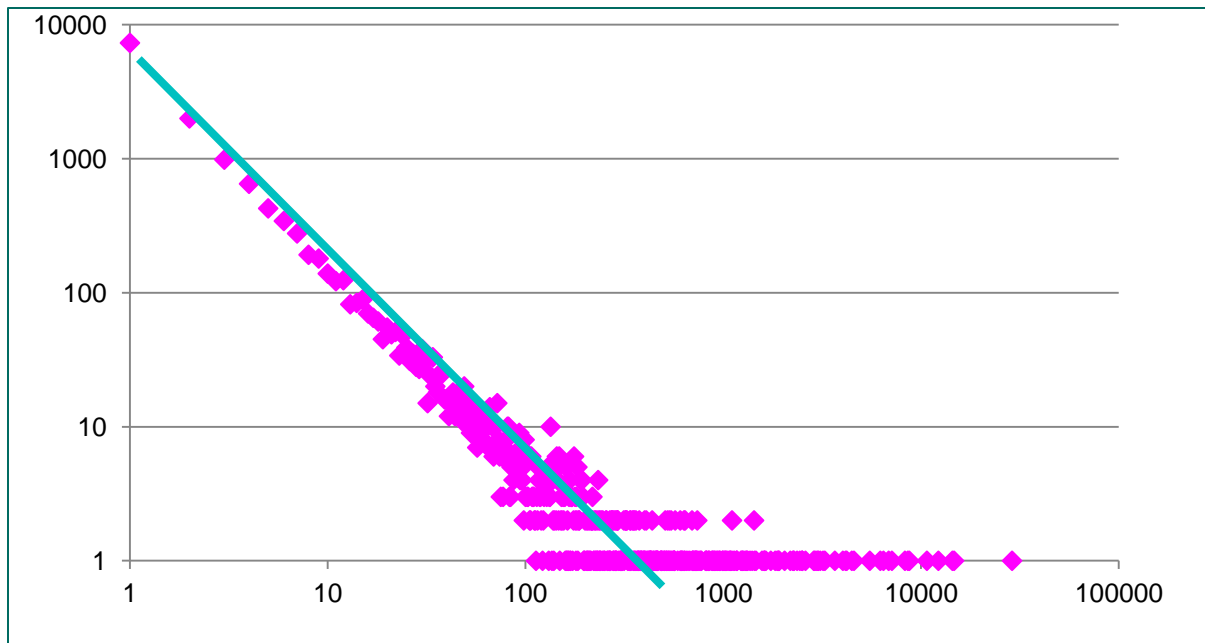
- A szavak száma az előfordulásuk gyakoriságának függvényében



- Ha ez normális eloszlást követne, akkor nem fordulhatna elő 28K gyakoriságú szó

Hatvány-eloszlás

- A szavak eloszlását akkor érthetjük meg ha a **log-log** grafikont vesszük

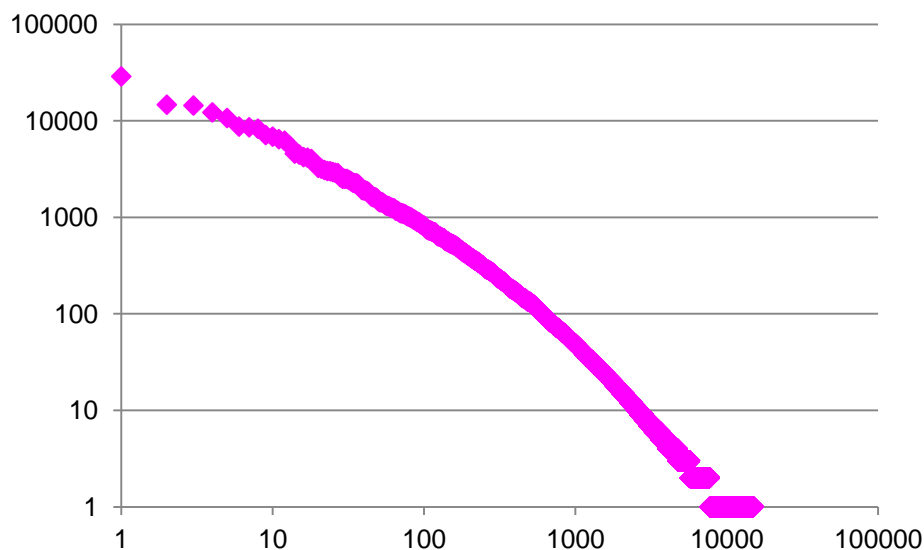


- Lineáris kapcsolat a log-log koordinátarendszerben

$$p(x = k) = k^{-a}$$

Zipf törvény

- A hatvány-eloszlást log-log koordinátarendszerben egy lineáris kapcsolatként lehet megfigyelni a **rang-gyakoriság** grafikonon

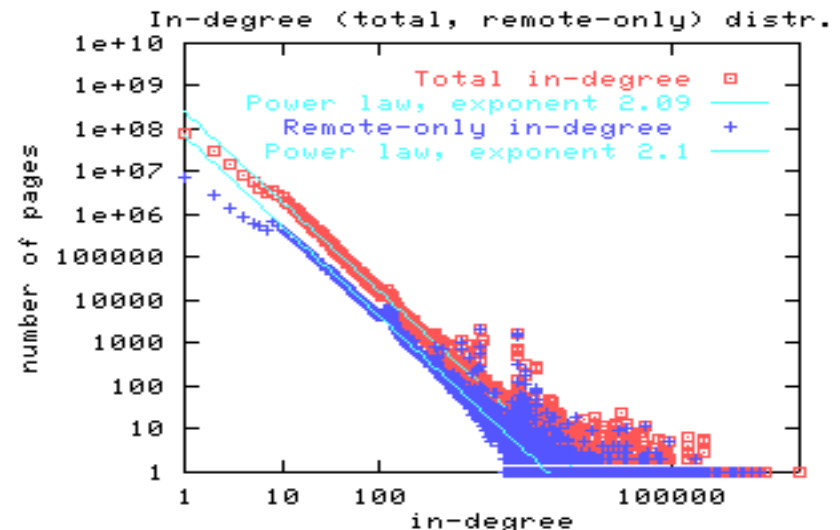


- $f(r)$: Gyakorisága az **r-th** leggyakoribb szónak

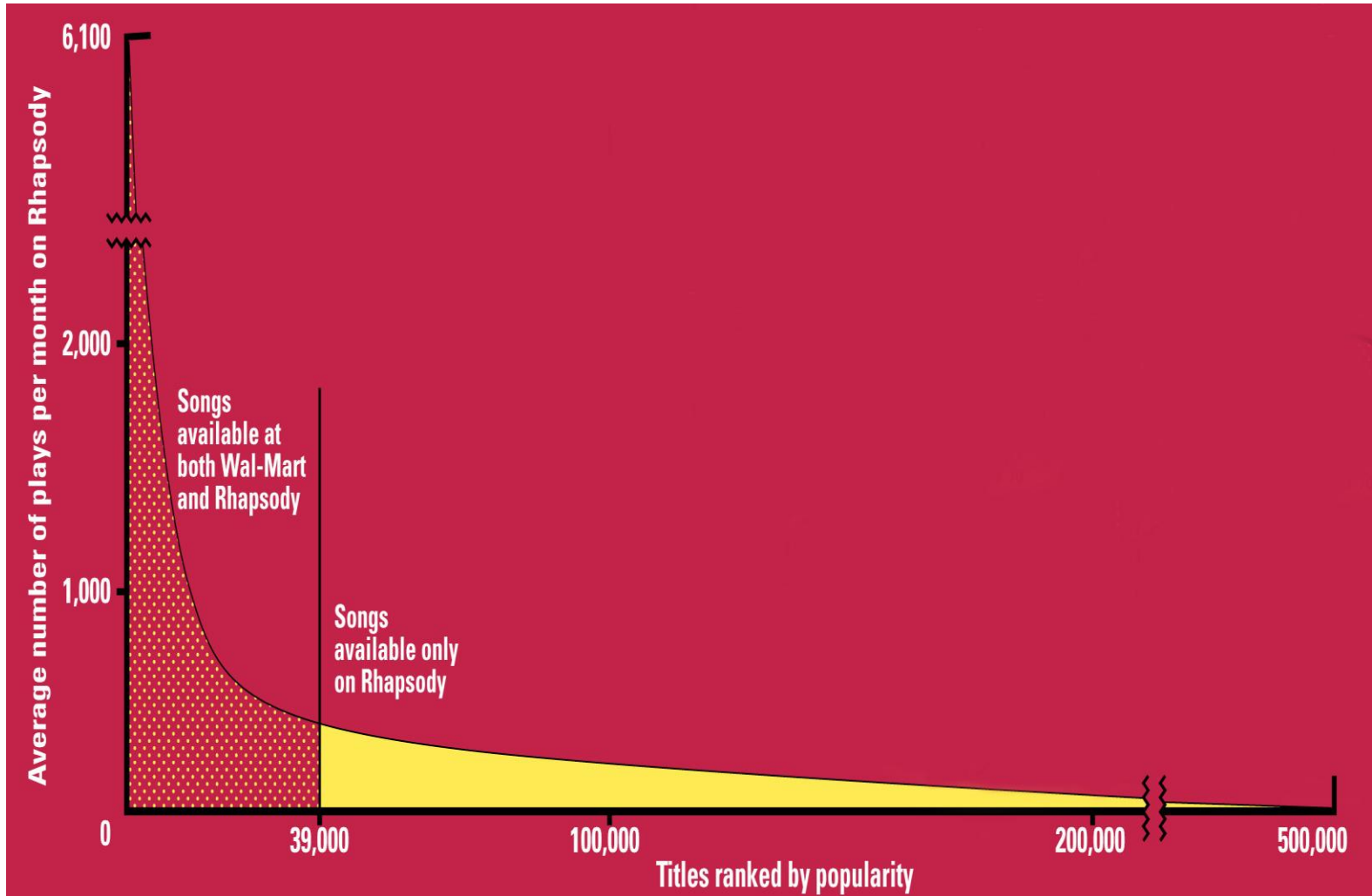
$$f(r) = r^{-\beta}$$

Hatvány-eloszlás mindenhol

- Bejövő és kimenő linkek web-lapokon, barátok száma közösségi hálózatokban, szavak gyakorisága, fájlok mérete, városok nagysága, jövedelem eloszlás, termékek vagy filmek népszerűsége
 - Az emberi tevékenység lenyomatait?
 - Egy szabály, amely mindent megmagyaráz?
 - A gazdagabb egyre gazdagabb lesz



A hosszú farok



Source: Chris Anderson (2004)

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks

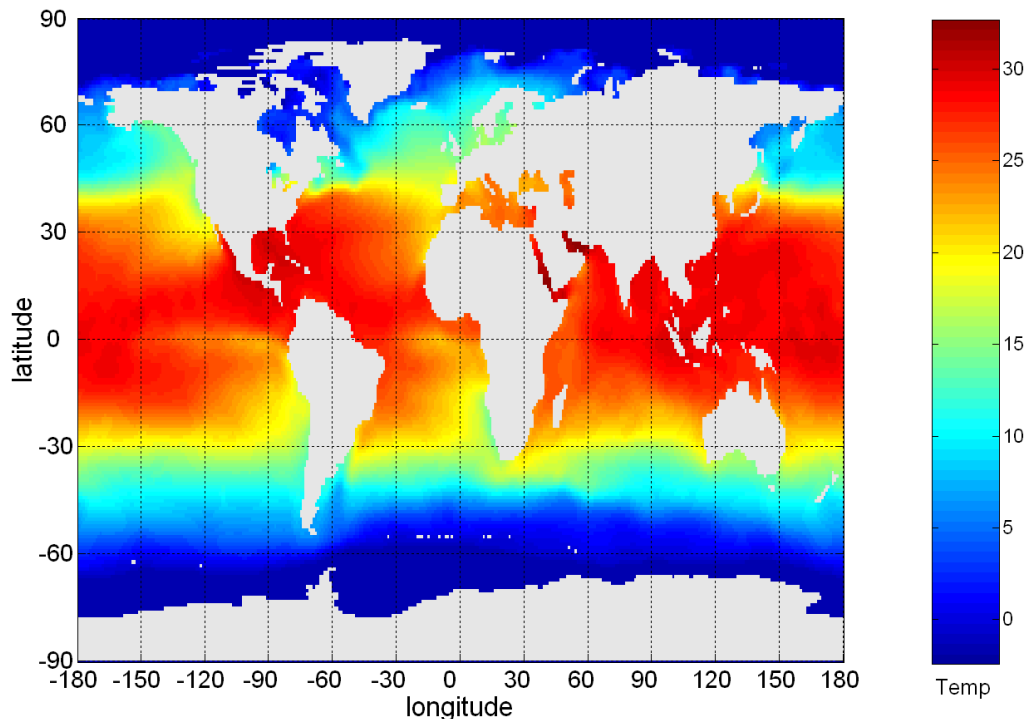
Megjelenítés

Megjelenítés az adatoknak vizuális vagy táblázatos formában való átalakítása a célból, hogy az adatok jellemzői és a közöttük lévő kapcsolat vizsgálható és elmondható legyen.

- Az adatok megjelenítése az adatfeltárás egyik legerősebb, leglátványosabb és legvonzóbb eszköze.
 - Az embernek jól kifejlett képessége, hogy képileg megjelenített nagy tömegű információt elemzzen.
 - Általános mintázatokat, trendeket észlelhetünk.
 - Kiugró értékeket és szokatlan mintázatokat találhatunk.

Példa: Tengerfelszín hőmérséklete

- Az alábbi ábra a tengerek felszínének hőmérsékletét mutatja 1982 júliusában
 - Mintapontok (rekordok) tízezreit összegeztük egy ábrában.



Reprezentáció

- Az információ vizuális formába való leképezése.
- Az objektumok, azok attribútumai és a közöttük lévő kapcsolatok grafikus elemekre, pl. pontokra, vonalakra, alakzatokra, színekre való leképezése.
- Példa:
 - Objektumokat gyakran pontokkal ábrázolunk.
 - Az attribútum értékek ábrázolhatóak a pontok koordinátaival vagy más jellemzőivel, pl. szín, méret, alak.
 - Ha a pozíciót tekintjük, akkor a pontok közötti kapcsolatok, pl. csoportokat alkotnak-e vagy egy pont kiugró-e, már könnyen észrevehető.

Elrendezés

- Vizuális elemek elhelyezése egy képernyőn.
- Nagyban befolyásolja, hogy milyen könnyű az adatainkat megérteni.
- Példa:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Szelekció

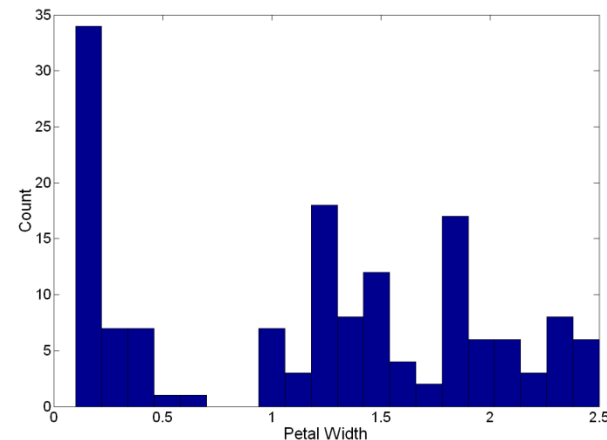
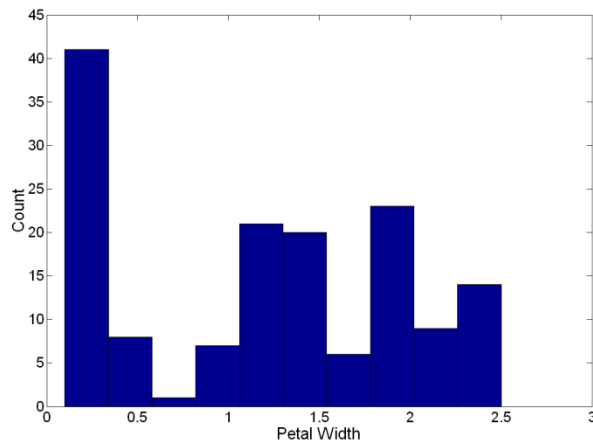
- Egyes objektumok és attribútumok elhanyagolása.
- A szelekció magába foglalhatja attribútumok egy részhalmazának kiválasztását.
 - Gyakran használunk dimenzió csökkentést, hogy a dimenziót kettőre vagy háromra redukáljuk.
 - Más megközelítés: vegyünk attribútum párokat.
- A szelekció magába foglalhatja objektumok egy részhalmazának kiválasztását.
 - A képernyő egyes részei túl sok pontot tartalmaznak.
 - Vegyünk mintát de ügyeljünk arra, hogy a ritkás területeken megtartsuk a pontokat.

Megjelenítési módszerek: hisztogramok

- Hisztogram

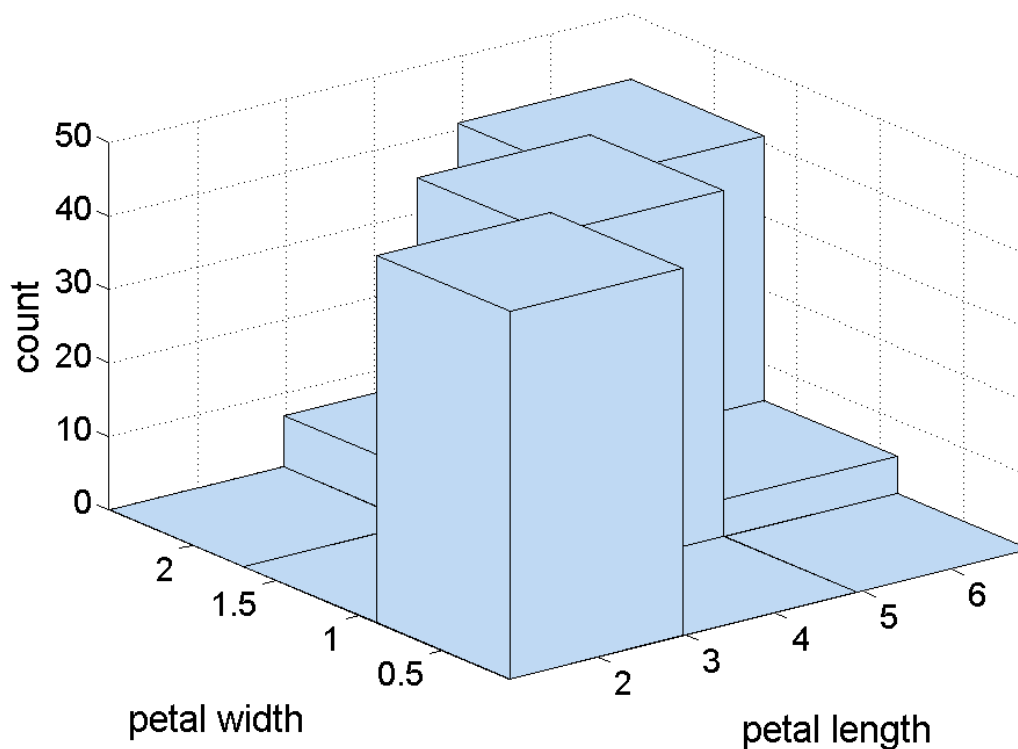
- Egy változó értékeinek eloszlását mutatja.
- Osszuk az értékeket diszjunkt intervallumokba és ábrázoljuk a gyakoriságokat egy oszlopgrafikonon.
- Az oszlopok magassága az intervallumba eső objektumok száma.
- A hisztogram alakja függ a beosztás finomságától.

- Példa: Szirom szélesség (10 illetve 20 beosztással)



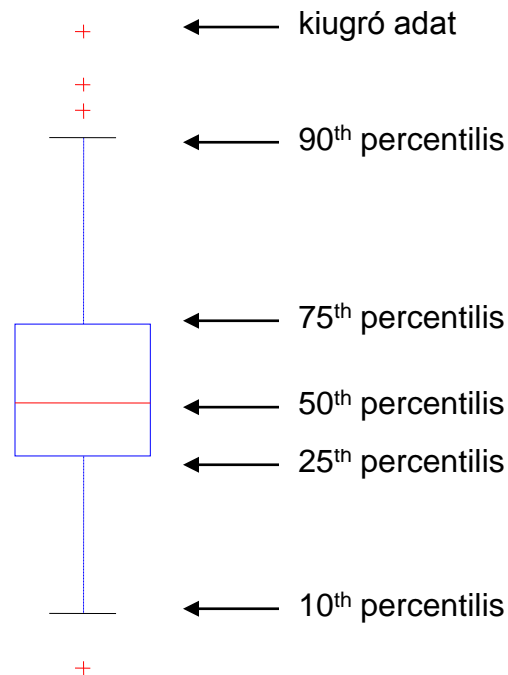
Kétdimenziós hisztogramok

- Két attributum értékeinek együttes eloszlását mutatja.
- Példa: szirm szélesség és szirm hosszúság
 - Mit mond ez nekünk?



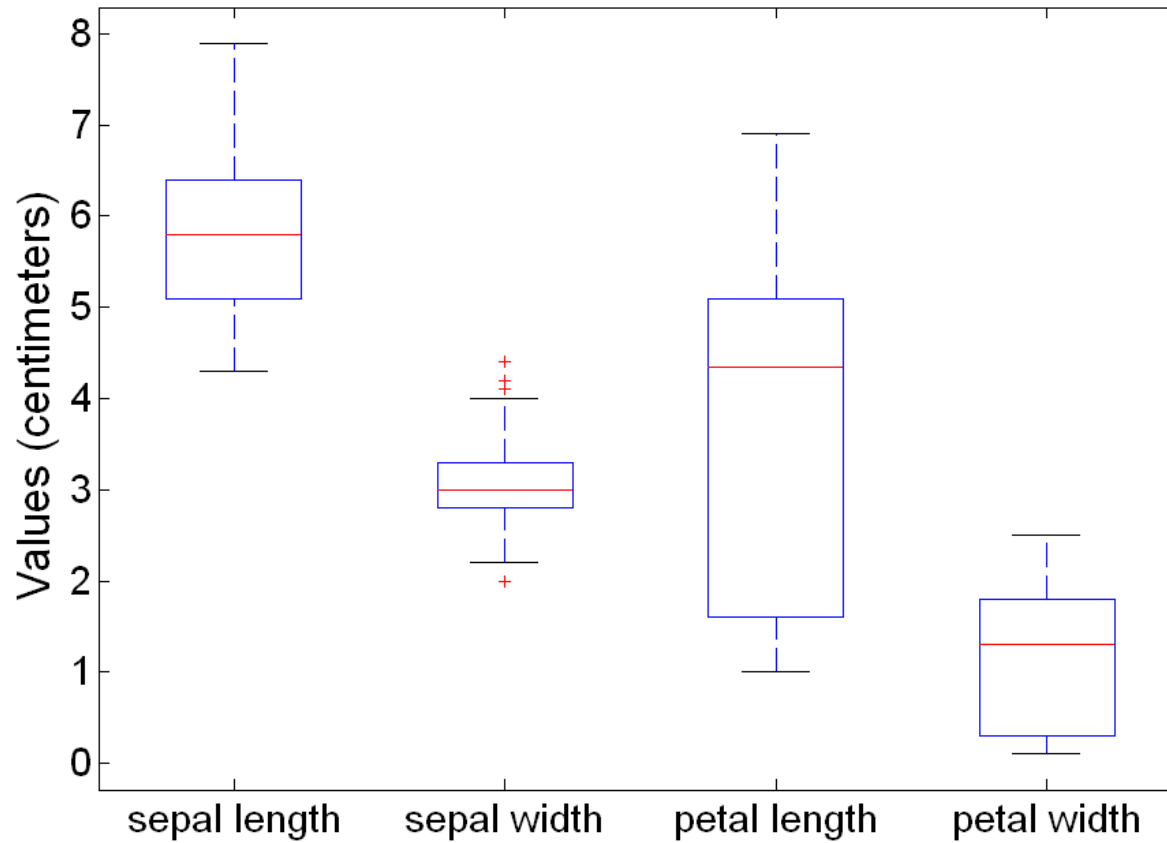
Megjelenítési módszerek: doboz ábra

- Doboz ábra
 - J. Tukey javasolta
 - Az adatok eloszlása szemléltetésének egy másik módja
 - A következő ábra a doboz ábra fő alkotó részeit mutatja



Példa doboz ábrákra

- Attributumok összehasonlítására használható.

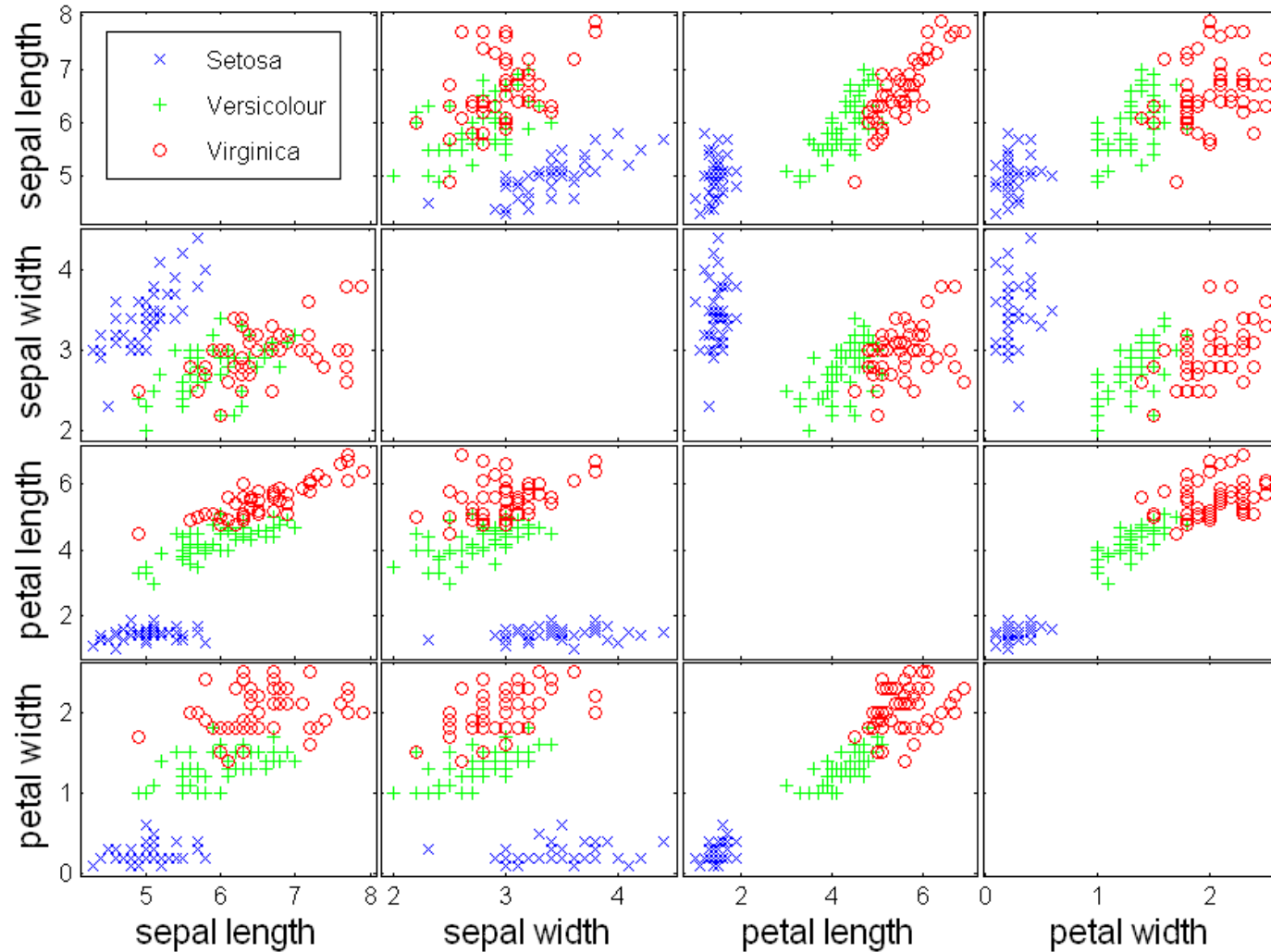


Megjelenítési módszerek : pontdiagram

● Pontdiagram

- Az attributum értékek pontokat határoznak meg a síkban (térben).
- A leggyakoribb a kétdimenziós pontdiagram de használnak háromdimenziós pontdiagramot is.
- Gyakran további attributumokat is meg kell jeleníteni, erre használhatjuk a méret, az alak vagy a szín markereket.
- Sokszor hasznos pontdiagramok egy mátrixát elkészíteni, amely több attributum pár kapcsolatát összegzi kompakt módon.
 - ◆ Lásd a következő oldali példát.

Írisz attributumok pontdiagramjai

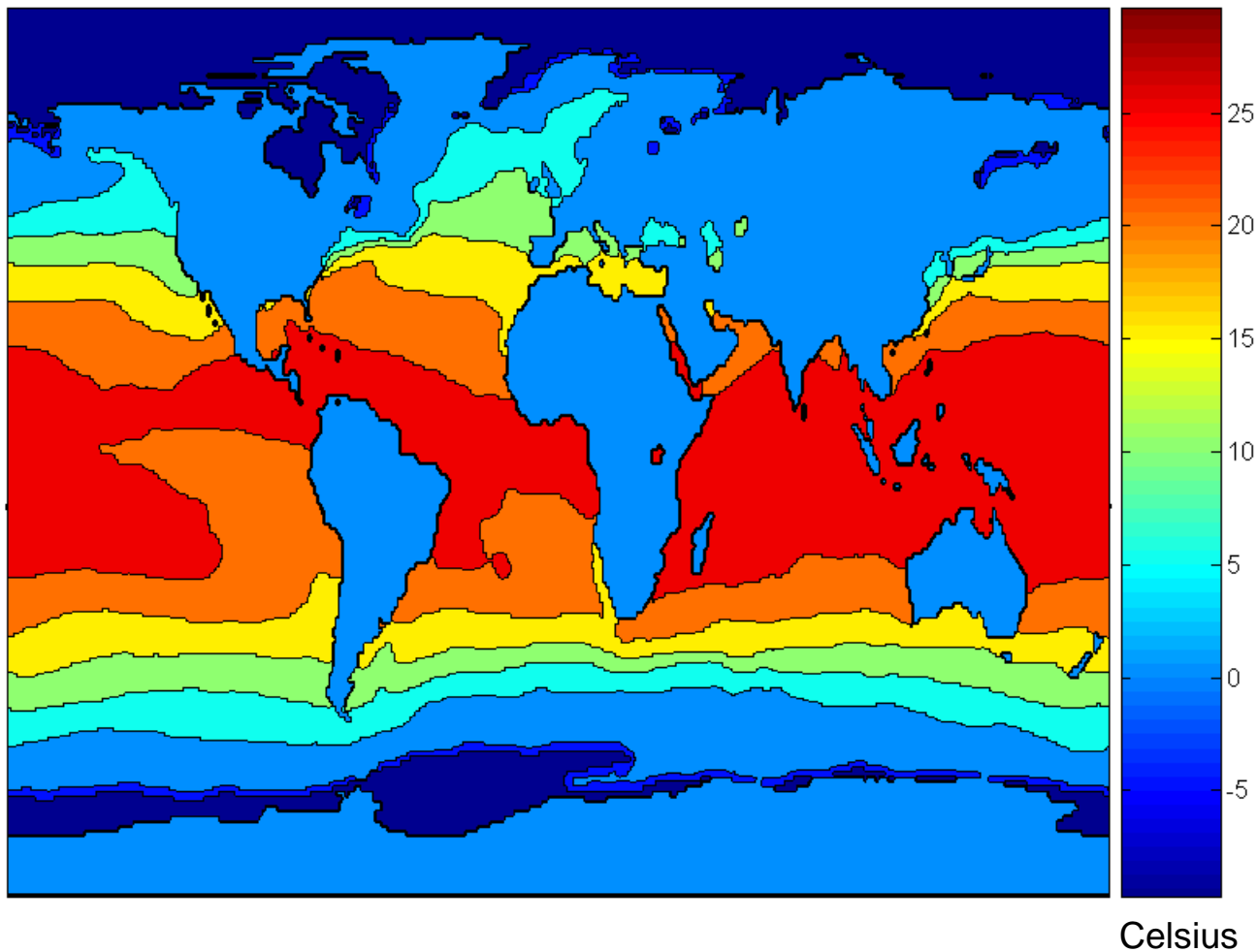


Megjelenítési módszerek : kontúrábra

- Kontúr ábra

- Hasznos amikor egy folytonos attributumot mérünk egy térbeli rácson.
- A síkot tartományokra bontjuk a hasonló értékek alapján.
- A kontúr vonalak, amelyek az egyenlő értékeket kötik össze, alkotják ezeknek a tartományoknak a határait.
- A legismertebb példa a tengerszint feletti magasság domborzati térképeken.
- Szintén megjeleníthetünk így hőmérsékletet, csapadékot, légnyomást stb.
 - ◆ Egy példa látható a következő oldalon.

Példa kontúr ábrára: SST, 1998 dec

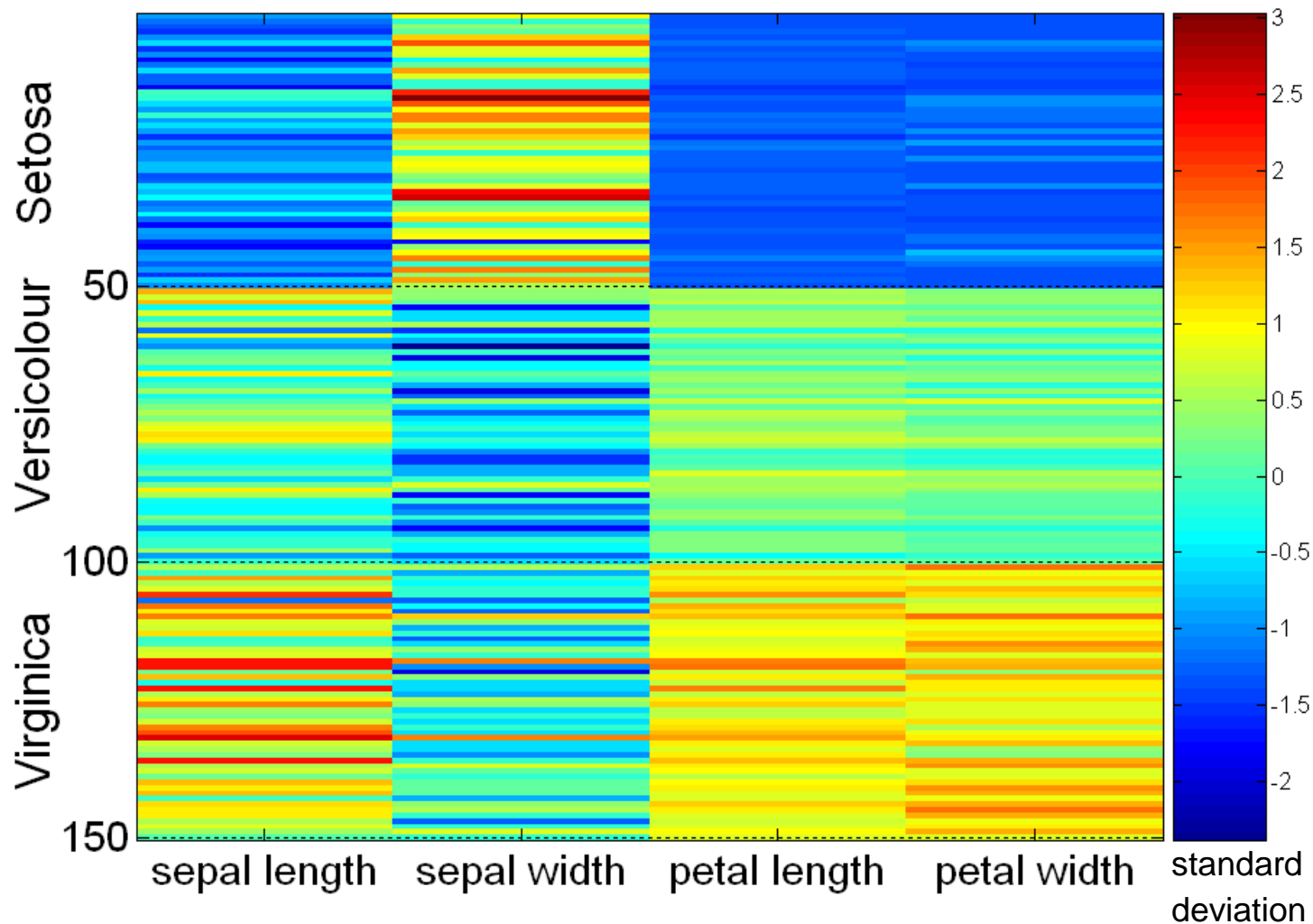


Megjelenítési módszerek : mátrix ábra

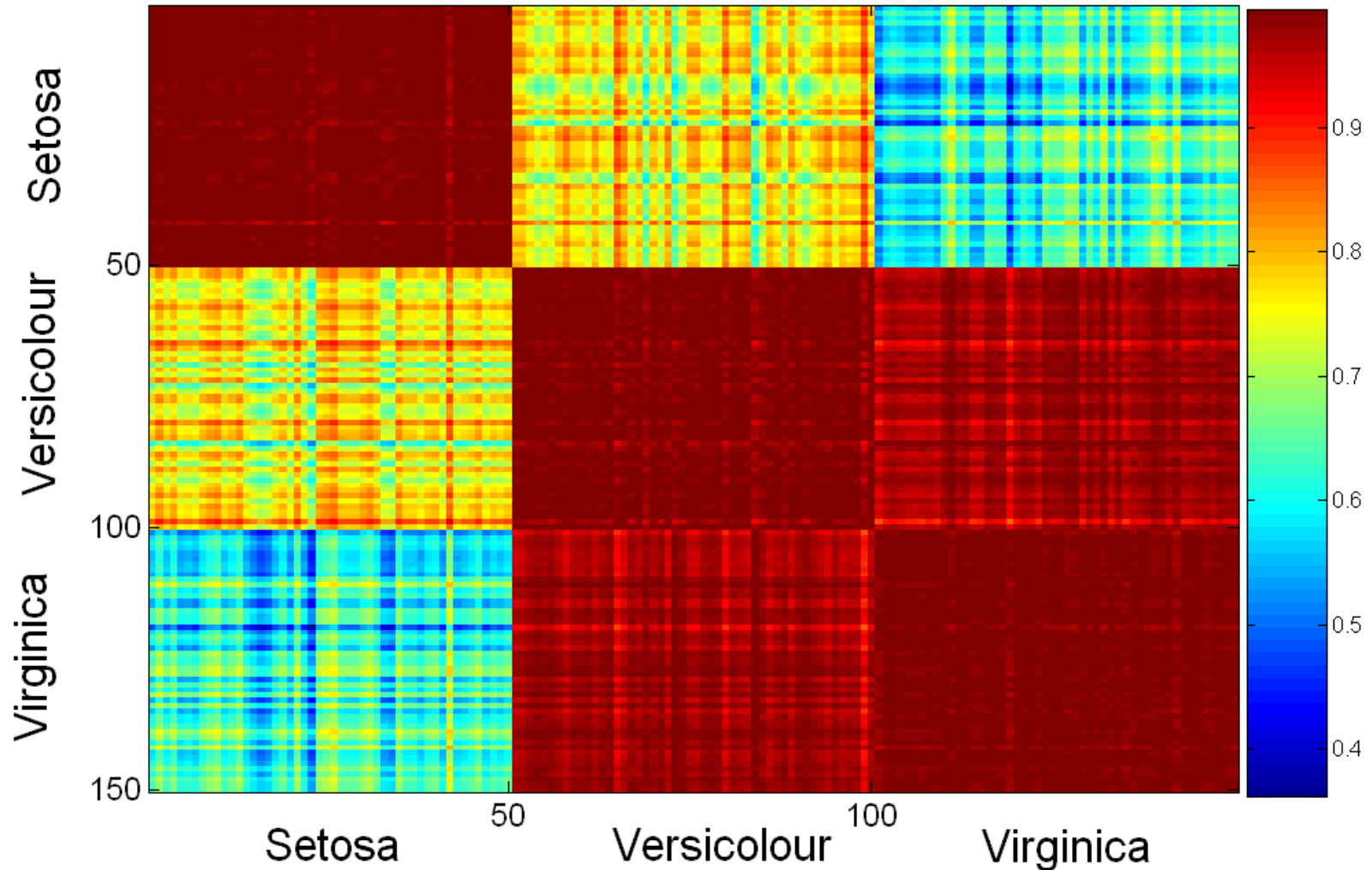
- Mátrix ábra

- Egy teljes adatmátrixot jeleníthetünk meg vele.
- Hasznos amikor az objektumok egy osztályozó változó szerint vannak rendezve.
- Általában az attributumokat normalizálni kell, hogy megelőzzük azt, hogy egy attributum domináljon.
- A hasonlóság és távolságmátrix ábrája szintén hasznos az objektumok közötti kapcsolatok megjelenítésére.
- A következő két oldalon található példa mátrix ábrára.

Az írisz adatmátrix megjelenítése



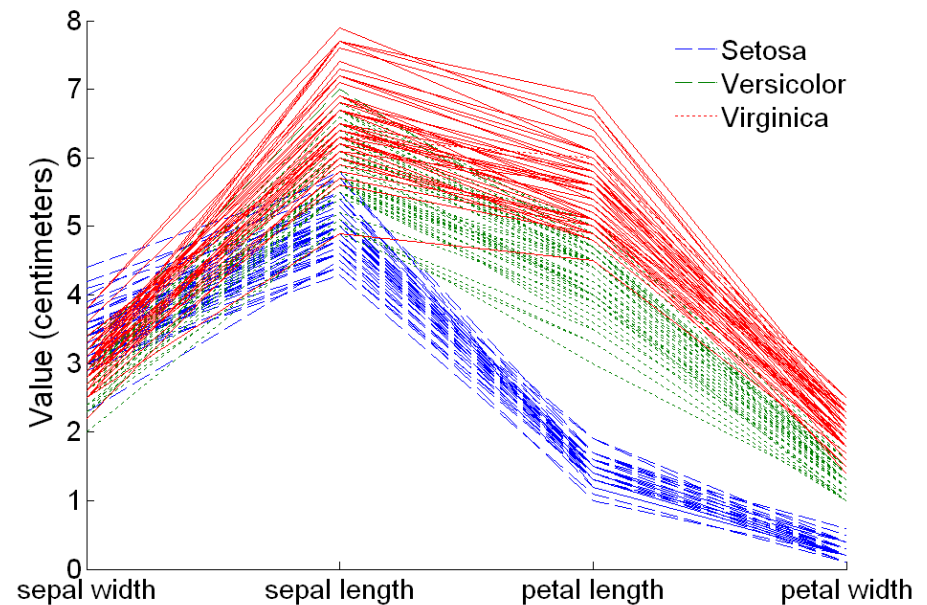
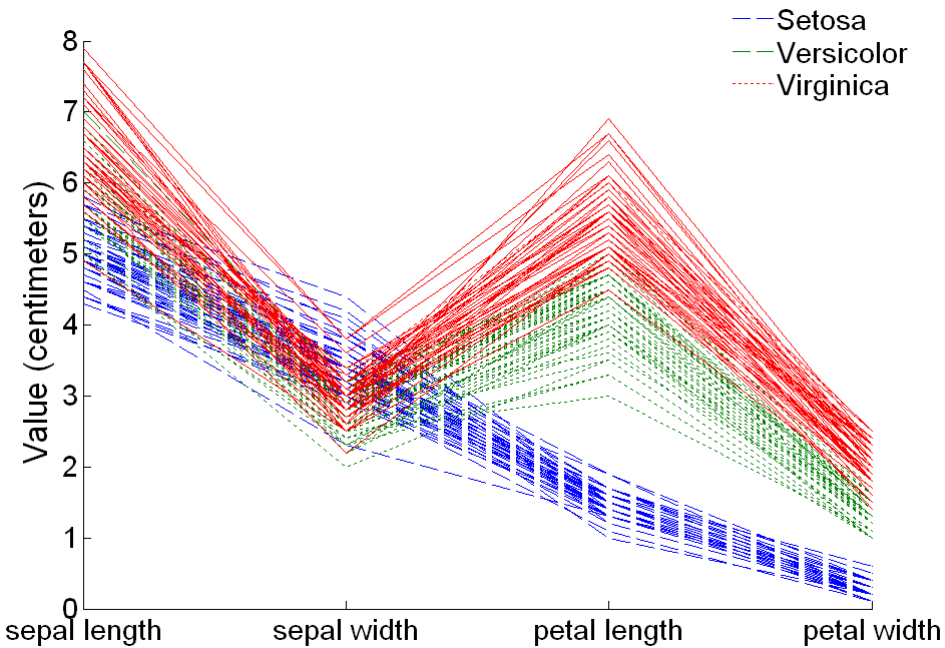
Írisz korrelációs mátrix



Megjelenítési módszerek: párhuzamos tengelyek

- Párhuzamos tengelyek
 - Magas dimenziós adatok attributum értékeinek megjelenítésére szolgál.
 - Merőleges koordinátatengelyek helyett használjunk párhuzamosakat.
 - Minden objektum attributum értékeit a megfelelő koordinátatengelyen egy pontként ábrázolva a pontokat vonallal kötjük össze.
 - Minden objektumot egy vonal reprezentál.
 - Gyakran a vonalak teljesen vagy egyes attributumok mentén csoportosulnak az objektumok különböző csoportjaira utalva.
 - Ennek felismerésére előbb rendezzük az attributumokat.

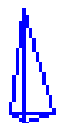
Párhuzamos tengelyek: írisz adatok



További megjelenítési módszerek

- Csillag ábra
 - A párhuzamos koordinátákhoz hasonló azzal az eltéréssel, hogy a koordináták egy centrumból sugarasan indulnak.
 - Egy objektum értékeit összekötő vonalak egy poligont alkotnak.
- Chernoff arcok
 - A módszer Herman Chernoff-tól származik.
 - Az attributumokhoz az arc egy-egy jellemzőjét kapcsoljuk.
 - Minden egyes attributum érték a megfelelő arc-jellemző megjelenését határozza meg.
 - Mindegyik objektum egy külön arc lesz.
 - Az emberek arcfelismerési képességére támaszkodik.

Az írisz adatok csillag ábrája



1



2



3

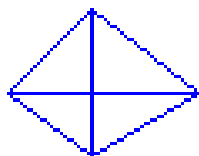


4

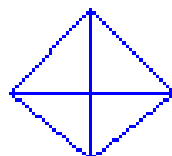


5

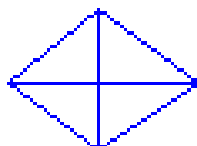
Setosa



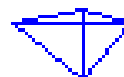
51



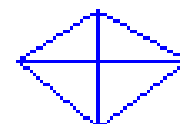
52



53

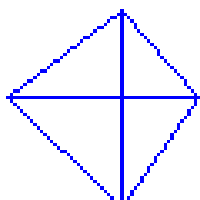


54

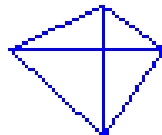


55

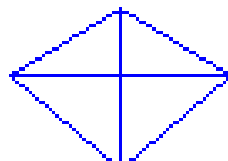
Versicolour



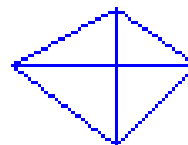
101



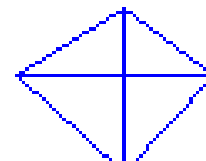
102



103



104



105

Virginica

Chernoff arcok az írisz adatokra



1



2



3

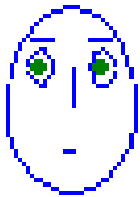


4



5

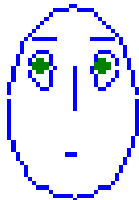
Setosa



51



52



53



54



55

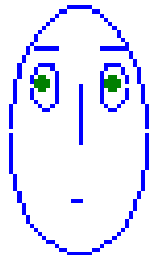
Versicolour



101



102



103



104



105

Virginica

OLAP

- A közvetlen analitikus feldolgozás (OLAP: On-Line Analytical Processing) módszerét E. F. Codd, a relációs adatbázisok atyja javasolta.
- A relációs adatbázisok az adatokat táblákban, míg az OLAP többdimenziós tömbökben tárolja.
 - Az adatok ilyen tárolása már korábban létezett a statisztikában és más területeken.
- Számos olyan adatelemzési és adatfeltárási módszer van, amely ezzel az adattárolási móddal könnyebbé válik.

Többdimenziós tömbök létrehozása

- A táblázatos adatok többdimenziós tömbökké való átalakításának két fő lépése.
 - Először határozzuk meg mely attributumok lesznek a dimenziók és mely attributum lesz a cél attributum, amelynek értékei a többdimenziós tömb elemei lesznek.
 - ◆ A dimenzió attributumoknak diszkrétnek kell lenniük.
 - ◆ A cél attributum általában a darabszám vagy egy folytonos változó, pl. egy tétel költsége.
 - ◆ Előfordulhat, hogy egyáltalán nincs cél attributum csak olyan objektumok darabszáma, melyeknek ugyanazok az attributum értékei.
 - Másodsor számoljuk ki a többdimenziós tömb minden elemének értékét a célattributum értékeinek összegzésével, vagy az összes olyan objektum összeszámolásával, amely attributum értékei megfelelnek az adott elemnek.

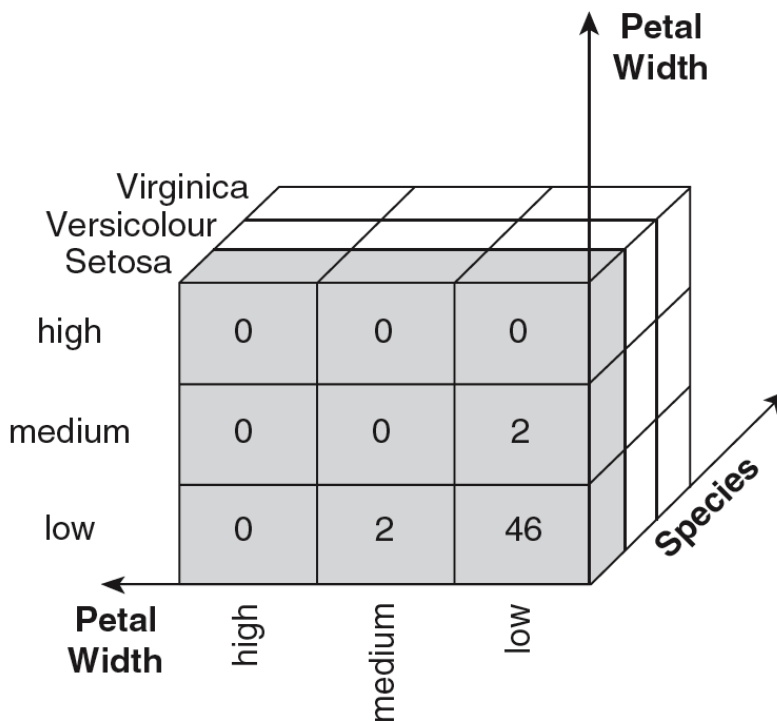
Példa: Írisz adatok

- Megmutatjuk, hogy a virág szélesség és hosszúság és az alfaj attribútumok hogyan alakíthatóak át többdimenziós tömbbé.
 - Először diszkrétizáljuk a virág szélességet és hosszúságot az alábbi értékek szerint: *low*, *medium* és *high*
 - A következő táblázatot kapjuk – a Count (db) új változó

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Példa: írisz adatok (folytatás)

- A virág szélesség és hosszúság és alfaj változók minden egyes egyértelmű rekordja a tömb egy eleme.
- Egy ilyen elemhez hozzárendeljük a megfelelő darabszámot.
- Az ábra mutatja az eredményt.
- Minden nem meghatározott elem 0.



Példa: írisz adatok (folytatás)

- A többdimenziós tömb szeleteit az alábbi kereszt-táblák mutatják.
- Mit mondanak ezek a táblák?

		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

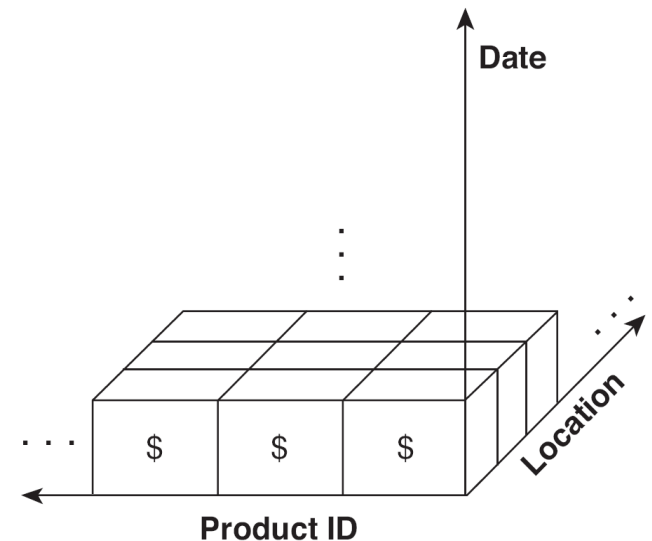
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

OLAP műveletek: adatkocka

- Az OLAP alapvető művelete az adatkocka létrehozása.
- Az adatkocka az adatoknak a többdimenziós megjelenítése az összes lehetséges összesítésükkel.
- Az összes lehetséges összesítés alatt azokat az összesítéseket értjük, melyeket úgy kapunk, hogy kiválasztjuk dimenziók egy részhalmazát és az összes többire összegzünk.
- Például ha az alfaj dimenziót választjuk az írisz adatoknál és az összes többi dimenzió mentén összegzünk, akkor az eredmény egy egydimenziós tömb lesz 3 elemmel, ahol az elemek az egyes alfajba tartozó virágok számát mutatják.

Példa adatkockára

- Tekintsünk egy olyan adatállományt, ahol a rekordok termékek boltokban különböző időpontokban eladott mennyisége.
- Ezek az adatok egy 3 dimenziós tömbbel reprezentálhatóak.
- A kétdimenziós összegzések száma 3 (3 alatt 2), az egydimenziós összegzések száma 3 és 1 db nulla-dimenziós összegzés van (ez a teljes összeg).



Példa adatkockára (folytatás)

- Az alábbi ábra a kétdimenziós összegzések egyikét mutatja két egydimenziós összegzéssel és a teljes összeggel együtt.

product ID	date				total
	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	
1	\$1,001	\$987	...	\$891	\$370,000
:	:			:	:
27	\$10,265	\$10,225	...	\$9,325	\$3,800,020
:	:			:	:
total	\$527,362	\$532,953	...	\$631,221	\$227,352,127

OLAP műveletek: szeletelés és kockázás

- A szeletelés cellák egy olyan csoportjának a kiválasztását jelenti a teljes többdimenziós tömbből, amelyet értékeknek egy vagy több dimenzió menti rögzítésével kapunk.
- A kockázás cellák egy olyan részalmazát jelenti, amelyet attributum értékek egy tartományának megadásával kapunk.
 - Ez ekvivalens azzal, hogy a teljes tömbből egy résztömböt választunk ki.
- A gyakorlatban mindkét művelet együttjárhat bizonyos dimenziók menti összegzéssel.

OLAP műveletek: göngyölítés és lefűrés

- Az attributum értékek gyakran hierarchikusan szerveződnek.
 - Minden dátumhoz tartozik év, hónap és nap.
 - A helyhez tartozik kontinens, ország, megye és település.
 - A termékek különféle osztályokba sorolhatóak, pl. ruházat, elektronika, bútor.
- Ezek az osztályok gyakran beágyazódnak egymásba és fát alkotnak (taxonómia)
 - Az év hónapokból, a hónap napokból áll.
 - Az ország megyéket, a megyék városokat tartalmaz.

OLAP műveletek: göngyölítés és lefűrés

- Ez a hierarchia teszi lehetővé a göngyölítés és lefűrés műveleteket.
 - Az eladási adatokat összegezhetjük (göngyölíthetjük) az összes dátumra egy hónapon belül.
 - Megfordítva egy olyan adattábla esetén, ahol az idő dimenzió hónapokra van bontva, a havi eladásokat bonthatjuk napi szintre (lefűrés).
 - Hasonlóan göngyölíthetünk vagy lefűrhatunk a hely vagy a termék azonosító mentén.