Facultatea de Electronică,
Telecomunicații și
Tehnologia Informației

SISTEME INTELIGENTE DE SUPORT DECIZIONAL

Ș.l.dr.ing. Laura-Nicoleta IVANCIU

Curs 4 – Tehnici de Data Mining

Cuprins

- Data mining – definiție
- Standarde și pași
- Tehnici de data mining
- Aplicații și provocări

Data Mining (*minarea/mineritul datelor*)

- prelucrarea cantităților mari de date cu scopul de a descoperi tipare ascunse, anomalii, legături între date
- prelucrare automată, computerizată
- *buzzword*
- denumiri alternative: data dredging/fishing/snooping

CRISP-DM (Cross-industry standard process for data mining)

1996, European Strategic Programme on Research in Information Technology (ESPRIT)

ASUM (Analytics Sollution Unified Method)

2015, IBM

- actualizează CRISP-DM

CRISP-DM (Cross-industry standard process for data mining)

Pași:

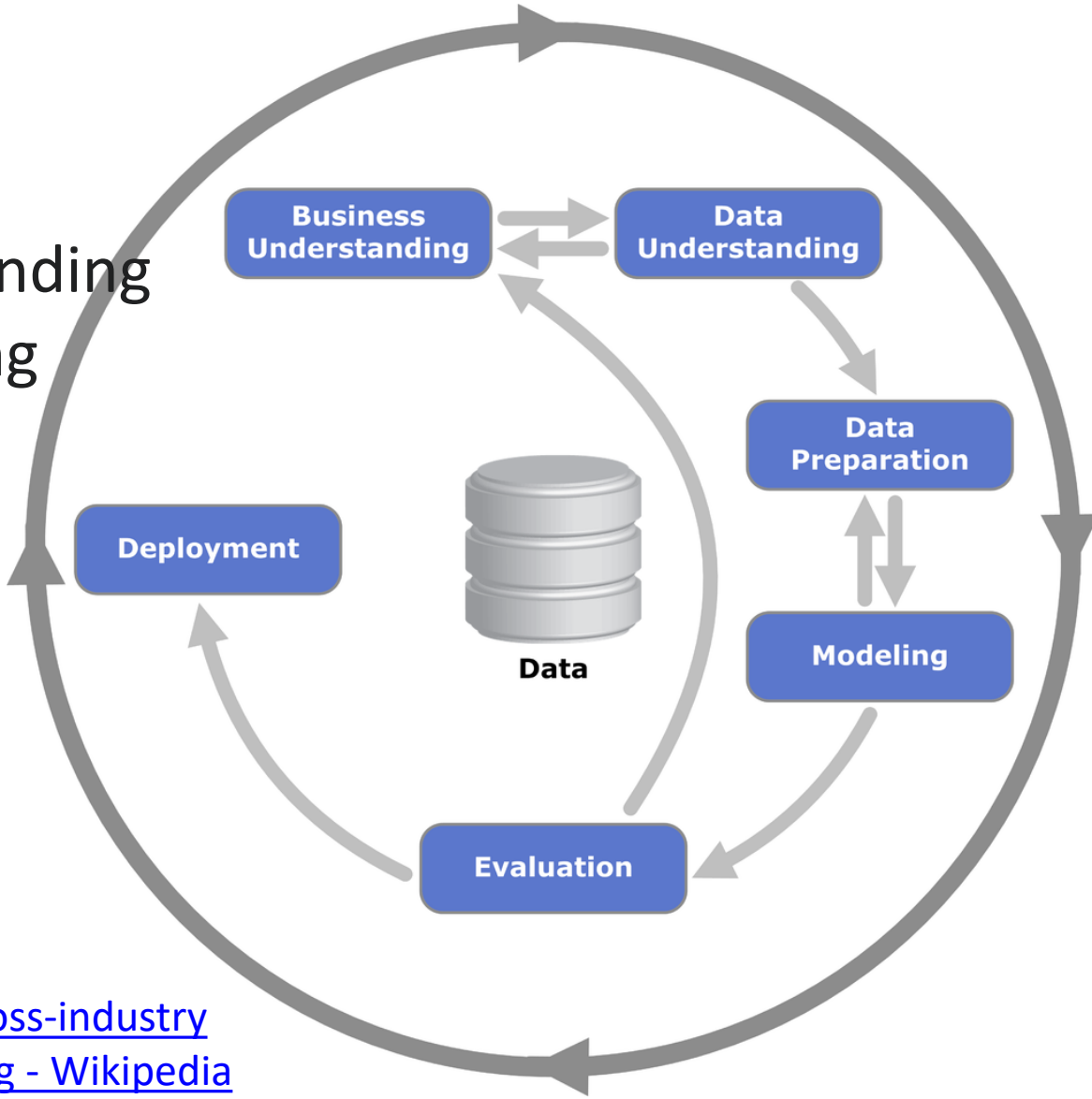
1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Versiune simplificată: (1) Pre-processing, (2) Data Mining, (3) Results Validation.

CRISP-DM

Pași:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment



[CRISP-DM Process Diagram - Cross-industry standard process for data mining - Wikipedia]

CRISP-DM

1. Pre-procesarea datelor

- setul de date trebuie să fie suficient de mare pentru a se putea descoperi tendințe/șabloane
- nu prea mare astfel încât să fie dificil de procesat
- *data cleaning*: eliminare date cu zgomot, date incomplete

CRISP-DM

2. Tehnici de data mining – categorii

Predictive Data Mining

- estimează un rezultat viitor (*future*) pe baza analizei valorilor curente
- nu oferă garanția unui rezultat corect

Descriptive Data Mining

- descrie datele pe baza analizei valorilor anterioare (*past*)

CRISP-DM

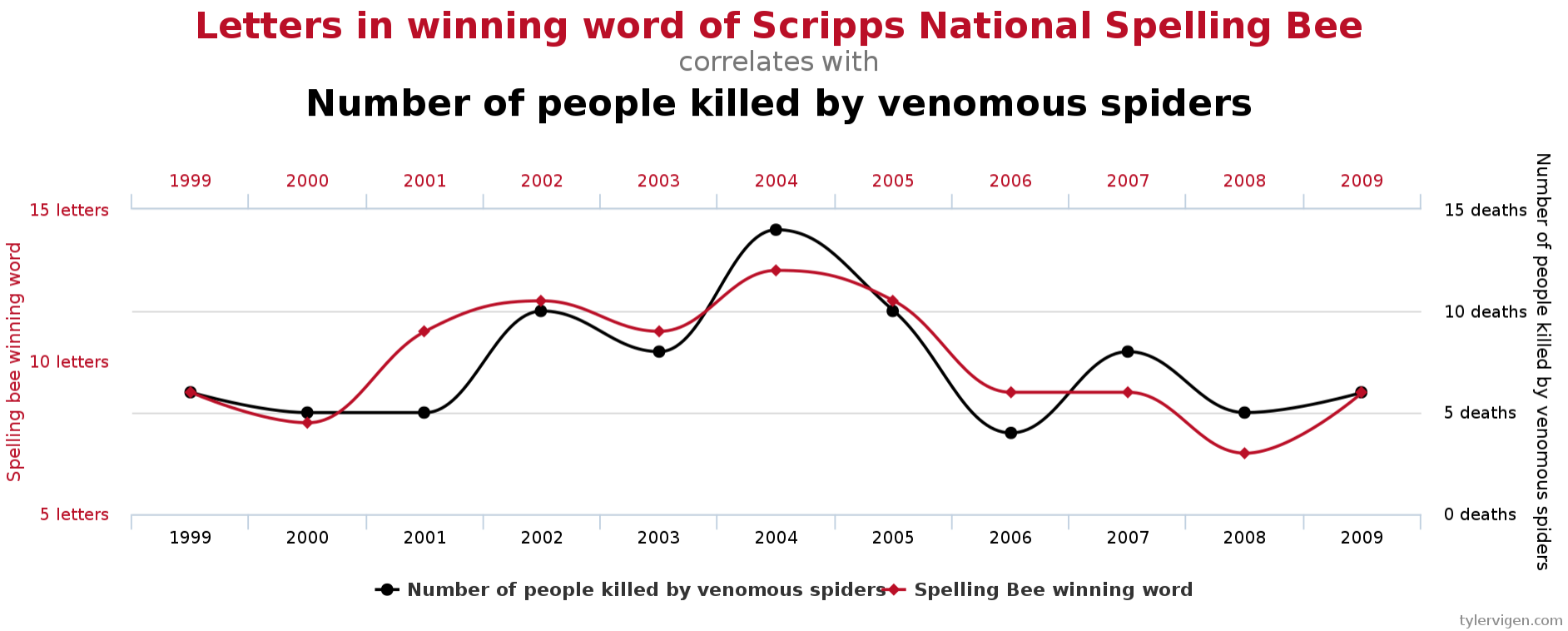
3. Validarea rezultatelor

- unele rezultate pot părea semnificative, dar se regăsesc exclusiv în setul de date analizat, deci nu pot fi utilizate ulterior pentru predicție/estimare

Ex. aruncarea unei monede de 5 ori, de 3 ori cap, de 2 ori stemă
Rezultat analiză: 3/5 și 2/5. Rezultatul este valabil exclusiv pentru datele analizate

CRISP-DM

3. Validarea rezultatelor



Evoluția datelor este similară, dar există vreo legătură?

Sursă

CRISP-DM

3. Validarea rezultatelor

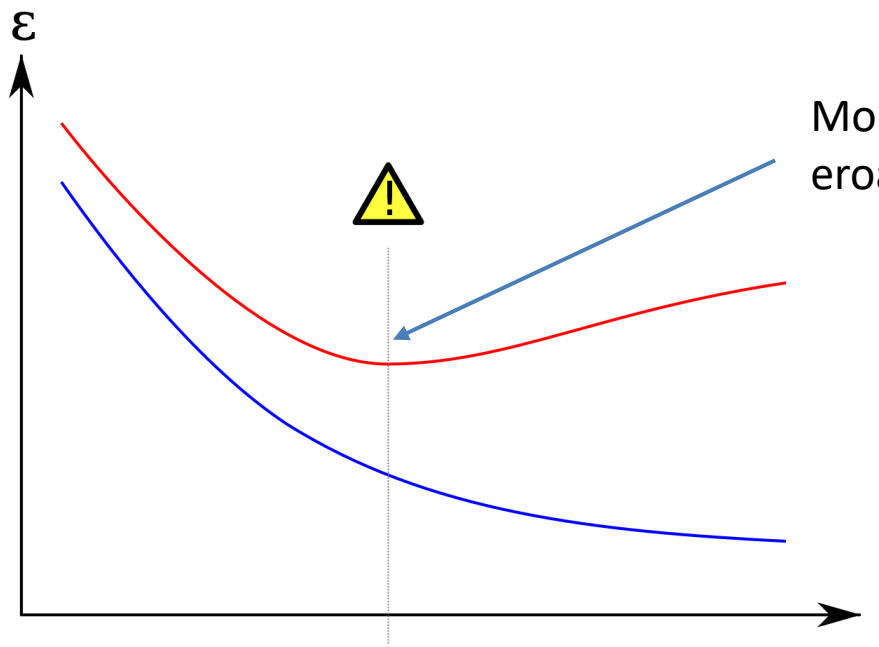
- over-fitting (suprapotrivire) – modelul este “supra-obișnuit” cu datele pe care a fost antrenat

Ex: un filtru spam care a fost antrenat cu anumite adrese de email, cuvinte-cheie din subiect/corp

CRISP-DM

3. Validarea rezultatelor

- over-fitting (suprapotrivire) – soluție: detecție cu set de date de validare/test, pe care algoritmul nu a fost antrenat



Momentul apariției suprapotrivirii, eroarea începe să crească

Sursă

Tehnici de Data Mining

Predictive Data Mining (P)

- Clasificare
- Regresie
- Serii temporale

Descriptive Data Mining (D)

- Detecție de anomalii
- Clustering
- Sumarizare (summarizing)
- Association rule learning

Tehnici de Data Mining

Clasificare (P)

- binară/non-binară (multi-clasă)

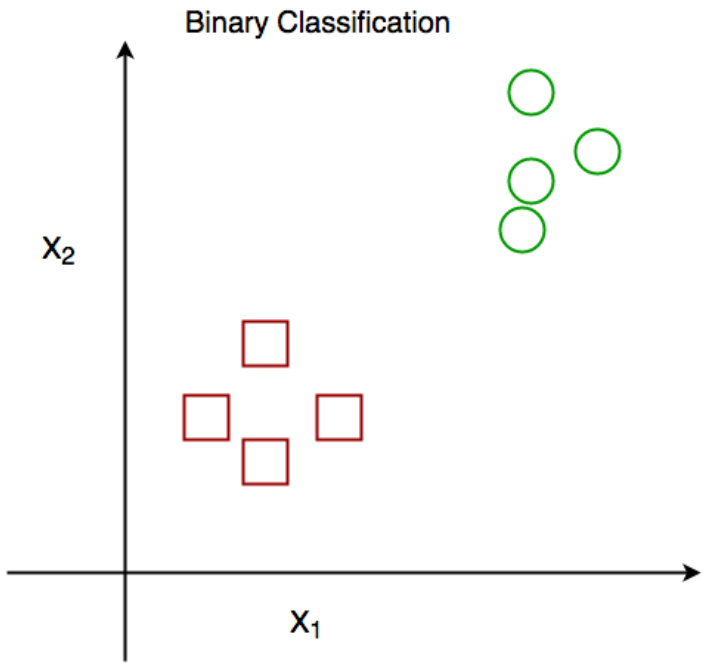
Ex. spam/non-spam, mașină/camion/motocicletă/bicicletă

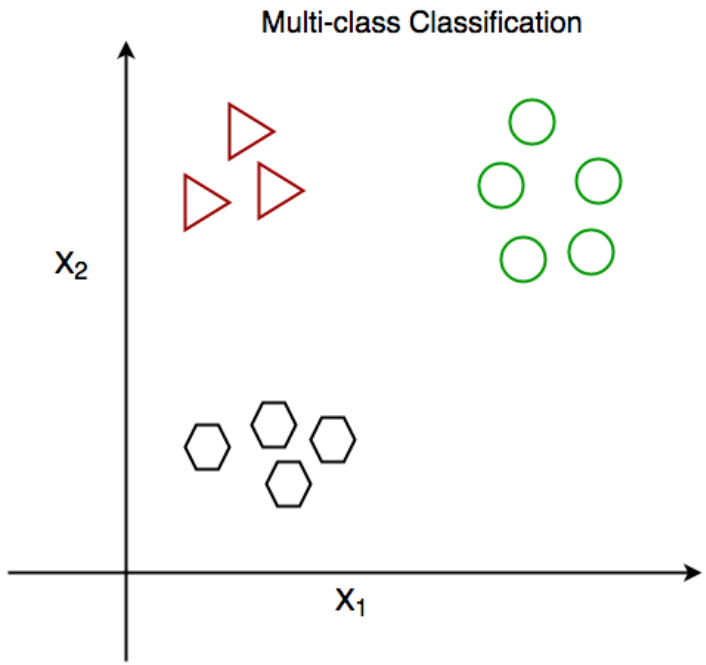
- împărțire pe categorii/clase, pe baza unor observații, proprietăți, trăsături (*explanatory variables, features*)

- algoritm de clasificare = clasificator

Tehnici de Data Mining

Clasificare (P)





Sursă

Tehnici de Data Mining

Regresie (P)

- estimarea relațiilor dintre o variabilă dependentă (rezultat, răspuns, etichetă) și una sau mai multe variabile independente (predictori, explanatory variables, features)

Regresie liniară
– descoperirea relației liniare care estimează cel mai bine distribuția datelor

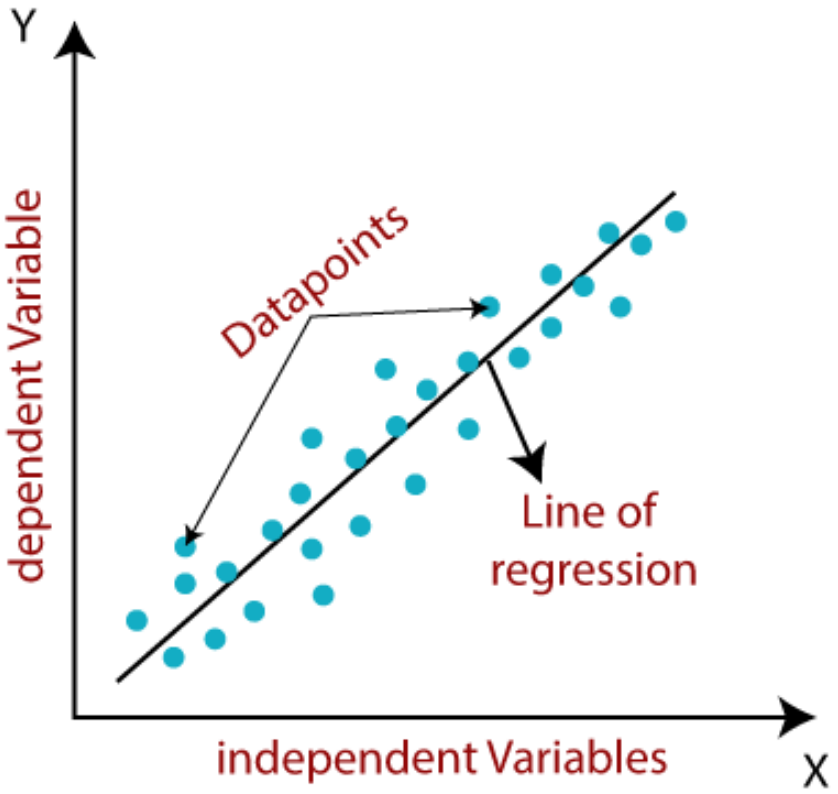
Tehnici de Data Mining

Regresie (P)

Regresie liniară

$$\widehat{y} = ax + b$$

X – variabilă independentă
Y- variabilă dependentă
a – panta
b – termen liber



Sursă

Evaluarea calității predicției prin regresie?

Tehnici de Data Mining

Serii temporale (time-series) (P)

- valori observate la momente de timp egal distanțate

Ex. valori de temperatură, presiune, trafic

Sursă

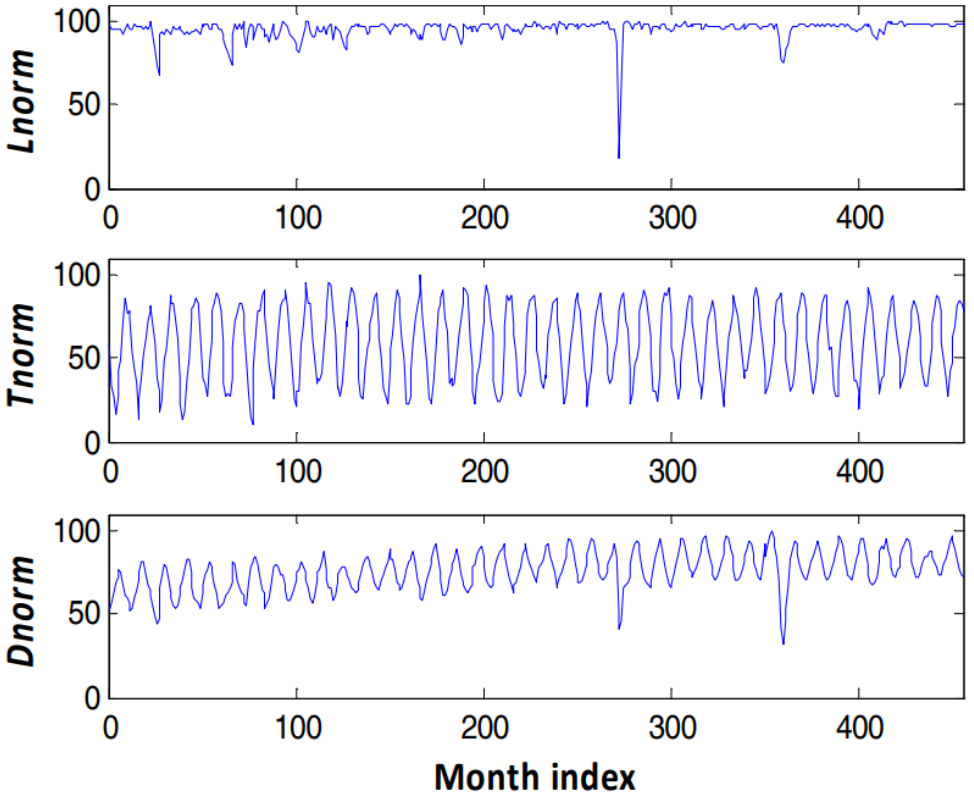
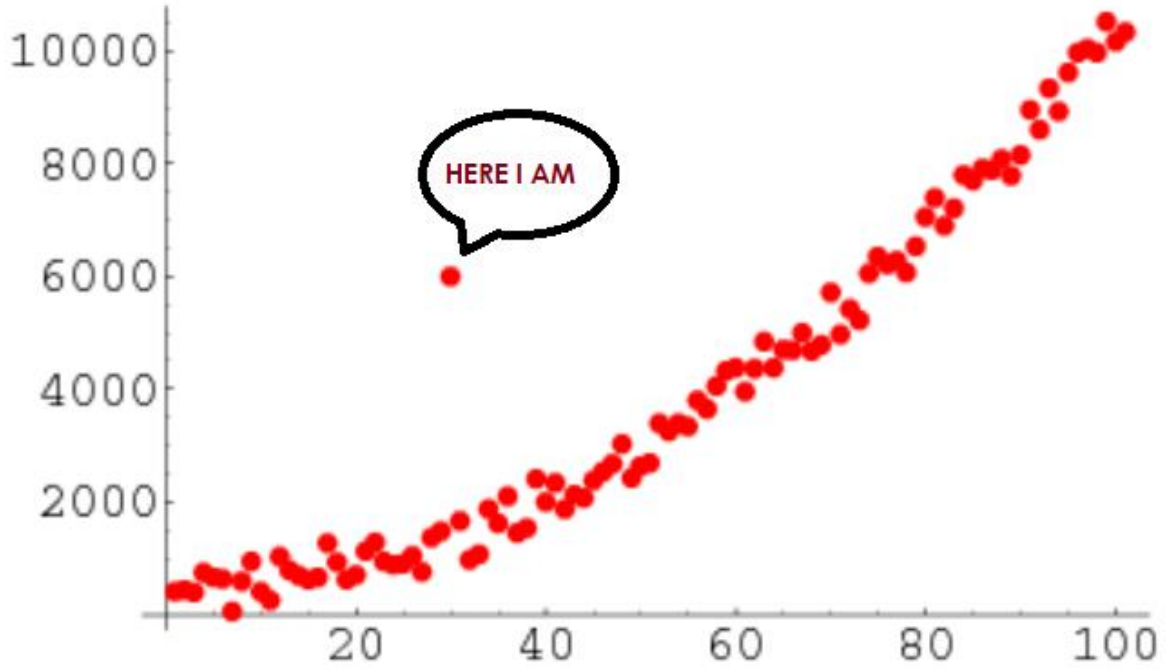


Figure 3. Normalised date series: *Lnorm* – normalized water level; *Tnorm* – normalized temperature; *Dnorm* – normalized horizontal displacement

Tehnici de Data Mining

Detecție de anomalii (D)

- outlier/change/novelty detection



Sursă

Tehnici de Data Mining

Detecție de anomalii (D)

- outlier/change/novelty detection

Ce înseamnă anomalie?

- single/collective outlier, contextual outlier

Aplicații: cyber-security intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, detecting ecosystem disturbances, defect detection in images, medical diagnosis [sursă]

Tehnici de Data Mining

Clustering (D)

- caută asemănări între date, și grupează datele cu trăsături comune în *clustere*

- clusterele nu sunt predefinite (cum era în cazul clasificării)

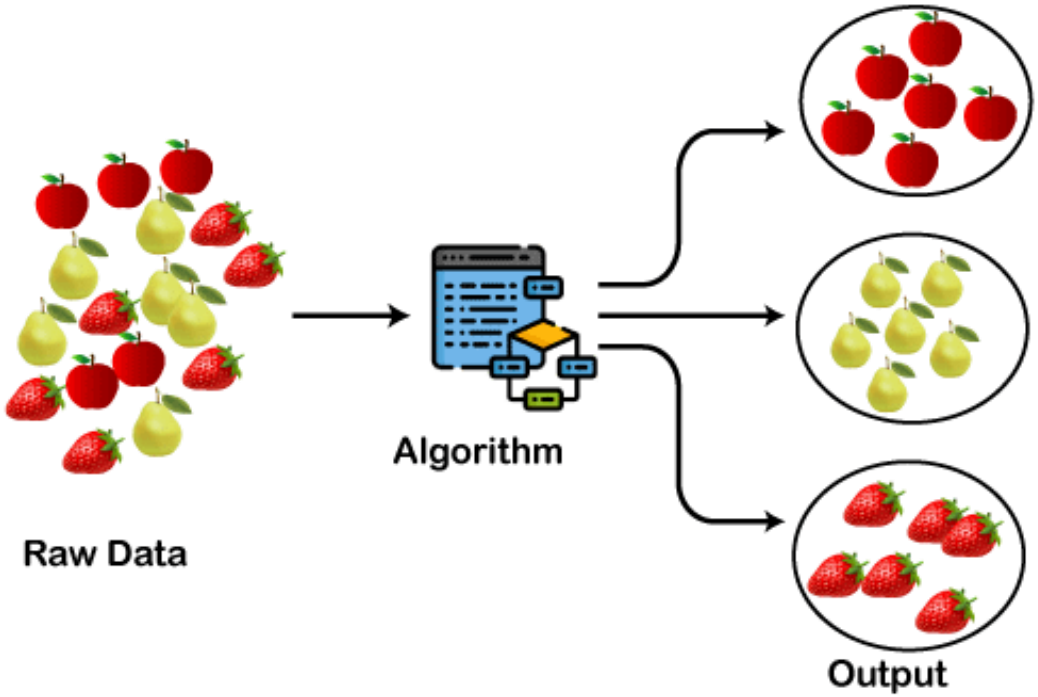
Aplicații: recomandări pe Amazon, Netflix

Sursă

Tehnici de Data Mining

Clustering (D)

- caută asemănări între date, și grupează datele cu trăsături comune în *clustere*



Sursă

Tehnici de Data Mining

Sumarizare (D)

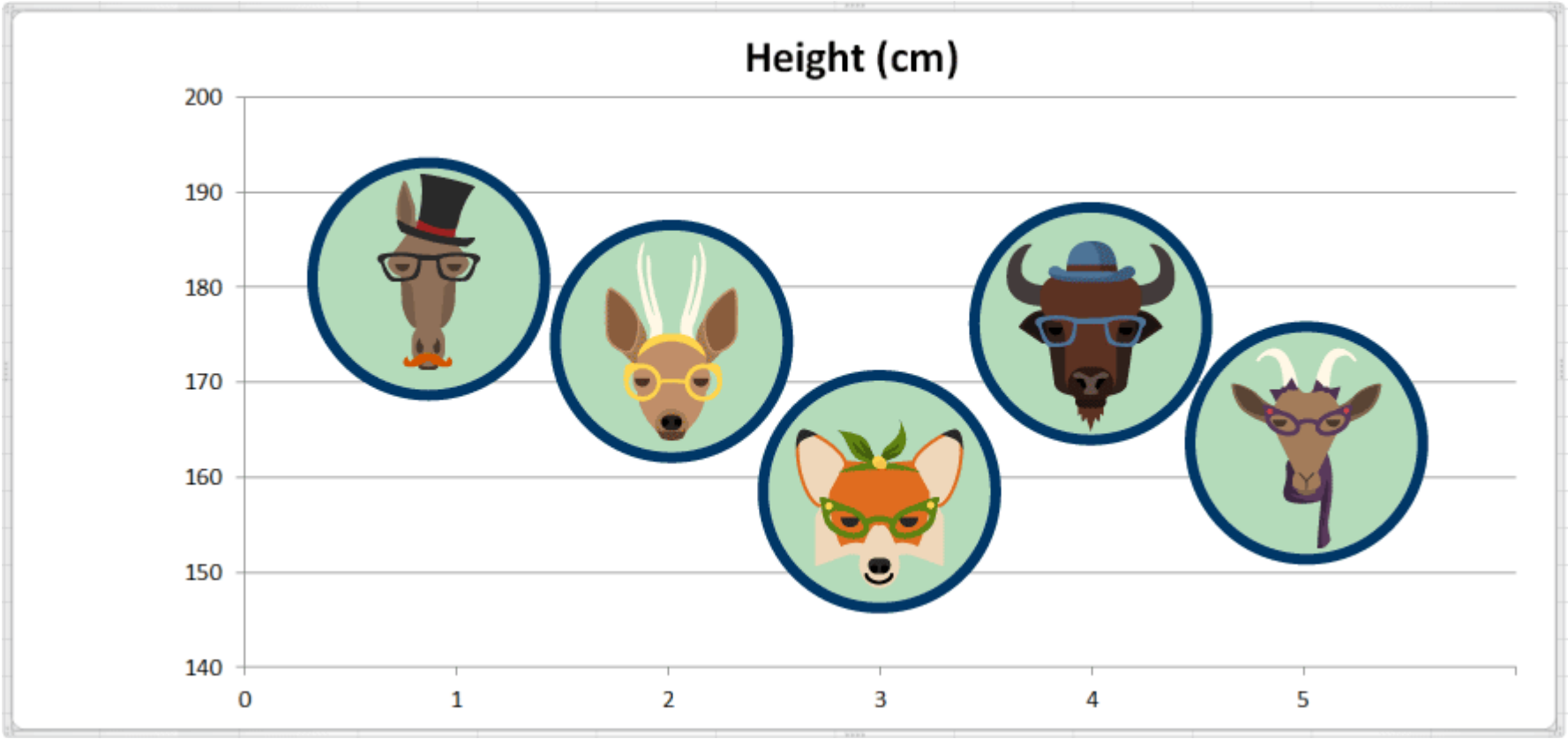
- cea mai accesibilă formă de Data Mining

	A	B
1	**Name**	**Height (cm)**
2	Harry the Horse	181
3	Dana the Deer	175
4	Fran the Fox	159
5	Bob the Buffalo	177
6	Gracie the Goat	165

Sursă

Tehnici de Data Mining

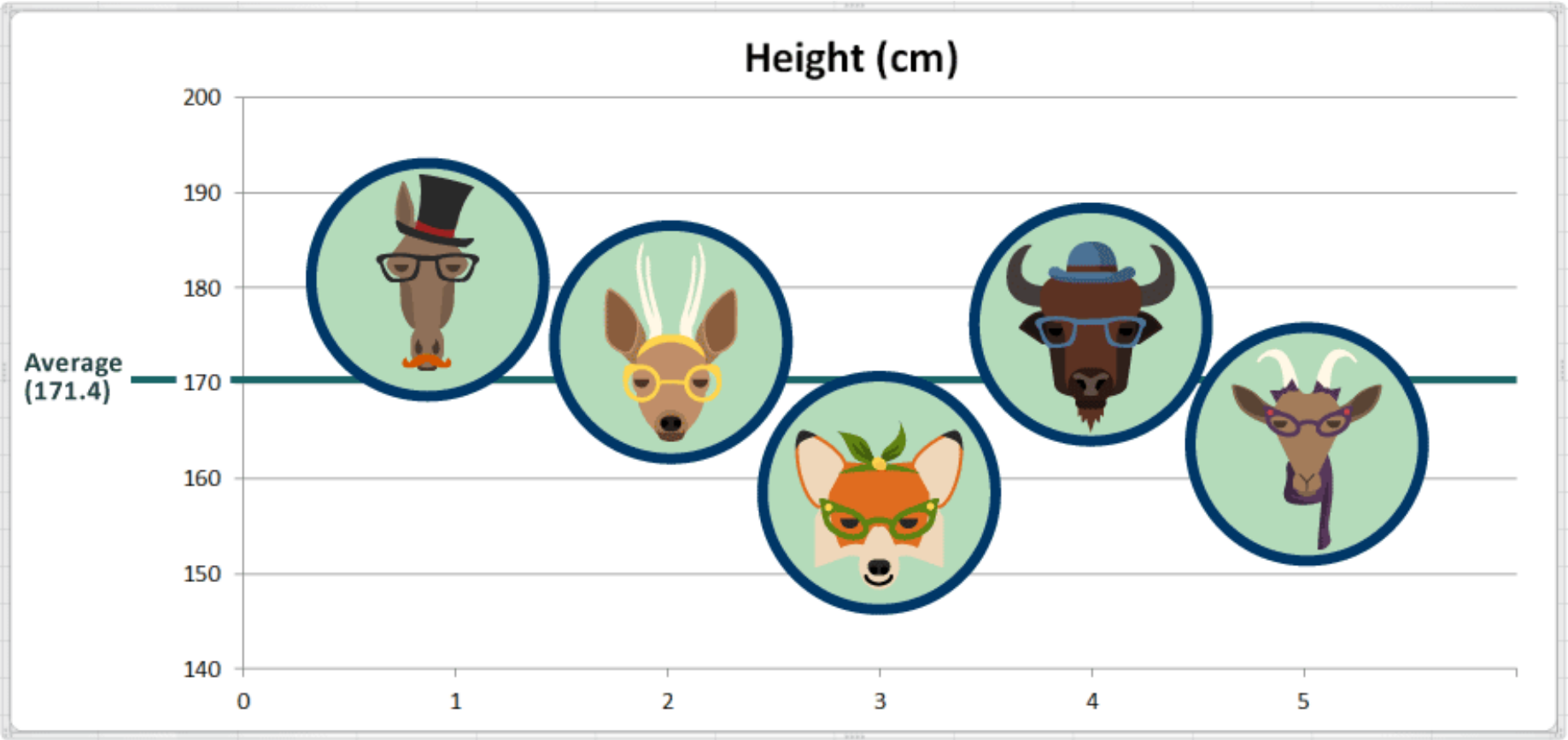
Sumarizare (D)



Sursă

Tehnici de Data Mining

Sumarizare (D)

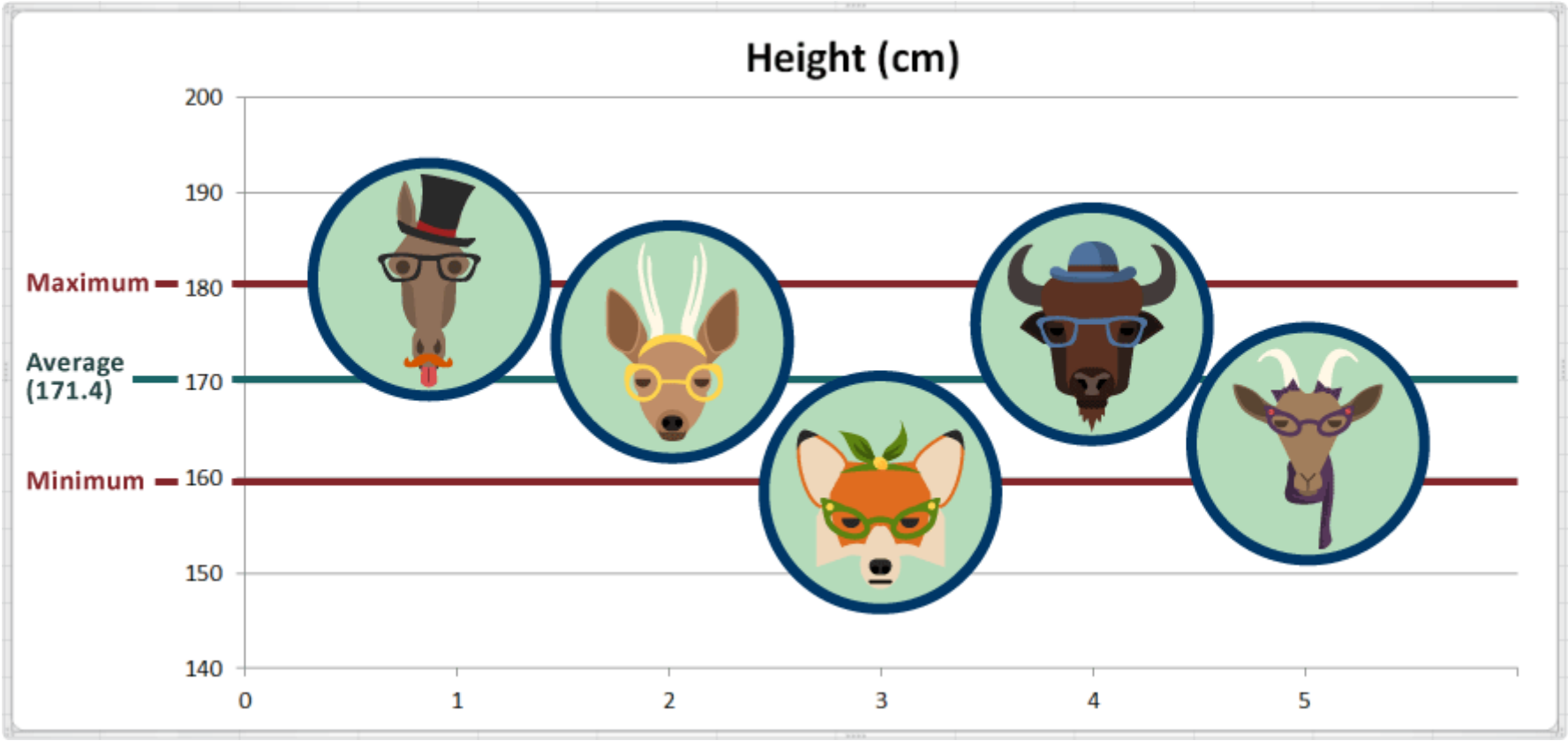


Average height = (181 + 175 + 159 + 177 + 165) ÷ 5 = 857 ÷ 5 = 171.4

Sursă

Tehnici de Data Mining

Sumarizare (D)



Sursă

Tehnici de Data Mining

Association rule learning (D)

- descoperirea de relații între variabilele unui set de date
- regulile care arată de ce și cum sunt conectate variabilele

Ex. analiza cumpărăturilor la supermarket (*market basket analysis*)

Tehnici de Data Mining

Association rule learning (D)

Ex. analiza cumpărăturilor la supermarket (*market basket analysis*)

No.	lapte	pâine	unt	bere	alune	ouă	fructe
1	1	1	0	0	0	0	1
2	0	0	1	0	0	1	1
3	0	0	0	1	1	0	0
4	1	1	1	0	0	1	1
5	0	1	0	0	0	0	0

Posibile reguli? Toate regulile rezultate sunt valide?

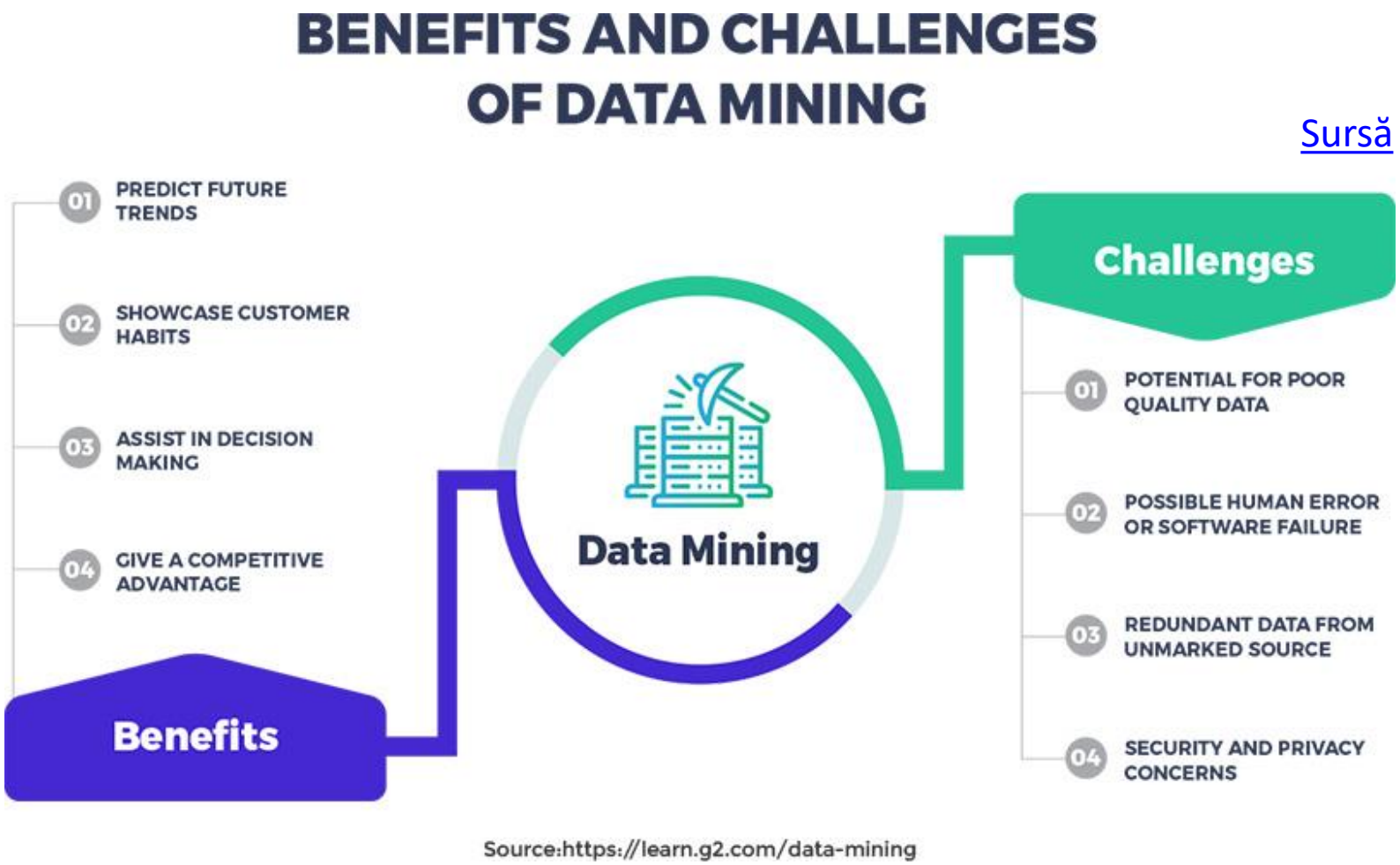
Aplicații ale Data Mining

- e-commerce (*People also viewed, Frequently bought together*)
- stabilirea dinamică a prețurilor pe baza cererii/ofertei
- asigurări (evaluare risc)
- acordare credite

Aplicații ale Data Mining

- CRM (customer relationship management)
  - mesaje și oferte pentru clienți
  - căutare de potențiali noi clienți și particularizare oferte
  - optimizare campanie (targeting)
- medicină (estimare diagnostic, evoluție, răspuns la tratament)
- securitate (fraude informatice, bancare,etc)
- educație (estimare risc de abandon)

Provocări ale Data Mining



Sursă

Provocări ale Data Mining

Probleme de etică

- protecția datelor personale (GDPR)

- discriminare pe baza datelor

ex. acordare de împrumut pe baza etniei; valoare poliță de asigurare pe baza vârstei/sexului

- campanii de marketing particularizate pe baza analizei obiceiurilor de cumpărare

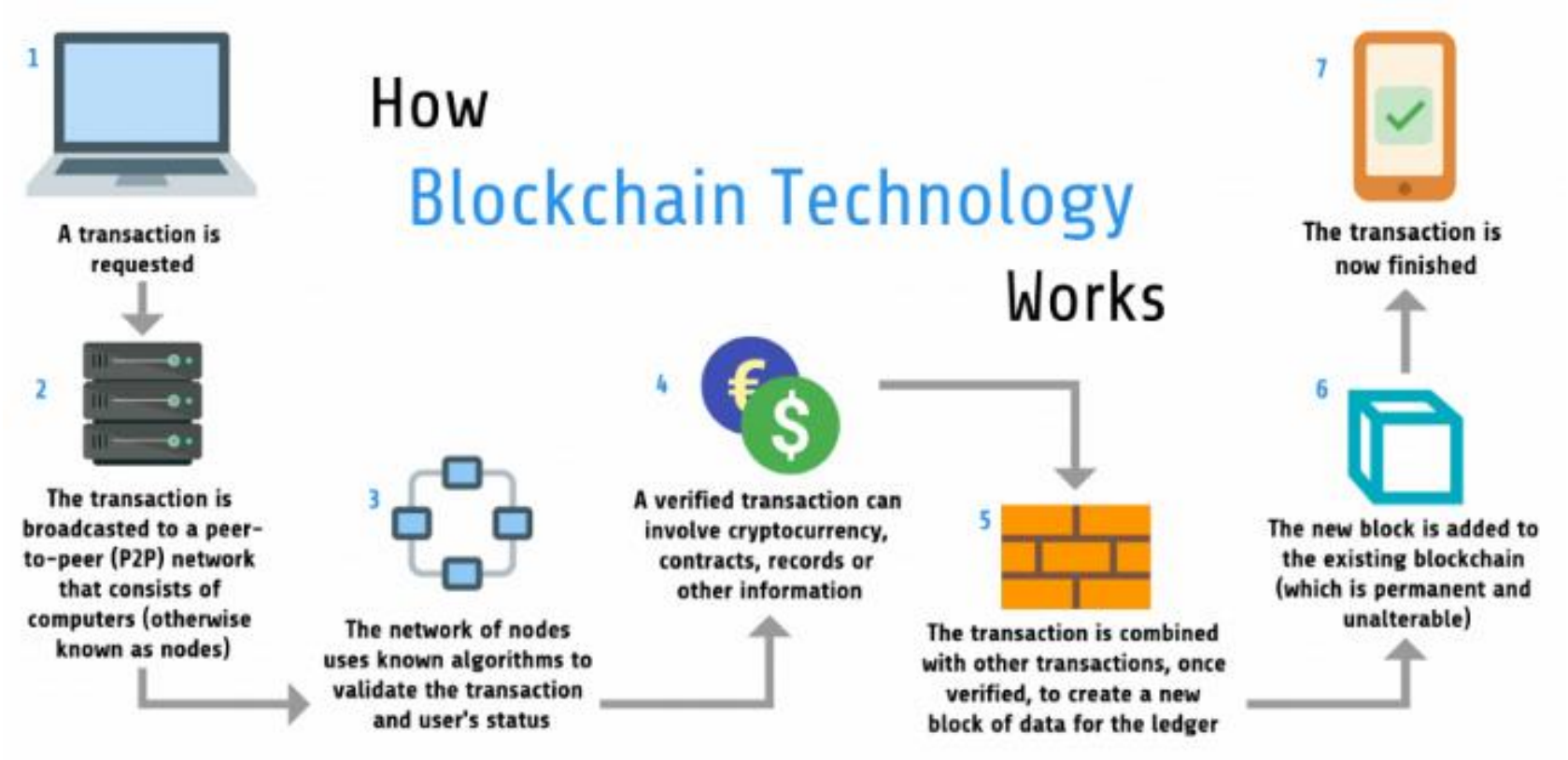
Bitcoin Mining

Sursă

Ce este bitcoin (BTC)?

- monedă virtuală (digitală)
- monedă descentralizată
- *cryptocurrency* – tranzacțiile sunt securizate și verificate prin criptografie
- 2009, Satoshi Nakamoto
- blockchain – distribuited ledger technology
- evoluție volatilă

Bitcoin Mining



Sursă

Bitcoin Mining

Bitcoin Mining

Bitcoin Mining

Bitcoin Mining

Bitcoin Mining

= rezolvare de probleme matematice complexe (soluție de 64 de caractere hexa), utilizând circuite și programe software specializate

ASIC (Application Specific Integrated Circuit)

- fiecare soluție descoperită se anunță în *blockchain*

Bitcoin Mining

Prin *bitcoin mining*, se verifică legitimitatea tranzacțiilor și se introduc noi bitcoin in circulație.

"Bitcoin miners use software to solve transaction-related algorithms that check bitcoin transactions. In return, miners are awarded a certain number of bitcoin per block. This entices them to keep solving the transaction-related algorithms, supporting the overall system."

Estimare Business Insider: 90% din totalul de bitcoin (21 mil. BTC) au fost deja minate. Toate bitcoin vor intra în circulație până în 2140 (bitcoin halving) – 120 ani pentru a mina restul de 10%. Sursă

The Bitcoin reward is divided by 2 every 210,000 blocks, or approximately four years.

Bitcoin Mining

Efecte negative

- consum excesiv de energie electrică
- resurse computaționale mari
- disipare de căldură

*Carbon emissions for mining a single bitcoin rose from **0.9 tons** in 2016 to **113 tons** in 2021—a 126-fold increase* Sursă

Discuție
- ❑ Legalitate
- ❑ Etică - The Ethics Of Crypto: Good Intentions And Bad Actors (forbes.com)
- ❑ Impact asupra mediului

Sumar

- Data mining – definiție ✓
- Standarde și pași ✓
- Tehnici de data mining ✓
- Aplicații și provocări ✓

În episodul următor: Raționament. Învățare/instruire.