Introducere în Data Mining

Curs 3: Explorarea datelor

Lucian Sasu, Ph.D.

Universitatea Transilvania din Brașov, Facultatea de Matematică și Informatică

April 7, 2014

Outline

Explorarea datelor

- Explorarea datelor reprezintă investigarea preliminară a datelor, cu scopul de a obține o înțelegere a caracteristicilor lor
- Pasul de explorare poate fi de folos în alegerea pașilor de preprocesare sau analiză
- Se poate folosi abilitatea naturală a oamenilor de a recunoaște pattern-uri
- Domeniul a fost introdus de către statisticianul John Tukey: *Exploratory Data Analysis*, Addison-Wesley
- AED este domeniu opus lui "Confirmatory Data Analysis", care are ca scop testarea ipotezelor statistice, calculul intervalelor de încredere etc.
- Curs de AED: aici

Explorarea datelor

- În AED, așa cum este definit de Tukey:
  - Focus-ul este pe vizualizare
  - Gruparea (clustering) și detectarea de anomalii sunt văzute ca tehnici exploratorii
  - Acestea două sunt subdomenii aparte ale DM, dincolo de analiză exploratorie
- Conținutul prezentării:
  - statistici de sumarizare
  - vizualizare
  - On-line Analytical Processing
- Primele două: clasice
- OLAP: util pentru explorarea datelor multidimensionale, cu scopul obținerii de sumarizări: pentru vânzări raportate în forma cantitate, locație, dată, produs, OLAP permite crearea de sumarizări care descriu vânzările pentru un anumit produs/locație/lună
- OLAP este inclus deseori ca auxiliar al SGBD-urilor actuale

Outline

Setul de date Iris

- Setul de date pe care se exemplifică în acest curs: Iris
- Constă în date măsurate pentru 150 de flori de iris, din 3 specii (Iris Setosa, Iris Versicolour, Iris Virginica, câte 50 de exemplare pe specie)
- Măsurătorile sunt pentru lungimea/lățimea petalelor/sepalelor în centimetri (4 coloane)
- A cincea coloană este specia florii – atribut nominal
- Datele se pot descărca de aici

Outline

Statistici de sumarizare

- Statisticile de sumarizare sunt numere care schițează caracteristicile unui set de valori
- Reprezintă manifestarea cea mai vizibilă a statisticii
- Exemple: frecvența, media, dispersia

Frecvența și valoarea modală

- Pentru un set de m date categoriale cu valorile $\{v_1, \ldots, v_i, \ldots, v_k\}$ **frecvența** unei valori v_i este:

$$frecventa(v_i) = \frac{\text{Numărul de obiecte cu valoarea } v_i}{m}$$

- **Valoarea modală** (sau **moda**) este valoarea cu cea mai mare frecvență:

$$moda = \arg\max_{v_i} frecventa(v_i)$$

- Atenție la situația când o anume valoare este folosită pentru a semnifica lipsa datelor: null-ul poate apărea ca modă
- Pot exista seturi de date pentru care frecvența maximă să fie atinsă pentru mai multe valori = seturi multimodale
- Pentru valori continue, conceptele de modă/frecvență nu sunt utile, cu excepția cazului când se aplică un pas de discretizare

Percentile

- Pentru cazul valorilor ordonate se pot considera **percentilele**
- Pentru un atribut continuu sau ordinal x și un număr p întreg între 0 și 100, a p-a percentilă x_p este o valoare din șirul de valori ale lui x astfel încât p% din aceste valori sunt mai mici decât x_p
- Nu există o definiție standardizată pentru percentile, cea de mai sus este luată pentru fixare
- Pentru cazul în care se calculează percentile pentru set mare de date, diferențele datorate diferitelor moduri de definire devin neesențiale
- Tradițional se consideră $x_{0\%} = \min(x)$ iar din definiție se poate arăta că $x_{100\%} = \max(x)$
- Mod de calcul pentru determinarea celei de a p-a percentile: pentru un set de n date se calculează valoarea întreagă k cea mai apropiată de $\frac{n}{100}p + \frac{1}{2}$ și se ia valoarea corespunzătoare acestui rang k în șirul x sortat

Măsurarea locației: media și mediana

- Pentru un set de valori $\{x_1, x_2, \ldots, x_m\}$ valoarea medie este:

$$\overline{x} = media(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Pentru aflarea medianei este nevoie să se facă sortarea valorilor inițiale, obținându–se mulțimea (permutarea) $\left\{x_{(1)}, x_{(2)}, \ldots, x_{(m)}\right\}$; mediana este

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

Măsurarea locației: media și mediana

- Media este valoare de mijloc doar dacă distribuția datelor este simetrică
- Dacă distribuția este asimetrică, atunci mediana este un indicator mai bun pentru valoare de mijloc
- Media este influențată de outliers, în timp ce mediana – nu
- Medie retezată (eng: trimmed mean) se utilizează pentru a exlude anomaliile: se fixează un procent p între 0 și 100; se elimină primele și ultimele $(p/2)\%$ din date; se calculează media pentru ceea ce rămâne
- media standard se obține din media retezată cu $p = 0$

Măsurarea locației: media și mediana

- Exemple:
  - Considerăm valorile $\{1, 2, 3, 4, 5, 90\}$. Media este 17.5, mediana este 3.5. Valoarea de trimmed mean pentru $p = 40\%$ este 3.5, considerabil diferită față de media setului întreg de date
  - Media, medianele și valoarea de trimmed mean pentru iris sunt:

Măsura	Lungimea sepalelor	Lungimea sepalelor	Lungimea petalelor	Lungimea petalelor
Media	5.84	3.05	3.76	1.20
Mediana	5.80	3.00	4.35	1.30
Trimmed mean (20%)	5.79	3.02	3.72	1.12

Exercițiu: dacă valoarea medianei este mai mică decât media, ce puteți spune despre date?

Măsurari ale împrăștierii datelor

- Sunt măsuri care cuantifică concentrarea datelor
- Diametrul domeniului de valori (eng: **range**) al unui set de date $\{x_1, x_2, \ldots, x_m\}$ corespunzător atributului x este

$$range(x) = \max(x) - \min(x) = x_{(m)} - x_{(1)}$$

- Range-ul este nerelevant, deoarece putem avea că majoritatea datelor sunt concentrate într–o zonă îngustă, dar câteva valori outlier măresc artificial raza setului
- **Varianța** (**dispersia**) unui set de date de m valori este:

$$varianta(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \overline{x})^2$$

- Utilizarea numitorului $m-1$ în loc de m este numită Corecția Bessel și are ca scop corectarea abaterii din estimarea variației de populație

Măsurari ale împrăștierii datelor

- Abaterea standard este $s_x = \sqrt{s_x^2}$ și are aceeași unitate de măsură ca și atributul x
- Deoarece media poate să fie distorsionată de outliers, rezultă că dispersia poate fi și ea influențată
- Se preferă considerarea altor trei măsuri:
  - absolute average deviation, AAD:

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

  - median absolute deviation, MAD

$$MAD(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

  - interquartile range

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Statistici de sumarizare a datelor multivariate

- Date multivariate: date cu mai multe atribute
- Pentru atributul x_i calculăm media $\overline{x_i}$
- Media setului de obiecte este $\overline{\mathbf{x}} = (\overline{x_1}, \ldots, \overline{x_n})$
- Analog se poate calcula dispersia, mediana etc. pe fiecare dimensiune
- **Matricea de covarianță**: elementul s_{ij} de pe linia i și coloana j este covarianța atributelor x_i și x_j:

$$s_{ij} = covarianta\,(x_i, x_j) = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})$$

  unde x_{pq} este a p-a valoare a atributului x_q
- s_{ij} este măsură a gradului în care două atribute variază împreună (mai precis: care este gradul lor de dependență liniară) și depinde de magnitudinea valorilor atributelor

Statistici de sumarizare a datelor multivariate

- $s_{ij} = 0$ înseamnă că atributele s_i și s_j nu sunt liniar dependente
- **Matrice de corelație**:

$$r_{ij} = corelatia(x_i, x_j) = \frac{covarianta(x_i, x_j)}{s_i s_j} \in [-1, 1]$$

- r_{ij} se mai numește corelația Pearson a atributelor x_i și x_j
- $r_{ij} = \pm 1$ indică faptul că x_i este în relație liniară cu x_j: $x_{ki} = a \cdot x_{kj} + b$ cu $sgn(a) = sgn(r_{ij})$

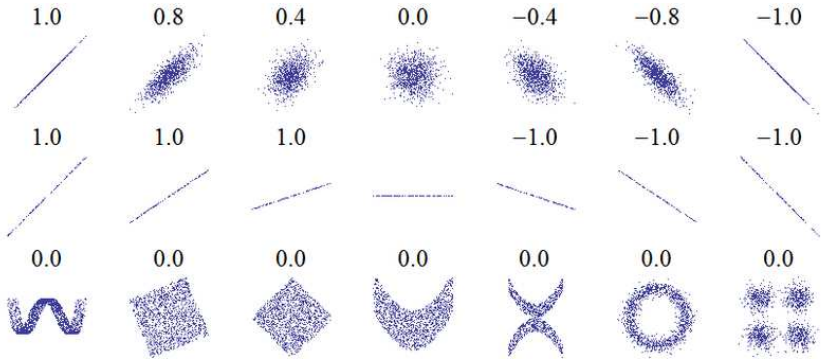


Figure 1: Seturi de date (x, y) împreună cu coeficientul de corelație. Coeficientul de corelație surprinde gradul în care un nor de puncte poate fi aproximat printr-o dreaptă (sus) precum și modul în care ele sunt legate liniar (creștere simultană sau evoluții în sensuri diferite), dar nu și panta acestei legături (figurile din mijloc) sau relații mai complexe între date (rândul de jos). Sursa: Wikipedia.

Statistici de sumarizare a datelor multivariate

Legat de coeficientul de corelație, câteva observații :

- "Corelația nu înseamnă cauzalitate" – nu se poate folosi o valoare absolută apropiată de 1 ca argument că între două atribute există o relație de cauzalitate. Corelație mare poate fi o condiție necesară pentru legătură de cauzalitate, dar nu asigură și suficiența. Cu toate acestea, corelația mare poate fi folosită ca punct de pornire în cercetarea unei legături între diferite fenomene.
- Corelația și liniaritatea – coeficientul Pearson reprezintă puterea unei relații liniare între două seturi de valori, dar nu caracterizează complet relația dintre date.
- Exemplu: 4 seturi de date cu două atribute; în toate situațiile media și dispersia lui y este aceeași, de asemenea avem același coeficient de corelație în fiecare caz (0.816); cu toate acestea, legătura dintre x și y e extrem de diferită de la un caz la altul.

Statistici de sumarizare a datelor multivariate

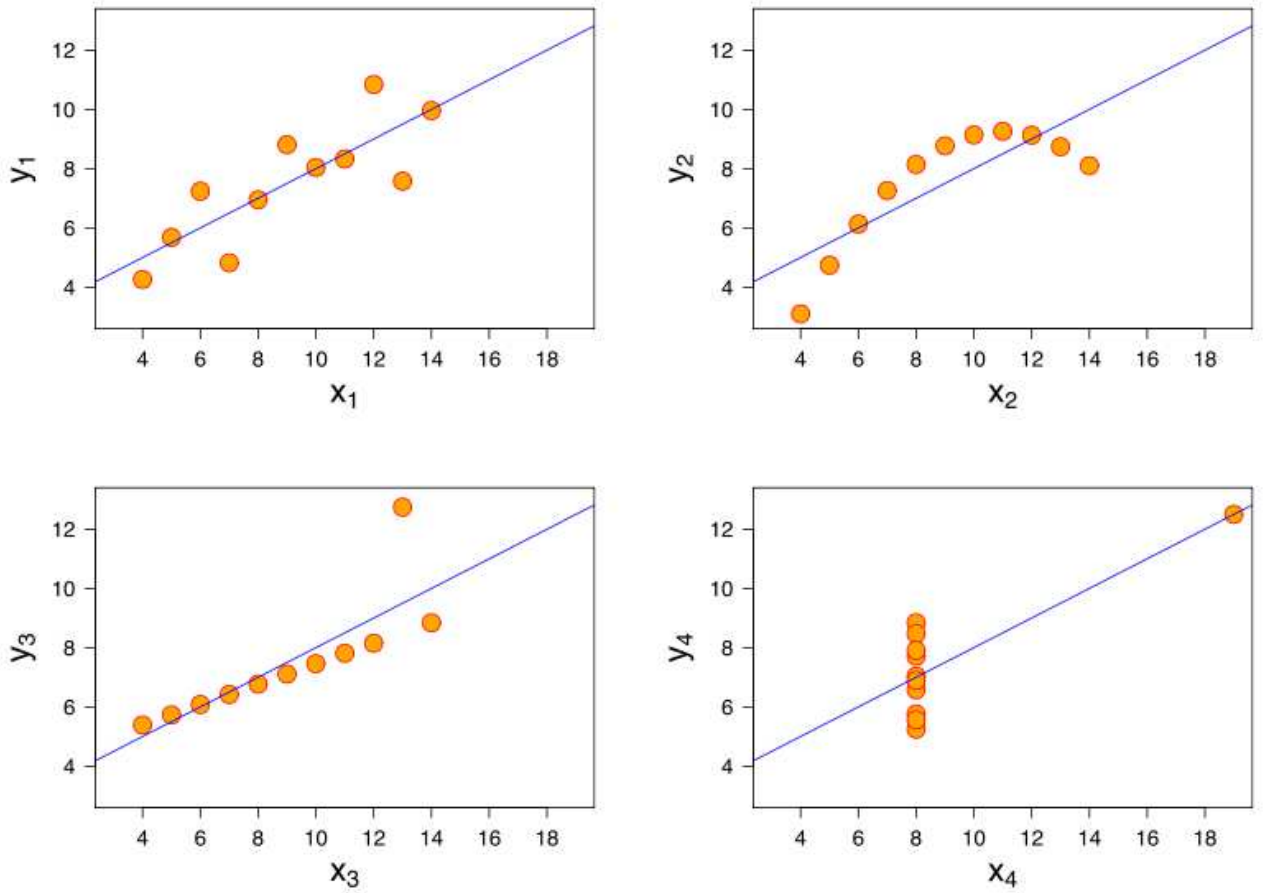


Figure 2: Date cu caracteristici numerice identice (medie, dispersie, corelație), dar esențial diferite ca natură: cvartetul lui Anscombe. Sursa: Wikipedia

Outline

Vizualizare

- Scopul vizualizării: reprezentarea informației într–un mod tabular sau grafic
- Caracteristicile datelor și relațiile dintre elemente pot fi analizate sau raportate
- Calități:
  - oamenii au o abilitate naturală de analiză pentru cantități mari de date prezentate vizual
  - oamenii pot detecta relativ ușor șabloane și tendințe
  - se pot detecta ușor outliers și grupări neobișnuite
- Altă utilizare: reprezentare a datelor obținute după analiză și confruntarea cu cunoștințele unor experți umani sau se pot elimina pattern-urile neinteresante

Vizualizare - exemplu

Exemplu: date reprezentând temperatura la suprafața apei în Iulie 1982 = zeci de mii de valori.

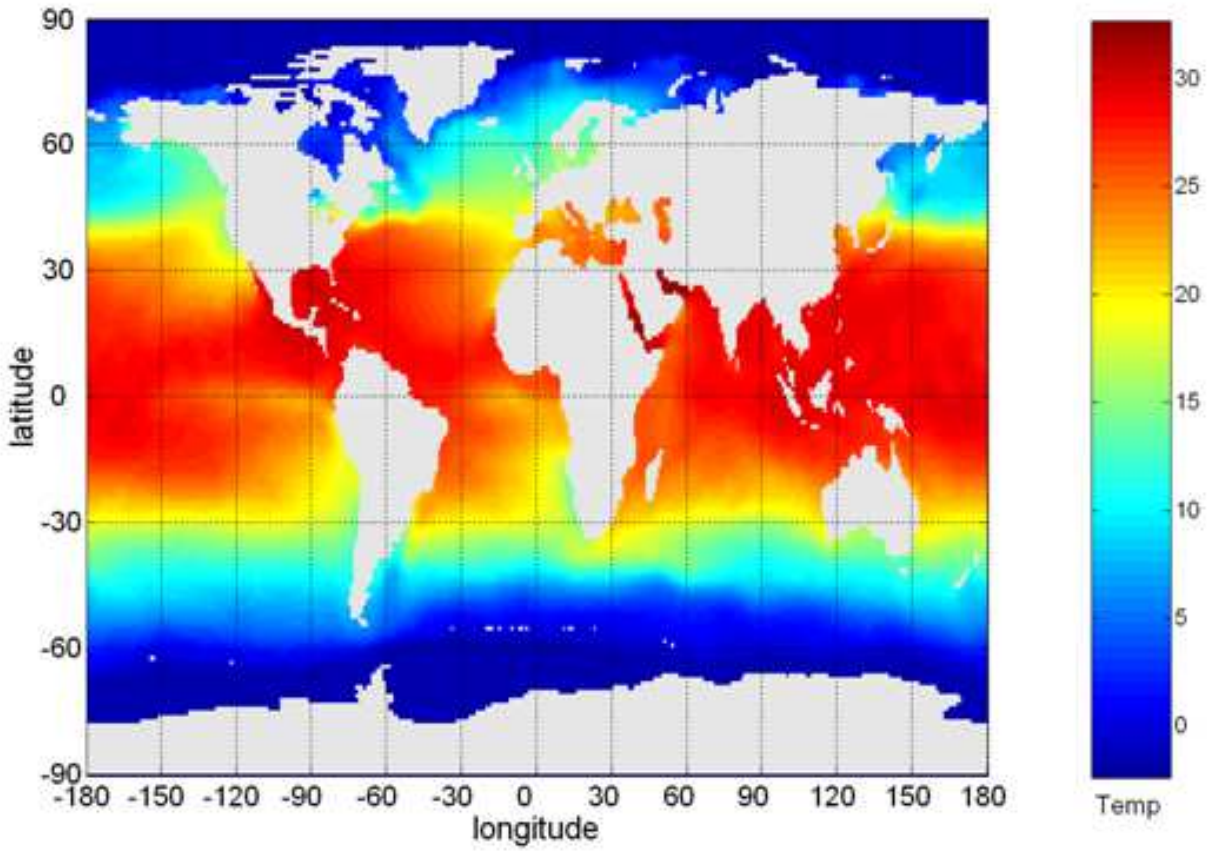


Figure 3: Rezultat ușor de înțeles și recunoscut: cu cât te îndepărtezi de ecuator, cu atât temperatura scade.

Vizualizare - reprezentarea

- Reprezentare = asocierea datelor cu elemente grafice
- Rezultat: obiectele, atributele și relațiile dintre ele sunt transformate în elemente grafice (puncte, linii, forme, culori)
- Exemple:
  - Obiectele sunt deseori reprezentate ca puncte în spațiul 2D sau 3D
  - Atributele pot fi asociate cu poziția punctelor sau cu atribute ale lor: culoare, formă, dimensiune
  - Dacă se folosește poziția punctelor atunci se poate percepe ușor o relație de grupare, disimilaritate sau un outlier

Vizualizare - aranjarea

- Se referă la plasarea elementelor vizuale pe display
- Rearanjarea datelor și a a atributelor poate să fie la fel de importantă ca alegerea reprezentării în sine
- Exemplu: reordonarea de atribute și obiecte

	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

Figure 4: Un tabel cu nouă obiecte și șase atribute binare.

	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

Figure 5: După efectuarea de permutări de obiecte și atribute, gruparea obiectelor în funcție de valori devine vizibilă.

Vizualizare - selectarea

- Selectarea = eliminarea sau deaccentuarea obiectelor sau a atributelor
- Beneficii: selectarea atributelor poate permite reprezentarea lor 2D sau 3D; eliminarea de înregistrări poate duce la obținerea unei reprezentări inteligibile
- Exemplu: se pot alege perechi de atribute care să se reprezinte grafic; dacă nu sunt prea multe atribute, atunci se pot reprezenta toate perechile de atribute
- Există și alte metode mai sofisticate de selectare a atributelor: analiza componentelor principale
- Eliminarea de obiecte: se poate face prin eșantionare, dar cu păstrarea datelor în regiuni slab populate; sau concentrarea doar pe un anumit subset al colecției inițiale (*e.g.* o clasă de obiecte: Iris Setosa)

Vizualizare - tehnici

- Metodele de vizualizare sunt deseori specializate pe tipurile de date
- Există și tehnici clasice ce sunt specializate după:
  - numărul de atribute
  - existența de legături de tip ierarhic sau graf între date
  - tipurile de atribute

Vizualizare: stem and leaf

- Stem and leaf (sau stemplot): utilă pentru reprezentarea distribuției de date întregi sau continue unidimensionale
- Mod de lucru pentru valori întregi: se împart valorile în grupuri, unde fiecare grup conține valori care sunt egale, abstracție făcând de ultima cifră
- Tulpinile sunt grupurile, iar frunzele sunt cifrele unităților
- Exemplu: pentru valorile 35, 36, 42, 51 avem tulpinile 3, 4, 5 iar frunzele sunt respectiv $\{5, 6\}$, $\{2\}$ și $\{1\}$.
- Reprezentare:

3	56
4	2
5	1

Vizualizare: stem and leaf

- Pentru Iris considerăm atributul 'lungimea sepalei' cu valorile înmulțite cu 10; se obține:

43, 44, 44, 44, 45, 46, 46, 46, 46, 47, 47, 48, 48, 48, 48, 48, 49, 49, 49, 49, 49, 49, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 51, 51, 51, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 53, 54, 54, 54, 54, 54, 54, 55, 55, 55, 55, 55, 55, 55, 56, 56, 56, 56, 56, 56, 57, 57, 57, 57, 57, 57, 57, 57, 58, 58, 58, 58, 58, 58, 58, 59, 59, 59, 60, 60, 60, 60, 60, 60, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64, 64, 64, 64, 64, 65, 65, 65, 65, 65, 66, 66, 67, 67, 67, 67, 67, 67, 67, 68, 68, 68, 69, 69, 69, 69, 70, 71, 72, 72, 72, 73, 74, 76, 77, 77, 77, 77, 79

- Reprezentarea prin stem and leaf duce la:

4	3444456666778888899999
5	000000000011111111122223444444555555566666677777778888888999
6	000000111111222233333333344444445555566777777778889999
7	0122234677779

- Utilitate:

Vizualizare: stem and leaf

- Pentru Iris considerăm atributul 'lungimea sepalei' cu valorile înmulțite cu 10; se obține:

43, 44, 44, 44, 45, 46, 46, 46, 46, 47, 47, 48, 48, 48, 48, 48, 49, 49, 49, 49, 49, 49, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 51, 51, 51, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 53, 54, 54, 54, 54, 54, 54, 55, 55, 55, 55, 55, 55, 55, 56, 56, 56, 56, 56, 56, 57, 57, 57, 57, 57, 57, 57, 57, 58, 58, 58, 58, 58, 58, 58, 59, 59, 59, 60, 60, 60, 60, 60, 60, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64, 64, 64, 64, 64, 65, 65, 65, 65, 65, 66, 66, 67, 67, 67, 67, 67, 67, 67, 68, 68, 68, 69, 69, 69, 69, 70, 71, 72, 72, 72, 73, 74, 76, 77, 77, 77, 77, 79

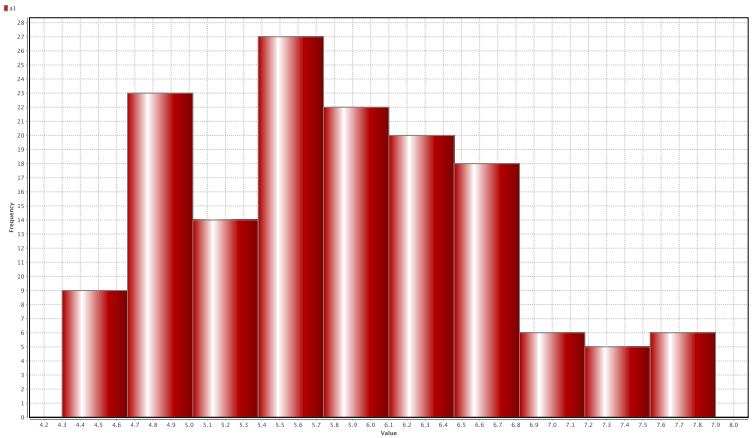
- Reprezentarea prin stem and leaf duce la:

4	34444566667788888999999
5	00000000001111111112222344444455555556666667777777888888999
6	00000011111122223333333334444444555556677777778889999
7	0122234677779

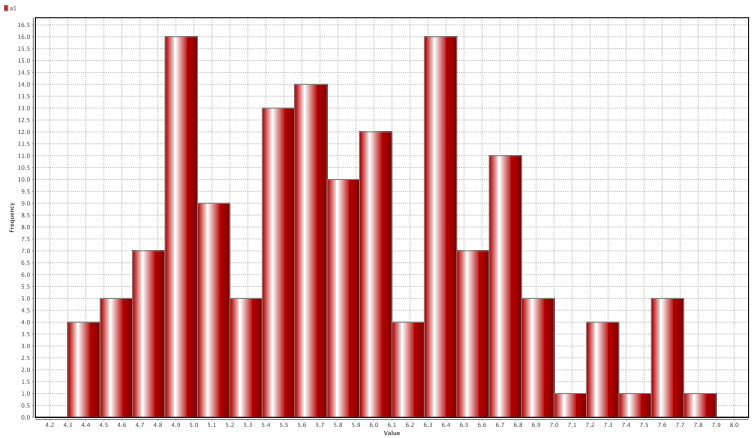
- Utilitate:
  - se poate vizualiza rapid densitatea relativă datelor; e.g. grupul cel mai numeros este între 5 și 6 cm.
  - se pot vedea rapid valorile outlier
- Restricție: pentru date în cantitate moderată, până la 200 de obiecte

Vizualizare: histograme

- Domeniul de valori este împărțit în subintervale; pentru fiecare subinterval se contorizează câte valori sunt incluse în el
- Pentru valori categoriale contorizarea se face pentru fiecare valoare; dacă sunt prea multe valori categoriale, atunci acestea se combină cumva
- Se construiește câte un dreptunghi aferent fiecărui interval/categorie cu înălțimea proporțională cu numărul de valori



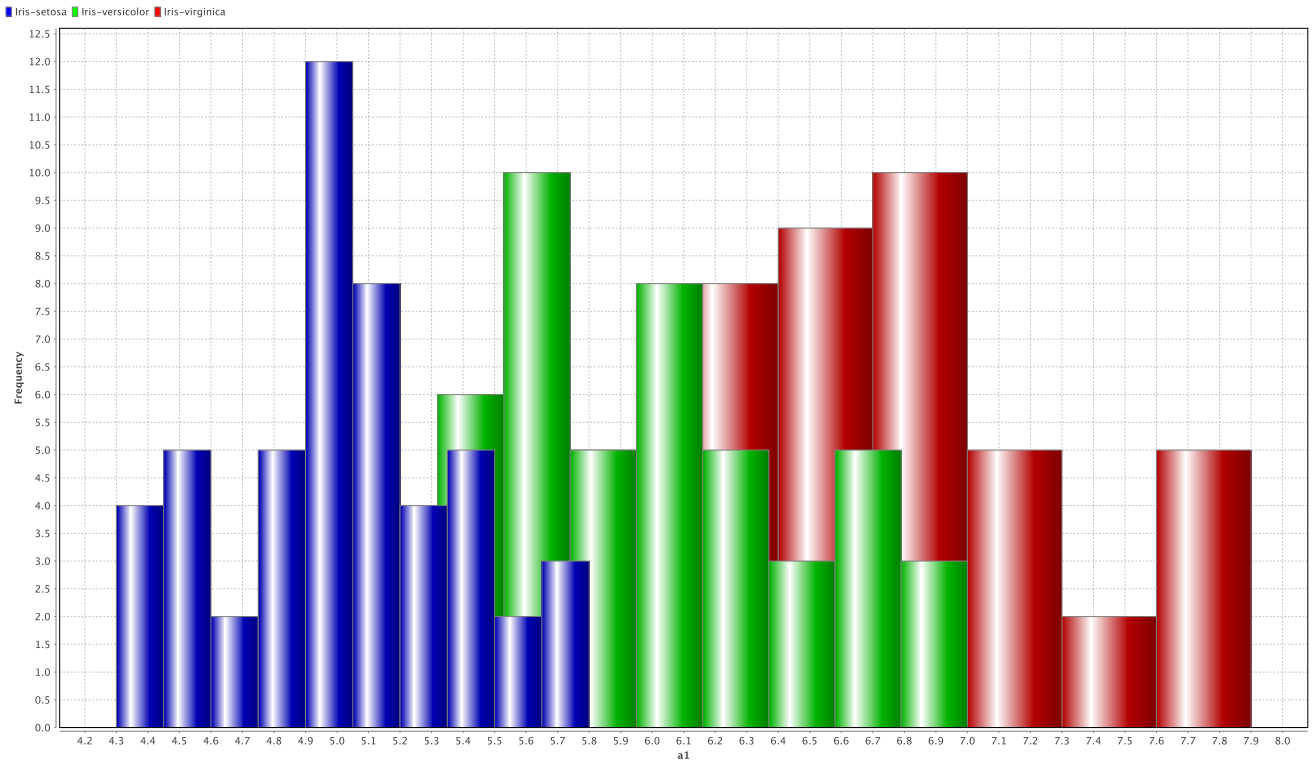
(a) Lungimea sepalelor, discretizare în 10 subintervale



(b) Lungimea sepalelor, discretizare în 20 de subintervale

Vizualizare: histograme

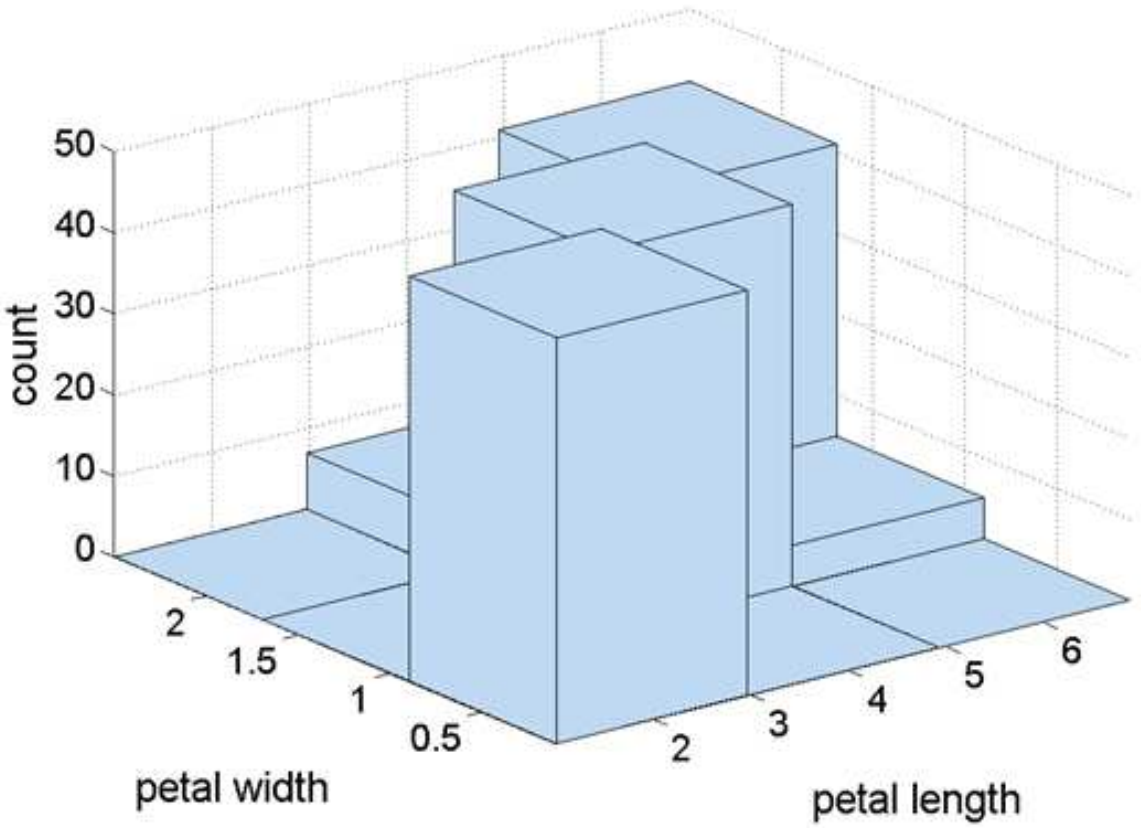
- Se pot reprezenta mai multe valori simultan pe o histogramă:



- Pentru cazul datelor categoriale, histograma Pareto este la fel cu histograma normală, dar categoriile sunt sortate în descrescător după numărul de obiecte conținute

Vizualizare: histograme bidimensionale

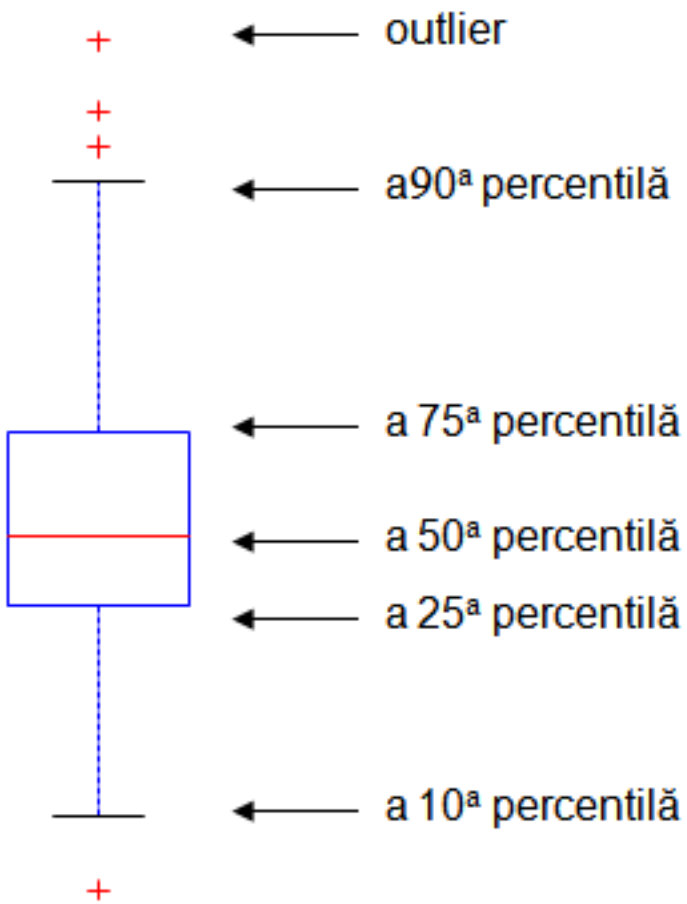
- Conțin contorizări pentru două dimensiuni
- Exemplu: lungimea și lățimea petalelor



- Ce arată histograma de mai sus? ce probleme pot fi la reprezentare?

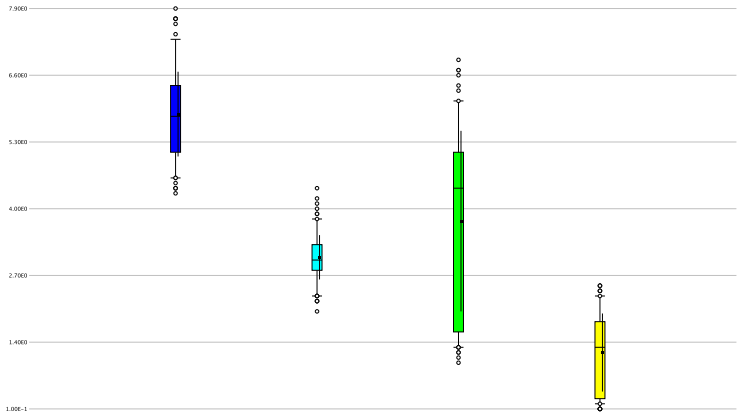
Vizualizare: boxplots

- Introduse de J. Tukey
- Arată distribuția valorilor pentru un singur atribut numeric
- Figura de mai jos explică componentele unui boxplot

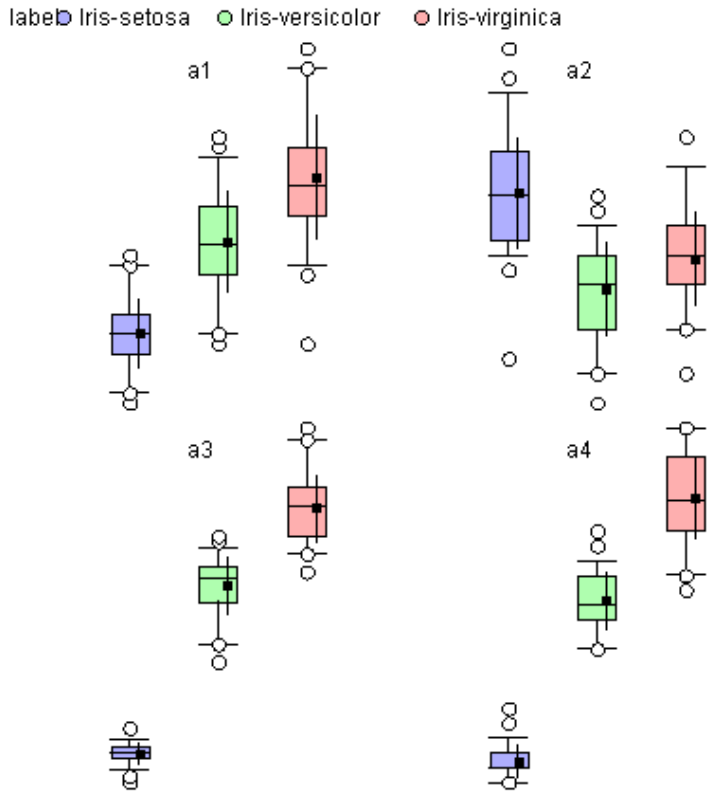


Vizualizare: boxplots

- Se pot compara mai multe serii de date
- a1=lungimea sepalei, a2=lățimea sepalei, a3=lungimea petalei, a4=lățimea petalei



(a) Boxplot pentru cele patru atribute ale setului de date Iris



(b) Matrice de boxplots

Vizualizare: pie charts

- Folosite de regulă pentru atribute categoriale cu puține valori distincte
- Ariile dau o idee asupra repartizării datelor în categorii
- Des folosite în lucrări de popularizare sau de raportare
- Rar folosite în scrierile tehnice, tocmai din cauză că e greu să se judece și să se compare aria zonelor
- În scrieri tehnice se preferă histogramele

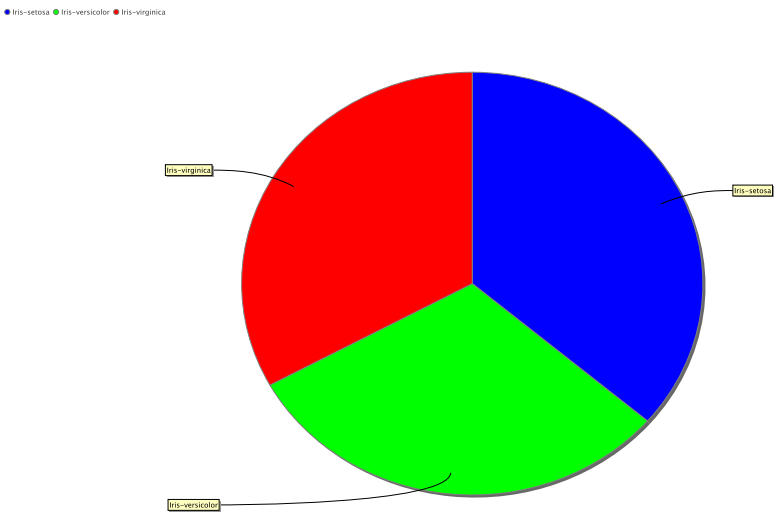


Figure 6: Piechart

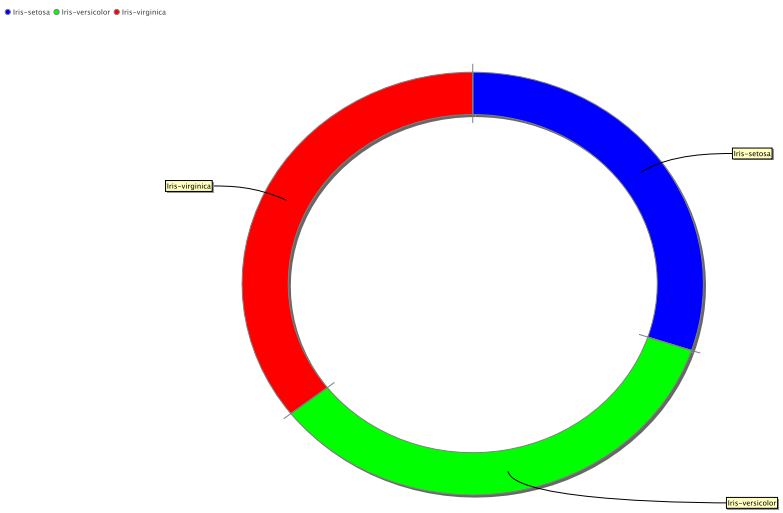
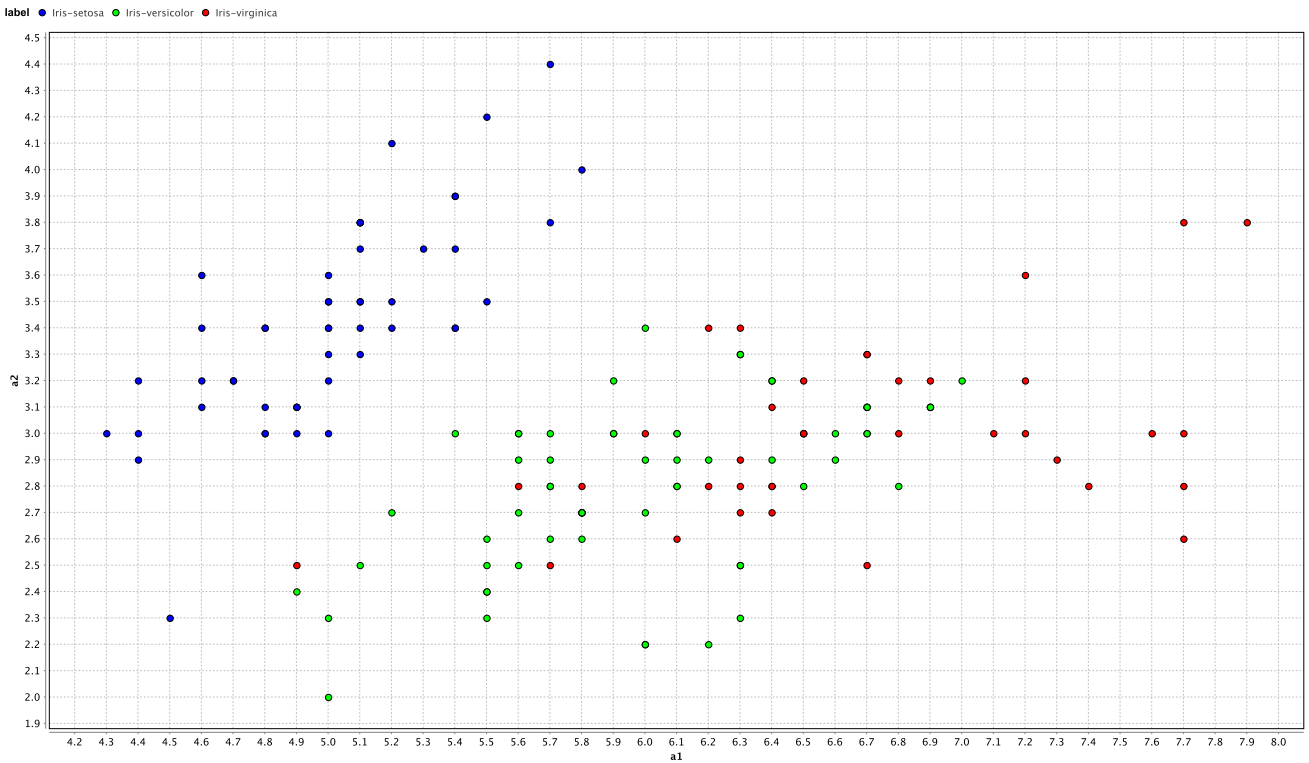


Figure 7: Ring

Vizualizare: Scatter plots

- Valorile atributelor determină poziția în plan
- Cel mai des folosite: scatter plots 2D, dar se pot realiza și 3D
- Atribute adiționale pot fi reprezentate folosind culori, forme, dimensiuni ale obiectelor grafice
- Cel mai des folosite: matrice de scatter plots care reprezintă perechi de atribute



Vizualizare: Matrix scatter plots

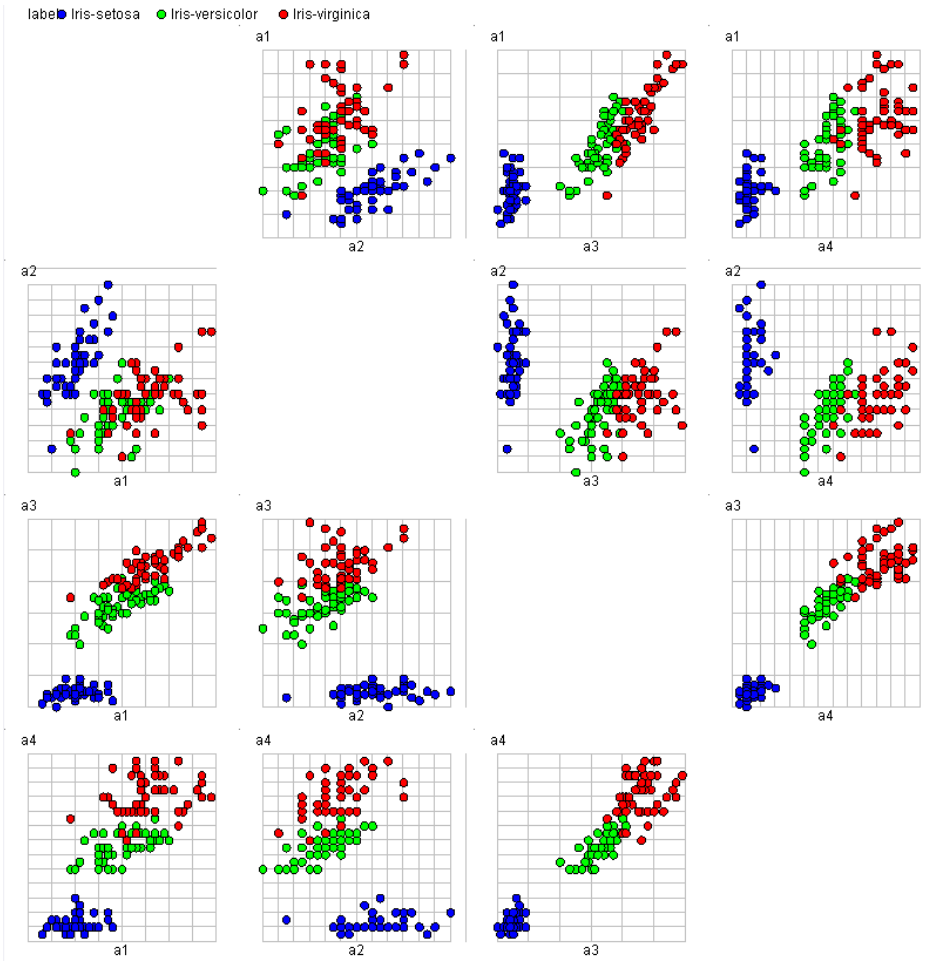
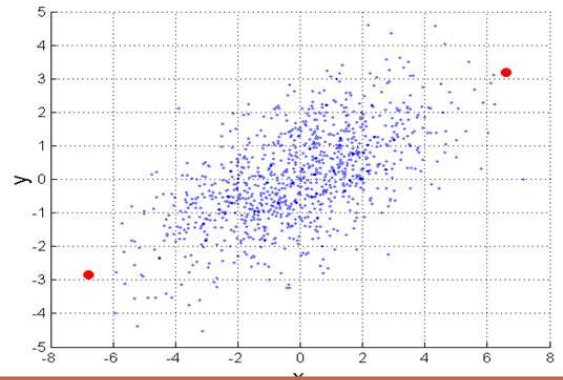


Figure 9: Matrice de scatter plots. a1=lungimea sepalei, a2=lățimea sepalei, a3=lungimea petalei, a4=lățimea petalei

Vizualizare: Scatter plots - utilitate

- Arată relația dintre două atribute; de exemplu, poate permite determinarea vizuală a gradului în care există o legătură liniară între valori (figura de mai jos)
- Dacă seturile de date sunt grupate pe clase, atunci se poate utiliza un scatter plot pentru a vedea în ce măsură două atribute separă clase — vezi în matricea de scatterplot, combinația $a3 - a4$ sau $a3 - a2$. Separabilitatea poate să fie liniară (o dreaptă produce două semiplane care conțin fiecare exclusiv câte o clasă) sau folosind o curbă mai complexă. Dacă nu se poate construi o astfel de curbă, atunci probabil că este nevoie de mai multe atribute care să permită discriminarea claselor, sau o altă metodă (*e.g.* kernel methods).



Vizualizare: Scatter plots - extindere multidimensională

- Scatter plot-urile pot fi extinse pentru a include încă niște atribute
- Pentru o reprezentare 3D se pot folosi atribute categoriale (*e.g.* clasa)

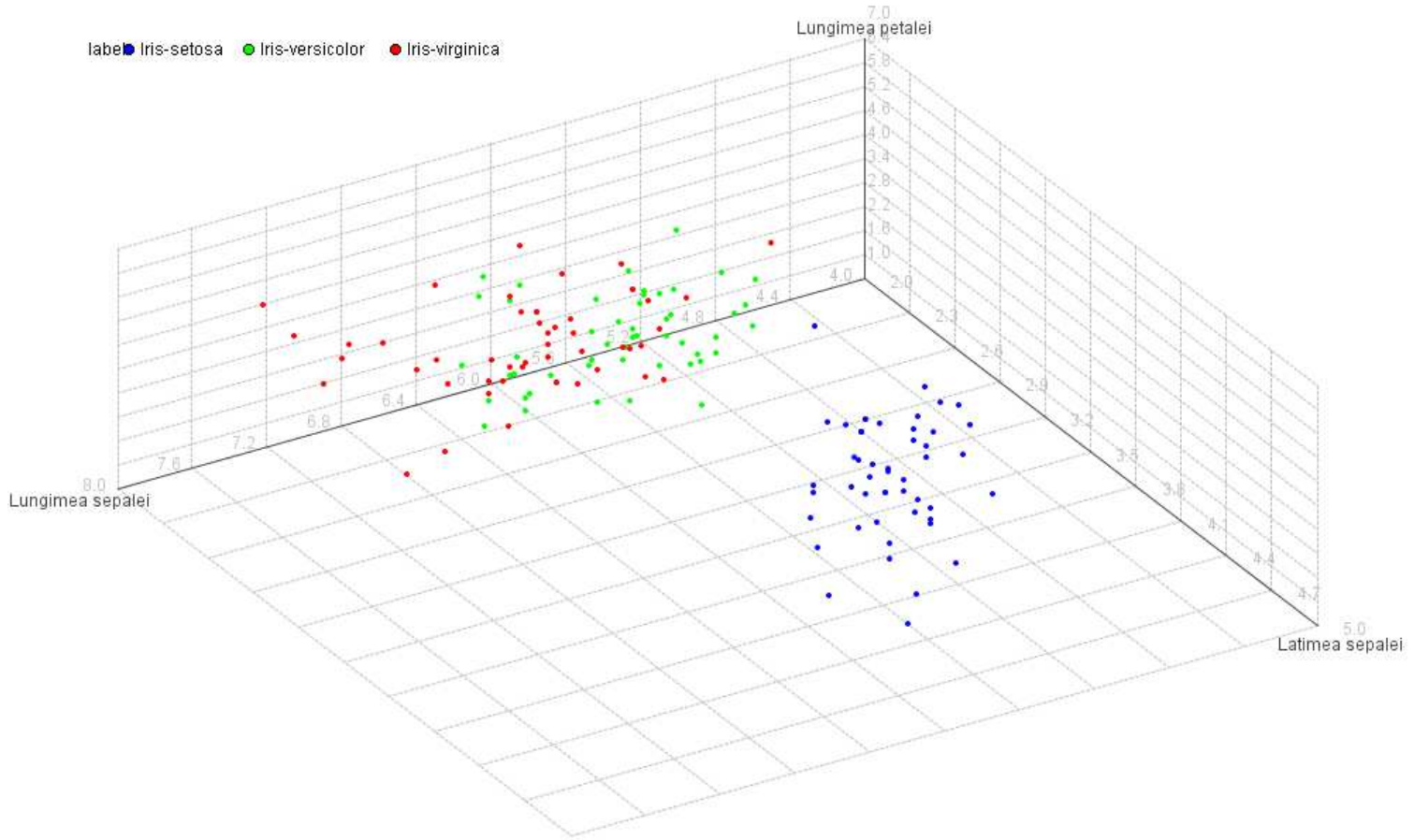


Figure 10: 4 dimensiuni reprezentate pe un scatter plot

Vizualizare: contour plots

- Utilizate atunci când un atribut continuu este măsurat peste un domeniu
- Se obține o partiționare a spațiului în zone pentru care valorile sunt aproximative egale
- Liniile de contur care separă regiuni diferite conectează valori egale
- Exemplu comun: hărți pe care se reprezintă altitudinea
- Pot de asemenea să reprezinte: temperatura, cantitatea de precipitații, presiunea aerului etc.

Vizualizare: contour plots

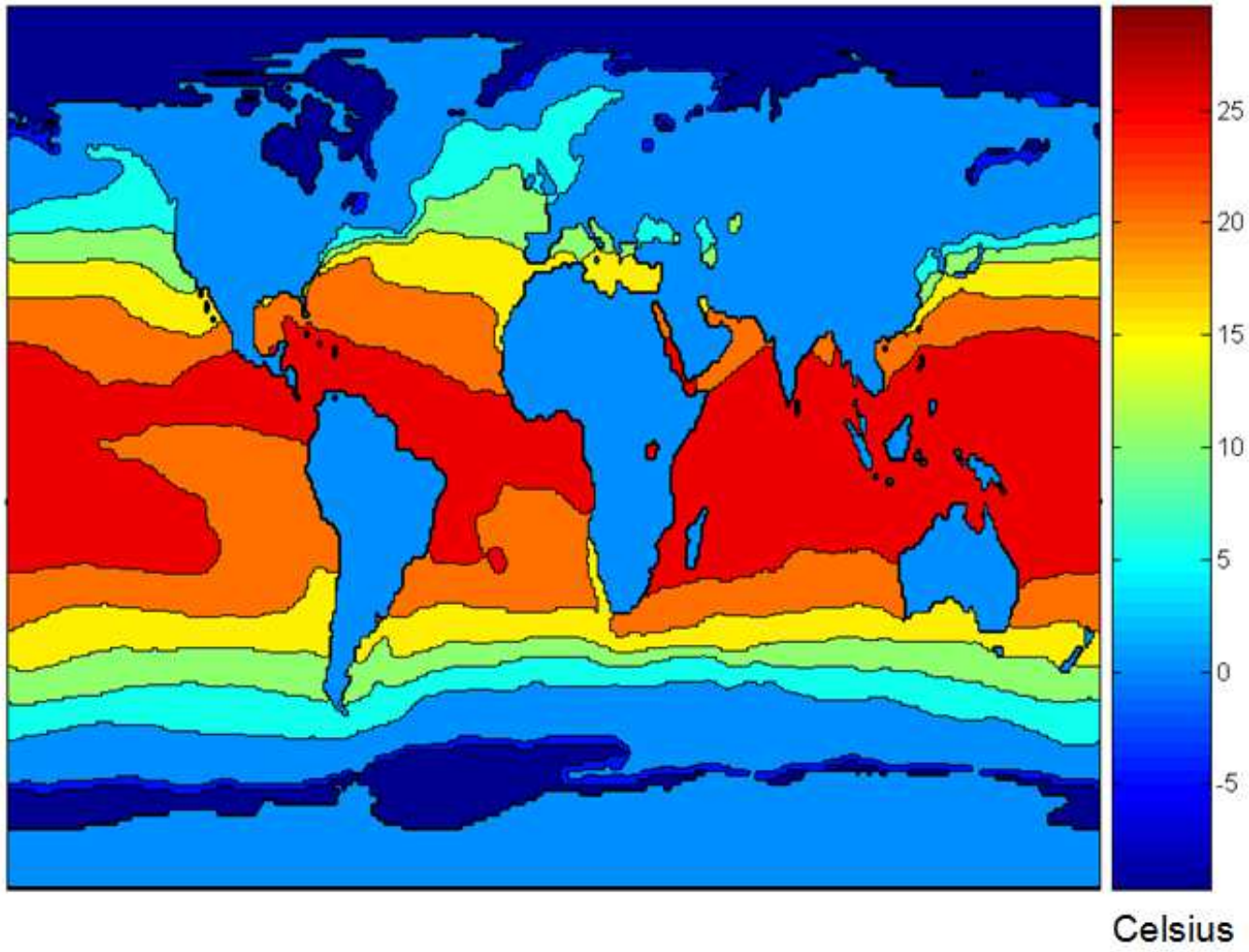


Figure 11: Temperatura medie, decembrie 1998

Alte modalități de vizualizare

- Surface plots
- Vector fields plot
- Lower dimensional slices
- Animații

Sursa: Introduction to Data Mining, cap 3

Vizualizarea datelor multidimensionale: matrice de imagini

- Utile când obiectele sunt grupate pe clase; se permite detectarea faptului că obiecte din aceeași clasă au valori similare
- O matrice de date este un tablou dreptunghiular de valori
- Valorile pot fi reprezentate prin puncte pe ecran, influențând culoarea și strălucirea punctelor
- Dacă atributele au domenii de valori diferite, atunci ele pot fi standardizate pentru a avea media 0 și dispersia 1; astfel se evită ca un atribut să domine reprezentarea grafică

Vizualizarea datelor multidimensionale: matrice de imagini

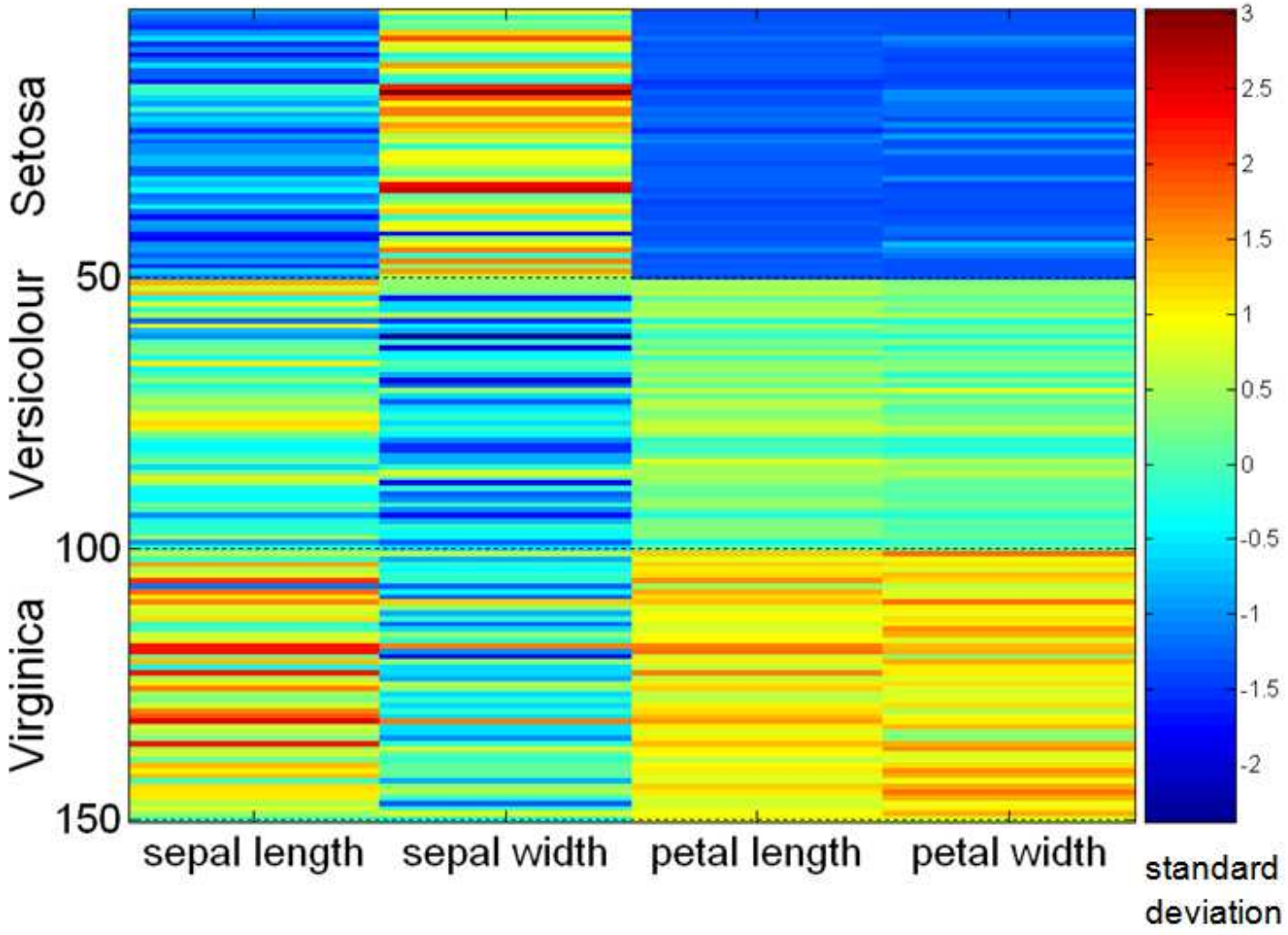


Figure 12: Vizualizarea matricei de date pentru setul Iris

Vizualizarea datelor multidimensionale: matrice de imagini

Florile din aceeași categorie sunt cele mai similare între ele, dar Versicolour și Virginica sunt mai similare între ele decât cu Setosa.

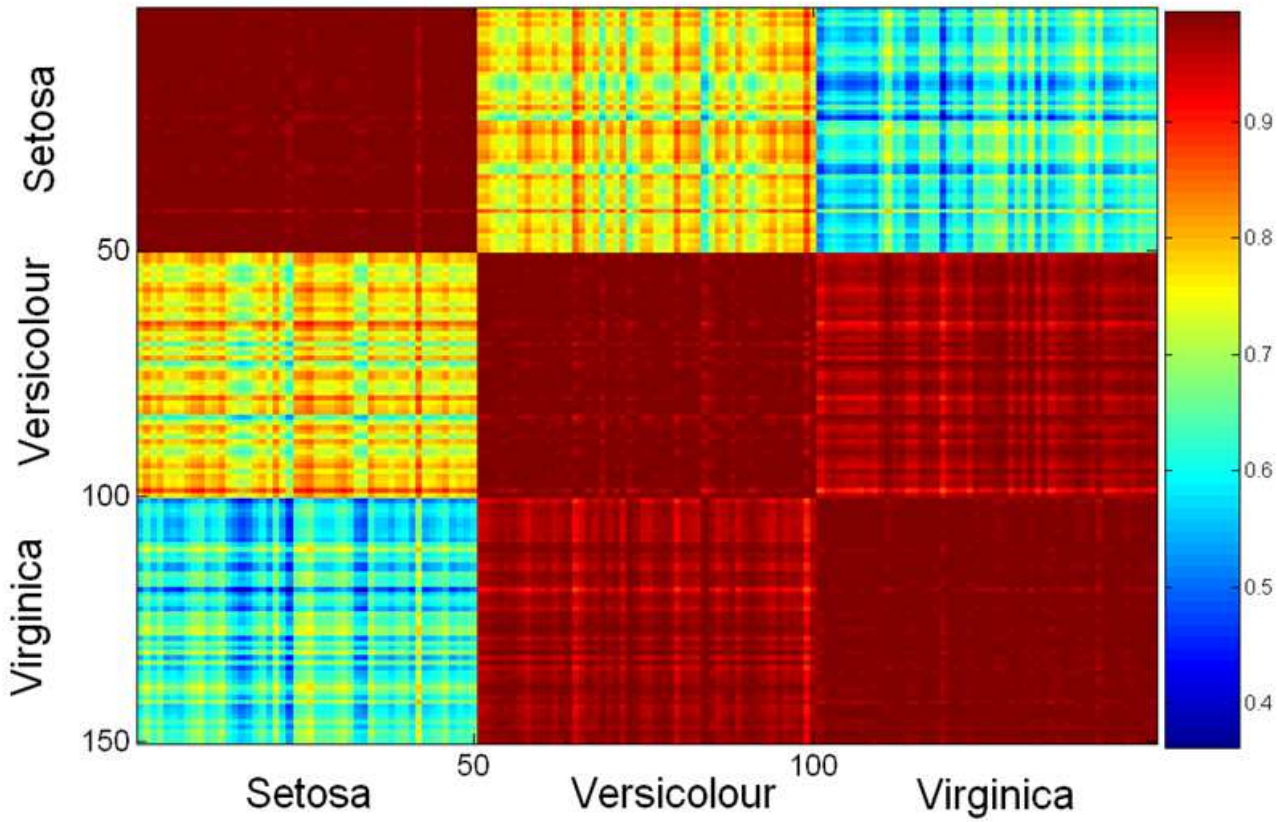


Figure 13: Vizualizarea matricei de corelație setul Iris

Vizualizarea datelor multidimensionale: coordonate paralele

- Au o axă verticală pentru fiecare din atribute; axele sunt paralele între ele
- Fiecare valoare a fiecărui atribut este asociată cu o poziție pe axă
- Dacă obiectele au tendința de a fi apropiate între ele în cadrul aceluiași grup, dar relativ bine separate pentru grupuri diferite, acest lucru se va vedea din reprezentare
- Funcționează bine cu un număr mediu de obiecte, până la 200

Vizualizarea datelor multidimensionale: coordonate paralele

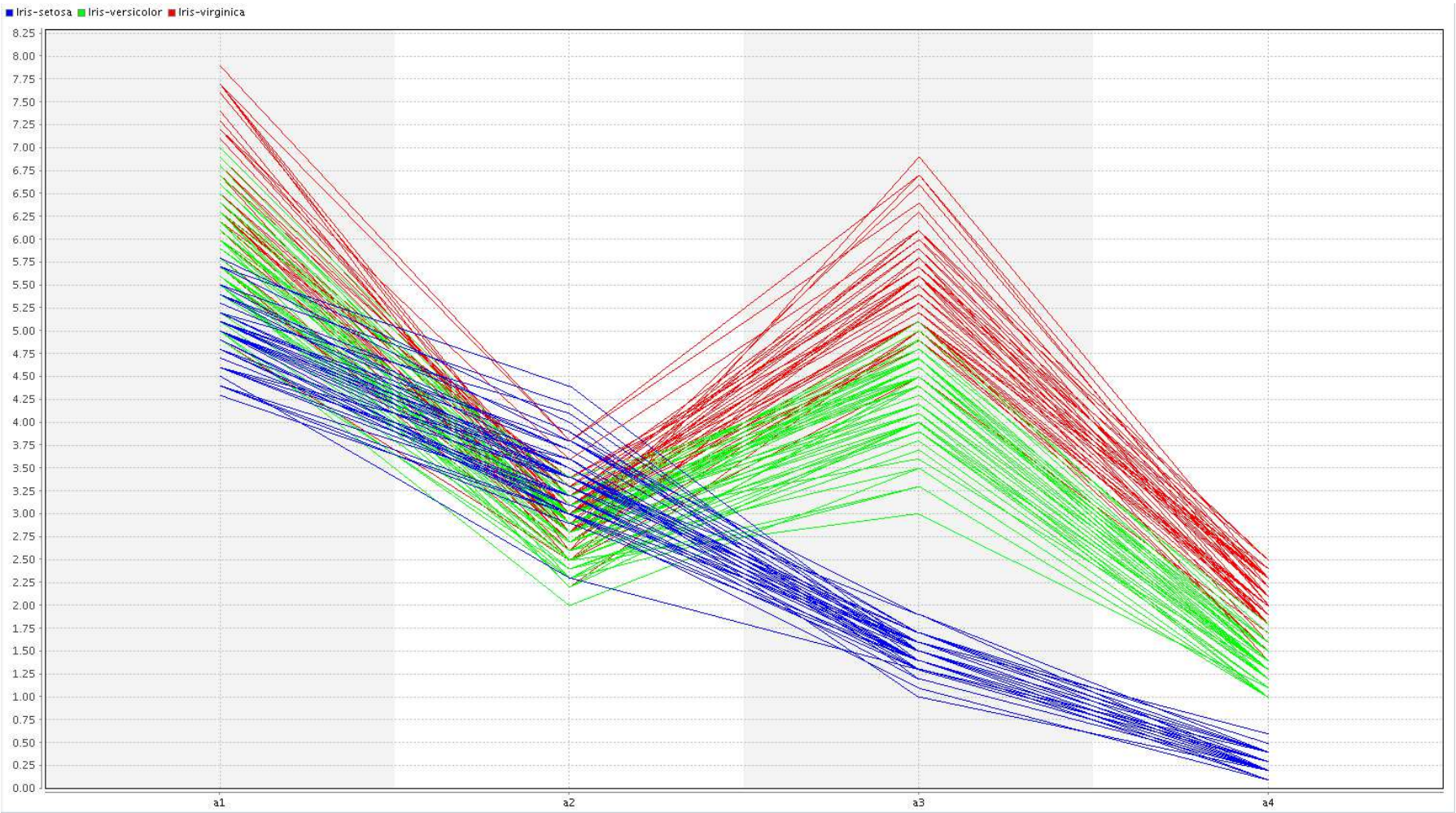


Figure 14: Reprezentare prin coordonate paralele pentru Iris

Vizualizarea datelor multidimensionale: coordonate paralele

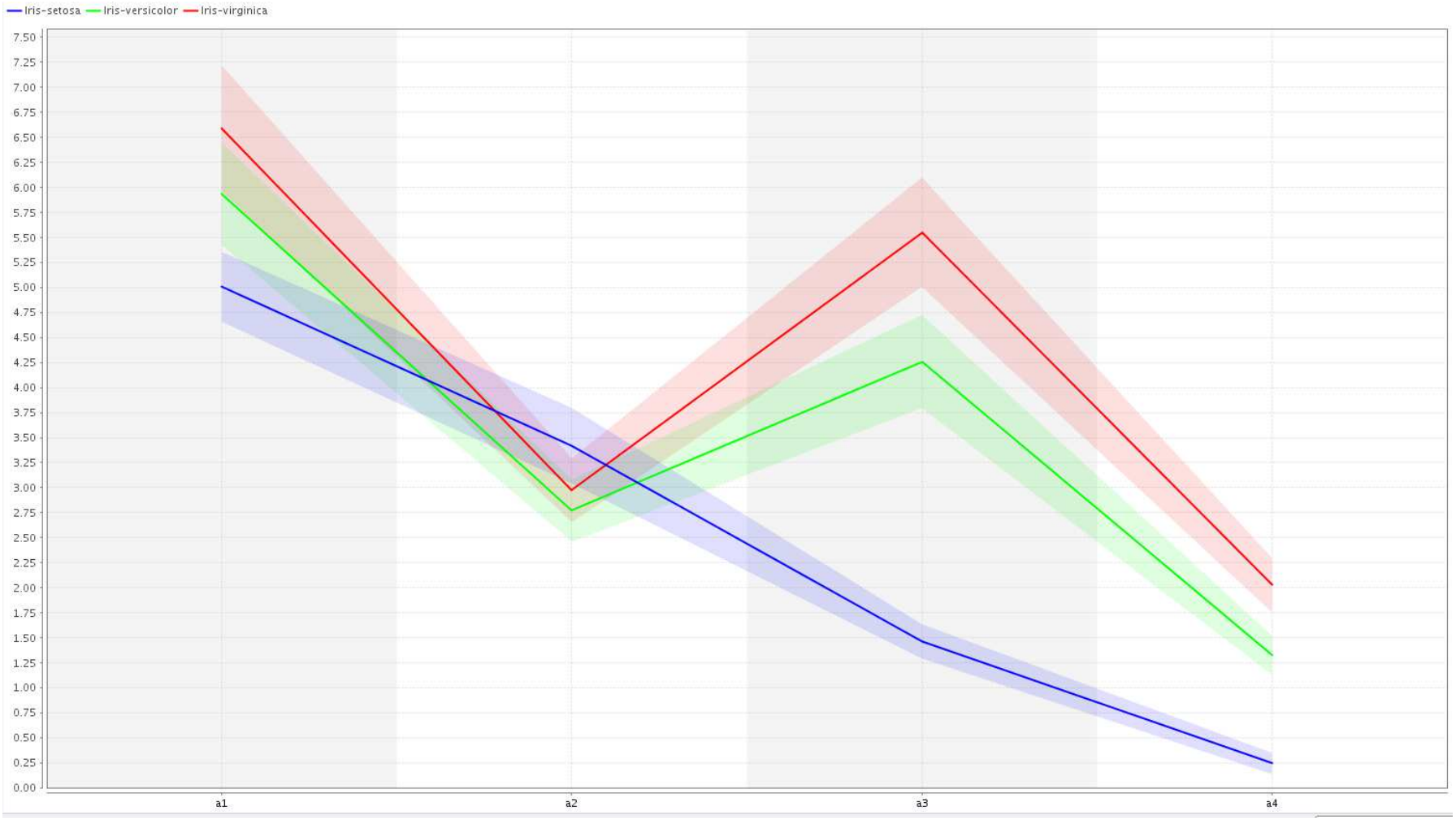
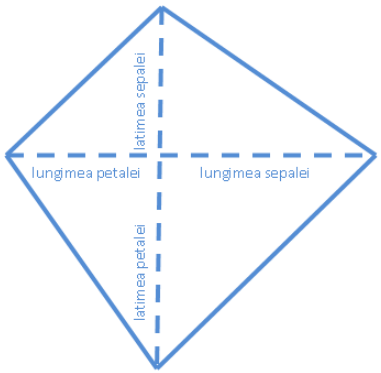


Figure 15: Variantă bazată pe coordonate paralele

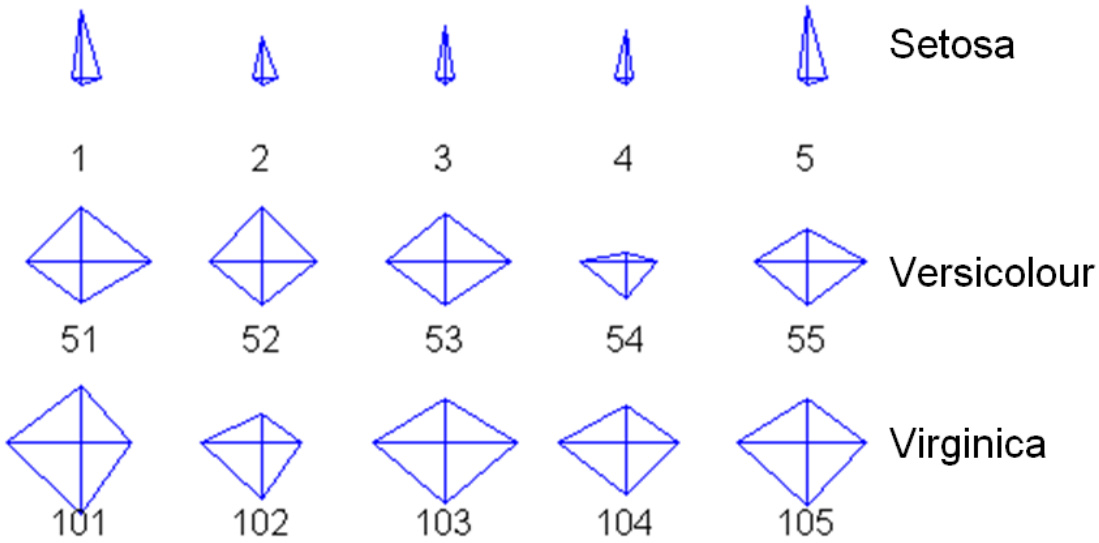
Vizualizarea datelor multidimensionale: alte variante

- Star plots
  - Similar cu coordonate paralele, dar axele radiază dintr–un punct central
  - Liniile care conectează valorile unui obiect creează un poligon
- Fețe Chernoff
  - Fiecare atribut este asociat cu o trăsătură facială
  - Valorile atributelor determină apariția trăsăturilor
  - Fiecare obiect devine o față separată
  - Metoda se bazează pe abilitatea de a distinge fețe

Vizualizarea datelor multidimensionale: Star plots



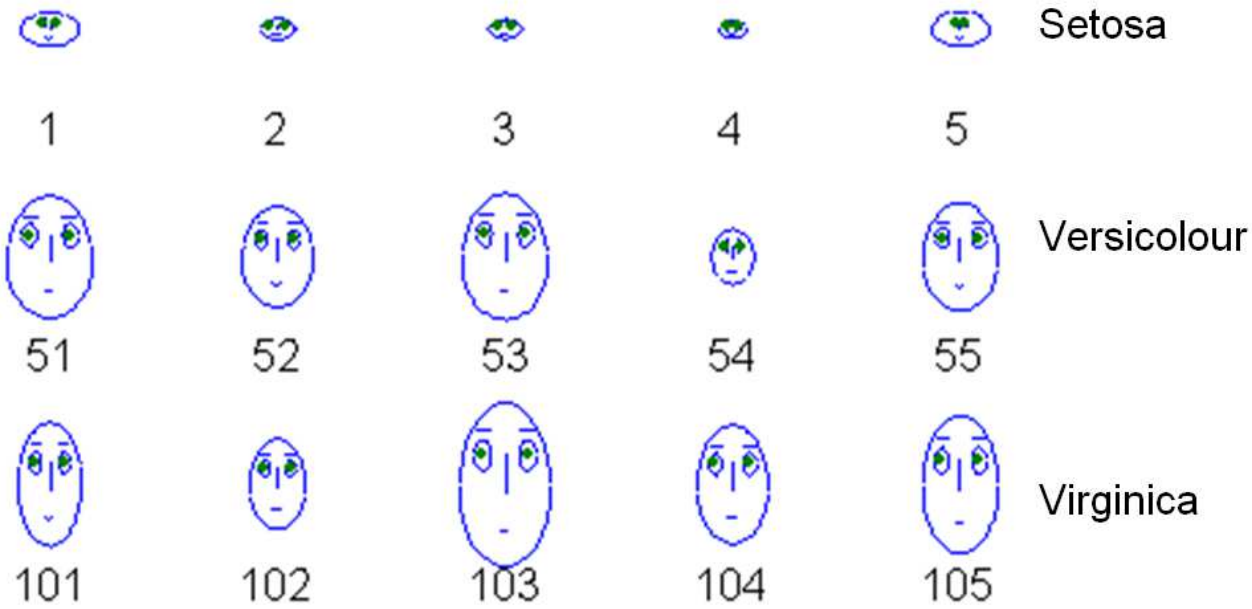
(a) Star plot: schema



(b) Star plot pentru 15 obiecte Iris

Vizualizarea datelor multidimensionale: fețe Chernoff

(a) O față Chernoff



(b) Fețe Chernoff pentru 15 obiecte iris

Outline

OLAP și analiza datelor multidimensionale

- On-Line Analytical Processing (OLAP) a fost propusă de E. F. Codd, părintele bazelor de date relaționale
- Bazele de date relaționale folosesc tabele pentru gruparea datelor, OLAP folosește tablouri multidimensionale
- Se prevede posibilitatea de a interacționa cu tabloul, de exemplu prin selectarea numărului de dimensiuni sau expandări/agregări pe anumite dimensiuni
- Există operații de analiză și explorare a datelor care lucrează ușor cu reprezentare OLAP

OLAP și analiza datelor multidimensionale

Pașii pentru convertirea datelor tabulare într–un tablou multidimensional:

1. Se identifică atributele care vor deveni dimensiuni și care vor deveni valori în cadrul tabloului – valori țintă
   - atributele folosite ca dimensiuni trebuie să aibă valori discrete
   - valoarea țintă este o valoare de contorizare sau o valoare reală exprimând cantitate, sumă, cost etc.
   - se poate să nu fie nicio variabilă țintă continuă și în acest caz se face numărarea obiectelor pe dimensiuni
2. Se calculează valorile din fiecare celulă a tabloului multidimensional prin însumări de valori sau prin numărări de obiecte

OLAP și analiza datelor multidimensionale

- Exemplu: pentru Iris se aleg lungimea, lățimea petalelor și tipul de floare ca atribute;
- Dimensiunile lungimea și lățimea petalelor se discretizează:
  - lungimea petalelor: low [0, 2.5), medium [2.5, 5), high [5, ∞)
  - lățimea petalelor: low [0, 0.75), medium [0.75, 1.75), high [1.75, ∞)
- Se obține tabelul:

Lungimea petalelor	Lățimea petalelor	Specia	Numărul
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

OLAP și analiza datelor multidimensionale

- Pentru orice combinație de valori ale atributelor este corespunzătoare o singură celulă în cadrul tabloului
- Acestei celule îi este asignata numărul de flori care respectă valorile corespunzătoare ale atributelor

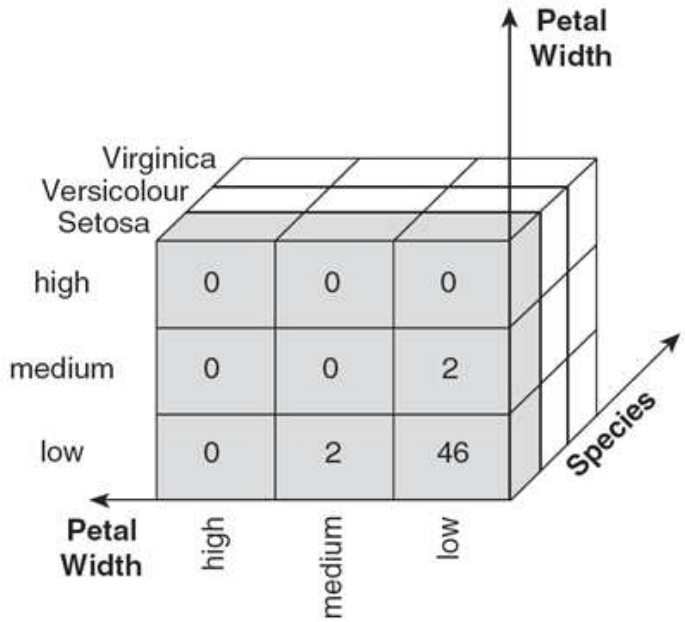


Figure 16: Reprezentare multidimensională pentru setul de date Iris

"Feliile" de tablou sunt arătate mai jos:

		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

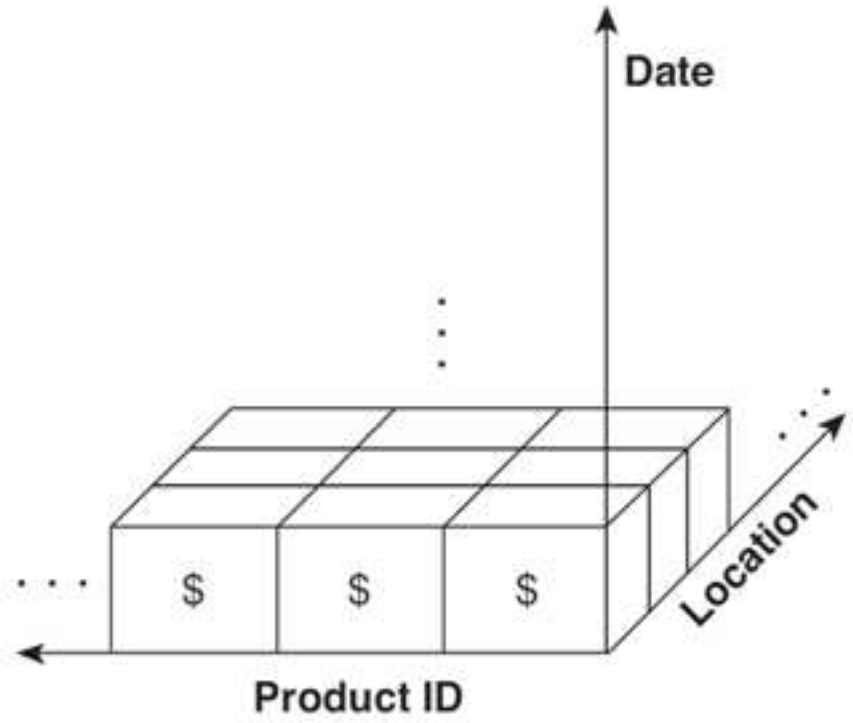
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

OLAP și analiza datelor multidimensionale

- Operația cheie în OLAP este crearea cuburilor de date
- Un cub de date este o reprezentare multidimensională, *împreună cu toate agregările posibile*
- Prin toate agregările posibile înțelegem agregările care se obțin prin alegerea unui subset propriu de dimensiuni și însumând valorile peste toate celelate dimensiuni
- Exemplu (banal): dacă se consideră dimensiunea "specie" și se fac contorizări peste celelate 4 dimensiuni (lungimi/lățimi . . .), atunci se obține un vector unidimensional care are ca valori numărul de plante din fiecare specie (50)

OLAP și analiza datelor multidimensionale

- Exemplu: fie un set de date în care se înregistrează vânzările de produse pentru niște companii, la date diferite
- Datele obținute pot fi reprezentate ca un tablou tridimensional
- Există 3 agregări bidimensionale (combinări de 3 luate câte 2), 3 agregări unidimensionale și o agregare fără dimensiune = totalul general



	date				
product ID	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
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

Figure 17: Tabelul reprezintă o agregare bidimensională, iar pe cele două margini sunt agregări unidimensionale. În colțul din dreapta jos se află agregarea fără dimensiune.

Operații OLAP: slicing, dicing

- Slicing: selectarea unui grup de celule prin specificarea unor valori concrete pentru anumite dimensiuni
- Dicing: selectarea unui subset de celule prin specificarea unui set de valori pentru atribute
- În practică, ambele operații pot fi acompaniate de agregare pe niște dimensiuni

Operații OLAP: roll-up, drill-down

- Datele au deseori o structură ierahică
  - o dată este asociată unei săptămâni, luni, an
  - o locație este asociată unui oraș, regiune, țară, continent
  - produsele pot fi divizate în câteva categorii: hrană, îmbrăcăminte etc.
- Categoriile deseori se conțin unele pe altele
- Roll-up: se poate face agregare a vânzărilor de la datele zilnice la luni sau ani
- Drill-down: invers față de roll-up; dacă se dau vânzările pe ani, se poate detalia la nivel de lună sau săptămână

Outline

Resurse

- Cărțile lui Edward Tufte: *The Visual Display of Quantitative Information* etc.
- Seven Basic Tools of Quality