Analiza și procesarea datelor prin tehnici de Învățare Automată

1. Introducere în Data Mining

Șef Lucrări Dr. Ing. Horia Modran

Contact: horia.modran@unitbv.ro / modranhoria@gmail.com

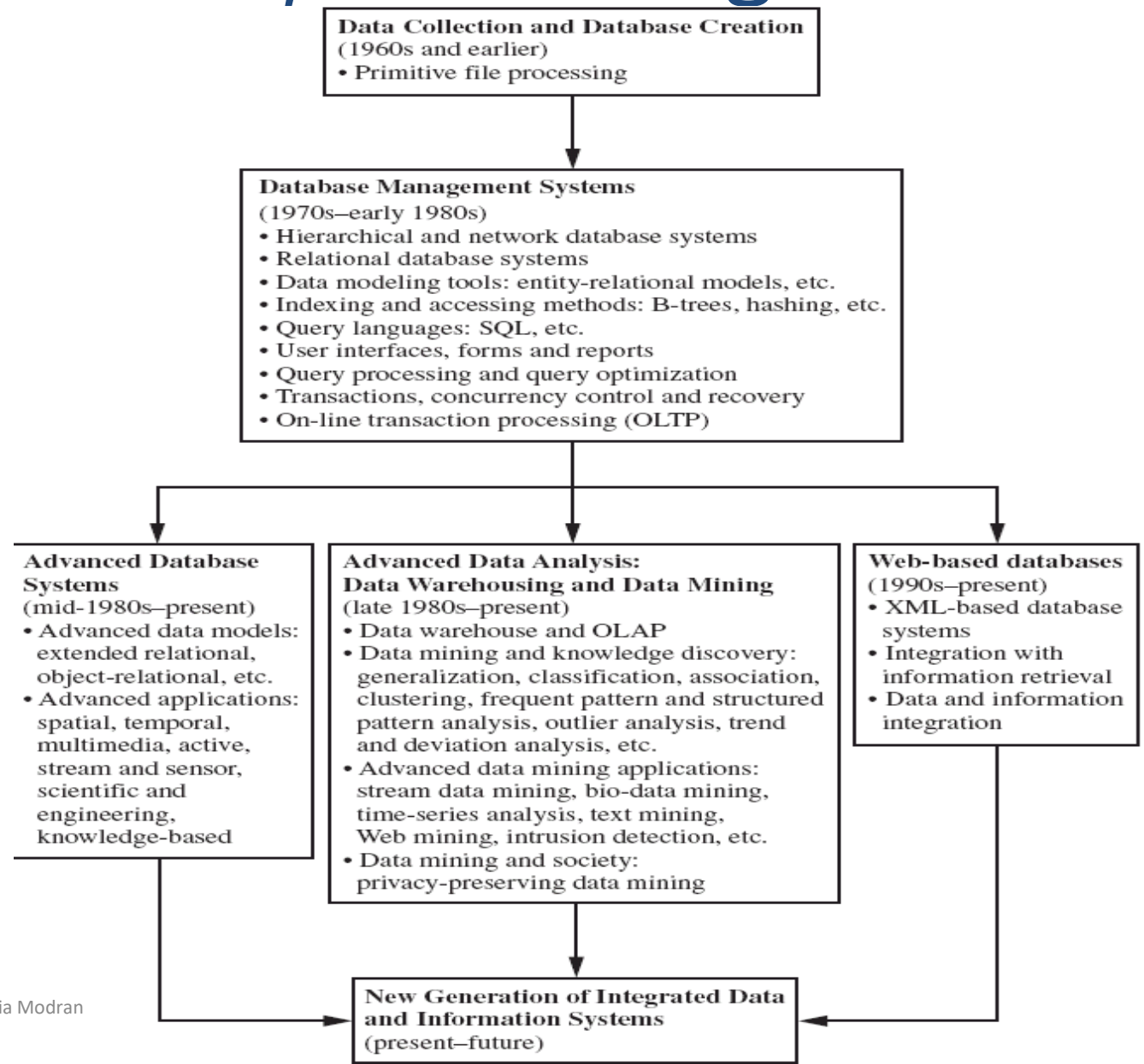
Tel: 0770171577

Universitatea Transilvania din Brașov

FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

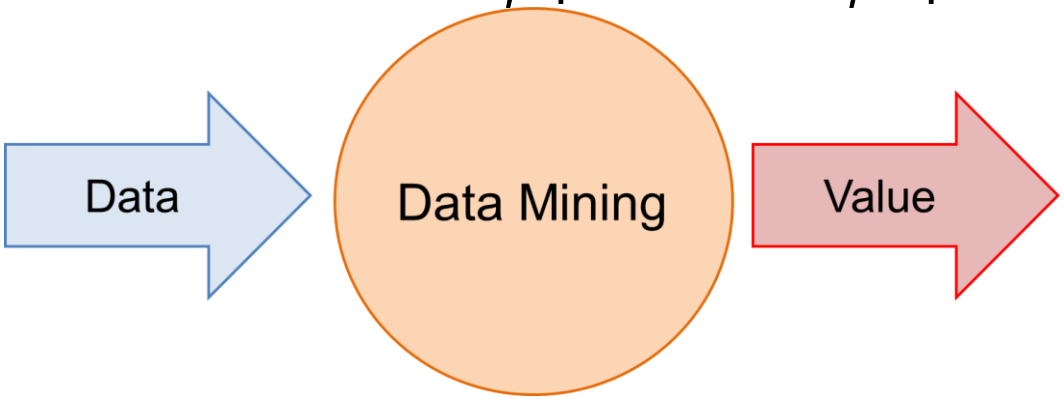
2024 - 2025

Evoluția tehnologiilor BD

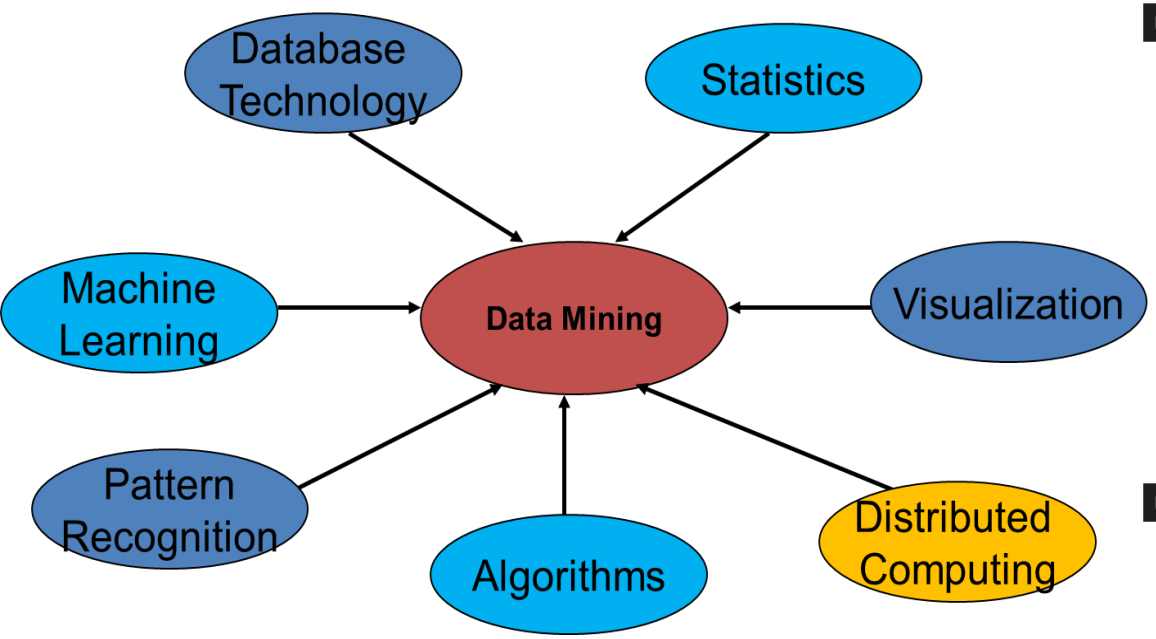


Ce este Data Mining?

- după ani de exploatare a datelor, încă nu există un răspuns unic la această întrebare
- termenul „Data Mining” nu a fost inventat decât în anii 1990
- definiție posibilă: „Data mining-ul este utilizarea unor tehnici eficiente pentru analiza unor colecții foarte mari de date și extragerea de structuri utile și posibil neașteptate în date.”



Data Mining



- **Data Science:** Datele sunt utile pentru a înțelege un proces și pentru a-l îmbunătăți
  - se concentrează pe aplicații imediate
- **Big Data:** Datele ar trebui procesate colectiv și interconectate. Este nevoie de infrastructură cloud
  - mai mult orientate spre sisteme.

- **AI/Învățare automată/Învățare profundă:** avem date pentru a realiza modele mai complexe, care sunt semnificativ mai puternice
  - accent pe descoperirile științifice

Data Mining – Motivație

- creșterea explozivă a datelor: de la terabytes (1000^4) la yottabytes (1000^8) -> cantitate foarte mare de date brute !!
  - colectarea datelor și disponibilitatea datelor
    - instrumente automate de colectare, sisteme DB, web
  - surse majore de date
    - business: Web, comerț electronic, tranzacții, stocuri, etc.
    - știință: bioinformatică, cercetare medicală
    - dispozitive mobile/IoT, camere digitale etc.
- Cum se analizează datele?
- Data mining - analiză automată a seturilor de date masive

Data Mining – Motivație

- cantități mari de date pot fi mai puternice decât algoritmii și modelele complexe
  - Google a rezolvat problemele de procesare a limbajului natural doar analizând datele: greșeli de ortografie, greșeli gramaticale, sinonime
- datele reprezintă cele mai mari active ale companiilor
- avem nevoie de o modalitate de a valorifica inteligența colectivă
- datele sunt foarte complexe: tabele, serii temporale, imagini, grafice, etc.

Ce sunt datele?

- colecție de obiecte și atributele acestora
- un atribut este o proprietate sau o caracteristică a unui obiect
  - exemple: culoarea ochilor, temperatura etc.
- o colecție de atribute descriu un obiect
- obiectul este cunoscut și ca înregistrare, punct, entitate, instanță

Atribute = coloanele tabelului

Obiecte = rândurile din tabel

Tid	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	95K	NULL
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	NULL	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Size: Number of objects
Dimensionality: Number of attributes
Sparsity: Number of populated object-attribute pairs

Tipuri de atribute

- Există diferite tipuri de atribute
  - **Categoriale**
    - exemple: culoarea ochilor, coduri poștale, cuvinte, clasificare (bun/rău), înălțime în {înalt, mediu, scund}
    - nominal (fără ordine sau comparație) vs ordinal (descrie o anumită oridine)
  - **Numeric**
    - exemple: data, temperatura, ora, lungimea, valoarea, etc.
    - discrete (ora) sau continue (temperatura)
    - caz special: atribute binare (da/nu, există/nu există)

Date numerice

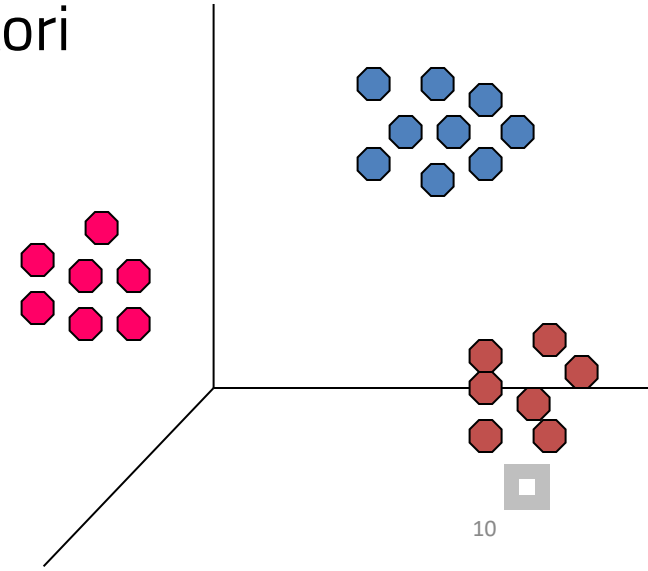
- dacă obiectele de date au același set fix de atribute numerice, atunci obiectele de date pot fi văzute ca puncte într-un spațiu multidimensional, unde fiecare dimensiune reprezintă un atribut distinct
- un astfel de set de date poate fi reprezentat printr-o matrice de date de dimensiune n x d, cu n rânduri (câte unul pentru fiecare obiect), și d coloane (câte una pentru fiecare atribut).

	Temperature	Humidity	Pressure
O1	30	0.8	90
O2	32	0.5	80
O3	24	0.3	95

30	0.8	90
32	0.5	80
24	0.3	95

Date numerice

- datelor numerice pot fi tratate ca puncte sau vectori
- pentru dimensiuni mici putem reprezenta grafic datele
- putem folosi analogi geometrici pentru a defini concepte precum distanța sau asemănarea
- putem folosi algebra liniară pentru a procesa matricea de date
- vom vorbi adesea despre puncte sau vectori



Date relaționale mixte

- Date care constau dintr-o colecție de înregistrări, fiecare dintre ele constând dintr-un set fix de atribute atât numerice, cât și categoriale

ia valori numerice, dar este de fapt categoric

ID	Cod poștal	Vârstă	Stare civilă	Salariu	Treaptă venit	Refund
1129842	45221	55	Single	250000	High	0
2342345	45223	25	Married	30000	Low	1
1234542	45221	45	Divorced	200000	High	0
1243535	45224	43	Single	150000	Medium	0

Atributele booleene pot fi considerate atât numerice, cât și categoriale
Când apar împreună cu alte atribute, ele au pot fi privite drept categoriale
Acestea sunt adesea reprezentate numeric.

Date relaționale mixte

- uneori este convenabil să se reprezinte atributele categoriale ca date de tip boolean
  - se adaugă un atribut boolean pentru fiecare valoare posibilă

ID	Zip 45221	Zip 45223	Zip 45224	Vârstă	Sin gle	Married	Divorced	Venit	Refund
1129842	1	0	0	55	1	0	0	250000	0
2342345	0	1	0	25	0	1	0	30000	1
1234542	1	0	0	45	0	0	1	200000	0
1243535	0	0	1	43	1	0	0	150000	0

Acum putem vedea întregul vector ca numeric

Date relaționale mixte

- în alte situații este convenabil să se reprezinte atributele numerice drept categoriale
  - se grupează valorile atributelor numerice în bin-uri

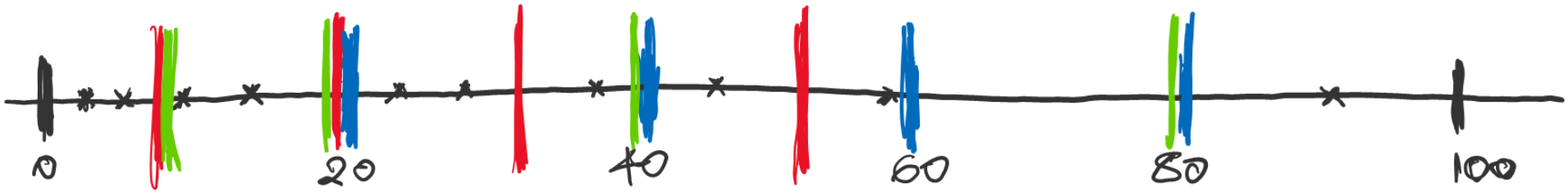
ID Number	Zip Code	Age	Marital Status	Income	Income Bracket	Refund
1129842	45221	50s	Single	High	High	0
2342345	45223	20s	Married	Low	Low	1
1234542	45221	40s	Divorced	High	High	0
1243535	45224	40s	Single	Medium	Medium	0

- Idee: se împarte intervalul domeniului atributului numeric în *bins* (intervale).

Bucketization

- **Equi-width bins**: toate intervalele au aceeași dimensiune
  - example: împărțirea timpului în decade
  - problemă: unele intervale pot conține puține puncte
- **Equi-size (depth)** bins: selectarea intervalelor astfel încât ele să conțină un număr egal de puncte
  - de exemplu, acest procedeu împarte datele astfel: primele 10%, următoarele 10%. etc.
  - problemă: unele intervale pot fi foarte mici
- **Equi-log** bins: log end–log $start$ este constant
  - dimensiunea zonei anterioare este o fracțiune din cea curentă

Bucketization - exemplu



Albastru: Equi-width [20,40,60,80]

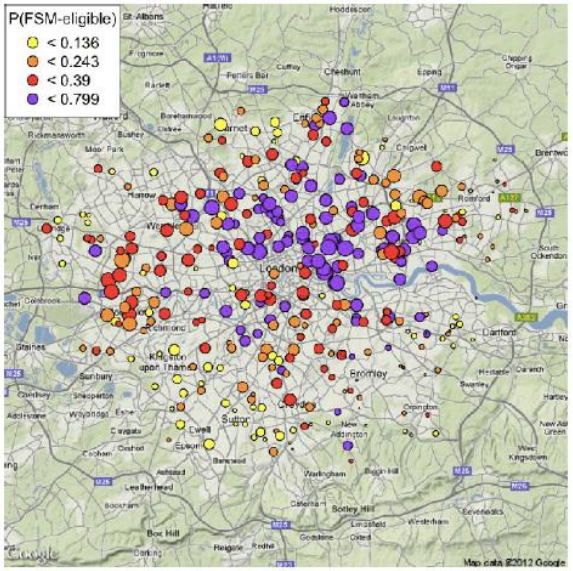
Roșu: Equi-depth (2 points per bin)

Verde: Equi-log ($\frac{end}{start} = 2$)

Tip de date - exemplu

TID	Articole
1	Pâine, Coca-Cola, Lapte
2	Bere, Pâine
3	Bere, Coca-Cola, Lapte
4	Bere, Pâine, Lapte
5	Coca-Cola, Lapte, Miere

Date tranzacționale



Date spațiale

Date ordonate

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Date documente

Tipuri de date

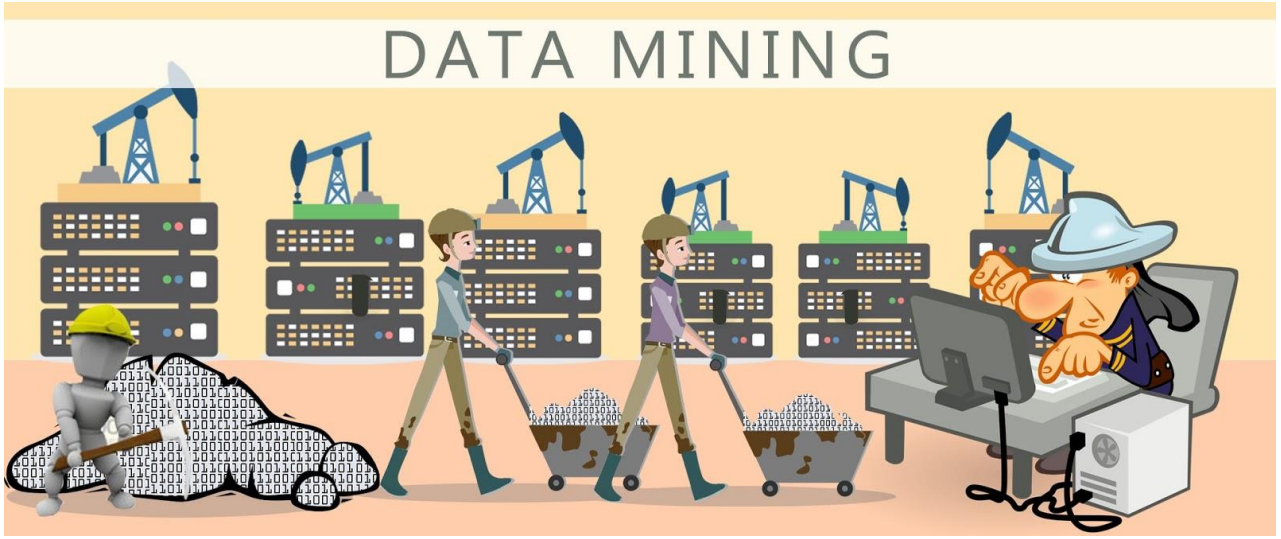
- Date numerice: fiecare obiect este un punct dintr-un spațiu multidimensional
- Date categoriale: fiecare obiect este un vector de valori categoriale
- Set de date: fiecare obiect este un set de valori (cu sau fără ordin de comparație)
- Secvențe ordonate: fiecare obiect este o secvență ordonată de valori
- Date grafice

Ce este data mining?

- **Punct de vedere comercial** (companii – Facebook, Google, etc.)
  - datele au devenit avantajul competitiv cheie al companiilor
  - capacitatea de a extrage informații utile din date este esențială pentru exploatarea lor comercială.
- **Punct de vedere științific**
  - poziție fără precedent- se colectează TB de date (date de la senzori, date bancare, date ale rețele sociale, etc.)
  - avem nevoie de instrumente specifice pentru a analiza astfel de date.

Utilizări ale Data Mining

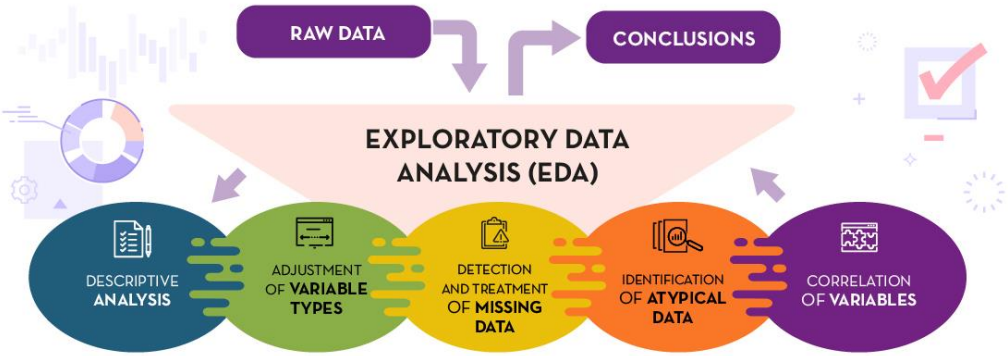
- Câteva utilizări uzuale ale Data Mining:
  - seturi frecvente de itemi (extragere de text, recomandări)
  - asociere și extragerea regulilor
  - analiza exploratorie
  - asemănări
  - grupare
  - clasificare



Analiză exploratorie

- Se realizează o analiză pentru a înțelege cum arată datele
- exemplu: postări social media
  - Cât de des postează utilizatorii, câte postări per utilizator, când postează, există o corelație între numărul de postări și numărul de prieteni, etc.
- Acesta este unul dintre primii pași după colectarea datelor
  - metrici: este important să se decidă **ce să măsoare**

Pașii de analiză



Explorarea similitudinilor

- Se consideră următoarele date despre utilizatori:
  - de câte ori au dat click pe postările din aceste pagini

Ce concluzii putem trage?

	NBA	ESPN	Sports.com	MSNBC	NY Times	Wall Street	Politico
A	100	50	73	10	1	1	4
B	500	200	400	20	10	4	1
C	80	100	60	1	3	1	1
D	4	2	1	12	90	100	80
E	9	3	4	9	100	80	70
F	3	4	5	30	300	200	500

Cum determinăm similaritatea?
Cum grupăm utilizatori similari? Clustering

Realizarea de predicții

- completarea unei valorii lipsă ~ sarcină de predicție
- Tipuri de sarcini de predicție:
  - prezicearea unei valori reale: Regresie
  - prezicerea unei valori binare (DA/NU): Clasificare binară
  - predicția pe mai multe clase : Clasificare
- Vă puteți gândi la sarcini de predicție/clasificare pentru o rețeaua de socializare?

Ad click prediction

Like prediction

Predict if a post is offensive

Predict if a photo contains nudity

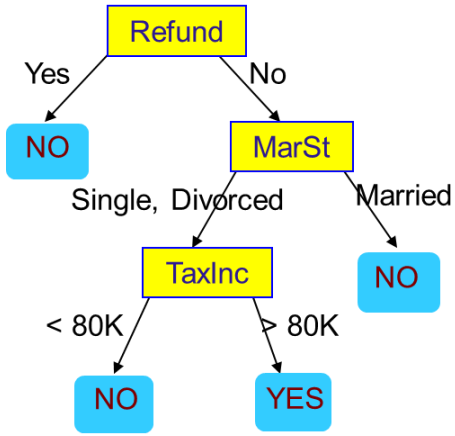
Ad clickthrough prediction

Predict if a user will like a post over another: Learning to rank

Clasificare

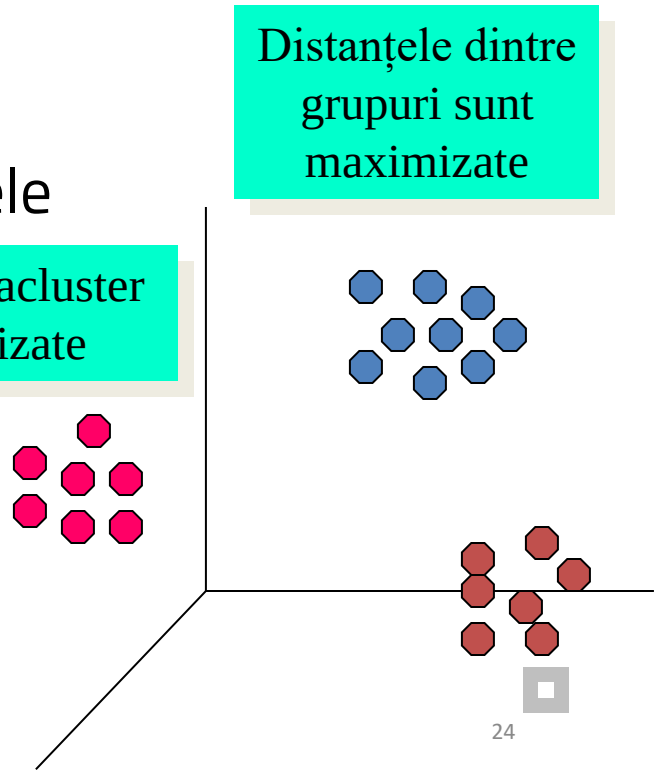
- Procesul de clasificare:
  - găsiți caracteristici care descriu o entitate
  - folosiți exemple de clase pentru a prezice
  - dezvoltă un model (funcție) care face redicția
- **Clasificarea este «motorul» din spatele Revoluției Inteligenței Artificiale (IA)**
  - utilizat în toate sistemele care iau decizii
  - a devenit foarte puternic cu Deep Learning
  - aplicații uriașe în *computer vision*

Tid	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	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



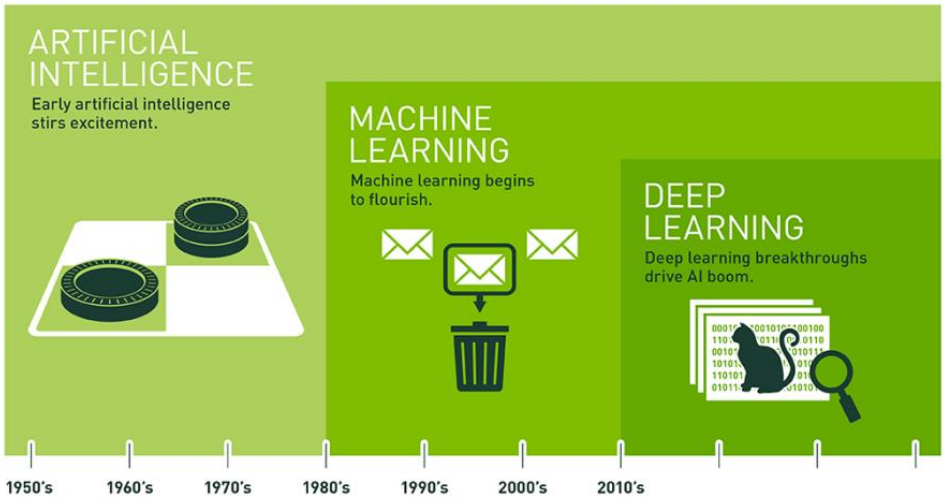
Clustering

- fiind dat un set de puncte, fiecare având un set de atribute și o măsură de similitudine între ele, găsiți grupuri astfel încât:
  - punctele de date dintr-un cluster să fie cât mai asemănătoare între ele
  - punctele de date din clustere diferite sunt cât mai puțin asemănătoare între ele
- Măsuri de similaritate?
  - distanța euclidiană
  - alte măsuri specifice problemei



Deep Learning

- Sisteme de învățare automată care utilizează rețele neuronale cu mai multe straturi și sunt antrenate pe cantități mari de date
  - capabil să învețe reprezentări complexe și modele puternice
  - aplicații în recomandări, analiză de rețea, analiză de text, recunoaștere de imagini, conducere autonomă, etc.

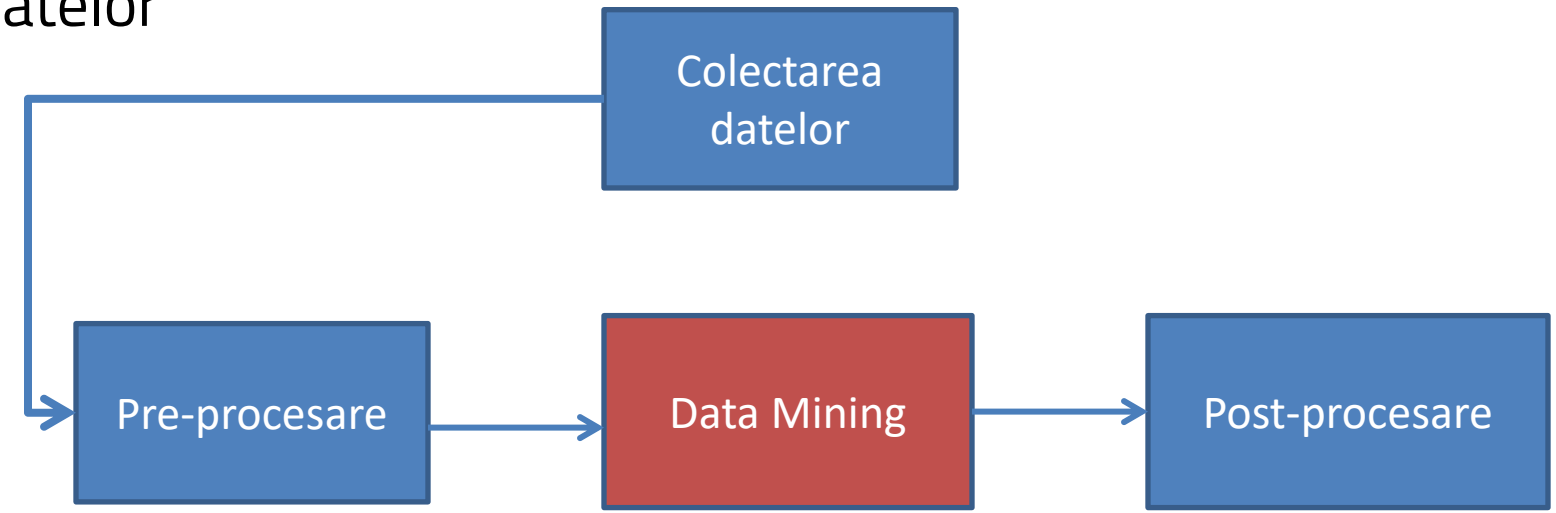


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

- necesită mai puțină procesare a datelor

Data Mining pipeline

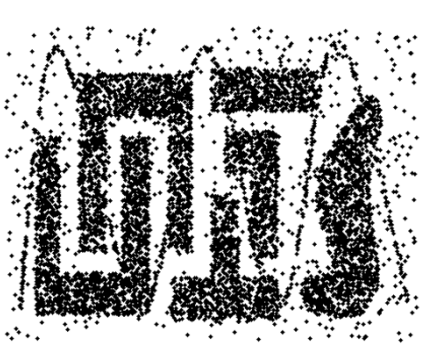
- Data Mining nu este singurul pas în procesul de analiză
- partea de data mining se referă la metodele analitice și algoritmii pentru extragerea informațiilor utile din date
- pre- și post-procesarea sunt adesea sarcini de extragere a datelor



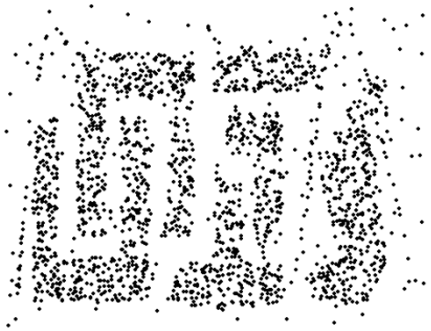
Colectarea datelor

- eșantionarea este tehnica principală folosită pentru selectarea datelor
  - este adesea folosit atât pentru investigarea preliminară a datelor, cât și pentru analiza finală a datelor
- statisticienii eșantionează deoarece obținerea întregului set de date de interes este prea costisitoare sau consumatoare de timp
  - exemplu: care este înălțimea medie a unei persoane în România?
- eșantionarea este utilizată în data mining
  - exemplu: Avem un set de 1M documente. Ce procent de perechi de documente are cel puțin 100 de cuvinte în comun?

Dimensiunea eșantioanelor



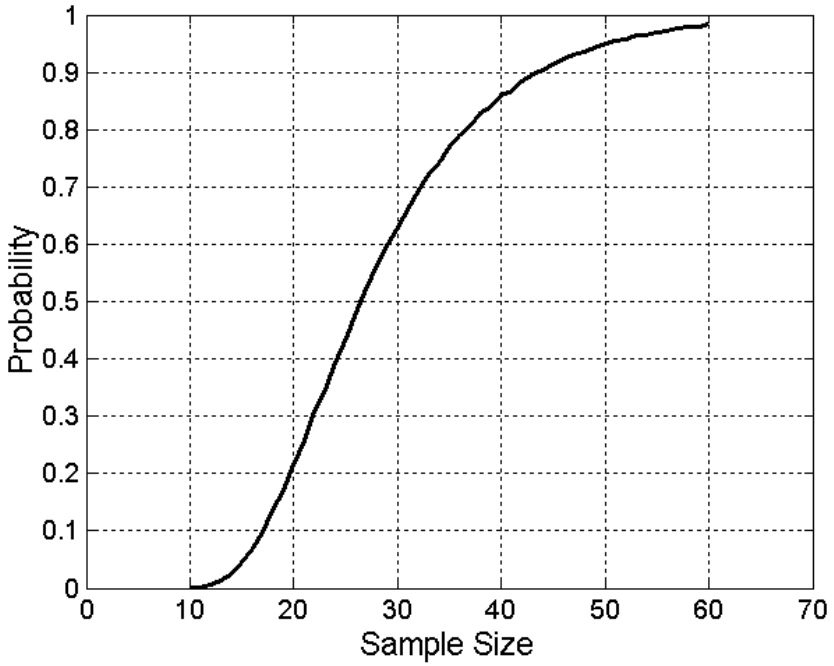
8000 points



2000 Points

500 Points

Ce dimensiune a eșantionului este necesară pentru a obține cel puțin un obiect din fiecare dintre cele 10 grupuri



Curățarea datelor

- datele trebuie **curățate**
- trebuie să extragem **caracteristicile reprezentative a datelor**

Examples of data quality problems:
- Noise and outliers
- Missing values
- Duplicate data

Greșeală sau milionar?

Valoare lipsă

Date inconsistente/duplicate

Tid	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
9	No	Single	90K	No

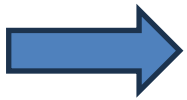
Extragerea caracteristicilor

- datele pe care le obținem nu sunt neapărat un tabel relațional
- datele pot fi într-un formă brută
  - exemple: text, vorbire, mișcări ale mouse-ului etc.
- trebuie să extragem caracteristicile din date
- extragerea caracteristicilor:
  - selectând caracteristicile reprezentative
  - necesită anumite cunoștințe de domeniu despre date
  - depinde de aplicație
- Învățare profundă elimină acest pas

Normalizarea datelor

- în multe cazuri este importantă normalizarea datelor
- tipul de normalizare pe care îl folosim depinde de ceea ce vrem să realizăm
- Normalizarea coloanelor
  - scădeți valoarea minimă și împărțiți la diferența dintre valoarea maximă și valoarea minimă pentru fiecare atribut
  - transformă valorile în intervalul [0,1]

Temperature	Humidity	Pressure
30	0.8	90
32	0.5	80
24	0.3	95



Temperature	Humidity	Pressure
0.75	1	0.66
1	0.40	0
0	0	1

new value = (old value – min column value) / (max col. value – min col. value)

Normalizarea rândurilor

- împărțiți valoarea la suma valorilor pentru fiecare document (rând din matrice)
- transforma un vector într-o **distribuție***

	Word 1	Word 2	Word 3
Doc 1	28	50	22
Doc 2	12	25	13

Sunt aceste documente similare?

	Word 1	Word 2	Word 3
Doc 1	0.28	0.5	0.22
Doc 2	0.24	0.5	0.26

new value = old value / Σ old values in the row

* de exemplu, valoarea celulei (Doc1, Word2) este probabilitatea ca un cuvânt ales aleatoriu din Doc1 să fie Word2

Normalizarea rândurilor

- Acești doi utilizatori evaluează filmele în mod similar?
- scădeți valoarea medie pentru fiecare utilizator (rând)
- capturează abaterea de la comportamentul mediu

	Movie 1	Movie 2	Movie 3
User 1	1	2	3
User 2	2	3	4

	Movie 1	Movie 2	Movie 3
User 1	-1	0	+1
User 2	-1	0	+1

new value = (old value – mean row value)

Post-procesare

- Vizualizare
  - ochiul uman este un instrument analitic puternic!!
- dacă vizualizăm datele în mod corespunzător, putem descoperi tipare și putem demonstra tendințe
- vizualizare - prezentați datele astfel încât modelele să poată fi văzute
  - histogramele și diagramele sunt o formă de vizualizare
  - există mai multe tehnici

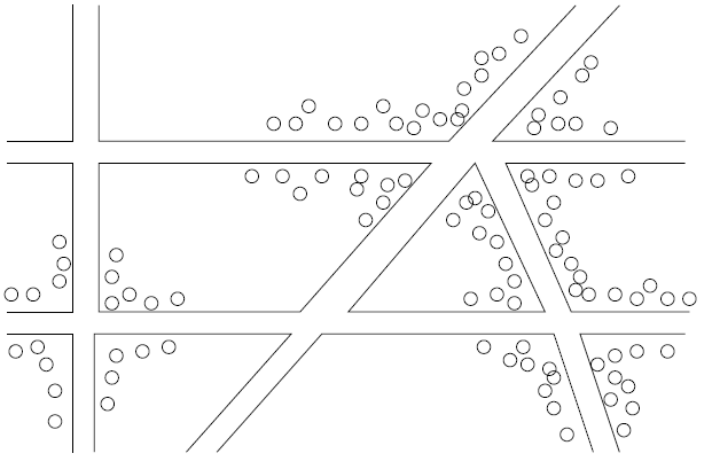


Figure 1.1: Plotting cholera cases on a map of London

Vizualizare pe o hartă

Reducerea dimensionalității

- ochiul uman este limitat la procesarea vizualizărilor în două (cel mult trei) dimensiuni
- una dintre marile provocări în vizualizare este vizualizarea datelor multi-dimensionale într-un spațiu bidimensional
  - reducerea dimensionalității
  - înglobări care păstrează distanța
- Reducerea dimensionalității este, de asemenea, o tehnică de **preprocesare**:
  - reduce cantitatea de date
  - extrageți informațiile utile

Reducerea dimensionalității

- Se considerară următorul set de date cu 6 dimensiuni:

$$D = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 & 0 \\ 2 & 4 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 2 & 4 & 6 \\ 1 & 2 & 3 & 1 & 2 & 3 \\ 2 & 4 & 6 & 2 & 4 & 6 \end{bmatrix}$$

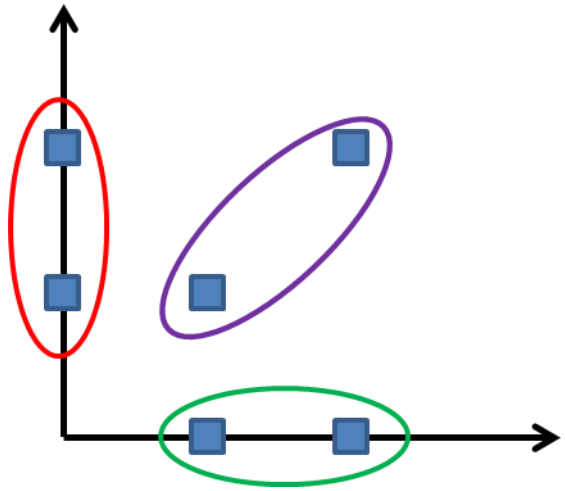
Fiecare rând este un multiplu a doi **vectori**

$$x = [1, 2, 3, 0, 0, 0]$$
$$y = [0, 0, 0, 1, 2, 3]$$

Ce observați? Putem reduce dimensiunea datelor?

Putem rescrie $\boldsymbol{D}$ ca:

$$D = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$



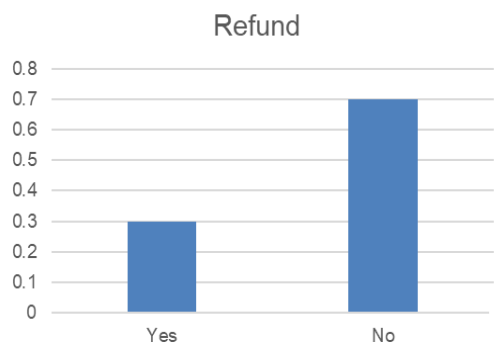
Analiza exploratorie

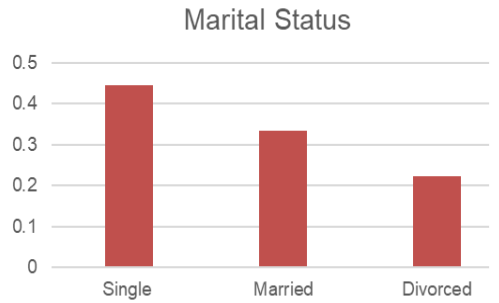
- Statistici sumare: numere care rezumă proprietățile datelor
- Proprietățile rezumate includ frecvența, locația și răspândirea
  - exemple: locație - medie

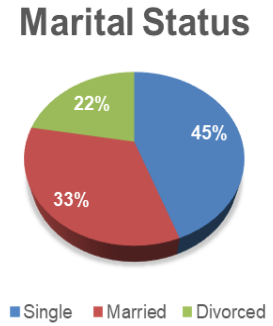
    răspândire (*spread*) - abatere standard
- frecvența unei valori de atribut este procentul de apariție a valorii în setul de date
  - de exemplu, având în vedere atributul „gen” și o populație reprezentativă de oameni, genul „feminin” apare în aproximativ 50% din timp
- **modul** unui atribut este cea mai frecventă valoare a atributului
- putem vizualiza frecvențele datelor folosind o histogramă

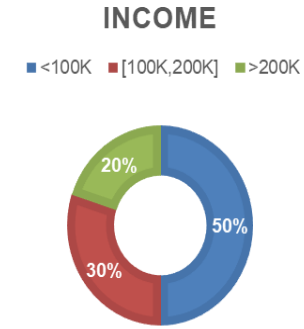
Exemple

Tid	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









Mod: Single

Single	Married	Divorced	NULL
4	3	2	1

Single	Married	Divorced
44%	33%	22%

Putem ignora valorile NULL!!

Percentile

- pentru datele continue, noțiunea de **percentile** este mai utilă
- dat fiind un oridinal sau continuu x și un număr p între 0 și 100, percentila p este o valoare x_p a lui x astel încât p% din valorile observate a lui x sunt mai mici sau egale decât x_p.
- de exemplu, percentila 80 este valoarea $x_{80\%}$ care este mai mare sau egală cu 80% din toate valorile lui x pe care le avem în setul de date.

$x_{80\%}$ = 125K

Taxable Income
10000K
220K
125K
120K
100K
90K
90K
85K
70K
60K

Valori medii și mediane

- **Valoarea medie** (*mean*) este cea mai comună măsură a locației unui set de puncte. $$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$
- **Mediana** este, de asemenea, frecvent utilizată

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- **Trimmed mean**: valoarea medie după eliminearea valorilor minime și maxime

Tid	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

Mean: 1096K

Trimmed mean (remove min, max): 112.5K

Median: (90+100)/2 = 95K

Relație între atribute

- în multe cazuri este interesant să privim împreună două atribute pentru a înțelege dacă sunt corelate
  - de ex., Cum legătura între starea civilă cu înșelăciunea
  - de exemplu, rambursarea se corelează cu venitul mediu?
  - Există o relație între anii de studiu și venit?

Tid	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	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

- Cum vizualizăm aceste relații?

Matricea de confuzie

	No	Yes
Single	3	1
Married	4	0
Divorced	1	1

Matricea de distribuție

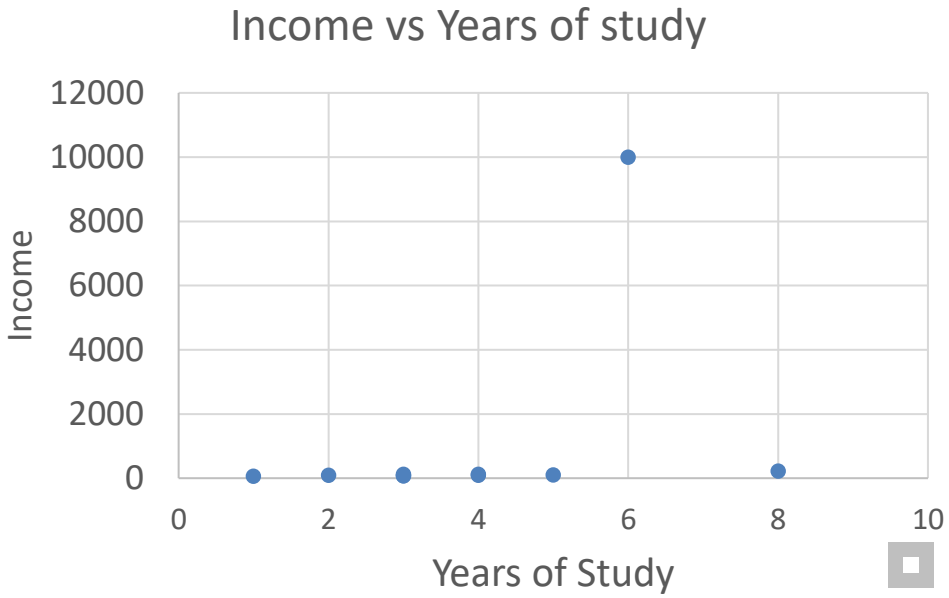
	No	Yes
Single	0,3	0,1
Married	0,4	0
Divorced	0,1	0,1

Corelare atribute

- Grafic de tip *Scatter plot*:
  - Axa X reprezintă un atribut, axa Y pe celălalt

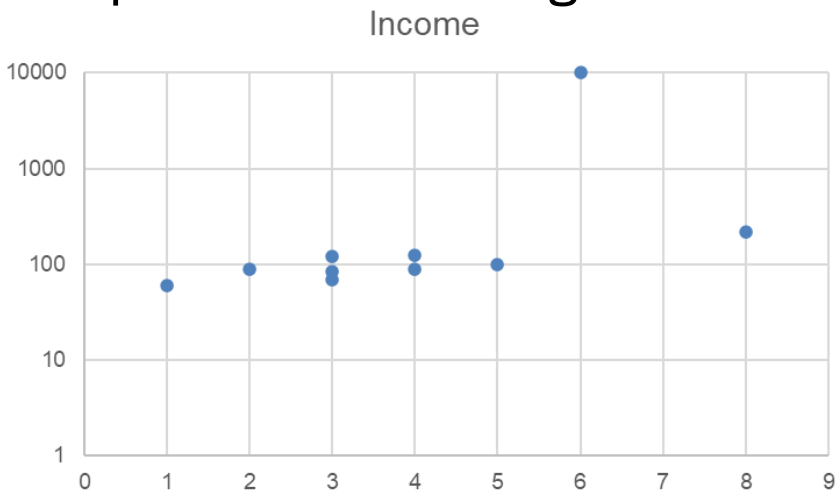
Tid	Refund	Marital Status	Taxable Income	Years of Study
1	Yes	Single	125K	4
2	No	Married	100K	5
3	No	Single	70K	3
4	Yes	Married	120K	3
5	No	Divorced	10000K	6
6	No	NULL	60K	1
7	Yes	Divorced	220K	8
8	No	Single	85K	3
9	No	Married	90K	2
10	No	Single	90K	4

- pentru fiecare intrare avem 2 valori
- intrările sunt reprezentate ca puncte bi-dimensionale

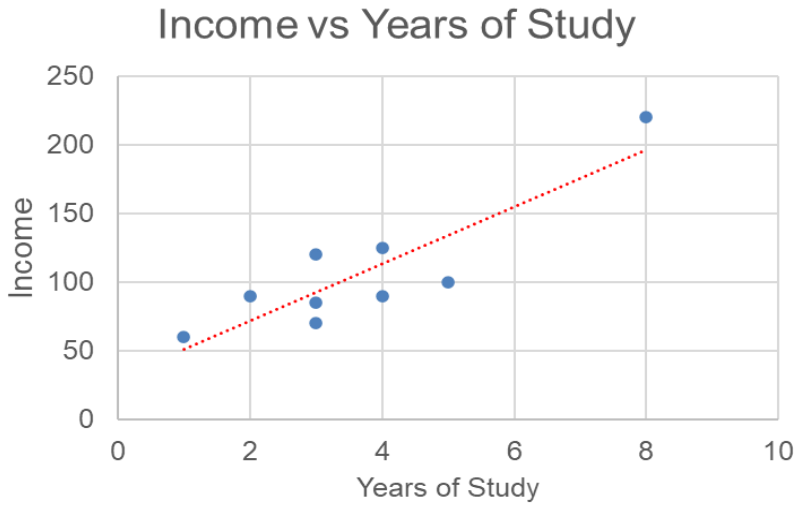


Corelare atribute numerice

- scara logaritmică pe axa Y face ca graficul să arate mai bine



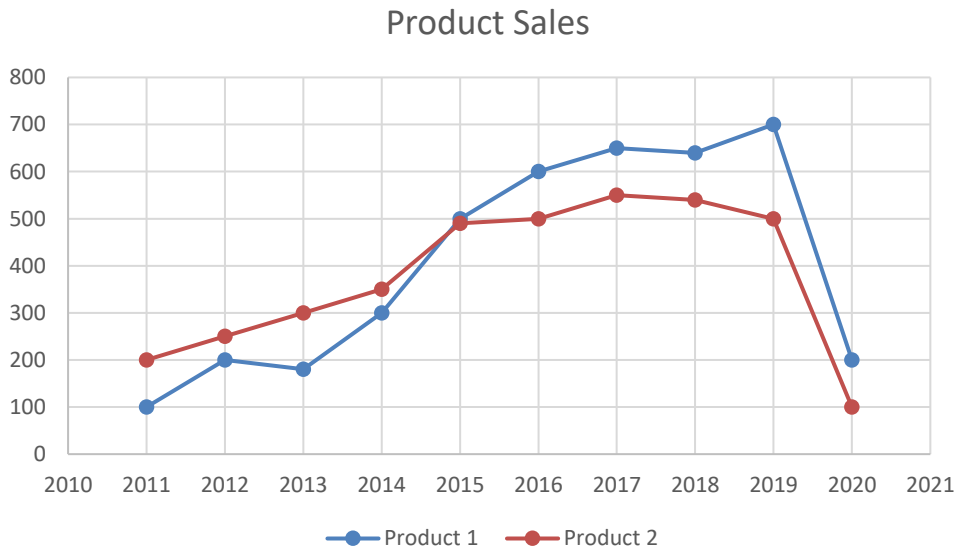
- **După eliminarea outlier-ului** există o corelație clară



Afișarea Atributelor

Year	Product 1	Product 2
2011	100	200
2012	200	250
2013	180	300
2014	300	350
2015	500	490
2016	600	500
2017	650	550
2018	640	540
2019	700	500
2020	200	100

- Cum ați vizualiza diferențele dintre vânzările de produse de-a lungul timpului?



ÎNTREBĂRI ?