BIG DATA

CURS 8

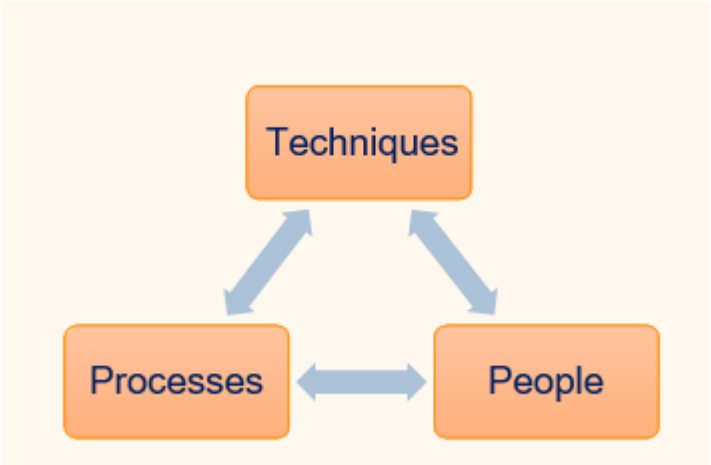
Concepte fundamentale ale analizei datelor

1. **Etapele unui proiect Data Science**
2. **Explorarea datelor**
3. **Analiza datelor**
4. **Descoperirea regulilor de asociere**

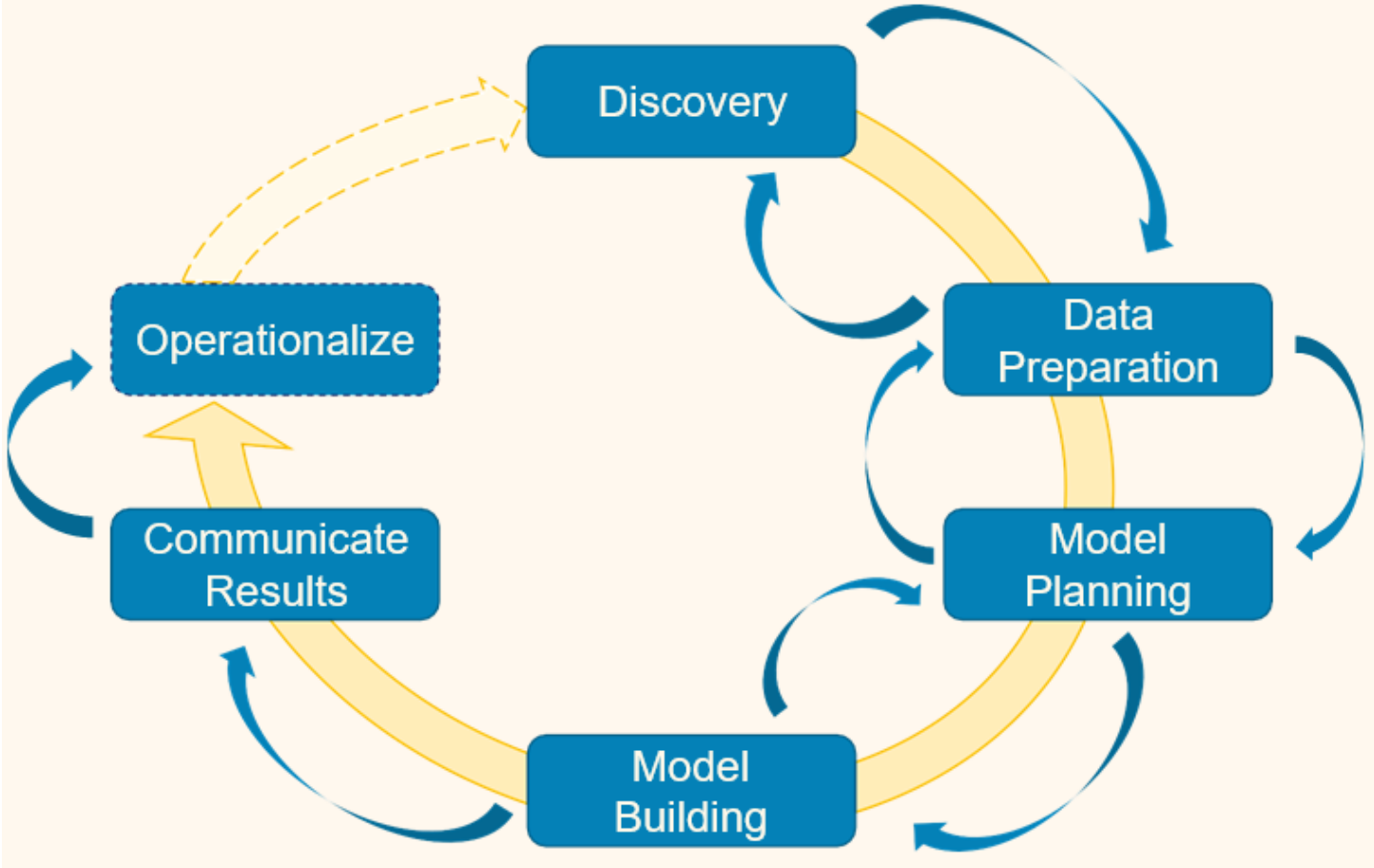
1. Etapele unui proiect Data Science

Desfășurarea unui proiect Data Science

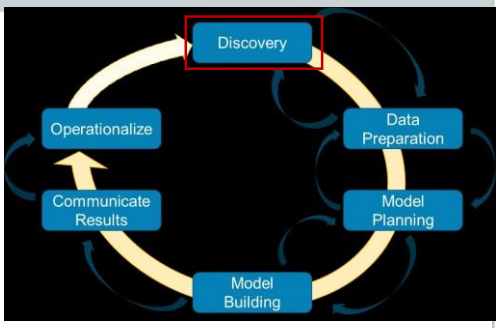
- Procesul desfășurării unui proiect DS
- În contextul unui proiect regăsim tehnici, persoane, procese
- Tehnicile presupun:
  - Limbaje, utilitare, metode
  - Trebuie să fie adecvate pentru problema dată
- Utilizatori (persoane):
  - Necesită învățarea tehnicilor
  - Trebuie ghidați în proiect printr-un proces
- Procesele:
  - Vin în sprijinul utilizatorilor
  - Trebuie să fie acceptate de către utilizatori
  - Trebuie să aibă un efect pozitiv măsurabil.



Procesul unui proiect DS



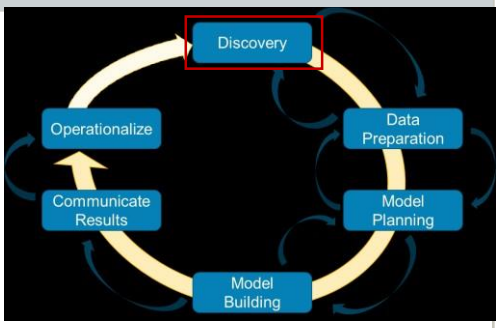
Etapele unui proiect DS



1. Descoperirea

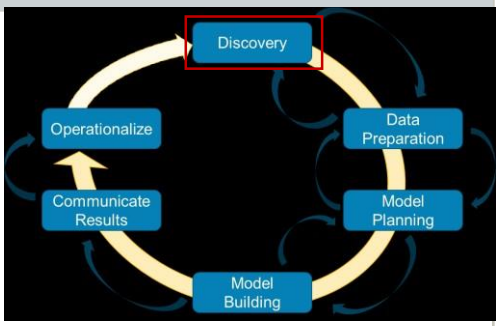
- Etapa inițială a proiectului
- Învățarea domeniului – însușirea de cunoștințe pentru:
  - înțelegerea datelor și a studiilor de caz ale proiectului
  - interpretarea rezultatelor
- Învățarea din experiența anterioară
  - Identificarea proiectelor din trecut ce tratează probleme similare
    - Diferențe, motive ale eșecurilor, puncte slabe

Etapele unui proiect DS



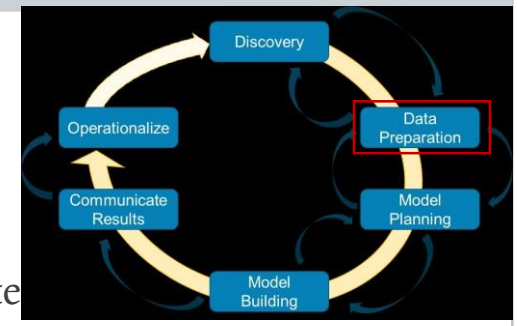
- Încadrarea problemei
  - Enunțarea problemei de analiză a datelor ce urmează a fi rezolvată
  - De ce este importantă, cine sunt beneficiarii, care este situația actuală și ce anume motivează derularea proiectului?
  - Care sunt obiectivele proiectului?
  - Ce trebuie făcut pentru a atinge obiectivele?
  - Care sunt riscurile proiectului?
- Învățarea datelor
  - Nivel înalt de înțelegere a datelor
    - Poate include statistici inițiale și vizualizări ale acestora
  - Determinarea cerințelor pentru structurile de date și a utilitarelor pemtru procesarea datelor

Etapele unui proiect DS



- Formularea ipotezelor
  - Aparţine lui „Science” din „Data Science”
  - Trebuie să definească aşteptările, de tipul:
    - O anumită caracteristică este adecvată pentru o anumită predicţie
    - Ce *pattern*-uri vor fi găsite în date
    - Arborii de decizie vor avea rezultate bune pentru anumite probleme.
- Analiza resurselor disponibile
  - Tehnologii
    - Resurse de calcul şi stocare
    - Framework-uri pentru analiză
  - Date
    - Suficiente pentru studiul de caz?
    - Sunt necesare şi alte date? Datele suplimentare pot fi obţinute?
  - Timp
  - Resurse umane, *skillset* adecvat

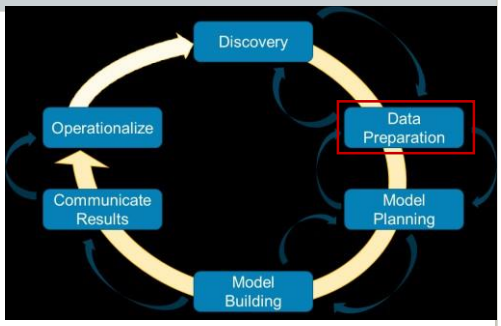
Etapele unui proiect DS



2. Pregătirea datelor

- Creează infrastructura proiectului
  - De obicei, diferită de infrastructura în care datele sunt furnizate
  - Warehouse/fişiere csv etc. → stocare distribuită ce permite analiza
    - Poate fi mai simplă, pentru datele de mici dimensiuni
- Extract-Transform-Load (ETL) asupra datelor
  - Defineşte modul în care se pot interoga bazele de date existente pentru a extrage datele necesare
  - Determină transformările necesare ale datelor brute
    - Verificarea calităţii (filtrarea datelor lipsă sau a celor neplauzibile)
    - Structurare (pentru datele nestructurate sau a celor având o structură diferită)
    - Conversii (timestamps, encodare de caractere etc.)
  - Încărcarea datelor în mediul folosit pentru analiza acestora

Etapele unui proiect DS

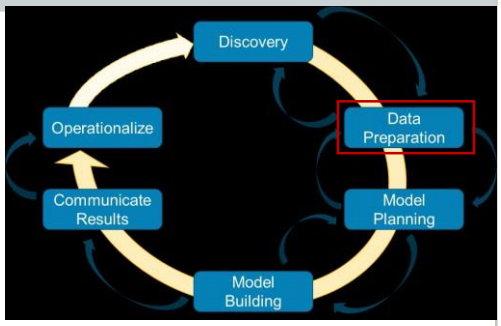


- ELT vs. ETL
  - Transformările pot fi mari consumatoare de timp în big data
  - Ar putea să nu fie posibile fără utilizarea infrastructurii pentru analiză

    => Se încarcă datele brute, ce vor fi transformate apoi

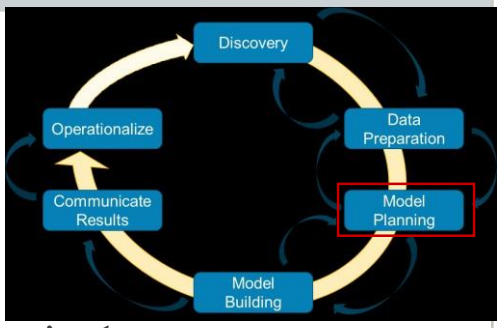
    -> ELT
  - Permite mai multă flexibilitate în privința transformărilor
    - De exemplu, testarea efectului diferitelor transformări
  - Permite accesul datelor brute
- Înțelegerea în profunzime a datelor din toate sursele
  - Ce conțin coloanele din bazele de date relaționale?
  - Cum se poate aplica o structură pe date semi/cvasi/nestructurate?

Etapele unui proiect DS



- Studiul şi vizualizarea datelor
  - Statistici descriptive
  - Analiza corelărilor
  - Vizualizări precum histograme, reprezentări grafice pentru densitate etc.
- Curăţarea şi normalizarea datelor
  - Eliminarea datelor ce nu sunt necesare
  - Normalizarea pentru a elimina problemele cauzate de diferenţele de scală
  - Poate face diferenţa între o infrastructură complexă şi o singură maşină folosită pentru analiză
- Exemplu:
  - 100.000.000 înregistrări cu câte 10 caracteristici în virgulă mobilă / înregistrare => 80 octeţi/ înregistrare
  - 3 caracteristici sunt utile => aprox. 24 octeţi/înregistrare
  - Pentru toate caracteristicile: 7.45 GB
  - Pentru caracteristicile utile: 2.23GB => se poate folosi o singură maşină

Etapele unui proiect DS

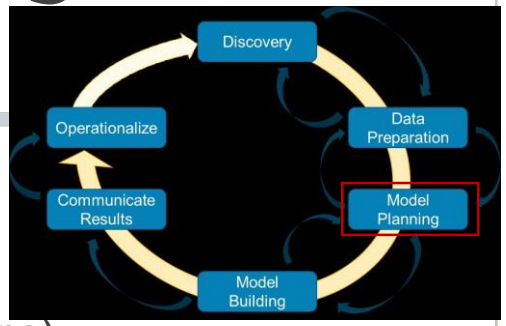


3. Planificarea modelului

- Determină metodele pentru analiza datelor
- Trebuie să fie adecvate pentru a permite realizarea obiectivelor propuse
  - Adesea, determină tipul de metodă
    - Clasificare, regresie, clustering, reguli de asociere etc.
  - Alți factori pot restricționa metodele disponibile
    - Dacă analiza este importantă, metodele blackbox nu pot fi utilizate
- Trebuie să fie adecvate datelor disponibile
  - Volum, structură etc.

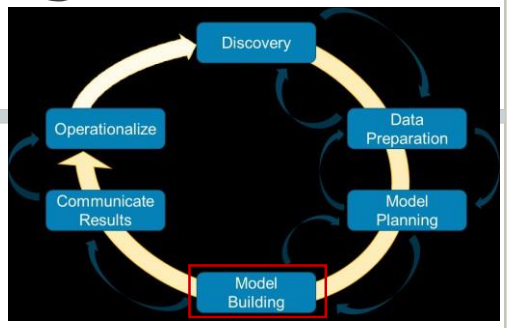
O metodă de tip blackbox este o metodă prin care doar se obțin rezultate, fără a fi înțeles modul în care s-a realizat calculul.
O metodă whitebox explică modul în care a fost obținut rezultatul.

Etapele unui proiect DS



- Metodele pentru analiza datelor pot acoperi:
  - Modelarea caracteristicilor (de exemplu, pentru text mining)
  - Selectarea caracteristicilor (de exemplu, pe baza câștigului de informație, corelări etc.)
  - Crearea modelului (pot exista modele diferite pentru un studiu de caz)
  - Metode statistice (pentru compararea rezultatelor)
  - Vizualizări (pentru prezentarea rezultatelor)
- Divizarea datelor în seturi de date diferite
  - Date de antrenare, date de validare, date de test
  - Considerarea unei mulțimi mici de date pentru a fi folosite local, în cazul big data
    - Date cu aceeași structura, dar cu volum mult mai mic

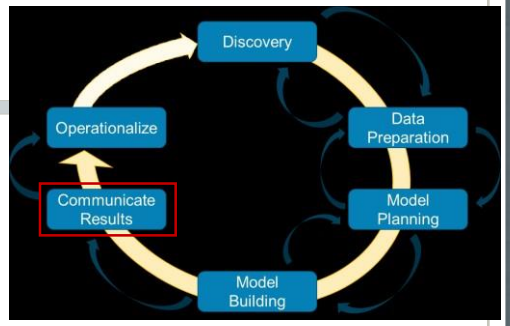
Etapele unui proiect DS



4. Construirea modelului

- Efectuează analiza folosind metodele planificate
  - De obicei, procesul este iterativ
- Etapă separată, deoarece poate dura foarte mult timp
  - Se pot folosi seturile de date mici la planificarea modelului
  - Se folosește setul real big data, potențial cu mulți hiperparametri ce pot fi optimizați în timpul construirii modelului
- Include calculul indicatorilor de performanță

Etapele unui proiect DS

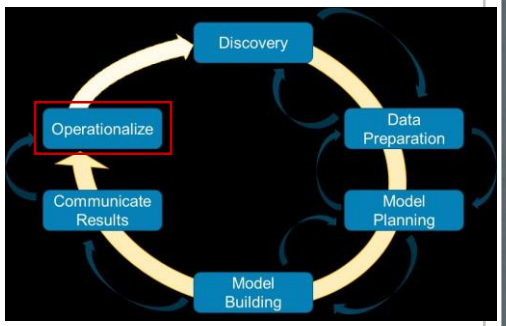


5. Comunicarea rezultatelor

- Proiectul a avut succes?
- Compararea rezultatelor cu ipotezele din etapa de descoperire
- Identificarea rezultatelor cheie, sumarizarea acestora pentru diferite tipuri de audiență
- Cuantificarea valorii rezultatelor – valoare de business (Return on Investment – ROI)

6. Operaționalizarea

- Implementarea rezultatelor
- Rularea unei versiuni pilot la început
- Definirea unui proces de actualizare și reantrenare a modelului
  - Datele se învechesc, modelele nu mai sunt actuale
  - Modelele *data driven* trebuie actualizate regulat



Livrabile

- Prezentarea pentru sponsori (audiență non-tehnică)
- Prezentarea pentru analiști
- Codul sursă
- Specificațiile tehnice

2. Explorarea datelor

Scopul explorării datelor

- Scop:
  - Înțelegerea caracteristicior de bază ale datelor
  - Exemple de caracteristici:
    - Structură
    - Dimensiune
    - Completitudine
    - Relații

Metode de explorare a datelor

- De obicei, interactiv și semiautomat
- Vizualizarea datelor brute – se utilizează editoare de text, comenzi ale sistemului (tail/head/more/less)
  - Ajută la înțelegerea structurii
- Statistici și vizualizări ce oferă informații despre distribuții și relații
- Explorarea include metadatele
  - Numele caracteristicilor, observarea legăturilor dintre date etc.

Statistici descriptive

- Sumarizează datele
- Nu fac nicio predicție asupra datelor (spre deosebire de statistica inductivă)
- Statistici de bază:
  - Tendința centrală (medie, mediană/mod)
  - Variabilitate (deviație standard, domeniu interquartil)
  - Domeniu de date (min/max)
- Alte statistici importante:
  - Indicele de aplatizare (*kurtosis*) și indicele de asimetrie (*skewness*) ale formei distribuțiilor
  - Mai multe măsuri pentru tendința centrală (medie redusă, medie armonică etc.)

Tendința centrală

- Valoarea „tipică“ a datelor
- Media aritmetică
  - $mean(x) = \frac{1}{n}\Sigma_{i=1}^{n} x$ unde $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$
- Mediană
  - Valoarea ce separă jumătatea superioară de datele din jumătatea inferioară
- Mod
  - Valoarea ce apare de cel mai mare număr de ori în date

Variabilitate

- Măsură pentru împrăștierea datelor
  - Se mai numește dispersie
- Deviația standard
  - Măsură pentru diferența dintre observație și media aritmetică
  - $sd(x) = \sqrt{\frac{\Sigma_{i=1}^{n}(x - mean(x))^2}{n}}$
- Domeniu interquartil (Inter-quartile range – IQR)
  - Percentilă: valoarea sub care se află un anumit procent al datelor
  - Diferența dintre percentila 75% și percentila 25%

Mediana este percentila 50%

Domeniul datelor

- Domeniul în care sunt observate valori
  - Poate fi infinit
- Minimum
  - Cea mai mică valoare observată
- Maximum
  - Ce mai mare valoare observată
- Pot fi distorsionate de către datele invalide
  - Pot constitui și un mod de a descoperi datele invalide

Exemplu

- Scriere random la tastatură:
  - x = (1,2,1,1,3,4,5,2,3,4,5,1,3,2,1,6,5,4,9,4,3,6,1,5,6,8,4,6,5,1,3,2,1,6,8,7,6,1,3,1,6,8,4,7,6,4,3,5,4,9,7,4,3,1,4,6,8,7,9, 1,4,6,1,3,8,6,7,4,9,6,5,1,3,6,8,7)
- Tendința centrală:
  - media: 4.46052631579
  - mediana: 4.0
  - modul (număr apariții): 1 (14)
- Variabilitate:
  - sd: 2.41944311488
  - IQR: 3.0
- Domeniu:
  - min: 1
  - max: 9

Vizualizări pentru explorarea datelor

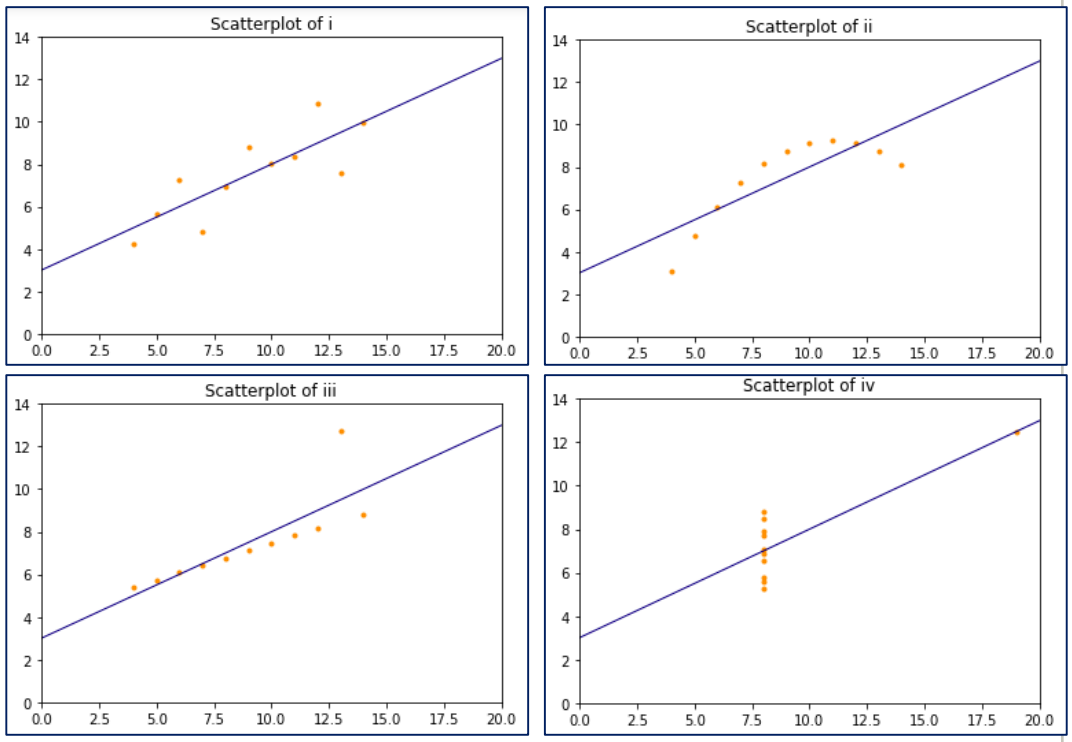
Statistici „înșelătoare”

Au aceeași:
- Medie
- Deviație standard
- Corelare între x și y
- Regresie liniară

i		ii	
X	y	x	y
10.00	8.04	10.00	9.14
8.00	6.95	8.00	8.14
13.00	7.58	13.00	8.74
9.00	8.81	9.00	8.77
11.00	8.33	11.00	9.26
14.00	9.96	14.00	8.10
6.00	7.24	6.00	6.13
4.00	4.26	4.00	3.10
12.00	10.84	12.00	9.13
7.00	4.82	7.00	7.26
5.00	5.68	5.00	4.74

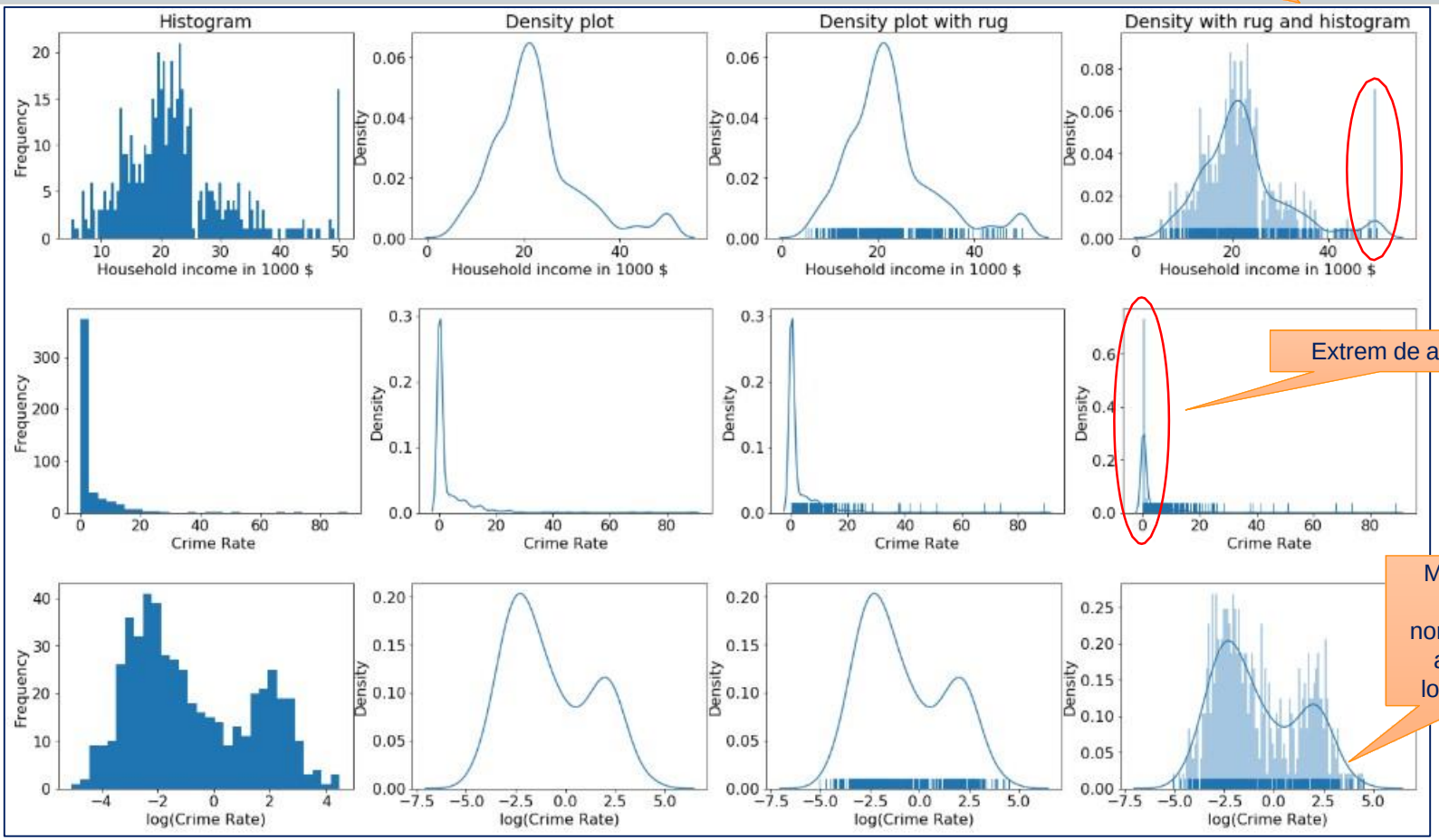
iii		iv	
x	y	x	y
10.00	7.46	8.00	6.58
8.00	6.77	8.00	5.76
13.00	12.74	8.00	7.71
9.00	7.11	8.00	8.84
11.00	7.81	8.00	8.47
14.00	8.84	8.00	7.04
6.00	6.08	8.00	5.25
4.00	5.39	19.00	12.50
12.00	8.15	8.00	5.56
7.00	6.42	8.00	7.91
5.00	5.73	8.00	6.89

Cvartetul lui Anscombe



Explorarea caracteristicilor

Pare o valoare mare în mod artificial -> Grupează toate veniturile mari

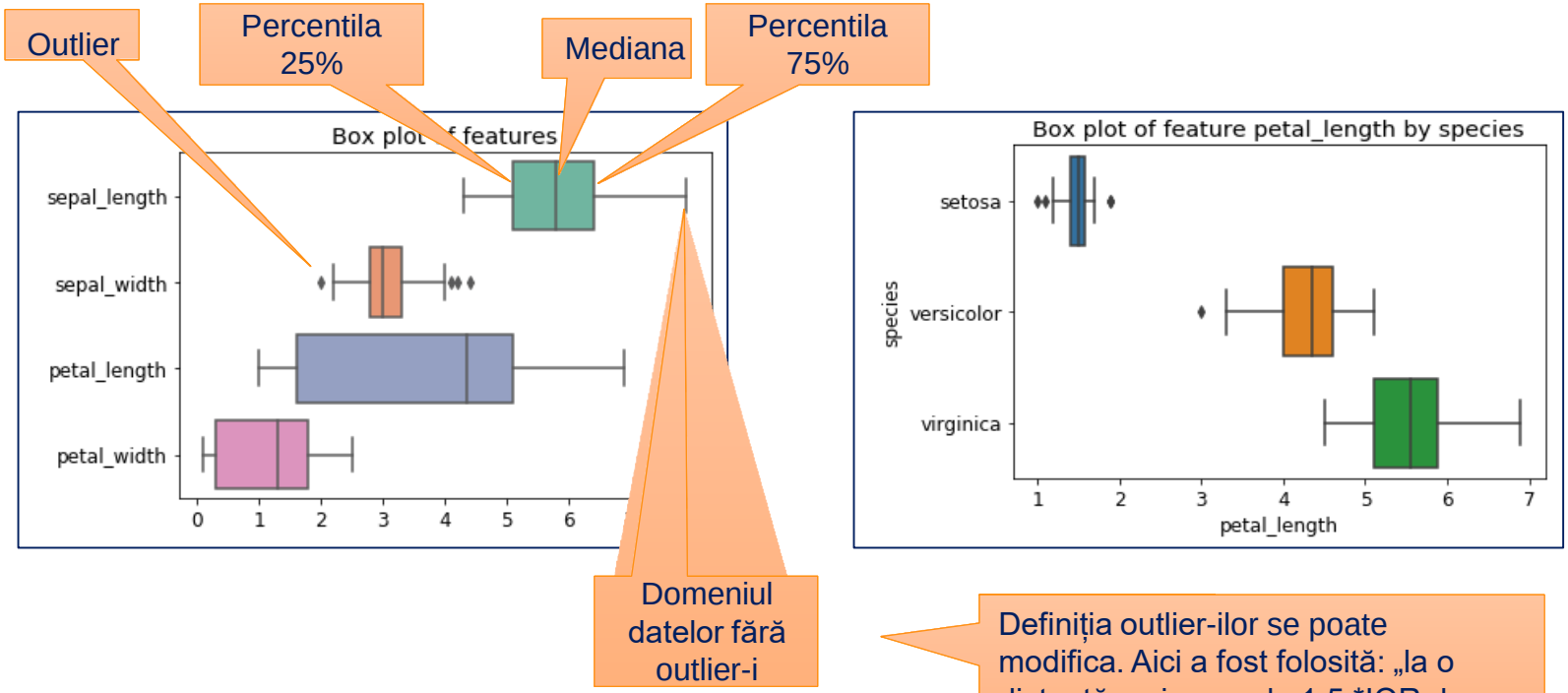


Extrem de asimetric

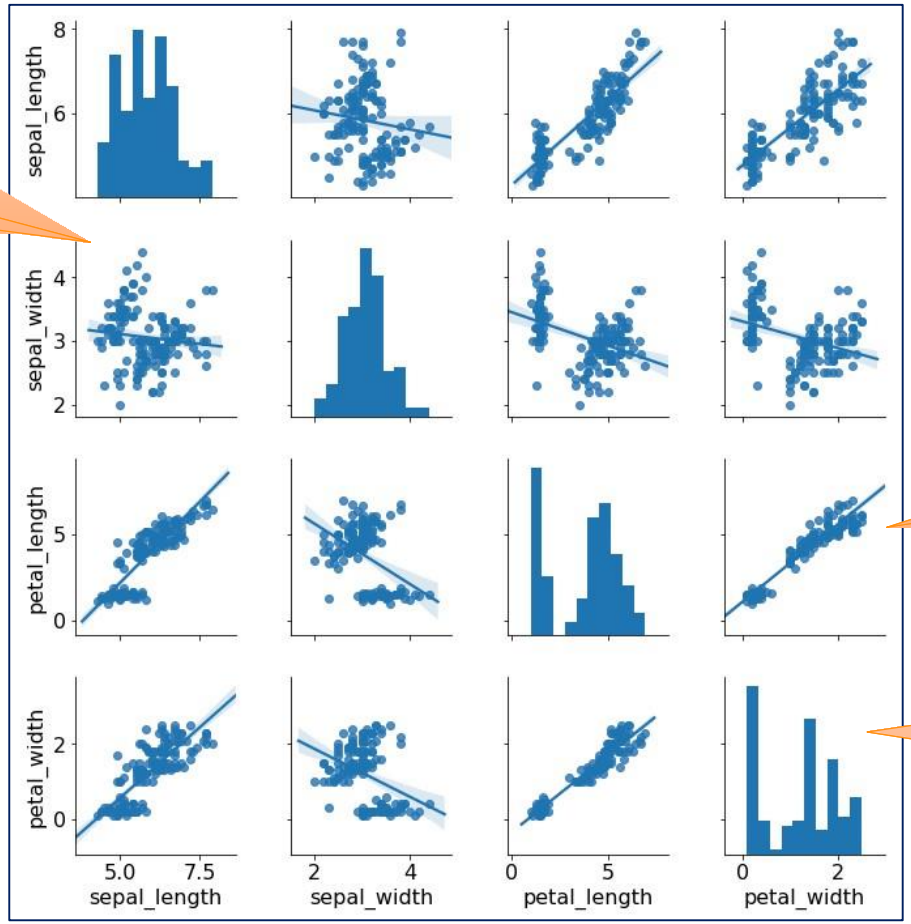
Mixtură a 2 repartiții normale după aplicarea logaritmului

Grafice ale setului de date „Boston house prices”
http://archive.ics.uci.edu/ml/datasets/Housing

Boxplots


Outlier
Percentila 25%
Mediana
Percentila 75%
Box plot of features
sepal_length
sepal_width
petal_length
petal_width
0
1
2
3
4
5
6
Box plot of feature petal_length by species
setosa
versicolor
virginica
species
1
2
3
4
5
6
7
petal_length
Domeniul datelor fără outlier-i
Definiția outlier-ilor se poate modifica. Aici a fost folosită: „la o distanță mai mare de 1.5 *IQR de percentila 25%/75%.”

Diagrame de împrăștiere bidimensionale (*Pairwise Scatterplots*) cu regresii



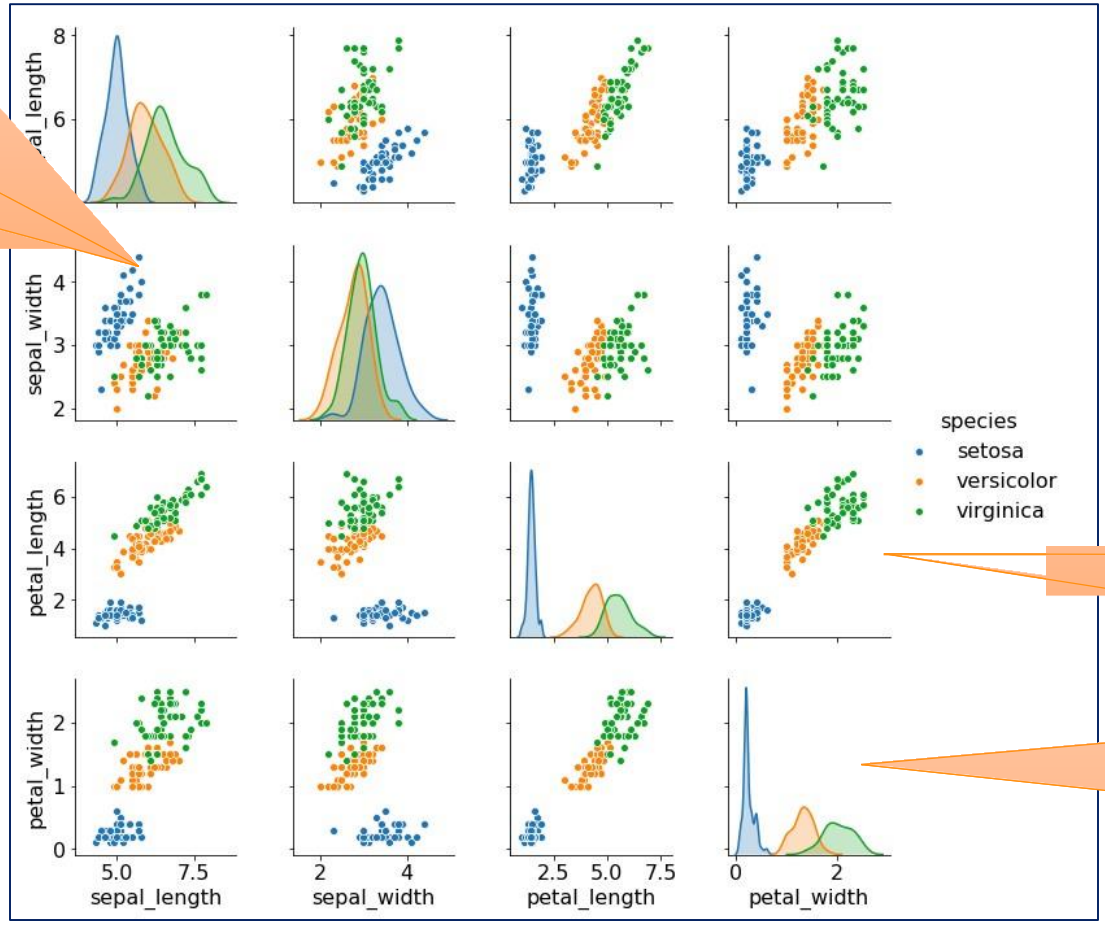
Nicio corelare vizibilă

Corelare liniară puternică

Histograma datelor din coloană

Diagrame bidimensionale (*Pairwise Plots*) cu clase

O bună separare a punctelor albastre, dar cele verzi și portocalii se suprapun



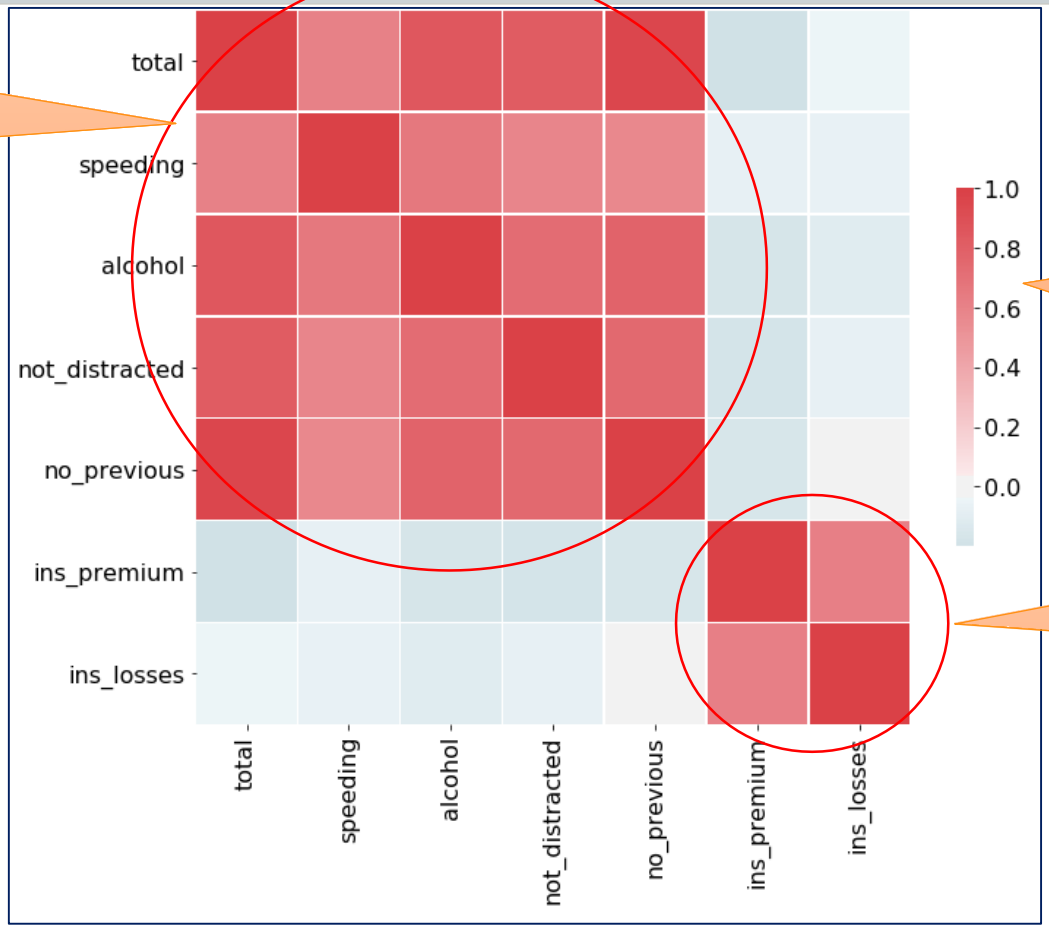
O bună separare a celor 3 clase

Grafice de densitate ale datelor din coloană, separate pe clase

Diagrama Heatmap de corelare

Există diferiți coeficienți de corelare. A fost folosit coeficientul Pearson, ce măsoară corelările liniare.

Corelare între motivele accidentelor

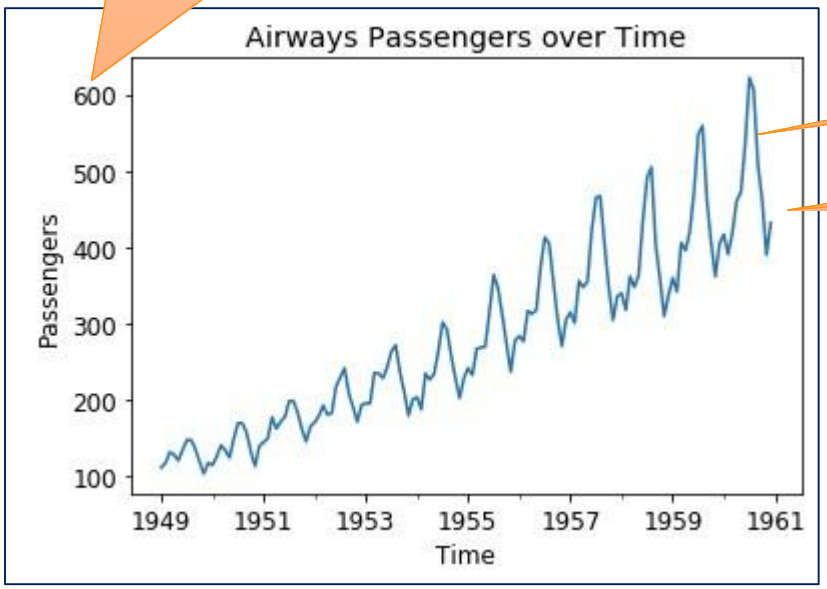


Culorile arată cât de puternică este corelarea

Corelare între prime și perderi

Grafice liniare (*line plots*) pentru serii de timp

Domeniul de valori



Pattern zgomot regulat → Periodic?

Trend liniar

Concluzii

- Important pentru înțelegerea datelor disponibile
- Statisticile de sumarizare oferă o bună perspectivă
  - Pot fi deceptive
- Vizualizarea reprezintă o modalitate puternică de înțelegere a datelor
- Înțelegerea metadatelor și modul în care experții din domeniu înțeleg datele sunt, de asemenea, importante

3. Analiza datelor

Teorema *No Free Lunch*

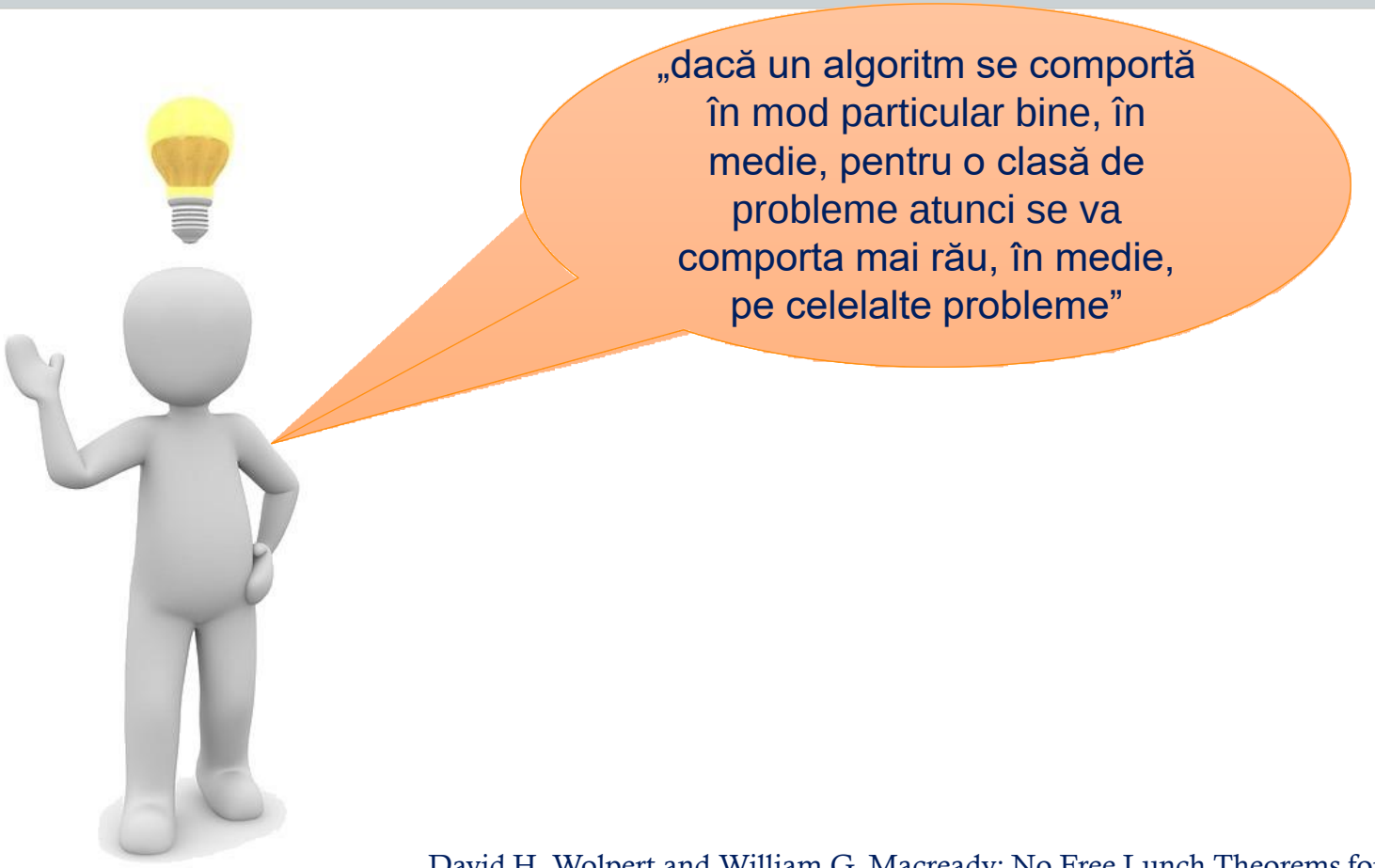
- d_m^y mulțimi ordonate de dimensiune m de valori de cost pentru $y \in Y$
- $f: X \rightarrow Y$ funcție de optimizat
- $P(d_m^y \mid f, m, a)$ probabilitatea condițională de a obține $d_m{}^y$ prin m execuții ale algoritmului a asupra funcției f

Teoremă: Pentru fiecare pereche de algoritmi a_1 și a_2:

$$\Sigma_f P(d_m^y \mid f, m, a_1) = \Sigma_f P(d_m^y \mid f, m, a_2)$$

Toți algoritmii sunt egali.

Implicații ale teoremei NFL



David H. Wolpert and William G. Macready: No Free Lunch Theorems for Optimization, IEEE Transactions on Evolutionary Computation, 1(1):67-82

No Silver Bullet

- Nu există un singur mod de a realiza analiza datelor
  - Există tehnici standard care, de cele mai multe ori, funcţionează bine

- Tehnicile potrivite sunt influenţate de mai mulţi factori:
  - Datele
  - Problema de rezolvat
  - Resursele disponibile
  - …

- Este necesar un portofoliu de tehnici de analiză a datelor

Categorii de tehnici de analiză a datelor

Categorie	Tehnici	Problema de rezolvat
Reguli de asociere	Apriori	Relaţii între item-uri
Clustering	K-Means Clustering DB Scan	Gruparea de item-uri similare Identificarea structurilor
Clasificare	K-nearest Neighbor Decision Trees Random Forests Logistic Regression Naive Bayes Support Vector Machines Neural Networks	Atribuirea de categorii obiectelor
Regresie	Linear Regression Ridge Lasso	Relaţii între ieşiri şi intrări
Analiza seriilor temporale	ARMA	Identificarea structurilor temporale Prognoza proceselor temporale
Text Mining	Bag-of-Words Stemming/Lemmatization TF-IDF	Analiza datelor text

Concepte fundamentale

Machine Learning

- Definiție (Tom M. Mitchel):
  - Spunem că un program de calculator învață din experiența *E* cu privire la o clasă de *task*-uri *T* și măsura de performanță *P*, dacă performanța task-urilor din *T*, măsurată de către *P*, se îmbunătățește prin experiența *E*.

- Legătura cu tehnicile de analiză a datelor:
  - Experiența *E*: datele noastre
  - Task-ul *T*: clustering/minarea asocierilor/clasificare/…
  - Măsura de performanță *P*: depinde de *task*-uri

T. M. Mitchel: Machine Learning, McGraw Hill, 1997

Descrierea unei fotografii a unei balene

- Cum putem descrie această fotografie folosind concepte generale?

Are o aripă dorsală

Background albastru

Corp oval

Partea de sus neagră, cea de jos albă

Caracteristicile obiectelor

Obiect

Funcție de mapare caracteristici
$\phi: O \rightarrow \mathcal{F}$

Caracteristici

hasFin	=	true
shape	=	oval
colorTop	=	black
colorBottom	=	white
background	=	blue

- O este spațiul obiect
- ϕ este funcția de mapare caracteristici
- $\mathcal{F}$ este spațiul caracteristicilor
  - $\mathcal{F} = \{\phi(o), o \in O\}$
- Exemplu:
  - Spațiu 5-dimensional cu dimeniunile de mai sus:
  - $\phi(\text{"whalepicture"}) = (true, oval, black, white, blue)$

Scalele caracteristicilor

- Nivelurile de măsurare ale lui Steven

Categorial

Scală	Proprietate	Operaţii permise	Exemplu
Nominal	Clasificare sau apartenenţă	=, ≠	Culori („black“, „white“, „blue“)
Ordinal	Comparare sau niveluri	=, ≠, >, <	Dimensiune („small“, „medium“, „large“)
Interval	Diferenţe sau afinităţi	=, ≠, >, <, +, −	Date, temperaturi, valori numerice discrete
Ratio	Magnitudini sau cantităţi	=, ≠, >, <, +, −, · , /	Dimensiunea în cm, durata în secunde, valori numerice continue

S. S. Stevens: On the Theory of Scales of Measurement, Science, 103(2684):677-680

Encodarea caracteristicilor categoriale

- Mulți algoritmi pot funcționa doar cu caracteristici numerice
- Caracteristicile categoriale pot fi encodate sub formă de caracteristici numerice binare
  - Exemplu: $x \in \{small, medium, large\}$
  - Encodăm sub forma a 3 variabile x^{small}, x^{medium}, x^{large}

$$x^{small} = \begin{cases} 1 & \text{dacă } x = small \\ 0 & \text{altfel} \end{cases}, \ldots$$

  - Se poate folosi cu o variabilă mai puțin, cazul rămas fiind encodat implicit din celelalte

- Acest tip de encodare se numește *One-Hot-Encoding*

Date de antrenare

- *Instanțe* de obiecte descrise prin intermediul caracteristicilor lor.

$\phi(o)$					
hasFin	shape	colorTop	colorBottom	background	**value of interest**
true	oval	black	black	blue	**whale**
false	rectangle	brown	brown	green	**bear**
…	…	…	…	…	**…**

- Învățare supervizată dacă valoarea de interes este cunoscută
  - Clasificare, regresie
- Altfel învățare nesupervizată
  - *Clustering*, Descoperirea (*mining*) regulilor de asociere

Date de test

- Datele pentru evaluarea rezultatelor analizei:
  - Aceeași distribuție ca datele de antrenare
- Datele de antrenare ≠ Datele de test
  - Evaluare generalizată
  - Evitarea *overfitting*-ului
    - Rezultatele analizei valide doar pe date de antrenare
    - Diferă și nu funcționează pe date neîntâlnite anterior
- Datele de test sunt adesea greu de obținut.

De unde obținem datele de test?

Date *Hold-out*

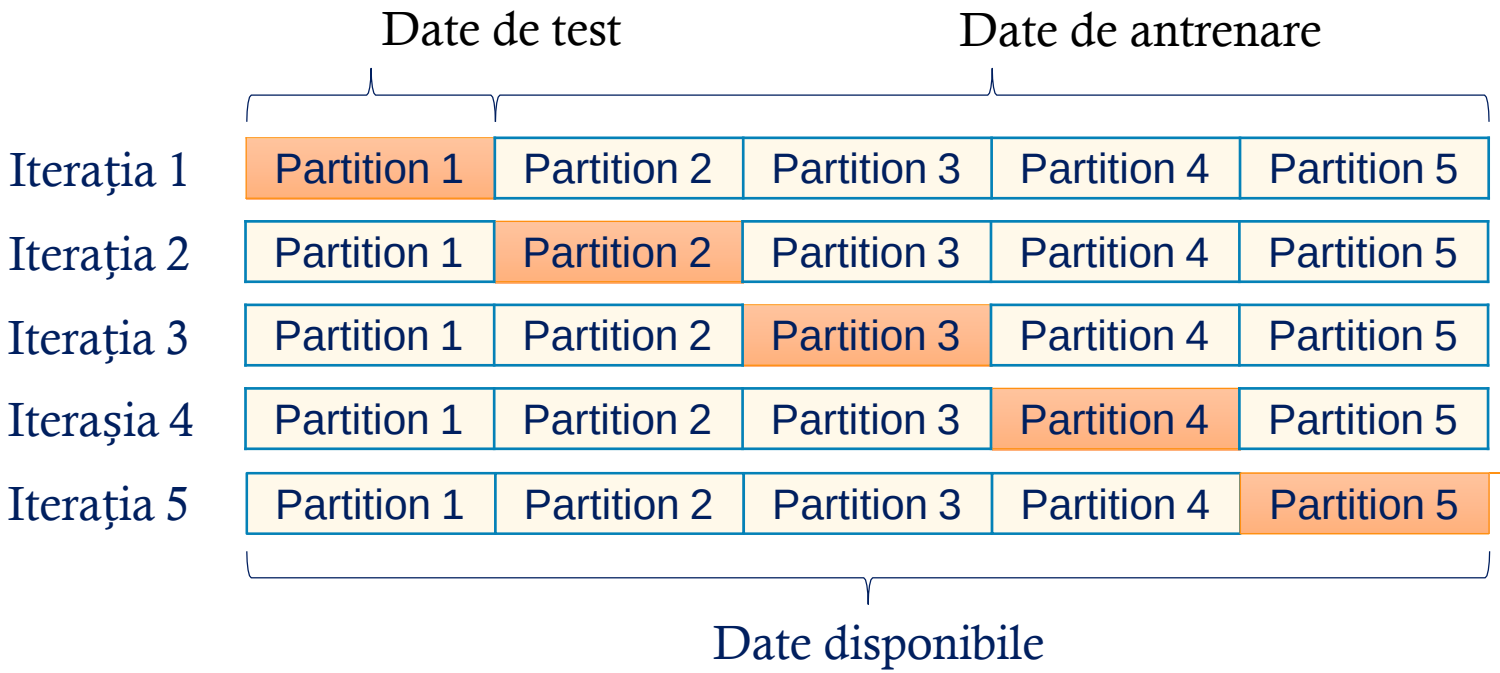
- Date ce nu sunt folosite deloc pentru antrenare
- Dimensiuni folosite frecvent pentru datele hold-out:
  - 50% din toate datele
  - 33% din toate datele
  - 25% din toate datele, în cazul în care este folosit și un set de validare

Depind de datele disponibile

- Exemplu:
  - Tranzacții ale clienților disponibile pentru 9 luni
  - Date de antrenare – primele 6 luni
  - Date de test – ultimele 3 luni

k-fold *Cross Validation*

- Se creează k partiții de date disponibile
- O partiție pentru testare, toate celelalte pentru antrenare
- Se estimează performanța combinând rundele unei iterații



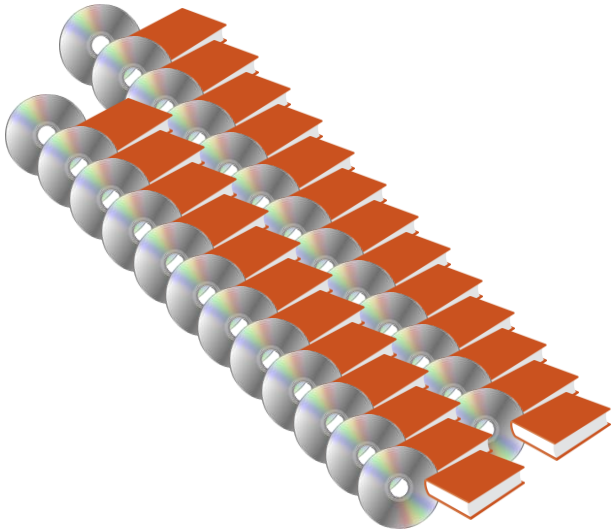
Concluzii

- Nu există un algoritm generic pentru toate problemele
- Obiectele sunt descrise prin caracteristici
- Caracteristicile sunt folosite în învățarea despre obiecte
- În general, datele se divizează în diferite mulțimi pentru diferite scopuri

4. Descoperirea regulilor de asociere

Exemplu de reguli de asociere

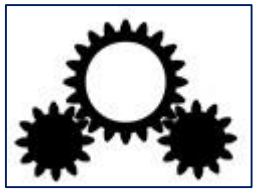
Item-uri deja în coș + Item-uri disponibile → Item probabil să fie adăugat



Problema Generală

Set de Tranzacții

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- …



Association Rule Mining

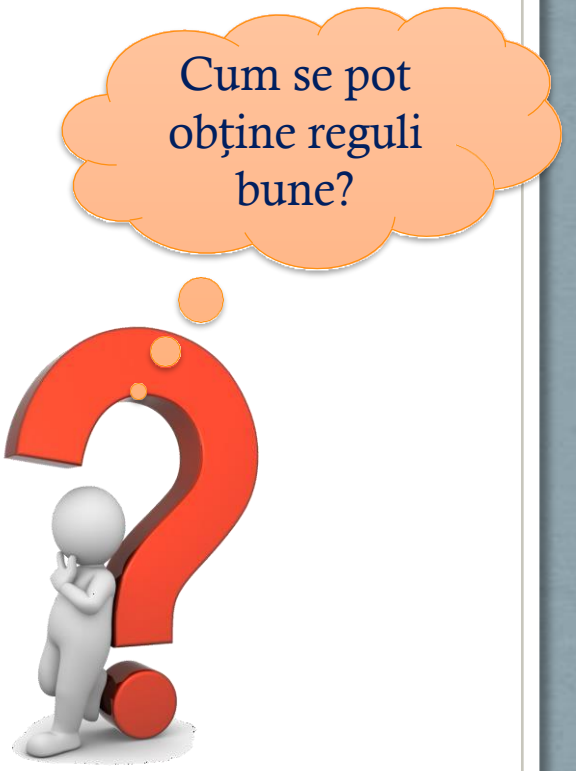
Reguli de asociere

- item2 → item3
- item2 → item4
- item2,item3 → item4
- …

Regulile descriu relații „interesante”

Problema formală

- Item-uri
  - $I = \{i_1, i_2, \dots, i_m\}$
- Tranzacții
  - $T = \{t_1, \dots, t_n\}$, unde $t_i \subseteq I$
- Reguli
  - $X \Rightarrow Y$ astfel încât $X, Y \subseteq I$ și $X \cap Y = \emptyset$
  - X se mai numește antecedent
  - Y se mai numește consecință



Ce înseamnă „relații interesante“?

Set de tranzacții

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- ...

item2, item3, item4 apar adeseori împreună

Interesant == adesea împreună

Algoritmul apriori

Suport și mulțimi de *item*-uri frecvente

- Suport
  - Procent de apariții al unui set de *item*-uri
  - $support(i) = \frac{|\{t \in T : i \subseteq t\}|}{|T|}$
- Mulțimi de item-uri frecvente (*Frequent item set*)
  - Mulțimi de *item*-uri ce apar împreună „suficient de des”
  - Se definesc cu ajutorul unui prag
  - $i \subseteq I$ este frecvent dacă $support(i) > minsupp$
- Regulile pot fi generate prin divizarea mulțimilor de *item*-uri:
  - $i = X \cup Y$

Example de generare a regulilor

$minsupp = 0.3$

- $\text{support}(\{item2, item3, item4\}) = \frac{3}{10} \geq 0.3$

Set de Tranzacții

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

- 8 reguli posibile:
  - $\emptyset \Rightarrow \{item2, item3, item4\}$
  - $\{item2\} \Rightarrow \{item3, item4\}$
  - $\{item3\} \Rightarrow \{item2, item4\}$
  - $\{item4\} \Rightarrow \{item2, item3\}$
  - $\{item2, item3\} \Rightarrow \{item4\}$
  - $\{item2, item4\} \Rightarrow \{item3\}$
  - $\{item3, item4\} \Rightarrow \{item2\}$
  - $\{item2, item3, item4\} \Rightarrow \emptyset$

Confidence, Lift, Leverage

- *Confidence*
  - Procentul de tranzacţii ce conţin antecedentul şi, de asemenea, consecinţa
  - $confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} = \frac{|\{t \in T : X \cup Y \subseteq T\}|}{|\{t \in T : X \subseteq T\}|}$
- *Lift*
  - Raportul probabilităţilor lui X şi Y împreună şi independent
  - $lift(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X) \cdot support(Y)}$
- *Leverage*
  - Diferenţa dintre probabilitatea lui X şi Y, considerate împreună şi independent
  - $leverage(X \Rightarrow Y) = support(X \cup Y) - support(x) \cdot support(Y)$

În general, *lift* favorizează itemset-urile cu suport mai mic, iar *leverage* pe cele cu suport mai mare

Exemplu: Confidența pentru reguli

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

- $confidence(\emptyset \Rightarrow \{item2, item3, item4\}) = \frac{0.3}{1} = 0.3$
- $confidence(\{item2\} \Rightarrow \{item3, item4\}) = \frac{0.3}{0.6} = 0.5$
- $confidence(\{item3\} \Rightarrow \{item2, item4\}) = \frac{0.3}{0.4} = 0.75$
- $confidence(\{item4\} \Rightarrow \{item2, item3\}) = \frac{0.3}{0.6} = 0.5$
- $confidence(\{item2, item3\} \Rightarrow \{item4\}) = \frac{0.3}{0.4} = 0.75$
- $confidence(\{item2, item4\} \Rightarrow \{item3\}) = \frac{0.3}{0.5} = 0.6$
- $confidence(\{item3, item4\} \Rightarrow \{item2\}) = \frac{0.3}{0.3} = 1$
- $confidence(\{item2, item3, item4\} \Rightarrow \emptyset) = \frac{0.3}{0.3} = 1$

Exemplu: *Lift* pentru reguli

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

- $lift(\emptyset \Rightarrow \{item2, item3, item4\}) = \frac{0.3}{1 \cdot 0.3} = 1$
- $lift(\{item2\} \Rightarrow \{item3, item4\}) = \frac{0.3}{0.6 \cdot 0.3} = 1.66$
- $lift(\{item3\} \Rightarrow \{item2, item4\}) = \frac{0.3}{0.4 \cdot 0.5} = 1.5$
- $lift(\{item4\} \Rightarrow \{item2, item3\}) = \frac{0.3}{0.6 \cdot 0.4} = 1.25$
- $lift(\{item2, item3\} \Rightarrow \{item4\}) = \frac{0.3}{0.4 \cdot 0.6} = 1.25$
- $lift(\{item2, item4\} \Rightarrow \{item3\}) = \frac{0.3}{0.5 \cdot 0.4} = 1.5$
- $lift(\{item3, item4\} \Rightarrow \{item2\}) = \frac{0.3}{0.3 \cdot 0.6} = 1.66$
- $lift(\{item2, item3, item4\} \Rightarrow \emptyset) = \frac{0.3}{0.3 \cdot 1} = 1$

Exemplu: Analiza tuturor valorilor

Regulă	*Confidence*	*Lift*	*Leverage*
$\emptyset \Rightarrow \{item2, item3, item4\}$	0.30	1.00	0.00
$\{item2\} \Rightarrow \{item3, item4\}$	0.50	1.66	0.12
$\{item3\} \Rightarrow \{item2, item4\}$	0.75	1.50	0.10
$\{item4\} \Rightarrow \{item2, item3\}$	0.50	1.25	0.06
$\{item2, item3\} \Rightarrow \{item4\}$	0.75	1.25	0.06
$\{item2, item4\} \Rightarrow \{item3\}$	0.60	1.50	0.10
$\{item3, item4\} \Rightarrow \{item2\}$	1.00	1.66	0.12
$\{item2, item3, item4\} \Rightarrow \emptyset$	1.00	1.00	0.00

Confidență perfectă, dar niciun câștig față de cazul aleator

Confidență perfectă și de 1.66 ori mai probabil decât cazul aleator

Itemsets și reguli = Exponențial

- Numărul de *itemset*-uri este exponențial
  - Toate *itemset*-urile posibile constituie mulțimea putere $\mathcal{P}$ a lui I
    - $|\mathcal{P}(I)| = 2^{|I|}$
  - Rămâne exponențial dacă restricționăm dimensiunea:
    - $|I|$ *itemset*-uri cu $k = 1$ *item*-uri
    - $\frac{|I| \cdot (|I| - 1)}{2}$ *itemset*-uri cu $k = 2$ *item*-uri
    - $\binom{|I|}{k} = \frac{|I|!}{(|I|-k)!k!}$ *itemset*-uri cu k *item*-uri
- Numărul de reguli per *itemset* este exponențial
  - Antecedenții posibili ai *itemset*-ului i sunt mulțimea putere $\mathcal{P}$ a lui i
    - $|P(i)| = 2^{|i|}$
- Exemplu: $|I| = 100, k = 3$
  - 161,700 *itemset*-uri posibile
    1,293,600 reguli posibile

Reducerea spațiului de căutare

- Proprietatea Apriori
  - **Toate submulțimile unui *itemset* frecvent sunt, de asemenea, frecvente**
  - $support(i') \geq support(i)$ pentru toți $i' \subseteq i, i \in I$
- Putem „crește“ *itemset*-urile și reduce spațiul de căutare aplicând proprietatea apriori
  - Începem cu *itemset*-urile de dimensiune $k = 1$
  - Eliminăm toate *itemset*-urile care nu au suport minimal
  - Construim toate combinațiile de dimensiune $k + 1$
  - Repetăm până când
    - Nu mai sunt găsite *itemset*-uri cu suport minimal
    - Este atins un prag pentru k
- Complexitatea este tot exponențială, dar mai limitată

Exemplu pentru creșterea *itemset*-urilor (k=1)

$minsupp = 0.3$

Set de tranzacții

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

Itemset	Suport	
$\{item1\}$	0.2	← Drop
$\{item2\}$	0.6	
$\{item3\}$	0.4	
$\{item4\}$	0.5	
$\{item5\}$	0.3	
$\{item6\}$	0.2	← Drop
$\{item7\}$	0.3	
$\{item8\}$	0.1	← Drop

Exemplu pentru creșterea *itemset*-urilor (k=2)

$minsupp = 0.3$

Set de tranzacții

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

Itemset	Suport	
$\{item2, item3\}$	0.4	
$\{item2, item4\}$	0.5	
$\{item2, item5\}$	0.1	← Drop
$\{item2, item7\}$	0.1	← Drop
$\{item3, item4\}$	0.3	
$\{item3, item5\}$	0.0	← Drop
$\{item3, item7\}$	0.1	← Drop
...	…	← Drop

Exemplu pentru creșterea *itemset*-urilor (k=3)

$minsupp = 0.3$

Set de tranzacții

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

Itemset	Suport
$\{item2, item3, item4\}$	0.3

Singurul *itemset* rămas, creșterea se încheie.

- S-au găsit următoarele *itemset*-uri frecvente cu cel puțin 2 *item*-uri:
  - $\{item2, item3\}, \{item2, item4\}, \{item3, item4\}$
  - $\{item2, item3, item4\}$

Candidați pentru reguli

- În general, nu sunt considerate toate regulile posibile
- Două restricții comune:
  - Fără antecedenți și consecințe vide
    - $X \neq \emptyset$ and $Y \neq \emptyset$
  - Un singur *item* drept consecință
    - $|Y| = 1$
- Exemplu:

~~$\emptyset \Rightarrow \{item2, item3, item4\}$~~
~~$\{item2\} \Rightarrow \{item3, item4\}$~~
~~$\{item3\} \Rightarrow \{item2, item4\}$~~
~~$\{item4\} \Rightarrow \{item2, item3\}$~~
$\{item2, item3\} \Rightarrow \{item4\}$
$\{item2, item4\} \Rightarrow \{item3\}$
$\{item3, item4\} \Rightarrow \{item2\}$
~~$\{item2, item3, item4\} \Rightarrow \emptyset$~~

Evaluarea regulilor de asociere

- Se folosesc diferite criterii, nu doar suport și confidență
  - *Lift* și *leverage* pot spune dacă regulile sunt doar coincidențe
- Validare dacă regulile sunt respectate de datele de test
  - Se verifică dacă regulile sunt găsite și pe datele de test
- Pentru predicția coșului de cumpărături, se folosesc itemset-uri incomplete pe date de test
  - Exemplu: se șterge *item4* din toate *itemset*-urile și se observă dacă regulile fac o predicție corectă asupra locului în care este asociat
- Inspectarea manuală a regulilor
  - Au sens?
  - Se pot consulta experții din domeniu

Concluzii

- Asocierile sunt relații interesante între *item*-uri
- „Interesant” înseamnă, de obicei, că item-urile apar împreună, dincolo de a fi o coincidență
- Numărul de *itemset*-uri / reguli posibile este exponențial
  - Se aplică proprietatea apriori pentru limitarea itemset-urilor
  - Se impun restricții pe structurile regulilor
- Date de test și inspectare manuală pentru validare