

Tehnici de explorare a datelor (Data mining)

Curs 1

De ce data mining?

Explozia de date

- Sistemul de urmarire orbitala de la NASA:
46 MB/sec

4.000.000 MB/zi

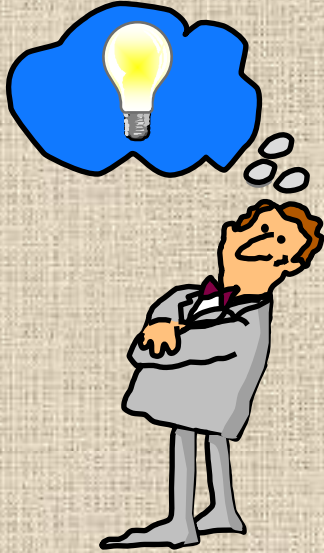
- biblioteca de imagini cu amprente a FBI :

200.000 GB bytes

- imagistica medicala

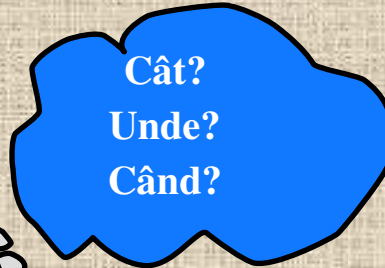
baze de date de ordinul 10TB

Verificare

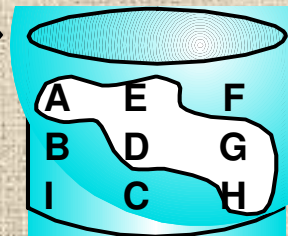
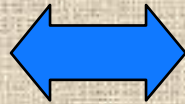


Soluție nouă pt o problemă

Utilizatorul explorează



Interogări

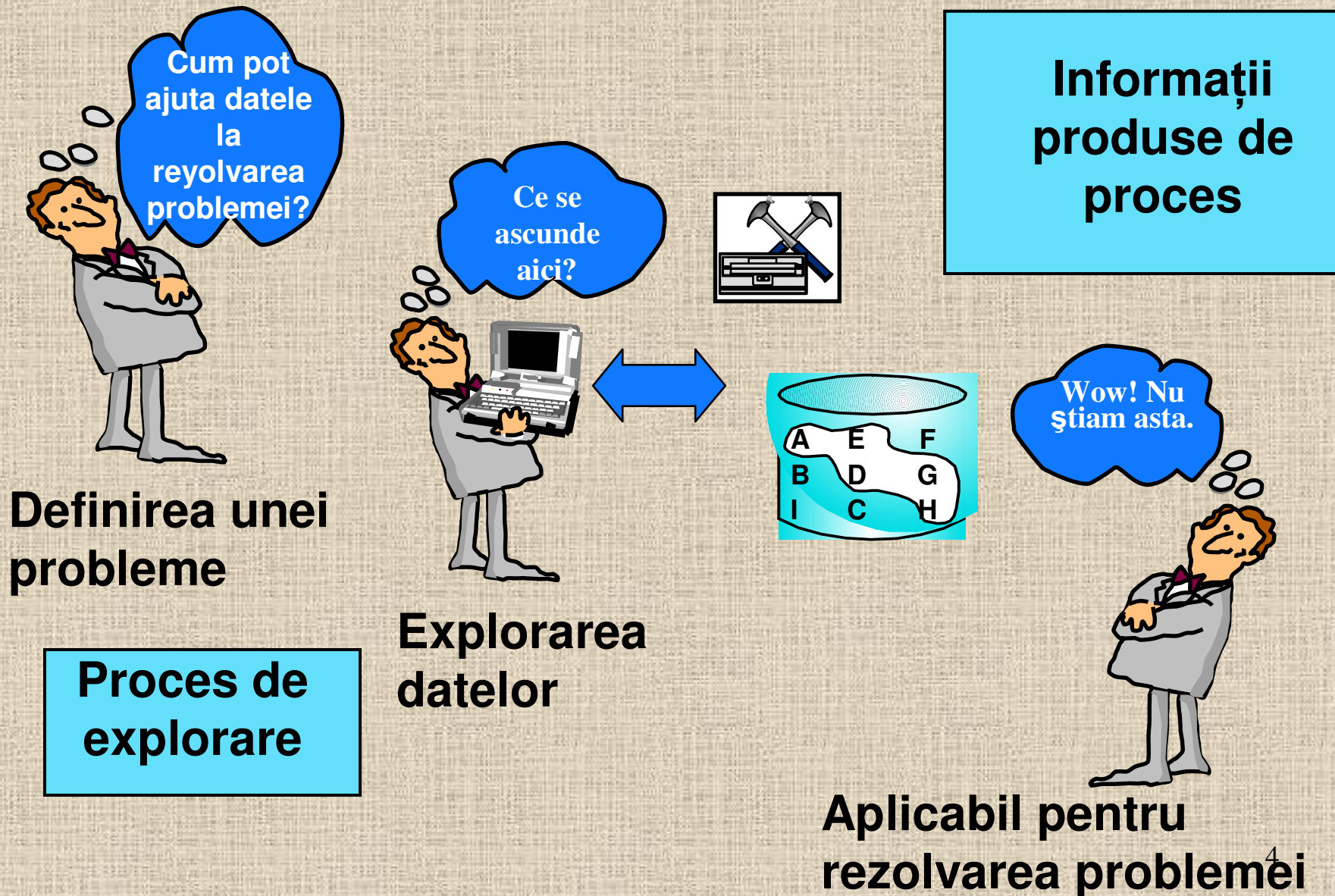


Informații determinate de utilizator



Validarea Soluției

Descoperirea



DE CE :

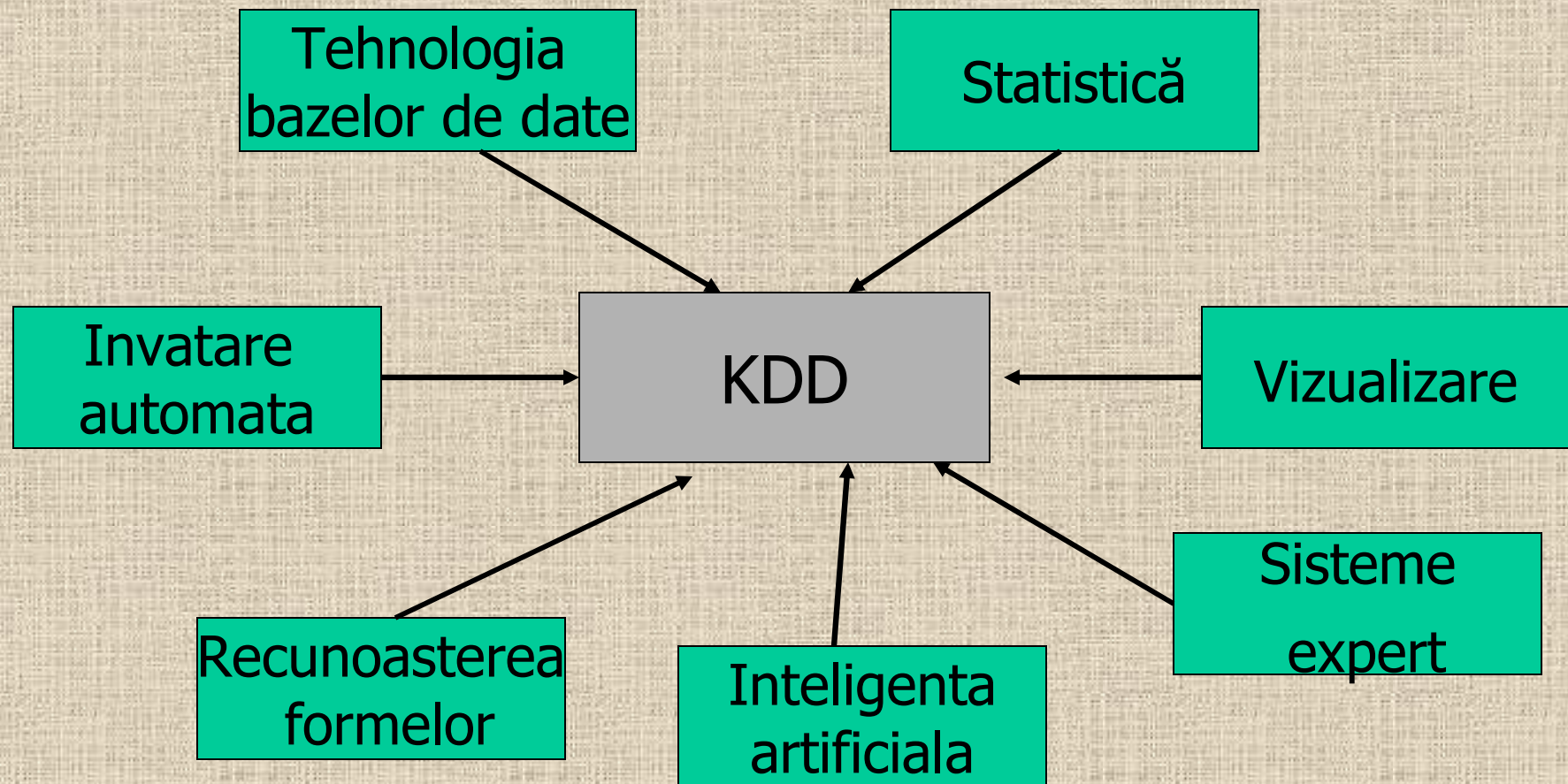
“Necesitatea este Mama Inventiei”

- Problemele generate de explozia de date
 - Căutarea în volume mari de date a tiparelor care prezintă interes dintr-un anumit punct de vedere, este o **necesitate** în condițiile în care în ultima perioadă de timp a avut loc o creștere exponențială a volumului de date stocate în baze de date, depozite de date sau alte depozite de informații, în paralel cu dezvoltarea capacităților și performanțelor sistemelor de calcul.
 - Mai mult decât atât, apariția a noi tipuri complexe necesită tehnici speciale de manipulare
- Soluția: Data warehousing și data mining
- Ce reprezintă explorarea datelor în contextul lumii reale?
 - Explorarea datelor (data mining) este procesul de *analiză a datelor brute* din bazele de date și de *sintetizare a informațiilor* utile în procesul luării deciziilor. Scopul este acela ca prin utilizarea informațiilor existente să fie obținute *noi fapte* și să fie descoperite *noi legături* anterior necunoscute chiar și pentru experții complet familiari cu datele respective.

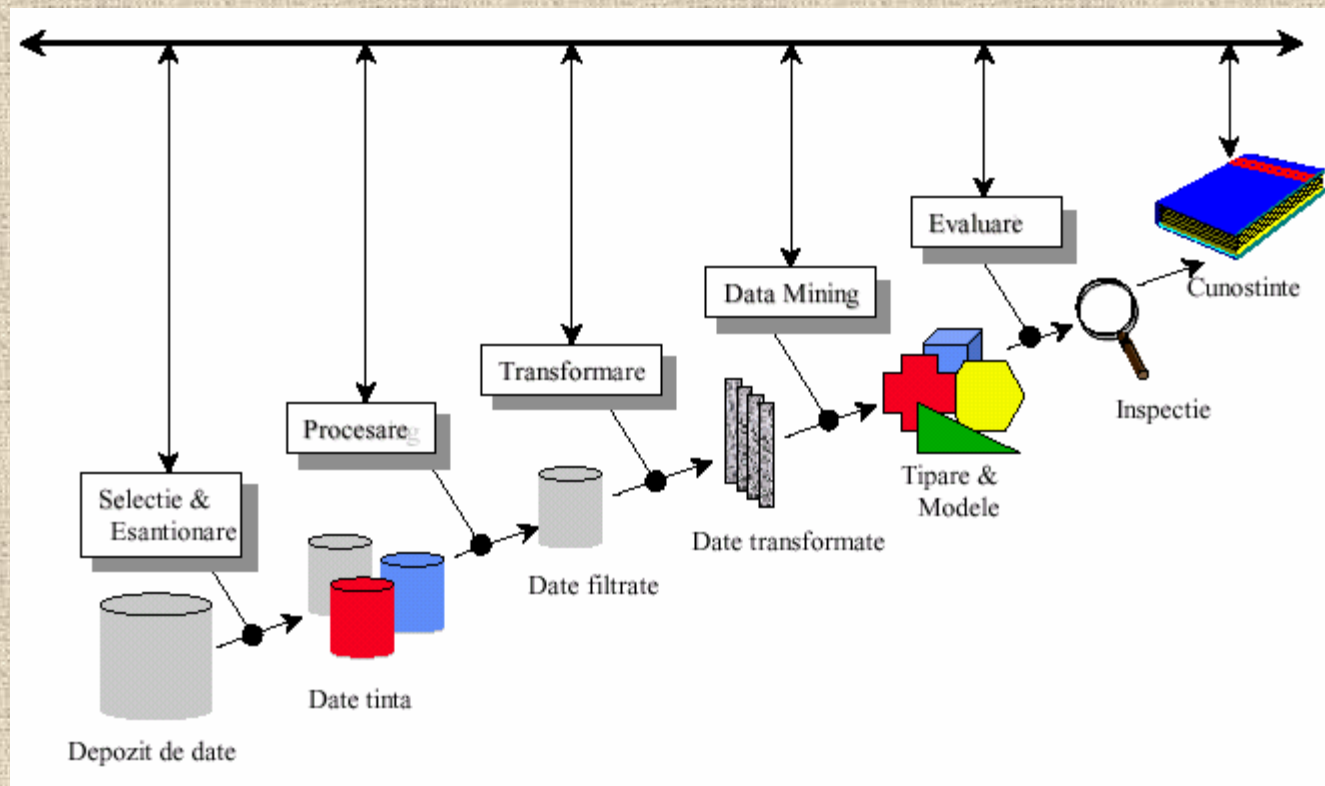
Descoperire a cunoștințelor din bazele de date (KDD) si explorarea datelor

- ***KDD - procesul netrivial de identificare a tiparelor din date, tipare valide, noi, potențial utile și inteligibile.***
 - *datele* - o mulțime de evenimente
 - *tiparul* - o expresie, într-un anumit limbaj care descrie un subset al datelor sau un model aplicabil acestui subset.
 - *proces* - descoperirea cunoștințelor din date este o succesiune de pași care implică iterații multiple pentru următoarele faze:
 - pregătirea datelor,
 - căutarea tiparelor,
 - evaluarea cunoștințelor și rafinarea, toate repetate în.
 - Tiparele descoperite trebuie să fie:
 - *valide* pe seturi de date noi, cu un anumit grad de certitudine,
 - *noi și potențial folositoare* în sensul că trebuie să conducă la anumite beneficii pentru utilizator
 - *inteligibile*, dacă nu imediat, cel puțin după anumite operații de post-procesare.
 - necesitatea definirii unor măsuri cantitative pentru evaluarea tiparelor extrase:
 - măsuri pentru certitudine (acuratețea predicției estimată pe date noi)
 - masuri pentru utilitate (câștigul, evaluat ca profit obținut de pe urma predicției mai bune sau a creșterii vitezei de răspuns a sistemului).
 - Noțiunile de noutate și inteligibilitate sunt mult mai subiective.
 - *Interesul* - este o măsură globală a valorii unui tipar care combină validitatea, noutatea, utilitatea și simplitatea.
 - Funcția de interes poate fi definită explicit sau se poate manifesta implicit prin intermediul unei ierarhii a tiparelor detectate sau a modelelor stabilită de sistemul de descoperire a cunoștințelor din date.

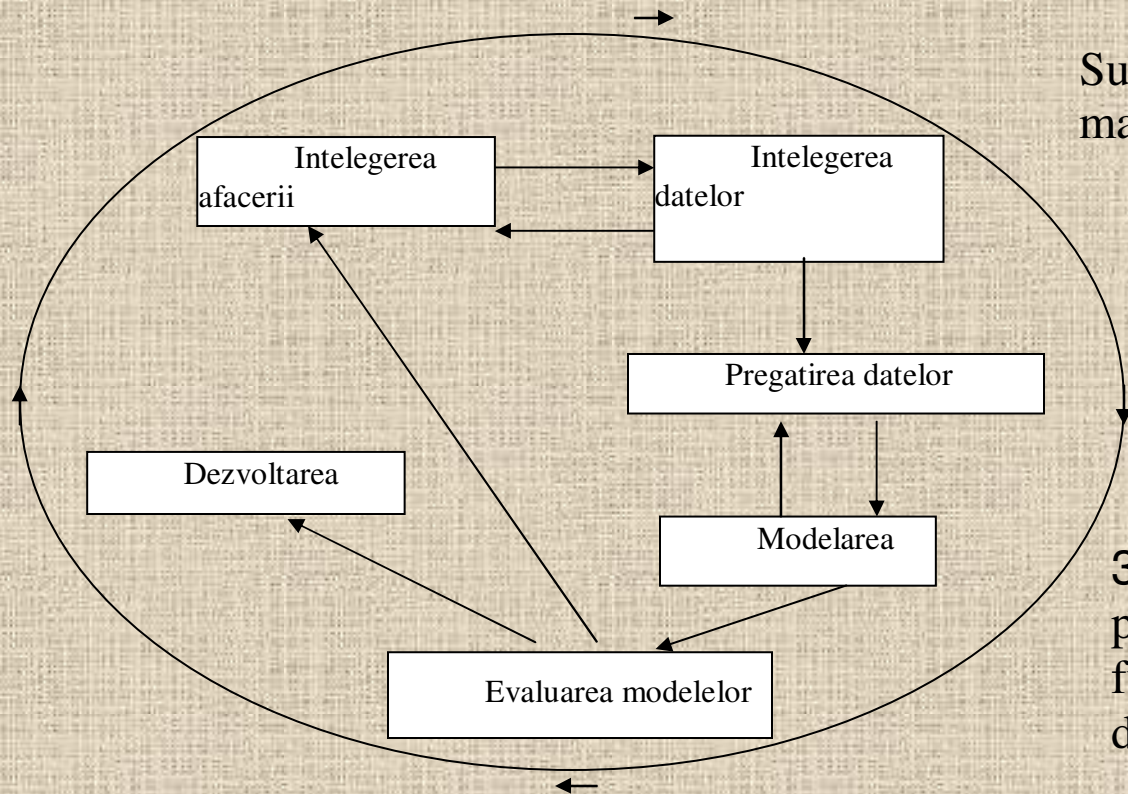
KDD: Confluența mai multor discipline



Descoperirea cunoștințelor din bazele de date (KDD) si explorarea datelor I Modelul academic (Fayyad)



Descoperirea cunoștințelor din bazele de date (KDD) și explorarea datelor II Modelul industrial (CRISP-DM)



Succesiune de pași, parcurși într-o manieră interactivă și iterativă

1. analiza scopurilor declarate de utilizatorul final și primirea tuturor cunoștințelor anterioare necesare.
2. datele țintă sunt pregătite și curățate de tot ceea ce înseamnă zgomote sau valori izolate.

3. se găsește caracteristica utilă pentru reprezentarea datelor, funcție de obiectivul sarcinii de descoperire.

4. se alege și se aplică un anumit algoritm de explorare a datelor în scopul de a prezice valorile viitoare ale variabilelor de interes sau de a găsi tiparele din date, interpretabile de factorul uman..

5. tiparele sunt interpretate și evaluate cu ajutorul unor instrumente specializate, cum ar fi cele de vizualizare.

Pregatirea datelor brute pentru explorare

- Datele reale contin erori
 - Incomplete: lipsesc valorii ale unor attribute, lipsesc attribute care pot fi de interes, pot contine doar valori agregate
 - Contin zgomote: contin erori sau date in afara domeniului
 - Inconsistente: contin discrepante in coduri sau in nume
- Sarcini majore in pregatirea datelor
 - Curatare - completarea valorilor lipsa, identificarea sau eliminarea datelor care nu sunt in gama admisa, rezolvarea inconsistentelor
 - Integrare -
 - Transformare - normalizare si agregare
 - Reducerea datelor - prin filtrare sau esantionare
 - Discretizarea datelor continui-

Explorarea datelor

- *Explorarea datelor (data mining) - aplicarea analizelor de date și descoperirea algoritmilor care, în limite acceptabile ale eficienței de calcul produc o enumerare particulară a tiparelor din date.*
- *Obiective:*
 - *verificarea, caz în care sistemul este folosit pentru a verifica ipotezele utilizatorului și*
 - *descoperirea atunci când sistemul găsește, în mod autonom noi tipare.*
 - *predicția - sistemul găsește tipare în scopul predicției comportamentului viitor pentru anumite entități*
 - *descrierea - este situația în care sistemul găsește tipare în scopul prezentării acestora unui utilizator, într-o formă inteligibilă*

Cerințe:

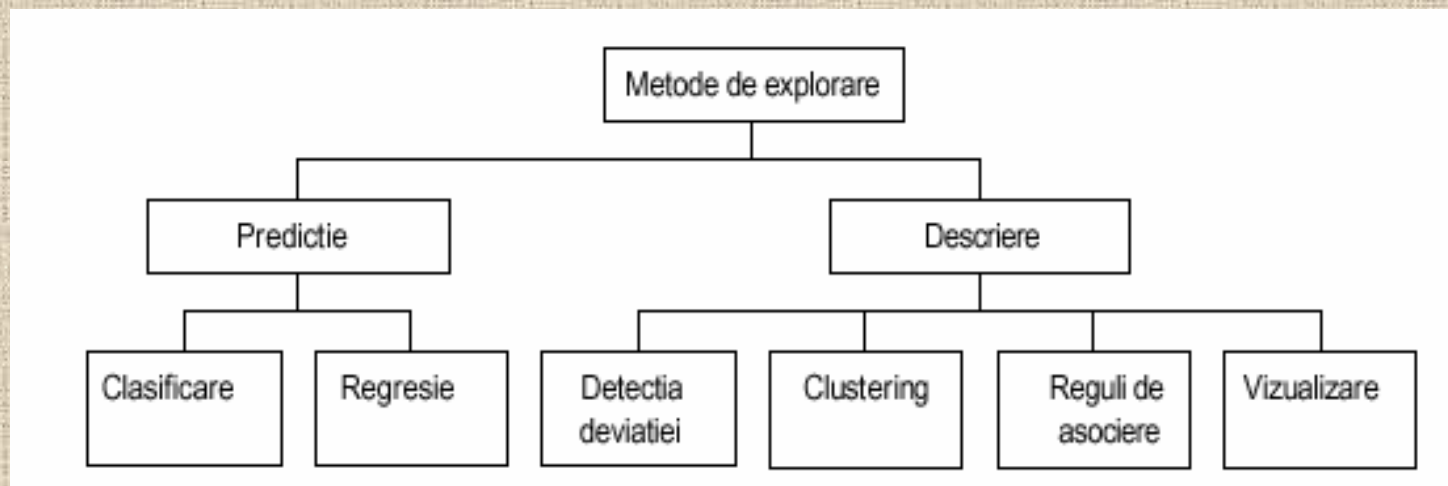
1. *necesitatea manipulării diferitelor tipuri de date*
2. *eficiență și scalabilitate pentru algoritmii de explorare a datelor*
3. *utilitate, încredere și expresivitate a rezultatelor obținute*
4. *expresii de tipuri variate pentru rezultatele explorării datelor*
5. *extragerea interactivă a cunoștințelor, pe niveluri multiple de abstractizare.*
6. *extragerea informațiilor din diferite surse de date*
7. *protecția intimității și asigurarea securității datelor*

Sarcinile explorarii datelor

- **previziunea / predicția** - realizarea unui model din exemplele analizate și utilizarea modelului dezvoltat pentru a prezice valorile viitoare ale variabilei țintă ;
- **clasificarea** – găsirea funcțiilor care grupează înregistrările într-una sau mai multe clase discrete. Prin această tehnică se alocă înregistrări noi claselor existente.
- **analiza legăturilor** - dezvoltarea regulilor de asociere între seturi de articole;
- **modelarea explicită** - găsirea formulelor explicite care descriu dependențele dintre diferite variabile;
- **clustering-ul** - gruparea articolelor în subseturi similare din punct de vedere statistic. Un cluster este definit ca un subset de date. Sarcina procesului de clustering este aceea de a diviza o bază de date în clustere de înregistrări similare.
- **detectarea deviațiilor** - determinarea schimbărilor semnificative a valorilor esențiale, obținute în urma măsurărilor, față de valorile anterioare sau față de cele așteptate.

Metode și tehnici generale de explorare a datelor

- **Clasificarea** : găsirea unei funcții care include un articol de date într-una din mai multe clase predefinite.
- **Regresia**: este utilizată la prezicerea unei valori a unei variabile continue bazată pe valorile altor variabile, presupunând un model de dependență liniar sau neliniar.
- **Gruparea** (clustering-ul): identifică o mulțime finită de categorii sau clustere pentru a descrie datele.
- **Rezumarea**: găsește o descriere compactă pentru o submulțime de date.
- **Modelarea dependențelor**: găsește un model care descrie dependențele semnificative dintre variabile.
- **Detectarea schimbărilor și a deviației** : descoperă cele mai semnificative schimbări produse în date în intervalul dintre două măsurări



Tipuri de probleme care se rezolvă prin data mining

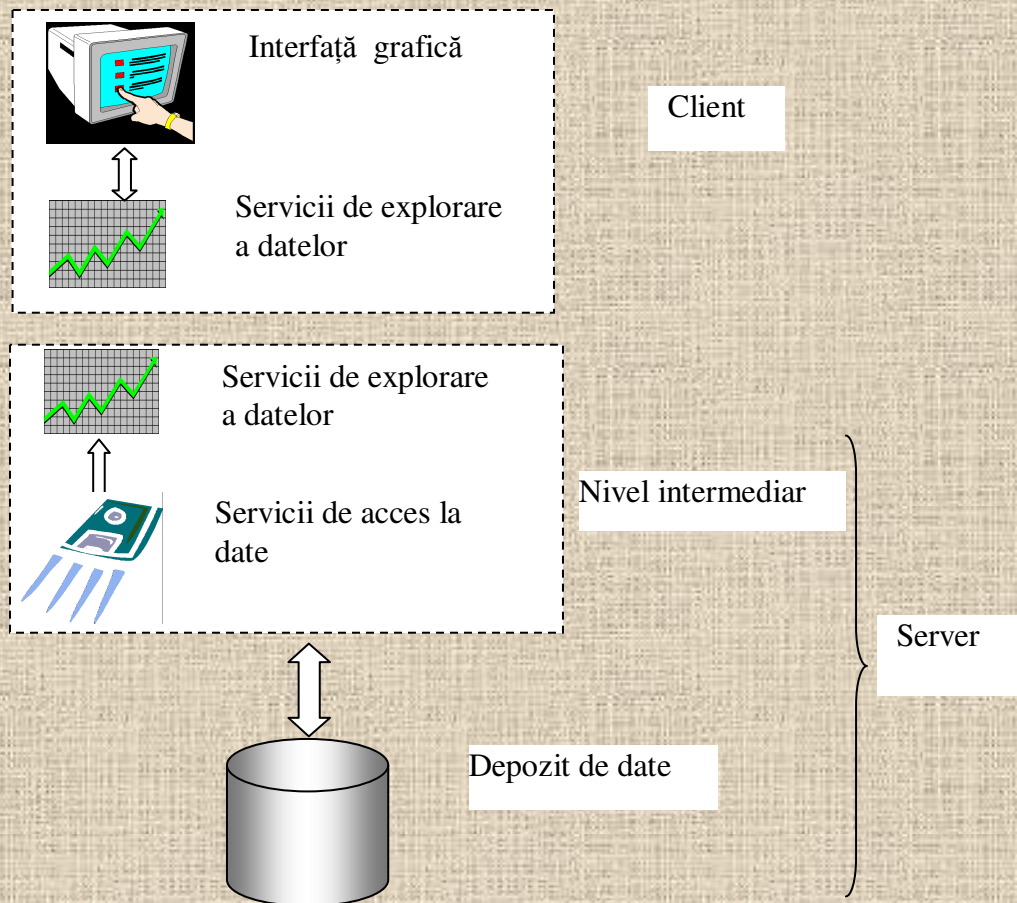
- 1. Ce le place clienților mei?
- **Clustering**
- 2. Ce clienți trebuie să țintesc într-o promoție?
- **Clustering**
- 3. Ce produse ar trebui să folosesc în promoție?
- **Asocieri sau tipare secvențiale**
- 4. Cum ar trebui să îmi plasez noile magazine?
- **Clustering și asocieri**
- 5. Cum pot detecta potențialele fraude?
- **Clustering plus asocieri**

Aplicatii pentru Data Mining

- Gestiunea relatiilor cu clientii - Customer Relationship Management (CRM)
 - pastrarea clientilor
- Analiza pietii
 - Gasirea pietelor tinta
 - Segmentarea pietei
 - Vanzari incrucisate
- Detectia fraudelor
 - Detectia fraudelor in domeniul sanatatii
 - Detectia fraudelor in cazul cartilor de credit
 - Detectia fraudelor in telecomunicatii
- Altele.....

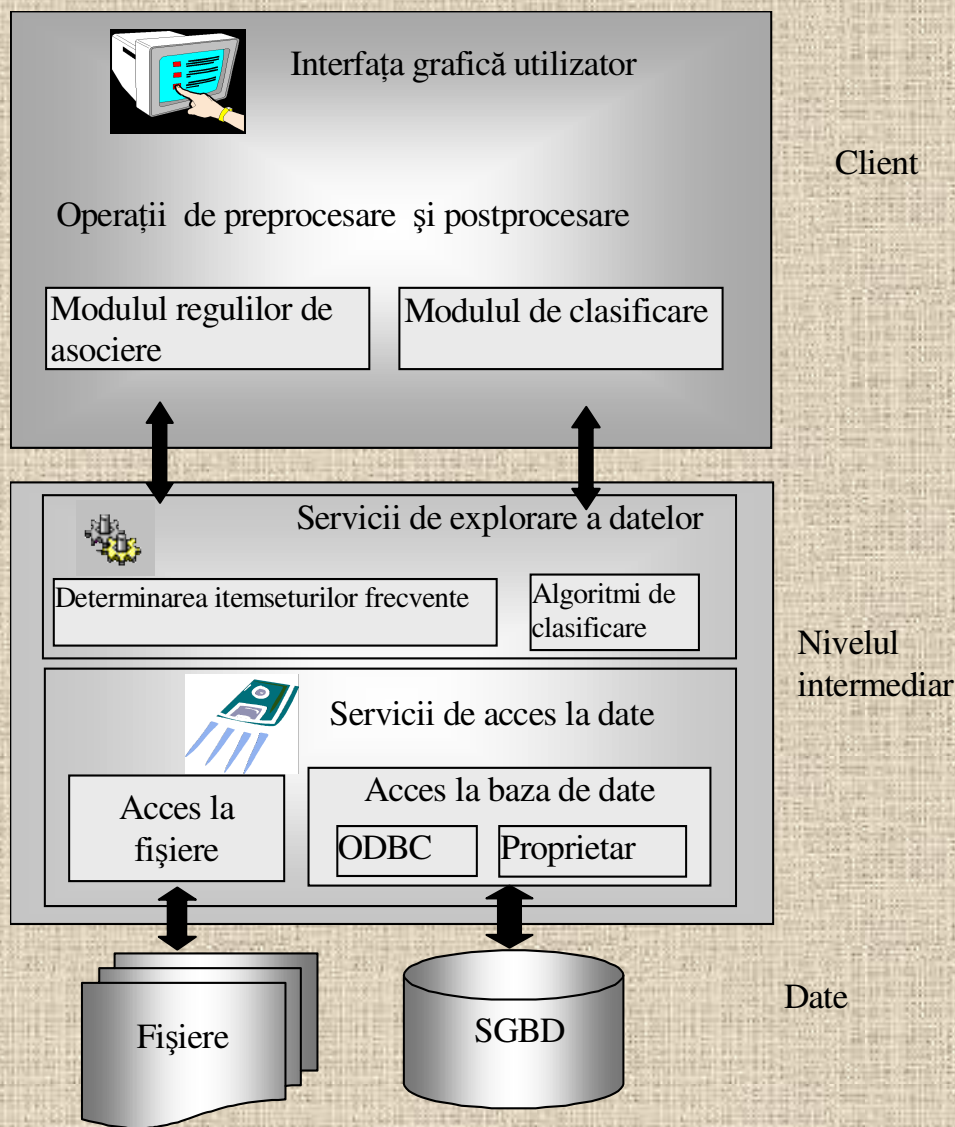
Arhitecturi de sisteme de explorare a datelor

- **caracteristici:**
 - **nu trebuie să se limiteze dimensiunea seturilor de date**
 - **se face optimizarea performanțelor pentru seturi mari de date**
 - **este necesară flexibilitatea față de diferitele tehnici de explorare a datelor**
 - **trebuie să se asigure suport pentru concurență și acces multi-user**
 - **trebuie realizat un control total al resurselor sistemului**
 - **este necesar un control total al accesului asupra datelor**
 - **să asigure administrarea și mentenanță de la distanță**
- **Componentele de bază ale unui sistem de explorare a datelor sunt:** *interfața utilizator, serviciile specifice de data mining, serviciile de acces la date și datele înseși*



Arhitectura pe trei niveluri

Model de sistem de explorare a datelor în arhitectură pe trei nivele



Servicii de conectare și acces: permit clienților să se conecteze la nivelul intermediar.

- este implicat un proces de „ascultare” care pornește servere de conexiuni dedicate. Fiecare server de conexiune tratează comunicarea cu un anumit client.

Servicii de administrare de la distanță: sunt accesate de către un client de administrare de la distanță.

- permite ca softul nivelului intermediar să fie configurat și controlat de la un calculator care este fizic separat de nivelul intermediar.

Serviciul de administrare a lucrărilor: este responsabil pentru execuția activităților de explorare da date.

Explorarea pentru reguli de asociere implică două activități.

- găsește toate seturile de articole frecvente ale căror suport este mai mare decât un prag minim de suport dat.

- generează regulile dorite din seturile de articole frecvente găsite în primul pas.

Explorarea pentru clasificare poate fi realizată folosind arbori de decizie sau metode asociative